IET Smart Grid

Research Article

The Institution of Engineering and Technology Journals

Reinforcement learning method for plug-in electric vehicle bidding

eISSN 2515-2947
Received on 28th December 2018
Revised 12th May 2019
Accepted on 21st June 2019
E-First on 1st August 2019
doi: 10.1049/iet-stg.2018.0297
www.ietdl.org

Soroush Najafi[1], *Miadreza Shafie-khah*[2], *Pierluigi Siano*[3], *Wei Wei*[4], *João P.S. Catalão*[5] ✉

[1]*Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran*
[2]*School of Technology and Innovations, University of Vaasa, 65200 Vaasa, Finland*
[3]*Department of Management & Innovation Systems, University of Salerno, Fisciano (SA), Italy*
[4]*State Key Laboratory of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing, People's Republic of China*
[5]*Faculty of Engineering, University of Porto and INESC TEC, Porto, Portugal*
✉ *E-mail: catalao@ieee.org*

Abstract: This study proposes a novel multi-agent method for electric vehicle (EV) owners who will take part in the electricity market. Each EV is considered as an agent, and all the EVs have vehicle-to-grid capability. These agents aim to minimise the charging cost and to increase the privacy of EV owners due to omitting the aggregator role in the system. Each agent has two independent decision cores for buying and selling energy. These cores are developed based on a reinforcement learning (RL) algorithm, i.e. Q-learning algorithm, due to its high efficiency and appropriate performance in multi-agent methods. Based on the proposed method, agents can buy and sell energy with the cost minimisation goal, while they should always have enough energy for the trip, considering the uncertain behaviours of EV owners. Numeric simulations on an illustrative example with one agent and a testing system with 500 agents demonstrate the effectiveness of the proposed method.

1 Introduction

Nowadays, the environmental problem caused by excessive consumptions of fossil fuels is one of the major challenges that human being is facing. Due to this problem, electrifying transportation becomes an emerging research trend in the fields of power system, traffic planning and urban development [1], lead to the proliferation of electric vehicles (EVs). The expected penetration of EVs in the near future creates prospects for a cleaner, more sustainable and more decarbonised future [2].

The charging impact of EVs is a recent issue on the power system. Safety and reliability of power system would be challenged with a high penetration of EVs [3] whose charging demand could be highly volatile. Also, with a proper management scheme, these EVs can be used as virtual energy storage for the power system [4].

Real energy storage systems can mitigate the real-time imbalance between generation and consumption, but their investments are also intensive. Thanks to the development of cutting-edge technology in smart grid, accumulation of a large quantity of EVs can support the power grid through the vehicle-to-grid (V2G) mode. In addition to helping the network for operating with low cost and high reliability, using V2G mode enables the EV owners to reduce their energy cost, because they can sell their surplus energy to the grid and make benefit. Due to impressive effects of EVs on electrical network and in order to increase EV owners' welfare, various kinds of researches have been carried out on this topic [5].

These researches can be categorised into two different subjects: (i) EVs effects on the operation and planning of power systems, (ii) EVs effects on cost reduction for owners.

There are two different paradigms for EV charging management, namely, centralised charging and decentralised charging. In the centralised methods, a central agent (e.g. an aggregator) has been determined to directly control the consumption of the end-users. The objective function to maximise is the aggregator's profit and the main constraint is that of buying enough energy for EV owners' trips.

The centralised method is based on bidirectional communication between end-users and the central agent/aggregator. In order to participate in the electricity market, all the required information about different EVs are transferred to a central level, where the energy consumption is determined and control signals are sent to individual EVs. By applying this approach, an optimum decision with a minimum cost is reached; however, it is not a suitable method for a large number of agents. The increasing number of EVs in a fleet, the corresponding optimisation problem would be complicated and challenging to solve.

When using decentralised approaches, a high communication level is not essential, but it should be enough for broadcasting signals in order to control consumption, as indicated in [6–9]. However, unpleasing outcomes may happen, such as simultaneous reactions, avalanche effects or errors in forecasting the consumer's attitude in relation to the price signals. This occurs due to the fact that the impact of the customers with demand response ability (DR-Enabled) bids is not taken into account. Hence, this type of strategy is probably only suitable at low penetrations [10].

A high communication level, with high speed, high reliability and bidirectional communication, should apply for the decentralised methods to avoid the aforementioned undesirable outcomes. Due to a high communication level, the load synchronisation does not represent anymore a problem when using bidding process strategy [11].

The problem of load synchronisation can also be solved by an iterative process as presented in [12, 13]. Each end-user must determine its bidding price based on the energy demand for given trip, and market clearing price signal is determined on the basis of end-users bid.

In Table 1, the advantages and disadvantages of different control methods are presented. Several references investigated effects of EV and renewable energy resource (RES), simultaneously. In [14], an EV charging policy is proposed that considers transmission and distribution integration issues and reacts to spatial and temporal market signals.

In [15], a stochastic optimisation model is investigated for optimal bidding strategies of EV aggregators in day-ahead energy and ancillary service markets with variable wind energy. A system integrating V2G, grid-to-vehicle and RESs with a converged fibre-wireless (FiWi) communication infrastructure is investigated in [16], and its performance is examined from perspectives in the communication level and physical (power system) level. The effects of EVs and RESs on the environment and cost are studied in [17]. If an aggregator manages EV energy demand appropriately, the uncertainty of RESs can be managed. In [18], a method is proposed for compensating the forecast errors of wind power plants by using EVs. In [2], an approach for bidding into the day-ahead electricity market is represented with the objective of minimising the charging costs while satisfying the EVs' flexible demand.

In [19], dynamic programming is proposed for developing a decision support algorithm and a market participation policy for charging scheduling of EVs connecting to a distribution network feeder. A multi-layer model based on multi-agent systems (MASs) and dynamic incomplete information game theory is developed in [20] with the aim of investigating the electricity markets in a smart grid environment, in which both the bidding strategy and demand response (DR) programs are considered.

In [21], a Bayesian neural network is employed for predicting the electricity prices. The primary goal of [21] is minimising the long-term cost of charging the battery of an individual EV. In [22], a strategy is proposed for implementing a MAS developed in Java Agent Development framework for distribution systems with intermittent distributed generators and EVs offering V2G. In [23], an agent-based framework for studying the behaviour of a retail market with DR is proposed by employing machine learning techniques to model the behaviour of the agents at different levels of a hierarchical framework, where Q-learning is employed to solve the decision-making problem of the consumers. In [24], a novel method for DR program based on Q-learning algorithm (QA) is presented in order to achieve cost reduction. In [25], a mobility-aware V2G control algorithm is proposed that considers the mobility of EVs, states of charge of EVs and the estimated/actual demands of MGs, and then determines charging and discharging schedules for EVs. For obtaining the mobility of EVs and the estimated/actual demand profiles of MGs, a reinforcement learning (RL) approach is also introduced. Improving regulation performance of EVs with an optimal control strategy for EVs based on RL has been proposed in [26]. A RL algorithm is proposed in [27] to learn an optimum cost reducing charging policy from a dataset of historical transition samples and then it exploited to make charging decisions in any situations.

Although various reports in the literature have been studied the control and scheduling of EVs, a decentralised model based on QA enabling V2G mode has not been addressed. In this paper, a novel method for minimising EV cost is presented. In this method, each agent has two decision cores for buying and selling energy. This method guarantees end users privacy during participation in the electricity market without any aggregator or central agent. An efficient aggregator has a lot of information from EV owners. Departure time, arrival time, fuel consumption, unexpected driving and so on. Meanwhile, for this massive information exchanging between the aggregator and EVs through communication networks, gigantic infrastructure is needed. The power networks and communication networks together compose a complicated network, which needs control strategies designed to handle the interaction between EVs and the smart grid [28]. Also, agents do not have any information about each other for the sake of privacy.

Table 1 Comparing advantages and disadvantages of different control methods

Control method		Advantages	Disadvantages
centralised	bidirectional communication	• achieve optimal outcomes • uncertainty for different consumers can be better managed	• scalability and complexity for the aggregator • high communication requirements • user privacy
decentralised	unidirectional communication	• low communication requirements • properties of battery modelled in a more detailed way	• undesirable outcomes, hence the impact of consumers demand on the prices is neglected • only effective for low agents penetration
	bidirectional communication	• work close to real-time and agents' demands are taken into account • user privacy is guaranteed	• high communication requirements

Moreover, users have the freedom to choose participating in only one market (sell or buy) or leaving the market. Another benefit of this method is that EV owners do not get involved in the bidding process.

As it said before, there are two decision cores and each core work with QA. It works by learning an action-value function that gives the expected utility of taking a given action in a given state and following a fixed policy thereafter. The most two significant strengths of the Q-learning are that it can compare the expected utility of the available actions without modelling the environment and it can be used on-line. Q-learning is well suited for solving sequential decision problems, where the utilities of actions depend on a sequence of decisions made and there exists uncertainty about the dynamics of the environment [29].

Thus, the novel contributions of the proposed method can be summarised as follows:

- In the proposed method, each agent has two decision cores to buy and sell energy;
- The method guarantees end users privacy due to the participation in the electricity market without any aggregator or central agent. Users have the right to choose the participation in only one market (sell or buy), or even to not participate in any markets.
- EV owners do not get involved in the bidding process.

The rest of this paper is organised as follows. Section 2 describes the problem formulation. In this section, the price of selling and buying and the amount of the required energy are determined. Furthermore, QA and the electricity market conditions are briefly explained. In Section 3, an illustrative numerical study is presented. This section has a simple example of one agent, as well as a market with a constant price to illustrate how each agent acts by using the proposed method. The main case study with 500 agents is described in Section 4, and simulation results are presented. Finally, Section 5 concludes the paper.

2 Problem formulation

In this section, the mathematical formulation for modelling and simulating EVs and the electricity market is presented. In general, each agent wants to optimise its cost. Hence, it participates in the electricity market to buy energy from it when the price is low (light load time). Also, it can sell energy to electricity market with high price (peak time) in order to make profit. So, agents can be buyer and seller simultaneously. At the first step, the prices for buying and selling energy are determined based on the previous state and previous cost for each vehicle. After that, these agents participate in the electricity market, and the amount of energy that each agent buys or sells is determined. Then, those prices and the amount of buying or selling energy are sent to the QA as input signals. This RL algorithm determines the next state and the next action for each agent. The specified reward for each agent is the total paying cost that means the cost of buying energy minus benefits for selling energy. Moreover, if the algorithm cannot provide enough energy at the departure time, a significant penalty would be assigned in the QA. With this penalty factor, the next state will be changed to avoid failing in the next departure.

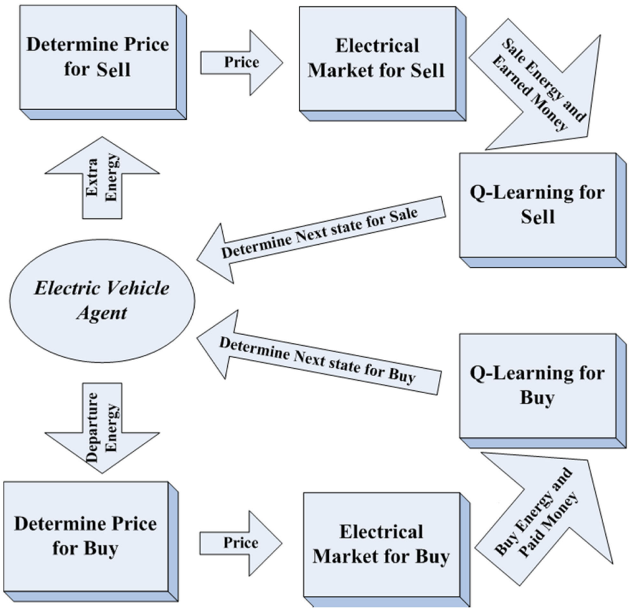

Determine Price for Sell
Price
Electrical Market for Sell
Sale Energy and Earned Money
Extra Energy
Q-Learning for Sell
Determine Next state for Sale
Electric Vehicle Agent
Determine Next state for Buy
Q-Learning for Buy
Departure Energy
Determine Price for Buy
Price
Electrical Market for Buy
Buy Energy and Paid Money

Fig. 1 *Diagram of participating in the electricity market*

Each agent tries to minimise its cost; however, this goal is hard to achieve, because of the time-varying environment of the electricity market, uncertain driving patterns and random energy consumption for each trip. Moreover, an agent performs no action between the departure and arrival times. It is assumed that EV owners do not access the electricity market and smart pile between the departure and arrival times.

The process of bidding and participating in the energy market is presented in Fig. 1.

The considered problem formulation can be represented by considering three phases. In the first phase, an agent should determine the required energy and bidding prices for buying energy in the local electricity market. Agents should buy enough energy for their trip before the departure time. Otherwise, a penalty factor is applied to them. The second phase is related to the selling energy to market for obtaining profit. EV owners should determine the price of energy at which they want to sell to the electricity market. The third phase that is somehow the main one, involves the QA. As has been mentioned before, the proposed model has two decision cores. Their functionalities are complementary: the one for buying energy wants to purchase enough for the trip and the one for selling wants to gain more profit; hence, it wants to determine a price so that it can sell more energy to the electricity market. They are related to each other with the penalty factor. If the agent did not have enough energy for the trip the reward of the QA for both cores is set to 10^{10}.

2.1 Demand energy of electric vehicle

EV energy demand depends on the physical battery constraints, left time to departure and a number of adjustable parameters that will be further explained. The maximum and minimum charging power for time step t are presented here with $P_{\min,v}$ and $P_{\max,v}$, respectively. They are calculated by taking both energy and power constraints into account [6]

$$P^t_{\max,v} = \min\left(\frac{C_v - E_v^{t-1}}{\eta_v \Delta t}\ ,\ P_{c,v}\right) \tag{1}$$

Battery capacity for vehicle v defines here with C_v and E_v^{t-1} indicates the energy content at the end of the previous time step, η_v the charging efficiency, Δt the duration of each time step, $P_{c,v}$ the maximum possible charging power and d_v^t is the left time to trip

$$P^t_{\max,v} = \max\left(\min\left(\frac{E^t_{\text{req},v}}{\eta_v \Delta t} - P_{c,v}(d_v^t - 1), P_{c,v}\right), 0\right) \tag{2}$$

$E^t_{\text{req},v}$ is the required energy for the trip with considering the present energy of battery, and can be calculated as

$$E^t_{\text{req},v} = \max\left(0,\ E^t_{\text{dep},v} - E_v^{t-1}\right) \tag{3}$$

In this part of formulation, demand energy and bidding price are determined. There are two possibilities for (2):

Possibility 1: $\dfrac{E^t_{\text{req},v}}{\eta_v \Delta t} > P_{c,v}\left(d_v^t - 1\right)$ In this case, the required energy is more than the maximum charging possibility and the minimum energy cannot be zero.

Possibility 2: $\dfrac{E^t_{\text{req},v}}{\eta_v \Delta t} < P_{c,v}\left(d_v^t - 1\right)$ In this case, the required energy is less than the maximum charging possibility and the minimum energy in this time step is zero. Hence, buying energy is not urgent and agents can wait for less price in the electricity market.

It is obvious from (2), the value of $P^t_{\min,v}$ is either zero or positive. If $P^t_{\min,v}$ is greater than zero, the charging is necessary in order to be able to depart with a desired battery content. $P^t_{\min,v}$ should not be lower than $E^t_{\text{dep},v}$, where $E^t_{\text{dep},v}$ is the energy that needs to be in the battery for the next departure. In this paper, the required energy for departure is determined randomly, and thus it is independent of the trip time. Each EV determines prices for buying and selling with two bid blocks. $P^t_{\min,v}$ is equal to zero when the charging of an EV is completely flexible. If the vehicle does not have any interactions with the electricity market at the given time step, then $P^t_{\max} = P^t_{\min,v} = 0$ clearly holds. This is similar for the agents who have full battery energy. In (4), the intermediate point from $P^t_{\min,v}$ and $P^t_{\min,v}$ is determined as the energy that agents want to buy from market. In this paper, the bid function has only two blocks, although it can have more blocks without loss of generality. The bid price of the second block, denoted as $\mathscr{P}^t_{b,v}$ is defined as a function of the charging urgency and two tuning parameters Pa_v, Pb_v. Each pair of $(Pa_v,\ Pb_v)$ is defined as a buy state in QA

$$P^t_{\text{int},v} = (P^t_{\max,v} + P^t_{\min,v})/2 \tag{4}$$

$$\mathscr{P}^t_{b,v} = Pa_v + Pb_v\left(\frac{E^t_{\text{req},v}}{\eta_v P_{c,v}}\right) \tag{5}$$

These two parameters are the fraction of the energy that needs to be charged for the next trip and the energy that could potentially be charged until departure [17]. Pa_v is an agent interest for buying energy from the market. If the vehicle has charged more than what it needs for the next trip, $E^t_{\text{req},v}$ is zero. However, it might still want to charge further if prices are low enough. This is the case when prices fall below parameter Pb_v.

There is a limitation for agents that want to buy energy from the market. Moreover, their bidding price should be higher than the market price. In this equation, $E_{\text{Buy}F,i,t}$ is the final energy that agent i buys in hour t. EL is the energy limitation assigned for these agents. If an agent cannot buy enough energy, its trip would fail and as a 'reward', it gets a high penalty. Hence, the agent learns to forecast price more precisely and bid appropriately

$$\sum_{i=1}^{n} E_{\text{Buy}F,i,t} \le EL \tag{6}$$

2.2 EV sell energy

If vehicle owners want to use V2G option, they should participate in the electricity market and sell their surplus energy to the grid.

QA core for selling energy should guarantee that the EV has enough energy for the departure. Moreover, if unpredictable evidence happens for EV owners and they have to drive more than what they predicted, the battery should have enough energy to

TABULAR ALGORITHM

```
1] Determine $t = 0$.

2] Agent choose a random state. Hence, the price is determined.

3] Agent participate in electricity market and the amount of energy that it buy or sell would determine.

4] The reward of agent would determine.

5] Begin $Q_i(a_i) = 0$ $\forall i \in \{1, \cdots, n\}$ and $\forall a_i \in A_i$.

6] Choose for each factor $i$ an action $a^t_i$ by applying a $\varepsilon$-Greedy policy.

7] Update for each factor $i$ its $Qi$-function according to:

    $Qi(a^t_i) \leftarrow Qi(a^t_i) + \alpha ti(r_i(a^t_1, \cdots, a^t_n) - Qi(a^t_i))$

8] The next state is determined.

9] $t \leftarrow t + 1$.

10] If enough number of episodes has been played, then stop. Otherwise, back to step 3.
```

Fig. 2 *Algorithm of RL agents interacting with a matrix game*

cover this extra energy usage. In this case, a security coefficient (α) is considered for selling less energy to the grid. With increasing or decreasing (α), the cost and the battery energy reliability for unpredictable events can be adjusted. As already mentioned, an agent does not take any action between departure and arrival times. The price that each agent bids and participates with it in the electricity market is based on the consumption level. Hence, each agent will have a multiple price rating according to its consumption level.

The formulation considers two different positive constant variables similar to Section 2.1. One is independent of any other variables, and the other is divided into the agents left time hour. $\mathscr{P}^t_{s,v}$ is the agent bidding price for selling energy in the electricity market and $E^t_{\text{sell},v}$ is the energy that EV can sell to the grid due to security coefficient

$$\mathscr{P}^t_{s,v} = \; Sa_v + Sb_v\left(\frac{E^t_{\text{req},v}}{\eta_v P_{c,v}}\right) \tag{7}$$

$$E^t_{\text{sell},v} = E^{t-1}_v - \left((1+\alpha)^* \, E^t_{\text{dep},v}\right) \tag{8}$$

Sa_v represents the agents' interest in selling energy, i.e. the price that agents consider to be high enough to sell energy even if there is an urgency for the next departure. Sb_v represents the sensitivity to the urge of buying energy of the willingness to buy electricity. Meanwhile, the state variables depend on the energy requirement; actions may have a wider range of values previously defined. Each pair of (Sa_v, Sb_v) is defined as a sell state in QA.

2.3 Bidding strategy and reinforcement learning

For running in an environment with different agents, RL can be beneficial. With RL these agents can take actions to maximise their cumulative rewards. With this objective function, they all take actions appropriately. RL is the problem faced by an agent that must learn behaviour through trial-and-error interactions within a dynamic environment [30]. Q-learning is a model-free RL algorithm, and it is typically easier to implement especially in a time-varying environment [31].

In this paper, each EV is defined as an agent. Agents try to place bids and determine the required energy for buying and selling optimally. Each agent has two decision cores, and each core learns how to act optimally based on the QA. Therefore, each agent can buy and sell energy in each time step simultaneously.

Before further explanation about the application of QA some terms should be defined:

- *Action:* The set of all possible moves for agents. It is obvious that there is a set of all possible actions and each agent should select its action among the list.
- *State (S):* A state is a specific place that an agent fined itself. States for each agent will change by using different actions.
- *Reward (R):* The success or failure of an agent is measured by reward. An agent sends output in the form of actions to the environment, and the environment returns the agent's new state (which resulted from acting on the previous state) as well as rewards, if there are any. Hence, evaluating agent's action is done by rewards.

When agent i is modelled by a QA, it keeps in memory a function $Q_i\!: A_i \rightarrow R$ such that $Q_i(a_i)$ means that it will obtain the expected reward if it plays action a_i. Agents always observe the environment and each agent plays the action that leads to the highest reward. After each action and calculation the obtained reward Q_i is going to be updated. For example, if agents play the game for tth time, the joint action $(a^t_1, \ldots, a^t_n)$ represents the actions that different agents have taken. After each episode, different rewards for each agent r_i are obtained and then each agent updates its Q_i-function according to the following equation:

$$Q_i(a^t_i) \leftarrow \; Q_i(a^t_i) + \; \lambda^t_i(r_i(a^t_1, \; \ldots, \; a^t_n) - \; Q_i(a^t_i)\,) \tag{9}$$

where $\lambda^t_i \in [0, 1]$ is the correction degree for QA. If $\lambda^t_i = 1$, the agent supposes that the expected reward that it gets by taking action $a_i = \; a^t_i$ in the next episode is equal to the reward it just observed. Hence, there is not any update in the Q_i value and agents always play the same action. Otherwise, if $\lambda^t_i = 0$, the agent does not take account its last observation to update the value of its Qi-function. In this paper, λ^t_i is always assumed to be 0.9.

Regarding the fact that the QA provides the agent information to know the most profitable actions, a policy for decision making is required. In this paper, ε-greedy policy has been used. With this policy, agents will not exploit just the available information, and with $(1-\varepsilon)$ probability a random available action is selected.

Fig. 2 shows the flowchart of the algorithm that simulates RL driven agents interacting with a matrix game. The number of games after which the simulation should be stopped (step 8 of the algorithm) depends on the purpose of the study.

2.4 Driving behaviours

Driving behaviour is highly uncertain. In this paper, we assumed three different variables for each agent. These variables consist of departure time, arrival time and energy consumption for the trip. Also, for the sake of simplicity, it is a general assumption that each agent has only one trip during a day. This is not a substantial barrier and multiple trips can be easily considered in the model. All variables are modelled by the probability density function of Gaussian distribution. The interval for departure time is between (4, 7) and for arrival time is (16, 22). The energy that each agent consumes during its trip is between (0, 16).

2.5 Electricity market

When considering a V2G approach for all agents, they should interact with the electricity market for buying and selling energy. These interactions benefit both the grid and the consumers. For the power grid, EV batteries play the role of energy storage and contribute DR capabilities which shave the peak and fill the valley.

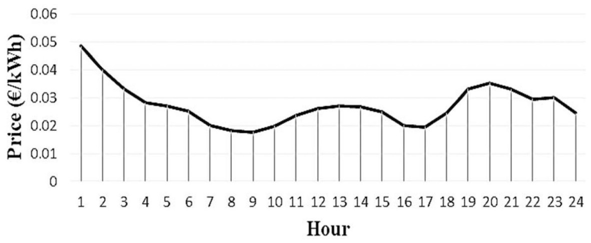


Fig. 3 *Daily electricity market price*

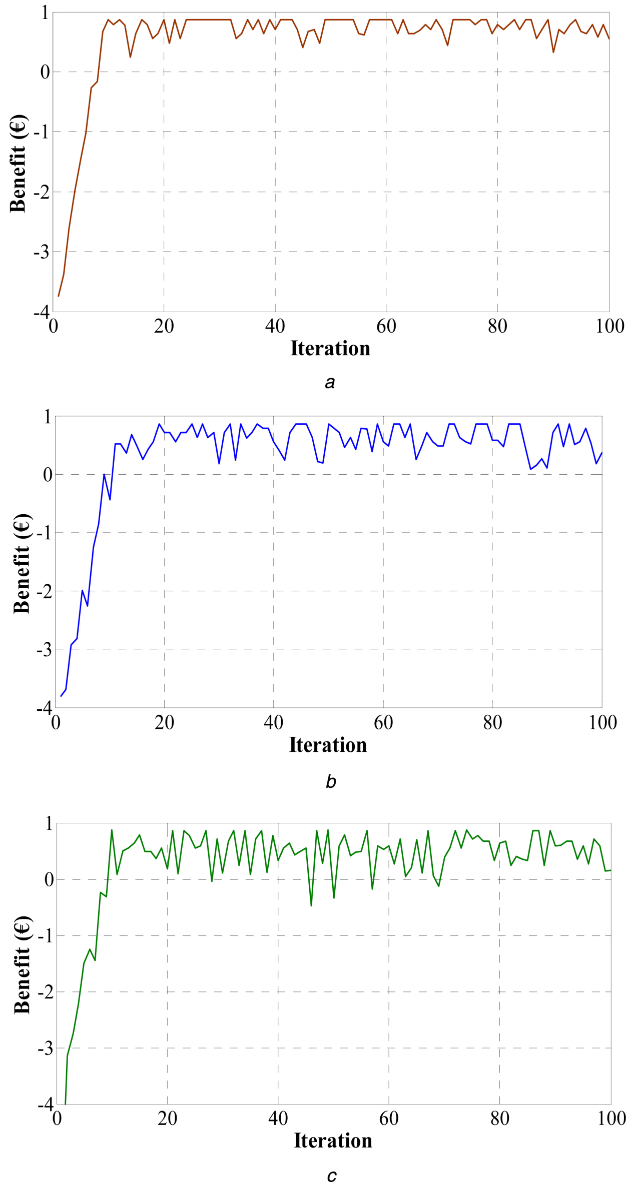


Fig. 4 *Simulation results for the illustrative example and comparing different ε on average benefit*

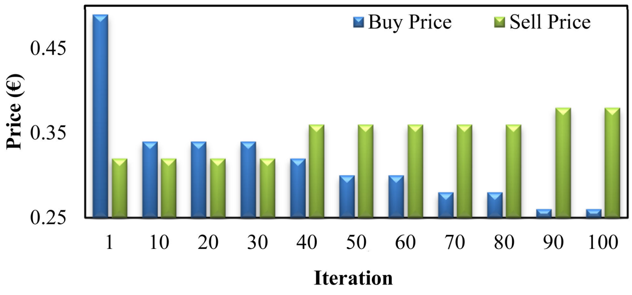


Fig. 5 *Agents bidding price for buy and sell energy*

For EV owners, they can buy cheap energy during off-peak hours and sell it to the grid during peak hours.

Two different kinds of electricity markets are assumed in this paper. The first one is for agents that wish to sell energy to the network. Due to many advantages that the buying energy bring for the network owner, there is not any limitation on that. Buying energy from the market increases the reliability, shaving peak power and also reduces the total loss. Hence, as long as an agent has extra energy for selling to the network, the electricity market buys it. The second market is for selling energy to the agents.

Unlike selling energy, buying energy has a limitation. This limitation is based on the total number of agents. So, each agent should bid appropriately. If the bidding price for buying energy from the market is always low, it may not be able to buy enough energy for completing its trip, which will be penalised in turn.

3 Illustrative numerical studies

In this section, a static environment with a single agent and specific electricity market price is presented for illustrating the strong ability of QA to learn how to act optimally. For simulating the simple static environment, there are some assumptions as follows:

i. Departure time is always at 6:00 AM.
ii. Arrival time is always at 5:00 PM.
iii. Energy usage for the trip is always 11.8 kW.
iv. The agent has no limitation for selling energy to the electricity market.
v. Price for buying energy from the market across the day is shown in Fig. 3, and remains unchanged during 1000 days.
vi. There are 100 different states for determining the buying and selling prices.
vii States for selling and buying are similar.
.

QA with ε-greedy policy is employed in this example with three different quantities. The total benefit for this agent and different amounts of ε is represented in Figs. 4*a*–*c*. In these figures, each iteration represents average benefit of 10 days. The strength of QA can be observed in these figures. With exploiting the environment, the benefit of this agent has ascending trend. As can be seen, with increasing ε, more exploiting has occurred, but the final results for all of these cases are not considerably different.

A more accurate study on ε is carried out in the following. It is a simple example of a static environment with a single agent who wants to maximise its benefit. In Fig. 5, bidding price for participating in the electricity market is plotted by the red line. The first price that the agent bids to the electricity market for buying energy is about 0.5 €/kWh. The core of buying energy has about 50% reduction. Also, the price for selling energy increases about 0.02 €/kWh.

Optimum ε can vary based on the nature of the problem. For determining the best amount of ε in this case study, a simple test has been carried out. In this test, the previous study repeats ten times and the result of the average cost for each iteration is demonstrated in Table 2. If $\varepsilon = 0.01$, due to less exploration, the agent cannot change its state during 1000 iterations. Hence, it was a risky decision for setting $\varepsilon = 0.01$.

Depending on the agent's initial state, either desirable or unfavourable outcomes could happen. If ε is set to 0.05, outcomes for ten repetitions are smooth due to high exploration ability. Although an agent should pay the electrical energy cost, the average cost is still higher than the one with $\varepsilon = 0.1$. Hence, it seems 0.1 is an appropriate value of ε in this case. Moreover, there is a limitation on the contribution of EVs in the electricity market. For determining this amount, another simple example with 100 agents has been performed. All seven assumptions that mentioned in the previous example are still authentic. In Fig. 6, the percentage of the agents which failed on their trips due to different EL is presented. Due to learning behaviour of the electricity market and other agents, these are descending diagrams.

If EL assigns to 100 kWh, about 90% of agents have failed on their trips. With increasing their contribution, the percentage of the

Table 2 Comparing different ε value in average cost for ten repetitions

No.	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$
1	0.264	0.462	0.041
2	0.252	0.524	0.257
3	0.306	0.299	0.517
4	0.287	0.417	0.609
5	0.190	0.512	−0.965
6	0.225	0.496	0.621
7	0.123	0.423	0.505
8	0.296	0.499	−0.339
9	0.207	0.534	−0.528
10	0.076	0.482	0.616

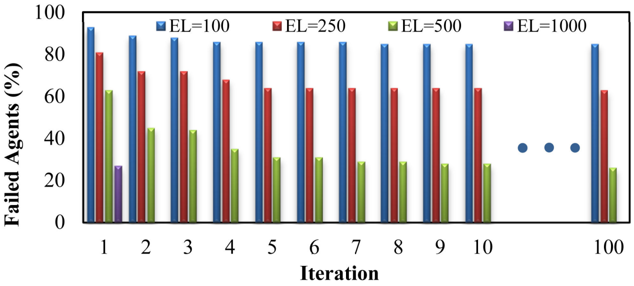


Fig. 6 *Percentage of fail agents due to different EL*

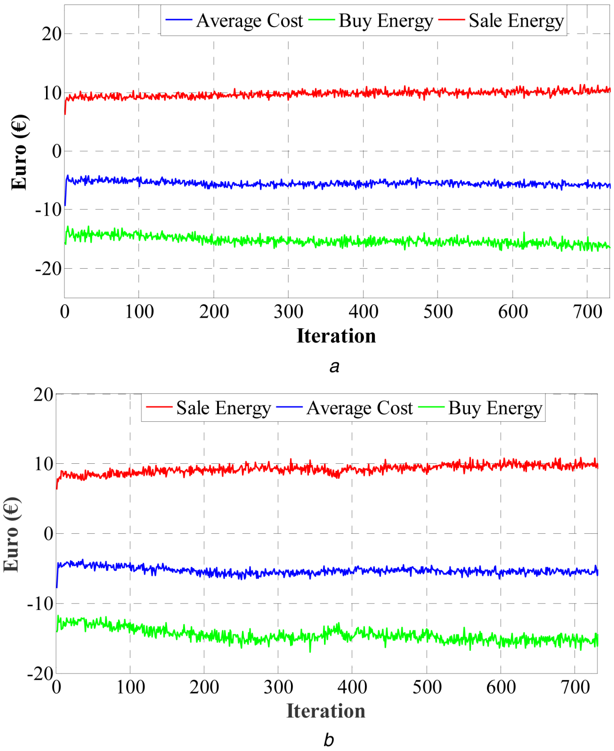


Fig. 7 *Average cost in two scenarios*

failed agents decreases. If EL determines 1000 kWh, only in the first iteration failing the agents has occurred, and in the rest of iterations there are no failed agents. Another notable occurrence is that agents learn to minimise their trip failure, so there is a descending trend for each share of energy. Hence, each agent should be prudent and use proper strategy for buying energy. If its bid is too low to buy energy from the market, there is not enough energy for the trip, and the trip can fail. For avoiding this incident, a substantial penalty factor is determined for each agent.

According to these results, it seems 800 kWh to be an appropriate amount for assigning to agents. With this limitation, the average benefit begins from −17€ to −0.7€.

4 Numerical study

In the case study with configurations of practical interests, 500 agents and dynamic electrical price retrieved from real data in the Spanish electricity market are considered. The aim of all agents is reducing their cost by implementing QA. It is assumed that all agents share the same battery capacity of 16 kWh. $P_{c,v}$ parameter that mentioned in (1) is 13.7 kW all the time. Sometimes unpredictable events can occur, thus there should be enough energy in the battery. Hence, a security coefficient is assumed in (8). With this coefficient, although they sell less energy to the market, there is always enough energy in the battery.

There are 100 different states for participating in the electricity market. Bidding processes for buying energy or selling energy is different. As mentioned before, we assigned two different cores for each agent that can choose two different states for buying and selling. Price set for Pa_v and Sa_v consists of 5 different prices, and for Pb_v and Sb_v consists of 20 different prices; hence, there are 100 different states for bidding and participating in the energy market.

As what was said in Section 3, the electricity market buys all of the energy that each agent wants to sell to the grid but it determines a limitation for selling to this agent. This limitation determines 3000 kWh for each hour. With increasing this limitation, the number of agents that failed is reduced. Based on this limitation, some agents cannot afford sufficient energy and their trips have failed. Five hundred agents are considered in this case study. It is not a limitation and number of agents can increase without increasing much complexity. Spanish electricity market is simulated here for 732 days (two years). Driving pattern for each agent is generated randomly for every day. Also, agent's driving pattern is not similar to each other. Three scenarios are specified in this paper as follows:

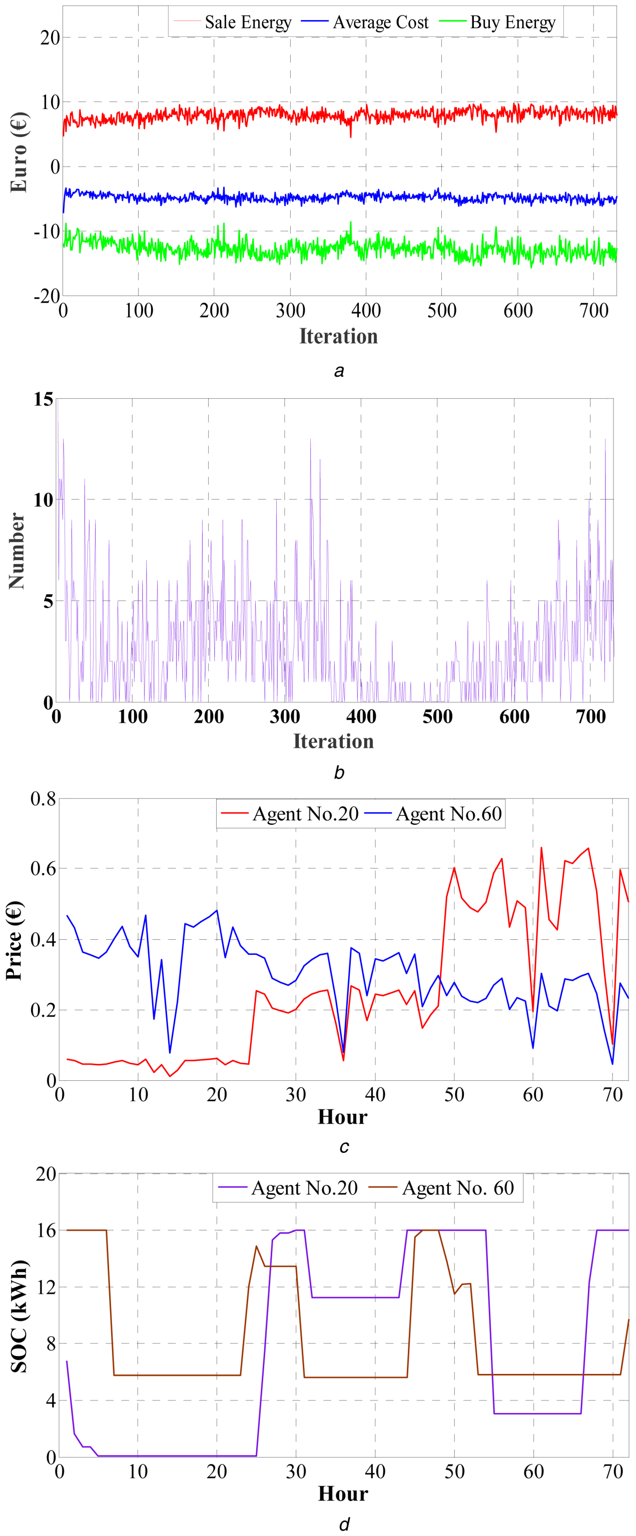
- *Scenario 1* – Bidding price for each agent is constant during a day,
- *Scenario 2* – Bidding price for each agent can change three times during a day,
- *Scenario 3* – Bidding price for each agent can change in every period during a day.

Scenario 1 represents the fixed-rate strategy. On this basis, in this scenario, each agent picks a state for buying energy and another one for selling energy to the market. Hence, the bidding price is constant in different hours for a day. After determining the price, they participate in the electricity market for providing enough energy for the trip. The cost of buying energy from the market, selling energy and absolute payment are presented in Fig. 7*a*. The average cost for each agent is about 5.5€ per day.

Scenario 2 represents the time of use strategy. Therefore, in this scenario, each agent determines a state for participating in the electricity market. Afterwards, an algorithm is run to calculate three different weights for determining the price in three time segments based on last week price. These time segments are [00:08, 08:16, 16:24]. Therefore, bidding prices remain constant for eight hours. These weights are similar for selling and buying, and they come from the previous week. The average cost for this scenario is decreased as presented in Fig. 7*b*. In the first day, 223 agents are failed on their trip, but after some iteration, the failed agents decrease to less than ten agents.

Scenario 3 represents the dynamic tariff strategy. In this scenario, after picking a state for buying and selling energy, 24 weights produced based on the previous week. This scenario has the minimum average cost within all scenarios. The result of this scenario is shown in Fig. 8. The average cost for each agent for two years is presented in Fig. 8*a*. The average cost for agents in this scenario is about 4.8€. In Fig. 8*b*, the number of agents that failed during each day is illustrated.

As aforementioned, the number of failed agents can increase or decrease by changing the market limitation. In Fig. 8*c*, the bidding price for two sample agents (agents no. 20 and no. 60) for buying energy is presented for 72 h (day 99 to day 101). State of the charge (SOC) for agent no. 20 and no. 60 is presented in Fig. 8*d* for



a

b

c

d

Fig. 8 *Simulation results for Scenario 3*

the same period. The minimum SOC is considered zero in this study.

The descending trend of failing agents is shown in Fig. 9 for three scenarios. The first scenario is shown with blue line. In the first day, more than 200 agents failed on their trip, but in all scenarios the number of failed agents decreases to <4% after some iterations.

Moreover, in Table 3, the average cost for different scenarios is presented. If the bidding price for buying and selling energy can change in every period during a day, the lowest cost is achieved.

5 Conclusion

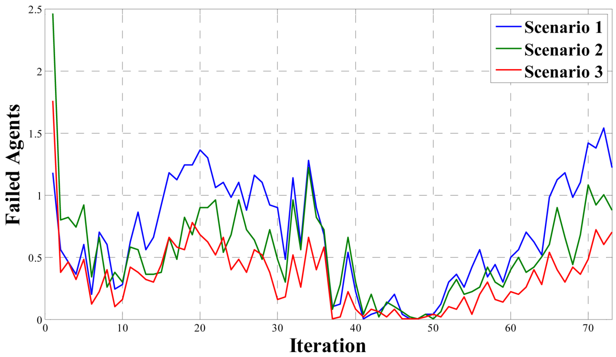


Fig. 9 *Percentage of failed agents during two years*

Table 3 Average cost for different scenarios

	1st scenario	2nd scenario	3rd scenario
average cost (€)	5.5133	5.3225	4.8223

In this paper, a novel method for minimising EVs energy cost with V2G ability was presented. In this method, each EV is defined as an agent that had two decision cores for buying and selling energy to the market. The first one tried to minimise its cost, while the second one maximised its benefit due to adjusting its price. Each core applied QA for decision making and each agent had individual driving behaviour. In this method, no central agents/aggregator was involved, thus the privacy of EV owners was completely guaranteed. Several numerical studies were investigated. At first, with an illustrative numerical study in a static environment, the strength of this method was demonstrated. The results showed that, by applying QA for participating in the electric market, the agent should not have paid for the energy and, in some cases, it could even gain monetary benefits. Then, the number of agents was increased and the ascending trend for the average benefit was noticeable. For evaluating this method in the real world, a dynamic electricity market was simulated with stochastic driving behaviour. Moreover, three different scenarios were defined, and agents could choose any one of them. In all the scenarios, the average cost for the agent was close to zero. Therefore, with two intelligent cores for participating in the electric market, the cost of energy exhibited significant reduction and in some cases the agents could benefit monetarily.

6 Acknowledgment

J.P.S. Catalão acknowledges the support by FEDER funds through COMPETE 2020 and by Portuguese funds through FCT, under POCI-01-0145-FEDER-029803 (02/SAICT/2017) and POCI-01-0145-FEDER-006961 (UID/EEA/50014/2019). The work of M. Shafie-khah was supported by FLEXIMAR-project (Novel marketplace for energy flexibility), which has received funding from Business Finland Smart Energy Program, 2017-2021 (https://www.businessfinland.fi/en/for-finnish-customers/services/programs/smart-energy-finland/).

7 References

[1] Neyestani, N., Damavandi, M.Y., Shafie-Khah, M., *et al.*: 'Plug-in electric vehicles parking lot equilibria with energy and reserve markets', *IEEE Trans. Power Syst.*, 2017, **32**, (3), pp. 2001–2016

[2] Vagropoulos, S.I., Balaskas, G.A., Bakirtzis, A.G.: 'An investigation of plug-in electric vehicle charging impact on power systems scheduling and energy costs', *IEEE Trans. Power Syst.*, 2017, **32**, (3), pp. 1902–1912

[3] Yang, W., Xiang, Y., Liu, J.: 'Agent-based modeling for scale evolution of plug-in electric vehicles and charging demand', *IEEE Trans. Power Syst.*, 2018, **33**, (2), pp. 1915–1925

[4] Jenkins, A.M., Patsios, C., Taylor, P.: 'Creating virtual energy storage systems from aggregated smart charging electric vehicles', *CIRED-Open Access Proc. J.*, 2017, **2017**, (1), pp. 1664–1668

[5] Vandael, S., Claessens, B., Ernst, D., *et al.*: 'Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market', *IEEE Trans. Smart Grid*, 2015, **6**, (4), pp. 1795–1805

[6] Vayá, M.G., Roselló, L.B., Andersson, G.: 'Optimal bidding of plug-in electric vehicles in a market-based control setup'. Proc. Power Systems Computation Conf., Wroclaw, Poland, 2014, pp. 1–8

[7] Rotering, N., Ilic, M.: 'Optimal charge control of plug-in hybrid electric vehicles in deregulated electricity markets', *IEEE Trans. Power Syst.*, 2011, **26**, (3), pp. 1021–1029

[8] Bashash, S., Moura, S.J., Forman, J.C*., et al.*: 'Plug-in hybrid electric vehicle charge pattern optimization for energy cost and battery longevity', *J. Power Sources*, 2011, **196**, (1), pp. 541–549

[9] Hoke, A., Brissette, A., Maksimović, D*., et al.*: 'Electric vehicle charge optimization including effects of lithium-ion battery degradation'. Proc. IEEE Vehicle Power Propuls. Conf., Chicago, USA, Sep. 2011, pp. 1–8

[10] Lopes, J.P., Soares, F., Almeida, P.R.: 'Identifying management procedures to deal with connection of electric vehicles in the grid'. IEEE Power Tech, Bucharest, Romania, 2009, pp. 1–8

[11] Kok, K., Scheepers, M., Kamphuis, R.: 'Intelligence in electricity networks for embedding renewables and distributed generation'. Intelligent Infrastructures, ser. Intelligent Systems, Control and Automation: Science and Engineering Series, New York, 2009

[12] Ma, Z., Callaway, D.S., Hiskens, I.A.: 'Decentralized charging control of large populations of plug-in electric vehicles', *IEEE Trans. Control Syst. Technol.*, 2013, **21**, (1), pp. 67–78

[13] Gan, L., Topcu, U., Low, S.H.: 'Optimal decentralized protocol for electric vehicle charging', *IEEE Trans. Power Syst.*, 2013, **28**, (2), pp. 940–951

[14] Foster, J.M., Trevino, G., Kuss, M*., et al.*: 'Plug-In electric vehicle and voltage support for distributed solar: theory and application', *IEEE Syst. J.*, 2013, **7**, (1), pp. 881–888

[15] Wu, H., Shahidehpour, M., Alabdulwahab, A*., et al.*: 'A game theoretic approach to risk-based optimal bidding strategies for electric vehicle aggregators in electricity markets with variable wind energy resources', *IEEE Trans. Sustain. Energy*, 2016, **7**, pp. 374–385

[16] Xu, D.Q., Joos, G., Levesque, M*., et al.*: 'Integrated V2G G2V and renewable energy sources coordination over a converged fiber-wireless broadband access network', *IEEE Trans. Smart Grid*, 2013, **4**, (3), pp. 1381–1390

[17] Gholami, A., Ansari, J., Jamei, M*., et al.*: 'Environmental/economic dispatch incorporating renewable energy sources and plug-in vehicles', *IET Gener. Transm. Distrib.*, 2014, **8**, (12), pp. 2183–2198

[18] Vayá, M.G., Andersson, G.: 'Self scheduling of plug-in electric vehicle aggregator to provide balancing services for wind power', *IEEE Trans. Sustain. Energy*, 2016, **7**, (2), pp. 886–899

[19] Foster, J.M., Caramanis, M.C.: 'Optimal power market participation of plug-in electric vehicles pooled by distribution feeder', *IEEE Trans. Power Syst.*, 2013, **28**, (3), pp. 2065–2076

[20] Shafie-khah, M., Catalão, J.P.S.: 'A stochastic multi-layer agent-based model to study electricity market participants behavior', *IEEE Trans. Power Syst.*, 2015, **30**, (2), pp. 867–881

[21] Chiş, A., Lundén, J., Koivunen, V.: 'Reinforcement learning-based plug-in electric vehicle charging with forecasted price', *IEEE Trans. Veh. Technol.*, 2017, **66**, (5), pp. 3674–3684

[22] Sivanand, H.S.V., Nunna, K., Battula, S*., et al.*: 'Energy management in smart distribution systems with vehicle-to-grid integrated microgrids', *IEEE Trans. Smart Grid*, 2018, **9**, (5), pp. 4004–4016, doi:10.1109/TSG.2016.2646779

[23] Dehghanpour, K., Nehrir, M.H., Sheppard, J.W*., et al.*: 'Agent-based modeling of retail electrical energy markets with demand response', *IEEE Trans. Smart Grid*, 2018, **9**, (4), pp. 3465–3475

[24] Najafi, S., Talari, S., Gazafroudi, A.S*., et al.*: '*Decentralized control of DR using a multi-agent method*' (Springer, Cham, Switzerland, 2018), pp. 233–249, ISBN:978-3-319-74411-7

[25] Ko, H., Pack, S., Leung, V.C.M.: 'Mobility-aware vehicle-to-grid control algorithm in microgrids'. *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**, (7), pp. 1–10

[26] Ye, X.Z., Ji, T.Y., Li, M.S*., et al.*: 'Optimal control strategy for plug-in electric vehicles based on reinforcement learning in distribution networks'. 2018 Int. Conf. on Power System Technology (POWERCON), Guangzhou, China, 2018, pp. 1706–1711

[27] Chiş, A., Lundén, J., Koivunen, V.: 'Reinforcement learning-based plug-in electric vehicle charging with forecasted price', *IEEE Trans. Veh. Technol.*, 2017, **66**, (5), pp. 3674–3684

[28] Wang, Q., Liu, X., Du, J*., et al.*: 'Smart charging for electric vehicles: A survey from the algorithmic perspective', *IEEE Commun. Surv. Tutor.*, 2016, **18**, (2), pp. 1500–1517

[29] Xin, S., Leung, H.-F.: 'A Q-learning based adaptive bidding strategy in combinatorial auctions'. Proc. of the 11th Int. Conf. on Electronic Commerce, Taipei, Taiwan, 2009

[30] Nguyen, N.D., Nguyen, T., Nahavandi, S.: 'System design perspective for human-level agents using deep reinforcement learning: a survey', *IEEE. Access.*, 2017, **5**, pp. 27091–27102

[31] Zhang, Q., Lin, M., Yang, L.T*., et al.*: 'Energy-efficient scheduling for real-time systems based on deep q-learning model', *IEEE Trans. Sustain. Comput.*, 2019, **4**, (1), pp. 132–141, doi:10.1109/TSUSC.2017.2743704