Applied Energy 255 (2019) 113755

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle

Quan Zhou[a], Ji Li[a], Bin Shuai[a], Huw Williams[a], Yinglong He[a], Ziyang Li[a], Hongming Xu[a,*], Fuwu Yan[b]

[a] Department of Mechanical Engineering, University of Birmingham, Birmingham B15 2TT, UK
[b] School of Automotive Engineering, Wuhan University of Technology, Wuhan 430070, China

HIGHLIGHTS

- Reinforcement Learning is researched for energy saving in a hybrid vehicle.
- Energy efficiency can be continuously improved by multi-step learning.
- The 'Recurrent-to-Terminal' strategy is shown the most effective learning strategy.
- The method can save energy by more than 7.8% in the selected real-time operations.

ARTICLE INFO

Keywords:
Model-free predictive control
Energy management
Multi-step reinforcement learning
Markov decision problem
Hybrid electric vehicle

ABSTRACT

The energy management system of an electrified vehicle is one of the most important supervisory control systems which manages the use of on-board energy resources. This paper researches a 'model-free' predictive energy management system for a connected electrified off-highway vehicle. A new reinforcement learning algorithm with the capability of 'multi-step' learning is proposed to enable the all-life-long online optimisation of the energy management control policy. Three multi-step learning strategies (Sum-to-Terminal, Average-to-Neighbour Recurrent-to-Terminal) are researched for the first time. Hardware-in-the-loop tests are carried out to examine the control functionality for real application of the proposed 'model-free' method. The results show that the proposed method can continuously improve the vehicle's energy efficiency during the real-time hardware-in-the-loop test, which increased from the initial level of 34% to 44% after 5 h' 35-step learning. Compared with a well-designed model-based predictive energy management control policy, the model-free predictive energy management method can increase the prediction horizon length by 71% (from 35 to 65 steps with 1 s interval in real-time computation) and can save energy by at least 7.8% for the same driving conditions.

1. Introduction

Climate change-associated environmental protection concerns are among the world's greatest challenges [1]. Allied with carbon dioxide (CO_2) mitigation through sustainable approaches [2], optimisation of energy usage can effectively reduce CO_2 emissions [3]. Driven by the increasingly stringent emission and fuel consumption regulations, conventional vehicle systems are undergoing electrification, e.g. regenerative electric brakes [4], electrified suspension [5,6], and electrified powertrain [7]. Hybrid electric vehicle (HEV) is an immediately available solution for vehicle electrification to help achieve the goal of CO_2 emission reduction in road transportation. The energy management system plays the most critical role for HEV CO_2 emission control and much effort is made to determine the distribution of power from each source (engine or motor) simultaneously satisfying the driver's demand, minimising energy consumption/emissions, and maintaining the battery's state-of-charge SoC [8].

The new real-world driving emissions (RDE) test requires original equipment manufacturers (OEMs) to optimise the performance of their products in real world driving [9]. Real world driving is a Markov decision process (MDP), which is a process for decision making in stochastic situations [10]. Advanced energy management systems should have the capability to operate online optimisation to determine the optimal power distribution in any transient and stochastic environment. Model-based predictive control (MPC) is a widely-used method for optimisation of the vehicle energy management system

* Corresponding author.
E-mail address: h.m.xu@bham.ac.uk (H. Xu).

https://doi.org/10.1016/j.apenergy.2019.113755
Received 14 June 2019; Received in revised form 6 August 2019; Accepted 17 August 2019
Available online 28 August 2019
0306-2619/

Nomenclature

P	power (W)
s	state
a	action
r	reward
p	prediction horizon
t	time
SoC	state of charge
R	resistance (Ω)
I	current (A)
$Loss$	power loss (W)
Π	control policy
Q	expected system performance
$\boldsymbol{S}$	set of states
$\boldsymbol{A}$	set of actions
$\boldsymbol{R}$	set of rewards
$\boldsymbol{Q}$	Q table
$\boldsymbol{D}$	set of data collected
Φ	learning strategy
OEM	original equipment manufacturer

Subscripts

dem	demand
h	history
eng	engine-generator set
$batt$	battery package
ini	initial
ref	reference

[8,11]. A two-step rolling optimisation process including 'prediction' and 'solving' is used for MPC to obtain the optimal control signal in each sampling interval. The performance of MPC is affected by three main aspects: the accuracy of predictive models [12,13]; the length of predictive horizon [14,15], and the optimisation ability of its algorithm [16,17].

Recently, the Vehicle to Everything (V2X) network has been developed for the connected vehicles to transmit a great amount of data and engage with networked computing resources. It allows advanced nonlinear model-based predictive control for further improvement of vehicle performance. The V2X network can extend the prediction horizon using the data from the global positioning system (GPS) [14,18]. Adaptive modelling is available with the V2X data and connected computing resources [19,20]. Online control optimisation with nonlinear prediction models can be carried out using the connected computing resources [21,22]. The development of prediction models and optimisation algorithms, however, still requires human knowledge to maximise control performance.

The recent progress on artificial intelligence (AI) has demonstrated that machines have the potential to beat human beings, in developing better strategies (control policy) for certain problems (e.g. playing board games [23]). AI has been widely used for automotive technology research, especially for autonomous and intelligent vehicles (e.g. lane change intention prediction [24], driver activity recognition [25], driver behaviour prediction [26], etc.). However, it is difficult to implement advanced AI algorithms in conventional vehicle controllers, because AI-based control needs a large memory space to store data and a powerful processor to run the learning algorithms [27]. With the help of the V2X network, which can provide big data and connected computing resources, the concept of 'model-free' brings automotive engineering research into a new era [28]. The control policies for model-free control are optimised by updating a merit function which is a function of the rewards representing the performance of the vehicle system. Vehicle system models are no longer required for model-free control, therefore, the model-free control can avoid the negative influence of the inaccuracy of the vehicle models [29].

Many energy management methods for hybrid vehicles have been developed with the standard 'one-step' reinforcement learning algorithms, which updates the merit/value function Q with the information of two neighbouring time instants [30]. To minimise the vehicle's fuel consumption over a predictive horizon, one-step reinforcement learning works with a Markov chain model which predicts the power demand over the prediction horizon [31,32]. The main drawback of using the one-step method is that obtaining a reward $r(t)$ only directly affects the values of the state-action pair $s(t)$ and $a(t)$ that led to the reward at a single step. The values of other state action pairs are affected only indirectly through the updated value$Q(s(t), a(t))$. This can make the learning process slow since many updates are required for the propagation of a reward to the relevant preceding states and actions [33]. Also from the aspect of predictive energy management control, one-step learning has to work with a Markov chain model to obtain the control action that can achieve the maximum reward over a predictive horizon [34,35]. On the other hand, the Markov model will significantly affect the vehicle performance and the computing effort [36].

Recent development in computer science suggests that multi-step reinforcement learning can achieve optimal model-free predictive control without extra training of Markov chain models [33]. Multi-step reinforcement learning is considered as a promising technology for optimisation of the energy systems with high uncertainty in power demand. Although many successful applications using this technology to solve complex decision making problems can be found from the literature, such as playing the game of Go [23], multi-step learning strategies should be developed for problems [37]. To the best of the authors' knowledge, no report is yet available in the literature concerning the research of multi-step reinforcement learning for the energy management of hybrid vehicles. Based on the above observations and the authors' recent research on an electrified off-highway vehicle [22,38], this paper proposes a new model-free predictive energy management method for an electrified off-highway vehicle. The following work was carried out for the first time: three multi-step learning strategies were developed for model-free predictive energy management of a hybrid vehicle; the learning performance of different learning strategies was investigated for optimisation of the energy management control policy; the computing efforts of model-free predictive control were studied for preparation of real-time control; and hardware-in-the-loop tests were performed to evaluate the feasibility of real application using the proposed method.

The rest of this paper is organised as follows: Section 2 introduces the off-highway vehicle, its model-free energy management system, and the problems associated with the model-free control. The multi-step reinforcement learning for model-free predictive energy management control is proposed in Section 3, followed by the description of the experimental system for real-time validation and evaluation in Section 4. In Section 5, the performance of the proposed energy management method is evaluated from the perspectives of learning performance, computing effort, and real-time control capability. The advantages of the proposed method are also demonstrated by comparing the vehicle performance with the conventional model-based method based on hardware-in-the-loop testing. Section 6 summarises the conclusions.

2. Energy management for the connected off-highway vehicle

2.1. The connected vehicle system

This paper develops a model-free predictive energy management method based on a connected hybrid aircraft-towing tractor, as shown

in Fig. 1. The off-highway vehicle works at an airport, which includes complex and interconnected traffic system involving aircraft, towing tractors, and ground support vehicles. The V2X network (connecting the vehicles, aircraft, and infrastructures) provides connected live data and computing resources for online optimization of the vehicle control system. This will save energy by organising the traffic and individual vehicle operation.

The aircraft towing scenario studied in this paper is one of the most typical individual vehicle operation scenarios at the airport [22], which consists of a towing tractor, an aircraft, an airport control, and V2X communications. The airport control firstly receives the tractor and aircraft's data via the V2X network; secondly, performs online control policy optimisation in the server computer; and finally, sends the optimised control policy back to the vehicle controller via the V2X network. This network with advanced intelligent algorithms enables online control policy optimisation for energy saving, which was previously limited by the performance of the local vehicle controller.

The hybrid aircraft towing tractor has a plug-in series hybrid configuration as shown in Fig. 2, which uses electricity from a battery package (consisting of 8200 lithium-ion cells) as the primary power to drive a 245 kW traction motor. An 86.2 kW engine-generator is equipped as an alternative power unit to provide extra power for vehicle operation and charging the battery package. The energy management system (EMS) determines the amount of power contributed by the engine-generator and the battery package to satisfy the power demand and maintain the state of charge (SoC) of the battery package.

2.2. The Model-free predictive energy management system

A model-free predictive energy management system is developed for the vehicle based on a layered and distributed control framework, as illustrated in Fig. 3. The system consists of two main layers: the control layer (located in the on-board vehicle controller) and the learning layer (located in the server computer). The two distributed control layers are connected via the V2X network. The control layer applies the EMS control policy to allocate the power-flow based on the driver's inputs and vehicle state (i.e. fuel consumption and battery's SoC). The learning layer executes the reinforcement learning algorithm and regularly updates the control policy, so that the control policy of the EMS can be adapted to real-world driving through multi-objective optimization over the predictive horizon.

The interaction between the energy management system (EMS) and its external environment (including the driver and the vehicle system) can be described as a Markov decision process (MDP) [31,34,39]. The main uncertainty of this MDP is caused by drivers in real-world driving conditions, which directly affects the power demand of the hybrid system. The four main components (i.e. state, action, reward, and control policy) within the MDP of this research are defined as follows:

(a) State

The state variables used in this study are defined as:

$$s(t) = \{\hat{P}_{dem}(t)\widehat{SoC}(t)\} \tag{1}$$

where $\hat{P}_{dem}(t) \in \mathcal{P}$ is the modified driver's power demand value based on actual measurement from the energy management controller; $\mathcal{P} = \{0, 0.5, 1, \cdots, 252.5, 253\}$; $\widehat{SoC} \in \mathcal{S}$ is the battery's modified SoC value based on battery data measured by the controller; and $\mathcal{S} = \{20, 20.1, 20.2, \cdots, 79.9, 80\}$. A nearest neighbourhood search method is used to project the measured data $\{P_{dem}(t)SoC(t)\}$ to the respective data in the state space $\{\hat{P}_{dem}(t)\widehat{SoC}(t)\}$. The SoC value is estimated based on its real-time current, battery capacity and initial SoC, the nearest neighbourhood search method works as a filter to isolate the noise and error of SoC value.

(b) Action

The action is the control signal, $u_{apu}(t)$, of the engine-generator:

$$a(t) = u_{apu}(t) \tag{2}$$

where $u_{apu}(t) \in \mathcal{U}$; $\mathcal{U} = \{0, 0.05, 0.1, \cdots, 0.95, 1\}$. Once the control signal, $u_{apu}(t)$, of the engine-generator is obtained, the power requirement for the battery package can be obtained by:

$$u_{batt}(t) = \frac{P_{req} - u_{apu}(t)\cdot P_{apu\_max}}{P_{batt\_max}} \tag{3}$$

where P_{batt} is the power supplied by the battery package; P_{apu} is the power supplied by the engine generator; P_{apu_max} is the maximum power that can be provided by the engine generator; and P_{req} is the required power for driving the traction motor and powering the on-boarded auxiliary devices.

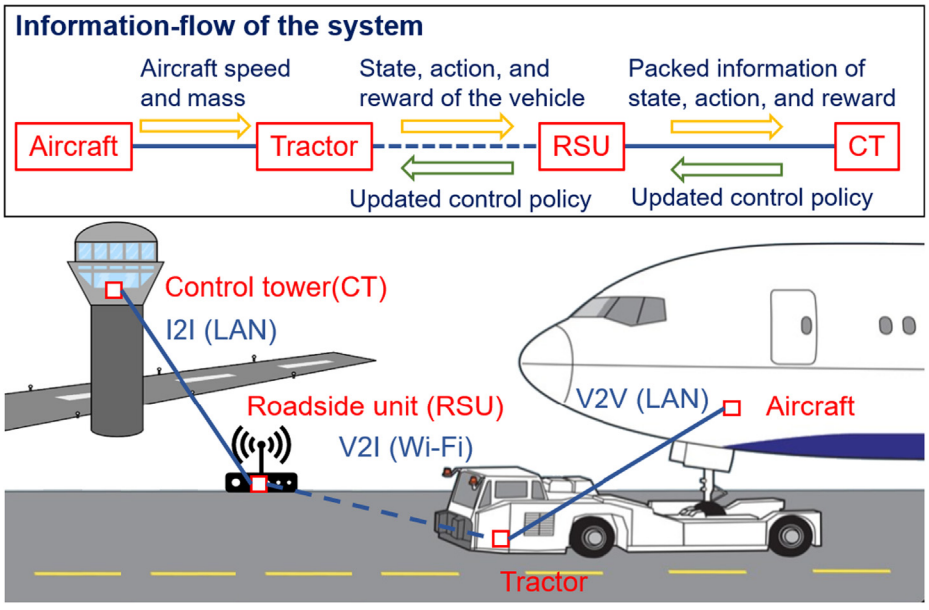

Fig. 1. Aircraft towing scenario with V2X communication.

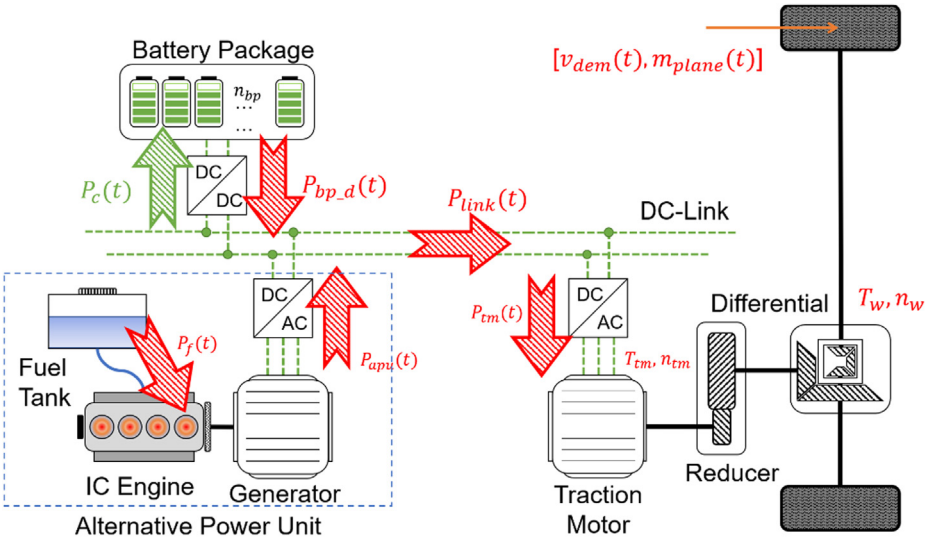

Fig. 2. System configuration and power-flow of the electrified off-highway vehicle.

(c) Reward

To minimise the vehicle power loss and maintain the battery SoC level simultaneously, the reward is defined as a function of the overall vehicle power loss P_{loss} and the remaining battery SoC, in which the power loss P_{loss} and absolute SoC value lower than the reference SoC and $|SoC_{ref} - SoC(t)|$ are added as penalties to the initial constant reward r_{ini} so that the learning system can remember which actions have been attempted and the rewards after these actions taken:

$$r(t) = \begin{cases} r_{ini} - P_{loss}(t) & SoC(t) \geq SoC_{ref}, \\ r_{ini} - P_{loss}(t) - \alpha\,|SoC_{ref} - SoC(t)| & SoC(t) < SoC_{ref}, \end{cases} \quad (4)$$

where 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 30%); α is a scale factor used to balance the consideration of SoC level and power efficiency; and $P_{loss}(t) = Loss_{eng}(t) + Loss_{batt}(t)$ is the total power loss of engine and battery, which comprises the power loss of the engine, $Loss_{eng}(t)$, and power loss of the battery, $Loss_{batt}(t)$. The latter two losses can be calculated by measuring the fuel consumption rate, $\dot{m}_f$, engine torque, T_{eng}, engine speed, n_{eng}, and battery current, I_{batt}, then using:

$$\begin{cases} Loss_{eng}(t) = \dot{m_f}(t) \cdot H_f - \frac{T_{eng}(t) \cdot n_{eng}(t)}{9550}, \\ Loss_{batt}(t) = R_{loss}(SoC) \cdot I_{batt}(t)^2, \end{cases} \quad (5)$$

where H_f is the heat value of the fuel (for diesel, $H_f = 44 \times 10^6$ J/kg) and R_{loss} is the equivalent internal resistance of the battery, which is a function of battery SoC.

(d) Control policy

A control policy Π is used to determine the action $\boldsymbol{a}(t)$ based on the current observation of vehicle state, $\boldsymbol{s}(t)$, which is a function of the state variables and Q values at the current state:

$$a(t) = \Pi(s(t) \;\; Q(\cdot)) \quad (6)$$

More specificity, a policy used for model-free control can be expressed as:

$$\Pi\text{: } \boldsymbol{a}(t) = \text{argmax}Q(s(t), :) \quad (7)$$

where $a(t)$ is the action value; $s(t)$ is the current state; Π is the control policy; and $Q(\cdot)$ is a database of the merit function values of all the possible states and actions in the MDP. The control policy is optimised by reinforcement learning through continuously updating the database of the merit function values$Q(\cdot)$.

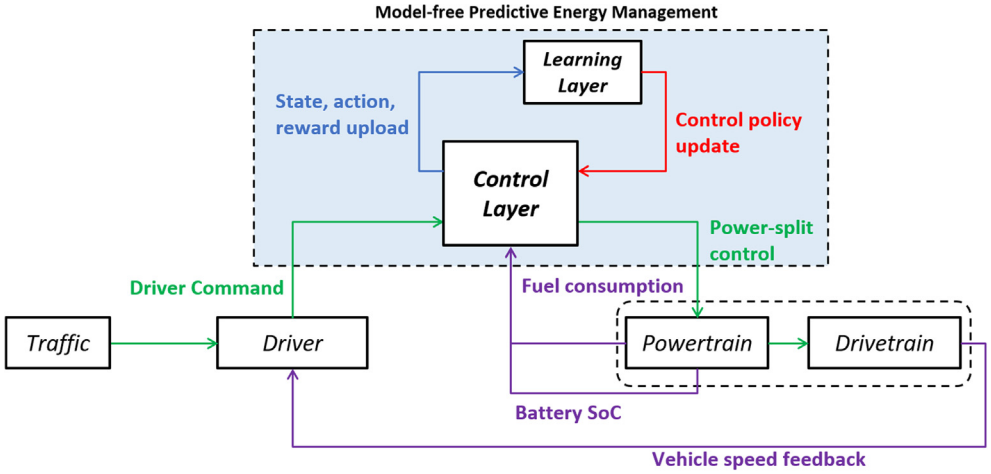

Fig. 3. Layered control framework for model-free predictive energy management.

3. Multi-step reinforcement learning

This paper proposes multi-step learning to optimise the control policy in Eq. (5), which will have the capability of predictive optimal energy management without any extra training of Markov chain models. Differently from one-step reinforcement learning, multi-step learning uses the history sets (data collected from $(t-p)^{th}$ time interval to the t^{th} time interval) of states$\boldsymbol{S}_h(t-p{:}\ t)$, actions$\boldsymbol{A}_h(t-p{:}\ t)$, and rewards $\boldsymbol{R}_h(t-p{:}\ t)$ to update the database of the merit function values Q($s(n)$, $a(n)$) at each time step t as follows:

$$Q(s(n), a(n)) \leftarrow \Phi(\boldsymbol{Q}, \boldsymbol{S}_h(t-p{:}\ t), \boldsymbol{A}_h(t-p{:}\ t), \boldsymbol{R}_h(t-p{:}\ t)) \quad n \in [t-p, t] \tag{8}$$

where $s(n) \in \boldsymbol{S}_h(t-p{:}\ t)$ and $a(n) \in \boldsymbol{A}_h(t-p{:}\ t)$ are state and action values at the n-th time step ($n = t-p, t-p+1, \cdots t$) of the history set collected at the t-th time interval; $\boldsymbol{S}_h(t-p{:}\ t)$; $\boldsymbol{A}_h(t-p{:}\ t)$, and $\boldsymbol{R}_h(t-p{:}\ t)$ are the history set of state, action and reward collected at the t-th time interval; $\boldsymbol{Q}$ is the Q-table before updating; and the Q-value is used to estimate the expected system performance. Here Φ is the n-step learning strategy, which aims to determine the optimal control policy.

Three multi-step learning strategies are investigated in this paper. The first learning strategy, named 'Sum-to-Terminal', is rooted in one-step Q-learning. The 'Sum-to-Terminal' strategy uses the sum of the rewards in the predictive horizon, the Q-value in the first step, and the terminal step to optimise the control policy. The second one ('Average-to-Neighbour') and the third one ('Recurrent-to-Terminal') are proposed to connect the Q-value in each step to the others within the predictive horizon using the rewards. The Average-to-Neighbour strategy updates the control policy using the average reward value and the Q-values of each of the neighbouring steps. The 'Recurrent-to-terminal' strategy builds a network of the Q-values of the current step with other steps within the predictive horizon. The three learning strategies will be described in detail in the following sections.

To clearly illustrate how the n-step learning strategies work, some concepts are defined as follows:

Definition 1. 'Data-set package' with notation $\boldsymbol{D}(t)$ defines the history sets of state, action and reward used for n-step learning at the t-th time interval as $\boldsymbol{D}(t) = \{\boldsymbol{S}_h(t-p{:}\ t), \boldsymbol{A}_h(t-p{:}\ t), \boldsymbol{R}_h(t-p{:}\ t)\}$.

Definition 2. 'Elements' with notation $d(i)$ defines the components within the data set package, where, $i \in [1, p]$ is the index of elements which will be defined in Definition 3. Each element includes the value of state, action and reward within the data-set package as $d(i) = \{s(i), a(i), r(i)\}$.

Definition 3. 'Order Index' with notation i defines the order of each element in the data-set package using for reinforcement learning at each time interval. For example, the order index of element $d = \{s(t-p+2), a(t-p+2), r(t-p+2)\}$ in the data-set package $\boldsymbol{D}(t)$ is 3, but in data-set package $\boldsymbol{D}(t+1)$, its order index will be 2.

Definition 4. 'Table-Lookup Index' with notation τ defines the position of each state and action in the set of state and action, which is used to find the respective Q value of each state and action from the Q table. For example, as the action set is A = {0, 0.05, 0.10, 0.15⋯.0.95, 1}, the table lookup index of action variable $a = 0.05$ is $\tau(a = 0.05) = 2$; a similar principle can be applied to the state variables.

3.1. Sum-to-terminal strategy (S2T)

The first n-step learning strategy named 'Sum-to-Terminal (S2T)' is a straightforward strategy as shown in Fig. 4.

With the data set $\boldsymbol{D}(t)$ packaged at each time step t, the S2T strategy connects the Q value of each elements $Q(s(i), a(i))$ to its terminal Q value $Q(s(p), a(p))$ directly with the sum of the reward during the time interval $\sum_{j=i}^{p} r(j)$:

$$Q(s(i), a(i)) \leftarrow Q(s(i), a(i)) + \alpha\left[\sum_{j=i}^{p} r(j) + maxQ(s(p), :) - Q(s(i), a(i))\right] \tag{9}$$

where $s(i) \in \boldsymbol{S}(t)$, $a(i) \in \boldsymbol{A}(t)$, $r(i) \in \boldsymbol{R}(t)$,are the state, action and reward with order index i respectively; $i = 1, 2\cdots p$ is the order index of each element within the data-set package; α is the learning rate; $maxQ(s(p), :)$ is the estimated terminal Q value obtained by look-up from the old Q table. The pseudo-code for the S2T strategy is provided in Fig. 5.

3.2. Average-to-neighbour strategy (A2N)

The 'Average-to-Neighbour (A2N)', as shown in Fig. 6, builds the relationship of each step by updating the Q values of neighbouring steps (i.e. $Q(s(i), a(i))$ and $Q(s(i+1), a(i+1))$) with the average reward of the predictive horizon $\frac{1}{p}\sum_{j=1}^{p} r(j)$.

The mathematical expression of A2N strategy for Q value optimisation is:

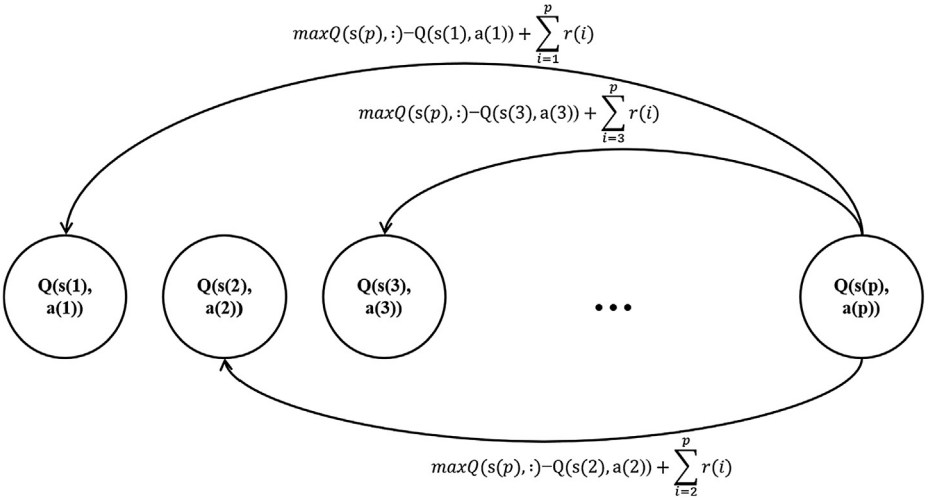

Fig. 4. Sum-to-Terminal (S2T) strategy for multi-step reinforcement learning.

Algorithm: n-step reinforcement learning with S2T strategy

Load Timer signal t
Load $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$, S, A, R, p
Choose action $a \in \boldsymbol{A}$ based on $\varepsilon-$ greedy policy
Recording s, a, r within predictive horizon p in $\boldsymbol{S_h}$, $\boldsymbol{A_h}$, $\boldsymbol{R_h}$
% start of the learning with S2T %
For $k = 1$: p
    Let the order index $i = k$
    Find the table lookup index of $s(i)$, $a(i)$ in S and A based on 'Nearest-Neighborhood policy'
    Find the table-lookup index of $s(p)$ in $\boldsymbol{S}$ based on 'Nearest-Neighbourhood policy'
    Update $Q(s(i), a(i))$ using Equation (9) with max Q(s(p)) and $\sum_{j=i}^{p} r(j)$
End for
% End of the learning with S2T%

Fig. 5. Pseudo-code for n-step reinforcement learning with S2T strategy.

$$Q(s(i), a(i)) \leftarrow Q(s(i), a(i)) + \alpha\left[\frac{\sum_{j=1}^{p} r(j)}{p} + maxQ(s(i+1), :) - Q(s(i), a(i))\right] \tag{10}$$

where $i = 1, 2\cdots p$ is the order index of each element within the data-set package. The pseudo-code for the A2N strategy is shown in Fig. 7.

3.3. Recurrent-to-terminal strategy (R2T)

The Recurrent-to-Terminal (R2T) strategy as shown in Fig. 8 is developed based on the Q(λ) learning algorithm [40].

The R2T strategy updates the Q-value recurrently step-by-step from the current step to the terminal step. A networked value function $\sum_{j=i}^{p} V(j)$ is introduced to update the Q value of each element $Q(s(i), a(i))$ using

$$Q(s(i), a(i)) \leftarrow Q(s(i), a(i)) + \alpha\cdot\sum_{j=i}^{p} V(j) \tag{11}$$

where $i = 1, 2\cdots p$ the order index of each element within the data-set package, $V(j)$ is the value function for R2T for the element with order index j, which is defined:

$$V(j) = \lambda^{j} r(j) + \lambda^{j+1}\text{maxQ}(s(j+1)) - Q(s(j), a(j)) \tag{12}$$

where λ is the discount factor, $r(j)$ is the reward at (j)-th time interval. The pseudo-code for the R2T strategy is given in Fig. 9.

4. Testing and validation set-up

The performance of multi-step reinforcement learning with various learning strategies is first investigated by tracking the evolution of vehicle's energy efficiency with the driving cycles defined by the tractor original equipment manufacturer (OEM). The vehicle's energy efficiency is the energy used for vehicle operation (e.g. driving and aircraft towing) divided by the equivalent energy in the fuel used from the tank and electricity supplied from the grid. The predictive capability is also investigated by evaluating the system performance with different predictive horizons. Next, the feasibility for real-time implementation is investigated by monitoring the computational time and performance for each learning algorithm. The model-free method for energy management of the off-highway vehicle real-time is validated in with two rounds of hardware-in-the-loop tests comprising 11 different scenarios. A comparison study with the model-based method is carried out to validate the advantages of the proposed method.

4.1. Driving cycles for machine learning and validation

Four driving cycles defined by the tractor OEM [22,38] were used in the present work and the power demand profiles of each driving cycle are shown in Fig. 10. The vehicle model following the driving cycle 1 (4252 s) for machine learning was first run repetitively for 24 h. The other three driving cycles were used to imitate the power demand diversity, and they were used to evaluate the model-free predictive energy management strategy. Driving cycles 2–4 represent three typical power demands of the aircraft towing tractors, which use the data collected from three vehicles working in different areas of the airport. Driving cycle 2 is built with a vehicle working with both large aircraft and small aircraft. Driving cycle 3 is built with a vehicle mainly working with large aircraft and driving cycle 4 uses the data of a vehicle mainly working with small aircraft. Further details in terms of the driving cycles can be found in the authors' previous research [28].

4.2. Hardware-in-the-loop testing facilities

The hardware-in-the-loop (HiL) test is used for testing the real-time performances of the energy management system. This will examine whether the energy management system can be used in real application

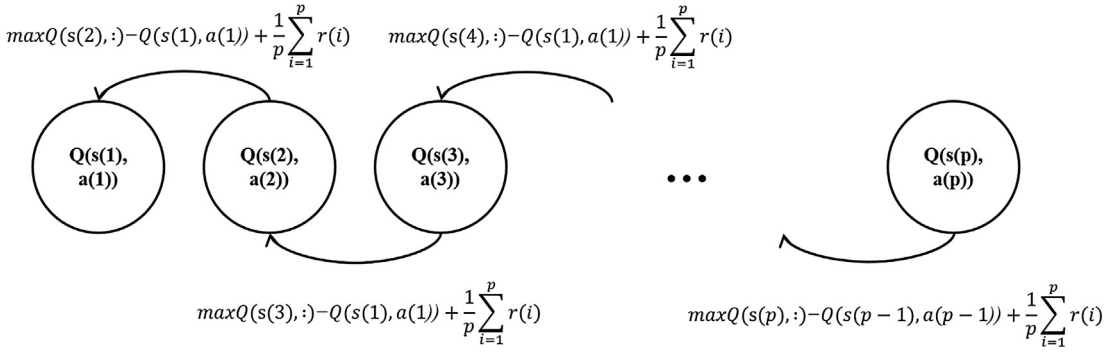

Fig. 6. Average-to-Neighbor (A2N) strategy for multi-step reinforcement learning.

Algorithm: n-step reinforcement learning with A2N strategy

Load Timer signal t
Load $\boldsymbol{Q(s,a)}$, S, A, R, p
Choose action $a \in \boldsymbol{A}$ based on $\varepsilon-$ greedy policy
Recording s, a, r within predictive horizon p in $\boldsymbol{S_h}$, $\boldsymbol{A_h}$, $\boldsymbol{R_h}$
% Start of the learning with A2N %
For $k = 1$: p
    Let the order index $i = k$
    Find the table lookup index of $s(i)$, $a(i)$ in S and A based on 'Nearest-Neighborhood policy'
    Find the table-lookup index of $s(p)$ in $\boldsymbol{S}$ based on 'Nearest-Neighbourhood policy'
    Update $Q(s(i), a(i))$ using Equation (10) with max $Q(s(i+1))$ and $\frac{\sum_{i=1}^{p} r(i)}{p}$
End for
% End of the learning with A2N%

Fig. 7. Pseudo-code for n-step reinforcement learning with A2N strategy.

for controlling the energy-flow of a hybrid electric powertrain. An Industrial standard testing system provided by the ETAS Group [41] will be used for the HiL test in this paper. The configuration of the HIL testing system is shown in Fig. 11. The layered control system for model-free predictive energy management runs on an ETAS ES910 prototype controller. The ES910 is configured with 1.5 GHz CPU, 4 GB RAM and a 1Gbps Ethernet port. The control algorithms are programmed in host PC-1 and flashed to ES910 by ETAS INTECRIO. The DESK- LABCAR performs as the hybrid aircraft towing tractor and communicates with the energy management controller (ES910) via CAN bus. The vehicle is modelled and compiled on host PC-2 and downloaded to DESK-LABCAR by ETAS Experiment Environment (EE) via Ethernet protocol. The performance of vehicle is monitored by ETAS EE in host PC-2. The real-time models for the HIL test are developed using Simulink as in the authors' previous work [22,38], and the models are verified by the test data from a prototype vehicle provided by the industrial partners [42].

5. Results and discussion

5.1. Learning performance

According the authors' recent research on energy management of the hybrid tractor using model-based predictive control, the maximum predictive length of model-based method using ES910 is 35 steps [22]. Therefore, the performance evaluation starts from the model-free predictive energy management with 35-step. The system efficiency of powertrain energy conversion is calculated for every 4252 s with the measured data of equivalent power loss for fuel consumption, battery power loss, and the power used by the traction motor, which is used to evaluate the learning performance.

The process of 'learning from scratch' (initial Q table is a zero-set) with all the 3 proposed learning strategies is monitored. The improvement of vehicle's energy efficiency during the learning process is shown in Fig. 12, which can be roughly classified into two stages, i.e., a rapid improving stage (at the beginning) and a slow improving stage (after sufficient experience for Q-table filling). The "$\varepsilon-$ greedy" [36] is used to generate an exponential reducing function for controlling the probability for self-exploration, which leads to a theoretical logarithmic improvement in system efficiency.

The model-free controller initially creates a control policy by itself and continuously updates the policy using feedback from the vehicle system. There are no human inputs for control policy modification nor is there any control parameter tuning during the whole learning processes. The purpose of the model-free predictive energy management system is to realise continuous optimisation of the energy management control policy without human intervention. This is evidenced by the fact the vehicle energy efficiency continuously increases with the time of learning, Fig. 12.

The discrete sampling of the system efficiency therefore generates a

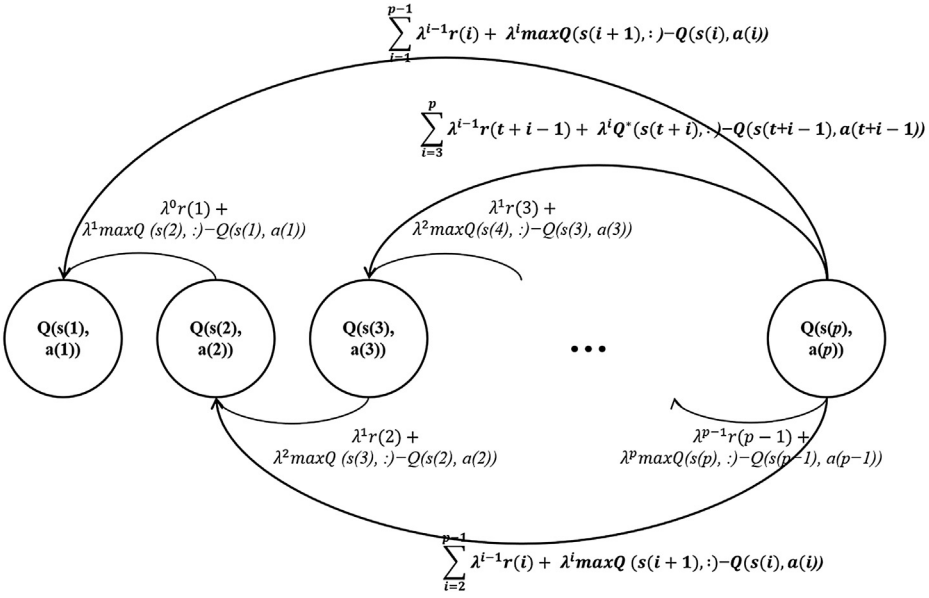

Fig. 8. Recurrent-to-Terminal (R2T) strategy for multi-step reinforcement learning.

Algorithm: n-step reinforcement learning with R2T strategy

Load Timer signal t
Load $\boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a})$, S, A, R, p
Choose action $a \in \boldsymbol{A}$ based on $\varepsilon-$ greedy policy
% Start of the learning with R2T %
For $k = 1{:}\,\mathrm{p}$
    Let the order index $i = k$
    Find the table lookup index of $s(i)$, $a(i)$ in S and A based on 'Nearest-Neighborhood policy'
    For $\mathrm{m} = 1{:}\,\mathrm{k}$
        Let the order index $j = m$
        Find the table-lookup index of $s(j)$, $a(j)$ in S and A based on 'Nearest-Neighborhood policy'
        Find the table-lookup index of $s(j+1)$ in S based on 'Nearest-Neighborhood policy'
        Calculate $V(j)$ using Equation (12) and reward value $r(j)$ with order index j
    End for
    Update $Q(s(i),a(i))$ using Equation (11)
End for
% End of the learning with R2T %

Fig. 9. Pseudo-code for multi-step reinforcement learning with R2T strategy.

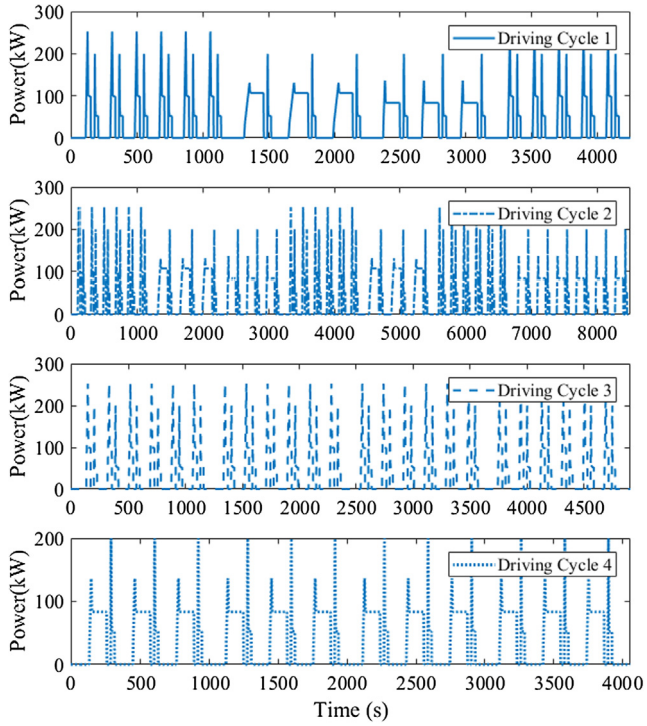

Fig. 10. Power demand of each driving cycle.

'knee' point at about 5 h for 35 step Q-learning. Among the three proposed learning strategies, the Recurrent-to-Terminal (R2T) strategy is the most efficient strategy for Model-free predictive energy management with the prediction horizon of 35 steps.

The global optimum solution for the energy management controller leads to the minimum total fuel consumption with an acceptable SoC value at the end of the trip. For predictive energy management of a HEV, a longer predictive horizon means the control will consider the vehicle performance with the information for a longer trip so it can make the system performance closer to its global optima for the trip [11]. The performance of the model-free predictive energy management was investigated by tracking the improvement of vehicle's energy efficiency with increasing prediction length and the results are presented in Fig. 13.

The system efficiencies after 24 h of reinforcement learning using Driving Cycle 1 with 4 different prediction step lengths are listed in Table 1. It shows that each of the 3 proposed learning strategies have the capability of improving the system performance by increasing the prediction length. More improvement on efficiency can be achieved with increased prediction horizon length when the prediction length is relatively low, for example, 18.67% energy efficiency improvement can be achieved when changing the prediction length from 35 steps to 55 steps using S2T strategy. When the prediction horizon is relatively high, the increasing of prediction length will no longer significantly improve the energy efficiency. The results also suggest that the R2T strategy outperforms the other strategies by achieving higher system efficiency at the end of the learning process, for all the selected prediction lengths.

5.2. Computing effort

The computational cost is a natural concern for real-time implementation and the prediction length is the most significant factor that affects the computational cost [20]. The computational cost of the proposed method with respect to prediction step size is investigated here using the ES910 prototype controller. The average computational cost per time step including the data communication is shown in Fig. 14. It indicates that while the augmented prediction size leads to increased computational load, prediction step size being less than 65 steps can make the controller with A2N and S2T strategy implementable in real-time while 60 steps can make the model-free energy management with R2T strategy implementable in real-time, as the computing time is less than the sampling time of 1 s. The average computing time for 65-step model-free method with A2N and S2T is 0.98 s and the average computing time for 60-step model-free method with R2T is 0.97 s. According to the preliminary research by the authors, the average computing time for 35-step model-based method is 0.92 s [22]. Comparing with the model-based method which has a maximum prediction horizon length of 35 steps for real-time control using the ES910 [22], the proposed method can help to increase prediction horizon length by at least 71%.

5.3. Capability for real time control

The model-free predictive energy management with R2T learning strategy outperforms other proposed strategies by achieving the best vehicle system efficiency with the same prediction length (as discussed in Section 5.1). The advantage of the R2T strategy is that it includes more iteration loops (as shown in Fig. 9) to build up a more complex learning system for storing more 'learning experience', and this additional complexity in contrast costs more in computational resources and shortens its maximum predictive step length in real-time.

The full performance in real-time of different learning strategies was

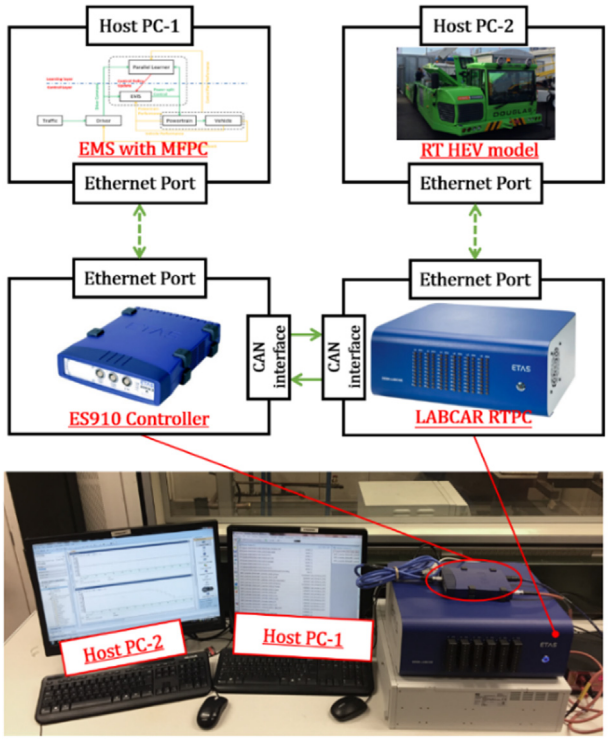

Fig. 11. Facilities for hardware-in-the-loop test.

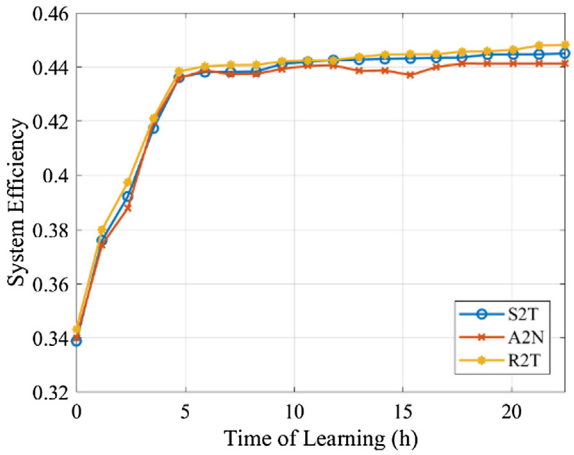

Fig. 12. Learning performance of different learning strategies (35-step).

Table 1
System performance of the proposed learning strategy with different prediction length.

	Prediction step length			
	35	55	85	125
R2T	44.66%	53.00%	55.47%	58.38%
A2N	44.12%	52.53%	55.46%	58.14%
S2T	44.55%	52.70%	54.99%	58.08%

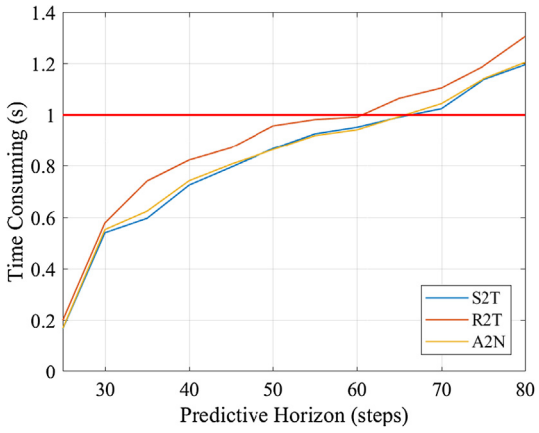

Fig. 14. Averange computation time per step for different learning strategy.

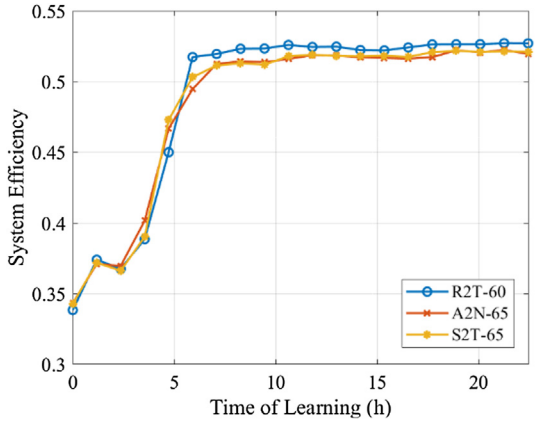

Fig. 15. The full performance in real-time of different learning strategy.

investigated by monitoring the learning performance of each strategy with its maximal step length in real-time (e.g. 60 steps for R2T, 65 steps for A2N and S2T) and the results are shown Fig. 15. In real-time,

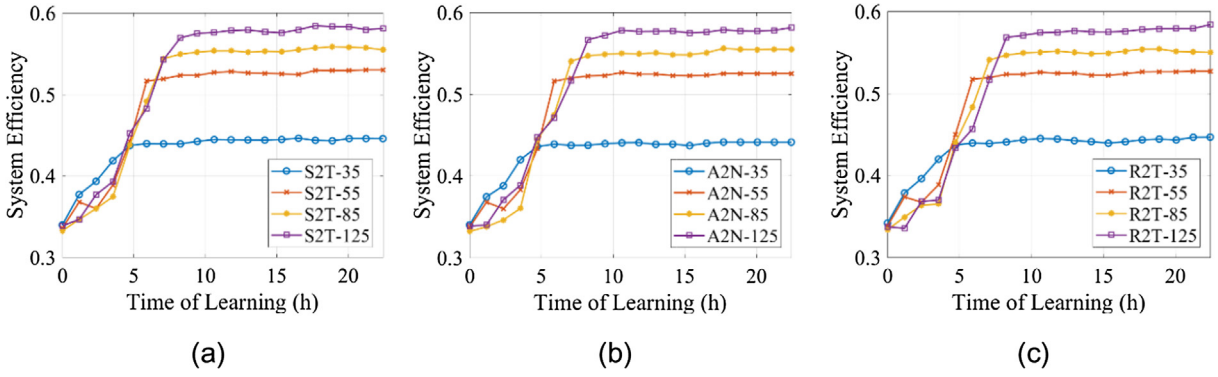

Fig. 13. Learning performance of different prediction length: (a)-S2T; (b)-A2N; (c)-R2T.

although the maximum prediction step size of R2T strategy is 5 steps shorter than other proposed strategies, it still outperforms other proposed strategies by achieving a better system efficiency at the end of the learning process and tends to continuously improve when it keeps learning from the real-world interaction.

5.4. Performance comparing with the model-based predictive method

The real-time performance of the model-free method with R2T learning strategy was compared with the model-based method via two rounds of hardware-in-the-loop (HiL) testing. In the first round HiL test, both methods control the vehicle's energy flow with a prediction horizon length of 35 steps. The real-time vehicle performance for different battery initial SoC values of 80% and 20% under driving cycle 1 are shown in Fig. 16(a) and (b) respectively.

The first round HiL testing results for the model-free method is shown in blue solid lines and the results of model-based method is presented in magenta dashed lines. The time history of energy loss, battery SoC, engine command, battery cell current/voltage are compared. The proposed control method can maintain the HEV's components working within the proper range in real-time. The model-free predictive energy management method outperforms model-based method in terms of both energy consumption and battery SoC maintenance.

The working condition varies between different scenarios; therefore, a test of robustness and repeatability is needed. The second round HiL tests were performed to evaluate the energy management methods in three driving cycles with different initial battery SoC values (80%, 50% and 20%). The full performance in real-time of the model-free method (with prediction length of 60 steps) was investigated and compared with model-based method (with prediction length of 35 steps [22]) in the second round test. The results in Table 2 indicate that the model-free method with its full performance capability leads to a significant improvement on energy saving. The energy saving rate is calculated by:

$$\Delta = \frac{E_{mb} - E_{mf}}{E_{mb}} \tag{13}$$

where Δ is the energy saving rate; E_{mb} is the total energy used for vehicle operation with the model-based method; E_{mf} is the total energy used for vehicle operation with the model-free method.

Comparing with model-based method, the model-free method can save at least 7.79% energy over the Driving cycle 2 when SoC is 20%; The highest energy saving range is obtained over the Real-world cycle-3 with the initial battery SoC of 50%, which can save 14.39% energy. The average energy saving is 10.68% for the 9 pairs of experiments.

From the comparison study with two rounds of HiL testing, the model-free method outperforms the model-based method in all the selected driving cycles, including the driving cycle for policy learning (driving cycle 1) and driving cycles for validation (driving cycle 2–4). There are three main reasons why the model-free method will perform better in real-time energy management control than the model-based method: (1) It is impossible to have a 'perfect model' to predict the performance of vehicles in all real world driving conditions; (2) the models used in control are fixed for the life of the vehicle without the ability to adapt to the changes in the vehicle itself e.g., aging; doing both online modelling and online model solving with model-based method is also very costly in terms of computational resource; (3) the model-free method has the capability of continuous control policy optimisation using the real-time feedback directly. This indicates that the proposed model-free predictive control method is a promising technology in the real application of hybrid vehicle energy management.

6. Conclusion

A new model-free predictive energy management method for a

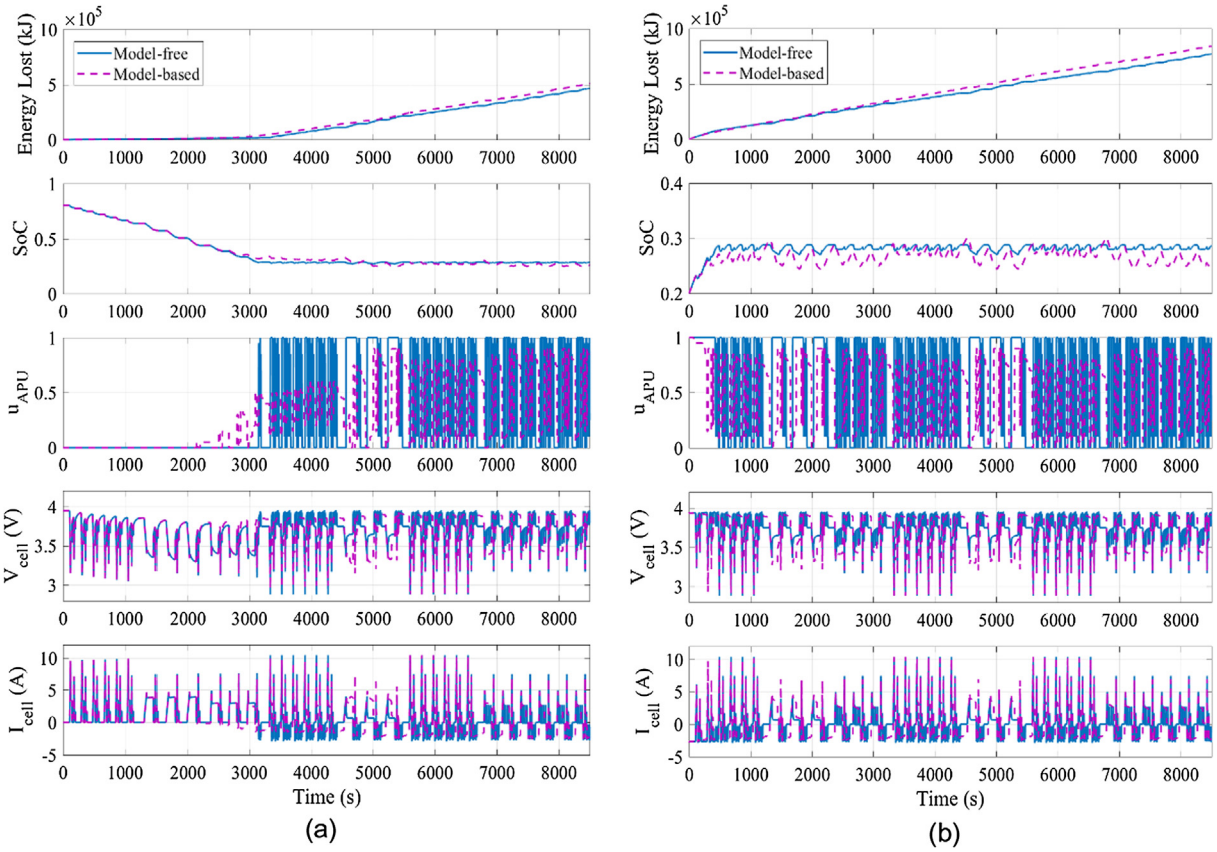

Fig. 16. Real-time performance of the vehicle: (a) Initial SoC = 80%; (b) Initial SoC = 20%

Table 2
Performance of model-free method and model-based method (full performance).

Cycle name	Initial SoC	Method	End SoC	Energy usage (MJ)	Saving rate
Driving Cycle 2	80%	Model-based	29.74%	844.08	–
	80%	Model-free	28.68%	778.25	7.79%
	50%	Model-based	29.69%	1016.96	–
	50%	Model-free	28.66%	922.41	9.29%
	20%	Model-based	29.73%	1177.86	–
	20%	Model-free	28.68%	1071.34	9.04%
Driving Cycle 3	80%	Model-based	29.74%	293.08	–
	80%	Model-free	28.81%	254.80	13.06%
	50%	Model-based	29.74%	469.34	–
	50%	Model-free	28.81%	401.78	14.39%
	20%	Model-based	29.74%	628.75	–
	20%	Model-free	28.81%	552.29	12.16%
Driving Cycle 4	80%	Model-based	29.74%	234.9	–
	80%	Model-free	28.81%	208.34	11.31%
	50%	Model-based	29.68%	394.2	–
	50%	Model-free	28.81%	354.75	10.08%
	20%	Model-based	29.68%	555.34	–
	20%	Model-free	28.81%	505.15	9.04%

hybrid vehicle has been studied and compared with the conventional model-based energy management method. Three different multi-step reinforcement learning strategies for the model-free predictive control were proposed and investigated for the first time. The learning performance and real-time implementation feasibility of the model-free method were evaluated in a hardware-in-the-loop testing system. The research has demonstrated the use of Artificial Intelligence (AI) and Internet of Things (IoT) for energy system optimisation in the transportation sector. The conclusions drawn from the investigation are as follows:

1. The Model-free predictive energy management method can achieve higher vehicle's energy efficiency after a certain time interval of vehicle operation through the proposed learning strategies. At least 29% of vehicle's energy efficiency improvement can be realised after 5 h real-time learning.
2. The proposed learning strategies can optimise the control policy in real-time with a maximum prediction length of 65 steps, and the optimal control policy can be obtained and implemented in the ES910 controller within 1 s; The maximum prediction horizon for real time control can be increased by at least 71% compared with the model-based method.
3. The proposed recurrent-to-terminal learning strategy is the most effective multi-step reinforcement learning strategy for model-free predictive energy management in the case study, and it outperforms in real-time the other proposed strategies with the same prediction step length.
4. The proposed model-free predictive energy management method is robust for energy saving and it outperforms the conventional model-based method through energy savings of at least 7.8%. An average 10.68% energy saving was achieved in the 9 pairs of hardware-in-the-loop tests.

Acknowledgement

The work is partially funded by Innovate UK (Grant 102253). The authors gratefully acknowledge the support from Textron Ground Support Equipment UK and Hyper-Drive Innovation.

References

[1] Fuso Nerini F, Sovacool B, Hughes N, Cozzi L, Cosgrave E, Howells M, et al. Connecting climate action with other Sustainable Development Goals. Nat Sustain 2019. https://doi.org/10.1038/s41893-019-0334-y.
[2] Sepehri A, Sarrafzadeh M-H. Activity enhancement of ammonia-oxidizing bacteria and nitrite-oxidizing bacteria in activated sludge process: metabolite reduction and CO_2 mitigation intensification process. Appl Water Sci 2019;9:131. https://doi.org/10.1007/s13201-019-1017-6.
[3] Pavlovic J, Marotta A, Ciuffo B. CO_2 emissions and energy demands of vehicles tested under the NEDC and the new WLTP type approval test procedures. Appl Energy 2016;177:661–70. https://doi.org/10.1016/j.apenergy.2016.05.110.
[4] Zhou Q, Guo X, Tan G, Shen X, Ye Y, Wang Z. Parameter analysis on torque stabilization for the eddy current brake: a developed model, simulation, and sensitive analysis. Math Probl Eng 2015;2015:1–10. https://doi.org/10.1155/2015/436721.
[5] Zhou Q, Guo X, Xu L, Wang G, Zhang J. Simulation based evaluation of the electro-hydraulic energy-harvesting suspension (EHEHS) for off-highway vehicles. SAE Tech Pap 2015 2015. https://doi.org/10.4271/2015-01-1494.
[6] Guo S, Chen Z, Guo X, Zhou Q, Zhang J. Vehicle interconnected suspension system based on hydraulic electromagnetic energy harvest: design, modeling and simulation tests. SAE Tech Pap 2014;1(2299). Submitted for publication.
[7] Lv C, Zhang J, Li Y, Yuan Y. Mechanism analysis and evaluation methodology of regenerative braking contribution to energy efficiency improvement of electrified vehicles. Energy Convers Manag 2015;92:469–82. https://doi.org/10.1016/j.enconman.2014.12.092.
[8] Marina Martinez C, Hu X, Cao D, Velenis E, Gao B, Wellers M. Energy management in plug-in hybrid electric vehicles: recent progress and a connected vehicles perspective. IEEE Trans Veh Technol 2016;66:1. https://doi.org/10.1109/TVT.2016.2582721.
[9] Continental Automotive GmbH. Worldwide emission standards and related regulations; 2017.
[10] Bellman RE. Dynamic programming. 2010th ed. Princeton, New Jersey: Princeton University Press; 2010.
[11] Huang Y, Wang H, Khajepour A, He H, Ji J. Model predictive control power management strategies for HEVs: a review. J Power Sources 2017;341:91–106. https://doi.org/10.1016/j.jpowsour.2016.11.106.
[12] Soriano F, Moreno-Eguilaz M, Álvarez-Flórez J. Drive cycle identification and energy demand estimation for refuse-collecting vehicles. IEEE Trans Veh Technol 2015;64:4965–73. https://doi.org/10.1109/TVT.2014.2382591.
[13] Sun C, Sun F, He H. Investigating adaptive-ECMS with velocity forecast ability for hybrid electric vehicles. Appl Energy 2017;185:1644–53. https://doi.org/10.1016/j.apenergy.2016.02.026.
[14] Huang Y, Khajepour A, Wang H. A predictive power management controller for service vehicle anti-idling systems without a priori information. Appl Energy 2016;182:548–57. https://doi.org/10.1016/j.apenergy.2016.08.143.
[15] Nuijten E, Koot MW, Kessels J. Advanced energy management strategies for vehicle power nets. EAEC 9th Int Congr Eur Automot Ind Driv Glob Chang; 2003. p. 1–9.
[16] Sampathnarayanan B, Serrao L, Onori S, Rizzoni G, Yurkovich S, Asme. Model predictive control as an energy management strategy for hybrid electric vehicles. Proc Asme Dyn Syst Control Conf 2009, Pts a B 2010. p. 1161–8. doi: 10.1115/DSCC2009-2671.
[17] Romijn TCJ, Donkers MCF, Kessels JTB, Weiland S. Receding horizon control for distributed energy management of a hybrid heavy-duty vehicle with auxiliaries. IFAC-PapersOnLine 2015;48:203–8. doi: 10.1016/j.ifacol.2015.10.029.
[18] Martínez CM, Cao D. iHorizon driver energy management for PHEV real-time control; 2019. doi: 10.1016/B978-0-12-815010-8.00006-5.
[19] Lv C, Wang H, Cao D. High-precision hydraulic pressure control based on linear pressure-drop modulation in valve critical equilibrium state. IEEE Trans Ind Electron 2017;0046:1. https://doi.org/10.1109/TIE.2017.2694414.
[20] Hu X, Wang H, Tang X. Cyber-physical control for energy-saving vehicle following with connectivity. IEEE Trans Ind Electron 2017;0046:1. https://doi.org/10.1109/TIE.2017.2703673.
[21] Liu T, Yu H, Guo H, Qin Y, Zou Y. Online energy management for multimode plug-in hybrid electric vehicles. IEEE Trans Ind Informatics 2018:1. https://doi.org/10.

1109/TII.2018.2880897.
[22] Zhou Q, Zhang Y, Li Z, Li J, Xu H, Olatunbosun O. Cyber-physical energy-saving control for hybrid aircraft-towing tractor based on online swarm intelligent programming. IEEE Trans Ind Inform 2018;14:4149–58. https://doi.org/10.1109/TII.2017.2781230.
[23] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550:354–9. https://doi.org/10.1038/nature24270.
[24] Xing Y, Lv C, Wang H, Wang H, Ai Y, Cao D, et al. Driver lane change intention inference for intelligent vehicles: framework, survey, and challenges. IEEE Trans Veh Technol 2019;68:1. https://doi.org/10.1109/TVT.2019.2903299.
[25] Xing Y, Lv C, Wang H, Cao D, Velenis E, Wang F-Y. Driver activity recognition for intelligent vehicles: a deep learning approach. IEEE Trans Veh Technol 2019:1. https://doi.org/10.1109/TVT.2019.2908425.
[26] Li J, Zhou Q, He Y, Shuai B, Li Z, Williams H, et al. Dual-loop online intelligent programming for driver-oriented predict energy management of plug-in hybrid electric vehicles dear editor and reviewers. Appl Energy 2019:253.
[27] Lv C, Hu X, Sangiovanni-Vincentelli A, Li Y, Martinez CM, Cao D. Driving-style-based codesign optimization of an automated electric vehicle: a cyber-physical system approach. IEEE Trans Ind Electron 2019;66:2965–75. https://doi.org/10.1109/TIE.2018.2850031.
[28] Qi X, Luo Y, Wu G, Boriboonsomsin K, Barth M. Deep reinforcement learning enabled self-learning control for energy efficient driving. Transp Res Part C Emerg Technol 2019;99:67–81. https://doi.org/10.1016/j.trc.2018.12.018.
[29] Liu T, Zou Y, Liu D, Sun F. Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle. IEEE Trans Ind Electron 2015;62:7837–46.
[30] Sutton RS, Barto AG. Reinforcement learning: an introduction The MIT Press; 2017. https://doi.org/10.1016/S1364-6613(99)01331-5.
[31] Zou Y, Kong Z, Liu T, Liu D. A real-time Markov chain driver model for tracked vehicles and its validation: its adaptability via stochastic dynamic programming. IEEE Trans Veh Technol 2017;66:3571–82. https://doi.org/10.1109/TVT.2016.2605449.
[32] Xiong R, Cao J, Yu Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Appl Energy 2018;211:538–48. https://doi.org/10.1016/j.apenergy.2017.11.072.
[33] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T, et al. Asynchronous methods for deep reinforcement learning. Proc 33th Int Conf Mach Learn 2016;48. https://doi.org/10.1177/0956797613514093.
[34] Liu T, Zou Y, Liu D, Sun F. Reinforcement learning-based energy management strategy for a hybrid electric tracked vehicle. Energies 2015;8:7243–60. https://doi.org/10.3390/en8077243.
[35] Wu J, He H, Peng J, Li Y, Li Z. Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus. Appl Energy 2018;222:799–811. https://doi.org/10.1016/j.apenergy.2018.03.104.
[36] Liu T, Hu X, Li S, Cao D. Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle. IEEE/ASME Trans Mechatronics 2017;4435:1. https://doi.org/10.1109/TMECH.2017.2707338.
[37] De Asis K, Hernandez-Garcia JF, Holland GZ, Sutton RS. Multi-step reinforcement learning: a unifying algorithm. Thirty-Second AAAI Conf Artif Intell; 2018. p. 2902–9.
[38] Zhou Q, Zhang W, Cash S, Olatunbosun O, Xu H, Lu G. Intelligent sizing of a series hybrid electric power-train system based on Chaos-enhanced accelerated particle swarm optimization. Appl Energy 2017;189:588–601. https://doi.org/10.1016/j.apenergy.2016.12.074.
[39] Hu Y, Li W, Xu K, Zahid T, Qin F, Li C. Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning. Appl Sci 2018;8:187. https://doi.org/10.3390/app8020187.
[40] Peng J, Williams RJ. Incremental multi-step Q-learning. Mach Learn 1996;22:283–90. https://doi.org/10.1007/BF00114731.
[41] ETAS products. ETAS Gr; 2017. < https://www.etas.com/en/index.php?langS = true > .
[42] Rachel Cooper. New hybrid aircraft push-back tractor on show at Inter Airport Europe Exhibition. Hyperdrive Innov Ltd; 2017. < https://hyperdriveinnovation.com/new-hybrid-aircraft-push-back-tractor-on-show-at-inter-airport-europe-exhibition/ > .