Reinforcement Learning based Pricing for Demand Response

Amir Ghasemkhani and Lei Yang
Computer Science and Engineering Department, University of Nevada Reno
Email: aghasemkhani@nevada.unr.edu, leiy@unr.edu

***Abstract*—Smart pricing based demand response (DR) programs can enable the system to shape load profiles to improve system reliability and performance. Existing works on pricing based DR often assume that users' response functions are available or predictable at the load serving entity (LSE) side. Due to divergent consumption habituates of the users, it is challenging to have an accurate estimate of users' responses. Clearly, any mischaracterization of users' responses would result in higher system costs. To tackle this challenge, we leverage reinforcement learning to learn users' response functions. Specifically, we formulate the DR problem as a stochastic optimization problem, in which random responses are considered due to users' volatile behaviors. We develop a reinforcement learning based algorithm to solve a pricing strategy for DR without assuming any specific forms of users' response functions. The proposed reinforcement learning algorithm is shown to converge to an equilibrium with near optimal performance, which is corroborated via numerical simulations.**

***Index Terms*–Smart Pricing; demand response; reinforcement learning; load serving entity (LSE)**

I. Introduction

A. Motivation

Traditional power grids are being transformed into smart grids using advanced information control and communication technologies to offer higher reliability, security and efficiency in power systems [1–4]. To address the grand challenge of a sustainable energy future, there has recently been a surge of interest in renewable energy resources, including wind and solar. However, the integration of wind/solar energy puts forth great operational challenges due to their high variability, made even more severe by their non-dispatchability [5–8]. As a vital component of smart grid, demand response (DR) plays a key role in reducing the peak load and incorporating renewables into the grid [9–11]. In the National Assessment of Demand Response Potential report [12], DR is identified as the potential approach that can reduce up to 20% of the total peak electricity demand across the country, and significantly ease the adoption of renewable energy into the grid.

Existing demand response programs can be generally categorized into direct load control [13–15] and smart pricing [16–18]. By using direct load control, the load serving entity (LSE) can control the operations and energy consumption based on an agreement between the LSE and the customers (users) which, however, raises serious privacy concerns at the user's side as discussed in [19]. By using smart pricing, users are encouraged to individually and voluntarily manage their loads. For example, critical-peak pricing (CPP), time-of-use pricing (ToUP), and real-time pricing (RTP) have been proposed to coordinate demand responses to the improvement of system reliability and performance. These pricing schemes allow dynamic adjustment of electrical elastic loads to adapt their consumption levels according to energy generation costs. It is worth noting that these pricing based DR programs are designed based on the assumption that users' response functions are available or predictable at the LSE side.

In general, acquiring an appropriate response function for each user is a challenging task for the LSE. Although many works [20–23] have considered different models (e.g., linear, exponential and etc.) for pricing based DR problems, these specific functions could not accurately characterize users' responses due to idiosyncratic and divergent habituates of users and therefore result in higher system costs. In order to model users' response functions more accurately, we leverage reinforcement learning to learn the users' response functions, instead of assuming any specific functions for the users. Using the learned response functions, the LSE can find a better price to minimize the system cost.

B. Related Works

There has been extensive research on various aspects of demand response programs. Comprehensive surveys on demand response programs and issues can be found in [24–26]. Different linear and non-linear mathematical load models have been considered in [20–22] in order to evaluate the demand side response to the offered prices. Recently, Z. Liu *et al.* [27] have proposed prediction-based pricing for data centers' DR, which highly depends on the prediction accuracy of users' responses to prices. Furthermore, a comprehensive DR model has been introduced in real-time pricing in [23]. Besides, a composite load response function has been introduced to model the load responses more accurately. All of these studies mathematically model the response functions to acquire an accurate estimation of responses to prices. However, these models still cannot accurately characterize idiosyncratic and divergent behaviors of customers. The problem of using reinforcement learning for DR program has been studied in [28]. This paper has considered financial cost and dis-satisfaction for delayed scheduling of appliances and used Q-learning to learn the behaviors of the consumers to make scheduling decisions. However, this paper is to learn the probabilistic

978-1-5386-4328-0/18/$31.00 ©2018 IEEE

influence of users' behaviors on the offered prices in order to find an optimal pricing scheme.

C. Summary of Major Contributions

In this paper, we study LSE's pricing strategy for DR program, aiming to minimize the system cost without assuming any specific response function of the users. Our main contributions are summarized as follows:

- We formulate the DR problem as a stochastic optimization problem, in which random responses are considered to model the users' volatile behaviors. The LSE aims to acquire the desired load profile of the system via pricing to increase the system robustness and minimize the incentive payments to the participants in the DR program.
- We develop a reinforcement learning based algorithm to solve a pricing strategy for DR without knowledge of response functions. Due to divergent consumption habituates of the users, it is challenging to have an accurate estimate of users' responses. Note that efficacy of the pricing scheme mainly depends on the accuracy of the estimated response functions. Using reinforcement learning approach could enable the LSE to learn the aggregated behaviors of the customers, in order to find the optimal pricing strategy.
- We prove that the proposed reinforcement learning algorithm converges to an equilibrium. In the proposed algorithm, a control parameter β is introduced to strike a trade-off between price exploration and exploitation. Then, we quantify the impact of choosing a proper β on the performance of the proposed algorithm. Finally, we show that by choosing a moderate β, we can achieve a balance between exploration and exploitation in the learning algorithm, which yields the minimum expected system cost.

II. Optimal Pricing for Demand Response

A. System Model

In this paper, we consider a discrete-time system, where in each time slot t, the LSE aims to procure a total amount $d(t)$ of load reduction from a set of users $\mathcal{N}$. The LSE's task is to choose a price $p(t)$ from a pricing plan $\mathcal{P}$ so that the LSE achieves the desired amount of curtailment. However, by announcing a price $p(t)$, the LSE loses revenue $p(t)s_n(t)$ when user n reduces consumption by $s_n(t) \geq 0$. Figure 1 shows the system model for the described system.

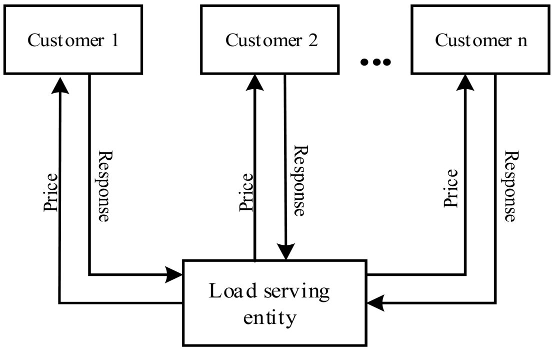

Fig. 1: System model

Note that the reduction $s_n(t)$ depends on user n's reaction to the price. We assume that the users' behaviors in different time slots are independent. Different from the existing studies, user n's response function to the price is assumed to be unavailable at the LSE. Let $r_n(\cdot)$ denote user n's response function to the price, which may depend on the energy usage state $e_n(t)$ of user n and other unknown parameters, e.g., weather conditions. For example, given the same price, a user may reduce less power for heating in a cold day, compared with a warm day. In practice, these variables of the response function represent the specific behaviors of the users and may not be available at the LSE and may differ across the users and change over the time. Therefore, given the price $p(t)$, the reduction $s_n(t)$ is a random quantity, i.e., $s_n(t) = r_n(p(t), e_n(t), x_n(t))$, where $x_n(t)$ denotes the unknown parameters to the LSE and is assumed to be independent and identically distributed.

Due to the uncertainty of users' responses, this curtailment may not match the demand response target $d(t)$. Any deviation from the target may incur some penalty to the LSE. Let $h(\cdot)$ denote the penalty function to capture the penalty of deviation from the target, which is assumed to be convex, non-negative, and has a global minimum $h(0) = 0$ (e.g., $h(x) = x^2$ is often considered in the literature [10], [27]). The penalty can be represented as $h(d(t) - \sum_n r_n(p(t), e_n(t), x_n(t)))$. Therefore, the LSE's overall system cost U at time slot t is equal to the sum of the penalty of deviation from the demand response target $d(t)$ and revenue loss, i.e.,

$$U(p(t), d(t), \mathbf{e}(t), \mathbf{x}(t)) = h\left(d(t) - \sum_n r_n(p(t), e_n(t), x_n(t))\right) + p(t)\sum_n r_n(p(t), e_n(t), x_n(t)) \tag{1}$$

where $\mathbf{e}(t) = e_n(t) \in \{\mathcal{E}\}$ denotes the energy usage state observed by the LSE, in which the energy usage $e_n(t)$ of each user n can be measured by smart meter, and $\mathcal{E}$ denotes the set of possible energy usage states. $\mathbf{x}(t) = \{x_n(t)\}$ are random variables to the LSE.

B. Problem Formulation

The objective of the LSE is to minimize the total expected system cost over time, as the LSE's overall system cost (1) is a random variable due to $\mathbf{x}(t)$. Note that we take the one-step pricing (i.e., greedy policy), such that in each time slot t, the LSE computes the best pricing policy for the current time only. For the problem under consideration, the one-step solution can be an efficient approximation for the stochastic learning based pricing problem, since the LSE has no knowledge of the future events. Therefore, the demand response problem can be formulated as the following stochastic optimization problem:

$$\min_{p(t)\in\mathcal{P}} \quad \mathbb{E}[U(p(t), d(t), \mathbf{e}(t), \mathbf{x}(t))]. \tag{2}$$

Solving the proposed demand response problem (2) requires to find a pricing strategy π that maps each state $\{\mathbf{e}(t), d(t)\}$ to a price $p(t)$. One key challenge of deriving the optimal pricing strategy is that $r_n(\cdot)$ is not available at the LSE, i.e., the LSE is not aware of users' response strategies, which is the major difference between this work and the existing studies, where users' response functions are often assumed to be availabile at the LSE or can be predicted accurately. Note that when $r_n(\cdot)$ is a linear function, it can be predicted from historical consumption data (see e.g., [27], [29–31]). However, such predictions are error prone and the prediction errors can be large when users' response functions are complicated. To tackle this challenge, we propose a reinforcement learning based approach such that the LSE can learn to adjust its pricing strategy adaptively based on users' responses.

C. Reinforcement Learning based Pricing Strategy

Specifically, the LSE determines the price based on the perception values. Each perception value corresponds to the LSE's current perception of the expected system cost when choosing a price at a given system state. Here the system state $S = \{d, \mathbf{e}\} \in \mathcal{S}$ includes the target $d \in \mathcal{D}$ and the energy usages of users $\mathbf{e} \in \mathcal{E}$, where $\mathcal{S}$ denotes the set of all possible states. Let $V_S^p(t)$ denote the perception of the expected system cost at time slot t when a price p is announced in state S. When a price p is announced, the LSE updates only the corresponding perception based on the perceived system cost in current state $S(t)$ from (1) as follows:

$$V_S^p(t) = \begin{cases} (1-\alpha_t)V_S^p(t-1) + \alpha_t U(t), \text{if } p(t) = p, \\ \qquad\qquad\qquad\qquad\qquad\qquad S(t) = S \\ V_S^p(t-1) \qquad\qquad\qquad\quad, \text{otherwise} \end{cases} \tag{3}$$

where α_t is a learning rate parameter satisfying that $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$. $U(t)$ is the system cost at time slot t, which can be computed based on users' responses $s_n(t)$ at time slot t. Simply put, based on (3), the LSE updates only the perception of system cost in the current state under price p and keeps the perceptions in other states unchanged.

Then, the price for the next time slot is announced based on the pricing strategy $\pi_S(t) = \{\pi_S^p(t)\}_{p\in\mathcal{P}}$, where $\pi_S^p(t)$ denotes the probability of announcing the price p at state S. Using the Boltzmann distribution, the pricing strategy is given as follows:

$$\pi_S^p(t) = \frac{\exp(-\beta V_S^p(t))}{\sum_{p'\in\mathcal{P}} \exp(-\beta V_S^{p'}(t))}, \tag{4}$$

where β is the parameter that makes the exploration versus exploitation tradeoff. When β is large, the LSE will choose these prices with currently lower perceptions with higher probabilities. When β is small, the LSE will explore these prices with currently higher perceptions. Intuitively, the choice of β would play a very important role in finding the optimal pricing strategy for the LSE. And, a moderately small β is needed to increase the randomness of the pricing strategy, in order to ensure sufficient exploration over the possible prices to guarantee the convergence of the algorithm as well as small performance loss. The proposed reinforcement learning based pricing algorithm is summarized in Algorithm 1.

Algorithm 1 Reinforcement learning based pricing algorithm

```
Initialization: Given the set of states $\mathcal{S}$ and the pricing plan $\mathcal{P}$, set the initial perception values $V_S^p(0) = 1$.
For each time slot $t$
1) Select a price $p(t)$ according to (4).
2) Compute the perceived system cost according to (1) using users' responses $s_n(t)$.
3) Update the perception values $V_S^p(t)$ according to (3).
```

D. Performance Analysis

We now study the convergence and optimality of the proposed reinforcement learning algorithm [32]. Proof of Lemma 1 is provided in [33] due to page limitation. All other proofs are relegated to appendix. First, We define the mapping function from the perception $\mathbf{V}(t) \triangleq (V_S^p(t), \forall S \in \mathcal{S}, p \in \mathcal{P})$ to the conditional expected cost value of the LSE given state S and price p, as

$$R_S^p(\mathbf{V}(t)) \triangleq \mathbb{E}\left[U(t)|\mathbf{V}(t), S(t) = S, P(t) = p\right] \tag{5}$$

where $\mathbb{E}[\cdot]$ is taken with respect to the random variable $x(t)$. We have the following result.

Lemma 1: If the parameter β satisfies the following condition,

$$\beta < \frac{1}{\max_{U^p(t)\in\mathcal{U}}\{\mathbb{E}[|U^p(t)|]\}} \tag{6}$$

where,

$$U^p(t) = h\left(d(t) - \sum_n r_n(p, e_n(t), x_n(t))\right) + p\sum_n r_n(p, e_n(t), x_n(t)),\ U^p(t) \in \mathcal{U}$$

Then, the mapping function $\mathbf{R}_S(\mathbf{V}(t)) \triangleq (R_S^p(\mathbf{V}(t)), \forall S \in \mathcal{S}, p \in \mathcal{P})$ forms a maximum norm contraction mapping, i.e.,

$$\|\mathbf{R}_S(\mathbf{V}) - \mathbf{R}_S(\hat{\mathbf{V}})\|_\infty \le \epsilon\|\mathbf{V} - \hat{\mathbf{V}}\|_\infty$$

where $\epsilon \triangleq \beta \max_{U^p(t)\in\mathcal{U}}\{\mathbb{E}[|U^p(t)|]\}$, $0 < \epsilon < 1$ and $\mathcal{U} = \{U(S,p) : \forall S \in \mathcal{S}, p \in \mathcal{P}\}$ is finite state-action space.

Lemma 1 implies that when the difference between the load reduction response and the reduction target is high, a smaller β is required to guarantee the convergence.

Now, we can show by using the property of contraction mapping that $\{\mathbf{V}(t), \forall t \ge 0\}$ is a sequence converging to a unique fixed point (i.e, the optimal point) $\mathbf{V}^*$.

Theorem 1: For Algorithm 1, if the parameter β follows (6), then the sequence $\{\mathbf{V}(t), \forall t \ge 0\}$ converges to an equilibrium point $\mathbf{V}^*$.

Now, we investigate the property of the fixed point $\mathbf{V}^*$ as an equilibrium of the reinforcement learning algorithm. Theorem 1 implies that the LSE can achieve an accurate estimation of objective value based on perception V_S^{p*} at the equilibrium. The following theorem shows an important result for perception value at $\mathbf{V}^*$ based on pricing strategy $\pi_S(t)$.

Theorem 2: For Algorithm 1, the pricing strategy $\pi_S^*(t) = \{\pi_S^{p*}\}_{p\in\mathcal{P}}$ at the equilibrium $\mathbf{V}^*$ approximately minimizes the expected cost function, i.e.,

$$\sum_{p\in\mathcal{P}} \pi_S^{p*} R_S^p(\mathbf{V}^*) \le \min_{\pi_S} \left\{ \sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) \right\} + \varphi$$

where the approximation gap φ is at most $\frac{1}{\beta}\ln(\mathcal{P})$.

Theorem 2 implies that a large β is required to have less performance gap in the reinforcement learning algorithm. On the other hand, Theorem 1 states that a small β is required to ensure the convergence of the reinforcement learning algorithm. Hence, Theorems 1 and 2 together indicate that a moderate β is required to establish a trade-off between the price strategy exploration and exploitation and ensure the convergence of the reinforcement learning algorithm to an equilibrium. It should be stated that the value of β is chosen off-line using historical data.

Remark. *The speed of convergence greatly depends on the size of the price set as the number of prices goes up, it takes longer time to explore the new pricing strategies. We use historical data to learn good initial pricing strategies off-line, in order to accelerate the convergence speed.*

III. Numerical Results

In this section, the performance of the pricing model will be evaluated via numerical simulations. Also, we will show the impact of the trade-off parameter β on the performance of the reinforcement learning algorithm.

A. Experimental Setup

We consider a case with 30 independent users to examine the performance of the proposed pricing mechanism. The hourly load profile for each user is generated based on a domestic electricity demand model proposed in [34] by taking mean values of each hour. Aggregated load profile of these loads before pricing is shown in Figure 2. We consider a composite random response consisting of linear, exponential, and logarithmic load models [23] associated with random variables to model the idiosyncratic reduction response, $s_n(t)$, to the prices. We also assume that the change of the load dynamics are slow which does not affect the learning process. Furthermore, based on the price range in Nevada electric services company (NV Energy), we assume that the approximate range of residential electricity rates is between 0.05 ($\$/KWh$) to 0.35 ($\$/KWh$). Hence, we consider our price set as $\mathcal{P} = \{0.05 : 0.05 : 0.35\}(\$/KWh)$ and initial price is assumed to be 0.15 ($\$/KWh$). Besides, the demand response target is assumed based on LSE's technical requirements (e.g., 30% load shifting from peak hours) as depicted in Figure 2. For this case, we generate the historical data in hourly time steps to train the algorithm.

B. Case Studies

1) Convergence speed: We show the convergence speed of the proposed reinforcement learning algorithm in Figure 3. The results are illustrated for different prices in the price set at $t = 1$ and $\beta = 5$. It is shown that the probabilities are converged in less than 15 iterations. The small variations in the probabilities after convergence are due to the characteristics of the Boltzmann distribution which tries to balance a trade-off between exploration and exploitation. However, these variation are within the range of convergence. Note that the converged probabilities are associated to a state. This means that price set probabilities would converge to different values in different states.

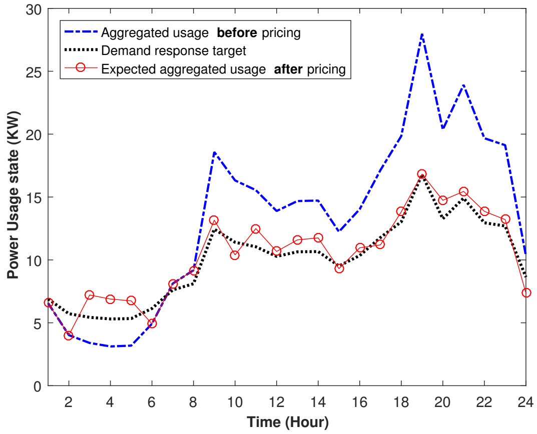

Fig. 2: Aggregated load profile before and after pricing for 30 independent loads

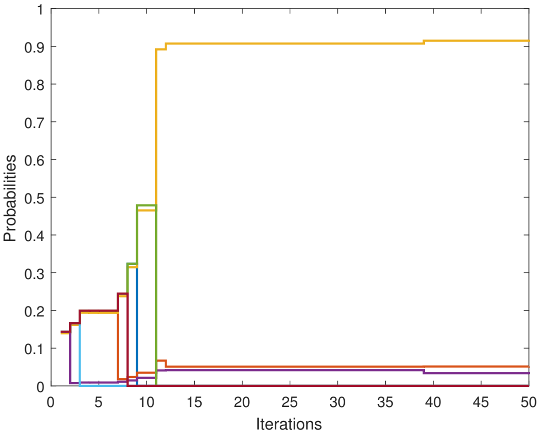

Fig. 3: Convergence of pricing strategy

2) Load profile shaping: We first evaluate the performance of our pricing methodology by showing the aggregated load profile of the users before and after pricing, as illustrated in Figure 2. As expected, the pricing strategy announces prices based on the difference between users' usage states and the desired demand response target. For example, if a higher reduction response is required, the LSE would announce a higher price to encourage the customers to reduce their consumption levels. Furthermore, the LSE could use the pricing mechanism to shape the desired load profile of the system. There are extreme load variations at peak times which could violate technical constraints of the system (e.g., ramping constraints) as shown in Figure 2. The LSE announces the high prices in order to smooth the system's spiky load profile. Figure 4 illustrates the prices in different time slots given by the proposed algorithm.

3) Response functions analysis: We also conduct a simulation to compare the results of the proposed approach with those using specific mathematical response functions as illustrated in [23]. The users' responses using different approaches (i.e., our approach, linear, logarithmic, and exponential) are illustrated in Figure 5. The results show that the expected average cost (Equation (2)) for random response is less than the cost for other assumed response functions. This indicates that using specific mathematical response functions could not accurately characterize users' responses and therefore could result in a sub-optimal solution and higher system cost. Hence, it is beneficial to use reinforcement learning to learn the users' behaviors, which can reduce the system cost.

4) Impact of β: We then evaluate the performance of the reinforcement learning algorithm for different choices of β. Figure 6 shows the system average cost for different values of β. It is indicated that with a small β, the LSE tends to explore more price rates, which results in a large performance gap according to Theorem 2. On the other hand, a large β may prevent the algorithm from finding the optimal pricing strategy due to over-exploitation, which can affect the system performance negatively. For example, by considering an extremely large β, the mapping distribution from perception to the policy space (i.e., Boltzmann distribution) boils down to $\pi_S^{p*} = \arg\max_{p \in \mathcal{P}}\{\pi_S^p\}$ which turns out to be a greedy paradigm. Then, the learning algorithm cannot sufficiently explore the policy space to find the globally optimal solution. As illustrated in Figure 6, the algorithm converges to sub-optimal solutions for large β. Hence, a moderate β needs to be chosen to strike a balance between exploration and exploitation in order to achieve the best performance. In our example, $\beta = 5$ is chosen, which yields the best trade-off between exploration and exploitation and provides the best system performance.

IV. Conclusions and Future Work

We have studied the optimal pricing scheme for DR program from LSE's point of view. Specifically, we developed a reinforcement learning algorithm to learn the behaviors of users, instead of using pre-defined response functions. Besides, convergence and optimality analyses for the learning algorithm were derived using contraction mapping theorem. The efficacy of the proposed pricing strategy was further evaluated through numerical simulations. Also, it was demonstrated that by choosing a proper β, we can achieve a balance between exploration and exploitation in the learning algorithm, which in turn results in a better system performance.

For the future work, we will study pricing strategies for privacy-preserving DR programs. Due to the growth of security and privacy concerns, customers tend to hide their real responses by adding noise. This will prevent external intruders to extract useful information of the customers. However, it also renders a significant challenge for the LSE to find good pricing strategies for DR programs. We believe that the findings in this paper would shed light on the development of privacy-preserving DR programs.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation under Grants CNS-1559696 and IIA-1301726.

References

[1] R. J. Hamidi, H. Livani, S. Hosseinian, and G. Gharehpetian, "Distributed cooperative control system for smart microgrids," *Electric Power Systems Research*, vol. 130, pp. 241–250, 2016.

[2] M. H. Amini, M. P. Moghaddam, and O. Karabasoglu, "Simultaneous allocation of electric vehicles parking lots and distributed renewable resources in smart power distribution networks," *Sustainable Cities and Society*, vol. 28, pp. 332–342, 2017.

[3] S. M. M. H. N., S. Heydari, H. Mirsaeedi, A. Fereidunian, and A. R. Kian, "Optimally operating microgrids in the presence of electric vehicles and renewable energy resources," in *2015 Smart Grid Conference (SGC)*, Dec 2015, pp. 66–72.

[4] A. Ghasemkhani, H. Monsef, A. Rahimi-Kian, and A. Anvari-Moghaddam, "Optimal design of a wide area measurement system for improvement of power network monitoring using a dynamic multiobjective shortest path algorithm," *IEEE Systems Journal*, 2015.

[5] M. He, L. Yang, J. Zhang, and V. Vittal, "A spatio-temporal analysis approach for short-term forecast of wind farm generation," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1611–1622, 2014.

[6] L. Yang, M. He, J. Zhang, and V. Vittal, "Support-vector-machine-enhanced markov model for short-term wind power forecast," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 791–799, 2015.

[7] L. Yang, M. He, V. Vittal, and J. Zhang, "Stochastic optimization-based economic dispatch and interruptible load management with increased wind penetration," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 730–739, 2016.

[8] V. Sarfi, I. Niazazari, and H. Livani, "Multiobjective fireworks optimization framework for economic emission dispatch in microgrids," in *North American Power Symposium (NAPS), 2016*. IEEE, 2016, pp. 1–6.

[9] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Optimal privacy-preserving energy management for smart meters," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 513–521.

[10] ——, "Cost-effective and privacy-preserving energy management for smart meters," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 486–495, 2015.

[11] M. Parvizimosaed, A. Anvari-Moghaddam, A. Ghasemkhani, and A. Rahimi-Kian, "Multi-objective dispatch of distributed generations in a grid-connected micro-grid considering demand response actions," 2013.

[12] F. E. R. Commission *et al.*, "A national assessment of demand response potential," *prepared by The Brattle Group, Freeman Sullivan, & Co, and Global Energy Partners*, 2009.

[13] N. Ruiz, I. Cobelo, and J. Oyarzabal, "A direct load control model for virtual power plant management," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 959–966, 2009.

[14] A. Gomes, C. Antunes, and A. Martins, "A multiple objective approach to direct load control using an interactive evolutionary algorithm," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1004–1011, 2007.

[15] D. Weers and M. Shamsedin, "Testing a new direct load control power line communication system," *IEEE transactions on power delivery*, vol. 2, no. 3, pp. 657–660, 1987.

[16] K. Herter, "Residential implementation of critical-peak pricing of electricity," *Energy Policy*, vol. 35, no. 4, pp. 2121–2130, 2007.

[17] C. Triki and A. Violi, "Dynamic pricing of electricity in retail markets," *4OR*, vol. 7, no. 1, pp. 21–36, 2009.

[18] P. Centolella, "The integration of price responsive demand into regional transmission organization (rto) wholesale power markets and system operations," *Energy*, vol. 35, no. 4, pp. 1568–1574, 2010.

[19] Y. Yuan, Z. Li, and K. Ren, "Modeling load redistribution attacks in power systems," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 382–390, 2011.

[20] F. C. Schweppe, M. C. Caramanis, R. D. Tabors, and R. E. Bohn, *Spot pricing of electricity*. Springer Science & Business Media, 2013.

[21] F. C. Schweppe, M. C. Caramanis, and R. D. Tabors, "Evaluation of spot price based electricity rates," *IEEE Transactions on Power Apparatus and Systems*, no. 7, pp. 1644–1655, 1985.

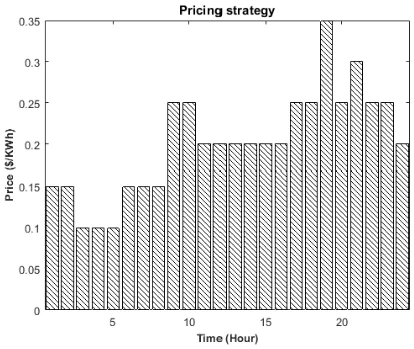

Fig. 4: Price rates in different time slots

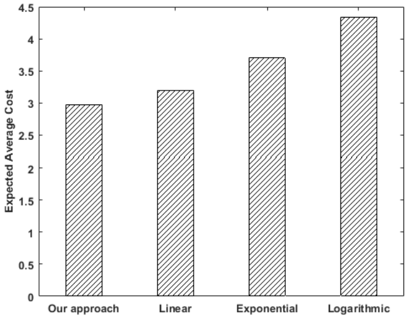

Fig. 5: Expected average cost for different models of users' responses

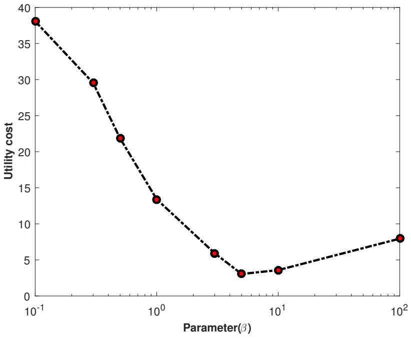

Fig. 6: System performance with different choices of β

[22] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 236–242, 2010.

[23] S. Yousefi, M. P. Moghaddam, and V. J. Majd, "Optimal real time pricing in an agent-based retail market using a comprehensive demand response model," *Energy*, vol. 36, no. 9, pp. 5716–5727, 2011.

[24] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, "A survey on demand response in smart grids: Mathematical models and approaches," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 570–582, 2015.

[25] J. S. Vardakas, N. Zorba, and C. V. Verikoukis, "A survey on demand response programs in smart grids: Pricing methods and optimization algorithms," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 152–178, 2015.

[26] M. H. Shoreh, P. Siano, M. Shafie-khah, V. Loia, and J. P. Catalão, "A survey of industrial applications of demand response," *Electric Power Systems Research*, vol. 141, pp. 31–49, 2016.

[27] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pp. 111–123, 2014.

[28] Z. Wen, D. ONeill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2312–2324, 2015.

[29] D. W. Cai and A. Wierman, "Inefficiency in forward markets with supply friction," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 5594–5599.

[30] F. Murphy and Y. Smeers, "On the impact of forward markets on investments in oligopolistic markets with reference to electricity," *Operations research*, vol. 58, no. 3, pp. 515–528, 2010.

[31] J. Yao, S. S. Oren, and I. Adler, "Two-settlement electricity markets with price caps and cournot generation firms," *european journal of operational research*, vol. 181, no. 3, pp. 1279–1296, 2007.

[32] X. Chen, X. Gong, L. Yang, and J. Zhang, "Amazon in the white space: Social recommendation aided distributed spectrum access," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 1, pp. 536–549, 2017.

[33] "Reinforcement learning based pricing for demand response." [Online]. Available: https://1drv.ms/b/s!AlY6wtnbo9UrlyQS7bYrDbZXjb8Z

[34] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy and buildings*, vol. 42, no. 10, pp. 1878–1887, 2010.

[35] A. Granas and J. Dugundji, *Fixed point theory*. Springer Science & Business Media, 2013.

Appendix

Proof of Theorem 1:

We can show that the sequence $\{\mathbf{V}(\mathrm{t})\}$ is a Cauchy sequence, hence it will converge to a limiting point based on fixed point theorem i.e., $\lim_{t\to\infty} \mathbf{V}(t) = \mathbf{V}^*$. The detailed convergence proof of a fixed point can be followed in [35].

Proof of Theorem 2:

First, we need to form an optimization problem that balances between pricing strategy exploitation and exploration. Hence, we consider the following problem:

$$\min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) + \frac{1}{\beta} \sum_{p\in\mathcal{P}} \pi_S^p \ln(\pi_S^p) \right)$$
$$\text{subject to} \sum_{p\in\mathcal{P}} \pi_S^p = 1, \pi_S^p \geq 0, \forall p \in \mathcal{P} \qquad (7)$$

the first term in (7) indicates to performance of the pricing strategy (i.e., exploitation) while the second term represents the entropy which measures the randomness of the pricing strategy (i.e., exploration). In other words, the minimization problem (7) tries to find the best trade-off between the price exploitation and exploration. Since the problem function is convex, using KTT conditions can result in the optimal solution as follows:

$$\tilde{\pi}_S^p = \frac{\exp(-\beta R_S^p(\mathbf{V}^*))}{\sum_{p'\in\mathcal{P}} \exp(-\beta R_S^{p'}(\mathbf{V}^*))}$$

we know by the concept of fixed point that at the point $\mathbf{V}^*$, $R_S^p(\mathbf{V}^*) = V_S^{p*}$. Hence, it can be shown that $\tilde{\pi}_S^p = \pi_S^{p*}$, i.e., the pricing strategy at $\mathbf{V}^*$ is the optimal solution to the problem (7). Also, we can show that,

$$\sum_{p\in\mathcal{P}} \pi_S^{p*} R_S^p(\mathbf{V}^*) = \min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) + \frac{1}{\beta} \sum_{p\in\mathcal{P}} \pi_S^p \ln(\pi_S^p) \right)$$
$$- \frac{1}{\beta} \sum_{p\in\mathcal{P}} \pi_S^{p*} \ln(\pi_S^{p*}) \qquad (8)$$

Furthermore, it is easy to verify that

$$\min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) + \frac{1}{\beta} \sum_{p\in\mathcal{P}} \pi_S^p \ln(\pi_S^p) \right) \leq \min_{\pi_S} \left(\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*) \right) \qquad (9)$$

Since the uniform distribution results in maximum entropy, we can say that

$$- \sum_{p\in\mathcal{P}} \pi_S^{p*} \ln(\pi_{S*}^p) \leq \ln(\mathcal{P})$$

then, it can be shown from (8) and (9) that

$$\sum_{p\in\mathcal{P}} \pi_S^{p*} R_S^p(\mathbf{V}^*) \leq \min_{\pi_S} (\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*)) - \frac{1}{\beta} \sum_{p\in\mathcal{P}} \pi_S^{p*} \ln(\pi_S^{p*})$$
$$\leq \min_{\pi_S} (\sum_{p\in\mathcal{P}} \pi_S^p R_S^p(\mathbf{V}^*)) + \frac{1}{\beta} \ln(\mathcal{P})$$