Proceedings of the 55th Hawaii International Conference on System Sciences | 2022

Efficient Representation for Electric Vehicle Charging Station Operations using Reinforcement Learning

Kyung-bin Kwon and Hao Zhu
Department of Electrical and Computer Engineering
The University of Texas at Austin
Emails: {kwon8908kr, haozhu}@utexas.edu

Abstract

Effectively operating an electric vehicle charging station (EVCS) is crucial for enabling the rapid transition of electrified transportation. By utilizing the flexibility of EV charging needs, the EVCS can reduce the total electricity cost for meeting the EV demand. To solve this problem using reinforcement learning (RL), the dimension of state/action spaces unfortunately grows with the number of EVs, which becomes very large and time-varying. This dimensionality issue affects the efficiency and convergence performance of generic RL algorithms. To this end, we advocate to develop aggregation schemes for state/action according to the emergency of EV charging, or its laxity. A least-laxity first (LLF) rule is used to consider only the total charging power of the EVCS, while ensuring the feasibility of individual EV schedules. In addition, we propose an equivalent state aggregation that can guarantee to attain the same optimal policy. Using the proposed aggregation scheme, the policy gradient method is applied to find the best parameters of a linear Gaussian policy. Numerical tests have demonstrated the performance improvement of the proposed representation approaches in increasing the total reward and policy efficiency over existing approximation-based method.

1. Introduction

Electrified transportation is drastically reshaping worldwide urban mobility as a key technology to enable a future low-carbon energy society. The number of electric vehicles (EVs) continues to grow rapidly [1], thanks to their high efficiency [2] and low pollution emissions [3]. This has propelled the popularity of EV charging stations (EVCS) in metropolitan areas, as supported by significant investment in urban electricity infrastructure.

Solving the problem of optimal operational strategies is crucial for maximizing the economic profit of EVCS owners while ensuring the quality-of-service for EV charging. In general, this problem aims to find the optimal policy for determining EV charging schedules to reduce the total electricity cost by utilizing the flexibility of EV charging needs [4–7]. In addition, several papers have accounted for co-located renewable generation or energy storage [8–10] or the coupling between EV traffic and electricity flow [11–13]. Nonetheless, one key challenge in formulating the EVCS problem lies in the randomness and uncertainty of EV arrivals and other inputs such as electricity market prices. It is possible to develop probabilistic models from actual data, such as the Gaussian distribution model of EV parking time and require demand in [14], or the representation of the charging demand as a mixed Gaussian model to be estimated in [10]. Although these models have led to efficient stochastic programming approaches for the EVCS problem, they could be prone to potential modeling mismatches or fail to capture the problem dynamics therein.

To tackle this challenge, this paper aims to develop a data-driven framework to solve the EVCS operation problem by leveraging reinforcement learning (RL) techniques [15]. Using actual data samples, RL has shown some success in solving this problem with no need for stochastic modeling [16–18]. Nonetheless, most existing approaches use the original problem representation of individual EVs' status and charging action. This leads to very *high* and *time-varying* dimensionality for both the state and action spaces, significantly affecting the efficiency and convergence of policy search by generic RL algorithms. By transforming the EV status to the so-termed *laxity* that measures the emergency level of its charging need, the work in [18] has proposed to consider the total charging power across the EVCS as the action instead. Furthermore, a least-laxity first (LLF) rule has been advocated to recover individual EVs' actions from the aggregated one, which can maintain the feasibility of the charging solutions. The dimensionality issue of state space is solved by approximating the action-value function, or Q-function, which lacks approximation guarantees.

To this end, our work has proposed a new state representation by aggregating the individual EV status into

URI: https://hdl.handle.net/10125/79772
978-0-9981331-5-7

the number of EVs in each laxity group. We have analytically shown that this aggregation scheme is equivalent to the original one and thus can lead to the same optimal policy by an RL algorithm. The main contribution of the present paper is two-fold. First, we have developed a comprehensive representation for both the state and action spaces of the EVCS operations problem, with guaranteed equivalence to the original model. Second, the proposed representation enjoys fixed and low problem dimensions, developing an efficient algorithm to search for the optimal policy. Our numerical results have validated the performance improvement of the proposed state representation compared to the existing approach of Q-function approximation and suggested additional state aggregation by further grouping the higher-laxity EVs with minimal performance degradation.

The rest of paper is organized as follows. Section 2 formulates the EVCS operations problem as a Markov Decision Process (MDP). Section 3 develops the LLF-based action reduction and our proposed equivalent state aggregation to deal with dimensionality issues. Based on this, Section 4 presents the reinforcement learning approach using policy gradient and linear Gaussian policy parameterization. Numerical tests using real-world data are studied in Section 5 to demonstrate the performance improvement of the proposed algorithm, and the paper is wrapped up in Section 6.

2. System Modeling

Consider the operations of an EV charging station (EVCS) as depicted in Fig. 1 over the time period $\mathcal{T} = [0, \ldots, T]$. For each time $t \in \mathcal{T}$, let $\mathcal{I}_t$ denote the set of parked EVs, with $\mathcal{J}_t$ and $\mathcal{L}_t$ denoting the sets of arriving and departing EVs, respectively. Hence, the set of EVs is updated by $\mathcal{I}_{t+1} = (\mathcal{I}_t \cup \mathcal{J}_{t+1}) \backslash \mathcal{L}_{t+1}$, thus time-varying. Upon the arrival of EV $i \in \mathcal{J}_t$, its remaining demand $d_{i,t}$ and parking time $p_{i,t}$ are determined by the owner. The goal of EVCS operations is to determine the charging action $a_{i,t}$ for every parked EV $i \in \mathcal{I}_t$, based on the real-time electricity prices $\{\rho_t\}$ received from the market operator. Each EV's status is updated according to the $\{a_{i,t}\}$ sequence, until its departure time $\tau \in \mathcal{T}$ such that either $d_{i,\tau} = 0$ or $p_{i,\tau} = 0$. For simplicity, all EVs are assumed to have the same charging power, with the possibility of extension to different charging rates as analyzed in [19]. In addition, this work assumes the charging actions will ensure each EV to be fully charged before departure; i.e., the departure time τ is the first slot with $d_{i,\tau} = 0$. This assumption is reasonable because the EVCS can always increase the total charging budget to meet all EV demands. In future, we will extend it to the general case of non-fully charged EVs by introducing a penalty cost.

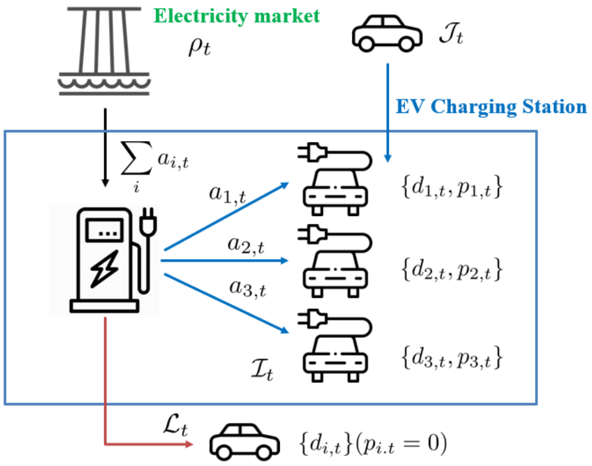

Figure 1: System Model of EV Charging Station

This work aims to develop efficient reinforcement learning (RL) algorithm for the EVCS operation problem. To this end, we model it as a Markov Decision Process (MDP) [15, Ch. 3] denoted as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, as detailed here.

State space $\mathcal{S}$ contains the set of feasible values for both the EV-internal and external status variables. This includes the remaining demand and parking time for each EV, as well as the electricity market price ρ_t. Hence, the state per time t is given by $s_t = [\rho_t, \{d_{i,t}, p_{i,t}\}_{i \in \mathcal{I}_t}]$.

Action space $\mathcal{A}$ includes the set of decisions that the active EVs can take. Without loss of generality (Wlog), consider a simple binary decision rule for each EV as given by $a_{i,t} \in \{0 \text{ (do nothing)}, 1 \text{ (charge)}\}$. It can be extended to a multi-level charging rate with $|\mathcal{A}| > 2$ or a continuous charging action. For simplicity, this paper focuses on the case of binary action.

Transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ captures the system dynamics under the Markov property [15, Ch. 3]. In the case of stochastic electricity market prices, we assume $\Pr(\rho_{t+1} | \{\rho_\tau\}_{\tau=1}^t) = \Pr(\rho_{t+1} | \rho_t)$. This is reasonable since the market price has short-term memory [20]. A longer memory is possible too; such as the prices that follow $\Pr(\rho_{t+1} | \{\rho_\tau\}_{\tau=1}^t) = \Pr(\rho_{t+1} | \rho_t,\ \rho_{t-1})$. In this case, both ρ_t and ρ_{t-1} are included as the part of the state per time t to satisfy the Markov transition property.

In addition, the EV status is updated according to the charging action in a deterministic fashion. For simplicity, let $d_{i,t}$ and $p_{i,t}$ denote the number of time slots for EV i to attain full charging and stay parked at time t, respectively. This way, their transitions are given by

$$d_{i,t+1} = d_{i,t} - a_{i,t}, \text{ and } p_{i,t+1} = p_{i,t} - 1. \quad (1)$$

This update rule also holds for general action spaces if $a_{i,t}$ is not binary.

Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ indicates the instantaneous reward used for defining the optimal actions. Wlog, assume all EVs have the same charging rate and thus the reward related to the total charging cost in time t is given by $r_t(s_t,\ a_t) = -\rho_t(\sum_{i \in \mathcal{I}_t} a_{i,t})$. The reward objective can also consider other economic factors such as peak demand reduction and load shaping benefits.

Discount factor $\gamma \in (0, 1]$ is a constant to accumulate the total reward along the time horizon. Smaller γ values imply that future rewards are less important than current ones at a discounted rate [15, Ch. 3]. For this finite time-horizon problem, $\gamma = 1$ will be used for simplicity.

For the MDP-based model, we can formulate the EVCS operation problem. The goal is to find the optimal policy π for mapping $a_t \sim \pi(s_t)$ with s_t. To simplify the policy search, we are particularly interested in the set of parameterized policies given by $\pi_\mu(\cdot) = \pi(\cdot; \mu)$, which optimizes over parameter μ as given by

$$\begin{aligned} \max_{\mu} J(\mu) =& V^{\pi}(s_0) \\ :=& \mathbb{E}_{a_t \sim \pi_\mu(s_t), \mathcal{P}} \left[\sum_{t=0}^{T} \gamma^t r_t(s_t,\ a_t) \Big| s_0 \right] \end{aligned} \quad (2)$$

where $V^{\pi}(s_0)$ denotes the value function for given initial state s_0. The formulation (2) allows for adopting popular RL algorithms. The parameterized model and problem set-up will be discussed with more details in Section 4 along with the policy gradient (PG) solution method [21]. Notably, the dimensions of state and action in (2) can be very high and are time-varying, making it challenging to search for an effective policy using RL. The following section will develop efficient state/action representation for the EVCS problem.

3. Efficient MDP Representation

Solving the MDP problem is challenged by the state/action representations of high dimension and time-varying. As the policy maps from state to action, the number of parameters in μ would grow with both state/action dimensions. This increasing rate would significantly slow down the search for an effective policy by generic RL algorithms. To tackle these issues, we propose considering the action reduction using the least-laxity first (LLF) rule and proposing an equivalent state aggregation through laxity-based grouping.

3.1 LLF-based action reduction

We can reduce the action space to $\mathcal{A}'$ that only consists of the total charging action $a_t = \sum_{i \in \mathcal{I}_t} a_{i,t}$. This way, the instantaneous reward becomes $r_t = -\rho_t \cdot a_t$. To recover each $a_{i,t}$ from a_t, we adopt the LLF rule proposed in [18] to rank the priority of EVs according to the laxity, as defined by $\ell_{i,t} := p_{i,t} - d_{i,t}$. The smaller $\ell_{i,t}$ is, the fewer flexible slots EV i can use to skip charging before departure, and thus the more emergent it is at time t compared to other EVs. If $\ell_{i,t} = 0$, or $p_{i,t} = d_{i,t}$, then EV i needs to be charged throughout its remaining parking time to be fully charged before departure. The LLF rule aims to increase the flexibility of EV charging by serving the least flexible ones first.

Time period		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
	a_t	2	1	0	2	0
EV_1	$d_{1,t}$	3	2	1	1	0
	$p_{1,t}$	4	3	2	1	0
	$\ell_{1,t}$	1	1	1	0	0
	$a_{1,t}$	1	1	0	1	0
EV_2	$d_{2,t}$	2	1	1	1	0
	$p_{2,t}$	4	3	2	1	0
	$\ell_{2,t}$	2	2	1	0	0
	$a_{2,t}$	1	0	0	1	0

Table 1: Two-EV example by following LLF rule.

Time period		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
	a_t	2	1	0	2	0
EV_1	$d_{1,t}$	3	2	2	2	1
	$p_{1,t}$	4	3	2	1	0
	$\ell_{1,t}$	1	1	0	-1	-1
	$a_{1,t}$	1	0	0	1	0
EV_2	$d_{2,t}$	2	1	0	0	0
	$p_{2,t}$	4	3	2	1	0
	$\ell_{2,t}$	2	2	0	0	0
	$a_{2,t}$	1	1	0	0	0

Table 2: Two-EV example not following the LLF rule.

To demonstrate the advantage of LLF-based action recovery, we use a simple example of only two EVs in the charging station as indexed by EV_1 and EV_2, respectively. A total horizon of $T = 4$ is considered, and a possible initial state is given in Table 1. Under a given sequence of total charging actions a_t, Table 1 lists the individual charging actions following the LLF rule, while Table 2 shows one case of not following it. In Table 2, EV_2 is charged at $t = 1$ instead of EV_1 even though $\ell_{2,1} > \ell_{1,1}$. As a result, EV_1 is not fully charged at the end, while the total charging sequence $\{a_t\}$ has led to both EVs being fully charged in Table 1. This comparison points out the importance of having the LLF rule in disaggregating the total a_t.

With the given total charging budget a_t, **Algorithm 1** demonstrates a procedure for selecting EVs to

Algorithm 1: Least-laxity first (LLF) rule

```
Inputs: Total charging power $a_t$, the set of EVs in $\mathcal{I}_t$ along with their remaining demand $d_{i,t}$ and parking time $p_{i,t}$.
Initialize: the allocated charging budget $a = 0$.
Compute the laxity for each EV $i \in \mathcal{I}_t$ as $\ell_{i,t} := p_{i,t} - d_{i,t}$ and set $a_{i,t} = 0$ to indicate that it is not yet selected for charging.
while $a \leq a_t$ do
    Search for the least-laxity EV $k = \arg\min_{i:a_{i,t}=0} \ell_{i,t}$ from the remaining unchosen EVs by arbitrarily breaking the tie if there is any.
    Set $a_{k,t} = 1$.
    $a \leftarrow a + 1$
end
```

charge at time t according to the LLF rule. The LLF based action reduction allows to recover feasible individual EV schedules, as shown in [18] and restated here for completeness.

Proposition 1. *If the EVCS total charging schedule $\{a_t\}_{t\in\mathcal{T}}$ is feasible, i.e., there exist corresponding feasible charging schedules for individual EVs that ensure each EV to be fully charged before departure, then the LLF procedure in **Algorithm 1** can produce such a feasible charging schedule for all the EVs.*

Instead of formally showing Proposition 1, we provide some intuition behind it. For given $\{a_t\}_{t\in\mathcal{T}}$, if there exist corresponding feasible individual EV schedules that do not follow the LLF rule, then we can transform the latter to feasible individual schedules that follow the LLF rule. Consider an arbitrary feasible EV schedule $\{a_{i,t}\}_i$ for each EV i that corresponds to the given $\{a_t\}_{t\in\mathcal{T}}$, i.e., it holds that $\sum_i a_{i,t} = a_t$ at every $t \in \mathcal{T}$. If the former does not follow the LLF rule, then there exist two EVs, say j and k, that violate the LLF rule at certain time t'. Specifically, we have $a_{j,t'} = 1$, and $a_{k,t'} = 0$ with the laxity $\ell_{j,t'} > \ell_{k,t'}$. The feasibility implies that $\ell_{j,t} \geq 0$ and $\ell_{k,t} \geq 0$, $\forall t \in \mathcal{T}$. Hence, let us switch the charging for those two EVs at time t', i.e., instead we pick EV k to charge by setting $a_{j,t'} = 0$, and $a_{k,t'} = 1$. First, this switch does not change the total charging action. Second, as $\ell_{j,t'} > \ell_{k,t'}$ at time t', this change still ensures feasibility or that the laxity values are always non-negative throughout the horizon $\mathcal{T}$. Hence, this example shows that by following the LLF rule, one can always recover the feasible individual EV schedules. Detailed proof for this result can be found in [18].

3.2 Equivalent state aggregation

In addition to action reduction, we also develop a state aggregation scheme to address the variable and high dimensionality issues of $\mathcal{S}$. We pursue the ideal *equivalent state aggregation* [22] such that the new state space $\mathcal{S}'$ can maintain the necessary information in $\mathcal{S}$. The aggregation needs to ensure that both $\mathcal{S}$ and $\mathcal{S}'$ attain the same value functions $V^\pi(\cdot)$ and thus the same optimal policies for any given action in $\mathcal{A}'$. Two conditions need to hold [22], as defined here.

Definition 1. *A state aggregation scheme $\mathcal{S} \to \mathcal{S}'$ satisfies **reward homogeneity** if for any pair of original states $\{s_t^{(i)}, s_t^{(j)}\}$ that will be aggregated into the same new state in $\mathcal{S}'$, it holds that*

$$r_t(s_t^{(i)}, a_t) = r_t(s_t^{(j)}, a_t),\ \forall a_t \in \mathcal{A}' \tag{3}$$

Definition 2. *A state aggregation scheme $\mathcal{S} \to \mathcal{S}'$ satisfies **dynamic homogeneity** if for any pair of original states $\{s_t^{(i)}, s_t^{(j)}\}$ that will be aggregated into the same new state in $\mathcal{S}'$, it holds that*

$$\Pr(s_{t+1}|s_t = s^{(i)}, a_t) = \Pr(s_{t+1}|s_t = s^{(j)}, a_t),$$
$$\forall s_{t+1} \in \mathcal{S},\ a_t \in \mathcal{A}' \tag{4}$$

To achieve these homogeneity conditions, we propose to aggregate the parked EVs at time t into the number of EVs for every integer-valued laxity level in $[0,\ L]$, where $L := \max_{i,t} \ell_{i,t}$ denotes the maximally possible laxity level at the EVCS. Note that as all EVs are assumed to be fully charged before departure, the laxity is always non-negative with the minimum equal to zero. Upon determining each EV's laxity as in Section 3.1, we define the aggregated state

$$s_t' = [\rho_t, n_t^{(0)}, n_t^{(1)}, \cdots, n_t^{(L)}] \in \mathcal{S}' \tag{5}$$

with $n_t^{(\ell)}$ denoting the number of EVs with laxity equal to ℓ. In order to show the new MDP is equivalent to the original one, let us consider the two homogeneity conditions. First, the reward homogeneity is easily satisfied as $r_t = -\rho_t a_t$ is not affected by the aggregation. Second, dynamic homogeneity also holds due to the LLF rule for action reduction. Upon recovering the individual EV actions $\{a_{i,t}\}_i$ from a_t, the original MDP transition in (1) states that $(d_{i,t+1}, p_{i,t+1}) = (d_{i,t} - a_{i,t}, p_{i,t} - 1)$ for each $i \in \mathcal{I}_t$. For the new MDP through aggregation, the state transition instead depends on the allocation of a_t to each subset of EVs of the same laxity. Specifically, if $a_{i,t} = 1$ or EV i is charged at time t, its laxity stays unchanged as $\ell_{i,t+1} = \ell_{i,t}$. Otherwise, its laxity is reduced by one as $\ell_{i,t+1} = \ell_{i,t} - 1$. We can update the subset of EVs with laxity ℓ for time $t + 1$ based on those of laxity ℓ at time t that are charged, those of laxity $(\ell + 1)$ that are not charged, along with the new arrival or departure at time $(t + 1)$, as given by

$$n_{t+1}^{(\ell)} = a_t^{(\ell)} + [n_t^{(\ell+1)} - a_t^{(\ell+1)}] + x_{t+1}^{(\ell)} - y_{t+1}^{(\ell)}, \ \forall l \tag{6}$$

where $a_t^{(\ell)}$ denotes the number of EVs of laxity ℓ that are charged in time t, while $x_{t+1}^{(\ell)}$ and $y_{t+1}^{(\ell)}$ representing the number of EVs of laxity ℓ that arrive/depart at time $(t + 1)$, respectively. Similar to the LLF-based action recovery in Section 3.1, we allocate the total charging budget a_t into each $a_t^{(\ell)}$ in an ordered fashion, as given by

$$a_t^{(\ell)} = \min\left\{n_t^{(\ell)}, \min\left\{a_t - \sum_{\ell=0}^{\ell-1} a_t^{(\ell)}, 0\right\}\right\}, \ \forall \ell. \tag{7}$$

Basically, starting from the smallest laxity level $\ell = 0$, we set $a_t^{(\ell)} = n_t^{(\ell)}$ until the total charging budget is met. Based on the two homogeneity conditions, we can formally establish the following proposition using the result from [22].

Proposition 2. *Consider the original MDP* $(\mathcal{S}, \mathcal{A}', \mathcal{P}, \mathcal{R}, \gamma)$ *and the new MDP* $(\mathcal{S}', \mathcal{A}', \mathcal{P}, \mathcal{R}, \gamma)$*. If* $\mathcal{S}'$ *is aggregated through* $s'_t = [\rho_t, n_t^{(0)}, n_t^{(1)}, \cdots, n_t^{(L)}]$ *with the transition following* (6) *and* (7)*, then it satisfies both reward homogeneity and dynamic homogeneity and thus the two MDPs are equivalent. As a result, the new MDP through aggregation can be used to obtain the optimal policies (determine the optimal actions) that are equivalent to the original ones.*

By guaranteeing the equivalence of the two MDPs, the aggregation maintains the same value function for any initial state as mentioned earlier. Hence, the optimal policy obtained by an RL algorithm for the new MDP would be the same as the original one. This state aggregation scheme can efficiently search for the best $\pi(\cdot)$, at no sacrifice of optimality. Note that the state aggregation can be further simplified in practice by merging the higher-laxity groups. If the maximum laxity L is very large, the equivalent aggregation can still be of quite large dimension. Our numerical experiences suggest that the groups of higher laxity values play similar role in determining the optimal action, as the LLF rule implies that the recovered action (or the transition) would mostly depend on the groups of smaller laxity values. Hence, we can cap the number of laxity groups at a value $L_{\max} < L$ such that $n^{(L_{\max})} = \sum_{\ell \geq L_{\max}} n^{(\ell)}$. Although this further simplification may not be equivalent, it can be effective in addressing the immense value of laxity in practice.

4. Learning the Optimal Policy

The proposed efficient MDP representation has successfully handled the dimensionality issue for state/action, and will be leveraged to efficiently solve for the optimal policy π in (2) using general RL algorithms. Recall that the unknown policy $\pi(\cdot)$ is assumed to follow certain parameterized model, and thus the problem is to find the optimal parameter μ for the mapping $a \sim \pi_\mu(s')$. The choice of parameterized model can affect the performance of RL algorithms. Without loss of generalizability, we consider a simple model of π_μ and adopt the policy gradient (PG) method [21] to search for the best μ. We use the linear Gaussian policy [23], which is popular for continuous spaces, as defined by the conditional distribution

$$a \sim \pi_\mu(s') = \pi_\mu(a|s') = \mathcal{N}(\mu_s^\top s' + \overline{\mu}, \sigma^2) \tag{8}$$

with parameter $\mu = [\mu_s; \ \overline{\mu}]$ relating s' to the mean for the Gaussian distributed action a. The variance σ^2 can be either part of the parameter or pre-determined as exploration noise. Equivalently, the random action in (8) can be simply generated by the following *linear policy*

$$a = \mu_s^\top s' + \overline{\mu} + e \tag{9}$$

where the additive noise $e \sim \mathcal{N}(0, \sigma^2)$. Using (8), the total reward function in (2) now becomes

$$J(\mu) = \int_{a \in \mathcal{A}'} \pi_\mu(a|s') Q_\mu(s', a) \mathrm{d}a, \tag{10}$$

with the Q-function, or action-value function, given by

$$\begin{aligned} &Q^\pi(s', a) \\ &:= \mathbb{E}_{a_t \sim \pi_\mu(s_t), \mathcal{P}} \left(\sum_{t=0}^{T} \gamma^t r_t | s_0 = s', a_0 = a \right). \end{aligned} \tag{11}$$

Before discussing the PG method, it is worth mentioning that other choices of π_μ can be readily applied as well. For example, one can use a nonlinear neural network to parameterize the Q-function, known as the Deep Q-Network (DQN) approach [15, Ch. 20]. The proposed state/action aggregation would be powerful for accelerating these nonlinear policy based RL methods too, which can be greatly affected by the dimensionality issue.

To maximize $J(\mu)$, we are interested to find its gradient over μ following from the *log-derivative trick*, as

$$\nabla_\mu J(\mu) = \mathbb{E}_{a \sim \pi_\mu(s)} \left[Q^\pi(s', a) \nabla_\mu \ln \pi_\mu(a|s')\right]. \tag{12}$$

Interestingly, this gradient computation boils down to that of the logarithmic term only, which can be easily obtained for Gaussian distribution as

$$\nabla_{\mu_s} \ln \pi_\mu(a|s') = \frac{a' - (\mu_s^\top s' + \overline{\mu})}{\sigma^2} s', \qquad (13)$$

$$\nabla_{\overline{\mu}} \ln \pi_\mu(a|s') = \frac{a' - (\mu_s^\top s' + \overline{\mu})}{\sigma^2}. \qquad (14)$$

To estimate this gradient, one can replace the expectation in (12) by the sample mean obtained from the trajectory $\{s_0', a_0, s_1', a_1, \cdots, s_T', a_T\}$:

$$\hat{\nabla}_\mu J(\mu) \propto \sum_{t=1}^{T} \hat{Q}_\mu(s_t', a_t) \nabla_\mu \ln \pi_\mu(a_t|s_t). \qquad (15)$$

with the samples $\hat{Q}_\mu(s_t', a_t) = \sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau(s_\tau', a_\tau)$ estimated from the trajectory. Note that the time window for approximating $\hat{Q}_\mu(s_t', a_t)$ decreases as t increases under the finite time-horizon setting of $\mathcal{T}$. For larger t values, fewer samples are used and the scale of Q-value is expected to decrease. To cope with this issue, one can normalize the approximated Q-function by subtracting the mean and dividing it with the standard deviation of all episode rewards [24]. This can generally improve the training stability under the high variance of the policy gradient estimator.

With a given learning rate (step-size) α, the policy gradient method uses the estimated gradient in (15) and implements the iterative gradient ascent updates of μ. Per iteration n, the update becomes

$$\mu^{n+1} = \mu^n + \alpha \hat{\nabla}_\mu J(\mu^n), \qquad (16)$$

until the parameters converge. To improve the gradient update, we can incorporate multiple training samples, each of which will produce a gradient estimate. Accordingly, the sum (or average) of the gradients estimated from each training sample will be used for the update in (16). **Algorithm 2** has detailed steps for solving the proposed MDP representation under LLF-based action reduction and the equivalent state aggregation.

5. Numerical Tests

We have tested the proposed **Algorithm 2** to demonstrate the effectiveness of our new MDP representation. To set up the EVCS operation problem, we have used the hourly data of electricity market prices from the ERCOT market portal [25] and the vehicle arrival data collected at the Richards Ave Station near downtown Davis, CA [26]. Three categories of EVs are considered: emergent, normal and residential uses, each having different initial demand and parking time distribution. Fig. 2 shows an example of the number of EVs in each category for a typical workday. Accordingly, the RL exploration time is the full-day period at 15-minute intervals, leading to a total horizon of $T = 96$. The EV data show the maximum laxity $L = 12$, and thus there are a total of 14 variables in s'.

Algorithm 2: Optimal EVCS policy search

```
Hyperparameters: discount factor $\gamma$,
  step-size $\alpha$, and exploration time period $T$.
Inputs: the price sequence $\{\rho_t\}_{t=0}^{T}$, and the
  EV arrivals in $\{\mathcal{J}_t\}_{t=0}^{T}$ along with the initial
  states of EVs
Initialize: $\mu^0$ at iteration $n = 0$.
while $\mu^n$ not converged do
    Initialize $t = 0$ with the original state $s_0$.
    for $t = 0, \cdots, T-1$ do
        Find the aggregated state $s_t'$ using (5);
        Sample $a_t \sim \pi_{\mu_n}(s_t')$ using (8);
        Use the LLF rule in Algorithm 1 to
          recover the individual EV charging
          actions $\{a_{i,t}\}_{i \in \mathcal{I}_t}$;
        Compute the instantaneous reward $r_t$;
        Update the new state $s_{t+1}$ using (1).
    end
    Use the sample trajectory to estimate
      gradient $\hat{\nabla}_\mu J(\mu^n)$ and perform the
      update in (16);
    Update iteration $n \leftarrow n + 1$.
end
```

We have compared **Algorithm 2** to the existing approach by estimating an approximate Q-function in [18], denoted by **Algorithm *QE***. In [18], the same LLF-based action reduction is used while four binary feature functions approximate the Q-function to deal with the state dimensionality issue. These feature functions correspond to the charging cost or constraints on EV charging for the EVCS problem, while the total Q-function is assumed to be a linear combination of them. Hence, the RL problem becomes to estimate the best linear coefficients as the parameter based on the Bellman optimality condition for Q-function. Although this approach can deal with time-varying states, the approximation therein is heuristic and could be inaccurate.

5.1 Training results

We have used 20 different daily profiles to train the RL algorithms. Fig. 3 and Fig. 4 plot the episode rewards and parameter values for the proposed **Algorithm 2** and **Algorithm *QE***, respectively. Clearly, both RL algorithms are shown to converge as rewards gradually increasing and parameter values stabilizing.

One important observation from the episode parameter values in Fig. 3 is that they are almost zero

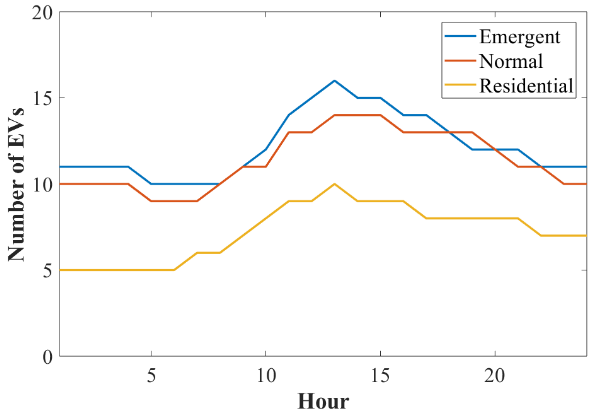

Figure 2: Hourly arrivals for the three categories of EVs during one day.

State	Parameter	**State**	Parameter
ρ_t	-1.9735	$n_t^{(6)}$	0.2021
$n_t^{(0)}$	1.8628	$n_t^{(7)}$	0.1404
$n_t^{(1)}$	0.5772	$n_t^{(8)}$	0.1386
$n_t^{(2)}$	0.3674	$n_t^{(9)}$	0.1592
$n_t^{(3)}$	0.2651	$n_t^{(10)}$	0.0975
$n_t^{(4)}$	0.3485	$n_t^{(11)}$	0.0693
$n_t^{(5)}$	0.1191	$n_t^{(12)}$	0.0797

Table 3: Parameter values obtained by **Algorithm 2**.

for most states, except for state ρ_t and $n_t^{(0)}$. Specifically, the negative most parameter is for ρ_t as the total charging budget a_t should decrease when the price is high. In addition, the positive most parameter is for $n_t^{(0)}$ as a_t should increase when there are many EVs with emergent charging needs. Compared to these two parameters, the states for other laxity groups have minimal parameter values, with the parameter value decreasing at larger laxity ℓ, as listed in Table 3. This learning result is very reasonable as this problem depends mainly on the EVs approaching their department deadlines. As mentioned in Section 3.2, it is possible to further reduce the number of states by merging the high-laxity EVs (larger than a threshold $L_{\max}$) into one single group. This simplification may violate the dynamic homogeneity condition, but it may not affect much the optimality of the resultant RL solution for practical systems based on this observation on minimal parameter values for high-laxity group states.

5.2 Testing results

Using the two policies obtained by the RL training, we have compared their testing performances using five additional daily profiles. Table 4 lists the total reward values attained by each of the two policies for each test trajectory. Clearly, the solution by **Algorithm 2** achieves higher total reward values, increasing those acquired by **Algorithm *QE*** by around 4.15% to 4.32%. Thanks to the equivalent state aggregation, **Algorithm 2** can effectively reduce the total charging cost for the EVCS. It enjoys high modeling accuracy as compared to the Q-function approximation in [18].

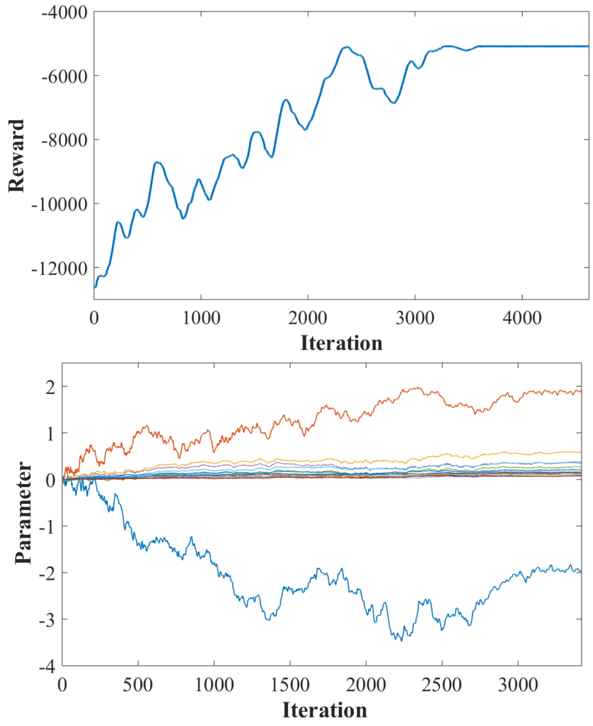

Figure 3: (Top) The episode reward and (bottom) episode parameter values for **Algorithm 2**.

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Average**
Alg. 2	-5016.2	-5022.6	-5009.5	-5012.8	-5007.8	-5013.8
Alg. QE	-5240.1	-5240.3	-5234.2	-5239.3	-5230.6	-5236.9
Increase (%)	4.27	4.15	4.29	4.32	4.26	4.26

Table 4: Testing reward values and percentage reward increases of the solution obtained by **Algorithm 2**, as compared to **Algorithm *QE***.

To better illustrate the improvement of **Algorithm 2**, Fig. 5 plots the daily total charging action comparisons along with the electricity market price. Interestingly, **Algorithm 2** is very sensitive to the price peaks and has chosen to dramatically reduce a_t. Meanwhile, **Algorithm QE** fails to reduce the charging needs over the peak-price period, as highlighted by the shaded area. This example further verifies that our proposed EVCS operation can improve the cost performance

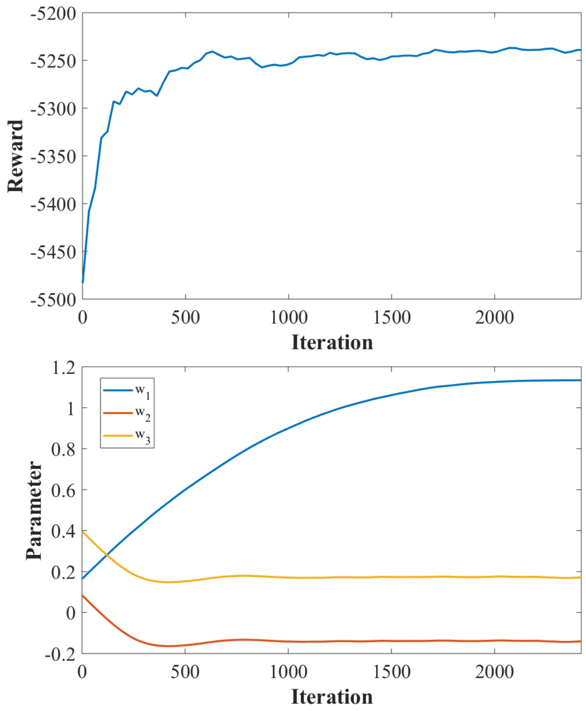

Figure 4: (Top) The episode reward and (bottom) episode parameter values for Algorithm QE.

while enjoying efficient RL solution time by considering the equivalent MDP problem.

6. Conclusions and Future Work

This paper has developed a practical modeling approach for the optimal EV charging station operation problem, allowing for efficient solutions using reinforcement learning (RL). To deal with the high and variable dimensions of states/actions, we propose to design efficient aggregation schemes by utilizing the EV's laxity that measures the emergency level of its charging need. First, the least-laxity first (LLF) rule has made it possible to consider only the total charging action across the EVCS, which is shown to recover feasible individual EV charging schedules if existing. Second, we propose aggregating the state into the number of EVs in each laxity group, which satisfies reward and dynamic homogeneities and thus leads to equivalent policy search. We have developed the policy gradient method based on the proposed MDP representation to find the optimal parameters for the linear Gaussian policy. Case studies based on real-world data have demonstrated the performance improvement of the proposed MDP representation over the earlier approximation-based approach for the EVCS problem. The RL parameter results imply that further state aggregation can deal with many laxity levels in practical systems at a minimal loss of optimality.

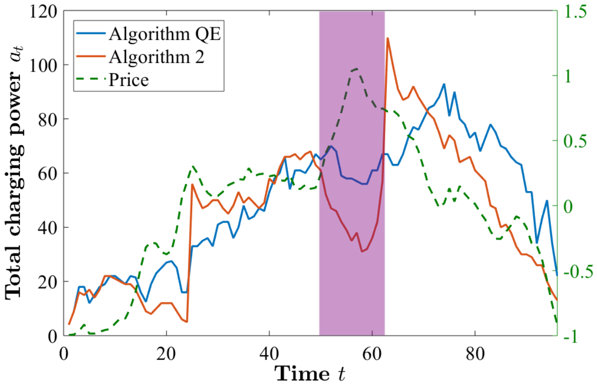

Figure 5: The daily profiles of total charging power respectively produced by Algorithms 2 and QE for one testing day as compared to the electricity market prices.

Exciting future research directions open up regarding more general EVCS problem set-ups such as penalizing non-fully charged EVs at departure, as well as variable EV charging rate and action. The former makes it relevant to consider a constrained RL formulation that limits the number (or total demand) of unsatisfied EVs at departure or the corresponding statistical risk, following from the safe RL framework [27]. As for the variable charging power, it would be interesting to pursue the connection to recent work [19] that uses a smoothed LLF approach to deal with different charging rates.

References

[1] I. E. Agency, "Global EV Outlook 2020," International Energy Agency, Tech. Rep., June 2020. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2020

[2] Z. Stevic and I. Radovanovic, "Energy Efficiency of Electric Vehicles," *New Generation of Electric Vehicles, edited by Z. Stevic (Intech, Rijeka, 2012)*, 2012.

[3] X. Hu, N. Chen, N. Wu, and B. Yin, "The Potential Impacts of Electric Vehicles on Urban Air Quality in Shanghai City," *Sustainability*, vol. 13, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/2/496

[4] Z. Ma, D. Callaway, and I. Hiskens, "Optimal charging control for plug-in electric vehicles," in *Control and Optimization Methods for Electric Smart Grids*. Springer, 2012, pp. 259–273.

[5] Y. Xu and F. Pan, "Scheduling for charging plug-in hybrid electric vehicles," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. IEEE, 2012, pp. 2495–2501.

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

[7] H. Zhang, Z. Hu, Z. Xu, and Y. Song, “Optimal Planning of PEV Charging Station With Single Output Multiple Cables Charging Spots,” *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2119–2128, 2017.

[8] Q. Yan, B. Zhang, and M. Kezunovic, “Optimized Operational Cost Reduction for an EV Charging Station Integrated With Battery Energy Storage and PV Generation,” *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2096–2106, 2019.

[9] Q. Chen, N. Liu, C. Hu, L. Wang, and J. Zhang, “Autonomous Energy Management Strategy for Solid-State Transformer to Integrate PV-Assisted EV Charging Station Participating in Ancillary Service,” *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 258–269, 2017.

[10] C. Luo, Y.-F. Huang, and V. Gupta, “Stochastic Dynamic Pricing for EV Charging Stations With Renewable Integration and Energy Storage,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1494–1505, 2018.

[11] M. Alizadeh, H.-T. Wai, M. Chowdhury, A. Goldsmith, A. Scaglione, and T. Javidi, “Optimal pricing to manage electric vehicles in coupled power and transportation networks,” *IEEE Transactions on Control of Network Systems*, 2015.

[12] F. He, D. Wu, Y. Yin, and Y. Guan, “Optimal deployment of public charging stations for plug-in hybrid electric vehicles,” *Transportation Research Part B: Methodological*, vol. 47, pp. 87–101, 2013.

[13] K. Zhang, L. Lu, C. Lei, H. Zhu, and Y. Ouyang, “Dynamic operations and pricing of electric unmanned aerial vehicle systems and power networks,” *Trans. Research Part C*, vol. 92, pp. 472–485, 2018.

[14] Q. Huang, Q.-S. Jia, and X. Guan, “Robust Scheduling of EV Charging Load With Uncertain Wind Power Integration,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1043–1054, 2018.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[16] Z. Wan, H. Li, H. He, and D. Prokhorov, “Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning,” *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246–5257, 2019.

[17] H. Li, Z. Wan, and H. He, “Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning,” *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2020.

[18] S. Wang, S. Bi, and Y. A. Zhang, “Reinforcement Learning for Real-Time Pricing and Scheduling Control in EV Charging Stations,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 849–859, 2021.

[19] N. Chen, C. Kurniawan, Y. Nakahira, L. Chen, and S. H. Low, “Smoothed least-laxity-first algorithm for EV charging,” *CoRR*, vol. abs/2102.08610, 2021. [Online]. Available: https://arxiv.org/abs/2102.08610

[20] T. Jónsson, “"Forecasting and decision-making in electricity markets with focus on wind energy",” Ph.D. dissertation, 2012.

[21] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*. MIT Press, 2000, vol. 12.

[22] R. Givan, T. Dean, and M. Greig, “Equivalence notions and model minimization in Markov decision processes,” *Artificial intelligence*, vol. 147, no. 1, pp. 163–223, 2003.

[23] K. Doya, “Reinforcement Learning in Continuous Time and Space,” *Neural Comput.*, vol. 12, no. 1, p. 219–245, Jan. 2000. [Online]. Available: https://doi.org/10.1162/089976600300015961

[24] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, “High-Dimensional Continuous Control Using Generalized Advantage Estimation,” 2018.

[25] Market Prices - ERCOT. http://www.ercot.com/mktinfo/prices. Accessed: 2021-06-07.

[26] UCDavis. Richards ave station arrivals, 2019. http://anson.ucdavis.edu/âĹijclarkf/richards.csv.gz. Accessed: 2021-06-07.

[27] J. Garcia and F. Fernández., “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, no. 1, p. 1437–1480, 2015.

Acknowledgments

This work has been supported by NSF Grants 1802319, 1807097, and 1952193.