IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR ELECTRIC POWER SYSTEMS

Received March 30, 2020, accepted April 21, 2020, date of publication May 4, 2020, date of current version May 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992062

Online Data-Driven Energy Management of a Hybrid Electric Vehicle Using Model-Based Q-Learning

HEEYUN LEE[1], CHANGBEOM KANG[1], YEONG-IL PARK[2], NAMWOOK KIM[3], AND SUK WON CHA[1]

[1]Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826, South Korea
[2]Department of Mechanical System Design Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea
[3]Department of Mechanical Engineering, Hanyang University–Ansan, Ansan 15588, South Korea

Corresponding authors: Namwook Kim (nwkim21@gmail.com) and Suk Won Cha (swcha@snu.ac.kr)

This work was supported in part by the Technology Innovation Program (Development of RDE DB and Application Source Technology for Improvement of Real Road CO2 and Particulate Matter) funded by the Ministry of Trade, Industry and Energy (MOTIE, South Korea) under Grant 20002762, and in part by the MSIT(Ministry of Science, ICT), Korea, through the High-Potential Individuals Global Training Program supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) under Grant 2019-0-01566.

ABSTRACT The energy management strategy of a hybrid electric vehicle directly determines the fuel economy of the vehicle. As a supervisory control strategy to divide the required power into its multiple power sources, engines and batteries, many studies have been conducting using rule-based and optimization-based approaches for energy management strategy so far. Recently, studies using various machine learning techniques have been conducted. In this paper, a novel control framework implementing Model-based Q-learning is developed for the optimal control problem of hybrid electric vehicles. As an online energy management strategy, a new approach could learn the characteristics of a current given driving environment and adaptively change the control policy through learning. Especially, for the proposed algorithm, the internal powertrain environment and external driving environment are separated so they can be learned via the reinforcement learning framework, which results in a simpler and more intuitive control strategy that can be explained using the vehicle state approximation model. The proposed algorithm is tested and verified through simulations, and the simulation results present near optimal solution. The simulation results are compared with conventional rule-based strategies and optimal control solutions acquired from Dynamic Programming.

INDEX TERMS Hybrid electric vehicle, optimal control, power management, Q-learning, reinforcement learning.

I. INTRODUCTION

Energy management strategies for hybrid electric vehicles (HEVs) are one of the most important factors determining the fuel economy performance of a vehicle. Coordinating multiple power sources, generally fossil fuel energy and electric energy in HEVs, the energy management strategy is a supervisory control method to operate each power source by determining when and how much energy to use according to the driving environment [1].

Simple and applicable rule-based approaches are mainly used for the controllers of real vehicles, which usually focus on obtaining the best efficiency for each powertrain component as well as calibration of the control parameters are based on heuristics or engineer's intuition. Examples of rule-based control can be found in [2], [3]. More mathematical approaches have also been conducted based on optimal control theories. One of the most widely known algorithms is Dynamic Programming (DP) [4]. The dynamic programming approach is a powerful tool that shows the best available fuel economy of the vehicle. Therefore, the results of the DP simulation for HEV can be used to obtain an intuition for the control policy of powertrain [5], [6]. However, DP is not available for real-time control since it needs the entire driving speed profile before vehicle departure.

The associate editor coordinating the review of this manuscript and approving it for publication was Canbing Li.

At the same time, optimization-based control strategies for real-time application have been developed in various ways. One of the most representative methods widely studied is a control strategy based on instantaneous optimization techniques such as Equivalent Consumption Minimization Strategy (ECMS) [7]–[9] and Pontryagin Minimum Principle (PMP) [10], [11]. ECMS and PMP have the advantage that they can be used as a real-time control strategy to achieve fuel efficiency optimization through equivalent calculations of the engine and fuel. However, similar to DP, these strategies need to reflect future driving information for control to achieve high fuel economy, which is given as an equivalent factor or co-state that represents the balance between fuel and electrical energy usage. As a result, to improve the fuel efficiency of hybrid vehicles as in DP, it is necessary to calculate an optimized solution that reflects the driving conditions of the vehicle [12], [13]. Accordingly, recent studies have been conducted in to predict and utilize future driving conditions. However, it is not easy to accurately predict these future driving speed profiles, and changing and learning the optimal control method according to the changing driving conditions of the vehicle requires a sophisticated algorithm and computational burden [14], [15]. Because of these problems, recent approaches have attempted to solve hybrid control problems using machine learning.

Reinforcement Learning (RL), a field of machine learning that has been actively researched in recent years, has a framework that can be applied to control problem suitably [16]. RL is one type of machine learning that has been developed based on the foundations of dynamic programming. Therefore, problems previously solved using DP such as the HEV optimal control problem are suitable for the control problem framework by applying RL. In fact, these RL techniques have been applied to HEV control, considering previous studies on stochastic dynamic programming (SDP) [17]–[19].

Much work has been done regarding RL for energy management strategies of HEV control, especially Q-Learning. In [20], RL was applied to the power management strategies of HEVs, in which a Temporal Difference (TD)-learning algorithm was used to derive the optimal control policy. In [21], RL was applied to the power management strategy for a Plug-in hybrid electric vehicle (PHEV), in which the remaining distance to travel was chosen as a state variable and the immediate reward was defined as the sum of the fuel consumption cost and battery energy usage cost. In [22], RL was used to optimize the power distribution between the battery and the ultra-capacitor for a PHEV. In this paper, the transition probability matrices were updated based on the driving cycle and Kullback-Leibler divergence rate. [23] presented the RL-based energy management strategy for a hybrid electric tracked vehicle, in which Q-learning and the Dyna algorithm were applied to generate the optimal control policy. [24] suggested a predictive energy management strategy based on RL and velocity prediction was applied to the parallel HEV. More recently, [25] utilized a Deep Q Network (DQN), which combined Q-learning and a deep neural network for HEV control.

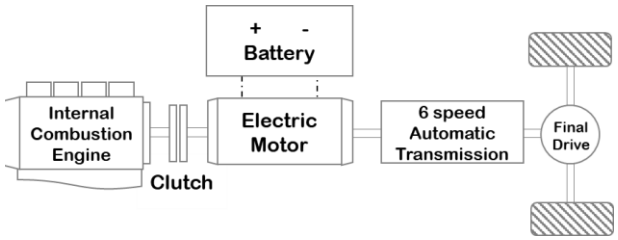


FIGURE 1. Vehicle simulation model.

In this paper, as in previous studies, we conducted a study on the HEV optimal control problem using RL. In particular, to apply RL to HEV control, we constructed a novel RL framework more suitable for the HEV optimal control problem based on previous studies [26], [27]. Especially, by separating the vehicle's internal powertrain environment and the vehicle's driving environment on the learning framework, we constructed a model-based Q-learning algorithm for energy management strategy of HEV, which is a more intuitive and explanatory learning framework for vehicle powertrain control. Accordingly, this approach was developed not just to find a generalized offline control policy according to many different driving patterns, but also to develop an online data-driven energy management strategy in which the vehicle controller is optimized with respect to the current given driving environment, thus allowing it to adaptively change the control policy according to change in the environmental data. The contribution of this paper is that by developing a novel optimal control framework using model-based Q learning applied to the HEV optimal control problem, the characteristics of the HEV optimal control problem and the intuition of the RL control technique for the HEV controller are better understood.

The remaining chapters are organized as follows. Chapter II gives a description of the HEV simulation model used in this paper. Subsequently, in Chapter III, the optimization problem to be solved in this paper is defined, and a novel algorithm using RL is proposed. Chapter IV discusses the feasibility and various features of the proposed algorithm based on simulation result, and finally, Chapter V gives the conclusions.

II. VEHICLE MODELING

In this study, the fuel efficiency performance and validity of the proposed algorithm are tested based on a vehicle simulation, thus it is very important to have a reliable vehicle powertrain model to perform simulations. In this study, we use a vehicle powertrain models consisting of each component model based on quasi-static modeling. For the powertrain structure, a parallel HEV is used, as given in Fig.1.

First, for engine modeling, a quasi-static engine fuel consumption model is utilized. It is assumed that the engine

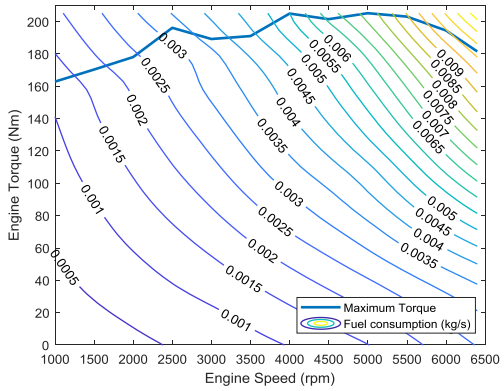


FIGURE 2. Engine fuel consumption map.

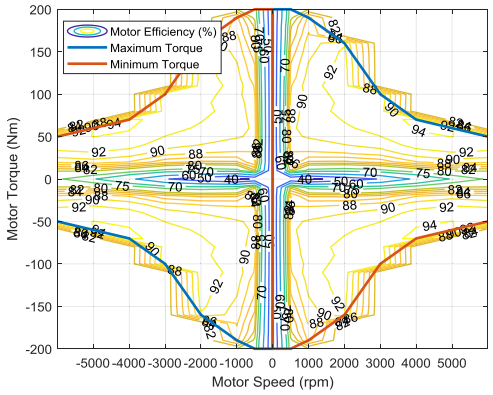


FIGURE 3. Motor efficiency map.

transient behavior such as the combustion dynamic are much faster than vehicle system level dynamics for energy flow analysis. The fuel consumption rate of the engine $\dot{m}$ is represented using map, as given in Fig.2 and (1), using the engine torque T_{eng} and engine speed ω_{eng}:

$$\dot{m} = f_{fuel}(T_{eng}, \omega_{eng}). \tag{1}$$

For the motor, the efficiency of the motor η_{mot} is calculated using the pre-determined map, and battery power output P_{bat} is also presented using the motor torque T_{mot} and motor speed ω_{mot}, as shown in (2).

$$P_{bat} = \eta_{mot}^{k} \cdot T_{mot} \cdot \omega_{mot} \tag{2}$$

The efficiency of the motor η_{mot} is a function of the motor torque T_{mot} and motor speed ω_{mot} as given in Fig. 3. If the machine is used as a motor, then $k = -1$, and if machine is used as a generator, $k = 1$. It is also assumed that the effects caused by the transient dynamics of the electric motor are sufficiently small, thus can be neglected. The battery power in (2) changes the State of Charge (SOC) in the battery, as modeled by the SOC dynamics described in (3), by considering an equivalent circuit model for the battery as shown in Fig. 4.

$$\dot{SOC} = -\frac{1}{Q_{bat}} \cdot \frac{V_{oc} - \sqrt{V_{oc}^2 - 4P_{bat}R_{bat}}}{2R_{bat}} \tag{3}$$

Here, the open circuit voltage of the battery is V_{oc}, the electric power consumed outside the battery is P_{bat}, the internal resistance is R and the battery capacitance is Q_{bat}. For the battery model, a simple internal resistance model is used. The open circuit voltage and internal resistance of the battery are determined by a pre-determined map, as shown in Fig. 5. For the powertrain, drivetrain dynamics from the transmission input shaft to the wheel can be expressed as shown in (4), (5), and (6) when a clutch is engaged.

$$T_{wh} = ((T_{eng} + T_{mot} - T_{gb\_loss}) \cdot \gamma_{gb} - T_{fd\_loss}) \cdot \zeta_{gb} \tag{4}$$

$$\omega_t = \gamma_{gb} \cdot \zeta_{gb} \cdot \omega_{wh} \tag{5}$$

$$T_t = T_{eng} + T_{mot} \tag{6}$$

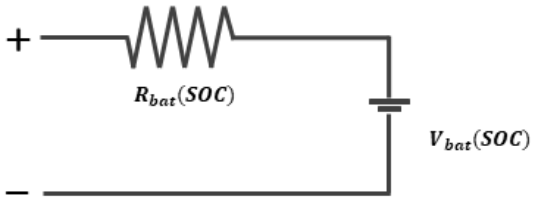


FIGURE 4. Equivalent battery model.

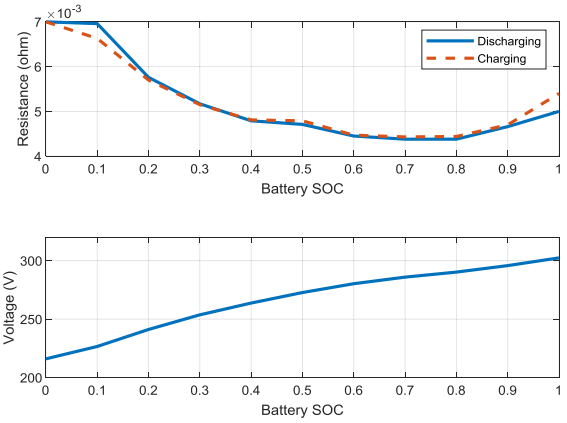


FIGURE 5. Battery resistance and open circuit voltage.

Here, T_{wh} is the wheel torque, T_{eng} is the engine torque, T_{mot} is the motor torque, T_{gb_loss} is the torque loss in the transmission, γ_{gb} is the gear ratio, T_{fd_loss} is the final drive torque loss, ζ_{gb} is the final drive gear ratio, ω_t is the transmission input speed, ω_{wh} is the wheel speed, and T_t is the transmission input torque. The loss for the gear box is given as a three-dimensional map, as given in (7), as a function of T_t, ω_t, and the gear step number i_{gb}.

$$T_{gb\_loss} = L_{gb}(T_t, \omega_t, i_{gb}) \tag{7}$$

For the final drive gear, T_{fd_loss} is given as function of the final drive input speed ω_{fd} and the final drive input torque T_{fd}.

$$T_{fd\_loss} = L_{fd}(T_{fd}, \omega_{fd}) \tag{8}$$

TABLE 1. Vehicle model parameters.

Component	Value
Internal Combustion Engine	Maximum power 122 (kW) @ 6000 (rpm)
Electric Motor	Permanent Magnet Synchronous Motor Rated power: 30 (kW)
Battery	Capacitance: 5.3 (Ah)
Final Drive Gear Ratio	3.23
Gear Box	6 speed Automatic Transmission (4.21, 2.64, 1.80, 1.39, 1.00, 0.77)
Vehicle Mass	1700 (kg)

The vehicle model can be described simply as (9) and (10) by considering only the longitudinal vehicle dynamics.

$$\dot{v} = \frac{T_{wh}R_{tire} - F_{brake} - F_{loss}}{(M_{veh} + M_{eq})} \tag{9}$$

$$F_{loss} = f_0 + f_1 \times v + f_2 \times v^2 \tag{10}$$

Here, v is vehicle speed, R_{tire} is the tire radius, F_{brake} is the brake force, and F_{loss} is the road load loss, which includes the road grade. M_{veh} is the vehicle mass and M_{eq} is the equivalent mass for the rotating inertia of the powertrain component. Finally, f_0, f_1, and f_2 are the driving resistance coefficients. Some of the vehicle model parameters are shown in Table 1. Based on these vehicle models, the algorithm presented in the paper was tested and verified. The next chapter describes the algorithm.

III. ONLINE DATA-DRIVEN ENERGY MANAGEMENT STRATEGY FOR HYBRID ELECTRIC VEHICLE

In this paper, RL is used for energy management of HEV. In RL, learning is accomplished through feedback, giving appropriate compensation for the outcomes of the learning. The difference between supervised learning and RL is that unlike supervised learning, in which it explicitly corrects undesired behaviors, RL focuses on the online performance, which is one of the advantages that it is more suitable for applications in real-time control strategies for HEVs. Among the RL algorithms, Q-learning is utilized in this study. Q-learning is a method that allows the learning of optimal control online, where the Q function is learned using the temporal difference method based on interactions between the controller and environment. Based on Q-learning, as mentioned in the introduction, a novel energy management strategy has been developed specifically for the optimal power distribution problem of HEV control. Prior to this, the optimal control problem is explained first, followed by the new energy management strategy.

A. OPTIMAL CONTROL PROBLEM

First, the optimal control problem is defined to minimize the expected total cost over an infinite horizon as shown in (11).

$$\min J_\pi(x_0) = \lim_{N\to\infty} \mathrm{E}_{d_k}\left\{\sum_{k=0}^{N-1} \gamma^k g(x_k, \pi(x_k))\right\} \tag{11}$$

constrained by

$$\begin{aligned} \omega_{eng,min} &\le \omega_{eng}(k) \le \omega_{eng,max} \\ T_{eng,min}(\omega_{eng}(k)) &\le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k)) \\ T_{mot,min}(\omega_m(k), SOC(k)) &\le T_{mot}(k) \\ &\le T_{mot,max}(\omega_m(k), SOC(k)) \\ SOC_{min} &\le SOC(k) \le SOC_{max} \end{aligned}$$

Here, x_k is the state variable, g is the instantaneous cost incurred, γ is the discount factor that represents the future cost as the expected value of the cost at current time step, $J_\pi(x_0)$ is the expected cost when the system starts at state x_0 and follows the policy π, and u is the engine power P_e, which is also discretized as

$$P_e \in \left\{P_e^1, P_e^2, \ldots, P_e^{N_u}\right\}, \tag{12}$$

where N_u is the number of discretized control inputs. The state variable x_k is composed of a four-dimensional state space as given below (13).

$$x_k = [SOC, P_{dem}, v, E_{on}] \tag{13}$$

Here, SOC is the battery state of the charge, and E_{on} is the engine on/off state. The engine on/off state is considered to avoid fuel consumption due to frequent engine changes to the on/off states. The instantaneous cost incurred g is defined as the equation below.

$$g = W_{fuel} + \zeta(SOC) + \beta \cdot \Delta E_{on} \tag{14}$$

Here, W_{fuel} is the instantaneous fuel consumption and β is the coefficient for the engine on/off penalty. $\zeta(SOC)$ is a term that penalizes the SOC deviation for charge sustenance as given below.

$$\zeta(SOC) = \begin{cases} \mu \cdot \left(SOC - SOC_{ref}\right)^2 & if\ SOC > SOC_{min} \\ C_{Penalty} & if\ SOC \le SOC_{min} \end{cases} \tag{15}$$

Here, μ and $C_{Penalty}$ are positive constant values for the SOC deviation. The underlying meaning of the optimal control problem is that the overall expectation of the cost for the infinite horizon is minimized instead of for a finite horizon, therefore the control policy result is time invariant, which can be easily implemented as a real-time vehicle controller. Note that the definition of this optimization problem is different from what the existing DP normally defines for the finite horizon or when using the Monte-Carlo method, which can learn from an episode of experience, and the final SOC constraint in DP is considered for the instantaneous cost.

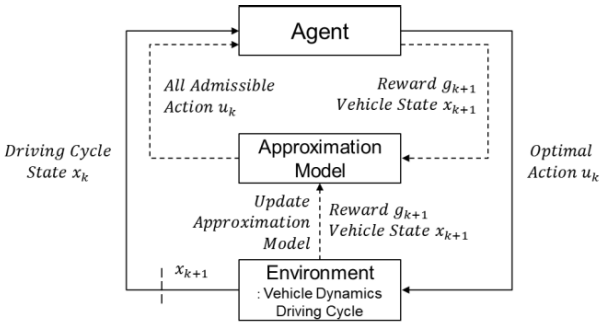


FIGURE 6. Concept of model-based reinforcement learning for an energy management strategy of hybrid electric vehicles.

B. MODEL-BASED Q LERNING

In this paper, to apply the Q-learning algorithm to the HEV control problem, a new energy management strategy based on the RL framework is developed. First, in Q-learning, the optimal cost $J^*(x_k)$ and optimal control policy $\pi^*(x_k)$ can be found as in the below equation using the Q-function:

$$J^*(x_k) = \min_u (Q^*(x_k, u)) \tag{16}$$

$$\pi^*(x_k) = \arg\min_u (Q^*(x_k, u)). \tag{17}$$

Further, the Q-function value can be updated as the below equation.

$$Q(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha \left(g_k + \gamma \min_u Q(x_{k+1}, u) - Q(x_k, u_k)\right) \tag{18}$$

When the system is in some state x_k, (i.e., in this HEV control problem, when the vehicle is in some state according to SOC_k, $P_{dem,k}$, v_k, and $E_{on,k}$), the control u_k is selected which has a minimum Q value. According to the action u_k, the state x_k changes to x_{k+1} with immediate reward g_k, then based on the Q value at the new state x_{k+1} and g_k, the Q-function value $Q(x_k, u_k)$ is updated with the Bellman equation. Equation (18) presents the baseline of the Q-learning algorithm.

Based on this algorithm, a new online data-driven energy management strategy using model-based Q-learning is proposed in this study. In the case of conventional Q-learning, the most important one is that of convergence. When applying Q-learning to various problems, including the HEV control problem, there is often difficulty considering the convergence properties or the state dimension is too large, thus taking a long time to converge. Additionally, there is the issue of the curse of dimensionality, as in DP. In this paper, we propose an algorithm that fits the framework of the HEV optimal control problem.

Fig. 6 and Fig. 7 present the concept of the algorithm and pseudo code, respectively. The idea of the algorithm presented in this paper is as follows. In HEV control problems, the states in (13) can be divided into stochastic and deterministic parts. That is, considering the driving environment

Algorithm for HEV control

Initialize $Q(x_k, u_k)$

Repeat each step $k = 1,2,3, \ldots$

1. Choose action optimal u_k based on $Q(x_k, u_k)$
2. Taking action u_k, observe reward $g(x_k, u_k)$, state x_{k+1}
3. 1 Update model based on observation
$$g(x_k, u_k) \leftarrow g(x_k, u_k) + \alpha(g_k - g(x_k, u_k))$$
$$x_{k+1}(x_k, u_k) \leftarrow x_{k+1}(x_k, u_k) + \alpha(x_{k+1} - x_{k+1}(x_k, u_k))$$
3. 2 Update Q using model for all admissible action u_k
$$Q(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha\left(g_k + \gamma \min_u Q(x_{k+1}, u) - Q(x_k, u_k)\right)$$
4. $x_k \leftarrow x_{k+1}$

FIGURE 7. Pseudo code of the algorithm.

of the vehicle (i.e., P_{dem} and v), the vehicle moves probabilistically with uncertainty, while the state of the vehicle (i.e., SOC and E_{on}) moves deterministically via the control input according to the given control policy with the given driving environment. Further, the fuel consumption W_{fuel} can be modelled deterministically for the given driving condition and control input. In many existing papers, when using Q-learning or DQN, the vehicle and environment states are grouped together and trained entirely free of the model. The advantage of this model-free approach is a feature of Q-learning. However, if we model the powertrain of the vehicle and the state in vehicle dynamics, we should consider model-based techniques. In this study, the algorithm is composed by approximating the vehicle model, as shown in Fig. 6, thus there is an inner-loop process in which a learning process can conducted separately. In the proposed algorithm, first the control u_k is chosen based on $Q(x_k, u_k)$. However, unlike the conventional Q-learning algorithm, in which an ε-greedy policy is used often, here the action u_k is selected only based on $Q(x_k, u_k)$ (i.e., minimum Q value) without any exploration strategy. Instead, the Q-function value is updated based on interactions between the agent and vehicle state approximation model using the driving cycle information. While the optimal action u_k is chosen and implemented in the environment, the agent updates the Q-function value by investigating all admissible actions u_k based on the vehicle model (considering the burdensome computation, the action number of u_k in the inner-loop can be reduced). The reward g_{k+1} and vehicle state x_{k+1} (which are SOC_k, and $E_{on,k}$) according to the action u_k is obtained using the vehicle state approximation model, and the Q-function value is updated by combining these data with the driving cycle state x_{k+1} (which are $P_{dem,k}$, and v_k).

The underlying meaning of this structure is that in terms of the exploration-exploitation dilemma, by separating the deterministic vehicle model state from the stochastic vehicle driving environment state, exploration of the control according to the vehicle driving environment is increased while exploitation of the control policy is secured. Thus,

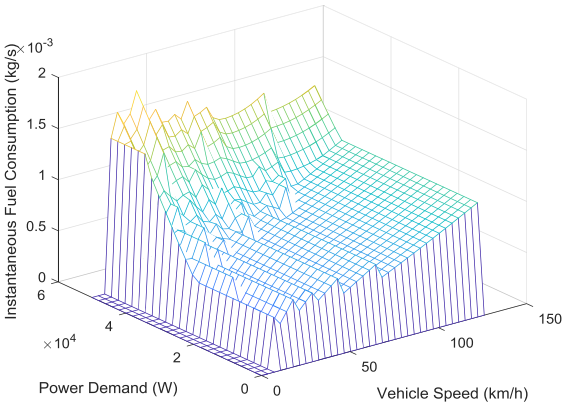


FIGURE 8. Example of the engine fuel consumption approximation model when the control input u_k (engine torque) is 37.7 Nm and the engine is on ($E_{on} = 1$).

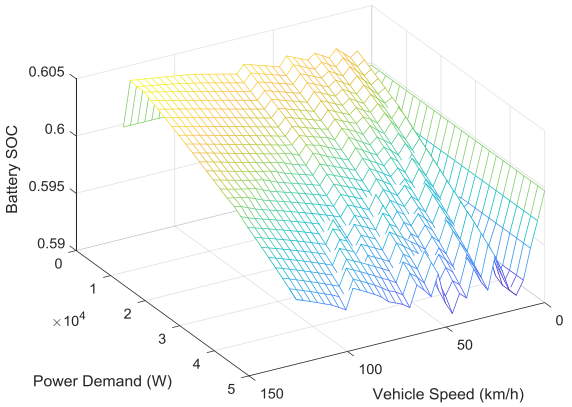


FIGURE 9. Example of the battery SOC approximation model when the current battery SOC is 0.60 and the control input u_k (engine torque) is 121.5 Nm.

the proposed algorithm works differently from existing Q-learning or DQN, where an ε-greedy policy is used often. That is, random selection of the control input for the HEV control problem decreases the fuel economy performance for exploration. Additionally, considering that these random control inputs in the exploratory strategy can cause undesirable behavior or even fatal errors in the vehicle system, the proposed algorithm has the advantage of stability and robustness, which is very important for vehicle control characteristics. Further, similar to DQN, experience replay could be conducted by updating Q using the vehicle model for different actions, which helps convergence.

On the other hand, the vehicle state approximation model is updated using the information obtained from interactions between the agent and environment as in the equation below.

$$g(x_k, u_k) \leftarrow g(x_k, u_k) + \alpha (g_k - g(x_k, u_k)) \tag{19}$$

$$x_{k+1}(x_k, u_k) \leftarrow x_{k+1}(x_k, u_k) + \alpha (x_{k+1} - x_{k+1}(x_k, u_k)) \tag{20}$$

The vehicle model is defined as above and updated using the results of the interaction between the actual agent and the environment. Note that it is still possible to have a model-free property, which is an advantage of Q-learning. The initial approximation of the vehicle model only helps faster learning and convergence of the algorithm. In other words, even if the model is not accurate, it can be modified by learning from the driving data, which allows optimal control to be explored. The vehicle approximation model (battery SOC and fuel consumption) is given as a four-dimensional look-up table that is a function of the state and control as written in equations below.

$$SOC_{k+1} = f_{soc}(SOC_k, P_{dem}, v, u) \tag{21}$$

$$W_{fuel} = f_{fuel}(P_{dem}, v, E_{on}, u) \tag{22}$$

Fig. 8 and Fig. 9 present examples of the battery SOC model and fuel consumption model, respectively.

The advantage of the proposed algorithm is that it separates the vehicle model from the environment differently from the existing Q-learning-based energy management strategies. In the case of the vehicle state approximation model, the future vehicle state and reward (battery SOC, engine on/off state, fuel consumption) can be derived when certain control inputs are given with along with the current vehicle state information. However, in the case of driving cycle information, it is not easy to accurately predict the change in vehicle speed and the required power demand. Therefore, for the vehicle powertrain, the state approximation model is considered in the control algorithm through modeling and is updated based on the initial value and learning. However, in the case of the vehicle driving cycle, the model is configured to learn the driving data based on interactions between the agent and environment as in the existing Q-learning based energy management strategy. Thus, based on the vehicle state model approximation, the uncertainty of the state transition model can be significantly reduced, and based on these vehicle state approximation, the decision making process could be explained more explicitly; this is unlike in conventional RL, which lacks visibility for the learning process.

Therefore, the state related vehicle model and control can be learned using full backups, and the driving environment can be learned using sample and shallow backups. Compared to the SDP algorithm, in which the driving cycle information is expressed as a transition probability matrix (TPM), the proposed algorithm updates instantaneous driving cycle information using the Q-function value and is stored based on the Bellman equation as if the TPM is updated at every moment. On the other hand, in the proposed algorithm, using the vehicle model, it is possible to derive the optimum control value by examining the vehicle state change and the compensation value according to all possible control inputs, as in SDP.

IV. SIMULATION ANALYSIS

The effectiveness of the proposed algorithm described above was verified through vehicle simulations. We investigated how well the learning process is actually conducted using the proposed algorithm, and how accurate the fuel economy

TABLE 2. Simulation conditions.

Parameter	Minimum value	Maximum value	Interval
Vehicle speed, v (m/s)	0	40	1
Battery SOC, SOC (%)	45	75	0.01
Power demand, P_{dem} (W)	0	96000	2000
Engine on/off, E_{on}	0(off)	1(on)	-
Engine Torque T_{eng} (Nm)	0	205.2	4.1

TABLE 3. Equivalent fuel economy (km/l) results.

Algorithm	Driving Cycle	
	UDDS	HWFET
Deterministic DP	26.1	26.2
RL-based	24.9	25.7
Rule-based	21.5	22.8

performance results based on learning are compared to the fuel economy of the DDP, which represents the optimal fuel economy. Additionally, simulation results with the conventional rule-based strategy are presented for comparison. First, discretization of the parameters is performed as in Table 2.

A. SIMULATION USING STANDARD DRIVIG CYCLE

Standard driving cycles for the Urban Dynamometer Driving Schedule (UDDS) and Highway Fuel Economy Test (HWFET) are used for the learning process and the vehicle simulation. Fig. 10 presents the learning curve for UDDS, in which the cumulative reward decreased rapidly as iterations were repeated. As the iterative learning continues, the cumulative reward value becomes smaller and convergence can be confirmed. Fig. 11 presents the battery SOC results for each simulation for UDDS. First, there is nothing previously learned, thus the battery SOC is decreased; this is because the controller will select the control to minimize the immediate fuel consumption and SOC deviation penalty without considering the discounted cost of the next state. Thus, the battery SOC value becomes smaller to reach the minimum boundary SOC value (0.45). However, as the learning process is repeated, the battery SOC is sustained near the target battery SOC value (0.60). In the same way, the simulation for HWFET is conducted. The strategy is simulated using the HWFET driving cycle repeatedly for learning, and the fuel efficiency performance is measured. Table 3 presents the equivalent fuel economy performance of the strategy for UDDS and HWFET, which are trained for each cycle separately. The simulation results show that

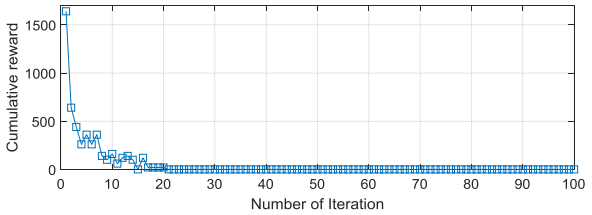


FIGURE 10. Cumulative reward accord to the number of iterations.

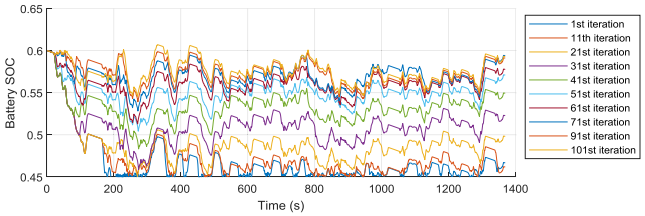


FIGURE 11. Battery SOC trajectory according to the number of iterations.

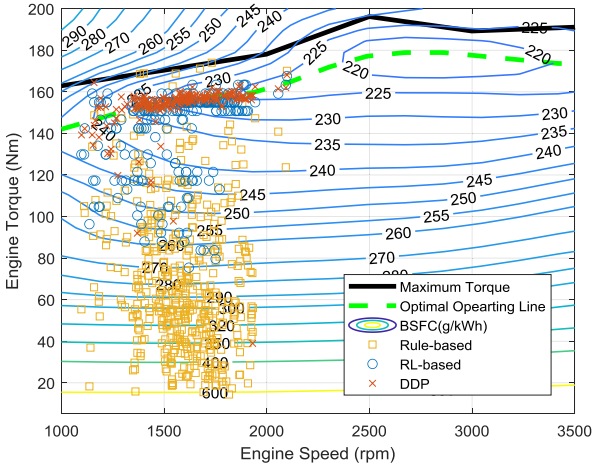


FIGURE 12. Engine operating point of the RL-based, rule-based and DDP strategies for vehicle simulation using UDDS.

in the case of UDDS, the RL-based strategy exhibits a fuel economy performance of 24.9 km/l, which is 95.4% of the optimal fuel efficiency for DDP. For the HWFET RL-based strategy, the fuel economy is 25.7km/l which is 98.1% of the DDP results. In both cases, we confirmed that the results of the RL-based strategy are better than the results of the rule-based strategy. The RL-based strategy presents a very similar behavior for the engine operating point with the DDP, as shown in Fig. 12. Fig. 12 shows that in both the DDP and RL-based strategies, the engine is operated near the optimal operating line, which has a relatively high Brake Specific Fuel Consumption (BSFC) efficiency. However, the fuel economy results for the RL-based strategy cannot reach that of DDP even though it is trained repeatedly using the driving cycle. This is because the optimization problem is defined with an infinite time horizon rather than a finite driving cycle, thus the derived optimal control is not suitable for the deterministic case.

TABLE 4. Equivalent fuel economy (km/l) for the learning ability simulations.

Algorithm		Driving Cycle	
		UDDS	HWFET
Deterministic DP		26.1	26.2
RL-based	UDDS	24.9	24.6
	HWFET	23.9	25.7
	HWFET+UDDS	24.9	25.4
	UDDS+HWFET	24.9	25.7
Rule-based		21.5	22.8

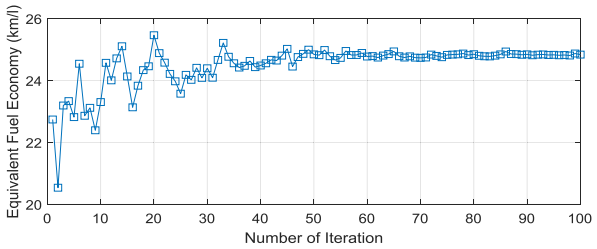


FIGURE 13. Equivalent fuel economy results of vehicle simulations using UDDS for a pre-learned setup with HWFET.

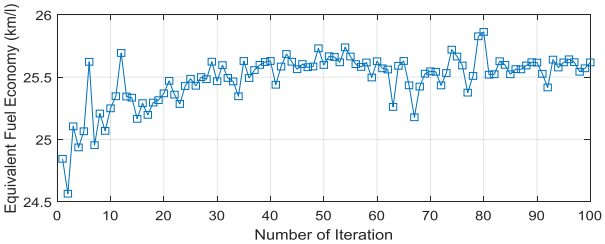


FIGURE 14. Equivalent fuel economy results of vehicle simulations using HWFET a pre-learned setup with UDDS.

B. ONLINE DATA-DRIVEN LEARNING

On the other hand, the learning ability of the proposed strategy was also tested through simulation. In these simulations, the UDDS and HWFET driving cycles are used for learning, and learning is performed again for different driving cycles (HWFET and UDDS) to determine whether new learning occurs with the existing learned data. Fig. 13 presents the equivalent fuel economy results as learning is performed for the UDDS driving cycle using pre-learned data with the HWFET driving cycle. It is seen that equivalent fuel economy is increased as iterations are repeated. Similarly, Fig. 14 shows the process for re-learning the HWFET driving cycle using the data learned during the UDDS driving cycle. Additionally, it can be confirmed that the fuel consumption value increases as learning is repeated, eventually converging to a constant value. Table 4 shows the fuel efficiency performance for the UDDS and HWFET cycles of the learned strategy for different cycles. In the case of the UDDS driving cycle, learning with only UDDS exhibits the best fuel economy, while learning with HWFET only shows the best fuel economy

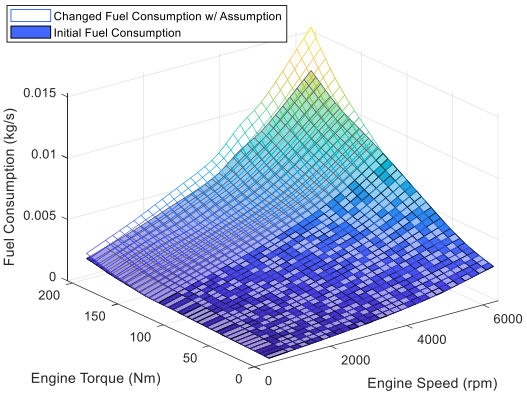


FIGURE 15. Vehicle fuel consumption model change based on assumption.

for the HWFET driving cycle. However, the fuel efficiency with different cycles exhibits reduced performance. When two cycles are learned, a similar fuel economy performance is seen compared to best fuel efficiency. The simulation results show that even when the proposed algorithm is implemented in a driving environment different from the initially trained driving environment (for example, from UDDS to HWFET, or HWFET to UDDS), a competitive fuel economy can be obtained based on the generalized control policy learned from existing learned data.

C. APPROXIMATION MODEL LEARNING

Finally, the model-free approach is tested via simulation. Generally, vehicles are exposed to various driving environments and performance deteriorates naturally. For example, aging of the engine or the performance degradation of the powertrain over time can happen in real vehicles, thus adaptation of the controller according to corresponding changes in the vehicle component performance is a necessary factor for minimizing the fuel efficiency reduction. One advantage of the proposed strategy is that the algorithms can learn by themselves and find the optimal control according to such environmental changes. In this case study, we deliberately changed the fuel consumption map of the engine model under the assumption that the engine consumes more fuel with a high torque area according to the performance reduction, and we verified that the proposed algorithm can work adaptively to learn and derive the optimal control rules according to this change. The fuel consumption map is intentionally modified as shown in Fig. 15, where the faint part of the high engine torque indicates the fuel consumption is rising. With this modified engine model, the RL-based energy management strategy is implemented with the HWFET driving cycle.

As a result, the strategy dynamically changes the existing set of parameters according to the change in elements to find the optimum control rule. Fig. 16 presents the vehicle fuel consumption approximation model in the RL-based energy strategy before and after changes in the fuel consumption map. Fig. 17 presents the BSFC map of the engine and the simulation results for the engine operating point with the

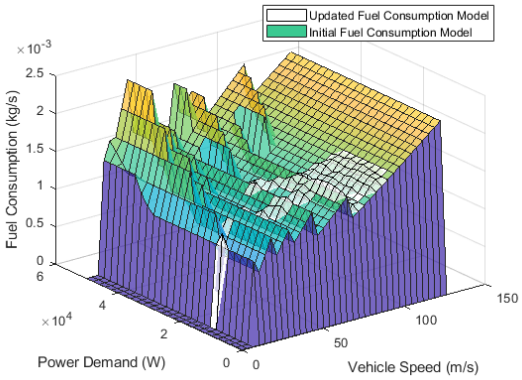


FIGURE 16. Updated vehicle fuel consumption model through learning when control input u_k (engine torque) is 121.5 Nm, and E_{on} (engine on/off signal) is 2.

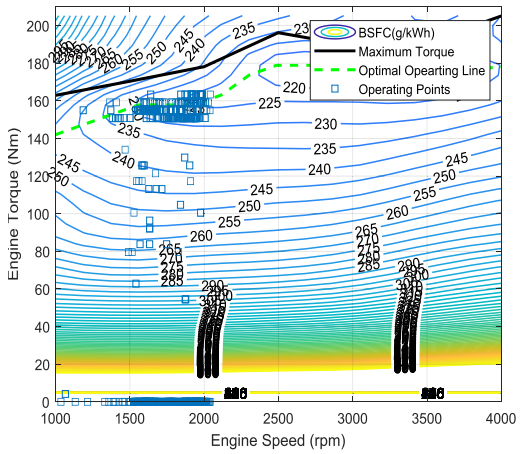
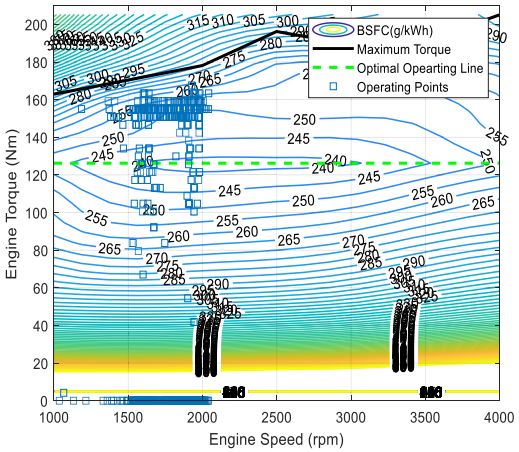


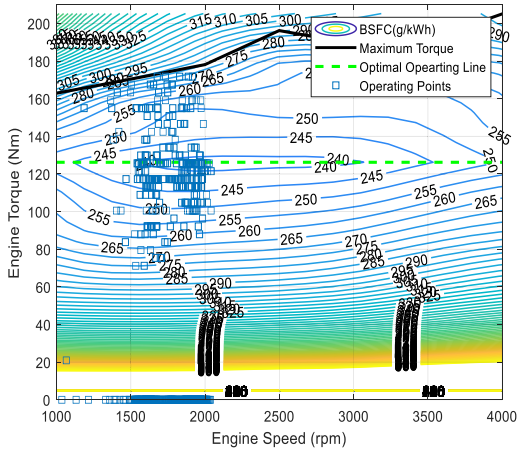
FIGURE 17. Engine operating point results with the original fuel consumption data.

original fuel consumption map data. On the contrary, Fig. 18 presents the simulation results for the engine operating points with the changed fuel consumption map data. It is seen that the BSFC map is changed as the fuel consumption map is modified intentionally; however, the engine operating point remains with the same area after the 1st iteration, as seen in Fig. 18 (a). After a few iterations, the engine operating point moves to the most efficient area of the BSFC, as seen in Fig. 18 (b), and (c). Additionally, according to the learning process, the fuel economy performance is also increased from 23.6 km/l for the 1st iteration to 24.6 km/l for the 20th iteration.

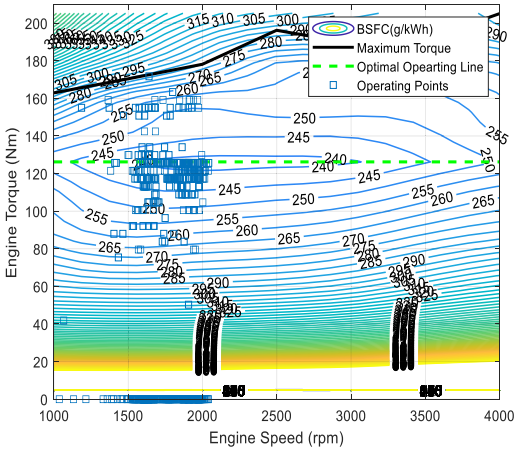
These results show that the control algorithm can adaptively find the optimal control policy when the performance or characteristics of the vehicle powertrain are changed, and those changes can be found using the vehicle state approximation model. This is possible by constructing the control framework for the powertrain model and the driving environment separately. Thus, the characteristics of the HEV optimal control and learning process of the RL control technique can be explained and understood more simply and intuitively through the vehicle powertrain engineering view, while the conventional RL-based strategy cannot explicitly describe the



(a) After learning with the 1st iteration. Equivalent fuel economy is 23.6 (km/l).



(b) After learning with the 10th iteration. Equivalent fuel economy is 24.1 (km/l).



(c) After learning with the 20th iteration. Equivalent fuel economy is 24.6 (km/l).

FIGURE 18. Engine operating point results according to the fuel consumption map data changes. Engine operating points move toward the efficient region as learning is conducted and equivalent fuel economy performance increases as a result.

decision making process. Thus an engineer also can check the model uncertainty in their powertrain model via reverse engineering or explain the control decision making process

considering component characteristic changes such as degradation of the battery performance or engine aging.

V. CONCLUSION

In this study, an RL-based control strategy was developed for the optimal control problem of HEVs. In the proposed RL- based control strategy, the transition probability of the vehicle's driving speed profile is learned online based on the driving data, and the control strategy is optimized based on model-based Q-learning. To obtain an improved fuel economy in HEVs, it is necessary not only to increase the efficiency of the vehicle powertrain, but also characterize the speed profile of the vehicle for use in the control strategy. The proposed control strategies in this paper have a powerful mathematical framework using reinforcement learning to model the driving cycle information from the stochastic view, and then solving the HEV supervisory control problem based on optimization using model-based approaches with an explainable and tunable vehicle state approximation model. As future work, experimental validation of the proposed control strategy is needed. Since the control strategy is verified based on simulations, it is necessary to verify the strategy based on experiments. Further, the tradeoff relationship of computational burdensome and fuel economy performance of the strategy should be investigated based on experimental evidence. Finally, combined with other practical issues such as emission or drivability, we expect that it is possible to advance the proposed strategy so that it is more practical and realistic.

REFERENCES

[1] F. R. Salmasi, "Control strategies for hybrid electric vehicles: Evolution, classification, comparison, and future trends," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 2393–2404, Sep. 2007.

[2] M. Sorrentino, G. Rizzo, and I. Arsie, "Analysis of a rule-based control strategy for on-board energy management of series hybrid vehicles," *Control Eng. Pract.*, vol. 19, no. 12, pp. 1433–1441, Dec. 2011.

[3] H. Banvait, S. Anwar, and Y. Chen, "A rule-based energy management strategy for plug-in hybrid electric vehicle (PHEV)," in *Proc. Amer. Control Conf.*, 2009, pp. 3938–3943.

[4] C.-C. Lin, H. Peng, J. W. Grizzle, and J.-M. Kang, "Power management strategy for a parallel hybrid electric truck," *IEEE Trans. Control Syst. Technol.*, vol. 11, no. 6, pp. 839–849, Nov. 2003.

[5] C. J. Mansour, "Trip-based optimization methodology for a rule-based energy management strategy using a global optimization routine: The case of the prius plug-in hybrid electric vehicle," *Proc. Inst. Mech. Eng., D, J. Automobile Eng.*, vol. 230, no. 11, pp. 1529–1545, Sep. 2016.

[6] H. Lee, J. Jeong, Y.-I. Park, and S. W. Cha, "Energy management strategy of hybrid electric vehicle using battery state of charge trajectory information," *Int. J. Precis. Eng. Manuf.-Green Technol.*, vol. 4, no. 1, pp. 79–86, Jan. 2017.

[7] G. Paganelli, S. Delprat, T. M. Guerra, J. Rimaux, and J. J. Santin, "Equivalent consumption minimization strategy for parallel hybrid powertrains," in *Proc. IEEE Veh. Technol. Conf.*, vol. 4, May 2002, pp. 2076–2081.

[8] S. Onori, L. Serrao, and G. Rizzoni, "Adaptive equivalent consumption minimization strategy for hybrid electric vehicles," in *Proc. ASME Dyn. Syst. Control Conf.*, vol. 1, 2010, pp. 499–505.

[9] G. Rizzoni, "A-ECMS?: An adaptive algorithm for hybrid electric vehicle energy management a-ECMS?: An adaptive algorithm for hybrid electric vehicle," *Eur. J. Control*, vol. 11, no. Dec. 2005, pp. 509–524, 2016.

[10] N. Kim, A. Rousseau, and D. Lee, "A jump condition of PMP-based control for PHEVs," *J. Power Sources*, vol. 196, no. 23, pp. 10380–10386, Dec. 2011.

[11] N. Kim, S. Cha, and H. Peng, "Optimal control of hybrid electric vehicles based on Pontryagin's minimum principle," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 5, pp. 1279–1287, Sep. 2011.

[12] C. Zhang and A. Vahidi, "Route preview in energy management of plug-in hybrid vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 2, pp. 546–553, Mar. 2012.

[13] C. Zheng, G. Xu, K. Xu, Z. Pan, and Q. Liang, "An energy management approach of hybrid vehicles using traffic preview information for energy saving," *Energy Convers. Manage.*, vol. 105, pp. 462–470, Nov. 2015.

[14] A. Rezaei, J. B. Burl, and B. Zhou, "Estimation of the ECMS equivalent factor bounds for hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 6, pp. 2198–2205, Nov. 2018.

[15] J. Zhang, C. Zheng, S. W. Cha, and S. Duan, "Co-state variable determination in Pontryagin's minimum principle for energy management of hybrid vehicles," *Int. J. Precis. Eng. Manuf.*, vol. 17, no. 9, pp. 1215–1222, Sep. 2016.

[16] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Oct. 2009.

[17] S. J. Moura, H. K. Fathy, D. S. Callaway, and J. L. Stein, "A stochastic optimal control approach for power management in plug-in hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 3, pp. 545–555, May 2011.

[18] C.-C. Lin, H. Peng, and J. W. Grizzle, "A stochastic control strategy for hybrid electric vehicles," in *Proc. Amer. Control Conf.*, 2004, pp. 4710–4715.

[19] T. Leroy, J. Malaize, and G. Corde, "Towards real-time optimal energy management of HEV powertrains using stochastic dynamic programming," in *Proc. IEEE Vehicle Power Propuls. Conf.*, Oct. 2012, pp. 383–388.

[20] X. Lin, Y. Wang, P. Bogdan, N. Chang, and M. Pedram, "Reinforcement learning based power management for hybrid electric vehicles," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2014, pp. 33–38.

[21] C. Liu and Y. L. Murphey, "Power management for plug-in hybrid electric vehicles using reinforcement learning with trip information," in *Proc. IEEE Transp. Electrific. Conf. Expo. (ITEC)*, Jun. 2014, pp. 1–6.

[22] R. Xiong, J. Cao, and Q. Yu, "Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle," *Appl. Energy*, vol. 211, pp. 538–548, Feb. 2018.

[23] T. Liu, Y. Zou, D. Liu, and F. Sun, "Reinforcement learning–based energy management strategy for a hybrid electric tracked vehicle," *Energies*, vol. 8, no. 7, pp. 7243–7260, 2015.

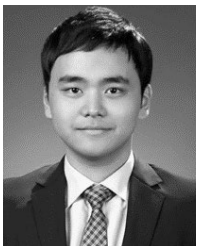
[24] T. Liu, X. Hu, S. E. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1497–1507, Aug. 2017.

[25] Y. Hu, W. Li, K. Xu, T. Zahid, F. Qin, and C. Li, "Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning," *Appl. Sci.*, vol. 8, no. 2, p. 187, 2018.

[26] H. Lee, "Stochastic optimal energy management based on Q-learning for hybrid electric vehicles," Ph.D. dissertation, Dept. Mech. and Aero. Eng., Seoul Nat. Univ., Seoul, South Korea, 2018.

[27] C. Song, H. Lee, K. Kim, and S. W. Cha, "A power management strategy for parallel PHEV using deep Q-Networks," in *Proc. IEEE Vehicle Power Propuls. Conf. (VPPC)*, Aug. 2018, pp. 1–5.

HEEYUN LEE received the B.S. degree in mechanical engineering from Sungkyunkwan University, South Korea, in 2013, and the Ph.D. degree in mechanical and aerospace engineering from Seoul National University, South Korea, in 2018. His current affiliation is with the Research and Development Division, Hyundai Motor Company, South Korea. His research interests include optimal control, reinforcement learning, modeling, and simulation of electrified vehicles.



CHANGBEOM KANG received the B.S. degree in mechanical aerospace from Seoul National University, South Korea, in 2012, and the M.S. degree in mechanical and aerospace engineering from Seoul National University, South Korea, in 2014, where he is currently pursuing the Ph.D. degree in mechanical and aerospace engineering. His research interests are energy management of hybrid electric vehicle and plug-in hybrid electric vehicles using Pontryagin's maximum principle and machine learning.

YEONG-IL PARK received the B.S. and Ph.D. degrees from Seoul National University, in 1981 and 1991, respectively. He is currently a Professor of mechanical system design engineering with the Seoul National University of Technology and Science. His research interests are novel hybrid powertrain and energy management strategy of hybrid electric vehicle, including plug-in hybrid electric vehicle.

NAMWOOK KIM received the B.S. and Ph.D. degrees from Seoul National University, in 2003 and 2009, respectively. He joined the Transportation Research Center, Argonne National Laboratory, in 2009, as a Postdoc and had worked from 2012 to 2015 as a Research Engineer. He is currently as an Associate Professor with Hanyang University–Ansan. His research interests include modeling and control for advanced vehicles, and he is also pursuing studies related large network behaviors of a transportation system.

SUK WON CHA received the B.S. degree in naval architecture and ocean engineering from Seoul National University, in 1994, and the M.S. and Ph.D. degrees in mechanical engineering from Stanford University, in 1999 and 2004, respectively. He is currently a Professor with the Department of Mechanical Engineering, Seoul National University. His current research interests include modeling of electric vehicle modules, and performance analysis of powertrain. He is also a Senior Editor of the *International Journal of Precision Engineering and Manufacturing–Green Technology*. He also serves as an Editor for the *International Journal of Automotive Engineering*.