IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 20, 2020, accepted November 4, 2020, date of publication November 9, 2020,
date of current version November 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036719

Model-Based Reinforcement Learning for Eco-Driving Control of Electric Vehicles

HEEYUN LEE[1], (Member, IEEE), NAMWOOK KIM[2], AND SUK WON CHA[1,3], (Member, IEEE)

[1]Department of Mechanical Engineering, Seoul National University, Seoul 08826, South Korea
[2]Department of Mechanical Engineering, Hanyang University, Ansan 15588, South Korea
[3]Institute of Advanced Machines and Design, Seoul National University, Seoul 08826, 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 National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant NRF-2019R1A4A1025848, and in part by the Technology Innovation Program (Development of Application Technologies For Heavy Duty Fuel Cell Electric Trucks Using Multi-Input Motor-Based 400kW Class Multi Speed Electrified Powertrain System) funded by the Ministry of Trade, Industry and Energy (MOTIE, South Korea) under Grant 20011834.

ABSTRACT With the development of autonomous vehicles, research on energy-efficient eco-driving is becoming increasingly important. The optimal control problem of determining the speed profile of the vehicle for minimizing energy consumption is a challenging problem that necessitates the consideration of various aspects, such as the vehicle energy consumption, slope of the road, and driving environment, e.g., the traffic and other vehicles on the road. In this study, an approach using reinforcement learning was applied to the eco-driving problem for electric vehicles considering road slopes. A novel model-based reinforcement learning algorithm for eco-driving was developed, which separates the vehicle's energy consumption approximation model and driving environment model. Thus, the domain knowledge of vehicle dynamics and the powertrain system is utilized for the reinforcement learning process, while model-free characteristics are maintained by updating the approximation model using experience replay. The proposed algorithm was tested via a vehicle simulation and compared with a solution obtained using dynamic programming (DP), and as well as conventional cruise control driving with constant speed. The simulation results indicated that the speed profile optimized using model-based reinforcement learning had similar behavior to the global solution obtained via DP and energy saving performance compared with cruise control.

INDEX TERMS Eco-driving control, electric vehicles, model-based reinforcement learning, optimal control, Q-learning, reinforcement learning.

I. INTRODUCTION

Recently, diverse technologies for autonomous vehicles have been developing rapidly, which has led to advancements in autonomous driving. In the future, vehicles can be operated with less intervention by human drivers. Without manipulation by the human driver, the vehicle becomes safer; additionally, vehicles will be able to move quickly with the aid of computational intelligence based on autonomous vehicle technologies in the near future. Another issue concerning future vehicles is the environmental aspect; diverse vehicles, such as hybrid electric vehicles (HEVs), electric vehicles (EVs), and fuel-cell EVs (FCEVs), are being developed to reduce emissions and increase the vehicular efficiency. The use of autonomous vehicles can also contribute to increasing the vehicular fuel efficiency. In an autonomous vehicle, when the level of driving automation increases, the intervention of the human driver can be minimized. The efficiency of these vehicles can be maximized while satisfying the desired travel time. This optimization of the vehicle speed profile can be very useful, as the vehicle efficiency can be increased without changes in the vehicle hardware, and this technology can be used in any type of vehicle. Additionally, considering that in the near future, many vehicles can be operated without a human driver, optimization of the vehicle speed profile, which is called an eco-driving strategy, is a very important problem.

Various studies have been conducted on eco-driving strategies. First, approaches based on an analytical solution derived from the optimal control problem have been proposed. In [1],

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco Perez-Pinal.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

a closed-form solution of the optimal problem was found for eco-driving of EVs. Here, the optimization problem was defined to minimize the fuel consumption, and the target traveling time for a given distance was given as the constraint of the problem. Then, the optimization problem was solved to obtain an explicit solution. In [2], an analytical state-constrained solution was derived considering vehicle safety constraints for EVs. Here, the minimum inter-vehicle distance and maximum road speed limit were defined as state constraints, and an analytical state constrained solution was derived for connected and automated vehicles.

Additionally, in many studies, approaches based on dynamic programming (DP) or Pontryagin's minimum principle (PMP) were utilized. In [3], look-ahead control was used to optimize the speed profile. Here, based on a global positioning system, the road geometry ahead of the vehicle was extracted, and DP was used in a predictive scheme to optimize the velocity trajectory for a heavy diesel truck. In [4], stochastic approaches based on the DP were employed. Here, a time-independent fuel-efficient control strategy based on stochastic DP was developed, which does not require preview information of the route or the road slope. Additionally, constraints on the vehicle-following distance are applied to develop a fuel-efficient vehicle-following control policy. In [5], PMP was applied to a passenger car with an internal combustion engine vehicle. Here, optimal periodic control was derived for cruise control, which is a hybrid system that includes gear shift and idle operation of the engine. In [6], the minimum fuel driving control was studied according to PMP. Here, the vehicle model was expressed as a point-mass vehicle with a quasi-static polynomial fuel-consumption model, and gear shifting, clutch disengagement, and brake control were modeled as simple on–off switches. More recently, in [7], PMP and DP were used together for the eco-driving of all-EVs. Here, PMP was first utilized to find the possible operating mode satisfying the necessary condition, and then, DP was used to solve the optimal control problem again in the distance domain, which reduced the computational burden of the DP calculation.

In [8], the traffic signal was included in the eco-driving control framework. Here, with the assumption of vehicle-to-infrastructure communication capabilities, an optimal speed profile was obtained to minimize the total fuel consumption while safely crossing an intersection. Additionally, combined with the energy management of HEVs, the speed profile control problem was defined in an all-inclusive manner in [9]. Here, a bi-level methodology was used for the predictive energy management of parallel HEVs, where the optimal velocity was calculated first in the outer loop using a Krylov subspace method, and in the inner loop, the optimal torque split and gear shift were determined using PMP based on the model predictive control (MPC) framework.

However, applying these eco-driving strategies to real-world driving situations is not easy and has many limitations. First, the environment changes frequently and has many disturbances. Thus, the deterministic algorithm has limitations in that it must predict future driving conditions precisely, or there are driving environments that are difficult to model, such as the driving behavior of the car ahead or a traffic jam. Additionally, implementing DP or PMP for an online eco-driving strategy is challenging because of the computational burden of DP or the co-state sensitive characteristics of PMP. The more practical method of MPC was used in [10], and [11]. Here, adaptive nonlinear MPC was utilized, and it was implemented in a vehicle with a standard production powertrain control module. To increase the prediction accuracy, a recursive least-squares algorithm was used for parameter adaptation, which was combined with MPC to obtain more reliable results under real-world driving conditions. In [12], the vehicle-following scenario was studied. In the automated car-following scenario, the pulse-and-gliding strategy was implemented based on the switching logic in a servo-loop controller to minimize the fuel consumption. However, these approaches also have limitations in that MPC and periodic control are focused on finding the local optimal for the near future, rather than the global optimal solution with entire travel distances. Thus, the fuel-economy improvement is limited, and consideration of complex driving environments is challenging, requiring an additional parameter calibration process.

Therefore, in this study, we conducted an eco-driving strategy based on reinforcement learning. Reinforcement learning is an algorithm that can learn the optimal control policy according to the interaction between the agent and the environment [13]. Reinforcement learning is very similar approach to the DP-based approach in that they can optimize the cost-to-go value function based on the Bellman equation, and it is possible to replace this DP-based approach with reinforcement learning-based approach. On the other hand, unlike DP, reinforcement learning can be used as a real time controller through learning in a stochastic manner, and it has a model-free feature by learning the optimal control policy through the interaction between the agent and the environment with adaptation. Accordingly, reinforcement learning approach is well suited to the eco-driving control problem in which an optimization solution must be found through a probabilistic point of view in various and complex road driving environments. Reinforcement learning has been used for eco-driving in several studies. In [14], multi-objective deep Q-learning was utilized for the eco-routing problem to identify the best route for minimizing the traveling time and fuel consumption. In [15], and [16], a reinforcement learning algorithm was studied for minimizing the fuel consumption in the vicinity of an isolated signal intersection. In [17], eco-driving control considering the car-following scenario using an actor–gear–critic network architecture was studied for a conventional vehicle equipped with an internal combustion engine and automated manual transmission. Here, a fuel economy with safe inter-vehicle distance constraints was considered as an objective function, but the road slope was not considered as a state variable and was set to zero.

In the present study, general speed profile optimization for an eco-driving strategy for longitudinal driving considering the road slope using model-based reinforcement learning (MBRL) was investigated. MBRL is a methodology that approximates the environment, including the system dynamics; thus, learning can be conducted with guaranteed stability [18], or few interactions [19]. In the case of vehicle control, MBRL was successfully applied to the optimal control problem of energy management of HEVs in our previous studies [20], [21]. The contribution of the present study is as follows: We developed a new algorithm for the eco-driving control problem using the reinforcement learning approach, and through this, we confirmed that the reinforcement learning method can be well applied to the eco-driving problem. In particular, in the eco-driving problem through optimization of the vehicle's speed profile reflecting the road slope, the reinforcement learning method was compared with the optimal solution using the existing DP method and the cruise control case with constant vehicle speed, demonstrating the excellence and feasibility of the reinforcement learning based approach. Especially, we developed an eco-driving strategy with model-based Q-learning and confirmed its effectiveness via a vehicle simulation. To the best of our knowledge, this was the first study in which the MBRL approach was applied to the eco-driving control problem. Even though only the road slope is considered among diverse driving environments for the eco-driving strategy, considering that the proposed approaches can be extended to diverse driving environment conditions, e.g., traffic signals and other vehicles on the road, thanks to the model-free characteristic of the algorithm, the approaches using the reinforcement learning technique can be powerful. Additionally, the trained optimal control policy can be used for real-time vehicle controllers. The remainder of this article is organized as follows. In Section II, the EV model used in this study is presented. In Section III, the optimization problem for the eco-driving strategy is presented, and the MBRL algorithm for eco-driving is explained. In Section IV, the vehicle simulation is presented, and in Section V, the conclusions are presented.

II. VEHICLE MODELING

In this study, a vehicle simulation was performed for training and testing the proposed algorithm. For the simulation, an EV was used. Compared with conventional internal combustion engine-based vehicles, EVs can recover energy from regenerative braking, making them more suitable for energy-efficient driving. However, the algorithm proposed in this article is not limited to EVs but is applicable to all vehicles.

For the EV modeling, a backward-looking vehicle simulation was performed via a quasi-static modeling technique, and only longitudinal vehicle dynamics are considered. The vehicle configuration is shown in Fig. 1, and the vehicle parameters used in the simulation are presented in Table 1. The efficiency of the motor including the efficiency of the converter $\eta_{elec}(T_{mot}, \omega_{mot})$, was calculated using a predetermined map, as shown in Fig. 2, and the electric power

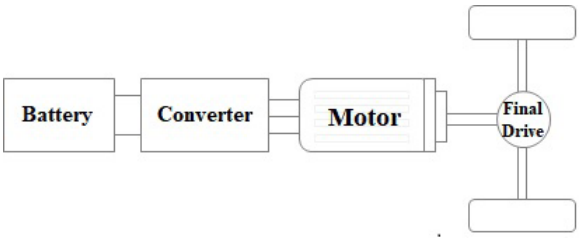

FIGURE 1. Simulation model of EVs.

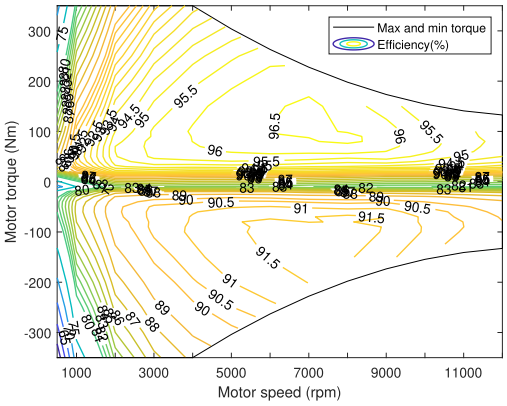

FIGURE 2. Efficiency of the motor, including the converter efficiency.

TABLE 1. Vehicle model parameters.

Parameter	Value
Vehicle mass	1800 kg
Wheel radius	0.322 m
Final gear ratio	9.5
Electric Motor	Permanent-magnet synchronous motor, Maximum torque: 350 Nm
Battery Capacity	Lithium-ion polymer, Capacity : 120 Ah
Driving coefficient	f_0=140 N f_1=-0.5N/km f_2=0.04N/km^2

consumed by the motor p_{bat}, was calculated using the following equation:

$$p_{bat} = \eta_{elec}^{-sgn(T_{mot})} \cdot T_{mot}\omega_{mot} \tag{1}$$

where T_{mot} represents the motor torque, and ω_{mot} represents the motor speed. The battery state of charge (SOC) dynamics can be expressed as follows:

$$\dot{SOC} = -\frac{V_{OC} - \sqrt{V_{OC}^2 - 4R_{bat}P_{bat}}}{2R_{bat}Q_{bat}} \tag{2}$$

where $V_{oc}(SOC)$ represents the open-circuit voltage, $R_{bat}(SOC)$ represents the internal resistance of the battery, and Q_{bat} represents the battery capacitance. The powertrain dynamic is given as follows:

$$T_{wh} = (T_{mot} - T_{fd,loss}) \cdot \gamma_{fd} \tag{3}$$

where T_{wh} represents the wheel torque, γ_{fd} represents the gear ratio of the final drive, $T_{fd,loss}(T_{fd}, \omega_{fd})$ represents the torque loss in the final drive, T_{fd} represents the input torque in the

final drive, and ω_{fd} represents the input speed in the final drive, which can be expressed as follows:

$$\omega_{fd} = \omega_{mot} = \frac{v \cdot \gamma_{fd}}{R_{tire}} \tag{4}$$

where R_{tire} represents the tire radius, and v represents vehicle's longitudinal speed. The vehicle dynamic is given as follows:

$$\dot{v} = \frac{T_{wh}/R_{tire} - F_{brake} - F_{load}}{M_{veh} + M_{eq}} \tag{5}$$

where F_{brake} represents the brake force, M_{veh} represents the vehicle mass, M_{eq} represents the equivalent mass of the rotating inertia in the vehicle component, and F_{load} represents the road load force, including the grading resistance, which can be expressed as follows:

$$F_{load} = f_0 + f_1 \cdot v + f_2 \cdot v^2 + M_{veh} g \sin\theta \tag{6}$$

where f_0, f_1, and f_2 are the road load coefficients, which have the units of N, N/km, and N/km^2, respectively, and θ represents the road slope. Using this vehicle model, a vehicle simulation was conducted to train and test the proposed algorithm, as explained in the following section.

III. MODEL-BASED REINFORCEMENT LEARNING FOR ECO-DRIVING STRATEGY

A. OPTIMAL CONTROL PROBLEM FORMULATION

The optimal control problem for an eco-driving strategy can be defined to minimize the battery electric energy consumption for state vector consists of the vehicle speed and the traveling distance $x = \{v, d\}$, while driving a given distance D for a given time T as follows:

$$\begin{aligned}
&\min \left(\int_0^T \dot{SOC}\left(v(t), u(t)\right) dt \right) \\
&subj.to\ \dot{v} = f(v(t), u(t)) \\
&\qquad \dot{d} = v \\
&\qquad v(t) \in [v_{min}(t), v_{max}(t)] \\
&\qquad u(t) \in [T_{mot,min}(\omega_{mot}(t)), T_{mot,max}(\omega_{mot}(t))] \\
&\qquad d(0) = 0, \quad d(T) = D \\
&\qquad v(0) = v_0, \quad v(T) = v_f
\end{aligned} \tag{7}$$

where f represents the nonlinear vehicle dynamics explained in the previous section, u represents the control variable, which is motor torque with the maximum torque $T_{mot,max}$ and the minimum torque $T_{mot,min}$. v_0 and v_f represent the vehicle initial speed, and the vehicle final speed at D respectively, which have the minimum and maximum speed of v_{min}, and v_{max}.

To simplify the optimal control problem, it can be expressed as one state formulation using $dt = dd/v$ as in [22], which is the weighted sum of the battery SOC usage and the traveling time according to the distance, as follows:

$$\begin{aligned}
&\min \left(\int_0^D \frac{\dot{SOC}}{v(d)} dd + \omega \left[\int_0^D \frac{1}{v(d)} dd - T \right] \right) \\
&subj.to\ \dot{v} = f(v(d), u(d)) \\
&\qquad v(d) \in [v_{min}(d), v_{max}(d)] \\
&\qquad u(d) \in [T_{mot,min}(\omega_{mot}(d)), T_{mot,max}(\omega_{mot}(d))] \\
&\qquad v(0) = v_0 \\
&\qquad v(D) = v_f
\end{aligned} \tag{8}$$

where ω represents the weighting factor to be tuned for satisfying the total driving time. By transferring the optimal control problem from (7) to (8), the problem is defined according to the traveling distance d, while the vehicle's initial and final speed constraints remain the same. As mentioned in the section I, the traffic-signal and car-following situations are not considered.

B. DETERMINISTIC DP

First, deterministic DP is used to solve the optimal control problem. The foregoing optimal control problem can be presented in discrete form as follows:

$$\begin{aligned}
&\min \left(\sum_{k=0}^{N-1} L\left(v(k), u(k)\right) \right) \\
&subj.to\ \dot{v} = f(v(k), u(k)) \\
&\qquad v(k) \in [v_{min}(k), v_{max}(k)] \\
&\qquad u(k) \in [T_{mot,min}(\omega_{mot}(k)), T_{mot,max}(\omega_{mot}(k))] \\
&\qquad v(0) = v_0 \\
&\qquad v(N) = v_f
\end{aligned} \tag{9}$$

where the index k represents the discretization step for N segments, which is equally divided with unit distance Δs, and $L(v(k), u(k))$ represents the instantaneous cost incurred, which can be expressed as follows:

$$L\left(v(k), u(k)\right)) = \Delta SOC(k) + \omega \cdot \Delta time(k) \tag{10}$$

$$\Delta SOC(k) = \dot{SOC}(k) \cdot \frac{2\Delta s}{v(k) + v(k+1)} \tag{11}$$

$$\Delta time(k) = \frac{2\Delta s}{v(k) + v(k+1)} \tag{12}$$

Then, the optimal solution can be obtained using the Bellman equation [23], as follows.

$$J^*_{k,N}(v(k)) = \min \left\{ L(v(k), u(k)) + J^*_{k+1,N}(v(k+1)) \right\} \tag{13}$$

Here, $J_{k,N}$ represents the cost function for traveling from step k to N, which can be expressed in the recursive form using the instantaneous cost $L(v(k), u(k))$, and the cost function for traveling from step $k+1$ to N, $J_{k+1,N}$. Using (13), an optimal speed profile based on the travel distance can be obtained, but deterministic DP is computationally inefficient and difficult to adapt under driving-condition changes. Thus, it has many limitations when used in real-time vehicle controllers.

C. MODEL-BASED REINFORCEMENT LEARNING

First, the optimal control problem can be expressed to minimize the expected total cost over an infinite horizon, as follows:

$$\min\left(J_\pi(x_0) = \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1}\gamma^k g(x_k, \pi(x_k))\right\}\right) \quad (14)$$

where $J_\pi(x_0)$ represents the cost with initial condition and control policy π, γ represents the discount factor, and g represents the instantaneous cost, which can be expressed as follows:

$$g_k = \Delta SOC(k) + \omega \cdot \Delta time(k) + \eta(v_k) \quad (15)$$

Here, $\eta(v_k)$ represents the penalty cost that is applied when the vehicle speed is higher than v_{max} or lower than v_{min}, as follows:

$$\eta(v_k) = \begin{cases} 0 & if\ v_{min}(k) \le v_k \le v_{max}(k) \\ cost_{penalty} & else \end{cases} \quad (16)$$

Here, $cost_{penalty}$ is a positive constant. The optimal control problem is defined in an infinite horizon; thus, the generated control policy is time-invariant, which can be easily implemented on a real-time vehicle controller. The state variable x is defined as follows:

$$x = \{v, h, \theta\} \quad (17)$$

where h represents the height, and θ represents the road slope. x is discretized as follows:

$$v \in \{v^1, v^2, v^3, \ldots, v^{N_v}\} \quad (18)$$
$$h \in \{h^1, h^2, h^3, \ldots, h^{N_h}\} \quad (19)$$
$$\theta \in \{\theta^1, \theta^2, \theta^3, \ldots, \theta^{N_\theta}\} \quad (20)$$

where N_v, N_h, and N_θ represent the number of the discretized speed, height and road slope respectively, and the control variable is also discretized as follows:

$$u \in \{u^1, u^2, u^3, \ldots, u^{N_u}\} \quad (21)$$

where N_u represent the number of the discretized control input.

Therefore, in this study, the eco-driving control policy was determined according to the current vehicle speed, height, and road slope. Among them, vehicle speed or road slope directly affect the cost values. However, in the case of height, it does not directly affect the cost value instantly, but it reflects the future cost concerning the road driving environment. Combined with the road slope, height can be expressed as a state of the Q function indicating the expected total cost of future energy use. That is, even in the same uphill situation, the optimal driving speed of the vehicle may vary according to the current height, and this is obvious when considering the relationship between the kinetic energy of the vehicle and the potential energy, and the energy consumption of driving the vehicle accordingly. Therefore, by considering the height as a state variable with the vehicle speed and the road slope, it is possible to better represent the value of the cost-to-go function in Q function and a probabilistic driving situation than when the road slope is considered only.

Based on Q-learning [24], The optimal cost $J^*(x_k)$ and optimal control policy $\pi^*(x_k)$ can be expressed as follows:

$$J^*(x_k) = \min_u(Q^*(x_k, u)) \quad (22)$$
$$\pi^*(x_k) = \arg\min_u(Q^*(x_k, u)) \quad (23)$$

Then, the Q function can be updated as follows:

$$Q(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha\left(g_k + \gamma \min_u Q(x_{k+1}, u) - Q(x_k, u_k)\right) \quad (24)$$

In this study, to solve the optimal control problem, a novel eco-driving strategy utilizing MBRL was developed on the basis of a previous study on MBRL for the HEV control case study in [21]. In (17), the state variable $x_k = [v_k, h_k, \theta_k]$ is partially stochastic (the driving environment h_k and θ_k can be considered stochastic without preview terrain information), but it is possible to expect v_k, and g_k deterministically based on the vehicle powertrain dynamics equations of (1)–(6), and the cost equation (15) for the given driving conditions of h_k, and θ_k and the given control input u. Thus, the domain knowledge of the known vehicle dynamic model and powertrain system can be used in reinforcement learning, while the remaining model uncertainty due to modeling error or other driving environments that are difficult to model can be learning via model-free still.

On the basis of this observation, we developed a new MBRL algorithm for the eco-driving strategy. The overall algorithm is presented in Fig. 3 and Algorithm 1. In the new algorithm, the agent's learning takes place based on approximation model using the deterministic variable, while stochastic variables are reflected in the agent's learning through the experience replay. Usually, in the Q-learning, the agent derives the action u_k using methods such as ϵ-greedy (exploitation and exploration) according to the Q function value and the current state x_k, and conducts learning by updating Q function value using the observation of the reward g_k and the next state x_{k+1}. Alternatively, in the new

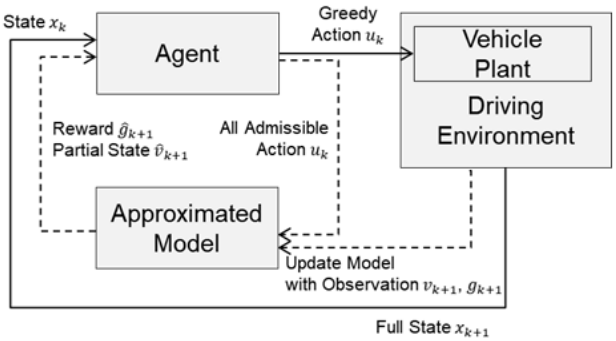

FIGURE 3. Model-based reinforcement learning algorithm for eco-driving.

Algorithm 1 Model-Based Reinforcement Learning Algorithm for Eco-Driving

Input: data x_k, size N
repeat
Observe $x_k = [v_k, h_k, \theta_k]$
Choose the greedy action
$u_k = \arg\min_u Q^*(x_k, u)$
Observe reward g_k, and state x_{k+1}
Update approximation model

$$\hat{g}(x|u) \leftarrow \hat{g}(x|u) + \alpha_g(g_k(x_k|u_k) - \hat{g}(x|u))$$
$$\hat{v}(x|u) \leftarrow \hat{v}(x|u) + \alpha_v(v_{k+1}(x_k|u_k) - \hat{v}(x|u))$$

Update Q using approximation model
for $m = 1$ **to** N_v **do**
for $l = 1$ **to** N_u **do**

$$Q \leftarrow (1-\alpha)Q(x_k = [v^m, h_k, \theta_k], u^l) + \alpha\left(\hat{g}_k + \gamma \min_u Q(\hat{x}_{k+1} = [\hat{v}_{k+1}, h_{k+1}, \theta_{k+1}], u)\right)$$

end for
end for
$k \leftarrow k + 1$
until Simulation stop

algorithm, the agent derives the greedy control input using the Q function and observes the reward and the next variable (exploitation), but observed information is used to make an approximation model, and by using it, various control inputs are tested for a given stochastic state transition (exploration). In other words, according to the experience of the driving environment of h_k and θ_k, learning with experience replay is conducted to optimize the Q function value as shown in "for" loop in the Algorithm 1. With estimation of $\hat{v}_{k+1}$ and estimation of the reward $\hat{g}_k$, the Q function value can be updated for different vehicle speeds $v \in \{v^1, v^2, v^3, \ldots, v^{N_v}\}$ and control inputs $u \in \{u^1, u^2, u^3, \ldots, u^{N_u}\}$, as follows:

$$Q \leftarrow (1-\alpha)Q(x_k = [v, h_k, \theta_k], u) + \alpha\left(\hat{g_k} + \gamma \min_u Q(\hat{x}_{k+1} = [\hat{v}_{k+1}, h_{k+1}, \theta_{k+1}], u)\right) \quad (25)$$

where $\hat{g}_k$, and $\hat{v}_{k+1}$ can be determined based on the vehicle powertrain model. Alternatively, $\hat{g}_k$, and $\hat{v}_{k+1}$ can be determined based on approximation model for keeping model-free characteristic of reinforcement learning (see [21]), that approximation could be done based on the experience as shown in following equation:

$$\hat{g}(x|u) \leftarrow \hat{g}(x|u) + \alpha_g(g_k(x_k|u_k) - \hat{g}(x|u)) \quad (26)$$
$$\hat{v}(x|u) \leftarrow \hat{v}(x|u) + \alpha_v(v_{k+1}(x_k|u_k) - \hat{v}(x|u)) \quad (27)$$

where α_g and α_v represent the learning rates. According to the experience of the stochastic state transition from (h_k, θ_k) to (h_{k+1}, θ_{k+1}), the deterministic state transition from v_k to $\hat{v}_{k+1}$ and reward $\hat{g}_k$ can be estimated to optimize the Q function value. In this study, to simplify the problem, the control input u was defined as the relative offset of the vehicle speed instead of the motor torque, as follows:

$$u \in \{-10\delta, -9\delta, -8\delta, \ldots, 0, \ldots, +8\delta, +9\delta, +10\delta\} \quad (28)$$

Here, δ represents the unit speed for discretization in (18). Then, the estimation of $\hat{v}_{k+1}$ is not required, as $\hat{v}_{k+1}$ is determined u directly, and the control policy can be expressed in a more intuitive form in the direction of reducing or increasing the speed of the vehicle in several steps.

The feature of this algorithm is to separate the model-based insight of the vehicle powertrain, which can be estimated relatively well, from various driving situations that are difficult to estimate, so that the control policy is extracted more effectively using reinforcement learning. Additionally, the proposed algorithm has the advantage of experience replay, which accelerates the convergence and enhances the stability. Furthermore, the control u can be tested in the nested "for" loop of the experience replay process before it is used in the real greedy action; thus, irrelevant control inputs can be excluded to prevent fatal system errors or optimize the controller performance. Additionally, to reduce the computational burden, the "for" loop in the algorithm can be alternated via prioritized sweeping without searching all the v and u values.

IV. VEHICLE SIMULATION

A simulation based on the vehicle model described in Section II was performed. For the vehicle simulation and development of the reinforcement learning algorithm, MATLAB was used. Information regarding the driving environment, i.e.,h_k and θ_k, was recorded during real-world driving and used in the simulation. The height and road slope profiles are shown in Fig. 4. The distance of the diving cycle was 10 km, and it was divided into equal intervals of 10 m.

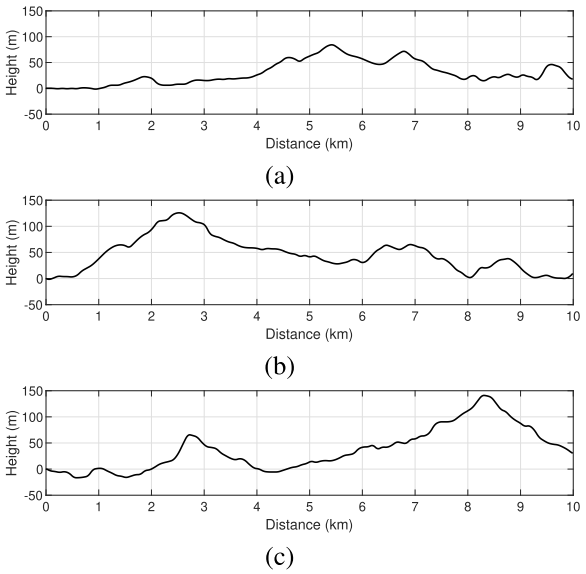

FIGURE 4. Driving environment: (a) driving cycle A; (b) driving cycle B; (c) driving cycle C.

The slope was assumed to be a piecewise constant. The parameters used in the reinforcement learning algorithm are presented in Table 2. For discretization, the nearest-neighbor method was used. The vehicle speed was discretized by 1 km/h from 0 to 100 km/h, and height was discretized by 5 m. The road slope was discretized by 1%. For the weighting factor, ω, which should be defined to satisfy the desired traveling time for a given driving environment, was assumed to be 0.004. For the battery SOC, the initial value was defined as 70% and the initial vehicle speed was defined as 60 km/h. The initial value of the approximation model was defined roughly using (1)–(6) and was updated during the learning process.

TABLE 2. Parameters for learning.

Parameter	Value
Learning rate, α	0.05
Discounted factor , γ	0.9995
Model learning rate, α_g	0.001
Model learning rate, α_v	0.001

A. LEARNING CURVE AND CONTROL POLICY

Using the proposed algorithm, a learning process was conducted to determine the energy-efficient speed trajectory using driving cycles A, B, and C, separately. With an initial speed of 60 km/h, the vehicle speed was generated for each driving cycle, and learning was performed 500 times. The learning curve resulting from the learning process utilizing driving cycle A is presented in Fig. 5. As shown, the sum of the instantaneous cost in (15) rapidly decreased and converged as the learning was repeated, indicating that the learning process was successful. The resulting speed trajectories and height with respect to the traveling distance, as well as the motor torque profiles for all the driving cycles, are presented in Fig. 6. Generally, the vehicle speed decreased when the vehicle traveled uphill and increased when the vehicle traveled downhill. Using the proposed algorithm, the optimal velocity profile utilizing the slope and height information of the terrain was learned. As an example, the control policy extracted from the Q function using driving cycle A is presented in Fig. 7, where the optimal speed command for either an increasing or decreasing speed is shown according to vehicle's current speed, and slope. Here, the trend of the control according to the current speed and slope value was confirmed: for a low speed, the control policy increased the speed actively, and higher slope values tended to reduce the

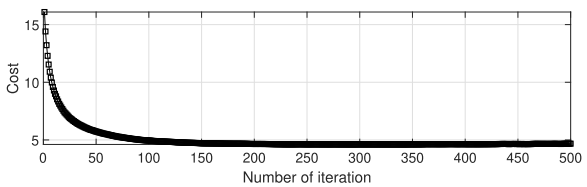

FIGURE 5. Learning curve for driving cycle A.

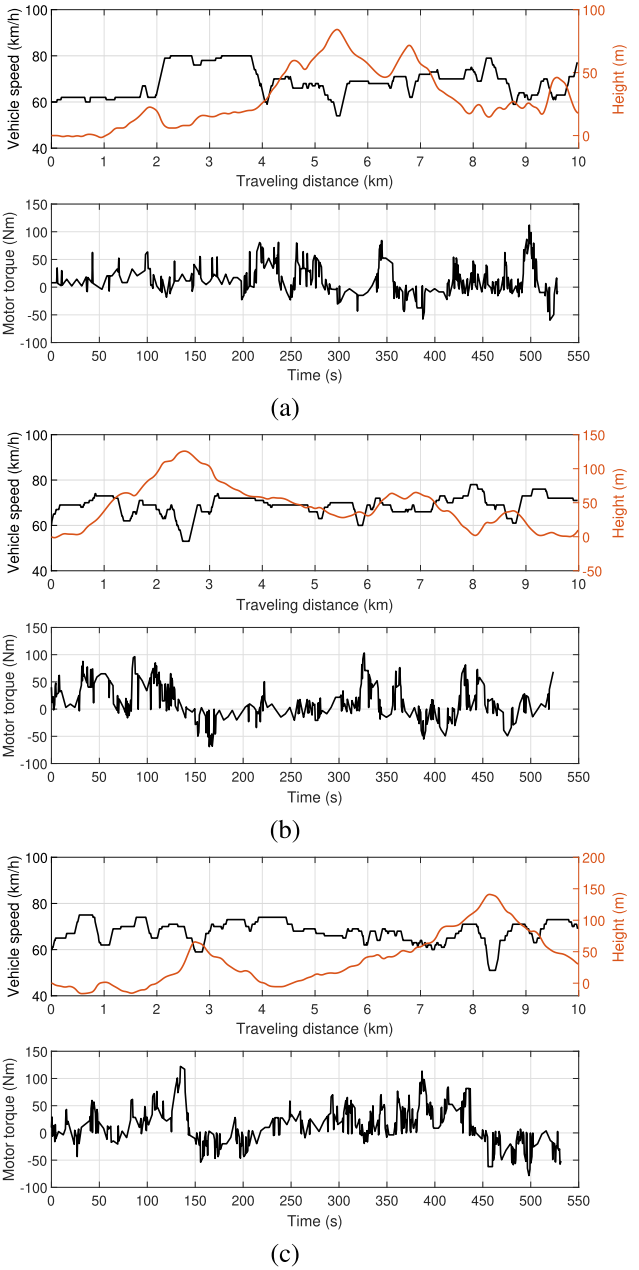

FIGURE 6. Optimized speed trajectory and height with respect to the traveling distance, as well as the motor torque. Results are presented for different driving cycles: (a) A; (b) B; (c) C.

speed. In all three cycles, the vehicle's speed was maintained between approximately 60 and 80 km/h. This is because the same weighting coefficient ω was used, and in the same way, the tendency to control with the target speed section can be confirmed in the extracted control policy.

B. COMPARISON WITH DETERMINISTIC DP RESULT AND CRUISE CONTROL RESULT

The energy saving performance of the given speed trajectory was evaluated by comparing it with the result of DP, as well as general cruise control in which the vehicle is driven at a

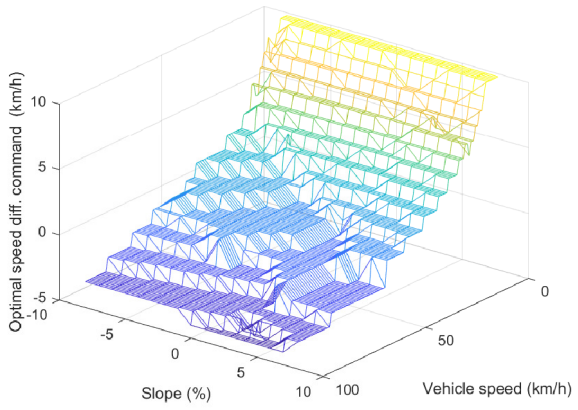

FIGURE 7. Control policy extracted from driving cycle A for a height of 50 *m*.

constant speed. As mentioned in Section III, DP can present the global optimal solution; thus, the solution of DP can be used as a benchmark. However, in contrast to DP, in which the initial and final condition of the state can be defined, in the proposed MBRL algorithm, the final speed cannot be determined in advance. Thus, according to the MBRL result, the initial and final speed results (v_0 and v_f, respectively) were set as constraints in DP, to compare the energy saving performance fairly by making the remaining kinetic energy of the vehicle identical between the two cases. For cruise control, the vehicle is driven at an average speed equal to the MBRL result v_{ave}, and the initial and final speed constraints are applied.

The simulation results are presented in Table 3, and Fig. 8. Table 3 presents the traveling time, SOC usage, and percent energy saving of SOC usage with respect to the cruise control result. The traveling time was an important factor, as driving the same distance slower tended to use less battery SOC. In the simulation, the traveling time results for DP, MBRL, and cruise control were close, with a maximum difference of 0.6%, which is negligible. With regard to the battery SOC use, DP was the most efficient of the three approaches,

TABLE 3. Simulation result for DP, MBRL, and cruise control.

Driving Cycle A ($v_0 = 60\ km/h$, $v_f = 77\ km/h$, $v_{ave} = 69\ km/h$)			
	DP	MBRL	Cruise
Traveling time (s)	528.5	527.9	528.7
△ SOC(%)	2.49	2.56	2.59
Energy saving (%)	3.9	1.2	-
Driving Cycle B ($v_0 = 60\ km/h$, $v_f = 71\ km/h$, $v_{ave} = 68\ km/h$)			
	DP	MBRL	Cruise
Traveling time (s)	523.6	523.5	521.0
△ SOC(%)	2.44	2.46	2.54
Energy saving (%)	3.7	3.0	-
Driving Cycle C ($v_0 = 60\ km/h$, $v_f = 69\ km/h$, $v_{ave} = 69\ km/h$)			
	DP	MBRL	Cruise
Traveling time (s)	528.5	531.6	528.9
△ SOC(%)	2.69	2.71	2.76
Energy saving (%)	2.6	1.8	-

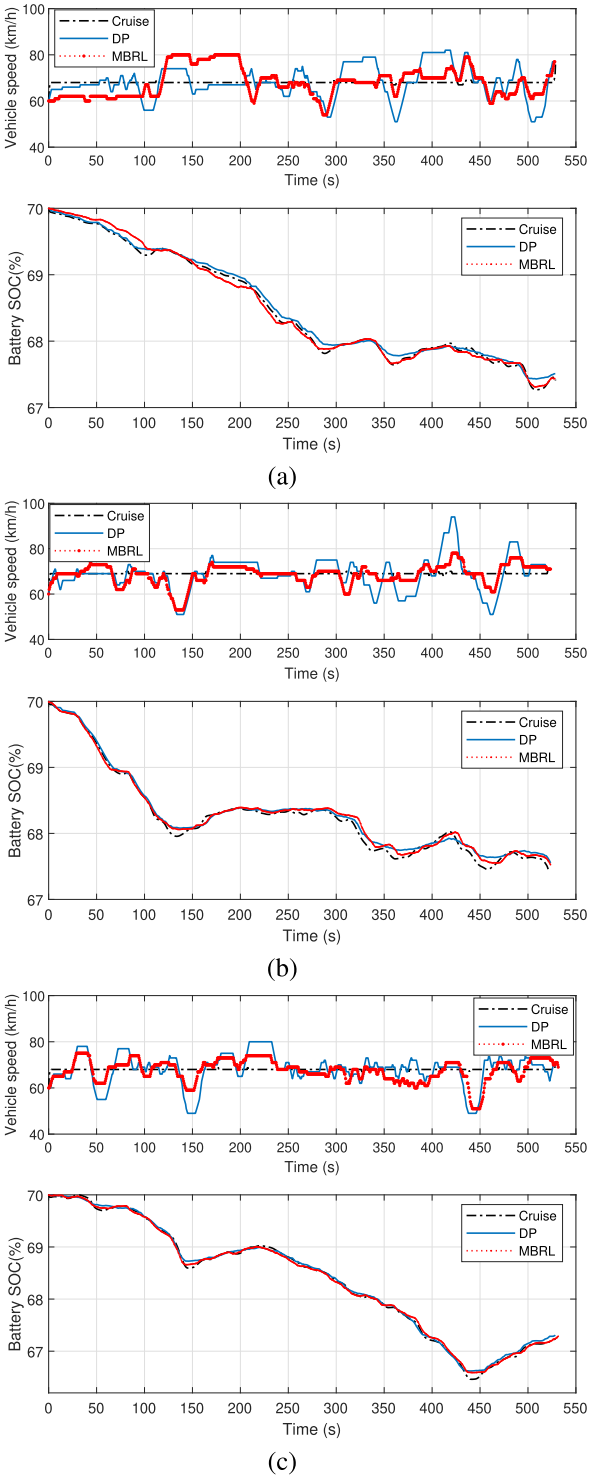

FIGURE 8. Optimized speed trajectory and battery SOC for cruise control, DP, and MBRL: (a) driving cycle A; (b) driving cycle B; (c) driving cycle C.

followed by MBRL. Compared with cruise control, DP and MBRL exhibited average energy saving of 3.4% and 2.0%, respectively. Fig. 8 shows the speed profiles and battery SOC trajectories. The speed profiles obtained from MBRL were similar to those obtained from DP, as expected from the energy saving performance results. However, there was a

difference between the DP and MBRL results, even though training was successfully conducted using the entire driving cycle information in MBRL. This can be explained by the problem definition: in MBRL, the optimal control problem is defined as minimizing the expected total cost in an infinite horizon, whereas in DP, it is defined in a finite horizon. This results in a higher energy saving for DP, while the control policy of MBRL can be used as an offline real-time controller in a stochastic manner. However, for various driving environment scenarios, MBRL shows good performance, because MBRL has learned the optimal behavior well using the transition probability of the vehicle's driving environment $\{h_k, \theta_k\}$ to $\{h_{k+1}, \theta_{k+1}\}$ through the learning process. Unlike deterministic DP in which entire driving environment information should be given in advance, or stochastic DP in which a model for the transition probability matrix of a driving environment should be given in advance [25], in reinforcement learning, the agent brings the transition probability distribution of the driving environment to the optimization problem through the interaction with the driving environment. Therefore, it is possible to derive optimal control through learning based on model-free characteristics.

C. PERFORMANCE FOR LEARNING WITH COMBINED DRIVING CYCLES

The performance of the algorithm when the learning process was conducted with various driving cycles was evaluated. Here, all the driving cycles (A, B, and C) were utilized for the training process, and the resulting Q function was employed as an offline control policy for simulation using each driving cycle. Thus, by testing the control policy obtained from the learning process using combined driving cycles, we checked whether the algorithm could show the performance when it visited an area previously learned. The simulation results are presented in Fig. 9 and Table 4. In Fig. 9, speed trajectories

TABLE 4. Simulation results for learning with all driving cycles and cruise control.

Driving Cycle A		
	MBRL w/ all driving cycles	Cruise (v_{ave} = 69.0 km/h)
Traveling time (s)	521.2	521.1
△ SOC(%)	2.60	2.62
Final speed (km/h)	77	77
Energy saving (%)	0.8	-
Driving Cycle B		
	MBRL w/ all driving cycles	Cruise (v_{ave} = 66.0 km/h)
Traveling time (s)	544.1	544.7
△ SOC(%)	2.37	2.43
Final speed (km/h)	68	68
Energy saving (%)	2.5	-
Driving Cycle C		
	MBRL w/ all driving cycles	Cruise (v_{ave} = 66.5 km/h)
Traveling time (s)	541.5	538.7
△ SOC(%)	2.72	2.76
Final speed (km/h)	75	75
Energy saving (%)	1.5	-

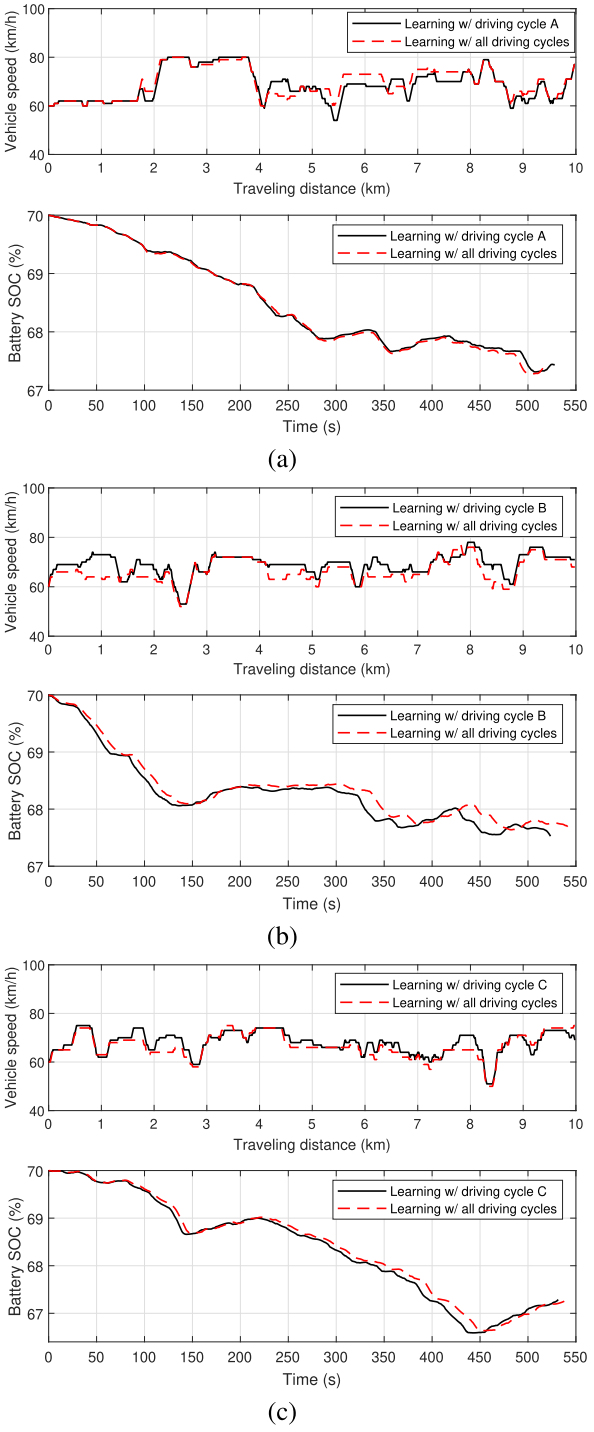

FIGURE 9. Optimized speed trajectory and battery SOC obtained from learning with a specific driving cycle only and leaning with all driving cycles: (a) driving cycle A; (b) driving cycle B; (c) driving cycle C.

obtained from learning with a specific driving cycle only and leaning with all driving cycles are compared, both using MBRL. As shown, the vehicle speed profiles were similar for all the driving cycles, indicating that the driving-cycle information (once learned in MBRL) could be stored in the Q function value and that improved performance could be

repeated when the vehicle revisited a driving environment; however, there was a small performance reduction in energy saving. In Table 4, the energy saving performance of the speed profile based on learning with all driving cycles was compared with that of cruise control. Similar to the previous comparison, the final speed and traveling time were applied as constraints for generating a cruise control speed profile. The results indicated that there was still meaningful energy saving, even though the percentage improvement was reduced compared with learning using a single driving cycle in Table 3. Therefore, the offline control policy (represented as the optimized Q function) can be utilized as a real-time controller for the eco-driving strategy.

V. CONCLUSION

A reinforcement learning algorithm was developed for the eco-driving of an EV, and through this, we confirmed that the reinforcement learning method can be well applied to the eco-driving problem. Especially, we showed that using MBRL, an energy saving optimal speed trajectory utilizing the road slope of the driving cycle can be acquired, and it was compared with the optimal solution using the existing DP method and the cruise control case, demonstrating the excellence and feasibility of the reinforcement learning based approach. The proposed MBRL algorithm separates the vehicle energy-consumption model from the driving environment; thus, learning can be conducted efficiently with the domain knowledge of vehicle dynamics and the powertrain model, while model-free characteristics are maintained by updating the approximation model with experience replay. The proposed algorithm exhibited an energy saving performance of 1.2% – 3.0% compared with cruise control and similar behavior to DP. Additionally, we showed that by training the control policy with combined driving cycles and testing for separated specific driving cycles, the control policy can be used as an offline real-time controller. The limitation of this study is that we only used road slope information to generate an optimal speed profile among diverse driving environments; other constraints associated with traffic signals or other vehicles on the road were not included in the learning process. However, the approaches based on reinforcement learning have advantages for dealing with model uncertainty using the interaction between the agent and the environment; thus, we expect that these constraints with disturbance can be modeled and an optimal control policy can be learned successfully in a stochastic manner. This should be investigated in a future study.

REFERENCES

[1] W. Dib, A. Chasse, P. Moulin, A. Sciarretta, and G. Corde, "Optimal energy management for an electric vehicle in eco-driving applications," *Control Eng. Pract.*, vol. 29, pp. 299–307, Aug. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066114000355

[2] J. Han, A. Sciarretta, L. L. Ojeda, G. De Nunzio, and L. Thibault, "Safe- and eco-driving control for connected and automated electric vehicles using analytical state-constrained optimal solution," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 2, pp. 163–172, Jun. 2018.

[3] E. Hellström, M. Ivarsson, J. Åslund, and L. Nielsen, "Look-ahead control for heavy trucks to minimize trip time and fuel consumption," *Control Eng. Pract.*, vol. 17, no. 2, pp. 245–254, Feb. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066108001251

[4] K. McDonough, I. Kolmanovsky, D. Filev, D. Yanakiev, S. Szwabowski, and J. Michelini, "Stochastic dynamic programming control policies for fuel efficient vehicle following," in *Proc. Amer. Control Conf.*, Jun. 2013, pp. 1350–1355.

[5] D. Shen, D. Karbowski, and A. Rousseau, "Fuel-optimal periodic control of passenger cars in cruise based on Pontryagin's minimum principle," *IFAC-PapersOnLine*, vol. 51, no. 31, pp. 813–820, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896318325874

[6] B. Saerens and E. Van den Bulck, "Calculation of the minimum-fuel driving control based on Pontryagin's maximum principle," *Transp. Res. D, Transp. Environ.*, vol. 24, pp. 89–97, Oct. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361920913000837

[7] H. Abbas, Y. Kim, J. B. Siegel, and D. M. Rizzo, "Synthesis of Pontryagin's maximum principle analysis for speed profile optimization of all-electric vehicles," *J. Dyn. Syst., Meas., Control*, vol. 141, no. 7, Jul. 2019, Art. no. 071004, doi: 10.1115/1.4043117.

[8] H. Rakha and R. K. Kamalanathsharma, "Eco-driving at signalized intersections using V2I communication," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 341–346.

[9] L. Guo, B. Gao, Y. Gao, and H. Chen, "Optimal energy management for HEVs in eco-driving applications using bi-level MPC," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2153–2162, Aug. 2017.

[10] O. Santin, J. Pekar, J. Beran, A. D'Amato, E. Ozatay, J. Michelini, S. Szwabowski, and D. Filev, "Cruise controller with fuel optimization based on adaptive nonlinear predictive control," *SAE Int. J. Passenger Cars-Electron. Electr. Syst.*, vol. 9, no. 2, pp. 262–274, Apr. 2016, doi: 10.4271/2016-01-0155.

[11] O. Santin, J. Beran, J. Pekar, J. Michelini, J. Jing, S. Szwabowski, and D. Filev, "Adaptive nonlinear model predictive cruise controller: Trailer tow use case," SAE International, SAE Tech. Paper 2017-01-0090, Mar. 2017, doi: 10.4271/2017-01-0090.

[12] S. E. Li, H. Peng, K. Li, and J. Wang, "Minimum fuel control strategy in automated car-following scenarios," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 998–1007, Mar. 2012.

[13] 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, Aug. 2009.

[14] X. Ma, Y. Xie, and C. Chigan, "Meta-deep Q-learning for eco-routing," in *Proc. IEEE 2nd Connected Automated Vehicles Symp. (CAVS)*, Sep. 2019, pp. 1–5.

[15] J. Shi, F. Qiao, Q. Li, L. Yu, and Y. Hu, "Application and evaluation of the reinforcement learning approach to eco-driving at intersections under Infrastructure-to-Vehicle communications," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2672, no. 25, pp. 89–98, Dec. 2018, doi: 10.1177/0361198118796939.

[16] A. Phan and H.-S. Yoon, "A study of using a reinforcement learning method to improve fuel consumption of a connected vehicle with signal phase and timing data," SAE International, SAE Tech. Paper 2020-01-0888, Apr. 2020, doi: 10.4271/2020-01-0888.

[17] G. Li and D. Görges, "Ecological adaptive cruise control for vehicles with step-gear transmission based on reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4895–4905, Nov. 2019.

[18] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 908–918. [Online]. Available: http://papers.nips.cc/paper/6692-safe-model-based-reinforcement-learning-with-stability-guarantees.pdf

[19] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski, "Model-based reinforcement learning for Atari," 2019, *arXiv:1903.00374*. [Online]. Available: https://arxiv.org/abs/1903.00374

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

[21] H. Lee, C. Kang, Y.-I. Park, N. Kim, and S. W. Cha, "Online data-driven energy management of a hybrid electric vehicle using model-based Q-Learning," *IEEE Access*, vol. 8, pp. 84444–84454, 2020.
[22] W. Dib, L. Serrao, and A. Sciarretta, "Optimal control to minimize trip time and energy consumption in electric vehicles," in *Proc. IEEE Vehicle Power Propuls. Conf.*, Sep. 2011, pp. 1–8.
[23] R. Bellman, R. Bellman, and R. Kalaba, *Dynamic Programming and Modern Control Theory* (Academic Paperbacks). Amsterdam, The Netherlands: Elsevier Science, 1965. [Online]. Available: https://books.google.co.kr/books?id=O_9QAAAAMAAJ
[24] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992, doi: 10.1007/BF00992698.
[25] H. Lee, C. Song, N. Kim, and S. W. Cha, "Comparative analysis of energy management strategies for HEV: Dynamic programming and reinforcement learning," *IEEE Access*, vol. 8, pp. 67112–67123, 2020.

HEEYUN LEE (Member, IEEE) 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 affiliated with the Research and Development Division of Hyundai Motor Company, South Korea. His research interests include optimal control, reinforcement learning, modeling, and simulation of electrified vehicles.

NAMWOOK KIM received the B.S. and Ph.D. degrees from Seoul National University, South Korea, in 2003 and 2009, respectively.

He joined the Transportation Research Center, Argonne National Laboratory, in 2009, as a Post-doctoral Researcher and from 2012 to 2015 as a Research Engineer. He is currently working as an Associate Professor with Hanyang University. His research interests include modeling and control for advanced vehicles. He is also pursuing studies related large network behaviours of a transportation system.

SUK WON CHA (Member, IEEE) 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 a Senior Editor of the *International Journal of Precision Engineering and Manufacturing – Green Technology*. He also serves as an Editor of the *International Journal of Automotive Technology*.