International Journal of Precision Engineering and Manufacturing-Green Technology (2022) 9:859–872
https://doi.org/10.1007/s40684-021-00379-8

Online ISSN 2198-0810
Print ISSN 2288-6206

REGULAR PAPER

A Speedy Reinforcement Learning-Based Energy Management Strategy for Fuel Cell Hybrid Vehicles Considering Fuel Cell System Lifetime

Wei Li[1,2] · Jiaye Ye[1] · Yunduan Cui[1] · Namwook Kim[3] · Suk Won Cha[4] · Chunhua Zheng[1]

Received: 16 April 2021 / Revised: 15 June 2021 / Accepted: 9 July 2021 / Published online: 29 July 2021

Abstract

A speedy reinforcement learning (RL)-based energy management strategy (EMS) is proposed for fuel cell hybrid vehicles (FCHVs) in this research, which approaches near-optimal results with a fast convergence rate based on a pre-initialization framework and meanwhile possesses the ability to extend the fuel cell system (FCS) lifetime. In the pre-initialization framework, well-designed power distribution-related rules are used to pre-initialize the Q-table of the RL algorithm to expedite its optimization process. Driving cycles are modeled as Markov processes and the FCS power difference between adjacent moments is used to evaluate the impact on the FCS lifetime in this research. The proposed RL-based EMS is trained on three driving cycles and validated on another driving cycle. Simulation results demonstrate that the average fuel consumption difference between the proposed EMS and the EMS based on dynamic programming is 5.59% on the training driving cycles and the validation driving cycle. Additionally, the power fluctuation on the FCS is reduced by at least 13% using the proposed EMS compared to the conventional RL-based EMS which does not consider the FCS lifetime. This is significantly beneficial for improving the FCS lifetime. Furthermore, compared to the conventional RL-based EMS, the convergence speed of the proposed EMS is increased by 69% with the pre-initialization framework, which presents the potential for real-time applications.

Keywords Energy management strategy · Fuel cell hybrid vehicle · Lifetime enhancement · Pre-initialization · Speedy reinforcement learning

1 Introduction

Increasingly serious energy shortage and environmental problems have triggered the revolution and innovation in the automotive industry. Fuel cell hybrid vehicles (FCHVs) are known as one of the ultimate solutions for transportation systems because they possess the renewable fuel, i.e. the hydrogen, and also achieve zero emissions compared with engine-motor hybrid vehicles [1]. To ensure all hybrid powertrain components work cooperatively and the fuel economy can be improved, various energy management strategies (EMSs) have been developed, which can be generally categorized into rule-based EMSs and optimization-based EMSs [2]. The former strategies can be realized easily, but the optimality of the control results cannot be guaranteed [3, 4]. The latter strategies are related to optimal control theories such as dynamic programming (DP) [5–7], Pontryagin's minimum principle (PMP) [8–11], and model predictive control (MPC) [12–14], most of which are only effective for predefined driving cycles (DCs), thus it is impractical to apply them online to unknown DCs.

Along with the rapid development of the artificial intelligence, learning-based EMSs such as neural network (NN)-based and reinforcement learning (RL)-based EMSs

✉ Chunhua Zheng
ch.zheng@siat.ac.cn

[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] Department of Mechanical Engineering, Hanyang University, 55 Hanyangdeahak-ro, Sangnok-gu, Ansan, Gyeonggi-do 15588, South Korea

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

have gained more and more attention recently. NNs have been applied to EMSs for the DC classifications [15, 16], the speed or route predictions [17, 18], and the parameters optimization [19, 20]. Apparently, intelligent algorithms are used as tools for the data processing in NN-based EMSs, the main disadvantage of which is the lack of the adaptability to different DCs. The RL algorithm is based on the data-driven and self-learning, which learns from historical data to reach the optimal results, thus RL-based EMSs can achieve better overall performance compared to other traditional EMSs in terms of the optimality and adaptability. An RL-based adaptive EMS for a hybrid electric tracked vehicle was proposed in [21], and the results showed that the RL-based EMS presents the strong adaptability, optimality, learning ability, and also the effectively reduced computational time compared to the stochastic DP (SDP).

The offline training and online application mode is usually adopted for the RL algorithm for hybrid vehicle EMS applications [22, 23]. In order to improve the RL convergence rate during the offline training, some researchers have improved the RL algorithm itself, for example a speedy Q-learning (SQL) algorithm [24] and a fast Q-learning algorithm [25] were developed to improve the convergence rate in the policy generation through adjusting the learning rate. In some research, new RL calculation frameworks have been developed to make the RL algorithm converges fast during training, for example the Dyna-H algorithm [26] and Dyna-Q learning algorithm [27] were proposed by combining the direct learning and indirect learning with a planning Dyna architecture while the multi-step learning algorithm [28] was also developed with different multi-step learning strategies. Refining the RL algorithm using different strategies to achieve the low computational cost is also a good idea, for example Equivalent Consumption Minimization Strategy was used to improve the RL algorithm to expedite the convergence in research [29] and an initialization strategy was introduced by combining the optimal learning with a properly selected penalty function in research [30]. For the online application, some researchers have proposed new methods in order to increase the adaptability of RL-based EMSs, in which the control strategy was updated in real-time according to the characteristic factors of the state transition probability matrix (TPM), such as the Kullback–Leibler (KL) divergence rate [22, 25, 27], the induced matrix norm (IMN) [24], and the cosine similarity [31].

In addition, in order to solve the curse of dimensionality problem which may be occurred in RL-based EMSs, some research proposed deep RL (DRL)-based EMSs by using deep NNs to fit the high-dimensional state-action spaces of the control systems. This method may present good control performances, however because of the large number of NN parameters, the parameter adjustment is challenging and this may lower the control stability accordingly.

Current research on RL-based EMSs commonly targets at traditional engine-motor hybrid vehicles and rarely considers FCHVs. In addition, for an FCHV, improving the fuel cell system (FCS) lifetime should also be one of the requirements for the energy management besides the energy saving. In this research, a speedy RL-based FCS lifetime enhancement (LE) EMS is proposed for FCHVs aiming at real-time applications, in which the convergence speed of the RL algorithm during offline training and the adaptability of the EMS during the online application are both considered. A pre-initialization framework is introduced to speed up the convergence of the RL algorithm, in which power distribution-related rules are designed and used to pre-initialize the Q-table of the RL algorithm. The FCS power difference between adjacent moments is used to evaluate the impact on the FCS lifetime. In order to validate the adaptability of the proposed strategy, it is trained offline under the Urban Dynamometer Driving Schedule (UDDS), Worldwide Harmonized Light Vehicles Test Cycle (WLTC), and New European Driving Cycle (NEDC) and then applied online to the Japan1015 cycle. Simulation results show that the proposed RL-based EMS presents good control performance on the fuel economy and the FCS lifetime prolonging effect and also the feasibility of the real-time application with the fast convergence speed and good adaptability.

The main contributions of this research are as follows: (1) there have been some different approaches for improving the convergence performance of the RL algorithm, and as a novel approach in the hybrid vehicle EMS area, power distribution-related rules are designed and used to pre-initialize the Q-table of the RL algorithm in order to expedite the offline training; (2) the RL-based EMS is proposed targeting at FCHVs in this research while FCHVs are rarely considered in the previous RL-based EMS research, and the FCS lifetime enhancement is also considered in the proposed RL-based EMS.

The remainder of this paper is organized as follows: the FCHV model including the FCS relative lifetime model is described in Sect. 2; the energy management problem of the FCHV is formulated and the proposed RL-based EMS is described in detail in Sect. 3; simulation results of the proposed EMS and discussions are provided in Sect. 4; finally, the conclusions are summarized in Sect. 5.

2 FCHV Model

In order to ensure high fidelity of the simulation model, the FCHV model and data from Autonomie [32] are used in this research, which is a commercial software for the vehicle performance analysis developed by the Argonne National Laboratory in the United States. The powertrain configuration of the FCHV is illustrated in Fig. 1, which is mainly composed

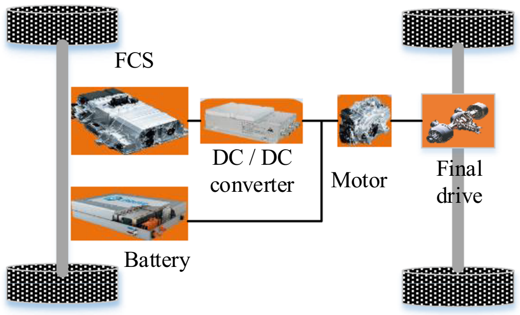


Fig. 1 Powertrain configuration of the FCHV

of the FCS, the DC/DC converter, the battery pack, and the motor. The vehicle data are provided in Table 1.

2.1 Power demand model

On a given DC, the power required for the vehicle P_{req} can be calculated by [13]:

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

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. The power balance relationship of the FCHV is as follows:

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

Table 1 The FCHV data

Items	Parameters	Value
Vehicle	Curb weight (kg)	1763
	Rolling resistance coefficient	0.013
	Tire radius (m)	0.301
	Aerodynamic drag coefficient	0.24
	Air mass density (kg/m^3)	1.2
	Vehicle frontal area (m^2)	2.5
FCS	Total cell number	400
	Max power (kW)	70
	Maximum efficiency	0.59
Battery	Maximum power (kW)	39
	Capacity (Ah)	31
Motor	Peak power (kW)	107
	Maximum efficiency	0.9

where P_{fcs} and P_{batt} represent the FCS power and the battery power respectively; η_{conv} and η_{mot} represent the DC/DC converter efficiency and the motor efficiency, respectively. When P_{req} is known, different EMSs will result in different power allocation between the FCS and the battery.

2.2 FCS model

An FCS is used as the main power source of the FCHV, which converts the chemical energy of the hydrogen and oxygen into the electrical energy by the electrochemical reaction. A physical and empirical FCS model is used in this research by considering the physical laws and the operating conditions. The hydrogen consumption rate of the fuel cell stack can be calculated based on the stack current as follows [33, 34]:

$$\dot{m} = \frac{N_{cell} \cdot M_{h2}}{2 \cdot F} \cdot I_{stack} \cdot \lambda \quad (3)$$

where N_{cell} represents the cell number of the stack, M_{h_2} represents the molar mass of the hydrogen, F is the Faraday constant, and λ is the hydrogen excess ratio, which is set to 1.05 in this research. The FCS efficiency is low when the FCS operates in the low-power area as the auxiliary components require relatively high power in order to start up the fuel cell stack. The FCS efficiency also reduces in the high-power area due to the physical nature of the fuel cell stack. The FCS efficiency can be calculated as follows:

$$\eta_{fcs} = \frac{P_{fcs}}{\dot{m} \cdot Lhv} \quad (4)$$

where $Lhv = 120000$ kJ/kg is the lower heating value of the hydrogen. Specific relationships for the FCS used in this research are illustrated in Fig. 2.

Regarding the FCS lifetime, some researchers have proposed prediction methods [35–37], in which a lot of physical experiments are usually required to obtain important parameters. In this research, the specific FCS lifetime values are

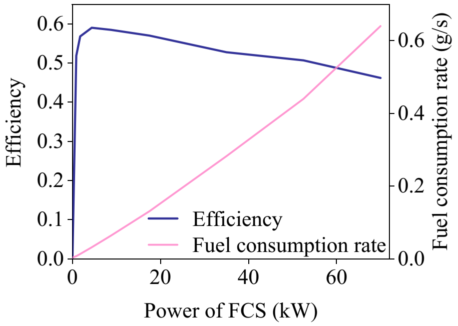


Fig. 2 Efficiency and hydrogen consumption rate of the FCS

not considered and the impact on the FCS lifetime is taken into account instead. Some research [38, 39] showed that the fluctuating load condition and the start-stop condition have greatest impacts on the operation of the fuel cell stack and further on the FCS lifetime. Considering the above, the average power difference of the FCS $\overline{\Delta P_{\text{fcs}}}$ under a certain DC is introduced to measure the impact of power fluctuations on the FCS lifetime, which are defined as follows:

$$\overline{\Delta P_{fcs}} = \frac{\sum_{t=1}^{N} \left| \left(P_{fcs}(t) - P_{fcs}(t-1) \right| \right.}{N} \tag{5}$$

where N represents the length of the DC.

2.3 Battery Model

In the FCHV, a battery is used as an auxiliary power source to assist the FCS during acceleration and recover the energy during braking. An Ni–MH battery and the partnership for a new generation vehicles (PNGV) [40] battery model shown in Fig. 3 are selected in this research.

As one of the most commonly used equivalent circuit models, the PNGV model has a clear physical meaning and guarantees high accuracy. The output voltage U_{batt} and power P_{batt} of the battery are expressed as follows:

$$\begin{cases} U_{batt} = V_{oc} - I_{batt}R_{int} - I_{pol}R_{pol} - \frac{1}{C_b}\int I_{batt}dt \\ P_{batt} = U_{batt}I_{batt} \\ \frac{dI_{pol}}{dt} = \frac{(I_{batt}-I_{pol})}{\tau} \end{cases} \tag{6}$$

where V_{oc} is the battery open circuit voltage (OCV), I_{batt} is the battery current, R_{int} is the battery internal resistance, I_{pol} is the polarization current, R_{pol} is the polarization resistance, C_b is a capacitance that accounts for the variation in the OCV with the time integral of the battery current I_{batt}, which is set to 386F, τ is the polarization time constant, namely $\tau = C_{pol}R_{pol}$, and C_{pol} is the polarization capacitance with a value of 480F. When the battery power is known, the battery current can be calculated from the Eq. (4). The battery state of charge (SOC) is an important factor, which is defined as the ratio of the remaining capacity to the total capacity Q_{cap}, as follow:

$$S\dot{O}C = -I_{batt}/Q_{cap} \tag{7}$$

Ignoring the influence of the temperature on the battery, the internal resistance, the polarization resistance, and the OCV of the battery are influenced by the battery SOC as illustrated in Fig. 4.

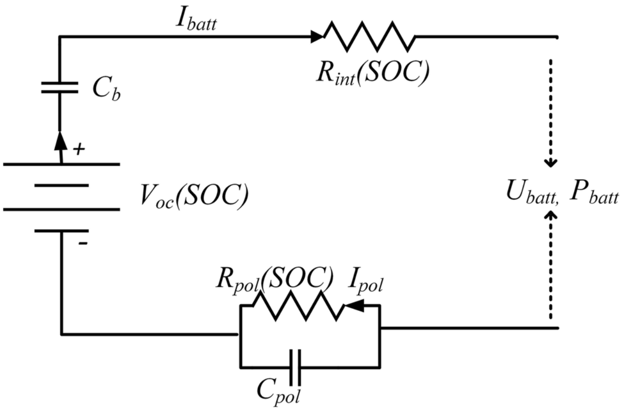


Fig. 3 Equivalent circuit of the PNGV battery model

2.4 Motor model

The electric motor is the only driver that converts the electrical energy of the FCS and battery into the mechanical energy to drive the FCHV. In the meantime, it can also act as a generator during regenerative braking to recover the braking energy and charge it into the battery. The maximum torque curves and the efficiency map of the motor used in this research are illustrated in Fig. 5.

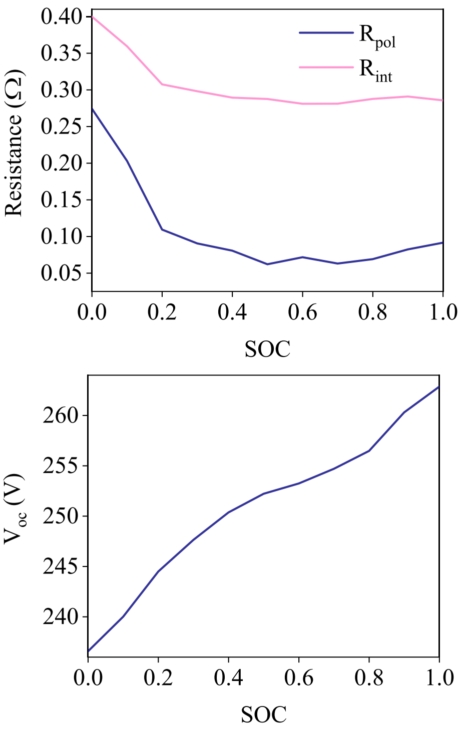


Fig. 4 Influences of the battery SOC on different battery parameters: **a** on the internal resistance and the polarization resistance; **b** on the OCV

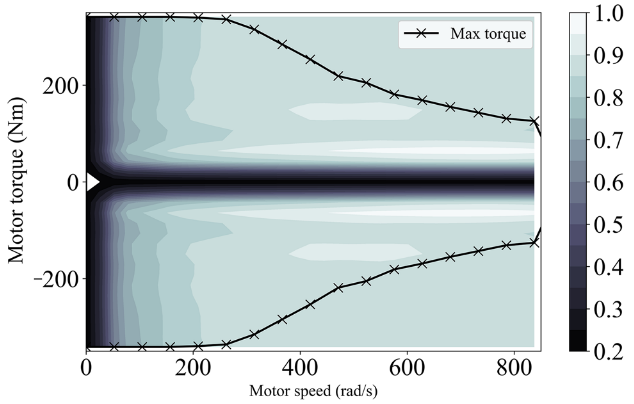


Fig. 5 Motor efficiency map and maximum torque curves

3 An EMS Design for the HESS

The RL algorithm is a learning frame structure that includes the agent, environment, state, action, and reward signal. To derive the optimal control action at a specific state by maximizing the reward function, the decision-maker called the agent continually interacts with the plant constrained by the external environment until the control strategy converges.

3.1 Problem Formulation

In this research, as the environment, the FCHV is modeled as a discrete-time dynamic system, which can be generally expressed by the following state equation:

$$x(t+1) = f(x(t), u(t)) \quad t = 0, 1, 2, \ldots, \mathrm{N} \tag{8}$$

where $x(t)$ and $u(t)$ represent the state variable and the control variable, respectively, at time t, which are selected as follows based on the analysis in Sect. 2, i.e. the power required for the vehicle and the battery SOC are the two state variables, and the battery power is the control variable.

$$\begin{cases} x(t) = \left(P_{req}(t), SOC(t)\right)^T \\ u(t) = \left(P_{batt}(t)\right)^T \end{cases} \tag{9}$$

In the RL algorithm, x and u above are usually named as the state and the action respectively. To ensure the FCS and battery working in safe and stable areas, the system should meet the following restrictions:

$$\begin{cases} SOC_{\min} \le SOC \le SOC_{\max} \\ P_{motor\_change\_max} \le P_{motor} \le P_{motor\_max} \\ P_{batt\_min} \le P_{batt} \le P_{batt\_max} \\ P_{fcs\_min} \le P_{fcs} \le P_{fcs\_max} \end{cases} \tag{10}$$

where parameters with min and max subscripts mean their corresponding minimum and maximum values, respectively. The SOC of the battery is maintained in [0.5, 0.75], in which the resistance is smaller; P_{motor_max} and P_{batt_max} are presented in Table 1, and $P_{motor\_change\_max}$, P_{batt_min} are the opposite values of them, respectively; P_{fcs_min} represents the minimum operating power of the FCS, which is set to 4 kW in this research, and P_{fcs_max} is also presented in Table 1.

If the state is updated by the "trial and error" randomization, there will be too many choices because of the discretization of the system, which will greatly prolong the simulation time. Therefore, the state TPM is introduced to simplify the state space in this research, i.e. the driving condition of the FCHV is considered as a Markov process, in which the next state is only related to the current state. The power required for the vehicle P_{req} can be calculated by Eq. (1), and then the TPM of it under various DCs can be obtained by the nearest neighbor method as follows:

$$p_{s,s'} = \frac{n_{s,s'}}{n_s} \tag{11}$$

where $p_{s,s'}$ represents the transition probability from the state s to s', $n_{s,s'}$ represents the occurrence number of the transition from the state s to s' in the whole DC, and n_s represents the total occurrence number of the state s in the whole DC. The UDDS, WLTC, and NEDC are selected as training DCs and the Japan1015 cycle is used as the validation DC in this research. These DCs cover the urban, suburban, and high-speed conditions and show big differences in the average speed, maximum speed, acceleration, and other characteristics, which will make the training strategy have the better adaptability. The corresponding velocity curves are shown in Fig. 6. Figure 7 shows the TPM of P_{req} on each DC, in which P_{next} represents the required power of the vehicle for the next moment. It can be observed that the values are generally limited to [0.2, 1], and most of them are concentrated on the diagonal because the required power of the vehicle rarely changes suddenly, which is in accordance with the actual situation.

The optimal strategy can be learned by maximizing the reward signal in the RL-based EMS. Thus, the form of the reward signal is significant for the RL-based EMS. In this research, the reward signal is designed by considering the control objectives and based on the reward signal forms sourced from previous research [41, 42]. To improve the fuel economy, maintain the battery SOC, and reduce the power fluctuation of the FCS, the reward signal is set as a function related to the hydrogen consumption rate $\dot{m}$, the battery SOC, and the FCS power difference between adjacent moments ΔP_{fcs}, as shown in (12), the specific form of which is determined based on multiple rounds of simulations and careful adjustments.

The reward signal is segmented based on the current battery SOC and the on–off status of the FCS, details of which

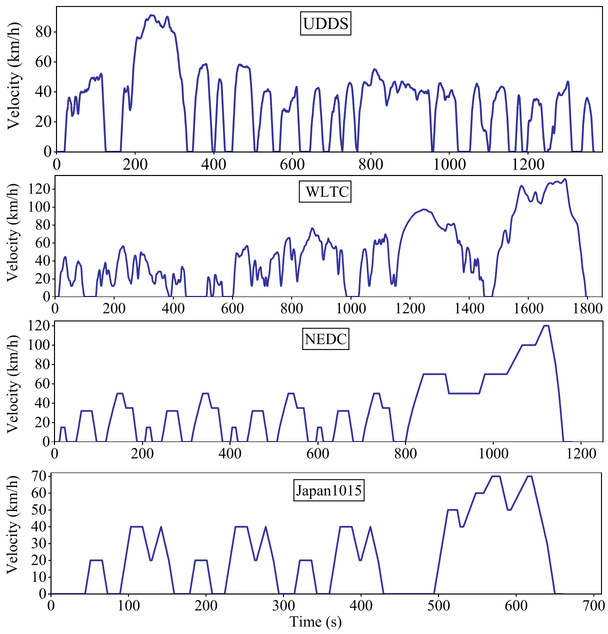


Fig. 6 Velocity curves of different DCs

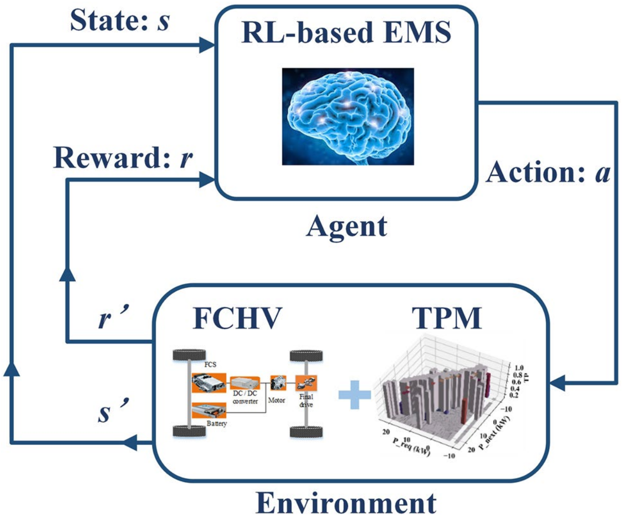


Fig.8 Schematic diagram of the interaction between agent and environment

are as follows: (1) when the FCS is working, different segments are set according to the battery SOC interval, and

Fig. 7 TPM of each DC: **a** UDDS; **b** WLTC; **c** NEDC; **d** Japan1015

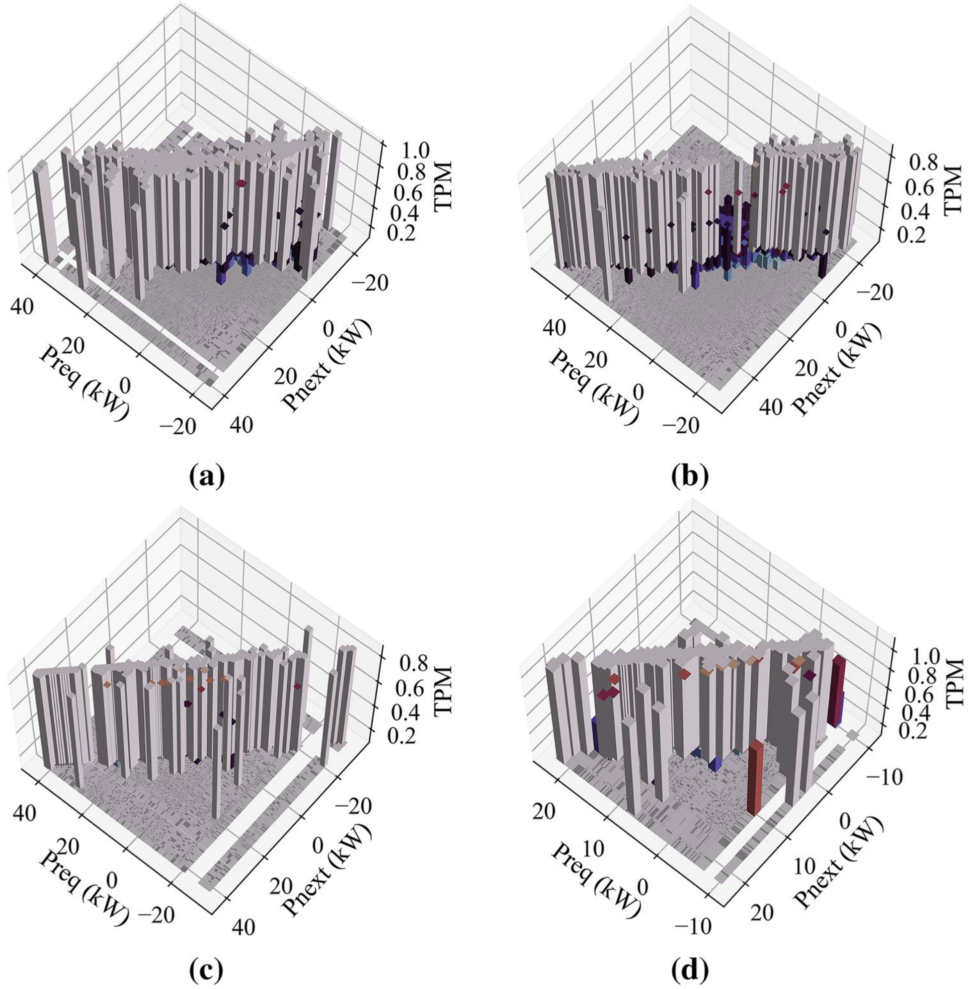


Table 2 Pseudocode of the proposed speedy Q-learning algorithm

```
Speedy Q-learning algorithm
Design rules
Pre-initialize the Q-table
Randomly select the initial $s, a$
Repeat each step:
    According to $Q(s, a)$ and current state $s$, choose $a$
    Take action $a$, observe $r(s, a)$ and $s'$
    $Q(s, a) \leftarrow Q(s, a) + \eta\left(r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)$
    $s \leftarrow s'$
Until $s$ is terminal
```

the reward signal is set to be inversely proportional to the instantaneous hydrogen consumption and the FCS power difference between adjacent moments (the first and second segments); (2) when the FCS is not working and the battery SOC is within a reasonable range, the battery should be used more to prevent on–off status changes of the FCS, thus the reward signal is set to be inversely proportional to the battery SOC (the third segment); (3) under the same condition with 2), if the battery is over-discharged, the reward signal should be reduced to prevent the battery from being used again, thus the reward signal is set to be inversely proportional to the battery power (fourth segment).

$$r = \begin{cases} \frac{1}{\dot{m}+\omega\times\left|\Delta P_{fcs}\right|} & \dot{m} \neq 0, SOC_{\min} \leq SOC \leq SOC_{\max} \\ \frac{1}{\dot{m}+10+\omega\times\left|\Delta P_{fcs}\right|} & \dot{m} \neq 0, SOC < SOC_{\min}, SOC > SOC_{\max} \\ \frac{1}{SOC+1} & \dot{m} = 0, SOC \geq SOC_{\min} \\ \frac{1}{P_{batt}} & \dot{m} = 0, SOC < SOC_{\min} \end{cases} \tag{12}$$

In (12), ω is set to 1 as the balance coefficient of the FCS power fluctuation after multiple rounds of simulations and careful parameter adjustments, and is set to 0 for the case where the FCS LE is not considered.

By summing up the above, the interaction process between the agent and the environment in the RL-based EMS is shown in Fig. 8, in which the strategy is used as the agent and the FCHV is the environment. The agent obtains the information on the state S_t and the reward R_t at time t and performs the appropriate action a_t to the environment. And then, the environment updates the state to S_{t+1} and feeds back the reward R_{t+1} at the next moment $t+1$ to the agent based on the current TPM. The interaction process is repeated until an optimal control strategy is learned.

3.2 Proposed RL-Based EMS

The objective function of the RL algorithm is defined as the total expectation of the cumulative reward in all future states, as follow:

$$V(s) = E\left(\sum_{t=0}^{\infty} \gamma^t r_t\right) \tag{13}$$

where γ is the discount factor set to 0.9 in this research, which is useful for guaranteeing the convergence during the learning process, E represents the expectation of cumulative returns, and $V(s_t)$ is a value function that satisfies Bellman's equation, as follows:

$$V(s) = r(s, a) + \gamma \sum_{s' \in S} P_{sa,s'} V\left(s'\right) \tag{14}$$

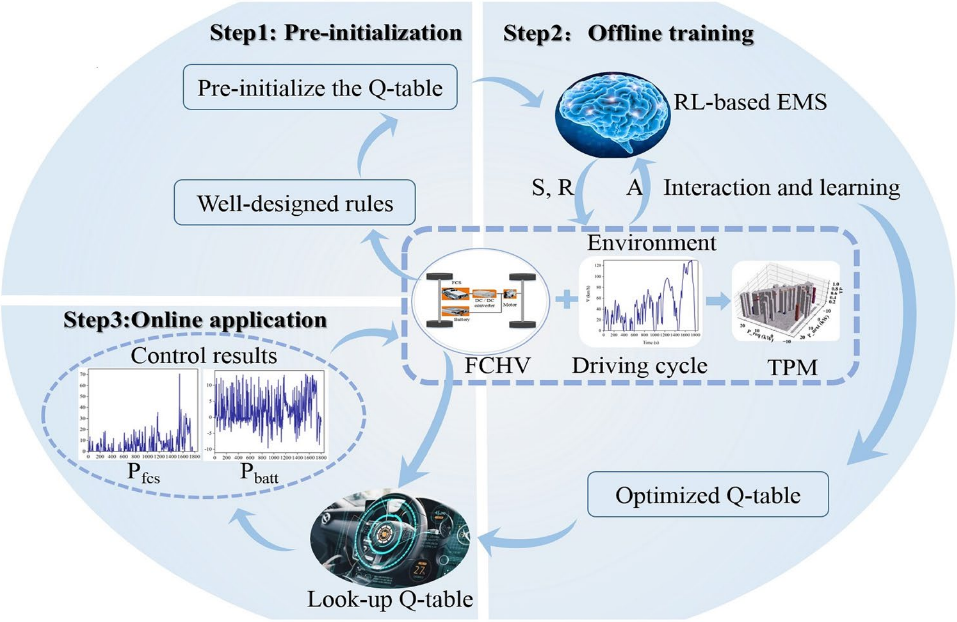


Fig. 9 Framework of the proposed RL-based EMS

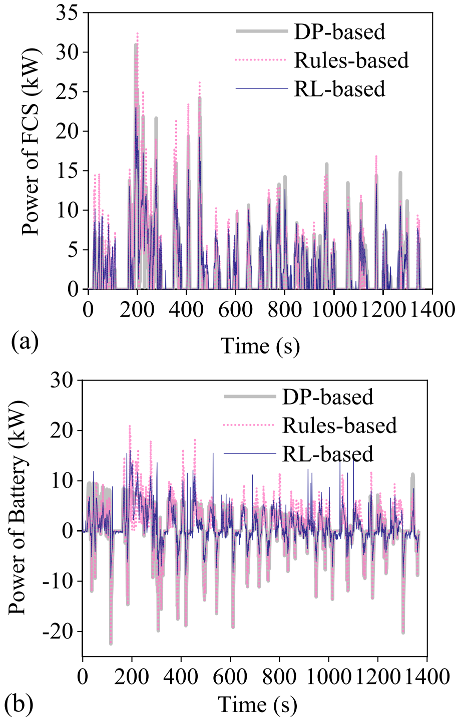


Fig. 10 FCS and battery output power of different strategies under UDDS: **a** FCS power; **b** battery power

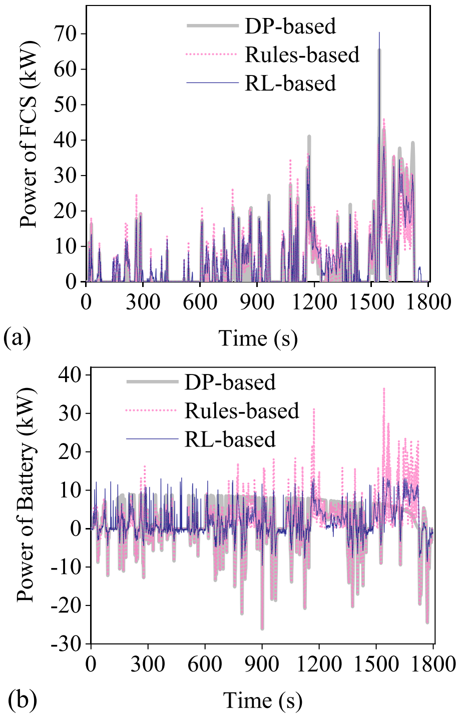


Fig. 11 FCS and battery output power of different strategies under WLTC: **a** FCS power; **b** battery power

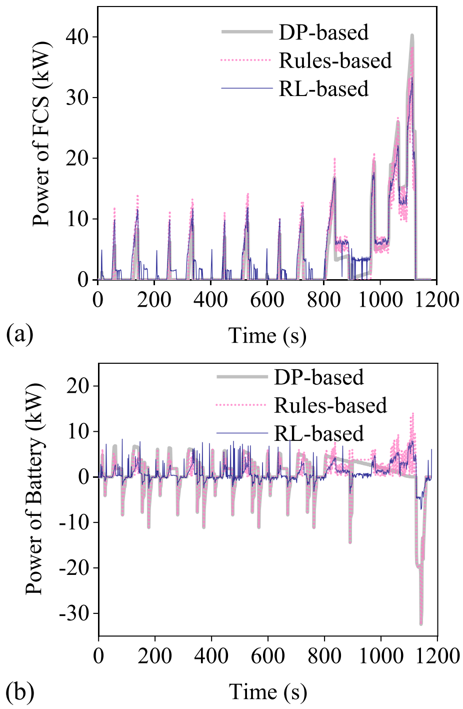


Fig. 12 FCS and battery output power of different strategies under NEDC: **a** FCS power; **b** battery power

where $P_{sa,s'}$ is the transition probability from the state s to s' with action a, which is the same with $P_{s,s'}$ in (9) in this research, and $r(s,a)$ is the instant reward in the current state s by taking action a.

The Q-learning algorithm is used for the RL-based EMS in this research, which is a commonly used RL algorithm and in which the maximum expected cumulative discounted reward is obtained by maximizing the Q-value function as follows:

$$\begin{cases} Q(s,a) = r(s,a) + \gamma \sum_{s' \in S} P_{sa,s'} Q(s',a') \\ Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in S} P_{sa,s'} \max_{a'} Q^*(s',a') \end{cases} \tag{15}$$

Eventually, the updating rule of the Q-learning algorithm is established as follows:

$$Q(s,a) \leftarrow Q(s,a) + \eta \Big(r(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a) \Big) \tag{16}$$

where η is the learning rate set to 0.001, the larger value of which will result in the faster RL convergence speed. However, too large value is likely to cause the learning oscillation and overfitting problems. The convergence and optimization of Q-learning algorithm have been widely proved, the optimal control strategy is given as follows [43, 44]:

$$\pi^*(s) = \arg\max_{a} Q^*(s,a) \tag{17}$$

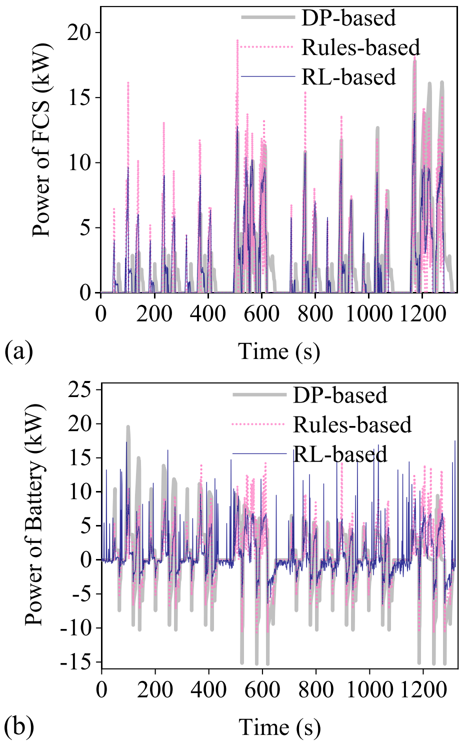


Fig. 13 FCS and battery output power of different strategies under Janpan1015: **a** FCS power; **b** battery power

A huge table called Q-table is built to store the Q-value for each state-action pair, which is continuously updated during the training process. In order to balance the relationship between exploration and exploitation in RL during training, the ε-greedy algorithm is used, i.e. the agent randomly chooses actions with a small probability $1-\varepsilon$ while selects actions maximizing the Q-function with a probability ε. Even if the TPM is used, a heavy offline training is still required because the Q-table is initialized to 0 in the conventional RL and the agent continually explores the environment randomly to update the Q-table, which will consume a lot of time. In this research, a speedy RL is proposed to reduce the offline training burden and accelerate the learning process. Table 2 shows the pseudocode of the proposed speedy Q-learning algorithm.

Rules are designed and utilized to obtain the preliminary Q-table, i.e. the preliminary reasonable action intervals are acquired according to the designed rules and the Q-values within these intervals are initialized to non-zero but extremely small positive numbers, so that the algorithm will have a higher exploration efficiency during offline training.

The framework of the proposed RL-based EMS is shown in Fig. 9, which can be divided into three steps, i.e. the pre-initialization, the offline training, and the online application. In the first step, rules are designed for pre-initializing the Q-table, which is helpful for the fast convergence of the Q-learning algorithm; the Q-table is continually updated according to the historical driving data and TPMs during the training process; during the online application, the Q-table performs a table lookup operation according to the current state and selects the action that can obtain the maximum

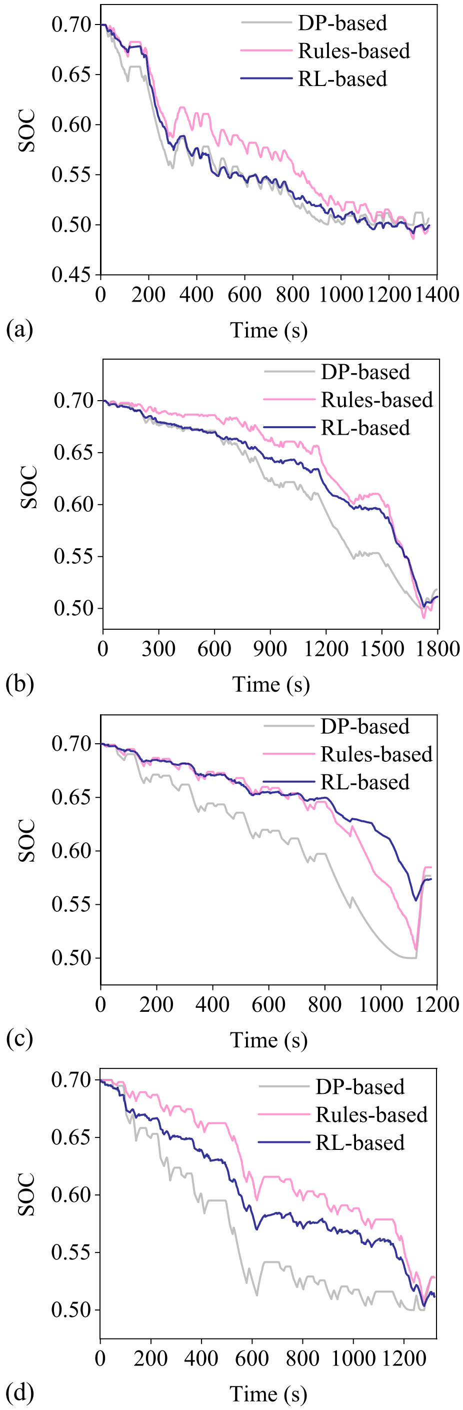


Fig. 14 Battery SOC curves of different strategies: **a** on UDDS; **b** on WLTC; **c** on NEDC; **d** on Japan1015

reward in real-time. If there is a state that is not covered by the Q-table, the proposed EMS will stay the same with the rules designed in the first step.

4 Simulation Results and Analysis

The proposed speedy RL-based EMS is implemented in the computer simulation environment and also trained and tested based on the contents introduced in Sects. 2 and 3. The proposed strategy would experience uncertainties in unfamiliar driving environments and untrained scenarios, as the training and optimization of the strategy are completed on the limited number of training DCs. In order to prove the effectiveness of the proposed strategy on untrained DCs, i.e. the adaptability, the UDDS, WLTC, and NEDC are selected as training DCs and Japan1015 cycle is used as the validation DC in this research. To ensure the duration of the validation DC is approximately equal to that of the training DC, Japan1015 cycle is continuously looped twice. Simulation results mainly focus on the aspects of the fuel economy, the battery SOC maintenance, the FCS LE, and the RL convergence speed. In order to prove the effectiveness of the proposed speedy RL-based EMS compared to other EMSs, simulation results of the proposed EMS are compared to those of the DP-based EMS, the rule-based EMS, and the conventional RL-based EMS respectively, among them the DP-based EMS is commonly used as the optimization benchmark.

4.1 Fuel Economy

The FCS LE is not considered in this subsection in order to compare the fuel economy of different EMSs purely, i.e. the weighting factor λ is set to 0 in the reward signal (10). The results on the FCS power and battery power on different DCs for different strategies are shown in Figs. 10, 11, 12 and 13. Taking the period of [880 s, 920 s] under UDDS as an example, the FCS is used as the main power source in both the RL-based and DP-based EMSs, while the battery is mostly used to drive the vehicle in the rule-based EMS. The difference in the power distribution will result in different fuel economy.

The corresponding battery SOC curves under different DCs are illustrated in Fig. 14. Although the battery SOC trajectories are different for each EMS on the training DCs, the final SOC values do not deviate much from each other due to the battery SOC maintaining function of each EMS. On the validation DC (Japan1015 cycle), there are no model constraints for the proposed strategy, and it directly looks up the table to select the optimal action, thus the final battery SOC difference is slightly larger.

Due to the slight difference on the final battery SOC for different strategies under the same DC, a final battery SOC correction method [45] is adopted for the fair comparison of the fuel economy. The equivalent hydrogen consumption results are listed in Table 3, in which the last column indicates the deviation from the result of the DP-based EMS. It can be observed that the fuel economy of the proposed strategy is deviated from that of the DP-based EMS by 3.58%, 6.35%, 4.71%, and 7.71%, while it is enhanced by 8.55%, 4.05%, 7.36%, and 6.04% compared to that of the rule-based EMS on the three training DCs and the validation DC, respectively. On the validation DC, the fuel economy of the proposed strategy is still better than that of the rule-based EMS and close to that of the DP-based EMS, which proves that the proposed strategy possesses the adaptability to different DCs.

Table 3 Equivalent hydrogen consumption results

DC	Strategy	Final SOC	Equivalent hydrogen consumption (kg/100 km)	Deviation from DP
UDDS	DP-based	0.5065	0.9000	–
	rule-based	0.4975	1.0119	12.43%
	RL-based	0.5101	0.9322	3.58%
WLTC	DP-based	0.5183	1.3824	–
	rule-based	0.5113	1.5322	10.84%
	RL-based	0.5112	1.4702	6.35%
NEDC	DP-based	0.5768	1.2095	–
	rule-based	0.5846	1.3671	13.03%
	RL-based	0.5737	1.2665	4.71%
Japan1015	DP-based	0.5291	1.0030	–
	rule-based	0.5283	1.1498	14.64%
	RL-based	0.5115	1.0803	7.71%

4.2 FCS LE

The consideration of the term ΔP_{fcs} in the reward signal makes the FCS output power smoother and helps to improve the FCS lifetime. To be consistent with 4.1, the strategy in which the weighting factor λ is set to 0 in the reward signal is named the RL-based EMS while the one in which λ is set to nonzero is named the RL-based EMS (with LE). The results on the adjacent power difference of the FCS ΔP_{fcs} on each DC for different EMSs are shown in Fig. 15. The $\overline{\Delta P}_{\text{fcs}}$ and the equivalent hydrogen consumption are shown in Table 4. It can be observed that the fuel economy of the RL-based EMS (with LE) is slightly reduced, however the power fluctuation on the FCS is reduced by more than 13% compared to the RL-based EMS, which is significantly beneficial for improving the FCS lifetime.

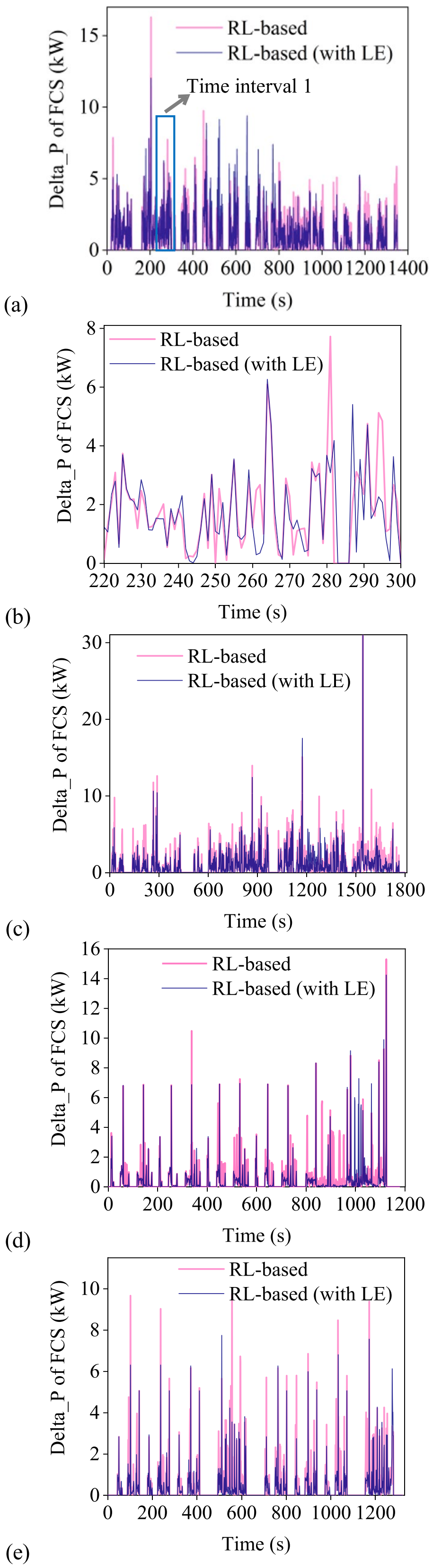
4.3 Convergence Speed

The TPM is adopted in the conventional RL-based EMS to simplify the calculation and greatly improve the training speed, but it is still difficult to meet the requirement for quickly updating parameters. The proposed speedy RL-based EMS, in which well-designed rules are used to pre-initialize the Q-table, further expedites the convergence speed on the basis of the conventional RL-based EMS. The above two EMSs are run separately on the UDDS and the corresponding mean square error (MSE) curves are shown in Fig. 16. It can be observed that the proposed speedy RL reaches the convergence with around 650 rounds while the conventional RL starts to converge with around 900 rounds, which means that the performance of the proposed speedy RL regarding to the convergence round is improved by 28% compared to the conventional RL. In addition, because the speedy RL consumes less calculation time per DC, its average time consumption is reduced by about 69% compared with the conventional RL.

5 Conclusion

The speedy RL-based FCS LE EMS is proposed for FCHVs in this research, which learns the optimized Q-table with a fast convergence rate during offline training on the training DCs by the pre-initialization framework. Simulation results show that the proposed strategy achieves good control results with respect to the fuel economy and the FCS LE and also presents the fast convergence speed and good adaptability. The following conclusions can be obtained from this research:

(1) The proposed strategy presents good fuel economy on the training DCs and the validation DC. The fuel economy difference between the proposed strategy and



(a)

(b)

(c)

(d)

(e)

Fig. 15 FCS adjacent power differences for the RL-based EMS and RL-based EMS (with LE): **a** on UDDS; **b** in time interval 1; **c** on WLTC; **d** on NEDC; **e** on Japan1015

Table 4 $\overline{\Delta P_{fcs}}$ and fuel economy of the RL-based EMS and RL-based EMS (with LE)

DC	Strategy	$\overline{\Delta P_{fcs}}$(kW)	Difference	Equivalent hydrogen consumption (kg/100 km)	Difference
UDDS	RL-based	1.13	–	0.9322	–
	RL-based (with LE)	0.98	13.27%	0.9434	1.20%
WLTC	RL-based	1.20	–	1.4702	–
	RL-based (with LE)	0.99	17.5%	1.5083	2.59%
NEDC	RL-based	0.65	–	1.2665	–
	RL-based (with LE)	0.53	18.46%	1.2938	2.16%
Japan-1015	RL-based	0.61	–	1.0803	–
	RL-based (with LE)	0.50	18.03%	1.1156	3.27%

the DP-based EMS is 3.58%, 6.35%, 4.71%, and 7.71% meanwhile the fuel economy of the proposed strategy is enhanced by 8.55%, 4.05%, 7.36%, and 6.04% compared to the rule-based EMS under the three training DCs and the validation DC, respectively.

(2) The FCS power fluctuation is reduced by more than 13% using the proposed strategy compared to the RL-based EMS which does not consider the FCS LE, and this is significantly beneficial for improving the FCS lifetime.

(3) The convergence speed of the proposed speedy RL-based EMS is increased by 69% with the pre-initialization framework compared to the conventional RL-based EMS, which presents the potential for real-time applications.

In learning-based EMSs, how to ensure sufficient data to make the strategy better adaptive to different driving cycles is a common problem to be explored. In the practical application, the proposed strategy needs more sufficient training and adjustment, which is a common challenge of learning-based methods. In addition, the curse of dimensionality problem may be occurred due to the discrete state space in the RL-based EMSs. From this perspective, DRL-based EMSs should also be developed at the same time, in which deep NNs are used to fit the high-dimensional state-action spaces of the control systems. This will be one of our future works in the hybrid vehicle EMS area.

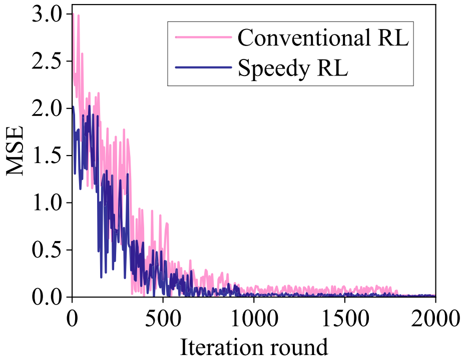


Fig. 16 MSE curves of speedy RL and conventional RL under UDDS

Acknowledgements This research was supported by Shenzhen Science and Technology Innovation Commission (Grant nos. KQJSCX20180330170047681 and JCYJ20180507182628567), Department of Science and Technology of Guangdong Province (Grant no. 2021A0505030056), CAS PIFI program (2021VEB0001), CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, 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. Liu, Y., Liu, J., Qin, D., Li, G., Chen, Z., & Zhang, Y. (2020). Online energy management strategy of fuel cell hybrid electric vehicles based on rule learning. *Journal of Cleaner Production, 260*, 121017.
2. Han, L., Jiao, X., & Zhang, Z. (2020). Recurrent neural network-based adaptive energy management control strategy of plug-in hybrid electric vehicles considering battery aging. *Energies, 13*(1), 202.
3. Hofman, T., Steinbuch, M., Van Druten, R., & Serrarens, A. (2007). Rule-based energy management strategies for hybrid vehicles. *International Journal of Electric and Hybrid Vehicles, 1*(1), 71–94.
4. Zheng, C., Wang, Y., Liu, Z., Sun, T., Kim, N., Jeong, J., & Cha, S. W. (2021). A hybrid energy storage system for an electric vehicle and its effectiveness validation. *International Journal of Precision Engineering and Manufacturing-Green Technology*. https://doi.org/10.1007/s40684-020-00304-5.
5. Bai, Y., He, H., Li, J., Li, S., Wang, Y. X., & Yang, Q. (2019). Battery anti-aging control for a plug-in hybrid electric vehicle with a hierarchical optimization energy management strategy. *Journal of Cleaner Production, 237*, 117841.
6. Lei, Z., Cheng, D., Liu, Y., Qin, D., Zhang, Y., & Xie, Q. (2017). A dynamic control strategy for hybrid electric vehicles based on parameter optimization for multiple driving cycles and driving pattern recognition. *Energies, 10*(1), 54.
7. Lei, Z., Qin, D., Zhao, P., Li, J., Liu, Y., & Chen, Z. (2020). A real-time blended energy management strategy of plug-in hybrid

electric vehicles considering driving conditions. *Journal of Cleaner Production, 252*, 119735.

8. Xu, L., Ouyang, M., Li, J., Yang, F., Lu, L., & Hua, J. (2013). Application of Pontryagin's Minimal Principle to the energy management strategy of plugin fuel cell electric vehicles. *International Journal of Hydrogen Energy, 38*(24), 10104–10115.
9. Yuan, Z., Teng, L., Fengchun, S., & Peng, H. (2013). Comparative study of dynamic programming and Pontryagin's minimum principle on energy management for a parallel hybrid electric vehicle. *Energies, 6*(4), 2305–2318.
10. Zheng, C., & 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.
11. Kim, N., Jeong, J., & Zheng, C. (2019). Adaptive energy management strategy for plug-in hybrid electric vehicles with Pontryagin's minimum principle based on daily driving patterns. *International Journal of Precision Engineering and Manufacturing-Green Technology, 6*(3), 539–548.
12. Zeng, X., & Wang, J. (2015). A parallel hybrid electric vehicle energy management strategy using stochastic model predictive control with road grade preview. *IEEE Transactions on Control Systems Technology, 23*(6), 2416–2423.
13. Wang, Y., Wang, X., Sun, Y., & You, S. (2018). Model predictive control strategy for energy optimization of series-parallel hybrid electric vehicle. *Journal of cleaner production, 199*, 348–358.
14. Bambang, R. T., Rohman, A. S., Dronkers, C. J., Ortega, R., & Sasongko, A. (2014). Energy management of fuel cell/battery/supercapacitor hybrid power sources using model predictive control. *IEEE Transactions on Industrial Informatics, 10*(4), 1992–2002.
15. Hu, Y., Yang, L., Yan, B., Yan, T., & Ma, P. (2015). An online rolling optimal control strategy for commuter hybrid electric vehicles based on driving condition learning and prediction. *IEEE Transactions on Vehicular Technology, 65*(6), 4312–4327.
16. Zhou, D., Al-Durra, A., Gao, F., Ravey, A., Matraji, I., & Simoes, M. G. (2017). Online energy management strategy of fuel cell hybrid electric vehicles based on data fusion approach. *Journal of Power Sources, 366*, 278–291.
17. Murphey, Y. L., Park, J., Chen, Z., Kuang, M. L., Masrur, M. A., & Phillips, A. M. (2012). Intelligent hybrid vehicle power control—part I: Machine learning of optimal vehicle power. *IEEE Transactions on Vehicular Technology, 61*(8), 3519–3530.
18. Xiang, C., Ding, F., Wang, W., & He, W. (2017). Energy management of a dual-mode power-split hybrid electric vehicle based on velocity prediction and nonlinear model predictive control. *Applied Energy, 189*, 640–653.
19. Qian, L. J., Gong, Z., & Zhao, H. (2006). Simulation of hybrid electric vehicle control strategy based on fuzzy neural network. *Journal of System Simulation, 18*(5), 1384–1387.
20. Xie, S., Hu, X., Qi, S., & Lang, K. (2018). An artificial neural network-enhanced energy management strategy for plug-in hybrid electric vehicles. *Energy, 163*, 837–848.
21. 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.
22. 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.
23. 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.
24. Liu, T., Wang, B., & Yang, C. (2018). Online Markov Chain-based energy management for a hybrid tracked vehicle with speedy Q-learning. *Energy, 160*, 544–555.
25. 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.
26. 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.
27. 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.
28. Zhou, Q., Li, J., Shuai, B., Williams, H., He, Y., Li, Z., et al. (2019). Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle. *Applied Energy, 255*, 113755.
29. 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.
30. 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.
31. 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*, 5013–5023.
32. Autonomie. https://www.autonomie.net/
33. Bernard, J., Delprat, S., Buechi, F., & Guerra, T. M. (2006). Global Optimisation in the power management of a Fuel Cell Hybrid Vehicle (FCHV). In: *2006 IEEE vehicle power and propulsion conference*, 1–6.
34. Lin, W. S., & Zheng, C. H. (2011). Energy management of a fuel cell/ultracapacitor hybrid power system using an adaptive optimal-control method. *Journal of Power Sources, 196*(6), 3280–3289.
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. 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.
37. Chen, H., Pei, P., & Song, M. (2015). Lifetime prediction and the economic lifetime of proton exchange membrane fuel cells. *Applied Energy, 142*, 154–163.
38. Pei, P., & Chen, H. (2014). Main factors affecting the lifetime of Proton Exchange Membrane fuel cells in vehicle applications: A review. *Applied Energy, 125*, 60–75.
39. Wu, P., Partridge, J., & Bucknall, R. (2020). Cost-effective reinforcement learning energy management for plug-in hybrid fuel cell and battery ships. *Applied Energy, 275*, 115258.
40. Daowd, M., Omar, N., Van Den Bossche, P., & Van Mierlo, J. (2011). Extended PNGV battery model for electric and hybrid vehicles. *International Journal of Revolution in Electrical and Electronic Engineering, 6*, 1264–1278.
41. 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.
42. Chen, Z., Mi, C. C., Xu, J., Gong, X., & You, C. (2013). Energy management for a power-split plug-in hybrid electric vehicle based on dynamic programming and neural networks. *IEEE Transactions on Vehicular Technology, 63*(4), 1567–1580.

43. Melo, F. S. (2001). Convergence of Q-learning: A simple proof. *Institute of Systems and Robotics*, Technical Report, pp 1–4.
44. Ribeiro, C., & Szepesvári, C. (1996). Q-learning combined with spreading: Convergence and results. In: *Proceedings of the ISRF-IEE International Conf. on Intelligent and Cognitive Systems (Neural Networks Symposium),* pp 32–36.
45. Clark, N., Xie, W., Gautam, M., Lyons, D. W., Norton, P., & Balon, T. (2000). Hybrid diesel-electric heavy duty bus emissions: Benefits of regeneration and need for state of charge correction (No. 2000-01-2955). *SAE Technical Paper*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wei Li received his B. S. from Southwest Jiaotong University in 2018. His research interest includes energy management strategies of hybrid vehicles.

Jiaye Ye received his Ph.D. from Chongqing University in 2020. He is currently working as an Assistant Professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research field includes electrochemical energy devices such as flow battery and fuel cell.

Yunduan Cui is currently undertaking an Associate Professor in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He received Ph.D. from Nara Institute of Science and Technology, Japan in 2017, M.E from Doshisha University, Japan in 2014, and the B.E. from Xidian University, China in 2012. His research interests include machine learning, especially reinforcement learning in various robots and unmanned systems.

Namwook Kim received his Ph.D. from Seoul National University in 2009. He is currently working as an Associate Professor in the Department of Mechanical Engineering, Hanyang University ERICA Campus at Ansan, South Korea.

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.

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.