SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2022, Vol. 65 122201:1–122201:16
https://doi.org/10.1007/s11432-020-3106-7

On large action space in EV charging scheduling optimization

Zhaoyu JIANG[1], Qing-Shan JIA[1*] & Xiaohong GUAN[1,2]

[1]*Center for Intelligent and Networked Systems, Beijing National Research Center for Information Science and Technology, Department of Automation, Tsinghua University, Beijing* 100084*, China;*
[2]*MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an* 710049*, China*

Received 6 March 2020/Revised 18 June 2020/Accepted 1 October 2020/Published online 18 May 2021

Abstract In order to reduce the air pollution by traditional fossil fuel and save the charging cost, much attention has been paid to the charging scheduling of electric vehicles (EVs) in the uncertain supplement of wind power and solar power. Owing to the randomness in the renewable power generation and the EV charging demand, simulation-based policy improvement (SBPI) has been well adopted for making charging decision. However, it is challenging to explore the large action space which grows exponentially with respect to the system scale. We consider this important problem in this paper and make the following contributions. First, we explore the structural property of the problem and develop an urgency index to rank the EVs. Second, we apply three methods to search in the action space. Third, we numerically demonstrate the performance of the urgency index and the search methods, and compare the SBPI methods with the CPLEX-based method. It shows that SBPI improves the base policies in all these cases.

Keywords EV, simulation-based policy improvement, large action space, sampling

Citation Jiang Z Y, Jia Q-S, Guan X H. On large action space in EV charging scheduling optimization. Sci China Inf Sci, 2022, 65(2): 122201, https://doi.org/10.1007/s11432-020-3106-7

1 Introduction

Electric vehicles (EVs) have become popular nowadays as they can alleviate the energy crisis and reduce the air pollution. Renewable energy such as wind power and solar power also help reduce the greenhouse gas emission [1] and save the charging cost. So it is of practical interest to schedule the charging of EVs to match the uncertain renewable power generation so that the charging demand is satisfied and the charging cost is minimized.

However, there are the following major difficulties. First, uncertainty. The generation for both wind power and solar power is highly uncertain [2, 3]. The charging demands from the EVs depend on the travel demands of the users and are also uncertain. Second, multi-stage decision making. The limited renewable power generation requires the EVs to coordinate the charging with each other. The finite battery capacity couples the charging scheduling at different time. Third, the curse of dimensionality. Both the state space and the action space increase exponentially fast with respect to the number of EVs. It soon becomes practically infeasible to enumerate every action candidate to decide the charging actions.

These challenges have attracted a lot of interest in the past few years [4, 5]. Relevant studies are briefly reviewed in Section 2. Among these proposed methods, simulation-based policy improvement (SBPI) [6] is convenient for making decisions. When the randomness is modeled and a base policy is given, the Q-factor, which is an action-utility function that measures the future performance after making a decision and helps compare the actions for a given state, can be evaluated by simulation. Then the action that maximizes the Q-factors for the current state is selected. However, it is difficult to evaluate all the Q-factors when there are a large number of action candidates. So an important question is how to improve from a base policy when there are a large number of action candidates.

* Corresponding author (email: jiaqs@tsinghua.edu.cn)

We consider this important problem in this paper and make the following major contributions. First, we explore the structural property of the EV charging scheduling problem and develop an urgency index to reduce the search region in the action space. It is shown that this urgency index provides a complete order among the actions and is consistent with the least-laxity-longer-processing-time-first (LLLP) principle [7]. Second, we apply three methods to search for the optimal action, namely model adaptive reference search (MRAS) [8, 9], model-based annealing random search (MARS) [10], and convergent optimization via most-promising-area stochastic search (COMPASS) [11], respectively. MRAS and MARS are model-based methods that can converge to the global optimum. COMPASS is a model-free method that can converge to the local optimum with probability 1. We present the parameterized distribution for action sampling in MRAS and MARS and show how to update the parameters optimally. We also show how to uniformly sample from the most promising area in COMPASS. Third, we numerically compare these search methods with uniform sampling (US) and compare the SBPI methods with the CPLEX-based (CPLEX-B) method both in the original action space and the action space reduced by the urgency index. The results show that using structural property of the system to reduce the action space does work and the search methods such as MRAS, MARS, and COMPASS outperform the US under the same computing budget. Meanwhile the SBPI improves the base policy in all these cases and performs better than the CPLEX-B method.

The rest of this paper is organized as follows. We give a brief literature review in Section 2 and formulate the problem in Section 3. We present the main results in Section 4, including the reduced action space using structural property of the problem in Subsection 4.1, the algorithms of applying MRAS, MARS, and COMPASS to search for the optimal action in Subsection 4.2, and the numerical comparison among MRAS, MARS, COMPASS, and US both in the reduced action space and the original action space in Subsection 4.3. We conclude the paper in Section 5.

2 Literature review

The EV charging process may be optimized for various objectives such as cost minimization [12–14], profit maximization for the charging station [15, 16] and the EV owners [17], peak load minimization in the distribution network [18, 19], service dropping rate minimization for the charging station with dual charging modes [20], and renewable energy penetration improvement [21]. Markov decision process (MDP) is a popular model in these studies, and also the problem formulation in this work.

It has attracted great interest in solving MDP in the past few decades. Policy iteration and value iteration are typical methods for solving the MDP [22]. However, they may still face the difficulties from the aforementioned three aspects. Both uncertainty and multi-stage decision making are difficult to handle. For uncertainty, Monte Carlo sampling [23] can be used to generate scenarios or scenario trees to depict the uncertainty [24, 25], while deep neural network can be used to optimize the decision making under uncertain scenarios [26]. Unknown transition probabilities can also be learned in the deep reinforcement learning without system model information [27]. For multi-stage decision making, event-based optimization [17] and multi-scale approaches [21] can alleviate the trouble by multi-stage decision making.

The curse of dimensionality is also difficult to handle. Although many studies focus on the large state space [27, 28], we focus on the large action space. The idea to handle the large action space mainly comes from three aspects. First, divide and conquer. One may use the structure of the problem to divide the objective function and the state [29]. Then the action space is curtailed in the small-scale problem. Second, reduce the action space. Aggregating the EVs by their properties [15, 21], and using a priority, such as that used in the LLLP principle [7], to rank the EVs are examples. Third, search for the optimal action. When the probabilities of the actions can be parameterized, model-based methods like MRAS and MARS can be implemented to search for the optimal action [30]. When the probabilities of the actions cannot be parameterized, model-free methods like COMPASS [11] and particle swarm optimization (PSO) [31] can be used to search for the optimal action. In this paper, we focus on how to find the optimal action in the EV charging problem when there is a large action space.

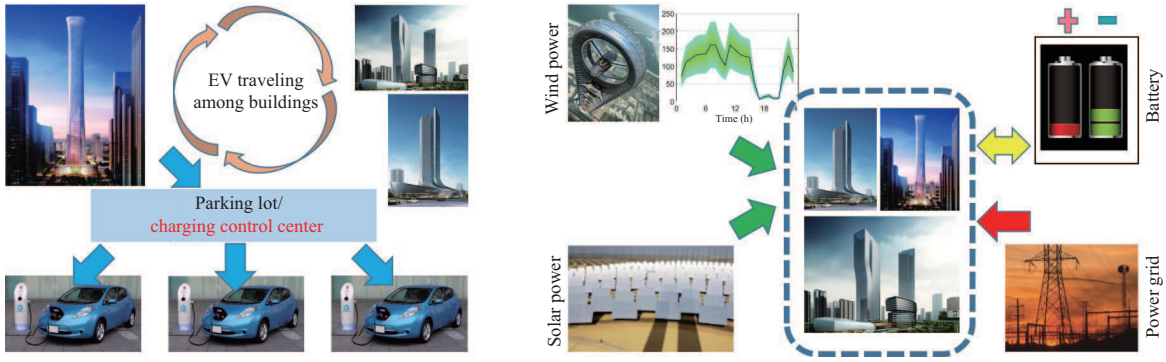

Figure 1 (Color online) The system structure.

3 Problem formulation

The system structure is shown in Figure 1. There are several buildings in the system and each of them has a parking lot. The EVs travel among these buildings. Whenever the EVs need to be charged in a parking lot, the building can obtain the charging information and manage the charging processes of the EVs. Besides the power grid, both building-mounted wind turbines and building-integrated photovoltaic panels can supply energy for the EVs. Owing to the cost and the degradation of the battery, we do not consider the discharging of the EVs. A large-capacity battery is equipped with the building to store energy when there is excess renewable energy or when the electricity price of the power grid is low, and release energy when there is little renewable energy or when the electricity price of the power grid is high. The goal is to minimize the total charging cost and at the same time satisfy the traveling demands of the EVs. We formulate the problem as a Markov decision process. We have the assumptions as follows.

Assumption 1. The charging power of the EVs, $P_{\rm ev}$, is constant and identical.

Assumption 2. The number of the charging piles in each parking lot is large enough for the EVs.

Assumption 3. Renewable energy from the building-mounted wind turbines and the building-integrated photovoltaic panels is free.

Assumption 4. Sufficient power can be obtained from the power grid to satisfy the charging demands of the EVs.

Assumption 1 is a typical assumption in [32] and the constant charging power is beneficial for the lifetime of the battery. Assumption 2 is used for formulating a general model. Assumption 3 is based on the fact that the costs for operation, maintenance, etc. on the renewable energy are much smaller than the fossil-based electricity generation process [33, 34]. Assumption 4 is similar to Assumption 2 and is reasonable for the urban power grid.

We define the system state at time $t \in \{0, 1, \ldots, T-1\}$ as $S_t = (P_{{\rm w},t}, P_{{\rm pv},t}, R_t, L_t, E_t, D_t)$, where T is the number of stages considered, $P_{{\rm w},t} = (P^1_{{\rm w},t}, P^2_{{\rm w},t}, \ldots, P^{N_{\rm b}}_{{\rm w},t})$ and $P_{{\rm pv},t} = (P^1_{{\rm pv},t}, P^2_{{\rm pv},t}, \ldots, P^{N_{\rm b}}_{{\rm pv},t})$ are the power generated by the building-mounted wind turbines and the building-integrated photovoltaic panels, respectively, $R_t = (R^1_t, R^2_t, \ldots, R^{N_{\rm b}}_t)$ is the state of charge (SOC) of the battery equipped with the buildings, $N_{\rm b}$ is the number of buildings in the system, $D_t = (D^1_t, D^2_t, \ldots, D^{N_{\rm ev}}_t)$ denotes the locations of the EVs, $L_t = (L^1_t, \ldots, L^{N_{\rm ev}}_t)$ and $E_t = (E^1_t, \ldots, E^{N_{\rm ev}}_t)$ denote the remaining parking time and charging demand of the EVs, respectively, and $N_{\rm ev}$ is the number of EVs in the system. Let $\mathcal{S}$ denote the state space. Then we have

$$\begin{cases} 0 < L^i_t \leqslant T-1-t, \\ 0 \leqslant E^i_t \leqslant E_{\rm cap}, \end{cases} \text{ if } 1 \leqslant D^i_t \leqslant N_b, \quad \text{and} \quad \begin{cases} L^i_t = 0, \\ E^i_t = 0, \end{cases} \text{ if } D^i_t = 0, \tag{1}$$

where $E_{\rm cap}$ is the battery capacity of the EVs.

The following equations [35] are used to estimate the wind power generation at each building:

$$P^j_{{\rm w},t} = \begin{cases} W_{\rm cap}, & \text{if } V_{\rm rated} \leqslant V^j_t < V_{\text{cut-off}}, \\ W_{\rm cap}\left(\frac{V^j_t}{V_{\rm rated}}\right)^3, & \text{if } V_{\text{cut-in}} \leqslant V^j_t < V_{\rm rated}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

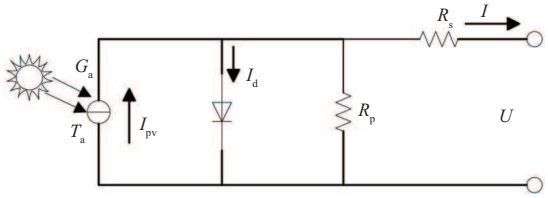

Figure 2 A typical model of two diode photovoltaic panels [36].

where W_{cap} is the capacity of the building-mounted wind turbine, V_t^j is the real-time wind speed at the blades of the wind turbine at building j, V_{rated}, $V_{\text{cut-in}}$, and $V_{\text{cut-off}}$ are the rated wind speed, cut-in wind speed, and cut-off wind speed of the wind turbine, respectively.

A typical model of two diode photovoltaic panels presented in [36] is used to estimate the solar power generation, which is shown in Figure 2. A building-integrated photovoltaic panel consists of several panels connected in series and in parallel. The main mathematical formulation shown in (3) is based on the property of P-N junction and the Thevenin's theorem.

$$\begin{aligned} &P_{\text{pv}} = U\left(I_{\text{pv}} - I_{\text{d}} - \frac{U + R_{\text{s}}I}{R_{\text{p}}}\right), \qquad I_{\text{pv}} = N_{\text{p}}(I_{\text{pv,n}} + K_{\text{i}}(T_{\text{a}} - T_{\text{n}}))\frac{G_{\text{a}}}{G_{\text{n}}}, \\ &I_{\text{d}} = I_0\left(\exp\left\{\frac{q(U + R_{\text{s}}I)}{akN_{\text{s}}T_{\text{a}}}\right\} - 1\right), \quad I_0 = N_{\text{p}}\frac{I_{\text{sc,n}} + K_{\text{i}}(T_{\text{a}} - T_{\text{n}})}{\{\frac{V_{\text{oc,n}} + K_{\text{v}}(T_{\text{a}} - T_{\text{n}})}{aV_{\text{t,n}}}\} - 1}, \end{aligned} \tag{3}$$

where N_{p} and N_{s} are the number of cells connected in parallel and in series, respectively, K_{i} and K_{v} are the short circuit current/temperature coefficient and open circuit voltage/temperature coefficient, respectively, $I_{\text{pv,n}}$ is the light-generated current at the nominal condition, T and G are temperature of the P-N junction and surface irradiance of the cell, respectively, T_{n} and G_{n} are the temperature and irradiance at the nominal condition, respectively, a is the diode ideality constant, $I_{\text{sc,n}}$ and $V_{\text{oc,n}}$ are the nominal short circuit and nominal open voltage, respectively, $V_{\text{t,n}}$ is the thermal voltage of N_{s} series-connected cells at the nominal condition, q is the electron charge ($1.60217646 \times 10^{-19}$ C), and k is the Boltzmann constant ($1.3806503 \times 10^{-23}$ J/K).

The action at time t is $A_t = (a_t^1, a_t^2, \ldots, a_t^{N_{\text{ev}}}, b_t^1, b_t^2, \ldots, b_t^{N_{\text{b}}}) \in \mathcal{A}$, where $\mathcal{A}$ is the action space, a_t^i is a Boolean variable which represents the charge action of the ith EV, b_t^j is a discrete variable which represents the action of the battery in the jth building. According to Assumption 1, the energy used for charging the ith EV at time t is $a_t^i \cdot P_{\text{ev}}$, where P_{ev} is the charging power of the EV.

The policy at time t, π_t, is a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$, i.e., $\pi_t : \mathcal{S} \to \mathcal{A}$. And the policy vector $\boldsymbol{\pi}$ is made up by the policy at each time, i.e., $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_t, \ldots, \pi_{T-1})$.

Suppose that the kth parking event of the ith EV happens at time $t_{k,\text{s}}^i$, i.e., $D^i_{t^i_{k,\text{s}}-1} = 0$ and $D^i_{t^i_{k,\text{s}}} > 0$. Then we have the following dynamics:

$$L_{t+1}^i = \begin{cases} L_t^i - 1, & \text{if } D_t^i > 0, \\ \mathcal{L}_k^i, & \text{if } t = t_{k,\text{s}}^i - 1, \\ 0, & \text{if } D_{t+1}^i = 0, \end{cases} \qquad E_{t+1}^i = \begin{cases} E_t^i - a_t^i P_{\text{ev}}, & \text{if } D_t^i > 0, \\ \mathcal{E}_k^i, & \text{if } t = t_{k,\text{s}}^i - 1, \\ 0, & \text{if } D_{t+1}^i = 0, \end{cases} \tag{4}$$

where $\mathcal{L}_k^i$ and $\mathcal{E}_k^i$ are the initial remaining parking time and the charging demand in the kth parking event of the ith EV, respectively. In Subsection 4.3, we show how to determine the values of these variables based on the real data [37].

Suppose that the kth parking event of the ith EV finishes at time $t_{k,\text{f}}^i$, i.e., $t_{k,\text{f}}^i = t_{k,\text{s}}^i + \mathcal{L}_k^i - 1$. And let $E_{k,\text{f}}^i$ denote the unfinished charging demand of the ith EV at $t_{k,\text{f}}^i$, i.e., $E_{k,\text{f}}^i = (\sum_{t=t_{k,\text{s}}^i}^{t_{k,\text{f}}^i} a_t^i - \mathcal{E}_k^i)$. Then one-stage cost at time t is defined as

$$C(S_t, A_t) = \beta_t \sum_{j=1}^{N_{\text{b}}} \max\left(P_{\text{ev}} \sum_{i=1}^{N_{\text{ev}}} a_t^i I(D_t^i = j) - P_{\text{w},t}^j - P_{\text{pv},t}^j - b_t^j P_{\text{b}}^j, 0\right) + \gamma \sum_{i=1}^{N_{\text{ev}}} I(t = t_{k,\text{f}}^i) E_{k,\text{f}}^i, \tag{5}$$

where β_t is the time-of-use (TOU) price of the power grid, $P_{\rm b}^j$ is the constant charging/discharging power of the battery equipped with the jth building, and $I(A)$ is the indicator function. The TOU price usually changes over time. Let γ denote the penalty coefficient for not fully charged by the departure time.

Consider the following objective function:

$$\begin{aligned} J(\boldsymbol{\pi}, S_0) =& \mathrm{E}\left\{\sum_{t=0}^{T-1} C(S_t, \pi_t(S_t))\middle| S_0\right\} \\ =& \mathrm{E}\left\{\sum_{t=0}^{T-1}\left[\beta_t \sum_{j=1}^{N_{\rm b}} \max\left(P_{\rm ev}\sum_{i=1}^{N_{\rm ev}} a_t^i I(D_t^i = j) - P_{{\rm w},t}^j - P_{{\rm pv},t}^j - b_t^j P_{\rm b}^j, 0\right)\right.\right. \\ &\left.\left. +\gamma \sum_{i=1}^{N_{\rm ev}} I(t = t_{k,{\rm f}}^i) E_{k,{\rm f}}^i\right]\middle| S_0\right\}, \end{aligned} \tag{6}$$

where S_0 is the initial state.

So the EV charging problem P1 is

$$\mathrm{P1}: \min_{\boldsymbol{\pi}\in\mathcal{P}} J(\boldsymbol{\pi}, S_0) \quad \text{s.t. (1)}, \tag{7}$$

where $\mathcal{P}$ is the policy space.

4 Main results

In this section, we focus on using the SBPI to improve from a heuristic policy in the EV charging problem. First, we use the urgency index to rank the EVs and reduce the action space. Then, we apply three search methods to find the optimal action in the action space. At last, we numerically compare these search methods with the US, and compare the SBPI methods with the CPLEX-B method both in the original action space and the reduced action space.

4.1 Reduced action space

We define the urgency index of the ith EV at time t, u_t^i as

$$u_t^i = \frac{E_t^i}{L_t^i}. \tag{8}$$

We know that the LLLP principle [7] can maintain the optimality. Now we show that the urgency index is consistent with the LLLP principle in the following theorem.

Theorem 1. The urgency index does not violate both of the two inequalities in the LLLP principle at the same time.

Note that if EV x has priority over EV y in LLLP principle, then $\omega_t^x \leqslant \omega_t^y$ and $E_t^x \geqslant E_t^y$, and one of these two inequalities strictly holds [7], where the laxity of the ith EV at time t, ω_t^i, is defined as

$$\omega_t^i = L_t^i - E_t^i. \tag{9}$$

Proof. For any two EVs, x and y, suppose that $u_t^x > u_t^y$. Then $\frac{E_t^x}{L_t^x} > \frac{E_t^y}{L_t^y}$. For laxity, let $L_t^x - E_t^x \geqslant L_t^y - E_t^y$; i.e., suppose that one of the inequalities in LLLP is violated. Then

$$E_t^x > \frac{L_t^x E_t^y}{L_t^y} \quad \text{and} \quad E_t^x \leqslant L_t^x + E_t^y - L_t^y. \tag{10}$$

So,

$$\begin{aligned} \frac{L_t^x E_t^y}{L_t^y} < L_t^x + E_t^y - L_t^y &\Rightarrow L_t^x L_t^y + E_t^y L_t^y - (L_t^y)^2 > E_t^y L_t^x \\ &\Rightarrow L_t^x(L_t^y - E_t^y) > L_t^y(L_t^y - E_t^y) \Rightarrow L_t^x > L_t^y. \end{aligned} \tag{11}$$

From $\frac{E_t^x}{L_t^x} > \frac{E_t^y}{L_t^y}$, we also have that $L_t^x < \frac{E_t^x L_t^y}{E_t^y}$. So,

$$\frac{E_t^x L_t^y}{E_t^y} > L_t^y \Rightarrow E_t^x > E_t^y; \tag{12}$$

i.e., it does not violate another inequality in LLLP.

Similarly, if $E_t^x \leqslant E_t^y$, we have that $L_t^x - E_t^x > L_t^y - E_t^y$. So the urgency index does not violate both of the two inequalities in the LLLP principle at the same time.

Using the urgency index, we can decide the number of EVs to be charged directly. The action at each stage is converted to $\hat{A}_t = (n_{\mathrm{ev},t}^1, \ldots, n_{\mathrm{ev},t}^{N_\mathrm{b}}, b_t^1, \ldots, b_t^{N_\mathrm{b}})$, where $n_{\mathrm{ev},t}^j$, $j = 1, 2, \ldots, N_\mathrm{b}$ is the number of EVs to be charged in the parking lot of the jth building. So the size of the action space is curtailed from $2^{N_\mathrm{ev}} \times 3^{N_\mathrm{b}}$ to $\prod_{j=1}^{N_\mathrm{b}}[3 \times (N_{\mathrm{ev},t}^j + 1)]$, where $N_{\mathrm{ev},t}^j$ is the number of EVs parking at the parking lot of the jth building at time t, and $\sum_{j=1}^{N_\mathrm{b}} N_{\mathrm{ev},t}^j = N_\mathrm{ev}$. It is obvious that the size of the new action space is smaller than that of the original action space. We refer to the new action space as the reduced action space.

4.2 Methods for searching

The action which maximizes the estimated Q-factors is selected in SBPI [38]. For each stage-action pair (S_t, A_t), the Q-factor is defined as

$$Q_t(S_t, A_t) = C_t(S_t, A_t) + \mathrm{E}[V(S_{t+1})|S_t, A_t], \tag{13}$$

where $V(S_{t+1})$ is the value function of state S_{t+1}, i.e.,

$$V(S_{t+1}) = \mathrm{E}\left[\sum_{\tau=t+1}^{T-1} C_\tau(S_\tau, \pi_t^*(S_\tau))\middle| S_{t+1}\right], \tag{14}$$

where π_t^* denotes the optimal policy at time t.

However, the optimal policy vector $\boldsymbol{\pi}^*$ is usually unavailable. So a heuristic policy $\boldsymbol{\pi}^{\mathrm{base}}$ is used to estimate the Q-factor:

$$\hat{Q}_t(S_t, A_t) = C_t(S_t, A_t) + \mathrm{E}[\hat{V}(S_{t+1})|S_t, A_t], \tag{15}$$

where

$$\hat{V}(S_{t+1}) = \mathrm{E}\left[\sum_{\tau=t+1}^{T-1} C_\tau(S_\tau, \pi_\tau^{\mathrm{base}}(S_\tau))\middle| S_{t+1}\right]. \tag{16}$$

Monte Carlo sampling is used to generate the sample paths for estimating the Q-factors in SBPI since the expectation in (16) is difficult to calculate:

$$\hat{Q}_t(S_t, A_t) \approx \tilde{Q}_t(S_t, A_t) = C_t(S_t, A_t) + \frac{1}{M}\sum_{m=1}^{M}\sum_{\tau=t+1}^{T-1}(C_\tau(S_\tau, \pi_\tau^{\mathrm{base}}(S_\tau))|\xi_m^\tau), \tag{17}$$

where M is the number of generated sample paths and ξ_m^τ is the randomness in the mth sample path at time τ. It is therefore of practical interest to discuss how to search for the optimal action under finite computing budget.

4.2.1 *MRAS*

We introduce the parameterized distribution for action sampling both in the original action space and the reduced action space, and show how to update the parameters optimally.

(1) In the original action space

In the original action space, let θ_t^i denote the probability that the ith EV charges at time t, so the probability distribution of the action for the EVs can be written as

$$f_\mathrm{ev}(a_t^i) = {\theta_{\mathrm{ev},t}^i}^{(a_t^i)}(1 - \theta_{\mathrm{ev},t}^i)^{(1-a_t^i)} = \exp\left(a_t^i \ln(\theta_{\mathrm{ev},t}^i) + (1 - a_t^i)\ln(1 - \theta_{\mathrm{ev},t}^i)\right), \tag{18}$$

where $a_t^i = 1$ means to charge the ith EV at time t, and $a_t^i = 0$ means not to charge the EV.

Let $\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}$ denote the probability that the battery in building j discharges at time t, and $\theta_{\mathrm{b},t}^{j_{\mathrm{v}}}$ denote the probability that the battery has no action. So the probability that the battery charges at time t is $1-\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}-\theta_{\mathrm{b},t}^{j_{\mathrm{v}}}$. Then the probability distribution of the action for the battery is described as

$$\begin{aligned} f_{\mathrm{b}}(\theta_{\mathrm{b},t}^{j_{\mathrm{u}}},\theta_{\mathrm{b},t}^{j_{\mathrm{v}}}) &= (\theta_{\mathrm{b},t}^{j_{\mathrm{u}}})^{I(b_t^j=1)}(\theta_{\mathrm{b},t}^{j_{\mathrm{v}}})^{I(b_t^j=0)}\times(1-\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}-\theta_{\mathrm{b},t}^{j_{\mathrm{v}}})^{(1-I(b_t^j=1)-I(b_t^j=0))} \\ &= \exp\left(I(b_t^j=1)\ln(\theta_{\mathrm{b},t}^{j_{\mathrm{u}}})+I(b_t^j=0)\ln(\theta_{\mathrm{b},t}^{j_{\mathrm{v}}})+(1-I(b_t^j=1)-I(b_t^j=0))\ln(1-\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}-\theta_{\mathrm{b},t}^{j_{\mathrm{v}}})\right), \end{aligned} \tag{19}$$

where $I(x)$ is the indicator function and $I(x)=1$ (or 0) if x is true (or false), $b_j^t=1$ means to discharge the battery in the jth building, $b_j^t=0$ means nothing happens, and $b_j^t=-1$ means to charge the battery.

Suppose that the action of each EV or battery is independent, and then the probability distribution of action A_t is

$$\begin{aligned} f(A_t) =& \prod_{i=1}^{N_{\mathrm{ev}}}\exp(a_t^i\ln(\theta_{\mathrm{ev},t}^i)+(1-a_t^i)\ln(1-\theta_{\mathrm{ev},t}^i)) \\ &\times\prod_{j=1}^{N_{\mathrm{b}}}\exp\Big[I(b_t^j=1)\ln(\theta_{\mathrm{b},t}^{j_{\mathrm{u}}})+I(b_t^j=0)\ln(\theta_{\mathrm{b},t}^{j_{\mathrm{v}}}) \\ &+(1-I(b_t^j=1)-I(b_t^j=0))\ln(1-\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}-\theta_{\mathrm{b},t}^{j_{\mathrm{v}}})\Big]. \end{aligned} \tag{20}$$

Applying this distribution to sample the actions in MRAS [8], we can update the parameters and find the optimal action as in Algorithm 1, where M_{total} is the total computing budget (counted by the number of simulation replications) available for allocation at each stage, and

$$\tilde{I}(\tilde{Q},\hat{\chi}_p)=\begin{cases}1, & \text{if } \tilde{Q}\leqslant\hat{\chi}_p,\\ \dfrac{\hat{\chi}_p-\tilde{Q}+\epsilon}{\epsilon}, & \text{if } \hat{\chi}_p<\tilde{Q}\leqslant\hat{\chi}_p+\epsilon,\\ 0, & \text{if } \hat{\chi}_p+\epsilon<\tilde{Q}.\end{cases} \tag{21}$$

The parameters are $\theta=(\theta_{\mathrm{ev},t}^1,\ldots,\theta_{\mathrm{ev},t}^{N_{\mathrm{ev}}},\theta_{\mathrm{b},t}^{1_{\mathrm{u}}},\theta_{\mathrm{b},t}^{1_{\mathrm{v}}},\ldots,\theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{u}}},\theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{v}}})$. We update θ in Step 6 by using the following theorem.

Theorem 2 (An optimal way to update the parameters in MRAS for the original action space)**.** For parameters $\theta=(\theta_{\mathrm{ev},t}^1,\ldots,\theta_{\mathrm{ev},t}^{N_{\mathrm{ev}}},\theta_{\mathrm{b},t}^{1_{\mathrm{u}}},\theta_{\mathrm{b},t}^{1_{\mathrm{v}}},\ldots,\theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{u}}},\theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{v}}})$ and action distribution $f(A_t)$, the optimal way to update the parameters in Step 6 of Algorithm 1 should follow

$$\begin{aligned} &\hat{\theta}_{\mathrm{ev},t,p+1}^i=\frac{\sum_{k=1}^{N_p}(c_{t,p}^k a_{t,p}^{i,k})}{\sum_{k=1}^{N_p}c_{t,p}^k}, \\ &\hat{\theta}_{\mathrm{b},t,p+1}^{j_{\mathrm{u}}}=\frac{\sum_{k=1}^{N_p}(c_{t,p}^k I(b_{t,p}^{j,k}=1))}{\sum_{k=1}^{N_p}c_{t,p}^k},\quad \hat{\theta}_{\mathrm{b},t,p+1}^{j_{\mathrm{v}}}=\frac{\sum_{k=1}^{N_p}(c_{t,p}^k I(b_{t,p}^{j,k}=0))}{\sum_{k=1}^{N_p}c_{t,p}^k}, \end{aligned} \tag{22}$$

where $c_{t,p}^k=\frac{[\mathcal{H}(\tilde{Q}(S_t,A_{t,p}^k))]^p}{\tilde{f}(A_{t,p}^k,\tilde{\theta_p})}\tilde{I}(\tilde{Q}(S_t,A_{t,p}^k),\hat{\chi}_p)$.

Proof. Take the partial derivative of $D_{t,p}$ for each component in θ. For parameters on the EV, $\theta_{\mathrm{ev},t}^i, i=1,\ldots,N_{\mathrm{ev}}$, we have

$$\frac{\partial D_{t,p}}{\partial\theta_{\mathrm{ev},t}^i}=\frac{1}{N_p}\sum_{A_{t,p}^k\in\Lambda_{t,p}}\left\{\frac{[\mathcal{H}(\tilde{Q}(S_t,A_{t,p}^k))]^p}{\tilde{f}(A_{t,p}^k,\tilde{\theta_p})}\times\ \tilde{I}(\tilde{Q}(S_t,A_{t,p}^k),\hat{\chi}_p)\frac{\partial\ln f(A_{t,p}^k,\theta)}{\partial\theta_{\mathrm{ev},t}^i}\right\}. \tag{23}$$

Let $c_{t,p}^k=\dfrac{[\mathcal{H}(\tilde{Q}(S_t,A_{t,p}^k))]^p}{\tilde{f}(A_{t,p}^k,\tilde{\theta_p})}\tilde{I}(\tilde{Q}(S_t,A_{t,p}^k),\hat{\chi}_p)$, and then

$$\frac{\partial D_{t,p}}{\partial\theta_{\mathrm{ev},t}^i}=\frac{1}{N_p}\sum_{k=1}^{N_p}\left[c_{t,p}^k\frac{\partial\ln f(A_{t,p}^k,\theta)}{\partial\theta_{\mathrm{ev},t}^i}\right]=\frac{1}{N_p}\sum_{k=1}^{N_p}\left[c_{t,p}^k\left(\frac{a_{t,p}^{i,k}}{\theta_{\mathrm{ev},t}^i}-\frac{1-a_{t,p}^{i,k}}{1-\theta_{\mathrm{ev},t}^i}\right)\right]. \tag{24}$$

Algorithm 1 Search for the optimal action using MRAS

Step 1: At time t, initialize $\rho_0 \in (0,1]$, N_0, M_0, $\epsilon \geqslant 0$, $\alpha > 1$, $\lambda \in (0,1]$, strictly decreasing function $\mathcal{H}(x) = \mathrm{e}^{-\kappa x}$, exponential distribution family $f(A_t, \tilde{\theta})$.
Step 2: Round $p = 0$, $\tilde{\theta_0} = \theta_0$, $M_{\mathrm{used},0} = 0$, loop flag $q = 0$.
Step 3: Sample N_p actions, $\Lambda_{t,p} = \{A_{t,p}^1, \ldots, A_{t,p}^{N_p}\}$ from the distribution $\tilde{f}(\cdot, \tilde{\theta_p}) = (1-\lambda) f(\cdot, \tilde{\theta_p}) + \lambda f(\cdot, \theta_0)$, where $A_{t,p}^k = (a_{t,p}^{1,k}, \ldots, a_{t,p}^{i,k}, \ldots, a_{t,p}^{N_{\mathrm{ev}},k}, b_{t,p}^{1,k}, \ldots, b_{t,p}^{j,k}, \ldots, b_{t,p}^{N_{\mathrm{b}},k})$. Use M_p sample paths for estimating the Q-factors $\tilde{Q}(S_t, A_{t,p}^1)$, $\tilde{Q}(S_t, A_{t,p}^2)$, $\tilde{Q}(S_t, A_{t,p}^{N_p})$.
if $q = 1$ **then**
  Select $A_{t,p}^{(1)}$ to implement, break.
end if
Step 4: Calculate the ρ_p quantile of the sample performance, $\tilde{\chi}_p(\rho_p, N_p) = \tilde{Q}_{(\lceil \rho_p \times N_p \rceil)}$, where $\tilde{Q}_{(k)}$ is the kth order statistic of $\tilde{Q}(S_t, A_{t,p}^{(k)})$, $k = 1, \ldots, N_p$. $M_{\mathrm{used},p+1} = M_p N_p + M_{\mathrm{used},p}$.
Step 5:
if $p = 0$ or $\tilde{\chi}_p(\rho_p, N_p) \leqslant \hat{\chi}_{p-1} + \epsilon$ **then**
  $\hat{\chi}_p = \tilde{\chi}_p(\rho_p, N_p)$, $\rho_{p+1} = \rho_p$, $N_{p+1} = N_p$, $A_{t,p}^* = A_{t,p}^{(\rho_p)}$, where $A_{t,p}^{(\rho_p)} \in \{A_{t,p}^k \in \Lambda_{t,p} : \tilde{Q}(S_t, A_{t,p}^k) = \tilde{Q}_{(\lceil \rho_p \times N_p \rceil)}\}$.
else
  Try to find the smallest $\hat{\rho}_p \in (\rho_p, 1]$ s.t. $\tilde{\chi}_p(\hat{\rho}_p, N_p) \leqslant \hat{\chi}_{p-1} + \epsilon$.
  if there exists such $\hat{\rho}_p$ **then**
    $\hat{\chi}_p = \tilde{\chi}_p(\hat{\rho}_p, N_p)$, $\rho_{p+1} = \hat{\rho}_p$, $N_{p+1} = N_p$, $A_{t,p}^* = A_{t,p}^{(\hat{\rho}_p)}$, where $A_{t,p}^{(\hat{\rho}_p)} \in \{A_{t,p}^k \in \Lambda_{t,p} : \tilde{Q}(S_t, A_{t,p}^k) = \tilde{Q}_{(\lceil \hat{\rho}_p \times N_p \rceil)}\}$.
  else
    $\hat{\chi}_p = \tilde{Q}(S_t, A_{t,p-1}^*)$, $\rho_{p+1} = \rho_p$, $N_{p+1} = \lceil \alpha N_p \rceil$, $A_{t,p}^* = A_{t,p-1}^*$.
  end if
end if
Step 6: Update the parameters θ_p by

$$\hat{\theta}_{p+1} \in \arg\max_{\theta \in \Theta} D_{t,p} = \frac{1}{N_p} \sum_{A_t \in \Lambda_{t,p}} \left\{ \frac{[\mathcal{H}(\tilde{Q}(S_t, A_t))]^p}{\tilde{f}(A_t, \tilde{\theta_p})} \bar{I}(\tilde{Q}(S_t, A_t), \hat{\chi}_p) \ln f(A_t, \theta) \right\}.$$

Step 7:
if $(2M_p N_p + M_{\mathrm{used},p}) > M_{\mathrm{total}}$ **then**
  $M_{p+1} = \lceil \dfrac{M_{\mathrm{total}} - M_{\mathrm{used},p}}{N_p} \rceil$, $q_{\mathrm{sign}} = 1$.
end if
$p = p + 1$, $\tilde{\theta}_p = \nu \hat{\theta}_p + (1-\nu) \tilde{\theta}_{p-1}$. Go to Step 3.

Similarly, for parameters on the equipped battery, $\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}, \theta_{\mathrm{b},t}^{j_{\mathrm{v}}}, j = 1, \ldots, N_{\mathrm{b}}$, we have

$$\frac{\partial D_{t,p}}{\partial \theta_{\mathrm{b},t}^{j_{\mathrm{u}}}} = \frac{1}{N_p} \sum_{k=1}^{N_p} \left[c_{t,p}^k \left(\frac{I(b_{t,p}^{j,k} = 1)}{\theta_{\mathrm{b},t}^{j_{\mathrm{u}}}} - \frac{1 - I(b_{t,p}^{j,k} = 1) - I(b_{t,p}^{j,k} = 0)}{1 - \theta_{\mathrm{b},t}^{j_{\mathrm{u}}} - \theta_{\mathrm{b},t}^{j_{\mathrm{v}}}} \right) \right] \tag{25}$$

and

$$\frac{\partial D_{t,p}}{\partial \theta_{\mathrm{b},t}^{j_{\mathrm{v}}}} = \frac{1}{N_p} \sum_{k=1}^{N_p} \left[c_{t,p}^k \left(\frac{I(b_{t,p}^{j,k} = 0)}{\theta_{\mathrm{b},t}^{j_{\mathrm{v}}}} - \frac{1 - I(b_{t,p}^{j,k} = 1) - I(b_{t,p}^{j,k} = 0)}{1 - \theta_{\mathrm{b},t}^{j_{\mathrm{u}}} - \theta_{\mathrm{b},t}^{j_{\mathrm{v}}}} \right) \right]. \tag{26}$$

Let $\dfrac{\partial D_{t,p}}{\partial \theta_{\mathrm{ev},t}^{i}}$, $\dfrac{\partial D_{t,p}}{\partial \theta_{\mathrm{b},t}^{j_{\mathrm{u}}}}$, and $\dfrac{\partial D_{t,p}}{\partial \theta_{\mathrm{b},t}^{j_{\mathrm{v}}}}$ be zero, and then we have (22).

By the end of the iteration, the algorithm outputs the action $A_{t,p}^{(1)}$, which has the smallest Q-factor in the final round.

(2) In the reduced action space

In the reduced action space, for normalization, a proportion $x_{\mathrm{ev},t}^j$ is used to tell the number of EVs to be charged, i.e., $x_{\mathrm{ev},t}^j N_{\mathrm{ev},t}^j = n_{\mathrm{ev},t}^j$.

Since we only need the optimal value of $\mu_{\mathrm{ev},t}^j$, a normal distribution is used to estimate the distribution of $x_{\mathrm{ev},t}^j$ as

$$f_{\mathrm{p}}(x_{\mathrm{ev},t}^j) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{ev},t}^j} \exp\left(\frac{(x_{\mathrm{ev},t}^j - \mu_{\mathrm{ev},t}^j)^2}{2(\sigma_{\mathrm{ev},t}^j)^2} \right), \tag{27}$$

where $0 \leqslant \mu_{\mathrm{ev},t}^j \leqslant 1$ and $0 \leqslant \sigma_{\mathrm{ev},t}^j \leqslant 1$. So we have the parameters $\theta = (\mu_{\mathrm{ev},t}^1, \sigma_{\mathrm{ev},t}^1, \ldots, \mu_{\mathrm{ev},t}^{N_{\mathrm{b}}}, \sigma_{\mathrm{ev},t}^{N_{\mathrm{b}}}, \theta_{\mathrm{b},t}^{1_{\mathrm{u}}}, \theta_{\mathrm{b},t}^{1_{\mathrm{v}}}, \ldots, \theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{u}}}, \theta_{\mathrm{b},t}^{(N_{\mathrm{b}})_{\mathrm{v}}})$.

Assume that all the parameters are independent, so the probability distribution of action $\hat{A}_t$ is

$$f(\hat{A}_t) = \prod_{i=1}^{N_{\mathrm{ev}}} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{ev},t}^j} \exp\left(\frac{(x_{\mathrm{ev},t}^j - \mu_{\mathrm{ev},t}^j)^2}{2(\sigma_{\mathrm{ev},t}^j)^2} \right)$$

$$\times \prod_{j=1}^{N_{\rm b}} \exp\Big[I(b_t^j=1)\ln(\theta_{{\rm b},t}^{j_{\rm u}}) + I(b_t^j=0)\ln(\theta_{{\rm b},t}^{j_{\rm v}}) + (1-I(b_t^j=1)-I(b_t^j=0))\ln(1-\theta_{{\rm b},t}^{j_{\rm u}}-\theta_{{\rm b},t}^{j_{\rm v}})\Big]. \tag{28}$$

Algorithm 1 may also be applied in the reduced action space. Take $A_{t,p}^k = (x_{t,p}^{1,k},\ldots,x_{t,p}^{i,k},\ldots,x_{t,p}^{N_{\rm ev},k}, b_{t,p}^{1,k},\ldots,b_{t,p}^{j,k},\ldots,b_{t,p}^{N_{\rm b},k})$. We update θ in Step 6 using the following theorem.

Theorem 3 (An optimal way to update the parameters in MRAS for the reduced action space)**.** For parameters $\theta = (\mu_{{\rm ev},t}^1, \sigma_{{\rm ev},t}^1, \ldots, \mu_{{\rm ev},t}^{N_{\rm b}}, \sigma_{{\rm ev},t}^{N_{\rm b}}, \theta_{{\rm b},t}^{1_{\rm u}}, \theta_{{\rm b},t}^{1_{\rm v}}, \ldots, \theta_{{\rm b},t}^{(N_{\rm b})_{\rm u}}, \theta_{{\rm b},t}^{(N_{\rm b})_{\rm v}})$ and action distribution $f(\hat{A}_t)$, the optimal way to update the parameters in Step 6 of Algorithm 1 should follow:

$$\begin{aligned} &\hat{\mu}_{{\rm ev},t,p+1}^j = \frac{\sum_{k=1}^{N_p}(c_{t,p}^k x_{{\rm ev},t,p}^{j,k})}{\sum_{k=1}^{N_p} c_{t,p}^k}, \qquad (\hat{\sigma}_{{\rm ev},t,p+1}^j)^2 = \frac{\sum_{k=1}^{N_p} c_{t,p}^k (x_{{\rm ev},t,p}^{j,k}-\mu_{{\rm ev},t}^j)^2}{\sum_{k=1}^{N_p} c_{t,p}^k},\\ &\hat{\theta}_{{\rm b},t,p+1}^{j_{\rm u}} = \frac{\sum_{k=1}^{N_p}(c_{t,p}^k I(b_{t,p}^{j,k}=1))}{\sum_{k=1}^{N_p} c_{t,p}^k}, \quad \hat{\theta}_{{\rm b},t,p+1}^{j_{\rm v}} = \frac{\sum_{k=1}^{N_p}(c_{t,p}^k I(b_{t,p}^{j,k}=0))}{\sum_{k=1}^{N_p} c_{t,p}^k}, \end{aligned} \tag{29}$$

where $c_{t,p}^k = \frac{[\mathcal{H}(\tilde{Q}(S_t,A_{t,p}^k))]^p}{\tilde{f}(A_{t,p}^k,\tilde{\theta_p})}\tilde{I}(\tilde{Q}(S_t,A_{t,p}^k),\hat{\chi}_p)$.

Proof. Calculate $\frac{\partial D_{t,p}}{\partial \mu_{{\rm ev},t}^j}$, $\frac{\partial D_{t,p}}{\partial \sigma_{{\rm ev},t}^j}$, $\frac{\partial D_{t,p}}{\partial \theta_{{\rm b},t}^{j_{\rm u}}}$, and $\frac{\partial D_{t,p}}{\partial \theta_{{\rm b},t}^{j_{\rm v}}}$. We can find that $\theta_{{\rm b},t,p}^{j_{\rm u}}$ and $\theta_{{\rm b},t,p}^{j_{\rm v}}$ can be updated identically as in Subsection 4.2.1(1). For $\mu_{{\rm ev},t}^j$ and $\sigma_{{\rm ev},t}^j$, we have

$$\frac{\partial D_{t,p}}{\partial \mu_{{\rm ev},t}^j} = \frac{1}{N_p}\sum_{k=1}^{N_p}\left[c_{t,p}^k \frac{(x_{{\rm ev},t,p}^{j,k}-\mu_{{\rm ev},t}^j)}{(\sigma_{{\rm ev},t}^j)^2}\right] \tag{30}$$

and

$$\frac{\partial D_{t,p}}{\partial \sigma_{{\rm ev},t}^j} = \frac{1}{N_p}\sum_{k=1}^{N_p}\left[c_{t,p}^k\left(\frac{-1}{\sigma_{{\rm ev},t}^j} + \frac{(x_{{\rm ev},t,p}^{j,k}-\mu_{{\rm ev},t}^j)^2}{(\sigma_{{\rm ev},t}^j)^3}\right)\right]. \tag{31}$$

Let $\frac{\partial D_{t,p}}{\partial \mu_{{\rm ev},t}^j}$ and $\frac{\partial D_{t,p}}{\partial \sigma_{{\rm ev},t}^j}$ be zero, combine the results in Subsection 4.2.1, and then we have (29).

By the end of the iteration, the algorithm outputs the action $A_{t,p}^{(1)}$, which has the smallest Q-factor in the final round.

4.2.2 *MARS*

MARS is different from MRAS in two aspects. The first is that Step 6 in Algorithm 1 is replaced by

$$\tilde{\theta}_{p+1} \in \arg\max_{\theta\in\Theta} D_{t,p} = \mathcal{D}(\hat{g}_{p+1}(A_t), f(A_t,\theta)), \tag{32}$$

where $\mathcal{D}(\cdot,\cdot)$ means to calculate the KL divergence, and $\hat{g}_{p+1}(A_t) = \upsilon_p \bar{g}_{p+1}(A_t) + (1-\upsilon_p) f(A_t,\tilde{\theta}_p)$.

In (32), $\bar{g}_{p+1}(A_t)$ is an empirical distribution that

$$\bar{g}_{p+1}(A_t) = \frac{\exp(-\frac{\kappa Q(S_t,A_t)}{T_{p+1}})/\tilde{f}(A_t,\tilde{\theta_p})}{\sum_{A_t\in\Lambda_{t,p}}\left[\exp(-\frac{\kappa Q(S_t,A_t)}{T_{p+1}})/\tilde{f}(A_t,\tilde{\theta_p})\right]}, \tag{33}$$

where T_{p+1} is a parameter in iteration $p+1$, and decreases towards a constant with the iterations.

The second is that the MARS uses all samples to update the parameters while the MRAS only uses 'elite' samples to update the parameters. We show the algorithm of MARS in Algorithm 2.

Algorithm 2 Search for the optimal action using MARS

Step 1: At time t, initialize N_0, M_0, $\lambda \in (0,1]$, $\alpha > 1$, parameterized distribution $f(A_t, \tilde{\theta})$.
Step 2: Round $p = 0$, $\tilde{\theta}_0 = \theta_0$, $M_{\text{used},p} = 0$, loop flag $q = 0$.
Step 3: Sample N_p actions, $\Lambda_{t,p} = \{A_{t,p}^1, \ldots, A_{t,p}^{N_p}\}$ from the distribution $\tilde{f}(\cdot, \tilde{\theta_p}) = (1-\lambda) f(\cdot, \tilde{\theta_p}) + \lambda f(\cdot, \theta_0)$. Use M_p sample paths for estimating the Q-factors $\tilde{Q}(S_t, A_{t,p}^1)$, $\tilde{Q}(S_t, A_{t,p}^2)$, $\tilde{Q}(S_t, A_{t,p}^{N_p})$. Find the optimal action $A_{t,p}^* = \arg\min_{A_{t,p}^k \in \Lambda_{t,p}} Q(S_t, A_{t,p}^k)$, $M_{\text{used},p+1} = M_p N_p + M_{\text{used},p}$.
if $q = 1$ **then**
  Select $A_{t,p}^*$ to implement, break.
end if
Step 4: Update the parameters θ_p by $\tilde{\theta}_{p+1} \in \arg\min_{\theta \in \Theta} D_{t,p} = \mathcal{D}(\hat{g}_{p+1}(A_t), f(A_t, \theta))$.
Step 5:
if $(2M_p N_p + M_{\text{used},p}) > M_{\text{total}}$ **then**
  $M_{p+1} = \lceil \frac{M_{\text{total}} - M_{\text{used},p}}{N_p} \rceil$, $q = 1$.
end if
$p = p + 1$, $\tilde{\theta}_p = \nu \hat{\theta}_p + (1-\nu)\tilde{\theta}_{p-1}$. Go to Step 3.

In the original action space, the parameters are $\theta = (\theta_{\text{ev},t}^1, \ldots, \theta_{\text{ev},t}^{N_{\text{ev}}}, \theta_{\text{b},t}^{1_{\text{u}}}, \theta_{\text{b},t}^{1_{\text{v}}}, \ldots, \theta_{\text{b},t}^{(N_{\text{b}})_{\text{u}}}, \theta_{\text{b},t}^{(N_{\text{b}})_{\text{v}}})$. Similar to the proof in MRAS, we update the parameters in Step 4 by

$$\hat{\theta}_{\text{ev},t,p+1}^i = \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) a_{t,p}^{i,k})}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)},$$

$$\begin{aligned} \hat{\theta}_{\text{b},t,p+1}^{j_{\text{u}}} &= \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) I(b_{t,p}^{j,k} = 1))}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}, \\ \hat{\theta}_{\text{b},t,p+1}^{j_{\text{v}}} &= \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) I(b_{t,p}^{j,k} = 0))}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}. \end{aligned} \tag{34}$$

In the reduced action space, the parameters are $\theta = (\mu_{\text{ev},t}^1, \sigma_{\text{ev},t}^1, \ldots, \mu_{\text{ev},t}^{N_{\text{b}}}, \sigma_{\text{ev},t}^{N_{\text{b}}}, \theta_{\text{b},t}^{1_{\text{u}}}, \theta_{\text{b},t}^{1_{\text{v}}}, \ldots, \theta_{\text{b},t}^{(N_{\text{b}})_{\text{u}}}, \theta_{\text{b},t}^{(N_{\text{b}})_{\text{v}}})$. Similar to the proof in MRAS, we update the parameters in Step 4 by

$$\begin{aligned} &\hat{\mu}_{\text{ev},t,p+1}^j = \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) x_{\text{ev},t,p}^{j,k})}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}, \qquad (\hat{\sigma}_{\text{ev},t,p+1}^j)^2 = \frac{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)(x_{\text{ev},t,p}^{j,k} - \mu_{\text{ev},t}^j)^2}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}, \\ &\hat{\theta}_{\text{b},t,p+1}^{j_{\text{u}}} = \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) I(b_{t,p}^{j,k} = 1))}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}, \quad \hat{\theta}_{\text{b},t,p+1}^{j_{\text{v}}} = \frac{\sum_{k=1}^{N_p} (\hat{g}_{p+1}(A_{t,p}^k) I(b_{t,p}^{j,k} = 0))}{\sum_{k=1}^{N_p} \hat{g}_{p+1}(A_{t,p}^k)}. \end{aligned} \tag{35}$$

4.2.3 *COMPASS*

The process of applying COMPASS to search for the optimal action is summarized in Algorithm 3.

Algorithm 3 Search for the optimal action using COMPASS

Step 1: At time t, the most-promising-area $\mathcal{R}_0 = \mathcal{A}$, randomly select $A_{t,0}^1 \in \mathcal{A}$, $\Lambda_{t,0} = A_{t,0}^1$, $A_{t,0}^* = A_{t,0}^1$, initialize $\tilde{M}$, $\tilde{N}$.
Step 2: Round $p = 1$, $M_{\text{used},p} = 0$, loop flag $q = 0$.
Step 3: Sample $\tilde{N}$ actions, $\{A_{t,p}^1, \ldots, A_{t,p}^{\tilde{N}}\}$ from $\mathcal{R}_p$ uniformly, $\Lambda_{t,p} = \Lambda_{t,p-1} \cup \{A_{t,p}^1, \ldots, A_{t,p}^{\tilde{N}}\}$. Using $\tilde{M}$ sample paths for estimating the Q-factors $\tilde{Q}(S_t, A_{t,p}^1)$, $\tilde{Q}(S_t, A_{t,p}^2)$, $\tilde{Q}(S_t, A_{t,p}^{p\tilde{N}+1})$, $M_{\text{used},p+1} = \tilde{M}(p\tilde{N}+1) + M_{\text{used},p}$, find the optimal action $A_{t,p}^* = \arg\min_{A_{t,p}^k \in \Lambda_{t,p}} Q(S_t, A_{t,p}^k)$.
if $q = 1$ **then**
  Select $A_{t,p}^*$ to implement, break.
end if
Step 4: Construct $\mathcal{R}_p$ using $A_{t,p}^*$ and $\Lambda_{t,p}$.
Step 5:
if $\{\tilde{M}[(2p+3)\tilde{N}+2] + M_{\text{used},p}\} > M_{\text{total}}$ **then**
  $\tilde{M}_{p+1} = \lceil \frac{M_{\text{total}} - M_{\text{used},p}}{(p+1)\tilde{N}+1} \rceil$, $q = 1$.
end if
$p = p + 1$. Go to Step 3.

In Step 4 of Algorithm 3, since constructing the most-promising-area $\mathcal{R}$ strictly is difficult, we uses Algorithm 4 to construct $\mathcal{R}$, which has been introduced in [11].

Algorithm 4 Uniformly sample from the most-promising-area in COMPASS

1: Step 1: At time t, in round p, initialize the start point $X_0 = A^*_{t,p}$, the warm-up length G, dimension of $A^*_{t,p}$ being D. Now $g = 0$, and the number of sampled actions $k = 0$.
2: Step 2: $g = g + 1$. Randomly select a dimension d. Let $l(X_{g-1}, I)$ be the line passing through the point X_{g-1} and parallel to the dth coordinate axis. This line interacts with the boundary of $\mathcal{R}$ at points y_1 and y_2. Let $\Upsilon(X_{g-1}, I)$ denote the set of integer points between y_1 and y_2.
3: Step 3: Sample X_g uniformly from $\Upsilon(X_{g-1}, I)$.
4: Step 4:
5: **if** $g < G$ **then**
6: Go to Step 2.
7: **else**
8: $\Lambda_{t,p} = \Lambda_{t,p} \cup X_G$, $k = k + 1$.
9: **if** $k = \tilde{N}$ **then**
10: Stop.
11: **else**
12: $g = 0$, $X_0 = X_G$. Go to Step 2.
13: **end if**
14: **end if**

In Step 2 of Algorithm 4, how to sample uniformly from $\Upsilon(X_{g-1}, I)$ is related to the specific problem. We give the following steps to demonstrate the process in the EV charging problem. It is obvious that y_1, y_2 and other points on the line between y_1 and y_2 have the same coordinates except the dth dimension. So only $c(X_g, d)$ needs to be calculated. Suppose the range of each dimension d in action space is $[r^d_{\text{down}}, r^d_{\text{up}}]$. The sampled points in $\Lambda_{t,p}$ besides $A^*_{t,p}$ are denoted by $\hat{\Lambda}^-_{t,p} = (\hat{A}^1_{t,p}, \hat{A}^2_{t,p}, \ldots, \hat{A}^{p\tilde{N}}_{t,p})$.

First, the square of Euclidean distance from each point $A^{\tilde{k}}_{t,p} \in \hat{\Lambda}^-_{t,p}$ to X_{g-1} is calculated and denoted by $z^{\tilde{k}}$. Let $z^{\tilde{k}}_d = z^{\tilde{k}} - (c(\hat{A}^{\tilde{k}}_{t,p}, d) - c(X_{g-1}, d))^2$, where $c(\hat{A}_t, d)$ denotes the coordinate of $\hat{A}_t$ in the dth dimension. Let z^* denote the square of Euclidean distance from X_{g-1} to $A^*_{t,p}$, and $z^*_d = z^* - (c(X_{g-1}, d) - c(A^*_{t,p}, d))^2$.

Second, for each sample $\hat{A}^{\tilde{k}}_{t,p} \in \hat{\Lambda}^-_{t,p}$, try to select a coordinate $\hat{c}(X^{\tilde{k}}_g, d)$ in dimension d, s.t. $\hat{z}^{\tilde{k}} = z^*$. So let $[\hat{c}(X^{\tilde{k}}_g, d) - c(A^*_{t,p}, d)]^2 - [\hat{c}(X^{\tilde{k}}_g, d) - c(\hat{A}^{\tilde{k}}_{t,p}, d)]^2 + (z^*_d - z^{\tilde{k}}_d) = 0$, and then we have

$$\hat{c}(X^{\tilde{k}}_g, d) = \frac{1}{2}\left[c(A^*_{t,p}, d) + c(\hat{A}^{\tilde{k}}_{t,p}, d) - \frac{z^{\tilde{k}}_d - z^*_d}{c(A^*_{t,p}, d) - c(\hat{A}^{\tilde{k}}_{t,p}, d)}\right]. \tag{36}$$

Third, let $[r^{(\tilde{k},d)}_{\text{down}}, r^{(\tilde{k},d)}_{\text{up}}]$ denote the area $\mathcal{R}$ bounded by the sample $\hat{A}^{\tilde{k}}_{t,p}$ in dimension d, and $[r^{(\tilde{k},d)}_{\text{down}}, r^{(\tilde{k},d)}_{\text{up}}]$ is updated by

$$\begin{cases} r^{(\tilde{k},d)}_{\text{up}} = \min(\lfloor \hat{c}(X^{\tilde{k}}_g, d) \rfloor, r^d_{\text{up}}),\ r^{(\tilde{k},d)}_{\text{down}} = r^d_{\text{down}}, & \text{if } c(\hat{A}^{\tilde{k}}_{t,p}, d) - c(\hat{A}^{\tilde{k}}_{t,p}, d) < 0, \\ r^{(\tilde{k},d)}_{\text{up}} = r^d_{\text{up}},\ r^{(\tilde{k},d)}_{\text{down}} = \max(\lceil \hat{c}(X^{\tilde{k}}_g, d) \rceil, r^d_{\text{down}}), & \text{if } c(\hat{A}^{\tilde{k}}_{t,p}, d) - c(\hat{A}^{\tilde{k}}_{t,p}, d) < 0, \\ r^{(\tilde{k},d)}_{\text{up}} = r^d_{\text{up}},\ r^{(\tilde{k},d)}_{\text{down}} = r^d_{\text{down}}, & \text{otherwise.} \end{cases} \tag{37}$$

After that, the upper bound of integral points in $\mathcal{R}$ for dimension d is $r^{(\mathcal{R},d)}_{\text{up}} = \min_{\tilde{k}}(r^{(\tilde{k},d)}_{\text{up}})$, and lower bound of integral points in $\mathcal{R}$ for dimension d is $r^{(\mathcal{R},d)}_{\text{down}} = \min_{\tilde{k}}(r^{(\tilde{k},d)}_{\text{down}})$. Then we can uniformly sample $c(X_g, d)$ from the interval and obtain X_g.

When the iteration stops, the optimal action $A^*_{p,t}$ in the final round is selected to implement.

4.3 Numerical results

In this subsection, we numerically compare the aforementioned three search methods with the US both in the original action space and the reduced action space.

In the numerical experiments, a problem of charging 100 EVs over 24 h is considered. Suppose there are three buildings and the parking lots of all the buildings are available for parking and charging. The mean value of wind power distribution is from a dataset recorded by a wind farm located in Inner Mongolia of

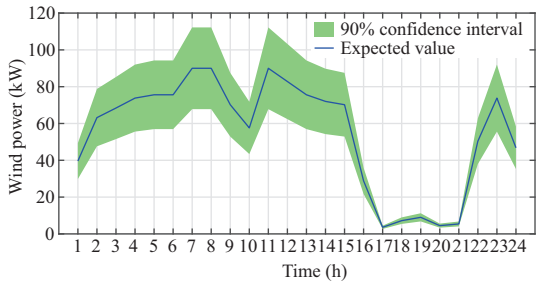

Figure 3 (Color online) The 90% confidence interval of the wind power distribution.

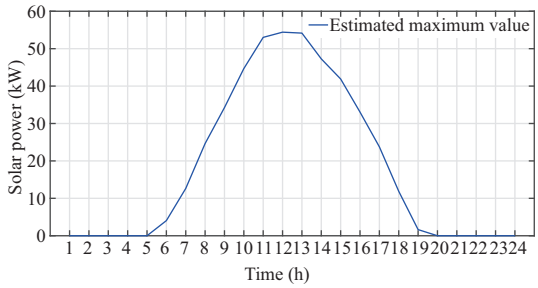

Figure 4 (Color online) An estimated maximum solar power generation curve.

Table 1 Parameters for parking events

t (h)	P_{park}	$\mu_{p,t}$	$\sigma_{p,t}$	t (h)	P_{park}	$\mu_{p,t}$	$\sigma_{p,t}$
0	0.0062	8.835	4.575	12	0.0584	4.812	5.746
1	0.0042	7.313	3.288	13	0.0495	5.604	6.776
2	0.0038	5.522	3.439	14	0.0543	6.029	6.717
3	0.0027	5.075	4.035	15	0.0698	7.395	7.207
4	0.0043	4.224	3.881	16	0.0909	8.190	7.236
5	0.0113	6.426	4.190	17	0.0788	8.423	7.008
6	0.0377	6.556	3.620	18	0.0641	9.035	6.715
7	0.0736	5.854	3.515	19	0.0532	10.125	6.144
8	0.0641	4.997	3.808	20	0.0457	10.496	5.336
9	0.0553	3.941	4.094	21	0.0338	10.152	5.031
10	0.0507	4.318	5.193	22	0.0189	9.981	4.791
11	0.0615	4.576	5.471	23	0.0115	9.676	4.916

China. The weather information recorded by the weather station located in Tsinghua University is from the web site[1)].

Suppose that all the buildings share the wind power equally, and we can obtain the true distribution of the wind power in the next 24 h. The normal distribution is used to approximate the wind power distribution. The ratio of the standard deviation (σ_{wind}) to the mean (μ_{wind}) in the distribution is 15% at each time. The total capacity of the wind power generators is 300 kW. We can sample from the distributions at each time continuously to generate different wind power curves. Figure 3 shows the 90% confidence interval of the wind power distribution at each time in the experiment. The mean value of the wind power distribution is the blue line in Figure 3.

Suppose that all the buildings considered in the system share the same solar radiation and the temperature. An estimated maximal solar power generation of each building is calculated and shown in Figure 4.

To capture the randomness in the generation of solar power, steps in (38) [6] are used to approximate the real supply of the solar power and generate different solar power curves.

$$\tilde{P}^j_{\text{pv},t} = \begin{cases} \eta^j_t P^j_{\text{pv},t-1}, & \text{w.p. } 1/3, \\ \eta^j_t P^j_{\text{pv},t}, & \text{w.p. } 1/3, \\ \eta^j_t P^j_{\text{pv},t+1}, & \text{w.p. } 1/3, \end{cases} \tag{38}$$

where $\eta^j_t \sim U[0.5, 1]$ is random discounted factor, and w.p. stands for with probability.

The vehicle data mentioned in [37] are applied and the parking events associated with home, working place, and commercial buildings are considered. We obtain the probability of parking, and the parameters of the distribution of parking duration at each time from the previous study [39], which are shown in Table 1. A χ^2 distribution is applied to depict the EV charging demand [40]. The charging power of the EV is set as 3 kW. We generate the parking events for each EV respectively using these distributions.

1) http://climate.dest.com.cn/.

Table 2 Performance improvements of SBPI methods using search methods and the US, and the CPLEX-B method when $N_{\text{ev}} = 100$

Method	Reduced action space (%)	Original action space (%)
SBPI + US	46.92 ± 1.45	36.61 ± 1.21
SBPI + MRAS	48.45 ± 1.41	48.25 ± 1.58
SBPI + MARS	48.51 ± 1.46	43.91 ± 1.50
SBPI + COMPASS	48.09 ± 1.46	46.29 ± 1.40
CPLEX-B	24.51 ± 1.48	20.50 ± 1.36

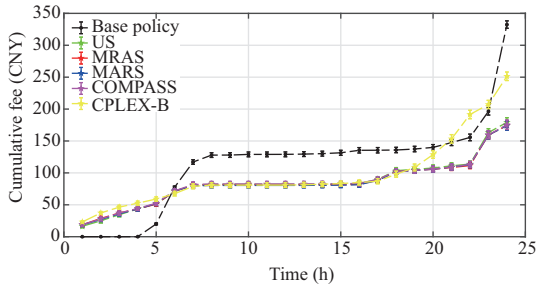

Figure 5 (Color online) Cumulative cost for 100 EVs within 24 h over different policies when using the reduced action space.

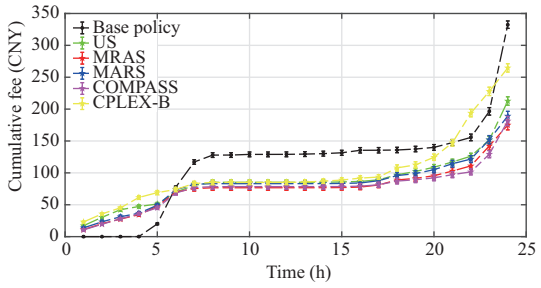

Figure 6 (Color online) Cumulative cost for 100 EVs within 24 h over different policies when using the original action space.

We use the myopic policy as the base policy in SBPI. The myopic policy charges the EVs by the descending order of their urgencies to match the renewable wind power and solar power. Meanwhile, a CPLEX-B method is used for comparison. This method considers M_s stochastic scenarios and uses the CPLEX to calculate the optimal action $\mathbb{A}_t^m$ in each scenario ξ_t^m. Let $\mathbb{A}_t^m = \left[\bar{A}_t^m, \ldots, \bar{A}_\tau^m, \ldots, \bar{A}_{T-1}^m\right]$, where $\bar{A}_\tau^m$ is the corresponding action at time τ. Since only the actions at time t, $\bar{A}_t^m, 1 \leqslant m \leqslant M_s$, are feasible in all of the M_s scenarios, we average them to obtain the decision at time t for the stochastic problem. For the action related to the EVs, we select N_{s} EVs to charge, where $N_{\text{s}} = \frac{1}{M_s}\sum_{m=1}^{M_s}\sum_{i=1}^{N}\bar{a}_t^{i,m}$. The EVs are selected by the descending order of their frequencies in action set $\bar{A} = \{\bar{A}_t^1, \ldots, \bar{A}_t^{M_s}\}$. For the action related to the batteries, we select the decision with the highest frequency in $\bar{A}$ for each building. It is easy to find that the obtained action is also feasible at time t.

Given an initial state S_0 and a base policy $\boldsymbol{\pi}^{\text{base}}$, the performance improvement $\tilde{G}$ is defined as

$$\tilde{G}(\boldsymbol{\pi}^{\text{base}}, S_0) = \frac{\tilde{J}(\boldsymbol{\pi}^{\text{base}}, S_0) - \tilde{J}(\hat{\boldsymbol{\pi}}, S_0)}{\tilde{J}(\boldsymbol{\pi}^{\text{base}}, S_0)}, \tag{39}$$

where $\tilde{J}(\boldsymbol{\pi}^{\text{base}}, S_0)$ is the total charging cost using the base policy $\boldsymbol{\pi}^{\text{base}}$ and $\tilde{J}(\hat{\boldsymbol{\pi}}, S_0)$ is the total charging cost using the policy $\hat{\boldsymbol{\pi}}$. Let $M_p = M_0 = \tilde{M} = 20$, and $M_{\text{total}} = 10000$, the performance improvements of SBPI methods using search methods and the US as well as the CPLEX-B method both in the original action space and the reduced action space over 50 repetitions are given in Table 2.

From the table, we can find that for the US, the performance improvement increases by about 10% when we reduce the action space, which shows that it is easier to make a better decision when we use the structural property to reduce the action space reasonably. Meanwhile, no matter what the action space is, all the search methods outperform the US when we use the SBPI, and the CPLEX-B method is even worse than the SBPI methods. From the values of performance improvement, we can find that all the SBPI methods can achieve an improvement no less than 36%, while searching for the optimal action can additionally support about 7%–12% of the improvement in the original action space and about 1%–2% of the improvement in the reduced action space. So the performance improvement is mainly brought by improving from a base policy rather than searching for the optimal action.

Figures 5 and 6 show the the statistical results of the cumulative cost for 100 EVs within 24 h over different policies when using the reduced action space and the original action space, respectively. We find that using the urgency index to reduce the action space is beneficial for making decisions. In the SBPI methods, we can store the electricity in advance and release the energy at the right time to charge the EVs. Since we only consider very few scenarios and purely average the actions in the CPLEX-B method, it has worse performance than others. This is reflected in the inability to store the electricity at the right

Table 3 Performance improvements of SBPI methods using search methods and the US, and the CPLEX-B method when $N_{\text{ev}} = 300$

Method	Reduced action space (%)	Original action space (%)
SBPI + US	15.67 ± 0.48	11.51 ± 0.25
SBPI + MRAS	16.64 ± 0.42	15.42 ± 0.35
SBPI + MARS	16.18 ± 0.38	14.14 ± 0.33
SBPI + COMPASS	15.93 ± 0.42	11.81 ± 0.26
CPLEX-B	5.26 ± 0.28	4.31 ± 0.30

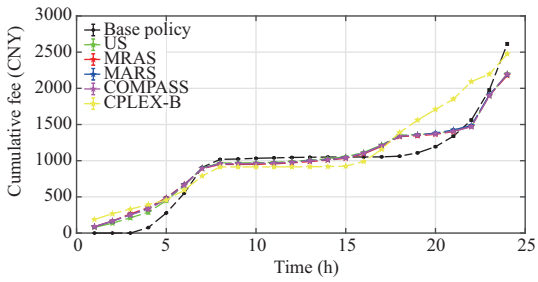

Figure 7 (Color online) Cumulative cost for 300 EVs within 24 h over different policies when using the reduced action space.

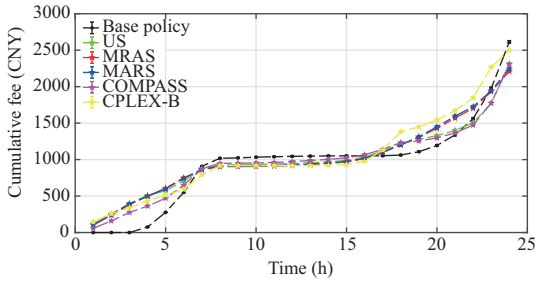

Figure 8 (Color online) Cumulative cost for 300 EVs within 24 h over different policies when using the original action space.

time. Meanwhile, searching for the optimal action in SBPI can help make a better plan for the electricity acquirement and avoid using more high-priced electricity. It is worth noting that the search methods are not limited to those mentioned in this paper. Methods such as PSO can be used to search for the optimal action while the performance may be similar. The CPLEX is not the only tool for programming as well. The GUROBI can be used instead of the CPLEX while the performance is not influenced.

When the number of the EVs grows to 300, the performance improvements of SBPI methods using search methods and the US as well as the CPLEX-B method both in the original action space and the reduced action space over 50 repetitions are given in Table 3, and the statistical results of the cumulative cost in both of the action space are shown in Figures 7 and 8, respectively. It demonstrates that the proposed approach is still appropriate while the performance improvements are lower than the case of 100 EVs. This is because the capacities of the renewable energy and the batteries do not change, which means that the regulating abilities of them become weak.

5 Conclusion

In this paper, we formulate the EV charging problem and explore the advantages of using structural property and search methods. Urgency index is used to reduce the action space and the rationality of the index is proved. Three search methods, MRAS, MARS and COMPASS, are used to find the optimal action in the action space. In MRAS and MARS, optimal ways to update the parameters are given and proved. In COMPASS, the method for uniformly sampling from the most-promising-area is specified in the EV charging problem. We numerically compare the performance among these SBPI methods and compare them with the CPLEX-B method. The results show all these methods perform better in the reduced action space than in the original action space. We also find that the search methods perform better than the US when we use the SBPI to improve the base policy, and all the SBPI methods perform better than the CPLEX-B method. This shows that the SBPI improves the base policy better, and the search methods make a little contribution in the performance.

In the future, we will focus on more general problems. For Assumption 1, if the charging power is not constant, the cost should be related to the SOC of the battery. For Assumption 2, if there is congestion, our approach may be still appropriate because it is equivalent to adding an upper bound of the number of the EVs charging in the parking lot. For Assumption 3, if the renewable energy is not free, the cost function should be modified while the proposed approach is not influenced. For Assumption 4, if the power from the grid is limited, it should be added to the system state. We may also consider the discharging process of the EVs and check whether we would charge the EVs with higher urgency index

and discharge the EVs with lower urgency index.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2016YFB0901900), National Natural Science Foundation of China (Grant No. 61673229), and the 111 International Collaboration Project of China (Grant No. BP2018006).

References

1 Wang H W, Zhang X B, Ouyang M G. Energy and environmental life-cycle assessment of passenger car electrification based on Beijing driving patterns. Sci China Technol Sci, 2015, 58: 659–668
2 Mohammadi K, Alavi O, Mostafaeipour A, et al. Assessing different parameters estimation methods of Weibull distribution to compute wind power density. Energy Convers Manage, 2016, 108: 322–335
3 Murata A, Ohtake H, Oozeki T. Modeling of uncertainty of solar irradiance forecasts on numerical weather predictions with the estimation of multiple confidence intervals. Renew Energy, 2018, 117: 193–201
4 Amjad M, Ahmad A, Rehmani M H, et al. A review of EVs charging: from the perspective of energy optimization, optimization approaches, and charging techniques. Transpation Res Part D-Transp Environ, 2018, 62: 386–417
5 Nimalsiri N I, Mediwaththe C P, Ratnam E L, et al. A survey of algorithms for distributed charging control of electric vehicles in smart grid. IEEE Trans Intell Transp Syst, 2020, 21: 4497–4515
6 Jia Q S, Shen J X, Xu Z B, et al. Simulation-based policy improvement for energy management in commercial office buildings. IEEE Trans Smart Grid, 2012, 3: 2211–2223
7 Xu Y, Pan F, Tong L. Dynamic scheduling for charging electric vehicles: a priority rule. IEEE Trans Automat Contr, 2016, 61: 4094–4099
8 Hu J Q, Fu M C, Marcus S I. A model reference adaptive search method for global optimization. Operations Res, 2007, 55: 549–568
9 Hu J Q, Fu M C, Marcus S I. A model reference adaptive search method for stochastic global optimization. Commun Inf Syst, 2008, 8: 245–276
10 Hu J Q, Hu P, Chang H S. A stochastic approximation framework for a class of randomized optimization algorithms. IEEE Trans Automat Contr, 2012, 57: 165–178
11 Hong L J, Nelson B L. Discrete optimization via simulation using COMPASS. Operations Res, 2006, 54: 115–129
12 Yan Q, Zhang B, Kezunovic M. Optimized operational cost reduction for an EV charging station integrated with battery energy storage and PV generation. IEEE Trans Smart Grid, 2019, 10: 2096–2106
13 Zhong Q W, Buckley S, Vassallo A, et al. Energy cost minimization through optimization of EV, home and workplace battery storage. Sci China Technol Sci, 2018, 61: 761–773
14 Zhang Y M, Wei Z, Li H, et al. Optimal charging scheduling for catenary-free trams in public transportation systems. IEEE Trans Smart Grid, 2019, 10: 227–237
15 Eldeeb H H, Faddel S, Mohammed O A. Multi-objective optimization technique for the operation of grid tied PV powered EV charging station. Electric Power Syst Res, 2018, 164: 201–211
16 Zhang Y M, Cai L. Dynamic charging scheduling for EV parking lots with photovoltaic power system. IEEE Access, 2018, 6: 56995–57005
17 Jia Q S, Wu J. On distributed event-based optimization for shared economy in cyber-physical energy systems. Sci China Inf Sci, 2018, 61: 110203
18 Xie S, Zhong W, Xie K, et al. Fair energy scheduling for vehicle-to-grid networks using adaptive dynamic programming. IEEE Trans Neural Netw Learn Syst, 2016, 27: 1697–1707
19 Wu Y, Ravey A, Chrenko D, et al. Demand side energy management of EV charging stations by approximate dynamic programming. Energy Convers Manage, 2019, 196: 878–890
20 Zhang Y M, You P C, Cai L. Optimal charging scheduling by pricing for EV charging station with dual charging modes. IEEE Trans Intell Transp Syst, 2019, 20: 3386–3396
21 Huang Q, Jia Q S, Guan X. A multi-timescale and bilevel coordination approach for matching uncertain wind supply with EV charging demand. IEEE Trans Automat Sci Eng, 2017, 14: 694–704
22 Puterman M, Puterman M L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Hoboken: John Wiley & Sons, 2014
23 Hastings W K. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 1970, 57: 97–109
24 Leterme W, Ruelens F, Claessens B, et al. A flexible stochastic optimization method for wind power balancing with PHEVs. IEEE Trans Smart Grid, 2014, 5: 1238–1245
25 Wang Z, Jochem P, Fichtner W. A scenario-based stochastic optimization model for charging scheduling of electric vehicles under uncertainties of vehicle availability and charging demand. J Cleaner Production, 2020, 254: 119886
26 Lopez K L, Gagne C, Gardner M A. Demand-side management using deep learning for smart charging of electric vehicles. IEEE Trans Smart Grid, 2019, 10: 2683–2691
27 Wan Z, Li H, He H, et al. Model-free real-time EV charging scheduling based on deep reinforcement learning. IEEE Trans Smart Grid, 2019, 10: 5246–5257

28 Qian T, Shao C, Wang X, et al. Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system. IEEE Trans Smart Grid, 2020, 11: 1714–1723

29 Yang Y, Jia Q S, Deconinck G, et al. Distributed coordination of EV charging with renewable energy in a microgrid of buildings. IEEE Trans Smart Grid, 2018, 9: 6253–6264

30 Chang H S, Hu J, Fu M C, et al. Simulation-based Algorithms for Markov Decision Processes. London: Springer, 2013

31 Sedighizadeh M, Mohammadpour A, Alavi S M M. A daytime optimal stochastic energy management for EV commercial parking lots by using approximate dynamic programming and hybrid big bang big crunch algorithm. Sustain Cities Soc, 2019, 45: 486–498

32 Tushar M H K, Assi C, Maier M, et al. Smart microgrids: optimal joint scheduling for electric vehicles and home appliances. IEEE Trans Smart Grid, 2014, 5: 239–250

33 Blanco M I. The economics of wind energy. Renew Sustain Energy Rev, 2009, 13: 1372–1382

34 Branker K, Pathak M J M, Pearce J M. A review of solar photovoltaic levelized cost of electricity. Renew Sustain Energy Rev, 2011, 15: 4470–4482

35 Grogg K. Harvesting the wind: the physics of wind turbines. Physics and Astronomy Comps Papers. Northfield: Carleton College, 2005

36 Ishaque K, Salam Z, Syafaruddin Z. A comprehensive MATLAB simulink PV system simulator with partial shading capability based on two-diode model. Sol Energy, 2011, 85: 2217–2227

37 Shahidinejad S, Bibeau E, Filizadeh S. Statistical development of a duty cycle for plug-in vehicles in a north american urban setting using fleet information. IEEE Trans Veh Technol, 2010, 59: 3710–3719

38 Bertsekas D P, Castanon D A. Rollout algorithms for stochastic scheduling problems. J Heuristics, 1999, 5: 89–108

39 Huang Q, Jia Q S, Qiu Z, et al. Matching EV charging load with uncertain wind: a simulation-based policy improvement approach. IEEE Trans Smart Grid, 2015, 6: 1425–1433

40 Lee T K, Bareket Z, Gordon T, et al. Stochastic modeling for studies of real-world PHEV usage: driving schedule and daily temporal distributions. IEEE Trans Veh Technol, 2012, 61: 1493–1502