A Bi-Level Control for Energy Efficiency Improvement of a Hybrid Tracked Vehicle

Teng Liu, *Member, IEEE*, and Xiaosong Hu, *Senior Member, IEEE*

***Abstract*—In this paper, a bi-level control framework is proposed to improve the energy efficiency for a hybrid tracked vehicle. The higher-level discusses how to accurately predict power demand based on the Markov Chain. Specially, fuzzy encoding predictor is used for power demand prediction, and a real-time recursive algorithm is applied to fuse the future power demand information into transition probability matrix (TPM) computation. Furthermore, the Kullback–Leibler (KL) divergence rate is employed to decide the alteration of control strategy. The lower-level computes the relevant energy management strategy, based on the updated TPM and a model-free reinforcement learning (RL) technique. Simulation results illustrate that the vehicular energy efficiency in the proposed scheme exceeds the common RL control by tuning the KL divergence value. Comparative results also show that the developed control strategy outperforms the common RL one, in terms of energy efficiency and computational speed.**

***Index Terms*—Energy management, hybrid tracked vehicle (HTV), Kullback–Leibler (KL) divergence rate, power demand prediction, reinforcement learning (RL).**

I. Introduction

DUE to a great importance of improving fuel economy and reducing pollutant emissions, hybrid electric vehicles (HEVs) have been being actively investigated over the world [1], [2]. Energy management strategy is an enabling technology in HEVs, in order to distribute power among multiple power sources for improving overall energy efficiency [3], [4]. One major difficulty to realize this goal lies in future driving condition prediction. Hence, an online, efficient predictive energy management strategy involving a preview of vehicular power demand is significant.

In order to attain desirable power split in HEVs, energy management strategies are often optimization-based techniques. These approaches are classified into global optimization and real-time optimization categories. As the message of driving cycle is previously given, dynamic programming (DP) could obtain theoretically global optimal control. For example, a novel efficient neural network module structure is compared with DP-based controls to declare its optimality in [5]. However, its online effectiveness cannot be guaranteed. Energy management strategy for a range extended electric vehicle was investigated by Chen [6] using DP, in which the driver comfort, battery life, and limitation of noise are considered in the cost function. Then a rule-based, multimode switch strategy that requires lower computation efforts was proposed.

However, the DP-based control strategy is typically inappropriate for real-world application because road topography is ususally unknown in advance [7]. As an alternative, convex programming (CP) was also adopted to acquire a global optimal solution via convex modeling and rapid solution search. Hu *et al.* [8] applied CP to a comparison framework of hybrid powertrains with three diffrernt energy storage systems, which allows hybrid powertrain designer to rapidly and optimally perform integrated component selection, sizing, and energy management.

Equivalent consumption minimization strategy (ECMS) [9], [10] and model predictive control (MPC) [11], [12] are two representative techniques in real-time optimization. ECMS focuses on the local optimization by exploring the accurate co-state value. Musardo *et al.* [13] presented an adaptive ECMS strategy to periodically refresh the co-state according to the current road load. Thus fuel consumption is minimized, while battery state of charge (SoC) is maintained within boundaries. Nevertheless, future driving condition is still not taken into account.

MPC benefits from the future driving information and could derive an energy management strategy through DP [14], quadratic programming [15], or nonlinear programming [16]. Markov chain (MC) [17] models and artificial neural networks (NNs) [18] were often utilized to forecast future driving condition information in MPC. Zeng *et al*. proposed a stochastic MPC-based energy management strategy using vehicle location, traveling direction, and terrain information for HEVs running in hilly regions with light traffic [19]. However, the performance of MPC control is highly dependent on the precision of future driving condition prediction [20].

The above model-based techiniques all require elaborate vehicle models [21]. This causes considerable expenditure of model parameter calibration [22]. To remedy this deficiency, reinforcement learning (RL) [23], [24] has been recently considered

Manuscript received September 1, 2017; revised December 31, 2017; accepted January 19, 2018. Date of publication January 24, 2018; date of current version April 3, 2018. This work was supported in part by the EU-funded Marie Sklodowska-Curie Individual Fellowships (IF) Project under Grant 706253-pPHEVH2020- MSCA-IF-2015. Paper no. TII-17-2035. (Teng Liu and Xiaosong Hu contributed equally to this work.) *(Corresponding authors: Teng Liu; Xiaosong Hu.)*

T. Liu is with Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada (e-mail: tengliu17@gmail.com).

X. Hu is with the Department of Automotive Engineering and the State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400044, China (e-mail: xiaosonghu@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2018.2797322

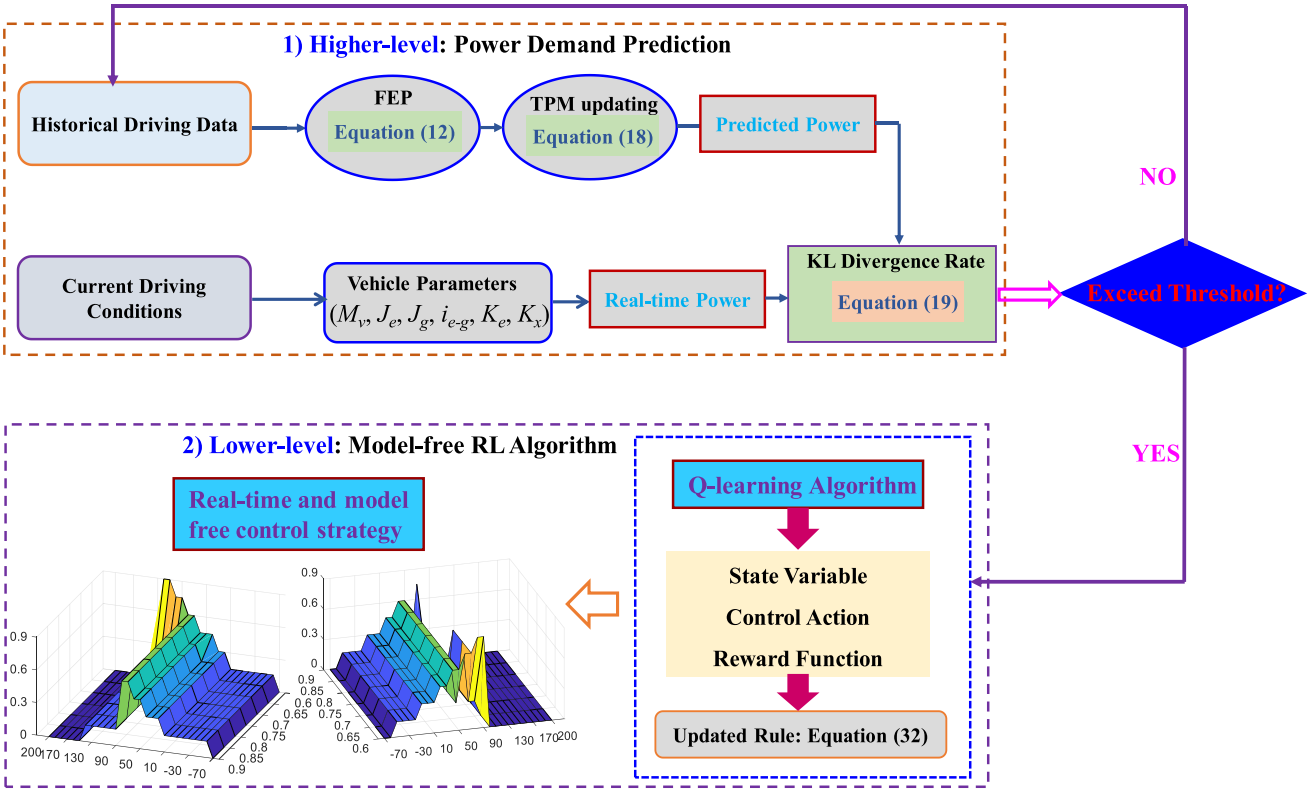


Fig. 1. Bi-level control framework for real-time and model-free energy management.

as a model-free method to search optimal control in energy management problem. Liu *et al.* [25], [32], [33] proposed a RL-based energy management strategy using Q-learning algorithm and MC models. The results indicated that fuel efficiency could be significantly improved by using the proposed RL-based energy management strategy. Nevertheless, it should be pointed out that the fuel economy of HEVs may be even degraded, if the associated control strategy is unsuitable for future driving conditions [26]. Hence, future driving information needs to be carefully considered during derivation of an energy management strategy for HEVs.

Recent studies reveal that predictive learning (PL) could be a useful tool for driving condition prediction for HEVs. Based on this concept, by sufficiently combining the predictive method in [3] and online updating technique in [25], this work proposes a novel PL and RL-based predictive, real-time and model-free control framework to improve energy efficiency for a hybrid tracked vehicle (HTV). PL makes the energy controls adapt to mutative future driving conditions, and RL enables the controls to be real-time implementable.

The main contribution of this paper is to present a bi-level control framework to formulate a predictive, real-time energy management strategy, which has not been discussed in our previous work (see Fig. 1 as an illustration). The higher-level discusses how to predict the power demand based on the MC. Specially, fuzzy encoding predictor (FEP) is employed to forecast power demand, and a real-time recursive algorithm is applied to fuse the future power demand information into transition probability matrix (TPM) computation. Furthermore, the Kullback–Leibler (KL) divergence rate is employed to decide the alteration of control strategy. The lower-level calculates the relevant energy management strategy, based on the updated TPM, using RL.

Through comparing with DP algorithm, the optimality of the proposed control is evaluated, and the influence of KL divergence rate is demonstrated by a comparison of a common RL method, in terms of energy efficiency. Simulation results underline that the proposed strategy leads to noticeable improved fuel economy and computational speed. These merits make it feasible for online application.

This rest paper is organized as follows: In Section II, the higher-level power demand prediction and real-time recursive algorithm are introduced. Section III describes the lower-level RL control of the HTV powertrain. In Section IV, tests are designed to evaluate the proposed approach, and simulation results are analyzed. Finally, conclusions and future work are described in Section V.

II. Higher-Level: Power Demand Prediction and Online Updating

The predicted approach and online updating expression for power demand are from our previous studies [3] and [25], and elucidated in this section for mathematical completeness. First, MC-based FEP for power demand prediction is introduced [3]. Then, we use a real-time recursive algorithm to fuse the future power demand information into online calculation of TPMs [25]. Finally, KL divergence rate is applied to evaluate

the availability of power demand prediction by comparing differences of multiple TPMs [31].

A. Fuzzy Encoding Predictor

In this paper, the power demand is modeled as a finite-state MC [27] and denoted as $P = \{p_{\text{dem}}^j | j = 1, \ldots, M\} \subset X$, where $X \subset R$ is bounded. The maximum likelihood estimator is used to estimate the transition probability of power demand by [32]

$$\begin{cases} p_{ij} = P(p_{\text{dem}}^+ = p_{\text{dem}}^j \,\big|\, p_{\text{dem}} = p_{\text{dem}}^i) = \frac{N_{ij}}{N_i} \\ N_i = \sum_{j=1}^{M} N_{ij} \end{cases} \tag{1}$$

where p_{dem} and p_{dem}^+ are the present and next one step-ahead power demands, respectively, and p_{ij} is the transition probability from p_{dem}^i to p_{dem}^j. Furthermore, N_{ij} indicates the transition counts from p_{dem}^i to p_{dem}^j, and N_i is the total transition counts initiated from p_{dem}^i.

The TPM Π is filled with elements p_{ij}. The one step-ahead probability vector of power demand taking one of finite values p_{dem}^j is linked as

$$(p^+)^T = p^T \Pi \tag{2}$$

and for $n > 1$ steps ahead as

$$(p^{+n})^T = p^T \Pi^n. \tag{3}$$

In the fuzzy encoding technique, X is divided into a finite set of fuzzy subsets $\Phi_j, j = 1, \ldots, M$ and the fuzzy subset Φ_j is a pair $(X, \mu_j(\cdot))$, where $\mu_j(\cdot)$ is a Lebesgue measurable membership function that satisfies the property

$$\mu_j : X \to [0, 1] \text{ s.t.} \forall p_{\text{dem}} \in X, \exists j, 1 \le j \le M, \mu_j(p_{\text{dem}}) > 0 \tag{4}$$

where $\mu_j(p_{\text{dem}})$ reflects the degree of membership of $p_{\text{dem}} \in X$ in μ_j. A continuous state $p_{\text{dem}} \in X$ in the fuzzy encoding may be associated with several states p_{dem}^j of the underlying finite-state MC model [28].

The FEP involves two transformations based on the theory of approximate reasoning [29]. The first transformation allocates an M-dimensional possibility (not probability) vector for each $p_{\text{dem}} \in X$ as follows:

$$\tilde{O}^T(p_{\text{dem}}) = \mu^T(p_{\text{dem}}) = [\mu_1(p_{\text{dem}}), \mu_2(p_{\text{dem}}), \ldots, \mu_M(p_{\text{dem}})]. \tag{5}$$

Notice that the sum of the elements in the possibility vector $\sim |!O(p_{\text{dem}})$ is unnecessary to equal 1. This transformation is named fuzzification and maps power demand in the space X to vector in M-dimensional possibility vector space $\tilde{X}$.

The second transformation is called the proportional possibility-to-probability transformation that converts the possibility vector $\sim O(p_{\text{dem}})$ to a probability vector $O(p_{\text{dem}})$ by normalization

$$O(p_{\text{dem}}) = \tilde{O}(p_{\text{dem}}) \Big/ \sum_{j=1}^{M} \tilde{O}_j(p_{\text{dem}}) \tag{6}$$

where this transformation maps $\tilde{X}$ to an M-dimensional probability vector space, $\bar{X}$. The probability distribution of the next state in $\bar{X}$ is computed by

$$(O^+(p_{\text{dem}}))^T = (O(p_{\text{dem}}))^T \Pi \tag{7}$$

where the element p_{ij} in the TPM Π is interpreted as a transition probability between Φ_i and Φ_j. To decode vectors in $\bar{X}$ back to X, the probability distribution $O^+(p_{\text{dem}})$ is utilized to aggregate the membership function $\mu(p_{\text{dem}})$ to encode the probability vector of the next state in X [27]

$$w^+(p_{\text{dem}}) = (O^+(p_{\text{dem}}))^T \mu(p_{\text{dem}}) = (O(p_{\text{dem}}))^T \Pi \mu(p_{\text{dem}}). \tag{8}$$

The expected value over the possibility vector leads to the next one-step ahead power demand using FEP

$$\begin{cases} p_{\text{dem}}^+ = \int_X w^+(y) y dy \Big/ \int_X w^+(y) dy \\ \int_X w^+(y) y dy = \sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij} \int_X y \mu_j(y) dy \\ \int_X w^+(y) dy = \sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij} \int_X \mu_j(y) dy. \end{cases} \tag{9}$$

Note that the centroid and volume of the membership function $\mu_j(p_{\text{dem}})$ is expressed as

$$\begin{cases} \bar{c}_i = \int_X y \mu_j(y) dy \\ V_j = \int_X \mu_j(y) dy. \end{cases} \tag{10}$$

Thus, the next one-step ahead power demand in expression (9) is rewritten as

$$\begin{cases} p_{\text{dem}}^+ = \dfrac{\sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij} V_j \bar{c}_j}{\sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij} V_j} \end{cases}. \tag{11}$$

Assuming that the membership functions have the same volume and using the fact $\sum_{j=1}^{M} p_{ij} = 1$ and $\sum_{i=1}^{M} O_i(p_{\text{dem}}) = 1$, (11) is further simplified as [3]

$$\begin{cases} p_{\text{dem}}^+ = \dfrac{\sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij} c_j}{\sum_{i=1}^{M} O_i(p_{\text{dem}}) \sum_{j=1}^{M} p_{ij}} = (O(p_{\text{dem}}))^T \Pi \bar{c} \end{cases} \tag{12}$$

where (12) is the next one-step ahead power demand using FEP. As can be seen, the probability distribution and centroid in (12) are related to the membership functions. In this paper, these functions are taken as Gaussian membership functions [30] with a standard deviation $\sigma = 1$ as follows:

$$q_i = e^{\frac{-(x - 2.5i + 1.25)^2}{2 \cdot \sigma^2}}, \quad i = 1, \ldots M. \tag{13}$$

B. TPM Online Updating

After predicting the future power demand, an online recursive algorithm is applied to integrate this information into TPM computation [25]. Assuming the length of the predicted power demand is K, it is convenient to modify expression (1) for on-board application as below:

$$p_{ij} = \frac{N_{ij}(K)}{N_i(K)} = \frac{N_{ij}(K)/K}{N_i(K)/K} = \frac{F_{ij}(K)}{F_i(K)} \tag{14}$$

where $F_i(K)$ is the total frequency rate of the transition events $f_i(K)$ initiated from $p^i_{\rm dem}$, and $F_{ij}(K)$ is the frequency rate of transition events $f_{ij}(K)$ from $p^i_{\rm dem}$ to $p^j_{\rm dem}$ within a specific window with K measurements [25]

$$\begin{cases} F_{ij}(K) = N_{ij}(K)/K = \frac{1}{K}\sum_{t=1}^{K} f_{ij}(t) \\ F_i(K) = N_i(K)/K = \frac{1}{K}\sum_{t=1}^{K} f_i(t) = \frac{1}{K}\sum_{t=1}^{K}\sum_{j=1}^{M} f_{ij}(t). \end{cases} \tag{15}$$

where $f_{ij}(t) = 1$, if a transition from $p^i_{\rm dem}$ to $p^j_{\rm dem}$ occurs at time instant t; $f_i(t) = 1$ if a transition initiated from the state $p^i_{\rm dem}$ at time instant t; otherwise, they take values to be zeros. The frequency rate can be iteratively deduced as

$$\begin{aligned} F_{ij}(K) &= \frac{1}{K}\sum_{t=1}^{K} f_{ij}(t) = \frac{1}{K}[(K-1)F_{ij}(K-1) + f_{ij}(K)] \\ &= F_{ij}(K-1) + \frac{1}{K}[f_{ij}(K) - F_{ij}(K-1)] \\ &= F_{ij}(K-1) + \varphi[f_{ij}(K) - F_{ij}(K-1)] \end{aligned} \tag{16}$$

$$\begin{aligned} F_i(K) &= \frac{1}{K}\sum_{t=1}^{K} f_i(t) = \frac{1}{K}[(K-1)F_i(K-1) + f_i(K)] \\ &= F_i(K-1) + \frac{1}{K}[f_i(K) - F_i(K-1)] \\ &= F_i(K-1) + \varphi[f_i(K) - F_i(K-1)] \end{aligned} \tag{17}$$

where $\varphi \in (0,1)$ is called forgetting factor, which is used for weighting the old power demand data with exponentially decreasing weights for online application. Finally, a recursive expression of the transition probability of power demand is formulated by integrating (14)–(17) as follows [25]:

$$p_{ij} = \frac{F_{ij}(K)}{F_i(K)} = \frac{F_{ij}(K-1) + \varphi[f_{ik,j}(K) - F_{ij}(K-1)]}{F_i(K-1) + \varphi[f_i(L) - F_i(K-1)]}. \tag{18}$$

C. KL Divergence Rate

To avoid the excessive updating of the TPM and improve the computational efficiency of the online recursive algorithm, a quantified parameter named KL divergence rate is proposed to measure the differences of multiple TPMs. The KL divergence rate is defined as [31]

$$D_{\rm KL}(P_1 \| P_2) = \sum_{x}\sum_{x^+}[P_1(x^+|x)P^*(x)]\log\left[\frac{P_1(x^+|x)}{P_2(x^+|x)}\right] \tag{19}$$

where P_1 and P_2 are two $M \times M$ TPMs of power demand, x and $x^+ \in [1, M]$ are the current and next indices of transition probability, respectively. P^* is the steady-state probability distribution of P, which can be represented by

$$P^* P_1 = P^*. \tag{20}$$

Obviously, P^* is an eigenvector of P_1 whose eigenvalue corresponds to 1. The logarithm operator in (19) requires the elements in P_1 and P_2 to be greater than zero and thus these two

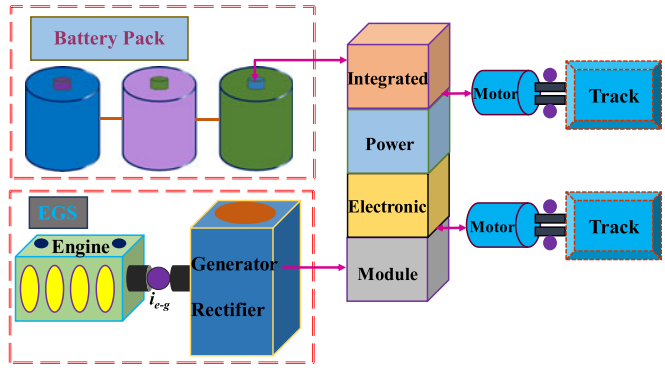


Fig. 2. Configuration of the series HTV powertrain [25].

TABLE I
ELEMENTARY PARAMETERS OF THE HTV POWERTRAIN [33]

Name	Value	Unit
Vehicle mass M_v	2500	kg
Generator inertia J_g	0.1	kg·m^2
Engine inertia J_e	0.2	kg·m^2
Gear ratio parameter i_{e-g}	1	/
Electromotive force parameter K_e	0.8092	Vsrad^{-2}
Electromotive force parameter K_x	0.0005295	NmA^{-2}
Minimum State of Charge $SoC_{\min}$	0.5	/
Maximum Sate of Charge $SoC_{\max}$	0.9	/
Battery capacity $C_{\rm bat}$	37.5	Ah

matrices are replaced by two equivalent one [31]

$$\begin{cases} P_1^{\rm reg} = (1-\delta)P_1 + \delta\frac{I}{M} \\ P_2^{\rm reg} = (1-\delta)P_2 + \delta\frac{I}{M} \end{cases} \tag{21}$$

where δ is a small constant ranging from 0 to 1, and I is the identity matrix with the same dimensions as the matrix P_1. Three characteristics are illuminated for the KL divergence rate: first, it is always nonnegative, $D_{\rm KL}(P_1 \| P_2) = 0$, if and only if $P_1 = P_2$; second, in general, it is a nonsymmetric measure, $D_{\rm KL}(P_1 \| P_2) \neq D_{\rm KL}(P_2 \| P_1)$; third, the closer it is to zero, the more similar P_1 is to P_2 [31].

III. Lower-Level: HTV Powertrain and RL

To improve the energy efficiency of a series HTV, an optimal control problem is formulated in this section. The powertrain configuration of the HTV is shown in Fig. 2 [25]. The vehicle powertrain mainly includes an engine generator set (EGS), a battery pack and two driving motors. The modeling of the EGS, battery pack and cost function are first introduced. Furthermore, the RL technique framework is constructed [32], and the Q-learning algorithm is harnessed to rapidly search the optimal control based on the online TPM of predicted power demand. The elementary parameters of the HTV powertrain are listed in Table I [33].

A. Vehicle Powertrain Model

In EGS, the engine rated power is 52 kW at the speed of 6200 r/min. The rated output power of the generator is 40 kW within the speed range from 3000 to 3500 r/min. The generator

speed is selected as the first state variable and can be calculated according to the torque equilibrium constraint

$$\begin{cases} \frac{dn_g}{dt} = \left(\frac{T_e}{i_{e-g}} - T_g\right)/0.1047\left(\frac{J_e}{i^2_{e-g}} + J_g\right) \\ n_e = n_g/i_{e-g} \end{cases} \tag{22}$$

where n_g and n_e are the rotational speeds, T_g and T_e are the torques of the generator and engine, respectively, and T_e is the sole control action in this work. J_e and J_g are the rotational moment of inertias of the engine and generator, respectively. i_{e-g} is the gear ratio between the engine and generator, and 0.1047 is the transformation factor that denotes 1 r/min = 0.1047 rad/s. The torque and output voltage of the generator can be derived as follows [33]:

$$\begin{cases} T_g = K_e I_g - K_x I_g^2 \\ U_g = K_e n_g - K_x n_g I_g \end{cases} \tag{23}$$

where K_e is the electromotive force coefficient, U_g and I_g are the generator voltage and current, respectively. Furthermore, $K_x n_g$ is the electromotive force, and $K_x = 3PL^g/\pi$, in which L^g is the armature synchronous inductance, and P is the poles number.

For the HTV, the SoC of the battery is chosen as another state variable, which is computed by

$$\frac{dSoC}{dt} = -\frac{I_{\text{bat}}(t)}{C_{\text{bat}}} \tag{24}$$

where I_{bat} and C_{bat} denote the current and rated capacity of battery, respectively. According to the internal resistance model [34], the derivative of SoC and battery output voltage can be computed by

$$\begin{cases} \frac{dSoC}{dt} = \frac{(V_{oc} - \sqrt{V_{oc}^2 - 4r_{ch}(r_{\text{dis}})P_{\text{bat}}(t))}}{2C_{\text{bat}}r_{ch}(r_{\text{dis}})} \\ U_{\text{bat}} = \begin{cases} V_{oc} - I_{\text{bat}}r_{ch}(SoC)(I_{\text{bat}} > 0) \\ V_{oc} - I_{\text{bat}}r_{dis}(SoC)(I_{bat} < 0) \end{cases} \end{cases} \tag{25}$$

where V_{oc} is the open circuit voltage and P_{bat} is the battery power. Furthermore, U_{bat} is the battery output voltage and $r_{\text{dis}}(SoC)$ and $r_{ch}(SoC)$ depict the internal resistances during discharging and charging, respectively.

The optimal control objective to be minimized is expressed as a tradeoff between the fuel consumption and charge sustenance as follows:

$$\begin{cases} J = \int_{t_0}^{t_f} [\dot{m}_f(t) + \alpha(\Delta_{SoC})^2]dt \\ \Delta_{SoC} = \begin{cases} SoC(t) - SoC_{\text{ref}} & SoC(t) < SoC_{\text{ref}} \\ 0 & SoC(t) \geq SoC_{\text{ref}} \end{cases} \end{cases} \tag{26}$$

where $[t_0, t_f]$ is the specific time interval, $\dot{m}_f$ is the fuel consumption rate, and α is a positive weight coefficient. In addition, SoC_{ref} is a preallocated constant to maintain charge-sustaining constraints [35]. To obtain the real-time energy management strategy using the future power demand information and RL,

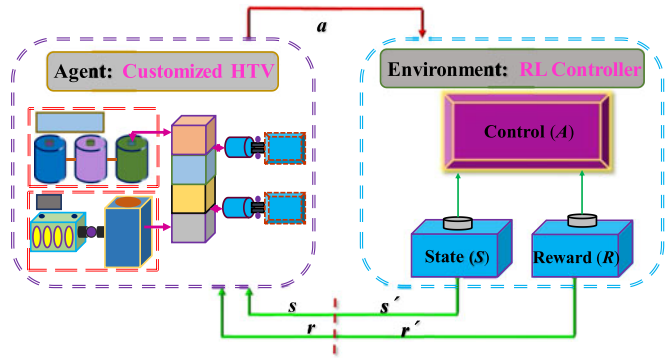


Fig. 3. RL interaction between environment and agent [25].

the following inequality constraints should be observed:

$$\begin{cases} SoC_{\min} \leq SoC(t) \leq SoC_{\max} \\ n_{g,\min} \leq n_g(t) \leq n_{g,\max} \\ T_{e,\min} \leq T_e(t) \leq T_{e,\max} \\ n_{e,\min} \leq n_e(t) \leq n_{e,\max} \\ I_{\text{bat},\min} \leq I_{bat} \leq I_{\text{bat},\max} \\ 0 \leq I_g \leq I_{g,\max}. \end{cases} \tag{27}$$

Since the core of this article focuses on discussing the control performance of the proposed real-time control strategy, both traction motors, as power conversion devices, are assumed to have an identical efficiency, and the battery aging is neglected in this study [32], [33].

B. RL Based Powertrain Control

The RL interaction between the environment and the agent is shown in Fig. 3 [25]. This interaction is modeled as a discrete discounted Markov decision process (MDP) that is a quintuple $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{\Pi}, \boldsymbol{R}, \beta)$, where $\boldsymbol{S}$ and $\boldsymbol{A}$ are the sets of state variables and control actions, $\boldsymbol{\Pi}$ is the TPM, $\boldsymbol{R}$ is the reward function, and $\beta \in (0, 1)$ is a discount factor.

Specially, the optimal control problem in this paper incorporates a set of state variables $s \in \boldsymbol{S} = \{(n_g(t), SoC(t))|1200 \leq n_{\text{g}}(t) \leq 3100, 0.5 \leq SoC(t) \leq 0.9\}$, a set of actions $a \in \boldsymbol{A} = \{T_e(t)|0 \leq T_e\,(t) \leq 92\}$, and a reward function $r \in \boldsymbol{R} = \{\dot{m}_f(s, a)\}$.

The control policy ψ is the distribution over the control actions a, given the current state s. The optimal value function is represented as the finite expected discounted sum of the rewards [36]

$$V^*(s) = \min_{\psi} E\left(\sum_{t=t_0}^{t_f} \beta^t r(s, a)\right). \tag{28}$$

Due to the uniqueness, (28) can be reformulated as a recursion expression

$$V^*(s) = \min_{a}\left(r(s, a) + \beta \sum_{s' \in S} \pi_{s'a,s} V^*(s')\right) \quad \forall s \in S \tag{29}$$

TABLE II
PSEUDOCODE OF THE Q-LEARNING ALGORITHM [3]

Algorithm: Q-Learning Algorithm.

1. Initialize $Q(s,a)$, s, and number of iteration N_k
2. Repeat each step $k = 1, 2, 3 \ldots$
3. Choose a, based on $Q(s,.)$ (ε-greedy policy)
4. Taking action a, observe r, s'
5. Define $a^* = \arg\max_a Q(s', a)$
6. $Q(s,a) \leftarrow Q(s,a) + \eta(r(s,a) + \beta \max_{a'} Q(s', a') - Q(s,a))$
7. $s \leftarrow s'$
8. until s is terminal

where $\pi_{s'a,s}$ denotes the transition probability from state s to state s' using action a. Given the optimal value function, the optimal control policy is determined as follows:

$$\psi^*(s) = \arg\min_a \left(r(s,a) + \beta \sum_{s' \in S} \pi_{s'a,s} V^*(s') \right). \quad (30)$$

In addition, the action value function $Q(s,a)$ and its optimal value $Q^*(s,a)$ are expressed as the following formula [32]:

$$\begin{cases} Q(s,a) = r(s,a) + \beta \sum_{s' \in S} \pi_{s'a,s} Q(s', a') \\ Q^*(s,a) = r(s,a) + \beta \sum_{s' \in S} \pi_{s'a,s} \min_{a'} Q(s', a'). \end{cases} \quad (31)$$

The variable $V^*(s)$ is the value of s, assuming that an optimal action is taken initially; therefore, $V^*(s) = Q^*(s,a)$ and $\psi^*(s) = \arg\min_a Q^*(s,a)$. The updated rule of action-value function in Q-learning algorithm is expressed as [36]

$$Q(s,a) \leftarrow Q(s,a) + \eta(r(s,a) + \beta \min_{a'} Q(s', a') - Q(s,a)) \quad (32)$$

where $\eta \in [0, 1]$ is a decaying factor in the Q-learning algorithm. The pseudocode of the Q-learning algorithm is described in Table II [3].

Fig. 1 shows the calculated flowchart of the proposed predictive real-time energy management strategy based on the updated TPM using RL approach. FEP is used to first forecast the future power demand. The TPM online updating algorithm integrates the future power demand information into online TPM computation. Then the KL divergence rate is applied to measure the differences between the current and future TPMs. As the KL divergence rate is larger than the pre-assigned threshold value, the control action computation is triggered, and the relevant energy management strategy is updated online. Otherwise, the current strategy maintains.

The state variables and control action are discretized as $n_g \in [1200 : 95 : 3100], SoC \in [0.5 : 0.02 : 0.9], T_e \in [0 : 4.6 : 92]$. The RL course is carried out in MATLAB using the MDP toolbox described in [37]. The decaying factor η is correlated with the time step k and taken as $1/\sqrt{k+2}$, the discount factor β is taken as 0.95, the number of iteration N_k is 10 000, and the sample time is 1 s. The optimality and adaptability will be discussed in Section IV.

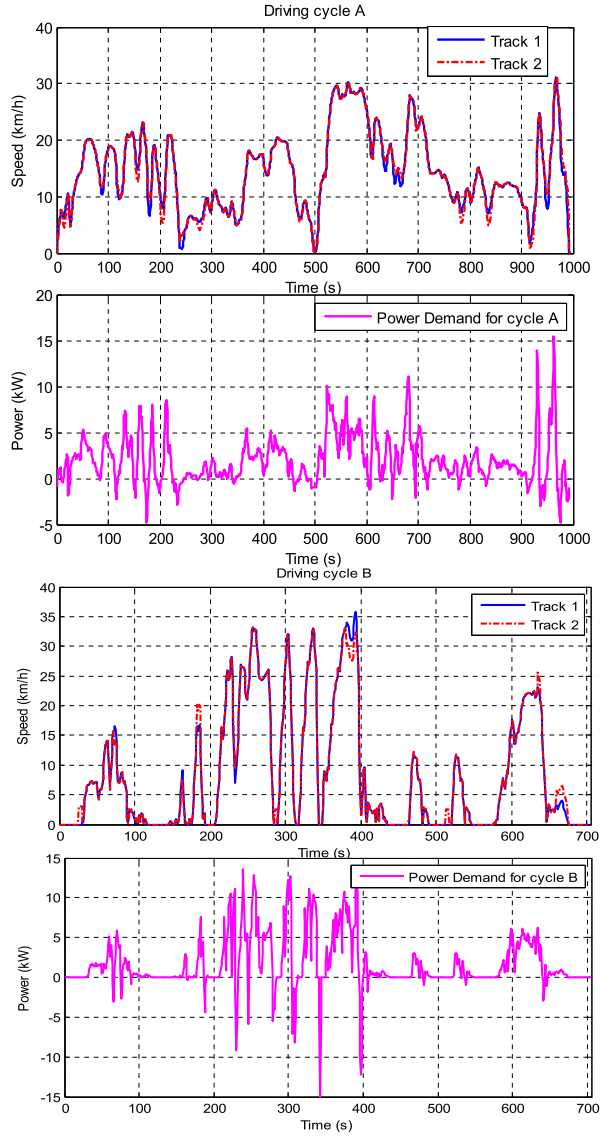


Fig. 4. Two driving cycles for power demand prediction [25].

IV. RESULTS AND DISCUSSION

The proposed PL and RL-enabled predictive real-time energy management strategy is evaluated in this section. The influences of KL divergence rate on fuel consumption are examined by comparing the control performance in different threshold value cases. Then, numerical tests illuminate that the energy efficiency improvement in our control strategy can exceeds the common RL method by tuning the KL divergence rate value.

A. Influence of KL Divergence Rate

In order to evaluate the performance of the FEP for power demand prediction, the proposed control strategy is compared with the common RL control strategy, in which the one-step ahead predicted power demands under the two representative driving cycles are considered (see Fig. 4 as an illustration) [25].

Fig. 5 illustrates the one-step ahead and 10-steps ahead power demand prediction trajectories for these two driving cycles. In

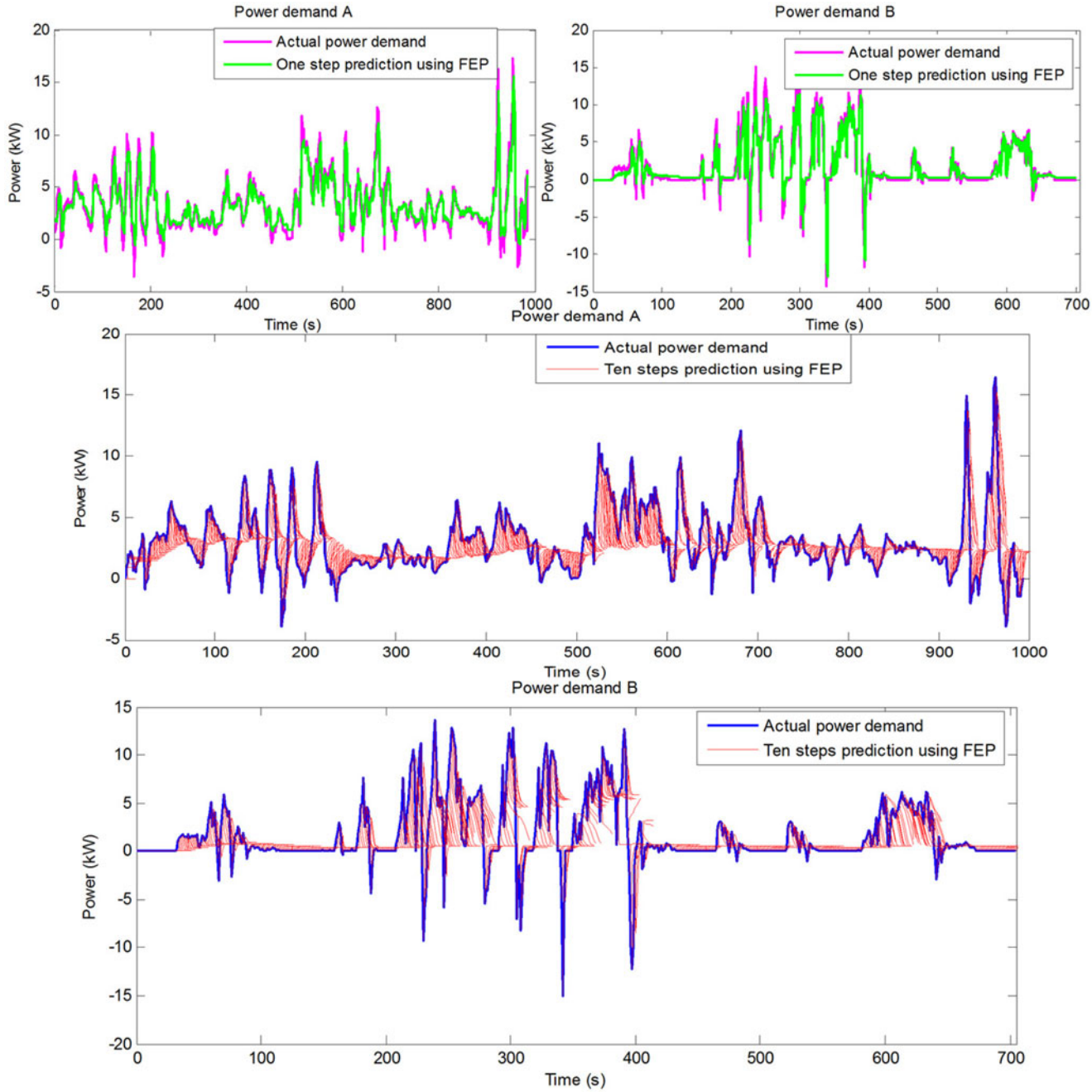


Fig. 5. One-step and 10-steps ahead power demand prediction for two driving cycles.

the proposed real-time control strategy, the online recursive algorithm and the KL divergence rate are implemented to integrate the predicted power demand information into the TPM computation. The KL divergence rate threshold values are defined as 0.3 and 0.5, and the forgetting factor $\varphi = 0.01$.

The SoC evolutions in different control cases for the two driving cycles are shown in Fig. 6. It can be discerned that the SoC trajectories in the three control cases are completely different for driving cycle A. However, for driving cycle B, the SoC trajectories in the proposed real-time predictive RL control are essentially the same and clearly differ from that of the common RL control. We attribute these differences to the alternation of the TPM of power demand, as shown in Fig. 7, in which the KL divergence rate values are calculated between two TPMs in every 100 s at different speeds. This alternation of TPM results in the updating of control strategy and the improvement of energy efficiency.

Since the KL divergence rate threshold values are defined as 0.3 and 0.5, the updating times of the TPM and control strategy are determined. Table III depicts the updating times and fuel consumption after SoC-correction [38] for different control strategies in the two driving cycles. The updating times for different KL divergence rate values are uniform in driving cycle B, which leads to the same SoC trajectories and fuel consumption

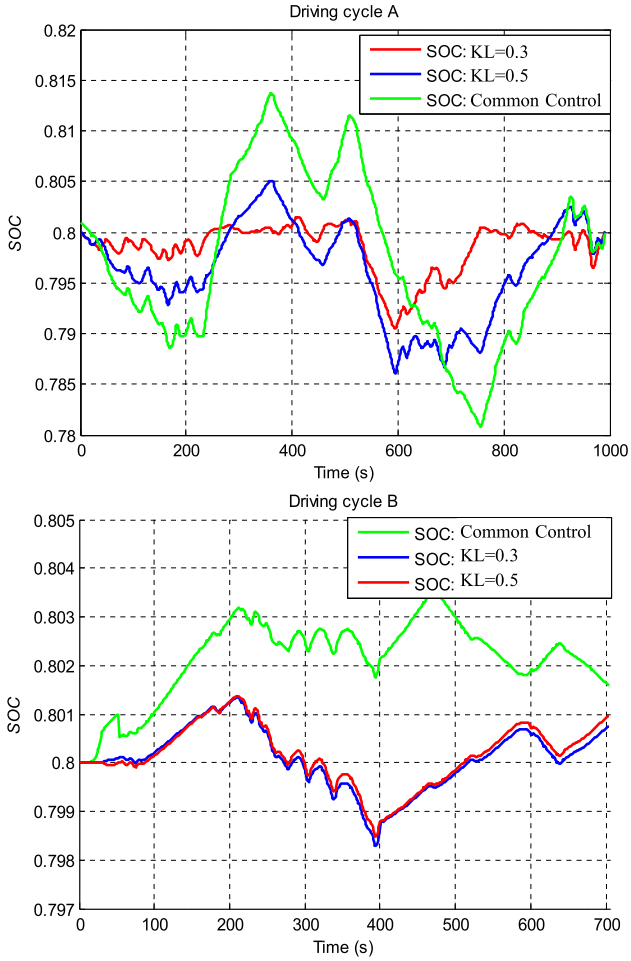


Fig. 6. SoC trajectories with different control strategies.

in the proposed predictive real-time control. Oppositely, they are diverse in driving cycle A. The proposed predictive real-time control strategy thus has better control performance than the common RL control strategy. Finally, 0.4 is chosen as the KL divergence rate threshold value when considering the performance and computation efficiency for comparing different control strategies in the next section.

B. Comparison of Different Control Strategies

To verify the optimality of the proposed predictive real-time control strategy, the common RL-based and DP-based strategies act as benchmark strategies for comparison purposes in this section. The simulation driving cycle is shown in Fig. 8, and the parameters setting for the proposed control strategy is depicted in Table IV.

Fig. 9 illustrates the SoC trajectories and power split results between the engine and the battery under the simulation cycle. As can be seen, the SoC trajectory in the proposed control straetgy is much closer to that of DP-based control strategy than the common RL control. An analogous result appears in the power split curves. This improvement can be ascribed to the online updating of control straetgy, as the predicted power demand is injected into the calculation of TPM.

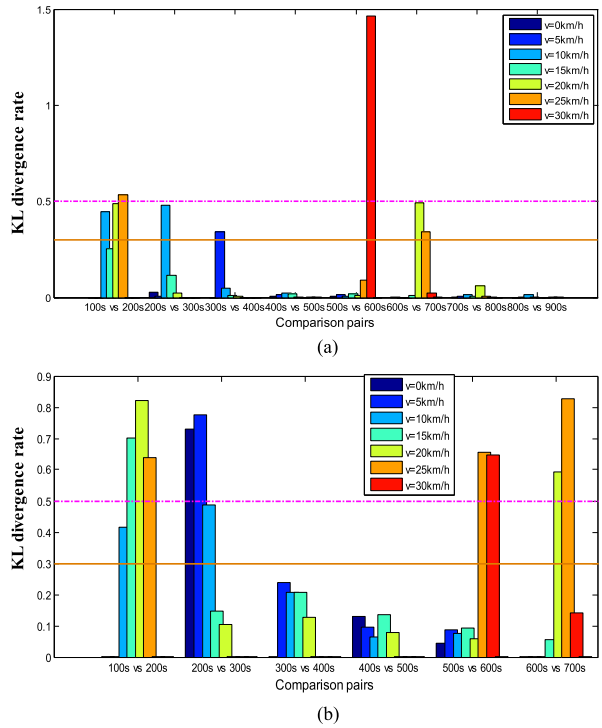


Fig. 7. KL divergence rate values for different driving cycles: (a) Driving cycle A and (b) driving cycle B.

TABLE III
UPDATING TIMES AND FUEL CONSUMPTION FOR DIFFERENT CONTROL STRATEGIES

Control strategies	Updated times^{A} (count)	Fuel (g)^{A}	Updated times^{B} (count)	Fuel (g)^{B}
Common control	0	476.3	0	361.0
Predictive real-time RL: KL = 0.3	5	429.2	4	317.4
Predictive real-time RL: KL = 0.5	2	453.9	4	317.4

^{A} denotes driving cycle A; ^{B} denotes driving cycle B.

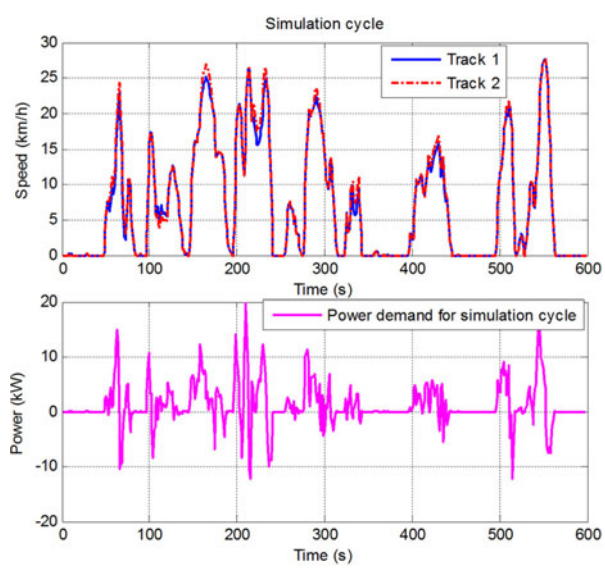


Fig. 8. Simulation drive cycle for control strategies comparison.

TABLE IV
PARAMETERS SETTING FOR THE PREDICTIVE REAL-TIME CONTROL

Parameters	Prediction step length	KL threshold value	Forgetting factor φ	Sample time
Value	5-step	0.4	0.01	1 s

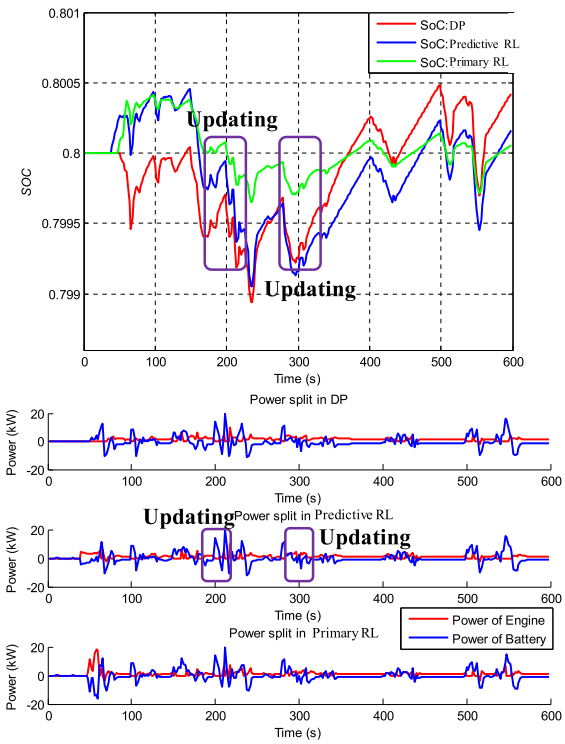


Fig. 9. SoC trajectories and power split with different control strategies under the simulation cycle.

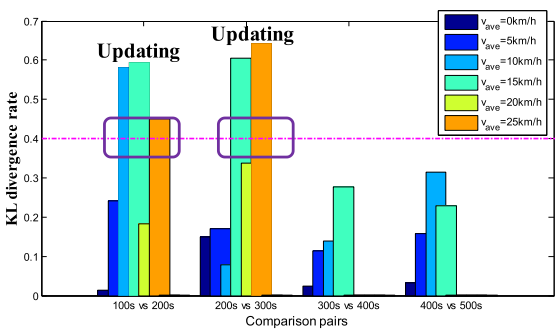


Fig. 10. KL divergence rate values with the simulation cycle.

Since the KL value surpasses the threshold value 0.4, the proposed control strategy is triggered to update at 200 and 300 s, as described in Fig. 10. The simulation results demonstrate that the MC-based predicted power demand is renovated effectively to make the predictive real-time and model-free control strategy similar to the DP-based control strategy.

Table V illustrates the fuel consumption results after SoC-correction for the three control strategies. It can be recognized that the fuel consumption of the proposed predictive real-time control strategy is lower than that of the common RL control by

TABLE V
FUEL CONSUMPTION COMPARISONAFTER SOC-CORRECTION

Control strategies	Fuel consumption (g)	Relative increase (%)
DP	260.3	–
Predictive real-time RL	266.8	2.5
Common RL	277.5	6.61

TABLE VI
COMPUTATION TIME IN DIFFERENT CONTROL STRATEGIES

Control strategies	Computational time[a] (h)	Relative increase (%)
Predictive real-time RL	1.42	–
DP	2.58	81.69
Common RL	4.65	227.46

[a]A 2.4 GHz microprocessor with 12 GB RAM was used.

4.11%, and is close to that of DP-based control. This demonstrates its optimality. The computational times of these control strategies are contrasted in Table VI. Note that the proposed solution is fastest among the three types of control strategies, which makes it online application easier.

V. CONCLUSION

In this paper, we seek energy efficiency improvement of a hybrid vehicle by synergizing PL with RL. Based on the study of FEP and realization of TPM online updating, the future power demand can be predicted and fused into real-time control strategy computation. With the acquaintance of predicted driving conditions, Q-learning algorithm is picked to derive rapidly the model-free energy management strategy.

Tests prove the optimality and availability of the proposed predictive real-time energy management strategy. In addition, the advantages in energy efficiency improvement and computational speed imply that the proposed real-time and model-free control can be applied in real-time situations.

To the best of our knowledge, this is the first attempt to synergistically leverage PL trick and reinforcement learning in HTV energy management field. It brings a new thought to design model-free control in real-time so as to improve energy efficiency online. In the future, more simulation and experimental investigations are underway to verify the proposed control strategy in different real-word working and driving conditions.

REFERENCES

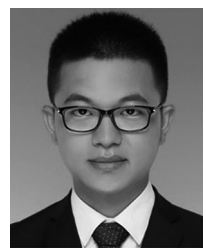
[1] J. Kang, R. Yu, X. Huang, S. Maharjan, and Y. Zhang, "Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains," *IEEE Trans Ind. Informat.*, 2017.

[2] Q. Zhang, W. Deng, and G. Li, "Stochastic control of predictive power management for battery/supercapacitor hybrid energy storage systems of electric vehicles," *IEEE Trans Ind. Informat.*, 2017.

[3] T. Liu, X. Hu, S. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1497–1507, Aug. 2017.

[4] Z. Zhou, J. Gong, Y. He, and Y. Zhang, "Software defined machine-to-machine communication for smart energy management," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 52–60, Oct. 2017.

[5] H. Tian, Z. Lu, X. Wang, X. L. Zhang, Y. Huang, and G. Y. Tian, "A length ratio based neural network energy management strategy for online control of plug-in hybrid electric city bus," *Appl. Energy*, vol. 177, pp. 71–81, 2016.
[6] B. C. Chen, Y. Y. Wu, and H. C. Tsai, "Design and analysis of power management strategy for range extended electric vehicle using dynamic programming," *Appl. Energy*, vol. 113, pp. 1764–1774, 2014.
[7] Y. Zou, T. Liu, F. Sun, and H. Peng, "Comparative study of dynamic programming and Pontryagin's minimum principle on energy management for a parallel hybrid electric vehicle," *Energies*, vol. 6, no. 4, pp. 2305–2318, 2013.
[8] X. S. Hu, N. Murgovski, L. M. Johannesson, and B. Egardt, "Comparison of three electrochemical energy buffers applied to a hybrid bus powertrain with simultaneous optimal sizing and energy management," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 3, pp. 1193–1205, Jun. 2014.
[9] P. Tulpule, V. Marano V, and G. Rizzoni, "Energy management for plug-in hybrid electric vehicles using equivalent consumption minimization strategy," *Int. J. Electr. Hybr. Veh.*, vol. 2, no. 4, pp. 329–350, 2010.
[10] J. Han, D. Kum, and Y. Park, "Synthesis of predictive equivalent consumption minimization strategy for hybrid electric vehicles based on closed-form solution of optimal equivalence factor," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5604–5616, Jul. 2017.
[11] S. Vazquez, J. I. Sergio, L. G. Franquelo, J. Rodriguez, and H. A. Young, "Model predictive control: A review of its applications in power electronics," *IEEE Trans Ind. Mag.*, vol. 8, no. 1, pp. 16–31, Mar. 2014.
[12] B. Stellato, T. Geyer, and P. Goulart, "High-speed finite control set model predictive control for power electronics," *IEEE Trans. Power Electron.*, vol. 32, no. 5, pp. 4007–4020, May 2017.
[13] C. Musardo, G. Rizzoni, Y. Guezennec, and B. Staccia, "A-ECMS: An adaptive algorithm for hybrid electric vehicle energy management," *Eur. J. Control*, vol. 11, no. 4, pp. 509–524, 2005.
[14] L. Johannesson, M. Asbogard, and B. Egardt, "Assessing the potential of predictive control for hybrid vehicle powertrains using stochastic dynamic programming," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 71–83, Mar. 2007.
[15] D. Rotenberg, A. Vahidi, and I. Kolmanovsky, "Ultracapacitor assisted powertrains: Modeling, control, sizing, and the impact on fuel economy," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 3, pp. 576–589, May 2011.
[16] H. Borhan H, A. Vahidi, A. M. Phillips, M. L. Kuang, and I. V. Kolmanovsky, "MPC-based energy management of a power-split hybrid electric vehicle," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 3, pp. 593–603, May 2012.
[17] W. Li, T. K. Lee, Z. S. Filipi, and X. Meng, "Development of electric machine duty cycles for parallel hybrid electric Beijing city bus based on Markov chain," *Int. J. Veh. Des.*, vol. 58, no. 2-4, pp. 348–366, 2012.
[18] C. Sun, X. Hu X, S. J. Moura, and F. C. Sun, "Velocity predictors for predictive energy management in hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1197–1204, May 2015.
[19] S. Onori, L. Serrao, and G. Rizzoni, "A parallel hybrid electric vehicle energy management strategy using stochastic model predictive control with road grade preview," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 6, pp. 2416–2423, Nov. 2015.
[20] L. Tribioli, R. Cozzolino, D. Chiappini, and P. Iora, "Energy management of a plug-in fuel cell/battery hybrid vehicle with on-board fuel processing," *Appl. Energy*, vol. 184, pp. 140–154, 2016.
[21] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, "Dependable demand response management in the smart grid: A stackelberg game approach," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 120–132, Mar. 2013.
[22] C. Dextreit and I. V. Kolmanovsky, "Game theory controller for hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 2, pp. 652–663, Mar. 2014.
[23] X. Qi, G. Wu, K. Boriboonsomsin, and M. J. Barth, "A novel blended real-time energy management strategy for plug-in hybrid electric vehicle commute trips," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, 2015, pp. 1002–1007.
[24] X. Qi, Y. Luo, G. Wu, K. Boriboonsomsin, and M. Barth, "Deep reinforcement learning-based vehicle energy efficiency autonomous learning system," in *Proc. IEEE Intell. Veh. Symp.*, 2017, pp. 1228–1233.
[25] Y. Zou, T. Liu, D. X. Liu, and F. C. Sun, "Reinforcement learning-based real-time energy management for a hybrid tracked vehicle," *Appl. Energy*, vol. 171, pp. 372–382, 2016.
[26] E. R. Stephens, D. B. Smith, and A. Mahanti, "Game theoretic model predictive control for distributed energy demand-side management," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1394–1402, May 2015.
[27] D. P. Filev and I. Kolmanovsky, "Generalized Markov models for real-time modeling of continuous systems," *IEEE Trans. Fuzzy. Syst.*, vol. 22, no. 4, pp. 983–998, Aug. 2014.
[28] D. P. Filev and I. Kolmanovsky, "Markov chain modeling approaches for on board applications," in *Proc. Amer. Control Conf.*, 2010, pp. 4139–4145.
[29] L. Johannesson, M. Asbogard, and B. Egardt, "Assessing the potential of predictive control for hybrid vehicle powertrains using stochastic dynamic programming," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 71–83, Mar. 2007.
[30] G. Grimmet and D. Stirzaker, *Probability and Random Processes*. London, U.K.: Oxford Univ. Press, 2004.
[31] Z. Rsched, F. Alajaji, and L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 917–921, May 2004.
[32] T. Liu, Y. Zou, D. Liu, and F. C. Sun, "Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7837–7846, Dec. 2015.
[33] T. Liu, Y. Zou, D. Liu, and F. C. Sun, "Reinforcement learning–based energy management strategy for a hybrid electric tracked vehicle," *Energies*, vol. 8, no. 7, pp. 7243–7260, 2015.
[34] X. Hu, S. Moura, N. Murgovski, B. Egardt, and D. Cao, "Integrated optimization of battery sizing, charging, and power management in plug-in hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 3, pp. 1036–1043, May 2015.
[35] L. Li, S. You, C. Yang, B. Yan, and J. Song, "Driving-behavior-aware stochastic model predictive control for plug-in hybrid electric buses," *Appl. Energy*, vol. 162, pp. 868–879, 2016.
[36] L. Kaelbing, M. Littman, and A. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.
[37] P. Shan, R. Li, S. Ning, and Q. Yang, "Markov decision process toolbox," in *Proc. IEEE Int. Workshop Open-Source Softw. Sci. Comput.*, 2009, pp. 123–128.
[38] C. Hou, M. G. Ouyang, L. F. Xu, and H. W. Wang, "Approximate Pontryagin's minimum principle applied to the energy management of plug-in hybrid electric vehicles," *Appl. Energy*, vol. 115, pp. 174–189, 2014.



Teng Liu (M'18) received the B.S. degree in mathematics from Beijing Institute of Technology, Beijing, China, 2011. He received the Ph.D. degree in automotive engineering from Beijing Institute of Technology (BIT), Beijing, in 2017.

His Ph.D. dissertation, under the supervision of Dr. Fengchun Sun, was entitled "Reinforcement learning-based energy management for hybrid electric vehicles." He is currently a Postdoctoral Fellow in the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada. He has more than 6 years research and working experience in renewable energy vehicles and autonomous vehicles. His current research focuses on parallel driving, parallel reinforcement learning, automated driving, and energy management of electrified vehicles. He has published more than 20 papers in these areas.

Xiaosong Hu (SM'16) received the Ph.D. degree in automotive engineering from Beijing Institute of Technology, China, in 2012.

He did scientific research and completed the Ph.D. dissertation in automotive research Center at the University of Michigan, Ann Arbor, USA, between 2010 and 2012. He is currently a Professor in the State Key Laboratory of Mechanical Transmissions and in the Department of Automotive Engineering, Chongqing University, Chongqing, China. He was a Postdoctoral Researcher in the Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA, between 2014 and 2015, as well as at the Swedish Hybrid Vehicle Center and the Department of Signals and Systems at Chalmers University of Technology, Gothenburg, Sweden, between 2012 and 2014. He was also a Visiting Postdoctoral Researcher in the Institute for Dynamic systems and Control, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, in 2014. His research interests include modeling and control of alternative powertrains and energy storage systems.

Dr. Hu has received several prestigious awards/honors, including the Emerging Sustainability Leaders Award in 2016, the EU Marie Currie Fellowship in 2015, the ASME DSCD Energy Systems Best Paper Award in 2015, and the Beijing Best Ph.D. Dissertation Award in 2013.