International Journal of Precision Engineering and Manufacturing-Green Technology (2022) 9:885–897
https://doi.org/10.1007/s40684-021-00403-x

Online ISSN 2198-0810
Print ISSN 2288-6206

REGULAR PAPER

A Deep Reinforcement Learning-Based Energy Management Strategy for Fuel Cell Hybrid Buses

Chunhua Zheng[1] · Wei Li[1,2] · Weimin Li[1] · Kun Xu[1] · Lei Peng[1] · Suk Won Cha[3]

Received: 15 May 2021 / Revised: 23 August 2021 / Accepted: 16 October 2021 / Published online: 21 December 2021

Abstract

An energy management strategy (EMS) plays an important role for hybrid vehicles, as it is directly related to the power distribution between power sources and further the energy saving of the vehicles. Currently, rule-based EMSs and optimization-based EMSs are faced with the challenge when considering the optimality and the real-time performance of the control at the same time. Along with the rapid development of the artificial intelligence, learning-based EMSs have gained more and more attention recently, which are able to overcome the above challenge. A deep reinforcement learning (DRL)-based EMS is proposed for fuel cell hybrid buses (FCHBs) in this research, in which the fuel cell durability is considered and evaluated based on a fuel cell degradation model. The action space of the DRL algorithm is limited according to the efficiency characteristic of the fuel cell in order to improve the fuel economy and the Prioritized Experience Replay (PER) is adopted for improving the convergence performance of the DRL algorithm. Simulation results of the proposed DRL-based EMS for an FCHB are compared to those of a dynamic programming (DP)-based EMS and a reinforcement learning (RL)-based EMS. Comparison results show that the fuel economy of the proposed DRL-based EMS is improved by an average of 3.63% compared to the RL-based EMS, while the difference to the DP-based EMS is within an average of 5.69%. In addition, the fuel cell degradation rate is decreased by an average of 63.49% using the proposed DRL-based EMS compared to the one without considering the fuel cell durability. Furthermore, the convergence rate of the proposed DRL-based EMS is improved by an average of 30.54% compared to the one without using the PER. Finally, the adaptability of the proposed DRL-based EMS is validated on a new driving cycle, whereas the training of the DRL algorithm is completed on the other three driving cycles.

Keywords Deep reinforcement learning · Energy management strategy · Fuel cell hybrid bus · Fuel cell degradation · Reinforcement learning

1 Introduction

As one of new energy vehicles, fuel cell hybrid vehicles (FCHVs) have gained increased attention worldwide, especially in China, fuel cell hybrid buses (FCHBs) have been developed rapidly in recent years. FCHBs use fuel cell systems (FCSs) and batteries as two power sources, which generates the energy management problem. Thus, an energy management strategy (EMS) is necessary for FCHBs, which determines the power distribution between FCSs and batteries and further influences the fuel economy of FCHBs and other relevant factors such as the performance degradation of FCSs.

Two types of EMSs were developed previously for hybrid vehicles, i.e. rule-based EMSs [1–4] and optimization-based EMSs [5–8]. The former is composed of some if–then control rules which are based on the expert knowledge. The real-time control performance of the former is good owing to the simplicity, however it still leaves rooms for the control optimality. The latter is usually based on different optimal control theories, which guarantees the control optimality but the real-time applications of which are limited mainly due to the dependency on the future driving cycles. Besides, in order to

✉ Suk Won Cha
swcha@snu.ac.kr

1 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen 518055, China

2 University of Chinese Academy of Sciences, 19 (A) Yuquan Road, Shijingshan District, Beijing 100049, China

3 School of Mechanical and Aerospace Engineering, Seoul National University, San 56-1, Daehak-dong, Gwanak-gu, Seoul 151742, South Korea

increase the adaptability to different driving conditions, relevant parameters should be adjusted appropriately in those two types of EMSs. Along with the rapid development of the artificial intelligence, the third type, i.e. learning-based EMSs have been gradually investigated for hybrid vehicles in recent years, learning algorithms adopted in which mainly include the reinforcement learning (RL) algorithm and the deep reinforcement learning (DRL) algorithm. RL-based and DRL-based EMSs are fully data-driven, which reach the optimal control results through interactions between the agent and the environment and the trial-and-error learning. In addition, RL-based and DRL-based EMSs do not rely on any predefined rules or optimal control theories and present good real-time performance and adaptability.

In the earlier research [9], the RL algorithm was adopted to the EMS for a hybrid electric tracked vehicle, and the results showed that the RL-based EMS presents the strong adaptability, optimality, learning ability, and also the effectively reduced computational time. In recent years, the RL algorithm has been widely adopted to different hybrid vehicles, such as engine-motor hybrid vehicles [9–17], FCHVs [18, 19], engine-ultracapacitor hybrid vehicles [20], and electric vehicles using hybrid energy storage systems [21]. In research [22], a parametric study on several key parameters of RL-based EMSs was conducted for hybrid vehicles, such as the state types and number of states, the state and action discretization, the exploration and exploitation, and the learning experience selection. The offline training & online application mode is commonly used for RL-based EMSs for hybrid vehicles. Besides the control optimality, the convergence rate during the offline training and the adaptability during the online application are also important performances of RL-based EMSs. In order to improve those performances of RL-based EMSs, some skills have been developed. Introducing the transition probability matrix (TPM) of the vehicle's required power to the RL algorithm framework is a common way to expedite the convergence of the offline training. Additionally, the learning rate was adjusted during the offline training in some research [12, 13] in order to improve the convergence rate. Furthermore, refining the RL algorithm using other strategies [19] and introducing initialization strategies to the RL algorithm using properly selected penalty functions [16] are also used to speed up the convergence of the offline training. For improving the adaptability of RL-based EMSs, some characteristic factors of the TPM, such as the kullback–leibler divergence rate [10, 13, 20], the induced matrix norm [12], and the cosine similarity [18], were introduced to the RL algorithm, and the control strategy was updated in real-time according to those characteristic factors, respectively.

The difference between the RL and DRL algorithms is on the expression form of the Q-value, which is an important factor for the decision-making for both algorithms, i.e. the Q-table and the Q-network. The RL algorithm is based on the state discretization, which will cause the rapid increase on the Q-table size and consequently the long computational time and the bad convergence ability when dealing with a higher-dimensional state space. On the other hand, the DRL algorithm uses a deep neural network (DNN), i.e. the deep Q-network (DQN), for fitting the Q-table, which is helpful for considering more state variables and also results in a more accurate identification of state variables as any continuous changes in state variables can be reflected in the DNN-based decision-making system. In earlier research [23], a DQN-based EMS was proposed for a power-split hybrid electric bus, and simulation results showed that the fuel economy of the proposed EMS approaches a 5.6% better performance than the RL-based EMS in a trained driving cycle and achieves nearly 90% level of the dynamic programming (DP) in an untrained driving cycle. In some research [24, 25], an extra DNN named the target network was created in order to improve the convergence performance, in which the target network was periodically updated by copying parameters from the original network. In research [26], a dueling network structure- DQN-based EMS was proposed for hybrid vehicles to further speed up the convergence, which is particularly useful in states where the actions do not affect the environment significantly. In addition to above DQN-based EMSs, the deep deterministic policy gradient (DDPG), which belongs to the actor-critic DRL framework, has also been adopted to DRL-based EMSs of hybrid vehicles [27–29].

Although DRL-based EMSs have presented the superiority compared to other types of EMSs, there are still some problems need to be solved to improve the performance. The leaning ability, i.e. the convergence speed of the DRL algorithm is the first key factor. Additionally, the control effect and the adaptability of the EMSs are also important factors. Those factors could be further improved by using different skills. Currently, most of research on RL-based and DRL-based EMSs is focused on traditional hybrid vehicles, i.e. engine-motor hybrid vehicles and rarely on FCHVs. In addition, for FCHVs, the fuel cell stack lifetime is an issue due to the high-cost, thus the fuel cell stack durability should be considered when designing EMSs.

In this research, a DQN-based EMS is proposed for FCHBs, in which the Prioritized Experience Replay (PER) is adopted in order to expedite the convergence of the DRL algorithm. In addition, the action space of the DRL algorithm is limited in the proposed EMS according to the efficiency characteristic of FCSs in order to improve the fuel economy of FCHBs. Furthermore, the fuel cell stack durability is considered in the proposed EMS based on a fuel cell degradation model. Finally, to validate the effectiveness of the proposed EMS, simulation

results of the proposed EMS for an FCHB are compared to those of an RL-based EMS and a DP-based EMS.

The remaining part of this paper is organized as follows: in Sect. 2, the target FCHB model is introduced including the fuel cell degradation model; in Sect. 3, the DRL-based EMS is proposed for the FCHB based on the introduction on the relevant algorithms; in Sect. 4, the effectiveness of the proposed EMS is validated in terms of the fuel economy, the fuel cell durability, the convergence performance, and the adaptability by comparing to an RL-based EMS and a DP-based EMS; at the end, conclusions are drawn from this research in Sect. 5.

2 The FCHB Model

The FCHB powertrain is mainly composed of the FCS, the DC/DC converter, the battery, the motor, and the final drive, as illustrated in Fig. 1. In this research, an FCHB is selected as the target bus, which is shown in the recommendation model lists for the new energy vehicle popularization and application of Ministry of Industry and Information Technology of China [30]. The relevant data of the FCHB are provided in Table 1.

2.1 FCHB Power Demand Model

The vehicle movement is determined by the tractive forces and resistances acting on the vehicle during driving. The power required for the vehicle during driving can be expressed as follows:

$$P_{req} = (fMg\cos\alpha + 0.5\rho_a AC_D v^2 + Mg\sin\alpha + \delta Ma)\cdot v \quad (1)$$

where f is the rolling resistance coefficient, M is the mass of the vehicle, g is the acceleration of gravity, α is the road slope which is set to 0 in this research, ρ_a is the air mass density, A is the vehicle frontal area, C_D is the aerodynamic drag coefficient, v is the vehicle velocity, δ is the mass factor which is set to 1 in this research, and a is the vehicle acceleration. For the FCHB, the power P_{req} is provided by the FCS and the battery, and the specific relationship on the power balance is as follows:

$$P_{req} = \left(P_{fcs}\cdot\eta_{\text{conv}} + P_{bat}\right)\cdot\eta_{mot}\cdot\eta_{final} \quad (2)$$

where P_{fcs} and P_{batt} represent the FCS power and the battery power respectively; η_{conv}, η_{mot}, and η_{final} represent the DC/DC converter efficiency, the motor efficiency, and the final drive efficiency respectively. Detailed values for a part of the parameters above are listed in Table 2, and the rest of parameters will be explained in the following parts.

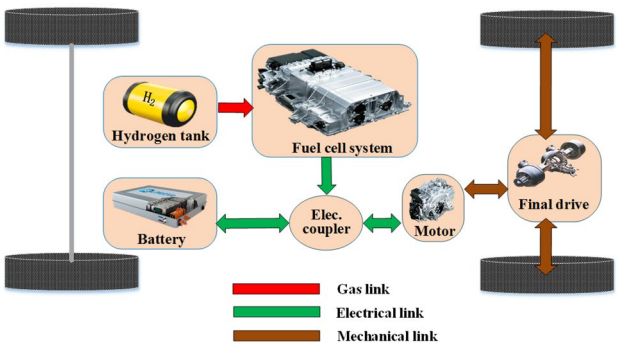

Fig. 1 Powertrain configuration of the FCHB

Table 1 Vehicle parameters of the FCHB

Parameter	Value
Vehicle total mass (kg)	14,000
FCS Max power (kW)	53
Battery power (kW)	170
Battery capacity (Ah)	300
Motor max power (kW)	150

2.2 FCS Model

An FCS consists of a fuel cell stack and some auxiliary components such as the air compressor, the cooler, and the humidifier. A part of power generated from the fuel cell stack is provided to the auxiliary components to keep the regular operation of the FCS. The fuel cell stack is composed of a number of single cells, and there are three types of losses which occur in every single cell, i.e. the activation loss, the ohmic loss, and the concentration loss. The voltage of the single cell v_{fc} can be expressed as follows:

$$v_{fc} = E - v_{act} - v_{ohm} - v_{conc} \quad (3)$$

where E is the open circuit voltage (OCV), v_{act}, v_{ohm}, and v_{conc} represent the activation loss, the ohmic loss, and the concentration loss respectively.

The hydrogen consumption rate of the stack $\dot{m}_{h_2}$ is related to the stack current I_{stack} according to the following equation:

Table 2 Parameter values of the FCHB

Parameter	Value
Rolling resistance coefficient f	0.007
Air mass density (kg/m^3)ρ_a	1.21
Vehicle frontal area (m^2)A	7.5
Aerodynamic drag coefficient C_D	0.6
Tire radius (m)	0.472
Final drive gear efficiency η_{final}	0.95
DC/DC converter efficiency η_{conv}	0.9

$$\dot{m}_{h_2} = \frac{N_{cell} \cdot M_{h_2}}{n \cdot F} \cdot I_{stack} \cdot \lambda \tag{4}$$

where N_{cell} represents the cell number of the stack, M_{h_2} represents the molar mass of the hydrogen, n represents the number of electrons acting in the reaction, F is the Faraday constant, and λ is the hydrogen excess ratio. For the FCS, the efficiency η_{fcs} is defined as follows:

$$\eta_{fcs} = \frac{P_{fcs}}{\dot{m}_{h_2} \cdot LHV} \tag{5}$$

where LHV is the lower heating value of the hydrogen. Further details on the FCS model can be found in our previous research [8, 31–33]. A 53 kW FCS is used in the FCHB, for which the hydrogen consumption rate and the efficiency of the FCS vary according to the FCS power as shown in Fig. 2.

In this research, an empirical fuel cell degradation model [34] is adopted in order to evaluate the effect of EMSs on the fuel cell durability, in which the fuel cell degradation is mainly caused by the load changing, the startup and shutdown, the idling, and the high power load operation conditions [34, 35]. The fuel cell degradation model is expressed as follows:

$$\Delta\phi_{\text{degrad}} = Kp\left(\left(k_1 t_1 + k_2 n_1 + k_3 t_2 + k_4 t_3\right) + \beta\right) \tag{6}$$

where $\Delta\phi_{\text{degrad}}$ represents the voltage decline percentage; t_1, n_1, t_2, and t_3 can be obtained from the driving condition of the FCHB, which represent the duration of the idle time, the start-stop count, the duration of rapid load variations, and the duration of high power loading conditions, respectively; k_1, k_2, k_3, and k_4 are the corresponding coefficients for the above each term, detailed values of which can be found in the research [34]; β is the natural decay rate; Kp is a modifying coefficient for on-road systems considering the durability difference between in the laboratory and on the road. Detailed values of β and Kp are also sourced from the research [34].

2.3 Battery Model

The battery is modeled by an equivalent circuit, which is composed of a voltage source U_{oc} and a resistance R_{int} connected to the voltage source in series, as illustrated in Fig. 3. The voltage source U_{oc}, which is also called the open circuit voltage (OCV), and the internal resistance R_{int} vary according to the battery state of charge (SOC), as shown in Fig. 4.

The battery SOC is derived from the ampere-hour integral method as follows:

$$\dot{SOC} = -\frac{I_{bat}}{q} \tag{7}$$

where q represents the battery capacity. The following relationship can be obtained from Fig. 3.

$$I_{bat} = \frac{U_{oc}(SOC) - \sqrt{U_{oc}(SOC)^2 - 4R_{\text{int}}(SOC) \cdot P_{bat}}}{2R_{\text{int}}(SOC)} \tag{8}$$

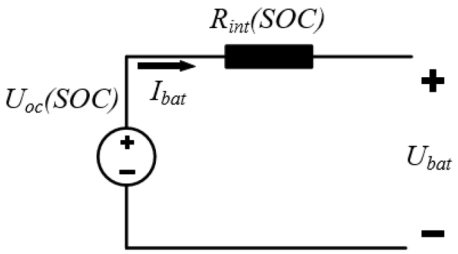

Fig. 3 Equivalent circuit diagram of the battery model

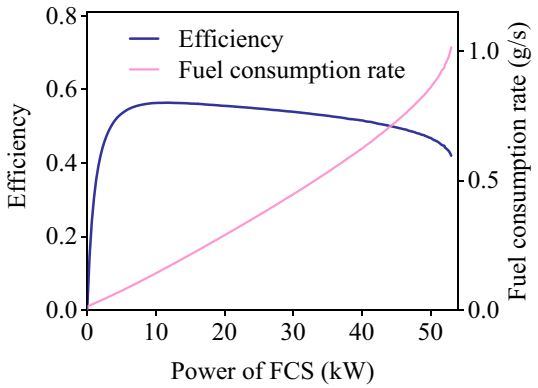

Fig. 2 Fuel consumption rate and efficiency of the FCS

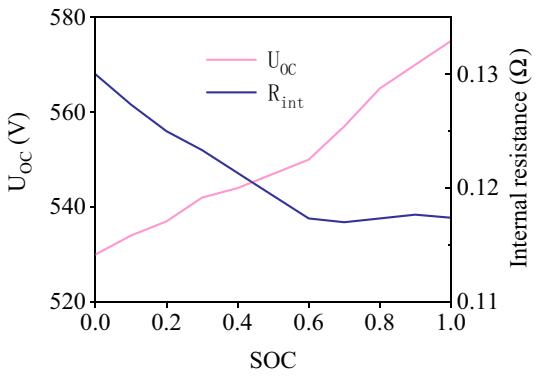

Fig. 4 OCV and internal resistance of the battery

2.4 Motor Model

The motor is modeled by an efficiency map, which indicates the relationship among the motor speed, torque, and efficiency, as illustrated in Fig. 5.

3 The Proposed DRL-based EMS

In this section, relevant algorithms including the RL and DRL algorithms are introduced first, and then the proposed DRL-based EMS is explained.

3.1 RL and DRL Algorithms

The RL algorithm is a main branch of machine learning algorithms, which contains several important factors including the agent, the environment, the state, the action, and the reward. The main concern of the RL algorithm is how the agent takes actions under a given environment in order to maximize the cumulative reward and finally reaches the optimal control results through interactions between the agent and the environment.

The Q-learning is the most commonly used RL algorithm, in which the Q-function that satisfies Bellman's equation is defined as follows:

$$Q(s_t, a_t) = E\left[R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) | s_t, a_t\right] \tag{9}$$

where Q is also called the value function; E represents the expectation of cumulative returns; s, a, and R represent the state, the action, and the reward, respectively; γ is a discount factor for the future value function, which is beneficial for the convergence during the learning process. The updating rule of the Q-learning is as follows:

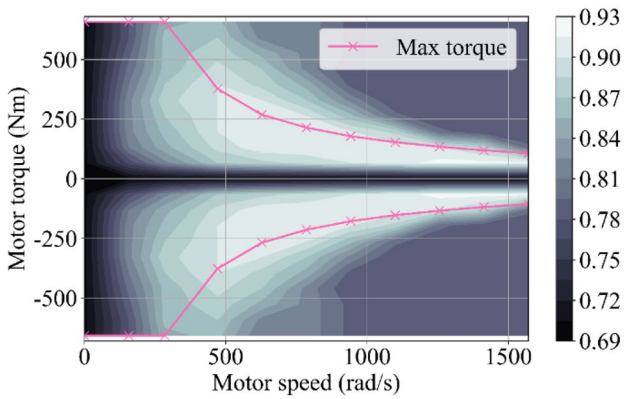

Fig. 5 Motor efficiency map

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta\left[R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right] \tag{10}$$

where η is the learning rate, which influences the convergence performance, i.e. the larger value results in the faster convergence speed but also causes the learning oscillation and overfitting problems. The relationship between the exploration and the exploitation during the learning process is decided by the ε-greedy algorithm, i.e. the agent randomly chooses actions with a small probability $1-\varepsilon$ while selects actions maximizing the Q-function with a probability ε. The optimal control strategy π can be finally acquired as follows after the Q-function is converged through the algorithm iterations.

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{11}$$

In the Q-learning algorithm, the Q-value for each state-action pair is stored in the huge Q-table. This will cause the rapid increase on the Q-table size when dealing with a higher-dimensional state space and consequently make the convergence difficult. The DRL algorithm uses DNNs to fit the Q-function, which is effective when dealing with higher-dimensional systems, as follows:

$$Q(s_t, a_t; \theta) \approx Q(s_t, a_t) \tag{12}$$

where θ represents the network parameter. In order to break the dependency between the target Q-value and the original DNN parameters and speed up and stabilize the convergence, an extra DNN named the target network is usually created with the network parameter of θ^-. The original DNN is called the evaluation network and used to select actions and the target network is periodically updated by copying parameters from the evaluation network. The evaluation network parameter θ is updated by implementing the back-propagation and gradient descent based on the loss function, which is defined as the mean squared error between the target Q-vale and the current Q-value derived from the evaluation network, as follows:

$$L(\theta) = E\left[\left(R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta)\right)^2\right] \tag{13}$$

In order to break correlations among training data, the experience replay is usually adopted during the training, where the experience $e_t = (s_t, a_t, R_t, s_{t+1})$ at each time step is stored in an experience pool $D_N = \{e_1, e_2, \ldots, e_N\}$ and mini-batches of data are sampled from the pool randomly for training. The experience replay is effective for cutting off the relationship among training data through the random sampling, however some important samples can be missed and this will influence the convergence [36]. In this research, the PER is adopted to

replay important samples more frequently, in which the absolute temporal difference (TD) error of each data sample is selected to assess its importance, i.e. the priority. The sampling probability of each sample is proportional to the sample priority, as follows [36]:

$$\begin{aligned} TD(s_t, a_t) &= Q(s_t, a_t;\theta) - \left(R_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1};\theta^-)\right) \\ p_i &= |TD_i| + \in \\ P_i &= \frac{p_i}{\sum_k p_k} \end{aligned} \tag{14}$$

where p_i and P_i are the priority and the sampling probability of the ith sample, respectively, $\in$ is a little positive number for avoiding the zero sampling probability.

3.2 The Proposed DRL-Based EMS

For the DRL-based EMS of FCHB, the agent is the EMS while the environment includes the FCHB status and the driving condition, as shown in Fig. 6. Important factors of the DRL algorithm, including the state, the action, the reward function, and the DNNs, should be set and designed first according to the control problem characteristics of the FCHB. Owing to the powerful fitting ability of DNNs, the discretization on the state variables is not necessary and considering more state variables is possible compared to the case of the RL algorithm.

In this research, the FCHB velocity, the acceleration, and the battery SOC are selected as the state variables, as follows:

$$S = \{v, a, SOC\} \tag{15}$$

The FCS power is set as the action variable, which is limited in an effective range according to the efficiency characteristic shown in Fig. 2, as follows:

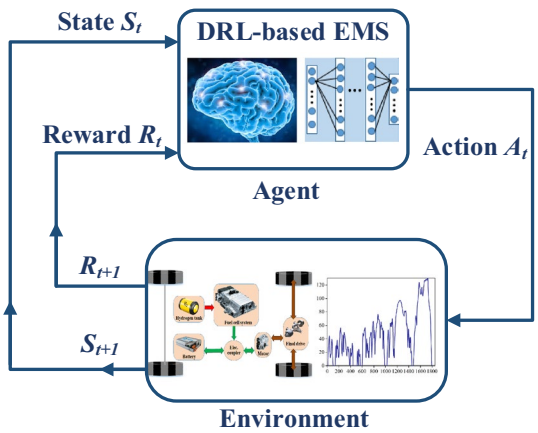

Fig. 6 Learning framework of DRL-based EMS for FCHB

$$A = \{2, 3, 4, 5, \ldots, 36, 37, 38, 39, 40\} \ \text{kw} \tag{16}$$

The reward function is significant for the performance of the DRL-based EMS as it directly influences the control effect and the convergence rate. Considering the control objective of improving the fuel economy and the fuel cell stack durability and the fuel cell degradation model introduced in 2.2, the reward function R is designed as follows:

$$R = -\left(\alpha \dot{m}_{h_2} + \mu(SOC_{ref} - SOC)^2 + \varphi\left|\Delta P_{fcs}\right| + \xi f(t_{life})\right) \tag{17}$$

where the first term is related to the fuel economy, the second term is related to the battery SOC sustaining, while the rest of terms are related to the fuel cell durability. α, μ, φ, ξ are weighting factors for each term, SOC_{ref} is the reference SOC value for sustaining which is set to 0.7 in this research, and f represents the sigmoid function, as follows:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{18}$$

ΔP_{fcs} and t_{life} are defined as follows:

$$\begin{aligned} \Delta P_{fcs}(t) &= P_{fcs}(t) - P_{fcs}(t-1) \\ t_{life} &= t_1 + n_1 + t_3 \end{aligned} \tag{19}$$

where t_1, n_1, and t_3 correspond to those in the fuel cell degradation model in (6), which are related to the idling, the startup and shutdown, and the high power load conditions respectively. Those three factors are considered together owing to the fact that the calculation time-step for the algorithm is one second in this research. According to the reward function (17) and the mechanism of the DRL algorithm, in order to maximize the cumulative reward, the agent will tend to minimize the fuel consumption, maintain the battery SOC to the reference, and minimize harmful operation conditions of the FCS.

The evaluation and target DNNs are designed with the same structure, where the input and the output of the network are related to the state variables and the action variable. Considering the network structure presented in research [37], there are three hidden layers except the input and output layers, where there are 200, 100, and 50 neurons in each layer. The ReLU function [38] is used as the activation function for each layer, which is defined as follows.

$$f(x) = \max(0, x) \tag{20}$$

The pseudocode of the DRL algorithm for the proposed EMS is presented in Table 3, where the first loop circulates different training episodes, i.e. driving cycles, which is processed once for every driving cycle, the second loop is processed once for every time-step within one driving cycle, and the third loop is processed nth at every

Table 3 Pseudocode of the DRL algorithm for the proposed EMS

DRL algorithm with PER

Parameters: M: number of episodes; T: length of one episode; n: size of mini-batch; C: target network updating period

Initialize experience pool D with capacity N;
Initialize evaluation and target networks with random weights $\theta = \theta^{-}$

1. for episode = 1: M do
2. Reset environment s_0
3. for t = 1:T do
4. With probability ε select a random action a_t
otherwise select $a_t = \underset{a \in A}{\text{argmax}}\, Q(s_t, a; \theta)$
5. Execute action a_t, observe reward R_t, update to next state s_{t+1}
6. Store (s_t, a_t, R_t, s_{t+1}) in D
7. Sample mini-batch of $\sum_{i=1}^{n} (s_t, a_t, R_t, s_{t+1})$ from D using PER
8. for i = 1:n do
9. if $t = T$, $y_i = R_i$
else $y_i = R_i + \gamma \underset{a_{i+1} \in A}{\max} Q(s_{i+1}, a_{i+1}; \theta^{-})$
10. Implement a gradient descent step based on loss function $L(\theta)$ and update evaluation network
11. Update target network by $\theta^{-} = \theta$ every C steps
12. end for
13. end for
14. end for

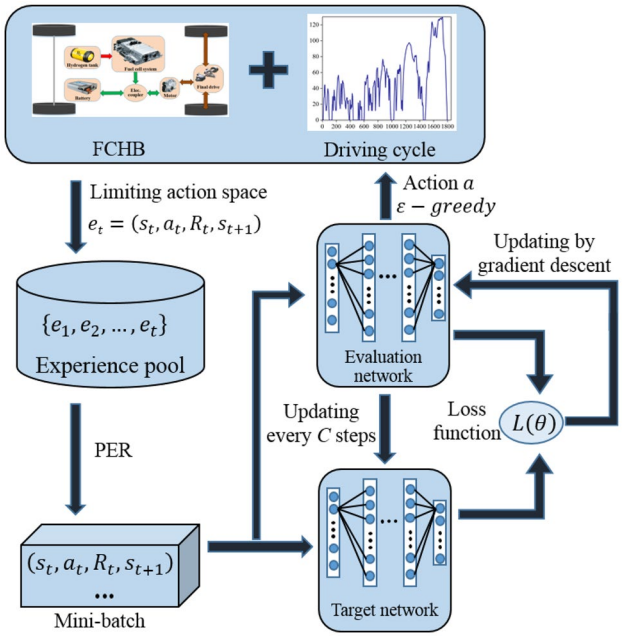

Fig. 7 Framework of the proposed DRL-based EMS

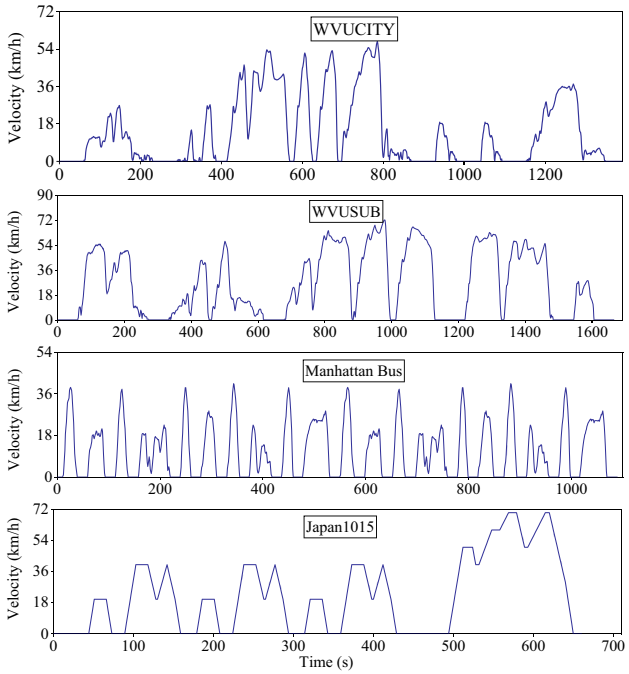

Fig. 8 Driving cycles used in this research

time-step, which is the size of the mini-batch from the experience pool. The framework of the proposed DRL-based EMS is illustrated in Fig. 7.

4 Simulation Results

Four driving cycles are utilized in this research as shown in Fig. 8, where the West Virginia University city cycle (WVUCITY) [39], the West Virginia University suburban

cycle (WVUSUB) [39], and the Manhattan Bus driving cycle are used in the algorithm training while the Japan 1015 driving cycle is used for the validation of the proposed DRL-based EMS. The effectiveness of the proposed DRL-based EMS is validated in terms of the fuel economy, the fuel cell durability, the adaptability, and the convergence performance, respectively. Specific values for the algorithm parameters are listed in Table 4.

4.1 Fuel Economy

The hydrogen consumption of the proposed DRL-based EMS is compared to that of an RL-based EMS and a DP-based EMS for the FCHB, where the DP is a global optimization method, the result of which is usually regarded as the benchmark for the evaluation of other control methods and details of which can be found in our previous research [40]. In order to focus on the fuel economy, only the first two terms in the reward function (17) are considered here. Figures 9, 10, and 11 show the comparison results of the FCS power and the battery power for the above three EMSs on the three training driving cycles respectively. Table 5 summarizes the hydrogen consumption comparison, where the differences on the final battery SOC are considered by the equivalent hydrogen consumption. The comparison results indicate that the fuel economy of the proposed DRL-based EMS is improved by 2.93%, 4.25%, and 3.72% compared to the RL-based EMS on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles respectively, while the difference to the DP-based EMS is within 5.53%, 5.67%, and 5.86% on the three driving cycles respectively.

4.2 Fuel Cell Durability

The voltage decline percentage $\Delta\phi_{\text{degrad}}$ in (6) is used to evaluate the fuel cell degradation rate in this research, which is obtained for the cases where only the first two terms in the reward function (17) and the whole reward function are considered respectively. Here, the former case corresponds

Table 4 Parameter values of the DRL algorithm

Parameter	Value
Episode circulation number M	1100
Experience pool capacity N	10,000
Mini-batch size n	64
Learning rate η	0.001
Discount factor γ	0.9
Weighting factor α	200
Weighting factor μ	1
Weighting factor φ	0.000035
Weighting factor ξ	1.5

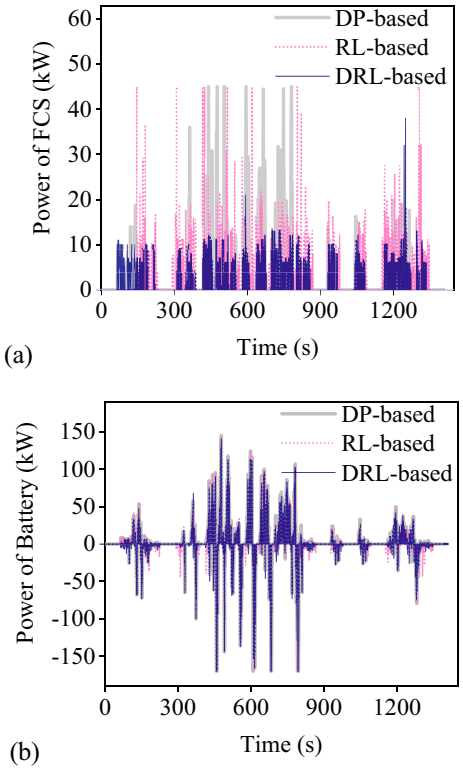

Fig. 9 Comparison on the WVUCITY: **a** FCS power; **b** battery power

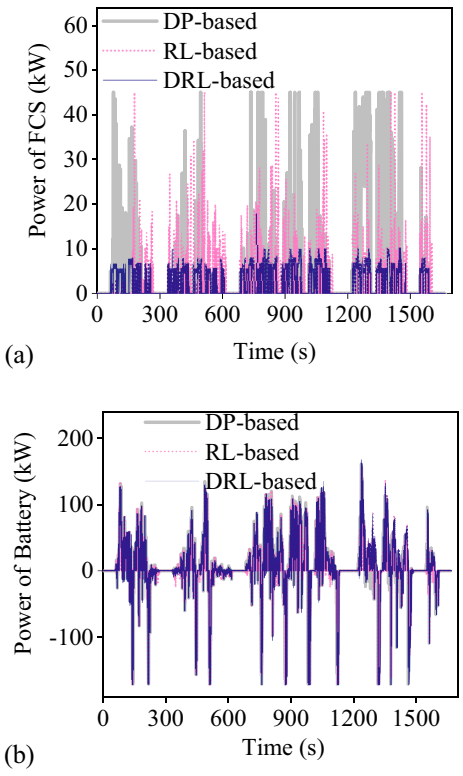

Fig. 10 Comparison on the WVUSUB: **a** FCS power; **b** battery power

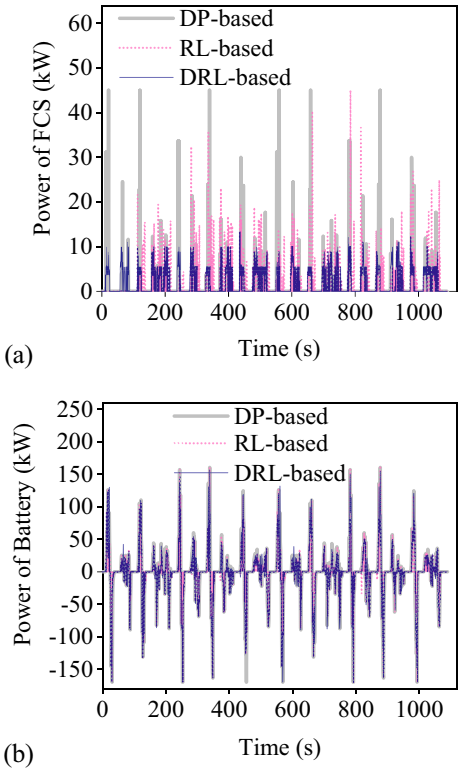

Fig. 11 Comparison on the Manhattan Bus: **a** FCS power; **b** battery power

to the DRL-based EMS in which the fuel cell durability is not considered. The results are provided and compared in Table 6, which show that the fuel cell degradation rate is decreased by 56.96%, 69.47%, and 64.03% using the proposed DRL-based EMS compared to the one without considering the fuel cell durability on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles respectively, while the fuel economy is almost not influenced.

4.3 Convergence Performance

In this research, the PER is adopted to replay important samples more frequently from the experience pool and expedite the convergence during training. In order to validate the effectiveness of the PER, the tendency of the average reward during training is compared for the cases of with the PER and without the PER on three driving cycles, as illustrated in Fig. 12. It can be observed that the DRL algorithm with the PER reaches the convergence with around 375, 365, and 466 rounds while the one without the PER starts to converge with around 420, 850, and 612 rounds on the three driving cycles respectively, i.e. the convergence performance of the proposed DRL-based EMS is improved by 10.71%, 57.06%, and 23.86% owing to the utilization of the PER on the three driving cycles respectively.

4.4 Adaptability

In order to validate the adaptability of the proposed DRL-based EMS to different driving cycles, it is applied to a new driving cycle after training, i.e. the Japan 1015 driving cycle. The simulation result of the fuel consumption on the Japan 1015 driving cycle is presented in Table 7 and compared to that of other two different EMSs, which reveals that the fuel economy of the proposed DRL-based EMS is improved by 4.18% compared to the RL-based EMS whereas the difference to the DP-based EMS is within 5.65%. Compared to Table 5, it is enough to prove that the proposed DRL-based EMS presents a good adaptability. Figure 13 illustrates comparison results of the FCS power and the battery power for the DP, RL, and DRL-based EMSs on the Japan 1015 driving cycle, where the driving cycle is repeated twice in order to observe more obvious results.

Table 5 Fuel consumption comparison results

Driving cycle	EMS	Final battery SOC	Equivalent hydrogen consumption (kg/100 km)	Deviation from DP (%)
WVUCITY	DP	0.6914	4.9548	–
	RL	0.6953	5.3742	8.46%
	DRL	0.6957	5.2288	5.53%
WVUSUB	DP	0.6795	5.1526	–
	RL	0.6810	5.6636	9.92%
	DRL	0.6786	5.4446	5.67%
Manhattan Bus	DP	0.6960	4.1464	–
	RL	0.6980	4.5434	9.58%
	DRL	0.6991	4.3894	5.86%

Table 6 Fuel cell durability comparison results

Driving cycle	EMS	Equivalent hydrogen consumption (kg/100 km)	Fuel cell degradation rate (%)	Decreasing percentage (%)
WVUCITY	Without considering fuel cell durability	5.2182	0.0869	–
	Proposed DRL-based EMS	5.2288	0.0374	56.96
WVUSUB	Without considering fuel cell durability	5.4394	0.0701	–
	Proposed DRL-based EMS	5.4446	0.0214	69.47
Manhattan Bus	Without considering fuel cell durability	4.3157	0.0734	–
	Proposed DRL-based EMS	4.2694	0.0264	64.03

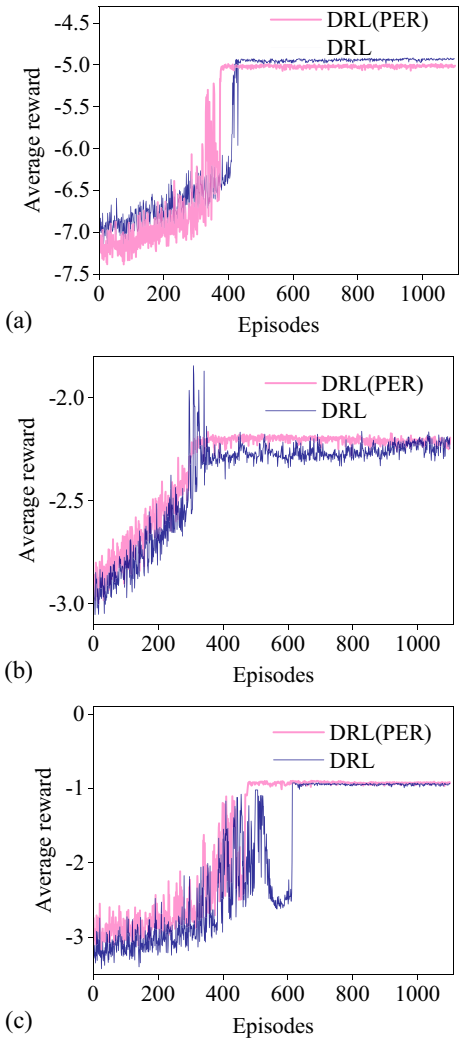

Fig. 12 Tendency of the average reward during training: **a** on the WVUCITY driving cycle; **b** on the WVUSUB driving cycle; **c** on the Manhattan Bus driving cycle

Table 7 Fuel consumption result on the Japan 1015 driving cycle

EMS	Final battery SOC	Equivalent hydrogen consumption (kg/100 km)	Deviation from DP (%)
DP	0.6826	5.3232	–
RL	0.6864	5.8466	9.83%
DRL	0.6853	5.6238	5.65%

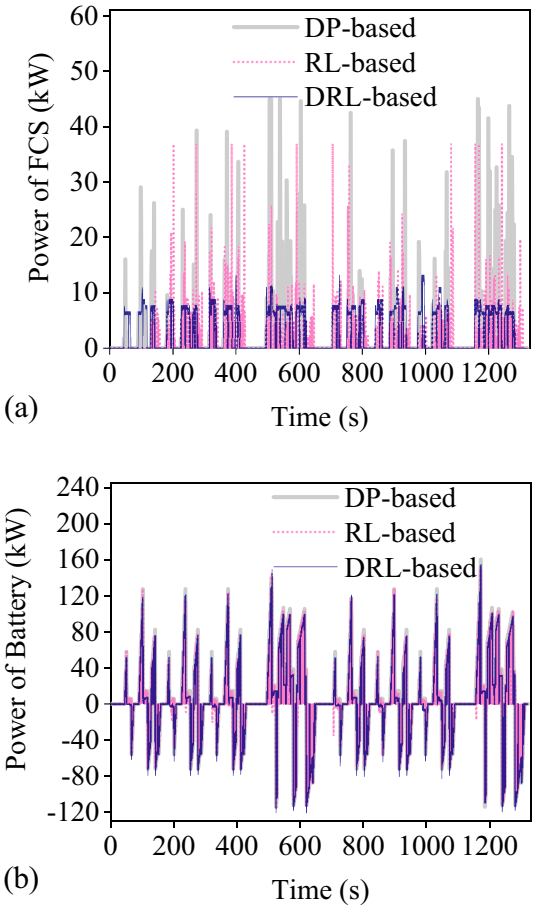

Fig. 13 Simulation results on the Japan 1015 driving cycle: **a** FCS power; **b** battery power

5 Conclusion

Considering the rapid development of FCHBs in China currently, a DRL-based EMS is proposed for FCHBs in this research, in which the fuel cell durability is considered based on a fuel cell degradation model. The PER is adopted for improving the convergence performance of the DRL algorithm and the action space of the DRL algorithm is limited for the better control effect. The effectiveness of the proposed DRL-based EMS for an FCHB is validated in terms of the fuel economy, the fuel cell durability, the convergence performance, and the adaptability by comparing the results of it to those of an RL-based and a DP-based EMSs. The following conclusions can be drawn from this research:

(1) The fuel economy of the proposed DRL-based EMS is improved by 2.93%, 4.25%, and 3.72% compared to the RL-based EMS on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles respectively, while the difference to the DP-based EMS is within 5.53%, 5.67%, and 5.86% on the three driving cycles respectively.
(2) The fuel cell degradation rate is decreased by 56.96%, 69.47%, and 64.03% using the proposed DRL-based EMS compared to the one without considering the fuel cell durability on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles respectively.
(3) The convergence performance of the proposed DRL-based EMS is improved by 10.71%, 57.06%, and 23.86% owing to the utilization of the PER on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles respectively.
(4) The adaptability of the proposed DRL-based EMS is validated on the Japan 1015 driving cycle, whereas the training of the DRL algorithm is completed on the WVUCITY, WVUSUB, and Manhattan Bus driving cycles, and the result proves that the proposed DRL-based EMS presents a good adaptability.

Acknowledgements This research was supported by Shenzhen Science and Technology Innovation Commission (Grant no. KQJSCX20180330170047681, JCYJ20210324115800002, JCYJ20180507182628567), Department of Science and Technology of Guangdong Province (Grant no. 2021A0505030056, 2021A0505050005), National Natural Science Foundation of China (Grant no. 62073311), CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Pengcheng Program, and Shenzhen Key Laboratory of Electric Vehicle Powertrain Platform and Safety Technology.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Gao, D. W., Jin, Z. H., & Lu, Q. C. (2008). Energy management strategy based on fuzzy logic for a fuel cell hybrid bus. *Journal of Power Sources, 185*(1), 311–317.
2. Zhang, Q., Deng, W., Zhang, S., & Wu, J. (2016). A rule based energy management system of experimental battery/supercapacitor hybrid energy storage system for electric vehicles. *Journal of Control Science and Engineering, 2016*, 1–17.
3. Yan, M., Li, M., He, H., Peng, J., & Sun, C. (2018). Rule-based energy management for dual-source electric buses extracted by wavelet transform. *Journal of Cleaner Production, 189*, 116–127.
4. Lee, H. S., Kim, J. S., Park, Y. I., & Cha, S. W. (2016). Rule-based power distribution in the power train of a parallel hybrid tractor for fuel savings. *International Journal of Precision Engineering and Manufacturing-Green Technology, 3*(3), 231–237.
5. Lin, C. C., Peng, H., Grizzle, J. W., & Kang, J. M. (2003). Power management strategy for a parallel hybrid electric truck. *IEEE Transactions on Control Systems Technology, 11*(6), 839–849.
6. Kim, N. W., Cha, S. W., & Peng, H. (2011). Optimal control of hybrid electric vehicles based on Pontryagin's minimum principle. *IEEE Transactions on Control Systems Technology, 19*(5), 1279–1287.
7. Hou, C., Ouyang, M. G., Xu, L. F., & Wang, H. W. (2014). Approximate Pontryagin's minimum principle applied to the energy management of plug-in hybrid electric vehicles. *Applied Energy, 115*, 174–189.
8. Zheng, C. H., Kim, N. W., & Cha, S. W. (2012). Optimal control in the power management of fuel cell hybrid vehicles. *International Journal of Hydrogen Energy, 37*(1), 655–663.
9. Liu, T., Zou, Y., Liu, D., & Sun, F. (2015). Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle. *IEEE Transactions on Industrial Electronics, 62*(12), 7837–7846.
10. Zou, Y., Liu, T., Liu, D., & Sun, F. (2016). Reinforcement learning-based real-time energy management for a hybrid tracked vehicle. *Applied Energy, 171*, 372–382.
11. Liu, T., & Hu, X. (2018). A bi-level control for energy efficiency improvement of a hybrid tracked vehicle. *IEEE Transactions on Industrial Informatics, 14*(4), 1616–1625.
12. Liu, T., Wang, B., & Yang, C. L. (2018). Online Markov chain-based energy management for a hybrid tracked vehicle with speedy Q-learning. *Energy, 160*, 544–555.
13. Du, G., Zou, Y., Zhang, X., Kong, Z., Wu, J., & He, D. (2019). Intelligent energy management for hybrid electric tracked vehicles using online reinforcement learning. *Applied Energy, 251*, 113388.
14. Liu, T., Hu, X., Hu, W., & Zou, Y. (2019). A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles. *IEEE Transactions on Industrial Informatics, 15*(12), 6436–6445.
15. Zhou, Q., Li, J., Shuai, B., Williams, H., He, Y., Li, Z., & Yan, F. (2019). Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle. *Applied Energy, 255*, 113755.
16. Liu, C., & Murphey, Y. L. (2019). Optimal power management based on Q-learning and neuro-dynamic programming for plug-in hybrid electric vehicles. *IEEE Transactions On Neural Networks and Learning Systems, 31*(6), 1942–1954.
17. Zhang, Q., Wu, K., & Shi, Y. (2020). Route planning and power management for PHEVs with reinforcement learning. *IEEE Transactions on Vehicular Technology, 69*(5), 4751–4762.
18. Lin, X., Zhou, B., & Xia, Y. (2020). Online recursive power management strategy based on the reinforcement learning

algorithm with cosine similarity and a forgetting factor. *IEEE Transactions on Industrial Electronics, 68*(6), 5013–5023.
19. Sun, H., Fu, Z., Tao, F., Zhu, L., & Si, P. (2020). Data-driven reinforcement-learning-based hierarchical energy management strategy for fuel cell/battery/ultracapacitor hybrid electric vehicles. *Journal of Power Sources, 455*, 227964.
20. Zhang, W., Wang, J., Liu, Y., Gao, G., Liang, S., & Ma, H. (2020). Reinforcement learning-based intelligent energy management architecture for hybrid construction machinery. *Applied Energy, 275*, 115401.
21. Xiong, R., Cao, J., & Yu, Q. (2018). Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. *Applied Energy, 211*, 538–548.
22. Bin, X., Dhruvang, R., Darui, Z., Adamu, Y., Xueyu, Z., Xiaoya, L., & Zoran, F. (2020). Parametric study on reinforcement learning optimized energy management strategy for a hybrid electric vehicle. *Applied Energy, 259*, 114200.
23. Wu, J., He, H., Peng, J., Li, Y., & Li, Z. (2018). Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus. *Applied Energy, 222*, 799–811.
24. Du, G., Zou, Y., Zhang, X., Liu, T., Wu, J., & He, D. (2020). Deep reinforcement learning based energy management for a hybrid electric vehicle. *Energy, 201*, 117591.
25. Hu, Y., Li, W., Xu, K., Zahid, T., Qin, F., & Li, C. (2018). Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning. *Applied Sciences, 8*(2), 187.
26. Qi, X., Luo, Y., Guoyuan, Wu., Boriboonsomsin, K., & Barth, M. (2019). Deep reinforcement learning enabled self-learning control for energy efficient driving. *Transportation Research Part C Emerging Technologies, 99*, 67–81.
27. Tan, H., Zhang, H., Peng, J., Jiang, Z., & Wu, Y. (2019). Energy management of hybrid electric bus based on deep reinforcement learning in continuous state and action space. *Energy Conversion and Management, 195*, 548–560.
28. Li, Y., He, H., Peng, J., & Wang, H. (2019). Deep reinforcement learning-based energy management for a series hybrid electric vehicle enabled by history cumulative trip information. *IEEE Transactions on Vehicular Technology, 68*(8), 7416–7430.
29. Liu, J., Chen, Y., Zhan, J., & Shang, F. (2019). Heuristic dynamic programming based online energy management strategy for plug-in hybrid electric vehicles. *IEEE Transactions on Vehicular Technology, 68*(5), 4479–4493.
30. http://www.miit-eidc.org.cn/module/download/downfile.jsp?classid=0&filename=b68c55c7356349629d0058c32e5f3474.pdf. Accessed 19 Dec 2021.
31. Zheng, C. H., Oh, C. E., Park, Y. I., & Cha, S. W. (2012). Fuel economy evaluation of fuel cell hybrid vehicles based on equivalent fuel consumption. *International Journal of Hydrogen Energy, 37*(2), 1790–1796.
32. Zheng, C. H., Xu, G. Q., Park, Y. I., Lim, W. S., & Cha, S. W. (2014). Prolonging fuel cell stack lifetime based on Pontryagin's Minimum Principle in fuel cell hybrid vehicles and its economic influence evaluation. *Journal of Power Sources, 248*, 533–544.
33. Zheng, C. H., & Cha, S. W. (2017). Real-time application of Pontryagin's Minimum Principle to fuel cell hybrid buses based on driving characteristics of buses. *International Journal of Precision Engineering and Manufacturing-Green Technology, 4*(2), 199–209.
34. Hu, Z., Li, J., Xu, L., Song, Z., Fang, C., Ouyang, M., & Kou, G. (2016). Multi-objective energy management optimization and parameter sizing for proton exchange membrane hybrid fuel cell vehicles. *Energy Conversion and Management, 129*, 108–121.
35. Pei, P., Chang, Q., & Tang, T. (2008). A quick evaluating method for automotive fuel cell lifetime. *International Journal of Hydrogen Energy, 33*(14), 3829–3836.
36. Li, Y., He, H., Peng, J., & Wu, J. (2018). Energy management strategy for a series hybrid electric vehicle using improved deep Q-network learning algorithm with prioritized replay. DEStech Trans Environ Energy Earth Sci. https://doi.org/10.12783/dteees/iceee2018/27794
37. Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research, 10*(1), 1–40.
38. Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315–323). JMLR Workshop and Conference Proceedings.
39. Kulikov, I., Kozlov, A., Terenchenko, A., & Karpukhin, K. (2020). Comparative study of powertrain hybridization for heavy-duty vehicles equipped with diesel and gas engines. *Energies, 13*(8), 2072.
40. Zheng, C. H., Xu, G. Q., Park, Y. I., Lim, W. S., & Cha, S. W. (2014). Comparison of PMP and DP in fuel cell hybrid vehicles. *International Journal of Automotive Technology, 15*(1), 117–123.

Chunhua Zheng received her Ph.D. from Seoul National University in 2012. She is currently working as an associate professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. Her research field includes energy management strategies of new energy vehicles and fuel cells.

Wei Li received his M. S. from University of Chinese Academy of Sciences in 2021. His is currently working in Huawei.

Weimin Li received his Ph.D. from Shanghai Jiaotong University in 2008. He is currently working as a professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research field includes new energy vehicles and robotics.

Kun Xu received his Ph.D. from University of Chinese Academy of Sciences in 2015. He is currently working as an associate professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research field includes intelligent control of new energy vehicles and autonomous vehicles.

Lei Peng received his Ph.D. from University of Electronic Science and Technology of China in 2009. He is currently working as an associate professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research field includes smart transportation and internet of things.

Suk Won Cha received his Ph.D. from Stanford University in 2004. He is currently working as a professor in the Department of Mechanical Engineering, Seoul National University, South Korea. His research field includes fuel cells and new energy vehicles.