Reinforcement Mechanism Design for Electric Vehicle Demand Response in Microgrid Charging Stations

Luyang Hou, Shuai Ma, Jun Yan, Chun Wang and Jia Yuan Yu
Concordia Institute for Information Systems Engineering (CIISE)
Concordia University
Montréal, QC H3G 1M8, Canada
luyang.hou@mail.concordia.ca; m_shua@encs.concordia.ca;
jun.yan@concordia.ca; chun.wang@concordia.ca; jiayuan.yu@concordia.ca

***Abstract*—Reinforcement learning has become an important scheduling solution with many successes in markets with dynamic pricing options, e.g., electric vehicle charging in a deregulated electricity market. However, the highly-uncertain requests and partially-unknown individual preferences remain major challenges to effective demand responses in the user-centric environment. For charging stations who aim to maximize the long-term revenue in this fast-growing market, an accurate estimate of user's sensitivity, or acceptance, of the prices they offered to the potential customers is the key to the success of dynamic pricing. While most existing pricing schemes assume users will consistently follow stable patterns that are observable or inferrable by the charging service provider, it remains crucial to consider how users may be influenced by historic prices they have observed and react strategically to decide optimal charging demands that can maximize their utilities. To overcome this limitation, this paper presents a new framework based on reinforcement mechanism design to determine the optimal charging price in a mechanism design setting, which can optimize the long-term revenue of charging stations as well as the social welfare of users with private utility functions. Specifically, the strategic interaction between the station and users is modelled as a discrete finite Markov decision process, a Q-learning-based dynamic pricing mechanism is proposed to explore how price affects users' demands over a sequence of time. The experiments demonstrate that our pricing mechanism outperforms the predetermined time-of-use pricing in maximizing the long-term revenue of the charging station.**

***Index Terms*—Charging station; electric vehicle; demand response; dynamic pricing; mechanism design; utility; Markov decision process; Q-learning.**

I. Introduction

Microgrids are advancing the management efficiency and security of power grids with the ability to integrate distribution renewable energies, energy storage systems and distributed controllers [1]. However, in microgrids, peak power demands at some specific times of the day may bring higher costs to end-users and instabilities to the electricity networks [2]. Recently, the high penetration of electric vehicles (EVs) may aggravate the peak loads, which also influences the energy prices in the electricity market and consequently the efficiency of charging scheduling [3]. This situation motivates microgrid to provide incentives for EV users to adjust the timing of charging [4]. In such a case, *demand response* (DR) enables users to manage their charging preferences through time-varying prices or incentives at different periods to help improve the grid stability by shifting on-peak charging demands towards off-peak periods [5], [6]. Typical pricing schemes in the existing literature include time-of-use, critical-peak and real-time mode [5].

However, two gaps exist in the current DR-based dynamic pricing mechanisms: First, most works neglect users' self-interested nature and their preferences on power demands, simply assuming that users' demands are predefined or drawn from a given distribution [7], [8]. In the literature, electric energy tariffs are the most common way to incentivize users to modify or predict their consumption habits in order to stabilize the grid loads with an assumption that the charging actions do not affect the electricity price. However, users should also participate in the price settlement acting both as a price taker and a price maker. In realistic scenarios, users' charging demands are flexible given their utility with reference to the price. Second, some incentive-based DR mechanisms that adopt game theoretical approaches focus on computing the Nash equilibrium-based solutions for the energy management at each hour or in a short period of time [9]–[11]. However, the Nash equilibrium solutions, based on user's best response strategy regarding the price signal, are always myopic and not optimal, especially in maximizing the long-term objectives.

In terms of user's strategic behaviors in a market environment, dynamic pricing should be formulated as a *mechanism design* problem, which can naturally capture the conflicting preferences of the self-interested users and obtain socially desirable outcomes, e.g., the maximal long-term revenue of charging station and the social welfare [12]. However, it is challenging to develop such a pricing mechanism for EV-based demand response in a charging market, where users are modelled as the self-interested agents who aim only to advance their own benefits rather than the system wide

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under grants RGPIN-2016-06691 and RGPIN-2018-06724, and in part by the Fonds de Recherche du Québec – Nature et Technologies (FRQNT) under grant 2019-NC-254971.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

efficiency. Particularly, the selfish users will take advantage of the energy-flexibility by adjusting their power demands for economic benefits [13]. In addition, their decisions are affected by multiple factors, making it inapplicable to assume user's demand information is single-dimensional, statistically known, and does not change over time [14]. In the charging market, users may not be fully rational due to information asymmetry and may not follow the price signal offered by the charging station. Moreover, there is no explicit utility model for users, whose private information is subject to stochastic changes over time. Add it all up, the information that affects the dynamic pricing is uncertain, unknown and changing dynamically over time, which is accumulated from users' random arrivals and changing preferences on charging demands. Such a strategic interaction between the charging stations and users will exclude many candidates from existing demand-dependent pricing schemes, especially when the demand-price profile and valuation function of users are not precisely known. Therefore, designing a pricing mechanism needs to address the stochastic process governing the agent's preferences with changing populations over time [15].

To this end, mechanism design can be integrated with various machine learning techniques in order to accommodate a variety of dynamic settings across periods and agents' feedback and their preferences, especially for dynamic pricing to obtain more profits than those possible from a single sale price [16]. Specifically, *reinforcement learning* has been widely used in decision-making under uncertain scenarios in energy systems control such as electric vehicles and smart appliances in the smart grid [17]–[20]. It is able to explore how the proposed demand response programs can be used for foresighted users in dynamic environments. For instance, a reinforcement learning algorithm is proposed in [19] to deal with dynamic pricing and energy consumption scheduling in microgrid. The service provider acts as a broker who purchases energy from the utility company and sells it to customers, while the customers schedule their energy demands following the retail charging price. Furthermore, an incentive-based DR algorithm that integrates reinforcement learning and deep neural network is proposed in [7] to purchase energy resources from its subscribed customers, in order to balance energy fluctuations and enhance grid reliability. However, most of these works model the electricity price as a component of state and assume users are price-takers whose actions do not affect the electricity price; moreover, users are assumed to consistently follow stable patterns that are observable. Different from them, we aim to estimate user's strategic response to the prices during a sequential decision-making process.

In order to address dynamics in mechanism design, a systematic approach called *automated mechanism design* solves the mechanism design problems as a search problem via artificial intelligence techniques [21]. It takes the input information of a set of agents and returns a mechanism that maximizes an objective such as expected revenue over the agents' valuation distribution. Within this context, P. Tang proposed a modelling and algorithmic framework, i.e., *reinforcement mechanism design* [14], to solve the mechanism design as a sequential decision-making problems and optimize the economic mechanisms in dynamic environments, where a designer can make use of the data generated in the process and automatically improve future design using reinforcement learning algorithms.

In this paper, we propose a novel reinforcement mechanism design framework based on [14] to address a DR-based dynamic pricing problem in an islanded microgrid charging station, taking EV users' strategic behaviors and other dynamics into account. This framework extends an one-time, static mechanism to a sequential, dynamic one, considering the characteristics of power loads, random EV arrivals, uncertain charging demands and the private preferences of the self-interested users. Different from the classic mechanism design, we solve the dynamic pricing as a sequential decision-making process, where the charging station adaptively sets the charging prices at each hour so as to maximize its long-term revenue as well as the social welfare across all users.

In such a decentralized and dynamic environment, users act as not only the price-taker, but also the price-maker. They are incentivized to flexibly adjust their charging demands and reduce the energy consumption of load peak periods by observing the charging price and the outcome or feedback that is relevant to them; meanwhile the charging station is interested in long-term objectives such as the cumulative revenue over time with different price parameters. The strategic interaction between the charging station and users is modelled as a finite Markov decision process (MDP) and solved by Q-learning which determines the optimal pricing for charging station over time and explores users' best response on the charging demands. To the best of our knowledge, this is the first work in the existing literature that adopts reinforcement mechanism design framework to address EV-based demand response problems via dynamic pricing.

The rest of this paper is organized as follows: Section II introduces the preliminaries and problem formulation. Section III illustrates the reinforcement mechanism design framework. Section IV presents the experimental study. Section V draws a conclusion and outlooks our future research.

II. EV-Based Demand Response in Microgrid

A. System Model

We set one day of 24 hours as the operation period $\mathcal{T} = 1, 2, ..., 24$, where the t-th hour is denoted by $t \in \mathcal{T}$. We consider an islanded microgrid where a charging station controls the energy allocated to each connected EV over time with an objective to maximize its long-term revenue. This station is connected with microgrid and installed with a solar panel and an energy storage system. Its power capacity is characterized by G_t^b and G_t^r, where G_t^b is the power offered by microgrid that is limited by the transformer, and G_t^r is the power of photovoltaic array and storage system connected to this station. The charging station has m identical chargers which can simultaneously charge at most m EVs at any time

t. It is noted that vehicle-to-grid paradigm is not considered in this system model.

Consider a set of users $\mathcal{I}$ who come and leave the charging station within $\mathcal{T}$, and each user $i \in \mathcal{I}$ has a charging request to be processed by this charging station. The request is defined as a 4-tuple: $\langle \overline{at}_i, \overline{dt}_i, SoE_i^{ini}, E_i \rangle$, where $\overline{at}_i$ and $\overline{dt}_i$ are user i's earliest arrival time and latest departure time, respectively. User i should complete her charge within time window $[\overline{at}_i, \overline{dt}_i]$. SoE_i^{ini} is the initial State-of-Energy (SoE) of user i when she plugs into a charger, and E_i is the battery capacity of her EV. Noting that $SoE_{i,t} = E_i * SoC_{i,t}$, where $SoC_{i,t}$ is the State-of-Charge (%) of EV at t.

Before plug-in, user i has a minimum energy demand $e_i^{min} \in [0, E_i - SoE_i^{ini}]$, and she should also decide her demand $x_{i,t} \in \mathbb{R}_+$ at for each $t \in [\overline{at}_i, \overline{dt}_i]$ and ensure that the total charged energy $\sum_t x_{i,t}$ does not exceed the maximum energy volume restricted by the battery capacity, i.e., $\sum_{t \in [\overline{at}_i, \overline{dt}_i]} x_{i,t} \in [e_i^{min}, E_i - SoE_i^{ini}]$. In addition, let $\mathcal{I}_t$ be the set of connected EVs at t, where $\mathcal{I}_t \subseteq \mathcal{I}$; and let n_t be the number of EVs plugged in at t, where $\forall t \in \mathcal{T},\ n_t \leq m$.

The charging station first sets the energy price $\lambda_t \in \Lambda$ per unit power at t and announces it to users, and then users respond to λ_t by demanding an optimal amount of power $x_{i,t}$. Then the station starts charging EVs and observes the outcome as well as the revenue at the end of t. These two events will continue to take place sequentially. The total charging demands X_t of all connected users at t is $\sum_{i \in \mathcal{I}_t} x_{i,t}$, and the energy-related revenue of station is $\lambda_t \sum_{i \in \mathcal{I}_t} x_{i,t}$. A user also has to pay a fixed parking fee τ^p every hour, and the parking-related revenue at t is $\tau^p n_t$.

The scheduling result (an outcome) at t satisfies all the charging demands of the connected EVs, maximizing the cumulative revenue of energy and parking, as follows:

$$R_{cs}^{total} = \sum_{t \in \mathcal{T}} (\lambda_t \sum_{i \in \mathcal{I}_t} x_{i,t} + \tau^p n_t - \tau^e [\sum_{i \in \mathcal{I}_t} x_{i,t} - G_t^b]^+). \quad (1)$$

If the total demands X_t exceeds the capacity G_t^b, the charging station has to start using the spare energy sources G_t^r and pay extra energy costs with the per unit price τ^e, i.e., $\tau^e [\sum_{i \in \mathcal{I}_t} x_{i,t} - G_t^b]^+$, where $[y]^+ = max\{0, y\}$. In our model, we assume the backup energy sources G_t^r are always enough for the excessive demands from users, i.e., $\forall t \in \mathcal{T},\ [\sum_{i \in \mathcal{I}_t} x_{i,t} - G_t^b]^+ \leq G_t^r$.

B. Pricing Mechanism Design

As users' valuation function and demand-price curve are not precisely known by the charging station. While the sequential decisions made by the station relies on the knowledge of users' charging demands at each hour, which come from the rough estimation of the maximum energy requirements according to the battery capacity of the vehicle model. To maximize the long-term revenue, charging station has to develop efficient mechanisms to elicit an estimated relation between the price and users' charging demands through the strategic interaction.

We first construct a mechanism design environment.

Definition 1 (Mechanism Environment): A mechanism environment $\Gamma = \{\mathcal{I}, \{\Theta_i\}_{i \in \mathcal{I}}, \{\mathcal{X}_i\}_{i \in \mathcal{I}}, \Phi, \{v_i\}_{i \in \mathcal{I}}\}$ consists of

- a set of users $\mathcal{I}$, where $\mathcal{I} = \{1, 2, ..., n\}$;
- for every user $i \in \mathcal{I}$, a set of types Θ_i;
- for every user $i \in \mathcal{I}$, a set of actions $\mathcal{X}_i$;
- a set of outcomes Φ and
- for every user $i \in \mathcal{I}$, a valuation function v_i.

Specifically, (i) the type of user encapsulates all the information possessed by users that is not publicly known. Type will affect user's valuation over the outcomes, and thus bring uncertainties in determining the charging demands. In our model, user i's type Θ_i is her current SoC level. (ii) Action set $\mathcal{X}_i$, a function of user i's type Θ_i at each hour, includes her all possible demands. An action profile $\boldsymbol{X}$ is denoted as the Cartesian product of the action set of all users: $\boldsymbol{X} = \prod_{i=1}^n \mathcal{X}_i$, and $\boldsymbol{x} = (x_1, x_2, ..., x_n) \in \boldsymbol{X}$. (iii) The set of outcomes Φ includes the energy allocation at each hour given the users' demands. (iv) User i's valuation v_i is the measurement on an outcome ϕ based on her type, i.e., $v_i(\phi; \theta) : \Theta_i \times \Phi \to \mathbb{R}_+$, which reflects user's demand-price curve. The system-wide goal of mechanism design is defined with a social choice function $f : \prod_{i=1}^n \Theta_i \to \Phi$, which maps the type profile of all users to a set of outcomes. Social choice selects the optimal outcome given agent types [22].

In this mechanism design environment, dynamic pricing mechanism is essentially the procedure through which achieves a desired social goal by providing incentives to users. This dynamic pricing mechanism contains a decision policy and a payment policy, as follows:

Definition 2 (Pricing Mechanism): A pricing mechanism $(\boldsymbol{x}, \{p_i\}_{i \in \mathcal{I}})$ over a mechanism environment Γ consists of

- A decision policy $\boldsymbol{x} : \Lambda \to \{x_{i,t}\}_{i \in \mathcal{I}}$, which maps the charging prices Λ to the charging demands of users at t;
- For each user i, a payment function $p_i : \boldsymbol{X} \to \mathbb{R}_+$, which maps the action profile $\boldsymbol{X}$ of all users to a real number.

In our study, user i pays $p_{i,t} = \lambda_t x_{i,t} + \tau^p$ at t. This pricing mechanism proceeds as follows: charging station sets the charging price λ_t at each hour t from the parameterized class Λ, and finds a policy that enjoys desirable cumulative revenue. Users observe the announced price signal at the end of time t, and then react strategically to determine their demands $x_{i,t+1}$ for the next hour. At the end of $t+1$, charging station receives an outcome as well as the associated immediate reward.

III. Reinforcement Mechanism Design Framework

To implement the pricing mechanism in sequential periods, we formalize the strategical interaction between the charging station and users as an MDP and solve the dynamic pricing with Q-learning, considering the uncertainties coming from the charging demands and random arrivals of EVs. The reinforcement mechanism design framework is illustrated in Fig. 1. In this section, we first introduce the preliminaries about MDP; and then present the detailed MDP formulation for the charging station and the Q-learning algorithm; finally, we analyze user's strategy in this dynamic pricing mechanism.

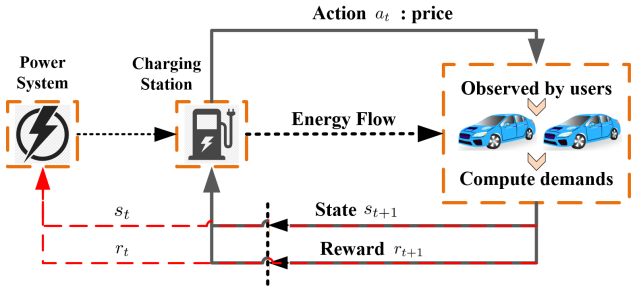


Fig. 1. MDP model for the interaction between charging station and users.

A. Preliminaries

The station-user interaction is formulated as an Markov decision process [23], which is typically characterized by a 5-tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\mathcal{S}$ is a finite set of states $s_t \in \mathcal{S}$ and $\mathcal{A}$ is a finite set of actions $a_t \in \mathcal{A}$. The function $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ defines the state transition probabilities, where $p(s_{t+1}|s_t, a_t)$ represents the transition probability from s_t to s_{t+1} after a_t is taken. The stochastic process satisfies the Markov property: $p(s_{t+1}|s_0, a_0, ..., s_t, a_t) = p(s_{t+1}|s_t, a_t)$. The function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the expected rewards for state–action pairs, where $r(s_t, a_t)$ is the immediate reward received when a_t is taken at s_t. Let R_t denote the discounted sum of rewards from the state s_t, then $R_t = \sum_{t \in \mathcal{T}} \gamma^t r(s_t, a_t)$, where $\gamma \in (0, 1]$ is the discount factor. In the case of charging scheduling, a station chooses a charging price from the given set in the current state, and users respond strategically based on the price. At the end of t, charging station receives an immediate reward associated with the outcome. Then the time progresses to $t + 1$ with all information updated accordingly.

B. Charging Station Side Analysis

In this MDP, a state consists of the base power capacity and battery capacity information of the connected EVs; the action for a charging station is to set the charging price; and the immediate reward is the total expected station revenue at the current hour. Specifically, the variables are defined as follows:

1) States: A state s_t is defined as a 3-tuple: $\langle G_t^b, E_t^{req}, n_t \rangle$, which consists of the base load G_t^b of the charging station, the total required energy E_t^{req} from all users, and the number of connected EVs n_t at t. In this study, $E_t^{req} \approx \sum_{i \in \mathcal{I}_t} (E_i - SoE_i^{ini})$, where E_t^{req} is an estimation of the total maximal energy that all connected EVs can charge based on each user's battery capacity and her initial SoE. The optimal action for charging station is determined by observing the current state.

2) Actions: An action taken by the charging station is the decision of charging price λ_t at t and the allocation of energy based on the limited energy supply G_t^b and the required user demands E_t^{req}. The price has three levels: off-peak λ_t^l, mid-peak λ_t^m and on-peak λ_t^h, where $\Lambda = \{\lambda_t^l, \lambda_t^m, \lambda_t^h\}, \lambda_t \in \Lambda$. After these actions are taken, s_t is updated according to the strategy of users with respect to the outcome $x_{i,t}$ of time t.

Algorithm 1 Q-learning based Demand Response

Input: The price set Λ, the maximum episode $\mathcal{H}$;
Output: The optimal policy π^*, $\forall t \in \mathcal{T}$;

```
for h = 1 → H do
    for each hour t ∈ T do
        Choose a_t by ε-greedy policy;
        Take action a_t;
        for each user i ∈ I do
            User i observes the price and submits
                their optimal demands x_{i,t};
        end for
        Charging station observes r(s_t, a_t), s_{t+1};
        Update the Q value;
    end for
end for
```

3) Reward: The immediate reward r_t at s_t of the charging station is defined as its expected revenue:

$$r_t = \lambda_t \sum_{i \in \mathcal{I}_t} x_{i,t} + \tau^p n_t - \tau^e [\sum_{i \in \mathcal{I}_t} x_{i,t} - G_t^b]^+. \quad (2)$$

To maximize the total reward, Q-learning is the most widely used model-free reinforcement learning algorithm due to its simplicity, in which the agents learn the optimal policy through their interaction with the environment [23]. In our study, charging station learns the optimal pricing through the strategic interaction with users. Q-learning uses the Q value $Q(s_t, a_t)$ as an expected reward for a state-action pair (s_t, a_t). While the real reward is represented by $Q'(s_t, a_t)$ and consists of the immediate reward $r(s_t, a_t)$ and the future expected Q value: $Q'(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$. And the Q value is updated by $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \sigma[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$, $\forall (s_t, a_t)$, where σ is the learning rate. As proven in existing literature [20], [24], Q-learning obtains a near-optimal policy by driving the action-value function towards the optimal action value $Q^*(s, a)$ through iterations.

Solving an MDP is to determine the optimal policy $\pi^*(a|s) : \mathcal{S} \to \mathcal{A}$ for the dynamic pricing, which is to select the optimal action (charging price) for each state $t \in \mathcal{T}$. Numerically, the optimal policy can be calculated by: $\pi^*(a_t|s_t) \leftarrow \arg\max_{a_t} \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t)[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})]$. The process of Q-learning–based demand response algorithm is shown in Algorithm 1.

Specifically, a charging station chooses the current action a_t with the ϵ-greedy strategy subject to the observations, which can avoid staying in the local optimum by balancing the exploitation and exploration during the learning process [24]. The ϵ-greedy algorithm continues to explore, with probability 1- ϵ of selecting the best action, and with probability ϵ of selecting a random action. In our study, the optimal action is used in about 90% of the price ($\epsilon = 0.1$), and takes a completely random action in about 10% of the cases to explore and meet bigger possible rewards.

C. User Side Strategy

EV users act both as a price-taker and a price maker, who observes the charging price and adaptively adjust their charging demands for each hour. While the charging station observes the outcome at the end of current hour and determines the price for the next. The bidding process can be automatically implemented on smart phones or other platforms, where users only need to set up their charging requests and the preference information. This section explores how users respond to the charging prices in order to achieve a maximal revenue by encouraging users to adapt their charging demands.

The final total energy $\sum_{t\in[\overline{at}_i,\overline{dt}_i]} x_{i,t}$ that user i will charge is not predetermined; instead it relies on the charging price λ_t and the current $SoE_{i,t}$ at each t. During t, users will consume the energy $x_{i,t}$ required at $t-1$, so that $SoE_{i,t+1} = SoE_{i,t} + x_{i,t}$, and then recompute their optimal demands for $t+1$ based on the updated charging price and SoE.

As the self-interested agents, users will always maximize their utilities when computing the optimal charging demands. In our model, we do not consider the strategic interaction and competition among users but focus on the station-user interaction, because users have no information about others' preferences and no conflicting interests with others. We then present the definition of user's utility.

Definition 3 (Quasi-linear Utility Function [15]): User i's utility is captured by the difference of her valuation $v_i(\cdot)$ for demand $x_{i,t}$ and the charging cost $p_{i,t}$ at t based on her type Θ_i and price λ_t, i.e.,

$$\begin{aligned} u_i(x_{i,t},\lambda_t;\theta_{i,t}) &= v_i(x_{i,t};\theta_{i,t}) - p_{i,t} \\ &= v_i(x_{i,t};\theta_{i,t}) - (\lambda_t x_{i,t} + \tau^p). \end{aligned} \tag{3}$$

The optimal demands $x^*_{i,t} \in \mathbb{R}_+$ for hour t are obtained by solving $\arg\max_{x_{i,t}} u_i(x_{i,t},\lambda_t;\theta_{i,t})$, in terms of their type $\theta_{i,t} \in \Theta_i$ and the charging price λ_t. And user i's charging cost $p_{i,t}$ includes the energy cost $\lambda_t x_{i,t}$ and parking fee τ^p. We assume that user's valuation function follows a *Logarithm* function in economics [25], [26]. Users are also assumed to have a decreasing marginal valuation as SoC increases, which implies the higher SoC level they have, the less satisfaction (lower valuation) they will get from the same amount of energy. Specifically, user's valuation is defined as the marginal value of obtaining a certain amount of energy $x_{i,t} = \Delta SoC_{i,t} * E_i$ given $SoC_{i,t-1}$, and $\Delta SoC_{i,t} = SoC_{i,t} - SoC_{i,t-1}$. Fig. 2 presents an illustrative example including two different SoC-valuation functions of user 1 and 2. And this general SoC-price curve also demonstrates that users always consume less energy when charging price is higher.

To analyze user's best response in generating the optimal demands, we first present the concept of individual rationality.

*Definition 4 (*Ex-ante *Individual Rationality):* The pricing mechanism is *ex-ante* individual rational if each user $i \in \mathcal{I}$ receives a non-negative utility by participation regardless of her type at t. That is, with user i's participation, we have

$$u_i(x_{i,t},\lambda_t;\theta_{i,t}) = v_i(x_{i,t};\theta_{i,t}) - p_{i,t} \geq 0, \ \forall t \in \mathcal{T}. \tag{4}$$

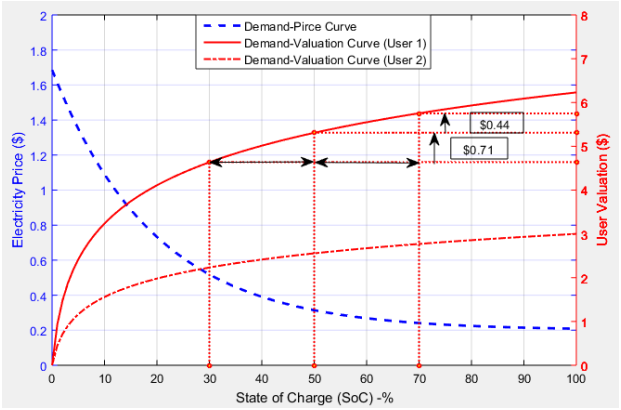


Fig. 2. An example of SoC-valuation/price curve of users. User 1 and 2 have different valuation functions, leading to different increase of values in terms of the same increase of SoC due to their individual types. It can be seen that user 1 is more sensitive than user 2 in terms of the increase of SoC. Moreover, the marginal valuation is decreasing with the increase of SoC. For instance, user 1 has an increased value of \$0.71 from 30% to 50% SoC; however, she has only \$0.44 for charging from 50% to 70% SoC.

In other words, *ex-ante* individual rationality holds if users can always achieve as much expected utility from participation as without participating, regardless of knowing her own type or other users' types [22].

Definition 5 (Best Response): User's best response $x^*_{i,t}$ is the charging demand that maximizes her utility based on her current $SoE_{i,t}$ and type Θ_i given the charging price λ_t. That is, the optimal demand is defined as $x_{i,t} : u_i(x_{i,t},\lambda_t;\theta_{i,t}) \geq max_{x'_{i,t}} u_i(x'_{i,t},\lambda_t;\theta_{i,t})$, $u_i(x_{i,t},\lambda_t;\theta_{i,t}) \geq 0$, $x_{i,t}, x'_{i,t} \in [0, E_i - SoE^{ini}_i - \sum_{t'\in[\overline{at}_i,t-1]} x_{i,t'}]$.

User i will stop charging under two conditions, which indicates $x^*_{i,t} = 0$ for t: First, for any charging demands that produce $u_i(x_{i,t},\lambda_t;\theta_{i,t}) < 0$, which indicates that continuing charging brings no more marginal values to her. Second, the current SoE reaches to EV's battery capacity limit, i.e., $E_i - SoE^{ini}_i - \sum_{t'\in[\overline{at}_i,t-1]} x_{i,t'} < x_{i,t}$.

Theorem 1: The dynamic pricing mechanism is *ex-ante* individual rational.

Proof 1: The set of outcomes Φ_{-i} that is achievable without user i is a weak subset of outcomes with user i, i.e., $\forall i$, $\Phi_{-i} \subseteq \Phi$. The utility u_i of user i is non-negative on all outcomes without her, i.e., $u_i(\phi';\theta_{i,t}) = 0$, $\forall \phi' \in \Phi_{-i}$. Noting that users are uncertain about their total demands $\sum_{t\in[\overline{at}_i,\overline{dt}_i]} x_{i,t}$ before charging, and their real demands are affected by the physical battery capacity and initial SoE. A rational user will stop charging when she obtains a negative utility, i.e., when the charging cost $p_{i,t}$ exceeds the valuation $v_i(x_{i,t};\theta_{i,t})$ brought by this amount of energy $x_{i,t}$. The parking fee is a constant cost in the utility function that can reduce a user's wait-and-see strategy to charge at a cheaper price in the future. Therefore, myopic users have no tendency to delay their charge. Therefore, user i's best response $x^*_{i,t} \leftarrow \arg\max_{x_{i,t}} u_i(\cdot)$ implies her optimal charging demands with the trade-off between valuation and cost, which admits a maximum utility under the current price λ_t. The expected utility accrued from the

rational users is always non-negative. Add it up, the proposed dynamic pricing mechanism is *ex-ante* individual rational.

Definition 6 (Weak Budget Balance): A mechanism is weakly budget balanced if all users make a non-negative payment to the charging station for all feasible type profiles, and the total payment is non-negative, i.e.,

$$\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t} p_{i,t} = \sum_{t\in\mathcal{T}}(\lambda_t \sum_{i\in\mathcal{I}_t} x_{i,t} + \tau^p n_t) \geq 0. \tag{5}$$

It can be seen that this pricing mechanism is weakly budget balanced. That is, there can only be a payment made from users to the station, but no payment from the station to users.

Followed by above properties, there exists a Nash equilibrium in this pricing mechanism if both of the charging station and users act on their best response based on the actions taken by the other side.

Definition 7 (Nash Equilibrium): The set $(\lambda_t^*, \boldsymbol{x_t^*})$ is the Nash equilibrium of this pricing mechanism [3], if charging station follows the equilibrium strategy $\lambda_t^* \in \Lambda$ given the best response $\boldsymbol{x_t^*}$ of all users at t, we have

$$r(\lambda_t^*, \boldsymbol{x_t^*}(\lambda_t^*)) \geq r(\lambda_t, \boldsymbol{x_t^*}(\lambda_t)), \quad \forall \lambda_t \in \Lambda. \tag{6}$$

where $\boldsymbol{x}$ is the action profile of all users, such that $\boldsymbol{x_t^*}(\lambda_t^*)$ is their collective best response, i.e., the optimal demands $(x_{i,t}^*)_{i\in\mathcal{I}}$ in terms of price λ_t^*. It can be inferred from the Theorem 4 in [26] that the set $(\lambda_t^*, \boldsymbol{x_t^*})$ is a Stackelberg equilibrium of the strategic interaction between the charging station and users, if the price set Λ is a non-empty, convex, and compact subset of an Euclidean space $\mathbb{R}$, and the utility function u_i of user i is continuous in Λ and concave in λ_t.

IV. Experimental Study

A. Experiment Setup

We design two experiments with different charging station sizes: $m = 10$ for Group 1 and $m = 30$ for Group 2. Both of them are Level-2 AC (240-volt) station supporting an output power of $> 3.7kW$ and $\leq 22kW$. We use the real-world 24-hour data of user power consumption at public charging stations[1], where the 20% and 50% of this commercial load are used as the base load supply $\{G_t^b\}_{t\in\{1,\dots,24\}}$ for 10 chargers (Group 1) and 30 chargers (Group 2), respectively.

The arrival rate of EVs at each hour t is assumed to follow a Possion distribution $p(k) = \frac{\delta^k}{k!}e^{-\delta}, k = 0, 1, ...,$ where $\delta = 4$ represents the Group 1 scenario, and $\delta = 6$ represents the Group 2 scenario. User i's latest departure time $\overline{dt}_i = t + \mathcal{U}[2, 5]$, where $\mathcal{U}$ is a uniform distribution, and her initial SoC is randomly distributed in $\mathcal{U}[10, 50]$ (%); then $SoE_i^{ini} = SoC_i^{ini} * E_i = 0.01 * \mathcal{U}[10, 50] * 30 = \mathcal{U}[3, 15]$. We assume that all EVs have an equivalent battery capacity $E_i = 30kWh$ and supports a maximum charging power $50kW$. The minimum energy demand e_i^{min} of user i is randomly drawn from $[0, E_i - SoE_i^{ini}]$.

[1]SCE load profiles, https://www.sce.com /regulatory/load-profiles, ID: GS-1, 08/20/2019

We build user's valuation function v_i based on the natural logarithm function following [26] and assume EV users share the same utility function, noting that our algorithm applies to heterogeneous utility functions with different α_i. The valuation-SoC function is shown as:

$$v_{i,t}^{SoC} = \begin{cases} \alpha_i \ln(\beta_i + SoC_{i,t}), & if\ 0 \leq SoC_{i,t} \leq \overline{SoC}_i \\ \alpha_i \ln(\beta_i + \overline{SoC}_i), & if\ \overline{SoC}_i \leq SoC_{i,t} \end{cases} \tag{7}$$

where $SoC_{i,t} = (SoE_{i,t-1} + x_{i,t-1})/E_i$. Noting that demand $x_{i,t}$ at every $t \in [\overline{at}_i, \overline{dt}_i]$ satisfies $x_{i,t} \in [0, E_i - SoE_i^{ini} - \sum_{t'\in[\overline{at}_i, t-1]} x_{i,t'}]$, such that the total energy charged will not exceed the battery capacity. α_i is randomly drawn from $0.2 * E_i * \mathcal{U}[0, 1]$ according to the different demand profile of users, and $\beta_i = 1$. $\overline{SoC}_i$ is the threshold of the marginal valuation (often set as 80%), because EVs' SoC or the charging voltage will not significantly increase at a saturation stage according to the battery charging profile[2]. The valuation is measured by the marginal gain for obtaining $x_{i,t}$ subject to the current SoC, i.e., $v_i(x_{i,t}; \theta_{i,t})$ for demand $x_{i,t}$, which is $E_i\alpha_i[\ln(\beta_i + SoC_{i,t}) - \ln(\beta_i + SoC_{i,t-1})]$. The optimal demands for t is computed by $x_{i,t} \leftarrow \arg\max_{x_{i,t}} u_i(x_{i,t}, \lambda_t; \theta_{i,t}) \pm \xi$, where ξ is an uncertain factor over user demands, $\xi \in [0.05E_i, 0.1E_i]$.

This experiment study uses the charging price in the U.S. public charging stations as the reference, which is around $\$0.15/kWh$ after tax[3]. Accordingly, the charging price of off-peak λ_t^l, mid-peak λ_t^m and on-peak λ_t^h hour in our model is set as $\$0.1/kWh$, $\$0.15/kWh$ and $\$0.2/kWh$, respectively. The parking cost τ^p is \$1 per hour. The extra energy purchasing fee τ^e is $\$0.35/kWh$. And we set 7:00 p.m. to 7:00 a.m. as off-peak hour, 7:00 a.m. to 11:00 a.m. and 5:00 p.m. to 7:00 p.m. as mid-peak hour, and 11:00 a.m. to 5:00 p.m. as on-peak hour in a general case[4].

We compare the pricing policy by the Q-learning with the uncontrolled and static strategy, namely the predetermined Time-of-Use (TOU) pricing, for these two groups of experiments. A user's best response and strategy under TOU pricing, as well as other experimental parameters, including random EV arrivals and user side information, etc., share the same setting as they are in the dynamic pricing mechanism.

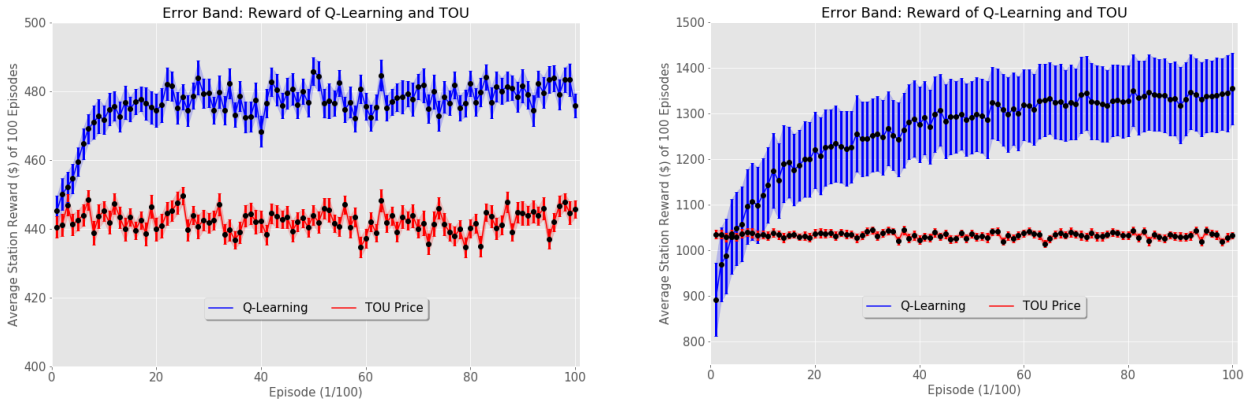
In this experiment, Q-learning algorithm and static TOU pricing have ten parallel experiments for each group, and one experiment iterates for 10,000 times (iterations); and the solutions are used to define a policy. Each iteration calculates the total rewards (revenue) R_{cs}^{total} (1) of a day ($24h$). To better display the performance of two methods in terms of the revenue, we take the average rewards of 100 iterations as one episode, and each experiment has totally 100 episodes.

We use the Q-learning algorithm to approximate $Q(s, a)$ which takes a state s as input and outputs a vector of Q-values corresponding to the actions of charging station:

[2]Battery University, https://batteryuniversity.com/learn/article/charging_lithium_ion_batteries.

[3]Global EV Outlook 2019: Scaling up the transition to electric mobility, https://www.iea.org/gevo2019/.

[4]TOU Pricing and Schedules, https://www.powerstream.ca/customers/rates-support-programs/time-of-use-pricing.html.



(a) (b)

Fig. 3. Error band by Q-learning (upper curve) and TOU pricing (lower curve) of 100 episodes. Group (1): with 10 chargers; Group (2): with 30 chargers. Each band takes the mean and standard deviation of the station reward of ten parallel experiments (y-axis) at each episode (x-axis).

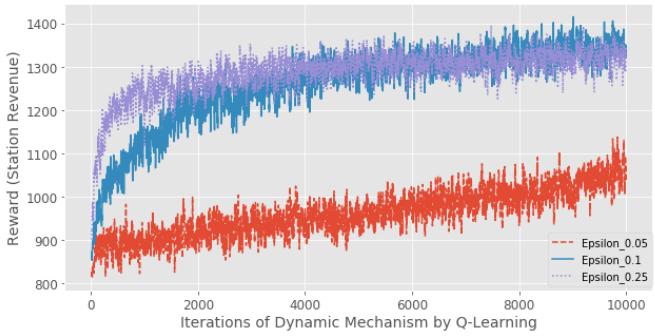


Fig. 4. (Smoothed) Rewards of Q-learning in training for Group 2 (with 30 chargers): one experiment example with 10,000 iterations. Three different epsilons (0.05, 0.1 and 0.25) are tested in this experiment study.

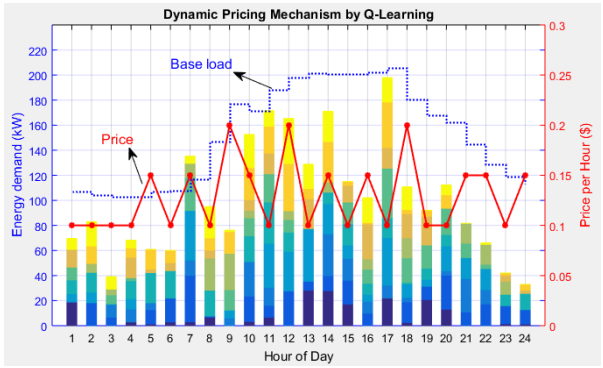


Fig. 5. User charging demands for charging in terms of the best charging price learned by Q-learning: one iteration example. The stack bar (left blue) shows the charging demands of each connected EV at each hour in terms of the charging price (right red). It can be seen that when the energy supply is low, the charging price can efficiently reduce the energy consumption from users and postpone the charging activities of some users from peak-hours to off-peak hours, such that the load stability can be well maintained.

$\lambda_t \in \{\lambda_t^l, \lambda_t^m, \lambda_t^h\}$. The pricing mechanism and Q-learning algorithm are coded in Python and use reinforcement learning environments from the OpenAI Gym. The experiments are carried out on a PC with a processor of Intel (R) Core (TM) i5-6500U CPU @ 3.2GHz, 8GB memory.

B. Results and Analysis

Fig. 3 demonstrates the performance of two groups using Q-learning algorithm and TOU pricing, respectively, which reports an error band-with the mean and standard deviation during training the cumulative reward (revenue). In Fig. 3 (a) (Group 1), the station revenue of these 100 episodes for the Q-learning and TOU are around \$476.44 and \$442.32, with an variance of \$6.81 and \$3.19, respectively. The average revenues of Group 2 are presented in Fig. 3 (b), which are \$1,321.25 with a variance of \$92.15 and \$1,032.74 with a variance of \$6.13 for dynamic pricing and TOU, respectively. The Q-learning with dynamic pricing mechanism can improve the station revenue for around 7.71% compared to the TOU pricing for Group 1 and around 27.93% for Group 2, which indicates charging station can make more \$34.12 (Group 1) and \$288.51 (Group 2) profits a day. Moreover, it can be seen that the dynamic mechanism presents a better performance for the larger charging station size with more users, as can be seen that Group 2 improves averagely 27.93% compared to TOU pricing in terms of the revenue. A larger station size implies more options for the demand response.

We pick one experiment with 10,000 iterations of Group 2 and present its rewards with three different epsilons (ϵ = 0.05, 0.1 and 0.25, respectively) in Fig. 4. It demonstrates that Q-learning algorithm converges to an average reward of \$1,326.61 with ϵ = 0.1. In addition, Fig. 5 presents one iteration of dynamic pricing under Q-learning, with the curve of users' charging demands and the electricity price from 1:00 a.m. to 12:00 p.m..

Nash equilibrium. It can be seen from Definition 7 that the Nash equilibrium exists if Λ is a non-empty, convex, and compact subset of an Euclidean space, while the utility function is continuous in Λ and concave in λ_t. We can easily see that the first condition holds. Combined with the utility

definition (3) and the valuation function (7), the second order derivative of user i's utility u_i is $\frac{\partial^2 u_i}{\partial \lambda_t^2} = 0,\ \forall t \in \mathcal{T}$. Hence, the second condition also holds. Therefore, Nash equilibrium exists if the best price setting can be learned for each hour under a lack of user-side information.

Since the reward obtained by the Q-learning algorithm is an expected value, the pair $(\lambda_t^*, \boldsymbol{x_t^*})$ is an approximation of Nash equilibrium at t after training the optimal policy π^* by Q-learning. Different from the identical-interest Nash equilibrium of stochastic game that computes the joint optimal policy of all players [27], the MDP model in this paper acts as a leader-follower mode, like [3], [26]. The equilibrium strategy exists in the supply and demand side where users have no conflicting interests with each other.

V. Conclusion and Future Research

This paper proposes a reinforcement mechanism design framework to solve a dynamic pricing problem of an islanded microgrid charging station in a dynamic charging market. The sequential strategic interaction between the charging station and users is modelled as an MDP and solved by the Q-learning algorithm. The optimal price settlement is learned by Q-learning considering the random arrivals of EVs and the uncertain charging demands of users in this sequential decision-making process. The experimental results show the charging station revenue by our approach can be improved by a maximum of 27.93% compared to the TOU pricing.

In our model, users are myopic agents who only care about their own utility in a short period of time (e.g., one hour), while computing the optimal charging demand needs more information about future parameters. For example, users may tend to wait for a better deal at a lower price in future and take the potential risk of an increased costs. Our future work will model user's decision-making as an MDP and explores the optimal joint policy of all users that gives them the maximal expected sum of discounted utilities. Moreover, more strict and detailed game theoretical proof should be developed to discuss the gap between the pair $(\lambda_t^*, \boldsymbol{x_t^*})$ and Nash equilibrium, as well as its convergence.

References

[1] Y. Yoldaş, A. Önen, S. Muyeen, A. V. Vasilakos, and İ. Alan, "Enhancing smart grid with microgrids: Challenges and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 205–214, 2017.

[2] Z. Wan, H. Li, and H. He, "Residential energy management with deep reinforcement learning," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[3] Z. Liu, Q. Wu, S. Huang, L. Wang, M. Shahidehpour, and Y. Xue, "Optimal day-ahead charging scheduling of electric vehicles through an aggregative game model," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5173–5184, 2018.

[4] S. Bahrami, V. W. Wong, and J. Huang, "An online learning algorithm for demand response in smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4712–4725, 2017.

[5] 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.

[6] M. Severini, S. Squartini, M. Fagiani, and F. Piazza, "Energy management with the support of dynamic pricing strategies in real micro-grid scenarios," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[7] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Applied energy*, vol. 236, pp. 937–949, 2019.

[8] A. Nazari and R. Keypour, "A two-stage stochastic model for energy storage planning in a microgrid incorporating bilateral contracts and demand response program," *Journal of Energy Storage*, vol. 21, pp. 281–294, 2019.

[9] F.-L. Meng and X.-J. Zeng, "An optimal real-time pricing for demand-side management: A stackelberg game and genetic algorithm approach," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 1703–1710.

[10] S.-G. Yoon, Y.-J. Choi, J.-K. Park, and S. Bahk, "Stackelberg-game-based demand response for at-home electric vehicle charging," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 4172–4184, 2016.

[11] Y. Dai, Y. Gao, H. Gao, and H. Zhu, "Real-time pricing scheme based on stackelberg game in smart grid with multiple power retailers," *Neurocomputing*, vol. 260, pp. 149–156, 2017.

[12] D. Muthirayan, D. Kalathil, K. Poolla, and P. Varaiya, "Mechanism design for demand response programs," *IEEE Transactions on Smart Grid*, 2019.

[13] B. Sun, X. Tan, and D. H. Tsang, "Eliciting multi-dimensional flexibilities from electric vehicles: A mechanism design approach," *IEEE Transactions on Power Systems*, 2018.

[14] P. Tang, "Reinforcement mechanism design." in *International Joint Conferences on Artificial Intelligence (IJCAI)*, vol. 17, 2017, pp. 26–30.

[15] D. Bergemann and J. Välimäki, "Dynamic mechanism design: An introduction," *Journal of Economic Literature*, vol. 57, no. 2, pp. 235–74, 2019.

[16] M.-F. Balcan, A. Blum, J. D. Hartline, and Y. Mansour, "Mechanism design via machine learning," in *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. IEEE, 2005, pp. 605–614.

[17] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, 2018.

[18] H. Li, Z. Wan, and H. He, "Constrained ev charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, 2019.

[19] B.-G. Kim, Y. Zhang, M. Van Der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2187–2198, 2015.

[20] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied Energy*, vol. 235, pp. 1072–1089, 2019.

[21] T. Sandholm, "Automated mechanism design: A new application area for search algorithms," in *International Conference on Principles and Practice of Constraint Programming*. Springer, 2003, pp. 19–36.

[22] D. C. Parkes and L. H. Ungar, *Iterative combinatorial auctions: Achieving economic and computational efficiency*. University of Pennsylvania, 2001.

[23] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 2, no. 4.

[24] M. Rahimiyan and H. R. Mashhadi, "An adaptive q-learning algorithm developed for agent-based computational modeling of electricity market," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 547–556, 2010.

[25] R. Srikant, *The mathematics of Internet congestion control*. Springer Science & Business Media, 2012.

[26] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, "Dependable demand response management in the smart grid: A stackelberg game approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 120–132, 2013.

[27] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal nash equilibrium in team markov games," in *Advances in neural information processing systems*, 2003, pp. 1603–1610.