Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

Optimization of multi-vehicle charging and discharging efficiency under time constraints based on reinforcement learning

Peng Liu [a], Zhe Liu [a], Tingting Fu [a], Sahil Garg [b,c], Georges Kaddoum [b,d], Mohammad Mehedi Hassan [e,*]

[a] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China
[b] Electrical Engineering Department, École de technologie supérieure, Montreal, H3C 1K3, Canada
[c] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India
[d] Artificial Intelligence & Cyber Systems Research Center, Lebanese American University, Beirut, 03797751, Lebanon
[e] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Electrical vehicle
Charging–discharging scheduling
Vehicle to grid (V2G)
Cooperative route planning
Reinforcement learning

ABSTRACT

In the Vehicle-to-Grid (V2G) scenario, a multitude of coordinated electric vehicles (EVs) equipped with high-capacity batteries actively participate in power grid dispatching as energy carriers, aiming to achieve a tripartite objective encompassing peak load reduction and valley filling, enhanced utilization of renewable energy sources, and added benefits for electric vehicle owners. To address the existing limitations in the charging–discharging decision-making process for electric vehicles based on V2G, such as the lack of consideration for charging pile constraints, EV profitability, EV transportation timeliness, and high costs associated with central servers, we proposed a reinforcement learning-based Multi-vehicle Joint Routing and Charging–Discharging Decision algorithm (MJRCDD). Firstly, the Markov decision process (MDP) was established to describe the problem, and the route selection and charging–discharging behavior of the vehicle were innovatively integrated in the vehicle action space. Secondly, the multi-vehicle joint route planning and charging–discharging decision problem was solved by multi-agent reinforcement learning. Finally, the effectiveness of MJRCDD was verified by simulation and comparison experiments based on PeMS.

1. Introduction

It is projected that the demand for oil will continue to experience a growth rate of 54 percent until 2035. However, conventional internal combustion engine vehicles typically exhibit an energy efficiency ranging from 20 to 30 percent, indicating that only a fraction of the fuel consumed by vehicles can be effectively converted into mechanical energy. On the other hand, the energy efficiency of electric vehicles is usually between 60 and 70 percent, so it is clear that electric vehicles are more energy efficient than internal combustion engines [1]. In addition to the continued exploitation of traditional energy sources, the continued development of new technologies has increased the efficiency of renewable energy sources such as wind and solar power. By 2050, solar power is expected to provide about 40 percent of the electricity supply, wind power about 31 percent, and other renewable energy sources such as hydro-power, geothermal energy, biomass, etc., about 29 percent of the electricity supply. This means that it is possible for renewable energy to replace traditional fossil fuels as the dominant source of electricity supply [2].

The global electric vehicle fleet is expected to reach 220 million by 2030. For example, an average car needs about 10L of gasoline, while a motorcycle needs about 5L of oil per hundred kilometers. Electric vehicles only need 20 kWh of electricity [3]. From the perspective of cost alone, electric vehicles are more economical than traditional cars and motorcycles. Furthermore, EVs are able to provide ancillary services to the power grid during high demand periods or outage situations [4].

Therefore, considering the advantages of electric vehicles mentioned above, they will definitely be the future trend. However, the large charging demand brought by the high penetration rate of electric vehicles will increase the load of the power grid, especially in peak times and for areas with dense charging stations, it will have a certain impact on the power grid [5]. Although the development of renewable energy power grid can alleviate the energy crisis to a certain extent, its output is affected by weather, season, geography and other factors, and such instability will affect the stability and security of power grid [6]. While many renewable energy generation facilities are built in remote

* Corresponding author.
E-mail address: mmhassan@ksu.edu.sa (M.M. Hassan).

https://doi.org/10.1016/j.aej.2024.08.002
Received 26 March 2024; Received in revised form 19 July 2024; Accepted 2 August 2024
Available online 21 August 2024

areas, the load centers of the grid are usually in urban and industrial areas, which makes it difficult to transfer and consume renewable energy, so much of renewable energy is still discarded.

In order to solve the above problems, Vehicle-to-Grid (V2G) technology [7] comes into being and becomes a very promising way. The technology takes advantage of two-way energy transfers between charging stations and electric vehicles, which can offload surplus energy into the grid [8]. This way of using the energy stored by a large number of EVs as a buffer between renewable energy and the grid can balance the grid load and improve the grid stability. Secondly, it can store renewable energy and reduce energy waste. Finally, it can create additional profit for EV users and further facilitate the popularization of EVs, achieving a win-win situation [9].

Considering the real V2G scenario, electric vehicles can buy and sell energy between Renewable-energy (R-type) charging stations and traditional grid (G-type) charging stations to gain profits and reduce their own travel costs. The fact is that the charging price at R-type station is usually less than the discharging price at G-type station, and is much less than the charging price at G-type station. Therefore, the owners of EVs may make profit by charging at R-type stations and discharging at G-type stations. However, the number of charging and discharging piles in the charging station is limited, and each electric vehicle is bounded by its departure and destination as well as its tolerable travel time. Although many papers have assumed there is an aggregator as a bridge between EVs and the smart grid [10], in this paper, we adopt a distributed method, regarding each electric vehicle in the system as an agent, which achieves the goal of maximizing the overall profits by selecting its actions. The main contributions of this paper are as follows:

- A system model is established to study the optimization of multi-vehicle charging and discharging profits under time constraints by integrating the interaction between vehicles. The Markov decision process (MDP) model is established to describe the state of the agent, the state transition equation is derived, and the appropriate reward function is given. In addition, the vehicle routing and charging and discharging behavior are combined to form the optional actions of the agent innovatively, and the action set and action value function are given.
- A reinforcement learning-based multi-vehicle joint route planning and charging–discharging decision algorithm (MJRCDD) is proposed in this study. The algorithm utilizes MADDPG, a multi-agent general model, and follows the framework of centralized training and decentralized execution. It encompasses predicting other vehicle strategies within the system and training vehicles with strategy sets. By continuously interacting with the road environment and accumulating experience, vehicles can learn the optimal strategy for each state. Finally, a simulation experiment of the algorithm is conducted on PeMS to verify its effectiveness in multi-vehicle and multi-constraint scenarios under time constraints for EV charging and discharging scheduling by comparing it with other algorithms.

The subsequent sections of this paper are structured as follows: Section 2 provides a comprehensive review and analysis of the relevant literature. In Section 3, we present the formulaic expression of the system model and problem. Section 4 introduces and elucidates the multi-vehicle joint routing and charging–discharging decision algorithm (MJRCDD). To evaluate its performance, we simulate the urban traffic map of Santa Clara, California in Section 5. Finally, in Section 6, we summarize our findings.

2. Related work

2.1. Effectiveness of V2G

Many studies have demonstrated the effectiveness of V2G in EV charging and discharging applications. In [11], with V2G as the goal to minimize losses in the system, the author makes a comprehensive cost analysis of strategies based on active power dispatching (APD) and reactive power dispatching (RPD), and verifies that integrating electric vehicles into distribution systems to support the grid by providing a flexible power management is a promising prospect. In [12], the author proposed the water cycle algorithm (WCA) of the EV charging model, and found that the bidirectional operation of the charging station could accommodate more EV without overloading the distribution transformer. Most importantly, using electric vehicles as reactive power compensators can improve grid voltage distribution without reducing battery life. The authors of Ref. [13] propose an intelligent framework based on Internet of Things (IoT) and edge computing to effectively manage V2G operations, which can handle distributed energy, contribute to grid stability, improve its reliability, and improve power efficiency.

2.2. Deficiencies in existing work

With more and more complicated problem scenes in reality, many researches have limitations in practical application. In literature [14], Guo Y et al. proposed a two-stage framework to effectively manage EV charging and discharging behaviors. However, this method is not suitable for real-time scenarios where EV charging demand changes and electricity price changes are more complex. Yang H [15] et al. proposed an optimization model for route selection and charging navigation of electric vehicles. However, the setting that electric vehicles can only sell surplus energy at the destination makes it unable to optimize the economic benefits of owners in real scenarios. Liu P [16] et al. proposed K shortest path joint routing scheduling algorithm (KSP-JRS) based on A* algorithm of artificial intelligence, considering various constraints and aiming at improving the overall economic benefits of electric vehicles. However, the whole algorithm is centralized. When it is applied to super large-scale cities, it is impractical to handle all scheduling tasks only through a limited number of central servers. Subsequently, the author proposed a distributed joint path planning and charge–discharge decision algorithm (JPPCDD) in literature [17] to solve this problem. Unfortunately, the algorithm adopts best-effort in maximum vehicle travel time and does not have accurate constraints.

2.3. Reinforcement learning in V2G

In recent years, reinforcement learning (RL) has been widely used in complex decision-making process [18], and it is often used to solve real-time charging–discharging scheduling problems in V2G field. Yao L et al. [19] coordinated charging tasks of multiple electric vehicles in a parking lot through binary programming strategy, which, however, depended on prior knowledge in the system. In contrast, model-free methods can learn good control strategies based on reinforcement learning and do not depend on any knowledge of the system [20]. Literature [21] uses Q table to estimate action value function through discrete electricity price and charging behavior, but this method can only deal with a small number of states and actions. Secondly, the decentralization procedure has a great influence on the performance. In order to compensate for its limitations, literature [22] employs linear basis functions as an approximation method for the action value function. However, linear approximators are inadequate when confronted with nonlinear problems. In addition, literature [23] uses nonlinear kernel mean regression operators to fit action value functions. However, this method relies too much on the determination of kernel function and its parameters. In general, the limited approximation ability of the above methods hinders their application in real world scenarios.

2.4. Deep learning in V2G

With the development of deep learning, neural networks have the potential to become universal approximators [24]. Deep neural networks have excellent performance in complex mapping learning of high dimensional data and in many complex decision applications. Zhang et al. proposed a new joint charging scheduling and computation offloading scheme (OCEAN) for Electric Vehicle-assisted Multi-access Edge Computing [25]. The paper formulated a cooperative two-timescale optimization problem to minimize the charging load and its variance subject to the performance requirements of computation tasks. In [26], the author proposed an EV charging navigation algorithm based on deep reinforcement learning, aiming to minimize the total driving time and charging cost. In literature [27], the author described EV charging–discharging scheduling problem as a constrained Markov decision process (CMDP) to find a scheduling strategy to minimize charging cost, and proposed a model-free method based on safety deep reinforcement learning (SDRL) to solve the CMDP problem. However, the problem is that these methods do not take into account the effects of multiple electric vehicles, making the study inapplicable to real-world scenarios.

Besides V2G technologies, some work proposed V2V energy trading [28], however, the small energy transferring bandwidth and strict site requirements, limited the application of this technology.

3. System model

3.1. Problem background

As trajectory planning has been a very important research topic in the area of IoV [29], in addition to participating in power grid dispatching as a mobile energy storage device, EV's travel requirements as a means of transportation should not be affected. Uncoordinated charging behavior will easily cause congestion at the charging stations [30]. In this paper, the joint routing and charging–discharging decision scheduling of electric vehicles are studied under the constraint of vehicle travel destination and maximum vehicle travel time.

In this scenario, ensuring that the one-time routing scheme planned for the vehicle and the subsequent charging–discharging decision based on this scheme can meet the time requirements poses a challenge for the game-based JPPCDD method proposed in literature [17]. On the contrary, the vehicle needs to “try” different schemes again and again, actually gets feedback by choosing the action and comparing the time spent after implementation to the maximum travel time, and then trains itself several times based on that feedback to achieve the given goal. The reinforcement learning is grounded in the “trial and error” behavior akin to that observed in animals, wherein the agent receives feedback subsequent to its actions [18], it subsequently learns and deliberates based on this feedback, thereby adapting its behavior more effectively to the environment. In this paper, based on the V2G scenario, each vehicle in the system, as an agent, can achieve the goal of maximizing the overall profits of the vehicle under the constraint conditions by selecting corresponding actions of route selection and charging–discharging.

3.2. Problem formulation

This section introduces the relevant instance definition, relevant variable setting, target problem description and relevant constraint formulation in the scenario. The physical objects in the system include electric vehicles and charging stations. As 6G [31] and privacy preserving [32] can be utilized to support IoV, the communication and data will not be considered in this paper. The symbolic variables related to the vehicle are given in Table 1. The charging–discharging station and its behavior here are the action set of the agent, and its attribute information is given in Table 2. Table 3 shows the related symbolic variables needed for system modeling.

According to the above variables, we can express the objective function as Eq. (1), where the constraints are shown as Eqs. (2), (3), (4), and (5) respectively.

$$Max \sum_{i \in \mathcal{V}} R_i^{final} \tag{1}$$

Table 1
Symbolic variable related to vehicle.

Notation	Description
$\mathcal{V}$	The set of all EV
$Total_i$	The total time specified for the journey of EV i
$Start_i$	The departure of the journey for EV i
End_i	The destination of the journey for EV i
$Einit_i$	The Electric quantity at the departure of EV i
$Emax_i$	Maximum battery capacity of EV i
loc_i	The station location of EV i in the current state
$time_i^{loc_i}$	The time when EV i left its current position
$energy_i^{loc_i}$	The electric charge of the EV i when it leaves its current position
$\mathcal{VN}$	Virtual node in the graph
$\mathcal{N}$	The set of all nodes in the graph

Table 2
Symbolic variable related to station.

Notation	Description
$attri$	The dimensional column vector, consisting of the following symbols, represents all the information about this station
j_{num}^{ch}	Number of charging piles at node j
j_{num}^{dis}	Number of discharge piles at node j
X_j^R	The 0/1 variable that indicates whether node j is a renewable energy charging station
X_j^G	The 0/1 variable that indicates whether node j is a grid charging station
Pow_j^c	Charging power of node i
Pow_j^d	Discharging power of node i
W_{jc}	Charging price of the charging station at node i
W_{jd}	Discharging price of the discharging station at node i
$in(j)$	The set of incoming edges of the node j
$out(j)$	The set of outcoming edges of the node j

Table 3
Symbolic variables related to system modeling.

Notation	Description
$\mathcal{S}$	The set of all the states
$\mathcal{A}$	The set of all actions
R_{t+1}	The corresponding reward value of (s_t, a_t)
R_{ij}	The 0/1 variable that indicates Whether node i can reach node j
j_c	The EV chooses to charge itself at node j
j_d	The EV chooses to discharge itself at node j
j_p	The EV chooses to pass by at node j
T_j^{wait}	The waiting time for EV at station j
αA	Electricity consumption per vehicle distance
v	Vehicle speed
d_{ij}	Distance from node i to node j
R_i^{init}	Initial cumulative reward for EV i
R_i^{final}	Final cumulative reward for EV i
G_t	The system eventually accumulates rewards
$Q_\pi(s,a)$	The value of action a at state s, Q_π for short

$$\sum_{m \in in(j)} X_m^{car} - \sum_{k \in out(j)} X_k^{car} = \begin{cases} -1 & j = Start_i \\ 0 & \forall j \in Z \setminus Start_i, End_i, VN \\ 1 & j = End_i, VN \end{cases} \tag{2}$$

$$0 \le energy_i^{loc_i} \le Emax_i, (\forall i \in V) \cap (\forall loc_i \in N) \tag{3}$$

$$time_i^{End_i} \le Total_i, (\forall i \in V) \tag{4}$$

$$\begin{cases} R_{loc_i k} = 1 \\ R_{k End_i} = 1 \end{cases} \tag{5}$$

R_i^{final} in Eq. (1) represents the final cumulative reward of the EV i, as described in Table 3. Its calculation is given in the reward calculation of MDP below. Since the system aims to maximize the overall vehicle profits, the goal is the sum of the final reward of all vehicles. Eq. (2) is the continuity condition of the route in graph theory, indicating that except for the departure and the destination, the inbound degree of other nodes must be equal to the outbound degree. In this paper, the vehicle stops only when it reaches its destination or virtual node, otherwise it will continue to select the next node, thus ensuring the constraint of Eq. (2). Eq. (3) is the energy constraint, which means that the electric quantity of any vehicle in the system should not be negative at any node, and should not exceed the maximum battery capacity. When discharging, the vehicle is limited to ensure 30% of the maximum electric quantity of it. Because the vehicle in the system is a mobile energy storage device with high battery capacity, 30% of the maximum electric quantity ensures that it will not stop on the road due to lack of electricity, thus meeting the requirements of Eq. (3).

Eq. (4) represents a crucial constraint on vehicle travel time, wherein the duration for any vehicle to reach its destination must not exceed the specified maximum travel time. To address this requirement, this study draws inspiration from reinforcement learning techniques regarding vehicles as agents, employing continuous attempts and final reward assignment. Eq. (5) is a constraint on alternative stations, respectively indicating that the nodes k to be reached by vehicles next are reachable at the current location loc_i and node k can reach the destination of vehicles End_i, which is also in line with common constraints. In contrast to the JPPCDD's scenario, this paper does not specify a particular waiting position as the constraints on the maximum travel time for vehicle charging–discharging behavior are more significant than those on the number of waiting positions. Instead, the waiting time is calculated into the final time expense of the vehicle and is restricted according to the time constraint.

4. Multi-vehicle joint routing scheduling and charging–discharging decision algorithm based on reinforcement learning

To solve the above problem and find out the strategy that satisfies the above constraints and maximizes the objective function under the multi-vehicle scenario, this section proposes the multi-vehicle joint routing scheduling and charging and discharging decision algorithm based on reinforcement learning (MJRCDD). The overall structure of the model is shown in Fig. 1. Firstly, the specific scenario of electric vehicles is abstractable and formulaic, which has been completed in the previous section. Action and other features are extracted to establish the basic Markov decision process of reinforcement learning. The MDP is chosen because it is able to accurately describe and solve decision problems in stochastic processes and many existing works such as [27,33,34] also use it to model the process of EV charging/discharging. Then, based on MDP, the multi-agent depth deterministic gradient strategy MJRCDD provides the joint routing scheduling and charge–discharge scheme under multi-vehicle and multi-constraint scenarios after multiple iteration training for the model environment.

4.1. Establish Markov decision process (MDP)

This section corresponds to the steps of establishing Markov decision process according to the formulated problem in Fig. 1. Markov decision process framework has become a framework suitable for solving most reinforcement learning problems. Simultaneously, the route planning and charging–discharging decision processing of the vehicle in this system involves multiple steps that necessitate comprehensive consideration, with each step's scheduling scheme determined by the preceding step's decision.

(1) **State Space**: The system state space is represented by $s_t = \{loc_t, energy_t^{loc_t}, time_t^{loc_t}, attri_{loc_t}\}$, where loc_t represents the position node of the vehicle at the time of the step t. $energy_t^{loc_t}$ represents the power storage when the vehicle travels to the current location node loc_t. $time_t^{loc_t}$ represents the time taken by the vehicle from the departure to the current location node loc_t. $attri_{loc_t}$ represents the attributes of the current node loc_t. It is a N dimensional column vector, including $X_{loc_t}^R$ which indicates whether the station is a R-type or G-type charging node.

(2) **Action Space**: Action refers to the behavior taken by electric vehicles in the system environment, which can be understood as a bridge between vehicle states. The existence of action makes the static environment move forward. In actual decision-making, vehicles will make corresponding actions in different states according to the guidance of algorithms, and finally acquire learning experience according to the reward feedback of the environment. a_t ($a_t \in A$, A represents the set of all actions) is used to represent the action taken by the electric vehicle in the state s at step t. Due to the limited charging and discharging stations that can be selected as the next step, as well as the three limited optional strategies (charging, discharging and passing through the station), the types of joint station selection and decision making are also limited. Therefore, the action space A can be defined as all the joint station selection and charging–discharging choices. For example, the decisions such as charging k_c, discharging k_d and passing k_p of a vehicle at a G-type station k will appear in the action set as three parallel actions. Of course, the action set also includes other nodes, such as three behaviors corresponding to the station, as shown in Eq. (8).

(3) **State transfer**: The state transfer of the system from step t to step $t+1$ is shown in Eq. (6), (7), (8), and (9), where Eq. (6) is the total state transfer, Eq. (7) is the position state transfer, k is the next station to be selected, and energy transfer is Eq. (8), when the next station k is selected and charged at the station, that is $a_t = kc$, the electric quantity of the vehicle in the step $t+1$ is the maximum electric quantity of the vehicle. Similarly, if the vehicle passes only at the selected station, the electric quantity changes as the electric quantity in the previous state minus the cost of the trip. k_d indicates that the vehicle discharges at the selected node. Formula (9) represents the transfer from step t to step $t+1$ time state, which also depends on the actions a_t taken by the vehicle step t in the state s, where T_k^{wait} represents the waiting time of the vehicle at the station. When there is no need to wait, it is 0, α represents the electric consumption per unit distance of the vehicle, and v represents the vehicle speed.

$$s_{t+1} = f(s_t, s_t) \tag{6}$$

$$loc_{t+1} = k \tag{7}$$

$$energy_{t+1} = \begin{cases} Emax_i & a_t = k_c \\ energy_t^{loc_t} - d_{loc_t k}/\alpha & a_t = k_p \\ Emax_i * 30\% & a_t = k_d \end{cases} \tag{8}$$

$$\mathrm{time}_{t+1}^{loc_{t+1}} = \begin{cases} \mathrm{time}_t^{loc_t} + \left[\mathrm{Emax}_i - \left(\mathrm{energ}_t^{\mathrm{loc}_t} - \frac{d_{loc_t k}}{\alpha}\right)\right] \times Pow_k^c + T_k^{\mathrm{wait}} \\ , a_t = k_c \\ \mathrm{time}_t^{loc_t} + \frac{d_{loc_t k}}{v}, a_t = k_p \\ \mathrm{time}_t^{loc_t} + \left[\ \mathrm{energy}_t^{loc_t} - \frac{d_{loc_t k}}{\alpha} - \left(\mathrm{Emax}_i \times 30\%\right)\right] \times Pow_k^d \\ + T_k^{\mathrm{wait}}, a_t = k_d \end{cases} \tag{9}$$

(4) **Reward function**: The reward R_{t+1} represents the outcome obtained after executing action a_t at time step t in state s. At the end of each action, the income or cost brought by the decision can be calculated as a reward function. Considering that the objective of this chapter is to maximize vehicle income by combining vehicle routing and charging–discharging behavior decision, By examining the reward function, it becomes evident that the lower the charging price or the higher the discharge price at station k, the greater the reward value the vehicle will receive. Meanwhile, the charging and discharging decision behavior of the same station will also affect the vehicle reward. The reward value corresponding to the agent in (s_t, a_t) is shown in Eq. (10):

$$R_{t+1} = \begin{cases} -\left[\ \mathrm{Emax}\ _i - \left(\ \mathrm{energy}\ _t^{\mathrm{loc}\ _t} - \frac{d_{\mathrm{loc}\ _t k}}{\alpha}\right)\right] \times W_{kc}, \dots, a_t = k_c \\ 0, \dots, a_t = k_p \\ \left[\ \mathrm{energy}\ _t^{\mathrm{loct}} - \frac{d_{\mathrm{loc}\ \ \mathrm{k}}}{\alpha} - \left(\ \mathrm{Emax}\ _i \times 30\%\right)\right] \times W_{kd}, \dots, a_t = k_d \end{cases} \tag{10}$$

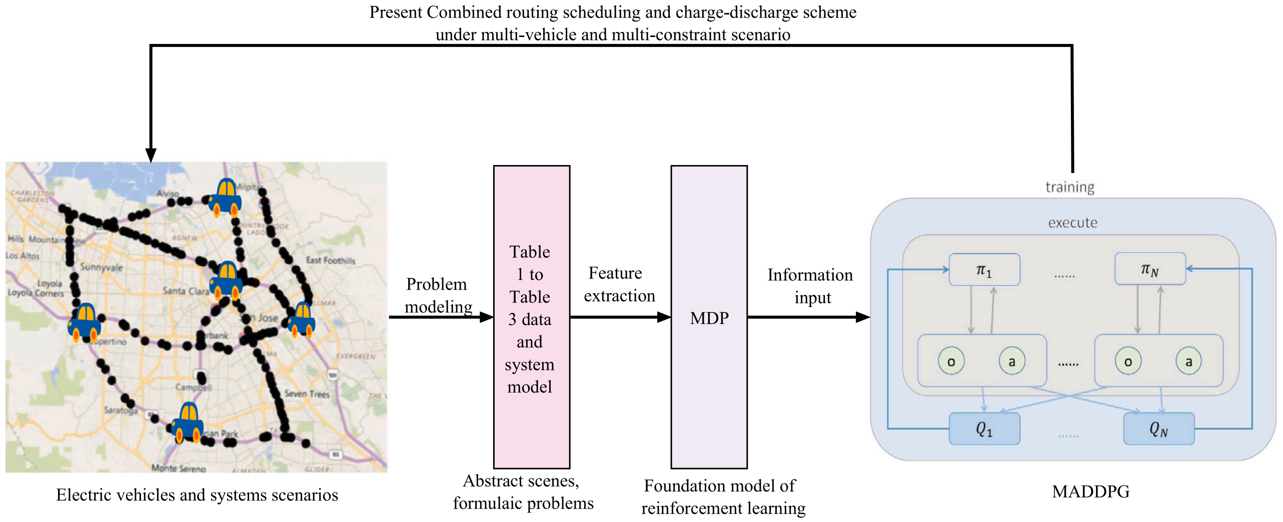

Present Combined routing scheduling and charge-discharge scheme
under multi-vehicle and multi-constraint scenario
Alviso
Milpitas
Sunnyvale
Los Altos
Loyola
Loyola Corners
Santa Clara
San Jose
East Foothills
Cupertino
Campbell
Saratoga
Seven Trees
Monte Sereno
Electric vehicles and systems scenarios
Problem modeling
Table 1 to Table 3 data and system model
Abstract scenes, formulaic problems
Feature extraction
MDP
Foundation model of reinforcement learning
Information input
training
execute
π_1
......
π_N
o
a
......
o
a
Q_1
......
Q_N
MADDPG

Fig. 1. Overview of the method model.

In each step t, a state–action sequence $seq = [s_t, a_t, s_{t+1}, a_{t+1}]$ can be obtained according to the state s and strategy π at the current moment. Here, the cumulative reward of all steps is shown as Eq. (11) :

$$R_i^{init} = R_{t+1} + \gamma R_{t+2} + \gamma^2 \mathrm{R}_{t+3} + \cdots + \gamma^{T-t-1} R_T = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} \tag{11}$$

γ is a constant between [0,1], representing the discount factor. Because the vehicle has to take into account not only the current reward, but also the future reward when calculating the total payoff. However, the longer this interval, the more uncertain the future rewards are, so we can reduce the proportion of future rewards in total revenue. The larger γ is, the more "far-sighted" the agent is. Furthermore, $\gamma = 0$ means that the strategy is "short-sighted" and only considers the current and timely rewards. $\gamma = 1$ means that future cumulative rewards can be obtained in advance without discount calculation. Considering the time requirement and the vehicle's own travel restrictions, as shown in Eqs. (4) and (5), Eq. (12) will be obtained on the basis of Eq. (11). I.e., only when the total time of the vehicle to the destination is less than the maximum travel time cost and it actually reaches the destination, the actual reward will be given to the vehicle. Among them, VN is a virtual node, which belongs to the exit degree of all nodes and will not have its own exit degree, in other words, once the vehicle chooses the station that does not meet Eq. (5) in the scene, after a finite number of steps, they will eventually reach and stop at VN. When the vehicle reaches its true destination, the state will end, and there will be no arrival at VN. Therefore, in addition to the timeout, when the vehicle is at VN, it also does not meet the travel requirements, and at this time, it also directly gives negative rewards to encourage the vehicle to conform to the constraints. The maximization goal in Eq. (12) is the same as that in Eq. (1).

$$R_i^{final} = \begin{cases} \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} \left(\mathrm{time}\, _T^{\mathrm{loc}\ T} \leq \mathrm{Total}_i \right) \wedge \left(\mathrm{loc}_T \neq VN \right) \\ -\sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1} \left(time_T^{loc_{T+1}} > \ \mathrm{Total}\ _i \right) \vee \left(\mathrm{loc}_T = VN \right) \end{cases} \tag{12}$$

On the basis of Eq. (12), further considering the multi-vehicle system of this model, the final cumulative return of the model can be updated into Eq. (13), i.e., the maximum income of all vehicles in Eq. (1) of this paper:

$$G_t = \sum_i^V R_i^{\mathrm{final}} \tag{13}$$

(5) **Action value function**: Because strategies π are probabilistically distributed, there may be many different seq, this paper uses G_t to evaluate the cumulative rewards of states $s_t = s$

$$V_\pi(s) = E_\pi \left[G_t \mid s_t = s \right] \tag{14}$$

V_π is a state-value function, and the corresponding state–action value function is:

$$Q_\pi(s, a) = E_\pi \left[G_t \mid s_t = s, a_t = a \right] \tag{15}$$

This paper aims to maximize the cumulative reward expectation, so the problem can be converted into a strategy π^* for solving Eq. (16) :

$$\pi^* = \mathrm{argmax}_\pi\, Q_\pi(s, a) \tag{16}$$

4.2. Multi-vehicle joint routing scheduling and charging–discharging decision algorithm

In this section, the multi-vehicle joint routing and charging-discharging decision algorithm MJRCDD is proposed based on the multi-agent depth deterministic strategy gradient algorithm based on reinforcement learning. MJRCDD adopts the framework of centralized training and distributed execution. As the actor–critic framework has been approved to be very efficient to handle the hybrid action [35], in the training phase, each vehicle has an actor network and a critic network, in which the critic network collects the state and action of all vehicles in the system and generates a value Q. The actor network of each electric vehicle makes decisions based on its own partial state. At the same time, MJRCDD extends the critic network to learn the strategies of other vehicles, so that each EV performs a function that approximates the strategies of other vehicles. For any $\pi_i \neq \pi_i'$, there is the transition probability shown in Eq. (17) :

$$\begin{aligned} & P\left(s' \mid s, a_1, \ldots, a_N, \pi_1, \ldots, \pi_N\right) \\ & \quad = P\left(s' \mid s, a_1, \ldots, a_N\right) \\ & = P\left(s' \mid s, a_1, \ldots, a_N, \pi_1', \ldots, \pi_N'\right) \end{aligned} \tag{17}$$

$\pi = \left\{\pi_1, \ldots, \pi_N\right\}$ represents a set of policies for all vehicles, $\pi' = \left\{\pi_1', \ldots, \pi_N'\right\}$ represents the deterministic strategy of a vehicle in the target strategy network. In other words, the MADDPG-based MJRCDD is different from most traditional reinforcement learning methods that cannot be directly applied to multi-agent scenarios. MJRCDD speculates the decisions of other vehicles through the extended critic network after learning other vehicles' strategies, and takes the speculated actions of other vehicles as the condition. On this basis, even if the strategy changes, the environment remains static. Secondly, in the multi-agent

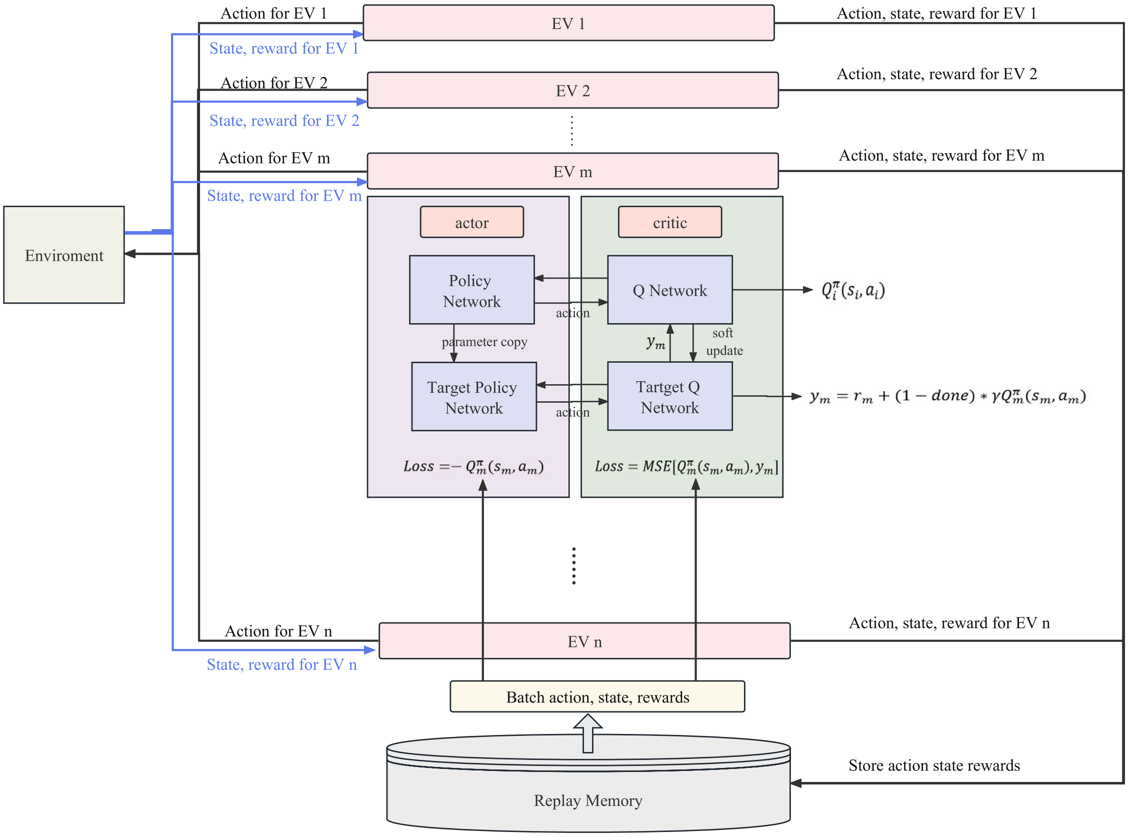


Fig. 2. MJRCDD overall structure diagram.

environment, the strategy of electric vehicles may overfit the actions of other vehicles, resulting in the problem that its own strategy may be ineffective when the strategy of other vehicles changes, so MJRCDD adopts electric vehicles with a strategy set to solve the problem. The overall structure of MJRCDD is shown in Fig. 2.

As can be seen from Fig. 2, MJRCDD is an extension of the actor–critic strategy gradient approach, with each electric vehicle or agent having its own actor–critic network. The overall structure of the algorithm is shown by the black and blue arrows on the left in Fig. 2. The electric vehicle takes actions against the system environment according to its network output, gets rewards from the environment feedback, and updates its state. At the same time, as shown by the arrow on the right, each vehicle stores information such as its motion state reward in a replay buffer represented by a gray cylinder $\mathcal{D}$, and updates the loss function accordingly. The player-critic method for each electric vehicle works as follows.

In this multi-electric vehicle actor–critic model, the deterministic strategy parameter of a vehicle in the actor strategy network in Fig. 2 is $\theta = \{\theta_1, \ldots, \theta_N\}$, the set of all vehicle strategies is $\pi = \{\pi_1, \ldots, \pi_N\}$, then the expected revenue of each vehicle $J(\theta_i)$ can be expressed as Eq. (18) :

$$\nabla_{\theta_i} J(\theta) = \mathbb{E}_{S\sim p^\mu, a_i\sim\pi_i}\left[\nabla_{\theta_i} \log \pi_i\left(a_i \mid s_i\right) Q_i^\pi\left(s, a_1, \ldots, a_N\right)\right] \tag{18}$$

Among them $Q_i^\pi\left(s, a_1, \ldots, a_N\right)$ is a centralized action value function output by a network of critics, which takes as input the action $a_1, \ldots, a_N$ and status information s of all vehicles and outputs the value Q of vehicles. Since each Q_i^π is learned separately, the reward structure of vehicles can be arbitrary. Further, it is extended to the deterministic strategy. For the continuous strategy μ_{θ_i} of an electric vehicle, the parameter is θ_i (abbreviated as μ_i), and the gradient can be written as Eq. (19) :

$$\nabla_{\theta_i} J\left(\mu_i\right) = \mathbb{E}_{s,a\sim\mathcal{D}}\left[\nabla_{\theta_i} \mu_i\left(a_i \mid s_i\right) \nabla_{a_i} Q_i^\mu\left(s, a_1, \ldots, a_N\right)\Big|_{a_i=\mu_i(s_i)}\right] \tag{19}$$

The replay buffer $\mathcal{D}$ here corresponds to the gray cylindrical replay buffer module in Fig. 2, which contains data tuples $\left(s, s', a_1 \ldots, a_N, r_1 \ldots, r_N, \text{done}\right)$ from N electric vehicles to record the experience of all vehicles. Q_i^μ, as mentioned earlier, $Q_i^\pi\left(s, a_1, \ldots, a_N\right)$ represents the concentrated action value function output by the critic network with all vehicle states and actions as inputs, as indicated by the output arrow of the critic Q network in Fig. 2. Q_i^μ is used to evaluate the quality of actor network output strategies. This article updates the critic policy network Q_i^μ by minimizing the loss function, and its update expression is Eq. (20), where y is expressed as Eq. (21):

$$\mathcal{L}\left(\theta_i\right) = \mathbb{E}_{s,a,r,rs'}\left[\left(Q_i^\mu\left(s, a_1, \ldots, a_N\right) - y\right)^2\right] \tag{20}$$

$$y = r_i + (1 - \text{ done }) \times \gamma Q_i^{\mu'}\left(s', a_1' \ldots, a_N'\right)\Big|_{a_j'=\mu_j'(s' j} \tag{21}$$

Among them $\mu' = \left(\mu_{\theta_1}{}', \ldots, \mu_{\theta' N}\right)$ is a target policy set with a delay parameter Q_i'. $Q_i^{\mu'}$ represents the target network based on deterministic policy set μ' and delay parameter θ_i'. The delay parameter θ_i' can be updated by Eq. (22), where τ is the soft update coefficient, corresponding to the soft update process from the critic Q network of electric vehicles to its target Q network in Fig. 2.

$$\theta_i' \leftarrow \tau\theta_i + (1 - \tau)\theta_i' \tag{22}$$

As mentioned above, MJRCDD adopts the method of learning strategies of other vehicles to meet the demand of taking actions of other vehicles as conditions mentioned in Eq. (20). For methods to infer the strategies of other agents, entropy regularizers can be used to learn by maximizing the logarithmic probability of the vehicle's action.

In this method, each vehicle i maintains an approximate value $\hat{\mu}_{\phi_i^j}$ of μ_j of vehicle j (where Φ is the approximate parameter, abbreviated as $\hat{\mu}_i^j$), as shown in Eq. (23) :

$$\mathcal{L}\left(\phi_i^j\right) = -\mathbb{E}_{s_j,a_j}\left[\log \hat{\mu}_i^j\left(a_j \mid s_j\right) + \lambda H\left(\hat{\mu}_i^j\right)\right], \tag{23}$$

where H is the entropy of the strategy distribution and the approximate strategy is used. The approximate value y in Eq. (20) can be replaced by the approximate value y calculated in Eq. (24) below, where $\hat{\mu}_i^{\prime j}$ is the target network of the approximate strategy $\hat{\mu}_i^j$.

$$\hat{y} = r_i + \gamma Q_i^{\mu}\left(s', \hat{\mu}_i'\left(s_1\right) \ldots, \hat{\mu}_i'\left(s_i\right), \ldots, \hat{\mu}_i^{\prime N}\left(s_N\right)\right) \tag{24}$$

In this scenario, there is the problem of non-stationary environment caused by constantly changing strategies of agents mentioned above in multi-agent reinforcement learning [36]. MJRCDD replaces the traditional method of over-matching behaviors of competitors by using vehicles with strategy sets, avoiding the vulnerability of traditional methods that fail when other vehicles change strategies. In order to obtain a multi-agent strategy with more robust policy variation against competing vehicles in the environment, MJRCDD trains k different set of sub-strategies. In each episode, a specific sub-strategy is randomly selected for each agent to execute.

Assume that the policy μ_i is a collection of K different subpolicies, in which the subpolicies k are represented $\mu_{\theta_i^{(k)}}$ (denoted as $\mu_i^{(k)}$). For each vehicle i, we maximize the set objective Eq. (25) :

$$J_e\left(\mu_i\right) = \mathbb{E}_{k\sim unif(1,K), S\sim p^{\mu}, a\sim\mu_i^{(k)}}\left[R_i(s,a)\right] \tag{25}$$

Since different subpolicies are executed in different episode, MJRCDD maintains a replay buffer $\mathcal{D}_i^{(k)}$ for each subpolicy $\mu_i^{(k)}$ of the vehicle i. Therefore, the gradient of the integration target relative to $\theta_i^{(k)}$ is shown in Eq. (26) :

$$\begin{aligned}\nabla_{\theta_i^{(k)}} J_e\left(\mu_i\right) = {} & \frac{1}{K}\mathbb{E}_{s,a\sim D_i^{(k)}} \\ & \times\left[\nabla_{\theta_i^{(k)}}\mu_i^{(k)}\left(a_i \mid s_i\right)\nabla_{a_i}Q^{\mu_i}\left(s,a_1,\ldots,a_N\right)\Big|_{a_i=\mu_i^{(k)}\left(s_i\right)}\right]\end{aligned} \tag{26}$$

The overall process of MJRCDD is shown in Alg. 1. Step 1 to 2 is the initialization of parameters. Step 3 defines the total episode of training. Step 6 to Step 26 is a complete episode, where each agent i will select an action a_t at time slot t. All EVs' states and actions, together with rewards and next states will be inserted into the environment and stored in the replay memory. Then, from Step 11 to 15, each EV will update its critic and actor network. If all EVs have reached their destinations, the current episode ends and the current profit is obtained. The training will be terminated when it reaches the $max\_episode$.

Algorithm 1 Multi-vehicle Joint Routing and Charging and Discharging Decision Algorithm(MJRCDD)

```
1:  Initialize actor policy network $\mu$, target policy network $\mu'$ with
    weight $\theta$ and $\theta'$ for all agents;
2:  Initialize replay memory as $\mathcal{D}$
3:  for each episode $ep = 1$ to $max\_episode$ do
4:    Randomly generate a process $\mathcal{N}$ for actions exploration
5:    Initialize the state $s_0$ and $step = 0$
6:    while done==False do
7:      Select actions $a_t$ for each agent $i$
8:      Enter EVs' states $s$ and actions $a$ into environment
9:      Obtain EVs' rewards $r$ and next states $s'$ by actions $a$
10:     Store $(s, a, r, s'$ , done $)$ in $\mathcal{D}$
11:     for EV $i = 1$ to $N$ do
12:       Sample a random batch of $\mathcal{S}$ as $\{(s^j, a^j, r^j, s'^j$ , done $^j)\}$
          from $\mathcal{D}$
13:       Set the target value $y^j$ as Eq. (21)
14:       Update critic by minimizing the loss:
          $\mathcal{L}(\theta_i) = \frac{1}{S}\sum_j \left(y^j - Q_i^{\mu}\left(s^j, a_1^j, \ldots, a_N^j\right)\right)^2$
15:       Update actor using the sampled policy gradient:
          $\nabla_{\theta_i} J = \frac{1}{S}\sum_j \nabla_{\theta_i}\mu_i\left(s_i^j\right)\nabla_{a_i}Q_i^{\mu}\left(s^j, a_1^j, \ldots, a_N^j\right)\Big|_{a_i=\mu_i\left(s_i^j\right)}$
16:     end for
17:     Update target network parameters for each agent $i$: $\theta_i' \leftarrow \tau\theta_i +$
        $(1-\tau)\theta_i'$
18:     $step = step + 1$
19:     for all EVs do
20:       if EV $i$ reaches the $End_i$ then
21:         $done_i = TRUE$
22:         Get Reward with Eq. (12)
23:       end if
24:       $done = all[done_1, \ldots, done_N]$
25:     end for
26:   end while
27:   Get $G_t$ with Eq. (13)
28: end for
```

Table 4
Type parameters of five electric vehicles in the experiment.

Vehicle type	Battery capacity (kWh)	Initial energy (kWh)	Maximum travel time	Start node	End node
1	10 000	7800	40	B	J
2	7500	6100	35	A	F
3	6800	5200	30	D	J
4	8200	8000	20	C	G
5	6000	5500	35	E	J

5. Experiments and analysis

5.1. Experiment setting

The simulation was based on a real-world traffic map around Santa Clara, California, derived from PeMS, which was a unified traffic data set collected by Caltrans on the California Highway. The data of the system is collected by thousands of sensors mounted on the side of the road every 15 min, marked as black dots in Fig. 3(a). As Directed Acyclic Graph (DAG) can describe many real-world applications [36], we model the traffic map into a directed acyclic traffic graph as shown in Fig. 3(b), which contains 5 G-type stations (black nodes), 3 reversible R-type stations (green nodes), and 2 ordinary intersections (blue nodes).

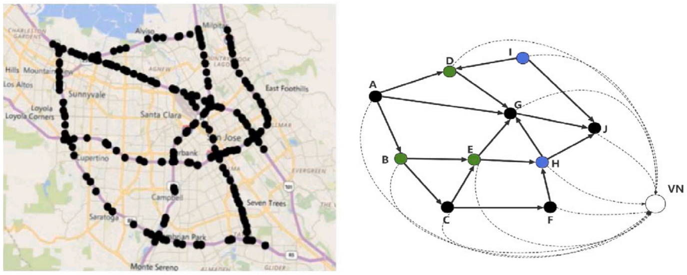
In the experiment, the price is 1 \$/kWh for purchasing electricity at R-type station, 10 \$/kWh for discharging electricity at G-type station, and 20 \$/kWh for charging electricity at G-type station. Without loss of generality, we select five types of vehicles, considering respectively maximum battery capacity, initial electricity quantity, starting node, destination, and maximum driving time, as shown in Table 4. Considering that each vehicle in the MJRCDD algorithm is associated with two actor networks and two critic networks, to limit the calculation scale, we assign 10 electric vehicles for each group.

For evaluation metrics, to better demonstrate the performance of the joint optimization method on total profit and individual transportation timeliness, the first one is the average profit of EVs, which can assess the public benefit. It is evaluated under different settings. The second one is the overtime vehicle ratio, which can evaluate the degree of the algorithm's guarantee of the user's service quality.

5.2. Performance evaluation

5.2.1. Comparison with basic methods

Under the above experimental settings, to increase the competitiveness of charging–discharging station resources among vehicles and



(a) Actual road traffic map

(b) Road traffic map model

Fig. 3. Experimental traffic map.

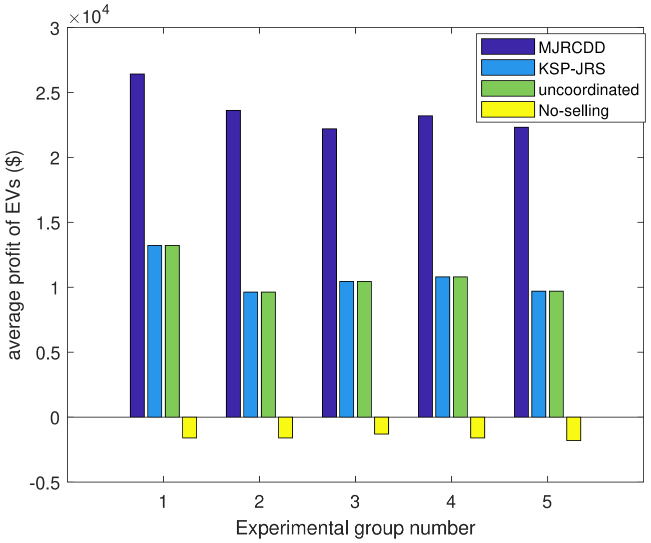


Fig. 4. Overall vehicle profits graph under different algorithms.

study the performance of scheduling algorithm under resource limitation, the 10 electric vehicles in the system come from the same type with the same starting place and destination in Table 4. 10 vehicles in the first group of experiments used the initial battery capacity of 7800 kWh and the maximum battery capacity of 10 000 kWh, which meant that the starting node was the end node and the time limit was 40 min. Another 10 vehicles in the second group of experiments used the second type of vehicles in Table 4. The 10 vehicles in the fifth group of experiments correspond to the third, fourth and fifth types of vehicles in Table 4, so as to verify the superiority and effectiveness of the algorithm in multiple groups of scenes. Similarly, to increase the competition between vehicles, the number of charging and discharging piles in the five groups of scenes is set at 3, and the experimental results are shown in Fig. 4.

From Fig. 4, it can be seen that MJRCDD performs better than other algorithms in five different types of vehicle scenarios. The superiority of MJRCDD over No-Selling, which does not perform charging and discharging behaviors, validates the economic value of vehicles charging and discharging through V2G. Compared with the uncoordinated algorithm directly applied to the optimal solution for a single vehicle in a multi vehicle scenario, MJRCDD demonstrates the benefits of considering the impact of multiple workshops. Compared with the KSP-JRS algorithm that uses the complete path as an optional strategy, the advantages of using a single site and charging/discharging behavior at the site as the action set have been demonstrated.

It is theoretically stated that in the multi-agent environment, the influence between each other will lead to the instability of the environment. In specific scenarios, such as the simplest third group of experiment scenarios, the optimal decision of single vehicle calculated from Uncoordinated is to adopt the route $D- > G- > J$ and discharge behavior is present only at J, so only three vehicles in the system can obtain the established profit, and the profit of other vehicles is 0 because they cannot discharge at the original node J. In MJRCDD, vehicles consider the options of other electric vehicles in the environment when making their own decisions. Some vehicles can discharge at point J to optimize the situation where all vehicles compete for the uncoordinated point. This greatly reduces the competitive relationship between vehicles and enables the full utilization of the "resources" in the entire system, which is also the reason why MJRCDD's experimental results are close to twice as the Uncoordinated algorithm.

Different from MJRCDD, which divides paths $D- > G- > J$ into two schemes ($D- > G(charge)- > J$ and $D- > G- > J(charge)$), KSP-JRS algorithm, which takes complete path $D- > G- > J$ as one of the schemes, has lower overall vehicle income than MJRCDD due to insufficient utilization of charging–discharging stations. In addition, due to the fact that vehicles with D as the starting point and J as the destination only have one route, $D- > G- > J$, KSP-JRS algorithm, compared to the Uncoordinated algorithm, cannot fully leverage the advantage of bringing overall benefits to vehicles by choosing other routes, resulting in a situation where the benefits are the same as the Uncoordinated algorithm. Furthermore, this is also the reason why applying KSP-JRS, the vehicle with the starting node A and the ending node F in the second group of experiment, and the vehicle with the starting node E and the ending node J in the fifth group of experiment, have the same profit with the Uncoordinated algorithm. In the first and fourth group of experiments, although the vehicle has other alternative paths, KSP-JRS can only give up the behavior of increasing the overall income of the vehicle through "detour" due to the constraints of maximum travel time, which also has the same effect on the income of the Uncoordinated algorithm. Despite the constraints of road and time in the experiment, KSP-JRS algorithm still has advantages compared with the Uncoordinated algorithm. This is evident in Fig. 6 when the maximum travel time of the first group of vehicles is extended from 40 min to 50 min. This also verifies the advantages of MJRCDD over KSP-JRS in more scenarios.

5.2.2. The relationship between the overall vehicle profits and the number of charging–discharging piles

In this subsection, the experiment studied the relationship between vehicle income and the number of charging–discharging piles. The first type of vehicle in Table 4 was used in the experiment. In addition, in order to realize that all 10 vehicles are still in a saturated state, the number of charging stations in this experiment is set from 1 to 4 to study the performance of different algorithms on vehicle joint routing planning and optimization of charging–discharging benefits based on the number of charging stations. The experimental results are shown in Fig. 5.

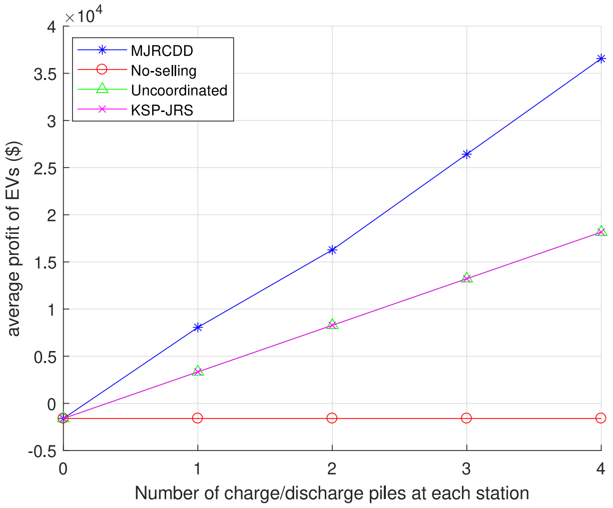


Fig. 5. The relationship between the overall vehicle revenue and the number of charging/discharging piles at the station.

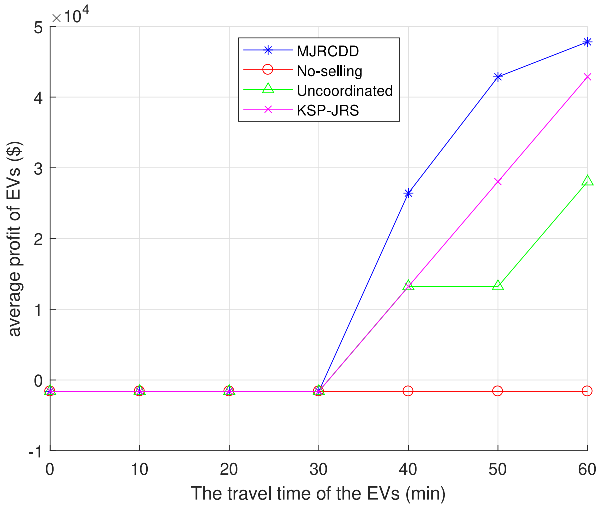


Fig. 6. The relationship between total vehicle revenue and maximum travel time.

As illustrated in Fig. 5, it is evident that No-Selling remains unaffected by the number of charging piles. However, the revenue of vehicles experiences a notable increase with the surge in the number of charging piles across all scheduling algorithms. This phenomenon can be attributed to the fact that the augmented availability of charging–discharging piles provides more vehicles with opportunities to engage in electricity trading through V2G technology, thereby augmenting their income. This trend aligns with common sense and provides a clear explanation for the observed data. It is worth noting that in this experiment scenario, the optimal decision of the bicycle calculated from Uncoordinated is to adopt $B- > E- > G- > J$ route and has discharging behavior only at node J, so when the number of charging piles is increased by one, one more vehicle among the 10 can obtain the path brought by this scheme, and the income of the other vehicles is 0. Therefore, the linear growth process shown in the figure appears. This also demonstrates the limitation of Uncoordinated as a vehicle scheduling algorithm that relies too much on the system scene environment.

In the specific scenario (Fig. 3), the 40-min maximum travel time requirement limits the KSP-JRS from selecting $B- > C- > E- > G- > J$ route, where it can charge and discharge in order to gain revenue at node C. Therefore, KSP-JRS has the same performance as Uncoordinated in Fig. 5. However, when the allowance time is extended, KSP-JRS will appear superior to Uncoordinated, as shown in Fig. 6.

Furthermore, in comparison to Uncoordinated scheduling, MJRCDD adopts the same detour mode as KSP-JRS and targets the path $B- > E- > G- > J$. Unlike all vehicles under Uncoordinated scheduling, MJRCDD offers two distinct schemes ($B- > E- > G- > J(charge)$ and $B- > E- > G(charge)- > J$), which include every site and its charging–discharging behaviors in the action set. This undoubtedly increases the number of charging piles capable of providing charging–discharging services again and reduces competition among vehicles within the system, aspects that are not considered in Uncoordinated scheduling. These factors are overlooked by the KSP-JRS algorithm based on complete routes, thereby explaining why the MJRCDD algorithm achieves higher average profits than KSP-JRS.

5.2.3. The relationship between total vehicle profits and maximum travel time

As mentioned earlier, MADDPG studies the optimization problem of multi electric vehicle charging and discharging under time constraints. Therefore, this experiment explores the impact of maximum vehicle travel time on vehicle revenue. Using 10 vehicles of the first type, with 3 charging and discharging stations, the maximum travel time of the vehicles was increased from 0 to 50 in 10 min. The experimental results are shown in Fig. 6.

As shown in Fig. 6, when the maximum journey time of a vehicle is limited to 20 min, which is less than the time required for the vehicle to travel according to the shortest path, the vehicle cannot reach the destination within the specified time even if it does not consider the charging and discharging behavior. Therefore, it appears that all the benefits of scheduling algorithms are journey costs. When the time is limited to 30 min, although the vehicle can complete the journey, the remaining time is not enough to support its charging and discharging behavior, so the situation also appears that the benefit is the travel expense. However, with the relaxation of time restrictions, in addition to No-selling, MJRCDD, KSP-JRS, and Uncoordinated have the time support to earn income through two-way flow of electricity by relying on V2G technology, and the phenomenon of income increase begins to appear.

In the scenario where the total allowed time consumption is 40 min, the vehicle's revenue is the result shown in Fig. 5 when the number of charging and discharging stations is 3. As analyzed in the previous text, this is because the "trade-off" of MJRCDD's multiple workshops brings better results than uncoordinated ones, as the charging and discharging behavior at a single station replaces the complete path, which brings better results than KSP-JRS.

When the time limit was relaxed to 50 min, an interesting phenomenon emerged that vehicle income increased under MJRCDD and KSP-JRS, while the income remained unchanged under Uncoordinated. This is because the additional ten minutes cannot make the vehicles in the system discharge again at this node after the original three vehicles discharging at node J, so there is no change in income for Uncoordinated. However, when the time continues to increase to 60 min, there is sufficient waiting time, so that the three charging piles at J can each complete two discharge services, that is, six vehicles can obtain the benefits brought by discharge. Therefore, under Uncoordinated scheduling, the benefits of vehicles increase by about two times compared with 40 min and 50 min limitation scenarios.

For MJRCDD and KSP-JRS, when the maximum travel time of vehicles is extended from 40 min to 50 min, although the increased time is not enough for the discharge station in the system to be used again by waiting vehicles, it provides time support for the vehicle to choose "detour" to other available stations to sell additional electricity

Table 5
Type parameters of five electric vehicles in the experiment.

Group number/method	1	2	3	4	5
MJRCDD	0%	0%	0%	0%	0%
JPPCDD	30%	0%	0%	70%	40%

and gain revenue. Therefore, three electric vehicles can obtain income by discharging at node C, which is also the reason for the increase in overall income when the maximum travel time of vehicles under JRCDD and KSP-JRS algorithm is 50 min compared with the maximum travel time of 40 min, and it also reflects the superiority of KSP-JRS scheduling algorithm to Uncoordinated.

From a vertical reading of the figure, it can be seen that MJRCDD results in higher overall vehicle revenue than KSP-JRS and Uncoordinated at the same time. Horizontally, when vehicles achieve similar revenue, the maximum travel time required by the MJRCDD algorithm is less than KSP-JRS and Uncoordinated. As time continues to increase, due to the fact that most vehicles in the system have already had the opportunity to discharge under the scheduling of MJRCDD, there has been a decrease in growth rate. For example, the increase in time from 50 min to 60 min has resulted in a smaller increase in revenue for vehicles than the increase from 40 min to 50 min. When the charging time is extended to 60 min, all vehicles in the system can benefit from it and reach a stable state. Based on this, further speculation and imagination can be made. As time continues to increase and even reaches a value that allows all vehicles to complete the optimal single vehicle path by waiting, MJRCDD will gradually approach this value at a slower and slower rate, KSP-JRS will approach it at a stable rate, and uncoordinated will continue to approach it with increasing amplitude. This also validates the effectiveness of MJRCDD scheduling under finite time constraints.

Effectiveness of MJRCDD under time constraints : This section conducted a comparative study on the effectiveness of MJRCDD in time-constrained scenarios by reducing the overall time cost of JPPCDD. The experiment is also divided into five groups. Each group corresponds to 10 vehicles of one type in Table 4. The number of charging piles is 3. Accordingly, in order to constrain the total JPPCDD time, the waiting position of each site is set to 0. Since JPPCDD only calculates the total time of the trip at the end of the journey, it is impossible to calculate the vehicle income within a given limited time. Combined with the analysis of the previous group of experiments, it is found that the income difference between MJRCDD and JPPCDD under no time constraint is mainly from the maximum discharge of greedy strategy algorithm. Therefore, this experiment directly takes the proportion of vehicles exceeding the given time as the index to evaluate the algorithm conforming to the time constraint scenario, and compares and analyzes the two algorithms. The experimental results are shown in Table 5.

As shown in Table 5, in the first set of data, MJRCDD will take the corresponding path and charging–discharging plan in strict accordance with the time constraint. Based on the decision of other vehicles, the subsequent vehicles in the system will abandon the charging–discharging behavior and only arrive at the end by the shortest path to ensure that all vehicles do not exceed the maximum travel time constraint. Accordingly, even if the waiting position is set to 0, vehicles are not allowed to increase revenue through time-consuming waiting behavior. However, since there is no specific time constraint, the JPPCDD algorithm aiming at maximizing the revenue of vehicle groups will still guide vehicles to increase revenue by detouring to other stations, which is allowed and even encouraged in JPPCDD. Therefore, in the ten-vehicle scenario, three vehicles took a detour to discharge at node C and then timeout occurred. However, for MJRCDD, since timeout would be negatively rewarded in each episode, the trained vehicles would choose to give up this behavior. As there is only one alternative path for the second and third groups of data, JPPCDD has no additional way to earn income compared with MJRCDD. Therefore, after the charging–discharging station in this path is occupied, vehicles are forced to drive from the starting node to the end node without charging and discharging. Hence, JPPCDD also has the situation that the proportion of time-out vehicles in the total vehicles is 0. In the fourth group, the vehicles with C as the starting node and J as the end node, due to the maximum travel time limit of 20, it can be guessed that the owners care much more about the time guarantee than the benefits. However, because there are more optional stations along the route from C to J, more vehicles exceed the maximum travel time limit while JPPCDD creates more benefits. This set of experiments clearly demonstrates the different emphases of JPPCDD and MJRCDD to maximize returns and ensure strict time constraints. Similarly, in the fifth group of experiments, the number of overtime vehicles in MJRCDD is still less than that in JPPCDD. As can be seen from Table 5, as the time allowance or the initial energy decrease, the number of vehicles exceeding the time limit increases. Also, the pair of starting node and destination will affect the result, i.e., the length of the route and whether there are charging stations along the road.

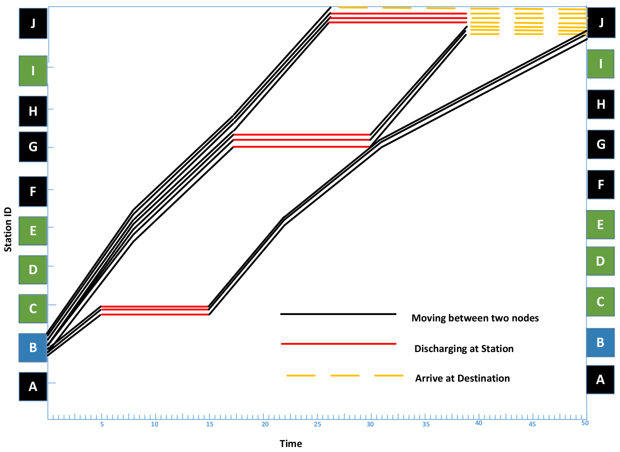


Fig. 7. Behavior diagram of vehicle under MJRCDD scheduling.

5.2.4. Behavior diagram of vehicle under MJRCDD scheduling

Fig. 7 shows the driving conditions of 10 vehicles of the first type in Table 4 when the number of charging piles is 3 and the travel time is 50 min, corresponding to the result of MJRCDD when the charging time is 50 in Fig. 6.

As shown in Fig. 7, there are four behaviors among the 10 vehicles: discharge at point C, discharge at point G, discharge at point J, and direct passing. Among them, three vehicles discharge at point J to generate revenue, which is also the solution chosen by the uncoordinated algorithm analyzed earlier. Unlike this, there are three vehicles in the system that can generate revenue by discharging at point G. Due to the given maximum travel time of 50, the vehicles can also generate revenue by taking a detour to discharge at point C. Finally, as all rechargeable and dischargeable stations are occupied, one vehicle can only abandon the charging and discharging behavior.

6. Conclusion

In this paper, joint routing planning and charging–discharging decision making of multiple electric vehicles under multiple constraints was studied. Utilizing V2G technology, a MJRCDD algorithm based on reinforcement learning was proposed. The algorithm is based on the MDP model which includes vehicle state space, action space, transition between states and reward function. It refers to the multi-agent deep deterministic strategy model, adopts the mode of centralized training and distributed execution, and adds the prediction of other vehicle decisions and vehicles with strategy set on the basis of the actor critic to adapt to the multi-vehicle scene. Vehicle routing and charging–discharging behaviors are combined to form actions based on vehicle motion space, so as to jointly schedule vehicle routing planning and charging–discharging decisions. Finally, simulation experiments and comparison experiments were conducted on PeMS data set and traffic data near Santa Clara, California, to verify the advantages of MJRCDD algorithm.

One of the most important practical concerns is the impact of repeated charging and discharging on the life of the battery. Therefore, in the future, we consider to include State of Charge (SOC) in the algorithm. The total profit will take into account the drain on the battery, so as to better meet the demand of electric vehicle owners.

CRediT authorship contribution statement

Peng Liu: Writing – original draft, Methodology. **Zhe Liu:** Writing – original draft, Methodology, Conceptualization. **Tingting Fu:** Writing – original draft, Validation, Investigation. **Sahil Garg:** Writing – review & editing, Resources. **Georges Kaddoum:** Writing – review & editing, Software, Resources. **Mohammad Mehedi Hassan:** Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62172134). The authors are grateful to King Saud University, Riyadh, Saudi Arabia for funding this work through Researchers Supporting Project Number (RSP2024R18).

References

[1] S. Alshahrani, M. Khalid, M. Almuhaini, Electric vehicles beyond energy storage and modern power networks: Challenges and applications, IEEE Access 7 (2019) 99031–99064, http://dx.doi.org/10.1109/ACCESS.2019.2928639.

[2] I. Aizpuru, A. Arruti, J. Anzola, U. Iraola, M. Mazuela, A. Rujas, Universal electric vehicle charging infrastructure analysis tool, in: 2020 IEEE Vehicle Power and Propulsion Conference, VPPC, 2020, pp. 1–5, http://dx.doi.org/10.1109/VPPC49601.2020.9330990.

[3] W. Zhang, J. Wang, Research on V2G control of smart microgrid, in: 2020 International Conference on Computer Engineering and Intelligent Control, ICCEIC, 2020, pp. 216–219, http://dx.doi.org/10.1109/ICCEIC51584.2020.00050.

[4] Y. Liu, P. Zhou, L. Yang, Y. Wu, Z. Xu, K. Liu, X. Wang, Privacy-preserving context-based electric vehicle dispatching for energy scheduling in microgrids: An online learning approach, IEEE Trans. Emerg. Top. Comput. Intell. 6 (3) (2022) 462–478.

[5] L. Zhang, Y. Liu, L. Hao, X. Zheng, X. Liu, J. Li, Multi-objective optimal scheduling strategy of microgrid based on V2g technology, in: 2022 12th International Conference on Power and Energy Systems, ICPES, 2022, pp. 597–601, http://dx.doi.org/10.1109/ICPES56491.2022.10073154.

[6] H. Ali, S. Hussain, H.A. Khan, N. Arshad, I.A. Khan, Economic and environmental impact of vehicle-to-grid (V2G) integration in an intermittent utility grid, in: 2020 2nd International Conference on Smart Power & Internet Energy Systems, SPIES, 2020, pp. 345–349, http://dx.doi.org/10.1109/SPIES48661.2020.9242992.

[7] T. Fu, C. Wang, N. Cheng, Deep-learning-based joint optimization of renewable energy storage and routing in vehicular energy network, IEEE Internet Things J. 7 (7) (2020) 6229–6241.

[8] D.A. Hagos, Comparing the role of V2G hydrogen fuel cell and V2G electric vehicles for increased integration of VRE in a low carbon neighbourhood, in: 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2022, pp. 1–5, http://dx.doi.org/10.1109/ISGT-Europe54678.2022.9960601.

[9] S. Li, C. Gu, J. Li, H. Wang, Q. Yang, Boosting grid efficiency and resiliency by releasing V2G potentiality through a novel rolling prediction-decision framework and deep-LSTM algorithm, IEEE Syst. J. 15 (2) (2021) 2562–2570, http://dx.doi.org/10.1109/JSYST.2020.3001630.

[10] X. Yao, F. Farha, R. Li, I. Psychoula, L. Chen, H. Ning, Security and privacy issues of physical objects in the IoT: Challenges and opportunities, Digit. Commun. Netw. 7 (3) (2021) 373–384.

[11] J. Singh, R. Tiwari, Cost benefit analysis for V2G implementation of electric vehicles in distribution system, IEEE Trans. Ind. Appl. 56 (5) (2020) 5963–5973, http://dx.doi.org/10.1109/TIA.2020.2986185.

[12] M. Mazumder, S. Debbarma, EV charging stations with a provision of V2G and voltage support in a distribution network, IEEE Syst. J. 15 (1) (2021) 662–671, http://dx.doi.org/10.1109/JSYST.2020.3002769.

[13] V. Chamola, A. Sancheti, S. Chakravarty, N. Kumar, M. Guizani, An IoT and edge computing based framework for charge scheduling and EV selection in V2G systems, IEEE Trans. Veh. Technol. 69 (10) (2020) 10569–10580, http://dx.doi.org/10.1109/TVT.2020.3013198.

[14] Y. Guo, J. Xiong, S. Xu, W. Su, Two-stage economic operation of microgrid-like electric vehicle parking deck, in: 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), 2016, p. 1, http://dx.doi.org/10.1109/TDC.2016.7519948.

[15] H. Yang, Y. Deng, J. Qiu, M. Li, M. Lai, Z.Y. Dong, Electric vehicle route selection and charging navigation strategy based on crowd sensing, IEEE Trans. Ind. Inform. 13 (5) (2017) 2214–2226, http://dx.doi.org/10.1109/TII.2017.2682960.

[16] P. Liu, C. Wang, J. Hu, T. Fu, N. Cheng, N. Zhang, X. Shen, Joint route selection and charging discharging scheduling of EVs in V2G energy network, IEEE Trans. Veh. Technol. 69 (10) (2020) 10630–10641, http://dx.doi.org/10.1109/TVT.2020.3018114.

[17] P. Liu, Z. Liu, N. Zhang, F. Lin, Cooperative game-based charging-discharging efficiency optimization of electric vehicles in 6G-enabled V2G, IEEE Trans. Green Commun. Netw. 7 (2) (2023) 1078–1089, http://dx.doi.org/10.1109/TGCN.2022.3191699.

[18] J. Wang, J. Hu, G. Min, A.Y. Zomaya, N. Georgalas, Fast adaptive task offloading in edge computing based on meta reinforcement learning, IEEE Trans. Parallel Distrib. Syst. 32 (1) (2021) 242–253.

[19] L. Yao, W.H. Lim, T.S. Tsai, A real-time charging scheme for demand response in electric vehicle parking station, IEEE Trans. Smart Grid 8 (1) (2017) 52–62, http://dx.doi.org/10.1109/TSG.2016.2582749.

[20] F. Ruelens, B.J. Claessens, S. Vandael, B. De Schutter, R. Babuška, R. Belmans, Residential demand response of thermostatically controlled loads using batch reinforcement learning, IEEE Trans. Smart Grid 8 (5) (2017) 2149–2159, http://dx.doi.org/10.1109/TSG.2016.2517211.

[21] Z. Wen, D. O'Neill, H. Maei, Optimal demand response using device-based reinforcement learning, IEEE Trans. Smart Grid 6 (5) (2015) 2312–2324, http://dx.doi.org/10.1109/TSG.2015.2396993.

[22] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, G. Deconinck, Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market, IEEE Trans. Smart Grid 6 (4) (2015) 1795–1805, http://dx.doi.org/10.1109/TSG.2015.2393059.

[23] A. Chiş, J. Lundén, V. Koivunen, Reinforcement learning-based plug-in electric vehicle charging with forecasted price, IEEE Trans. Veh. Technol. 66 (5) (2017) 3674–3684, http://dx.doi.org/10.1109/TVT.2016.2603536.

[24] IEEE Draft Framework and Process for Deep Learning Evaluation, IEEE P2841/D2, 2022, pp. 1–30, February 2022.

[25] Y. Zhang, J. Hu, G. Min, X. Chen, N. Georgalas, Joint charging scheduling and computation offloading in EV-assisted edge computing: A safe DRL approach, IEEE Trans. Mob. Comput. (2024) 1–16, http://dx.doi.org/10.1109/TMC.2024.3355868.

[26] T. Qian, C. Shao, X. Wang, M. Shahidehpour, Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system, IEEE Trans. Smart Grid 11 (2) (2020) 1714–1723, http://dx.doi.org/10.1109/TSG.2019.2942593.

[27] H. Li, Z. Wan, H. He, Constrained EV charging scheduling based on safe deep reinforcement learning, IEEE Trans. Smart Grid 11 (3) (2020) 2427–2439, http://dx.doi.org/10.1109/TSG.2019.2955437.

[28] Z. Liu, Y. Xu, C. Zhang, H. Elahi, X. Zhou, A blockchain-based trustworthy collaborative power trading scheme for 5G-enabled social internet of vehicles, Digit. Commun. Netw. 8 (6) (2022) 976–983.

[29] R. Liu, Z. Qu, G. Huang, M. Dong, T. Wang, S. Zhang, A. Liu, DRL-UTPS: DRL-based trajectory planning for unmanned aerial vehicles for data collection in dynamic IoT network, IEEE Trans. Intell. Veh. 8 (2) (2023) 1204–1218.

[30] N. Aung, W. Zhang, K. Sultan, S. Dhelim, Y. Ai, Dynamic traffic congestion pricing and electric vehicle charging management system for the internet of vehicles in smart cities, Digit. Commun. Netw. 7 (4) (2021) 492–504.

[31] H. Li, K. Ota, M. Dong, Learning IoV in 6G: Intelligent edge computing for internet of vehicles in 6G wireless communications, IEEE Wirel. Commun. 30 (6) (2023) 96–101.

[32] T. Li, S. Xie, Z. Zeng, M. Dong, A. Liu, ATPS: An AI based trust-aware and privacy-preserving system for vehicle managements in sustainable VANETs, IEEE Trans. Intell. Transp. Syst. 23 (10) (2022) 19837–19851.

[33] J. Qian, Y. Jiang, X. Liu, Q. Wang, T. Wang, Y. Shi, W. Chen, Federated reinforcement learning for electric vehicles charging control on distribution networks, IEEE Internet Things J. 11 (3) (2024) 5511–5525.

[34] Y. Yang, H. Xu, Z. Jin, T. Song, J. Hu, X. Song, RS-DRL-based offloading policy and UAV trajectory design in F-MEC systems, Digit. Commun. Netw. (2024) http://dx.doi.org/10.1016/j.dcan.2023.12.005.

[35] Z. Cheng, M. Liwang, N. Chen, L. Huang, N. Guizani, X. Du, Learning-based user association and dynamic resource allocation in multi-connectivity enabled unmanned aerial vehicle networks, Digit. Commun. Netw. 10 (1) (2024) 53–62.

[36] J. Wang, J. Hu, G. Min, W. Zhan, A.Y. Zomaya, N. Georgalas, Dependent task offloading for edge computing based on deep reinforcement learning, IEEE Trans. Comput. 71 (10) (2022) 2449–2461.