Dynamic Pricing for Differentiated PEV Charging Services Using Deep Reinforcement Learning

Ahmed Abdalrahman, *Student Member, IEEE*, and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—With the increasing popularity of plug-in electric vehicles (PEV), charging infrastructure becomes widely available and offers multiple services to PEV users. Each charging service has a distinct quality of service (QoS) level that matches user expectations. The charging service demand is interdependent, i.e., the demand for one service is often affected by the prices of others. Dynamic pricing of charging services is a coordination mechanism for QoS satisfaction of service classes. In this paper, we propose a differentiated pricing mechanism for a multiservice PEV charging infrastructure (EVCI). The proposed framework motivates PEV users to avoid over-utilization of particular service classes. Currently, most of dynamic pricing schemes require full knowledge of the customer-side information; however, such information is stochastic, non-stationary, and expensive to collect at scale. Our proposed pricing mechanism utilizes model-free deep reinforcement learning (RL) to learn and improve automatically without an explicit model of the environment. We formulate our framework to adopt the twin delayed deep deterministic policy gradient (TD3) algorithm. The simulation results demonstrate that the proposed RL-based differentiated pricing scheme can adaptively adjust service pricing for a multiservice EVCI to maximize charging facility utilization while ensuring service quality satisfaction.**

***Index Terms*—PEV charging infrastructure, service differentiation, dynamic pricing, deep reinforcement learning.**

NOMENCLATURE

Indices and Sets

$\mathcal{A}$ Set of all RL agent's actions.

$\mathbf{T}$ Set of time segments over a day, indexed by t.

$\mathbf{N}$ Set of all facilities in EVCI, indexed by n.

$\mathbf{M}$ Set of service classes in the EVCI, indexed by m.

$\mathbf{P}$ Set of all parking lots in the EVCI.

$\mathbf{F}$ Set of all fast-charging stations in the EVCI.

$\mathbf{O}$ Set of all on-road wireless chargers in the EVCI.

Parameters

q_m Targeted minimum QoS level of service class m.

c_n Number of chargers allocated in facility n.

K_n Maximum number of PEVs in facility n including charging and waiting.

β_n Self-elasticity parameter with respect to the service price at facility n.

γ_m Self-elasticity parameter with respect to the service QoS at class m.

$\beta_{n,\hat{n}}$ Cross-elasticity parameter with respect to the service price at facility $\hat{n}$.

$\gamma_{m,\hat{m}}$ Cross-elasticity parameter with respect to the service QoS at class $\hat{m}$.

γ Discount factor.

τ Soft target update factor.

Variables

$\lambda_{t,n}$ PEV charging demand at facility n at time t.

$\mathcal{N}_{t,n}$ The average number of admitted PEVs to charging facility n at time t.

π Pricing policy.

$p_{n,t}$ Normalized charging price at facility n at time t.

$\mathcal{Q}_{n,t}$ Service performance metric at facility n at time t.

Q^{π} Action-value function for each state-action pair following policy π.

I. INTRODUCTION

TRANSPORTATION electrification through the adoption of plug-in electric vehicles (PEVs) is gaining more popularity due to the increased awareness of its environmental and economic benefits. The number of PEVs around the globe exceeded 5.1 million in 2018, with an approximately 60% year-on-year growth rate [1]. Along with the growing number of PEVs, public PEV charging infrastructures (EVCIs) have become widely available to accommodate the raising PEV charging demand. EVCIs consist of various types of charging technologies, which can be either with a plug-in cable or wireless charging [2], [3]. Plug-in chargers are classified into three main charging levels, which are AC level 1, AC level 2, and DC fast charging [4]. Charging levels vary in terms of charging power capability; hence charging time of a PEV depends on the type of used charger. On the other hand, wireless chargers can dynamically charge PEVs on-roads through wireless power transfer technology [5].

The increasing PEV penetration rate comes with challenges. One challenge results from of the high PEV charging power, which introduces a substantial load on the power system. This load can negatively impact the power distribution system in various aspects, especially with voltage deviations and power losses [6]. Another challenge comes from the long charging duration of PEVs at charging facilities. As users spend a long time charging their PEV batteries at a charging facility, other PEV users need to wait for a longer time before getting a charging service at that facility. Also, PEV users can be blocked from a charging facility if the facility is at its full capacity. Long waiting times and a high blocking rate degrade PEV user satisfaction and hence charging quality-of-service (QoS).

To overcome the challenges, coordination of PEV charging demands has become indispensable to meet the PEV user

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. (e-mail: aabdalra@uwaterloo.ca; wzhuang@uwaterloo.ca). This work was financially supported by a research grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

requirements, while minimizing the negative impacts on the power grid. PEV charging demand management is classified into two main categories: centralized and decentralized (distributed) coordinations [7]. In the centralized coordination, a central controller directly controls the charging process of the participated PEVs. This approach can optimize the PEV charging process. However, it requires sophisticated communication and control systems to monitor and coordinate the charging process of a large number of PEVs. In the decentralized coordination, a dynamic pricing mechanism can be leveraged to coordinate the PEVs charging process and influence the behaviors of PEV users. The price-setting should simultaneously achieve the following objectives [8]: 1) to preserve the QoS of charging facilities to maximize users' satisfaction; 2) to alleviate the negative impacts on the power system; and 3) to maximize the utilization of charging facilities.

Recently, significant progress has been made in developing pricing schemes that account for the uncertainty of PEV charging demand, while considering the fluctuations in electricity price and power grid conditions. The preliminary pricing schemes do not deal with the multiple services offered by EVCIs and the multiple QoS classes associated with these services. In a multiservice EVCI, a set of charging services is offered, and each service is provided by a type of charging facility with a certain QoS class. Different service classes vary in the PEV charging rate, average waiting time, and charging method (i.e., either with a plug-in cable or wireless charging). Thereby, users' dwelling time during the PEV charging depends on the charging service class. Each service class has a minimum level of QoS that must be preserved during the operation to ensure the PEV user satisfaction and the profitability of charging service.

In addition to the mentioned challenges, two issues remain to be addressed in the contemporary modeling approach of dynamic pricing for PEV charging demand. First, existing dynamic pricing models rely on simplified assumptions and can be unrealistic, such as full knowledge of the customer-side information including the current charging demand and the influence of pricing decisions on future user behaviors. Even if the demand is modeled as a random variable, the assumption of complete information about the expected demand is unrealistic. Second, PEV charging demand coordination is a complex non-stationary and stochastic process. Typically, PEV charging demand can change over time-of-day, day-of-the-week, seasons, or due to an increased or decreased desirability of particular charging technology. Applying abstract dynamic pricing models in this environment cannot ensure optimality, as any change of variables can lead to model misspecification, resulting in unreliable estimation of the system operation and/or revenue loss. To overcome the limitations, reinforcement learning (RL) emerges as one of the most promising tools for the decision-making problem in an unknown environment. RL is capable of learning from the interactions with a dynamically changing environment. Moreover, RL can provide locally optimal solutions for complex and non-linear optimization problems without requiring a pre-specified model of the environment.

Driven by the challenges, we propose a new way to use deep RL algorithms in the context of dynamic pricing of charging services. The proposed pricing mechanism preserves different QoS classes at a multiservice EVCI. Additionally, the proposed approach is able to learn a pricing policy, while the complete customer-side information is not available. The major contributions of this study are as follows:

1) We propose a differentiated pricing mechanism that discourages over-utilization of a charging service, in addition to enhancing the performance of charging facilities in meeting the expectation of PEV users. Towards this end, the problem is formulated as a social welfare maximization problem, where the objective is to maximize the demand for charging services while maintaining the targeted QoS in all service classes;
2) The proposed framework is based on the twin-delayed deep deterministic policy gradient (TD3) algorithm, which is a model-free RL approach using actor-critic methods. In the proposed approach, deep neural networks are trained to determine a pricing policy while interacting with the unknown environment. The neural networks take the current EVCI state as input and generate pricing signals that coordinate the anticipated PEV charging demand.

The rest of this paper is organized as follows. The related work is discusses in Section II. Section III presents system model, along with a discussion of the PEV charging demand and charging facility models. We formulate the dynamic pricing problem in Section IV, and present an RL approach to solve this problem in Section V. Numerical results are given in Section V to evaluate the proposed framework. Finally, Section VI draws some conclusions from this study and presents a possible future extension for our framework.

II. Related work

This section provides an overview of the research areas relevant to this study, which is divided into two parts. The first part discusses dynamic pricing mechanisms for PEV charging demand coordination, while the second part presents a survey of adopting RL algorithms in optimizing the dynamic pricing policies.

A. *Dynamic Pricing for PEV Charging Demand Coordination*

Dynamic pricing is a decentralized coordination and load management method for the PEV charging demand. In literature, significant efforts have been devoted to the coordination of PEV charging demand through dynamic pricing. Dynamic pricing signals for PEV charging coordination can be derived based on three modeling approaches, namely, game-theoretic approaches, stochastic optimization methods, and queuing-based models.

Game theory-based dynamic pricing is widely employed to map the relation among multiple entities, where each entity maximizes its own profit [8]–[11]. For instance, the relation between charging station operator and PEV users can be modeled as a single-leader-multi-follower Stackelberg game [8]. The station operator is modeled as a leader whose main interest is to optimize the service price to maximize its profit with the same amount of energy resources. The PEV users are modeled as followers who maximize their own level of satisfaction by selecting a nearby charging station with low charging cost. Game theory is then used to derive a dynamic pricing scheme that balances the PEV load demand among the adjacent charging stations. Game theory-based approaches assume that there exists a high fidelity communication infrastructure, in which each charging facility can communicate with PEV users to influence their preferences.

Optimization models can be used to determine the dynamic pricing of the PEV charging service [12] [13]. For instance, stochastic dynamic programming can be used to determine charging prices [12]. This model considers multiple uncertainties, such as in PEV charging demand, the intermittency of renewable energy sources, and the electricity price fluctuation. Optimization-based models assume that the charging service provider can perform an online prediction of charging demand functions, which characterize customers response to price fluctuations.

Queuing models can be used to estimate the PEV charging demand and then derive a dynamic pricing expression to coordinate the PEV charging service [14] [15]. For instance, the impact of wireless charging load demand and PEV mobility on the location marginal price (LMP) of electricity is investigated in [14]. The BCMP queuing network model is used to estimate the charging demand of PEVs, considering the PEVs mobility. Then, a dynamic pricing scheme is optimized to adjust the retail price of wireless charging. Queuing-based models assume the availability of accurate statistics regarding PEV mobility patterns, energy consumption, and battery characteristics.

Based on the preceding discussion, the existing works study various aspects of the dynamic pricing of charging services. The existing pricing schemes treated all types of charging facilities equally. In practice, however, a PEV charging infrastructure includes various types of facilities such as on-road wireless chargers, fast charging stations, and slow chargers at parking lots. Each type of charging facility has a unique QoS capability and usage pattern. In order to provide differential QoS, a differential pricing is needed to discourage overutilization of a certain type of charging service. Thus, a differential pricing mechanism should be developed to set the price for each charging service, which offers the necessary incentives for PEV users to choose the charging service that matches their requirements. Consequently, the QoS levels of charging facilities are maintained.

B. Reinforcement Learning for Dynamic Pricing

RL is an area of machine learning, which deals with goal-directed learning based on the interaction between an active decision-making agent and its unknown environment, while the agent seeks to maximize a numerical reward signal [16]. Using deep RL algorithms, neural networks are trained off-line in a simulated environment. Then, the neural networks can be exploited on-line in the practical system.

Recently, there has been a collection of research works studying how to optimize dynamic pricing policies using reinforcement learning. For example, dynamic pricing of interdependent perishable products can be optimized using Q-learning [17]. Given an initial inventory for the products, this approach is to maximize the revenue by dynamically adjusting the pricing policy over a finite sale horizon when the demand function is stochastic and unknown. RL algorithms can be used in the context of demand response [18]–[21]. For instant, Q-learning can help to reduce supply-demand mismatches by dynamically deciding the retail electricity price, considering both the service provider profit and customers' cost [18]. In [19], deep Q-learning and deep policy gradient are used for on-line optimization of building energy consumption, where the objective is either to minimize the energy cost or to flatten the net energy profile.

Comparing with conventional dynamic pricing methods, RL-based methods offer two main advantages [21], [22]: 1) RL-based methods do not require strict models of the system. RL can provide a locally optimal solution to the decision problem when the full knowledge of the system is not available; 2) RL-based methods are capable of responding to a dynamically changing environment. When the operating state changes, RL can dynamically change the decision policy without requiring any additional information.

In this work, inspired by the recent research outcomes, we present a dynamic pricing algorithm for differentiated PEV charging services using deep RL.

III. System Model

Consider an EVCI with M classes of charging services managed by a charging service provider (CSP). Let $\mathbf{M} = \{1, 2, \ldots, M\}$ denote the available service class in the system, where $i < j$ indicates i is a higher service class than j (i.e., i has more strict QoS requirements than j). Each service class is specified by a minimum QoS level that is to be maintained all the time. The EVCI consists of a set charging facilities denoted by $\mathbf{N} = \{1, 2, \ldots, N\}$, which is composed of three subsets: subset $\mathbf{P}$ for parking lots (PLs) with AC chargers, subset $\mathbf{F}$ for DC fast-charging stations (FCSs), and subset $\mathbf{W}$ for on-road wireless chargers (OWCs). Each service class is offered by a number of charging facilities, where $n \rightarrow m$ indicates that charging facility n provides charging service of class m. The business hours of charging facilities can be partitioned into T time slots with equal duration Δt, where $\mathbf{T} = \{1, 2, \ldots, T\}$ denotes the set of all time slots over a business day.

Charging facility n ($n \in \mathbf{N}$) with service class m ($m \in \mathbf{M}$) provides a charging service with minimum QoS $q_m \in [0, 1]$ and associated with normalized charging price $p_{n,t} \in [0, 1]$ for all $t \in \mathbf{T}$. The CSP announces one-time (stationary) minimum QoS level of each service class at the beginning of the planning horizon, $\mathbf{q} = (q_1, \ldots, q_M)$. To maintain these QoS classes, the CSP adjusts the pricing policy of charging services periodically at the beginning of each time slot. Thereby, at time slot t, the CSP adopts a pricing policy represented in the price vector $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$.

PEV users compare prices, QoS, and other service attributes (e.g., location of the charging facility, type of charging technology) of all services and choose a charging service at one charging facility offering the particular service class. The offered charging services are substitutable because the demand for one service class not only depends on its own price but also depends on the prices of other services in the EVCI. Also, the charging service classes are vertically differentiated, which means that customers always prefer a higher service class if the charging prices are the same among different classes [23]. Thereby, any change in the price of one service can impact the demand for other services in the system.

A. PEV Charging Demand Model

PEV demand for charging services is assumed sensitive to both the charging service price and the QoS guarantees. This assumption is only used in the simulated environment for RL agent training. However, the proposed RL approach is adaptive and able to learn the actual PEV user behaviors based on the interaction with the environment. To represent the PEV charging demand at each charging facility, we use the linear demand model [24]. Based on this model, the time-varying charging demand (arrival rate), $\lambda_{n,t}$, at charging facility n in time slot t is a linear function of all prices and service levels of charging facilities, given by

$$\begin{aligned} \lambda_{n,t}(\mathbf{p},\mathbf{q}) &= \Lambda_{n,t} - \beta_n p_{n,t} + \sum_{\hat{n}\neq n} \beta_{n,\hat{n}} p_{\hat{n},t} + \gamma_m q_{m,t} \\ &- \sum_{\hat{m}\neq m} \gamma_{m,\hat{m}} q_{\hat{m},t}, \quad n \to m,\ \forall n,\hat{n} \in \mathbf{N},\ \forall m,\hat{m} \in \mathbf{M} \end{aligned} \tag{1}$$

where $\Lambda_{n,t}$ (> 0) denotes PEV arrival rate to charging facility n at time t at the base price, which is accepted by all users. This arrival rate can be estimated based on the traffic volume intercepted at the charging facility [2]; β_n and γ_m are positive parameters denoting self-elasticity, which indicate the relative change in the demand for a charging service that would result from a change in the service price and quality, respectively; $\beta_{n,\hat{n}}$ and $\gamma_{m,\hat{m}}$ are positive parameters representing cross-elasticity, which reflect the change in the service demand as a result of the change in the prices and service levels of other charging facilities, respectively. The demand for charging service at facility n should be more sensitive to its own price changes than those for the other services. Thus, the elasticity parameters have the relation $\beta_n > \sum_{\hat{n}\neq n} \beta_{n,\hat{n}}$. The demand function is assumed to satisfy the monotonicity properties [25], as follows

$$\begin{aligned} \frac{\partial \lambda_{n,t}(\mathbf{p},\mathbf{q})}{\partial p_{n,t}} \leq 0,\ \frac{\partial \lambda_{n,t}(\mathbf{p},\mathbf{q})}{\partial p_{\hat{n},t}} \geq 0,\ \frac{\partial \lambda_{n,t}(\mathbf{p},\mathbf{q})}{\partial q_{m,t}} \geq 0, \\ \frac{\partial \lambda_{n,t}(\mathbf{p},\mathbf{q})}{\partial q_{\hat{m},t}} \leq 0, \quad \forall n,\hat{n} \in \mathbf{N}. \end{aligned} \tag{2}$$

The assumption means that, if the CSP increases $p_{n,t}$ (or decreases q_m), the demand for charging service at facility n decreases; however, if the CSP increases $p_{\hat{n},t}$ (or decreases $q_{\hat{m},t}$), the demand at charging facility n increases. Note that PEV users' demand at facility n for service class m is assumed depending on the announced QoS level of service class q_m rather than its actual service level in the facility.

B. Charging Station Model

The capacity of any charging facility in the EVCI is finite. Thereby, charging facilities can be modeled as a finite queueing system [3], where arriving PEV users are rejected (blocked) at times when the charging facility is full. Also, PEV users may wait until service becomes available if all chargers are busy and there is waiting space available. PEVs arrivals to a charging facility follow a non-homogeneous Poisson arrival process, which is a non-stationary counting process with a deterministic arrival rate $\lambda_{n,t}$ [3]. As discussed in Subsection III-A, this arrival rate depends on various factors such as traffic volume and users' response to changes in $\mathbf{p}$ and $\mathbf{q}$. The charging time of PEVs at a charging facility is assumed independently and exponentially distributed, with service rate μ that depends according to the chargers' power capability at the charging facility [2]. Each charging facility has c_n independent and identical chargers (servers) that serve PEV users on the first-come-first-served rule. The maximum number of PEVs that can be admitted to charging facility n is denoted by K_n.

Let $\mathcal{N}_{t,n}$ denotes the average number of admitted PEVs to charging facility n ($n \in \mathbf{N}$) at time t, which reflects the facility utilization. The number of admitted PEVs can be expressed as

$$0 \leq \mathcal{N}_{t,n} \leq K_n. \tag{3}$$

Due to the finite capacity of charging facilities, high congestion (overload) may occur at a facility if the number of admitted PEVs approaches the maximum facility capacity. Then, the newly arrived PEVs may suffer from a low QoS level in terms of long waiting time or service rejection (blocking). Thereby, we define QoS metric $\mathcal{Q}_{n,t} \in [0,1]$ to measure the service performance at charging facility n at time t, which is related to the weighted sum of the blocking and delay rates. We define the QoS metric as

$$\mathcal{Q}_{t,n} = 1 - \alpha \mathbb{P}_{K_{t,n}} - \beta \mathbb{P}\{W_{t,n} > 0\} \tag{4}$$

where α and β are weighting factors to reflect the impact of service blocking and service delay respectively on the satisfaction level of PEV users, with $\alpha + \beta = 1$; $\mathbb{P}_{K_{t,n}}$ and $\mathbb{P}\{W_{t,n} > 0\}$ denote the blocking and delay probabilities of the charging facility, respectively. Both $\mathcal{N}_{t,n}$ and $\mathcal{Q}_{n,t}$ represent the state of the charging facility in terms of the facility utilization and service quality.

IV. Dynamic Pricing for Differentiated PEV Charging Services

Price-based coordination mechanisms can be leveraged to dynamically meet the QoS requirements of charging facilities, while maximizing social welfare which is the sum of utilities over all users and service providers. In a multiservice charging infrastructure, a differentiated pricing scheme is required to provide different QoS classes. Thus, a PEV user has incentives to use a charging service that matches their needs. The differentiated pricing scheme can enhance the performance of charging facilities in meeting the expectation of PEV users by discouraging the over-utilization of some charging services.

Our objective in this research is to develop a differentiated dynamic pricing scheme that accounts for the interactions between two players, which are the CSP and the non-cooperative PEV users. On one hand, PEV users want to choose the charging service that maximizes their utilities, when making the charging decisions. The utilities depend on various random variables, including the current state-of-charge (SoC) of PEV battery, the charging price, and the user's value of time, which indicates how much a user appreciates time-saving during the charging process. On the other hand, the CSP wants to maximize revenue while achieving targeted QoS. Towards this goal, the CSP aims to shape charging demand profiles by incentivizing PEV users to behave in ways that improve the overall utilization and performance of the EVCI. The most effective incentive comes in the form of lower service prices charged to PEV users. The CSP can offer lower charging prices at a less congested facility to incentivize PEV users to be rerouted to that facility. Thus, using a differentiated pricing policy, the CSP dynamically adjusts the charging prices for all facilities based on the current and anticipated demand patterns.

Choosing the *right* price for a PEV charging service is challenging. Determining a pricing policy requires information not only about how much the current PEV user values each charging service but also about what the future demand will be. In order to develop the differentiated pricing scheme, some assumptions are necessary: 1) The PEV charging infrastructure includes a limited set of service classes, and each service is offered by a set of charging facilities as discussed in our system model; 2) There is no competition between charging facilities with different QoS classes; 3) PEV charging demand is elastic and price-responsive; 4) The CSP periodically collects performance statistics regarding the utilization and service quality from all charging facilities in the EVCI; 5) PEV users are informed about the real-time service prices at all charging facilities in the EVCI, which can be facilitated by web applications offered to PEV users; 6) The targeted minimum QoS level of each class of charging services is predetermined based on user preferences, which can be collected from market surveys.

The differentiated pricing policy can be represented as a social welfare maximization problem, where the objective is to maximize the utilization of charging facilities while maintaining the minimum targeted QoS in all service classes. The differentiated pricing problem can be formulated as

$$\max_{\mathbf{p}} \sum_{t=1}^{T}\sum_{n=1}^{N} \lambda_{n,t}(\mathbf{p},\mathbf{q}) \tag{5a}$$

$$s.t.\ \mathcal{Q}_{t,n} \ \le \ q_m, \quad n \rightarrow m, \forall n \in \mathbf{N},\ \forall m \in \mathbf{M},\ \forall t \in \mathbf{T}. \tag{5b}$$

Optimizing this pricing policy must explicitly incorporate the stochasticity and non-stationarity of the PEV charging demand. Also, the pricing policy must be forward-looking by setting the price signals in anticipation of future demand patterns. As indicated in (5), determining the pricing policy requires full knowledge of the customer-side information including the current charging demand and the influence of pricing decisions on future user behaviors. However, due to the lack of complete information and the non-stationarity of system variables, conventional abstract models cannot be adopted to determine the pricing policy. Thereby, we propose an RL approach to learn the pricing mechanism. Based on the proposed approach, the CSP can adjust the pricing signals in real-time without requiring a pre-specified model of the environment.

V. RL Approach for Differentiated Pricing

Determining a differentiated pricing policy is a real-time decision-making problem in an unknown environment. Here, we present an RL approach to decide the pricing policy based on learning while interacting with the environment. Towards this goal, the differentiated pricing problem is firstly formulated as a discrete finite-horizon Markov decision process (MDP) [16]. Then, the twin delayed deep deterministic policy gradient (TD3) algorithm is used to train neural networks that generate the pricing policy, without requiring the full knowledge of system dynamics and uncertainties. Finally, we present the implementation details of the TD3 algorithm along with the associated hyperparameter, neural network architectures, and the reward function design.

A. Markov Decision Process

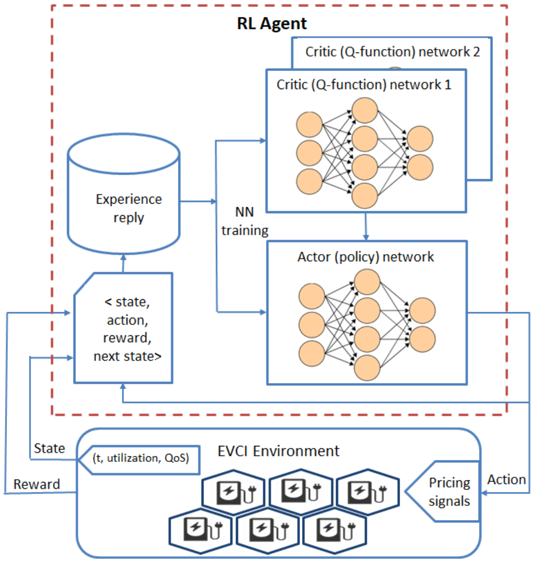


Fig. 1: MDP framework for differentiated pricing.

As shown in Fig. 1, MDP for the differentiated pricing problem is a formalization of the interaction between an agent and the environment [16]. The agent is the learner and decision-maker that selects actions. The environment responds to the agent's actions, presents new situations to the agent, and gives a numerical value to the agent as a reward to evaluate the agent's actions. The MDP is defined by the following key components:

- A set of states, $\mathcal{S}$, that reflects the current state of the EVCI. At each discrete time step t, system state $s_t \in \mathcal{S}$ is denoted as $s_t = (t, \mathcal{N}_{t,1}, \mathcal{Q}_{t,1}, \ldots, \mathcal{N}_{t,N}, \mathcal{Q}_{t,N})$. As discussed in Subsection III-B, the system state represents the utilization and QoS of all charging facilities in the EVCI;
- A set of actions, $\mathcal{A}$, that is selected by the agent based on the current system state and the anticipated future PEV

charging demand. The selected actions affect the charging demand on the next time slot. At discrete time t, the agent selects an action, $a_t \in \mathcal{A}$, based on its policy $\pi : \mathcal{S} \mapsto \mathcal{A}$. This action is a vector of length equal to N with elements normalized to the range $[-1, 1]$. The CSP maps this action into pricing vector $\mathbf{p}_{t+1}$ that sets the charging price (money value) for all facilities at time slot $t + 1$;

- State-transition probabilities $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, where $p(s_{t+1}|s_t, a_t)$ indicates the likelihood that action a_t results in the transition from state s_t to next state s_{t+1};
- Reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, where $r(s_t, a_t)$ computes an immediate reward signal for executing action a_t when the system is at state s_t. The agent's goal is to maximize not only the immediate rewards but also the cumulative discounted future rewards $R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$, where $\gamma \in [0, 1)$ is a discount factor indicating the priority of short-term rewards.

One episode of the MDP forms a finite sequence or a trajectory in the form $(s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$. Determining optimal policy π^* can be done using the action-value function $Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, which is defined as the discounted expected total reward when taking action a_t at state s_t and thereafter following policy π [16]. The Q-function can be formulated as

$$Q^\pi(s, a) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a \right]. \tag{6}$$

The optimal Q-function for each state-action pair is $Q^*(s, a) = \max_\pi Q^\pi(s, a)$, and the optimal policy that returns the highest valued action can be obtained by [26]

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a). \tag{7}$$

B. Adopting TD3 Algorithm for Differentiated Pricing

Optimizing a policy based on the MDP can be done using either policy iteration or value iteration if system transition dynamics (probabilities) are known [20]. However, system dynamics are unknown and need to be estimated through interactions with the environment. RL can adopt the model-free approach, in which the RL agent learns to optimize an action for each state without requiring a complete and perfect model of the environment. One of the most common model-free algorithms in RL is Q-learning, which uses a table to store and update Q values while exploring the environment. However, Q-learning is only applicable when the action space is finite and discrete [27]. In the context of differentiated pricing for charging services, both the state and action spaces are continuous, and discretization of the states and actions introduces a dimensionality problem [20]. Instead, here we resort to the TD3 algorithm [28], which is a model-free off-policy actor-critic algorithm that can learn policies in a high-dimensional continuous action space.

The TD3 algorithm builds on the deterministic policy gradient (DPG) algorithm, along with deep neural networks. Different from Q-learning, Q-table is replaced by deep neural networks (NN) that act as a function estimator, which achieves better generalization for continuous state and action spaces via deriving unknown correlations from previous experience. However, NN training can be unstable, if the used RL-algorithm continuously overestimates Q-values. These estimation errors build up over time and can lead to suboptimal policy updates and divergent behavior. TD3 addresses this issue by reducing the overestimation bias problem of Q-values. As a result, TD3 achieves higher performance and improves the learning speed when compared with other deep RL algorithms [28]. TD3 belongs to the actor-critic methods, which directly optimize policy $\pi(s)$ in addition to learning Q-function $Q^\pi(s, a)$. Policy optimization, known as the actor, directly maps the states to actions. The Q-function, known as the critic, assigns a value that evaluates the policy's action given the system state and the selected action.

As shown in Fig. 1, the TD3 architecture consists of two critic (Q-value) networks and one actor (policy) network. Each NN is characterized by a set of parameters that consist of the NN weights and biases. The parameters of the critic networks are denoted by ϕ_1 and ϕ_2, and the parameters of the actor network are denoted by vector θ. Since learning of NN can be unstable, target networks are needed to slowly keep track of the updates in the (online) critic and actor networks. Thereby, TD3 uses two target critic networks with parameters $\grave{\phi}_1$ and $\grave{\phi}_2$, and a target actor network with parameters $\grave{\theta}$. During learning, the RL agent collects and stores a set of transitions in experience replay buffer $\mathcal{R}$. Each transition has the form of 4-tuple $(s, a, r, \grave{s})$, which denotes state, action, reward, and next state, respectively. Then, a mini-batch is uniformly sampled at each step to train the actor and critic networks. Training NN using mini-batches ensures that the selected samples are independently and identically distributed, which in turn facilitates an efficient optimization of NN parameters.

The TD3 concurrently updates two critic networks, Q_{ϕ_1} and Q_{ϕ_2}, using the recursive Bellman equation

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1})). \tag{8}$$

To approximate the optimal Q-function, the mean-squared Bellman error function is utilized to indicate how closely Q_{ϕ_1} and Q_{ϕ_2} satisfy the Bellman equation, as follow

$$L(\phi_i, \mathcal{R}) = \mathbb{E}_{(s,a,r,\grave{s}) \backsim \mathcal{R}} \left[\left(Q_{\phi_i}(s, a) - y(r, \grave{s}) \right)^2 \right], \; i = 1, 2 \tag{9a}$$

$$y(r, \grave{s}) = r(s, a) + \gamma \min_{i=1,2} Q_{\grave{\phi}_i}(\grave{s}, \pi_{\grave{\theta}}(\grave{s}) + \epsilon). \tag{9b}$$

Note that, in (9b), the TD3 uses the smallest of the two Q-values to form the targets in the Bellman error loss functions. This practice, which is called clipped double Q-learning, helps in reducing the overestimation bias problem of Q-values [28]. Also, the target policy is smoothed by adding a small Gaussian noise component, ϵ, to the target action, which prevents over-fitting on the narrow peaks of Q-values. Optimizing the policy can be done by training the actor network to give the action that maximizes the expected Q-function as

$$\max_\theta = \mathbb{E}_{s \backsim \mathcal{R}} \left[Q_{\phi_1}(s, \pi_\theta(s)) \right]. \tag{10}$$

The parameters of the actor network are updated through the gradient ascent of the expected return $\nabla_\theta J(\theta)$ with respect to actor parameters only, as given by [29]

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \backsim \mathcal{R}} \left[\nabla_a Q_{\phi_1}(s, a)|_{a=\pi(s)} \nabla_\theta \pi_\theta(s) \right]. \tag{11}$$

To improve training stability and reduce the accumulation of errors resulting from temporal difference learning, the actor network is updated once every two updates of the critic networks [28]. The stability of NN learning is also improved by adopting soft target update [27], in which the parameters of the target networks are slowly updated to track the changes in the online actor and critic networks by some portion τ, as follows

$$\grave{\phi}_i \leftarrow \tau\phi_i + (1-\tau)\grave{\phi}_i, \quad i = 1,2 \tag{12a}$$

$$\grave{\theta} \leftarrow \tau\theta + (1-\tau)\grave{\theta}. \tag{12b}$$

To remove the dependence on the randomly initialized parameters of NN, actions are sampled uniformly from the action space (pure exploratory policy) for limited time steps at the beginning of the training process. Subsequently, the TD3 starts exploiting what NN learned and exploring the environment by adding an uncorrelated mean-zero Gaussian noise to the selected action. The added noise is clipped to conform with the action space bounds. It is proved, in [28], that the clipped double Q-learning approach converges in the finite MDP setting. As with most deep RL algorithms, the convergence of TD3 is not guaranteed. This is because TD3 uses NNs as non-linear function approximators, which nullify any convergence guarantees [27]; however, experimental results on various test environments demonstrate that the TD3 algorithm converges without the need for any modifications [28]. Algorithm 1 summarizes the TD3 algorithm as adopted in our application.

Algorithm 1 TD3 for differentiated pricing

```
Randomly initialize critic networks parameters $\phi_1$,$\phi_2$, and actor network parameters $\theta$
Set the target NN parameters equal the online NN parameters
  $\grave{\phi}_1 \leftarrow \phi_1$, $\grave{\phi}_2 \leftarrow \phi_2$, $\grave{\theta} \leftarrow \theta$
Initialize empty replay buffer $\mathcal{R}$ with size 100k
for Episode $= 1, 15k$ do
    Receive initial observation state $s$
    for $t = 1, T$ do
        if Episode $\leq 100$ then
            Randomly select actions $a = \mathcal{U}(-1, 1)$
        else
            Select actions according to the current policy and add exploration noise $a \leftarrow \pi_\theta(s) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1)$
        end if
        Execute action $a$, observe reward $r$ and next state $\grave{s}$
        Store transition tuple $(s, a, r, \grave{s})$ in $\mathcal{R}$
        Randomly sample a mini-batch of $N$ transitions from $\mathcal{R}$
        Compute greedy actions for next states using target actor network and add clipped Gaussian noise
        $\grave{a} \leftarrow \pi_{\grave{\theta}}(\grave{s}) + \epsilon$, where $\epsilon \sim \texttt{clip}(\mathcal{N}(0, 0.2), -0.5, 0.5)$
        Compute targets $y \leftarrow r(s, a) + \gamma \min_{i=1,2} Q_{\grave{\phi}_i}(\grave{s}, \grave{a}))$
        Update critic networks parameters using gradient descent
        $\phi_i \leftarrow \arg\min_{\phi_i} \frac{1}{N} \sum_{(s,a,r,\grave{s}) \backsim N} (y - Q_{\phi_i}(s, a))^2$, $i = 1, 2$
        if Episode mod $2 = 0$ then
            Update actor network parameters using gradient ascent
            $\nabla_\theta J(\theta) = \frac{1}{N} \sum_{s \backsim N} \Big[ \nabla_a Q_{\phi_1}(s, a)|_{a=\pi(s)} \nabla_\theta \pi_\theta(s) \Big]$.
            Update target networks
            $\grave{\phi}_i \leftarrow \tau\phi_i + (1-\tau)\grave{\phi}_i, \quad i = 1, 2$
            $\grave{\theta} \leftarrow \tau\theta + (1-\tau)\grave{\theta}$
        end if
    end for
end for
```

C. Implementation Details

1) States and actions: As described in Section III and Subsection V-A, the system state (observation space) is a combination of state variables that include time-of-day t, and the utilization and current QoS of each charging facility in the EVCI. The agent's action represents continuous control of the charging service prices in all charging facilities. At each time slot t, the RL agent generates a pricing action, a_t, based on the current state of the EVCI, s_t, and the expected future charging demand. The selected pricing action, a_t, influences the behaviors of PEV users in the next time slot $t + 1$. The length of the state tuple is $2N + 1$, and all elements in the state tuple are normalized to range $[0, 1]$. The action space is a tuple of size N, and all elements in the tuple are normalized to range $[-1, 1]$. Normalization of action and observation spaces facilitates the convergence of the TD3 algorithm.

2) Network architecture and hyper-parameters: The TD3 has a pair of critic networks along with a single actor network. Each neural network consists of two fully-connected hidden layers with 400 and 300 units, respectively. Rectified linear (`ReLU`) activation units are used for all hidden units. For the critic networks, the size of input layers is equal to the sum of the observation space size and the action space size. Critic network outputs consist of a single linear unit per network, representing the Q-value. For the actor network, the input layer size is equal to the observation space size, and the output layer size is equal to the action space size. Actor network output consists of `tanh` activation units.

Adam optimizer [30] is used to optimize the parameters of actor and critic networks, with a learning rate of 10^{-4} and 10^{-3} for actor and critic networks, respectively. The mini-batch size is chosen to be 64, and the experience replay memory can hold up to 10^5 state transitions. We use a discount factor of $\gamma = 0.99$, and a soft target update factor of $\tau = 0.005$. For policy exploration, we use uncorrelated additive Gaussian action space noise $\mathcal{N}(0, 0.1)$ with zero mean and 0.1 standard deviation. Target policy is smoothed by adding Gaussian noise $\mathcal{N}(0, 0.2)$, clipped to $(-0.5, 0.5)$, to the selected action from the target network.

3) Reward function: The performance of RL algorithms is highly impacted by the reward function. To achieve the desired behavior, a reward function must be designed in a way to guide the agent towards the goal. Reward functions can be designed to follow two main forms: sparse reward and shaped reward [31]. In sparse reward functions, the agent is given a positive reward if it achieves the desired goal and zero rewards otherwise. Although sparse reward functions are easy to design for most of the tasks, it does not motivate the RL agent to learn and may need a lot of training to converge to an acceptable policy. To motivate the agent's learning, reward shaping is usually used to give more rewards to the agent in the states that are closer to the target state. Shaped reward functions are difficult to design. This is because a shaped reward function can bias learning towards undesirable behaviors if it is not carefully designed. To achieve the proposed objective in (5), we design the following reward function

$$r_t(s_t, a_t) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{N}_{t,n} \prod_{n=1}^{N} \mathcal{C}_{t,n} \tag{13a}$$

$$\mathcal{C}_{t,n} = \begin{cases} \exp\left[-300(q_m + \zeta - \mathcal{Q}_{t,n})^2\right] & \text{, if } q_m + \zeta < \mathcal{Q}_{t,n} \\ 1 & \text{, otherwise.} \end{cases} \tag{13b}$$

The reward function consists of two parts representing the differentiated pricing problem in (5). The first part imitates the objective function in (5a) and incentivizes the RL agent to increase facility utilization. As $\mathcal{N}_{t,n}$ increases, the RL agent receives more rewards. The second part mimics constraint (5b), where $\mathcal{C}_{t,n}$ penalizes the reward signal if the current QoS at a charging facility is less than the targeted QoS. To encourage learning of a pricing policy that guarantees the targeted QoS all the time, a safeguard constant $\zeta = 2\%$ is added to the targeted QoS.

4) Training and validation processes: The training proceeds episodically for 15×10^3 episodes. Each episode represents one business day that is simulated from an initial state at time $t = 1$ until the end of the daily time horizon at $t = T$. To prevent overfitting and to achieve good generalization, neural networks are trained on simulated environments that vary in each training run. Each simulated environment differs in the random seeds that is sampled uniformly from 50 random seeds. The performance of the training process is evaluated using daily cumulative rewards, which is the total rewards that the RL agent receives over a day. As expected, the cumulative reward rapidly increases at the beginning of training, then increases at a much slower rate as the training goes on. During training, the pricing policy is evaluated periodically without exploration noise. Different from the training process, the validation process always uses an environment with the same random seed, which is different from the training seeds. The neural network that achieves the best performance (maximum cumulative reward) in the evaluation process is selected to form the final pricing policy.

VI. Numerical Results

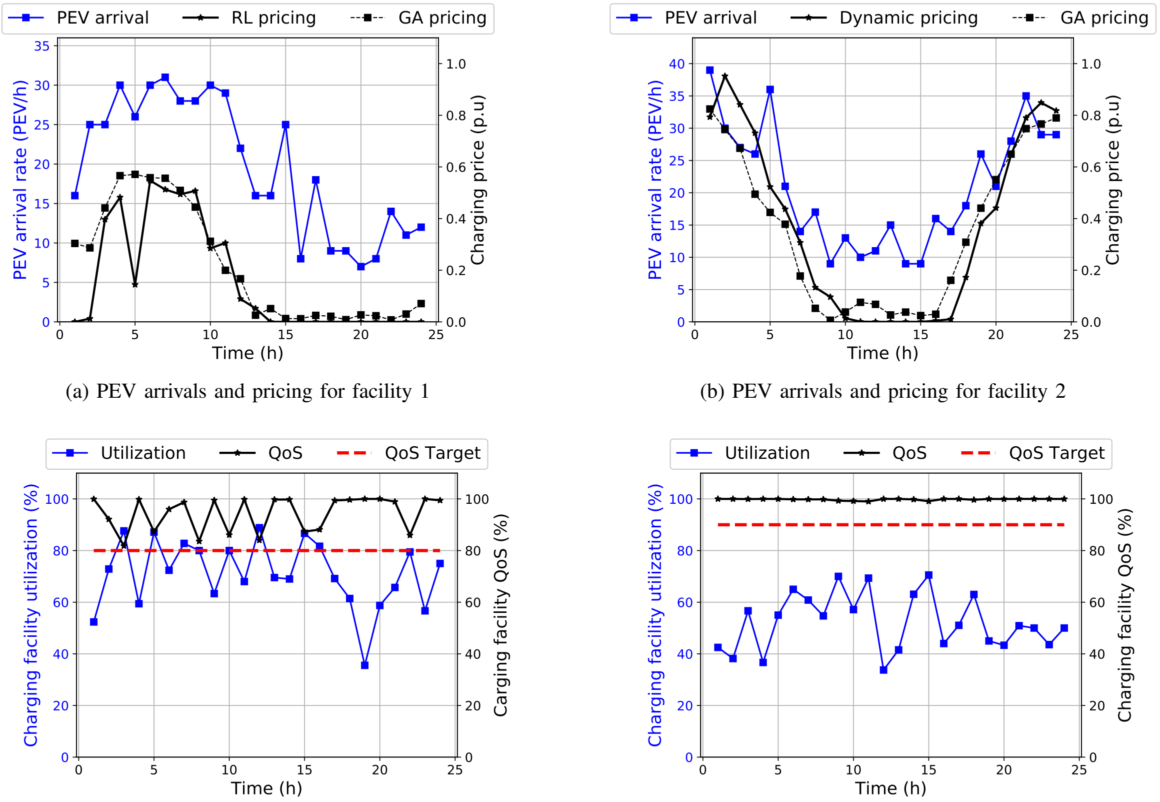
In this section, we evaluate our proposed differentiated pricing framework and demonstrate the applicability of the TD3 algorithm in optimizing dynamic pricing policies. Firstly, we present a numerical example that highlights the relationship among the pricing signals, and charging facility utilization and QoS. Then, we demonstrate the scalability of our proposed framework on a larger EVCI, with more realistic architecture and demand properties. The daily time horizon is divided into 24 time-slots each of which lasts for one hour. The CSP determines a dynamic pricing policy for each charging facility to maximize the utilization of charging facilities, while maintaining the targeted QoS. The CSP does not make any assumptions regarding the PEV charging demand or the impact of pricing signals on PEV user behaviors (beyond the assumptions made in Section IV). In the numerical examples, the RL agent selects pricing actions to impact charging demand in a simulated EVCI environment. The simulated environment responds to the agent's actions by presenting the new state of the environment, in addition to a numerical reward signal that evaluates the agent's actions. Thereby, the RL agent does not have any predefined model of the environment. Instead, NNs are trained to optimize the pricing decisions based on interactions with the environment. The simulations are implemented under Python 3.7 environment on a laptop computer with a 2.3-GHz Intel(R) Core(TM) i5-8300H CPU, 8 GB of memory, and NVIDIA GeForce GTX 1050 Ti GPU unit.

A. Example 1: Two Charging Facilities

In this example, the CSP finds a pricing policy for two charging facilities. For clarity of illustration, PEV charging demands in these facilities are assumed sensitive only to the pricing signals. The numbers of chargers and waiting positions in both facilities are the same, with seven chargers and three waiting-positions. The service rate in both charging facilities equals to 3 PEV/h. As discussed in Subsection III-B, PEV arrivals to charging facilities are modeled as nonhomogeneous Poisson process, and the arrival rates are given by $\Lambda_{1,t} = 21 + 10\sin(2\pi t/24)$ and $\Lambda_{2,t} = 21 + 10\cos(2\pi t/24)$. For the two facilities, the charging price is normalized to range $[0, 1]$, where zero represents the base price and one is the maximum allowable charging price. The CSP objective is to optimize a differentiated pricing policy that maintains two QoS classes, with PEV charging service completion targets at 80% and 90% for the first and second charging facilities, respectively.

To evaluate the efficacy of the proposed approach, we use a genetic algorithm (GA) as a benchmark to solve the optimization problem in (5). GA is a heuristic search algorithm that finds a high-quality solution to the optimization problems by randomly evaluating several points in the search space. GA obtains a population of the evaluated points and converges to a local optimum by relying on biologically inspired operators such as mutation, crossover, and selection [32], [33]. In the GA-based solution, the CSP is assumed to have complete information regarding customer-side information and the influence of price signals on the future EV charging demand. Moreover, the GA uses a fitness function similar to the reward function in (13). We use the normalized cross-correlation (NCC) coefficient to evaluate the degree of similarity between the RL-based pricing and the GA-based solution. The NCC is confined to the range $[-1, 1]$, where 1 indicates a perfect correlation and -1 indicates anticorrelation [34].

We first consider an independent demand scenario, where PEV charging demand in a facility is only dependent on the charging price at that facility. In this scenario, self-elasticity and cross-elasticity parameters are set to $\beta_1 = \beta_2 = 20$ and $\beta_{1,2} = \beta_{2,1} = 0$, respectively. The simulated environment is represented by two finite queuing systems, as described in Section III. For each facility, a QoS index is calculated based on (4), with parameters $\alpha = 0.75$ and $\beta = 0.25$. The reward function is given in (13), with $\zeta = 0.02$. Figs. 2a and 2b show the pricing policy based on both our RL approach and the GA solution. For both charging facilities, the RL agent optimizes pricing policies that anticipate the nonstationary stochastic charging demand. When evaluating the similarity between the



(a) PEV arrivals and pricing for facility 1

(b) PEV arrivals and pricing for facility 2

(c) Facility 1 utilization and QoS

(d) Facility 2 utilization and QoS

Fig. 2: Pricing policy for independent demand scenario

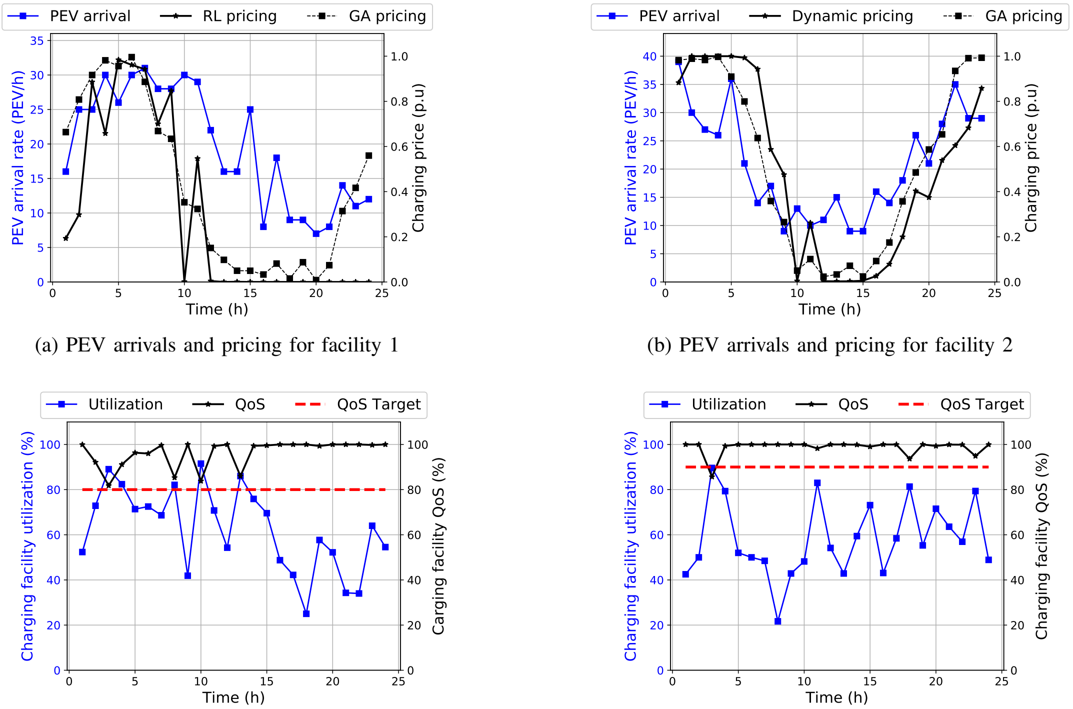
RL-based pricing signals and the GA-based pricing, we can find NCC coefficients of 0.92 and 0.99 for facility 1 and facility 2, respectively. The impact of RL pricing policy on the performance of charging facilities is evaluated in Figs. 2c and 2d, where the resulting facility utilization and QoS are shown. It can be noted that the generated pricing policy preserves the targeted QoS in both facilities. Furthermore, facility 1 has a higher utilization than facility 2, because the targeted QoS in facility 1 is less than facility 2.

Another scenario is considered, where PEV charging demand in the charging facilities are interdependent. Thereby, the charging demand at one facility not only depends on its price but also on the charging price in the other facility. Self-elasticity and cross-elasticity parameters are set to $\beta_1 = \beta_2 = 20$ and $\beta_{1,2} = \beta_{2,1} = 10$, respectively. As shown in Figs. 3a and 3b, the RL-based pricing policies are adapted to address the interdependence in charging demand between the two facilities. In Fig. 3b, it can be noted that the charging price of facility 2 is higher than that in Fig. 2b in response to the increasing demand in facility 1. To evaluate the similarity between the RL-based pricing and the GA solution, NCC coefficients are observed to be 0.90 and 0.98 for facility 1 and facility 2, respectively. The generated RL pricing policy is capable of maximizing the facility utilization while maintaining the targeted QoS levels, as shown in Figs. 3c and 3d. Thus, our RL-based pricing can adjust the pricing policies automatically according to the changes in customer behaviors without any preliminary settings.

It is observed in Fig. 3d that the achieved QoS level is slightly less than the targeted QoS level. This is because the RL algorithm is tested on a stochastic environment, which in turn can be in a state that is far away from the long-term stochastic average. Throughout the RL training, the NNs are optimized to generate a general policy that can always select locally optimal actions. However, the selection of optimal actions does not necessarily hold for any arbitrary state in the stochastic environment, which can exhibit unusual behaviors, making the selection of optimal actions problematic.

The convergence of the TD3 algorithm is shown in Fig. 4. At the beginning of the training, the RL agent explores the environment by sampling the selected actions uniformly from the action space. The pure exploratory policy removes the dependence on the randomly initialized NN parameters. Then, the RL agent exploits what NNs learned from interactions with the environment by adding an uncorrelated Gaussian noise to the selected NN action. As shown in Fig. 4, as the iteration goes by, the cumulative reward gradually increases and converges.

This example demonstrates that the RL approach can learn and optimize the dynamic pricing strategies, while interacting with an unknown environment. Based on the proposed approach, deep neural networks are trained to receive the current state of the EVCI and generate pricing signals that coordinate the future PEV charging demand.



(a) PEV arrivals and pricing for facility 1

(b) PEV arrivals and pricing for facility 2

(c) Facility 1 utilization and QoS

(d) Facility 2 utilization and QoS

Fig. 3: Pricing policy for interdependent demand scenario

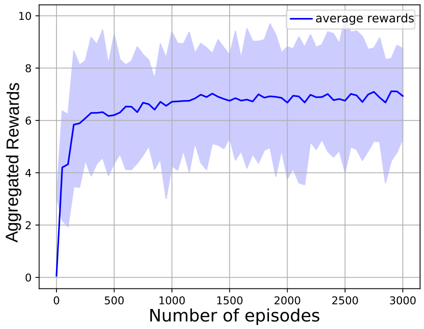


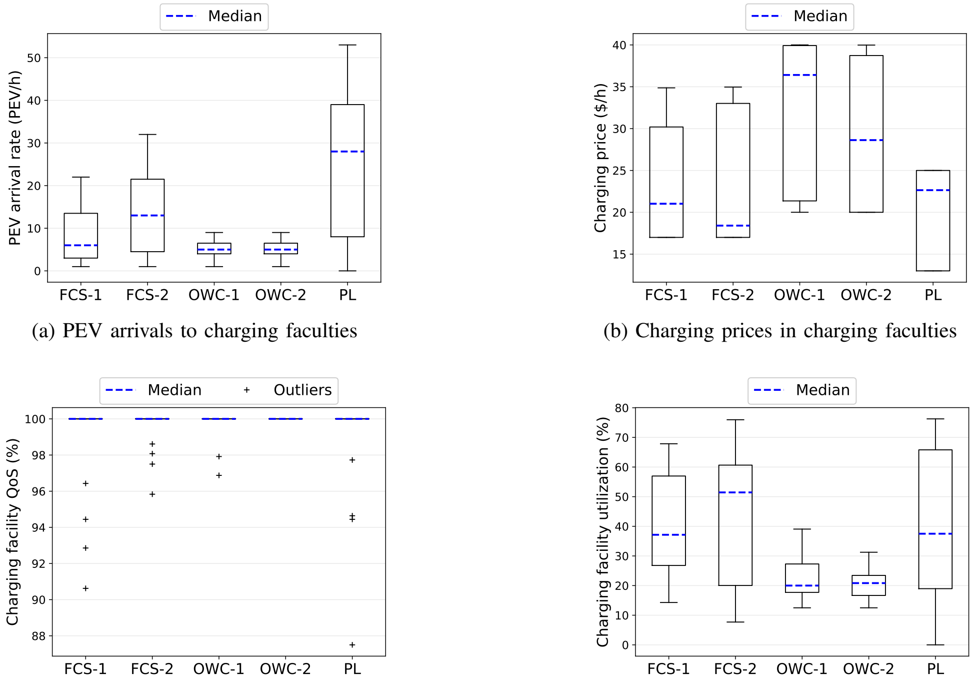
Fig. 4: Convergence process of TD3 algorithm

B. Example 2: Large-Scale Scenario

To validate our approach on a relatively larger EVCI, with more realistic charging demand properties, we select the well-known Nguyen-Dupuis road transportation network [35]. The topology and traffic demand attributes of the Nguyen-Dupuis network are introduced in [2]. We consider EVCI that consists of three service classes:

- The first service class is offered by two OWCs, called OWC-1 and OWC-2, which are located at links 5 and 16, respectively. The targeted QoS for the first class is 95%, and 4 chargers are allocated in each OWC. The mean charging time in OWCs is set to 15 minutes. Usually, charging service price is set as an hourly rate, which is billed by the minute based on the usage time and the type of charging service [36]. The charging price for the first class service is set to range [20, 40] \$/h, with self-elasticity parameter setting to 0.15;
- The second service class is offered by two FCSs, with targeted QoS levels set at 90%. One FCS, called FCS-1, is placed at node 10, and contains 5 chargers and 2 waiting positions. The second FCS, called FCS-2, is located at node 12, and contains 11 chargers and 2 waiting positions. The mean charging time in FCSs is set to 30 minutes. The charging price for the second class service is set to range [17, 35] \$/h, with self-elasticity parameter set to 0.6;
- The third service class is offered by one PL, located at node 3. The targeted QoS for PL is 85%. The PL contains 30 chargers and 10 waiting positions. The mean charging time in PL is set to 3 hours. Charging price for the third class service is set to range [13, 25] \$/h, with self-elasticity parameter set to 2.5.

Service quality elasticity parameters, γ_m and $\gamma_{m,\hat{m}}$, for all service classes are set to 3 and 1, respectively. We implement our RL approach to determine the charging prices for each facility in the EVCI. For compactness, we utilize box-plots to visualize the variability in the pricing policy, i.e., the distribution of the EVCI prices and performance over a day are visualized instead of the time-varying visualization. Box-plots provide a simplified approach for ease of performance comparison among multiple charging facilities. Fig. 5a shows the PEV arrival distributions to charging facilities, which represent the demand for the charging services. Fig. 5b shows the pricing signal distribution for each charging facility. It can be noted that charging prices change in the whole specified price range, and the median prices (shown in dashed line) for OWCs are higher than that in FCSs and PL. As shown in Fig. 5c, the pricing policy achieves a 100% QoS level for all service classes most of the time, except for a few outliers



(a) PEV arrivals to charging faculties

(b) Charging prices in charging faculties

(c) Achieved QoS level in charging facilities

(d) Achieved utilization in charging faculties

Fig. 5: Pricing policy for EVCI

of low QoS (shown in + sign). However, for each charging facility, the lowest achieved QoS level is higher than the specified QoS target. Thus, our proposed approach preserves the targeted QoS for multiple service classes without requiring a strict model of PEV charging demand. Fig. 5d shows that the utilization at OWCs is less than that of both FCSs and PL. This is because the targeted QoS at OWCs is higher than that in FCSs and PL. The generated pricing policy offers incentives in the form of low charging prices that motivate PEV users to use a charging service matching their needs. Furthermore, our pricing mechanism increases the prices to discourage over-utilization of charging services with high QoS targets.

The simulation results indicate that the proposed approach dynamically decides the pricing signals for a multiservice EVCI based on the targeted QoS and usage patterns of each service class. Our differentiated pricing mechanism provides insight into how to make a trade-off between facility utilization and service QoS. Offering lower service prices at charging facilities incentivizes more PEV users to be admitted for service and therefore increases facility utilization, at the cost of potentially reduced QoS due to service congestion. On the other hand, service prices can be increased to achieve higher QoS at charging facilities, but high prices may lead to a decrease in facility utilization. The proposed approach dynamically adjusts the pricing signals to incentive PEV users in adapting their service requests according to the EVCI usage pattern. Also, the generated pricing signals maintain the targeted QoS levels for multiple service classes when the PEV charging demand is a nonstationary stochastic process.

VII. Conclusions and Future Work

In this paper, we study decentralized coordination of PEV charging demand and propose a differentiated pricing mechanism for a multiservice EVCI. The objective of the proposed pricing mechanism is to maximize the utilization of charging facilities while preserving the targeted QoS level for each service class. A deep RL approach is leveraged to generate pricing policies without requiring a pre-specified model of PEV charging demand. The adopted RL approach can learn and optimize pricing strategies while interacting with the environment. Firstly, the differentiated pricing problem is formulated as a finite-discrete MDP. Then, TD3 algorithm is employed to train neural networks to encode the current EVCI state into pricing signals. The numerical results show that our differentiated pricing mechanism can coordinate the operation of a multiservice EVCI by preserving the targeted QoS for each service class. Thereby, the CSP can dynamically adjust pricing signals to incentivize PEV users to behave in ways that improve the overall utilization and performance of the EVCI.

The proposed differentiated dynamic pricing mechanism can be extended to minimize the negative impacts on the power grid. Various charging services can impact the power system differently. This is because each charging service class has a unique charging rate, thereby the load demand of the facilities varies according to the facility type. To minimize the negative impacts on the power system, differentiated pricing should motivate the PEV users to follow the valley-filling strategy. Consequently, PEV users are encouraged to shift the request of charging services to the time periods and the service classes which can be safely accommodated by the power grid.

References

[1] The International Energy Agency (IEA), "Global EV outlook 2019, scaling-up the transition to electric mobility," 2019.

[2] A. Abdalrahman and W. Zhuang, "PEV charging infrastructure siting based on spatial-temporal traffic flow distribution," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6115–6125, 2019.

[3] A. Abdalrahman and W. Zhuang, "QoS-aware capacity planning of networked PEV charging infrastructure," *OJVT*, vol. 1, pp. 116–129, 2020.

[4] A. Khaligh and M. D'Antonio, "Global trends in high-power on-board chargers for electric vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3306–3324, 2019.

[5] S. Y. Choi, S. Y. Jeong, B. W. Gu, G. C. Lim, and C. T. Rim, "Ultraslim s-type power supply rails for roadway-powered electric vehicles," *IEEE Trans. Power Electron.*, vol. 30, no. 11, pp. 6456–6468, 2015.

[6] A. Abdalrahman and W. Zhuang, "A survey on PEV charging infrastructure: Impact assessment and planning," *Energies*, vol. 10, no. 10, pp. 1650–1674, 2017.

[7] Q. Wang, X. Liu, J. Du, and F. Kong, "Smart charging for electric vehicles: A survey from the algorithmic perspective," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1500–1517, 2016.

[8] I. S. Bayram, G. Michailidis, and M. Devetsikiotis, "Unsplittable load balancing in a network of charging stations under QoS guarantees," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1292–1302, 2015.

[9] Y. Liu, R. Deng, and H. Liang, "A stochastic game approach for PEV charging station operation in smart grid," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 969–979, 2018.

[10] W. Tushar, W. Saad, H. V. Poor, and D. B. Smith, "Economics of electric vehicle charging: A game theoretic approach," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1767–1778, 2012.

[11] W. Yuan, J. Huang, and Y. J. A. Zhang, "Competitive charging station pricing for plug-in electric vehicles," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 627–639, 2017.

[12] C. Luo, Y.-F. Huang, and V. Gupta, "Stochastic dynamic pricing for EV charging stations with renewable integration and energy storage," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1494–1505, 2018.

[13] A. S. B. Humayd and K. Bhattacharya, "Design of optimal incentives for smart charging considering utility-customer interactions and distribution systems impact," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1521–1531, 2017.

[14] C.-H. Ou, H. Liang, and W. Zhuang, "Investigating wireless charging and mobility of electric vehicles on electricity market," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 3123–3133, 2015.

[15] Y. Zhang, P. You, and L. Cai, "Optimal charging scheduling by pricing for EV charging station with dual charging modes," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3386–3396, 2018.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[17] R. Rana and F. S. Oliveira, "Dynamic pricing policies for interdependent perishable products or services using reinforcement learning," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 426–436, 2015.

[18] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, 2018.

[19] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, July 2019.

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

[21] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE J. Power Energy*, vol. 5, no. 1, pp. 1–10, 2019.

[22] S. Zhou, Z. Hu, W. Gu, M. Jiang, M. Chen, Q. Hong, and C. Booth, "Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach," *Int. J. Elec. Power*, vol. 120, p. 106016, 2020.

[23] M. Chen and Z.-L. Chen, "Recent developments in dynamic pricing research: multiple products, competition, and limited demand information," *Prod. Oper. Manag.*, vol. 24, no. 5, pp. 704–731, 2015.

[24] F. Bernstein and A. Federgruen, "A general equilibrium model for industries with price and service competition," *Oper. Res.*, vol. 52, no. 6, pp. 868–886, 2004.

[25] J. Huang, M. Leng, and M. Parlar, "Demand functions in decision modeling: A comprehensive survey and research directions," *Decision Sciences*, vol. 44, no. 3, pp. 557–609, 2013.

[26] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 2765–2771.

[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[28] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.

[29] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *2014 International Conference on Machine Learning (ICML)*.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. Irpan, "Deep reinforcement learning doesn't work yet," https://www.alexirpan.com/2018/02/14/rl-hard.html, 2018.

[32] F. R. Cruz and T. Van Woensel, "Finite queueing modeling and optimization: A selected review," *J. Appl. Math.*, vol. 2014, 2014.

[33] A. F. Gad, "Implementing genetic algorithm in Python PyGAD," https://pypi.org/project/pygad, accessed: June 14, 2020.

[34] A. Kaso, "Computation of the normalized cross-correlation by fast fourier transform," *PLoS One*, vol. 13, no. 9, p. e0203434, 2018.

[35] S. Nguyen and C. Dupuis, "An efficient method for computing traffic equilibria in networks with asymmetric transportation costs," *Transport. Sci.*, vol. 18, no. 2, pp. 185–202, 1984.

[36] Electric Circuit Canada, "Charging stations and prices," https://lecircuitelectrique.com/charging-stations-and-rates, accessed: Mar 11, 2020.

Ahmed Abdalrahman received the Ph.D. (2020) degree in Electrical and Computer Engineering from the University of Waterloo, Canada, the M.Sc. (2013) degree in Electrical Engineering from Ain Shams University, Egypt, and the B.Sc. (2008) (First Class Hons.) degree in Electronics and Communications Engineering from Thebes Higher Institute of Engineering, Egypt. Currently, he is a Specialist, Market Analysis, Independent Electricity System Operator (IESO), Ontario, Canada. Dr. Ahmed is also the recipient of the 2020 Jon W. Mark Graduate Scholarship in Communication, and a recipient of several Faculty of Engineers Awards form the University of Waterloo. His current research interests include smart grids, mathematical optimization, and artificial intelligence.

Weihua Zhuang (M93-SM01-F08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks.

Dr. Zhuang was a recipient of the 2017 Technical Recognition Award from the IEEE Communications Society Ad Hoc and Sensor Networks Technical Committee, and a co-recipient of several Best Paper Awards from IEEE conferences. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2013, the Technical Program Chair/Co-Chair of IEEE VTC 2017 Fall and 2016 Fall, and the Technical Program Symposia Chair of IEEE Globecom 2011. She is an elected member of the Board of Governors and Vice President - Publications of the IEEE Vehicular Technology Society. She was an IEEE Communications Society Distinguished Lecturer from 2008 to 2011. Dr. Zhuang is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.