Reinforcement Learning for Real-Time Pricing and Scheduling Control in EV Charging Stations

Shuoyao Wang, Suzhi Bi, *Senior Member, IEEE*, and Yingjun Angela Zhang, *Senior Member, IEEE*

***Abstract*—This article proposes a reinforcement-learning (RL) approach for optimizing charging scheduling and pricing strategies that maximize the system objective of a public electric vehicle (EV) charging station. The proposed algorithm is "online" in the sense that the charging and pricing decisions made at each time depend only on the observation of past events, and is "model-free" in the sense that the algorithm does not rely on any assumed stochastic models of uncertain events. To cope with the challenge arising from the time-varying continuous state and action spaces in the RL problem, we first show that it suffices to optimize the total charging rates to fulfill the charging requests before departure times. Then, we propose a feature-based linear function approximator for the state–value function to further enhance the efficiency and generalization ability of the proposed algorithm. Through numerical simulations with real-world data, we show that the proposed RL algorithm achieves on average 138.5% higher charging-station profit than representative benchmark algorithms.**

***Index Terms*—Dynamic programming, machine learning, pricing and scheduling, reinforcement learning, state–action–reward–state–action (SARSA).**

Nomenclature

Set

$\mathcal{A}$	Action space.
$\mathcal{I}_t$	Set of electric vehicles (EVs) that arrive at the charging station during the time t.
$\mathcal{J}_t$	Set of all parking EVs at time t.
$\mathcal{K}_t$	Set of EVs yet to be charged in time slot t after the arrival of $\mathcal{I}_t$.
$\mathcal{S}$	State space.

Notation

α_t	Learning rate of the proposed algorithm at tth time slot.
$\boldsymbol{w}$	Weights of the feature values.
γ	Discount factor of Markov decision process (MDP) formulation.
θ_1, θ_2	Parameters that define the feature functions.
$\tilde{d}_i^t$	Residual demand of EV i at time t (kWh).
$\tilde{p}_i^t$	Residual parking time of EV i at time t.
A_t	Charging station action at time t.
c_t	Real-time electricity price at time t (\$kWh).
d_i	Demand of EV i (kWh).
$D_i(r)$	Demand of EV i in response to the charging price r at the arrival time slot (kWh).
$e^{\max}$	Maximum total charging rate of the charging station (kWh).
$g(S_t, A_t, \mathcal{I}_t, c_t)$	State transition function.
$L(\mathcal{J}_t)$	Latest residual parking time among the EVs in t.
p_i	Parking time of EV i.
$Q(S_t, A_t)$	Q-value given state–action pair (S_t, A_t).
S_t	System state at time t.
t_i^a	Arrival time of EV i.
$v_t(S_t, A_t)$	Reward function given state–action pair (S_t, A_t) (\$).
$x^{\max}$	Maximum charging rate of individual EVs (kWh).

Variable

e_t	Total charging rate of the charging station at time t (kWh).
r_t	Unique public charging price at time t (\$kWh).
x_{it}	Charging rates of EV i at time t (kWh).

I. Introduction

BEING one of the fastest growing sources of energy demand and greenhouse gas emission, the transportation sector is under great pressure to be decarbonized through deploying electric vehicles (EVs). High penetration of EVs is expected to change the power load profile significantly in distribution

Manuscript received May 31, 2019; revised September 23, 2019; accepted October 27, 2019. Date of publication October 31, 2019; date of current version November 18, 2020. The work of S. Bi was supported in part by the National Natural Science Foundation of China (Project 61871271), in part by the Guangdong Province Pearl River Scholar Funding Scheme 2018, in part by the Department of Education of Guangdong Province (Project 2017KTSCX163), in part by the Foundation of Shenzhen City (Project JCYJ20170818101824392), and in part by the Science and Technology Innovation Commission of Shenzhen (Project 827/000212). This work was supported in part by the General Research Funding (Project 14200315) established by Hong Kong Research Grant Council. Paper no. TII-19-2226. *(Corresponding author: Suzhi Bi.)*

S. Wang is with the Department of Risk Management, CDG, Tencent, Shenzhen 610041, China (e-mail: yorksywang@tencent.com).

S. Bi is with the College of Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: bsz@szu.edu.cn).

Y. A. Zhang is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, and also with the Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong (e-mail: yjzhang@ie.cuhk.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2950809

networks, creating potential threats to the power grid [1]. Establishing a conveniently available public charging infrastructure is essential to accommodating more clean energy, reducing carbon emissions, and alleviating peak charging loads. In the past decade, various EV charging control and scheduling schemes have been proposed to improve grid reliability [2], reduce charging operation cost [3], [4], offer auxiliary services [5], and promote the integration of renewable generation in commercial Microgrids [6], etc. The authors in [7] and [8] presented comprehensive surveys on the efficient online charging scheduling algorithms for the power grid performance enhancement.

Other than charging scheduling control, there have been increasing research efforts on designing proper demand response (DR) mechanisms to improve the overall system efficiency [9]–[11]. Therein, EVs adjust their charging demands according to the charging price announced by the charging stations or utilities. For instance, the authors in [10] and [11] considered dynamic pricing DR mechanisms for EV charging stations and distributed EVs, respectively. Overall, it has been widely accepted that an effective pricing and scheduling policy benefits both EV users and the grid system.

Most existing studies assume that besides the observation of past events, certain knowledge of the future EV arrivals and electricity prices are also known to the charging station [9], [12]–[14]. Such noncausal knowledge is broadly classified into two categories, namely the knowledge of the exact realization of future events and the knowledge of stochastic distributions of future events. For example, Sarker *et al.* [12] optimized the DR of EV aggregators with the assumption that the aggregator knows the future electricity prices perfectly in a noncausal manner. The authors in [13] and [14] proposed Markov decision process (MDP) based algorithms assuming that the stochastic distributions of future events are known. In practice, however, neither exact realizations nor distributional information of future events is easy to estimate precisely in a cost-effective manner. Moreover, the distribution of future events is likely to be time-varying in practice, making the estimation much harder. Alternatively, learning-based approaches that are driven by real-world data observations are good candidates to deal with this issue. For example, Chiş *et al.* [15] adopted an reinforcement-learning (RL) algorithm to schedule the home charging of an individual EV, such that the long-term electricity costs of EV owners are reduced. Likewise, an RL approach was adopted in [16] to learn the day-ahead hourly planning of EV fleet charging. However, most of the previous learning based algorithms focus on day-ahead planning schedule and, thus, ignore the random arrival and departure of EVs in real time. This is because random EV arrival and departure causes the state and action spaces of the RL problem to vary with time, rendering standard solution algorithms inapplicable.

To complement the previous work, we propose an RL algorithm to obtain the optimal pricing and charging scheduling strategy when random EV arrivals and departures are considered. To tackle the difficulty of time-varying state and action spaces due to random EV arrivals and departures, we propose to represent the state–action function using a linear combination of a set of carefully designed feature functions. The algorithm is practical in the sense that it is model-free and online. Here, by model-free, we mean that the decision does not depend on any assumed stochastic model of uncertain future events. By online, we mean that the algorithm is based on only the past events, including the arrival and departure process of EVs, that have already arrived and the electricity prices that have already been observed. Our main contributions are detailed as follows.

1) To the best of the authors' knowledge, this article is among the first to develop a model-free data-driven method for joint pricing and charging scheduling at an EV charging station with random EV arrivals and departures.
2) Through rigorous analysis, we show that it suffices to optimize the total charging rates of EVs instead of the individual charging rate to fulfill the charging demands before their departure times. Based on this, we greatly reduce the dimension of the action space without compromising the performance of the proposed algorithm.
3) To deal with the time-varying state and action spaces caused by random EV arrivals and departures, we approximate the state–action function with four carefully designed feature functions based on the features of the underlying physical system. The feature functions not only greatly reduce the dimension of the state space but also convert the decision in a time-varying action space to that of four time-invariant constants.
4) We evaluate the performance of the proposed method, referred to as the hyperopia state–action–reward–state–action (SARSA) based algorithm (HSA), through extensive simulations performed on real-world data. Our results show that HSA provides 20.2% and 132.8% extra profit than the truncated sample-average approximation (SAA) approach [17], [18] and the greedy policy on average, respectively. Moreover, we also study experiments to show the low computation time and peak shaving effect for HSA.

The rest of this article is organized as follows. We introduce the system model and formulate the problem as an MDP problem in Section II. Through rigorous analysis in Section III, we show that it suffices to optimize the total charging rates rather than the individual charging rates. In Section IV, we propose the HSA algorithm with feature function approximation. The simulation results are presented in Section V. Finally, Section VI concludes this article.

II. System Model and MDP

A. System Model

We consider the operation of an EV charging station (see Fig. 1) over a time horizon that is divided into T time slots. EVs arrive at the charging station at random times. We denote by $\mathcal{I}_t$, the set of EVs that arrive at the charging station at the beginning of time slot t, and by $\mathcal{J}_t$ the set of EVs that are 1) already at the charging station before the new EVs $\mathcal{I}_t$ arrive at time slot t, and 2) have not finished their charging. For notation simplicity, we denote $\mathcal{K}_t := \mathcal{J}_t \cup \mathcal{I}_t$ to be the set of EVs yet to be charged in time slot t after the arrival of $\mathcal{I}_t$. For all EVs

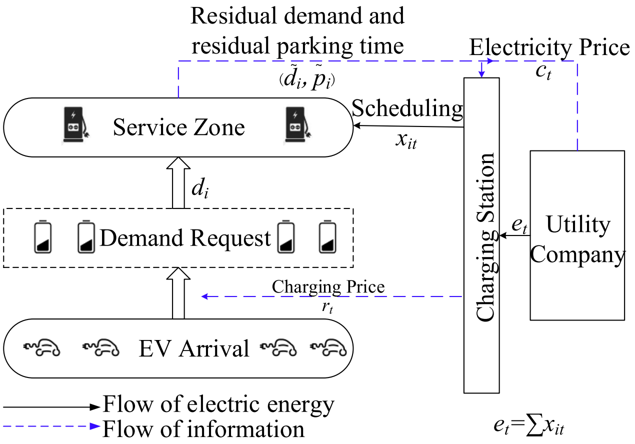

Fig. 1. EV charging station interaction system. Control signal: x_{it} and r_t.

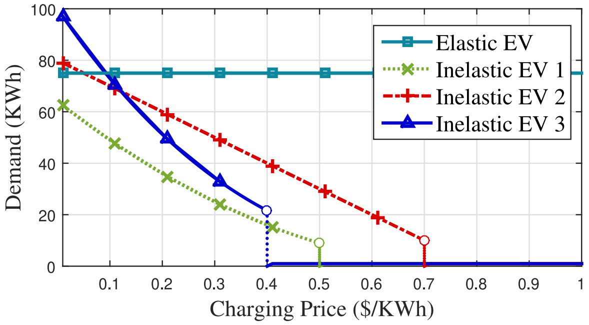

Fig. 2. Few examples of DR functions.

$i \in \mathcal{I}_t$, let t_i^a, p_i, and d_i denote its arrival time, parking time, and charging demand, respectively. In particular, the demand d_i must be satisfied before the departure of EV i at time $t_i^a + p_i$.

At each slot t, the charging station determines the charging rate of each EV $i \in \mathcal{K}_t$, denoted as x_{it} kWh. The charging rates are constrained by

$$x_{it} \leq x^{\max}, \; t = 1, \ldots, T, \quad \forall i \in \mathcal{K}_t \tag{1a}$$

$$\sum_{i \in \mathcal{K}_t} x_{it} \leq e^{\max}, \; t = 1, \ldots, T \tag{1b}$$

$$\sum_{t=t_i^a}^{t_i^a + p_i} x_{it} \geq d_i \quad \forall i \tag{1c}$$

where $x^{\max}$ and $e^{\max}$ are the maximum individual and aggregate charging rates, respectively. Moreover, (1 c) guarantees that the charging demand of each EV is satisfied before its departure time.

In return, the charging station also determines an unique public charging price r_t \$/kWh for all EVs that arrive at time t. The price then remains constant for the entire parking time. On the other hand, prices for EVs arriving at different times are different. This is consistent with the practical situation, where the EV owners accept the listed price when entering the parking lot. The EVs are assumed to be price sensitive. In response to r_t, an EV $i \in \mathcal{I}_t$ sets its charging demand as $d_i = D_i(r_t)$ kWh, where $D_i(\cdot)$ is the DR function of EV i.

Remark 1: The DR functions of different EVs can differ from each other greatly. For instance, the DR functions of inelastic users are constants. Moreover, when the charging price is too high, the EV users may choose another charging station, i.e., the demand equals to zero. As a result, for elastic user i, $D_i(r)$ always equals to zero as long as r is larger than a threshold r_i^{thres}. In Fig. 4, we plot several examples of DR functions.

As a result, at each time t the charging station collects a payment of $\sum_{i \in \mathcal{I}_t} r_t D_i(r_t)$ from the EVs, and pays an electricity bill of $c_t \sum_{i \in \mathcal{K}_t} x_{it}$ to the utility company. The industrial electricity price charged to the charging station, i.e., c_t \$/kWh, varies over time under the real-time pricing scheme [19].[1] Due to the uncertainty of the EV arrival process and electricity price, the charging station only knows the charging profiles of the EVs that have already arrived. Likewise, only the past and current electricity prices are known.

B. Markov Decision Process

At the beginning of each time slot t, the charging station determines the charging price and charging schedule based on its observation of the past and current events, including the charging demand and the departure time of EVs that have already arrived, and the electricity prices. The decision, in turn, affects the residual charging demands left for future time slots. Thus, the optimal decision is naturally a solution of an MDP [20]. In the following, we discuss the state, action, transition function, and reward function of the MDP.

1) System State: The system state at time t is described by $S_t = (\mathcal{J}_t, \tilde{d}_i^t|_{\forall i \in \mathcal{J}_t}, \tilde{p}_i^t|_{\forall i \in \mathcal{J}_t})$, where $\tilde{d}_i^t$ and $\tilde{p}_i^t$ are the residual charging demand and parking time of EV i at time t.

2) Action and Transition Function: Based on S_t and the observation of the real-time electricity price c_t and the EV arrivals $\mathcal{I}_t$, the charging station decides the charging price r_t and the charging rate of each EV x_{it} at time t. As such, the action A_t at time t is described by a high-dimensional vector, i.e., $A_t = (r_t, x_{it}|_{\forall i \in \mathcal{K}_t})$. Under the charging schedule x_{it}, we have

$$\begin{aligned} \tilde{d}_i^{t+1} &= \tilde{d}_i^t - x_{it} \quad \forall i \in \mathcal{J}_t \\ \tilde{d}_i^{t+1} &= D_i(r_t) - x_{it} \quad \forall i \in \mathcal{I}_t \end{aligned} \tag{2}$$

at the beginning of time slot $t+1$. Meanwhile, the residual parking time decreases as time increases from t to $t+1$, i.e.,

$$\begin{aligned} \tilde{p}_i^{t+1} &= \tilde{p}_i^t - 1 \quad \forall i \in \mathcal{J}_t \\ \tilde{p}_i^{t+1} &= p_i - 1 \quad \forall i \in \mathcal{I}_t \end{aligned} \tag{3}$$

where $\tilde{p}_i^{t+1} = 0$ indicates that EV i departures before time slot $t+1$ and thus $i \notin \mathcal{J}_{t+1}$. Therefore, $\mathcal{J}_{t+1}$ and the state transition function are calculated as follows:

$$\mathcal{J}_{t+1} = \mathcal{K}_t \{ i \in \mathcal{K}_t | \tilde{p}_i^{t+1} = 0 \text{ or } \tilde{d}_i^{t+1} = 0 \} \tag{4}$$

[1] We consider the real-time pricing scheme in our problem according to the rule of California ISO for the high power consumption companies.

and

$$\begin{aligned} S_{t+1} &:= g(S_t, A_t, \mathcal{I}_t, c_t) \\ &= \left(\mathcal{J}_{t+1}, \tilde{d}_i^{t+1}|_{\forall i\in\mathcal{J}_{t+1}}, \tilde{p}_i^{t+1}|_{\forall i\in\mathcal{J}_{t+1}}\right). \end{aligned} \tag{5}$$

As we can see, the problem now involves a high-dimensional action space $\mathcal{A}_t$. Fortunately, as we will prove in Section III-A, the action space can be equivalently reduced to a two-dimensional space $A_t = (r_t, e_t)$ without compromising the feasibility of the scheduling decision. Here, $e_t := \sum_{i\in\mathcal{K}_t} x_{it}$ is the total charging rate at time t.

3) Reward Function and Decision Problem: The reward function is closely related to the objective of the charging station, e.g., the profit of the charging station, the benefit of EV customers, the social welfare, etc. Without loss of generality, we suppose that the objective is to maximize the profit of the charging station in this article.[2] The reward function observed by the charging station at time t is the difference between the payment it collects and the electricity bill it pays. That is,

$$v_t(S_t, A_t) := \sum_{i\in\mathcal{I}_t} r_t D_i(r_t) - c_t e_t. \tag{6}$$

At each time t, the charging station aims to find the optimal action A by solving the following MDP problem:

$$Q_t(S_t) = \max_A v_t(S_t, A) + \gamma E_{c_{t+1},\mathcal{I}_{t+1}}\left[Q_{t+1}\left(g\left(S_t, A, \mathcal{I}_t\right)\right)\right] \tag{7}$$

subject to the physical and deadline constraints in (1 a)–(1 c). Here, $\gamma \in (0, 1)$ is a discount factor.

III. Problem Formulation

Random EV arrival and departure causes the state and action spaces of problem (7) to vary with time, rendering standard solution algorithms inapplicable. Therefore, before we tackle problem (7) by an RL approach in Section IV, we prove in Section III-A that to fulfill EV's charging demands before their departure times, it suffices to optimize the total charging rate, instead of individual charging rates, at time t. Then, the dimension of the action space is greatly reduced from $|\mathcal{K}_t| + 1$ to two.

A. Action Reduction

To satisfy the EVs' charging demands, the charging schedule x_{it} must satisfy constraints (1). Note that if we sum (1 c) over the EVs with the same departure time, we can conclude that the summation of the residual demands with departure time no later than $t + k$ must be no more than the summation of the charging rates from t to $t + k$. Meanwhile, (1 a) and (1 b) state that the total charging rate is bounded by not only the total charging rate upper bound $e^{\max}$ but also the number of parking EVs as the individual charging rate is also bounded by $x^{\max}$, i.e.,

$$e_t \le \min(e^{\max}, |\mathcal{K}_t| x^{\max}) \quad \forall t = 1, \ldots, T. \tag{8}$$

Substituting the charging rate bound into the summation of deadline constraints, we have the following Theorem 1.

For notational simplicity, we denote the longest residual parking time of the set $\mathcal{K}_t$ as $L(\mathcal{K}_t) := \max_{i\in\mathcal{K}_t} \tilde{p}_i^t$.

Theorem 1: If the total charging rate e_t satisfies the following inequalities for each time t

$$e_t \ge \sum_{i\in\mathcal{I}_t, p_i\le k} D_i(r_t) - \sum_{\tau=1}^{k} e_{t+\tau}^{up} + \sum_{i\in\mathcal{J}_t, \tilde{p}_i^t\le k} \tilde{d}_i^t$$

$$\forall k = 0, \ldots, L(\mathcal{K}_t) \tag{9a}$$

$$e_t \le \min(e^{\max}, |\mathcal{K}_t| x^{\max}) \tag{9b}$$

then there exists at least a set of x_{it}s that is feasible to (1). Here, we denote the maximum total charging rates in future time slots as $e_{t+\tau}^{up} = \min(e^{\max}, |\mathcal{K}_{t+\tau}| x^{\max})$, $\tau = 0, \ldots, L(\mathcal{K}_t)$. One such set of x_{it}s can be obtained by $\boldsymbol{e}$-supervised least laxity first (LLF) scheduling in Appendix A.

The detail proof of Theorem 1 could be find in Appendix B.

B. Problem Formulation

Thanks to Theorem 1, we can reduce the action space $\mathcal{A}_t$ from $(r_t, x_{it}|_{i\in\mathcal{K}_t})$ to (r_t, e_t). Problem (7) can be recast as

$$\begin{aligned} &Q_t(S_t) \\ &= \max_A \quad v_t(S_t, A) + \gamma E_{c_{t+1},\mathcal{I}_{t+1}}\left[Q_{t+1}\left(g\left(S_t, A, \mathcal{I}_t\right)\right)\right] \\ &\text{s.t.} \\ &e_t \ge \sum_{i\in\mathcal{I}_t, p_i\le k} D_i(r_t) - \sum_{\tau=1}^{k} e_{t+\tau}^{\mathrm{up}} + \sum_{i\in\mathcal{K}_t, \tilde{p}_i^t\le k} \tilde{d}_i^t \\ &\forall k = 0, \ldots, L(\mathcal{K}_t) \\ &e_t \le \min(e^{\max}, |\mathcal{K}_t| x^{\max}). \end{aligned} \tag{10}$$

If the distributions of $\mathcal{I}_t$ and c_t are explicitly and accurately known, (10) can be solved by the conventional numerical methods, such as SAA based on Monte Carlo sampling techniques methods, although the complexity could be prohibitively high. In practice, however, precise distributional information is rarely available.

IV. RL Approach

In this section, we employ an RL algorithm to solve (10) based on SARSA, which is data driven and requires no distributional information about EV arrivals and electricity prices in the future. Note that the traditional SARSA algorithm [21] cannot be directly applied to our problem. This is because the dimension of the state space S_t, which is proportional to the cardinality of $|\mathcal{K}_t|$, keeps varying with the random arrival and departure of EVs. To address the problem, we propose an SARSA algorithm with binary linear feature function approximation in this section.

[2] Changing the objective does not affect the structure of the problem, and our analysis remains unchanged. By adapting the feature functions according to the reward functions, our proposed algorithm still works.

A. SARSA Algorithm

Since the distributions of $\mathcal{I}_t$ and c_t are unknown, we replace the state–value function $Q_t(S_t)$ in (10) by a state–action value function $Q(S_t, A_t)$, which is estimated from the observation of the reward $v_t(S_t, A_t)$ and the transition to the next state (S_{t+1}, A_{t+1}). More specifically, it is updated as

$$\begin{aligned} Q(S_t, A_t) \leftarrow & (1-\alpha_t)Q(S_t, A_t) \\ & + \alpha_t \left[v_t(S_t, A_t) + \gamma Q(S_{t+1}, A_{t+1})\right] \end{aligned} \tag{11}$$

where α_t is the learning rate. Moreover, to balance between exploration and exploitation, we adopt an ϵ-greedy policy, where the charging station adopts the optimal action that maximizes $Q(S_t, A_t)$ with probability $1-\epsilon$ and randomly chooses an action with probability ϵ. That is,

$$\Pr[A_t = \arg\max_{A_t \in \mathcal{A}_t} Q(S_t, A_t)] = 1 - \varepsilon,\, 0 < \varepsilon < 1. \tag{12}$$

B. Linear Function Approximation

The conventional SARSA method requires $Q(S_t, A_t)$ to be learned and stored for all (S_t, A_t) pairs. This is infeasible in our problem due to the following two reasons. First, the state and action spaces are continuous. Discritizing them leads to a prohibitively large Q table, if the quantization level is reasonably small. Second, the space of S_t varies with time due to EV arrivals and departures. To address the abovementioned issues, a different table must be created and stored for each time t, which is not practical at all. As such, we propose to approximate $Q(S_t, A_t)$ by a linear combination of Y feature functions $\hat{f}_y(S_t, A_t)$, $y = 1, \ldots, Y$. The approximate of $Q(S_t, A_t)$, referred to as $\hat{Q}(S_t, A_t)$, is calculated as the weighted sum of the feature values. Specifically

$$\hat{Q}(S_t, A_t, \boldsymbol{w}) = \sum_{y=1}^{Y} w_y \hat{f}_y(S_t, A_t) \tag{13}$$

where $\boldsymbol{w} := (w_1, \ldots, w_Y)$ are the weights of feature values. Then, the proposed SARSA algorithm only needs to learn $\boldsymbol{w}$ instead of the Q-table for all pairs (S_t, A_t). The feature functions reduce the look-up space from a time-varying $Q(S_t, A_t)$-table to Y values.

1) Feature Functions: In this section, we introduce four feature functions based on the fundamental attributes of the problem. In particular, we construct the feature functions based on the objective and constraint functions of the problem. First of all, the total profit is the summation of the profits in each time slot. At time t, the profit of the charging station consists of the payments from the EV customers and the electricity bill. Accordingly, the first and second feature functions are defined as follows:

$$f_1(S_t, A_t) = \sum_{i \in \mathcal{I}_t} D_i(t) r_t \tag{14}$$

and

$$f_2(S_t, A_t) = -c_t e_t. \tag{15}$$

Meanwhile, from (2), the charging rate at time t directly affects the residual demand $\tilde{d}$ in future time slots. Less residual demands and later deadlines indicate more flexibility and, hence, more potential future profit. The third and fourth feature functions are proposed to prevent overly aggressive scheduling decisions that may compromise the rewards in future time slots. The same amount of demand with different deadlines result in different limitations of future action spaces. To quantify the deadline-biased demand, we use the arithmetic sequence weights and geometric sequence weights to define the biases. In particular, we define $f_3(S_t, A_t)$ as follows:

$$f_3(S_t, A_t) = -\sum_{\tau=0,\ldots,L(\mathcal{K}_t)-1} (L(\mathcal{K}_t) - \tau)\theta_1 \sum_{i \in \mathcal{K}_t, p_i \le \tau+1} \tilde{d}_i^t \tag{16}$$

where θ_1 is the predetermined common ratios of the arithmetic sequence weights. Meanwhile, the weights may not be linearly increasing. As a result, we further define $f_4(S_t, A_t)$ as follows:

$$f_4(S_t, A_t) = -\sum_{\tau=1,\ldots,L(\mathcal{K}_t)} \theta_2^\tau \sum_{i \in \mathcal{K}_t, p_i \le \tau} \tilde{d}_i^t \tag{17}$$

where θ_2 is the predetermined common ratios of the geometric sequence weights. By "negative sign," we mean that we prefer smaller f_3 and f_4. The selection of the value of θ_1 and θ_2 are numerically studied, and set as 0.1 and 0.9 in our simulation experiments.

2) Binary Feature Functions: It has been shown in [22] that if the learning rate α_t in (11) satisfies $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 \le \infty$, the SARSA algorithm with linear function approximation converges to a bounded region with probability one under the condition that $||f_y||_\infty = 1$.[3,4] However, the feature functions defined previously can potentially be unbounded. Moreover, the feature functions can take any real numbers, which may incur high computation and storage costs. To enforce convergence and reduce the computation and storage costs, we propose to transform the feature functions defined previously into binary feature functions. In particular, we denote $\bar{f}_1$, $\bar{f}_2$, $\bar{f}_3$, and $\bar{f}_4$ as the moving average of f_1, f_2, f_3, and f_4, respectively.[5] If the current reward is no less than the moving average $\bar{f}_1$, we set $\hat{f}_1(S_t, A_t)$ to 1. Otherwise, it is set to 0. That is,

$$\hat{f}_1(S_t, A_t) = \begin{cases} 1, \text{if } \sum_{i \in \mathcal{I}_t} D_i(t) r_t \ge \bar{f}_1 \\ 0, \text{otherwise.} \end{cases} \tag{18}$$

Likewise, $\hat{f}_2(S_t, A_t)$, $\hat{f}_3(S_t, A_t)$, and $\hat{f}_4(S_t, A_t)$ are given by

$$\hat{f}_2(S_t, A_t) = \begin{cases} 1, \text{if } -c_t e_t \ge \bar{f}_2 \\ 0, \text{otherwise} \end{cases}$$

$$\hat{f}_3(S_t, A_t)$$

[3] By "converges to a bounded region," we mean that there exits a bounded region of the weight vector such that the weight vector of the algorithm converges to the region with probability 1.

[4] $||f_y||_\infty := \max(|f_1|, \ldots, |f_Y|)$.

[5] In this article, we use the moving average over the last 20 time slots. For the first 20 time slots, $\bar{f}_y$ is simply the average feature value.

Algorithm 1: HSA (Hyperopia SARSA Algorithm).

```
Input: $\gamma$
Output: $A_t$
  $\boldsymbol{w} \leftarrow$ arbitrary values; Observe $S_1$.
  for $t := 1$ to $\infty$ do
     Based on the current state $S_t$, execute action $A_t$.
    Obtain the immediate reward $v_t(S_t, A_t)$ and the new
    system state $S_{t+1}$.
    Select $A_{t+1} \leftarrow \begin{cases} \max_{A \in \mathcal{A}_t} Q(S_{t+1}, A, \boldsymbol{w}) & \text{w.p } 1-\varepsilon \\ \text{a random action} & \text{w.p } \varepsilon \end{cases}$
    Update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha_t \big[ v_t(S_t, A_t) + \gamma \hat{Q}(S_{t+1}, A_{t+1}, \boldsymbol{w}) - \hat{Q}(S_t, A_t, \boldsymbol{w}) \big] \nabla_{\boldsymbol{w}} \hat{Q}(S_t, A_t, \boldsymbol{w})$.
    Update $S_t \leftarrow S_{t+1}$, $A_t \leftarrow A_{t+1}$, and $\bar{f}_y$
end for
```

$$= \begin{cases} 1, \text{if } - \sum\limits_{\tau=0,\ldots,L(\mathcal{K}_t)-1} \left(L(\mathcal{K}_t) - \tau\right) \theta_1 \sum\limits_{i \in \mathcal{K}_t, p_i \le \tau+1} \tilde{d}_i^t \ge \bar{f}_3 \\ 0, \text{otherwise} \end{cases}$$

$$\hat{f}_4(S_t, A_t)$$

$$= \begin{cases} 1, \text{if } - \sum\limits_{\tau=1,\ldots,L(\mathcal{K}_t)} \theta_2^{\tau} \sum\limits_{i \in \mathcal{K}_t, p_i \le \tau} \tilde{d}_i^t \ge \bar{f}_4 \\ 0, \text{otherwise.} \end{cases} \tag{19}$$

When the binary feature function approximation is well-defined, the updates are performed on the weights $\boldsymbol{w}$ because they control the contribution of each feature function on $\hat{Q}(S_t, A_t, \boldsymbol{w})$. To make the approximation precise, we aim to minimize the mean-squared error over the dynamic arrival and electricity price, i.e.,

$$E_{\mathcal{I}_t, c_t}\left[\left(Q(S_t, A_t) - \hat{Q}(S_t, A_t, \boldsymbol{w})\right)^2\right]. \tag{20}$$

In (20), the expectation is taken over $\mathcal{I}_t$ and c_t. Recall that we do not assume any distribution of I_t and c_t. Instead, we adopt a data-driven approach to update $\boldsymbol{w}$ based on the observations of I_t and c_t. Taking the derivative of (20) over $\boldsymbol{w}$, we get the update direction as

$$\triangle \boldsymbol{w} = \alpha \big[v_t(S_t, A_t) + \gamma \hat{Q}(S_{t+1}, A_{t+1}, \boldsymbol{w}) - \hat{Q}(S_t, A_t, \boldsymbol{w}) \big] \nabla_{\boldsymbol{w}} \hat{Q}(S_t, A_t, \boldsymbol{w}). \tag{21}$$

In particular, at time t, the vector $\boldsymbol{w}$ is adjusted in the direction (21) that reduces the error between $Q(S_t, A_t)$ and $\hat{Q}(S_t, A_t, \boldsymbol{w})$.

The proposed algorithm, referred to as HSA, is shown in Algorithm 1. By "hyperopia," we mean that the algorithm aims at the average profits, which prevents the charging station from making too aggressive decisions that sacrifice future profits for current profits. Without loss of generality, α_t is selected as $\alpha_t = \frac{1}{\sqrt{t}}$. This, together with the fact that $||\hat{f}_y||_\infty = 1, y = 1, \ldots, Y$, indicates that HSA converges to a bounded region with probability one [22]. The convergence is also verified by simulations in Section V-D.

Before leaving this section, we would like to emphasize that $\boldsymbol{w}$ is updated based on the current realization of I_t and c_t, but not on any distributional information or noncausal information about the future. If the distribution of $\mathcal{I}_t$ and c_t remains the same, the online learning HSA algorithm obtains a stationary $\boldsymbol{w}$-policy at convergence. Meanwhile, our numerical results show that even if the distribution is time varying, HSA is able to quickly adapt the policy in response to the time varying distribution, as long as the variation is no faster than the learning process.

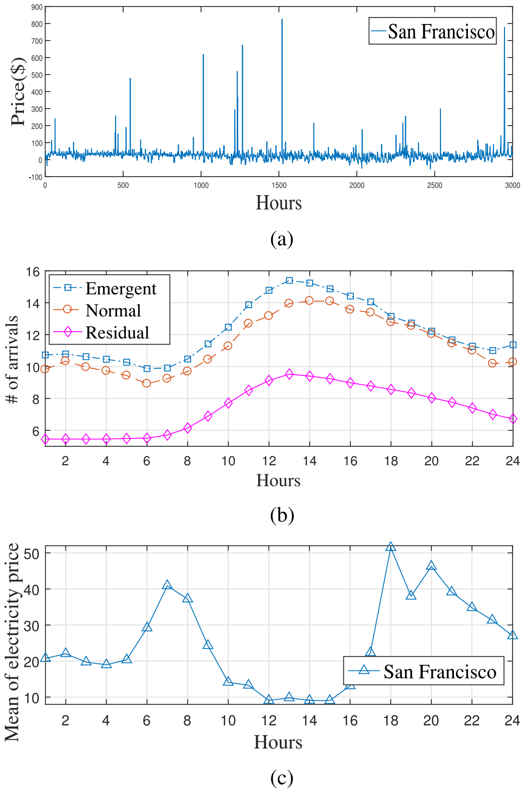

Fig. 3. Data for conducting simulations. (a) Hourly electricity prices of San Francisco. (b) Average hourly arrivals of different types of EVs. (c) Average hourly electricity prices of San Francisco.

V. Experiments

In this section, we investigate the performance, computation time, convergence, and impact to the grid of HSA. All the computations are executed in MATLAB on a computer with an Intel Core i7-3770 3.40 GHz CPU and 16 GB of memory.

A. Experimental Setup

We base our simulations on the historic hourly data, including the day-ahead electricity prices [see Fig. 3(a) and (c)] of San Francisco in California ISO [23] and the number of vehicle arrivals for Richards Ave station near downtown Davis [see Fig. 3(b)] [24]. The data of arrivals are the total number of the passing-by vehicles. We use the scaled number of arrivals to model the number of EVs that enter the charging station. The DR function is modeled as $D(r) = \beta_1 r + \beta_2$. The EVs are divided into three types, namely emergent, normal, and residential uses. The parameters of the three are listed in Table I.

TABLE I
PARAMETERS OF DIFFERENT TYPES OF EV CHARGING PROFILES

Type	Emergent	Normal	Residential
Variance	4.47	3.96	2.63
β_1 (KWh/$)	-1	-4	-25
β_2 (KWh)	6	15	100
Parking Time (minutes)	30	120	720

The dataset spans from July 1, 2016 to July 31, 2016. The length of a time slot is one minute in our simulations. Unless specified otherwise, we set the discount factor as $\gamma = 0.9$. Suppose that the total charging capacity of the charging station ranges from 200 to 600 kW.[6]

The regulations of existing markets differ in their clearing frequencies. Here, we consider two clearing frequencies, i.e., slowly varying and fast varying electricity prices. Under the slowly varying case, the electricity price remains unchanged during the same hour. Under the fast varying case, the electricity price changes every 5 min.

In each experiment, we compare the performance of our proposed HSA algorithm with the following benchmark algorithms.

1) Robust simulation-based policy improvement (RSPI): Huang *et al.* [14] proposed the RSPI algorithm to stochastically match the EV charging load with the wind supply. Through extensive simulation, Huang *et al.* [14] demonstrated that the RSPI can obtain a policy that is much better than the state-of-the-art benchmark algorithms.
2) SAA: As discussed in Section III, the SAA method is typically adopted as the optimal online algorithm [17], [18]. However, since the complexity of SAA is too high, truncation is often applied in SAA to reduce the complexity at a cost of performance loss. We refer to the SAA with a truncation period of k time slots as the SAA-k method. In our simulations, the number of samples in each stage is 10 000.
3) Greedy policy: The charging station makes the charging and scheduling decision based on the assumption that there is no future EV arrivals. Moreover, the future electricity prices are assumed to be equal to the average price.

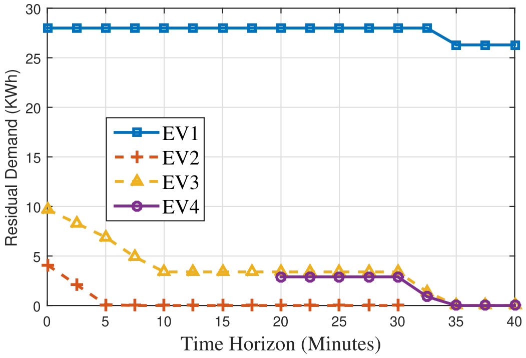

Fig. 4. Curve of charging SOC. The departure times of EV1, EV2, EV3, and EV4 are time 720, time 30, time 120, and time 50, respectively.

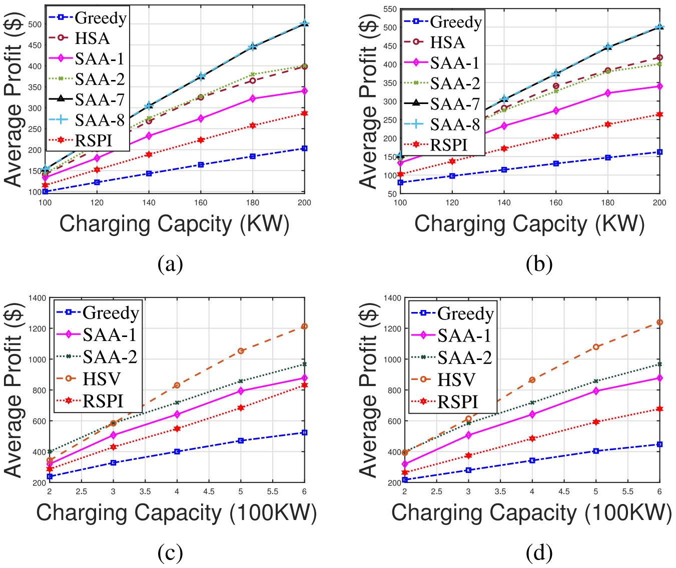

Fig. 5. Profit performance comparison versus capacity. (a) Slowly varying price. (b) Fast varying price. (c) Slowly varying price. (d) Fast varying price.

B. Profit Performance

In the first experiment, we investigate how the charging station pricing and scheduling strategy is and how the profit of the charging station changes versus the total charging capacity.

In Fig. 4, we show the charging state of charge (SOC) using data from July 1, 2016 as an example. At the beginning of the plotted time horizon, three EVs arrive at the charging station and request their charging demands in response to the charging price. Because of the earliest departure time, EV 2 has the highest priority, and thus, the charging SOC increases rapidly. From time 0 to 10, the charging rate of EV 1 is always zero. This is because EV 1 will park at the charging station for a long time and the charging station aims to charge EV 1 when the electricity price is very low. From time 10 to 20, the charging capacity is occupied by other EVs with tighter deadlines, and thus, the charging rate of EV 1 and EV 3 are both zero. Due to the high electricity price, the charging station stops all the charging from time 20 to 30. Overall, the EV with the tightest deadline has the highest priority for the charging service. The EVs that park at the charging station for a very long time are only charged when the electricity price is very low.

In Fig. 5, the profit of the charging station is evaluated under both fast and slowly varying electricity prices. The average profit per hour over 30 days are plotted in Fig. 5. We compare the average profit performances in Fig. 5(a) and (b) when the charging capacity increases from 100 to 200 kW, where we can apply SAA to high order of k, e.g., SAA-7 and SAA-8. From our simulations, we notice that the SAA algorithms with truncation periods of more than 7 time slots achieve almost the same average profit. Accordingly, we consider in practice the results

[6] [Online]. Available: https://www.tesla.com/destination-charging

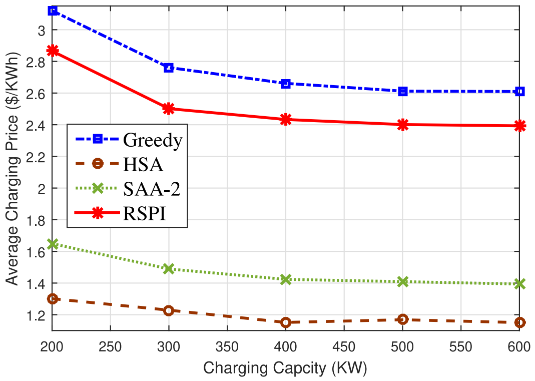

Fig. 6. Average charging fee per kWh versus capacity.

of SAA-7 as the optimal online solution. On average, the HSA algorithm achieves 79.80% and 83.79% profit of the optimal solution under fast and slowly varying electricity prices, respectively. However, the SAA algorithm not only takes long time to compute but also requires the distributional future information, which may not be practically, especially when k is large, in a real scenario.

In Fig. 5(c) and (d), we consider charging stations with larger charging capacities. Specifically, we vary the charging capacity from 200 to 600 kW. In this case, the computation time of SAA with a truncation period of more than 3 time slots can be as long as several weeks. As such, we only simulate the performance of the Greedy method, SAA-1, SAA-2, and RSPI for performance comparison. Therefore, we only compare HSA with the Greedy method, SAA-1, SAA-2, and RSPI when the charging capacity increases from 200 to 600 kW. Overall, the profits increase as the charging capacity increases. In comparison, the gaps between the HSA, and the Greedy method and RSPI, widen when the charging capacity becomes large. This is because with a large charging capacity, the HSA method has more flexibility to shift the charging demands to the time slots with low electricity prices. In contrast, benchmark algorithms, especially the Greedy method, always charges the EVs as fast as possible as long as the marginal profit is positive.

From Fig. 5, we see that the proposed HSA method significantly outperforms the benchmark methods under different charging station capacity setting. The performance advantage is especially notable when the charging capacity is large.

C. Charging Fee Performance

In the second experiment, we compare the performance achieved by HSA with the Greedy method, SAA, and RSPI from the EV owners perspective. The average charging fee per kWh over 30 days are plotted in Fig. 6.

From Fig. 6, we can observe that HSA also achieves lower average charging fee than the Greedy method, SAA, and RSPI when the electricity price increases from 200 to 600 kWh. This, together with the abovementioned performance evaluations from the perspective of the charging station, implies that the proposed charging station algorithm benefits both the charging station and EV owners.

TABLE II
NORMALIZED COMPUTATION TIME VERSUS CAPACITY

	200 KW	300 KW	400 KW	500 KW	600 KW
HSA	1	1.32	1.73	1.91	2.00
SAA-1	6.93×10^3	8.95×10^3	1.11×10^4	1.38×10^4	1.64×10^4
SAA-2	9.17×10^6	1.13×10^7	1.27×10^7	2.85×10^7	1.57×10^7

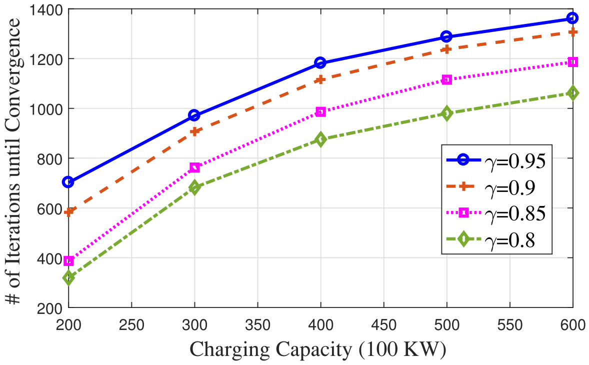

Fig. 7. Convergence of HSA versus capacity.

D. Computational Time

In the third experiment, we compare the computation time of HSA and the SAA algorithms under different charging capacities in Table II. We normalize the computation time of different schemes to that of the HSA method under 200 kW capacity, which is 0.14 s in our simulation. Each number in the table is an average over 30 days under both slow and fast varying cases. Table II shows that the average computation time of HSA increases almost linearly with the charging capacity. In contrast, for SAA-1 and SAA-2, the average computation times grow much faster as the charging capacity increases. We notice that the average computation time of SAA-2 is far greater than that of SAA-1, which could be several weeks in our simulations. In contrast, the proposed HSA only takes several seconds to compute a result. It is foreseeable that the computational complexity of SAA will become extremely high as we further increase the charging capacity.

E. Convergence of HSA

In the fourth experiment, we plot the average number of iterations to convergence versus the total charging capacity of the charging station in Fig. 7. Unless otherwise stated, each point in the figure is an average performance of 100 random opening hours of the charging station with random initial weights. For all discount factors, the number of iterations until convergence increases as the charging capacity increases, as expected. This is because that the larger the total charging capacity is, the more actions the charging station could make. The larger number of feasible actions leads to longer exploration and convergence time. Similarly, for all charging capacities, the number of iterations increases as the discount factor increases. An intuitive explanation is that the large discount factors value the future profits a lot and thus the charging station use more iterations to learn the future profits more precisely.

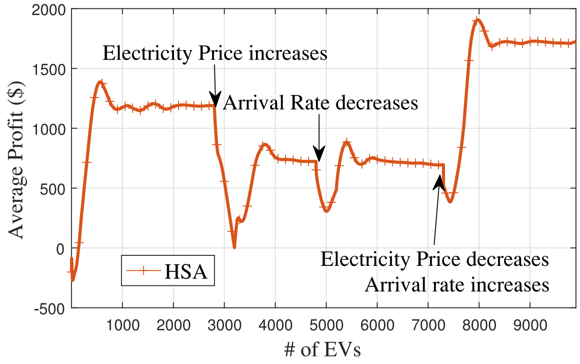

Fig. 8. Convergence of HSA.

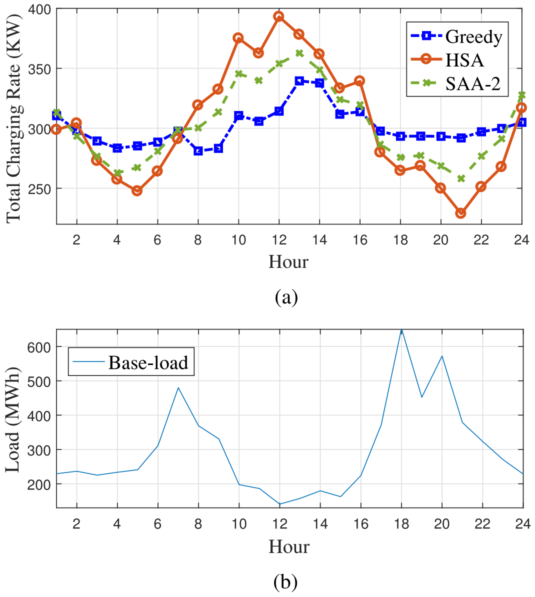

Fig. 9. Impact of the charging load to the Grid. (a) Average hourly total charging rate. (b) Average hourly base load.

Overall, the average numbers of iterations to convergence are 497.6, 830.4, 1039.7, 1154.9, and 1229.0 for total charging capacity 200 kWh, 300 kWh, 400 kWh, 500 kWh, and 600 kWh, respectively.

Although the number of iterations increases as the system size increases, we observe from Fig. 7 that the iteration number increases slowly with the charging capacity, indicating that HSA is scalable and efficient when applied by large-size charging station.

In addition, we also plot the convergence of HSA in Fig. 8 when the distribution is time varying. In our simulation, the mean of the electricity prices doubles after the 2800th arrival, the arrival rate of emergent EVs reduces by one half after 4800th arrival, and the mean of the electricity prices and the arrival rate of normal EVs both double after 7300th arrival. We observe that the profit converges to a new value quickly after the distribution changes. Overall, when the electricity price increases or the arrival rate decreases, the converged profit decreases, and vice versa. This shows that even if the distribution is time varying, HSA is able to quickly adapt the policy in response to the time varying distribution.

F. Impact to the Grid

In the fifth experiment, we evaluate the average total charging rate of the proposed HSA algorithm in Fig. 9(a). To facilitate the comparison, we also plot the average total charging rate of SAA-2 and the Greedy method as benchmark algorithms. Overall, we observe that when the base-load is high [see Fig. 9(b)], the charging load is relatively small, and vice versa. This is also known as peak shaving, which can, in turn, reduce costs by eliminating the need for peaking power plants. This is because the real-time price is highly related to the base-load. In particular, when the base-load is high, the grid issues a higher electricity price that encourages the customers to use less power, and vice versa, in the hope to reduce the peak load. Thus, there exists a inverse correlation between the charging rate and the base-load. The inverse correlation indicates that the charging station not only maximizes its own profit but also reduces the peak load of the local grid.

VI. Conclusion

In this article, we proposed a model-free data-driven method for charging station pricing and scheduling strategies that maximizes the objective of a charging station. The algorithm was model-free in the sense that the decision does not depend on any assumed distribution of uncertain events. We formulated the pricing and scheduling problem into an MDP. In absence of any noncausal information or distributional information, we solved the MDP problem using an RL algorithm called SARSA. To address the challenge arising from the time-varying continuous state and action spaces in the RL problem, we first showed that it suffices to optimize the total charging rates to fulfill the charging requests before departures. Then, we proposed a feature-based linear function approximator for the state–value function to further enhance efficiency and generalization ability of the proposed algorithm. The experimental substantiation performed on a real-world data showed that HSA considerably outperforms the benchmark algorithms in terms of the profit, while owning a much lower computational complexity, reducing the average charging fee of the EV owners, and shrinking the peak load of the grid.

Appendix A
Definition of e-Supervised LLF

Before we introduce $\boldsymbol{e}$-supervised LLF, we first denote the laxity of EV i in Definitions 1.

Definition 1: The laxity of EV i at tth time slot is the remaining parking time minus the minimum charging time needed to fulfill the remaining demand, i.e., $l_{it} := \tilde{p}_i^t - \tilde{d}_i^t/x^{\max}$.

Accordingly, we define $\boldsymbol{e}$-supervised LLF as follows.

Definition 2: The $\boldsymbol{\epsilon}$-supervised LLF policy determines the charging rates starting from time $t = 1$ as follows.

1) *Step 1:* $\boldsymbol{\epsilon}$-supervised LLF searches for the EV $\hat{i}$ with the least laxity.
2) *Step 2:* The policy increases $x_{\hat{i}t}$ from 0 until one of the following conditions satisfies: 1) $\tilde{d}_{\hat{i}}^t = 0$, 2) $x_{\hat{i}t} = x^{\max}$, 3) $e_t = \epsilon_t$, 4) the current EV is not the least laxity one.

3) *Step 3:* If the total charging rate equals to the supervised value ϵ_t, $t \leftarrow t+1$; If the total charging rate equals to ϵ_t and $t \leq T$, the policy turns back to Step 1; otherwise, we finish the schedule.

Appendix B
Proof of Theorem 1

Proof: (a) We show that if there exists a feasible schedule with the total charging rate $\boldsymbol{e} = (e_1, \ldots, e_T)$, then there always exists a feasible schedule that follows $\boldsymbol{e}$-supervised LLF policy. Let $\hat{x}_{it}$ denote a feasible schedule that does not follow $\boldsymbol{e}$-supervised LLF policy. Then, there must exist a (j, k, τ_1) that $\hat{x}_{j\tau_1} > \hat{x}_{k\tau_1}$ and $l_{j\tau_1} > l_{k\tau_1}$. Because $l_{j\tau_1} > l_{k\tau_1}$ and both EVs fulfill their demand before deadline, there must exist τ_2 that $\hat{x}_{j\tau_2} \leq \hat{x}_{k\tau_2}$. Let $\triangle l := l_{j\tau_1} - l_{k\tau_1}$. We can generate a new charging schedule as $\tilde{x}_{it} \leftarrow \hat{x}_{it}$, $\tilde{x}_{j\tau_1} \leftarrow \hat{x}_{j\tau_1} + \min(\hat{x}_{j\tau_2}, \hat{x}_{k\tau_2}, \frac{\triangle l}{2})$, $\tilde{x}_{k\tau_1} \leftarrow \hat{x}_{k\tau_1} - \min(\hat{x}_{j\tau_2}, \hat{x}_{k\tau_2}, \frac{\triangle l}{2})$, $\tilde{x}_{j\tau_2} \leftarrow \hat{x}_{j\tau_2} - \min(\hat{x}_{j\tau_2}, \hat{x}_{k\tau_2}, \frac{\triangle l}{2})$, and $\tilde{x}_{k\tau_2} \leftarrow \hat{x}_{k\tau_2} + \min(\hat{x}_{j\tau_2}, \hat{x}_{k\tau_2}, \frac{\triangle l}{2})$. The new schedule $\tilde{x}_{it}$s is still a feasible one and the total charging rates of $\tilde{x}_{it}$s is the same with the one of $\hat{x}_{it}$s. Repeating abovementioned, we finally reach an $\boldsymbol{e}$-supervised LLF schedule.

(b) We use the inductive method to show that if the total charging rates $\boldsymbol{e} = (e_1, \ldots, e_T)$ for a set of charging requests $\mathcal{C}$ from time 1 to T satisfy the following inequalities:

$$\sum_{\tau=1}^{k} e_\tau \geq \sum_{i \in \mathcal{C}, t_i^a + p_i \leq k} d_i,\ k = 1, \ldots, T$$
$$e_t \leq \min(|\{i \in \mathcal{C} | t_i^a \leq t, t \leq t_i^a + p_i\}| x^{\max}, e^{\max})$$
$$t = 1, \ldots, T \tag{22}$$

then there exists at least a set of x_{it}s that is feasible to (1).

1) For $|\mathcal{C}| = 1$, $x_{it} = e_t$ is a feasible schedule.

2) We suppose that for all $|\mathcal{C}_k| = k$ sets, under the condition that

$$\sum_{t=1}^{k} e_t \geq \sum_{i \in \mathcal{C}_k, t_i^a + p_i \leq k} d_i,\ \ k = 1, \ldots, T \tag{23a}$$
$$e_t \leq \min(|\{i \in \mathcal{C}_k | t_i^a \leq t, t \leq t_i^a + p_i\}| x^{\max}, e^{\max})$$
$$t = 1, \ldots, T \tag{23b}$$

there exists at least a set of x_{it}s that satisfies (1).

Then, for any set whose $|\mathcal{C}_{k+1}| = k+1$ and e_t that satisfies

$$\sum_{t=1}^{k} e_t \geq \sum_{i \in \mathcal{C}_{k+1} | t_i^a + p_i \leq k} d_i$$
$$e_t \leq \min(|\{i \in \mathcal{C}_{k+1} | t_i^a \leq t, t \leq t_i^a + p_i\}| x^{\max}, e^{\max})$$
$$t = 1, \ldots, T \tag{24}$$

we can construct an vector $\boldsymbol{e}^{'}$ and y_t as follows:

$$e_1^{'} = \sum_{i \in \mathcal{C}_{k+1} \backslash i^{'}, t_i^a + p_i \leq 1} d_i$$
$$e_t^{'} = \sum_{i \in \mathcal{C}_{k+1} \backslash i^{'}, t_i^a + p_i \leq t} d_i - e_{t-1}^{'}, t = 2, \ldots, T$$
$$y_t = e_t - e_t^{'} \tag{25}$$

where $i^{'} = \arg\max_i t_i^a + p_i$.

i) From (23 a), we have $\sum_{t=t_i^a}^{t_i^a + p_i} y_t \geq d_{i^{'}}$. Therefore, the charging request $i^{'}$ is fulfilled before its deadline.

ii) From (25), we have

$$\sum_{t=1}^{k} e_t^{'} \geq \sum_{i \in \mathcal{C}_{k+1}, t_i^a + p_i \leq k} d_i$$
$$e_t^{'} \leq \min(|\{i \in \mathcal{C}_{k+1} | t_i^a \leq t, t \leq t_i^a + p_i\}| x^{\max}, e^{\max})$$
$$t = 1, \ldots, T \tag{26}$$

As a result, the charging requests $\mathcal{C}_{k+1} \backslash i^{'}$ are fulfilled without deadline violation.

Combining (i) and (ii), we get that for any set where $|\mathcal{C}_{k+1}| = k+1$ and any e_t that satisfies (24), there exist at least a set of x_{it}s that satisfy

$$x_{it} \leq x^{\max} \quad \forall t = 1, \ldots, T,\ i \in \mathcal{K}_t$$
$$\sum_{i \in \mathcal{C}_{k+1}} x_{it} \leq e^{\max} \quad \forall t = 1, \ldots, T$$
$$\sum_{t=t_i^a}^{t_i^a + p_i} x_{it} = d_i \quad \forall i \in \mathcal{C}_{k+1}. \tag{27}$$

Combining 1) and 2) concludes the statement of (b).

(c) Considering a small time period from time slot t to $t + L(\mathcal{K}_t)$ and the charging requests $\{(\tilde{d}_i^t, \tilde{p}_i^t) | i \in \mathcal{K}_t\} \cup \{(D_i(r_t), \tilde{p}_i^t) | i \in \mathcal{I}_t\}$, we can derive from (b) that if (r_t, e_t) satisfies

$$e_t \geq \sum_{i \in \mathcal{I}_t, p_i \leq k} D_i(r_t) - \sum_{\tau=1}^{k} e_{t+\tau}^{\text{up}} + \sum_{i \in \mathcal{J}_t, \tilde{p}_i^t \leq k} \tilde{d}_i^t$$
$$\forall k = 0, \ldots, L(\mathcal{K}_t) \tag{28a}$$
$$e_t \leq \min(e^{\max}, |\mathcal{K}_t| x^{\max}) \tag{28b}$$

then there exists at least a set of x_{it}s that is feasible to (1). Moreover, we derive from (a) that one such set of x_{it}s could be obtained by $\boldsymbol{e}$-supervised LLF. □

References

[1] O. V. Vliet *et al.*, "Energy use, cost and CO$_2$ emissions of electric cars," *J. Power Sources*, vol. 196, no. 4, pp. 2298–2310, 2011.

[2] M. Tursini, F. Parasiliti, G. Fabri, and E. D. Loggia, "A fault tolerant e-motor drive system for auxiliary services in hybrid electric light commercial vehicle," in *Proc. Int. Elect. Veh. Conf.*, 2014, pp. 1–6.

[3] W. Tang, S. Bi, and Y. J. Zhang, "Online coordinated charging decision algorithm for electric vehicles without future information," *IEEE Trans. Smart Grid*, vol. 5, no. 6, pp. 2810–2824, Nov. 2014.

[4] H. Zhang, Z. Hu, Z. Xu, and Y. Song, “Optimal planning of PEV charging station with single output multiple cables charging spots,” *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2119–2128, Sep. 2017.

[5] N. Zou, L. Qian, and H. Li, “Auxiliary frequency and voltage regulation in microgrid via intelligent electric vehicle charging,” in *Proc. Int. Conf. Smart Grid Commun.*, Venice, Italy, Nov. 2014, pp. 662–667.

[6] N. Liu *et al.*, “A heuristic operation strategy for commercial building microgrids containing EVs and PV system,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2560–2570, Apr. 2015.

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

[8] W. Tang, S. Bi, and Y. J. Zhang, “Online charging scheduling algorithms of electric vehicles in smart grid: An overview,” *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 76–83, Dec. 2016.

[9] E. Akhavan-Rezai, M. F. Shaaban, E. F. El-Saadany, and F. Karray, “Managing demand for plug-in electric vehicles in unbalanced LV systems with photovoltaics,” *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1057–1067, Jun. 2017.

[10] A. Ghavami and K. Kar, “Nonlinear pricing for social optimality of PEV charging under uncertain user preferences,” in *Proc. Inf. Sci. Syst.*, Princeton, NJ, USA, 2014, pp. 1–6.

[11] D. A. Chekired, L. Khoukhi, and H. T. Mouftah, “Decentralized cloud-SDN architecture in smart grid: A dynamic pricing model,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1220–1231, Mar. 2018.

[12] M. R. Sarker, M. A. Ortega-Vazquez, and D. S. Kirschen, “Optimal coordination and scheduling of demand response via monetary incentives,” *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1341–1352, May 2015.

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

[14] Q. Huang, Q.-S. Jia, and X. Guan, “Robust scheduling of EV charging load with uncertain wind power integration,” *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1043–1054, Mar. 2018.

[15] A. Chiş, J. Lundén, and V. Koivunen, “Reinforcement learning-based plug-in electric vehicle charging with forecasted price,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3674–3684, May 2017.

[16] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, “Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market,” *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1795–1805, Jul. 2015.

[17] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. Berlin Germany: Springer, 2011.

[18] B. Defourny, D. Ernst, and L. Wehenkel, “Multistage stochastic programming: A scenario tree based approach,” *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*, Hershey, PA, USA: IGI Global, 2011, p. 97.

[19] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, “A survey on demand response in smart grids: Mathematical models and approaches,” *IEEE Trans. Ind. Informat.*, vol. 11, no. 3, pp. 570–582, Jun. 2015.

[20] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1. Cambridge, MA, USA: MIT Press, 1998, no. 1.

[22] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL, USA: CRC Press, 2010.

[23] CAISO. Load settlement price report for 2017, 2017. [Online]. Available: http://www.caiso.com/Documents/LoadSettlementPriceReportfor2017.xlsx

[24] UCDavis. Richards ave station arrivals, 2019. [Online]. Available: http://anson.ucdavis.edu/~clarkf/richards.csv.gz

Shuoyao Wang received the B.Eng. and Ph.D degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2013 and 2018, respectively.

Since 2018, he has been with the Department of Risk Management, CDG, Tencent, Shenzhen, China, where he is currently an Senior Researcher. His research interests include nature language processing, adversarial (reinforcement) learning, graph neural networks, optimization theory, dynamic programming, deep learning and reinforcement learning algorithm in both Smart Grid and WeChatPay Risk Management.

Suzhi Bi (S’10–M’14–SM’19) received the B.Eng. degree in communications engineering from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2013.

From 2013 to 2015, he was a Postdoctoral Research Fellow with the Electrical and Computer Engineering Department, National University of Singapore, Singapore.

Since 2015, he has been with the College of Electronic and Information Engineering, Shenzhen University, Shenzhen, China, where he is currently an Associate Professor. His research interests mainly include the optimizations in wireless information and power transfer, mobile computing, and smart power grid communications.

Dr. Bi was a corecipient of the IEEE SmartGridComm 2013 Best Paper Award, was a two time recipient of the Shenzhen University Outstanding Young Faculty Award in 2015 and 2018, and the Guangdong Province “Pearl River Young Scholar” Award in 2018.

Yingjun Angela Zhang (S’00–M’05–SM’10) received the Ph.D. degree in electrical and electronics engineering from the Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2004.

She is currently an Associate Professor with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. She was a Visiting Scholar with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, from July to August 2007 and from July to August 2009, and a Visiting Scholar with the California Institute of Technology during the summer of 2014. Her research interests focus on optimization in wireless communication systems and smart power grids.

Dr. Zhang is currently a Fellow of The Institution of Engineering and Technology (IET) and a Distinguished Lecturer of IEEE ComSoc. She is the co-recipient of 2014 IEEE Comsoc Asia Pacific Outstanding Paper Award, 2013 IEEE SmartGridComm Best Paper Award, 2011 IEEE Marconi Prize Paper Award on Wireless Communications and a recipient of 2011 Young Researcher Award of The Chinese University of Hong Kong. As the only winner from Engineering Science, she has won the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science. She was the Chair of the Executive Editor Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. She is also the Chair of IEEE Technical Committee on Smart Grid Communications and a Member of the Steer Committee of IEEE SmartGridComm conference. She is a Guest Editor of the IEEE INTERNET OF THINGS Journal Special Issue on Internet-of-Things for Smart Energy Systems and a Guest Editor of the IEEE COMMUNICATIONS MAGAZINE, Feature Topic on New R&D Tools for Communications Research. She has served on the Organizing Committees of many top conferences, including IEEE GLOBECOM, IEEE International Conference on Communications, IEEE Vehicular Technology Society, SmartGridComm IEEE Consumer Communications and Networking Conference, IEEE International Congress on Cognitive Computing, IEEE International Conference on Computer Communication and Networks, IEEE International Conference on Wireless Communications and Signal Processing, IEEE International Conference on Communication Systems, IEEE International Conference on Mobile Ad-hoc and Sensor Systems, etc.