Dynamic Pricing for EV Charging Stations: A Deep Reinforcement Learning Approach

Zhonghao Zhao, *Graduate Student Member, IEEE*, and Carman K. M. Lee, *Senior Member, IEEE*

***Abstract*— Dynamic pricing, which aims to dynamically adjust the charging price in a timely fashion to unlock the flexibility of electric vehicle (EV) customers, has been extensively studied with the rapid development of charging technologies. Many existing works on dynamic pricing focus on maximizing the social welfare of charging service providers and EV customers. Cases of high-dimensional charging environments, which are often encountered with the rapid growth of EV market penetration, have been rarely considered to date. This article proposes a new dynamic pricing framework for EV charging stations that can offer multiple charging options to customers over a finite-time horizon. The charging price can be dynamically adjusted to maximize the quality of service (QoS) with a differentiated service requirement level (SRL) whenever the arrival rates and queuing system capacities of the charging systems are given at the end of a time period. The dynamic pricing problem is formulated as a finite-discrete horizon Markov decision process (MDP) with a mixed state space. A customized deep reinforcement learning (DRL) approach is employed to solve the examined EV dynamic pricing problem. The simulation results demonstrate the effectiveness of the proposed method.**

***Index Terms*— Deep reinforcement learning (DRL), dynamic pricing, electric vehicle (EV) charging station, quality of service (QoS).**

I. Introduction

THE resurgence of electric vehicles (EVs) powered by electricity provides an opportunity to reduce fossil oil dependence and mitigate the deterioration of roadside air quality [1]. With the increasing market penetration of EVs and the rapid development of charging technologies, public charging stations are considered to be an essential recharging source for EV customers. Compared with home charging, public charging stations can offer relatively lower charging prices for EV customers, as power can be purchased at a lower rate from the wholesale power market [2]. Hence, many large charging stations are being designed and deployed to offer multiple charging options for EV customers [3]. In this context, many governments have set climate-friendly policies for the deployment of EV chargers. A battery of executive orders has been released by the U.S. government in 2021 to stem the worsening impacts of catastrophic global warming [4]. In parallel, the U.K. government has announced an ambitious new target to achieve "net-zero" greenhouse gas emissions by 2050 to stave off the global security challenges posed by climate change [5]. Hence, the deployment of public charging stations has become a key priority to promote the use of EVs, having regard to the resulting energy efficiency and environmental benefits [6].

Admittedly, the increasing EV market penetration rate comes with many challenges. For instance, disordered charging service will inevitably lead to the line and transformer overloading [7]. Dynamic pricing, which refers to dynamically adjust the real-time charging price, is considered as a promising tool to overcome the challenges related to the increasing penetration rate of EVs [8]. Charging stations with an efficient dynamic policy can offer better charging services by providing economic incentives to control the charging demand of EV customers [9]. This is mainly because the dynamic pricing policy can reasonably reflect the situation of market supply and demand, shift the charging load of EVs, and reduce the negative impact of disorderly charging. Moreover, the relationship between charging price and demand can be explicitly defined since EV customers are price-sensitive to the charging service [10]. The existing dynamic pricing techniques in the literature can be mainly grouped into two categories: model-based algorithms and model-free algorithms. The former requires a deterministic model of the charging environment, mainly including evolutionary algorithms (EAs) and game theory (GT). EAs use mechanisms inspired by biological evolution. In [11], a particle swarm optimization (PSO) algorithm was employed to search for the optimal dynamic parking energy pricing policies at each time period through a stochastic framework. In [12], a heuristic algorithm considering malicious customers and unstable energy providers was proposed to determine the dynamic electricity price and the load capacity. In [13], a heuristic solution was introduced to determine the optimal coordinated dynamic pricing policy for EV charging stations so as to minimize the overlap between the EV charging and the residential peak load periods. However, there is no guarantee that the global optimum can be identified under a complex charging environment. On the other hand, dynamic pricing approaches based on GT are also widely used to map the relationship between charging service providers and EV customers, where the objective is to maximize the total social welfare [14]. A dynamic pricing scheme based on a stochastic

Manuscript received May 28, 2021; revised August 30, 2021 and November 1, 2021; accepted December 23, 2021. Date of publication December 30, 2021; date of current version April 20, 2022. This work was supported in part by the Laboratory for Artificial Intelligence in Design through the Innovation and Technology Fund under Grant RP 2-2. *(Corresponding author: Carman K. M. Lee.)*

The authors are with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, and also with the Laboratory for Artificial Intelligence in Design, Hong Kong (e-mail: zhongh.zhao@connect.polyu.hk; ckm.lee@polyu.edu.hk).

Digital Object Identifier 10.1109/TTE.2021.3139674

TABLE I
COMPARISON OF DIFFERENT DYNAMIC PRICING METHODS

Method	Pros	Cons
EA	performs well in uncomplicated problems; simple to understand and implement	model-based; lacks a reliable theoretical system; curse of dimensionality; poor performance with large state and action spaces
GT	performs well in situations of conflicting interests;	model-based; curse of dimensionality; low applicability; requires some unrealistic assumptions
DP	approximate optimum is guaranteed; wide applicability	model-based; curse of dimensionality; time-consuming
Q-learning	performs well in uncomplicated problems; real-time decision making; model-free	curse of dimensionality; time-consuming
DRL	real-time decision making; model-free; performs will in high-dimensional problems; good execution speed	selecting appropriate hyperparameters can be tricky

game was proposed in [15] based on the Nash equilibrium solution to minimize power distribution losses and guarantee the system reliability. In [16], a noncooperative Stackelberg game was adopted to set the price for optimizing the profit while ensuring the EV customers' participation. The existence and uniqueness of the dynamic pricing equilibrium can be achieved based on a low complexity game algorithm [17]. However, the assumption that players have knowledge about their own pay-offs and the pay-offs of others is not practical.

Although many authors have investigated the dynamic pricing problem for an EV charging station, the aforementioned model-based approaches are limited to decision-making problems under a deterministic charging environment [18]. Hence, as a "model-free" and "no need of expert knowledge" algorithm based on control theory, reinforcement learning (RL) techniques have gained popularity in recent years [19]–[21]. A state–action–reward–state–action (SARSA)-based dynamic charging pricing framework with a feature-based linear function approximator was used in [22] to optimize the total charging rates. In [23], a real-time price-based demand response algorithm was developed based on Q-learning. An online reservation system with a real-time charging pricing strategy was studied in [24] to motivate EV customers to use the designated charging station for services. Nevertheless, the computational efficiency of original RL-based algorithms sharply decreases under a high-dimensional charging environment since a lookup table is required to store the transition information of state–action pairs, soon rendering the problem intractable. Some studies use the deep reinforcement learning (DRL) approach with a nonlinear approximator to tackle the dimensionality challenges. In [25], a deep Q-network (DQN) with a priority experience replay memory was utilized to optimize the pricing decisions from the charging service provider's perspective. The state space and action space were defined as the wholesale market price and retail price, whereas the reward was set to be the overall profit. In [26], a dynamic pricing mechanism was investigated for a multiservice EV charging station, where the predetermined service quality was assumed to be maintained all the time. The comparison of different dynamic pricing methods is shown in Table I.

Although the DRL algorithms have been developed rapidly and have achieved excellent performance in many challenging tasks, the implementation of DRL in dynamic pricing is still limited due to the complexity of the DRL-based framework. In this study, the dynamic pricing problem is investigated from the EV customers' standpoint. The market share of EVs in recent years is still low compared with the traditional liquid-fuel powered vehicles [27], and thus, many governments have adopted economic incentive policies to increase the market penetration [4]–[6]. Obviously, the most important task in the early development stage before wide adoption is to improve EV customers' satisfaction [28]. Therefore, we propose a new DRL-based dynamic pricing framework considering the quality of service (QoS) from the EV customers' standpoint. In the literature, QoS is predominantly modeled based on indicators that result from queuing models [29], [30]. However, the connotation of the existing QoS evaluation models is mainly derived based on some unrealistic assumptions. A key assumption in the literature is that the customers' service requirement level (SRL) is assumed to be maintained all the time. In real life though, it is suggested that customers who anticipate a congested queuing system can become more impatient [31]. To overcome this challenge, we propose a new QoS evaluation model based on the universal generating function (UGF) technique, where the differentiated SRL is incorporated into the model so as to provide more flexibility to decision-makers.

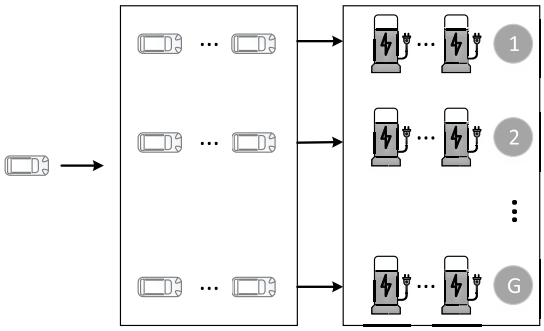

Fig. 1. Schematic view of the charging station.

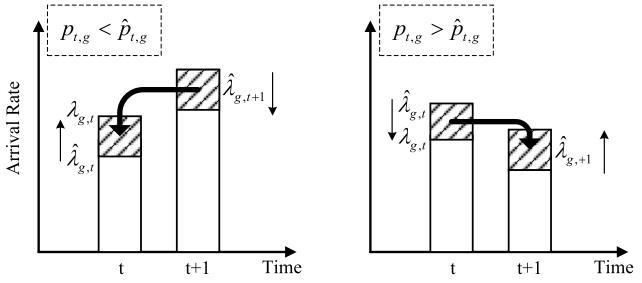

Fig. 2. Price-based charging demand response model.

To address the fundamental limitations of previous works, we present a customized DRL approach to resolve the studied dynamic pricing problem, where a novel QoS evaluation model is proposed to represent the service quality from the EV customers' perspective. To conclude, the contributions of this article are summarized as follows.

1) We formulate the dynamic pricing problem as an Markov decision process (MDP) from the EV customers' standpoint, where the $M/M/s/N$ queuing theory is utilized to model the charging environment.
2) We propose a novel QoS evaluation model considering the differentiated service requirements to formalize the charging service quality based on queuing theory and UGF.
3) We develop a customized DRL approach to resolve the dynamic pricing problem, where the state space consists of the queuing system capacities and arrival rates of all charging systems. By dynamically adjusting the charging price, the service quality can be maximized through a time horizon.

The remaining parts of this article are organized as follows. After the introduction in Section I, Section II presents the system model of the charging environment, along with a discussion of the queuing behavior and QoS evaluation. The MDP and a DRL-based dynamic pricing framework are given in Section III to resolve the examined dynamic pricing problem. Simulation results are reported in Section IV to examine the effectiveness of the DRL-based approach. Section V draws conclusions and future research for this study.

II. System Model

In many charging environments, charging stations are designed to provide multiple charging options to EV customers with different preferences and demands. In this study, queuing theory [32] is introduced to describe the EV charging behavior. As shown in Fig. 1, the EVs join in a first-input–first-output (FIFO) queue if all suitable chargers are occupied by other customers with the same charging option. In this study, the charging environment is outlined in the following.

1) We consider a charging station that provides G distinct charging options for EV customers with a specific option $g, g \in \{1, 2, \ldots, G\}$, which can be distinguished by their personal preferences and technical specifications.
2) Time is divided into T periods, where $t \in \{1, 2, \ldots, T\}$ represents the set of all time periods over a normal day. We assume that the arrival rate of each charging system does not change significantly in each time period.
3) Since the unified charging standard is still in the early development stage, we assume that each charger can only provide one type of charging service and that the charging station can be viewed as the combination of G charging systems. In the tth time period, the EV arrivals at each charging system are according to a Poisson process with arrival rate $\lambda_{g,t}$. Furthermore, the charging service follows an exponential distribution with a service rate μ_g that is to be maintained at all time.
4) All charging systems share one waiting area, i.e., there is a competitive relationship between each queue. In the tth time period, the charging service is provided by s_g chargers in each charging system with a distinct number of waiting positions $w_{g,t}$, queuing system capacity $N_{g,t}$, and charging price $p_{g,t}$.
5) The charging price can be adjusted at the end of each time period to schedule the charging behavior of EV customers and thus provide better service in the following time periods. For example, the peak load problem of the charging station can be mitigated by setting a higher charging price, whereas the service utilization can be increased by cutting the charging price during nonpeak hours [33], [34].

A. Charging Demand Response Model

In general, the charging demand profile is largely shaped by the EV customers' preferences. Therefore, the fundamental operation of an EV charging station is mainly governed by economic incentives. Reasonable adjustment of the charging price can help to align the charging demand to the service quality. Hence, it is essential to analyze the relationship between the charging demand and the real-time price. We denote the predetermined base charging price and arrival rate of charging system g in the tth time period as $\hat{p}_{g,t}$ and $\hat{\lambda}_{g,t}$. We assume that the charging request from EV customers will be advanced or delayed at most by one time period, as shown in Fig. 2. If an EV customer tends to defer the charging service, the price should satisfy $p_{g,t} > \hat{p}_{g,t}$ and vice versa. In this study, we adopt a price-based charging demand response model

based on a linear relationship, i.e.,

$$\lambda_{g,t} = \begin{cases} \hat{\lambda}_{g,t} + \hat{\lambda}_{g,t+1}\dfrac{\hat{p}_{g,t} - p_{g,t}}{\hat{p}_{g,t}}, & \textbf{if } 0 \le p_{g,t} < \hat{p}_{g,t} \\ \hat{\lambda}_{g,t}\left(1 + \dfrac{\hat{p}_{g,t} - p_{g,t}}{\hat{p}_{g,t}}\right), & \textbf{if } \hat{p}_{g,t} \le p_{g,t} < 2\hat{p}_{g,t} \\ 0, & \textbf{if } p_{g,t} \ge 2\hat{p}_{g,t} \end{cases} \quad (1)$$

where the first item of (1) indicates that the relative change of EV customers in the current time period is determined by the base arrival rate of the next time period and the second item shows that a high charging price will reduce the current charging demand. The third item restricts that the arrival rate should be a positive number. If $t = T$, that is, no future time period remains, the corresponding charging price is assumed to be $\hat{p}_{g,T} = p_{g,T}$. If more than one future time period is left, the base arrival of the $(t+1)$th time period can be updated by

$$\hat{\lambda}_{g,t+1} = \hat{\lambda}_{g,t+1} + \left(\hat{\lambda}_{g,t} - \lambda_{g,t}\right). \quad (2)$$

As a matter of fact, the price-based charging demand can be influenced by many stochastic and dynamic factors, including psychological phenomena, economic development levels, and personal arrangements. Consequently, it is difficult to mathematically model the customers' charging demand response in this context [35]. As a result, we adopt a DRL-based approach to resolve the dynamic pricing problem, where the charging demand can be learned through continuous interaction with the charging environment in an actual charging station [26]. Therefore, the proposed model is only used in the simulated charging environment in this study.

B. Queuing Model

The QoS [36] of the charging station serves as the scalar reward signal of the customized DRL framework. In this study, a novel QoS evaluation model is presented to examine the service quality from the standpoint of EV customers. In order to formalize the QoS of a charging station, a queuing theory is introduced to investigate the relationship between the service quality and the waiting time [37].

Unquestionably, a long waiting time has an adverse effect on the customer experience and vice versa. To this end, the $M/M/s_g/N_{g,t}$ queuing model is employed to describe the charging behavior, where M denotes a finite-state ergodic Markov chain. The state description of the Markov chain corresponds to the number of EV customers in the charging system $g, g \in \{1, 2, \ldots, G\}$. The Markov chain moves to another queuing state when a new EV arrives or leaves the charging system. Furthermore, the queuing state of the overall charging station can be further defined as the combination of the queuing states of all charging systems. The state transition process is shown in Fig. 3.

A Markov chain with finite ergodic state space $\xi = \{0, 1, \ldots, N_{g,t}\}$ is defined as a stochastic process that exhibits the memory-less property, i.e., the next queuing state of the Markov chain strictly depends on the current queuing state.

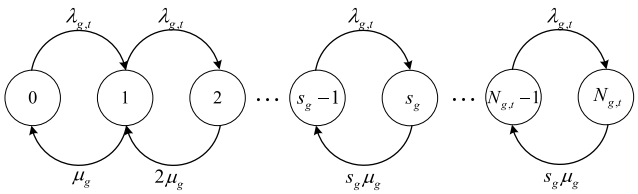

Fig. 3. Queuing state transition diagram of a charging system.

The matrix form of the transition process can be expressed as

$$\Lambda_{g,t} = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda & \cdots & 0 & 0 \\ 0 & 2\mu & -(\lambda+2\mu) & \cdots & \cdots & 0 \\ \cdots & & & \cdots & & \\ 0 & 0 & \cdots & s\mu & -(\lambda+s) & \lambda \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad (3)$$

where the elements in the matrix represent the one step infinitesimal rate of the state transition process. Let $\boldsymbol{\Gamma}_{g,t} = \{\Gamma_{g,t}(0), \Gamma_{g,t}(1), \ldots, \Gamma_{g,t}(N_{g,t})\}$ denote the steady state distribution of charging system g in the tth time period; then, we have

$$\boldsymbol{\Gamma}_{g,t} \cdot \Lambda_{g,t} = \mathbf{0}^T. \quad (4)$$

Stated in more general terms, the recursive form of the k_gth and (k_g+1)th steady-state probabilities can be written as

$$\left(k_g + 1\right)\mu_g \Gamma_{g,t}\left(k_g + 1\right) = \lambda_{g,t}\Gamma_{g,t}\left(k_g\right). \quad (5)$$

Denoting the occupancy rate of the queuing system as $\rho_g = \lambda_{g,t}/\mu_g$, correspondingly, the steady state distribution is given by the following:

$$\Gamma_{g,t}\left(k_g\right) = \begin{cases} \dfrac{1}{k_g!}\rho_g^{k_g}\Gamma_{g,t}(0), & \text{for } 0 \le k_g \le s_g \\ \dfrac{\rho^{s_g}}{s_g! s_g^{k_g - s_g}}\Gamma_{g,t}(0), & \text{for } s_g \le k_g \le N_{g,t}. \end{cases} \quad (6)$$

Due to the fact that $\sum_{k_g=0}^{N_{g,t}} \Gamma_{g,t}(k_g) = 1$, the probability that there is no customer in the charging system g is calculated by

$$\Gamma_{g,t}(0) = \left(\sum_{k_g=0}^{s_g-1}\frac{1}{k_g!}\rho^{k_g} + \frac{\rho^{s_g}\left[1 - \left(\rho/s_g\right)^{N_{g,t}-s_g+1}\right]}{s_g!\left(1 - \rho/s_g\right)}\right)^{-1}. \quad (7)$$

We denote the queuing length vector corresponding to the steady-state distribution as $\boldsymbol{L}_{g,t}$. According to Fig. 3, the queuing length is 0 if the queuing state satisfies $k_g < s_g$ since a suitable charger is available in such a case. Otherwise, the EV has to join a queue if $k_g \ge s_g$ since all available chargers are occupied by other EV customers. Hence, the mean queue length is given by

$$E\left(\boldsymbol{L}_{g,t}\right) = \sum_{k_g=s_g}^{N_{g,t}}\left(k_g - s_g\right)\cdot\Gamma_{g,t}\left(k_g\right). \quad (8)$$

The mean waiting time of charging system g in the tth time period is further computed by

$$E\left(\boldsymbol{W}_{g,t}\right) = \frac{E\left(\boldsymbol{L}_{g,t}\right)}{\lambda_{g,t}\left(1 - \Gamma_{g,t}N_{g,t}\right)} \quad \forall g \in \{1, 2, \ldots, G\}. \quad (9)$$

The waiting time is a natural evaluation metric of QoS for a charging system from the standpoint of EV customers. In what follows, a novel QoS evaluation model of the overall charging station is proposed based on EV customers' waiting time and the UGF.

C. QoS Evaluation

The overall charging station can be viewed as the combination of multiple charging systems characterized by a distinct queuing model and charging price. Viewed in this perspective, the service quality of the overall charging station is determined by all charging systems. From the EV customers' standpoint, a short waiting time always represents a better service provided by the charging station. We denote the waiting time in queuing state $N_{g,t}$ as $\boldsymbol{W}_{g,t}(N_{g,t})$. By recalling that the charging waiting time can be calculated whenever the arrival rate and queuing system capacity become known in a time period, the QoS of charging system g can be defined in the following:

$$H_{g,t} = \mathbf{min}\left\{\frac{\alpha(p_{g,t})}{\beta}\left(1 - \frac{E(\boldsymbol{W}_{g,t})}{\boldsymbol{W}_{g,t}(N_{g,t})}\right), 1\right\} \tag{10}$$

where $0 \le \alpha(p_{g,t}) \le 2$ represents the price penalty factor, which indicates that the EV customers' satisfaction with the charging price. Based on the price-based charging demand response model, $\alpha(p_{g,t})$ can be given in a linear form as

$$\alpha(p_{g,t}) = \mathbf{min}\left\{1 + \frac{p_{g,t} - \hat{p}_{g,t}}{\hat{p}_{g,t}}, 2\right\}. \tag{11}$$

Besides, $\beta \ge 1$ is the weight of the price penalty factor that can be predetermined by the charging service provider. The charging system, therefore, exhibits a finite number of states in a time period denoted as $H_{g,t} \in \boldsymbol{H}_{g,t} = \{H_{g,t}(0), H_{g,t}(1), \ldots, H_{g,t}(N_{g,t})\}$, where $H_{g,t}(k_g)$ is the QoS of charging system g that corresponds to queuing state k_g in the tth time period. Obviously, $H_{g,t}(k_g)$ is a discrete random variable that can have any value ranging from $H_{g,t}(0)$ to $H_{g,t}(N_{g,t})$. Then, the QoS vector can be further written as

$$\boldsymbol{H}_{g,t} = \mathbf{1} - \frac{\boldsymbol{W}_{g,t}}{\boldsymbol{W}_{g,t}(N_{g,t})}. \tag{12}$$

Let $\tilde{H}_t$ be the weighted average QoS of all charging systems. The relationship between $\tilde{H}_t$ and $H_{g,t}$ can be expressed as

$$\tilde{H}_t = \frac{1}{\lambda_t}\sum_{g=1}^{G}\lambda_{g,t}H_{g,t} \tag{13}$$

where $\lambda_t = \sum_{g=1}^{G}\lambda_{g,t}$ represents the arrival rate of the overall charging station. To differentiate the charging service requirement from the EV customers' perspective, UGF [38], [39] is adopted to represent the QoS state distribution. The UGF of charging system g in the tth time period, denoted as $u_{g,t}$, can be represented by

$$u_{g,t} = \sum_{k_g=0}^{N_{g,t}}\Gamma_{g,t}(k_g)z^{H_{g,t}(k_g)} \tag{14}$$

where $\Gamma_{g,t}(k_g)$ is the probability of charging system g sojourning in QoS state k_g in the tth time period. The auxiliary parameter z is used to distinguish the value of the discrete random variable and its corresponding probability. Given the arrival rate and the charging price of charging system g, i.e., $\lambda_{g,t}$ and $p_{g,t}$, the UGF of the charging station, denoted as U_t, can be written as follows:

$$U_t = \phi\left[\sum_{k_1=0}^{N_{1,t}}\Gamma_{1,t}(k_1)z^{H_{1,t}(k_1)}, \sum_{k_2=0}^{N_{2,t}}\Gamma_{2,t}(k_2)z^{H_{2,t}(k_2)}, \ldots, \sum_{k_G=0}^{N_{G,t}}\Gamma_{G,t}(k_G)z^{H_{G,t}(k_G)}\right] \tag{15}$$

where ϕ is a composition operator, which represents the structure relationship between the charging systems and the charging station. Let $K = \{k_{1,t}, k_{2,t}, \ldots, k_{G,t}\}$ denote the QoS state of the charging station, which represents a combination of the QoS state of all charging systems. Let $\Gamma_{J,t}$ be the probability of the QoS being equal to $H_{g,t}$ in the tth time period, and the UGF of the charging station can be further rewritten as

$$\begin{aligned} U(t) &= \sum_{k_1=0}^{N_{1,t}}\sum_{k_2=0}^{N_{2,t}}\cdots\sum_{k_G=0}^{N_{G,t}}\left[\prod_{g=1}^{G}\Gamma_{g,t}(k_g)z^{\phi\left[H_{1,t}(k_1),\ldots,H_{G,t}(k_G)\right]}\right] \\ &= \sum_{k_1=0}^{N_{1,t}}\sum_{k_2=0}^{N_{2,t}}\cdots\sum_{k_G=0}^{N_{G,t}}\left[\prod_{g=1}^{G}\Gamma_{g,t}(k_g)z^{\tilde{H}_t}\right] \\ &= \sum_{K}\Gamma_{J,t}z^{\tilde{H}_t}. \end{aligned} \tag{16}$$

Clearly, it is unreasonable to simply use $\tilde{H}_t$ to evaluate the QoS in many real-world applications. For instance, 15 min of waiting time will cause customers' dissatisfaction at noon, while at other times, people may have more patience. Hence, it is crucial to include the differentiated service requirement in the QoS evaluation model. To shed more light on this, the QoS of a charging station is further defined as the probability that $\tilde{H}_t$ is greater than a predetermined differentiated SRL $d_t (0 \le d_t \le 1)$ in the tth time period. Let $H_t^{d_t}$ denote the QoS when the SRL is set as d_t. Based on the aforementioned model, $H_t^{d_t}$ can be expressed as

$$H_t^{d_t} = \sum_{\hat{H}_t \ge d_t}\Gamma_{K,t}. \tag{17}$$

A toy example is presented in this section to put some of the details into perspective. In this example, the capacity of each queue is set as 2. The charging price of each charging system is set as $p = \hat{p}$. The arrival rate, QoS, and steady-state distribution of each charging system are given in Table II.

Consequently, the UGF of each charging system is

$$\begin{aligned} u_{1,t} &= \Gamma_{1,t}(0)z^1 + \Gamma_{1,t}(1)z^{0.65} + \Gamma_{1,t}(2)z^0 \\ &= 0.25z^1 + 0.65z^{0.65} + 0.1z^0 \end{aligned} \tag{18}$$

$$\begin{aligned} u_{2,t} &= \Gamma_{2,t}(0)z^1 + \Gamma_{2,t}(1)z^{0.85} + \Gamma_{2,t}(2)z^0 \\ &= 0.1z^1 + 0.85z^{0.65} + 0.05z^0 \end{aligned} \tag{19}$$

TABLE II
PARAMETERS OF THE TOY EXAMPLE

ID	Arr. Rate	State	QoS	Trans. Prob.
		0	1	0.25
1	5	1	0.65	0.65
		2	0	0.10
		0	1	0.10
2	10	1	0.75	0.85
		2	0	0.05

and the resulting UGF of the overall charging station is

$$\begin{aligned} U_t = {} & 0.005z^0 + 0.0325z^{0.2167} + 0.0125z^{0.3333} \\ & + 0.085z^{0.4333} + 0.01z^{0.6667} + 0.2125z^{0.7667} \\ & + 0.5525z^{0.7833} + 0.065z^{0.8833} + 0.025z^1. \end{aligned} \tag{20}$$

The SRL of the charging station in a time period can be determined by the charging service provider based on historical data and expert knowledge. In such a case, if the SRL of the charging station is $d_t = 0.6$, the QoS is computed to be 0.865. If the SRL is $d_t = 0.75$, the corresponding QoS is 0.6425.

III. PROPOSED DRL FRAMEWORK

As a matter of fact, EVs will be the only choice for customers in the future. The future charging station will be heterogeneous and high-dimensional. Therefore, the DRL framework is employed in this work to tackle the curse of dimensionality problem.

A. Markov Decision Process

In the tth period, the system state that includes information about the arrival rates and queuing system capacities can be observed by the decision-maker. Based on the observation, the charging service provider will choose a pricing action that will be performed for the $(t+1)$th time period. The details of the MDP formulation are given as follows.

1) *State:* For the dynamic pricing problem in this study, the system state encapsulates three types of information: 1) the arrival rates, $\boldsymbol{\lambda}_t = \{\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{G,t}\}$; 2) the queuing system capacities, $\mathbf{N}_t = \{N_{1,t}, N_{2,t}, \ldots, N_{G,t}\}$; and 3) the completed time periods, $t, t \in \{1, 2, \ldots, T\}$. Consequently, the state space can be expressed as

$$\mathbf{S}_t = \{s | \boldsymbol{s}_t = (\boldsymbol{\lambda}_t, \mathbf{N}_t, t)\}. \tag{21}$$

2) *Action:* At the end of a time period, the dynamic pricing action, denoted as $\mathbf{A}_t$ ($t \in \{1, 2, \ldots, T-1\}$), represents the charging price for each charging system, and $\mathbf{A}_t$ can be expressed as

$$\mathbf{A}_t = \left\{ \mathbf{a}_t | \sum_{t=1}^{T} \sum_{g=1}^{G} C_{g,t} \geq C_{\min}, \sum_{g=1}^{G} w_{g,t} = w_{\max} \right\} \tag{22}$$

where $C_{g,t}$ denotes the profit of the charging system g in the tth time period. The total profit requirement of the charging station is $C_{\min}$, which indicates that the selected pricing actions should satisfy the charging service provider's minimal expected profit. Furthermore, all charging systems share the same waiting area that consists of multiple waiting positions. Let Δ_t denote the period length of each time period. The waiting position allocation in the $(t+1)$th time period is completely determined by the expected profit of each charging system, i.e.,

$$C_{g,t} = \hat{\lambda}_{g,t}\left(1 - \Gamma_{g,t}\left(N_{g-1,t}\right)\right)\Delta_t p_{g,t} \tag{23}$$

$$w_{g,t+1} = \left[\frac{C_{g,t}}{\sum_{g=1}^{G} C_{g,t}} w_{\max} \right]. \tag{24}$$

3) *Reward:* The scalar reward signal is defined as the QoS of the next time period, i.e., $H_{t+1}^{d_{t+1}}$. In the following, the superscript d_{t+1}, which indicates the SRL, is omitted for the simplicity of notation.

The DRL-based approach combines the advantages of RL with deep learning, which does not seek to model the uncertainty of the state transition. All the information can be obtained after interacting with the charging environment [40]. Given the arrival rates and queuing system capacities of all the charging systems, i.e., $\boldsymbol{\lambda}_t = \{\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{G,t}\}$ and $\mathbf{N}_t = \{N_{1,t}, N_{2,t}, \ldots, N_{G,t}\}$, let $V(\boldsymbol{\lambda_t}, \mathbf{N_t}, t)$ denote the cumulative QoS of the future time periods. If $t = T-1$, i.e., only one time period is left, $V(\boldsymbol{\lambda_t}, \mathbf{N_t}, t)$ is equivalent to the QoS of the last time period and can be given by the following:

$$V(\boldsymbol{\lambda}_{T-1}, \mathbf{N}_{T-1}, T-1) = \max_{\mathbf{a}_{T-1} \in \mathbf{A}_{T-1}} H_{K-1}. \tag{25}$$

If there is more than one remaining time period, we have

$$\begin{aligned} V(\boldsymbol{\lambda}_t, \mathbf{N}_t, t) &= \max_{\pi} E\left[\sum_{n=t+1}^{T} H_n | \boldsymbol{\lambda_t}, \mathbf{N_t}, t \right] \\ &= \max_{\mathbf{a}_t \in \mathbf{A}_t} \{H_t + V(\boldsymbol{\lambda_{t+1}}, \mathbf{N_{t+1}}, t+1)\} \end{aligned} \tag{26}$$

where π is the pricing policy, which maps from a specific state to the dynamic pricing actions. Once the arrival rates, the queuing system capacities, and the selected charging pricing actions of all the charging systems in a charging station, i.e., $\boldsymbol{\lambda}_t, \mathbf{N}_t$, and $\mathbf{a}_t$, are given, the cumulative QoS of the future remaining $T-t$ time periods is called a Q-function of the proposed MDP, denoted as $Q_\pi(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t)$. The Q-function can be written as

$$Q_\pi(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t) = E_\pi\left[\sum_{n=t+1}^{T} \gamma^{n-t} \cdot H_n | \boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t \right] \tag{27}$$

where γ is the discount factor, which is introduced to balance the importance between the immediate reward and the long-term reward. In this study, the objective of the dynamic pricing problem is to find the optimal pricing policy, denoted as π^*, over all feasible policies to maximize the Q-function as

$$Q_\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t) = H_{t+1} + \max_{\pi} E\left[\sum_{n=t+2}^{T} H_n | \boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t \right]. \tag{28}$$

It is worth noting that the discount factor is set as 1 in this study since the simulated horizon is relatively short.

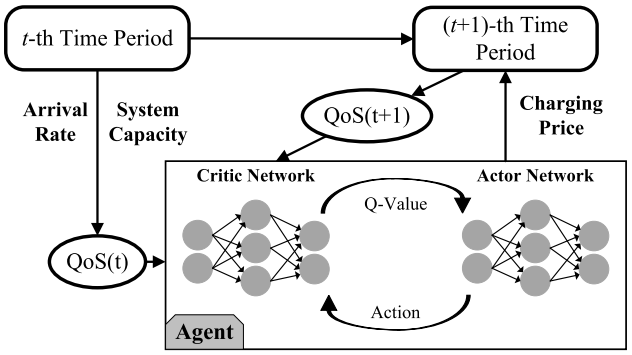

Fig. 4. Illustration of the proposed DRL framework.

Then, we have

$$V(\boldsymbol{\lambda}_t, \mathbf{N}_t, t) = \max_{\mathbf{a}_t \in \mathbf{A}_t} Q_\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t). \tag{29}$$

As the number of iterations approaches infinity, the Q-function will converge to $Q_\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t)$, and the optimal dynamic pricing policy can be therefore determined by the following:

$$\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t) = \operatorname*{argmax}_{\mathbf{a}_t \in \mathbf{A}_t} Q_\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t). \tag{30}$$

B. DRL for Dynamic Pricing

The realization of the model-free characteristic of the DRL-based framework mainly depends on the effective connection between customers, vehicles, and charging stations. In recent years, the continuous development of transportation cyber physical systems (CPSs) makes this connection possible. Moreover, high market penetration is critical to the collection of training data. Given the fact that the governments of many countries or regions have released ambitious policies to promote the adoption of EVs [4]–[6], it can be predicted that the market share of EVs will increase unceasingly in the near future.

With the continuous interaction between the agent and the charging environment, the value of the Q-function, denoted as Q value, is required to be stored in a lookup table. However, it is intractable to update the table in a high-dimensional environment. Hence, as a nonlinear Q value approximator, deep neural network (DNN) is introduced to address the "curse of dimensionality" problem, and the corresponding framework is defined as a DRL approach, as shown in Fig. 4. DRL is considered as a model-free approach to solve sequential decision-making problems. In real-world cases, the agent receives the current state (arrival rate and queuing capacity) and reward (QoS) at the end of each time period. The agent then selects an action (price) from the set of available actions, which is subsequently sent to the charging environment. Hence, the demand–price pairs can be observed and collected for agent training, and thus, our approach does not use the transition probability distribution associated with the MDP. The actor-critic algorithm [19] is employed to enhance the performance of the DRL-based approach. More specifically, the critic network (also called Q-Network) with parameter θ_Q is defined as a Q value approximator that evaluates the QoS of the overall charging station, i.e., $Q_\pi(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t) \approx Q_\pi^*(\boldsymbol{\lambda}_t, \mathbf{N}_t, \mathbf{a}_t, t)$. The structure of the critic network is shown in Fig. 5.

Based on the observed arrival rates and queuing system capacities, the actor network with parameter θ_b determines the dynamic pricing action for all charging systems. Likewise, in lieu of the conventional EAs (such as genetic algorithm (GA) and PSO algorithm), the actor network is constructed to output a pricing action under a large action space. The structure of the actor network is shown in Fig. 6. The input features of the actor network comprise the arrival rates and queuing system capacities, i.e., $\boldsymbol{\lambda}_t$ and $\mathbf{N}_t$,. In this study, the charging price is defined as a discrete variable since each circulating currency has its own minimum unit. For instance, the minimum unit of the Hong Kong dollar is "ten cents," namely, 0.10 HKD. However, it is a challenging work to find the optimal dynamic pricing policy with a minimal profit requirement under such a case. To overcome this problem, the Wolpertinger architecture [41], [42], which serves as a postprocess, is employed to find the optimal pricing actions in a computational efficient manner. More specifically, given the pricing action generated by the actor network, all the possible solutions with the Euclidean distance to this action being less than a predetermined length R are labeled as the neighbors of this action. If no neighboring solution satisfies the minimal profit requirement, the solution with the maximal profit is selected as a new pricing action. The process of selecting neighbors is repeated until a feasible solution complying with the profit requirement is found. Finally, the selected action will be put into the critic network for further evaluation. An illustration of the Wolpertinger architecture-based postprocess for a charging station consisting of two charging systems is shown in Fig. 7. Let $Q(\cdot)$ and $b(\cdot)$ denote the actor network and critic network, respectively. The feasible solution selecting process can be further represented by the policy function $\pi(\boldsymbol{\lambda}_t, \mathbf{N}_t, t | Q, b)$.

C. Training of Deep Neural Networks

In this study, learning the optimal dynamic pricing policy from a scalar reward signal, i.e., the QoS of the charging station, is difficult since the reward is considered to be frequently sparse, noisy, and delayed. Besides, the correlation between sequential samples is problematic for the proposed DRL-based approach that assumes a fixed underlying distribution. Therefore, a capacitated experience replay memory [43] is utilized to improve the stability of the agent training process. Let $\mathcal{D}$ denote the experience replay buffer. After performing a simulation iteration, a seven-tuple $(\boldsymbol{\lambda}_t, \mathbf{N}_t, t, \mathbf{a}_t, H_{t+1}, \boldsymbol{\lambda}_{t+1}, \mathbf{N}_{t+1})$ can be recorded in the replay buffer, and a batch of recorded tuples, denoted as $\mathcal{F}$, is randomly sampling from the replay buffer to update the critic network and actor network.

At each agent training step, the outputs of the critic network and the actor network inevitably shift, and if the constantly shifting outputs are used to update the networks, the subsequent outputs can easily spiral out of control. Hence, the target network technique is employed to address this

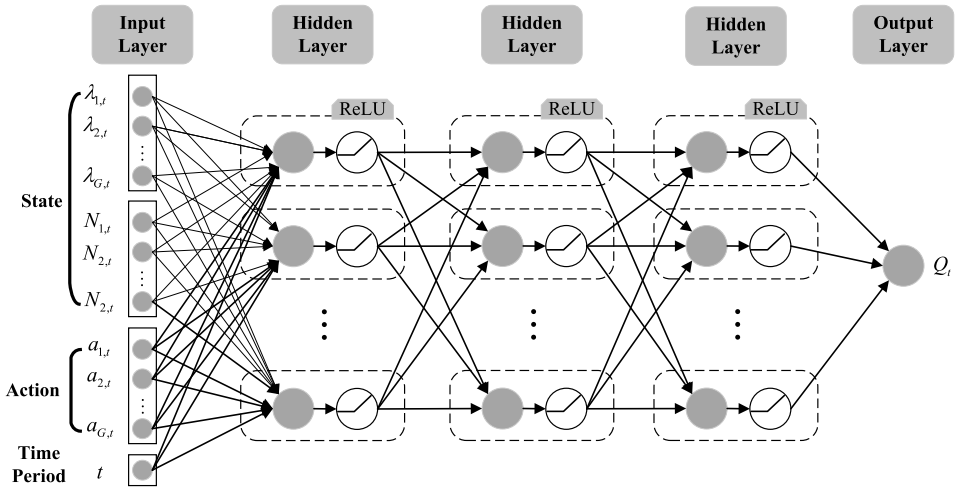

Fig. 5. Structure of the critic network.

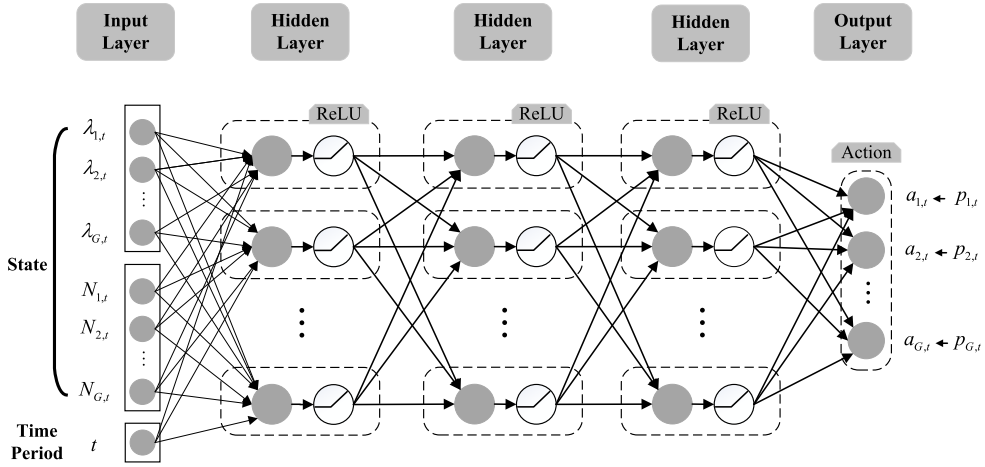

Fig. 6. Structure of the actor network.

fundamental limitation. Specifically, two target networks whose parameters are fixed and updated according to a pre-specified frequency O are introduced with the same structure of the original networks. Let $\hat{Q}(\cdot)$ and $\hat{b}(\cdot)$ denote the target critic network and target actor network with parameters $\hat{\theta}_Q$ and $\hat{\theta}_b$, respectively. The ith transition record in the minibatch uniformly sampled in the replay buffer, denoted as $\mathcal{F} = \{(\boldsymbol{\lambda}_{t_i}, \mathbf{N}_{t_i}, t_i, \mathbf{a}_{t_i}, H_{t_i}, \boldsymbol{\lambda}_{t_i+1}, \mathbf{N}_{t_i+1})\}_{i=1}^{\#\mathcal{F}}$, is utilized to estimate the target Q value, i.e.,

$$q_i = \begin{cases} H_{t_i+1}, & \textbf{if } t_i = T-1 \\ H_{t_i+1} + \hat{Q}\big(\boldsymbol{\lambda}_{t_i+1}, \mathbf{N}_{t_i+1}, t_i+1, \pi\big(\cdot|\hat{Q}, \hat{b}\big)\big), & \textbf{if } t_i < T-1 \end{cases} \tag{31}$$

where $\pi(\cdot|\hat{Q}, \hat{b}))$ is the postprocess for the selected actions from the target critic network. Then, the loss function and corresponding updating rule can be further derived as

$$L\big(\theta_Q\big) = \sum_{i=1}^{\#\mathcal{F}} \big[q_i - Q\big(\boldsymbol{\lambda}_{t_i}, \mathbf{N}_{t_i}, t_i, \mathbf{a}_{t_i}; \theta_Q\big)\big]^2 \tag{32}$$

$$\theta_Q = \theta_Q - \eta_Q \bigtriangledown_{\theta_Q} L\big(\theta_Q\big) \tag{33}$$

where η_Q is the learning rate of the critic network. In the same fashion, the action selected by the target actor network, i.e., $\hat{\mathbf{a}}_{t_i}$, is given by

$$\hat{\mathbf{a}}_{t_i} = \pi\big(\boldsymbol{\lambda}_{t_i}, \mathbf{N}_{t_i}, t_i|\hat{Q}, \hat{b}\big). \tag{34}$$

Likewise, the loss function and updating rule of the actor work can be written as follows:

$$L(\theta_b) = \sum_{i=1}^{\#\mathcal{F}} \big\|\hat{\mathbf{a}}_{t_i} - b\big(\boldsymbol{\lambda}_{t_i}, \mathbf{N}_{t_i}, t_i; \theta_b\big)\big\|_2 \tag{35}$$

$$\theta_b = \theta_b - \eta_b \bigtriangledown_{\theta_b} L(\theta_b). \tag{36}$$

The agent executes the pricing actions with a ε-greedy policy, that is, performing the selected pricing action with a probability $1-\varepsilon$ and randomly choosing an action for all charging systems with a probability ε. Finally, an uncorrelated mean-zero Gaussian noise is added to improve the performance of the training model. The proposed DRL-based approach for the examined dynamic pricing is presented in Algorithm 1.

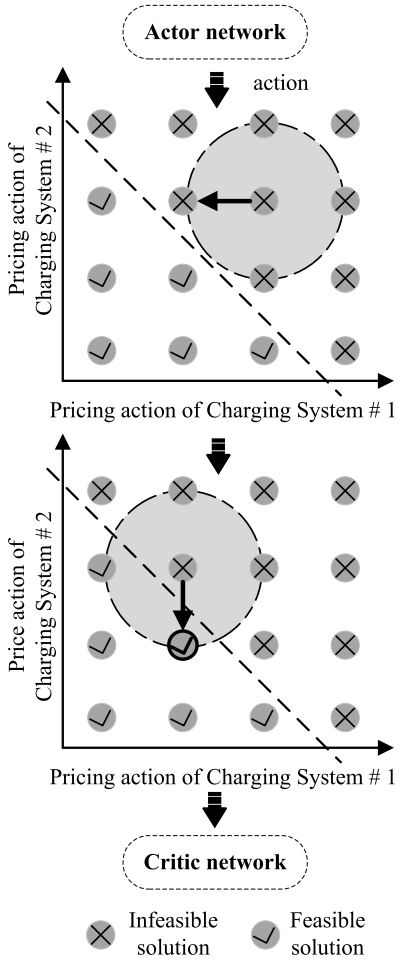

Fig. 7. Illustration of the feasible solution searching strategy.

Algorithm 1 DRL for Dynamic Pricing

Input: Maximum number of episodes: $\mathrm{I}_{\max}$
Output: Target critic network and target actor network

Randomly initialize the critic network parameters θ_Q and actor network parameters θ_b
Initialize the targets networks $\hat{\theta}_Q \leftarrow \theta_Q, \hat{\theta}_b \leftarrow \theta_b$
Initialize the replay memory $\mathcal{F} \leftarrow \emptyset$
for $i = 1 \rightarrow \mathrm{I}_{\max}$ **do**
  Initialize the state of the charging station
  Receive the observed state of the first time period
  for $t = 1 \rightarrow T - 1$ **do**
    if $I \leq 0.15\ \mathrm{I}_{\max}$ **then**
      Randomly select a pricing action
    else
      Select the pricing action $\pi(\cdot) + \varsigma \sim \mathcal{N}(0, \sigma^2))$ based on ε-greedy policy
    end if
    Execute action $\mathbf{a}_t$ and observe reward H_t
    Receive the observed state of the next time period
    Store transition $(\boldsymbol{\lambda}_t, \mathbf{N}_t, t, \mathbf{a}_t, H_{t+1}, \boldsymbol{\lambda}_{t+1}, \mathbf{N}_{t+1})$ in replay memory $\mathcal{D}$
  end for
  Sample a random minibatch of $\mathcal{F}$ transitions from $\mathcal{D}$
  if $t_i = T - 1$ **then**
    Set $q_i = H_{t_i+1}$
  else
    Set $q_i = H_{t_i+1} + \hat{Q}(\boldsymbol{\lambda}_{t_i+1}, \mathbf{N}_{t_i+1}, t_i + 1, \pi(\cdot|\hat{Q}, \hat{c}))$
  end if
  Perform gradient descent steps to update the target networks with a frequency R
end for

IV. Numerical Results

This section reports the simulation results to evaluate the feasibility and effectiveness of the proposed dynamic pricing framework for a charging station that provides multiple charging options to EV customers. In Section IV-A, a small-scale charging station that consists of three charging systems is presented to evaluate the efficiency of the proposed dynamic pricing model, whereas a large-scale charging station is designed in Section IV-B to examine the effectiveness of the DRL-based approach under a real-world charging environment.

A. Case Study I

In this section, we consider a charging station that offers three types of charging option to customers. The initial condition, i.e., the base arrival arrivals, base queuing system capacities, and base charging prices, of the overall charging station is prespecified in the simulated environment. An episode is partitioned into three time periods with equal length $\Delta_t = 8$ h ($t = 1, 2$, and 3), and the total profit requirement is $C_{\min} = 3.8 \times 10^3$ U.S. dollars. The SRL of each time period is set to $\{0.5, 0.9, 0.7\}$. Furthermore0, the maximal number of waiting positions is set as 14. All the parameter settings, including the parameters of the corresponding queuing model and the QoS evaluation model of all charging systems in the first numerical case, are elaborated in Table III.

TABLE III
Queuing Parameters of the First Numerical Example

ID	$\hat{\lambda}_{g,1}$	$\hat{\lambda}_{g,2}$	$\hat{\lambda}_{g,3}$	$N_{g,1}$	μ_g	$\hat{p}_g$	s_g
1	3	9	5	9	1.5	13	5
2	6	15	7	12	2.1	8	8
3	6	13	8	12	1.6	10	6

In the representation networks, all neural networks are constructed with three hidden layers. The number of neurons in each hidden layer is set as 12. The capacities of the experience replay memory and minibatch are set at $\mathcal{D} = 2048$ and $\mathcal{F} = 32$, respectively. The weight of the price penalty factor is set as 4. The distance of the feasible solution searching postprocess is set at $R = 1$. The standard deviation of the Gaussian noise is set as 0.1. The parameters of the target networks are updated every 15 iterations. Finally, the proposed algorithm is trained for $\mathrm{I}_{\max} = 1000$ episodes to learn the optimal dynamic pricing policy. The evolution of the cumulative rewards over 1000 episodes is shown in Fig. 8. The training process takes 2.727×10^3 s on the computer with

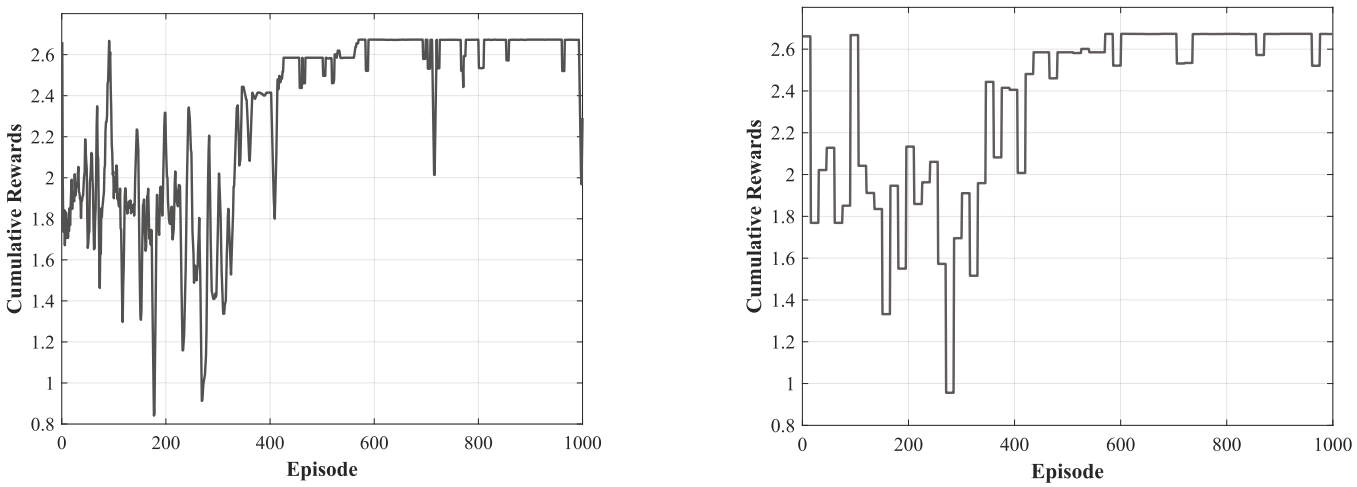

Fig. 8. Evolution of the cumulative rewards. (a) Training process of the critic network. (b) Training process of the target critic network.

an Intel Core i7-1065G7 CPU and 8-GB RAM. In the first 150 episodes, the dynamic pricing policy is randomly selected from all feasible solutions. Then, from episode 150, the action is selected based on a ε-greedy policy where the probability ε decays from 1 to 0.05 and remains 0.05 afterward. The cumulative rewards, i.e., the cumulative QoS of the charging station, are 2.672 as estimated by the critic network. The results demonstrate that the DRL-based approach succeeds in finding a pricing policy to maximize the cumulative QoS. To numerically justify the benefit of employing a DRL-based dynamic pricing framework, we compare the computation time and cumulative rewards for dynamic programming (DP), GA, DRL, and Q-learning under the same parameter setting, where DP serves as a benchmark since an approximate optimal solution can always be guaranteed. Please note that we still use the term "cumulative reward" to represent the final fitness of GA for convenience. The crossover probability, mutation probability, population size, number of iterations, and generation gap are set as 0.7, 0.01, 150, 1000, and 0.9, respectively. The results are tabulated in Table IV. We can see that DRL and GA learn considerably faster since DP and Q-learning take a lot of memory to store the results without ensuring if the stored Q value will be used or not. It indicates that with the increasing dimension of the charging environment, the computational efficiency of the methods, which require a lookup table, is unacceptable. We also observe that the cumulative reward of GA is significantly lower than other methods, whereas a relatively higher accuracy can be achieved by using a DRL-based method. This is because GA can only find the better solutions simply based on a binary encoding. It can be seen that our DRL-based method clearly outperforms the state-of-the-art methods. In addition to its advantages regarding computational time, the proposed method can also guarantee a high-quality solution.

TABLE IV
COMPARISON OF COMPUTATIONAL TIMES AND CUMULATIVE REWARDS FOR DIFFERENT METHODS

Method	Runtime (s)	Rewards	Relative Error
DP	5.190×10^4	2.647	-
DRL	2.727×10^3	2.672	0.94%
GA	2.133×10^3	2.254	14.85%
Q-learning	4.994×10^4	2.633	0.53%

B. Case Study II

In the second illustrative example, the presented DRL-based approach will be evaluated in a large-scale charging station. The simulated horizon is partitioned into eight time periods with equal length $\Delta_t = 3$ h ($t = 1, 2, \ldots, 8$). Four charging options (distinguished by the connector type), CHAdeMO, combined charging system (CCS), Tesla Supercharger, and ac Level-II, are offered by the service provider. It should be noted that our DRL-based approach is general enough to have wide applicability to different charging techniques. The base arrival rate and SRL are given according to the share of arriving cars per hour of a workday on Bornholm [44], as shown in Fig. 9. Based on previous studies on queuing patience [31], we assume that the SRL is proportional to the arrival rate, which is computed to be $\{0.503, 0.507, 0.686, 0.716, 0.757, 0.888, 0.719, 0.561\}$. The queuing parameters of each charging system is given in Table V. Based on the charging price report (U.S. dollar per kWh) in USA [45], four mainstream EV models, including Tesla Model S (40 kWh), Nissan Leaf (40 kWh), KONA Electric (39.2 kWh), and VM E-Golf (35.8 kWh), are selected to calculate the charging price of each charging option. The maximal number of waiting positions and the profit requirement are set at U.S. \$37 and U.S. $\$7.2 \times 10^3$, respectively. The values of all training parameters are tabulated in Table VI.

The training process of the cumulative rewards, i.e., the total QoS of the charging station, is shown in Fig. 10. The average results, including the charging prices, arrival rates, and queuing system capacities, of the last 40 episodes are shown in Fig. 11, where the dashed line represents the fixed pricing policy.

The cumulative reward is 7.2422 when the episode reaches $I_{\max}$. If a fixed pricing policy is adopted, i.e., the charging prices remain unchanged for all time periods,

TABLE V
QUEUING PARAMETERS OF THE SECOND NUMERICAL EXAMPLE

Charging System	$\hat{\lambda}_{g,1}$	$\hat{\lambda}_{g,2}$	$\hat{\lambda}_{g,3}$	$\hat{\lambda}_{g,4}$	$\hat{\lambda}_{g,5}$	$\hat{\lambda}_{g,6}$	$\hat{\lambda}_{g,7}$	$\hat{\lambda}_{g,8}$	μ_g	$\hat{p}_g$	$N_{g,1}$	s_g
Tesla Supercharger	2.1	2.5	9.1	11.1	14.1	16.5	11.2	5.6	2	14	14	7
CCS	1.8	2.3	8.4	10.3	12.5	14.5	9.8	4.3	1.5	12	19	9
CHAdeMO	1.6	3.1	7.3	9.8	11.2	13.4	8.1	3.6	1.2	10	25	13
Level-II	3.3	2.5	0.9	2.6	3.1	4.2	5.1	4.9	0.33	6.2	18	10

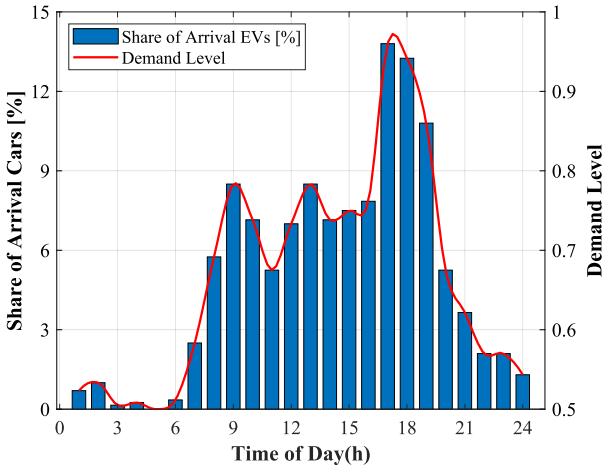

Fig. 9. Share of arriving EVs per each hour and the corresponding SRL.

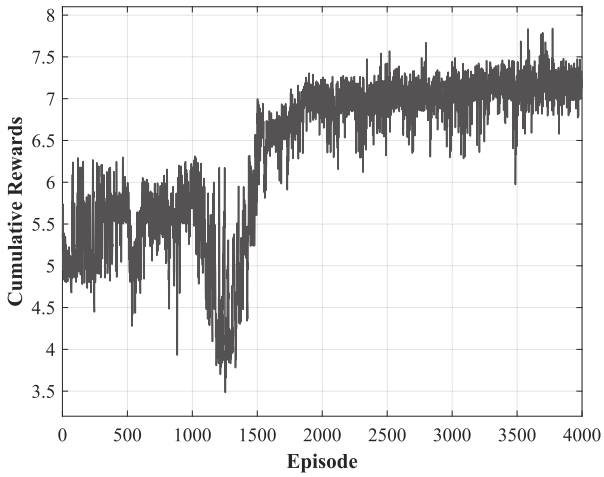

Fig. 10. Cumulative rewards of the second illustrative example.

TABLE VI
SIMULATION PARAMETERS

Parameter	**Value**
$I_{\max}$: number of episodes	4000
$\mathcal{D}$: replay memory size	10000
$\mathcal{F}$: minibatch size	64
O: update frequency	40
R: distance of the searching postprocess	1
number of hidden layers	3
number of hidden neurons	16
ε: ε-greedy policy	$1 \rightarrow 0.05$

the cumulative QoS is computed to be 5.8213. Therefore, compared with a fixed pricing policy, the cumulative rewards are increased by 24.4%. Under the proposed dynamic pricing framework, the charging price can continuously change in different time periods, and thus, the EV customers could respond to the dynamic charging prices provided by the charging service provider and thus adjust their charging request from peak hours to off-peak hours so as to maximize the cumulative reward. In Fig. 11(a) and (b), the arrival rates of the peak hours can be reduced by setting large charging price differentials between peak and off-peak time periods, resulting in a higher prominent flattening effect on the arrival rates. While in time periods where the base arrival rates are neither too high nor too low, the offered charging prices do not substantially change. In the off-peak hours, the charging service provider decreases the offered prices in order to attract EV customers from the following time period since the service utilization rate is low. In addition, the fluctuation of the charging prices is always kept in a reasonable range since the price penalty factor is included in the QoS evaluation model. In Fig. 11(c), the queuing system capacities are balanced based on the charging price to maximize the cumulative QoS of all time periods.

It is worth noting that the arrival rate distribution of the Level-II charging system over a normal day is different from other charging systems since the customers with this charging option tend to recharge their EVs at night. For the Level-II charging system, it is difficult for the charging service provider to find a reasonable pricing action in the first two periods by shifting the charging demand since the lower base arrival rate of the third time period limits the strategic potential of exploiting the EV customer's charging flexibility. Hence, more waiting positions are allocated to the Level-II charging system to guarantee a high QoS. Besides, the optimal dynamic pricing policy tends to set a higher charging price for sixth and seventh time periods and thus significantly increase the arrival rates of the last period. It can be explained by recalling that the SRL of the last period is only 0.561, which indicates that a higher QoS can be easily achieved by reasonably reallocating the waiting positions for each charging system in the last period. Going further, the queuing system capacities of Tesla Supercharger, CCS, and CHAdeMO charging systems in the last period are relatively lower. This is because the SRL of the last period is easily satisfied and the service rates are higher than that of the Level-II charging system.

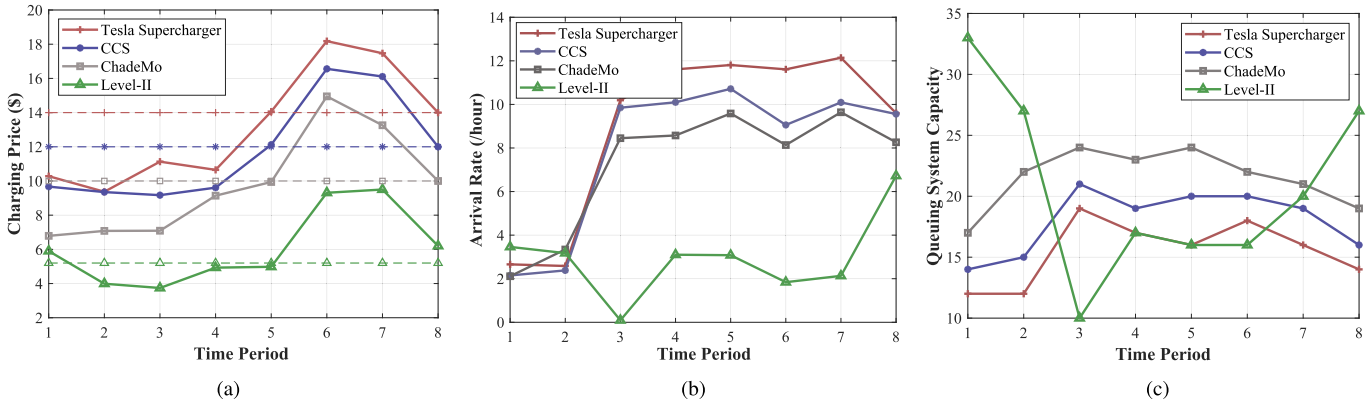

Fig. 11. Average results over a simulated horizon of the last 40 episodes. (a) Charging price. (b) Arrival rate. (c) Queuing system capacity.

The results demonstrate that our DRL-based approach is able to help the decision-maker to dynamically adjust the charging price to schedule the EV customers' charging request and reallocate the queuing system capacity, so as to provide better charging service to EV customers. It is worthwhile to note that not all charging stations can meet the requirements (e.g., effective customer feedback) of the proposed DRL framework, so the results obtained should be transferable. Fortunately, most of the current charging technologies are universal and tend to be unified. In addition, the price-based demand response in similar environments is also transferable, which provides a necessary prerequisite for the application of DRL.

V. Conclusion

In this study, we investigate the dynamic pricing problem for a charging station that can provide multiple types of charging services. In contrast with the existing literature, a novel QoS evaluation model considering differentiated SRL is proposed to evaluate the service quality based on the $M/M/s_g/N_{g,t}$ queuing model and UGF from the standpoint of the EV customers. To maximize the cumulative QoS of all time periods, the charging price of each charging system can be dynamically adjusted at the end of a time period. The dynamic pricing problem was formulated as a finite-discrete MDP. In this setting, motivated by the computational limitations of state-of-the-art approaches, a DRL-based approach is leveraged to find the optimal dynamic pricing policy.

The findings of the presented numerical examples are twofold. In the first numerical example, the DRL approach has been compared against DP, which requires a lookup table to store the transition information. The simulation results show that the presented DRL-based approach can determine the optimal dynamic pricing policy for all charging systems in a computationally efficient manner. The proposed approach is further applied to a real-world scenario in order to examine the effectiveness of the DRL-based framework under a high-dimensional charging environment. The simulation results show that the proposed DRL-based dynamic pricing approach achieves 24% higher QoS for the overall charging station than a fixed pricing policy. Thereby, the charging service provider can dynamically adjust the charging price to maximize the cumulative QoS of all time periods.

Future research aims at considering the stochastic case of the state and action features. In this study, the base arrival rates and charging prices are assumed to be deterministic, which is unrealistic in an EV charging network. To minimize the negative impacts on the dynamic pricing policy, the original DRL-based method should be further extended to a multiagent framework.

References

[1] Z. A. Needell, J. McNerney, M. T. Chang, and J. E. Trancik, "Potential for widespread electrification of personal vehicle travel in the United States," *Nature Energy*, vol. 1, no. 9, pp. 16112–16118, Aug. 2016.

[2] W. Lee, L. Xiang, R. Schober, and V. W. S. Wong, "Analysis of the behavior of electric vehicle charging stations with renewable generations," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Oct. 2013, pp. 145–150.

[3] B. Wang, P. Dehghanian, S. Wang, and M. Mitolo, "Electrical safety considerations in large-scale electric vehicle charging stations," *IEEE Trans. Ind. Appl.*, vol. 55, no. 6, pp. 6603–6612, Nov. 2019.

[4] The Guardian. (2021). *Biden Signals Radical Shift From Trump Era With Executive Orders on Climate Change*. [Online]. Available: https://www.theguardian.com/us-news/2021/jan/27/joe-biden-climate-change-executive-orders

[5] Committee on Climate Change. (2019). *Reaching Net Zero in the U.K.* [Online]. Available: https://www.theccc.org.U.K./U.K.-action-on-climate-change/reaching-net-zero-in-the-U.K./

[6] Steering Committee on the Promotion of Electric Vehicles. (2021). *Promotion of Electric Vehicles in Hong Kong*. [Online]. Available: https://www.epd.gov.hk/epd/english/environmentinhk/air/probsolutions/promotion_ev.html

[7] I. S. Bayram, A. Tajer, M. Abdallah, and K. Qaraqe, "Capacity planning frameworks for electric vehicle charging stations with multiclass customers," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1934–1943, Jul. 2015.

[8] S. Saharan, S. Bawa, and N. Kumar, "Dynamic pricing techniques for intelligent transportation system in smart cities: A systematic review," *Comput. Commun.*, vol. 150, pp. 603–625, Jan. 2020.

[9] J. Liu, G. Lin, S. Huang, Y. Zhou, Y. Li, and C. Rehtanz, "Optimal EV charging scheduling by considering the limited number of chargers," *IEEE Trans. Transport. Electrific.*, vol. 7, no. 3, pp. 1112–1122, Sep. 2021.

[10] Q. Chen *et al.*, "Dynamic price vector formation model-based automatic demand response strategy for PV-assisted EV charging stations," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2903–2915, Nov. 2017.

[11] S. Bagherzade, R.-A. Hooshmand, P. Firouzmakan, A. Khodabakhshian, and M. Gholipour, "Stochastic parking energy pricing strategies to promote competition arena in an intelligent parking," *Energy*, vol. 188, Dec. 2019, Art. no. 116084.

[12] Q. Tang, K. Yang, D. Zhou, Y. Luo, and F. Yu, "A real-time dynamic pricing algorithm for smart grid with unstable energy providers and malicious users," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 554–562, Aug. 2016.
[13] Z. Moghaddam, L. Ahmad, D. Habibi, and M. Masoum, "A coordinated dynamic pricing model for electric vehicle charging stations," *IEEE Trans. Transport. Electrific.*, vol. 5, no. 1, pp. 226–238, Mar. 2019.
[14] 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, May 2015.
[15] 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, Mar. 2018.
[16] G. Graber, V. Calderaro, P. Mancarella, and V. Galdi, "Two-stage stochastic sizing and packetized energy scheduling of BEV charging stations with quality of service constraints," *Appl. Energy*, vol. 260, Feb. 2020, Art. no. 114262.
[17] W. Yuan, J. Huang, and Y. Zhang, "Competitive charging station pricing for plug-in electric vehicles," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 627–639, Mar. 2017.
[18] M. Yang *et al.*, "Dynamic charging scheme problem with Actor–Critic reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 370–380, Jan. 2021.
[19] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[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, Feb. 2019.
[21] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 203–214, Jan. 2020.
[22] S. Wang, S. Bi, and Y. A. Zhang, "Reinforcement learning for real-time pricing and scheduling control in EV charging stations," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 849–859, Feb. 2021.
[23] 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, Jun. 2018.
[24] Z. Su, T. Lin, Q. Xu, N. Chen, S. Yu, and S. Guo, "An online pricing strategy of EV charging and data caching in highway service stations," in *Proc. 16th Int. Conf. Mobility, Sens. Netw. (MSN)*, Dec. 2020, pp. 81–85.
[25] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5901–5912, Sep. 2020.
[26] A. Abdalrahman and W. Zhuang, "Dynamic pricing for differentiated PEV charging services using deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 1, 2020, doi: 10.1109/TITS.2020.3025832.
[27] International Energy Agency. (2020). *Global EV Outlook 2020*. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2020
[28] M. Ebrahimi, M. Rastegar, M. Mohammadi, A. Palomino, and M. Parvania, "Stochastic charging optimization of V2G-capable PEVs: A comprehensive model for battery aging and customer service quality," *IEEE Trans. Transport. Electrific.*, vol. 6, no. 3, pp. 1026–1034, Sep. 2020.
[29] 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, Sep. 2019.
[30] B. Sun, X. Tan, and D. H. K. Tsang, "Optimal charging operation of battery swapping and charging stations with QoS guarantee," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4689–4701, Sep. 2018.
[31] S. Veeraraghavan and L. Xiao. (2018). *Impatience and Learning in Queues*. [Online]. Available: https://faculty.wharton.upenn.edu/wp-content/uploads/2018/11/Impatience-Learning-Queues.pdf
[32] U. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Boston, MA, USA: Birkhäuser, 2018.
[33] Y. Yang, M. Wang, Y. Liu, and L. Zhang, "Peak-off-peak load shifting: Are public willing to accept the peak and off-peak time of use electricity price?" *J. Cleaner Prod.*, vol. 199, pp. 1066–1071, Oct. 2018.
[34] M. Uddin, M. F. Romlie, M. F. Abdullah, S. A. Halim, A. H. A. Bakar, and T. C. Kwang, "A review on peak load shaving strategies," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 3323–3332, Feb. 2018.
[35] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.
[36] E. Ucer, I. Koyuncu, M. C. Kisacikoglu, M. Yavuz, A. Meintz, and C. Rames, "Modeling and analysis of a fast charging station and evaluation of service quality for electric vehicles," *IEEE Trans. Transport. Electrific.*, vol. 5, no. 1, pp. 215–225, Mar. 2019.
[37] T. S. Bryden, G. Hilton, B. Dimitrov, C. P. de Leon, and A. Cruden, "Rating a stationary energy storage system within a fast electric vehicle charging station considering user waiting times," *IEEE Trans. Transport. Electrific.*, vol. 5, no. 4, pp. 879–889, Dec. 2019.
[38] G. Levitin, *The Universal Generating Function in Reliability Analysis and Optimization*. London U.K.: Springer, 2005.
[39] S. Destercke and M. Sallak, "An extension of universal generating function in multi-state systems considering epistemic uncertainties," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 504–514, Jun. 2013.
[40] Y. Gao, J. Yang, M. Yang, and Z. Li, "Deep reinforcement learning based optimal schedule for a battery swapping station considering uncertainties," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5775–5784, Sep. 2020.
[41] G. Dulac-Arnold *et al.*, "Deep reinforcement learning in large discrete action spaces," 2015, *arXiv:1512.07679*.
[42] Y. Liu, Y. Chen, and T. Jiang, "Dynamic selective maintenance optimization for multi-state systems over a finite horizon: A deep reinforcement learning approach," *Eur. J. Oper. Res.*, vol. 283, no. 1, pp. 166–181, May 2020.
[43] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[44] L. Calearo, A. Thingvad, K. Suzuki, and M. Marinelli, "Grid loading due to EV charging profiles based on pseudo-real driving pattern and user behavior," *IEEE Trans. Transport. Electrific.*, vol. 5, no. 3, pp. 683–694, Sep. 2019.
[45] National Renewable Energy Laboratory. (2020). *News Release: Research Determines Financial Benefit From Driving Electric Vehicles*. [Online]. Available: https://www.nrel.gov/news/press/2020/research-determines-financial-benefit-from-driving-electric-vehicles. html

Zhonghao Zhao (Graduate Student Member, IEEE) received the B.S. degree from Northeast Forestry University, Harbin, China, in 2016, and the M.S. degree from Beihang University, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree in system engineering with the Industrial and Systems Engineering Department, The Hong Kong Polytechnic University, Hong Kong.

His research interests include reinforcement learning, plug-in electric vehicles, and charging infrastructure planning and operation.

Carman K. M. Lee (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong.

She is currently an Associate Professor with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong. She is also the Program Leader of [B.Sc. (Hons.)] Enterprise Engineering with Management. Her main research areas include industrial engineering, enterprise resource planning (ERP), logistics and supply chain management, the Industrial Internet of Things (IIoT), wireless sensor and actuator networks (WSANs), cloud computing, and big data analytics.