

UNIVERSITY OF
BIRMINGHAM

REAL-TIME ENERGY MANAGEMENT OF THE HYBRID OFF-HIGHWAY VEHICLE USING REINFORCEMENT LEARNING

BY

BIN SHUAI

A THESIS SUBMITTED TO
THE COLLEGE OF ENGINEERING AND PHYSICAL SCIENCE OF
THE UNIVERSITY OF BIRMINGHAM

DOCTOR OF PHILOSOPHY

Department of Mechanical Engineering
School of Engineering
College of Engineering and Physical Science of
The University of Birmingham

Feb 2022

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Rapid developments in transportation and mobility systems including off-highway vehicles have a significant impact on the environment and energy. Electrification of off-highway vehicles makes a significant contribution to decarbonisation with advancements in vehicle technologies. This Ph.D. research focuses on the development of an advanced energy management strategy (EMS) for the hybrid off-highway aircraft-towing tractor, harnessing reinforcement learning (RL), one of the most promising artificial intelligence technologies. Based on real-time modelling of the vehicle system using MATLAB/Simulink, the dedicated RL algorithms are developed to enhance the exploration capability and reduce the overestimation of the reward-action pairs so that they can improve the energy efficiency of the hybrid vehicle in learning cycles and real-world driving. The real-time control functionalities of the proposed energy system are tested and verified on the hardware-in-the-loop platform.

By introducing three decay functions to the epsilon-greedy process of the RL-based EMS, the impacts of the exploration-to-exploitation ratio (E2E) on improving vehicle energy efficiency are investigated. A step-based decay function is proposed as a unified decay function incorporated with the vehicle EMS. Its contribution to the improvement of trustworthiness is demonstrated by a comparison with standard Q learning. An improvement rate of up to 6.21% higher than the standard Q-learning is

achieved.

Based on the performance analysis of the RL-based EMS with different decay functions, information-fused and synchronised ensemble learning frameworks are developed to enhance the optimality and robustness of the RL-based EMS. The random-based, maximum-based and weighted-based methods are studied and compared under the ensemble learning frameworks. The study suggests that the weighted-based method is the best of the three ensemble methods. This research leads to a new synchronised ensemble learning framework founded on the weighted-based method, with up to 1.06% and 5.06% higher vehicle energy efficiency than the maximum-based method and random-based method, respectively. Additionally, the weighted-based method achieves 1.15% higher vehicle energy efficiency than the standard Q-learning.

Finally, by observing overestimation phenomena in the RL-based EMS under disturbed learning and undisturbed learning scenarios, a predictive double Q-learning framework with backup models and random action execution policy, respectively is developed and used to increase the efficiency of the learning experience and determine the final action to improve the robustness of the RL-based EMS for the hybrid off-highway vehicles.

Keywords: Hardware-in-the-Loop (HiL), Reinforcement Learning (RL), Energy Management Strategy (EMS), Hybrid Electric Vehicle (HEV), Real-time Simulation

ACKNOWLEDGEMENTS

First of all, I would like to thank my lead supervisor, Prof. Hongming Xu, for giving me an opportunity to follow my dream of obtaining a Ph.D. degree and providing excellent guidance throughout my Ph.D. study. Thank you for your time, patience, and advice in teaching me physically and mentally over the past four years. Your passion for research and life is inspirational and working alongside you has heightened my enthusiasm for science and technology. My thanks to you that you introduced many great scientists & friends to me. I hope we could have many opportunities to work together and enjoy drinking beers in the future. Additionally, I would like to express my sincere gratitude to Dr Quan Zhou, who taught me to courageously move forward to overcome the difficulties of scientific research. And He constantly shared his experience with me on seeking innovation to help me avoid shoddy work. I am deeply impressed by his sunny life philosophy and broad, which will surely be in my future life trajectory.

I am very grateful for the contribution of the world-leading collaborators with whom I had the fortune to work together during my Ph.D. research. I hope we could still have a long successful partnership in the future. Thanks to the staff in the university, thank you for maintaining the facilities and expert generosity assistance in giving me ‘free reign’ to finish my experiments in the laboratory.

Special thanks to group fellow for ensuring my daily life is constantly active in Birmingham. We have a great time travelling, exercising, and having meals. I want to deliver my sincere thanks to my housemates, Dr Yinglong He and Mr Zeyu Sun. Also, I want to extend my gratitude to my closest friend, Dr Haoye Liu. We spent many happy moments playing badminton and video games. To Mr Cetengfei Zhang, Dr Chongming Wang, Dr Ji Li, Dr Ziyang Li and Dr YunFan Zhang, you guys shared a lot of research experience and gave me many suggestions from the perspective industry. Thank you, Mr Fanggang Zhang and Mr Chengqing Wen, for bringing me so much fun when I felt stressed, and we went to many places in the UK together to get stressed out.

Special thanks to my colleague, Miss Min Hua, for helping me clarify my thesis logic, patiently listening to me to talk about my thesis repeatedly, and consoling me when I feel down and anxious during the latter part of my PhD progress.

Also, I would like to deliver my heartfelt thanks to my girlfriend, Dan Luo. Thanks for your accompany and support in the UK for two years in my PhD study.

Last but certainly not least, I want to dedicate this thesis to my warmly family. Thanks to my parents for instilling in me the work ethic necessary to finish this long journey. Thank you for motivating me during the tough times and for your financial support, which has kept my head above water through the years.

Finally, I would like to thank those underappreciated staff who manage the day-to-day running of the mechanical engineering building and provide a fantastic service to the university. Their collective presence and enthusiasm gave me to experience during my stay at the University of Birmingham.

Bin Shuai

Feb. 2022

TABLE OF CONTENT

Abstract I

Acknowledgements III

Table of Content VI

List of Publications i

List of Abbreviations iii

List of Symbols vi

List of Figures x

List of Tables xv

1 Introduction 1

1.1 Research background 1

1.1.1 Urgent demand for transportation electrification 1

1.1.2 The application of artificial intelligence in electric vehicles 13

1.2 Motivations 17

1.3 Objectives 21

1.4 Thesis outline 22

2 Literature review 26

2.1 Modelling and validation method for HEVs 26

2.1.1 Modelling approaches of powertrain systems 27

2.1.2 Validation approaches of powertrain systems 29

2.2 Classification of EMSs 33

2.2.1 Offline EMSs for HEVs 34

2.2.2 Online EMSs for HEVs 38

2.2.3 Learning-based EMSs for HEVs 40

2.3 Recent development of Q-learning algorithm and its application of EMSs for HEVs 44

2.4 Challenges 48

2.5 Summary 49

3 Research methodology and facilities 50

3.1 Research methodology 50

3.2 Benchmark EMS 52

3.2.1 Dynamic programming 52

3.2.2 Standard Q-learning algorithm 55

3.3 Research platform 58

3.3.1 Driving Cycles 58

3.3.2 Modelling of the series powertrain system and subsystems 63

3.3.3 HiL testing platform 70

3.3.4 Software for the HiL testing 72

3.4 Summary 73

4 Exploration-to-exploitation ratio in reinforcement learning control 74

4.1 Problem formulation 74

4.1.1 Reward function 75

4.1.2 Exploration to exploitation ratio 76

4.2 Framework of the model-free supervisory control system 79

4.2.1 States perception module 80

4.2.2 Action execution module 80

4.2.3 Reward assessment module 82

4.3 Three decay functions 83

4.4 Results and discussion 86

4.4.1 Impact of different reward functions 87

4.4.2 Learning performance of RL-based EMS with three decay methods 89

4.4.3 Different initial epsilon values 91

4.4.4 Real-time performance 93

4.4.5 Robustness testing 94

4.5 Summary 95

5 Ensemble learning with E2E ratio in RL-based energy management 97

5.1 Problem statement 97

5.2 Supervisory ensemble learning-based system 98

5.2.1 Learning automata module 99

5.2.2 Synchronised-based ensemble learning 100

5.2.3 Information fusion-based ensemble learning 104

5.3 Performance and discussions 107

5.3.1 Improvement in energy efficiency 107

5.3.2 Real-time performance 114

5.3.3 Robustness testing 115

5.4 Summary 117

6 Reduction of overestimation in RL-based energy management control 119

6.1 Problem statement 119

6.1.1 Recording learned experience in two knowledge bases 120

6.1.2 Action execution in conventional double Q-learning 121

6.2 Double Q-learning module in layered control framework 121

6.2.1 Action execution policies for double Q-learning..... 122

6.2.2 Double Q-learning with backup models..... 125

6.3 Results and discussions..... 129

6.3.1 Learning performance of double Q-learning methods..... 129

6.3.2 Real-time operation of double Q-learning methods..... 139

6.3.3 Robustness testing 144

6.4 Summary 148

7 Conclusion and future Work 150

7.1 Conclusions 150

7.2 Future work 153

8 Reference..... 154

LIST OF PUBLICATIONS

Journal Papers (6)

1. ***Bin Shuai***., Quan Zhou, Ji Li, Yinglong He, Ziyang Li, Huw Williams, Shijin Shuai, "Heuristic action execution for energy efficient charge-sustaining control of connected hybrid vehicles with model-free double Q-learning." Applied energy 267 (2020): 114900. (IF: 9.746)
2. ***Bin Shuai***, Yanfei Li, Quan Zhou, Hongming Xu, Shijin Shuai, "Model-free supervisory control of the off-highway hybrid electric vehicle for fuel economy improvement using predictive double Q-learning with backup models" (Accepted by the Journal of Central South University of Technology). (IF:1.733)
3. ***Bin Shuai,*** Quan Zhou, Min Hua, Cetengfei Zhang, Huw Williams, Yanfei Li, Shijin Shuai, Lun Hua, Hongming Xu, "An ensemble reinforcement learning with optimal exploration-to-exploitation ratio for energy management of hybrid vehicles". (To be submitted the Applied Energy.) (IF: 9.746)

4 Quan Zhou**,** Ji Li, ***Bin Shuai***, Yinglong He, Ziyang Li, Yunfan Zhang, Hongming Xu. "Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle." Applied Energy 255 (2019): 113755. (IF: 9.746)

5 Ji Li, Quan Zhou, Ziyang Li, Yunfan Zhang, Yinglong He, ***Bin Shuai***, Hongming Xu. "Dual-loop online intelligent programming for driver-oriented predict energy management of plug-in hybrid electric vehicles." Applied Energy 253 (2019): 113617. (IF: 9.746)

6 Quan. Zhou, Dezhong. Zhao, ***Bin Shuai***, Yanfei. Li, Huw Williams, Hongming. Xu, "Knowledge implementation and transfer with an adaptive learning network for real-time power management of the plug-in hybrid vehicle." IEEE Transactions on Neural Networks and Learning Systems 32.12 (2021): 5298-5308. (IF:10.451)

Conference Papers and Presentation (2)

1. ***Bin Shuai***, Yanfei Li, Quan Zhou, Huw Williams, Hongming Xu , Shijin Shuai. “Model-free supervisory engine control in the heavy-duty plug-in hybrid electric vehicle using predictive double Q-learning with backup models”; The 2nd world congress on internal combustion engines (WICE 2021). 21-24 April 2021, Jinan, China. (Best paper)
2. ***Bin Shuai***, Yanfei Li, Lun Hua, Quan Zhou, Huw Williams, Hongming Xu, “Impact of exploration-to-exploitation ratio on energy saving potential of plug-in hybrid vehicle controlled by reinforcement learning” Applied Energy Symp. 2021 Low carbon cities urban energy systems, 2021. (Invited to submit Applied Energy of CUE2021)

LIST OF ABBREVIATIONS

APC	Advanced Propulsion Centre
AI	Artificial Intelligence
BEV	Battery Electric Vehicle
CO2	Carbon Dioxide
CAGR	Compound Annual Growth Rate
CT	Control Tower
DQL-MEP	Double Q-learning with maximum based execution policy
DQL-REP	Double Q-learning with random execution policy
DQL-CON	Double Q-learning with conventional policy
DQN	Deep Q Network
DDPG	Deep Deterministic Policy Gradient
DRL	Deep Reinforcement Learning
ECU	Electric Control Unit
EGU	Engine Generator Unit
EF	Equivalent Factor
FCEV	Fuel Cell Electric Vehicle

GHG	Green House Gas
HEV	Hybrid Electric Vehicle
HiL	Hardware-in-the-loop
IEA	International Energy Agency
ICE	Internal Combustion Engines
IoT	Internet-of-Things
IFE	Information fusion-based ensemble learning framework
LAM	Learning Automaton Module
ML	Machine Learning
MC	Monte Carlo
MPC	Model Predictive Control
MDP	Markov Decision Process
NN	Neural Network
NOx	Nitrogen Oxides
OEM	Original Equipment Manufacturer
PM	Particulate Matters
PBDC	Push Back Driving Cycle

PNGV	Partnership for a New Generation Vehicle
PHEV	Plug-in Hybrid Electric Vehicle
PDQL	Predictive double Q-learning method with backup models
RL	Reinforcement Learning
RSU	Roadside Unit
RDE	Real-world Driving Emission
R&D	Research and Development
SiL	Software-in-the-loop
SoC	State-of-Charge
SGD	Stochastic Gradient Descent
SYE	Synchronized-based ensemble learning framework
SQL	Standard Q-learning
TD	Temporal Difference
V2X	Vehicle-to-Everything

LIST OF SYMBOLS

a_{acc}	Vehicle acceleration
a_{final}	Final action from two Q-tables
A	Vehicle fronted area
a_A	The selected action from Q-table A
a_B	The selected action from Q-table B
C_1	Short transient capacity
C_2	Long transient resistance
C	A random value
C_d	Aerodynamic coefficient
F	A factor control the epsilon value drops
F_t	Driving Force
F_t	Grade resistance
F_f	Rolling resistance
F_w	Aerodynamics resistance
F_j	Acceleration resistance
g	Gravity coefficient

H_f	Fuel heat value
I_{bc_cmax}	Rated battery cell charge current
I_{bc_dmax}	Rated battery cell discharge current
I_{bc}	Battery cell current
I_{batt}	Current of battery package
P_{fuel}	Power of fuel
$P_{loss,trm}$	Power loos of the traction motor
P_{ef}	Equivalent power fuel
P_{bp}	Output power of battery pack
P_{trm}	Output power from tranction motor
$P_{egu,max}$	Maximum power of engine-generator
$P_{bp,max}$	Maximum power of battery package
P_{dem}	Driver's power demand
$P_{loss_{tot}}$	Total power loss
$P_{loss_{egu}}$	Power loss of engine-generator
$P_{loss_{bp}}$	Power loss of battery pack
P_{egu}	Engine-generator output power
Q_{bc}	Capacity of battery cell

Q_{c_max}	Capacity of battery cell
Q_A	Knowledge base A (Q-table A)
Q_B	Knowledge base B (Q-table B)
R_0	Battery cell internal resistance
R_1	Short transient resistance
R_2	Long transient resistance
SoC_{now}	Current battery state-of-charge
SoC_{past}	Past battery state-of-charge
$SoC_{initial}$	Initial state-of-charge of battery
T_{trm}	Traction motor torque
T_{egu}	Torque of engine-generator
U_{bc}	Voltage provided by battery cell
V_{bc_max}	Rated battery cell voltage
V_{oc}	Open voltage
v_f	Rate of fuel consumption
ψ	Mass factor
n_{trm}	Rotation speed of traction motor
$\dot{m}_f$	Fuel consumption rate

ρ_f	Fuel density
η_{egu}	The energy efficiency of EGU
n_{bc}	Number of battery cell
θ	Epsilon greedy value
π_{exe}	Execution policy
u_{egu}	Control signal to engine-generator
$\dot{m}_f$	Fule rate kg/s
ω	A random number
θ_1	Initial epsilon value
π_{max}	Maximum-based action execution policy
π_{Random}	Random-based action execution policy
π_{Weight}	Weighted-based action execution policy
η_{EXD}	Vehicle energy efficiency using the EXD
η_{Com}	Vehicle energy efficiency using the WBD or IFBS
η_{final}	The vehicle energy efficiency at the 1st learning interation
η_{final}	The final vehicle energy effieicny
η_{veh}	Vehicle energy efficiency

LIST OF FIGURES

Fig.1-1 Overall UK GHG emissions keeping a downward trend (International Energy Agency, 2019) 2

Fig.1-2 Global CO2 emission rebound by nearly 5% in 2021, approaching the 2018-2019 peak (Global Energy Review 2019, 2020) 4

Fig.1-3 Global off-highway electric market (Recent trends and demands through 2030, 2022) 8

Fig.1-4 The off-highway hybrid vehicles market in application type (Mordor intelligence, 2019).....10

Fig.1-5 Classification of HEVs according to the hybridisation degrees.....11

Fig.1-6 One case of the series hybrid electric powertrain (Zhou et al , 2017).....12

Fig.1-7 Aircraft towing operation scenario.....18

Fig.1-8 A connected vehicle system (Shuai et al , 2020).....19

Fig.1-9 The main research content of this thesis and the structure of each chapter..23

Fig.2-1 Forward-looking modelling approach (Huang et al , 2018).....28

Fig.2-2 Backward-looking modelling approach (Huang et al , 2018)…….29

Fig.2-3 The workflow of designing new functions of the ECU (Schaffnit et al , 2000)……30

Fig.2-4 Three common validation approaches (Schaffnit et al , 2000)……..31

Fig 2-5 Online rapid control prototyping testing platform (Zhou et al , 2022)……..31

Fig.2-6 Hardware-in-the-loop test platform (Wang et al , 2020)……...32

Fig.2-7 Driver-in-the-loop testing platform (Li et al , 2019)……33

Fig.2-8 Classification of EMSs for HEVs…….34

Fig.3-1 The research methodology for a real-time energy management system….51

Fig.3-2 Discretisation of the state and time space in DP……...53

Fig.3.3 The calculation process of DP algorithm……..54

Fig.3-4 The interaction process of RL…….56

Fig.3-5 Control architecture based on Q-learning algorithm for the studied off-highway vehicle……..57

Fig. 3-6 The Q-learning algorithm based on MDP (Zhou et al , 2019)…….58

Fig.3-7 The signal flow of the studied hybrid aircraft towing tractor…….63

Fig.3-8 Motor power map……..66

Fig.3-9 The two RC equivalent circuit model for a battery cell……...69

Fig.3-10 HiL testing platform (Shuai et al , 2020)……71

Fig.3-11 Software for HiL testing system (Zhou et al , 2019)…….72

Fig.4-1 The exploration and exploitation process in RL…….....77

Fig.4-2 One example of exploration to exploitation……78

Fig.4-3 Layered control framework for model-free supervisory control (Shuai et al ,2020)…….....79

Fig.4-4 The variation of epsilon value the exponential decay function (initial $\theta_1 = 0.8 \text{ and } 0.5$)……..84

Fig.4-5 The variation of epsilon value with step-based decay function (initial θ = **0.8**, ***D*** = **5** and **10**)……..85

Fig.4-6 The variation of epsilon value with reciprocal function (initial θ_1=**0.8** and **0.5**)……...86

Fig.4-7 Learning performance of two reward functions……88

Fig.4-8 Vehicle system efficiency curves obtained by the RL algorithms with three decay functions (initial battery SoC is 50%) ……...90

Fig.4-9 Value of E2E ratios over the different epsilon values and rounds of learning (the initial value of θ is 0.8)……..91

Fig.4-10 Real-time performance of RL-based EMSs with different decay functions (initial battery SoC = 50%)…….94

Fig.5-1 Learning automata working process……..99

Fig.5-2 The framework of synchronised-based ensemble learning system…….100

Fig.5-3 The first step of the information fusion-based ensemble learning……...104

Fig.5-4 The second step of the information fusion-based ensemble learning system……106

Fig.5-5 Learning performance of three competing methods……..108

Fig.5-6 The learning performance of four model-free control methods…….112

Fig.5-7 Real-time performance of two model-free methods……...115

Fig.6-1 Two problems in conventional double Q-learning……..…120

Fig.6-2 The two loops in double Q-learning with backup models……..……127

Fig.6-3 The working diagram of the predictive double Q-learning method with backup models……...128

Fig.6-4 Learning performance of three control strategies……...130

Fig.6-5 Vehicle's energy efficiency obtained in undisturbed learning and disturbed learning a) conventional Q-learning; b) double Q-learning method with the max-value-based execution policy; c) double Q-learning method with the random execution policy……133

Fig.6-6 Evolution of vehicle energy efficiency obtained by three model free method……...138

Fig.6-7 HiL testing results of three double Q-learning methods……141

Fig.6-8 HiL testing results of DQL-REP and SQL, and the offline results of DP…...142

Fig.6-9 Energy flow in driving cycle 1: (a) initial SoC = 25%; and (b) initial SoC = 30%.....143

Fig.6-10 HiL test results of three control methods under PBDC-1……144

LIST OF TABLES

Table. 1-1 Countries for banning the sale of ICE powered vehicles 5

Table. 1-2 The main architecture of the hybrid electric vehicle system..... 12

Table. 1-3 RL-based EMS for hybrid electric vehicles 15

Table. 3-1 Cycle component profile 59

Table. 3-2 Detailed profiles of all driving cycles..... 61

Table. 3-3 Parameters of studied vehicle 65

Table. 3-4 Engine generator unit profile..... 67

Table. 3-5 Parameters of the studied battery cell 70

Table. 4-1 Vehicle energy efficiency for the two reward equation at different numbers of iterations..... 88

Table. 4-2 Different initial value for E2E with decay functions 92

Table. 4-3 Investigation of $\theta = 0.8$ under PBDC-2 and PBDC-3 with an initial battery SoC of 50% 93

Table. 4-4 Real time performance of E2E with the EXD, RBD and SBD under three predefined driving conditions 95

Table. 5-1 Vehicle energy efficiency obtained by ensemble RL-based EMSs with different determining factors 109

Table. 5-2 Different proportions of two parallel agents 111

Table. 5-3 The vehicle energy efficiency at different learning rounds 113

Table. 5-4 Comparison of two ensemble learning systems 114

Table.5-5 Performance of the EXD and WBD methods under PBDC 2-4..... 118

Table. 6-1 Stability of learning at stage two with different deciding factor values 138

Table. 6-2 Correlation of the results and performance of the models with the deciding variable 136

Table.6-3 Mean efficiency and variance of 20 samples 137

Table.6-4 Comparison of different model-free control methods in total time..... 139

Table.6-5 Performance of double Q-learning method and standard Q-learning method……...146

Chapter One

INTRODUCTION

Off-highway vehicles (e.g., aircraft towing tractors) work in very different conditions compared to highway vehicles (e.g., passenger cars) and the electrification of off-highway vehicles can lead to significant energy saving and emissions reduction if the vehicles are controlled properly. This thesis focuses on the energy management control of a hybrid off-highway vehicle. This chapter presents the research background, motivation, objectives and outline of the thesis.

1.1 Research background

1.1.1Urgent demand for transportation electrification

The rapid development of the transportation industry has raised the requirements for energy sources (Li *et al.*, 2020,Li *et al.*, 2020; Su *et al.*, 2021; Tran *et al.*, 2021). A significant proportion of energy sources comes from the commonly used fossil energy, which is one of the key contributors for global climate change (Aolou *et al.* 2019; Aldhafeeri *et al.*, 2020; Peksen *et al.*, 2021). Since 1970s, the global emissions from fossil energy combustion have increased by 90%, which have led to excessive

greenhouse gas (GHG) emissions. Many countries have set their ambitious target for reducing GHG emissions as a national priority to mitigate the global warming (Du *et al.*, 2018; Li *et al.*, 2021; Shamsi *et al.*, 2021). For instance, the UK's total GHG emissions were over 550 million tonnes of carbon dioxide (CO2) equivalent in 2019 with a decrease of approximately 3% from the previous year's emissions, as shown in Fig.1-1.

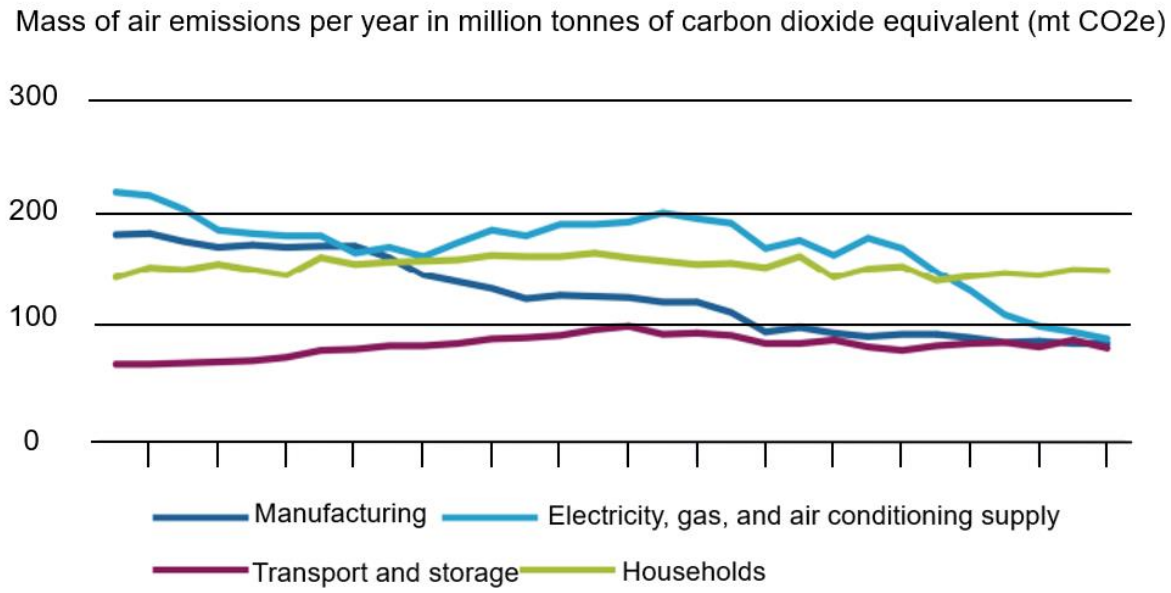


Fig. 1-1 Overall UK GHG emissions keeping a downward trend (International Energy Agency, 2019)

The top four contributors to the UK's GHG consist of households and the electricity (plus gas, steam, and air conditioning supply), manufacturing, and transport & storage sectors in 2019, accounting for over 72% of the total emissions for the UK. However, there is much potential for improvement with the developments and advancements of intelligent and electrified transport technologies. Consequently,

vehicles with internal combustion engines (ICE) can be gradually replaced with green powers, which use renewable energies that not only improve the environment without producing GHG, but also reduce fuel consumption (Archer *et al.*, 2018).

According to the latest collected data from the International Energy Agency (IEA), the global energy-related CO2 emissions is heading for their second-largest annual increase in 2021, as shown in Fig 1-2. The global CO2 emissions decreased by 5.8% in 2020, the biggest drop, and almost five times decline in 2009 due to the global financial crisis. The decline of CO_2 emissions in 2020 was due to a hard acquisition of oil and coal demands during the Covid-19 pandemic. In 2021 global energy-related CO2 emissions are projected to rebound and grow by 4.8% since the demand for coal, oil and gas rebounds. The increase of over 1500 Mt CO2 would be the largest single increase; it leaves global emissions in 2021 at around 400 Mt CO2, or 1.2%, below the 2019 peak(‘Global Energy Review 2019’, 2020).

The transportation electrification, as an essential contributing factors, has become one of the promising alternatives to ICE-powered vehicles to lessen the GHG and energy crisis (Huang *et al.* 2021). In this regard, searching for safe, clean, and highly efficient alternative solutions is the goal for automobile manufacturers to integrate carbon emission reduction technologies into ICE-powered vehicles (Tran *et al.*, 2020). The impact of ICE-powered vehicles on climate change due to emissions and the depletion of fossil fuels has contributed to the progress of electrified powertrains.

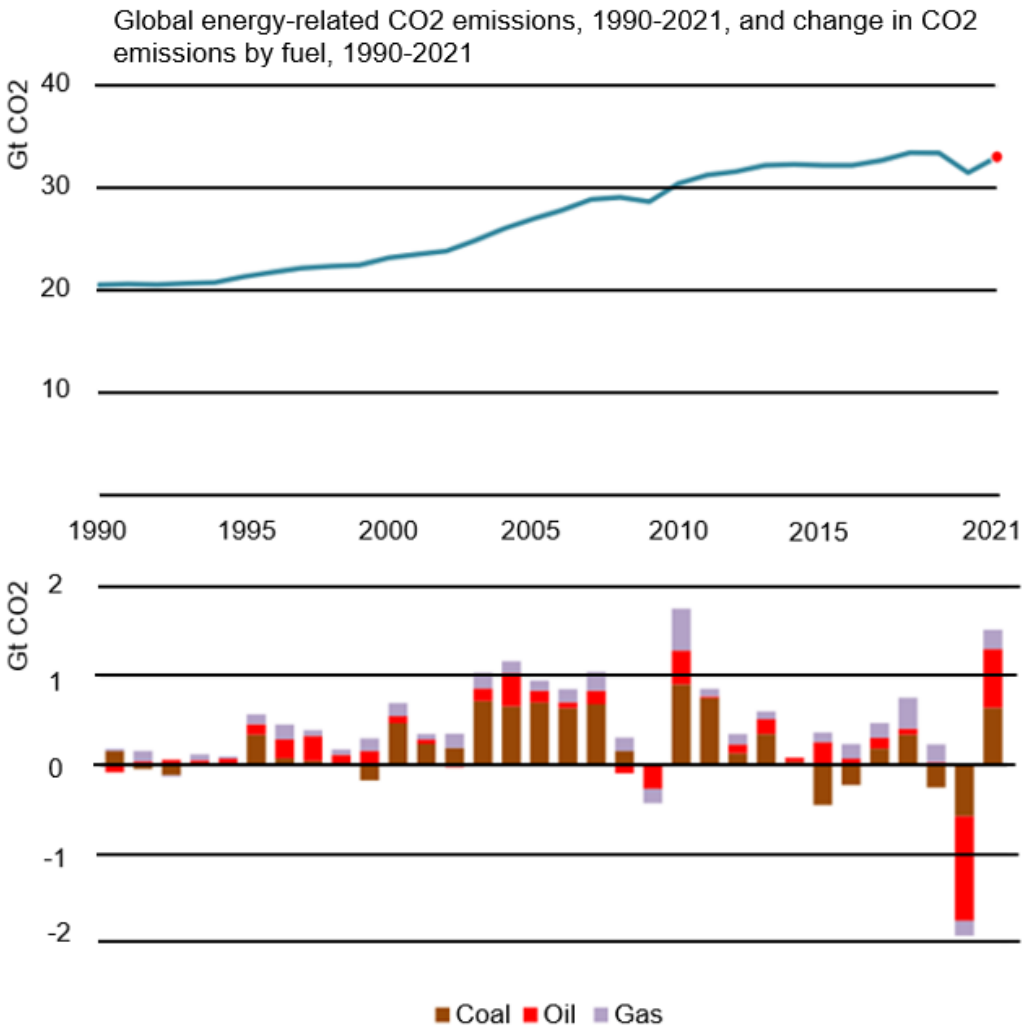


Fig. 1-2 Global CO2 emissions rebound by nearly 5% in 2021, approaching the 2018-2019 peak (Global Energy Review 2019, 2020)

To decrease the resource consumption and environment damage, the automotive industry is undergoing a huge change from traditional fossil fuel vehicles to new energy vehicles, such as battery electric vehicles (BEVs), hybrid electric vehicles (HEVs) and fuel cell electric vehicles (FCEVs). Countries in Europe, America, China have formulated a series of regulations and policies to promote the development of

new energy vehicles successively. In November 2017, the European Commission issued a proposal to update CO2 emission standards with a new target of a 15% reduction in CO2 emissions per kilometre for new passenger cars and light commercial vehicles by 2025 (Europen Commission, 2019). Besides the EU, a number of other governments have announced their own timetables for fuel vehicle bans, as shown in Table 1-1, to achieve the goal with more and more new energy vehicles.

Table.1-1 Countries for banning the sale of ICE-powered vehicles

Country	2025	2030	2032	2040	2045
France				○	
Ireland		○			
Netherlands		○			
Norway	○				
Slovenia		○			
Sri Lanka				○	
Sweden					○
United Kingdom				○	

The legislations are also aimed at the off-highway vehicles. These vehicles contributed 27% of ground transport CO2 emissions and 5% of EU GHG emissions in 2016 (Gregor *et al* , 2019). The EU Commission has proposed that the average

CO2 emissions from off-highway vehicles in 2025 should be 15% lower than that in 2019, and a further 30% reduction should be achieved in 2030 (European Commission, 2017). These vehicles require high-quality performance with reliability and cost-effectiveness while facing other important restrictions, including low and zero-emission zones, running limits, urban circulation restrictions (APC, 2020). Especially in the urban areas, the indoor operation and airport-zoning have strict legislation with tailpipe pollutant requirements. For instance, the peak power requirement is relatively lower, thus, the battery power can substitute part of the power from the ICE. As a result, the application of new energy vehicles, especially HEVs in the area, is an ideal solution to meet the environment and the legislations demands.

The powertrain system of HEVs consists of an ICE and an electric motor. The motor depletes electricity energy stored in the battery to generate torque for driving, and it can also be used a generator for energy recovery from braking. Compared to conventional ICE-powered vehicles, HEVs can significantly enhance fuel economy and reduce emissions (Muhammad *et al*, 2021). Moreover, HEVs are firstly, less subject to battery-related issues due to their hybrid drive architecture with an engine, compared to BEVs. Secondly, the use of an ICE helps the HEVs significantly extend the range of mileage. Thirdly, the ICE operates in a highly efficient and clean area, resulting in a great reduction of the fuel consumptions and emissions. The low

emissions and low fuel consumption of the HEVs heavily depend on the applied EMS; which is used to optimise the energy distribution based on real-time ground information and make the engine and motor work simultaneously at the highest efficiency point (Johnson *et al*, 2000). Therefore, it is necessary to carry out a promising and practical research for optimising the hybrid drivetrains, such as the engine and electric machine working solely or simultaneously according to the real driving environment conditions of the vehicles.

a) Electric off-highway vehicles markets

Traditional off-highway vehicles are powered by diesel engines, especially the older models that can emit high levels of particulate matters (PM), nitrogen oxides (NOx), and other pollutants causing chronic disease and premature death, especially in urban areas with the high density of populations (Ouren, *et al*, 2007).. Therefore, the hybridization for off-highway vehicles is creating a new interest due to the driving power provided by multi-energy sources (Vandenhove *et al.*, 2020); which can enhance controllability and reduce GHG through advanced control strategies in real-world driving (Yang *et al.*, 2020). From the Fig. 1-3, The revenue of the off-highway electric vehicles market in 2030 will be $97,754.1 million, compared to an estimated $16,908.6 million in 2021, and it will witness a massive 21.5% CAGR between 2021 and 2030. The key reason behind it would be the technological advancements in electric propulsions, which have already been widely tested and adopted in on-road

vehicles (Recent trends and demands through 2030, 2022). And the off-highway vehicles registrations for energy consumption and emission reduction are increasing in China, Europe, and North America in recent years (INTERNATIONAL ENERGY AGENCY, 2021).

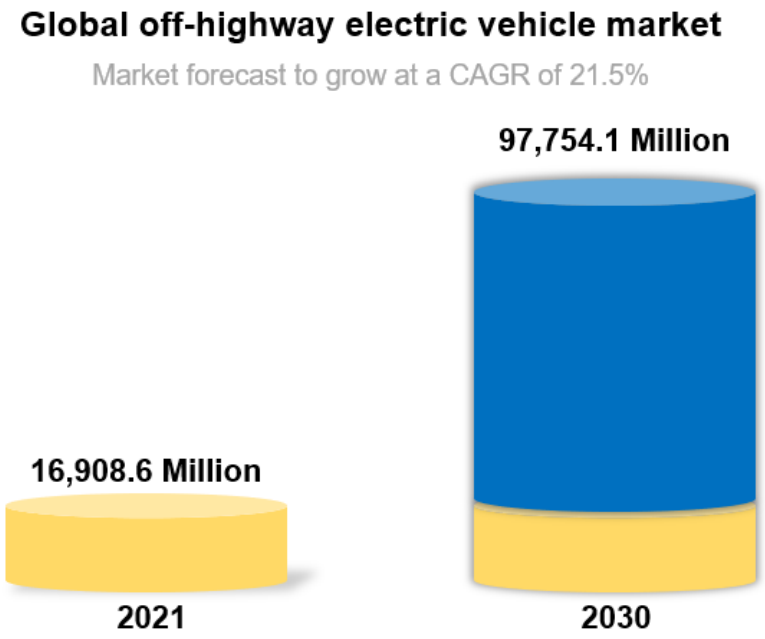


Fig. 1-3 Global off-highway electric vehicle market (Recent trends and demands through 2030, 2022)

From the aspect of the application as described in Fig.1-4, The construction segment accounted for the highest share of around 38.2% in 2019. The segment includes electric construction equipment, such as electric loaders, electric excavators, and more. One of the key factors driving the market demand for electric construction equipment is the growing investments by governments worldwide to build public infrastructure. In order to comply with the tougher emission requirements, OEMs are

introducing electric off-highway equipment. In the future, legislation could mandate zero emissions on the heavy-duty construction equipment to pave the way for OEMs to drop diesel-powered engine output and go all-electric (Mordor Intelligence, 2019). The main reason is attributed to the growth in the number of huge construction projects in Europe and the Asia-Pacific region (Mordor intelligence, 2018). The agriculture off-highway electric vehicle segment is expected to grow at a CAGR of 22.1%. The segment covers electric agricultural equipment such as electric tractors, electric harvesters, etc. Consumers have begun embracing new farm machinery and advanced agricultural processes to achieve higher yields in regions such as North America and Europe. Thus, the demand for electric and hybrid tractors and another farm equipment is expected to rise in the forecasted timeline. Then the Belt and Road established by China, which involves the construction of ports, roads, railways, airports and energy projects is expected to significantly propel the demand for hybrid construction vehicles and machineries since China is rapidly adopting electric and hybrid vehicles (Mordor intelligence, 2018). This is a major factor of market growth along with increased farm mechanization on a global level.

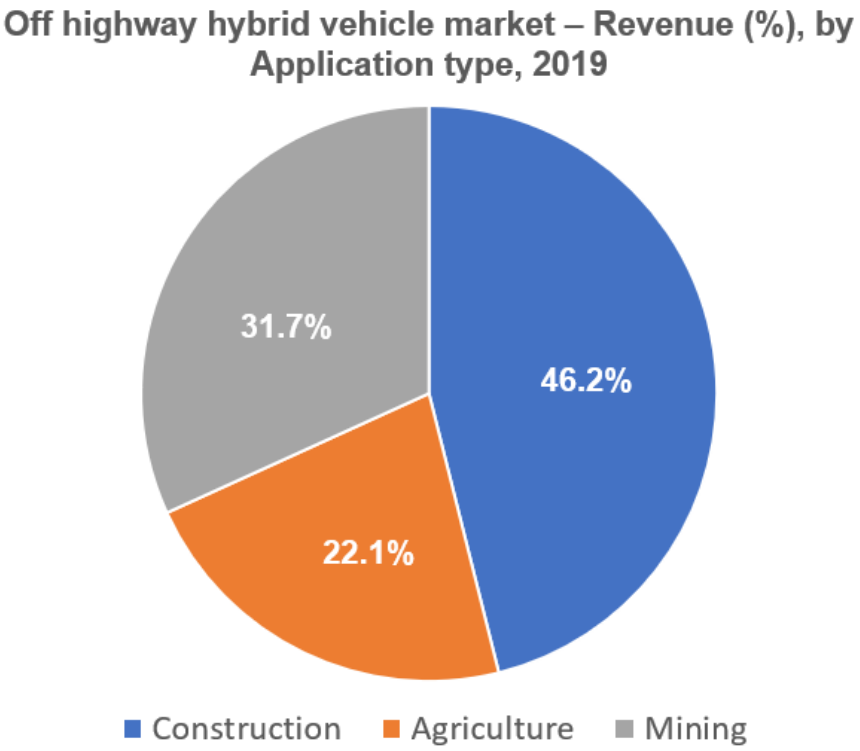


Fig. 1-4 The off-Highway hybrid vehicles market in application type (Mordor Intelligence, 2019)

b) Hybrid electric powertrain for off-highway vehicles applications

HEVs are in an important position during the transport electrification and there is a large investment of human, financial and material resources into the development of specified technologies for HEVs (Millo *et al*, 2011). The degree of hybridisation is the ratio of power by an electric motor in a hybrid vehicle to the total power consumed by the vehicle. The degree of the hybridisation for vehicles can be categorised as micro, mild, full HEVs, as shown in Fig.1-5.

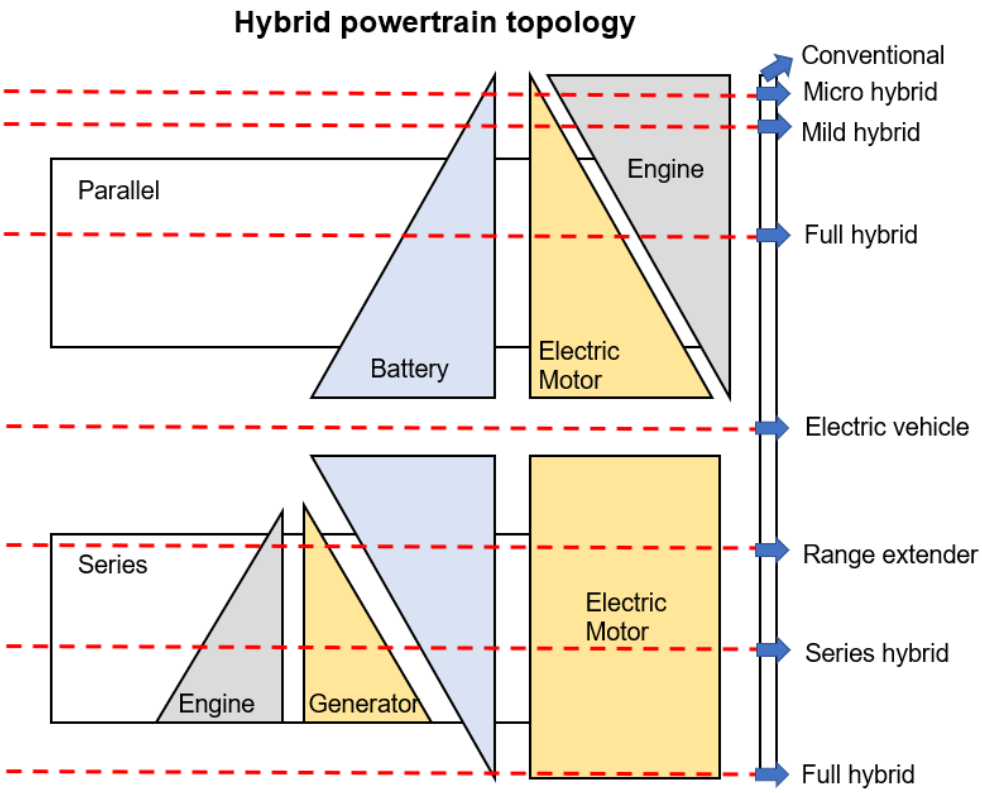


Fig. 1-5 Classification of HEVs according to the hybridisation degree

The architecture of the HEV is more complex than BEV and FCEV because multiple energy sources can be used to drive the vehicle. According to the different topology combination of mechanical traction and electric traction, the hybrid powertrain can be divided into three basic types: 1) series, 2) parallel and 3) multi-mode. The HEVs can be classified into series, parallel and multi-mode HEVs. The details for each architecture are shown in Table.1-2.

The series hybrid powertrain is commonly used for heavy trucks and tractors (Okui *et al.*, 2015; Zou *et al.*, 2016). It allows a more flexible allocation of traction motors and takes full advantage of electric drive in high-torque and low-speed domains. For

example, researchers from the University of Birmingham have developed an aircraft-towing tractor using a series hybrid powertrain, as shown in Fig.1-6. The tractor is powered by a tractor motor with electric energy generated by an 86.20 kW diesel engine and a battery pack made up to 8200 battery cells with NCR-18650 type (Zhou *et al.*, 2017).

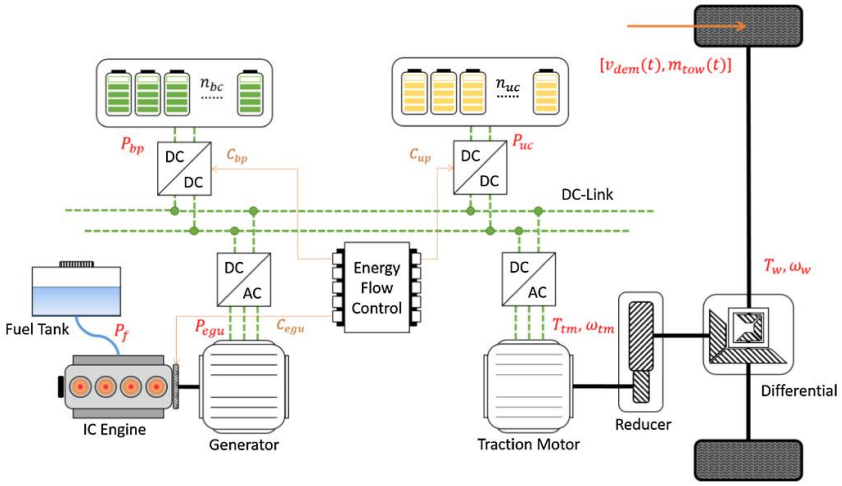
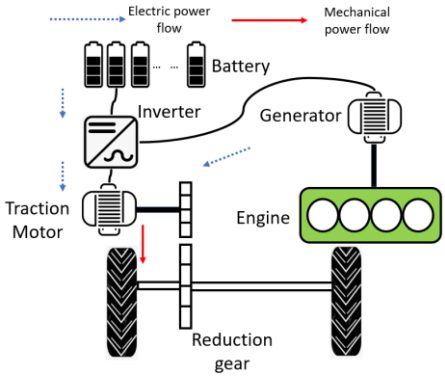
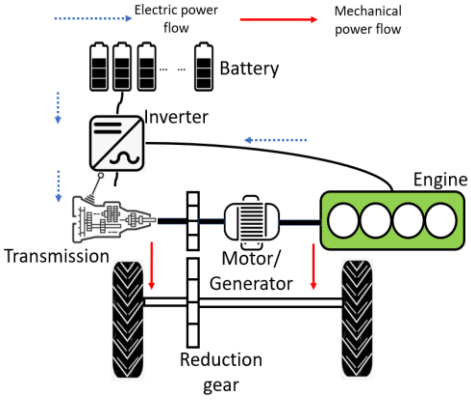
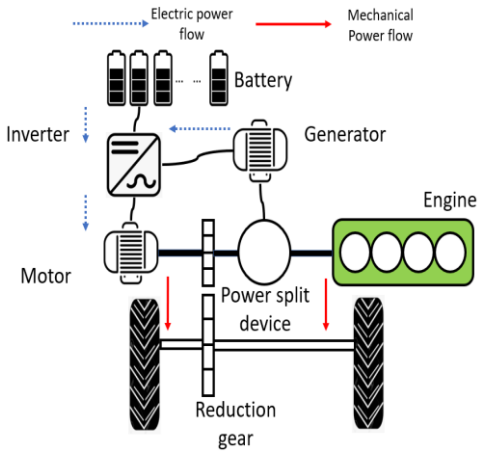


Fig. 1-6 One case of the series hybrid electric powertrain (Zhou *et al.*, 2017)

Table.1-2 The main architecture of the hybrid electric vehicle system

Architecture of HEV	Mainstream Diagram	Characteristics

The series		1) The engine-generator can operate at most efficient points. 2) Simple structure and easy control. 3) More energy loss and low efficiency in the two times of energy conversions.
The parallel		1) The ICE can directly drive the vehicle, reducing the loss caused by multiple energy conversion. 2) The dynamic vehicle performance can be improved, due to the electric motor and ICE can work together to drive vehicle. 3) The structure of the power system is complex and the control is difficult. Because there are mechanical connection between engine and electric motor.
The multi-mode		1) With more working modes, the ICE can always work in the high efficiency range. 2) More combinations of ICE, motor and generator to achieve much low fuel consumption and emissions. 3) The structure of the power system is more complex and the control is more difficult.

1.1.2The application of artificial intelligence in electric vehicles

Vehicle networking and automation harnessing artificial intelligence (AI) is another important trend in the automotive industry. In recent years, due to the great progress in machine vision, deep learning and other technologies, the development of advanced assisted driving and energy management control has been accelerated. At present, there are abundant studies focusing on vehicle perception, intelligent driving, and interaction among drivers, vehicles and environments, Different from conventional methods, the learning-based strategies in AI algorithms including neural-network-based and reinforcement-learning-based, have gradually become the research hotspots in EMSs.

A neural-network-based (NN-based) strategy is generally capable of handling nonlinear modelling and classification prediction, working condition classification, speed prediction, parameter optimisation, etc. The core idea of the reinforcement learning (RL) algorithm comes from behaviourist psychology. And the agent produces the exploration to optimise control strategies, interacts to the external environment, updates the strategy according to the feedback state, and obtains the maximum cumulative reward, which can be achieved model-free. Furthermore, the RL-based algorithms in the researches on the EMS of electrified vehicles including hybrid electric vehicles, plug-in hybrid electric vehicles, hybrid electric tracked vehicles, battery/fuel cell/ultra-capacitor vehicles and so on have been developed in recent years as shown in Table. 1-3 (Ganesh *et al.*, 2022).

For different driving styles and changes in common driving routes, the controller can adaptively adjust the control strategy through online learning to obtain more benefits on energy and emissions. The EMSs based on RL can realise on-board optimisation. From above-mentioned research studies, it has been found that the Q-learning algorithm (described in section 3.22) with a tabular method dominates in RL algorithms, followed by the Deep Q Network. The main reason is attributed to the computation cost and algorithmic complexity, as the Q tabular method requires larger computing effort in terms of memory, but can be updated much faster than the NN-based RL. and is easily applied to practical driving cycles. Therefore, this study designs an EMS controller for a plug-in hybrid electric vehicle (PHEV) based on the tabular Q learning algorithms and verifies their superiority in the PHEV optimisation.

Table.1-3 RL-based EMS for hybrid electric vehicles

Author`s	Configuration	RL algorithm	Action	State	Reward
Xu et al. (Xu *et al.*, 2019) (Xu *et al.*, 2021)(Xu, *et al.*, 2021)	HEVs (parallel)	Q-learning	Power split	v, P_{dem}, a, SoC	m_{fuel}
Liessner et al.(Liessner	HEVs	Deep Q Network	Power output for	w_{wh}, SoC	m_{fuel}

et al., 2018)		(DQN)	electric motor		
Liu et al. (Liu *et al.*, 2018)	Hybrid Electric Tracked Vehicles, Plug-in Hybrid Electric Vehicles	Q-learning, Dyna H	Power output for electric motor	v, SoC	m_{fuel}
;Hu et al. (Hu *et al.*, 2018)	HEVs	Deep Q Network (DQN)	T_{eng}	Total required torque, SoC	m_{fuel}
Zhou et al.(Zhou *et al.*, 2020)	Fuel Cell Vehicles	Deep Deterministic Policy Gradient (DDPG)	Upper and lower limits of SOC	SoC	m_{fuel}
Song et al.(Song *et al.*, 2019)	Plug-in Hybrid Electric Vehicles (parallel)	Deep Q Network (DQN)	T_{eng}	SoC, P_{dem}, v	m_{fuel}
Hsu et al. (Hu *et al.*, 2018);	Fuel Cell Hybrid Electric Vehicle	Q-learning (Tabular)	Degree of Hybridization	SoC, pedal position	m_{fuel}

Sun et al. (Sun *et al.*, 2020)	Fuel Cell Hybrid Electric Vehicle	Q-learning (Tabular)	P_{batt}	SoC	m_{fuel}
Reddy et al. (Reddy *et al.*, 2019);	Fuel Cell Hybrid Electric Vehicle	Q-learning (Tabular)	Change in fuel cell power	SoC	m_{fuel}
Wu et al. (Wu *et al.*, 2018)	Hybrid Electric Bus	Deep Q Network (DQN)	Power of internal combustion engine	SoC, P_{dem}, v	m_{fuel}
Han et al. (Han *et al.*, 2019)	Hybrid Electric Tracked Vehicles	Double DQN	Rotational speed	SoC, P_{dem}, v	m_{fuel}

Abbreviation.

SoC Battery state of charge.

v Vehicle speed

P_{dem} Power demand

m_{fuel}Mass of fue

T_{eng}Engine torquel

1.2 Motivations

This research is partially supported by the Innovate UK project for the development of a plug-in hybrid electric aircraft-towing tractor. The tractor, as shown in Fig. 1-7,

was powered by a solo diesel engine and has been upgraded with a series power-train configuration including a diesel engine-generator and a battery package.

The studied off-highway vehicle is working in a connected environment and consists of an aircraft, an aircraft-towing tractor, an airport control tower and the V2X communications, as illustrated in Fig. 1-8. This connected environment provides an ideal platform for the study of learning-based control method because the vehicle operation data can be synchronised to a powerful workstation harnessing the



Fig. 1-7 Aircraft-towing operation scenario

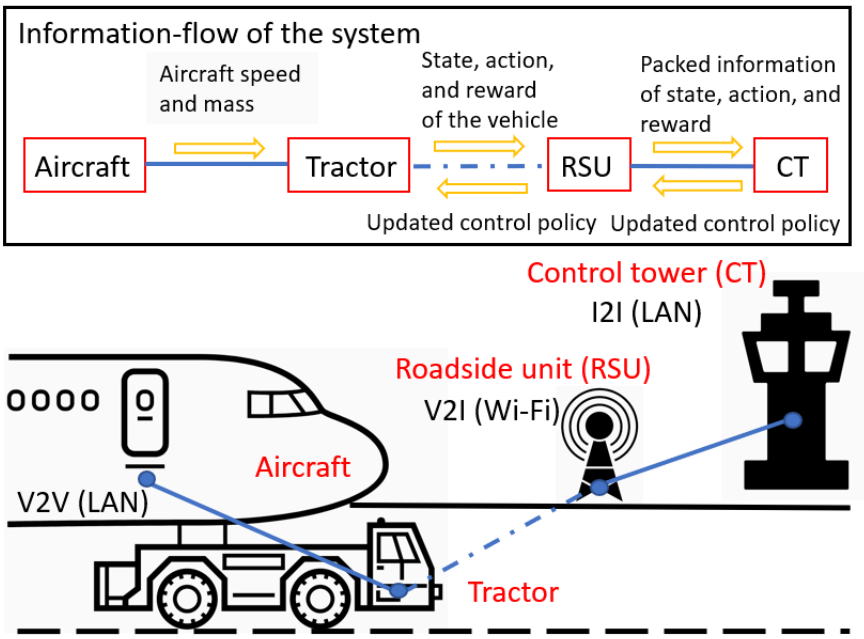


Fig. 1-8 A connected vehicle system (Shuai *et al.*, 2020)

Internet-of-Things (IoT), which provides larger memory space and greater computing power than the onboard controller for the online vehicle control optimization. The signal flow of the connected vehicle system is illustrated at the top of Fig. 1-8. The control tower (CT) sends the command signal (updated control policy) to the roadside unit (RSU) and receives the vehicle information from the RSU through the Ethernet for online RL. The optimised control policy will be downloaded to the tractor for real-time supervisory control.

This industrial-oriented project has promoted new AI methodologies to develop an EMS for a novel electrified aircraft-towing tractor, improving the fuel economy in real-time control. By analysing the literature from Table 1-3, rule-based and model-based predictive control methods are traditionally used for energy management control of

hybrid electric vehicles. However, the robustness of these methods is not satisfactory because these control methods are only calibrated based on the standard driving cycle. The RL is a promising option because the scientific novelty of Q-learning is the implementation of the knowledge base to allow online optimisation in an unknown environment based on Bellman's theory. The working scenario is simple and the state space is not very huge. Therefore, this thesis has been conducted based on the following four observations:

1) Strict legislation for off-highway vehicles lead to manufacturers seeking an advanced control method that allows vehicle to develop its control strategy/policy through real-world interactions to achieve the fuel economy and reduce the GHG emission.

2) With the rapid development of AI, RL has been developed as the one of most promising optimal control methods because it enables the control policy being optimised in real-world operations, but few publications demonstrated RL application in off-highway vehicle control.

3) Through the analysis of RL-based EMS applications for electric vehicles, Q-learning with a tabular method dominates the field of EMS for HEVs.

4) Through the analysis of Q learning as an effective RL method, the exploration and the overestimation during the Q learning process are two main barriers

that limited the optimization performance, and there has yet been any research providing empirical knowledge for reducing overestimation in RL-based EMS for off-highway vehicles.

1.3 Objectives

This Ph.D. research aims to develop a RL-based EMSs for the hybrid off-highway vehicles. It will provide a methodology that will help the OEMs pass legislation in the R&D stage and continuously improve the vehicle's energy efficiency in real-world driving. Meanwhile, this study also promotes the development of dedicated AI for automotive application with a case study on the electrification of off-highway vehicles. The specific objectives of this Ph.D. research are as follows:

1) To build a research platform based on real-time modelling and the hardware-in-the-loop (HiL) testing facility to study the hybrid off-highway vehicle controlled by RL, and the problem of energy management is proposed, that is to say, how to split the total power demand into the engine power and the motor power is to minimize the total fuel consumptions.

2) To solve the exploration and exploitation dilemma of the RL-based EMS through investigating the reward function and the exploration-to-exploitation ratio in reinforcement learning with different decay functions, the learning

process has been enhanced to increase the energy efficiency of the RL-based EMS.

3) To improve energy efficiency of the RL-based EMS by developing an ensemble learning system for the hybrid off-highway vehicle in real-time control, ensemble methods to determine the final control signal are conducted to enhance the optimality, which is a very important indicator to demonstrate how much energy efficiency improvement can be achieved by the proposed RL methods compared to the baseline methods.

4) To reduce the overestimation in the learning process and develop a RL-based EMS for the hybrid off-highway vehicle in real-time control, introducing an extra Q-table to improve the robustness of the desired control strategy has been studied.

1.4 Thesis Outline

This thesis focuses on the development of an advanced energy management strategy (EMS) for the hybrid off-highway vehicle based on reinforcement learning (RL). To tackle the challenges in RL-based EMS, the specific contributions of this thesis are (1) Three decay functions, exponential-based, reciprocal-based and step-based are used to generate an appropriate trajectory for the first time to improve the

exploration of reinforcement learning; (2) A new ensemble learning system based on a combination of two E2E methods has been developed for the first time to improve the vehicle energy efficiency of an off-highway vehicle under real-time operation; (3) Three ensemble methods (maximum, random, and weighted-based) have been investigated for the first time to determine the final action;(4) A new predictive double Q-learning framework is proposed by introducing into conventional double Q-learning two action execution policies and an asynchronous learning method with backup models. And this thesis is organized into seven further chapters which can be summarised in Fig.1-9.

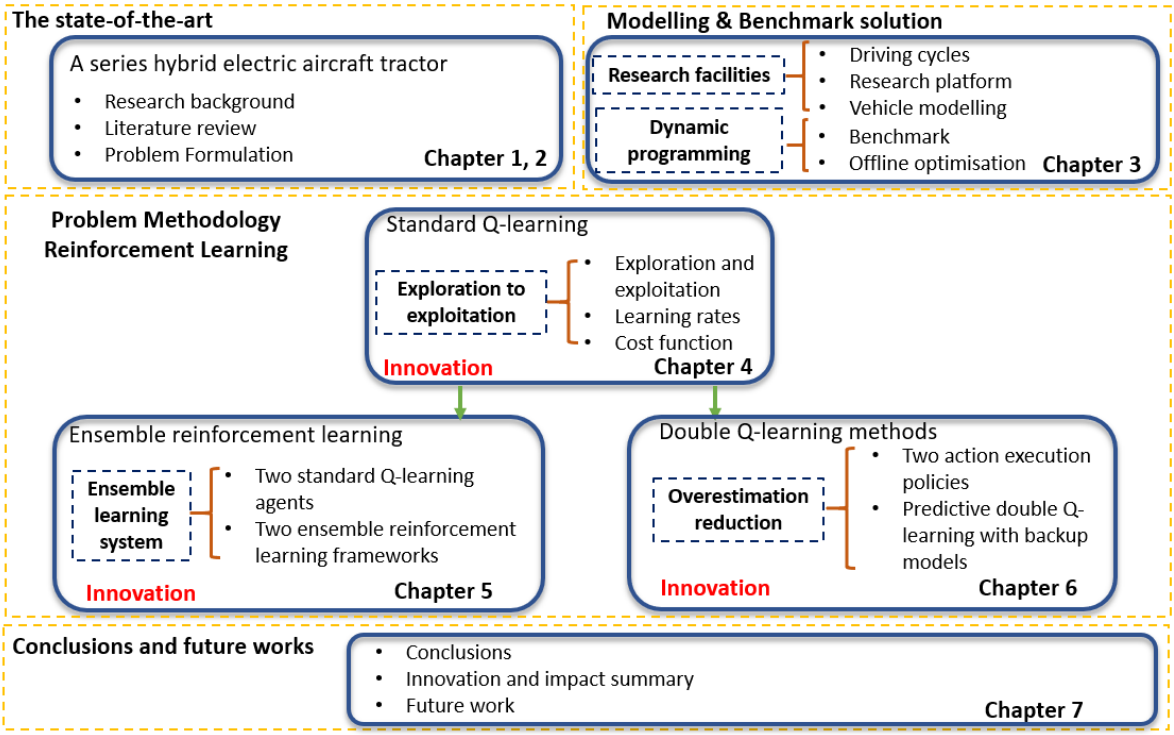


Fig. 1-9 The main research content of this thesis and the structure of each chapter

Chapter One introduces the research background, motivations, objectives and thesis outline.

Chapter Two illustrates a comprehensive literature review of: 1) fundamental modelling methods for HEVs and validation approaches for desired EMS; 2) The state-of-art methods of EMSs for HEVs. 3) And recent development of RL.

Chapter Three depicts the research methodology and facilities used for this research. Firstly, the research methodology is introduced, followed by presenting the research platform, including modelling of powertrain for a hybrid off-highway vehicle, the HiL test platform and software for the HiL testing.

Chapter Four presents research of online EMS using a Q-learning method. Firstly, the mathematical EMS model for a hybrid off-highway vehicle is demonstrated, and two problems of standard Q-learning are introduced, followed by description on the model-free supervisory control system. Next, three decay functions are proposed to generate appropriate trajectories for exploration probability. Finally, the feasibility of three decay functions is carried out based on the HiL.

Chapter Five studies an ensemble learning system for HEVs. Two frameworks of an ensemble learning system are firstly implemented in the optimisation of vehicle energy efficiency. Next, three ensemble methods are studied to improve the optimality. Finally, the feasibility and optimality of the proposed EMSs are simulated

and evaluated by HiL testing.

Chapter Six investigates the model-free supervisory control with double Q-learning for a hybrid off-highway vehicle in a charge-sustaining stage. Firstly, two problems in double Q-learning methods are presented. The two new heuristic action execution policies are proposed for the double Q-learning methods, and a framework of predictive double Q-learning with backup models is introduced. Ultimately, the HiL is carried out to explore energy-saving potential under real-world vehicle operations.

Chapter Seven summarises the outcomes and innovations of the research conducted in this thesis. A recommendation for future work is also given, followed by a recommendation for future work.

Chapter Two

LITERATURE REVIEW

Hybrid powertrains have been widely used in highway vehicles because of their potential to reduce fuel consumption and emissions significantly. They minimise the vehicle's energy consumption and emissions with the help of advanced EMSs in real-world driving. Although the off-highway vehicles work in quite different conditions, powertrain architecture and control strategies are similar to those of highway vehicles. This chapter illustrates a comprehensive literature review about: 1) modelling methods of HEVs and validation approaches for the desired EMSs; 2) the current development of EMSs for HEVs; and 3) the recent developments and applications to the Q-learning algorithm of HEVs.

2.1 Modelling and validation method for HEVs

Modelling is necessary for the research and development of HEVs because accurate vehicle system models can evaluate the performances of the vehicles with a minimised cost. The simulation of a vehicle’s performance (e.g. fuel economy, cost, emission) with different design parameters and control strategies before a physical prototype is built to verify the presented algorithms (Huang *et al.*, 2018). Currently,

the graphical software such as using MATLAB/Simulink and ADVISOR, is primarily used for HEVs modelling(Gao *et al.*, 2007). This section introduces two modelling methods for HEVs: the forward-looking and backward-looking modelling approach (Wipke *et al.*, 1999).

The learning performances of distinct proposed RL-based algorithms will be evaluated, such as using the real-time performance, the stability, the optimality, and the robustness through rapid control prototyping and the HiL method to demonstrate that the presented algorithms are effective and efficient(Q. Zhou *et al.*, 2021).

2.1.1Modelling approaches of powertrain systems

According to the main components and the power flow, forward-looking and backward-looking are conducted to develop the modelling of the different powertrain systems for HEVs(Huang *et al.*, 2018). A driver model is included in the forward-looking modelling approach as shown in Fig. 2-1, where the current vehicle speed and required speed are taken into consideration to develop an appropriate control command. More specifically, the so-called forward-looking simulation means that the driver first controls the accelerator pedal or the brake pedal, to generate the control signal flow, and then this signal flow passes through various parts of the transmission system to the wheels. Therefore, the computation is complex, and it takes much time dealing with numerous variables from different vehicle components.

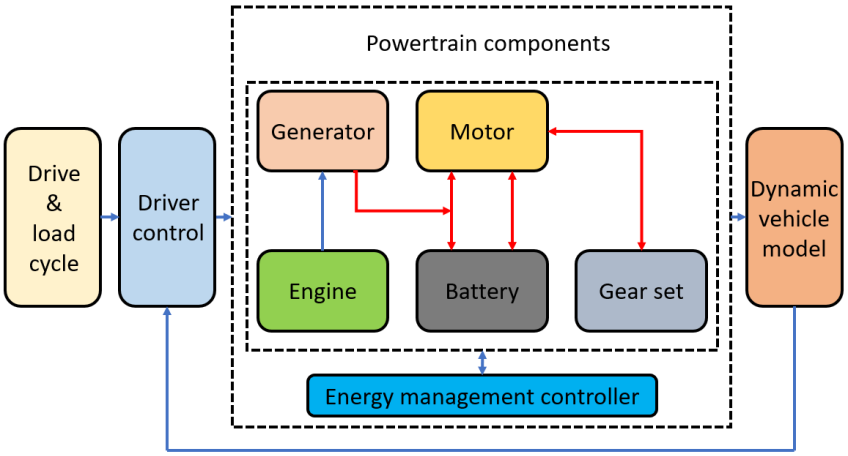


Fig. 2-1Forward-looking modelling approach (Huang *et al.*, 2018)

The latter is the backward-looking modelling approach, as shown in Fig. 2-2, assumes that the vehicle has already met the drive power requirement; and then, the vehicle controller deals with the power coming from the energy sources wheels (Gao *et al.*, 2007). The wheel torque, wheel speed, vehicle acceleration will be firstly obtained from the predefined driving cycles. Then, the energy consumption (e.g. battery usage and fuel consumption) and emissions will be calculated through each vehicle component (Enang *et al.*, 2017). Compared with the forward-looking simulation, the backward-looking approach does not have a driver model, so the dynamic process of clutch and transmission is ignored. Therefore, the calculation speed becomes faster and can be calculated many times to achieve a high performance (Guzzella *et al.*, 2007). It is suitable for pre-evaluation of the vehicle's dynamic and fuel economy performance.

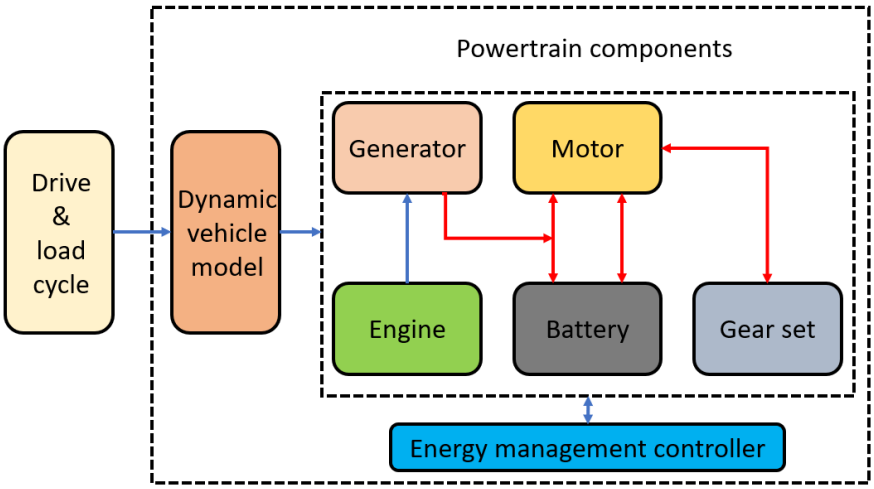


Fig. 2-2 Backward-looking modelling approach (Huang *et al.*, 2018)

2.1.2Validation approaches of powertrain systems

It is essential for the automobile industries to validate the feasibility of real-time driving when designing the engine, electronics, and other mechanical components via an efficient development method. In (Schaffnit *et al.*, 2000), the author presented a workflow for designing and testing the new vehicle electric control unit (ECU) functions, as shown in Fig.2-3. The common validation approaches of powertrain system are demonstrated in Fig.2-4, which contains the software-in-loop (SiL), rapid control prototyping (RCP) and hardware-in-loop (HiL). Many research publications in terms of designing new EMS are firstly developed based on SiL with forward-looking (Qi *et al.*, 2022) or backward-looking approaches (Zhou *et al.*, 2022).

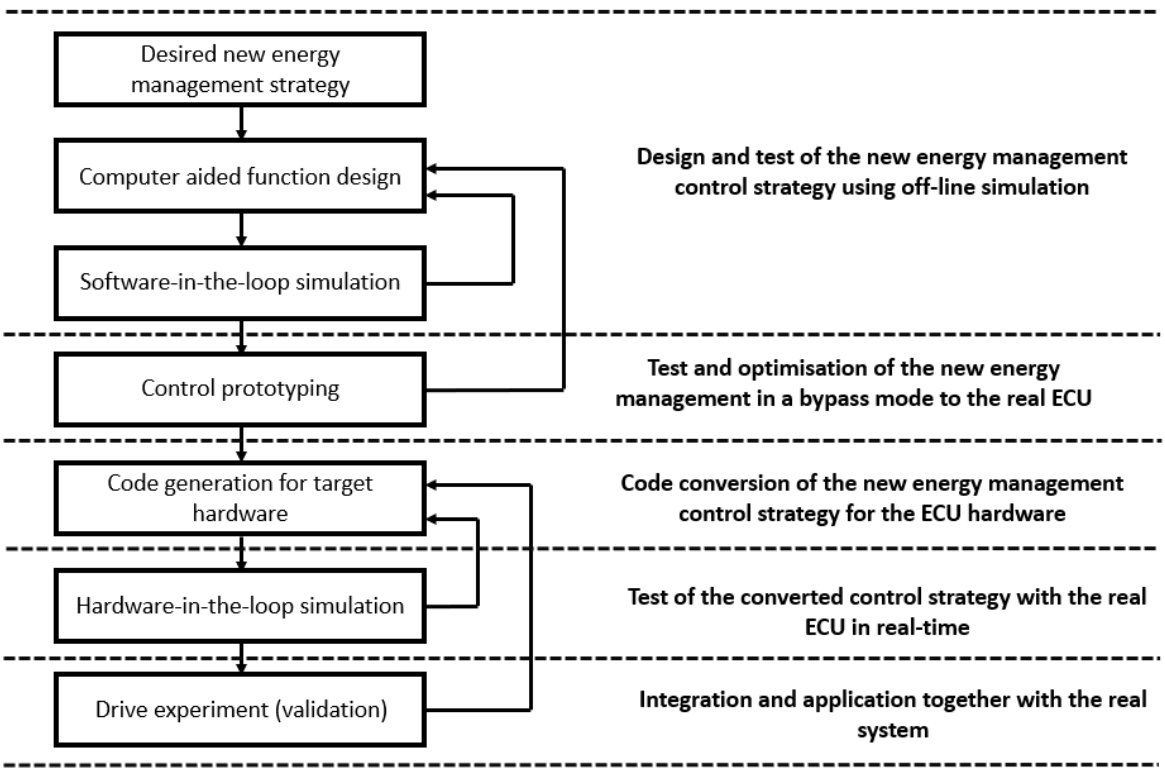


Fig. 2-3 The workflow of designing new functions of the ECU (Schaffnit *et al.*, 2000)

Next, RCP is used to investigate new functions of the ECU through bypassing the mode to the real ECU under simulation process or actual process. It is alternative way for validating the capability of EMS under real-time simulation(J. Liu *et al.*, 2019), in which the real-time target machine only calculates the selected functions in a bypass mode to the real ECU. For example, the University of Birmingham has built an RCP system for cyber-physical control of a hybrid off-highway vehicle, as shown in Fig.2-5, in which the desired EMS will be downloaded into the Speedgoat machine to evaluate the real-time feasibility.

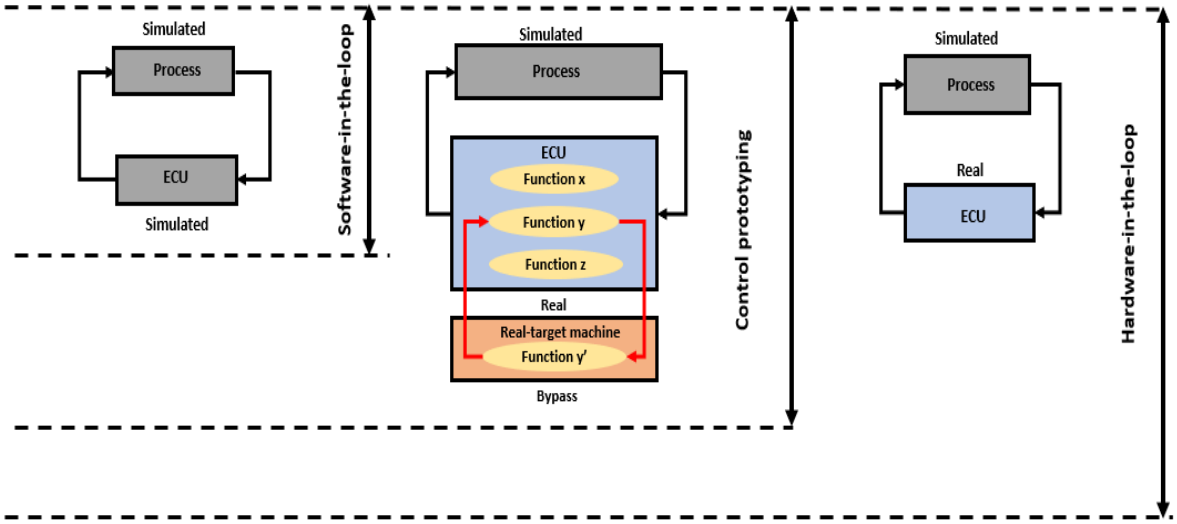


Fig. 2-4 Three common validation apporaches (Schaffnit and Isermann, 2000)

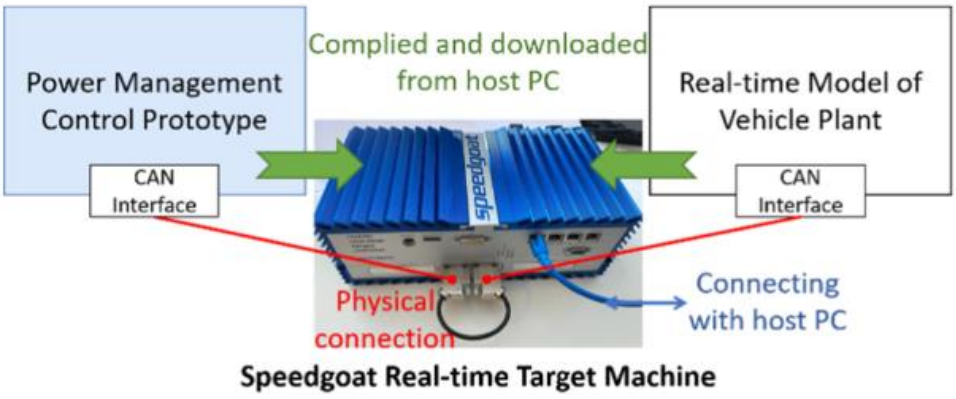


Fig. 2-5 Online-rapid control prototyping testing platform (Zhou *et al.*, 2022)

Testing control algorithms in real world, can be time-consuming, expensive, and potentially unsafe. Therefore, HiL approach is used to operate one or more real components in connection with real-time simulated components (Yi *et al.*, 2016). The real-time target machine will work together with the simulation process in real-time (Schaffnit *et al.*, 2000). The Fig.2-6 demonstrated an example of HiL system. The

dSPACE simulator is used to provide a real-time circumstance, and the desired EMS is downloaded into vehicle controller. And the dSPACE simulator and vehicle controller are connected by CAN bus (Wang et al, 2020). In order to improve the trustworthiness of the HiL, the driver' model can be involved by driving controlling the accelerator and brake pedal (J. Li *et al.*, 2019)he driver-in-loop system is necessary, as shown in Fig.2-7.

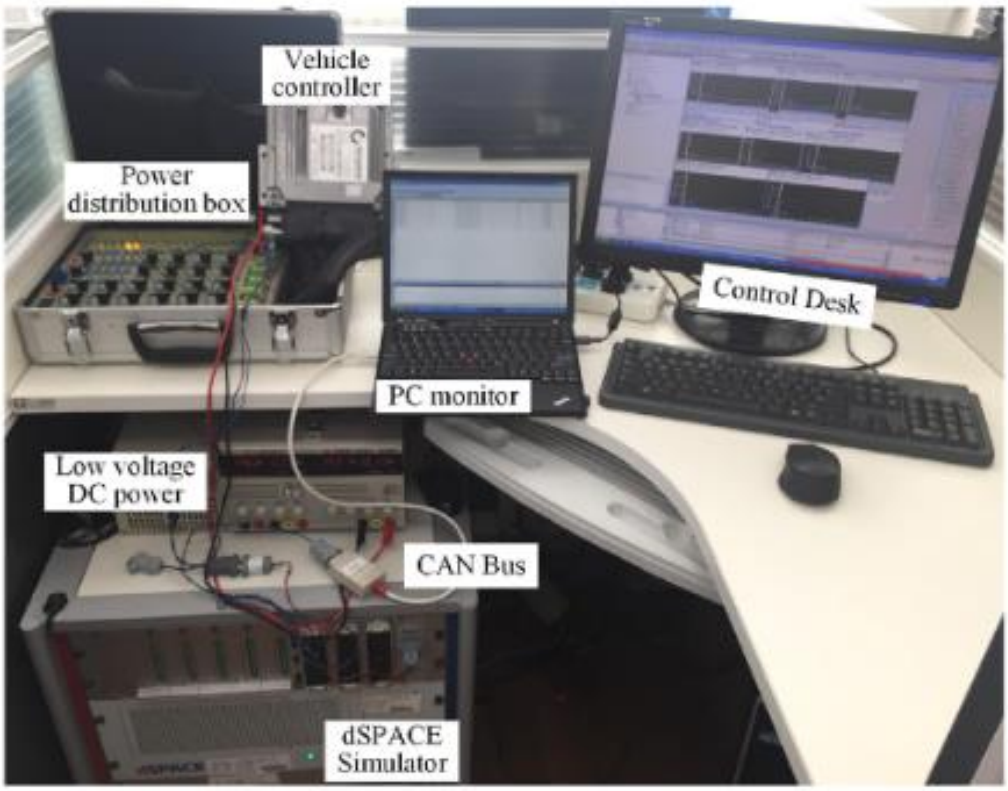


Fig. 2-6 Hardware-in-the-loop test platform (Wang et al, 2020)

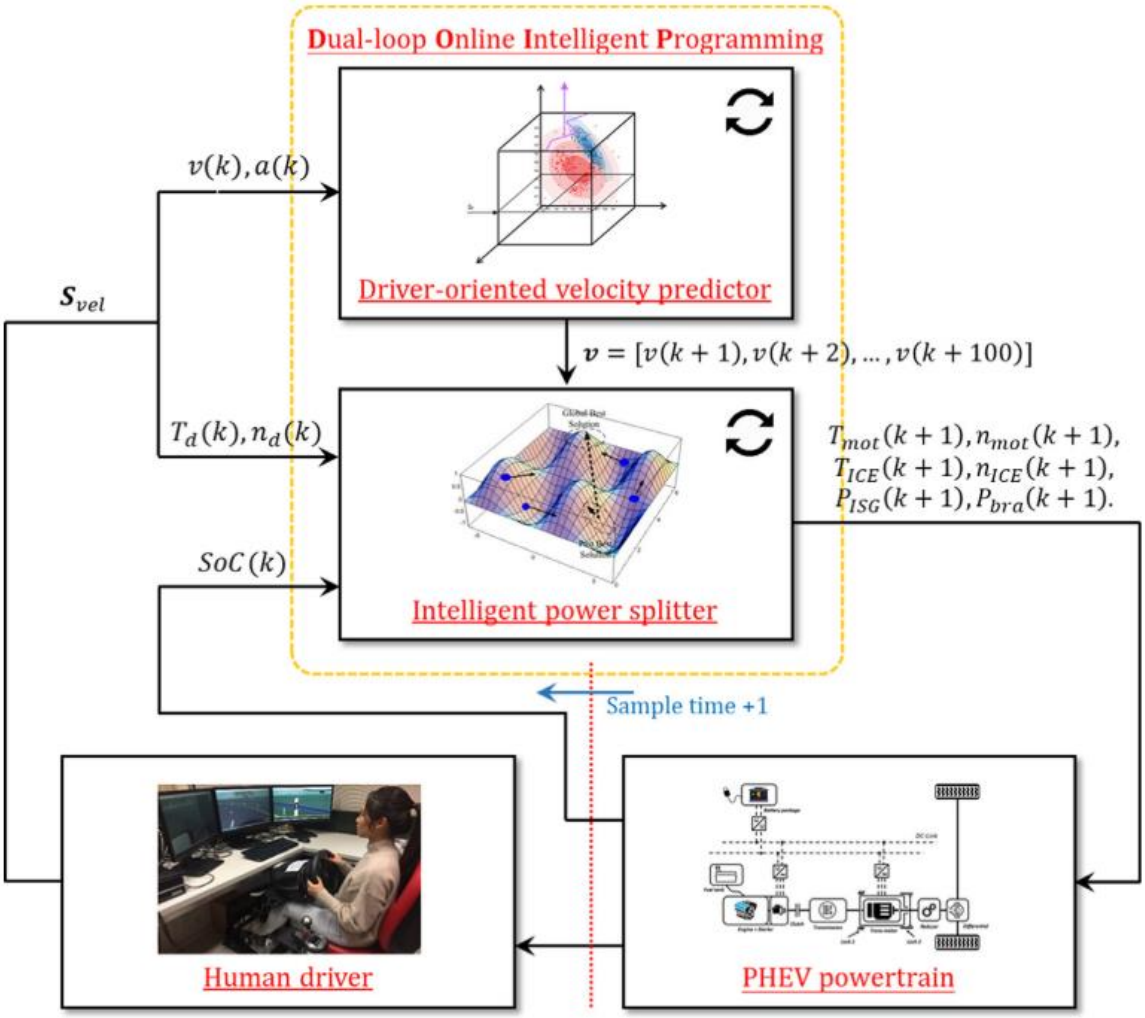


Fig. 2-7 Driver-in-the-loop testing platform (Li *et al.*, 2019)

2.2 Classification of EMSs

The EMSs for HEVs are an indispensable part, which coordinates the power flow within the system of multiple power components (e.g. ICE, tractor motors) (F. Zhang *et al.*, 2019). The goal of developing HEVs is to reduce fuel consumption by investigating appropriate EMSs while meeting the drive requirements. Regardless of the topology of the powertrain, the improvements in the fuel economy for HEVs

depend on the EMSs, which efficiently and instantaneously disturb the energy flow of multiple power components to achieve the control objectives (Sabri *et al.*, 2016). Various EMSs for HEVs have been implemented in recent years. Much similar work has been done, e.g. for passenger cars, buses, and trucks, and there are three main categories of EMSs, as shown in Fig. 2-8, including offline EMSs, online EMSs and learning-based EMSs.

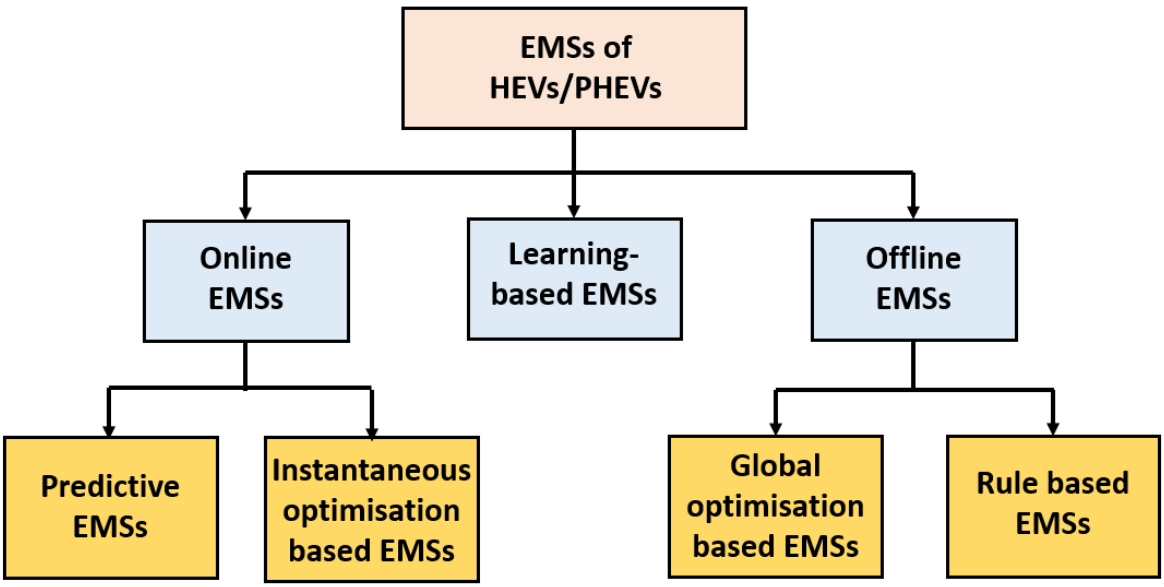


Fig. 2-8 Classification of EMSs for HEVs

2.2.1Offline EMSs for HEVs

The offline EMSs are layered based on the level of driving information under rule-based EMSs and optimisation-based EMSs. Generally, these types of EMSs are unable to implement real-time control because they require prior knowledge of all information (i.e. the whole driving cycle).

a) The optimisation-based EMSs for HEVs

The optimisation-based EMSs seek a control policy using prior knowledge of the typical driving cycle. There are many optimisation-based EMSs which can be applied for solving the optimal problem of power allocation for HEVs, such as stochastic dynamic programming (SDP) (Xiong *et al.*, 2021); game theory (GT) (Javanmard *et al.*, 2021); and Pontryagin's minimum principle (PMP) (Shi *et al.*, 2021). Dynamic programming (DP) is one of the most popular methods in offline optimisations, which seek a globally optimal policy based on prior knowledge of the typical driving cycles (Lee *et al.*, 2020; Tian *et al.*, 2021). In order to improve the energy-saving possibilities of off-highway vehicles, Hu et al. (2021), presented a two-step framework based on dual-engine system. An optimal power distribution strategy between two diesel engines based on their efficiency maps is utilized. This distribution strategy is incorporated with the DP algorithm. In Murphey *et al.*, (2013), an improved rule-based control was developed by analysing the DP results for HEVs. The results indicate that the improved rule-based control policy can reduce its performance (fuel economy and emission evaluation) gap to the theoretically by 50%-70%. A weighted improved DP was proposed to distribute power for HFCEVs, which considers the battery Q-maps. Weights are incorporated in the fitness function in-order to improve convergence rate of such a method with long duration driving cycles (Fares *et al.*, 2020). Other optimisation-based EMSs, including a genetic algorithm (GA) (Ding and

Prasad, 2020); simulated annealing algorithm (SA) (Hafez *et al.*, 2021); particle swarm optimisation (PSO) (Mamun *et al.*, 2019); and non-dominated sorting genetic algorithm (NSGA-II) (Li *et al.*, 2021) are suitable to solve the offline optimisation problems, in which the driving conditions are fixed and known in advance. These EMSs use a stochastic search to seek the optimal control policy by taking the iterative method. For example, Zhou et al. proposed a modified accelerated particle swarm optimisation algorithm (MAPSO) which determines the optimal combination of component specification (Zhou *et al.*, 2020). Hwang et al. implemented PSO for a power-split HEV to improve fuel economy (Optimization *et al.*, 2020). Boujneh et al. converted a multi-objective problem to a single-objective problem and then used PSO to conceive appropriate control parameters for a parallel HEV (Boujneh *et al.*, 2020). However, the above-mentioned offline EMSs, cannot be directly implemented in real-time control. Because they suffer from a considerable computational effort, and it is impossible to know the future driving conditions (road slope, traffic dynamics) (J. Zhou *et al.*, 2021). Therefore, these methods are usually employed to extract some heuristic rules and evaluate the performance of other EMSs' algorithms.

b) Rule-based EMSs for HEVs

Rule-based methods are fundamental control schemes. They need less memory for onboard computing and are easily implemented in real-time vehicle control (Liu *et al.*, 2008). The core of rule-based methods is the predefined rules based on human

experience and mathematical models. The rule-based EMSs can be classified into deterministic rule-based (Salmasi *et al.*, 2007) and fuzzy logic strategies. The former has been widely implemented in commercial vehicle controllers, where the rules are designed based on the specific driving cycles. The fuzzy logic-based EMSs aim to allocate the power flow with a set of linguistic rules. The fuzzy relations greatly rely on the similarity or the degree of similarity between data sets. Usually, the input of the fuzzy logic control (FLC) are the battery SoC, and the drive power demand; while the output is the power from the energy sources (Salmanx *et al.*, 2000; Faizrizwan *et al.*, 2018; Ma *et al.*, 2019). For example, Singh et al. presented a forward-looking vehicle model for a power-split HEV and implemented a fuzzy logic-enabled energy management strategy for the hybrid electric vehicle based on torque demand, battery state of charge and regenerative braking is designed and implemented. The proposed energy management strategy allows engine and motor to maneuver in their efficient operating regions (Veer *et al.*, 2020). In Essoufi, Hajji and Rabhi, (2020), an FLC is established in MATLAB/Simulink, and it allocates the power distribution for a FCEV to decrease hydrogen consumption while increasing the durability of power sources by respecting dynamic constraints as well as the battery's state of charge. In Montazeri-gh and Mahmoodi-k, (2015) a power-split hybrid powertrain with a multi-input FLC is introduced and compared to the rule-based EMSs through emissions and fuel consumption. The results indicate that proposed controller is reduced fuel consumption in real and modal driving cycles about 21% and 6%

respectively.

However, the aforementioned EMSs usually require an informative prior knowledge of all the driving conditions, less adaptability and easily fall into the local optimal solution, which is not able to fully optimise the fuel consumption for the whole trip (Sabri *et al.*, 2016). Therefore, online optimisation is necessary for real-time hybrid powertrain control with the limited information.

2.2.2Online EMSs for HEVs

The online EMSs are presented with instantaneous optimisation-based EMSs and predictive EMSs. These EMSs do not require a priori knowledge of all of the future vehicle information (Huang *et al.*, 2018). Therefore, they can be utilised in real-time by changing the global optimisation problem to an instantaneous optimisation problem.

a) Instantaneous optimisation-based EMSs

These EMSs optimise the power distribution with the current vehicle performance (e.g. emissions and drivability) to reduce the instantaneous fuel consumption at each time step (Yanjun Huang *et al.*, 2017).

One of the representatives belongs to the equivalent consumption minimisation

strategies (ECMSs). In the ECMSs, the equivalent factor (EF) is the crucial factor in reducing energy consumption (Wirasingha *et al.*, 2011; Panday *et al.*, 2014; Zhang *et al.*, 2015), which balances the fuel consumption and electricity based on online optimisation. Under different driving cycles , a value for the optimal EF obtained offline is selected online in real-time control (Martinez *et al.*, 2017; Sun *et al.*, 2017). This is due to the future trip information being unknown. The fuzzy control is also employed to dynamically regulate the EF with an appropriate trajectory widely mixed different driving conditions (Wang *et al.*, 2019; Lei *et al.*, 2020). Other instantaneous EMSs, including the adaptive equivalent consumption minimisation strategy (A-ECMS) (Li *et al.*, 2020), are also used to address the optimal solution for HEVs. However, these types of EMSs still have weakness: 1) they easily fall into suboptimal solutions instead of global optimisation; 2) there is the massive computation cost.

b) Predictive EMSs for HEVs

Considering road conditions, preceding vehicles, and traffic conditions, predictive EMSs with accurate prediction models can adapt various driving conditions in real-time implementation(Shen *et al.*, 2018)(He *et al.*, 2021). Model-based predictive control (MPC) has been widely adopted as one of the representatives for the online optimisation of HEVs (Jinquan *et al.*, 2019). The MPC operates a rolling optimisation process in the vehicle controller, based on predicting the vehicle's future power demands to effectively deal with the distribution of the energy sources over a

prediction horizon under the unpredictable future driving conditions (Afram *et al.*, 2014). Thus, precise predictive models, developed with mathematic functions, are the keys to enabling MPC-based methods to achieve the optimal performance in real-world driving (C. Yang *et al.*, 2020). However, it is tough to obtain the precise models including detailed real-world dynamics (e.g. battery ageing, driver's behaviours). Therefore, learning-based control methods need to be developed to address the limitations of both offline and online optimisation-based methods.

The primary challenge of the control strategies is that most of the optimal solutions are generally designed under the standard driving cycles, which means it is impossible to achieve optimal results from the real-world driving cycles. Besides, the current EMSs, such as instantaneous optimisation-based and predictive optimisation-based, are hard to implement in real-time due to the computational complexity.

2.2.3Learning-based EMSs for HEVs

With the development of AI, learning-based control is promising because it can break through the limitations of the existing offline and online optimisation methods in EMSs. The learning-based EMSs, including NN-based and RL-based, can self-learn to adapt to the different traffic conditions without an accurate model, and the control parameters can be updated by the interaction with an unknown environment and

learning from it(Zou *et al.*, 2021).

a) Neural network-based EMSs for HEVs

NN-based learning strategies are employed to take historical information into consideration concerning various performances gained in the past testing conditions, and then online learning methods are used to adapt current data from the preceding cases (Cotroneo *et al.*, 2013). NNs are the typical representative method of offline learning algorithms (Loganathan *et al.*, 2019). The NN-based method combined with dynamic programming (DP) is formulated to effectively predict road and traffic information (Murphey *et al.*, 2012). During the real driving conditions, the energy management system controller needs the ability of self-learning to face changeable driving conditions and vehicle status; therefore, combining optimisation methods with NNs or machine learning have been presented and have become central issues in recent years (Gao *et al*, 2021). There are neuro-fuzzy networks which Leandro L. S. Linhares (Linhares *et al.*, 2015) adopted to follow the optimal results in real time based on offline training. It is advantageous to employ the flexible usage of data from the process and the rapid adaptation of network parameters to changing non-linear process behaviour by using fairly standard optimisation procedures.

b) RL-based EMSs for HEVs

Among the online learning-based methods, RL has gained the most popularity in

many research areas, such as: driver-behaviour classification (Lu *et al.*, 2018); robotics (Apolinarska *et al.*, 2021); gaming (Liu *et al.*, 2021); autonomous vehicles (Liu *et al.*, 2018); and lane detection (Xing *et al.*, 2020). The RL algorithms obey the interactions between the agent and environment. The learning process of RL includes six essential elements in RL: 1) **Agent**: the learner and decision-maker; 2) **Environment**: the agent learns and decides what actions to perform; 3) **Action**: a set of actions which the agent can perform; 4) **State:** the state of the agent in the environment; 5) **Reward**: for each action selected by the agent the environment provides a reward; 6) **Policy**: the decision-making function (control method) of the agent, which represents a mapping from situations to actions.

Unlike the MPC-based methods that require the accurate physical model of the vehicle platforms, the RL-based methods are physical model-free, which has already been used in EMSs for HEVs (Liu *et al.*, 2018; Zhao *et al.*, 2018a; Liu, *et al.*, 2020; Nyong Bassey *et al.*, 2020). The optimisation process is conducted by an RL agent that follows the Bellman's optimality theorem to update its knowledge base (a model mapping the state variables and action variables to the expected rewards), through rolling interactions with the unknown environments.

Thanks to the rapid advancement in RL techniques and the highly flexible qualities of RL-based methods, abundant RL-based methods of interdisciplinary integration have emerged. Monte Carlo (MC) is one of the early RL algorithms, which depends

on actual data rather than the estimation when updating every step, but solely updating until the episode is completed. Different from MC, temporal difference (TD) is capable of updating when the episode is not finished, and the estimation accuracy determines whether the learning performance is efficient and effective. Then, based on the method and principle of DP, the Q-learning method was presented in EMSs to learn different driving conditions(Han *et al.*, 2019). It is model-free compared with DP; however, 1) the overestimation problem of the action values is produced and 2) this method is limited to discrete and finite state and action space. To solve the first problem, double Q-learning is employed to underestimate action values, but it brings about a high computation cost compared with standard Q-learning (Van Hasselt *et al.,* 2015; Zhang *et al.*, 2018).. Then, deep Q-learning is introduced to deal with the second problem and can be applied to highly complicated systems with state and action space(Zhang *et al.*, 2021) (Yu *et al.*, 2021). However, it is inevitable that overfitting phenomena occurs and there is the large computation cost during learning. Combined with respective characteristics of these methods, like Q-learning and TD, multi-step Q-learning is proposed to enhance the learning performance. With the complication of the practical problems, multi-agents system with more flexible architectures is presented to enhance the efficiency of data exchange and sharing. The inverse RL and safe RL have gradually developed in recent years. Therefore, the classification of RL-based EMSs is various according to different standards. Based on the research background in chapter one, Q-learning is drawing more

attention to the research of the application of EMSs for HEVs.

2.3 Recent development of Q-learning algorithm and its application of EMSs for HEVs

The Q-learning algorithm is a widely used model-free RL technique, unlike modern control algorithms, like MPC, there is no need to provide the precise model in control algorithms. (Baoxian *et al.*, 2013). A Q-table represents a 'map' for the agent to determine an optimal action, in which the state-action values and Q-values are recorded for trial and error. Every entry to the table represents a state-action pair, and all Q-values are updated as:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r(s,a) + hmax_{a'}Q(s',a') - Q(s,a) \qquad (2\text{-}1)$$

where $\alpha \in [0,\ 1]$ is the learning rate, which explicitly defines to what degree new knowledge overrides old knowledge; for α = 0, no learning occurs; while for α = 1, all prior knowledge is lost; $h \in [0,1]$ is the discount factor, which determines the importance of a future reward. The scientific novelty of the Q-learning algorithm is implementing the knowledge base to allow online optimisation based on Bellman's theory (Bellman, 1954). Based on the storage method of value and policy functions, there are two typical categories, tabular-based and approximate-based methods.

a) Tabular-based methods

The tabular-based methods are very straightforward and are usually implemented in solving optimal problems with a small number of states and action variables. The approximate value functions can be represented as tables or arrays (Sutton *et al.*, 2012). In the standard Q-learning, a Q-table is always used for evaluating the value function and seeking the optimal control policy. For example, Liu et al. presented some types of research of EMSs based on the standard Q-learning method of various topologies for hybrid powertrains, including the series hybrid powertrain (Liu *et al.*, 2015; Liu *et al.*, 2018; Liu *et al.*, 2018), the parallel hybrid powertrain (Liu *et al.*, 2017; Liu *et al.*, 2020) and the power-split powertrain (Zhu *et al.*, 2015; T. Liu *et al.*, 2019). Three multi-step RL-based methods based on tabular Q-learning were proposed by Zhou et al. The results show that the vehicle energy efficiency can be increased from the initial level of 34% to 44% after 5h'35 with 35-step learning (Zhou *et al.*, 2019). Cao et al. optimised the energy use with a dual-energy storage system based on tabular Q-learning (Cao *et al.*, 2017). While Li et al. researched a power management method based on Q-learning with continuous state-action space. The results indicate that the total cost of a trained AC based method is only 2.76% higher than that of DP, while saving 88.7% of calculation time than that DP takes (Li *et al.*, 2017).

b) Hybrid-based methods

Two or more advanced artificial intelligent algorithms and more vehicle information

can be considered as integrating with the RL framework to achieve better energy efficiency compared to tabular-based methods for HEVs. Ensemble methods show more potential ability to benefit from all algorithms (Krawczyk *et al.*, 2017), and have been successfully implemented in supervised learning (Richman *et al.*, 1996) and semi-supervised learning (Zhang *et al.*, 2013). In RL, ensemble methods have been used for representing and learning the value function (Sun *et al.*, 1999). The critical challenge of the ensemble methods is how to distribute the proportion between different base learners because all base learners are at the same level of separation power. There is not much research which has been done in energy management for HEVs. One of the first attempts was by (Xu *et al.*, 2020), in which two frameworks of RL-based ensemble methods were investigated for a power-split HEV.

The hybrid-based methods are also popular with a rapid development in computation. The deep reinforcement learning (DRL) technique is more powerful to deal with higher dimensional states and action than classic tabular-based EMSs, considering more vehicle information (Zhao *et al.*, 2018b; Wu *et al.*, 2019; Du *et al.*, 2020). Several researchers attempted to utilise the DRL models to achieve optimal solutions for HEVs. Wu et al. enabled continuous optimisation of a power-split hybrid system based on a deep Q network. The main contribution of this study is to explore a novel reinforcement learning methodology into energy management for HEB which solve the curse of state variable dimensionality, and the techniques can be adopted to

solve similar problems (Wu *et al.*, 2018). While Qi et al. enabled self-learning control of a PHEV based on DRL It is a fully data-driven and self-learning model that does not rely on any prediction, predefined rules or even prior human knowledge. The experiment results show that the proposed model is capable of achieving 16.3% energy savings (with the designed PHEV simulation model) on a typical commute trip, compared to conventional binary control strategies (Qi *et al.*, 2019). Then, Li et al. implemented DRL for the EMSs of a series hybrid bus considering its trip history information. The simulation results indicate that after further applying output frequency adjustment, a mean gap of 8.7%, which is comparable with MPC-based EMS with mean prediction error of 1 m/s, is maintained with concurrently noteworthy improvement in reducing engine start times (Li *et al.*, 2019). However, standard DRL uses the same values for selecting and evaluating, making it more likely to have a bias resulting in over-optimistic value estimates.

Many researchers have focused on EMSs for HEVs. They aimed to reduce fuel consumption and emissions to achieve global results. And the RL method can solve the EMSs for the HEV including: 1) in every driving cycle, vehicle speed, power demand, battery SOC values etc. are changing, and the EMS controller needs to determine the output power of the engine and battery respectively in the next state, which is related to the current control action. 2) The optimisation goal of the EMS problem for the HEV is the total fuel consumption and charge sustained over the

whole trip, instead of a certain step. 3) The current vehicle speed, power demand and other states are obtained then the fuel consumption and SOC of the current action are calculated; that is to gain feedback, it is unnecessary to grasp the specific information of the entire trip in advance. 4) Actual driving conditions are unstable and changeable. Based on the above literature reviews, the reinforcement learning method can improve the vehicle's energy efficiency by at least 5% compared to the conventional MPC-based optimal control method (Liu *et al.*, 2017)

The future trends of EMSs for HEVs can be predicted from the existing research: 1) the real-time dynamic traffic condition and future information are collected from the intelligent transportation system(ITS) through the V2X network (Martínez *et al.*, 2019); 2) simplify the vehicle model and develop 'model-free' control with a self-learning capability of EMSs through RL (Bouktif *et al.*, 2018; Xiong *et al.*, 2018).

2.4 Challenges

Based on the literature review in previous sections, literature is rich in the RL-based EMS area and is still growing. Even though considerable achievements have been made in this area, there still exist many challenges for off-highway vehicle. 1) There are still lack of study on the impact of vehicle energy consumption by changing ratio of exploration and exploitation in RL. 2) And a few literature reviews research on EMS based ensemble reinforcement learning. 3) Overestimation phenomena is still

a crucial problem in RL which will highly have impact on reducing vehicle energy efficiency.

2.5 Summary

This chapter firstly reviewed two modelling methods, forward-looking and backward-looking, to build a modular and flexible hybrid vehicle model, followed by an introduction about the validation approach for the control strategy. This chapter also conducted a comprehensive review of current EMSs for HEVs with three categories: offline optimisation, online optimisation, and learning-based optimisation methods. The above contributions can be summarised as follows:

- Compared with the forward-looking modelling approach, the backward-looking modelling approach is used in a known driving cycle to reduce the time during the simulation process significantly. And it can be used for real-time simulation based on the HiL platform for validation of the EMS control functionality.

- The RL-based EMSs provide model-free predictive control to use self-learning with the unknown driving conditions. And they are promising to deal with the adaptive optimisation for HEVs.

- The Q-learning method dominates the field of RL, bringing about two main problems of 1) how to balance the exploration and exploitation; and 2) how to reduce the overestimation.

Chapter Three

RESEARCH METHODOLOGY AND FACILITIES

This chapter provides an introduction about the research methodology employed in this study, in which the DP method and standard Q-learning are developed as baseline. Next, a description of the research platform is presented, including 1) the real-time models of hybrid off-highway vehicle system and subsystems; 2) the HiL testing platform; and 3) the software for HiL testing.

3.1 Research methodology

The research methodology for developing the RL-based EMS in the hybrid off-highway vehicle is shown in Fig.3-1. The basic RL algorithm is the tabular Q-leaning, in which the learned experience is stored in the Q-table to obtain optimal EMS control sequence of modelling for a hybrid aircraft-towing tractor. In general, four work packages are design to achieve the four objectives as illustrated in Chapter One. A research platform consisting of the real-time vehicle model and the RL-based EMS is developed based on HiL testing platform to achieve Objective one (details in this Chapter). Three decay functions are investigated to solve the dilemma of balancing the exploration probability in standard Q-learning; and two types of reward functions

are studied to improve the accuracy of selecting optimal actions to achieve the Objective two (details in the Chapter Four). An ensemble reinforcement learning system is presented to consider the strength of each decay function, in which three ensemble methods are proposed to determine the final action to achieve the Objective three (details in Chapter Five). To achieve objective four, the double Q-learning methods are investigated (details in Chapter Six), where two execution policies are firstly studied to reduce the overestimation; furthermore, a double Q-learning with a backup model is developed to improve the learning performance. The vehicle performances with the proposed RL-based methods are evaluated on the HiL platform by using the results obtained by DP and standard Q-learning as the benchmark.

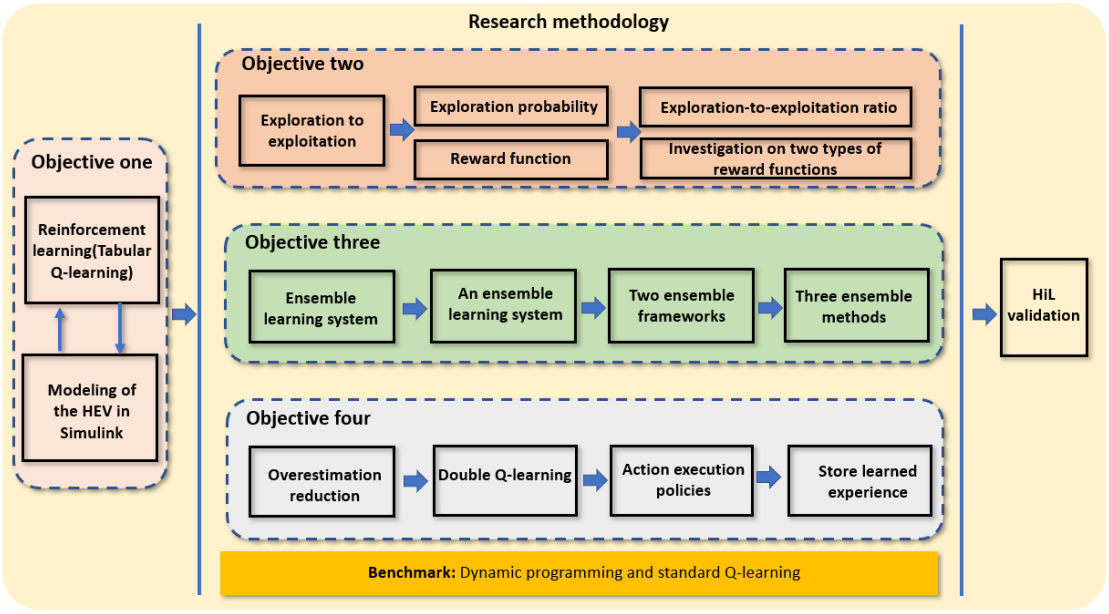


Fig. 3-1 The research methodology for a real-time energy management system

3.2 Benchmark EMS

In order to quantify the energy saving potential of the developed energy management system, DP and standard Q-learning are used as benchmarks for comparison.

3.2.1Dynamic programming

DP is a global optimisation algorithm widely used in EMS control. It is based on the Bellman optimality principle and decomposes the overall optimisation process into a sequence of minimisation problems, they can be solved recursively. This method is categorised as an off-line energy management technique, and it is tough to apply in online control. Therefore, a benchmark approach is provided to evaluate the performance of the proposed RL schemes.

In this research, the state variable of the hybrid off-highway vehicle is the State-of-Charge (SoC) of the battery since the vehicle speed depends on the given driving cycle. The engine control is regarded as the control variable; and the fuel consumption is calculated as different cost-to-go values at every point. The DP approach is considered as numerical. Firstly, the discretisation is conducted to deal with the continuous control problem, as shown in Fig 3-2. The horizon time (t_0 to t_T) is discretised as 1 s. The battery SoC is divided into N points (N=201, step size =

0.05, from SoC = [20%,30%], in this study). Secondly, a backward search is used to find the optimal cost function of each stage (minimum fuel consumption in this study) and the corresponding optimal control over a given driving cycle. Eventually, forward calculation is employed to formulate the optimal trajectory by forward iteration of the optimal control obtained by the backward search.

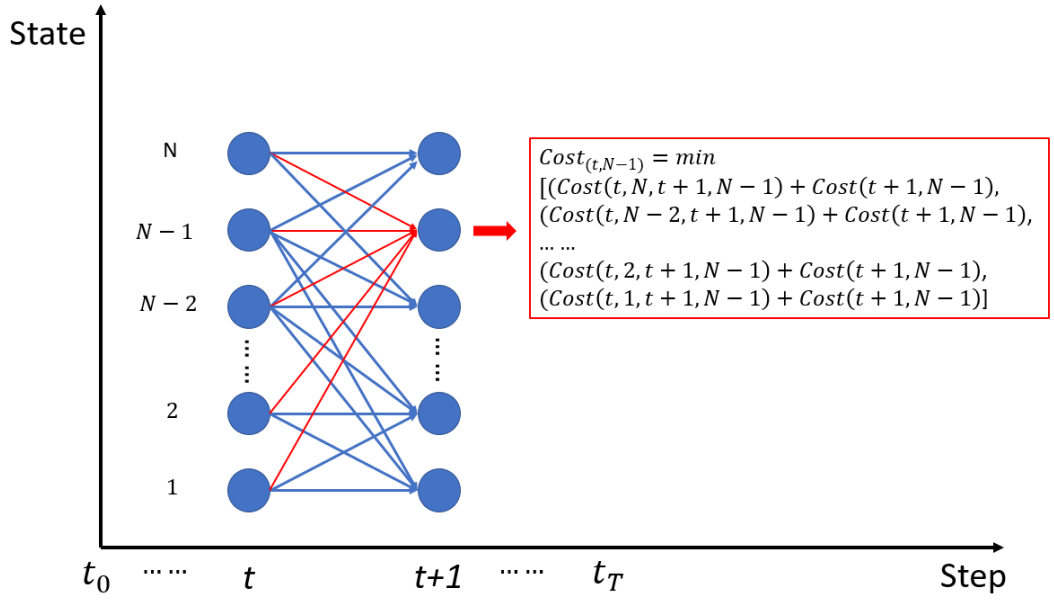


Fig. 3-2 Discretisation of the state and time space in DP

The purpose of backward computation is to find the optimal sequence of the whole process. This process starts from the last time t_T of the driving condition and ends at the initial time. Forward calculation refers to the process from the start of stage t_0 to the end of stage t_T. The optimal solution at each step can be calculated by restoring the optimal trajectory. The specific process is in Fig 3-3:

Algorithm : Dynamic Programming (DP)

% Start forward calculation%

Initial size of the P_{requre} (win), SoC_{set}

For $t = 1{:}1{:}win$

For $k = 1{:}1{:}\ length(SoC_{set})$

For $j = 1{:}1{:}\ length(SoC_{set})$

$Compute\ I_{batt}(t,k,j),\ loss_{batt}(t,k,j),\ P_{batt}\ (t,k,j);$

$Compute\ u_{egu}(t,k,j), loss_{egu}\big(loss_{batt}(t,k,j)\big),\ Cost(t,k,j);$

end

end

end

% Start backward calculation%

$V = (win\ ,\ win) = 0;$

$num_{states} = linspace\ (SoC_{set}, win);$

For $u = win$:-1:1

For $n = 1 : num_{states}$

$B_{cost}(n) = Cost\big(u, n, index(u+1)\big) + V(u, win);$

end

$V(u, n) = min(B_{cost});$

$clear\ \ B_{cost}$

end

end

Fig. 3-3 The calculation process of DP algorithm

In the forward calculation, the cost-to-go value between two nodes (discrete SoC) is calculated according to battery model and engine-generator model using the backward-looking method, and the cost represents the total energy loss and the change of SoC. The control variable is the engine-generator control signal u_{egu}, as shown in Fig.3-3, which decides power distribution among engine-generator and traction motor; then the battery pack loss $P_{loss_{bp}}$ and the engine-generator loss $P_{loss_{egu}}$ are obtained, and the third term in Eq.(3-1) is set to keep charge-sustaining, where, ω is 6000 and $SoC_{initial}$ is 25% in this study.

$$Cost(t) = P_{loss_{bp}}(t) + P_{loss_{egu}}(t) + \omega(SoC(t) - SoC_{initial}) \quad (3\text{-}1)$$

In the backward calculation, ${J_{cost}}^*(n+1)$ is optimal cost function at step $n+1$ and can be solved in the backward process to get optimal control law.

$${J_{cost}}^*(n+1) = min_u\big[Cost\big(u, n, index(u+1)\big) + {J_{cost}}^*(n)\big] \quad (3\text{-}2)$$

$${J_{cost}}^*(n) = min_u[V(u, win)] \quad (3\text{-}3)$$

3.2.2Standard Q-learning algorithm

RL is one of the essential machine learning (ML) algorithms. The basic idea of RL is to learn how to select an optimal action to obtain the maximum expected reward based on the environment. Agents are trying to learn and discover which actions can bring the greatest rewards, since good actions not only bring immediate rewards, but also affect the next state and bring all subsequent rewards. Obviously, trial-and-error search and delayed return are two critical characteristics of RL. The interaction process of RL is usually described as shown in Fig. 3-4.

The agent can perceive the state of the external environment and the reward, and then can conduct learning and decision-making. The action is the behaviour of the agent, which can be discrete or continuous, and its action space refers to all actions taken by the agent at this state. The policy is the function of the agent to decide the

next action according to the environment. Random strategies are generally used due to the better performance of exploring the environment. For the environment outside the agent, the environment changes its state under the agent's action, and then feeds back the corresponding rewards to the agent. The state is the representation of the environment which can be discrete or continuous, and its state space is all the possible states. The state transition of the environment to the next state occurs, and the immediate reward is obtained after the agent makes an action according to the current state.

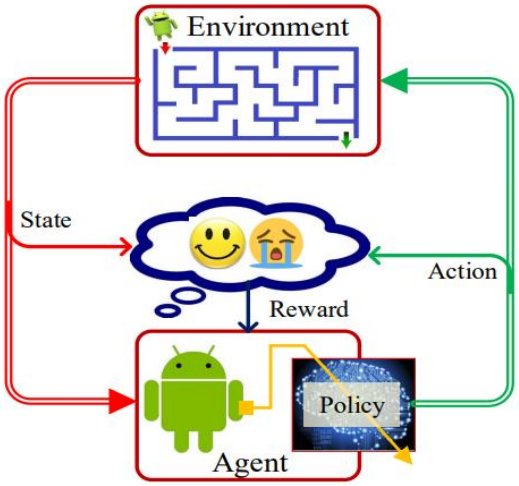


Fig. 3-4 The interaction process of RL

Through analyzing the basic principle of RL, RL can be employed in the EMS for the hybrid off-highway vehicle. And the RL-based framework of the EMS is developed, as shown in Fig. 3-5. The tabular Q-learning, which is one of the representatives of RL, is commonly used for the EMS of a series HEV because it normally requires no more than three state variables and the Q-table is capable of mapping the merit

function values with the state variable inputs.

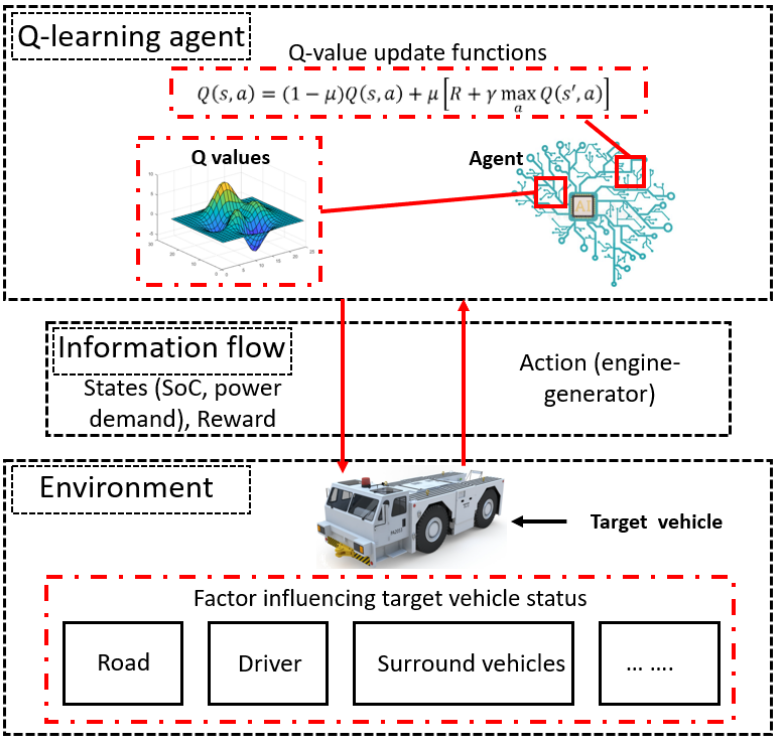


Fig. 3-5 Control architecture based on Q-learning algorithm for the studied off-highway vehicle

In this research, the basic RL algorithm is Q-learning which is used for investigating the learning performance of EMS for the studied hybrid off-highway vehicle. The vehicle information (e.g. battery SOC, driver's power demand P_{dem}) controller represents an agent, which directly takes actions (e.g. control command a) and makes interactions with the unknown driving environments (e.g. road, driver behaviour, surrounding vehicles etc.). The tasks of RL for the EMS of the HEV are usually described by using the Markov decision process (MDP), as shown in Fig. 3-6. According to the state observation, such as the power demand P_{dem} , the agent

(EMS unit) takes the action to HEV based maximum Q value policy, the reward (considering the fuel consumption from HEV) will be obtained.

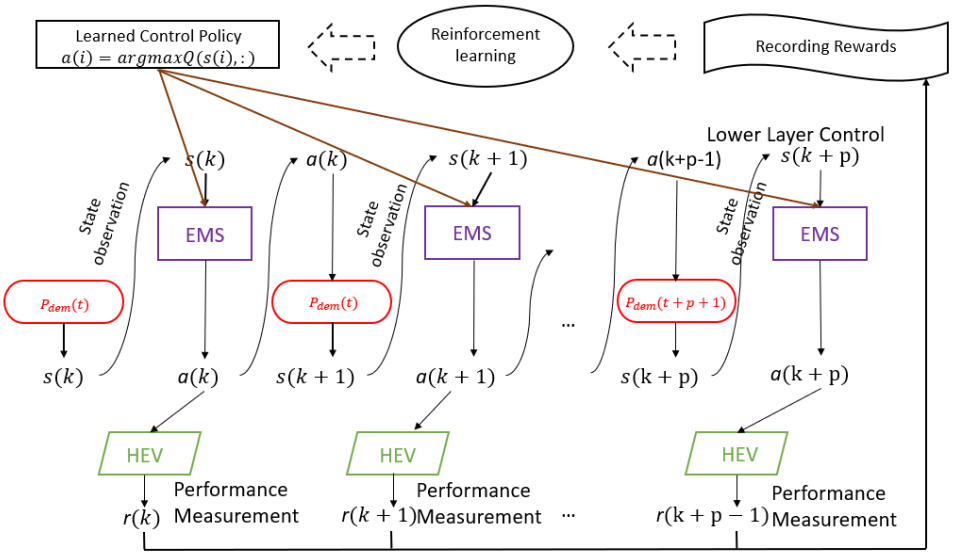


Fig. 3-6 The Q-learning algorithm based on MDP (Zhou *et al.*, 2019)

3.3 Research platform

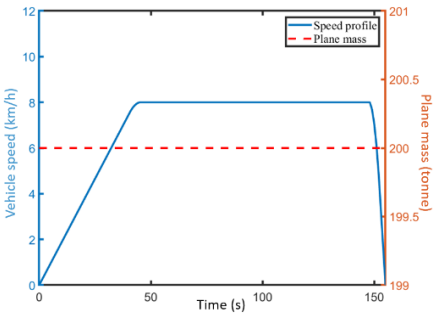
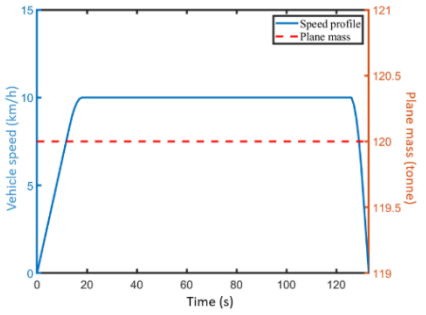
In this section, research driving cycles are firstly established; then the related facilities for developing real-time control methods are introduced, including the models of the powertrain system and subsystems, both software-in-the-loop (SiL) and hardware-in-the-loop (HiL) test platforms. The details are shown as follows:

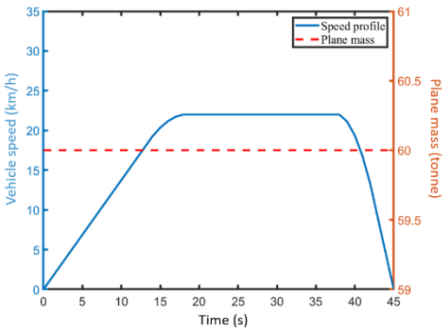
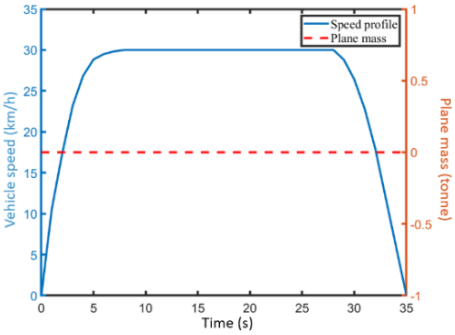
3.3.1Driving Cycles

The push-back driving cycle (PBDC) was provided by the OEM based on statistical data collected at London Heathrow airport. The PBDC consists of four typical

operating scenarios: (a) heavy pushback, (b) medium pushback, (c) light pushback, (d) solo run. Each operation of each scenario's profile is presented in Table.3-1, which includes the maximum vehicle speed, the push-back load (plane mass), acceleration time (Acc time) and driving cycle length.

Table.3-1 Cycle component profile

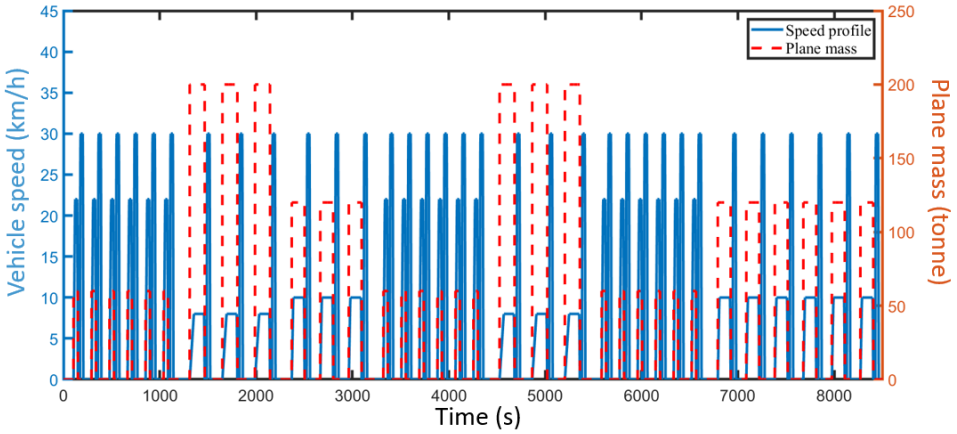
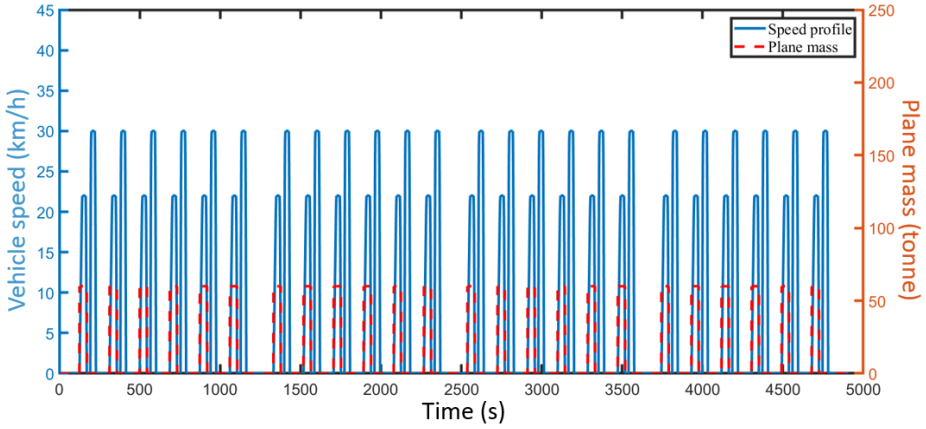
Typical working scenarios	Vehicle maximum. speed	Plane Mass	Acc. time	Cycle length
 (a)	8 km/h	200tonne	43s	155s
 (b)	10 km/h	120tonne	16s	133s

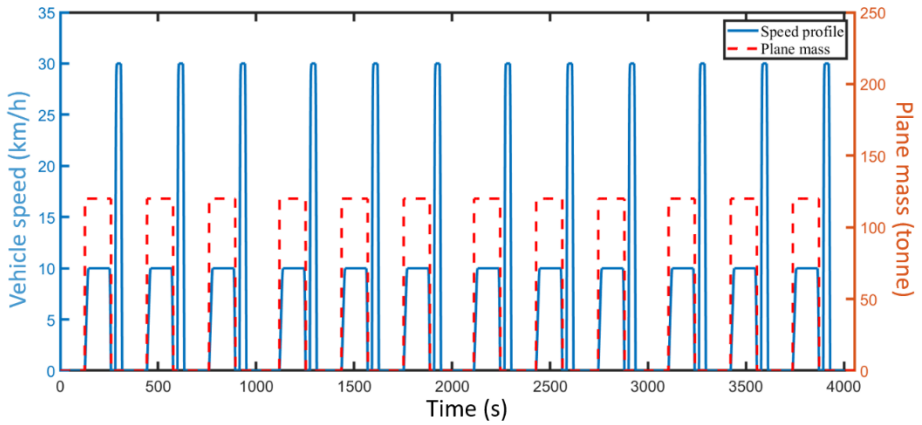
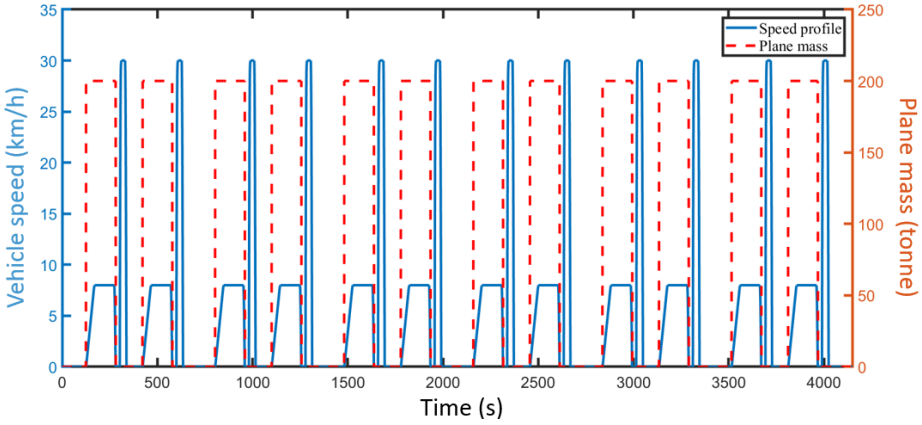
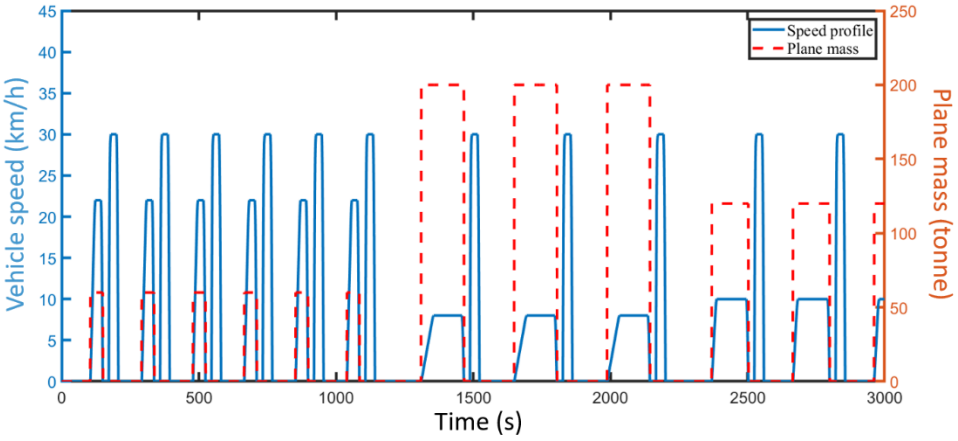
 (c)	22 km/h	60tonne	17s	45s
 (d)	30 km/h	0tonne	7s	35s

The PBDC-I gives a comprehensive driving cycle including all the possible scenarios, i.e. the tractor may pushback a heavy aircraft to the track and then come back solely (a&d scenario), or pushback a medium aircraft and then come back solely (b&d scenario), or pushback a light aircraft and then come back solely (c&d scenario). The detailed driving cycle profiles are shown in Table 3-2. Push back driving cycle 1 (PBDC-I) consists of 6 repeated a&d scenarios, 3 repeated a&d scenarios, 3 repeated b&d scenarios, 6 repeated c&d scenarios, 3 repeated a&d scenarios, 6 repeated c&d scenarios, and 6 repeated b&d scenarios, and other push back driving cycles are similar. And the PBDC-II to PBDC-IV is used to mimic unpredictable driving

conditions and the learning driving cycle is created for model-free control.

Table.3-2 Detailed profiles of all driving cycles

Cycle name	Profile of predefined driving cycles
Push back driving cycle 1 (PBDC-I)	6 c&d+3 a&d+3 b&d+6 c&d+3 a&d+6 c&d+6 b&d
	
Push back driving cycle 2 (PBDC-II)	6 c&d+ 6 c&d+ 6 c&d+ 6 c&d+ 6 c&d
	
Push back	3 b&d+ 3 b&d+ 3 b&d+ 3 b&d

driving cycle 3 (PBDC-III)	
Push back driving cycle 4 (PBDC-IV)	2 a&d+2 a&d+ 2 a&d+ 2 a&d+ 2 a&d+ 2 a&d
	
Learning driving cycle	6 c&d+3 a&d +2 b&d
	

3.3.2Modelling of the series powertrain system and subsystems

This research uses a backward-looking modelling method, which models in the opposite direction to the signal flow. And the model parameters are calibrated by our industry partner and its accuracy is acceptable for control functionality development. The modelling procedure of the studied hybrid aircraft-towing tractor is shown in Fig.3-7. The main components include vehicle dynamic modelling, EMS controller, engine-generator and battery pack, as shown below:

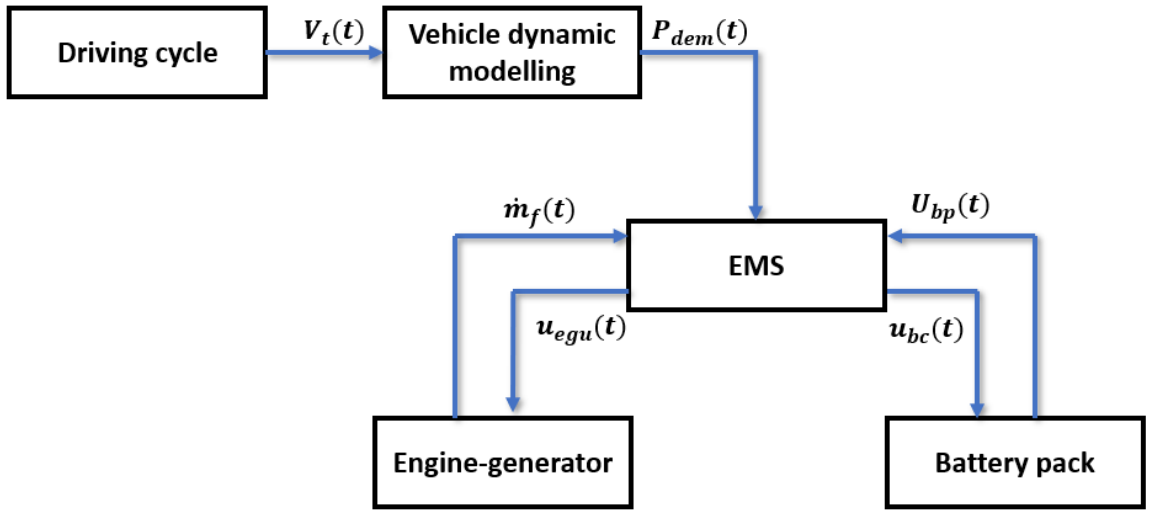


Fig. 3-7 The signal flow of the studied hybrid aircraft-towing tractor

a) Vehicle dynamic modeling

The backward-looking vehicle model is built according to Fig. 3-7, for vehicle dynamic modeling, the resistance of the vehicle mainly includes rolling resistance, aerodynamics resistance, acceleration resistance, and grade resistance. According

to Newton's second law, the studied vehicle is in equilibrium at any time, and the equation of state is:

$$F_t = \sum F \tag{3-4}$$

where, F_t is the driving force generated by the engine; $\sum F$ is the sum of the vehicle resistance, which can be calculated by:

$$\sum F = F_f + F_w + F_j + F_i \tag{3-5}$$

where, F_i is the grade resistance which is ignored in this case; F_f is the rolling resistance; F_w is the aerodynamics; F_j is the acceleration resistance. The three types of environment resistance can be calculated by

$$\begin{cases} F_w = \left(\frac{C_d A}{21.15}\right) v^2 \\ F_j = \psi m a_{acc} \\ F_f = mg(f_1 v + f_2 v^2) \end{cases} \tag{3-6}$$

where δ is the mass factor caused by the moment of inertia of four wheels and the powertrain rotating components. Here the ψ=1; m is the sum of the vehicle mass m_1 and plane mass according to above different driving cycles; a_{acc} is the acceleration; and C_d is the aero-dynamic coefficient. Furthermore, A is the fronted area; g is the gravity coefficient; and f_1, f_2 are the coefficient of rolling resistance. The detailed parameters of the studied vehicle are shown in Table 3-3.

Table.3-3 Parameters of studied vehicle

Parameters	Values
Vehicle mass m_1	17000kg
Drag coefficient C_d	0.7
Vehicle front area *A*	7.95 m^2
Gravity coefficient g	9.8
Rolling resistance coefficient f_1	0.000056
Rolling resistance coefficient f_2	0.0076

b) Traction motor modeling

For traction motor modelling, the traction motor is provided by TM4 electrodynamic system Ltd. As for the studied aircraft-towing tractor, the tractor generally works in low speed and high-load driving conditions, which means the regenerative braking system is not cost-efficient. Therefore, the traction motor is only operating in traction mode. It can be modelled as Eq. 3-6, furthermore, the tractor motor loss can be calculated by a quadratic function of the motor torque (Hu *et al.*, 2013). The efficiency of the traction motor is described Eq.3-7:

$$P_{trm}(t) = \frac{T_{trm} \times n_{trm}(t)}{9550} \tag{3-7}$$

$$\eta_{trm}(t) = \frac{P_{trm}(T_{trm}(t), n_{trm}(t))}{P_{trm}(T_{trm}(t), n_{trm}(t)) + P_{loss,trm}(T_{tm}(t), n_{trm}(t))} \tag{3-8}$$

where, the P_{trm} is the output power from the traction motor; T_{trm} and n_{trm} are the torque and rational speed of the traction motor, $P_{loss,trm}$ is the power loss of the traction motor, which can also be described as a function of motor torque. In this research, the traction motor power output is interpolated from the map shown in Fig.3-8 with tractor torque and speed as the inputs, then tractor electrical loss is integrated into the map as well.

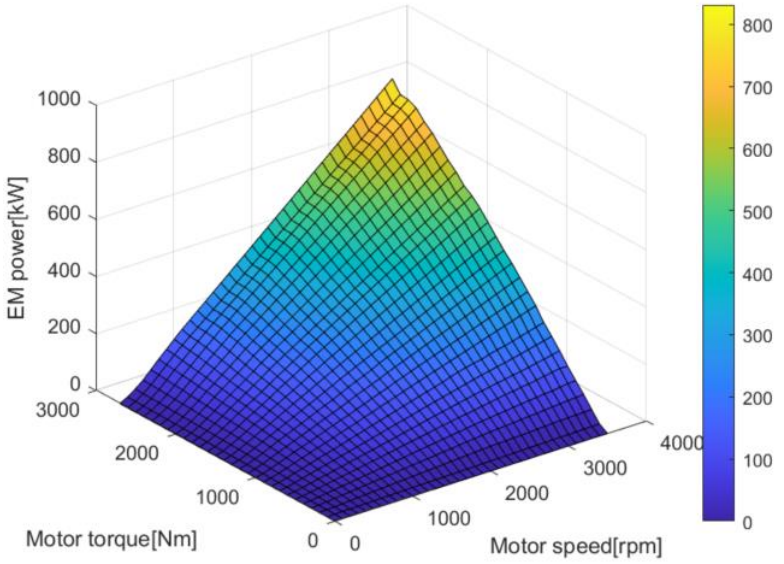


Fig. 3-8 Motor power map

c) Engine-generator modeling

For the engine-generator unit (EGU) of the aircraft-towing tractor is provided by JCB. The specification of the EGU is shown in Table 3-4. The power provided by the EGU can be modelled as a quadratic function (Hu *et al.*, 2013)

$$P_{ef} = b_2 \times P_{egu}^2 + b_1 \times P_{egu} + b_0 \quad (3\text{-}9)$$

where P_{egu} is the EGU's output power; the coefficient b_0, b_1 and b_2 are decided with curve fitting using the fuel rate data from Table 3-4. And there is the trade-off between the accuracy and the computation, thus, the second time fitting is conducted to obtain the coefficient $b_2 = -0.00035$, $b_1 = 0.2757$, $b_0 = 0.0136$ .The P_{ef} is the equivalent power fuel consumed, which can be evaluated by:

$$P_{ef} = \frac{\dot{v_f}\rho_f H_f}{3600} \tag{3-10}$$

where, v_f is the rate of fuel consumption in L/h; $\rho_f = 0.87kg/L$ is the fuel density; $H_f = 44 \times 10^6 J/kg$ is the heat value of the fuel.

Table.3-4 Engine-generator unit profile

Specifications	Value
Maximum power	86.20 kW
Frequency	50 Hz
Voltage	230V
Phase	3-AC
Fuel type	Diesel
50% load fuel rate	13.00 L/h
75% load fuel rate	18.60 L/h
100% load fuel rate	24.10L/h
Fuel tank capacity	285 L

The energy efficiency of the EGU can be calculated as

$$\eta_{egu}(t) = \frac{P_{egu}(t)}{P_{fuel}(P_{egu}(t))} \tag{3-11}$$

where the main parameters of the function P_{fuel} are calibrated by calculating the real-time $P_{fuel}(t)$ with the measured real-time fuel consumption rate $\dot{m}_f$.

d) Battery pack modeling

A lithium-ion battery pack is selected for this Ph.D. research, which contains 8200 individual battery cells provided by Panasonic Automotive & Industrial Systems Ltd. Battery package model outputs the battery state of charge to the hybrid system for energy management control. Battery cell current and voltage are used for iterative calculation to simulate the dynamics of battery cells. From the beginning of each iteration, the battery cell current I_{bc} is firstly calculated by:

$$I_{bc} = \frac{P_{bp}}{n_{bc} \cdot U_{bc}} \tag{3-12}$$

where, U_{bc} is the available voltage which can be provided by the battery cell. For lithium-ion battery package. Usually, the equivalent circuit model of a lithium-ion battery contains the Thevenin model, the Rint model(Seaman *et al.*, 2014), the Partnership for a New Generation Vehicle (PNGV) model and the multi-order RC loop model. The two RC model is chosen for modelling the current-voltage dynamics of a lithium-ion battery cell (Guo *et al.*, 2019), as shown in Fig.3-9.

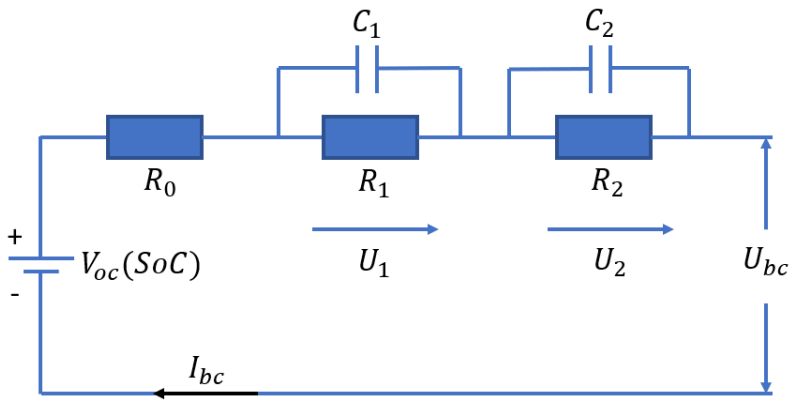


Fig. 3-9 The two RC equivalent circuit model for a battery cell

This circuit model obeys the following functions (Hu *et al.*, 2014):

$$\begin{cases} V_{oc}(SoC) = I_{bc}R_0(SoC) + U_1 + U_2 + U_{bc} \\ I_{bc} = \frac{U_1}{R_1(SoC)} + C_1(SoC)\frac{dU_1}{dt} \\ I_{bc} = \frac{U_2}{R_2(SoC)} + C_2(SoC)\frac{dU_2}{dt} \end{cases} \tag{3-13}$$

where, the V_{oc} is the open voltage of the battery cell; R_0 is the ohmic resistance (internal resistance of the battery cell). The R_1 and R_2 are the short transient resistance and long transient resistance, respectively. And the C_1 and C_2 are two capacitors for simulating the internal dynamic of the battery cell (e.g., polarization), in which, the C_1 is the short transient capacity and C_2 is the long transient capacity. And the C_1, C_2, R_1, R_2 are functions of battery's state-of-charge (Min Chen *et al*, 2006). The battery's SoC can be evaluated as follows:

$$SoC_{now} = SoC_{past} - \frac{\int_{t_0}^{t_T} I_{bc}\,dt}{Q_{bc_maz}} \tag{3-14}$$

where, SoC_{now} is current battery SoC; SoC_{past} is the past battery SoC; and Q_{bc} is the

capacity of the battery cell. The parameters of the studied battery cell are presented in Table.3-5.

Table.3-5 Parameters of the studied battery cell

Parameter	Description	Value
V_{bc_max}	Rated battery cell voltage (V)	3.6
I_{bc_cmax}	Rated battery cell charge current (A)	2.25
I_{bc_dmax}	Rated battery cell discharge current (A)	11.25
Q_{bc_max}	Rated battery cell capacity (mAh)	2900
V_{oc}	$1.031 \cdot e^{-35 \cdot SOC} + 3.685 + 0.2156 \cdot SOC - 0.1178 \cdot SOC^2 + 0.3201 \cdot SOC^3$	
R_1	$0.1562 \cdot e^{-24.37 \cdot SOC} + 0.07446$	
R_2	$0.3208 \cdot e^{-29.14 \cdot SOC} + 0.04669$	
R_0	$6.603 \cdot e^{-155.2 \cdot SOC} + 0.04984$	
C_1	$-752.9 \cdot e^{-13.51 \cdot SOC} + 703.6$	
C_2	$-6056 \cdot e^{-27.12 \cdot SOC} + 4475$	

3.3.3HiL testing platform

In order to validate the real-time feasibility of studied control strategies, the HiL test is need to ensure that the optimal control rules based on reinforcement learning can

be calculated within 1 second. The HiL platform is built with ETAS vehicle control development (Shuai *et al.*, 2020), as shown in Fig 3-10. The control strategies were developed using MATLAB/Simulink initially, and then be compiled into C-code using ETAS' INTECRIO/INCA software in Host PC-1. The complied control strategies are deployed into an ETAS ES910 prototype controller through a USB-to-CAN interface. DESK-LABCAR implements the real-time model of the hybrid aircraft towing tractor on host PC-2 via Ethernet. The real-time vehicle model was developed in Simulink and was verified using the testing data from a prototype vehicle(Zhou *et al.*, 2017). The LABCAR and the ES910 were connected via a CAN bus to emulate the communication between the controller and the vehicle plant in real-time. The performance of the vehicle was monitored and recorded by ETAS' Experimental Environment.

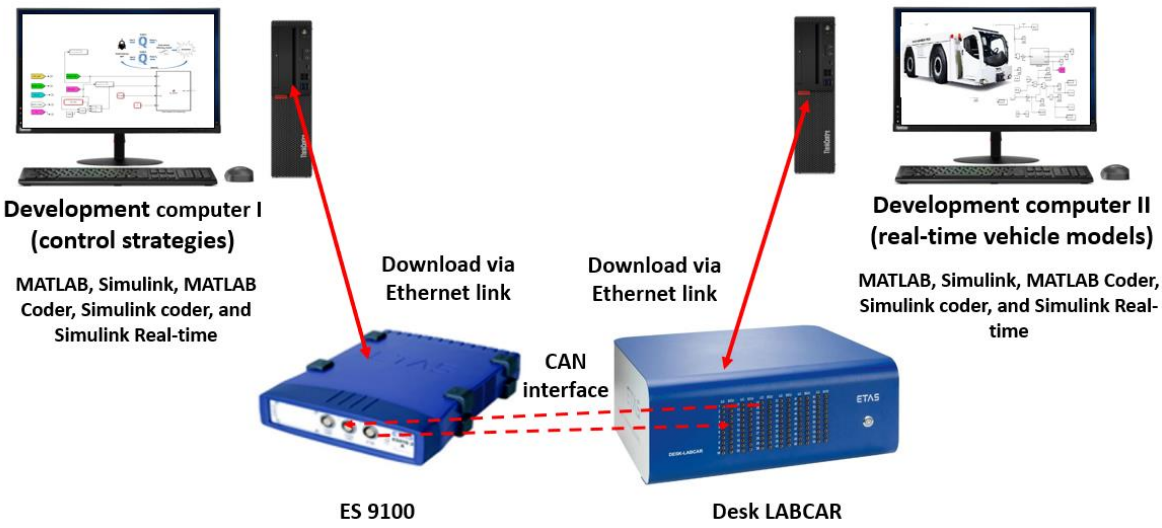


Fig. 3-10 HiL testing platform (Shuai *et al.*, 2020)

3.3.4Software for the HiL testing

The software for the HiL testing is used mainly for modelling, compiling and implementation, as shown in Fig.3-11. The software is installed in host-PCs which communicate with ETAS hardware via Ethernet. The real-time models for the hybrid vehicle and energy management controller are built by MATLAB/Simulink. The real-time models are connected with the hardware interface (e.g. CAN) and compiled into C-code for real-time computing in the compiling procedure. The ETAS EE and INTECRIO are the model compiling software used for LABCAR and ES910, respectively. Complied models of vehicles and controllers are implemented in the LABCAR and ES910 through ETAS Experiment Environment (EE) and INCA, respectively. The model performance in LABCAR and ES910 can be monitored in the host-PCs through EE and INCA.

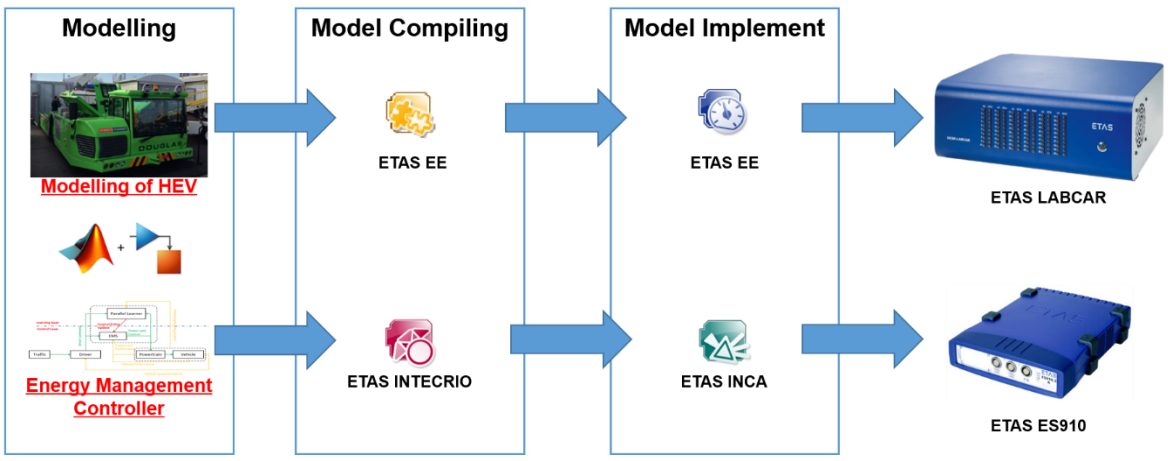


Fig 3-11 Software for HiL testing system (Zhou *et al.*, 2019)

3.4 Summary

The work conducted in this chapter includes:

- The research methodology of this thesis is introduced, in which the RL-based is the tabular Q-learning algorithm. And the DP and standard Q-learning is set as benchmarks for the presented EMS. Then the RL in terms of these advantages can be employed in EMS effectively compared with other optimization algorithms.

- The research platform is described, including the predefined driving cycles, the detailed series powertrain system and subsystems; the HiL testing platform; and software for the HiL testing.

- The real-time vehicle models with backward-looking method, including vehicle dynamic modeling, traction motor modeling, engine-generator modeling and battery pack modeling, is built to design EMS with RL-based method and validate the performance of the proposed algorithms on the HiL testing platform.

Chapter Four

EXPLORATION-TO-EXPLOITATION RATIO IN REINFORCEMENT LEARNING CONTROL

This chapter studies the impact of the internal characteristics of Q-learning on the performance of hybrid off-highway vehicles, one of which is how to balance the exploration and exploitation during learning process. The influencing factors studied in this chapter include: 1) different types of the reward function in RL settings; 2) three decay functions, ‘exponential decay function', ‘reciprocal function-based' and 'step-based', generating different trajectories of the epsilon values for exploration and exploitation in RL; and 3) balancing of the 'exploration-to-exploitation ratio' (E2E) for RL-based EMS in the hybrid off-highway vehicle.

4.1 Problem formulation

As for energy management system of hybrid off-highway vehicles, it can be considered as a constraint in the optimization model. Therefore, the energy allocation problem based on energy efficiency and battery SoC is transformed into a nonlinear optimization problem, the energy optimization function, and constraints of which are described mathematically in the following section.

The problem of energy management is the use of a series of discrete controls to optimise the performance over a specific time horizon. When employing RL-based method, the state transition is governed by the discrete time equation $s(t+1) = f(s(t), a(t))$, $a(t)$ is the control actions at the state $s(t)$. Then an immediate cost $P_{loss_{tot}}(t)$ occurs at each step, the total energy cost is the sum of all of each immediate cost, the problem of minimizing the expected total cost J in energy management is described as follows:

$$min\, J = \sum_{i=1}^{\infty} c(s(t)\,, a(t)) \tag{4-1}$$

$$s.t. \quad s(t+1) = f(s(t), a(t)) \tag{4-2}$$

$$SoC \geq SoC_{hrd} \tag{4-3}$$

where $(s(t)\,, a(t))$ is the state-action pair at the time step t. $c(s(t)\,, a(t))$ is the immediate cost incurred at each step. SoC_{ref} is set to 0.28 in this study. At the same time, the entire vehicle powertrain needs to meet physical constraints, including the battery pack and the engine-generator.

4.1.1Reward function

The reward function describes how an agent 'should' behave. It has 'normative' content that specifies what you want the agent to accomplish. There is no absolute

limit to this. If the agent learns better, the reward function will be 'better behaviour'. In practice, this means convergence speed and not getting stuck in local minima. Therefore, determining an appropriate reward function in the EMS of the hybrid off-highway vehicle plays an essential role in saving energy. In order to comply with the rules of Q-learning, the agent selects the action with the maximum reward in standard Q-learning. The criteria for developing the reward function should consider key factors, such as the maximum reward value for the current state corresponding to the minimum total power loss (battery loss and engine-generator loss) for the current state.

4.1.2Exploration to exploitation ratio

Exploration of new control settings and exploitation of the existing control policy are two critical procedures in RL (Hu *et al.*, 2020). The agent makes a trade-off between exploration and exploitation based on the epsilon-greedy policy to achieve better learning performance (Carroll *et al.*, 2001).The process of exploration and exploitation is shown in Fig 4-1.

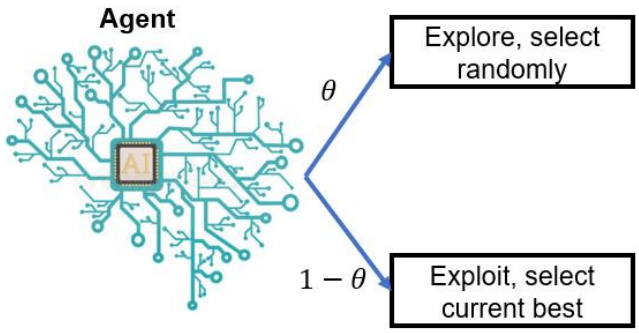


Fig. 4-1The exploration and exploitation process in RL

During the action execution process, there is θ probability that the agent takes actions with maximum value at the current state. Moreover, there is a 1- θ probability that the agent takes random actions. If the majority of the experience from exploration is not as good as the best existing returns, over-exploration may impact the learning performance of RL. On the contrary, without enough exploration, the exploitation may lead the agent to be trapped with only sub-optimal solutions (Wang *et al.*, 2020). Therefore, a reasonable value for θ is necessary to achieve a good balance and reduce the bias during the learning process (Wang *et al.*, 2013). Some research has been carried out by setting the value of θ as a constant, e.g. θ = 0.25 (Xu *et al.*, 2019); θ = 0.7 (Qi *et al.*, 2015); θ =0.1 (Liu *et al.*, 2014); or as a gradually decreasing value, e.g. θ = 0.8/θ_{end} = 0.6 (Qi *et al.*, 2016); θ = 1.0/θ_{end} = 0.1 (Qi *et al.*, 2019). However, exploration and exploitation in the whole learning process still suffer from the dilemma of balancing the probability of exploration (Wang *et al.*, 2013)

The exploration-to-exploitation ratio (E2E) is proposed to evaluate exploration

probability by taking the whole learning process into consideration. The E2E can be calculated by:

$$E2E = \frac{P_{exploration}}{P_{exploitation}} \tag{4-5}$$

where $P_{exploration}$ is a probability of exploration at each iteration and $P_{exploitation}$ is a probability of exploitation at each iteration. Fig. 4-2 also illustrates how to define E2E in the whole learning process. The value of θ follows the trajectory of the exponential decay function with 35 iterations of learning. For example, at the 10th learning iteration, the probability of exploitation and exploration can be divided by the solid purple line. And the agent needs to choose appropriate actions in various states based on previous experience to gain better rewards by exploiting known actions; in the meanwhile, requisite action exploration is required to discover these actions. Therefore, E2E represents the trade-off proportion between the exploitation and exploration process.

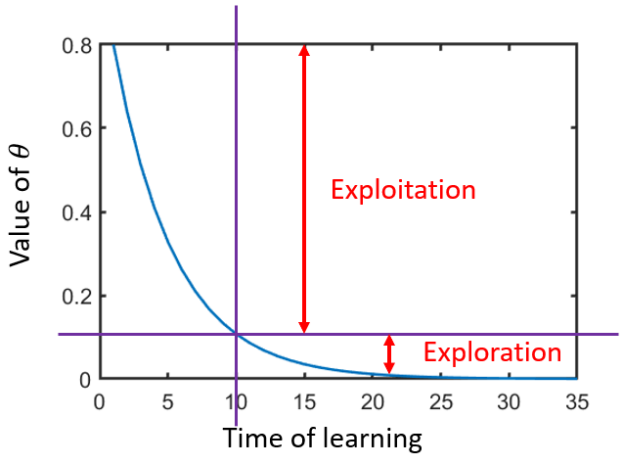


Fig. 4-2 One example of exploration to exploitation

4.2 Framework of the model-free supervisory control system

The model-free control system is developed with a layered energy management system (Zhou *et al.*, 2019) as shown in Fig.4-3; which includes two layers: the control layer (installed in the onboard tractor controller) and the learning layer (deployed in the server computer). The two layers communicate through the V2X network. There are four function modules to enable the model-free charge-sustaining control. Three function modules, the state's perception module, the action execution module, and the reward assessment module, are located in the control layer. The RL module is located in the learning layer. The inputs and outputs of each layer and algorithms used to enable the functionalities are described as follows:

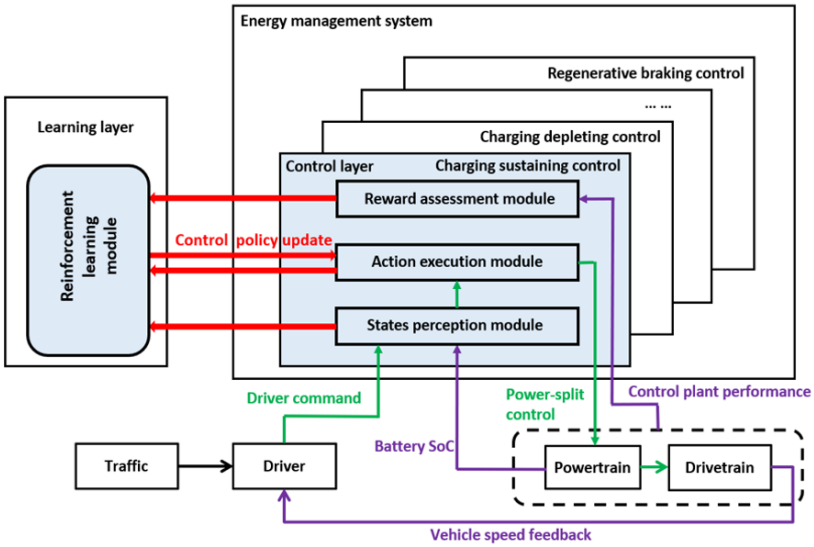


Fig. 4-3 Layered control framework for model-free supervisory control (Shuai *et al.*, 2020)

4.2.1States perception module

The states perception module is responsible for determining the current state of the vehicle based on the feedback. The battery SoC and the driver's power demand are selected as the state variables for the best performance of the learning system with the minimum computational effort (Zhou *et al.*, 2019). The precepted state variables will be sent to the RL module for the control policy and the action execution module, to control the downstream controllers. The variables are measured at each sampling time and discretised into the finite state vector,

$$s(t) = [P_{dem}(t), SoC(t)] \tag{4-6}$$

where, $s(t)$ is the current state at the t^{th} time step; $P_{dem}(t) \in \{0\text{kW} \leq P_{dem} \leq 253\text{kW}\}$ is the driver's power demand value at the t^{th} time step; $SoC(t) \in \{20\% \leq SoC \leq 80\%\}$ is the battery SoC value at the t^{th} time step.

4.2.2Action execution module

The action execution module connects with the state perception module, the RL module and the downstream controllers including the engine-generator controller and the battery management controller. The primary purpose of the action execution module follows the action execution policy, to implement the control signal to the downstream controllers (e.g. engine controller). The action taken in each sampling

time will be measured and forward to the RL module. The execution policy π_{exe} is used to determine the action $a(t) \in \boldsymbol{A}$ based on the current vehicle state $s(t)$, which is defined below:

$$a(t) \leftarrow \pi_{exe}(\mathbb{Q}(s(t), \boldsymbol{A})) \tag{4-7}$$

where, $\mathbb{Q}$ is the knowledge base that stores the merit-function values corresponding to different state and action variables; In standard Q-learning, there is only one Q-table. The policy π_{exe} is the action execution policy, which will be continuously optimised via RL; $a(t)$ is the action value at the time t^{th} time step, and it will be used to control the power rate of the engine-generator:

$$u_{egu}(t) = a(t) \tag{4-8}$$

where, $u_{apu}(t) \in \boldsymbol{A} = \{0\% \leq u_{apu} \leq 100\%\}$. After obtaining the control signal from the controller, the requirement for the battery pack can be obtained using:

$$u_{bp}(t) = \frac{P_{dem} - u_{egu}(t) \cdot P_{egu_max}}{P_{bp_max}} \tag{4-9}$$

where, P_{bp} is the power supplied by the battery package; P_{egu_max} = 86.2 kW is the maximum power that can be supplied by the engine generator; P_{bp_max} = 365 kW is the maximum power that can be supplied by the battery pack;

4.2.3Reward assessment module

The reward assessment module measures the powertrain performance, including energy consumption and the remaining battery SoC. It uses a merit function to evaluate vehicle performance after taking each action. The merit function calculated at each sampling time will be sent to the RL module. This helps the training of the optimal control policy which minimises the vehicle total power loss P_{loss_tot} while maintaining battery SoC simultaneously. In this chapter, two types of reward functions are compared in terms of learning performance. They are shown as follows:

The Reward Function A is formulated as(Qi *et al.*, 2015):

$$r(t) = \begin{cases} \frac{1}{P_{loss_{tot}}(t)+1} & SoC(t) \geq SoC_{ref} \\ \frac{1}{P_{loss_{tot}}(t)-\tau\left|SoC_{ref}-SoC(t)\right|} & SoC(t) < SoC_{ref} \end{cases} \quad (4\text{-}10)$$

and the Reward Function B is designed as (Zhou *et al.*, 2019):

$$r(t) = \begin{cases} r_{ini} - P_{loss_{tot}}(t) & SoC(t) \geq SoC_{ref} \\ r_{ini} - P_{loss_{tot}}(t) - \tau\left|SoC_{ref} - SoC(t)\right| & SoC(t) < SoC_{ref} \end{cases} \quad (4\text{-}11)$$

where, r_{ini} is set to 1; SoC_{ref} is the reference battery SoC value that is chosen to maintain the battery SoC within an acceptable range (for the best performance and health of the battery SoC_{ref} should be 28%). Then τ is a scale factor to balance the consideration of the battery SoC level and power efficiency; $P_{loss_{tot}}(t) = P_{loss_{egu}}(t) +$

$P_{loss_{batt}}(t)$ is the total power loss of the engine and battery. The power loss of the engine $P_{loss_{egu}}(t)$ and the power loss of the battery $P_{loss_{egu}}(t)$ can be calculated by:

$$\begin{cases} P_{loss_{egu}}(t) = \dot{m_f}(t) \cdot H_f - \frac{T_{eng}(t) \cdot n_{eng}(t)}{9550} \\ P_{loss_{bp}}(t) = R_0(SoC) \cdot I_{batt}(t)^2 \end{cases} \quad (4\text{-}12)$$

where, $\dot{m_f}$ is the real-time measurement of the fuel rate in kg/s; T_{egu} and n_{egu} are the engine torque and speed; I_{batt} is the current of the battery pack; H_f is the heat value of fuel (for diesel, $H_f = 44 \times 10^6$J/kg); R_0 is the equivalent internal resistance of the battery and a function of the battery SoC.

4.3 Three decay functions

In this research, three common decay functions, **exponential decay, step-based decay,** and **reciprocal function-based decay** are chosen for seeking an appropriate trajectory E2E schedule. The details of each decay function are shown below:

a) Exponential decay function

The exponential decay (EXD) function is a common method used to generate a reasonable set of values. The mathematical exponential decay function can be formed as follows:

$$\theta = \theta_1^k \tag{4-13}$$

where k is the current learning iteration; θ_1 is the initial value for epsilon value. The Fig.4-4 illustrates the variation of θ with exponential decay function in initial values $\theta_1 = 0.8$ and 0.5 under 40 times of learning iterations.

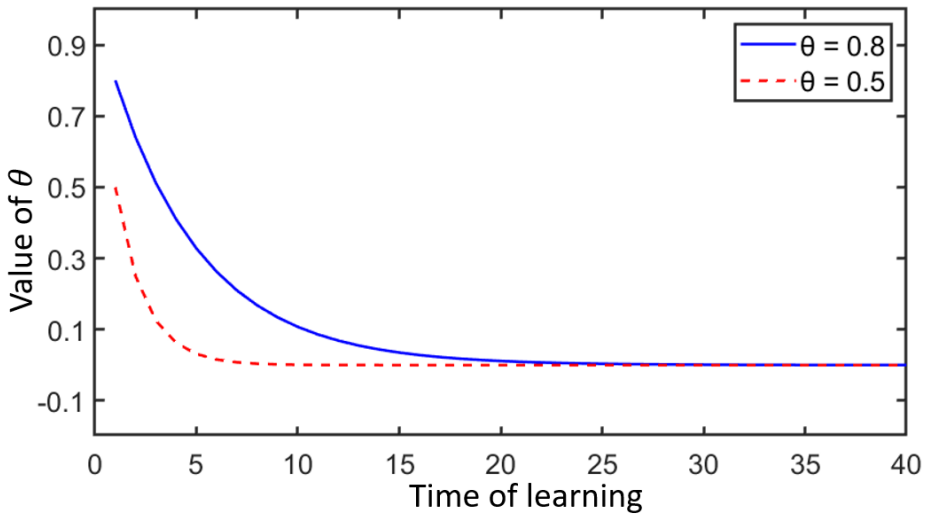


Fig. 4-4 The variation of the exponential decay function (initial $\theta_1 =$ 0.8 and 0.5)

b) Step-based decay function

The step-based decay (SBD) function systematically drops the value of epsilon like a 'ladder', which can be seen as a piecewise function. The step-based decay function can be defined as below:

$$\theta = \theta_1 \times F^{round(\frac{1+k}{D})} \tag{4-14}$$

where F is a factor value controlling the rate in which the epsilon value drops. For example, the larger factor F is, the slower the epsilon value will decay. Conversely, the smaller the factor F, the faster epsilon value will decay. D is the scale factor (typically the θ decreases after ten times of learning) that controls how many learning iterations the θ drops. For example, the Fig.4-5 presents the variation of the step-based decay function with the initial value of $\theta_1 = 0.8$ and $D = 5$ and 10 under 125 times of learning.

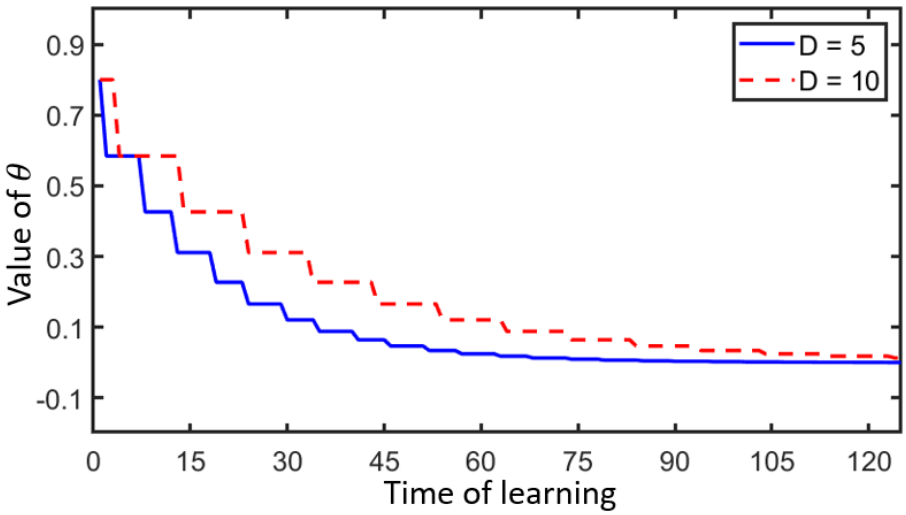


Fig. 4-5 The variation of epsilon value θ with step-based decay function (initial θ = 0.8, D = 5 and 10)

c) Reciprocal function-based decay method

The mechanism of the reciprocal function-based decay (RBD) method is very similar to the stochastic gradient descent method (SGD). It drops the fastest than the

exponential decay and step-based decay method. The reciprocal-decay is described as follow:

$$\theta = \frac{\theta_1}{1+R_{decay}*D} \tag{4-15}$$

where R_{decay} is a decay rate; If the $R_{decay} = 0$, there is no influence on epsilon value; If the R_{decay} is specified, which will decrease R_{decay} by the given fixed amount. Fig.4-6 visualises the reciprocal function-based decay methods under 125 times of learning with $\theta_1 = 0.8$ and 0.5, respectively.

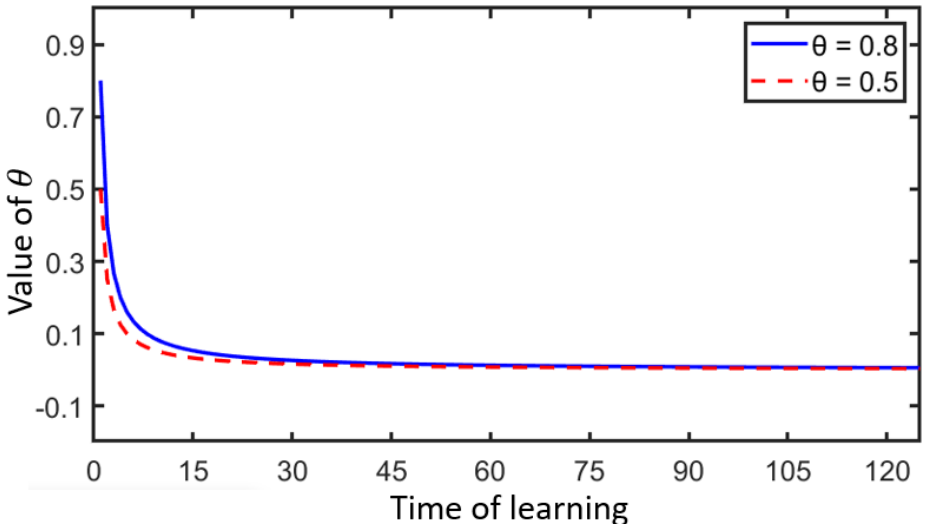


Fig. 4-6 The variation of epsilon value with reciprocal function- (initial θ_1=0.8 and 0.5)

4.4 Results and Discussion

The learning performances of RL-based EMS with two reward functions, and three

decay functions are investigated by tacking the vehicle energy efficiency in the MATLAB/Simulink. Each of them is conducted under the PBDC-1, repeated for 20 times with an initial battery SoC of 50%. A model-free charge-sustaining control performance is initially investigated by monitoring the vehicle's energy efficiency at different learning stages on the SiL platform. The vehicle's energy efficiency is defined by:

$$\eta_{veh} = \frac{\Sigma(P_d)}{\Sigma(P_{tot_loss})+\Sigma(P_d)} \quad (4\text{-}16)$$

where, the η_{veh} is the vehicle's energy efficiency; P_d is the demand power; and the P_{tot_loss} is the total power loss.

4.4.1Impact of different reward functions

In order to find a better form of the reward function, the learning performance of the designed reward function A and reward function B is compared as shown in Fig 4-7, where the solid pink line represents reward function A and the solid blue line represents reward function B.

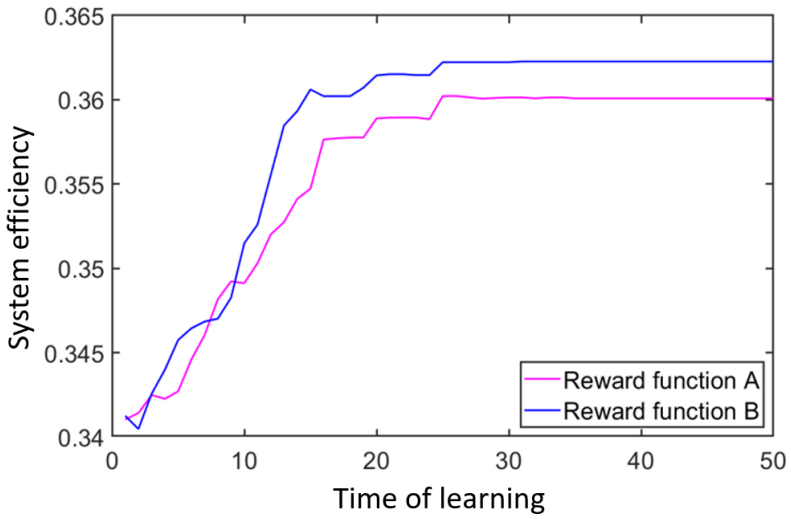


Fig. 4-7 Learning performance of two reward functions

At the 1st learning round, both reward functions achieved the vehicle energy efficiency of approximately 34.08%. Then, the energy efficiency of value function B increases faster in the following iterations, surpassing reward function A by the 10th time of learning. At the end, the reward function B achieved an energy efficiency for the vehicle of 36.23%, up to 0.06% higher than the reward function A, and the improvement rate was 6.31%. To be specific, the vehicle energy efficiency at the 10th, 20th, 30th, 40th and 50th iterations is shown in Table 4-1.

Table.4-1 Vehicle energy efficiency for the two reward equations at different numbers of iterations

Reward functions	10th	20th	30th	40th	50th
A	34.91%	35.89%	36.01%	36.01%	36.01%
B	35.15%	36.14%	36.22%	36.23%	36.23%

The reward function B surpassed A in the 10th learning round and kept this advantage

until the 50th learning round, which means the reward function B converged faster than reward function A. Therefore, the reward function B will be selected as a study for the later of E2E with three decay functions.

4.4.2Learning performance of RL-based EMS with three decay methods

The first attempt to investigate the learning performance of E2E with three decay methods was conducted under the PBDC-1, with an initial battery SoC of 50%, and an initial epsilon value of 0.8. The results of the learning performance, including the SBD, RBD and EXD methods, are demonstrated in Fig. 4-8. In order to further analyse the critical difference between the three methods, the E2E trajectories of the three methods are shown in Fig. 4-9. Before the 20th learning iteration, the vehicle energy efficiencies of three model-free control methods increase at the same speed. At the end of the learning process, the SBD method achieved the highest energy efficiency among the three action selection methods, up to 1.69% and 0.72% higher than the EXD and RBD methods, respectively. Additionally, the SBD method achieved the highest improvement rate of 6.21% (final vehicle energy efficiency compared with initial vehicle energy efficiency).

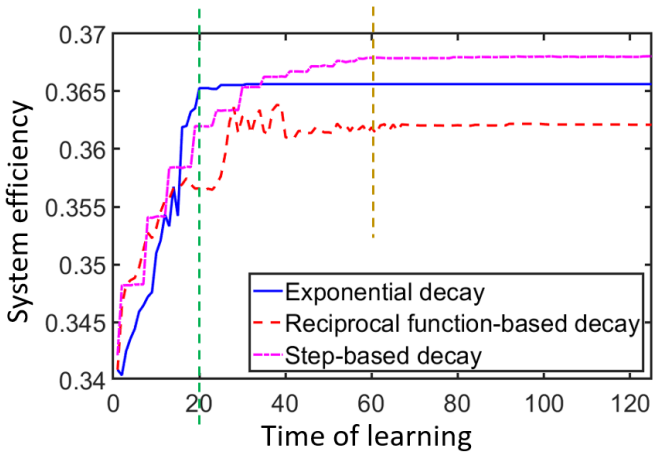


Fig. 4-8 Vehicle system efficiency curves obtianed by the RL algorithms with three decay functions (initial battery SoC is 50%)

By analysing the variation of the three decay methods in Fig 4-9, the SBD balanced the E2E slightly. More specifically, between the 20th and 60th learning iterations, the E2E of the SBD indicates that the RL agent still had tremendous potential to explore new engine control signals. These new control signals are likely to be better than the actions stored in the current Q-table, which will efficiently reduce energy consumption. On the contrary, the E2E trajectory generated by the EXD and RBD methods decreased rapidly. After the 20th learning round, both the RL agents of the EXD and RBD almost stopped exploring new engine control signals and repeatedly used their existing experience to control the engines.

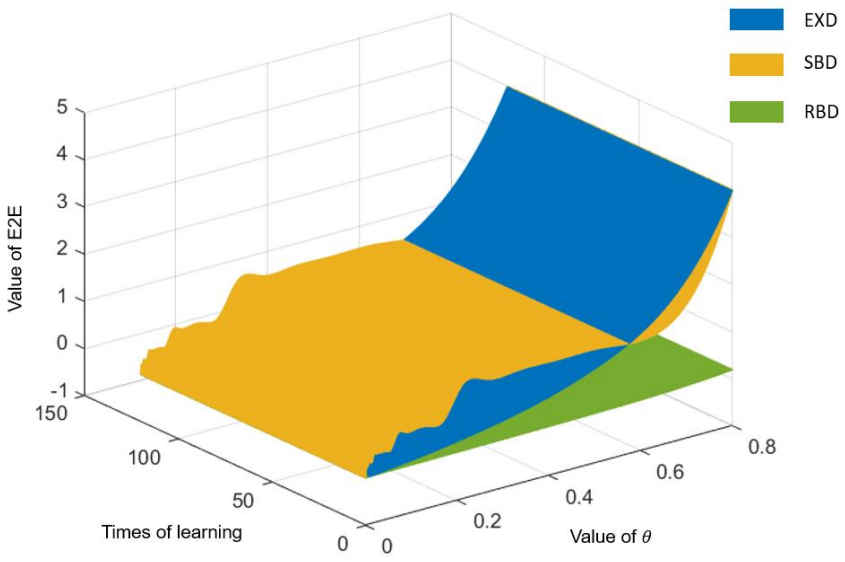


Fig. 4-9 Value of E2E ratios over the different epsilon values and rounds of learning (the intial value of θ is 0.8)

4.4.3Different initial epsilon values

This investigation was carried out in order to explore the best initial epsilon value for the exploration probability in the studied vehicle. The influence of the different initial epsilon values was studied with the three decay functions, from 0.1 to 0.9 via experimental methodology. The experiment was conducted by repeating the PBDC-1 for 20 times with an initial battery SoC of 50%. Table. 3-7 compares the vehicle energy efficiency obtained by the three decay methods. The results indicate that the initial epsilon value has little impact on vehicle performance. All three decay methods can achieve their highest performance when the initial epsilon value is 0.8, which demonstrates the initial epsilon value has an essential impact on the energy

efficiency. The average energy efficiency values obtained by the RL with the EXD, SBD and RBD methods are 36.59%, 36.77% and 36.44%, respectively. The SBD outperforms the EXD and RBD when the epsilon value is 0.8, with the highest energy efficiency of 36.84% achieved at the end of learning. Therefore, the initial epsilon value is set to 0.8 in the subsequent validation.

Table.4-2 Different initial values for E2E with three decay functions

Epsilon values	EXD	SBD	RBD
0.1	36.71%	36.71%	36.51%
0.2	36.74%	36.73%	36.40%
0.3	36.72%	36.76%	36.31%
0.4	36.70%	36.73%	36.27%
0.5	36.67%	36.74%	36.37%
0.6	36.09%	36.81%	36.35%
0.7	36.63%	36.78%	36.51%
0.8	**36.77%**	**36.84%**	**36.67%**
0.9	36.31%	36.83%	36.61%
Average	**36.59%**	**36.77%**	**36.44%**

In order to explore whether this decay method still has the protentional ability to achieve the highest vehicle performance, a further investigation was carried out regarding the learning performance of the step-based decay method under different driving conditions. The RL-based EMSs with EXD and SBD functions were studied under PBDC-2 and PBDC-3 with a battery SoC of 50%, repeating each of them for 20 times. The results are shown in Table.4-3.

Table.4-3 Investigation of $\theta = 0.8$ under PBDC-2 and PBDC-3 with an initial battery SoC of 50%

Epsilon = 0.8	PBDC-2 (4908s)	PBDC-3 (4152s)	Improvement (PBDC-2)	Improvement (PBDC-3)
EXD	35.54%	35.45%	-	-
SBD	36.24%	36.94%	1.97%	4.21%

The RL-based EMS with the SBD still outperforms the one with EXD under both PBDC-2 and PBDC-3. It achieved 36.24% and 36.94% energy efficiency under PBDC-2 and PBDC-3, respectively, compared to the EXD, it can obtain a 1.97% improvement in PBDC-2 and a 4.21% improvement in PBDC-3.

4.4.4Real-time performance

The real-time performance of the RL-based EMS with three decay functions are evaluated based on the HiL testing system developed in Chapter 3. The standard Q learning (using EXD decay function) was used as the benchmark for comparison. The real-time vehicle performance data, including total vehicle energy loss, battery SoC and total battery energy loss, was collected from the HiL system and compared in Fig.4-10. The SBD can save more than 1.52% total energy loss from the benchmark. The primary energy savings is contributed from the battery total loss, which is 12.63% lower than the benchmark.

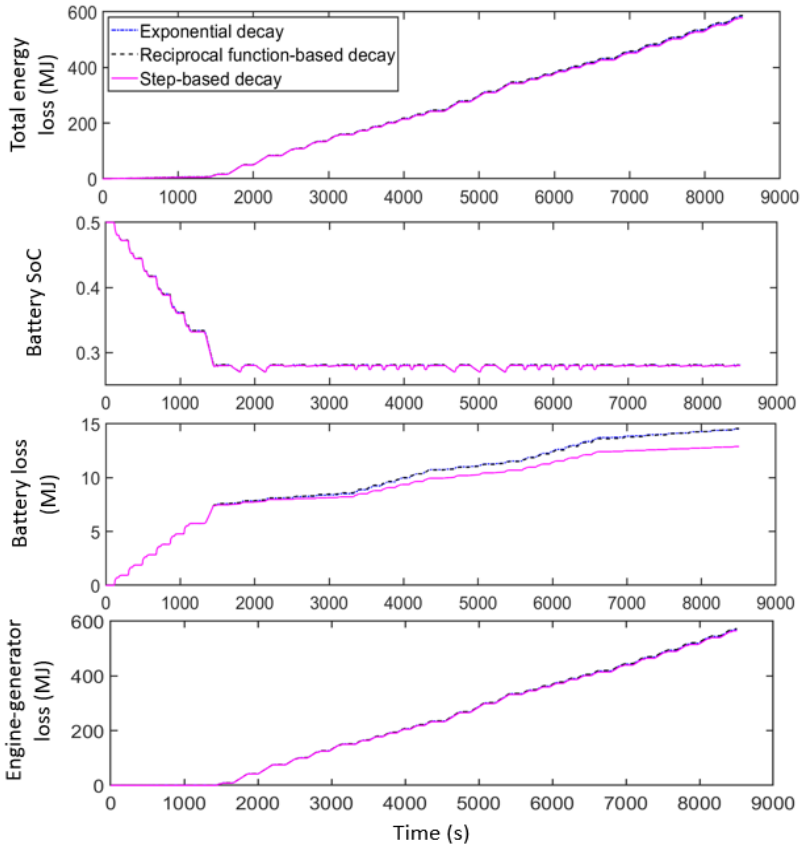


Fig. 4-10 Real-time performance of RL-based EMSs with different decay functions (Initial battery SoC = 50%)

4.4.5Robustness testing

In actual driving situations, the EMSs for HEVs will encounter various driving conditions. Therefore, a robustness test is required to explore the real-time feasibility of the three decay function methods. The results are given in Table.4-4, including the initial battery SoC, the battery SoC at the end of the learning iteration (End battery SoC), the total energy usage and the saving rate (Savings). The RL-based EMS with SBD function is shown more adaptive than the ones with RBD and EXD in all

predefined driving cycles (i.e PBDC 2-4). Compared to the RL-based EMS with EXD function, the RL-based EMS with SBD can save at least 1.25% energy, and the highest saving rate was 2.04% in PBDC-2.

Table.4-4 Real time performance of E2E with the EXD, RBD and SBD under three predefined driving conditions

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

4.5 Summary

Two types of reward functions in the RL-based EMS are investigated to improve the vehicle’s energy efficiency. Three decay functions, 'exponential decay' (EXD), 'step-based decay' (SBD), and 'reciprocal function-based decay' (RBD) are introduced to study the impacts of E2E and epsilon values to the learning performance of the RL-based EMS. The learning performances of the RL-based EMS with the three decay methods were compared by monitoring the vehicle’s energy efficiency based on SiL

and HiL platforms under four driving cycles. The conclusions drawn from this work are as follows:

- The SiL testing of RL-based EMS using different reward function suggested that the negative value of the vehicle energy consumption (Reward Function B) achieved higher vehicle energy efficiency than putting the vehicle energy consumption on the denominator of (Reward Function A) to minimise the fuel loss. The RL-based EMS with Reward Function B can lead to an efficiency improvement rate (the energy efficiency from the 1st learning to the end of the learning round) of 6.31%.

- The investigation on RL-based EMS with different decay functions on the SiL platform suggests that the RL-based EMS with SBD function can achieve the highest efficiency improvement rate of 6.21%, which is 1.36% and 0.33% higher than the RBD and EXD methods, respectively.

- By evaluation the vehicle performance under the HiL platform, the real-time control functionalities of the RL-based EMSs are validated, then the robustness testing is conducted, and the results show that the RL-based EMS with SBD-function is the most robust over the other two. It can save more than 1.25% efficiency improvement compared to the standard Q learning with EXD function in the predefined driving cycles, i.e. PBDC 2-4.

Chapter Five

ENSEMBLE LEARNING WITH E2E RATIO IN RL-BASED ENERGY MANAGEMENT

Following the study on the impact of decay functions on E2E ratio and learning performance, this chapter aims to incorporate the advantages of different decay functions and two ensemble RL schemes based on learning automaton modules (LAM) are studied to improve the energy efficiency of RL-based EMS. Based on the comparison of study on three ensemble approaches, the weighted-based approach with synchronised-based framework is proposed as the unified ensemble learning approach for RL-based EMS.

5.1 Problem statement

Based on the analysis of E2E ratio change and conventional RL-based algorithms, several issues such as instability in Q-learning and balancing exploration and exploitation has brought about. To mitigate these problems, ensemble methods have been shown to be effective and efficient in combining single learning process in a system, leading to a higher accuracy than obtainable with a single learning process. For instance, the idea of bagging in machine learning is a simple method that trains

multiple classifiers using a different partitioning of the training set and combines them by majority voting. Therefore, this chapter explores the learning performance of an ensemble learning system for a plug-in hybrid off-highway vehicle. In RL-based learning, ensemble methods have been used for calculating function approximators to store the value function and update Q values more accurately. Following the investigation on the E2E ratio with three decay functions: i.e., 'exponential decay' (EXD), 'step-based decay' (SBD) and 'reciprocal decay functions' (RBD), in Chapter 4, the exponential decay and step-based decay methods outperformed the reciprocal decay method. Therefore, these two methods were selected and implemented into the ensemble learning system.

5.2 Supervisory ensemble learning-based system

Two frameworks are presented based on learning automata to study the learning performance of an ensemble RL-based system. Two base learners are located in the LAM for providing distinct control policies: Learner A follows the exponential decay method, and Learner B follows the step-based decay method. The detailed information on the learning automata and the two frameworks of the ensemble learning-based system is shown below.

5.2.1Learning automata module

The LAM contains all base learners that provide optimal solutions, as shown in Fig. 5-1. The LAM acts as a classifier to identify the global solutions from all base learners. In this research, there are two parallel base learners in the LAM. Each of them follows a distinct E2E ratio method.

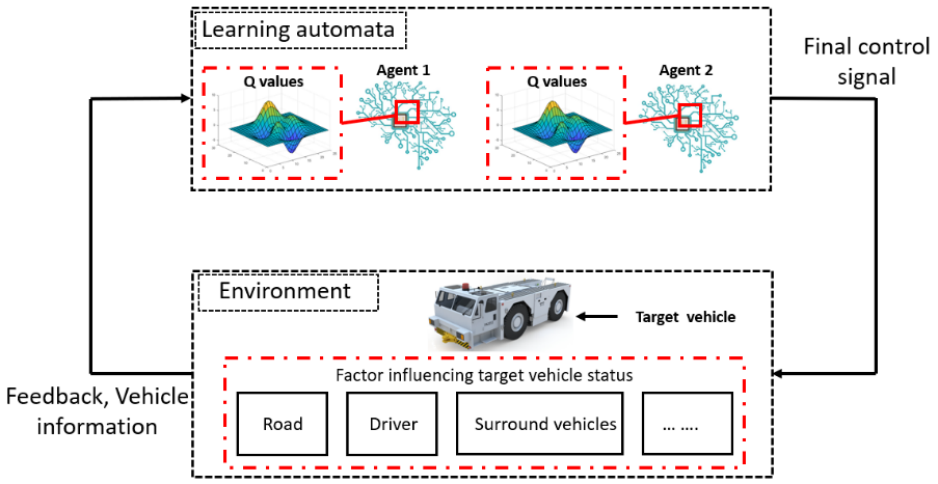


Fig. 5-1 Learning automata working process

There are three essential components in the LAM, i.e., state, action and reward, which are the same as the model-free supervisory control system (as discussed in Chapter 4) and can be modelled as:

$$LAM = \{\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}\} \tag{5-1}$$

This thesis focuses on the ensemble of two learning agents within the LAM, and there are only two possible frameworks for ensemble learning systems. The details

of each framework are introduced as follows.

5.2.2Synchronised-based ensemble learning

The synchronised-based ensemble learning (SYE) framework is described in Fig. 5-2. The difference between it and the IFE is that: 1) the learning and ensemble processes can be run synchronously; 2) three ensemble methods are conducted in the LAM. The detailed working step of the synchronised-based ensemble learning system is described as follows:

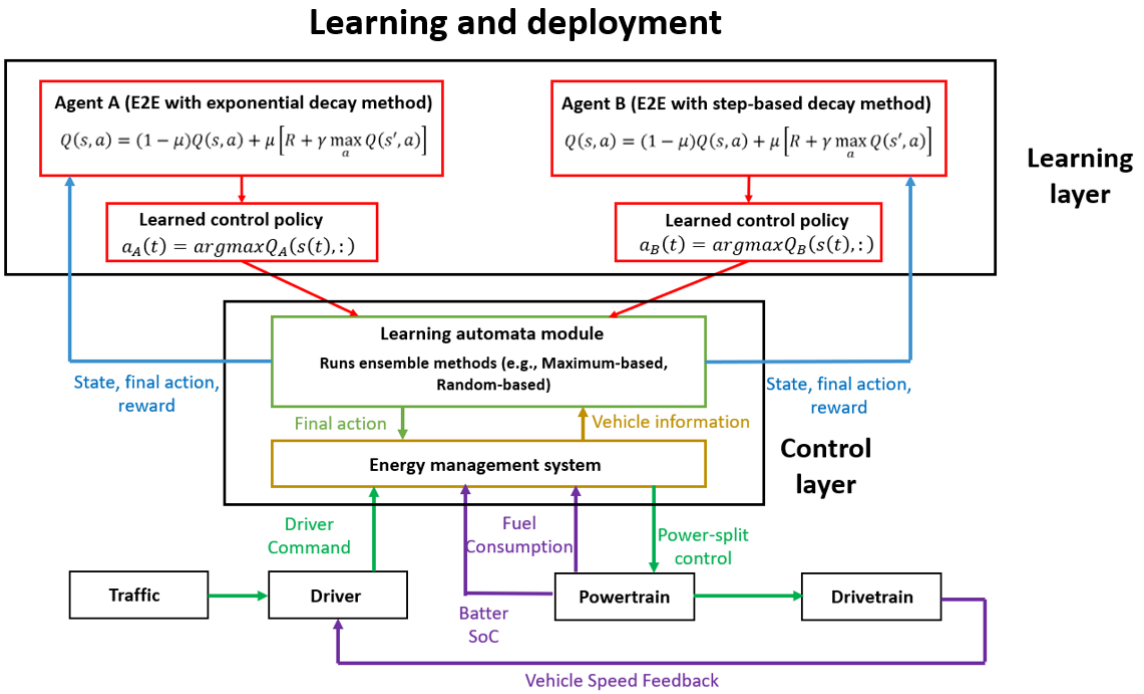


Fig. 5-2 The framework of synchronised-based ensemble learning system

Step 1: at the time step t, similar to the IFE, the energy management system collects

vehicle information and transmits it to the LAM. Then the two base learners take the observation with the current vehicle states, $s(t)$, and select its optimal action $a_A(t)$ and $a_B(t)$ based on the epsilon policy. The $a_A(t)$ can be selected as follows:

$$a_A(t) \leftarrow \begin{cases} arg \max\limits_{a(t)\in A}[Q^A(s(t),:)] & if\ T \geq \theta \\ randomly\ pick\ from\ A & if\ T < \theta \end{cases} \quad (5\text{-}2)$$

and the action $a_B(t)$ can be determined as follows:

$$a_B(t) \leftarrow \begin{cases} arg \max\limits_{a(t)\in A}[Q^B(s(t),:)] & if\ T \geq \theta \\ randomly\ pick\ from\ A & if\ T < \theta \end{cases} \quad (5\text{-}3)$$

where θ is the epsilon greedy value determined by the E2E decay functions (as discussed in Chapter 4); *T* is a random value between 0 and 1.

Step 2: next, the LAM collects all the base learners' optimal control signals and uses three ensemble methods: 1) maximum-based, 2) random-based, and 3) weighted-based. The description for each ensemble method is described as follows:

a) Maximum-based method

The final action is selected according to the current state-action pair value in the maximum-based method. If the corresponding value of base learner A is higher than base learner B, the final action is executed from base learner A. On the contrary, if the corresponding value of base learner B is higher than base learner A, the final

action is selected from learner B. The action execution policy of the maximum-based method is formulated by:

$$\pi_{max}: a_{final}(t) \leftarrow \begin{cases} a_A(t) & if\ Q^A\big(s(t), a(t)\big) \geq Q^B(s(t), a(t)) \\ a_B(t) & if\ Q^A\big(s(t), a(t)\big) < Q^B(s(t), a(t)) \end{cases} \quad (5\text{-}4)$$

where, $a_A(t)$ and $a_B(t)$ are two actions picked from each base learner with the maximum value in the current vehicle state $s(t)$; π_{max} is the execution policy for the maximum-based method; $Q^A\big(s(t), a(t)\big)$ and $Q^B\big(s(t), a(t)\big)$ are the accumulated reward from the knowledge bases of base learner A and base learner B.

b) Random-based method

The random-based ensemble method has two critical factors: a deciding variable and a random variable. Each of them uses comparisons to determine the final actions between the two base learners. The action execution policy for the random-based method is described as follows:

$$\pi_{Random}: a_{final}(t) \leftarrow \begin{cases} a_A(t) & if\ \ Y \geq T \\ a_B(t) & if\ Y < T \end{cases} \quad (5\text{-}5)$$

where π_{Random} is the action execution policy of the random-based method; *Y* is the random number between 0 and 1; *T* is the determining factor, which is a constant between 0 and 1, that helps determine the final action between the two base learners. If $Y \geq T$, the final action $a_{final}(t)$ is selected from the base learner A; otherwise, the final action $a_{final}(t)$ is executed from base learner B.

c) Weighted-based method

The weighted-based method considers the advantages of different E2E methods. The final action is determined by taking the consideration of each agent's proportion, which can be calculated by

$$\pi_{Weight}: a_{final} \leftarrow \mu a_{A(t)} + \delta a_B(t) \quad (5\text{-}6)$$

where $\pi_{Weighted}$ is the action execution policies of the weighted-based method; μ and δ are the coefficient factors determining the proportions of action $a_A(t)$ and action $a_B(t)$, respectively.

Step 3: ultimately, the final action is sent to the engine generator through the energy management system. The powertrain feeds back to the LAM with a reward, R. All the related information is uploaded to the learning layer. The control policies for each base learner are used in this final action to update its knowledge base by:

$$Q^A(s(t), a_A(t)) \leftarrow Q^A(s(t), a_A(t)) + [\boldsymbol{R} + maxQ^A(s(t+1), :) - Q^A\big(s(t), a_{final}\big)] \quad (5\text{-}7)$$

$$Q^B(s(t), a_B(t)) \leftarrow Q^B(s(t), a_B(t)) + [\boldsymbol{R} + maxQ^B(s(t+1), :) - Q^B\big(s(t), a_{final}\big)] \quad (5\text{-}8)$$

where $maxQ^A(s(t+1), :)$ and $maxQ^B(s(t+1), :)$ are the maximum value from base learner A and base learner B at vehicle state $s(t+1)$.

5.2.3Information fusion-based ensemble learning

The first framework is called the information fusion-based ensemble learning (IFE), which decouples the learning and ensemble processes into two individual steps. The first step is shown in Fig. 5-3 (a),(b), and the second step is shown in Fig. 5-4.

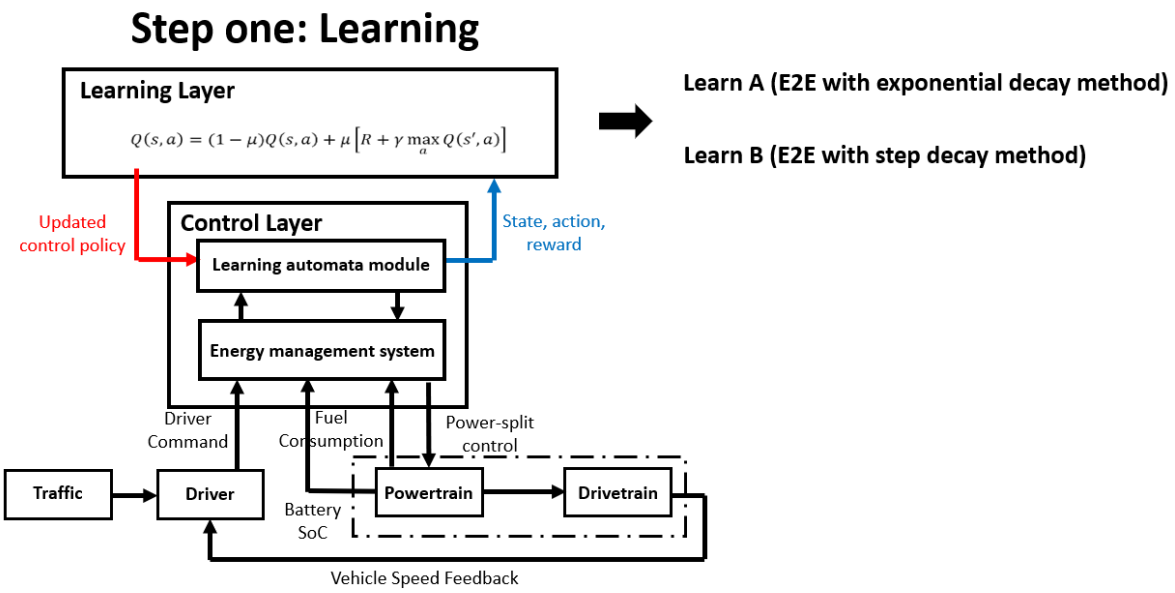


Fig. 5-3 The first step of the information fusion-based ensemble learning

The detailed working steps are shown below:

Step 1: the first step (learning stage) is shown in Fig. 5-3. There is no ensemble part. Learner A and learner B update their own experience individually with distinct E2E methods based on vehicle state $\boldsymbol{S}$, action set $\boldsymbol{A}$, and reward $\boldsymbol{R}$, under the same driving conditions.

Step 2: after the first step, the trained control policies from each agent are flashed into the learning automata module (LAM) as two base learners. The two optimal control policies for each learner can be formulated as:

$$\begin{cases} a_A(t) = arg \max\limits_{a(t)\in A}[Q^A(s(t),:)] \\ a_B(t) = arg \max\limits_{a(t)\in A}[Q^B(s(t),:)] \end{cases} \quad (5\text{-}9)$$

Step 3: the LAM starts the ensemble learning process, as shown in Fig. 5-4. At each time step t, the energy management system collects vehicle information (e.g. battery SoC, drive power requirement) and sends it to the LAM. The LAM starts observing the current vehicle states and transferring them to each base learner. Then, the two base learners select optimal actions, $a_A(t)$ and $a_B(t)$, from each knowledge base with maximum value. Next, the corresponding value for the selected action is extracted by the LAM as follows:

For action $a_A(t)$, its Q values for different learner can be retrieved by

$$[Q_A\big(s(t), a_A(t)\big), Q_B\big(s(t), a_A(t)\big)] \quad (5\text{-}10)$$

For action $a_B(t)$, its Q values for different learner can be retrieved by

$$[Q_B\big(s(t), a_B(t)\big), Q_A\big(s(t), a_B(t)\big)] \quad (5\text{-}11)$$

where $Q_A\big(s(t), a_A(t)\big)$ and $Q_B\big(s(t), a_A(t)\big)$ are the values from the knowledge base of learner A and the knowledge base of learner B, based on action $a_A(t)$. Additionally,

$Q_B\big(s(t), a_B(t)\big)$ and $Q_A\big(s(t), a_B(t)\big)$ are the corresponding values of the knowledge base of learner B and the knowledge base of learner A, based on action $a_B(t)$.

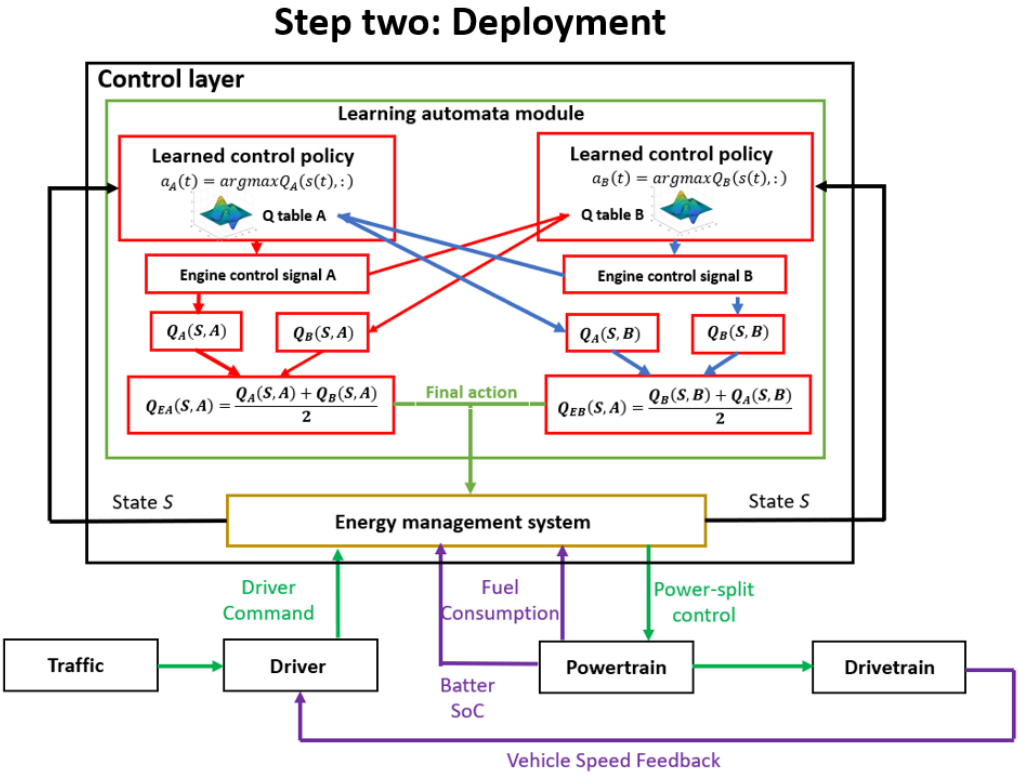


Fig. 5-4 The second step of the information fusion-based ensemble learning system

Step 4: next, the evaluation of the two actions is carried out by calculating the average of the corresponding Q values of $a_A(t)$ and $a_B(t)$. The two optimal actions can be evaluated by:

$$Q_{EA} = \frac{Q_A(s(t),a_A(t))+Q_B(s(t),a_A(t))}{2} \tag{5-12}$$

$$Q_{EB} = \frac{Q_B(s(t),a_B(t))+Q_A(s(t),a_B(t))}{2} \tag{5-13}$$

where Q_{EA} is the evaluation value for $a_A(t)$; Q_{EB} is the evaluation value for $a_B(t)$.

Step 5: next, the LAM determines the final action based on two evaluation values, and the final action can be selected by:

$$\pi_{exe}: a_{final} \leftarrow \begin{cases} a_A(t) & if\ Q_{EA} > Q_{EB} \\ a_B(t) & if\ Q_{EB} > Q_{EA} \end{cases} \quad (5\text{-}14)$$

where a_{final} is the final action which is sent to the engine generator for optimal control in real-time operations.

5.3 Performance and Discussions

In ensemble RL-based methods, firstly, different ensemble methods will be analysed with synchronised-based system from the perspectives of leaning performance and sensitivity analyses; then two different RL-based frameworks with singe agent and parallel agent are validated under predefined driving cycles; finally, the fusion-based system is obtained and tested.

5.3.1Improvement in energy efficiency

a) Efficiency improvement of the synchronised-based system with different ensemble methods

The efficiency improvement performance of three ensemble methods, including random-based, maximum-based, and weighted-based methods, are shown in Fig.

5-5.

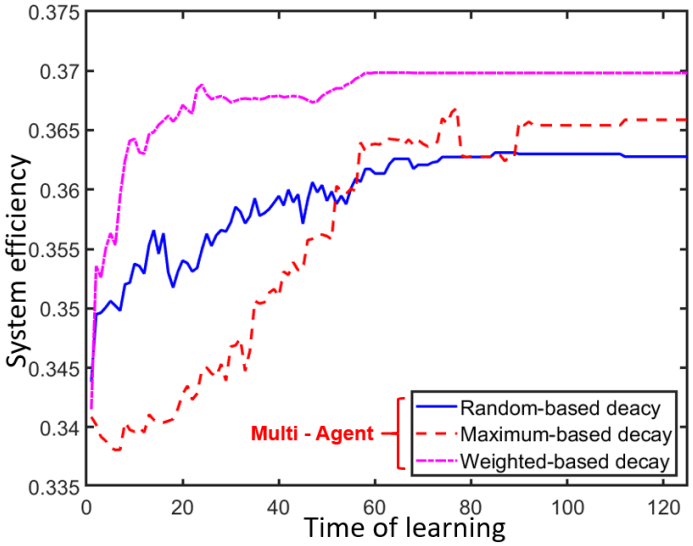


Fig. 5-5 Learning performance of three competing methods

The experiment was conducted for 125 times of learning under PBDC-1, with an initial battery SoC of 50% based on the SiL test platform. The energy efficiency values obtained by the maximum-based and random-based methods were unstable and converged very slowly throughout the learning process. Moreover, the vehicle energy efficiency obtained by the weighted-based method improved very smoothly and achieved the highest energy efficiency at the end of the learning process: up to 1.06% and 5.06% higher than the maximum-based method and random-based method, respectively.

b) Impact of determining factor values on vehicle energy efficiency

The default value of the determining factor is 0.5, which means each agent has a

50% chance of becoming the final winner to generate action for the engine. Therefore, it is necessary to study the impact of different values on determining factors by switching the value from 0.1 to 0.9. The test was conducted by 125 rounds of PBDC-1 with an initial battery SoC of 50%. Agent A follows the EXD method, and agent B follows the SBD method. If the random number Y is greater than T (the determining factor), agent A will be chosen for action execution. Otherwise, agent B will be selected for action execution. The results of different values of determining factors in the random-based method are shown in Table. 5-1.

Table.5-1 Vehicle energy efficiency obtained by ensemble RL-based EMSs with different determining factors

Determining factor	Vehicle's energy efficiency
0.1	35.75%
0.2	35.79%
0.3	35.93%
0.4	35.82%
0.5	35.90%
0.6	35.98%
0.7	36.22%
0.8	**36.28%**
0.9	36.21%

c) Efficiency improvement with different proportions of actions from the two learners in the weighted-based method

In the weighted-based method, the proportion of actions from the two agents may

affect the learning performance, therefore, this section investigates its impact on vehicle energy efficiency with different proportion settings. The default proportion of each agent was 50% assuming that the final action was determined by half of agent A (following SBD method) and half of agent B (following EXD). The influence of each agent was investigated by changing the proportion from 10% to 90%. The test was conducted based on 125 rounds of learning under predefined driving cycle (PBDC-1) with an initial battery SoC of 50%. Each test with a different proportion of agent A and agent B was repeated 25 times, and the results are compared in Table. 5-2. The vehicle's energy efficiency achieved the highest of 36.98% when the control action is calculated by the sum of 10% of agent A’s output and 90% of agent B’s output. That is to say, in the ensemble system, two agents with different methods have the different impacts on the final action, agent A’s 10% output and agent B’s 90% output represents different importance in determining the final action, and agent A cannot be ignored, the results shows that the agent B is more important.

Table.5-2 Different proportions of two parallel agents

Agent A (EXD)	Agent B (SBD)	Vehicle's energy efficiency
10%	**90%**	**36.98%**
20%	80%	36.85%
30%	70%	36.85%
40%	60%	36.68%
50%	50%	36.51%
60%	40%	36.84%
70%	30%	36.80%
80%	20%	36.59%
90%	10%	36.80%

d) Comparison of efficiency improvement with the ensemble learning system

The energy efficiency of the ensemble learning system using the weighted-based method (90% of learner A and 10% of learner B) is compared to results obtained by the single-learner based supervisory control system (SBD, RBD, EXD). The results are shown in Fig. 5-6. The weighted-based method achieved 1.15% higher vehicle energy efficiency than the baseline method (EXD). Moreover, the weighted-based method outperformed all single-learner methods at the end of the learning by at least 0.5% vehicle energy efficiency. The whole learning process can be divided into two stages. In the early learning stage (before the 60^{th} time of learning), the weighted-based method took the lead, performing consistently higher than the single-agent methods. This was due to the LAM taking the strength of 1) the SBD method in high probability to explore the new attempts; and 2) taking the 'suggestion' of the EXD to generate more new attempts of an engine control signal. Then, in the latter learning

stages (after the 60th time of learning), all the methods slowed down in improvement after a 'knee point'. This is because the E2E is very close to zero, which leads the LAM to determine the final action from the existing knowledge base.

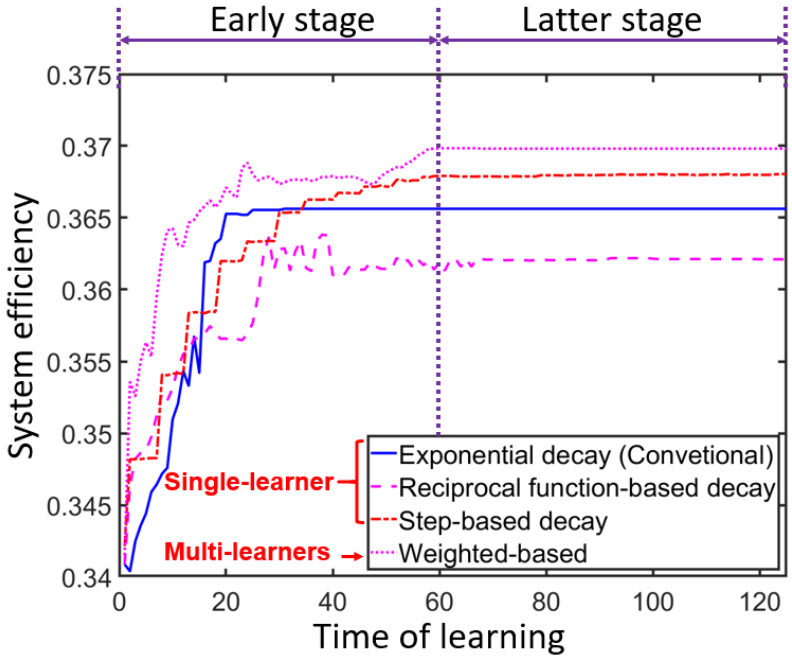


Fig. 5-6 The learning performance of four model-free control methods

e) Learning performance of information fusion-based system

For the investigation of the information fusion-based ensemble system (IFBS), the initial values of the epsilon for the two decay functions (exponential decay and step-based decay functions) were set at 0.8. Next, the control strategy was optimised separately under the PBDC-1 with a battery SoC of 50%. The vehicle energy efficiency learned by the two agents at the learning iterations of the 25th, 50th, 75th, 100th, 125th rounds is shown in Table 5-3.

Table.5-3 The vehicle energy efficiency at different learning rounds

Decay methods	25th	50th	75th	100th	125th
Exponentional decay	36.55%	36.56%	36.56%	36.56%	36.56%
Step-decay	36.33%	36.72%	36.79%	36.80%	36.80%

Then, the two learned control strategies are ensembled based on Eq. (5-7) in the LAM. The final vehicle energy efficiency obtained by the IFES is compared with the synchronous system with the weighted-based method and the E2E with EXD by the improvement rate in Table 5-4, through the improvement rate and the improving rate from the first learning rounds (IRF).

The improvement rate is calculated by:

$$\Delta_{ir} = \frac{\eta_{EXD} - \eta_{Com}}{\eta_{EXD}} \tag{5-15}$$

where Δ_{ir} is the improvement rate; η_{EXD} is the vehicle energy efficiency using the EXD method; η_{Com} is the vehicle energy efficiency using the WBD or IFBS. The improving rate from the first learning round can be calculated by:

$$\Delta_{irr} = \frac{\eta_{1st} - \eta_{final}}{\eta_{1st}} \tag{5-16}$$

where Δ_{irr} is the improving rate from the first learning rounds of each model-free method; η_{final} is the final vehicle energy efficiency of each model-free method. The average vehicle energy efficiency at the first learning round is 34.13%. After the ensemble process, the WBD achieves 36.98% at the end of learning, which is 1.17%

higher than the baseline method (EXD). The WBD method also has the highest IRF among the three methods, reaching 8.41%.

Table.5-4 Comparison of two ensemble learning systems

Control methods	1st learning rounds' vehicle energy efficiency	Final vehicle energy efficiency	Δ_{ir}	Δ_{irr}
EXD	34.09%	36.55%	-	7.21%
WBD	34.11%	36.98%	1.17%	8.41%
IFE	34.09% (EXD) 34.21% (SBD)	36.89%	0.09%	8.21%(EXD) 7.83%(SBD)

5.3.2Real-time performance

In ensemble reinforcement learning methods, the real-time performance based on the HiL test platform for the weighted-based method is carried out under the PBDC-1 with an initial battery SoC of 50% and compared to the results obtained by the EXD. The results are demonstrated in Fig. 5-7. The real-time performance of these two methods is compared by using the total energy loss of the vehicle, the variation of the battery SoC and the equivalent energy loss in the battery and engine generator. The weighted-based method (WBD) outperformed the conventional EXD method by reducing the total energy loss. This can be seen from the trajectory of the battery SoC. Meanwhile, the WBD can maintain the battery SoC close to the predefined threshold ($\mathrm{SoC}_{ref} = 28\%$). This led to significant energy savings in the battery with the proposed method.

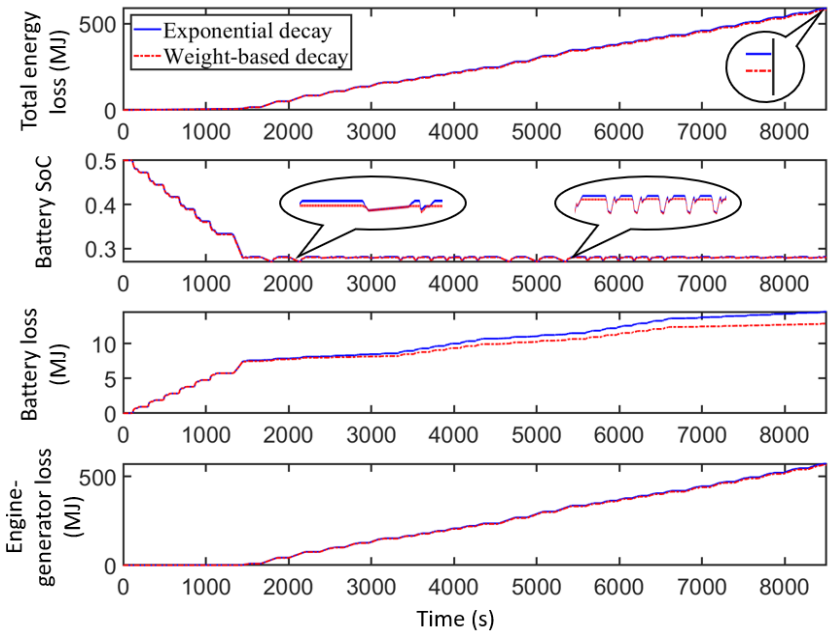


Fig. 5-7 Real-time performance of two model-free methods

5.3.3Robustness testing

In ensemble RL-based methods, the hybrid off-highway vehicle's energy management system will encounter various driving conditions in actual driving situations. Therefore, a robustness test is required for the proposed ensemble learning system. Three predefined driving cycles (PBDC-2 to 4) are selected to simulate unknown real-world driving conditions. The 'End SoC' (battery SoC after running a whole driving cycle) and 'OEC' (overall energy usage after running an entire driving cycle) are given in Table 5-5. The ensemble learning with the WBD is compared to the baseline method with the EXD. The rate of total energy savings can be calculated as follows:

$$\Delta = \frac{Ef_{EXD} - Ef_{WBD}}{Ef_{EXD}} \quad (5\text{-}17)$$

where Δ is the rate of energy-saving for each model-free method; Ef_{EXD} is the vehicle operation total energy cost using the EXD method; Ef_{WBD} is the total energy cost of the vehicle operation using the WBD. In order to keep the battery pack for a long life cycle and for safety, a threshold is set to 28%: if the battery SoC is below 28%, the engine-generator will automatically get involved to charge the battery pack. Compared to the EXD, the WBD can save at least 1.04% energy. The highest overall energy-saving rate of the WBD is 2.61%.

Table.5-5 Performance of the EXD and WBD methods under PBDC 2-4

Driving Cycle`	Initial SoC	Method	End SoC	OEC (MJ)	Savings
PRDC-2	30%	EXD	28.13%	333.98	-
	30%	WBD	28.03%	327.51	1.94%
	50%	EXD	28.13%	239.25	-
	50%	WBD	28.02%	234.38	2.03%
	70%	EXD	28.13%	142.07	-
	70%	WBD	28.01%	138.36	2.61%
PRDC-3	30%	EXD	28.13%	295.67	-
	30%	WBD	28.02%	292.03	1.23%
	50%	EXD	28.13%	201.75	-
	50%	WBD	28.01%	199.12	1.31%
	70%	EXD	28.13%	105.11	-
	70%	WBD	28.03%	103.37	1.65%
PRDC-4	30%	EXD	28.13%	395.41	-
	30%	WBD	28.02%	391.31	1.04%
	50%	EXD	28.13%	305.05	-
	50%	WBD	28.02%	301.11	1.31%
	70%	EXD	28.13%	211.09	-

	70%	WBD	28.03%	207.43	1.76%

5.4 Summary

Two ensemble learning frameworks, synchronised-based (SYE) and information fusion-based (IFE), were studied to combine the advantages of different E2E trajectories for the learning automata module (LAM). Three ensemble learning methods: maximum-based, random-based, and weighted-based methods (WBD), are developed based on the IFE framework and compared with the SYE framework for energy management of the vehicle in various operation conditions on the SiL and HiL platforms. The conclusions drawn from this work are as follows:

- The weighted-based ensemble method with the SYE framework dominates all the studied ensemble learning methods, and it can achieve 1.06% and 5.06% higher vehicle energy efficiency, compared to the random-based and maximum-based ensemble methods.

- For the weighted-based ensemble method, the investigation of different proportions in the weighted-based method has revealed that giving 90% confidence on the SBD outputs and 10% confidence on the EXD method outputs achieve the highest vehicle energy efficiency of 36.98%, compared to other ratio of the weighted-based ensemble method.

- The HiL testing under PBDC 2-4 conditions for validation of the control functionalities suggested that the WBD method is robust. It can save more than 1.04% of energy compared to the standard EXD method.

Chapter Six

REDUCTION OF OVERESTIMATION IN RL-BASED ENERGY MANAGEMENT CONTROL

To explore practically robust RL-based methods for the hybrid off-highway vehicles, this chapter carries out theoretical and experimental studies on reduction of overestimation based on conventional double Q-learning, including: 1) how to record the learned experience and 1) how to choose the action execution. Moreover, charge-sustaining control of the battery is studied in the presented EMSs owing to the consideration of the battery cycle life.

6.1 Problem statement

The agent of standard Q-learning always selects the action with the maximum expected future values. This characteristic brings the overestimation in its Q-values. Therefore, double Q-learning was proposed to reduce the overestimation. Two essential problems in the conventional double Q-learning method, including: 1) how to record the learned experience efficiently; and 2) how to determine the final action among two Q-tables, will be studied as illustrated in Fig. 6-1.

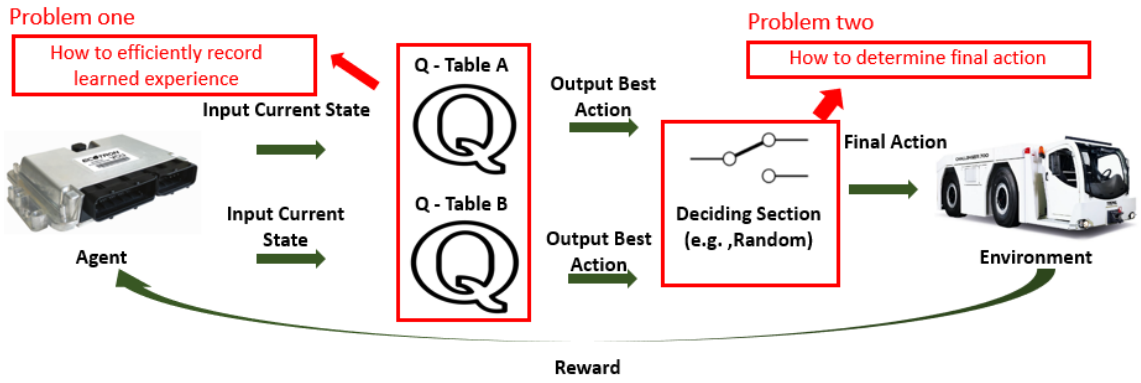


Fig. 6-1 Two problems in conventional double Q-learning

6.1.1Recording learned experience in two knowledge bases

Generally, the RL agent follows two update policies, including on-policy and off-policy, to record the learned experience from the environment (Zhu *et al.*, 2021). The behaviour policy (generating an appropriate action) and the improved policy (improving the agent experience) are critical differences between these two update policies. Usually, both the behaviour policy and the improved policy are the same in on-policy methods. In contrast, the behaviour policy and the improved policy are different in off-policy methods. Most existing RL algorithms solely conduct on-policy or off-policy methods to optimise their control policy. To the best of the author's knowledge, there are few publications on combining on-policy and off-policy methods based on double NQ-learning. Therefore, a new predictive double Q-learning method with backup models (PDQL) is proposed; where the AI agent records learned experiences based on the combination of on-policy and off-policy

methods.

6.1.2Action execution in conventional double Q-learning

The double Q-learning method can be used for EMSs of HEVs. Double Q-learning is chosen because it enables a straightforward learning process and is capable of EMSs for HEVs that only need two-state variables. Apart from the practical study using the double Q-learning algorithms with default settings, there are still some theoretical issues, e.g. how action execution policies impact the learning performance, this needs to be solved for specified engineering applications. Therefore: 1) two heuristic action execution policies are implemented for the control of a hybrid off-highway vehicle with double Q-learning; 2) the concept of disturbed and undisturbed learnings is used for the first time to study the potential of energy efficiency improvement with the double Q-learning methods.

6.2 Double Q-learning module in layered control framework

The double Q-learning module receives the state, action, and reward variables from the other three function modules. It implements a double Q-learning algorithm to optimise the action execution policy by updating the merit-function values in the knowledge bases. Two knowledge bases (Q^A and Q^B) are used to predict the merit

function value with the observation of the state variable. Theoretically, this mechanism can reduce the overestimation between the actual value and the approximation (Q. Zhang *et al.*, 2019). The learning process of double Q-learning is mathematically modelled as the updating of the two Q-tables using

$$\begin{cases} Q^A(s(t),a(t)) \leftarrow Q^A(s(t),a(t)) + \alpha\big(r(t) + Q^B(s(t+1),a') - Q^A(s(t),a(t))\big) \\ Q^B(s(t),a(t)) \leftarrow Q^B(s(t),a(t)) + \alpha\big(r(t) + Q^A(s(t+1),b') - Q^B(s(t),a(t))\big) \end{cases} \quad (6\text{-}1)$$

where, $Q^x(s(t),a(t))$ is the element in the knowledge bases Q^x ($x = A$ or B); the element value is indexed by the state variable $s(t)$ and action variable $a(t)$; a' and b' are the action for the next step predicted with the knowledge bases Q^A and Q^B with the maximum merit-function values; α is the learning rate, and the default setting $\alpha = 0.5$ is chosen for this research (Wang, Li and Lin, 2013). Conventional double Q-learning applies a rolling alternate progressing policy (Han *et al.*, 2019), i.e. Q^A and Q^B will change the role as the leading knowledge base in each time cycle, when Q^A is used as the leading knowledge base for action execution, the feedback merit-function value will be used to update the Q^B. A similar policy is applied if Q^B is used as the leading knowledge base for action execution.

6.2.1Action execution policies for double Q-learning

Action execution is the most important procedure for model-free charge sustaining control as described above. It determines the control signals to the downstream

controllers so that the powertrain feedbacks with a reward. Both action and reward variables significantly affect the optimisation performance in the double Q-learning module. Deterministic exploration with the maximum merit function value and random exploration are the two key elements working collaboratively to allow continuous evolution in RL. They are only used for the learning module for standard Q-learning and double Q-learning (Han *et al.*, 2019; Zhou *et al.*, 2019). Two new action execution policies are proposed with the consideration of these two elements as follows:

a) Maximum-based action execution policy

The max-value-based execution policy is proposed based on the deterministic exploration for system evolution. The action is picked up with the maximum merit-function value in both Q-tables:

$$\pi_{exe}: a(t) \leftarrow arg \max_{a(t)\in U} [Q^A(s(t),:)\ Q^B(s(t),:)] \tag{6-2}$$

where, $Q^A(s(t),:)$ and $Q^B(s(t),:)$ are two arrays of the merit function values; they are indexed by the state variables $s(t)$from the two knowledge bases $Q^A$and Q^B, respectively.

After the action $a(t)$ is executed, the following rules will be used for updating the knowledge bases with the reward feedback $r(t)$ from the vehicle:

$$\mathrm{Q}^{\mathrm{A}}\left(s(t),a(t)\right) \leftarrow \mathrm{Q}^{\mathrm{A}}\left(s(t),a(t)\right) + \alpha\left(r(t) + Q^{B}(s(t+1),a') - Q^{A}\left(s(t),a(t)\right)\right) \; if \; Y \geq 0.5 \tag{6-3}$$

$$\mathrm{Q}^{B}\left(s(t),a(t)\right) \leftarrow \mathrm{Q}^{B}\left(s(t),a(t)\right) + \alpha\left(r(t) + Q^{A}(s(t+1),b') - Q^{B}\left(s(t),a(t)\right)\right) \; if \; Y < 0.5 \tag{6-4}$$

where *Y* is a random number between 0 and 1. Because the action execution policy is deterministic based on the maximum value in $Q^{A}(s(t),:)$ and $Q^{B}(s(t),;)$, therefore, the random number *Y* is needed to include a stochastic process to generate new features for the system evolution.

b) Random-based action execution policy

The random execution policy is proposed based on random exploration. A deciding variable *D* is introduced to compare with a random comparison variable *C*. The action is executed using the comparison result as:

$$\pi_{exe}: a(t) \leftarrow \begin{cases} arg\ \max\limits_{a(t)\in U} Q^{A}(s(t),:) & if \;\; C \geq D \\ arg\ \max\limits_{a(t)\in U} Q^{B}(s(t),:) & if \; C < D \end{cases} \tag{6-5}$$

where *C* is the random comparison variable that is between 0 and 1; *D* is the selecting actions. If *C*≥ *D*, the action execution module will select the deciding

variable, which determines the Q-table that will be used for action $a(t)$ from Q^A. Otherwise, an action $a(t)$ will be collected from Q^B. After the action $a(t)$ is executed, the following rules will be used for updating the Q-tables with the reward feedback $r(t)$ from the vehicle:

$$Q^A\left(s(t),a(t)\right) \leftarrow Q^A\left(s(t),a(t)\right) + \alpha\left(r(t) + hQ^B(s(t+1),a') - Q^A\left(s(t),a(t)\right)\right) \; if: \; a(t) \leftarrow \pi_{exe}(Q^A) \tag{6-6}$$

$$Q^B\left(s(t),a(t)\right) \leftarrow Q^B\left(s(t),a(t)\right) + \alpha\left(r(t) + hQ^A(s(t+1),b') - Q^B\left(s(t),a(t)\right)\right) \; if: \; a(t) \leftarrow \pi_{exe}(Q^B) \tag{6-7}$$

where, $a(t) \leftarrow \pi_{exe}(Q^X)$ means the action $a(t)$ is executed based on the knowledge bases Q^X $(X = A \; or \; B)$. Since the random policy involves uncertainties in action execution as in Eq. (6-5), the updated policy for both knowledge bases should be deterministic; that is if action $a(t)$ is randomly executed from a knowledge base (e.g. Q^A), the vehicle's system feedback will update the merit-function value in another knowledge base (e.g. Q^B).

6.2.2Double Q-learning with backup models

By introducing backup models in the standard double Q-learning loops (Schilperoort

et al., 2019), a predictive double Q-learning method with the backup models (PDQL) is proposed to improve control policy in real-world driving in this section. As illustrated in Fig. 6-2, based on the off-policy method (external loop), a new loop (internal loop) is added to imitate the on-policy method. The external loop starts running one step before the internal loop, and each loop predicts an action with the epsilon-greedy policy in a rolling process. The working steps of the proposed method are described as follows:

- **Step one** (external loop): at each time interval t, the agent observes the current state $s(t)$ and predicts an action $a(t)$ following the epsilon-greedy policy by randomly selecting Q-table A or Q-table B.
- **Step two** (external loop): the agent executes the action $a(t)$ and measures the state variable $s(t+1)$ and reward variable $r(t)$ from the system. Then, the agent starts to update its knowledge $Q^S\big(s(t)\,, a(t)\big)$ by taking the maximum Q-value of action in the next state from the selected Q-table (improved policy for external loop) with the following equations:

$$b^* = arg\,max Q^S\,(s(t+1)\,, :); \tag{6-8}$$

$$Q^S\big(s(t), a(t)\big) \leftarrow Q^S\big(s(t), a(t)\big) + \alpha\left(r(t) + hQ^M(s(t+1)\,, b^*) - Q^S\big(s(t), a(t)\big)\right); \tag{6-9}$$

where, Q^S is the selected table either Q-table A or Q-table B in time step t; Q^M

($M = A\ or\ B$) represents the backup model, which is not used to predict the action at time step t; b^* is the action from the Q^M with the maximum merit function value at state $s(t+1)$; $\alpha = 0.5$ is the learning rate; and h = 1 is the discount factor.

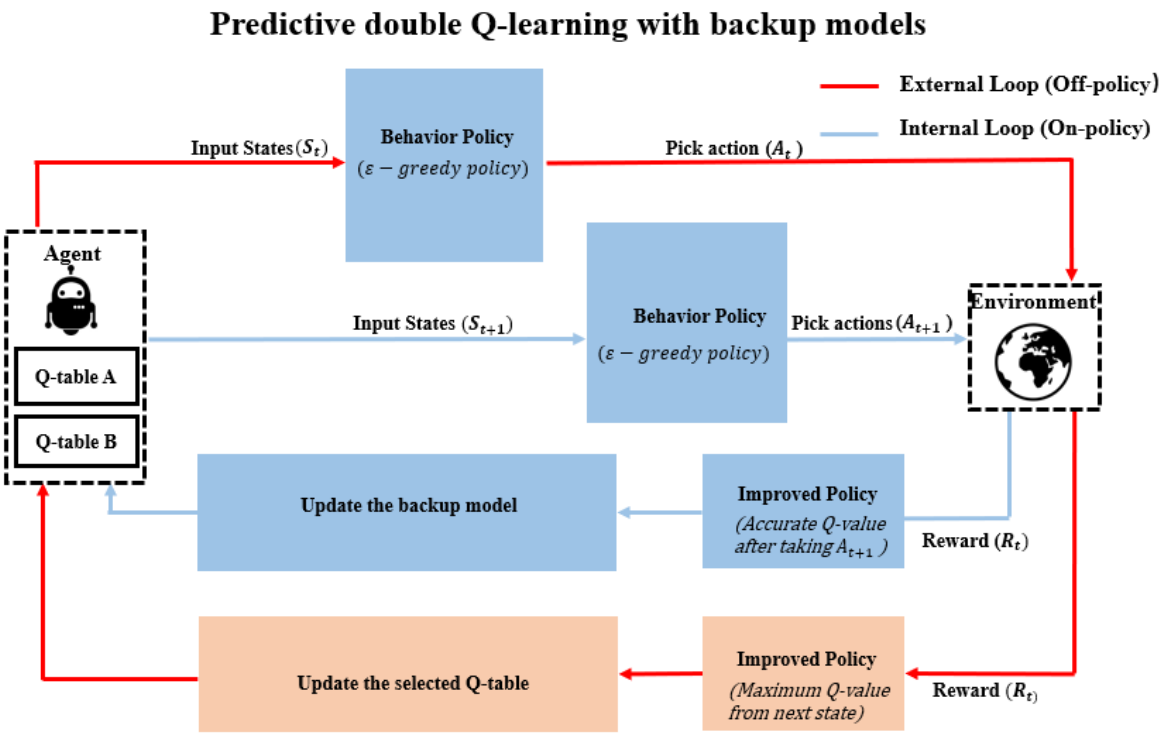


Fig. 6-2 The two loops in double Q-learning with backup models

- **Step three** (internal loop): the agent determines an action $a(t+1)$ based on the epsilon-greedy policy (improved policy for internal loop) from the two Q-tables after obtaining an observation in $s(t+1)$ and obtaining the reward $r(t+1)$. Then the agents start to update the backup model $Q^M\big(s(t), a(t)\big)$ by

$$Q^M\big(s(t), a(t)\big) \leftarrow Q^M\big(s(t), a(t)\big) + \alpha\Big(r(t+1) + hQ^S\big(s(t+1), a(t+1)\big) -$$

$$Q^M\big(s(t), a(t+1)\big)\Big) \tag{6-10}$$

- **Step four** (external loop): repeating step one, the agent moves to S_{t+2} and starts to update $Q^S(S_{t+1}, A_{t+1})$ with the following functions:

$$b^* = arg\, max Q^S\,(s(t+2), :); \tag{6-11}$$

$$Q^S\big(s(t+1), a(t+1)\big) \leftarrow Q^S\left(s(t+1),\ a(t+1)\right) + \alpha\Big(r(t+1) + hQ^M(s(t+2), b^*) - Q^S\big(s(t+1), a(t+1)\big)\Big) \tag{6-12}$$

The functional diagram for the proposed predictive double Q-learning method with the backup models is visualised in Fig. 6-3.

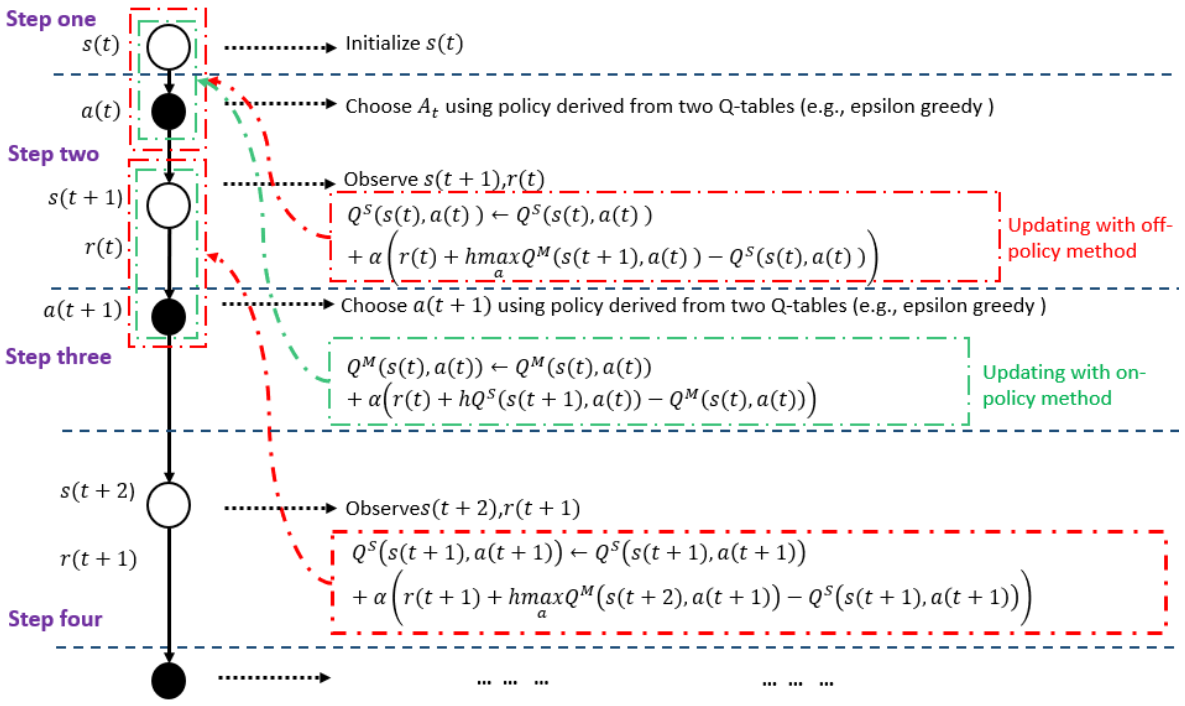


Fig. 6-3 The working diagram of the predictive double Q-learning method with backup model

6.3 Results and discussions

6.3.1Learning performance of double Q-learning methods

When double Q-learning method is employed, the improvement in energy efficiency, stability and learning performance are validated under different conditions, like different execution policies, disturbed and undisturbed learning scenarios.

a) Improvement in the energy efficiency of double Q-learning with two execution policies

A model-free charge-sustaining control performance with double Q-learning is initially investigated by monitoring the vehicle's energy efficiency at different learning stages on the SiL platform. This will demonstrate how the double Q-learning with the proposed action execution policies outperform the conventional double Q-learning method. The results obtained with conventional double Q- learning methods and the double Q-learning with the heuristic action execution policies are compared in Fig. 6-4.

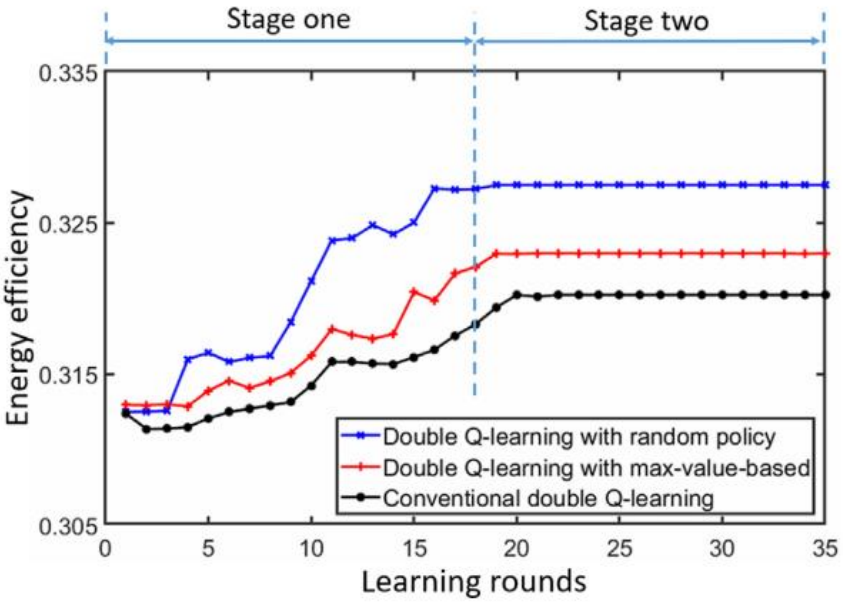


Fig. 6-4 Learning performance of three control strategies

The black line with circles shows the improvement in the vehicle's energy efficiency with conventional double Q-learning. The blue crossed line monitors the vehicle's performance using the random action execution policy, and the red plus-sign line shows the vehicle's performance using the max-value-based action execution policy. The learning processes can be classified into two stages based on the improving rate of the vehicle's energy efficiency. In stage one, the energy efficiency improvement fluctuates dramatically. This is due to the range of attempts of the differing actions for RL. Stage two slows down the improvement of the vehicle's energy efficiency after a 'knee point'. This is due to the progressive reduction in the probability of new action exploration over time, which leads to a theoretical logarithmic improvement in the vehicle's energy efficiency.

All the double Q-learning algorithms can achieve an improvement in the vehicle's energy efficiency, starting with an average initial vehicle energy efficiency of 31.29% at around one by learning from scratch (all the knowledge bases are zero sets). Then 2.56%, 3.23%, and 4.80% improvements in vehicle energy efficiency are realised by conventional double Q-learning, double Q-learning with the max-value-based execution method, and double Q-learning with the random execution method, respectively. The two proposed double Q-learning control methods can obtain a better vehicle performance than the standard Q-learning at the end of the learning process. The double Q-learning with the proposed action execution policies outperforms the conventional double Q-learning method by additionally improving the vehicle's energy efficiency by at least 1.09%. The model-free charge-sustaining control with the random execution policy can further increase the vehicle's energy efficiency by at least 3% compared to the max-valued-based method.

b) Efficiency improvements in disturbed and undisturbed learning scenarios

In real-world driving, the learned experience stored in the knowledge base may negatively affect the learning progress. This motivates a study of the model-free charge-sustaining control in the undisturbed and disturbed learning scenarios. Three groups have been used for offline learning: Standard Q-learning (Group A); double Q-learning method with maximum action execution (Group B); and double Q-

learning with random action execution (Group C), to observe the adaptivity under turbulence (defects from the algorithm). DP is used to obtain the global optimisation results offline to justify the maximum efficiency that can be achieved ideally. Based on the authors' previous development (Ji Li *et al.*, 2020), the DP algorithm uses the battery SoC as a state variable, the cost for the transient from one state to another is calculated by Eq. (4-5). In each group, the RL algorithm operates two individual optimisations for model-free charge-sustaining control under the same driving condition (repeating driving cycle 1 for 35 rounds) with an initial battery SoC of 30%. The first optimisation is learning from scratch (undisturbed learning), and the second optimisation is learning from a pre-defined knowledge based on a poor control policy (disturbed learning). Disturbed learning introduces some turbulences into the learning process by simulating bad attempts. Figs 6-5 (a), (b) and (c) show the vehicle's energy efficiency (calculated at the end of each round) based on standard Q-learning, the double Q-learning method with the max-value-based action execution policy, and the double Q-learning method with the random action execution policy, respectively. The vehicle's energy efficiency obtained in undisturbed learning (red crossed line) and in disturbed learning (blue circles' line) is compared.

The global optimal energy efficiency calculated by the DP for the driving cycle is 32.81% (shown as the solid pink line) with the assumption that the power demand

of the driving cycle is known in advance. This can only provide a reference to show the effectiveness of the online RL-based control methods (achieving more than 31.53%). But it is impossible to use the DP in real-time practice because it requires actuating 100% of the future power demand profiles in real-world driving applications. The model-free energy management method is not limited by the driving conditions and less computation (after the training process). It can be practically applied online for achieving optimal energy efficiency through continuously interacting with the characteristics of unknown driving cycles.

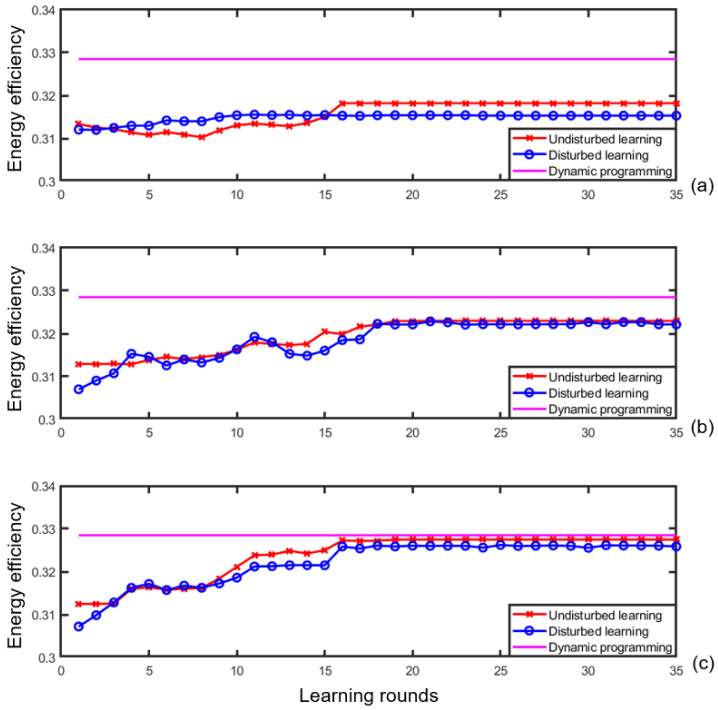


Fig. 6-5 Vehicle's energy efficiency obtained in undisturbed learning and disturbed learning a) conventional Q-learning; b) double Q-learning method with the max-value-based execution policy; c) double Q-learning method with the random execution policy

Generally, undisturbed learning achieves better vehicle energy efficiency than disturbed learning because a poor control policy is applied in the pre-defined knowledge base. For the standard Q-learning (Group A), undisturbed learning experiences some oscillations during the learning process and finally achieves an energy efficiency of 31.82%. The disturbed Q-learning gradually improves energy efficiency and finally converges to the energy efficiency of 31.53%. Both double Q-learning methods (Group B and C) can achieve higher vehicle energy efficiency than the standard Q-learning with undisturbed or disturbed learning. More specifically, for the double Q-learning method with the max-value-based execution policy (Group B), the vehicle energy efficiency can achieve 32.21% in disturbed learning and a similar energy efficiency level of 32.30% in undisturbed learning. The random action execution policy (Group C) has more potential to underestimate or reduce the value of each action in both Q-tables. It obtained the highest vehicle energy efficiency with 32.59% and 32.75% in disturbed learning and undisturbed learning, respectively.

Considering the improvements (from the initial energy efficiency to the end energy efficiency) in disturbed learning, the standard Q-learning gives an improvement of 1.05%, much less than that in undisturbed learning (3.09% improvement). This is because the pre-defined knowledge base in disturbed learning leads to the overestimation of the actions, which limits the probability of attempting other actions for global searching. The double Q-learning with the max-value-based execution

policy achieves an improvement of 3.22%; while 4.80% energy efficiency improvement is achieved by the random execution method. The double Q-learning methods can achieve at least three times higher energy efficiency improvement than the conventional Q-learning during the disturbed learning process.

c) Stability of learning in the vehicle's energy efficiency improvement

Double Q-learning involves many stochastic processes that may affect learning stability. This is more significant for the method using random execution policy, in which an additional random number *C* is initially compared to the deciding factor *D*. If *C* is larger than *D,* then action from Q-table A is executed. The default value of the deciding factor is 0.5, which assumes an equal probability for attempting an action from one of the two tables. The influence of the deciding factor on the learning level was investigated by changing the value from 0.1 to 0.5. The tests were carried out based on repeating 35 rounds of driving cycle 1 with an initial battery SoC of 30%. Each test with a different deciding variable value was repeated 20 times. Table 6-1 compares the level of learning at stage two for different deciding factor values, where the vehicle energy efficiency at the end of the learning process and the variance of results in stage two are used to evaluate the stability of learning.

Table.6-1 Stability of learning at stage two with different deciding factor values

Deciding variable	Energy efficiency	Variance of results
0.10	32.65%	8.94e-8
0.15	32.63%	3.46e-8
0.20	32.73%	8.73e-8
0.25	32.56%	4.20e-7
0.30	32.57%	7.68e-8
0.35	32.75%	2.13e-8
0.40	32.74%	1.12e-8
0.45	32.65%	6.50e-8
0.50	32.74%	3.20e-8

The data from Table 6-1 was analysed in the statistical software package, Minitab, to ascertain any correlation between the outcomes and the level of the deciding variable.

Table.6-2 Correlation of the results and performance of the models with the deciding variable

x	y	Correlation Coefficient	P-value
Deciding variable	Energy efficiency	0.386	0.304
	Variance of results	-0.250	0.517

The results in Table. 6-2 show that (at say 5% significance) the null hypothesis (that there is no correlation) cannot be rejected for either the energy efficiency or the variance, so it is concluded that there is no significant correlation with the level of the deciding variable. The stability of learning for the control policy optimisation based on Q-learning, the double Q-learning method with the max-value-based execution policy, and the double Q-learning method with the random execution policy is

compared in Table 6-3.

Table.6-3 Mean efficiency and variance of 20 samples

Algorithm	Execution method	Average efficiency	Variance of results
Q-Learning	-	31.47%	4.78e-7
Double Q-Learning	Max-value-based	32.35%	1.03e-6
	Random (best: $\mathcal{D}$ =0.35)	32.75%	2.13e-8
	Random (worst: $\mathcal{D}$ =0.25)	32.56%	4.20e-7

The variance of the results obtained by all the methods is minimal, which shows that the learning process in stage two is very stable. Among these methods, the double Q-learning with the random execution policy has the lowest variance of results. This includes both the best and the worst average energy efficiency obtained by the random execution policy with different deciding factor values in Table 6-3. Regarding the average vehicle energy efficiency in stage two, all double Q-learning methods outperform the Q-learning method by achieving higher energy efficiency. The double Q-learning method with the random execution policy can achieve at least 32.56% energy efficiency, marginally better than the double Q-learning method with the max-value-based policy (32.35%).

d) Learning performance of predictive double Q-learning with backup models

The level of learning for the PDQL is investigated with offline learning and compared

to the DQL-REP and the SQL through monitoring the evolution of vehicle energy efficiency. Both the equivalent energy loss in fuel consumption and the energy dissipated in the internal resistance of the batteries are necessary performance measures for the plug-in hybrid off-highway vehicle. The results of offline learning obtained by SQL, DQL-CON and PDQL with battery SoC of 30% are shown in Fig.6-6.

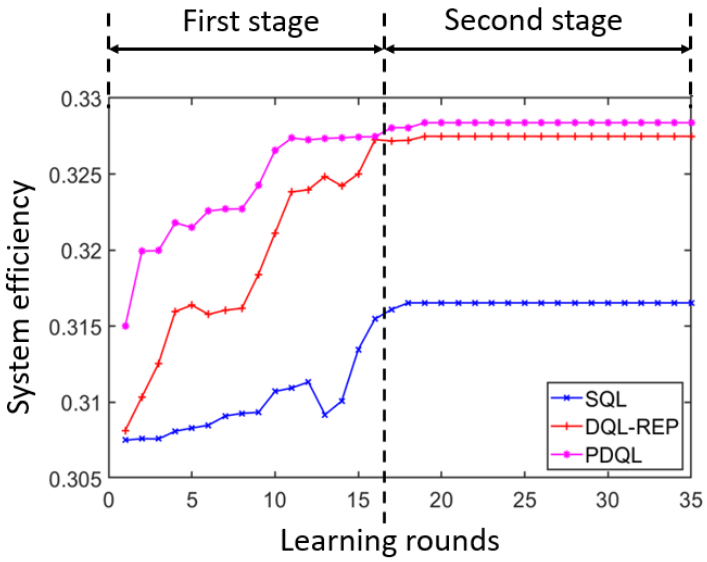


Fig. 6-6 Evolution of vehicle energy efficiency obtained by three model-free methods

The PDQL outperforms the DQL-REP and SQL. The PDQL learns faster and outperforms the other two control methods. For example, after the 7^{th} round of learning, compared with the other model-free methods, the vehicle energy efficiency of the PDQL tends to be more stable and reaches 32.09%, which is higher than SQL and DQL-REP with 2.08%, and 4.24%, respectively. Next, at the second stage, the PDQL achieves the highest vehicle energy efficiency with 32.84%, which is higher

than SQL and DQL-CON with 3.85% and 0.28%, respectively.

Based on the results obtained after 35 rounds of undisturbed learning, the learning performances considering vehicle energy efficiency (VEE) at the end of learning, total training time (Time), the average computing time per control interval (CLT), and the energy efficiency improvement are compared in Table 6-4. All three RL-based methods need less than one second to proceed with the action execution, data collection, and Q-table updating procedures.

The model used for HiL testing was provided by our industry partner and the fidelity is acceptable for control functionality verification with the real-time EMSs that requires 1 second sampling time (Zhou *et al.*, 2019) compared with the baseline methods, the PDQL needs longer computing time but can improve energy efficiency by 3.85% in offline learning conditions.

Table.6-4 Comparison of different model-free control methods in total time

Control methods	VEE(%)	Time (h)	CLT (s)	Improvement (%)
SQL	31.82%	5.87	0.021	-
DQL-REP	32.75%	7.51	0.257	2.92%
PDQL	32.84%	7.80	0.267	3.85%

6.3.2Real-time operation of double Q-learning methods

a) Real-time operation of two action execution policies

The energy flow in real-world operation with battery initial SoC values of 30% was investigated on the HiL platform under real-world cycle-1 for two stages. Firstly, we compared all the double Q-learning methods to select the most effective double Q-learning method that achieves the minimum total energy loss at the end of the driving cycle. In the second stage, the selected double Q-learning method with the standard Q-learning is compared and the results were obtained offline with dynamic programming. The results obtained by the double Q-learning (DQL) methods, including the DQL with conventional policy (DQL-CON), DQL with max-value-based action execution policy (DQL-MEP), and DQL with random action execution policy (DQL-REP), are compared in Fig. 6-7.

The real- time data, including the accumulated energy loss, battery SoC, equivalent energy loss in battery and engine generation, is collected in each subfigure. The vehicle with DQL-CON method (magenta dashed line) achieves almost the same value of total energy loss as that with the DQL-MEP (blue solid line) method. The DQL-REP method (red solid line) outperforms the other two methods by losing less energy in real-world driving. According to the real-time battery SoC, the DQL-REP enables a smoother operation of battery charge and discharge, and maintains the battery SoC close to the predefined level, i.e. 28%. This results in significant energy saving from the battery for the vehicle with DQL-REP.

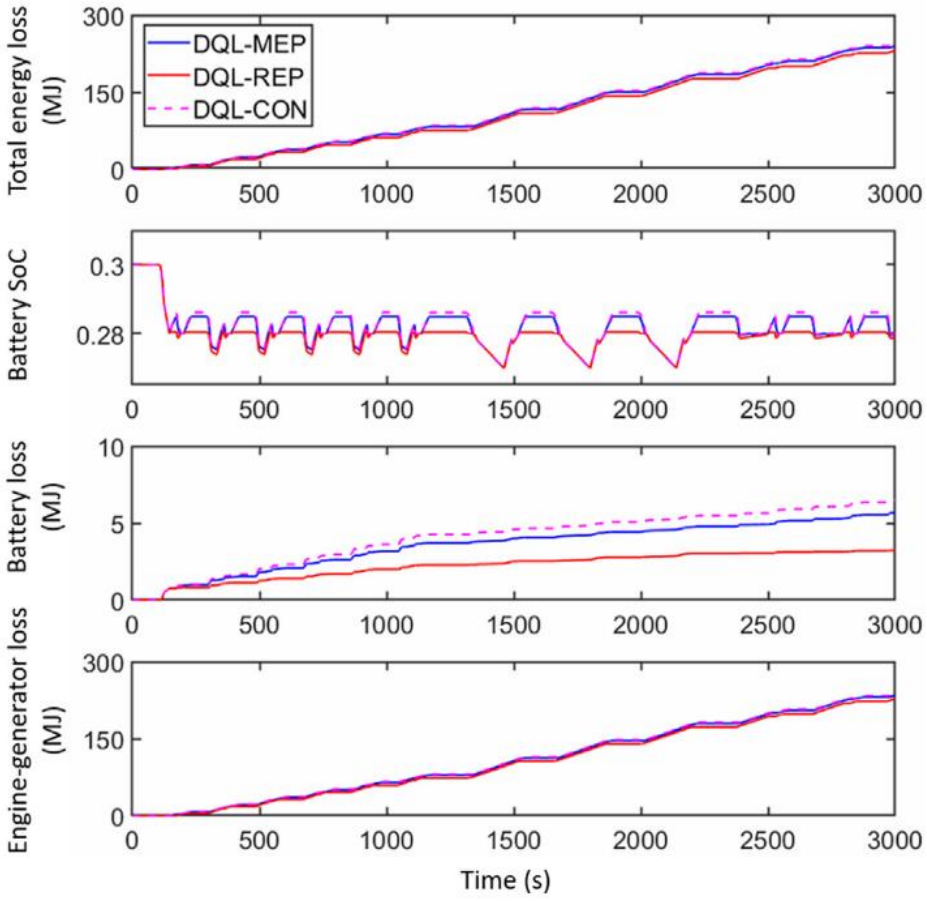


Fig. 6-7 HiL testing results of three double Q-learning methods

The vehicle performance with the DQL-REP method is then compared with the vehicles using the standard Q-learning (SQL) method, and the offline results obtained by DP. The vehicle performances are presented in Fig. 6-8. The DQL-REP (red solid line) outperforms the SQL (blue dash- dot line) method by achieving less total energy loss at the end of the driving cycle. The energy consumption curve obtained with the DQL-REP is closer to that obtained with the DP (cyan dashed line) method, especially for the energy loss in the battery. The battery SoC obtained with

the DP remains at a relatively high level compared to the SQL and DQL-REP method, because DP plans each action from a global optimisation perspective with consideration of the all the driving information. However, the DP results are practically inaccessible because it is impossible to obtain information about the entire trip in advance in real- world operations.

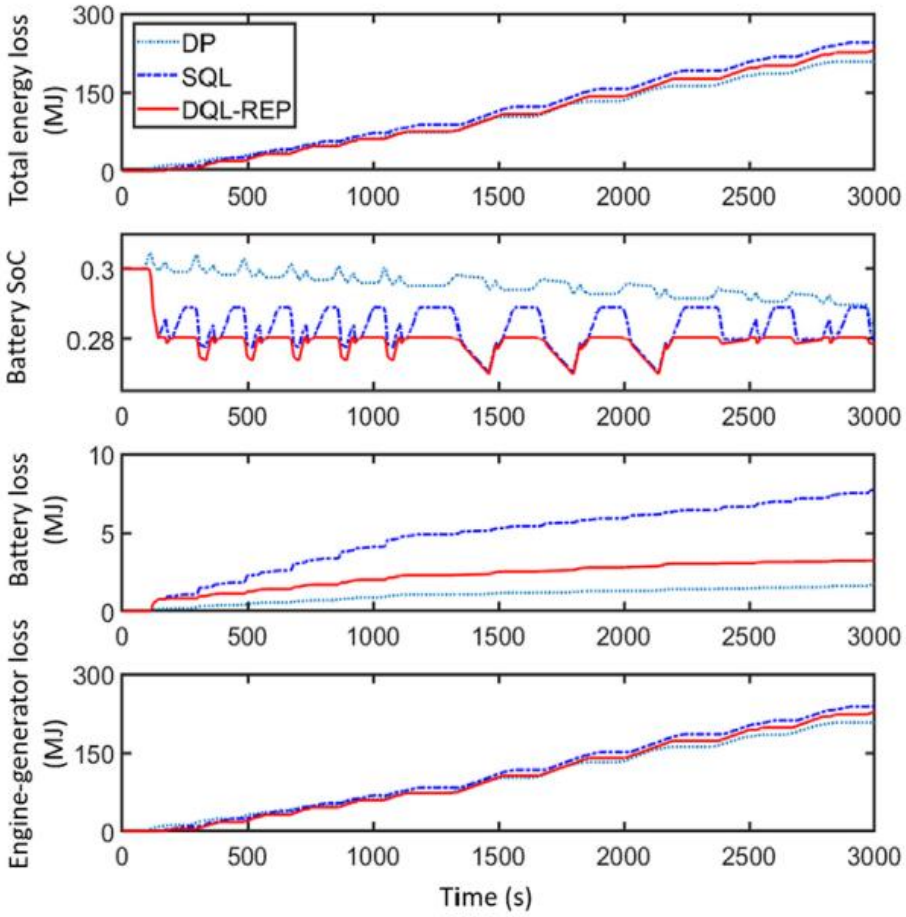


Fig. 6-8 HiL testing results of DQL-REP and SQL, and the offline results of DP

The energy flows of the vehicle with different energy management methods are further compared in Fig. 6-9. The accumulated total energy losses achieved by SQL,

DQL-CON, DQL-MEP, DQL-REP, and DP are 246.84 MJ, 243.33 MJ, 240.61 MJ, 231.45 MJ, and 209.19 MJ, respectively. The DP achieves the minimum energy losses with offline optimisation. It can be seen that DQL-REP is the best online learning method and has the lowest total energy loss compared to SQL and other DQL methods. The DQL-REP saves 15.39 MJ energy compared to the SQL method. This is contributed by 4.49 MJ energy saving in battery loss (58.20% lower than SQL method) and 10.90 MJ energy saved in the engine generator (4.56% lower than SQL method). In charge-sustaining control scenarios, the energy saving in the battery makes a bigger contribution to the total energy saving.

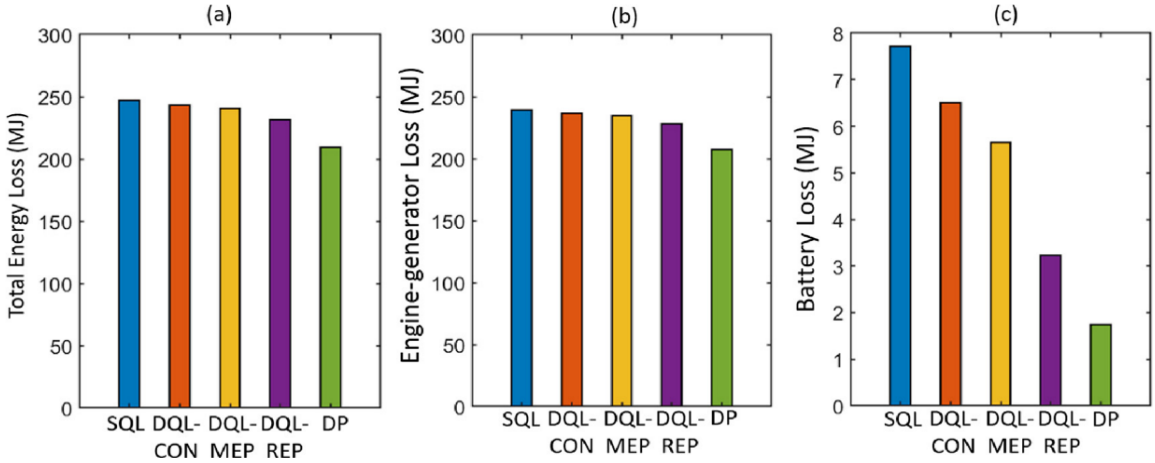


Fig. 6-9 Energy flow in driving cycle 1: (a) initial SoC = 25%; and (b) initial SoC = 30%

b) Implementing the PDQL method under real-world operation

The real-time performance of the PDQL is studied and compared to the SQL and DQL-REP via HiL testing with the initial battery SoC value of 30% under PRDC-1.

The results obtained by SQL, DQL-REP and PDQL are shown in Fig.6-10. The data of overall energy loss, battery SoC, battery energy loss and alternative power unit (APU) energy loss is collected in each subfigure. Three model-free control methods can maintain the battery SoC close to 30%, meeting the design target for charge-sustaining control.

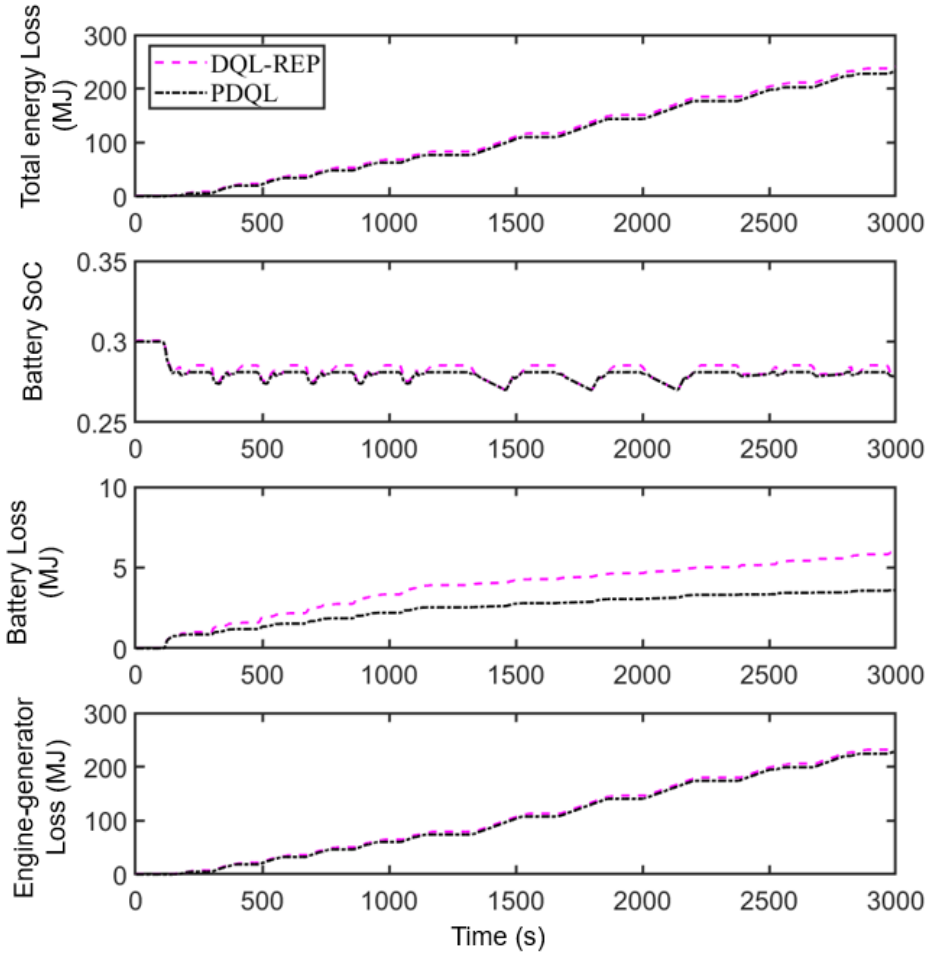


Fig. 6-10 HiL test results of three control methods under PBDC-1

6.3.3Robustness testing

A robustness test is carried out based on the HiL platform to validate the adaptability

of the proposed charge-sustaining control method in real-world driving. Three driving cycles, comprising nine scenarios with different initial battery SoC values (30%, 28%, and 25%), are used to emulate the unknown driving conditions in real-world driving. None of the driving conditions is used for offline learning of the control policy, the online learning capability will be evaluated in these unknown conditions. Both the battery SoC level at the end of each cycle (end SoC) and total energy usage is listed in Table 6-5. The total energy usage is the sum of the energy loss and the effective energy used for vehicle operation (e.g. driving and aircraft-towing). The proposed double Q-learning methods with the max-value-based execution policy (DQL-MEP), random execution policy (DQL-REP) and predictive double Q-learning (PDQL) are compared with standard Q-learning (SQL) and with double Q-learning with conventional policy (DQL-CON). The standard Q-learning method is chosen as the baseline method. The battery SoC is maintained at around 30% to keep the battery pack for a long-life cycle and for safety. When the battery SoC is below 28% in the proposed EMSs, the engine-generator will work at full load. This will bring a significant increase in fuel consumption, and thus the algorithm will be aware of the penalty when the strategy is too greedy to use the battery power. The energy-saving rate (savings) is calculated by:

$$\Delta = \frac{E_Q - E_{DQ}}{E_Q} \tag{6-14}$$

where, Δ is the energy-saving rate; E_Q is the total energy used for vehicle operation

with the standard Q-learning method; and E_{DQ} is the total energy used for vehicle operation with the two double Q-learning methods. According to the real-time variation of battery SoC, the saving is achieved by charging and smoothly discharging the battery following the power demand. The double Q-learning method with the random policy (DQL-REP) can save more than 4.55% of additional energy compared to the standard Q-learning method (SQL). The highest energy saving of 7.78% is obtained on real-world cycle-3 with the initial battery SoC set at 28%. The average energy saving is 5.78% for the nine pairs of experiments. The highest saving rates of the other two double Q-learning methods are 2.07% (DQL-CON) and 3.46% (DQL-MEP). It should be noticed that the Q-learning algorithm used here can save at least 7.8% energy compared to the model-based method (Zhou *et al.*, 2019).

Table.6-5 Performance of double Q-learning method and standard Q-learning method

Test driving cycles	Battery state of charge	Control methods	End battery state of charge	OEC (MJ)	Rate of total energy savings
PRDC-2	25%	SQL	28.80%	1082.51	-
	25%	DQL-CON	28.61%	1070.61	1.06%
	25%	DQL-MEP	28.49%	1062.32	1.85%
	25%	DQL-REP	28.04%	1032.46	4.55%
	25%	**PDQL**	**28.01%**	**1028.31**	**5.01%**
	28%	SQL	28.80%	1066.84	-
	28%	DQL-CON	28.61%	1054.13	1.21%
	28%	DQL-MEP	28.49%	1045.42	2.03%
	28%	DQL-REP	28.04%	1015.61	4.85%
	28%	**PDQL**	**28.01%**	**1012.56**	**5.36%**
	30%	SQL	28.80%	1056.13	-
	30%	DQL-CON	28.61%	1044.31	1.11%

	30%	DQL-MEP	28.49%	1035.92	1.93%
	30%	DQL-REP	28.04%	1006.44	4.76%
	30%	**PDQL**	**28.01%**	**1002.13**	**5.11%**
PRDC-3	25%	SQL	28.90%	556.22	-
	25%	DQL-CON	28.61%	545.95	1.88%
	25%	DQL-MEP	28.49%	538.79	3.17%
	25%	DQL-REP	28.04%	515.97	7.24%
	25%	**PDQL**	**28.01%**	**513.43**	**7.69%**
	28%	SQL	28.90%	539.75	-
	28%	DQL-CON	28.61%	528.93	2.07%
	28%	DQL-MEP	28.49%	521.57	3.46%
	28%	DQL-REP	28.04%	498.67	7.73%
	28%	**PDQL**	**28.01%**	**497.32**	**7.86%**
	30%	SQL	28.90%	529.64	-
	30%	DQL-CON	28.61%	519.25	2.06%
	30%	DQL-MEP	28.49%	512.06	3.45%
	30%	DQL-REP	28.04%	489.54	7.78%
	30%	**PDQL**	**28.01%**	**487.51**	**7.95%**
PRDC-4	25%	SQL	28.90%	494.56	-
	25%	DQL-CON	28.61%	489.98	1.02%
	25%	DQL-MEP	28.49%	486.94	1.68%
	25%	DQL-REP	28.04%	471.86	4.82%
	25%	**PDQL**	**28.01%**	**469.84**	**4.98%**
	28%	SQL	28.90%	487.99	-
	28%	DQL-CON	28.61%	469.94	1.13%
	28%	DQL-MEP	28.49%	457.05	1.86%
	28%	DQL-REP	28.04%	454.58	5.13%
	28%	**PDQL**	**28.01%**	**452.15**	**7.34%**
	30%	SQL	28.90%	467.33	-
	30%	DQL-CON	28.61%	462.83	1.08%
	30%	DQL-MEP	28.49%	459.81	1.81%
	30%	DQL-REP	28.04%	445.22	5.12%
	30%	**PDQL**	**28.01%**	**441.13**	**5.61%**

6.4 Summary

Two heuristic action execution policies and a predictive double Q-learning framework were proposed for the improvement of the energy efficiency of hybrid off-highway vehicles by reducing the overestimation of the merit-function values. SiL and HiL tests were used to evaluate the optimisation performance in the charge-sustaining control scenarios, using a standard Q-learning algorithm as the baseline, for both offline and online investigations. Four driving cycles were defined in the study based on real-world vehicle operation data; where driving cycle 1 was repeated 35 times in offline training (the data started to converge after 25 times). The conclusions drawn from the investigation are as follows:

- The heuristic action execution policies can improve the energy efficiency with at least 1.09% after 35 rounds of offline undisturbed learning. The double Q-learning with a random execution policy is robust to the turbulence and achieves an energy efficiency of 32.59% in disturbed learning.

- The double Q-learning method with the random execution policy (DQL-MEP) has the lowest variance and the highest average energy efficiency after 35 learning rounds in the defined learning cycle.

- Under undisturbed learning (i.e. the agent is learning from scratch), the PDQL achieves the highest energy efficiency, which is up to 3.35% and 1.58% higher than the SQL and DQL-REP.

- For online real-time control in driving cycles PBDC 1–4 with differing initial battery SoC, the performance of double Q-learning with the random execution policy is shown robust, saving at least 4% of total energy usage over the baseline methods.

Chapter Seven

CONCLUSION AND FUTURE WORK

This Ph.D. research developed a RL-based energy management system for off-highway vehicles and improved energy efficiency in push-back driving cycles and simulated real-world conditions. Following the research conducted in chapters four to chapter six. This chapter draws the conclusions, summarises the main contributions and discusses some potential future work.

7.1 Conclusions

1. **Following the study into the impact of the E2E ratio on improving vehicle energy efficiency, the step-based decay control method is shown to be capable of improving the energy efficiency of an off-highway vehicle**. And the main novelty, which is:

Three decay functions, exponential-based, reciprocal-based and step-based are used to generate an appropriate trajectory for the first time to improve the exploration of reinforcement learning. The vehicle energy efficiency improvement rate of the E2E method with the step-decay method is the highest efficiency improvement rate of 6.21%, which can achieve 1.36% and 0.33% higher than the RBD and EXD methods,

respectively.

2. **Taking the characteristics of E2E with three decay functions into consideration, the proposed ensemble learning system with the weighted-based ensemble method demonstrated the capability to improve the learning performance of the RL-based EMS**. And the main novelty, which is:

   A new ensemble learning system based on a combination of two E2E methods has been developed for the first time to improve the vehicle energy efficiency of an off-highway vehicle under real-time operation. Three ensemble methods (maximum, random, and weighted-based) have been investigated for the first time to determine the final action. The weighted-based ensemble method outperformed the random-based and maximum-based methods, by achieving 1.06% and 5.06% higher vehicle energy efficiency. Moreover, for the weighted-based ensemble method, the investigation of different proportions in the weighted-based method is 90% step-based outputs and 10% exponential-based outputs, which can achieve the highest vehicle energy efficiency of 36.98%.

3. **By introducing double Q-learning with two action execution policies and the predictive double Q-learning with backup models, the proposed EMS for a hybrid off-highway vehicle has achieved the ability to enhance the robustness and trustworthiness, due to the reduction of the overestimation in the RL-**

based EMS. And the main novelty, which is:

- A new predictive double Q-learning framework is proposed by introducing into conventional double Q-learning two action execution policies and an asynchronous learning method with backup models. 1) Two action execution policies (max-value based and random-based) were used in online control of an off-highway vehicle for the first time. The learning performance of the random policy with different values of the deciding factor is investigated from 0.1 to 0.5. Two learning scenarios named undisturbed learning and disturbed learning are proposed to study the overestimation phenomena in RL. The results show that the heuristic action execution policies can improve the energy efficiency with at least 1.09% after 35 rounds of offline undisturbed learning. 2) The capability of predictive double Q-learning with backup models (PDQL) for overestimation is conducted under an experienced learning scenario. The mechanism of storing the learned experience in two Q-tables is combined with the on-policy and off-policy for the first time. The double Q-learning with the max-value-based execution policy achieves an improvement of 3.22%; while 4.80% energy efficiency improvement is achieved by the random execution method. And the double Q-learning with a random execution policy is robust to the turbulence and achieves an energy efficiency of 32.59% in disturbed learning. The double Q-learning methods can achieve at least three times higher energy efficiency improvement

than the conventional Q-learning during the disturbed learning process.

7.2 Future Work

This research has provided some useful results for development of the RL-based EMS for hybrid off-highway vehicles. Following the work conducted in this thesis, there are several potential future directions to addressing the following issues:

- The traffic information can be considered in the optimisation process based on RL.

- Advanced deep neural network and neuro-fuzzy models can be developed and incorporated with the RL-based system for improving its robustness in different driving conditions.

- Data from V2X and near-field sensors can be used to obtain the prediction information which can help improve the vehicle energy efficiency and traffic efficiency simultaneously based on RL.

REFERENCE

Afram, A. and Janabi-Sharifi, F. (2014) ‘Theory and applications of HVAC control systems - A review of model predictive control (MPC)’, *Building and Environment*, 72, pp. 343–355. doi: 10.1016/j.buildenv.2013.11.016.

Aldhafeeri, T. *et al.* (2020) ‘A review of methane gas detection sensors: Recent developments and future perspectives’, *Inventions*, 5(3), pp. 1–18. doi: 10.3390/inventions5030028.

APC (2020) ‘Roadmap 2020 Heavy Goods >3.5t and Off-highway Vehicle’, (February 2021).

Apolinarska, A. A. *et al.* (2021) ‘Robotic assembly of timber joints using reinforcement learning’, *Automation in Construction*, 125, p. 103569. doi: 10.1016/j.autcon.2021.103569.

Apostolou, D. and Xydis, G. (2019) ‘A literature review on hydrogen refuelling stations and infrastructure. Current status and future prospects’, *Renewable and Sustainable Energy Reviews*, 113(May), p. 109292. doi: 10.1016/j.rser.2019.109292.

Archer, G. *et al.* (2018) ‘CO 2 Emissions From Cars: The facts’, *Transport & Environment*, (August), p. 53.

Baoxian, C. *et al.* (2013) ‘Robot Q- learning coverage algorithm in unknown environments’. doi: 10.14177/j.cnki.32-1397n.2013.06.032.

Bellman, R. (1954) ‘Dynamic Programming and a New Formalism in the Calculus of Variations’, *Proceedings of the National Academy of Sciences*, 40(4), pp. 231–235. doi: 10.1073/pnas.40.4.231.

Boujneh, M. *et al.* (2020) ‘Optimal Control Strategy Parameters of Parallel Hybrid Electric Vehicles Based on Particle Swarm Optimization’, pp. 107–112. doi: 10.1109/STA50679.2020.9329308.

Bouktif, S. *et al.* (2018) ‘Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches’, *Energies*, 11(7). doi: 10.3390/en11071636.

Cao, J. and Xiong, R. (2017) ‘Reinforcement Learning-based Real-time Energy Management for Plug-in Hybrid Electric Vehicle with Hybrid Energy Storage System’, in *Energy Procedia*. Elsevier, pp. 1896–1901. doi: 10.1016/j.egypro.2017.12.386.

Carroll, J. L., Peterson, T. S. and Owens, N. E. (2001) ‘Memory-guided exploration in reinforcement learning’, *Proceedings of the International Joint Conference on Neural Networks*, 2(January), pp. 1002–1007. doi: 10.1109/ijcnn.2001.939497.

Chen, Z. *et al.* (2020) ‘Stochastic model predictive control for energy management

of power-split plug-in hybrid electric vehicles based on reinforcement learning', *Energy*, 211, p. 118931. doi: 10.1016/j.energy.2020.118931.

Chomchai, C., Sonjaipanich, S. and Cheewaisrakul, S. (2011) 'Powertrain System Optimisation for a Heavy-duty Hybrid Electric Bus', *International Journal of Automotive Technology*, 12(1), pp. 131–139. doi: 10.1007/s12239.

Cotroneo, D., Pietrantuono, R. and Russo, S. (2013) 'A learning-based method for combining testing techniques', *Proceedings - International Conference on Software Engineering*, pp. 142–151. doi: 10.1109/ICSE.2013.6606560.

Ding, N. and Prasad, K. (2020) 'Design of a hybrid energy management system using designed rule-based control strategy and genetic algorithm for the series-parallel plug-in hybrid electric vehicle', (July), pp. 1627–1644. doi: 10.1002/er.5808.

Du, G. *et al.* (2020) 'Deep reinforcement learning based energy management for a hybrid electric vehicle', *Energy*, 201, p. 117591. doi: 10.1016/j.energy.2020.117591.

Du, H. *et al.* (2018) 'Who buys New Energy Vehicles in China? Assessing social-psychological predictors of purchasing awareness, intention, and policy', *Transportation Research Part F: Traffic Psychology and Behaviour*, 58, pp. 56–69. doi: 10.1016/j.trf.2018.05.008.

Enang, W. and Bannister, C. (2017) 'Modelling and control of hybrid electric vehicles

(A comprehensive review)', *Renewable and Sustainable Energy Reviews*, 74(January), pp. 1210–1239. doi: 10.1016/j.rser.2017.01.075.

Essoufi, M., Hajji, B. and Rabhi, A. (2020) 'Fuzzy Logic based Energy Management Strategy for Fuel Cell Hybrid Electric Vehicle'. doi: 10.1109/ICEIT48248.2020.9113162.

European Commission (2017) *Proposal for post-2020 CO2 targets for cars and vans | Climate Action*.

Eurpean Commission (2019) 'CO 2 emission standards for heavy-duty vehicles Mobility Packages I-III : an integrated approach'. Available at: https://ec.europa.eu/clima/sites/clima/files/transport/vehicles/heavy/docs/hdv_co2_standards_en.pdf.

Faizrizwan, M. and Sabri, M. (2018) 'Improved Fuel Economy of Through-the-Road Hybrid Electric Vehicle with Fuzzy Logic-Based Energy Management Strategy', *International Journal of Fuzzy Systems*, 20(8), pp. 2677–2692. doi: 10.1007/s40815-018-0521-4.

Fang, Y. *et al.* (2015) 'An energy management strategy for hybrid electric bus based on reinforcement learning', *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, pp. 4973–4977. doi:

10.1109/CCDC.2015.7162814.

Fares, D. *et al.* (2020) ‘ScienceDirect Dynamic programming technique for optimizing fuel cell hybrid vehicles’, *International Journal of Hydrogen Energy*, 40(24), pp. 7777–7790. doi: 10.1016/j.ijhydene.2014.12.120.

Ganesh, A. H. and Xu, B. (2022) ‘A review of reinforcement learning based energy management systems for electrified powertrains: Progress, challenge, and potential solution’, *Renewable and Sustainable Energy Reviews*, 154(April 2021), p. 111833. doi: 10.1016/j.rser.2021.111833.

Gao, D. W., Mi, C. and Emadi, A. (2007) ‘Modeling and Simulation of Electric and Hybrid Vehicles’, *Proceedings of the IEEE*, 95(4), pp. 729–745.

Gao, T. and Lu, W. (2021) ‘Machine learning toward advanced energy storage devices and systems’, *iScience*, 24(1), p. 101936. doi: 10.1016/j.isci.2020.101936.

‘Global Energy Review 2019’ (2020) *Global Energy Review 2019*. doi: 10.1787/90c8c125-en.

Gregor, E. (2019) ‘EU Legislation in Progress CO 2 emission standards for heavy-duty vehicles’, (December 2018).

Guo, X. *et al.* (2019) ‘SOC estimation with an adaptive unscented Kalman filter based

on model parameter optimization', *Applied Sciences (Switzerland)*, 9(19). doi: 10.3390/app9194177.

Guzzella, L. and Sciarretta, A. (2007) 'Vehicle Energy and Fuel Consumption-Basic Concepts', in *Vehicle propulsion systems: introduction to modeling and optimization*, pp. 13–46.

Hafez, A. A. *et al.* (2021) 'Optimal sizing of off-line microgrid via hybrid multi-objective simulated annealing particle swarm optimizer', *Computers and Electrical Engineering*, 94(September 2020), p. 107294. doi: 10.1016/j.compeleceng.2021.107294.

Han, X. *et al.* (2019) 'Energy management based on reinforcement learning with double deep Q- learning for a hybrid electric tracked vehicle', *Applied Energy*, 254(January), p. 113708. doi: 10.1016/j.apenergy.2019.113708.

van Hasselt, H., Guez, A. and Silver, D. (2015) 'Deep Reinforcement Learning with Double Q-learning'. doi: 10.1016/j.artint.2015.09.002.

He, H. *et al.* (2021) 'An improved MPC-based energy management strategy for hybrid vehicles using V2V and V2I communications', *Energy*, 225, p. 120273. doi: 10.1016/j.energy.2021.120273.

Hu, C. and Xu, M. (2020) 'Adaptive Exploration Strategy with Multi-Attribute

Decision-Making for Reinforcement Learning', *IEEE Access*, 8, pp. 32353–32364. doi: 10.1109/ACCESS.2020.2973169.

Hu, X. *et al.* (2013) 'Energy efficiency analysis of a series plug-in hybrid electric bus with different energy management strategies and battery sizes', *Applied Energy*, 111, pp. 1001–1009. doi: 10.1016/j.apenergy.2013.06.056.

Hu, X. *et al.* (2014) 'Comparison of three electrochemical energy buffers applied to a hybrid bus powertrain with simultaneous optimal sizing and energy management', *IEEE Transactions on Intelligent Transportation Systems*, pp. 1193–1205. doi: 10.1109/TITS.2013.2294675.

Hu, Y. *et al.* (2018) 'Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning', *Applied Sciences (Switzerland)*, 8(2). doi: 10.3390/app8020187.

Huang, S., Yang, L. and Ji, L. (2021) 'Current Status and Development Trends of Sludge Disposal Technology', *E3S Web of Conferences*, 290(2), p. 03002. doi: 10.1051/e3sconf/202129003002.

Huang, Y. *et al.* (2018) 'A review of power management strategies and component sizing methods for hybrid vehicles', *Renewable and Sustainable Energy Reviews*, 96, pp. 132–144. doi: 10.1016/J.RSER.2018.07.020.

INTERNATIONAL ENERGY AGENCY (2021) ‘Global EV Outlook 2021 - Accelerating ambitions despite the pandemic’, *Global EV Outlook 2021*, p. 101.

Javanmard, B. *et al.* (2021) ‘Energy management of multi-microgrids based on game theory approach in the presence of demand response programs , energy storage systems and renewable energy resources’, *Journal of Energy Storage*, 42(April), p. 102971. doi: 10.1016/j.est.2021.102971.

Jinquan, G. *et al.* (2019) ‘A novel MPC-based adaptive energy management strategy in plug-in hybrid electric vehicles’, *Energy*, 175, pp. 378–392. doi: 10.1016/j.energy.2019.03.083.

Johnson, V. H. V. H., Wipke, K. B. K. B. and Rausen, D. J. D. J. (2000) ‘HEV control strategy for real-time optimization of fuel economy and emissions’, *SAE Transactions*, pp. 1677–1690. doi: 10.4271/2000-01-1543.

Krawczyk, B. *et al.* (2017) ‘Ensemble learning for data stream analysis: A survey’, *Information Fusion*, 37, pp. 132–156. doi: 10.1016/j.inffus.2017.02.004.

Lee, H. *et al.* (2020) ‘Comparative Analysis of Energy Management Strategies for HEV : Dynamic Programming and Reinforcement Learning’, *IEEE Access*, 8, pp. 67112–67123. doi: 10.1109/ACCESS.2020.2986373.

Lei, Z. *et al.* (2020) ‘An adaptive equivalent consumption minimization strategy for

plug- in hybrid electric vehicles based on traf fi c information', *Energy*, 190, p. 116409. doi: 10.1016/j.energy.2019.116409.

Li, J. *et al.* (2019) 'Dual-loop online intelligent programming for driver-oriented predict energy management of plug-in hybrid electric vehicles', 253(March). doi: 10.1016/j.apenergy.2019.113617.

Li, Ji *et al.* (2020) 'Driver-identified Supervisory Control System of Hybrid Electric Vehicles based on Spectrum- guided Fuzzy Feature Extraction', 6706(c). doi: 10.1109/TFUZZ.2020.2972843.

Li, Jizi *et al.* (2020) 'Dual credit policy: Promoting new energy vehicles with battery recycling in a competitive environment?', *Journal of Cleaner Production*, 243, p. 118456. doi: 10.1016/j.jclepro.2019.118456.

Li, J. *et al.* (2021) 'Impact of the latent topics of policy documents on the promotion of new energy vehicles: Empirical evidence from Chinese cities', *Sustainable Production and Consumption*, 28, pp. 637–647. doi: 10.1016/j.spc.2021.06.023.

Li, Y. *et al.* (2017) 'Power Management for a Plug-in Hybrid Electric Vehicle Based on Reinforcement Learning with Continuous State and Action Spaces', *Energy Procedia*, 142, pp. 2270–2275. doi: 10.1016/j.egypro.2017.12.629.

Li, Y. *et al.* (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), pp. 7416–7430. doi: 10.1109/tvt.2019.2926472.

Li, Y. *et al.* (2021) 'Multi-objective energy management for Atkinson cycle engine and series hybrid electric vehicle based on evolutionary NSGA-II algorithm using digital twins', *Energy Conversion and Management*, 230(September 2020), p. 113788. doi: 10.1016/j.enconman.2020.113788.

Li, Y. and Jiao, X. (2020) 'Energy management strategy for hybrid electric vehicles based on adaptive equivalent consumption minimization strategy and mode switching with variable thresholds', 103(1), pp. 1–20. doi: 10.1177/0036850419874992.

Liessner, R. *et al.* (2018) 'Deep reinforcement learning for advanced energy management of hybrid electric vehicles', *ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2(January), pp. 61–72. doi: 10.5220/0006573000610072.

Linhares, L. L. S. *et al.* (2015) 'Fuzzy Wavelet Neural Network Using a Correntropy Criterion for Nonlinear System Identification', *Mathematical Problems in Engineering*, 2015. doi: 10.1155/2015/678965.

Liu, C. and Murphey, Y. L. (2014) ‘Power management for plug-in hybrid electric vehicles using reinforcement learning with trip information’, *2014 IEEE Transportation Electrification Conference and Expo: Components, Systems, and Power Electronics - From Technology to Business and Public Policy, ITEC 2014*. doi: 10.1109/itec.2014.6861862.

Liu, J. *et al.* (2019) ‘Heuristic Dynamic Programming Based Online Energy Management Strategy for Plug-In Hybrid Electric Vehicles’, *IEEE Transactions on Vehicular Technology*, 68(5), pp. 4479–4493. doi: 10.1109/TVT.2019.2903119.

Liu, J. and Peng, H. (2008) ‘Modeling and Control of a Power-Split’, 16(6), pp. 1242–1251.

Liu, R. Z. *et al.* (2021) ‘Efficient Reinforcement Learning for StarCraft by Abstract Forward Models and Transfer Learning’, *IEEE Transactions on Games*, 14(8), pp. 1–14. doi: 10.1109/TG.2021.3071162.

Liu, T. *et al.* (2015) ‘Reinforcement Learning of Adaptive Energy Management with Transition Probability for a Hybrid Electric Tracked Vehicle’, *IEEE Transactions on Industrial Electronics*, 62(12), pp. 7837–7846. doi: 10.1109/TIE.2015.2475419.

Liu, T., Hu, X., Li, S., *et al.* (2017) ‘Reinforcement Learning Optimized Look-Ahead Energy Management of a Parallel Hybrid Electric Vehicle’, *IEEE/ASME Transactions*

on Mechatronics, 4435(c), pp. 1–1. doi: 10.1109/TMECH.2017.2707338.

Liu, T., Hu, X., Li, S. E., *et al.* (2017) 'Reinforcement Learning Optimized Look-Ahead Energy Management of a Parallel Hybrid Electric Vehicle', *IEEE/ASME Transactions on Mechatronics*, pp. 1497–1507. doi: 10.1109/TMECH.2017.2707338.

Liu, T. *et al.* (2018) 'Reinforcement Learning-Based Predictive Control for Autonomous Electrified Vehicles', *IEEE Intelligent Vehicles Symposium, Proceedings*, 2018-June(Iv), pp. 185–190. doi: 10.1109/IVS.2018.8500719.

Liu, T. *et al.* (2019) 'A Heuristic Planning Reinforcement Learning-Based Energy Management for Power-Split Plug-in Hybrid Electric Vehicles', *IEEE Transactions on Industrial Informatics*, pp. 6436–6445. doi: 10.1109/TII.2019.2903098.

Liu, T., Tang, X., Hu, X., *et al.* (2020) 'Human-like Energy Management Based on Deep Reinforcement Learning and Historical Driving Experiences', *arXiv*, pp. 1–8.

Liu, T., Tang, X., Chen, J., *et al.* (2020) 'Transferred energy management strategies for hybrid electric vehicles based on driving conditions recognition', *2020 IEEE Vehicle Power and Propulsion Conference, VPPC 2020 - Proceedings*. doi: 10.1109/VPPC49601.2020.9330856.

Liu, T. and Hu, X. (2018) 'A Bi-Level Control for Energy Efficiency Improvement of a Hybrid Tracked Vehicle', *IEEE Transactions on Industrial Informatics*, 14(4), pp.

1616–1625. doi: 10.1109/TII.2018.2797322.

Liu, T., Wang, B. and Yang, C. (2018) ‘Online Markov Chain-based energy management for a hybrid tracked vehicle with speedy Q-learning’, *Energy*, 160, pp. 544–555. doi: 10.1016/j.energy.2018.07.022.

Loganathan, N., Mayurappriyan, P. S. and Lakshmi, K. (2019) ‘Demand side EMS using artificial neural network-based LP method’, *IOP Conference Series: Materials Science and Engineering*, 623(1), pp. 0–8. doi: 10.1088/1757-899X/623/1/012001.

Lu, C. *et al.* (2018) ‘Learning Driver-Specific Behavior for Overtaking: A Combined Learning Framework’, *IEEE Transactions on Vehicular Technology*, 67(8), pp. 6788–6802. doi: 10.1109/TVT.2018.2820002.

M. Sabri, M F, Danapalasingam, K. A. and Rahmat, M. F. (2016) ‘A review on hybrid electric vehicles architecture and energy management strategies’, *Renewable and Sustainable Energy Reviews*, 53, pp. 1433–1442. doi: 10.1016/j.rser.2015.09.036.

M. Sabri, M. F., Danapalasingam, K. A. and Rahmat, M. F. (2016) ‘A review on hybrid electric vehicles architecture and energy management strategies’, *Renewable and Sustainable Energy Reviews*, 53, pp. 1433–1442. doi: 10.1016/j.rser.2015.09.036.

Ma, K. (2019) ‘ScienceDirect ScienceDirect Numerical Investigation on Logic Control Energy Numerical Investigation on Fuzzy Fuzzy Logic Control Energy Management

Strategy of Parallel Hybrid Electric Vehicle Management Strategy of Parallel Hybrid Electric Vehicle Assessi', *Energy Procedia*, 158, pp. 2643–2648. doi: 10.1016/j.egypro.2019.02.016.

Mamun, A. Al *et al.* (2019) 'An integrated design and control optimization framework for hybrid military vehicle using lithium-ion battery and supercapacitor as energy storage devices', *IEEE Transactions on Transportation Electrification*, 5(1), pp. 239–251. doi: 10.1109/TTE.2018.2869038.

Martinez, C. M. *et al.* (2017) 'Energy Management in Plug-in Hybrid Electric Vehicles: Recent Progress and a Connected Vehicles Perspective', *IEEE Transactions on Vehicular Technology*, 66(6), pp. 4534–4549. doi: 10.1109/TVT.2016.2582721.

Martínez, C. M. *et al.* (2019) 'Conclusions, discussion and future direction for research', *Ihorizon-Enabled Energy Management for Electrified Vehicles*, pp. 335–374. doi: 10.1016/B978-0-12-815010-8.00008-9.

Millo, F., Rolando, L. and Andreat, M. (2011) 'Numerical Simulation for Vehicle Powertrain Development', *Numerical Analysis - Theory and Application*, (September). doi: 10.5772/24111.

Min Chen, Student Member, IEEE, and Gabriel A. Rinc´on-Mora, Senior Member, I. (2006) 'Accurate Electrical Battery Model Capable of Predicting Runtime and I–V

Performance’.

Montazeri-gh, M. and Mahmoodi-k, M. (2015) ‘Development a new power management strategy for power split hybrid electric vehicles’, *Transportation Research Part D*, 37, pp. 79–96. doi: 10.1016/j.trd.2015.04.024.

Mordor intelligence (no date) *OFF-HIGHWAY HYBRID VEHICLES MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2022 - 2027)*. Available at: https://www.mordorintelligence.com/industry-reports/off-highway-hybrid-vehicles-market.

Muhammad, A. and Haruna, I. S. (2021) ‘Hybrid Electric Vehicles: A mini Overview’, *Journal of Modern Manufacturing Systems and Technology*, 5(1), pp. 27–36. doi: 10.15282/jmmst.v5i1.6055.

Murphey, Y. L. *et al.* (2012) ‘Intelligent hybrid vehicle power control part I: Machine learning of optimal vehicle power’, *IEEE Transactions on Vehicular Technology*, 61(8), pp. 3519–3530. doi: 10.1109/TVT.2012.2206064.

Murphey, Y. L. *et al.* (2013) ‘Intelligent hybrid vehicle power control - Part II: Online intelligent energy management’, *IEEE Transactions on Vehicular Technology*, 62(1), pp. 69–79. doi: 10.1109/TVT.2012.2217362.

Nyong-Bassey, B. E. *et al.* (2020) ‘Reinforcement learning based adaptive power

pinch analysis for energy management of stand-alone hybrid energy storage systems considering uncertainty', *Energy*, 193, p. 116622. doi: 10.1016/j.energy.2019.116622.

Okui, N. and Kobayashi, M. (2015) 'A Study on Hybrid Control Method for Improvement of Fuel Economy and Exhaust-gas Emission of Hybrid Trucks', in *JSAE/SAE 2015 International Powertrains, Fuels & Lubricants Meeting*. Kyoto. doi: 10.4271/2015-01-1780.

Optimization, S. and Hwang, H. (2020) 'Optimized Fuel Economy Control of Power-Split Hybrid Electric Vehicle with Particle'.

Ouren, B. D. S., Haas, C. and Melcher, C. P. (2007) 'Environmental Effects of Off-Highway Vehicles on Bureau of Land Management Lands: A Literature Synthesis, Annotated Bibliographies, Extensive Bibliographies, and Internet Resources'.

Panday, A. and Bansal, H. O. (2014) 'A review of optimal energy management strategies for hybrid electric vehicle', *International Journal of Vehicular Technology*, 2014. doi: 10.1155/2014/160510.

Peksen, M. (2021) 'Hydrogen technology towards the solution of environment-friendly new energy vehicles', *Energies*, 14(16). doi: 10.3390/en14164892.

Qi, C. *et al.* (2022) 'Hierarchical reinforcement learning based energy management

strategy for hybrid electric vehicle', *Energy*, 238, p. 121703. doi: 10.1016/j.energy.2021.121703.

Qi, X. *et al.* (2015) 'A Novel Blended Real-Time Energy Management Strategy for Plug-in Hybrid Electric Vehicle Commute Trips', *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2015-Octob, pp. 1002–1007. doi: 10.1109/ITSC.2015.167.

Qi, X. *et al.* (2016) 'Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles', *Transportation Research Record*, 2572, pp. 1–8. doi: 10.3141/2572-01.

Qi, X. *et al.* (2019) 'Deep reinforcement learning enabled self-learning control for energy efficient driving', *Transportation Research Part C: Emerging Technologies*, 99(December 2018), pp. 67–81. doi: 10.1016/j.trc.2018.12.018.

Reddy, N. P. *et al.* (2019) 'An Intelligent Power and Energy Management System for Fuel Cell/Battery Hybrid Electric Vehicle Using Reinforcement Learning', *ITEC 2019 - 2019 IEEE Transportation Electrification Conference and Expo*. doi: 10.1109/ITEC.2019.8790451.

Richman, R. and Wüthrich, M. V. (1996) 'Nagging predictors', *Risks*, 8(3), pp. 1–26. doi: 10.3390/risks8030083.

Salmanx, M., Schouten, N. J. and Kheir, N. A. (2000) ‘Control strategies for parallel hybrid vehicles’, (June).

Salmasi, F. R. (2007) ‘Control Strategies for Hybrid Electric Vehicles: Evolution, Classification, Comparison, and Future Trends’, *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, 56(5), pp. 2393–2404. doi: 10.1109/tvt.2007.899933.

Schaffnit, J. and Isermann, R. (2000) ‘Hardware-in-the-Loop Simulation and Rapid Prototyping for the Development of Control Functions for Diesel Engines’, *IFAC Proceedings Volumes*, 33(26), pp. 811–817. doi: 10.1016/S1474-6670(17)39245-5.

Schilperoort, J. *et al.* (2019) ‘Learning to Play Pac-Xon with Q-Learning and Two Double Q-Learning Variants’, in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*. IEEE, pp. 1151–1158. doi: 10.1109/SSCI.2018.8628782.

Seaman, A., Dao, T. S. and McPhee, J. (2014) ‘A survey of mathematics-based equivalent-circuit and electrochemical battery models for hybrid and electric vehicle simulation’, *Journal of Power Sources*, 256, pp. 410–423. doi: 10.1016/j.jpowsour.2014.01.057.

Shamsi, H. *et al.* (2021) ‘Macro-Level optimization of hydrogen infrastructure and supply chain for zero-emission vehicles on a canadian corridor’, *Journal of Cleaner*

Production, 289, p. 125163. doi: 10.1016/j.jclepro.2020.125163.

Shen, P. *et al.* (2018) 'Optimal energy management strategy for a plug-in hybrid electric commercial vehicle based on velocity prediction', *Energy*, 155(4800), pp. 838–852. doi: 10.1016/j.energy.2018.05.064.

Shi, D. *et al.* (2021) 'Pontryagin ' s minimum principle based fuzzy adaptive energy management for hybrid electric vehicle using real-time traffic information State of Charge', *Applied Energy*, 286(August 2020), p. 116467. doi: 10.1016/j.apenergy.2021.116467.

Shuai, B. *et al.* (2020) 'Heuristic action execution for energy efficient charge-sustaining control of connected hybrid vehicles with model-free double Q-learning', *Applied Energy*, 267(November 2019). doi: 10.1016/j.apenergy.2020.114900.

Song, C. *et al.* (2019) 'A power management strategy for parallel PHEV using deep Q-networks', *2018 IEEE Vehicle Power and Propulsion Conference, VPPC 2018 - Proceedings*, (2), pp. 1–5. doi: 10.1109/VPPC.2018.8604984.

Su, C. W. *et al.* (2021) 'Can new energy vehicles help to achieve carbon neutrality targets?', *Journal of Environmental Management*, 297(March), p. 113348. doi: 10.1016/j.jenvman.2021.113348.

Sun, C., Sun, F. and He, H. (2017) 'Investigating adaptive-ECMS with velocity

forecast ability for hybrid electric vehicles', *Applied Energy*, 185, pp. 1644–1653. doi: 10.1016/j.apenergy.2016.02.026.

Sun, H. *et al.* (2020) 'Data-driven reinforcement-learning-based hierarchical energy management strategy for fuel cell/battery/ultracapacitor hybrid electric vehicles', *Journal of Power Sources*, 455(March), p. 227964. doi: 10.1016/j.jpowsour.2020.227964.

Sun, R. and Peterson, T. (1999) 'Multi-agent reinforcement learning: Weighting and partitioning', *Neural Networks*, 12(4–5), pp. 727–753. doi: 10.1016/S0893-6080(99)00024-6.

Sutton, R. S. and Barto, A. G. (2012) 'Reinforcement learning: An Introduction Second edition', *Learning*, 3(9), p. 322.

Teng Liu, Hong Wang, Li Li, Fellow, IEEE, Dongpu Cao, and Fei-Yue Wang, Fellow, I. A. (2018) 'Reinforcement learning-based predictive control framework for autonomous vehicle.pdf'.

Tian, Y. *et al.* (2021) 'Optimal control strategy for parallel plug-in hybrid electric vehicles based on dynamic programming', *World Electric Vehicle Journal*, 12(2). doi: 10.3390/wevj12020085.

Tran, D. D. *et al.* (2020) 'Thorough state-of-the-art analysis of electric and hybrid

vehicle powertrains: Topologies and integrated energy management strategies', *Renewable and Sustainable Energy Reviews*, 119, p. 109596. doi: 10.1016/j.rser.2019.109596.

Tran, M. K. *et al.* (2021) 'A review of range extenders in battery electric vehicles: Current progress and future perspectives', *World Electric Vehicle Journal*, 12(2). doi: 10.3390/wevj12020054.

Vandenhove, T. *et al.* (2020) 'Electrical variable transmission for hybrid off-highway vehicles', *Proceedings - 2020 International Conference on Electrical Machines, ICEM 2020*, pp. 2176–2182. doi: 10.1109/ICEM49940.2020.9270723.

Veer, K. *et al.* (2020) 'Feed‐forward modeling and real‐time implementation of an intelligent fuzzy logic‐based energy management strategy in a series – parallel hybrid electric vehicle to improve fuel economy', *Electrical Engineering*, 102(2), pp. 967–987. doi: 10.1007/s00202-019-00914-6.

Wang, H., Zariphopoulou, T. and Zhou, X. Y. (2020) 'Reinforcement learning in continuous time and space: A stochastic control approach', *Journal of Machine Learning Research*, 21, pp. 1–34.

Wang, S. *et al.* (2019) 'Fuzzy Adaptive-Equivalent Consumption Minimization Strategy for a Parallel Hybrid Electric Vehicle', 7. doi:

10.1109/ACCESS.2019.2941399.

Wang, W., Zeng, X. and Wang, Q. (2003) ‘Develop Hybrid Transit Buses for Chinese Cities’, in *2003 SAE World Congress*. Detroit, Michigan: SAE International. doi: 10.4271/2003-01-0087.

Wang, Y. H., Li, T. H. S. and Lin, C. J. (2013) ‘Backward Q-learning: The combination of Sarsa algorithm and Q-learning’, *Engineering Applications of Artificial Intelligence*, 26(9), pp. 2184–2193. doi: 10.1016/j.engappai.2013.06.016.

Wang, Y., Zeng, X. and Song, D. (2020) ‘Hierarchical optimal intelligent energy management strategy for a power-split hybrid electric bus based on driving information *’, *Energy*, 199, p. 117499. doi: 10.1016/j.energy.2020.117499.

Wipke, K. B., Cuddy, M. R. and Burch, S. D. (1999) ‘ADVISOR 2.1: A User-Friendly Advanced Powertrain Simulation Using a Combined Backward/Forward Approach’, *IEEE Transactions on Vehicular Technology: Special Issues on Hybrid and Electric Vehicles*.

Wirasingha, S. G. and Emadi, A. (2011) ‘Classification and review of control strategies for plug-in hybrid electric vehicles’, *IEEE Transactions on Vehicular Technology*, 60(1), pp. 111–122. doi: 10.1109/TVT.2010.2090178.

Wu, J. *et al.* (2018) ‘Continuous reinforcement learning of energy management with

deep Q network for a power split hybrid electric bus', *Applied Energy*, 222(January), pp. 799–811. doi: 10.1016/j.apenergy.2018.03.104.

Wu, Y. *et al.* (2019) 'Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus', *Applied Energy*, 247(September 2018), pp. 454–466. doi: 10.1016/j.apenergy.2019.04.021.

Xing, Y. *et al.* (2020) 'An ensemble deep learning approach for driver lane change intention inference', *Transportation Research Part C: Emerging Technologies*, 115(March), p. 102615. doi: 10.1016/j.trc.2020.102615.

Xiong, R., Cao, J. and 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(5), pp. 538–548. doi: 10.1016/j.apenergy.2017.11.072.

Xiong, X., Sha, J. and Jin, L. (2021) 'Optimizing coordinated vehicle platooning : An analytical approach based on stochastic dynamic programming', *Transportation Research Part B*, 150(March 2020), pp. 482–502. doi: 10.1016/j.trb.2021.06.009.

Xu, B. *et al.* (2019) 'Real-Time reinforcement learning optimized energy management for a 48V mild hybrid electric vehicle', *SAE Technical Papers*, 2019-April(April), pp. 1–9. doi: 10.4271/2019-01-1208.

Xu, B. *et al.* (2020) ‘Ensemble Reinforcement Learning-Based Supervisory Control of Hybrid Electric Vehicle for Fuel Economy Improvement’, *IEEE Transactions on Transportation Electrification*, 6(2), pp. 717–727. doi: 10.1109/TTE.2020.2991079.

Xu, B., Hou, J., *et al.* (2021) ‘Learning Time Reduction Using Warm-Start Methods for a Reinforcement Learning-Based Supervisory Control in Hybrid Electric Vehicle Applications’, *IEEE Transactions on Transportation Electrification*, 7(2), pp. 626–635. doi: 10.1109/TTE.2020.3019009.

Xu, B., Tang, X., *et al.* (2021) ‘Q-Learning-Based Supervisory Control Adaptability Investigation for Hybrid Electric Vehicles’, *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10. doi: 10.1109/TITS.2021.3062179.

Yang, C. *et al.* (2020) ‘A Stochastic Predictive Energy Management Strategy for Plug-in Hybrid Electric Vehicles Based on Fast Rolling Optimization’, *IEEE Transactions on Industrial Electronics*, 67(11), pp. 9659–9670. doi: 10.1109/TIE.2019.2955398.

Yang, T. *et al.* (2020) ‘Reinforcement learning in sustainable energy and electric systems: a survey’, *Annual Reviews in Control*, 49, pp. 145–163. doi: 10.1016/j.arcontrol.2020.03.001.

Yanjun Huang a, Hong Wang a, b, *, Amir Khajepour a, Hongwen He b, J. J. c (2017)

‘Model predictive control power management strategies for HEVs: A review’, pp. 91–106.

Yi, L., He, H. and Peng, J. (2016) ‘Hardware-in-Loop Simulation for the energy management system development of a Plug-in hybrid electric bus’, 88, pp. 950–956. doi: 10.1016/j.egypro.2016.06.118.

Yu, Y., Cai, Z. and Liu, Y. (2021) ‘Double deep Q-learning coordinated control of hybrid energy storage system in island micro-grid’, *International Journal of Energy Research*, 45(2), pp. 3315–3326. doi: 10.1002/er.6029.

Zhang, F. *et al.* (2019) ‘Energy management strategies of connected HEVs and PHEVs: Recent progress and outlook’, *Progress in Energy and Combustion Science*, 73, pp. 235–256. doi: 10.1016/j.pecs.2019.04.002.

Zhang, J., Jiao, X. and Yang, C. (2021) ‘A Double-Deep Q-Network-Based Energy Management Strategy for Hybrid Electric Vehicles under Variable Driving Cycles’, *Energy Technology*, 9(2), pp. 1–12. doi: 10.1002/ente.202000770.

Zhang, M. L. and Zhou, Z. H. (2013) *Exploiting unlabeled data to enhance ensemble diversity*, *Data Mining and Knowledge Discovery*. doi: 10.1007/s10618-011-0243-9.

Zhang, P., Yan, F. and Du, C. (2015) ‘A comprehensive analysis of energy management strategies for hybrid electric vehicles based on bibliometrics’,

Renewable and Sustainable Energy Reviews, 48(205), pp. 88–104. doi: 10.1016/j.rser.2015.03.093.

Zhang, Q. *et al.* (2019) 'A Double Deep Q-Learning Model for Energy-Efficient Edge Scheduling', *IEEE Transactions on Services Computing*, 12(5), pp. 739–749. doi: 10.1109/TSC.2018.2867482.

Zhang, W. *et al.* (2020) 'Reinforcement learning-based intelligent energy management architecture for hybrid construction machinery', *Applied Energy*, 275(5988), p. 115401. doi: 10.1016/j.apenergy.2020.115401.

Zhang, Y. *et al.* (2018) 'Human-like Autonomous Vehicle Speed Control by Deep Reinforcement Learning with Double Q-Learning', *IEEE Intelligent Vehicles Symposium, Proceedings*, 2018-June(15), pp. 1251–1256. doi: 10.1109/IVS.2018.8500630.

Zhao, P. *et al.* (2018a) 'A deep reinforcement learning framework for optimizing fuel economy of hybrid electric vehicles', *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC*, 2018-Janua, pp. 196–202. doi: 10.1109/ASPDAC.2018.8297305.

Zhao, P. *et al.* (2018b) 'A Deep Reinforcement Learning Framework for Optimizing Fuel Economy of Hybrid Electric Vehicles', pp. 196–202.

Zhou, J. *et al.* (2021) ‘A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning’, *Energy*, 224, p. 120118. doi: 10.1016/j.energy.2021.120118.

Zhou, M., Yu, Y. and Qu, X. (2020) ‘Development of an Efficient Driving Strategy for Connected and Automated Vehicles at Signalized Intersections: A Reinforcement Learning Approach’, *IEEE Transactions on Intelligent Transportation Systems*, 21(1), pp. 433–443. doi: 10.1109/TITS.2019.2942014.

Zhou, Q. *et al.* (2017) ‘Cyber-Physical Energy-SavIPing Control for Hybrid Aircraft-Towing Tractor based on Online Swarm Intelligent Programming’, *IEEE Transactions on Industrial Informatics*. doi: 10.1109/TII.2017.2781230.

Zhou, Q. *et al.* (2019) ‘Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle’, *Applied Energy*, 255(August). doi: 10.1016/j.apenergy.2019.113755.

Zhou, Q. *et al.* (2020) ‘Modified Particle Swarm Optimization with Chaotic Attraction Strategy for Modular Design of Hybrid Powertrains’, *IEEE Transactions on Transportation Electrification*, 7782(c). doi: 10.1109/TTE.2020.3014688.

Zhou, Q. *et al.* (2021) ‘Knowledge Implementation and Transfer with an Adaptive Learning Network for Real-Time Power Management of the Plug-in Hybrid Vehicle’,

IEEE Transactions on Neural Networks and Learning Systems, 32(12), pp. 5298–5308. doi: 10.1109/TNNLS.2021.3093429.

Zhou, Q. *et al.* (2022) 'Transferable representation modelling for real-time energy management of the plug-in hybrid vehicle based on k-fold fuzzy learning and Gaussian process regression', *Applied Energy*, 305, p. 117853. doi: 10.1016/j.apenergy.2021.117853.

Zhu, X. *et al.* (2015) 'Optimization approach to depot location in car sharing systems with big data', *2015 IEEE International Congress on Big Data*, pp. 335–342. doi: 10.1109/BigDataCongress.2015.57.

Zhu, Z. *et al.* (2021) 'Safe Model-based Off-policy Reinforcement Learning for Eco-Driving in Connected and Automated Hybrid Electric Vehicles', pp. 1–14.

Zou, R. *et al.* (2021) 'DQL energy management: An online-updated algorithm and its application in fix-line hybrid electric vehicle', *Energy*, 225, p. 120174. doi: 10.1016/j.energy.2021.120174.

Zou, Y. *et al.* (2016) 'Reinforcement learning-based real-time energy management for a hybrid tracked vehicle', *Applied Energy*, 171, pp. 372–382. doi: 10.1016/j.apenergy.2016.03.082.