**Applied Energy Symposium 2021: Low carbon cities and urban energy systems**
**September 4-8, 2021, Matsue, Japan**
**Paper ID:56**

# IMPACT OF EXPLORATION-TO-EXPLOITATION RATIO ON ENERGY SAVING POTENTIAL OF PLUG-IN HYBRID VEHICLES CONTROLLED BY REINFORCEMENT LEARNING

Bin Shuai[1,] Quan Zhou[1] Huw Williams[1], Hongming Xu[1*], Yanfei Li[2], Lun Hua[3]

1 Department of Mechanical Engineering, University of Birmingham, B15 2TT, UK

2 State Key Laboratory of Automotive Energy and Safety, Tsinghua University, 100083, China

3 Suzhou Automotive Research Institute, Tsinghua University, 215200, China

(*Corresponding Author: h.m.xu@bham.ac.uk )

**ABSTRACT**

Electric vehicles, including hybrids, will domain the road transport in future cities. Reinforcement learning has shown its capacity in online optimization of energy management strategy for hybrid vehicles. Exploration of new control settings and exploitation of the existing control policy are two key procedures in reinforcement learning but there is a lack of study on how the exploration-to-exploitation (E2E) ratio affects the energy efficiency improvement for hybrid vehicles. This paper introduces two decay functions, 'Reciprocal function-based decay' (RBD) and 'Step-based decay' (SBD), to generate E2E ratio trajectories for reinforcement learning algorithm which is conventionally based on Exponential decay (EXD) function. By monitoring the improving rate of vehicle energy efficiency in the learning process, the vehicle controlled by Q-learning algorithm based on the SBD function has shown the best compared with the vehicles based on the RBD function and the EXD function. The improving rate can be more than 6.21%. In the HiL testing, the SBD can save 1.52% energy compared to the EXD in real-time control under a predefined driving cycle.

**Keywords:** Vehicle energy efficiency, Reinforcement learning, Hybrid electric vehicle, exploration-to-exploitation ratio

## 1. INTRODUCTION

With the aim to eliminate greenhouse gas (GHG) and reduce energy consumption [1], transportation has been deemed to take responsibility for this problem due to almost automobile derived by the internal combustion engines (ICE). Recently, the hybrid electric vehicle (HEV) is considered a booming solution for acquiring low-recourse consuming and environment-friendly due to their high vehicle's energy efficiency and low CO2 emission [2]. It generally consists of two energy sources, including an internal combustion engine (ICE) and a battery pack, which provide a better mileage in normal life driving. Energy management system (EMS) plays a crucial part of the HEV, which efficiently disturbs the energy flow of the powertrain within the system of multiple power components [3].

Rule-based methods are widely applied to commercial HEVs because of the less memory requirements. The control rules are mainly designed based on intuition, human experience, and mathematical models. However, owing to the uncertain driving conditions, the control rules are required calibrations more frequently to achieve better adaptatively under various worldwide driving conditions. As an alternative, optimization-based methods are necessarily needed to guarantee the energy distribution under any driving conditions [4]. Other offline optimization algorithms (e.g. modified accelerated particle swarm optimization (MAPSO) [5], swarm optimisation (PSO) [6] genetic algorithm (GA) [7] and dynamic programming (DP) )are typical global algorithms as a benchmark for the HEVs in given driving cycles [8]. However, the above optimal algorithms usually require an informative prior knowledge of entire driving conditions, which is not able to carry out in online real-time control. Therefore, online optimization-based control methods have been become an urge concern for

HEVs. The model predictive control (MPC) has been implemented for HEVs to obtain an optimal solution in real-time control [9]. However, a precise models (e.g., ageing, changing of drivers) and huge computational capability are obvious drawbacks for the MPC [10].

In recent years, learning-based optimal algorithms have been widely applied in many research areas such as Robotics controls [11] driver-classification [12], imaging process [13] and Gaming [14]. Reinforcement learning (RL) methods have been validated that it can lead to significant improve in the energy efficiency of hybrid electric vehicles due to its model-free and self-learning ability [15].

Q-learning is a widely used reinforcement learning method, and it has been implemented in optimal control strategy for the EMS of HEVs. Liu et al. implemented the Q-learning method to EMS in various topologies (e.g. parallel hybrid powertrain [16], power-split powertrain [18] ) of hybrid powertrains. Zhou et al. proposed a multi-step learning-based control method [19]and a novel adaptive learning network [20] to improve the HEVs control strategy in real-time driving. The deep reinforcement learning method is also applied to EMS for Eco-driving [21], Du et al. proposed a new framework for achieving experience sampling more reasonable [22]. Zou et al. carried out a combination between model predictive control and deep network to accelerate the learning speed [23]. However, these two exploration-exploitation policies are not completely biased, they are still having some difficulty in balancing the exploration-exploitation process [24], especially on how to define the value of the deciding factor in epsilon-greedy policy to achieve an optimal control strategy. Conventional Q-learning usually implements an exponential decay (EXD) function to control the exploration to exploitation (E2E) ratio [25]. To best of author's knowledge, there is a lack of study on how the exploration-to-exploitation (E2E) ratio affects the energy efficiency improvement for hybrid vehicles.

This paper has two main contributions: 1) two new E2E ratio decay functions, 'Reciprocal function-based decay' (RBD) and 'Step-based decay' (SBD), are introduced to generate E2E ratio trajectories for reinforcement learning algorithm which is conventionally based on Exponential decay (EXD) function; and 2) the energy saving potentials of different E2E ratio decay functions are quantified by monitoring the improving rate of vehicle energy efficiency during the reinforcement learning process.

The remainder of the paper is organized as follows: the studied hybrid electric vehicle system is described in Section 2. The framework of the energy management system is introduced in Section 3. Section 4 introduces the two new E2E ratio decay functions. Section 5 analyses the results of experimental evaluations. Conclusions are drawn in Section 6.

## 2. THE HYBRID ELECTRIC POWERTRAIN SYSTEM

The hybrid powertrain of the studied vehicle is illustrated in **Fig.1**. The traction motor is mainly powered by the battery pack as the primary power unit. The alternative power unit is consisted of an ICE and a generator (charging the battery pack) for normal vehicle operations. The key vehicle specifications are summarized in Table 1.

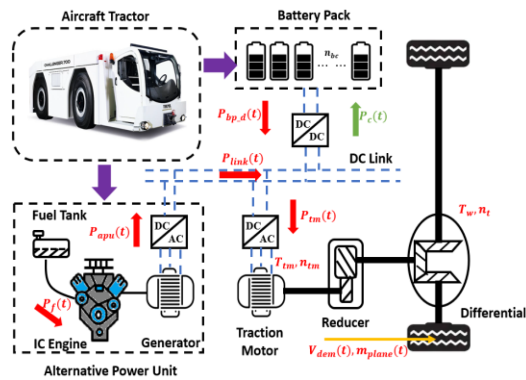


Fig. 1 Power flow of the hybrid powertrain system

Table I Vehicle parameter

| Parameter | Description | Value |
|---|---|---|
| $m_{\mathrm{veh}}$ | Vehicle mass | 16t |
| $r_{\mathrm{whl}}$ | The radius of the wheels | 0.75m |
| $f_f$ | The friction coefficient | 0.02 |
| $\rho_{\mathrm{air}}$ | The density of the air | 1.2258 |
| $C_d$ | Aerodynamic drag coefficient | 0.8 |
| $A_f$ | Effective front area | 6.8m$^2$ |

## 3. THE ENERGY MANAGEMENT SYSTEM

The supervisory energy management system contains two interconnected layers [26], as shown in **Fig.2**. The two layers can be connected via local control network or the Internet of vehicles for different application scenarios. In the control layer, the energy management system continuously sends the engine control signal, $u_{egu}(t)$), to the alternative power unit and transmits the vehicle states (power demand, $P_d(t)$ and battery SoC, $SoC(t)$ ) and performance (energy loss, $P_{loss}(t)$) to the learning layer. In the learning layer, the action selection policies module receives the vehicle information and runs action selection policy to optimize the control strategy. Then the control policy will be transferred to the control layer to enable the best energy economy for the hybrid electric tractor driving in real-world driving.

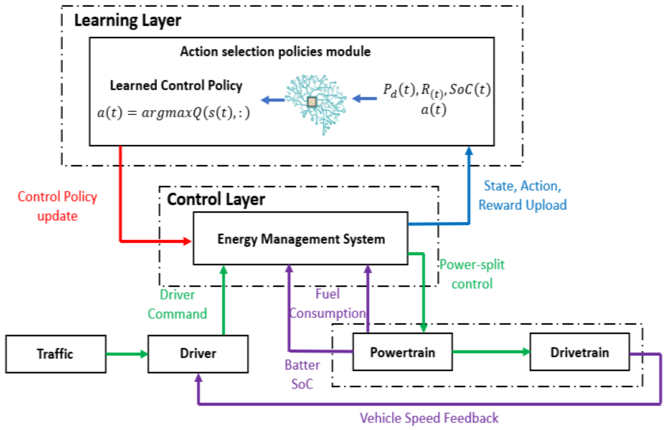


Fig. 2 Framework of supervisory engine control

### *3.1 States*

At each time interval $t$, driver's power demand, $P_d(t)$, and battery state-of-charge, $SoC(t)$, are selected as the reinforcement learning 'States' for this research, as shown below:

$$S(t) = [P_d(t), SoC(t)] \tag{5}$$

### *3.2 Action*

The outputs of the energy management system are the engine power demand, $u_{egu}(t)$, and battery power demand, $u_{batt}(t)$. This paper uses the engine power demand as the action variable and it is determined by the control policy, $\mathbf{\Pi}$, by

$$a(t) = u_{egu}(t) = \mathbf{\Pi}(S(t), \mathbf{Q}) \in [0,1] \tag{6}$$

where, $\mathbf{Q}$ is the knowledge based that is developed by reinforcement learning algorithm. Once the power demand of engine is obtained, the power provided by BP, $P_{batt}$, can be calculated by

$$P_{batt}(t) = \frac{P_r - u_a(t) \cdot P_{a_{max}}}{P_{b_{max}}} \tag{7}$$

where, $P_{b_{max}}$ = 365kw and $P_{a_{max}}$ = 86.2 kW are the maximum power available of BP and engine generator unit, respectively; and $P_r$ is the power requirement for driving the vehicle move forward.

### *3.3 Reward*

A merit function is used for evaluating the vehicle performance after taking selected control signal, which can be calculated by [19]

$$r(t) = \begin{cases} r_{ini} - P_{loss}(t) & SoC(t) \geq SoC_{ref} \\ r_{ini} - P_{loss}(t) - \mu |SoC_p - SoC(t)| & SoC(t) < SoC_{ref} \end{cases} \tag{8}$$

where, a constraint factor $SoC_p$ =30% is needed to keep a longer battery usage; a scale factor $\mu$ balances the power efficiency and BP's SoC level; $P_{loss}(t)$ is considered as a total power loss from the alternative power unit, it can be calculated by:

$$P_{loss}(t) = L_{eng}(t) + L_{batt}(t) \tag{9}$$

$L_{eng}(t)$ and $L_{batt}(t)$ are the equivalent power loss of the diesel engine, and the power loss of the BP, respectively. They can be calculated by:

$$\left. \begin{aligned} L_{eng}(t) &= \dot{m_f}(t) \cdot H_f - \frac{T_{eng}(t) \cdot n_{eng}(t)}{9550} \\ L_{batt}(t) &= R_{loss}(SoC) \cdot I_{batt}(t)^2 \end{aligned} \right\} \tag{10}$$

where, $\dot{m_f}$ is the fuel rate in real-time. $T_{eng}$ and $n_{eng}$ are torque and running speed of the diesel engine, respectively; $I_{batt}$ is BP's current; The equivalent internal battery resistant $R_{loss}$ is a function of battery SoC; and diesel fuel heat value is $H_f = 44 \times 10^6$J/kg.

## 4. EXPLORATION TO EXPOLITIATION RATIO DECAY FUNCTIONS

Exploration of new control settings and exploitation of the existing control policy are two key procedures in reinforcement learning. Exploration and exploitation will generate different action signals for same vehicle state condition thus will obtain different rewards to building the knowledge base, $\mathbf{Q}$, as:

$$\mathbf{Q}(\boldsymbol{S}_t, a_t) \leftarrow \mathbf{Q}(\boldsymbol{S}_t, a_t) + \alpha \cdot [r_t + \gamma \cdot \max \mathbf{Q}(\boldsymbol{S}_{t+1}, :) - \mathbf{Q}(\boldsymbol{S}_t, a_t)] \tag{11}$$

where $\alpha \epsilon [0,1]$ is the learning rate; and $\gamma$ is a discount factor weighting the maximum estimated reward from the knowledge base, $\max \mathbf{Q}(\boldsymbol{S}_{t+1}, :)$.

In conventional Q learning algorithms, exploration and exploitation is control based on the comparation between a random number, $\varepsilon$, with the value of an explorational decay function, $\theta = b^k$ ,where $k$ is the current learning iteration; If $\varepsilon > \theta$ the algorithm will do exploitation based on the current control policy, otherwise it will explore possible better control policy by randomly pick up an action value. To study the Q-learning performance with different E2E ratio (E2ER), this paper introduces two new decay functions to generate different E2E trajectories for Q-learning based vehicle energy management control.

### *4.1 Step-based decay policy*

The E2ER will be decreased as the people go down the stairs. This mechanics can lead to gain more opportunities to interact with the unknown environment. The E2ER ($\theta$) of Step-based policy can be calculated as below:

$$\theta = f \times drop^{(\frac{1+LR}{LR_{drop}})} \tag{11}$$

where, $f$ is previous deciding factor value; and $drop$ is a scale factor; $LR$ is current learning iteration; and $LR_{drop}$ controls the deciding factor dropping after how many learning iterations. Typically, E2ER will be dropped after 10 learning iterations.

### *4.2 Reciprocal function-based decay policy*

The E2ER of reciprocal function-based decay policy decreases as number of learning iterations increasing. The different from traditional exponential decay (EXD) is that the E2ER will drop faster than EXD. The E2ER ($\theta$) of reciprocal-decay is described as follow:

$$\theta = \frac{f}{Sf + R_{rate} * LR} \tag{12}$$

where, $Sf$ is a scale factor, usually equals to 1, $R_{rate}$ is a decay rate which will decrease $\theta$ by the given fixed amount[27].

## 5. EXPERIMENTAL EVALUATIONS

### *5.1 Testing driving cycle and platform*

Four predefined driving cycles are selected to evaluate the real-time performance of proposed policies, as shown in **Fig.3**. The proposed two action selection polices is firstly carried out in MATLAB/Simulink platform with initial battery state-of-charge of 50% under the predefined driving cycle one. And the results will be compared with traditional exponential decay (EXD) policy.

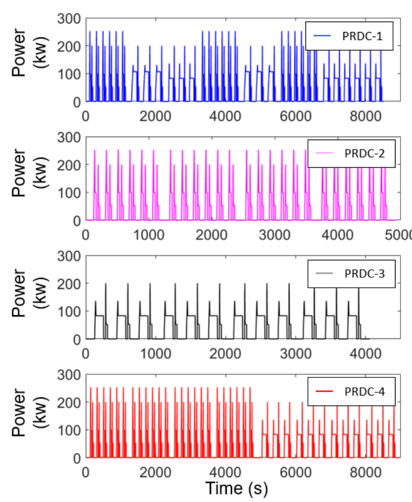


Fig. 3 Power of the predefined driving cycle

Next, the proposed policies will be implemented on the hardware-in-the-loop-test (HiL). A Desk LABCAR is used for HiL testing, as shown in Error! Reference source not found.**Fig.4**. The control prototype and the real-time vehicle model are compiled in a development PC, downloaded onto the DESL LABCAR through Ethernet. The result can be monitored through development computer.

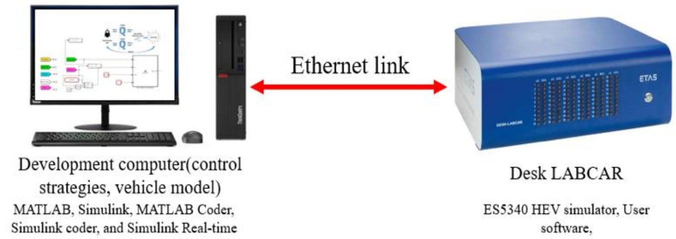


Fig. 4 Hardware-in-the-loop platform for real-time testing

### *5.2 learning performance of Step-based decay policy and Reciprocal function-based decay policy*

The learning evaluation of step-based decay (SBD) policy, reciprocal function-based decay (RBD) policy and exponential decay (EXD) policy was demonstrated in **Fig.5**. And the variation $\theta$ of three policies during the learning process was shown in **Fig.6**. The EXD was set as a benchmark. The EXD tend to be more stable after the 25th learning iterations and finally reached to 36.56%. The learning performance of the RBD was not stable and achieved 36.21% at 125th learning iteration. This because the deciding factor dropped rapidly, which did not leave enough time to the agent for understanding the unknown driving conditions. The SBD performed the best among three action selection policies reached to 36.82%. It still had huge probability to explore new vehicle engine control signals between the 20th and 60th of learning iterations to improve the vehicle energy efficiency. The SBD achieved the highest improvement rate of 6.21% (final vehicle energy efficiency compared with initial vehicle energy efficiency), which was higher than RBD with 1.36% and 0.33% than the EXD policy, respectively.

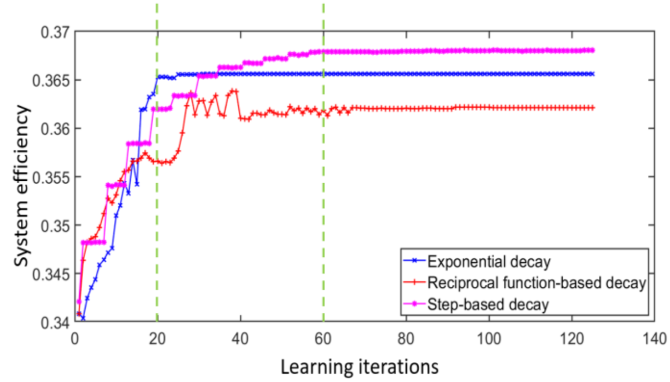


Fig. 5 Learning performance of three model-free methods

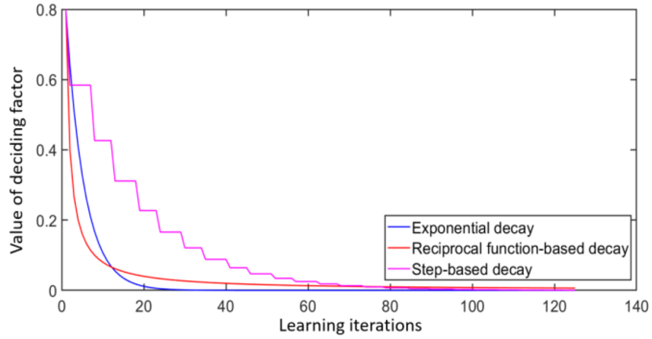


Fig. 6 The variation curve of the deciding factor

### *5.3 Real-time performance*

By deploying the two proposed policies in the hardware-in-loop testing system, the real-time performance was obtained and compared with the results from EXD through total vehicle energy loss, battery SoC and total battery energy loss as shown in Fig. 1**Fig.7**. The LBD achieved a 1.52% lower total energy loss than the benchmark policy EXD. The SBD can save more energy among three methods, the main energy saving came from the battery total loss, the 11.62%, 12.63% lower than the RBD and the EXD policy, respectively.

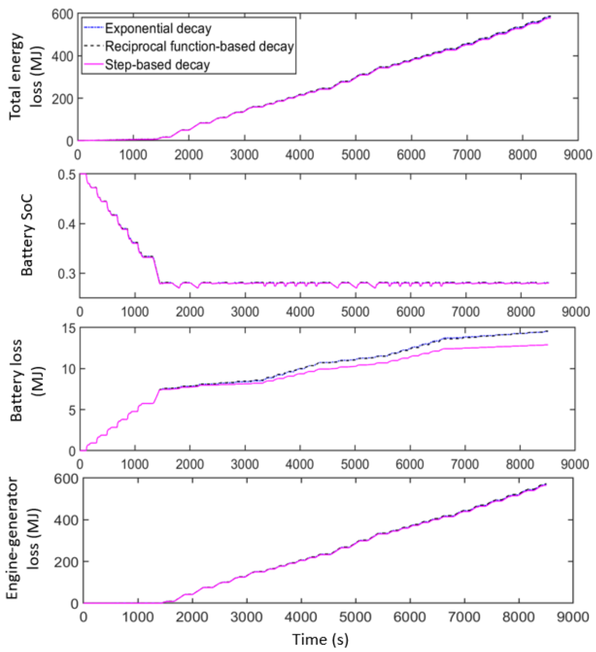


Fig. 7 Real-time performance of three model-free methods

### *5.4 Robustness testing*

In actual driving situations, the hybrid electric vehicle energy management system will encounter various driving conditions. Therefore, a robustness test is required for step-based decay (SBD) and reciprocal function-based decay (RBD) policies. The traditional exponential decay (EXD) will be set as a benchmark. The results are given in Table II, including the initial battery state-of-charge (SoC)the battery SoC at the end of learning iteration (End battery SoC), the total energy usage (TEU) and the saving rate (savings). The adaptivity of SBD outperform than the RBD and the EXD in all predefined driving cycles (2-4). Compared to the EXD, the SBD can save at least 1.25% energy, and the highest saving rate was 2.04% in predefined cycle 2.

Table II Performance of the CBD, FDB and LBD under predefined driving conditions

| Driving cycles | Battery SoC | Policy | End battery SoC | TEU (MJ) | savings |
|---|---|---|---|---|---|
| PRDC-2 | 50% | EXD | 28.13% | 239.22 | - |
| | 50% | RBD | 28.16% | 240.23 | -0.42% |
| | **50%** | **SBD** | **28.04%** | **234.42** | **2.04%** |
| PRDC-3 | 50% | EXD | 28.13% | 201.72 | - |
| | 50% | RBD | 28.16% | 203.13 | -0.69% |
| | **50%** | **SBD** | **28.02%** | **199.22** | **1.25%** |
| PRDC-4 | 50% | EXD | 28.13% | 543.52 | - |
| | 50% | RBD | 28.16% | 546.64 | -0.57% |
| | **50%** | **SBD** | **28.01%** | **536.21** | **1.36%** |

## 6. CONCLUSIONS

To study the influence of exploration-to-exploitation (E2E) ratio to the energy efficiency improvement for hybrid vehicles, this paper introduces two decay functions, 'Reciprocal function-based decay' (RBD) and 'Step-based decay' (SBD), to generate E2E ratio trajectories for reinforcement learning algorithm that is conventionally based on exponential decay (EXD) function. Learning performance of two proposed policies were monitored from hybrid vehicle energy efficiency based on both software-in-the-loop (SiL) and hardware-in-the-loop (HiL) platforms. The conclusions drawn from this work are as follows:

- For vehicle optimization using reinforcement learning, the SBD function can achieve the highest vehicle energy efficiency improvement rate of 6.21% compared with RBD and EXD.
- Reinforcement learning algorithm based on the SBD function can contribute to 11% energy saving rate from the battery system compared with RBD-based and EXD based systems in predefined driving cycle one.
- The performance of the SBD-based system is robust. It can save more than 1.25% energy compared to the conventional system (EXD) in the predefined driving cycles, i.e., PRDC 2-4.

## ACKNOWLEDGEMENT

The authors are grateful to the Innovate UK (Grant 102253) and the State Key Laboratory of Automotive Safety and Energy (KF2029).

## REFERENCE

[1] C. E. Sandy Thomas, "Transportation options in a carbon-constrained world: Hybrids, plug-in hybrids, biofuels, fuel cell electric vehicles, and battery electric vehicles," *Int. J. Hydrogen Energy*, vol. 34, no. 23, pp. 9279–9296, 2009, doi: 10.1016/j.ijhydene.2009.09.058.

[2] F. Zhang, X. Hu, R. Langari, and D. Cao, "Energy management strategies of connected HEVs and PHEVs: Recent progress and outlook," *Progress in Energy and Combustion Science*, vol. 73. pp. 235–256, 2019, doi: 10.1016/j.pecs.2019.04.002.

[3] M. F. M. Sabri, K. A. Danapalasingam, and M. F. Rahmat, "A review on hybrid electric vehicles architecture and energy management strategies," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 1433–1442, 2016, doi: 10.1016/j.rser.2015.09.036.

[4] J. Zhou, S. Xue, Y. Xue, Y. Liao, J. Liu, and W. Zhao, "A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning," *Energy*, vol. 224, p. 120118, 2021, doi: 10.1016/j.energy.2021.120118.

[5] Q. Zhou *et al.*, "Modified Particle Swarm Optimization with Chaotic Attraction Strategy for Modular Design of Hybrid Powertrains," *IEEE Trans. Transp. Electrif.*, vol. 7782, no. c, 2020, doi: 10.1109/TTE.2020.3014688.

[6] A. Al Mamun, Z. Liu, D. M. Rizzo, and S. Onori, "An integrated design and control optimization framework for hybrid military vehicle using lithium-ion battery and supercapacitor as energy storage devices," *IEEE Trans. Transp. Electrif.*, vol. 5, no. 1, pp. 239–251, 2019, doi: 10.1109/TTE.2018.2869038.

[7] X. Lü *et al.*, "Energy management of hybrid electric vehicles: A review of energy optimization of fuel cell hybrid power system based on genetic algorithm," *Energy Convers. Manag.*, vol. 205, no. January, p. 112474, 2020, doi: 10.1016/j.enconman.2020.112474.

[8] G. Du, Y. Zou, X. Zhang, T. Liu, J. Wu, and D. He, "Deep reinforcement learning based energy management for a hybrid electric vehicle," *Energy*, vol. 201, p. 117591, 2020, doi: 10.1016/j.energy.2020.117591.

[9] Y. Zhou, A. Ravey, and M. C. Péra, "Real-time cost-minimization power-allocating strategy via model predictive control for fuel cell hybrid electric vehicles," *Energy Convers. Manag.*, vol. 229, no. December 2020, 2021, doi: 10.1016/j.enconman.2020.113721.

[10] S. Zhang, R. Xiong, and F. Sun, "Model predictive control for power management in a plug-in hybrid electric vehicle with a hybrid energy storage system," *Appl. Energy*, vol. 185, pp. 1654–1662, 2017, doi: 10.1016/j.apenergy.2015.12.035.

[11] A. A. Apolinarska *et al.*, "Robotic assembly of timber joints using reinforcement learning," *Autom. Constr.*, vol. 125, p. 103569, 2021, doi: 10.1016/j.autcon.2021.103569.

[12] C. Lu, H. Wang, C. Lv, J. Gong, J. Xi, and D. Cao, "Learning Driver-Specific Behavior for Overtaking: A Combined Learning Framework," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 6788–6802, 2018, doi: 10.1109/TVT.2018.2820002.

[13] N. Zeng *et al.*, "Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip," *Neurocomputing*, vol. 425, pp. 173–180, 2021, doi: 10.1016/j.neucom.2020.04.001.

[14] R. Z. Liu *et al.*, "Efficient Reinforcement Learning for StarCraft by Abstract Forward Models and Transfer Learning," *IEEE Trans. Games*, vol. 14, no. 8, pp. 1–14, 2021, doi: 10.1109/TG.2021.3071162.

[15] W. Li *et al.*, "Deep reinforcement learning-based energy management of hybrid battery systems in electric vehicles," *J. Energy Storage*, vol. 36, no. June 2020, 2021, doi: 10.1016/j.est.2021.102355.

[16] T. Liu, X. Hu, S. Li, and D. Cao, "Reinforcement Learning Optimized Look-Ahead Energy Management of a Parallel Hybrid Electric Vehicle," *IEEE/ASME Trans. Mechatronics*, vol. 4435, no. c, pp. 1–1, 2017, doi: 10.1109/TMECH.2017.2707338.

[17] T. Liu, X. Hu, W. Hu, and Y. Zou, "A Heuristic Planning Reinforcement Learning-Based Energy Management for Power-Split Plug-in Hybrid Electric Vehicles," *IEEE Trans. Ind. Informatics*, vol. PP, no. 3, pp. 1–1, 2019, doi: 10.1109/TII.2019.2903098.

[18] T. Liu, Y. Zou, D. Liu, and F. Sun, "Reinforcement Learning of Adaptive Energy Management with Transition Probability for a Hybrid Electric Tracked Vehicle," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7837–7846, 2015, doi: 10.1109/TIE.2015.2475419.

[19] Q. Zhou *et al.*, "Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle," *Appl. Energy*, vol. 255, no. 2019, 2019, doi: 10.1016/j.apenergy.2019.113755.

[20] Q. Zhou, D. Zhao, B. Shuai, Y. Li, H. Williams, and H. Xu, "Knowledge Implementation and Transfer With an Adaptive Learning Network for Real-Time Power Management of the Plug-in Hybrid Vehicle," *IEEE Trans. Neural Networks Learn. Syst.*, no. July, 2021, doi: 10.1109/TNNLS.2021.3093429.

[21] Z. Zhu, N. Pivaro, S. Gupta, A. Gupta, and M. Canova, "Safe Model-based Off-policy Reinforcement Learning for Eco-Driving in Connected and Automated Hybrid Electric Vehicles," pp. 1–14, 2021, [Online]. Available: http://arxiv.org/abs/2105.11640.

[22] G. Du, Y. Zou, X. Zhang, L. Guo, and N. Guo, "Heuristic Energy Management Strategy of Hybrid Electric Vehicle Based on Deep Reinforcement Learning with Accelerated Gradient Optimization," *IEEE Trans. Transp. Electrif.*, vol. 7782, no. c, pp. 1–1, 2021, doi: 10.1109/tte.2021.3088853.

[23] R. Zou, L. Fan, Y. Dong, S. Zheng, and C. Hu, "DQL energy management: An online-updated algorithm and its application in fix-line hybrid electric vehicle," *Energy*, vol. 225, p. 120174, 2021, doi: 10.1016/j.energy.2021.120174.

[24] Y. H. Wang, T. H. S. Li, and C. J. Lin, "Backward Q-learning: The combination of Sarsa algorithm and Q-learning," *Eng. Appl. Artif. Intell.*, vol. 26, no. 9, pp. 2184–2193, 2013, doi: 10.1016/j.engappai.2013.06.016.

[25] Q. Zhou *et al.*, "Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle," *Appl. Energy*, vol. 255, 2019, doi: 10.1016/j.apenergy.2019.113755.

[26] B. Shuai *et al.*, "Heuristic action execution for energy efficient charge-sustaining control of connected hybrid vehicles with model-free double Q-learning," *Appl. Energy*, vol. 267, no. November 2019, 2020, doi: 10.1016/j.apenergy.2020.114900.

[27] Y. Ding, "The Impact of Learning Rate Decay and Periodical Learning Rate Restart on Artificial Neural Network," pp. 6–14.

 