IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR ELECTRIC POWER SYSTEMS

Received March 21, 2020, accepted April 1, 2020, date of publication April 7, 2020, date of current version April 21, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2986373*

# Comparative Analysis of Energy Management Strategies for HEV: Dynamic Programming and Reinforcement Learning

**HEEYUN LEE[1], CHANGHEE SONG[1], NAMWOOK KIM[2], AND SUK WON CHA[1]**
[1]Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826, South Korea
[2]Department of Mechanical Engineering, Hanyang University, Ansan 15588, South Korea

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

This work was supported by the Ministry of Trade, Industry, and Energy (MOTIE), South Korea, through the Technology Innovation Program (Development of RDE DB and Application Source Technology for Improvement of Real Road CO2 and Particulate Matter), under Grant 20002762.

**ABSTRACT** Energy management strategy is an important factor in determining the fuel economy of hybrid electric vehicles; thus, much research on how to distribute the required power to engines and motors of hybrid vehicles is required. Recently, various studies have been conducted based on reinforcement learning to optimally control the hybrid electric vehicle. In fact, the fundamental control approach of reinforcement learning shares many control frameworks with the control approach by using deterministic dynamic programming or stochastic dynamic programming. In this study, we compare the reinforcement learning based strategy by using these dynamic programming-based control approaches. For optimal control of hybrid electric vehicle, each control method was compared in terms of fuel efficiency by performing simulation by using various driving cycles. Based on our simulations, we showed the reinforcement learning-based strategy can obtain global optimality in the optimal control problem with an infinite horizon, which can also be obtained by stochastic dynamic programming. We also showed that the reinforcement learning-based strategy can present a solution close to the optimal one using deterministic dynamic programming, while a reinforcement learning-based strategy is more appropriate for a time variant controller with boundary value constraints. In addition, we verified the convergence characteristics of the control strategy based on reinforcement learning, when transfer learning was performed through value initialization using stochastic dynamic programming.

**INDEX TERMS** Dynamic programming, hybrid electric vehicle, optimal control, reinforcement learning, power management.

## I. INTRODUCTION

In recent years, because of the growing concern for the environment and consequent regulations in many countries, research and development of environment friendly vehicles have been conducted actively. Active research and development of hybrid electric vehicle (HEV), electric vehicle (EV), and fuel cell electric vehicle (FCEV) have resulted in the commercialization and production of these vehicles. HEV is a combination of an EV and a conventional internal combustion engine based vehicle, and can be driven by using the fossil energy produced by the internal combustion engine and the electric energy supplied by the battery. However, because of the recent increase in environmental regulations, research and development of EV or FCEV have been in the spotlight. Thus, for HEV to be competitive with EV and FCEV, further improvement of fuel economy in terms of efficiency is essential.

To improve the fuel efficiency of HEV, fundamental research such as weight reduction of the vehicle and enhancement of the efficiency of individual component such as the engine and motor could be conducted. In addition, in terms of the system level, study on the structure of HEV powertrain or the component sizing according to a given powertrain structure could be conducted. Further, another approach to improve the fuel economy is through a supervisory control of multiple power sources of HEV; this is called energy management strategy. HEV uses two power sources: an internal

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

combustion engine and electric battery as aforementioned; thus, the fuel economy varies substantially depending on how the power required by the vehicle is transferred from each power source [1], [2].

The research trends on the energy management strategy can be divided into rule-based strategy and optimization-based strategy. Rule-based strategy sets the power distribution of HEV according to the rules designed based on the expert's intuition or heuristic, and it offers a real-time control of the vehicle; however it is difficult to obtain a high fuel economy in various driving situations [3], [4]. In contrast, the optimization-based strategy is based on mathematical theory [5]–[9]. Compared with rule-based strategy, optimization-based strategy provides higher fuel economy, but it is difficult to use this strategy directly for real-time vehicle control because of the problems such as computational cost and causality problem for which the *a priori* knowledge of the driving conditions is required.

More recently, the research on vehicle control by using machine learning is being conducted actively. So far, an extensive research has been conducted mainly on autonomous driving; however, recently, many studies on the energy management strategy of HEV are being conducted by using reinforcement learning (RL). RL is one of the machine learning algorithms, which can be used to control the vehicle, and is based on the learning structure that uses interaction between controller and environment [10]. Various studies that are conducted by using RL are mentioned in the following. In [11], temporal difference learning, which is one of the RL algorithms, is utilized to optimally control the HEV because of its relatively higher convergence property and better performance in non-Markovian environment. Energy management strategy for a plug-in HEV is studied in [12]; here, an option to charge along the way is considered. In [13], RL is utilized to minimize the entire cost of gasoline fuel and use of electric battery in a plug-in HEV. Here, remaining trip distance is used for the control strategy, with an assumption that the remaining distance is highly correlated with the energy consumption in the future. In addition, in [14], RL is used for predictive energy management strategy to control a parallel HEV; here, vehicle speed is predicted based on the historic driving cycle data using nearest neighbor predictor and fuzzy encoding predictor, and Q-learning algorithm is used for energy management. In [15] and [16], a Transition Probability Matrix (TPM) is updated based on the forgetting factor and Kullback-Leibler divergence rate, and RL is utilized for control based on the TPM. In [17], a fast Q-learning algorithm is developed to accelerate the convergence rate, and cloud computation is suggested to lessen the computational burden in hardware-in-loop simulations. In addition, the Fuzzy Q-learning algorithm is developed in [18], where action-value function estimation is conducted using a neural network while fuzzy parameters are tuned based on the Q function value. In [19], online correction predictive energy management is developed; here, RL is combined with a fuzzy logic controller to eliminate the influence of prediction error. In [20], the weighted sum of the fuel and battery electric energy use is defined as the cost function of the Q-learning algorithm, which is then applied to a 48V mild HEV. More recent techniques such as Deep Q Networks are utilized in [21]–[24], which combine Q-learning with a deep neural network to obtain fast convergence and improve the learning performance; similarly, gradient-based methods such as Deep Deterministic Policy Gradient (DDPG) are utilized for HEV control in [25].

Although the recent studies on HEV energy management strategy use this RL technique, recent studies that are based on RL can be related with previous studies that are based on dynamic programming (DP) because RL, in fact, has been developed based on the DP expressed by the Bellman equation. DP is one of the optimization-based strategies, which can provide an optimal control of a given HEV system with a given speed profile [5]. This technique has been studied for various reasons: to assess the fuel economy performance of a new vehicle powertrain system [26], or to provide an insight to the control methodology of the rule-based strategy by investigating the optimal control results acquired by using it [27], [28]. In addition to this deterministic approach, a stochastic approach for the real time control of HEV, namely stochastic dynamic programming (SDP), has been investigated; this approach can be differentiated from previous DP approaches, namely deterministic dynamic programming (DDP). In [29], the control strategy of a parallel HEV is studied based on SDP, and in [30], the SDP is applied to HEV powertrain of Toyota Prius, which uses power-split system. In [31], the control strategy is validated experimentally by using SDP; here, the algorithm is implemented in the electronic control unit of the real vehicle.

DDP and SDP are known as global optimal control strategies [29], [32]; specifically, DDP is a time-variant optimal control policy over a specific driving cycle, while SDP is a time-invariant optimal control policy for a given transition probability, TPM. Both approaches are based on Bellman's principle of optimality, which RL is based on. However, there are few studies that compare RL-based approaches with SDP-based and DDP-based approaches. Especially, to optimize RL, RL- and SDP-based strategies can utilize the same optimal value function. In other words, in RL-based strategies, the value function for SDP can be converted to an equivalent Q function. However, the optimality the of RL-based strategy compared to SDP-based strategies is not presented clearly in the literature. Most papers, including [11], [13], [16], [21], and [24], compare the RL-based strategy only with the rule-based strategy. In [13], [18], [20], and [25], the reward function for RL includes the weighted sum of the fuel and electric energy use, thus an equivalent factor or co-state is needed to achieve optimality. In [14]–[19] and [33], RL-based strategies are developed based on the driving prediction method or as a time-variant controller. Here, the optimality of the RL methodology may be a result of DDP. However, the optimality here can only be shown with assumption that driving prediction is perfect, i.e., when

all future driving information is known in advance. Further, optimality of the RL strategy is not presented clearly due to problems with the data. These problems include a different driving cycle or TPM being used in the learning process than in testing as shown in [21], [34], and [35], or the approximation via a neural network as shown in [21], [18], [22], [23].

In this paper, we conducted a comparative study on HEV energy management strategies that use DDP, SDP, and RL based on a previous study [36]. The contribution of this paper is that by comparing the optimization processes of HEV control that use the three different algorithms of DDP, SDP, and RL, the recent studies on RL based energy management strategy can be understood better, and the performance of the algorithm can be thoroughly analyzed and compared. Based on the simulations, we showed that the RL-based strategy can obtain global optimality for the optimal control problem with an infinite horizon and the given TPM, as obtained by SDP. We also demonstrated that the RL-based strategy can obtain a solution close to the optimal result obtained by DDP, where the RL-based strategy is defined as a time-variant controller with boundary value constraints. Further, based on this comparative study, a new methodology is implemented by utilizing RL and transfer learning of SDP. The remaining sections of this study are organized as follows. In section II, different algorithms of DDP, SDP, and RL for HEV control are presented and compared. Then, in section III, vehicle simulation is conducted by using these algorithms, and the results are analyzed. Finally, in section IV, the conclusions are presented.

## II. ENERGY MANAGEMENT STRATEGY FOR HEV

In the HEV control problem, a model of the HEVs can be expressed in a discrete-time format as shown in the following:

$$x(k+1) = f(x(k), u(k)), \quad k = 0, 1, \ldots N-1 \tag{1}$$

where $x(k)$ is the state variable of system at time $k$, $u(k)$ is the control variable, $f(x(k), u(k))$ is the system dynamics, and $N$ is the duration of driving cycle. The optimization problem for fuel economy of HEV can be defined to find the control input $u(k)$, which minimizes the cost function $J$ as shown in the following:

$$\begin{aligned} &\min J = \sum_{k=0}^{N-1} L(x(k), u(k)) \\ &\text{subjectto } SOC(0) = SOC(N) \\ &\qquad \omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max} \\ &\qquad T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k)) \\ &\qquad T_{mot,min}(\omega_{mot}(k)) \le T_{mot}(k) \\ &\qquad \le T_{mot,max}(\omega_{mot}(k)) \\ &\qquad SOC_{min} \le SOC(k) \le SOC_{max} \end{aligned} \tag{2}$$

where $L$ is the instantaneous cost, $\omega_{eng}$ is the engine speed, $T_{eng}$ is the engine torque, $T_{mot}$ is the motor torque, and $SOC$ is the battery state of charge (SOC). Note that generally, the final SOC value of the battery, $SOC(N)$ should be equal to the initial SOC value, $SOC(0)$; thus, only the fuel consumption can be evaluated. The optimization problem for HEV control is a constrained nonlinear optimal control problem.

### A. DDP BASED STRATEGY

As aforementioned, the DP is a well-known algorithm that aids in solving a complicated optimization problem effectively. For the optimization problem of HEV, DDP has been used in many studies to control the power distribution between engine and motor. DDP presents a global optimal solution by searching for all the possible control options; thus, it can be used to find the best fuel economy of the given vehicle system over a given driving cycle. In this study, the DDP is built up for comparing and suggesting the best available solution for the given optimal control structure. General optimal control problem defined in (2) can be defined for DDP as per the following equation:

$$\begin{aligned} &\min J = \sum_{k=0}^{N-1} (L(x(k), u(k)) + \beta \cdot \Delta E_{on}) \\ &\text{subject to } SOC(0) = SOC(N) \\ &\qquad \omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max} \\ &\qquad T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k)) \\ &\qquad T_{mot,min}(\omega_{mot}(k)) \le T_{mot}(k) \le T_{mot,max}(\omega_{mot}(k)) \\ &\qquad SOC_{min} \le SOC(k) \le SOC_{max} \end{aligned} \tag{3}$$

where state $x_k$ is the battery SOC, $SOC(k)$, control $u$ is the engine torque, $L$ is the instantaneous fuel consumption, $W_{fuel}$, and $\beta$ is the coefficient for the engine on/off event, $\Delta E_{on}$. A penalty term for the engine on/off event is added to the cost function to avoid frequent engine on/off. To consider the engine on/off, the engine on/off state can be added to the system state $x_k$; however, in this case, immense computations are required to be conducted. Therefore, in this study, the aforementioned penalty term is used. Thus, the result of DDP becomes more practical and comparable with the other simulation results.

In DDP, the optimization problem, (3) can be broken down into sub equations recursively, as shown in (4) and (5):

$$\begin{aligned} &For\ k = N-1, \\ &J_k^*(x(k)) = \min_{u(k)} [g(x(k), u(k))] \end{aligned} \tag{4}$$

$$\begin{aligned} &For\ 0 \le k < N-1, \\ &J_k^*(x(k)) = \min_{u(k)} [g(x(k), u(k)) + J_{k+1}^*(x(k+1))] \end{aligned} \tag{5}$$

where $J_k^*(x(k))$ is the optimal value function at time $k$, and $g$ is the instantaneous cost as shown in the following.

$$g = W_{fuel} + \beta \cdot \Delta E_{on} \tag{6}$$

Schematic of DP calculation is shown in Fig. 1. For a given driving duration time, the state variable, which is the battery SOC, $SOC(k)$ is discretized. For each time step, the instantaneous cost $g$ is calculated according to the state transition; thus, the optimal cost-to-go value of each state $J_k^*(x(k))$ can

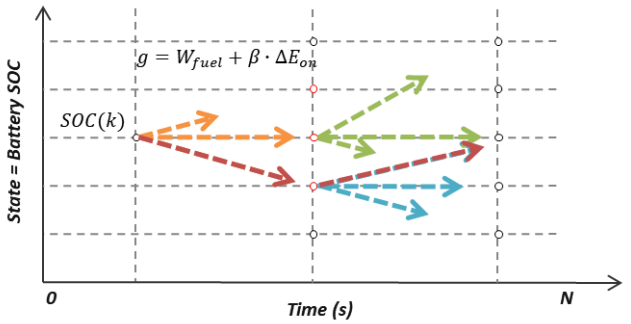



**FIGURE 1.** Schematic of the calculation process of DDP approach.

be calculated. For HEV optimal control problem, DDP provides an optimal power distribution between the ICE engine and electric motor, after considering the vehicle model, powertrain characteristic, and given driving cycle speed profile. DDP is one of the promising solutions, because it guarantees optimality. However, the application of DDP into real-time vehicle control is not feasible owing to its computational burdensome and non-causal characteristic where entire future speed profile of the vehicle should be known in advance before the trip; this means that the optimality of DDP cannot be applied in real world. SDP and RL based strategies are a variation of DDP; here, the result of optimized control policy can be used as a real-time vehicle controller, and a near-optimal result can be acquired. Firstly, the SDP based strategy is investigated as described in the following section.

### B. SDP BASED STRATEGY

In SDP, instead of a finite horizon problem, an infinite horizon problem is defined to minimize the expected total cost over an infinite horizon.

$$\min J_\pi(x_0) = \lim_{M\to\infty d_k} E\left\{\sum_{k=0}^{N-1} \gamma^k g(x_k, \pi(x_k))\right\}$$
$$\text{subject to } \omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max}$$
$$T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k))$$
$$T_{mot,min}(\omega_{mot}(k)) \le T_{mot}(k) \le T_{mot,max}(\omega_{mot}(k))$$
$$SOC_{min} \le SOC(k) \le SOC_{max} \quad (7)$$

where $x_k$ is the state variable, $g$ is the instantaneous cost incurred, $\gamma$ is the discount factor used to include the future cost into the expected value of the cost in the current time step, and $J_\pi(x_0)$ is the expected cost when the system starts at state $x_0$ and follows the policy $\pi$. State variable, $x_k$ is defined as a four-dimensional state space in this study, as shown in the following equation:

$$x_k = [SOC, P_{dem}, v, E_{on}] \quad (8)$$

where $SOC$ is the battery SOC, and $E_{on}$ is the engine on/ off state. Similar to DDP, engine on/off state is considered in SDP to avoid the fuel consumption due to frequent engine on/off. Instantaneous cost incurred, $g$ is defined as per the following equation:

$$g = W_{fuel} + \zeta(SOC) + \beta \cdot \Delta E_{on} \quad (9)$$

where $W_{fuel}$ is the instantaneous fuel consumption and $\zeta(SOC)$ is the term used for penalizing the SOC deviation for the charge sustenance as shown in the following.

$$\zeta(SOC) = \begin{cases} \mu \cdot (SOC - SOC_{ref})^2 & if\ SOC > SOC_{min} \\ C_{Penalty} & if\ SOC \le SOC_{min} \end{cases} \quad (10)$$

where $\mu$ and $C_{Penalty}$ are the positive constant values used for SOC deviation. The underlying concept of the SDP is that in SDP, overall expectation of the cost is minimized over an infinite horizon instead of a finite horizon; therefore, the control policy result is time-invariant and can be easily implemented as a real-time vehicle controller. Note that the final SOC constraint used in DDP is moved into the instantaneous cost in SDP, because the optimal control problem is defined as an infinite-horizon problem.

Unlike DDP, in SDP, the driving cycle is not provided as a speed profile, but as a transition probability matrix (TPM) by using Markov process. Based on the driving cycle information such as the historic driving data, the speed profile of driving cycle can be modeled as a TPM, in which the power demand of the vehicle at a given speed is expressed in terms of probability. An example of TPM is shown in Fig. 2; here, the TPM has a high value in the diagonal area, because the power demand of the driver, $P_{dem}$ is not changed suddenly. Therefore, there is a high chance that $P_{dem}$ will move to a near value. The driving pattern of the driver can be reflected into the TPM in this way; for example, the transition probability is widely distributed for a driving pattern with severe acceleration and deceleration.

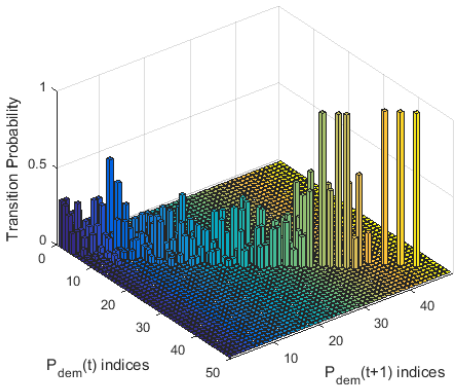



**FIGURE 2.** Example of TPM for a given vehicle speed of 30 km/h.

By using this TPM, we solved the SDP. Various approaches can be employed to solve this SDP problem; however, the most typical ones are value iteration and policy iteration. Policy iteration is expressed as a two-step process: value function approximation through policy evaluation step and control policy search through policy improvement step.

$$J_\pi(x_k) = g(x_k, \pi(x_k)) + E\{\gamma J_\pi(x_{k+1})\} \quad (11)$$

$$\pi'(x_k) = \underset{u \in U(x_k)}{\operatorname{argmin}} \left[ g(x_k, u) + E\left\{ \gamma J_\pi (x_{k+1}) \right\} \right] \quad (12)$$

In the case of value iteration, the policy evaluation and improvement steps are combined into one step to perform the process of value update and control policy search simultaneously.

$$J_{\pi'}(x_k) = \min_{u \in U(x_k)} \left[ g(x_k, u) + E\left\{ \gamma J_\pi (x_{k+1}) \right\} \right] \quad (13)$$

In this study, the optimization process that uses SDP was performed by using value iteration, in which policy evaluation and improvement steps are conducted one after another. Fig. 3 shows the schematic of the calculation process of SDP. Unlike the deterministic approach of DDP, in SDP, although the control is chosen, next state is not determined; instead it is determined stochastically based on the TPM. Thus, the expected cost $J_{\pi'}(x_k)$ is estimated based on the instantaneous cost $g$ and the sum of expectation values of discounted future cost $\gamma J_\pi (x_{k+1})$. The result of the generated control policy is time-invariant, and the control policy is based on the statistical approaches that use the probability transition of the driving cycle. Thus, it can be easily implemented on a real time vehicle controller.

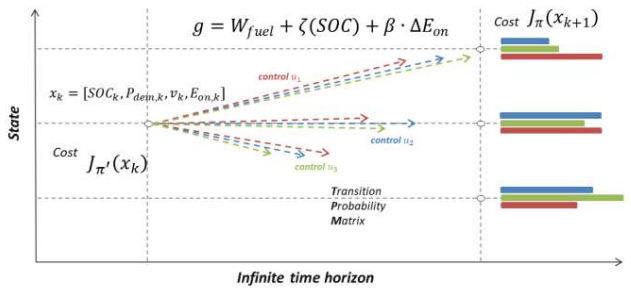



**FIGURE 3.** Schematic of the calculation process of SDP approach.

However, the drawbacks of SDP are that it requires TPM, and the optimization is conducted based on the TPM, such that the optimality of the control policy provided as a result is only valid for the given TPM. Therefore, if the characteristic of current driving speed profile changes, TPM should be updated and iterative optimization should be conducted again to obtain a new control policy relevant to the current driving condition. In addition, the iteration processes used in SDP are computationally burdensome, and is difficult to be used for online optimization. Further, similar to DDP, SDP requires system modeling, which is a model-based approach. Thus, if such modeling is not done correctly, an error between real vehicle powertrain and vehicle model used in the optimization may occur, and this could degrade the optimality of the control obtained through the SDP. RL based strategy is based on an online learning structure, and has a model-free property; thus, RL is one of the good approaches to solve the problems of SDP and DDP.

### C. RL BASED STRATEGY

In this study, Q-learning is employed for HEV optimal control among many RL algorithms. Q-learning is a method that allows the learning of optimal control online [37]. In Q-learning, Q function is learned by using the temporal difference method, which is based on the interaction between the controller and environment. Similar to SDP, the optimization goal of the control problem is to find the control input $u(k)$, which minimizes the cost function in an infinite horizon problem as shown in (7)-(10).

For Q-learning, the action and value can be defined as a Q function value, which is an action-value function as shown in the following equation:

$$Q_\pi (x_k, u_k) = g(x_k, u_k) + \gamma J_\pi (x_{k+1}) \quad (14)$$

Equation (14) means that $Q_\pi (x_k, u_k)$ includes immediate reward $g(x_k, u_k)$, which is the immediate cost when the state is $x_k$, the control $u_k$ is chosen, and the discounted cost of next state $x_{k+1}$ is included, which follow control policy $\pi$. Then, by using the Q function, the optimal cost $J^*(x_k)$ and optimal control policy $\pi^*(x_k)$ can be obtained as per the following equation:

$$J^*(x_k) = \min_u (Q^*(x_k, u)) \quad (15)$$

$$\pi^*(x_k) = \arg\min_u (Q^*(x_k, u)) \quad (16)$$

In addition, the Q function value can be updated as Bellman equation as shown in the following equation:

$$Q(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha(g_k + \gamma \min_u Q(x_{k+1}, u) - Q(x_k, u_k)) \quad (17)$$

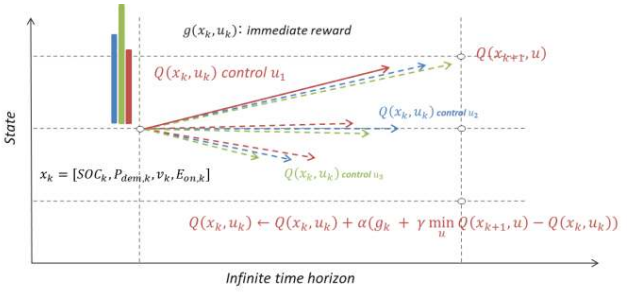



**FIGURE 4.** Schematic of the calculation process of RL approach.

where $\alpha$ is learning rate. In this study, reinforcement learning algorithm developed in previous study [36] is used. Fig. 4 shows the concept of Q-learning calculation. When the system is in some state $x_k$ (when the vehicle is in some state of $SOC_k$, $P_{dem,k}$, $v_k$, and $E_{on,k}$), control $u_k$, which has a minimum Q value is selected. Similar to SDP, although the control is chosen, the next state is not determined; instead it is determined stochastically on the environment. According to action $u_k$, the state $x_k$ moves to $x_{k+1}$ along with the immediate reward $g_k$, and then, based on the Q value at the

new state $x_{k+1}$, and $g_k$, the Q value $Q(x_k, u_k)$ is updated to hold Bellman equation.

The difference between the proposed and existing SDP algorithms lies in the model-based approach. In SDP, the driving cycle information is expressed as TPM, which is defined in the optimal control problem as driving cycle information. In RL, as the learning process is repeated, the transition between states accumulates and learns these models through Q learning to find an optimal control without TPM modeling. In addition, the vehicle powertrain model need not be built in Q-learning. RL is particularly well-suited to problems that include a long-term versus short-term reward trade-off. In this HEV problem, the battery SOC should be sustained as long-term, and the fuel consumption should be minimized as short-term at the same time; thus, the HEV problem is well suited to the RL approach.

## III. VEHICLE SIMULATION

In this study, based on the different strategies of DDP, SDP, and RL, vehicle simulations are conducted on various driving cycles, and their results are compared. In addition, a rule-based strategy is simulated for comparison. For the rule-based strategy, power follower control strategy is used. First of all, vehicle model used in this study is explained briefly in the following section.

### A. VEHICLE MODELING

Parallel HEV is used as the vehicle model. A gasoline engine with a maximum power of 122 kW is used, and a 30 kW electric motor with 5.3 Ah Li-ion battery is used. For transmission, a six speed automatic transmission is used, and the vehicle total mass is 1700 kg. The structure of the HEV is shown in Fig. 5. For the engine, the fuel consumption $\dot{m}$ is calculated based on the quasi-static assumption, as shown in the following equation:

$$\dot{m} = w_{fuel}(T_{eng}, \omega_{eng}) \tag{18}$$

where $w_{fuel}$ is a function of the engine torque $T_{eng}$ and engine speed $\omega_{eng}$. For the motor, the battery power $P_{bat}$ is derived using the motor efficiency $\eta_{mot}$, which is also a static function of the motor torque $T_{mot}$ and motor speed $\omega_{mot}$, as shown in the following equation:

$$P_{bat} = \eta_{mot}^{-sgn(T_{mot})} \cdot T_{mot} \cdot \omega_{mot} \tag{19}$$

The battery power $P_{bat}$ determines the change in battery SOC $\dot{SOC}$, as shown in the following equation:

$$\dot{SOC} = -\frac{1}{Q_{bat}} \cdot \frac{V_{oc} - \sqrt{V_{oc}^2 - 4P_{bat}R_{bat}}}{2R_{bat}} \tag{20}$$

where $Q_{bat}$ is the battery capacity, $V_{oc}$ is the open circuit voltage of the battery, $R_{bat}$ is the internal resistance, and $Q_{bat}$ is given as a constant. Note that $V_{oc}$ and $R_{bat}$ are assumed to be nonlinear functions of the battery SOC.

The powertrain dynamics are as shown in the following equation:

$$T_{wh} = ((T_t - T_{gb\_loss}) \cdot \gamma_{gb} - T_{fd\_loss}) \cdot \gamma_{fd} \tag{21}$$

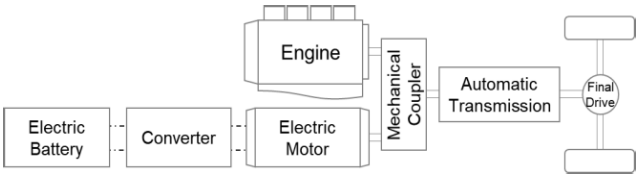



**FIGURE 5.** Vehicle simulation model.

$$\omega_t = \gamma_{gb} \cdot \gamma_{fd} \cdot \omega_{wh} \tag{22}$$

Here, $T_{wh}$ is the wheel torque, $T_t = T_{eng} + T_{mot}$ is the transmission input torque, $T_{gb\_loss}(T_t, \omega_t, i_{gb})$ is the gear box loss, $\omega_t$ is the transmission input speed, $i_{gb}$ is the gear step number, $\gamma_{gb}$ is the gear ratio, $\gamma_{fd}$ is the final drive gear ratio, $T_{fd\_loss}(T_{fd}, \omega_{fd})$ is the final drive loss, $T_{fd}$ is the final drive input torque, $\omega_{fd}$ is final drive input speed, and $\omega_{wh}$ is the wheel speed. The vehicle model can be expressed as shown in the following equation:

$$\dot{v} = \frac{T_{wh}R_{tire} - F_{brake} - F_{loss}}{(M_{veh} + M_{eq})} \tag{23}$$

where $R_{tire}$ is the tire radius, $F_{brake}$ is the friction brake force, $F_{loss} = f_0 + f_1 \times v + f_2 \times v^2$ is the road load loss with road load coefficients $f_0$, $f_1$, and $f_2$, $M_{veh}$ is the mass of the vehicle, and $M_{eq}$ is the equivalent mass of rotating inertias of the vehicle components. Based on this vehicle model, simulations are conducted to compare different energy management strategies as described in the following sections.

### B. VEHICLE SIMULATION ON STANDARD DRIVING CYCLES

First, different energy management strategies are simulated with various standard driving cycles. Here, SDP and RL are trained using relatively long real-world driving cycle data, which is enough data for SDP and RL to learn generalized optimal control policy. These methods are also tested with standard driving cycles. Considering that the learning processes of SDP, and RL are time consuming and that the future driving cycle information is not known in advance, these simulation results accurately reflect the performance of general control rules for energy management using RL and SDP, which does not depend on the specific driving cycle. For DDP, optimal control is found for specific driving cycles.

For testing, standard driving cycles of Urban Dynamometer Driving Schedule (UDDS), Highway Fuel Economy Test (HWFET), Japan 10-15 (JN1015), Worldwide harmonized Light Vehicles Test Cycles (WLTC), and New European Driving Cycle (NEDC) are used, which are assumed as unknown driving cycles that are not used in the training. For the training, the real-world driving cycle is used. Real-world driving cycle data is obtained from the digital tachographs of taxi in the city of Seoul. Fig. 6 shows the real-world driving cycle A, in which the maximum speed of the driving cycle is less than 60 km/h, and the length of the driving cycle is approximately 4000 s. Fig. 7 shows the real-world driving cycle B, which is also a real-world driving cycle with a

**TABLE 1.** Equivalent fuel economy(km/l) result for different algorithms on standard driving cycles (Relative percent to deterministic DP result).

| Algorithm | Driving Cycle | | | | | |
|---|---|---|---|---|---|---|
| | UDDS | HWFET | JN1015 | WLTC | NEDC | Average |
| Deterministic DP | 26.1 | 26.2 | 26.3 | 24.6 | 24.6 | 25.6 |
| RL-based | 24.6 (94.3) | 25.5 (97.3) | 24.7 (93.9) | 23.0 (93.9) | 23.2 (94.3) | 24.2 (94.5) |
| Stochastic DP | 24.5 (93.9) | 25.1 (95.8) | 24.7 (93.9) | 22.8 (92.7) | 23.2 (94.3) | 24.1 (94.1) |
| Rule-based | 21.5 (82.4) | 22.8 (87.0) | 21.8 (82.9) | 21.0 (85.4) | 20.5 (83.3) | 21.5 (84.0) |

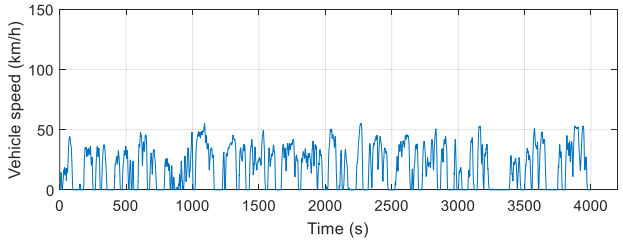



**FIGURE 6.** Real-world driving cycle A.

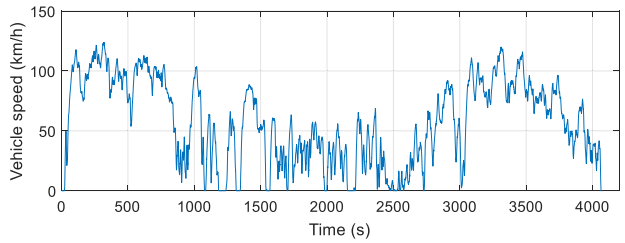



**FIGURE 7.** Real-world driving cycle B.

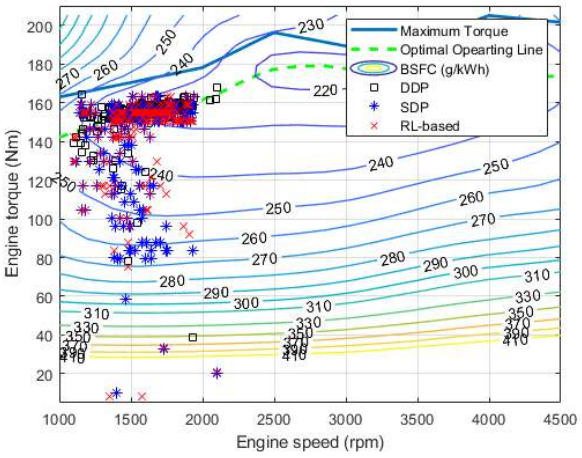



**FIGURE 8.** Engine operating points for the UDDS driving cycle.

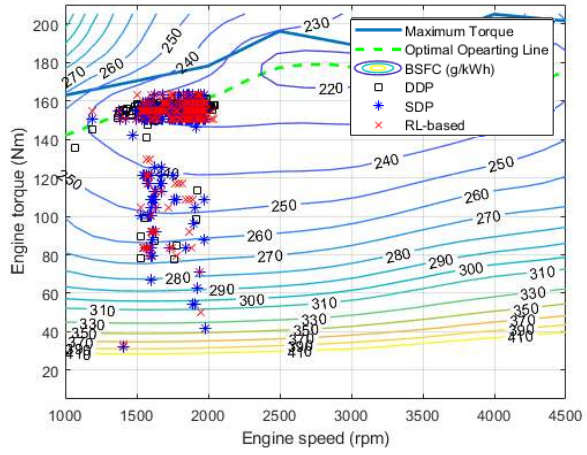



**FIGURE 9.** Engine operating points for the HWFET driving cycle.

maximum vehicle speed of more than 120 km/h, and the length of the driving cycle is also approximately 4000 s. For learning process in SDP, it is necessary to have sufficient driving data to construct TPM, thus real world driving cycle data is used.

Simulation results are shown in Table 1. The numbers in the bracket of Table 1 represent relative percentage values toward equivalent fuel economy of optimal result of DDP. The results indicate that the average fuel economy of the RL-based strategy is 24.2 km/l and the average fuel economy of the SDP is 24.1 km/l. Compared with the rule-based strategy, both the RL-based strategy and SDP exhibit increased fuel economy performance for all driving cycles. This is accomplished through the learning process, and the generally optimized control policy of the proposed algorithms using driving cycles A and B. The operating points of the engine for UDDS and the HWFET driving cycle are shown in Fig. 8 and Fig. 9, respectively. Note that the engine operating points for the SDP and RL-based strategies are concentrated on the optimal operating line, which is the same as DDP.

RL-based strategy exhibits a fuel economy similar to SDP, with a difference of equivalent fuel economy of less than 0.2 km/l between the RL-based and SDP approaches for different driving cycles, with the exception of the HWFET driving cycle; this is because it uses the same framework that is based on the Bellman equation for both SDP and RL-based strategies. The simulation results of the battery SOC and engine torque are shown in Fig. 10 for the UDDS driving cycle and Fig. 11 for the HWFET driving cycle. We found that the battery SOC and engine torque trajectory of the SDP and RL-based approaches are very close to one another, except for a few points. This implies that the SDP and

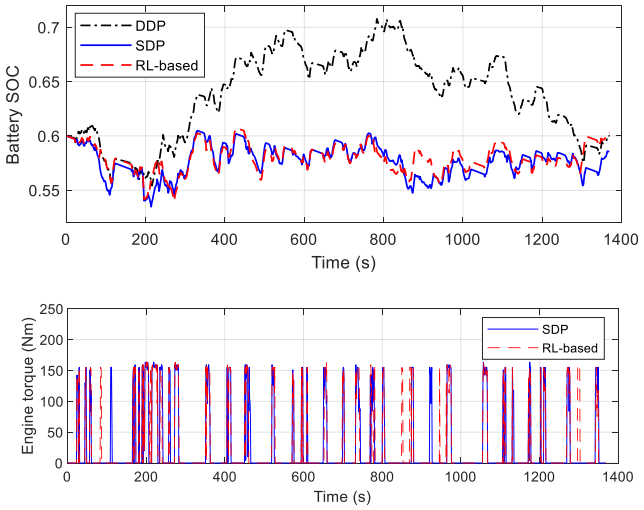



**FIGURE 10.** Simulation results for the UDDS driving cycle.

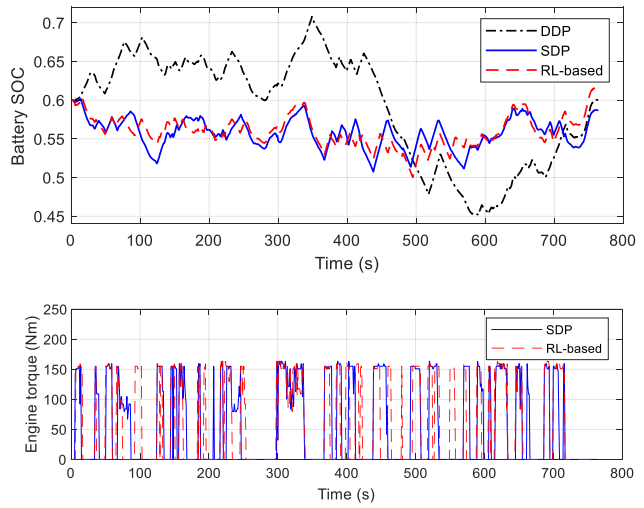



**FIGURE 11.** Simulation results for the HWFET driving cycle.

RL-based strategies can implement the same control policy when optimized for the same TPM. In this paper, the SDP and RL-based strategies are implemented based on the vehicle model simulations, and the fuel economy performances are very similar, except for a small gap due to numerical errors from SDP during the discretization process. However, in real-world scenarios, the RL-based strategy relies heavily on the model-free characteristic, thus may exhibit better performance than SDP if there is model uncertainty.

### C. OPTIMAL PERFORMANCE OF THE RL-BASED STRATEGY

In Table 1, compared with the fuel economy of DDP, SDP and RL-based strategies exhibit approximately 94% of the fuel economy of DDP. In the case of DDP, all the given driving cycle information is known in advance and is optimized in the finite horizon; this results in an optimal fuel economy result. In the case of SDP and RL-based strategies, optimization is performed on an infinite horizon to secure real-time control, thus the time-invariant control policy obtained from the SDP and RL-based strategies show decreased fuel economy performance compared to DDP. This difference in fuel economy can also be explained by the fact that the training data is different from the standard driving cycle, which is simulated with the general control policy acquired from the optimization result obtained using the real-world driving cycle.

The simulation results for data dependency learning for the SDP and RL-based strategies are shown in Table 2. In these simulations, the RL-based and SDP strategies are trained and tested using the real-world driving cycles A and B, unlike previous simulations in which real-world driving cycles are used for training and standard driving cycles are used for testing. For both the RL-based and SDP strategies, the fuel economy performance is improved when the same driving cycle is used for learning and testing. Especially, when testing driving cycle B while using cycle A to train the RL-based and SDP strategies, the fuel economy performance is decreased severely; this is expected considering that driving cycle A includes a low vehicle speed compared to driving cycle B.

**TABLE 2.** Equivalent fuel economy(km/l) result for different algorithms on real-world driving cycles (Relative percent to deterministic DP result).

| Algorithm | | Driving Cycle | |
|---|---|---|---|
| | | Real-world Driving Cycle A | Real-world Driving Cycle B |
| Deterministic DP | | 23.5 | 24.6 |
| RL-based | Learning w/ Driving Cycle A | 22.4 (95.3) | 22.2 (90.2) |
| | Learning w/ Driving Cycle B | 22.2 (94.5) | 24.0 (97.6) |
| SDP | Learning w/ Driving Cycle A | 22.3 (94.9) | 22.0 (89.4) |
| | Learning w/ Driving Cycle B | 22.0 (93.6) | 23.9 (97.2) |
| Rule-based | | 19.1 (81.3) | 21.7 (88.2) |

Therefore, in both RL and SDP, if the driving speed information of the environment to be driven is known in advance as in the DDP, and the optimization process can be performed, then a good fuel economy can be obtained. However, in the case of SDP, it is necessary to know the future driving environment information expressed in terms of TPM in the optimization process, and optimization must be performed by using this; eventually, to obtain a good fuel economy as in DDP, the future information must be known in advance. In the case of RL-based strategy, it is possible to actively improve the fuel efficiency by identifying and learning the characteristics of the driving environment while driving; however, this process could be time-consuming.

However, as mentioned above, there is still a gap in fuel economy performance between DDP and the time-invariant control of the RL-based and SDP strategies. In this paper, we define the optimal control problem with a finite horizon for the RL-based strategy and shows that the corresponding

optimal control policy converges to that for DDP. For the RL-based strategy, instead of (7), we define the optimal control problem as shown in the following equation:

$$min\ J_\pi\ (x_0) = \sum_{k=0}^{N-1} \gamma^k g\ (x_k, \pi\ (x_k)) \tag{24}$$

where the discount factor is $\gamma = 1$ and $x_k$ is the state variable, which includes the time $t_k$ as state information as follows:

$$x_k = [SOC, E_{on}, t_k] \tag{25}$$

Additionally, for the final battery SOC constraint in the RL-based strategy, which is the same as for DDP, $Q\ (x_N, u)$ is defined to have a reward for the target SOC, $SOC(N)$ as shown in following equation:

$$Q\ (x_N = [SOC, E_{on}, t_N]\ , u) = \begin{cases} C_{reward}, & \text{SOC} = SOC_{final} \\ 0, & \text{SOC} \neq SOC_{final} \end{cases} \tag{26}$$

Here, $C_{reward}$ is a negative value so that we define a minimization problem.

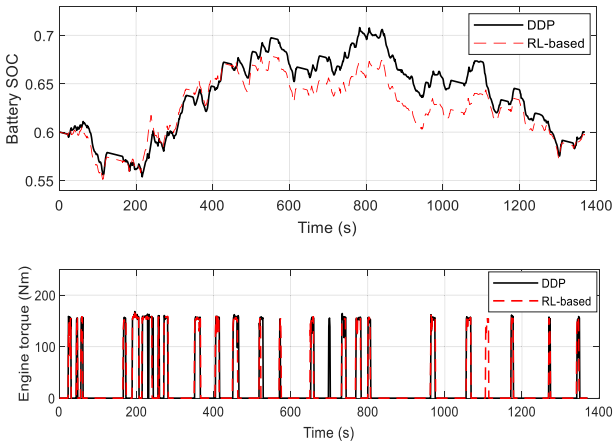



**FIGURE 12.** Simulation result for the UDDS driving cycle using the time-variant RL-based strategy.

Simulation results for the RL-based strategy are compared with those for DDP considering the UDDS and HWFET driving cycles, and are shown in Fig. 12, and Fig. 13. The simulation results show that the battery SOC and engine torque values for the RL-based strategy are very close to those of DDP. Especially, unlike the RL-based results shown in Fig. 10 and Fig. 11, the battery SOC range is expanded due to *a priori* information of the driving cycle. Additionally, note that the final battery SOC constraint is suitably satisfied as a result of Q-learning. The fuel economy performance for the time-variant RL-based strategy is shown in Table 3, which is either close to the DDP results or shows improved performance. Note that this is due to DDP exhibiting numerical errors during approximation in the discretization process. This implies that the RL-based strategy can achieve global optimality similar to DDP when the driving cycle is perfectly

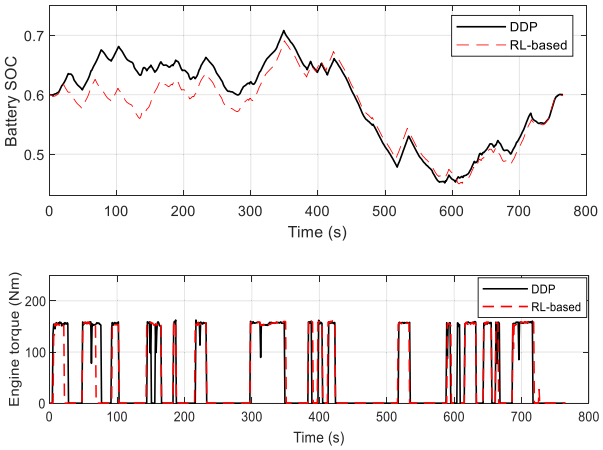



**FIGURE 13.** Simulation result for the HWFET driving cycle using the time-variant RL-based strategy.

**TABLE 3.** Equivalent fuel economy (km/l) results for time-variant control using the RL-based strategy (percent relative to the DDP results).

| Algorithm | Driving Cycle | |
|---|---|---|
| | UDDS | HWFET |
| DDP | 26.1 | 26.2 |
| RL-based | 26.2 (100.4) | 26.2 (100.0) |

known in advance and learning is successful. However, this RL-based strategy cannot be used as a vehicle controller since it only works when the driving cycle predictions perfectly match the current driving environment, similar to DDP.

**TABLE 4.** Computation time for different algorithm.

| Algorithm | DDP | SDP | RL |
|---|---|---|---|
| Computation time [a] | 0.5 h | 4 h | 6 h |

[a] Simulation was conducted based on MATLAB using Intel(R) Core(TM) i7-8700k CPU @ 3.70GHz, with 32.0GB RAM

### D. TRANSFER LEARNING

The RL-based strategy can improve the fuel economy performance. However, practical application to online real-time vehicle controllers is still distant considering this is a time-consuming process. The computation times for the DDP, SDP, and RL-based algorithms using driving cycle A both for training and testing are shown in Table 4. Estimations of the equivalent fuel economy performance according to the computation time for SDP and RL are given in Fig. 14. Note that different measures could be used to determine the convergence of SDP and RL, but in this paper, the equivalent fuel economy is considered. Compared to DDP, it takes a long time to converge for SDP and RL. The computation time varies according to driving cycle, methodology, and the conditions applied to DDP, SDP, and RL. However, the

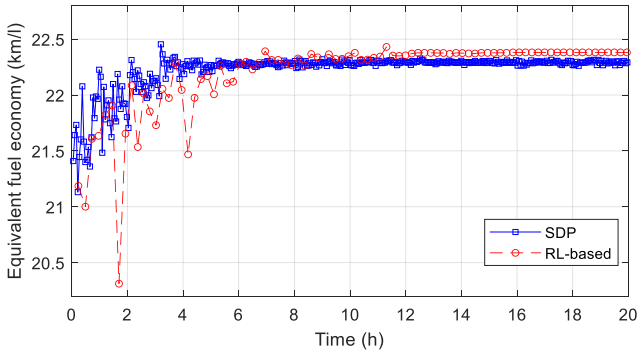



**FIGURE 14.** Fuel economy performance according to the computation time for SDP and RL.

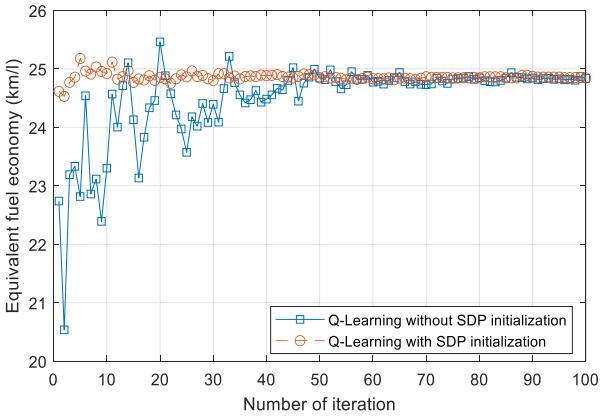



**FIGURE 15.** Comparison of Q-learning with and without transfer learning based on SDP.

computational burden of RL is large, even compared with SDP, such that optimal control is learned through interaction.

Considering the advantages and disadvantages of the SDP and RL-based strategies, this paper studies and simulates the transfer learning process. To employ an RL-based strategy such as Q-learning, it is necessary to perform learning with immense data at the beginning. This time-consuming learning is a limitation in utilizing RL-based strategies as real-time control. One way to reduce this computation time is to use SDP result for initializing Q function value. Thus, based on driving cycle characteristic, pre-optimized SDP result could be used to make Q-learning converge faster. In this paper, we confirmed how the influence of convergence performance can be affected when the initial Q table value of Q-learning is specified using SDP initialization. Optimal $Q^*(x_k, u_k)$ value can be defined based on the optimal cost $J^*(x_k)$, which is acquired from SDP as shown in the following:

$$Q^*(x_k, u_k) = g(x_k, u_k) + \gamma J^*(x_{k+1}) \tag{27}$$

Therefore, even when the driving cycle characteristic of the vehicle changes suddenly, high fuel efficiency performance can be still obtained by using the Q value derived from the initial SDP value with a similar historic driving cycle data, which is optimized in advance and fuel economy performance could converge faster.

The results of the control strategy that uses SDP initialization and the one that does not use SDP initialization are shown in Fig. 15. In the case of SDP initialization, the result of SDP that uses real world driving cycles (which were presented in the previous simulation using the standard driving cycle), is defined as the initial value of Q value for the control strategy. In the case that does not use SDP initialization, the Q value was learned only by using the HWFET cycle. When these two control strategies are simulated in the UDDS driving cycle, it can be confirmed that the fuel efficiency of the vehicle converges faster when SDP initialization is performed as shown in Fig. 15. The simulation results show that even if the driving cycle changes suddenly, it is possible to improve the fuel efficiency of the proposed strategy by using the information optimized through the SDP, and it is also possible to reduce the learning time for a specific driving cycle by using pre-calculated SDP result.

This knowledge transfer process has several advantages. First, the reduction of computation burden of the RL-based strategy, based on transfer learning, is important considering the process based on RL can require an extremely high computation performance from the hybrid control unit in the vehicle, which could be a barrier for the RL-based approach for vehicle control application. Second, in terms of the robustness of RL-based strategy, when there is a problem in the learning process owing to a lack of data, pre-calculated optimal results through SDP can be used to ensure the robust performance of RL-based strategy. For example, by expressing the historic driving cycle data by region, time, or person to represent each characteristic, similar historic data can be used for optimization through SDP, and this optimized knowledge can be transferred to the RL-based strategy.

## IV. CONCLUSION

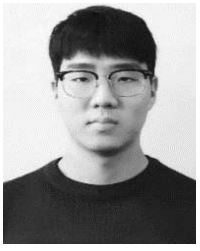
In this paper, DDP, SDP, and RL-based strategies for HEV's energy management strategy were studied. Based on vehicle simulations, we showed that the RL-based strategy can obtain optimal performance in the optimal control problem with an infinite horizon, as can also be obtained by stochastic dynamic programming. We also showed that the RL-based strategy can achieve a solution close to that of DDP when defined as a time-variant controller with boundary value constraints. In conclusion, SDP and RL-based strategies can be suggested as an approach to utilize DDP for a real vehicle control. In the case of SDP, the future driving environment must be modeled through TPM in advance, similar to DDP. By simply processing various driving cycles statistically, SDP can be used to draw out general control strategies from the optimization results, or it can be used in HEV controller with repeated driving patterns. The RL-based strategy is able to learn adaptively in this respect; however, because it takes immense time to learn, a process such as transfer learning that uses SDP can be established. As future work, it is necessary to study the framework for real-time control of such RL-based strategy, and experimentally validate it based on

the real vehicle. In addition, a comparative research that uses various RL algorithm-based optimal control frameworks will be studied along with the previous studies on DDP and SDP.

## 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] F. Zhang, X. Hu, R. Langari, and D. Cao, "Energy management strategies of connected HEVs and PHEVs: Recent progress and outlook," *Prog. Energy Combustion Sci.*, vol. 73, pp. 235–256, Jul. 2019.
[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.*, Jun. 2009, pp. 3938–3943.
[4] N. J. Schouten, M. A. Salman, and N. A. Kheir, "Fuzzy logic control for parallel hybrid vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 10, no. 3, pp. 460–468, May 2002.
[5] 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.
[6] 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.
[7] N. Kim, A. Rousseau, and D. Lee, "A jump condition of PMP-based control for PHEVs," *J. Power Sour.*, vol. 196, no. 23, pp. 10380–10386, Dec. 2011.
[8] N. Kim, S. Won Cha, and H. Peng, "Optimal equivalent fuel consumption for hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 3, pp. 817–825, May 2012.
[9] A. Rezaei, J. B. Burl, B. Zhou, and M. Rezaei, "A new real-time optimal energy management strategy for parallel hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 2, pp. 830–837, Mar. 2019.
[10] 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, 3rd Quart., 2009.
[11] 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.
[12] X. Qi, G. Wu, K. Boriboonsomsin, M. J. Barth, and J. Gonder, "Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles," *Transp. Res. Rec.*, vol. 2572, no. 1, pp. 1–8, 2016.
[13] 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.
[14] 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.
[15] Y. Zou, T. Liu, D. Liu, and F. Sun, "Reinforcement learning-based real-time energy management for a hybrid tracked vehicle," *Appl. Energy*, vol. 171, pp. 372–382, Jun. 2016.
[16] 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.
[17] G. Du, Y. Zou, X. Zhang, Z. Kong, J. Wu, and D. He, "Intelligent energy management for hybrid electric tracked vehicles using online reinforcement learning," *Appl. Energy*, vol. 251, Oct. 2019, Art. no. 113388.
[18] Y. Hu, W. Li, H. Xu, and G. Xu, "An online learning control strategy for hybrid electric vehicle based on fuzzy Q-Learning," *Energies*, vol. 8, no. 10, pp. 11167–11186, 2015.
[19] J. Wu, Y. Zou, X. Zhang, T. Liu, Z. Kong, and D. He, "An online correction predictive EMS for a hybrid electric tracked vehicle based on dynamic programming and reinforcement learning," *IEEE Access*, vol. 7, pp. 98252–98266, 2019.
[20] B. Xu, F. Malmir, D. Rathod, and Z. Filipi, "Real-Time reinforcement learning optimized energy management for a 48V mild hybrid electric vehicle," SAE Tech. Paper 2019-01-1208, Apr. 2019, pp. 1–9.
[21] 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.
[22] 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.
[23] J. Wu, H. He, J. Peng, Y. Li, and Z. Li, "Continuous reinforcement learning of energy management with deep q network for a power split hybrid electric bus," *Appl. Energy*, vol. 222, pp. 799–811, Jul. 2018.
[24] P. Zhao, Y. Wang, N. Chang, Q. Zhu, and X. Lin, "A deep reinforcement learning framework for optimizing fuel economy of hybrid electric vehicles," in *Proc. 23rd Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2018, pp. 196–202.
[25] R. Liessner, C. Schroer, A. Dietermann, and B. Bäker, "Deep reinforcement learning for advanced energy management of hybrid electric vehicles," in *Proc. 10th Int. Conf. Agents Artif. Intell.*, 2018, pp. 61–72.
[26] Y. Yang, X. Hu, H. Pei, and Z. Peng, "Comparison of power-split and parallel hybrid powertrain architectures with a single electric machine: Dynamic programming approach," *Appl. Energy*, vol. 168, pp. 683–690, Apr. 2016.
[27] L. Lai and M. Ehsani, "Dynamic programming optimized constrained engine on and off control strategy for parallel HEV," in *Proc. IEEE Vehicle Power Propuls. Conf. (VPPC)*, Oct. 2013, pp. 1–5.
[28] 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.
[29] C.-C. Lin, H. Peng, and J. W. Grizzle, "A stochastic control strategy for hybrid electric vehicles," in *Proc. Amer. Control Conf.*, Jan. 2004, pp. 4710–4715.
[30] J. Liu and H. Peng, "Modeling and control of a power-split hybrid vehicle," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 6, pp. 1242–1251, Nov. 2008.
[31] T. Leroy, F. Vidal-Naquet, and P. Tona, "Stochastic dynamic programming based energy management of HEV's: An experimental validation," *IFAC Proc. Volumes*, vol. 47, no. 3, pp. 4813–4818, 2014.
[32] C. Vagg, S. Akehurst, C. J. Brace, and L. Ash, "Stochastic dynamic programming in the real-world control of hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 3, pp. 853–866, May 2016.
[33] 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.
[34] Z. Chen, H. Hu, Y. Wu, R. Xiao, J. Shen, and Y. Liu, "Energy management for a power-split plug-in hybrid electric vehicle based on reinforcement learning," *Appl. Sci.*, vol. 8, no. 12, p. 2494, 2018.
[35] T. Liu, Y. Zou, D. Liu, and F. 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.
[36] H. Lee, "Stochastic optimal energy management based on Q-learning for hybrid electric vehicles," Ph.D. dissertation, Dept. Mech. Aero. Eng., Seoul Nat. Univ., Seoul, South Korea, 2018.
[37] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

**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. He is currently 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.



**CHANGHEE SONG** received the B.S. degree in mechanical engineering from Yeonsei University, Seoul, South Korea, in 2014. He is currently pursuing the Ph.D. degree in mechanical engineering with the Graduate School, Seoul National University. He is interested in energy management for a hybrid electric vehicle and artificial intelligence. His graduation thesis is focused on the energy management for hybrid electric vehicles with deep reinforcement learning.

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

**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*.