An Energy Management Strategy of Hybrid Electric Vehicles based on Deep Reinforcement Learning

Chengzhao Yang, Changcheng Zhou

Abstract—**An energy management strategy (EMS) plays important roles on the performance of hybrid electric vehicles (HEVs). However, the EMS based on rules is difficult to achieve an optimized result, whereas the EMS based on optimization theories cannot achieve the adaptability to different driving cycles. An energy management strategy of HEVs based on the deep reinforcement learning (DRL) is proposed in this research, which is fully data-driven and learning-enabled and does not rely on any experience of experts and accurate mathematical models. The proposed strategy is verified under a co-simulation environment, and the result shows that the proposed strategy achieves a 3.51% fuel saving compared to a rule-based strategy.**

***Index Terms*—hybrid electric vehicle, energy management strategy, deep reinforcement learning, simulation verification, fuel saving**

I. INTRODUCTION

HEVs have many advantages over traditional internal combustion engine (ICE) vehicles and electric vehicles due to the combination of the motor drive system and the engine drive system. At the same time, since HEVs have the characteristics of multi-power source and complicated drive system, how to adopt an efficient and reasonable energy management strategy and manage the energy flow between power sources is particularly important. At present, the existing energy management strategies are mainly divided into the following three categories: 1) Rule-based strategies rely on the experience of experts and experiment results, which are simple and effective. However, it is difficult to achieve an optimized result on the energy management using this type of strategies [1]. 2) Optimization-based strategies optimize the energy management for the entire trip based on the known or predicted future driving cycles using different optimal control theories [2]. The obvious disadvantage of this type of strategies is the dependency on the accurate trip information. 3) Learning-based strategies do not depend on future driving cycles, and the energy management strategy parameters can be adjusted to achieve a good adaptability to different driving cycles [3]. Nevertheless, this type of strategies are mostly based on the reinforcement learning (RL), and the strategies based on DRL are not fully developed yet [4].

In this research, an energy management strategy of HEVs based on the DRL is proposed, which combines the concept of the RL and a deep neural network to form a deep Q network, and obtains control actions directly from the input states. The key concept of the DRL-based energy management strategy is introduced, and the deep network for the estimation of Q function is established, and the algorithm calculation steps of the DRL-based energy management strategy is also introduced. The proposed strategy is verified by the computer simulation. Compared to a rule-based strategy, the proposed strategy performs better in both maintaining the stability of the state of charge (SOC) of battery and the fuel saving.

II. MODELING OF THE HEV

Establishing an accurate simulation model of the HEV improves the accuracy of the energy management algorithm and benefits for the subsequent energy management strategy research. In this research, the vehicle model is obtained by the backward modeling, in which the effects of dynamic response and temperature are ignored in various parts of the vehicle. Figure 1 shows the process of backward modeling.

In this research, a single-axle parallel HEV is adopted, and the power system structure is shown in Figure 2. It consists of the engine, motor, battery, automatic transmission and so on. According to the direction of the energy flow, the parallel HEV studied in this research has five main working modes: the electric mode, the engine mode, the hybrid driving mode, the charging mode, and the brake recovery mode. The vehicle parameters of the HEV are shown in TABLE 1.

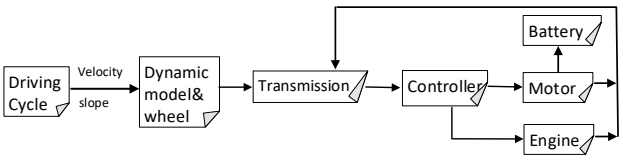

Figure 1. Process of backward modeling

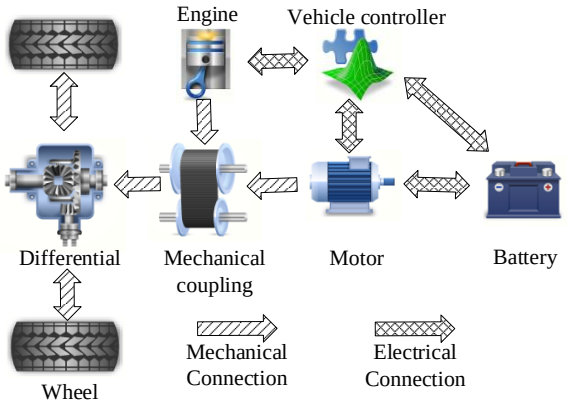

Figure2. Structure of a parallel HEV powertrain

TABLE 1. Vehicle parameters

A. Vehicle Dynamics

By ignoring the lateral dynamic model of the vehicle, the torque demand on the wheel $T_{rq_{whl}}$ is calculated from the

longitudinal dynamic model of the vehicle when the speed v and the slope α are given, as follows:

$$T_{rq_{whl}}=r(\frac{C_D A v^2}{21.15}+fmgcos\alpha+mgsin\alpha+\delta\frac{dv}{dt}) \quad (1)$$

The rotational speed of the wheel n_{whl} is calculated by:

$$n_{whl}=\frac{v}{2\pi r} \quad (2)$$

where r is the wheel radius, C_D is the air resistance coefficient, A is the frontal area, f is the rolling resistance coefficient, δ is the rotational mass conversion coefficient, m is the curb weight of the vehicle.

B. ICE Model

A quasi-static model is utilized to evaluate the fuel economy of an engine. The average engine efficiency is defined as：

$$\eta_e=\frac{T_e\omega_e t}{9550 m_f q} \quad (3)$$

where T_e is the engine torque, ω_e is the angular speed of the engine, t is the engine working time, m_f is the fuel consumption, q is the specific heat of the fuel.

The fuel consumption per second of an engine is defined as:

$$Q_e=\frac{T_e\omega_e b_e}{367g\rho_g} \quad (4)$$

where b_e is the fuel consumption rate, ρ_g is the gasoline density.

C. Motor Model

Ignoring the electromagnetic characteristics of the motor, the motor power is defined as:

$$P_m=T_m\omega_m\eta_m^{-sgn(T_m)} \quad (5)$$

where T_m is the motor torque, ω_m is the angular speed of the motor, η_m is the motor efficiency.

D. Battery Model

SOC is an important parameter of the battery pack. The effect of the temperature on the battery pack is ignored in this research, and the internal resistance model of the battery is used in this research, as follows:

$$SOC(t+1)=SOC(t)-\frac{V_{OC}-\sqrt{V_{OC}^2-4R_{int}P_m(t)}}{2(R_{int}+R_t)Q_{max}} \quad (6)$$

where V_{OC} is the open-circuit voltage, R_{int} is the internal resistance, R_t is the terminal impedance, Q_{max} is the maximum charging capacity, $P_m(t)$ is the power provided by the battery at the time step t.

III. DRL-Based Energy Management Strategy

According to the HEV model described in Section II, a Q-learning based EMS is proposed in this section.

A. DRL Theory

Curb weight ***m***	Gross weight ***G***	Wheel radius ***r***
1350kg	**1690kg**	**0.36m**
Rotational mass conversion coefficient ***δ***		Frontal area ***A***
1.05		**3.2m²**
Friction coefficient	Wind resistance coefficient C_D	
0.015	**0.63**	

The agent and the environment are two essential elements of the DRL. The agent needs to learn how to make decisions and the environment provides the external support of learning for the agent. The agent interacts with the environment in the process of executing tasks, which is an interactive process of the RL [5]. The RL is a type of learning that maps from the environmental states to the action (shown in Fig 3). S is taken as the set of all environmental states, $s_t \in S$ represents the state at time step t, A represents the set of executable actions, and the policy π: S→A represents a mapping from the state space to the action space. At each time step t, the agent chooses an action a_t from A, and the environment feeds back a corresponding reward r_t to the agent, and the state correspondingly transfers to the new state s_{t+1}. The agent adjusts the strategy and makes a new decision according to the reward. In the process of learning interaction, the cumulative reward from time t to the end time T is defined as:

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \quad (7)$$

where $\gamma\in[0,1]$ is used to weigh the impact of the future reward on the cumulative reward.

The goal of the RL is to find an optimal strategy π^*, so that the agent gains the maximum cumulative reward for any state and at any time, that is:

$$\pi^*=argmax_\pi E_\pi(\sum_t^T\gamma^k r_{t+k}|s_t=s) \quad (8)$$

The most commonly used algorithm of the RL is the Q-learning algorithm $Q^\pi(s_t,a_t)$, which indicates that the agent selects the action a_t on the state s_t according to a certain strategy π. The state-action value function is formulated by:

$$Q(s_t,a_t)=r_{t+1}+\gamma\max_a Q(s_{t+1},a) \quad (9)$$

In the traditional RL, the state-action value function is generally formulated by iterating Berman equation, and the state-action value function is finally converged by continuous iteration, so that the optimal strategy is obtained. However, the environmental state is usually with high dimensions in practice, thus solving the optimal strategy by evaluating each action on each state independently is not feasible. The feature of the state can be learned directly from the original input by the deep learning (DL) represented by the deep neural network, so the deep Q-learning network (DQN) combined with the deep neural network and the Q learning is proposed (shown in Fig3). The agent directly learns the control strategy from the high-dimensional input by the DQN. The optimal value function is approximated by the Q-valued function with the parameter θ in the deep Q network:

$$Q(s_t,a_t,\theta)\approx Q^*(s_t,a_t) \quad (10)$$

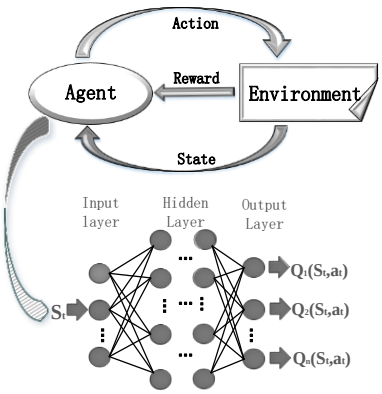

Figure 3. Model of DRL

B. Problem Formulation

The goal of the energy management strategy of HEVs is to minimize the fuel consumption under the constraints of the system while the power performance of the vehicle is

satisfied and the charge-discharge balance of the battery is maintained. Therefore, the energy management is a multi-objective optimization problem with constraints. The objective of the optimization problem is to minimize the cumulative cost function. The cost function is composed of the weighted sum of the fuel consumption and emission functions in the following form:

$$J_\pi = lim_{N\to\infty} E\left\{\sum_{t=0}^{N-1}\gamma^t \cdot L(t)\right\} \quad (11)$$

where N is the duration of the driving cycle, $L(t)$ is the cost function. In this research, the reward function R(t) is used to define the cost function $L(t)$.

C. State, Action, and Reward Determination

State: In the DQN algorithm, the control action is directly dependent on the state of the system. In this research, the driving state of the HEV can be defined by a 3-dimension vector:

$$S(t)=\{v(t),\alpha(t),SOC(t)\}^T \quad (12)$$

Action: The core of the energy management of HEVs is how to optimize the torque distribution ratio between the engine and the motor. Therefore, the output torque of the motor T_m is selected as the control action: A(t)= T_m . Meanwhile, the T_m is discretized into 36 points, and the torque of engine be subtracted from the total demand torque by the motor torque.

Reward: The reward function directly affects the adjustment of parameters of the deep neural network. The DQN algorithm tends to maximize the reward function at each time step. In this work, the reciprocal of the instantaneous battery SOC is selected as the reward function, as follows:

$$R(t)=\frac{1}{soc(t)} \quad (13)$$

D. DQN Framework

In this research, a DQN-based framework (in Fig 4) for the EMS of HEVs is designed to automatically learn the optimal control actions from input states without any predictive informations and presetted rules. The neural network has one input layer, one output layer with the linear activation function, and two hidden layers with 350 neurons in each layer and with the Rectified Linear Unit (ReLU) activation function. The network is trained within the iteration of a conventional RL. When choosing the action, the relationship between "exploration" and "exploitation" needs to be dealt correctly. This work uses the strategy of ε-greedy to explore the environment, i.e. selecting random actions with the ε probability and choosing actions based on the maximum Q value with the 1-ε probability.

The loss function is defined as:

$$L(\theta)= \zeta\,[r+\gamma maxQ\left(s',a';\theta^-\right)-Q(s,a;\theta)] \quad (11)$$

where $r+\gamma maxQ\left(s',a';\theta^-\right)$ is the temporal difference (TD) target, ζ is the learning rate, θ^- represents the parameter of the TD target network and θ represents the parameter of the main network. Building an independent TD target network, which is different from the current main network in parameters, and calculating the loss function accelerates the convergence and improves the stability of the algorithm.

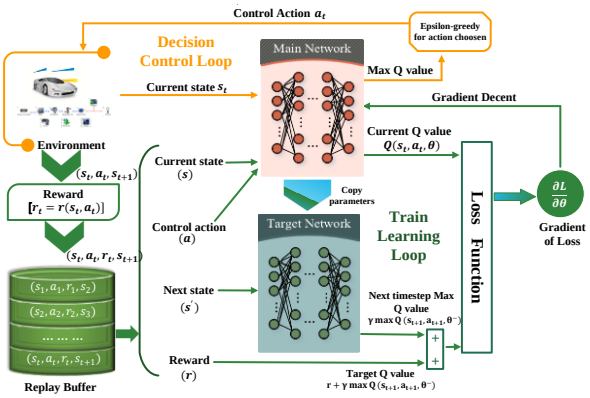

Figure 4. DQN-based framework for HEV EMS

When training the neural network, the independent-identic al-distribution condition needs to be satisfied for the training data. However, there are strong correlations among the traini ng samples in a short time period when training the network using the RL. In this research, the experience replay is adopt ed to break correlations among data. The pseudocode of the DRL is given in Algorithm 1.

IV. Simulation Setup And Results

The DQN algorithm is trained and evaluated under the UDDS driving cycle. The hyper parameters of the algorithm are shown in the TABLE 2.

TABLE 2. Hyper parameters

Hyper parameters	Value
Mini-batch size	48
Replay buffer size	1500
Buffer start size	150
Learning rate	0.9
Discount factor γ	0.95
Initial ε	1
Final ε	0.1

Algorithm 1：DRL-based Energy Management

Initialize replay buffer D to capacity N
Initialize action-value function Q with random weight θ
Initialize target action-value function Q' with weight $\theta^-=\theta$
For epoch=1, M do
  Initialize sequence $s_1=(v_1,\alpha_1,SOC_1)^T$
  For t=1, T do
    With probability ε select a random action C
    Otherwise select $a_t = maxQ(s_t,a;\theta)$
    Execute action a_t and observe reward r_t and s_{t+1}
    Store transition(s_t,a_t,r_t,s_{t+1})
    Sample random minibatch $(s_j,a_j,r_j,s_{j+1})$from D
    set $y_j = \begin{cases} r_j & \text{if terminal } s_{j+1} \\ r_j+\gamma maxQ'\left(s_{j+1},a';\theta^-\right) & \text{otherwise} \end{cases}$
    Perform a gradient descent step on $y_j - Q(s_{j+1},a';\theta)$ with respect to the network parameters θ
    Every K steps reset $Q'=Q$
  End for
End for

The effectiveness of the DQN algorithm is evaluated firstly. The track of the loss function is illustrated in Fig 5. It is obvious that the loss function drops rapidly but fluctuates at the beginning of the training. When the training is going on,

the loss function tends to be stable. Fig 6 shows the track of the total reward. It is clear that the reward increases when

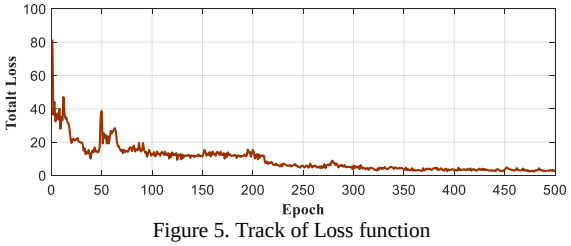

Figure 5. Track of Loss function

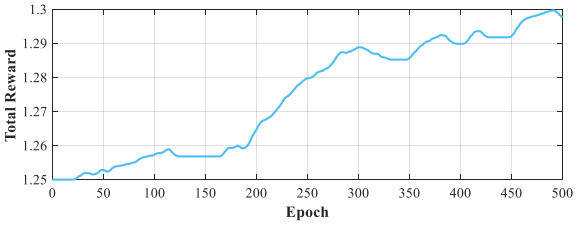

Figure 6. Track of Total reward function

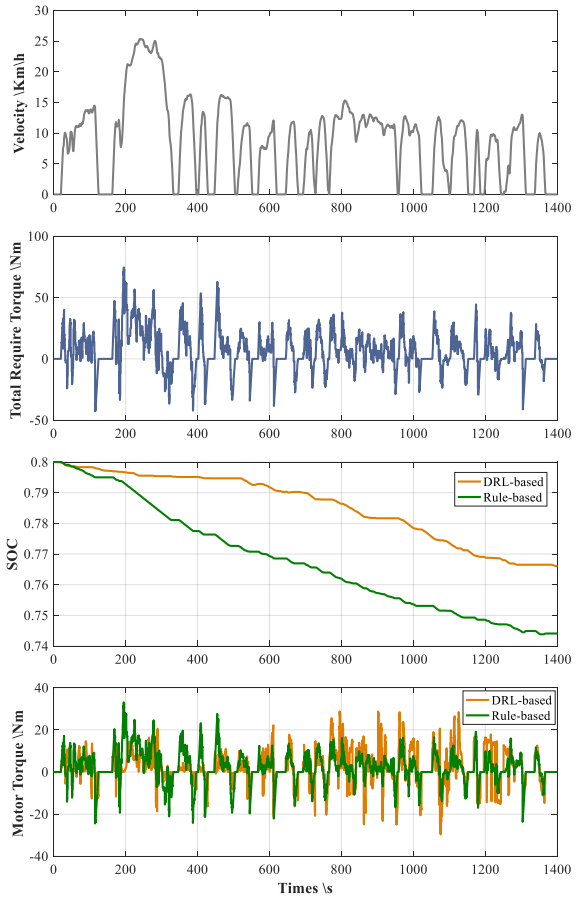

Figure 7. Comparison of simulation results

the number of training increases. In order to evaluate the proposed EMS better, its results are compared to those of a rule-based strategy under the UDDS condition. The comparison result is shown in Fig 7. For a fair comparison, the electric energy consumption is converted to the equivalent fuel consumption. The result shows that the DRL-based EMS achieves a 3.51% equivalent fuel consumption reduction compared to the rule-based strategy.

V. CONCLUSIONS

The energy management strategy based on the DRL, which combines the RL and the deep neural network to generate the deep Q network and obtains control actions directly from the input states, is proposed in this research. The simulation training is carried out under the UDDS driving cycle, and a better fuel-saving effect is achieved compared to a rule-based strategy.

REFERENCES

[1] X. Zhao, and G. Guo, “Survey on energy management strategies for hybrid electric vehicles,” J. Acta Automatica Sinica, vol.42, pp. 321–333, March 2016.

[2] R. He, and J.M. Li, “Survey on power coupling system and energy management strategy for hybrid electric vehicles,” J. Journal of Chongqing University of Technology(Nature Science), vol.32, pp.1-16, October 2018.

[3] J. Wu, H. He, J. Peng, Y. Li, and Z. Li, “Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus,” J. Applied Energy, vol.222, pp.799-811, April 2018.

[4] Z. Chen, C. Mi, R, Xiong, J. Xu, and C. You, “Energy management of a power-split plug in hybrid electric vehicle based on genetic algorithm and quadratic programming,” J. Power Sources, vol.248, pp.416-426, August, 2014

[5] D. Silver, H. Aja, J.C. Maddison, A, Guez, L. Sifre, J. Schrittwieser, and et al, “Mastering the game of go with deep nerual networks and tree search,” J. Nature, vol.518, pp.484-489, 2016.

Chengzhao Yang, he was born in December 1994 in Shandong Province, China. he is a graduate student at the Shandong University of Technology, and major in Vehicle Engineering. His research direction is the Energy Management of the Hybrid Electric Vehicle.

Changcheng Zhou, he was born in August 1965 in Shandong Province, China. He is a professor at Shandong University of Technology. Her research direction is the Vehicle System Dynamics and Energy Management.