공학박사학위논문

Q-Learning에 기반한 하이브리드 차량의 확률론적 최적 에너지 관리

Stochastic Optimal Energy Management based on Q-Learning for Hybrid Electric Vehicles

2018년 8월

서울대학교 대학원

기계항공공학부

이 희 윤

Stochastic Optimal Energy Management based on Q-Learning for Hybrid Electric Vehicles

by

Heeyun Lee

A Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department of Mechanical and Aerospace Engineering

Seoul National University

August 2018

Q-Learning에 기반한
하이브리드 차량의 확률론적 최적 에너지 관리

Stochastic Optimal Energy Management
based on Q-Learning
for Hybrid Electric Vehicles

지도교수 차 석 원

이 논문을 공학박사 학위논문으로 제출함

2018년 4월

서울대학교 대학원

기계항공공학부

이 희 윤

이희윤의 공학박사 학위논문을 인준함

2018년 6월

위 원 장 : 민 경 덕

부위원장 : 차 석 원

위 원 : 이 동 준

위 원 : 박 영 일

위 원 : 임 원 식

ABSTRACT

Stochastic Optimal Energy Management based on Q-Learning for Hybrid Electric Vehicles

Heeyun Lee

Department of Mechanical and Aerospace Engineering

Seoul National University

Student number: 2013-20707

Hybrid Electric Vehicles(HEVs) have been widely studied recently with growing concern over sustainability development of the global environment. HEVs use multiple power sources of internal combustion engine and electric battery generally, thus Energy Management Strategy (EMS) determining the split between two power sources needs to be coordinated to maximize the entire vehicle system efficiency. In this study, an EMS based on Stochastic Dynamic Programming(SDP) and reinforcement learning is developed.

SDP is an approaches based on probability, in which optimization problem is defined infinite time horizon, therefore obtained control policy can be used as real-time controller of the vehicle. In order to apply SDP to vehicle control, the characteristics of the driving cycle are expressed as a Transition Probability Matrix (TPM) through the

Markov process, in which the vehicle speed and the power demand are discretized. The objective function of the optimization problem in this study is defined as minimizing the expected value of vehicle fuel consumption, deviation of battery state of charge, and frequent engine on / off. As result power split ratio is given as function of vehicle speed, battery SOC, power demand and engine on/off status.

However, SDP has limitation that it is offline policy considering it required TPM, thus in this study, reinforcement learning technique is used to compensate this problem. In the newly proposed control strategy, based on the Q-learning algorithm, the probability information of the driving cycle is updated to the Q value, that is expected cost value of each state variable and control input, and the control rule is updated by calculating the cost function for the all admissible control input at each time step using vehicle model.

To verify control strategy, backward-looking is developed for parallel type HEVs. Simulation results show that fuel economy is improved compared to rule-based strategy and near optimal solution is obtained.

Key words: Hybrid Electric Vehicles, Energy Management Strategy, Optimal Control, Stochastic Dynamic Programming, Reinforcement Learning

Heeyun Lee

School of Mechanical and Aerospace Engineering

Seoul National University

Student number:2013-20707

CONTENTS

LIST OF FIGURE

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF TABLES

Chapter 2

Chapter 5

CHAPTER 1 INTRODUCTION

1.1 Motivation

Automobile has significantly contributed to the development of the modern society by satisfying human needs for mobility. Combined with the invention of the internal combustion engine, the automobiles with the advanced technologies have become faster and more powerful. However, due to growing concern for environmental problem and regulations according to it, a necessity of replacing conventional vehicles to eco-friendlier vehicles arises, and fuel economy of a vehicle has been mainstream research area for automotive last decade. Figure 1.1 shows efficiency and power of light-duty vehicles in US over time. Average peak horsepower of the vehicles has grown continuously. In case of fuel economy, it declined from the late 1980s through the mid-2000s, partly due to the rise in popularity of light trucks such as trucks, SUVs, and vans. However, since 2004, average fuel efficiency has been increasing largely due to Corporate Average Fuel Economy (CAFE) standards. Figure 1.2 presents vehicle fuel efficiency requirement by year, in which target miles per gallon value of passenger cars and light duty trucks are defined as rapidly increasing for next decades. To satisfy CAFE requirements, industry and governments need to increase sale share of eco-friendly vehicles.

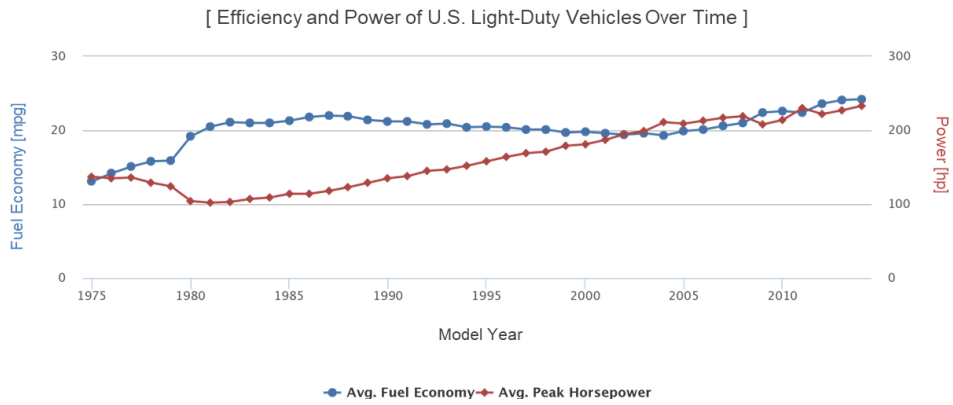


Figure 1.1 Efficiency and power of U.S. light-duty vehicles over time [1]

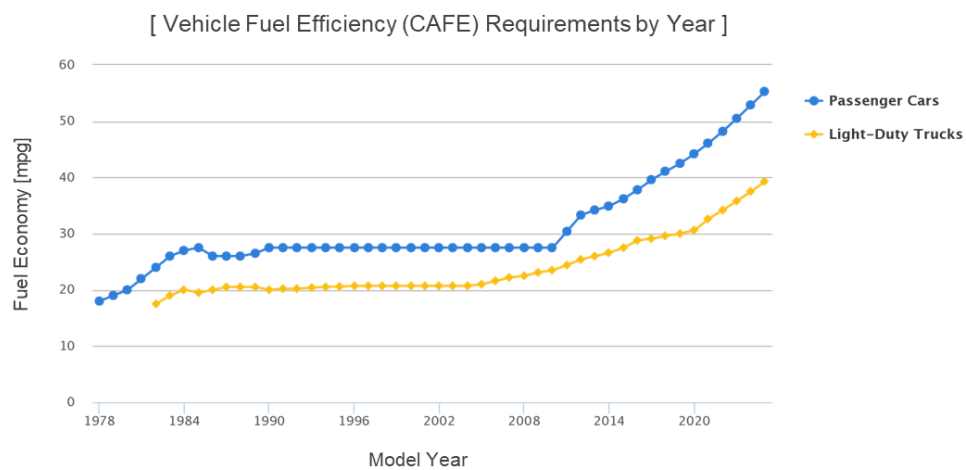


Figure 1.2 Vehicle fuel efficiency(CAFE) requirements by year [2]

Hybrid Electric Vehicles (HEVs) or Plug-in Hybrid Electric Vehicles(PHEVs) are becoming increasingly popular recently thanks to those underlying thirst. HEVs use multiple power sources as internal combustion engine and electric battery, thus can be coordinated to maximize fuel efficiency of the vehicle powertrain system. HEVs emit few pollutants compared with conventional Internal Combustion Engine (ICE) based vehicle and HEVs have much longer driving distance compared to electric vehicles. Also, just like conventional ICE based vehicle, fuel recharge could be done very quickly and easily, and unlike Fuel Cell Hybrid Vehicle (FCHV) and EV, additional infrastructures are not needed.

There has been many researches to make more fuel efficient HEVs[3]-[9]. Basically, there are many researches to improve the efficiency or performance of each part of HEVs, such as a more efficient engine, an electric motor with low loss, or a powertrain with good transmission efficiency. Also, designing powertrain structure of HEVs would be one of the ways[4], [10]-[14]. Typical powertrain structures of HEVs are series, parallel, and power split type. Series HEVs are driven only by the electric motor, and the engine is not connected to drivetrain mechanically. Instead, the engine is operated when the battery energy is not sufficient for the power demand. Parallel HEVs use the electric motor and the engine either individually or together to supply propulsion power. Power split HEVs drive train generally use two motors, in contrast to the parallel HEVs, which typically only one motor is used for. The power splitter, which is usually planetary gear set, is used to distribute power. Power split HEVs have gained its popularity recent decades thanks

to their capability to take diverse engine operating points. On the other hand, since parallel HEVs has advantage of high transfer efficiency, thus different types of powertrain structure have been created recently to take advantage of both power split and parallel type.

The other main issue for developing HEVs is control strategy. The control strategy is mainly an Energy Management Strategy(EMS) to decide power distribution between multiple power sources of HEVs while maintaining drivability and satisfying constraints such as powertrain component's physical operating limitation or battery charge sustenance[15]–[17]. HEVs have complicate powertrain structure, thus fuel economy performance of HEVs is tremendously depends on operating strategy of powertrain.

In order to achieve high fuel efficiency, an EMS should be set up, thus the energy source of the vehicle, that is, the fossil fuel and the electric energy, should be used in its most efficient operating regions of the engine and the motor, respectively. Also, it is necessary not only to consider the efficiency of the individual vehicle component but also to increase the overall efficiency of the vehicle powertrain and appropriately distribute the power to the engine and the motor according to the driving environment of the vehicle over traveling time.

In this paper, a study on the EMS of HEVs is conducted and a new strategy is developed. In the next section, we investigate literature reviews of the EMS previously studied.

1.2 Background Studies

Energy management strategies for HEVs can be classified generally into two main research trends of rule-based control strategy and optimization-based control strategy according to their methodologies as shown in figure 1.3 [18],[19]. In rule-based control strategy, powertrain is controlled based on rule which is made upon heuristics or intuition, which is mainly focused on operating its power source such as the engine or electric motor in efficient area [20]-[29]. Usually, rule-based control strategy is more applicable into an implementation as a real-time vehicle controller, and it shows robust performance. However, the fuel economy performance of rule-based control strategy is lower than that of optimization-based strategy generally. Rule-based strategy could be classified as a Fuzzy rule-based and deterministic rule-based strategy.

Optimization-based control strategy is the strategy based on optimization theory and presents outstanding fuel economy performance[28], [30]-[38]. There have been many researches for optimization-based strategy recent decades, which could be divided into global optimization strategy and real-time optimization strategy. Usually, global optimization strategy requires the entire driving cycle information to get optimal solution, while real-time strategy is based on intuitive concept of optimal control. Equivalent Consumption Minimization Strategy (ECMS) is based on optimal concept of relationship between fuel consumption and electric power

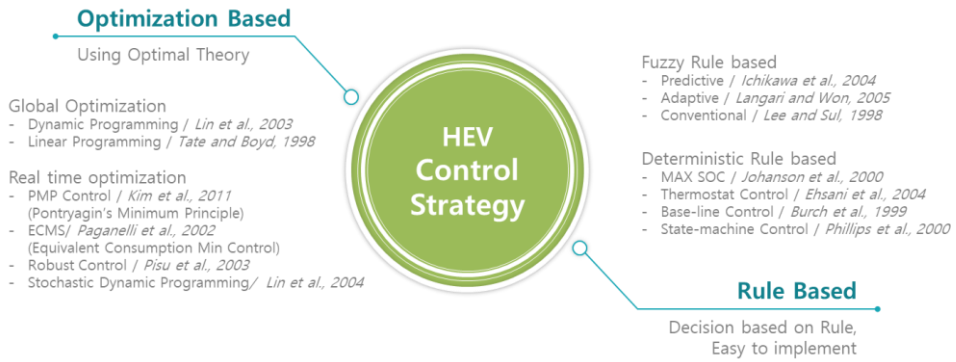


Figure 1.3 Overview of HEV control strategy research trends

consumption, while its underlying concept of ECMS is proved by Pontryagin's Minimum Principle(PMP). In ECMS, or PMP, coefficient called equivalent factor, or co-state, which decide relative value of fuel consumption over electric power consumption should be determined according to driving cycle, thus predictive or adaptive control algorithm based on ECMS or PMP is now developing.

On the other hand, Dynamic Programming (DP) is an effective optimization algorithm, presenting outstanding fuel economy performance. DP is one of the optimization-based control strategies using Bellman's principle of optimality and it has been investigated for EMS of various HEV types in many previous researches[39]-[45]. Despite these advantages, DP is considered non implementable as a real-time supervisory controller of the vehicle, since it requires a computation burdensome and an additional rule extraction process is needed. Above all, DP has a non-causal property, as it needs a

future driving schedule to obtain global optimality and the acquired optimality is only valid for a specific driving cycle. Thus optimality for the other driving cycles is not guaranteed. Therefore, the use of DP is mainly limited for estimating vehicle system's maximum fuel economy performance or for finding general rules to control the vehicle's power sources.

There have been studies about the rule-based strategy of HEV optimized using DP indirectly for real-time implementation of vehicle control. [46] optimized the rule-based control strategy by using DP, which was developed based on the maximum SOC control strategy and the engine on-off control strategy. In another research, [21] used DP as a benchmark to find optimal power distribution of DC motor and hydraulic pump for a hydraulic electric hybrid vehicle. Rule-based control strategy is easier to implement as a real-time control supervisory and has the advantage of being highly robust, but the major drawbacks of the control approaches in the above mentioned studies, were that their rule-based control strategies could not be optimized for the whole trip. Even though parameters in the rule-based control strategy were optimized, they were only valid for a specific driving cycle, and efficiency of the entire drivetrain was not optimized for diverse driving cycles.

To overcome the drawback of DP and rule-based strategy, Stochastic Dynamic Programming(SDP) has been studied recently as shown in figure 1.4. [47] suggested the concept of SDP for the vehicle controller of gear shifting and engine control. For energy management problem of HEVs, several types of vehicle have been studied using SDP. In [40], an infinite-horizon stochastic dynamic

optimization problem for parallel HEV is formulated, and [48] uses SDP for Toyota Hybrid System which contains power-split planetary gear system. [49] studied Plug-in HEV in which SDP is used to optimize power management of PHEV by examining the tradeoff relationship of fuel and electricity usage base on relative fuel to electricity price. and [50] applied SDP for the fuel cell vehicle, in which a reduced-order fuel cell model is created and SDP approach is used to optimize the power management strategy of the FCHV. Also, more practical approaches have been studied for SDP. [51] used traffic driving data to present target vehicle speed to minimize average weighted sum of fuel use and time for conventional ICE based vehicle. and in [52], implementation of the SDP controller in hardware is presented with considering practical issues such as real-time computability, and driver perception. Experimental

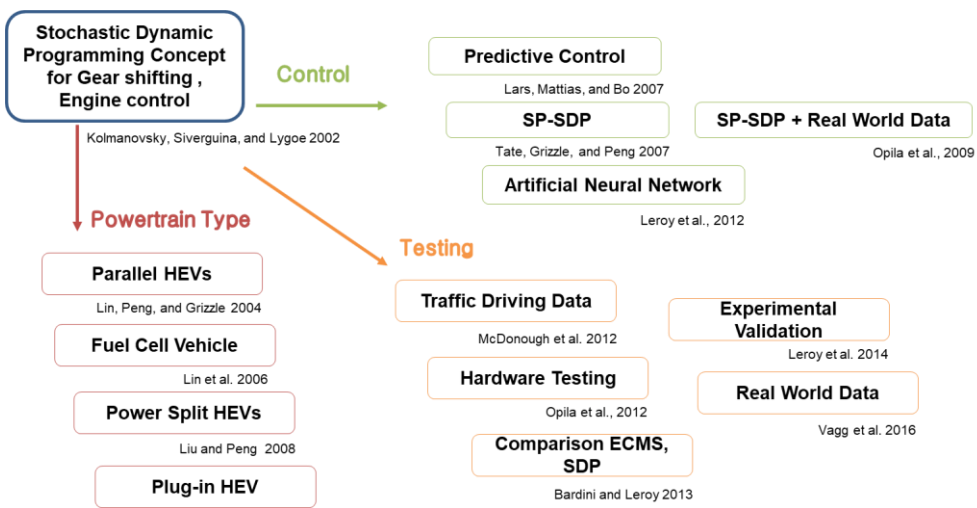


Figure 1.4 Previous researches for stochastic dynamic programming

validation for SDP has been done by [53] in which SDP controller is implemented in the electronic control unit of the vehicle and in [54], SDP with practical consideration for robust implementation is addressed for dynamometer testing. [55] did a comparison study for ECMS and SDP with consideration regarding engine utilization and dwell time. On the other hand, [7] applied SDP based on predictive control in which predictive powertrain control is assessed for different level of information supplied by the vehicle navigation and traffic flow information system. [56] developed shortest path SDP considering battery SOC condition in which deviation of SOC is penalized only at key off , and [57] developed SDP combined with artificial neural network to overcome drawbacks of SDP that it requires large storage memory and the calibration task of SDP is time-consuming process.

However, the limitation of SDP is that it is an offline policy fundamentally. SDP also models the characteristics of the driving cycle and uses it for optimization. Therefore, it is possible to use general control rules derived from SDP for real-time control, but another calculation process is required to extract control rules in accordance with actual driving speed profile being driven if characteristic of current driving cycle is different with previous one, used for the optimization process.

Reinforcement Learning (RL), which is one of the field of machine learning that has been under active research in recent years, has advantage that it can solves the Bellman optimal equation of DP online in real time unlike SDP. As shown in figure 1.5, RL is a more advanced

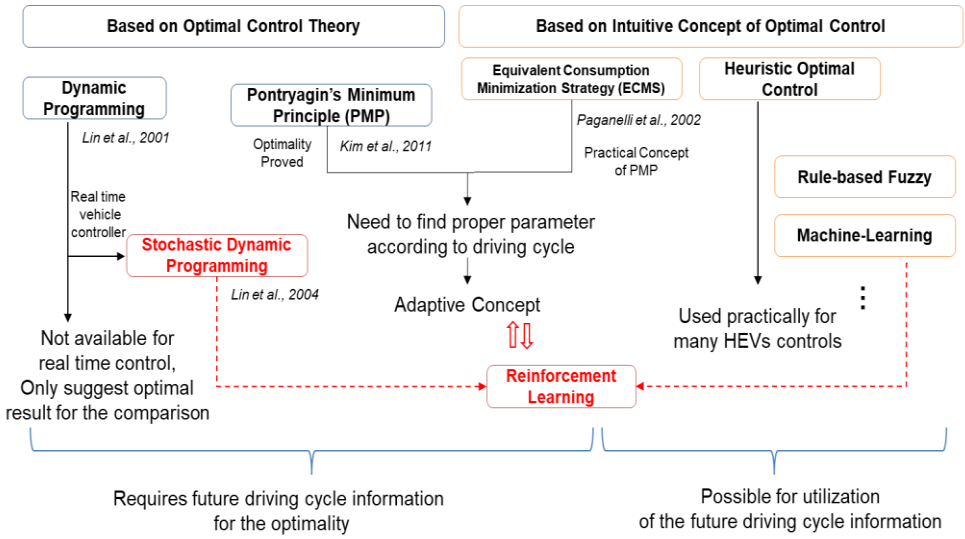


Figure 1.5 Optimization-based energy management strategy for HEVs

form than the existing SDP (which is also called as Markov decision process), while the learning process and the problem solving are represented in the form of optimal control problems, which is different from the existing studies on the EMS of HEVs based machine learning. RL is closely related to optimal control and adaptive control, which is an interesting point that the RL-based control strategy has similar advantages to the adaptive ECMS or adaptive PMP concept aforementioned because it is based on learning process. A few works have been done to use technique of RL for EMS of HEVs, especially Q-learning. [58] is the first work that applies the RL to the power management strategy of HEV, in which TD-learning algorithm is used to derive optimal control policy. In [59], RL is applied to power management strategy for PHEV, in which the

remaining distance to travel is chosen as state variable and the immediate reward is defined as the sum of the fuel consumption cost and battery energy usage cost. In [60], RL is used to optimize power distribution between the battery and the ultra-capacitor for a PHEV. In this paper, the transition probability matrices are updated based on driving cycle and Kullback-Leibler divergence rate. [61] presented the RL-based EMS for a hybrid electric tracked vehicle, in which Q-learning and Dyna algorithm are applied to generate optimal control policy. [62] suggested predictive EMS based on RL and velocity prediction is applied to parallel HEV. More recently, [63] utilized Deep Q Network, which combines Q-learning and a deep neural network for HEV control.

However, the researches in these papers are limited to case studies only that apply the RL algorithm to HEVs control problems, and the correlation between existing fuel efficiency results with DDP or SDP is not suggested. In addition, since it is based on learning process, there are limitations in using it for actual vehicle control, or there is a case where the optimality of the presented algorithm is inferior.

In this paper, a new EMS based on stochastic optimal control is investigated for parallel type HEVs to overcome these limitations. Firstly, SDP is developed for EMS of HEVs, and based on SDP framework, RL method is used to develop a new EMS which can be used as real-time vehicle controller based on Q-learning algorithm.

1.3 Contributions

Contribution of this paper is that development of EMS based on SDP and RL which enables real-time vehicle control. Firstly, the process for applying SDP into EMS of HEVs is established. We have derived an EMS that can be utilized as real-time controller by applying SDP to optimal control problem of HEVs. In this study, the optimization based on SDP was performed considering the engine on / off cost especially in order to improve the practical application of the algorithm, and the numerical technique was used to reduce the calculation load of the SDP.

In addition, a new algorithm based on RL has been proposed to overcome the shortcomings of SDP, which is limitations as an offline policy. In the new algorithm, we developed an EMS that enables online control by applying the Q-learning algorithm while using the SDP framework. In the newly developed control strategy, it is possible to adaptively control by learning the characteristic of current driving speed profile, and it is possible to derive the control rule based on the vehicle component's operating condition thanks to its model-based and self-learning characteristics.

The contribution of this paper is to apply and verify the real time control algorithm idea based on the Bellman equation from DP to SDP, and RL for the EMS of HEVs.

1.4 Thesis Outlines

This thesis concentrates on stochastic optimal control of HEVs. The main body of this dissertation composed of 6 chapters. Each chapter organizes as follows:

Chapter 2 describes a HEV powertrain model development. As a target vehicle, Hyundai Sonata Hybrid 2011 was used, which is parallel type HEVs. For the vehicle simulation, a backward-looking vehicle simulator are developed.

Chapter 3 describes stochastic optimal control of HEVs. Process for stochastic optimal control is presented, in which Markov process is used for modeling driving cycle and control policy is extracted from stochastic dynamic programming.

Chapter 4 describes a new reinforcement learning based EMS. Stochastic optimal control approach is proposed for EMS based on Q-learning, which can be used when driving cycle information is unknown.

Chapter 5 describes simulation analysis. Diverse driving cycles are used for the evaluation of the proposed EMS.

Chapter 6 describes concluding remarks. Conclusion and future work of this paper is presented.

CHAPTER 2 VEHICLE MODEL DEVELOPMENT

In this study, we use simulation verification method to test and verify the proposed algorithm. It is very important to have reliable vehicle powertrain model to perform vehicle simulations. This chapter describes the development process of vehicle simulation models and explains the mathematical basis for them.

2.1 Target Vehicle – Hyundai Sonata Hybrid

Target vehicle model is 2011 Hyundai Sonata Hybrid. 2011 Hyundai Sonata Hybrid is Hyundai's first hybrid and a version of the mid-size Sonata sedan as shown in figure 2.1. Sonata Hybrid is a front-wheel drive 4-door sedan and it is a parallel type HEV, which combines a 2.4-liter engine with a 30 kW electric motor, thus the vehicle can move via the engine, electric motor, or combination of the two. For the transmission, six-speed automatic transmission is used. The fuel economy of Sonata Hybrid is 37 miles per US gallon in the city and 40 miles per US gallon on the highway. The vehicle parameter and characteristic data of the vehicle are listed in Table 2.1. The schematic of the vehicle powertrain is shown in figure 2.2.

Table 2.1 Vehicle simulation parameter and characteristic data

Components	Value
Gasoline engine	Maximum power 122 $[kW]$ @ 6000 $[rpm]$
Electric motor	Permanent Magnet Synchronous Motor Rated power: 30 $[kW]$
Battery	Capacitance: 5.3 $[Ah]$
Final drive ratio	3.23
Transmission	6 speed Automatic Transmission [4.21, 2.64, 1.80, 1.39, 1.00, 0.77]
Load road coefficient	f_0 : 26.8 $[lb]$, f_1 : 0.15 $[lb/mph]$, f_2 : 0.0145 $[lb/mph^2]$
Vehicle mass	1700 $[kg]$

Figure 2.1 2011 Hyundai Sonata hybrid

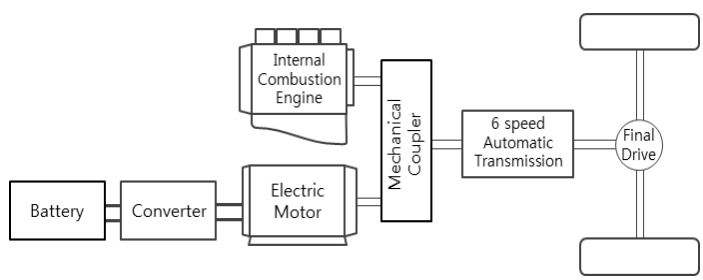


Figure 2.2 The schematic of the vehicle powertrain

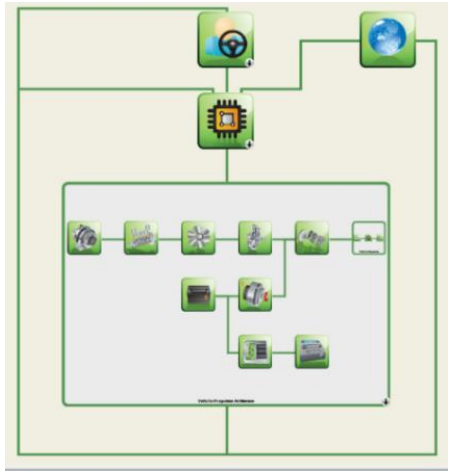


Figure 2.3 Autonomie software

Vehicle powertrain model is developed based on 2011 Hyundai Sonata Hybrid. The control strategy developed in this study is not limited for parallel hybrid, but in this study, parallel hybrid type is used. For developing vehicle model, characteristic data of each vehicle component from vehicle simulation software Autonomie is used [64]. Autonomie is a vehicle system simulator for energy consumption and performance analysis developed by Argonne national laboratory in U.S. It is a MATLAB based software and is validated based on Advanced Powertrain Research Facility(APRF) dynamometer test data in Argonne national laboratory. In this study, characteristic data of battery and electric motor are used from Autonomie. These data are estimated characteristic data of the permanent magnet electric motor and Li-ion battery from test data of the 2011 Hyundai Sonata Hybrid. In case of engine data, 2L gasoline engine data is modified and scaled to meet maximum power of known characteristic data of the engine to use for simulation.

2.2 Vehicle Modeling

Vehicle model for simulation is developed based on 2011 Hyundai Sonata Hybrid data as aforementioned. In this study, a backward-looking vehicle simulator is developed. Forward-looking vehicle simulator is a usual vehicle simulator, in which driver model commands accelerator and brake pedal signal to follow given target speed profile, and each vehicle component react according to the command. On the other hand, backward-looking vehicle simulator is the vehicle simulator to analyze how each vehicle components are operated. In backward-looking vehicle simulator, driving cycle profile is given as same as forward-looking vehicle simulator, but it is interpreted backward from wheel to power source without the driver model. Backward-looking vehicle simulator is used mainly for developing control algorithm or to check fuel economy performance of the vehicle thanks to its advantage of time efficient compared to forward-looking vehicle simulator. In backward-looking vehicle simulator, quasi-steady model is used, thus transient dynamics of the vehicle powertrain are neglected. In this study, backward-looking vehicle simulator is developed. Optimization-based strategy is evaluated upon backward-looking vehicle simulator to validate the effectiveness of the proposed control strategy.

2.2.1 Sub-System Modeling

Vehicle components are modeled based on mathematical models. For engine modeling, a quasi-static engine fuel consumption model is utilized. It is assumed that the engine transients such as combustion dynamic are much faster than vehicle system level dynamics for energy flow analysis. Fuel consumption rate of engine, $\dot{m}$ is represented using map as given in figure 2.4 and equation from the engine torque T_{eng} and engine speed ω_{eng}

$$\dot{m} = f_{fuel}(T_{eng}, \omega_{eng}) \tag{2.1}$$

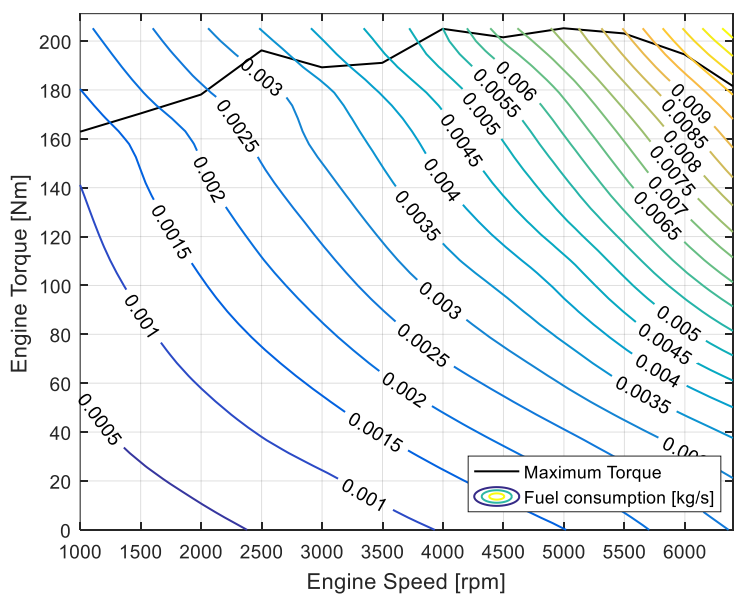


Figure 2.4 Engine fuel consumption map

In case of motor modeling, the efficiency of motor, η_{mot} is calculated using pre-determined map and battery power output, P_{bat} is also presented from motor torque T_{mot} and motor speed ω_{mot} as equation (2.2).

$$P_{bat} = {\eta_{mot}}^{k} \cdot T_{mot} \cdot \omega_{mot} \tag{2.2}$$

The efficiency of motor, η_{mot} is function of motor torque T_{mot} and motor speed ω_{mot} as given in figure 2.5. If motor is used as a motor $k = -1$, and if motor is used as generator $k = 1$. It is also assumed that the effects caused by transient dynamics of electric motor are sufficiently small, thus can be neglected.

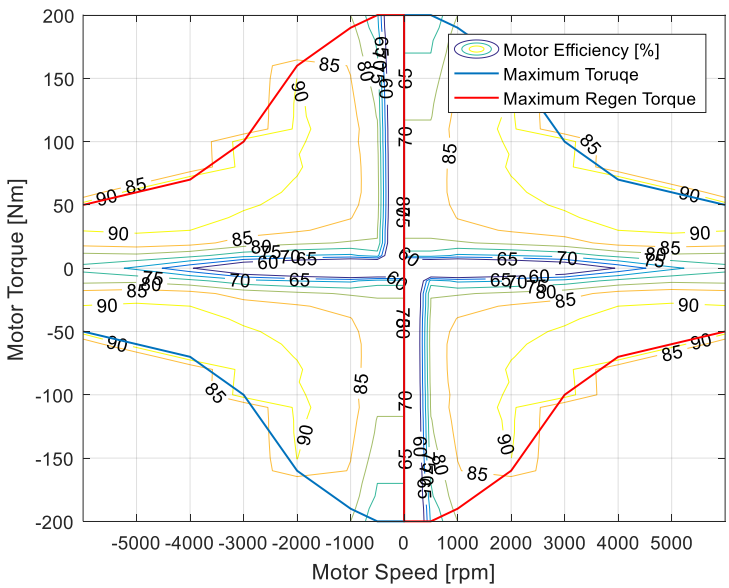


Figure 2.5 Motor efficiency map

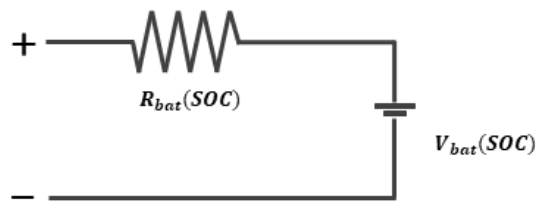


Figure 2.6 An equivalent circuit model of a battery

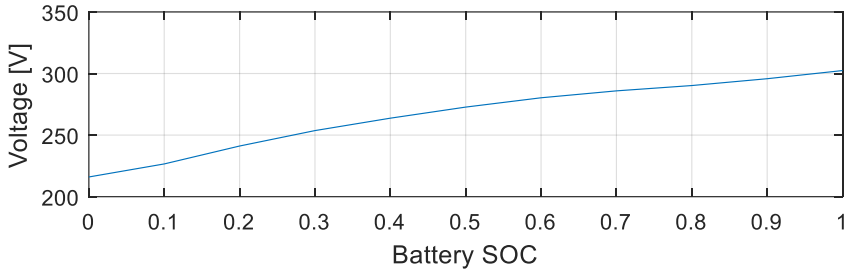


Figure 2.7 Voltage of battery

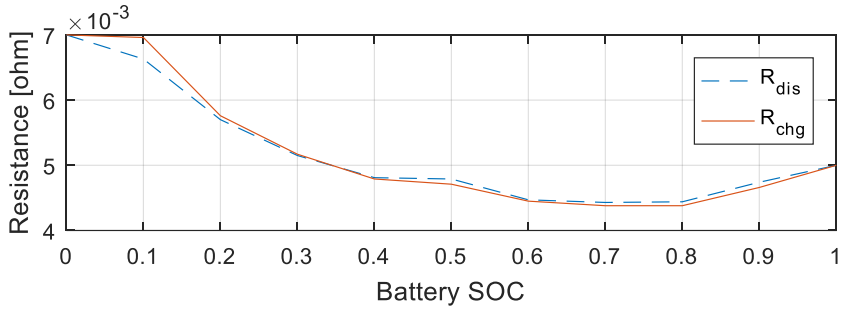


Figure 2.8 Resistance of battery

The battery power in (2.2) changes State of Charge (SOC) in the battery as modeled with SOC dynamics described by equation (2.3) considering an equivalent circuit model of a battery as shown in figure 2.6.

$$\dot{SOC} = -\frac{1}{Q_{bat}} \cdot \frac{V_{oc} - \sqrt{{V_{oc}}^2 - 4P_{bat}R_{bat}}}{2R_{bat}} \tag{2.3}$$

where open circuit voltage of battery is V_{oc}, electric power consumed at battery outside is P_{bat} , internal resistance is R and the battery capacitance is Q_{bat}. For the battery model, simple internal resistance model is used. Open circuit voltage and internal resistance of the battery are determined by pre-determined map as shown in figure 2.7 and figure 2.8.

For the powertrain, drivetrain dynamics from the transmission input shaft to the wheel can be expressed simply as equation (2.4) - (2.6) when a clutch is engaged.

$$T_{wh} = ((T_{eng} + T_{mot} - T_{gb\_loss}) \cdot \gamma_{gb} \ - T_{fd\_loss}) \ \cdot \zeta_{gb} \tag{2.4}$$

$$\omega_t = \gamma_{gb} \cdot \zeta_{gb} \cdot \omega_{wh} \tag{2.5}$$

$$T_t = T_{eng} + T_{mot} \tag{2.6}$$

where T_{wh} is wheel torque, T_{gb_loss} is torque loss in transmission, γ_{gb} is gear ratio, T_{fd_loss} is final drive torque loss, ζ_{gb} is final drive gear ratio, ω_t is transmission input speed, ω_{wh} is wheel speed, and T_t is transmission input torque. Loss for the gear box is given as a 3−dimensional map as given in equation (2.7), which is a function of T_t , ω_t , and gear step number i_{gb}.

$$T_{gb\_loss} = L_{gb}(T_t, \omega_t, i_{gb}) \tag{2.7}$$

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

$$T_{fd\_loss} = L_{fd}(T_{fd}, \omega_{fd}) \tag{2.8}$$

In this paper, gear shift is conducted based on gear shifting schedule as shown in figure 2.9. Vehicle model can be described simply as equation (2.9) and (2.10) by considering only the longitudinal vehicle dynamics.

$$\dot{v} = \frac{T_{wh}R_{tire} - F_{brake} - F_{loss}}{(M_{veh} + I_{eq})} \tag{2.9}$$

$$F_{loss} = f_0 + f_1 \times v + f_2 \times v^2 \tag{2.10}$$

where R_{tire} is the tire radius, F_{brake} is brake force, F_{loss} is the road

load loss, which includes road grade, f_0, rolling resistance, f_1 , and aerodynamic loss, f_2.

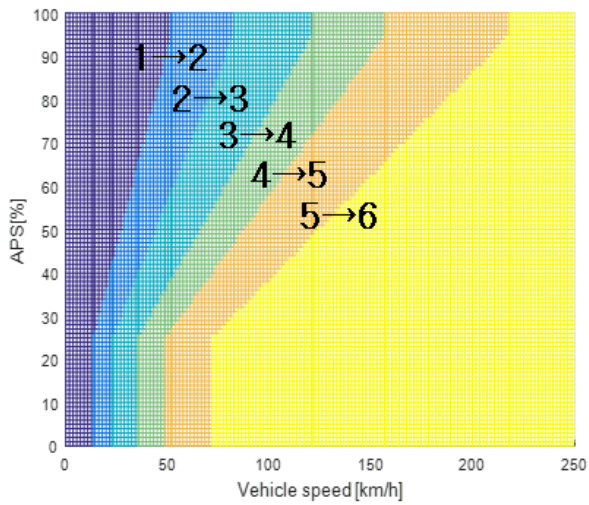


Figure 2.9 Gear shifting map

CHAPTER 3 STOCHASTIC DYNAMIC PROGAMMING BASED ENERGY MANAGEMENT STRATEGY

3.1 Introduction

In the discrete−time format, a model of the HEVs can be expressed as

$$x(k+1) = f\big(x(k), u(k)\big), k = 0,1, \dots N-1 \tag{3.1}$$

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

$$\begin{aligned} \min \quad & \mathrm{J} = \sum_{k=0}^{N-1} L\big(x(k), u(k)\big) \\ \text{subject to} \quad & \end{aligned} \tag{3.2}$$

$$SOC(0) = SOC(N)$$

$$\omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max}$$

$$T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k))$$

$$T_{mot,min}(\omega_m(k), SOC(k)) \le T_{mot}(k) \le T_{mot,max}(\omega_m(k), SOC(k))$$

$$SOC_{min} \le SOC(k) \le SOC_{max}$$

where L is instantaneous fuel consumption. Note that generally final SOC value, $SOC(N)$ is required to be same as initial SOC value, $SOC(0)$, thus only fuel consumption could be evaluated. The optimization problem for HEV control has non-convex property and it is a constrained nonlinear optimal control problem. Deterministic Dynamic programming is one of the promising solution, since it guarantees optimality. However, the application of deterministic dynamic programming into real-time vehicle control is not available due to its computational burdensome and non-causal characteristic that entire future speed profile of the vehicle should be known in advance before trip, which mean optimality of DDP does not hold anymore in real world.

SDP is a variation of DDP, in which optimized control policy result can be used as real-time vehicle controller, while near-optimal result can be acquired. Figure 3.1 presents overall process of SDP. Firstly, from driving cycle information speed profiles are collected. Based on vehicle powertrain modeling, the vehicle speed profiles are interpreted into transition probability matrix(TPM), in which the transition of vehicle speed state and power demand state are modeled. Using TPM, and vehicle powertrain modeling, SDP conducts optimization process and optimal control policy is extracted. Finally, the acquired optimal control policy is implemented on the vehicle simulator directly. The concept of the SDP is also based on the Bellman equation. However, unlike DDP, SDP uses Markov process to model vehicle driving cycle, thus stochastic characteristic of real-world driving cycle information can be modeled and those

characteristic can be defined in the framework of optimal control problem. DDP is powerful tool to handle complicated problem or the problem which has non-linear property. SDP also has the same approach based on the Bellman equation, therefore it can be applied into the optimal control problem of HEV. The aim of the SDP is to compute optimal control policy that behaves optimally in the face of uncertainty of vehicle driving environment. In this chapter, SDP algorithm and application into HEV control will be explained. Before SDP, DDP will be introduced shortly in a subsequent section, which is developed to build up a benchmark for SDP.

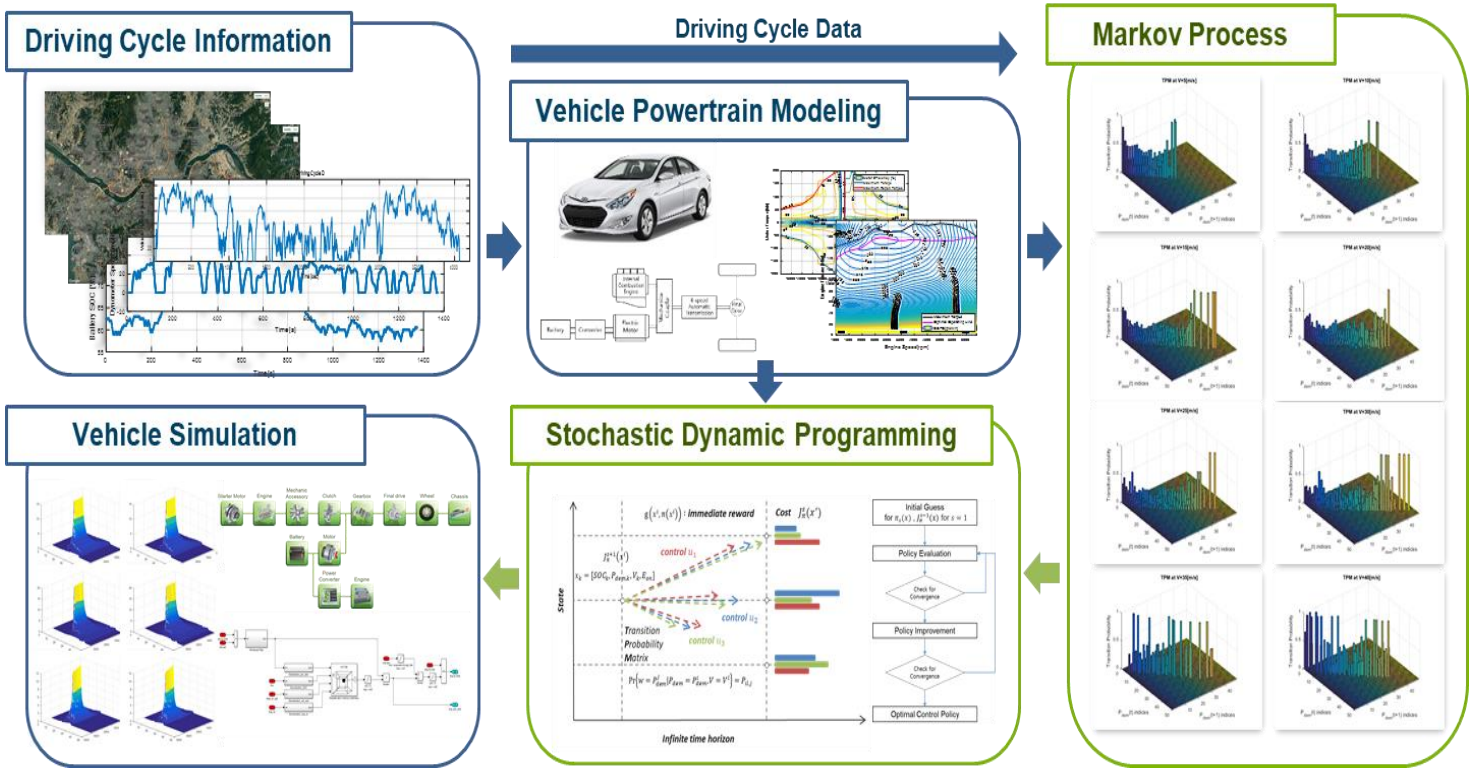


Figure 3.1 Overview of stochastic dynamic programming process

3.2 Deterministic Dynamic Programming

As mentioned above, Dynamic programming is well known algorithm to solve a complicate optimization problem effectively. For optimization problem of HEV, DDP has been used in many research to control power distribution between engine and motor [40],[65]. DDP presents a global optimal solution by searching all possible control options, thus it could be used to find the best fuel economy of the given vehicle system over a given driving cycle. In this study, DDP is built up for the purpose of suggesting best available solution of the EMS. For the optimization problem, equation (3.2) can be broken down into sub equations recursively, as equation (3.3) and (3.4)

$For\ \ k = N-1,$

$$J_k^*(x(k)) = \min_{u(k)}[L\big(x(k), u(k)\big) + \mu(SOC(N) - \mathrm{SOC}(0))^2] \quad (3.3)$$

$For\ \ 0 \le k < N-1,$

$$J_k^*(x(k)) = \min_{u(k)}[L\big(x(k), u(k)\big) + J_{k+1}^*(x(k+1))] \quad (3.4)$$

Where $J_k^*(x(k))$ is the optimal value function at time k, and μ is term for penalizing SOC deviation. For HEV optimal control problem, DDP presents the optimal power distribution among the ICE engine

and the electric motor, after considering the vehicle model, powertrain characteristic and the driving cycle speed profile.

In this study, DDP is developed for the comparison study with the proposed EMS. Gear shift is conducted based on a conventional gear shift map, which is a function of the vehicle speed, and power demand as shown in Figure 3.2. For the optimality, gear shift schedule also should be optimized, however the optimized gear shift schedule considering only fuel consumption is unrealistic on the practical side of vehicle control when it comes to a drivability. Therefore, gear shift is conducted based on the shift map in this study. On the other hand, a penalty term for engine on/off event is also added to cost function to avoid frequent engine on/off. The optimization problem (3.2) can be written again as equation (3.5)

$$\begin{aligned} \min \quad & \mathrm{J} = \sum_{k=0}^{N-1} \left(L\big(x(k), u(k)\big) + \beta \cdot \Delta E_{on}\right) \\ \text{subject to} \quad & \end{aligned} \tag{3.5}$$

$$SOC(0) = SOC(N)$$

$$\omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max}$$

$$T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k))$$

$$T_{mot,min}(\omega_m(k), SOC(k)) \le T_{mot}(k) \le T_{mot,max}(\omega_m(k), SOC(k))$$

$$SOC_{min} \le SOC(k) \le SOC_{max}$$

where β is a coefficient for the engine on/off event, ΔE_{on}. Therefore, result of DDP becomes more practical solution and comparable with other simulation results. Figure 3.3 −3.5 present simulation results using DDP regarding the engine on penalty cost change. The number engine on/off event decrease as the penalty value increase. Figure 3.6 presents DDP simulation result for UDDS and HWFET.

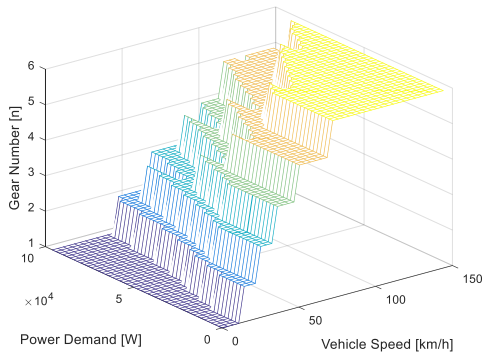


Figure 3.2 Gear number matrix with respect to power demand and vehicle speed

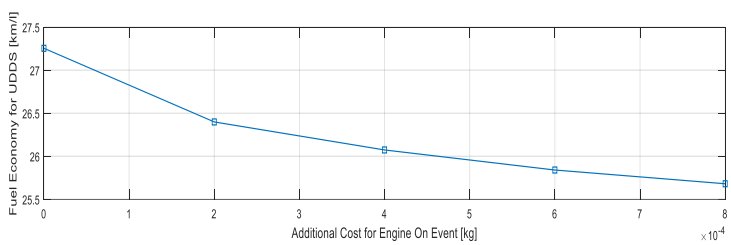


Figure 3.3 Number of engine on according to engine of penalty cost

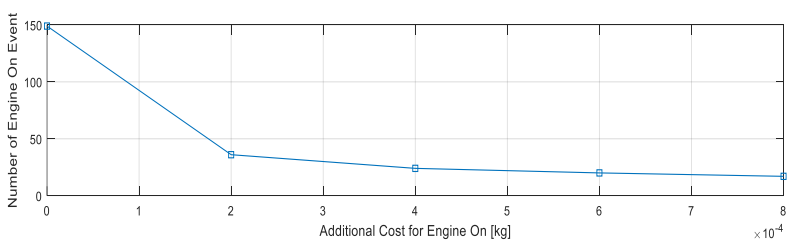


Figure 3.4 Fuel economy for UDDS according to engine on penalty cost

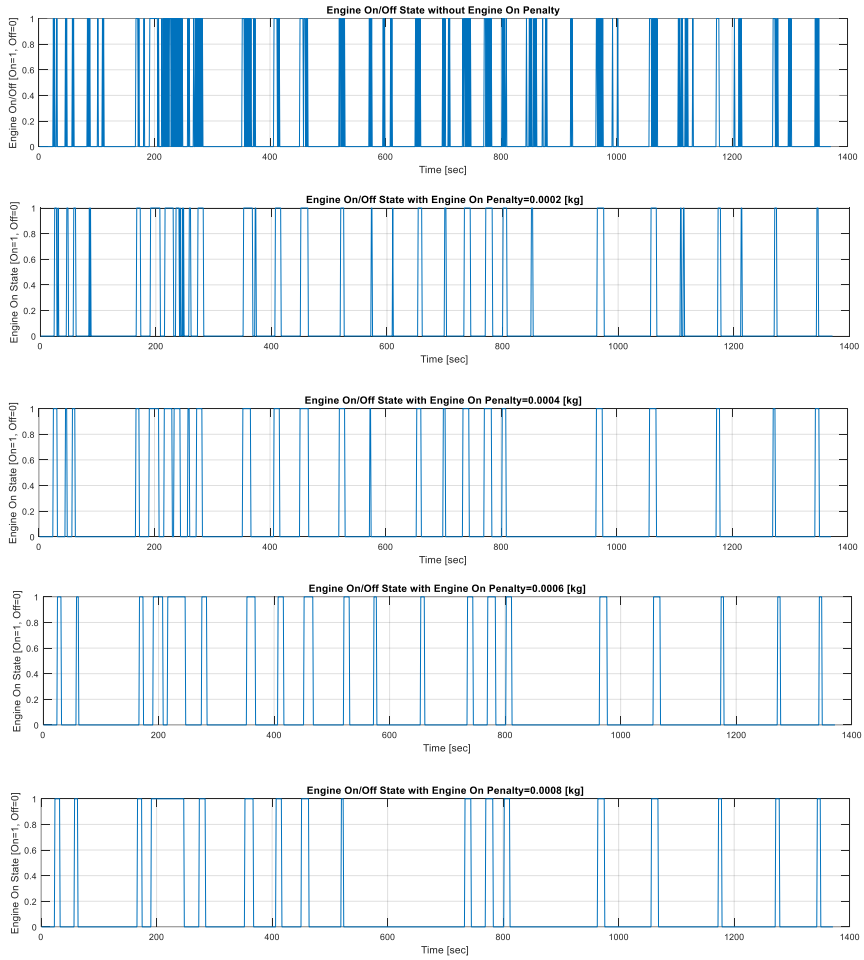
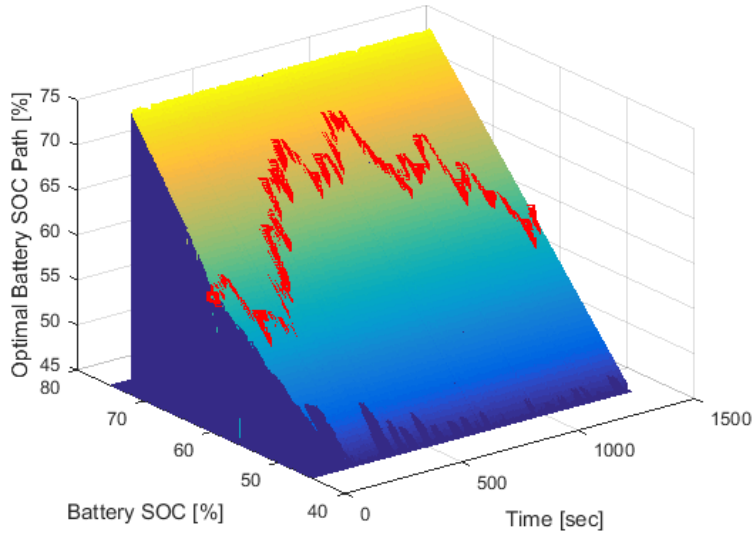
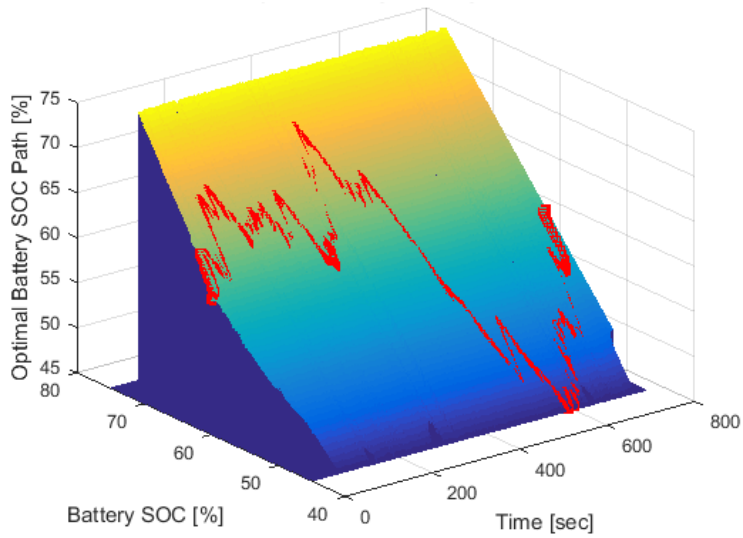


Figure 3.5 Engine on/off result according to engine on penalty cost



(a) UDDS



(b) HWFET

Figure 3.6 DDP simulation results, optimal SOC path

3.3 Stochastic Modeling of Driving Cycle Information

In optimization-based strategy, future driving cycle information is required to be used in order to increase fuel economy performance of the vehicle. There has been many research that utilize future driving cycle information. [66] developed rule-based strategy using navigation information. [67] uses PMP with the estimation of co-state value for driving cycle. For DDP, [68] presents charge-depletion strategy based on DDP, and [46] suggests constrained engine on-off strategy, which is optimized using DDP upon a driving cycle. The way to utilize driving cycle information is a very important problem for HEV control that not only configuration of vehicle powertrain, or efficiency of each vehicle component, but also the estimation of the future speed profile of the vehicle and utilization of it into EMS are key factor deciding fuel economy of the vehicle.

In SDP, Markov process is used to utilize driving cycle information. In DDP, power demand and vehicle speed are given as constant value at each time step. However, in SDP, whole vehicle speed trajectory is not a given information, but it is given as TPM in average sense. In SDP, Markov chain modeling is used to extract the TPM from driving cycle information. A Markov chain (or A Markov process) is mathematical framework to model future driving cycle information. A Markov chain is a stochastic process with the Markov property, which

is the conditional probability distribution of current state depends only upon the previous state, not on the sequence of past state. In the Markov chain, the state of system is presented by the finite number of values, and it evolves from a state to the another based on TPM. In this study, power demand and vehicle speed are presented by Markov chain model.

Power demand, P_{dem} is defined from driving cycle and vehicle parameters

$$P_{dem} = v \cdot (F_{loss} + F_{accel}) \tag{3.6}$$

$$F_{accel} = (M_{veh} + I_{eq}) \cdot a_{veh} \tag{3.7}$$

where F_{accel} is a force for vehicle acceleration, a_{veh}, and I_{eq} is equivalent rotating inertia in the vehicle powertrain, and F_{loss} is road load loss. Power demand and vehicle speed are discretized into a finite number of values,

$$P_{dem} \in \left\{P_{dem}^1, P_{dem}^2, \dots, P_{dem}^{N_p}\right\} \tag{3.8}$$

$$v \in \{v^1, v^2, \dots, v^{N_v}\} \tag{3.9}$$

where N_p is number of the power demand, and N_v is number of speed discretized. Nearest-neighborhood method is used to map continuous value to the quantized state, which is the closest to the

continuous value. Then, the dynamics of power demand can be expressed as equation (3.10) – (3.11)

$$P_{dem,k+1} = d_k \tag{3.10}$$

$$P_{il,j} = \Pr\{d = P_{dem}^{j} | P_{dem} = P_{dem}^{i},\ v = v^{l}\} \tag{3.11}$$

$$\text{for i, j} = 1,2,\dots,N_p,\ l = 1,2,\dots,N_\omega$$

$$\hat{P}_{il,j} = \begin{cases} \frac{n_{il,j}}{n_{il}} & if\ n_{ij} \neq 0 \\ 0 & if n_{ij} = 0 \end{cases} \tag{3.12}$$

where $n_{il,j}$ is the total number of times that power demand state is transited from P_{dem}^{i} to P_{dem}^{j} when the vehicle speed state is v^{l}, and n_{il} is the total number of times that power demand is P_{dem}^{i} and vehicle speed state is v^{l}. Figure 3.7 shows an example of TPM obtained from a real–world driving cycle with sampling frequency 10 Hz. It presents that TPM mainly has high value in diagonal area, in which P_{dem} value at current step and P_{dem} value at next step are similar. This is obvious that power demand of driver is not changed suddenly, therefore there is high chance for P_{dem} will transit to near value. The driving pattern of the driver could be reflected in this way into the TPM, e.g. the transition probability is widely distributed for the driving pattern of severe acceleration and deceleration and on the contrary, it will be mainly concentrated on the diagonal direction for the driving pattern of less acceleration and deceleration.

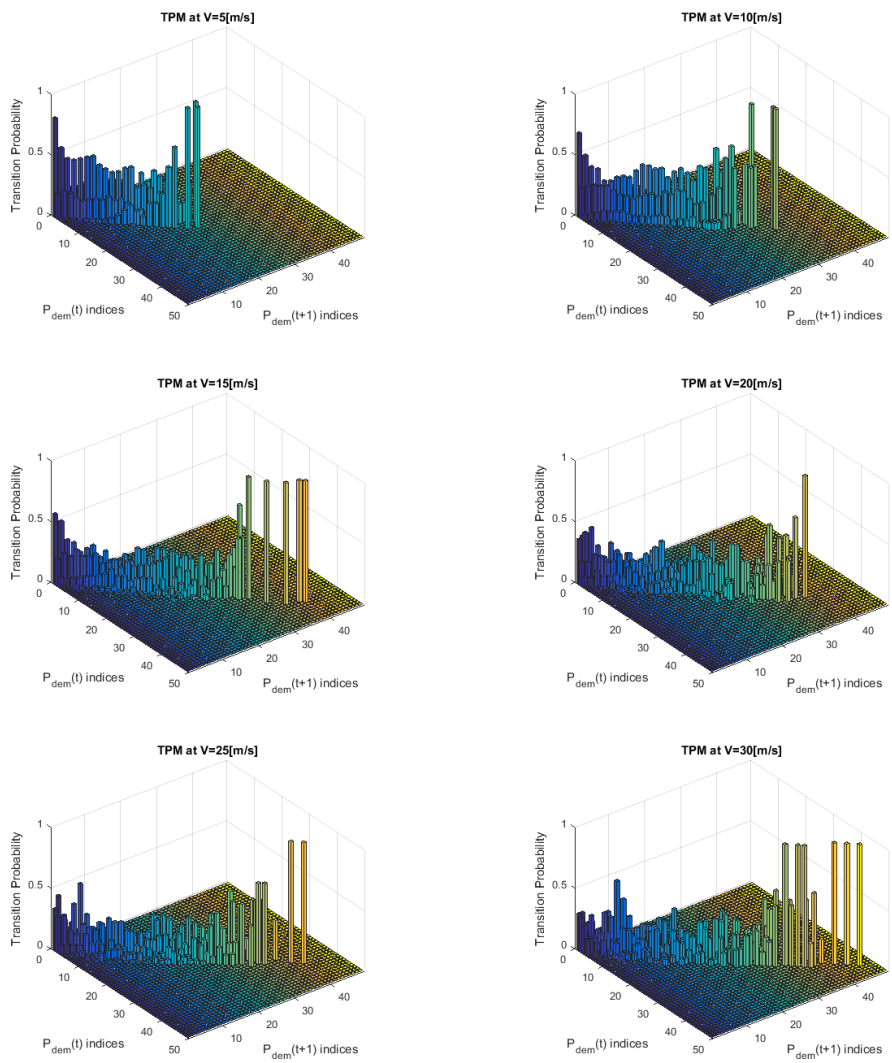


Figure 3.7 Transition probability matrix at different vehicle speeds

However, TPM is formed from historic data, which is not future speed profile. The assumption in SDP is that the TPM in the future driving cycle will be not totally different with that of historic speed profile, and the control policy obtained from the TPM of the historic speed profile could generally fit into future speed profile as well.

3.4 Stochastic Dynamic Programming

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

$$\begin{aligned} \min \quad & J_\pi(x_0) = \lim_{N\to\infty} \mathop{\mathrm{E}}_{d_k}\left\{\sum_{k=0}^{N-1} \gamma^k g\big(x_k, \pi(x_k)\big)\right\} \\ \text{subject to} \quad & \\ & \omega_{eng,min} \le \omega_{eng}(k) \le \omega_{eng,max} \\ & T_{eng,min}(\omega_{eng}(k)) \le T_{eng}(k) \le T_{eng,max}(\omega_{eng}(k)) \\ & T_{mot,min}(\omega_m(k), SOC(k)) \le T_{mot}(k) \le T_{mot,max}(\omega_m(k), SOC(k)) \\ & SOC_{min} \le SOC(k) \le SOC_{max} \end{aligned} \tag{3.13}$$

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

$$P_e \in \left\{P_e^1, P_e^2, \dots, P_e^{N_u}\right\} \tag{3.14}$$

where N_u is the number of control input discretized. State variable x_k is composed of a 4−dimensional state space as given equation

$$x_k = [SOC, P_{dem}, v, E_{on}] \tag{3.15}$$

where SOC is battery state of charge, E_{on} is engine on/ off state. Same as DDP, engine on/off state is considered to avoid fuel consumption due to frequent engine on/off. Instantaneous cost incurred g is defined as equation

$$g = W_{fuel} + \zeta(SOC) + \beta \cdot \Delta E_{on} \tag{3.16}$$

where W_{fuel} is instantaneous fuel consumption and $\zeta(SOC)$ is term for penalizing SOC deviation for the charge sustenance as below

$$\zeta(SOC) = \begin{cases} \mu \cdot \left(SOC - SOC_{ref}\right)^2 & if\ SOC > SOC_{min} \\ C_{Penalty} & if\ SOC \le SOC_{min} \end{cases} \tag{3.17}$$

where μ and $C_{Penalty}$ positive constant value for SOC deviation. The underlying meaning of the SDP is that in SDP, overall expectation of the cost in infinite horizon is minimized instead of the finite horizon,

therefore control policy result is time-invariant, which can be easily implemented as a real-time vehicle controller. Note that final SOC constraint in DDP is moved into instantaneous cost, since optimal control problem is defined as infinite-horizon problem.

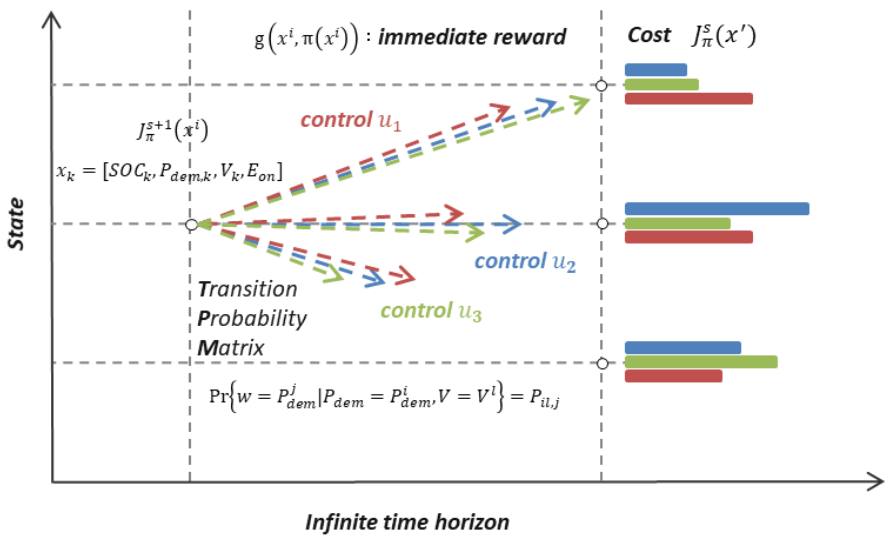


Figure 3.8 Calculation concept of stochastic dynamic programming

Figure 3.8 presents the calculation process of SDP. When control input is decided, a state transit to other states according to probability in TPM. With respect to the current state x_k, and the control input u, different immediate reward values are given. However even if same control input u is given, a state can be transferred to different states due to uncertainty caused by TPM. In this study, power demand P_{dem}, and vehicle speed v are transited according to TPM, which is not affected by control input. Other states,

SOC, and E_{on} is determined based on control input, thus state transition is partly random and partly deterministic. Cost of the state, $J_{\pi}^{s+1}(x^i)$ is decided by immediate reward $g\left(x^i, \pi(x^i)\right)$ and the expectation value of the cost of the possible next states. Control input u is determined to minimize the cost of the state, therefore, acquired control policy $\pi'(x^i)$ is determined. It is assumed that at each time step, P_{dem} is satisfied by the combination of engine power P_e, and motor power P_{mot} as shown in equation (3.17),

$$P_{dem} = P_e + P_{mot} \tag{3.18}$$

Therefore, the motor power P_{mot} is a dependent variable due to the power balance requirement assumption.

To solve optimization problem, the iterative method such as Policy Iteration or Value Iteration algorithm can be used [69]–[71]. For Policy Iteration method, firstly initial control policy should have stabilizing control policy $\pi_0(x)$, and in the policy evaluation step, the value of the current control policy is determined using Bellman equation as below

$$J_{\pi}^{s+1}(x^i) = g\left(x^i, \pi_s(x^i)\right) + \underset{d}{E}\{\gamma J_{\pi}^{s+1}(x')\} \tag{3.19}$$

where s is the number of iteration and x' is a next state. Equation

(3.19) can be solved using iterative method. Once the value converges, an improved policy is determined based on the value in policy improvement step as below equation

$$\pi'(x^i) = \underset{u \in U(x^i)}{\operatorname{argmin}} \left[g(x^i, u) + \underset{d}{\mathrm{E}}\{\gamma J_\pi(x')\} \right] \tag{3.20}$$

where $\pi'(x^i)$ is the updated control policy. Equation (3.19) and (3.20) is repeated until cost value converges. It has been proved that under certain conditions it converges to optimal control solution [70], [72], [73].

However, solving the equation (3.19) using the iterative method requires computation burdensome, thus Value Iteration method can be used instead of Policy Iteration. For Value Iteration method, initially optimal policy $\pi_s(x)$ and the discounted infinite horizon cost $J_\pi^{s-1}(x)$ for $s = 1$ are have random values. In the value update step, cost value are updated based on equation (3.21)

$$J_\pi^{s+1}(x^i) = g\left(x^i, \pi(x^i)\right) + \underset{d}{E}\{\gamma J_\pi^s(x')\} \tag{3.21}$$

where s is the number of iteration. Note that old value, $J_\pi^s(x')$ is used compared to the equation (3.19) and iterative method to solve (3.21) are not required. According to the updated cost value, an improved policy is determined in the next step, which is policy improvement

step, as same as Policy Iteration. This iterative method is repeated until cost value $J_{\pi}^{s}(x)$ converges as given equation (3.22)

$$|J_{\pi+1}(x) - J_{\pi}(x)| \leq \delta \tag{3.22}$$

where δ is a tolerance level.

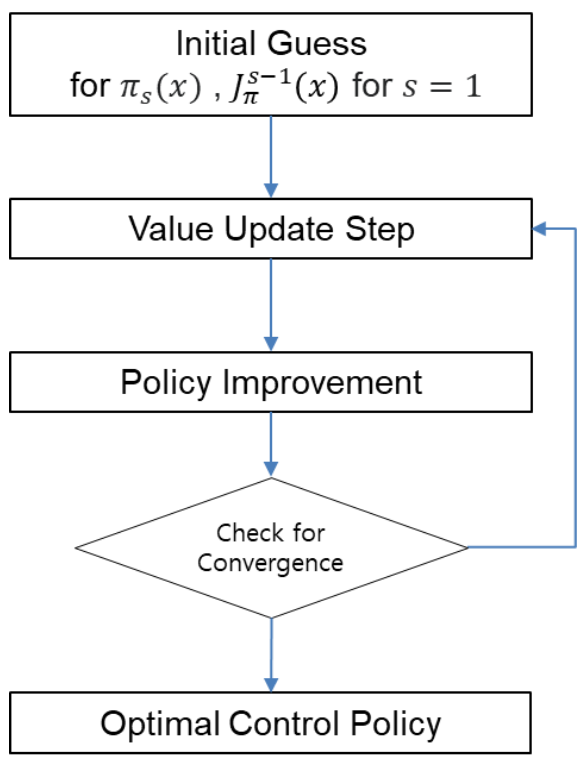


Figure 3.9 Value iteration algorithm

Figure 3.9 present Value Iteration algorithm procedure. In this paper, Value Iteration method is used. The underlying meaning of this process is that the Bellman equation and the Bellman optimality equation are fixed point equations, there is a unique fixed point $J_\pi(x^i)$, and equation (3.21) can be iterated with any initial value $J^0(x^i)$, and it will converge to $J_\pi(x^i)$. In the practical view, it is a time consuming process to evaluate the immediate reward, $g\left(x^i, \pi(x^i)\right)$ and the new state x' at every step of value iteration. Therefore, in this study, pre-calculated maps such as vehicle state, and fuel consumption are utilized based on vehicle modeling developed in the previous chapter. Figure 3.10 presents map for the new vehicle speed state according to current vehicle speed and power demand. Figure 3.11 shows that maximum engine torque and minimum engine torque map. Note that minimum engine torque is defined to satisfy a given power demand. According to this maximum and minimum engine torque, fuel consumption for a given control u can be calculated as given in figure 3.12. Also, since P_{mot} is determined according to a given control u, battery SOC at next step can be calculated as shown figure 3.13. according to constraints for powertrain, admissibility of each control u can be calculated also as presented figure 3.14. Based on these pre-calculated look-up tables, Value Iteration can be conducted to find optimal control policy with less time consumed.

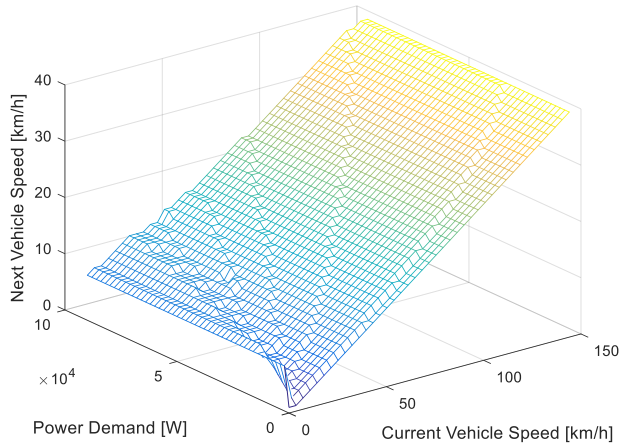
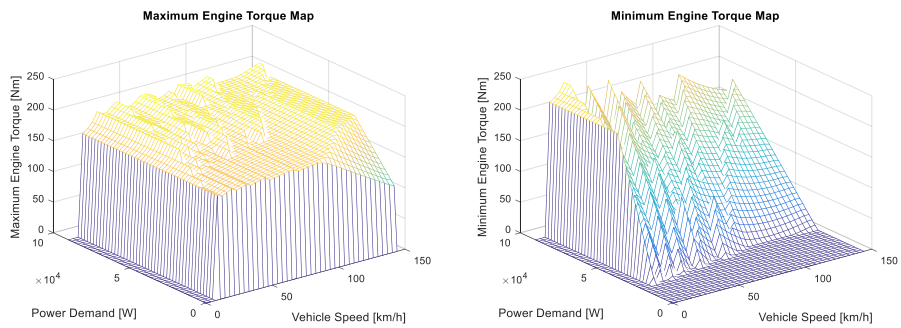
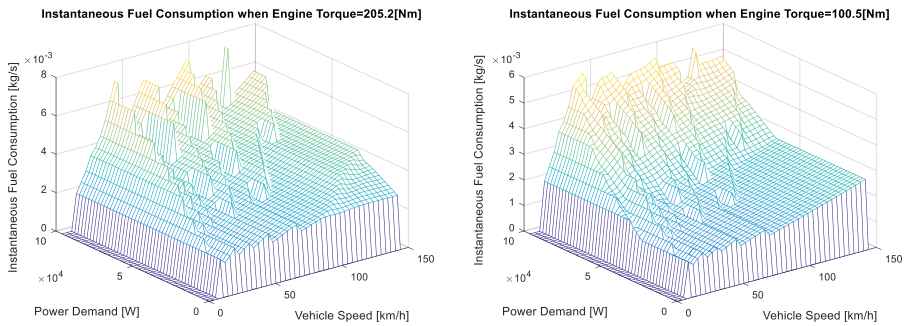


Figure 3.10 Vehicle speed map with respect to power demand and current speed



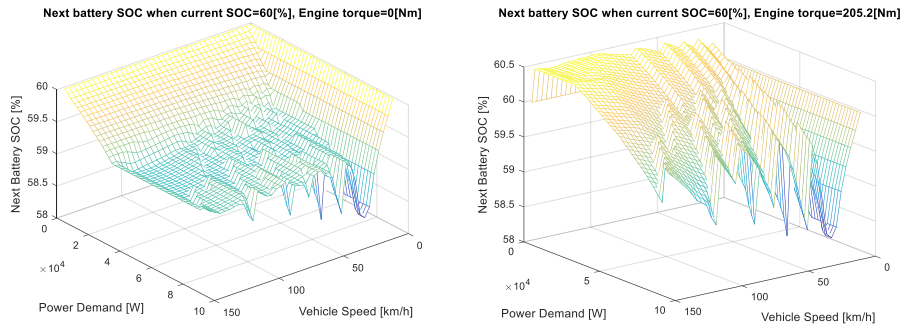
(a) (b)

Figure 3.11 Maximum engine torque map (a), and minimum engine torque map (b)



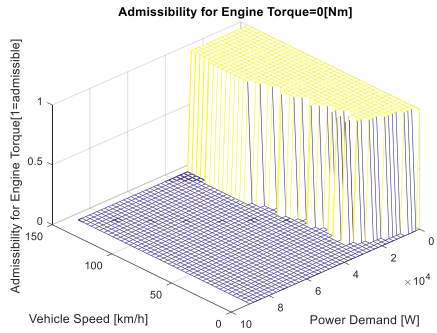
(a) $T_{eng} = 206.2\ [Nm]$ (b) $T_{eng} = 100.5\ [Nm]$

Figure 3.12 Instantaneous fuel consumption maps with respect to power demand and vehicle speed for different engine torques

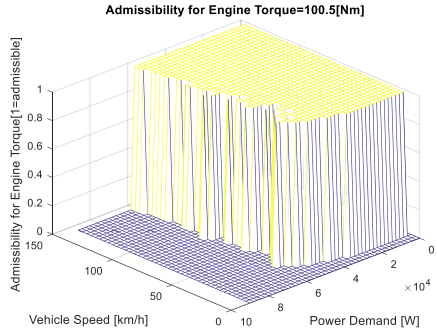


(a) $SOC = 60\ [\%],\ T_{eng} = 0[Nm]$ (b) $SOC = 60\ [\%],\ T_{eng} = 0[Nm]$

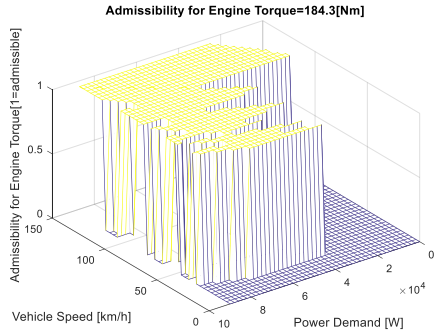
Figure 3.13 Next battery SOC map with respect to power demand and current vehicle speed for the current SOC and the engine torque value



(a) $T_{eng} = 0 \ [Nm]$



(b) $T_{eng} = 100.5 \ [Nm]$



(c) $T_{eng} = 184.3 \ [Nm]$

Figure 3.14 Admissible engine control input matrix with respect to power demand and vehicle speed for the given engine torque

As a result of SDP, Power-Split-Ratio (PSR) can be obtained as a function of power demand, battery SOC, vehicle speed, and engine on-off status. PSR is defined as equation (3.23). Note that in this study, parallel type HEV is used, therefore PSR can be also defined as equation (3.24)

$$PSR \equiv P_e / P_{dem} \tag{3.23}$$

$$PSR \equiv T_{eng} / T_{whl} \tag{3.24}$$

Typical PSR line according to power demand is as shown in Figure 3.15. PSR could be divided into 4 parts regarding power demand. Generally, when power demand is low, engine is not used, thus PSR values is zero. Once power demand is above a certain value, engine is turned on and it provides more power than needed for propelling, thus PSR becomes more than 1. On the engine mode, only engine is operated to fulfill required power demand, and when power demand is high, motor assist engine to produce needed power, thus PSR becomes less than 1.

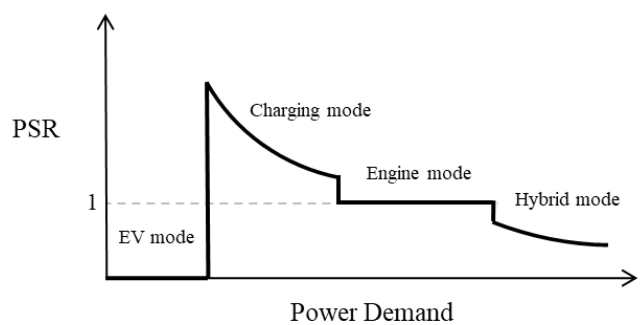
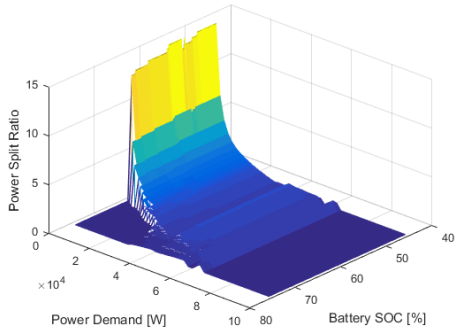
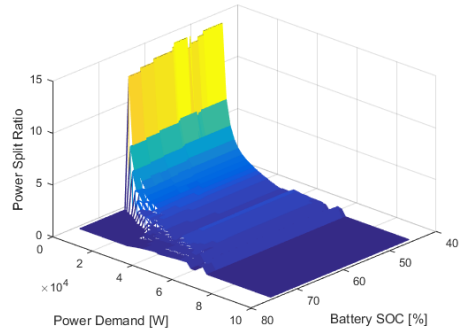


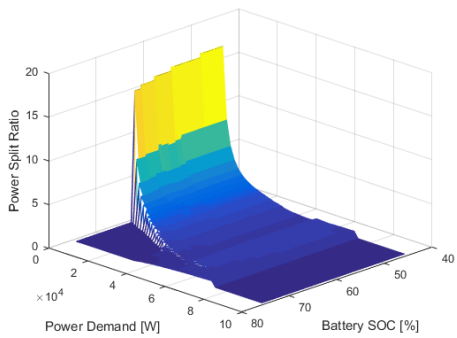
Figure 3.15 Typical PSR line according to power demand



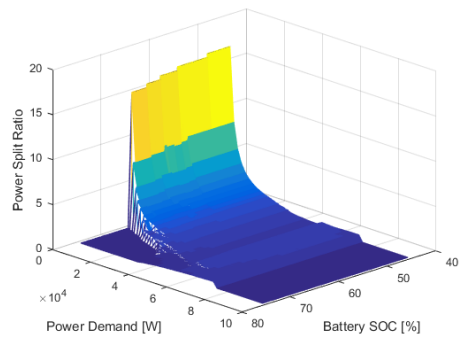
(a) $v = 10\ [m/s],\ E_{on} = 0$



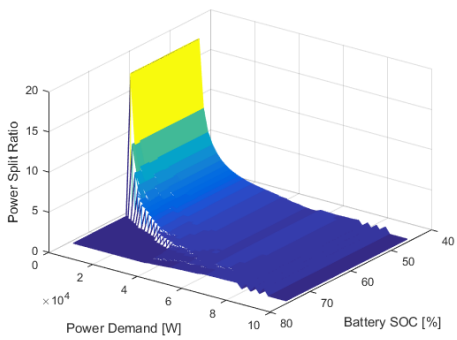
(b) $v = 10\ [m/s],\ E_{on} = 1$



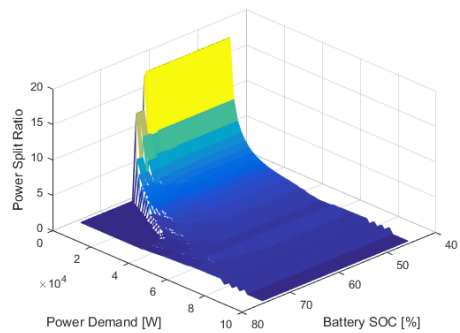
(c) $v = 20\ [m/s],\ E_{on} = 0$



(d) $v = 20\ [m/s],\ E_{on} = 1$



(e) $v = 30\ [m/s],\ E_{on} = 0$



(f) $v = 30\ [m/s],\ E_{on} = 1$

Figure 3.16 PSR line extracted from SDP

Figure 3.16 presents example of PSR map, given as result from SDP. PSR map is given as 4−dimensional map which is function of battery SOC, SOC power demand, P_{dem}, vehicle speed, v and engine on/off status, E_{on}.

$$PSR = L(SOC, P_{dem}, v, E_{on}) \tag{3.25}$$

Note that when P_{dem} is low, PSR value is 0, which is EV mode, and as power demand increases engine is turned on, which is charging mode. When SOC value is high, PSR value becomes 0 for charge sustenance. Also, according to engine of/off status, PSR values changes, which give engine on/off hysteresis to avoid frequent engine on/off reflecting cost for turning engine on.

Figure 3.17 present battery SOC and power split ratio line. According to battery SOC value, power split ratio line changes that the power demand of engine turn on point tends to increase as battery SOC is high. Therefore, when battery SOC is high, engine hardly turns on. Also vehicle speed and power split ratio line is presented in Figure 3.18. As vehicle speed increases, engine power tends to increase, but not necessarily for high power demand. Engine turn on point with respect to power demand hardly changes according to vehicle speed change. Engine on/off and power split ratio line is shown in Figure 3.19. Engine on/off cost are considered in SDP, thus when engine is turned off, Power split ratio for low power demand is

zero, while once engine turns on, power split ratio value is high as engine is operated even though power demand is low.

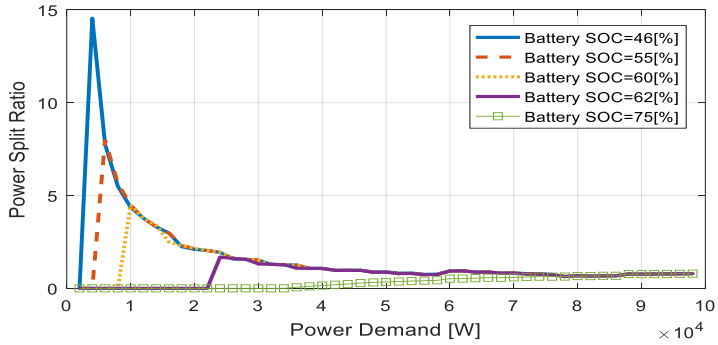


Figure 3.17 Battery SOC and power split ratio line

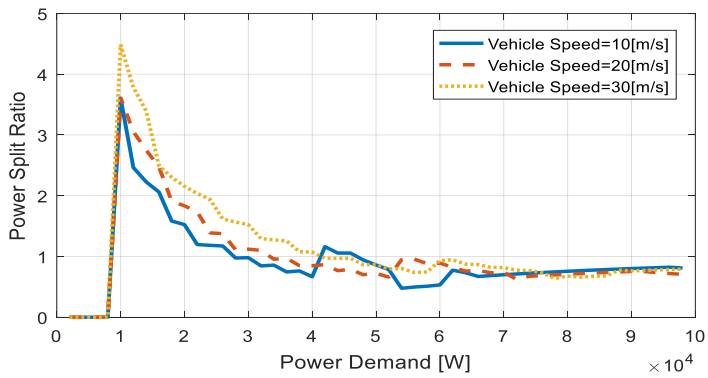


Figure 3.18 Vehicle speed and power split ratio line

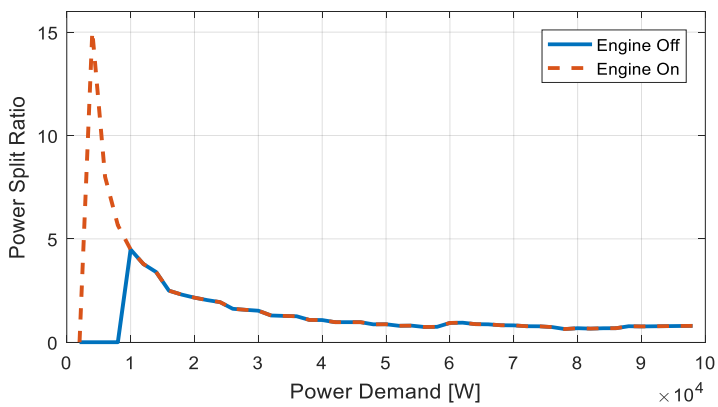


Figure 3.19 Engine on/off and power split ratio line

Once control policy is obtained, it can be implemented on the real-time vehicle simulator directly without any post-processing methodology unlike DDP, which needs analysis for extracting control law. In SDP, only state value of power demand, vehicle speed, battery SOC, engine on/off status needs to be detected and should be discretized to utilize PSR map.

SDP is a powerful tool that acquired optimal control policy can be implemented, and it utilizes driving cycle information as a form of probability matrix, thus can shows robust performance generally. The disadvantage of SDP is that the optimization problem defined in the given Markov chain modeling of driving cycle. Therefore, fuel economy performance of SDP cannot reach to that of DDP, even though entire driving cycle information is known in advance. Therefore, SDP result for a specific driving cycle scenario is not optimal. Also, if TPM has different characteristic with that of future driving cycle, the effectiveness of control strategy will be decrease. However, it is impossible to know entire future driving cycle in advance or it is hard to predict driving cycle precisely due to the change in traffic flow, or driving pattern of different drivers, thus SDP, which presents near-optimal performance, is practical and robust control strategy under uncertainties of the real-world driving cycle.

CHAPTER 4 REINFORCEMENT LEARNING BASED ENERGY MANAGEMENT STRATEGY

In the previous chapter, SDP algorithm was developed to derive the optimal control rules of the vehicle using statistical and probabilistic methods according to various driving situations. However, a drawback of SDP is that it requires TPM and optimization is conducted based on TPM, such that optimality of the control policy given as result is only valid for the given TPM. Therefore, if characteristic of current driving speed profile changes, TPM should be updated and iterative optimization process should be conducted again to get new control policy which is relevant to current driving condition. However, iteration processes used in SDP are computationally burdensome which cannot be used for online EMS. In this chapter, Reinforcement Learning (RL) technique is proposed to update control policy according to the change in vehicle driving condition. A new algorithm is proposed to overcome the drawbacks of SDP. In the newly proposed algorithm, the SDP framework is used as it is, but the information about the speed profile of the vehicle is updated without the TPM using the Q-learning techniques, which is one of the RL algorithm.

4.1 Introduction

In this study, RL algorithm is utilized for HEV control problem. RL is a type of machine learning. Machine learning is actively researched recently and it could be divided into supervised learning, semi-supervised learning, unsupervised learning, and RL depending on the learning method. Supervised learning utilizes pre-built training data to learn models, and semi-supervised learning is a method of using both training data and unorganized data for training. Unsupervised learning does not construct specific learning data, but it learns by clustering or analyzing the data. In case of RL, learning is done through feedback, giving appropriate compensation for the outcomes of the learning. The difference between supervised learning and RL is that unlike supervised learning in which it explicitly corrects undesired behaviors, RL is focuses on online performance more, which is one of the advantage that is suitable for application into real-time HEV control strategy.

RL is action-based learning, in which action is modified based on interactions with environment. RL is an area of machine learning inspired by behaviourist psychology and it also called approximate dynamic programming or neuro-dynamic programming in the control literature [74],[75]. The idea of RL is that control decision which presents good performance should be remembered through means of

reinforcement signal, such that it should be used again next time. RL is a more goal-oriented learning method than other machine learning techniques and well-suited especially to problems which have a long-term reward and short-term reward together. RL is an algorithm that has been actively used for robot control, elevator scheduling, network problem.

In many cases, the problem of RL is expressed in the Markov Decision Process (MDP). The implication of this is that the RL algorithm is closely related to dynamic programming. Figure 4.1. present concept of RL. At each time step t, the system is in some state s_t, and the agent, which is the controller, chooses any action a_t, which is admissible in state s_t. The environment responds by moving into a new state s', at the next time step $\mathrm{t}+1$ and agent receives corresponding reward $R_a(s,s')$ according to action a_t. The framework of RL is also based on Bellman equation as same as DDP or SDP. However, in RL, MDP is not necessarily fully known as modeled in SDP while Bellman equation could be used.

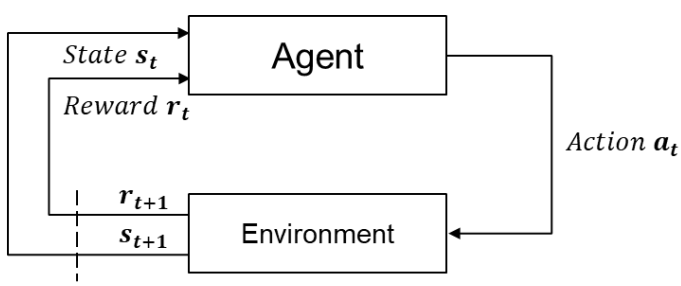


Figure 4.1 Concept of reinforcement learning

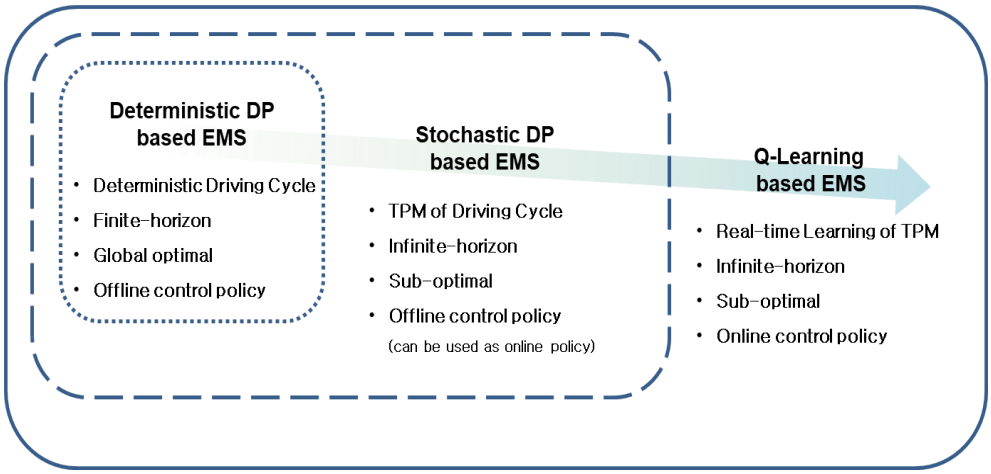


Figure 4.2 Comparison of DDP, SDP, and Q-learning based EMS.

In this study, Q-learning is used among many RL algorithms for HEV optimal control. Q-learning is a method that allows the learning of optimal control online, which is developed by Watkins and Werbos [76]-[78]. In Q-learning, Q function is learned using temporal difference method [79] based on interaction between controller and environment. Figure 4.2 shows a brief description of EMS based on DDP, SDP and Q-learning. The Q-learning based EMS studied in this paper can be regarded as a more advanced form in the existing SDP based EMS. In case of SDP, it can be used as an online control policy differently from DDP, but it has a disadvantage as an offline control policy in fundamentally as shown in the previous chapter. Q-learning based EMS is fundamentally an online control policy, which, like SDP, can define an optimal control problem and obtain a sub-optimal

solution in the infinite horizon. However, unlike SDP, it does not require an iterative optimization process using TPM, and has the advantage of obtaining an optimal control solution obtained from the SDP through a learning process adaptively. The subsequent section describes detail of Q−learning and development of EMS based on it.

4.2 Q-Learning based Energy Management Strategy

Infinite horizon problem can be written again as equation below

$$J_\pi(x_k) = \sum_{i=k}^{\infty} \gamma^{i-k} g(x_i, u_i) \tag{4.1}$$

Then, based on Bellman equation, equation (4.1) can be expressed recursively as equation (4.2) and (4.3)

$$J_\pi(x_k) = g(x_k, u_k) + \gamma \sum_{i=k+1}^{\infty} \gamma^{i-(k+1)} g(x_i, u_i) \tag{4.2}$$

$$J_\pi(x_k) = g\big(x_k, \boldsymbol{\pi}(x_k)\big) + \gamma J_\pi(x_{k+1}) \tag{4.3}$$

For optimality, right hand side and left hand side of equation (4.3) should be equal, or at least difference between them should be minimized. Discrete−time Hamiltonian can be defined as difference between them as equation as below

$$H(x_k, \boldsymbol{\pi}(x_k), \Delta J_k) = g\big(x_k, \boldsymbol{\pi}(x_k)\big) + \gamma J_\pi(x_{k+1}) - J_\pi(x_k) \tag{4.4}$$

where $\Delta J_k = \gamma J_\pi(x_{k+1}) - J_\pi(x_k)$. Discrete−time Hamiltonian is also called as Temporal Difference (TD) Error in temporal difference learning. TD learning is a type of model−free RL method, in which it samples from the interaction with the environment, such as Monte Carlo methods, and simultaneously update is conducted based on current estimates, which is based on Dynamic programming methods. TD error e_k can be written again as equation below

$$e_k = g\big(x_k, \boldsymbol{\pi}(x_k)\big) + \gamma J_\pi(x_{k+1}) - J_\pi(x_k) \tag{4.5}$$

Then $J_\pi(x_k)$ can be updated based on this TD error with learning rate α as below equation

$$J_\pi(x_k) \leftarrow J_\pi(x_k) + \alpha(g\big(x_k, \boldsymbol{\pi}(x_k)\big) + \gamma J_\pi(x_{k+1}) - J_\pi(x_k)) \tag{4.6}$$

Learning rate α is from 0 to 1, which means for 0, it learns nothing while for 1, only the most recent learning knowledge is used for the update. In this way, TD error can update the difference between the estimated reward and the actual reward. To utilize this update method directly, action and value fair could be defined as Q function value, which is action−value function as equation below

$$Q_\pi(x_k, u_k) = g(x_k, u_k) + \gamma J_\pi(x_{k+1}) \tag{4.7}$$

Equation (4.7) means that $Q_\pi(x_k, u_k)$ is the value including immediate reward $g(x_k, u_k)$, which is the immediate cost when state is x_k, and control u_k is chosen, and discounted cost of next state x_{k+1}, which follow control policy π. Also, optimal $Q^*(x_k, u_k)$ value can be defined as below

$$Q^*(x_k, u_k) = g(x_k, u_k) + \gamma J^*(x_{k+1}) \tag{4.8}$$

Then using Q function, optimal cost $J^*(x_k)$ and optimal control policy $\pi^*(x_k)$ can be found as below equation

$$J^*(x_k) = \min_u(Q^*(x_k, u)) \tag{4.9}$$

$$\pi^*(x_k) = \arg\min_u(Q^*(x_k, u)) \tag{4.10}$$

Also, Q function value can be updated as same way as equation (4.6) as below

$$Q(x_k, u_k) \leftarrow Q(x_k, u_k) + \alpha(g_k + \gamma \min_u Q(x_{k+1}, u) - Q(x_k, u_k)) \tag{4.11}$$

Equation (4.11) presents Q-learning algorithm. Q-learning is a specific TD algorithm used to learn Q function. The convergence of Q-learning algorithm for finite MDP has been proven by Watkins [76] based on stochastic approximation method.

Figure 4.3 presents concept of Q−learning calculation and figure 4.4 presents the pseudo code of Q−learning algorithm. When system is in some state x_k, (i.e. in this HEV control problem, when vehicle is in some state of SOC_k, $P_{dem,k}$, v_k, and $E_{on,k}$) control u_k is selected which has minimum Q value. According to action u_k, state x_k moves x_{k+1} with immediate reward g_k, then based on Q value at new state x_{k+1}, and g_k, Q value $Q(x_k, u_k)$ is updated to hold Bellman equation.

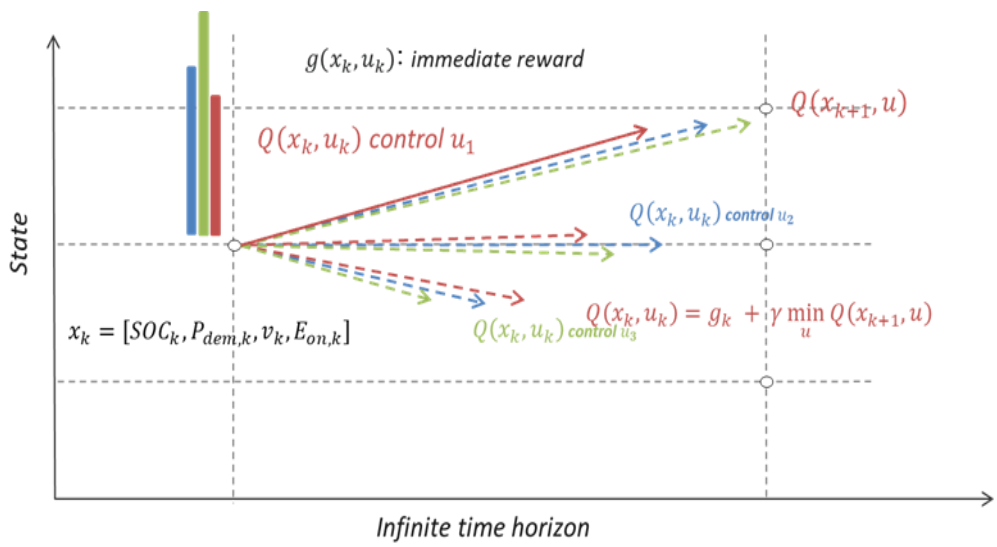


Figure 4.3 Concept of Q-learning calculation

Q-learning Algorithm

Initialize $Q(x_k, u_k)$

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

1. Choose action u_k based on $Q(x_k, u_k)$ (ϵ-greedy policy)
2. Taking action u_k, observe reward $g(x_k, u_k)$, state x_{k+1}
3. Update Q

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

4. $x_k \leftarrow x_{k+1}$

Figure 4.4 Pseudo code of Q-learning algorithm

When action u_k is chosen, generally an exploration strategy is needed for an agent to learn optimal control within all possible control input, since an agent requires a good control policy to deal optimally with all possible states in an environment, while these control policy could be learned only by being exposed to as many of those states as possible. One of the way to balancing exploration and exploitation is $\varepsilon-$greedy policy. $\varepsilon-$greedy policy is the policy that when action is chosen, with probability $1-\varepsilon$, best long-term effect action is selected and with probability of ε, uniformly random action is chosen.

Based on this Q-learning algorithm, this paper developed a power distribution control strategy for HEV available in real time. For the control problem definition, same as SDP, the optimization goal is to find the control input $u(k)$ to minimize cost function as below

$$min \quad J_\pi(x_0) = \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \gamma^k g\left(x_k, \pi(x_k)\right)\right\} \qquad (4.12)$$

subject to

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

where $x_k = [SOC_k,\ P_{dem,k},\ v_k,\ E_{on,k}]$, Control u is engine power P_e, g is instantaneous cost including fuel consumption, W_{fuel}, and penalty for SOC deviation and engine on/off penalty as same as SDP.

Figure 4.5 and Figure 4.6 presents concept of the new energy management strategy for HEV control and Pseudo code of it respectively.

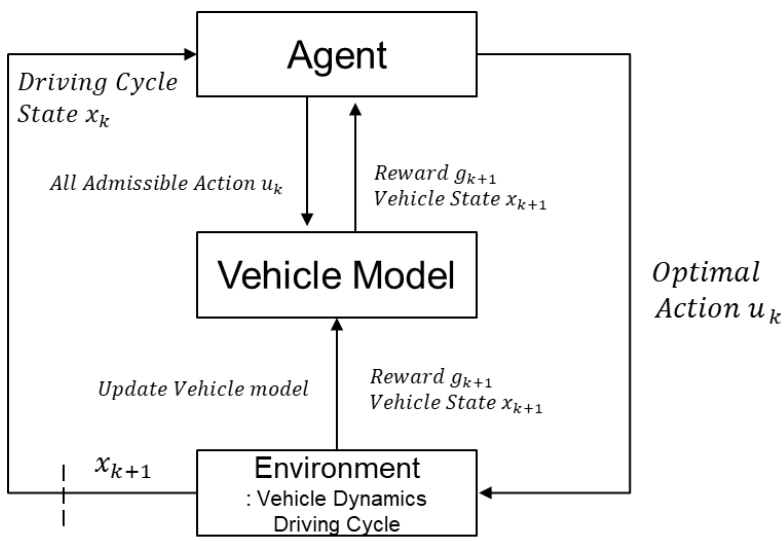


Figure 4.5 Concept of the new strategy for HEV control

Algorithm for HEV control

Initialize $Q(x_k, u_k)$

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

1. Choose action optimal u_k based on $Q(x_k, u_k)$

2. Taking action u_k, observe reward $g(x_k, u_k)$, state x_{k+1}

3.1 Update model based on observation

$$g(x_k, u_k) \leftarrow g(x_k, u_k) + \alpha\big(g_k - g(x_k, u_k)\big)$$
$$x_{k+1}(x_k, u_k) \leftarrow x_{k+1}(x_k, u_k) + \alpha\big(x_{k+1} - x_{k+1}(x_k, u_k)\big)$$

3.2 Update Q using model for all admissible action u_k

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

4. $x_k \leftarrow x_{k+1}$

Figure 4.6 Pseudo code of the new strategy for HEV control

In order to apply the Q−learning algorithm to the HEV control problem, the new energy management strategy based on stochastic optimal control framework is developed. In ε −greedy policy, it includes random property that for HEV control problem, fuel economy performance of the vehicle is decreased with this random selection of control input. Therefore, in newly proposed algorithm, control u_k is chosen based on $Q(x_k, u_k)$ value, however unlike conventional Q−learning algorithm, action u_k is selected only based on $Q(x_k, u_k)$ without any exploration strategy. Instead, Q function value is updated based on interaction between agent and vehicle model. In SDP, optimal control policy is acquired by searching all admissible control and state in next time step. Just like SDP, in newly

proposed algorithm, while optimal action u_k is chosen and implemented on the environment, agent updates Q function value by investigating all admissible action u_k based on vehicle mode. Reward g_{k+1} and vehicle state x_{k+1} (which are SOC_k, and $E_{on,k}$) according to set of action u_k is obtained using vehicle model and Q function value is updated by combining these data with driving cycle state x_{k+1} (which are $P_{dem,k}$, and v_k). On the other hand, vehicle model is updated by using information obtained from interaction between agent and environment as equation below

$$g(x_k, u_k) \leftarrow g(x_k, u_k) + \alpha\big(g_k - g(x_k, u_k)\big) \tag{4.13}$$

$$x_{k+1}(x_k, u_k) \leftarrow x_{k+1}(x_k, u_k) + \alpha\big(x_{k+1} - x_{k+1}(x_k, u_k)\big) \tag{4.14}$$

By defining the vehicle model in this way and updating it using the results of the interaction between the actual agent and the environment, it is possible to have a model-free property that is an advantage of the existing Q-learning. In other words, even if the model is not accurate, it can be modified through learning, which allows the optimal control to be explored. Vehicle model (battery SOC and fuel consumption) is given as 4-dimensional look-up table which is function of state and control as written in equations below

$$SOC_{k+1} = f_{soc}(SOC_k,\ P_{dem}, v, u) \tag{4.15}$$

$$W_{fuel} = f_{fuel}(P_{dem}, v, E_{on}, u) \tag{4.16}$$

Figure 4.7 and figure 4.8 present example of battery SOC model and fuel consumption model respectively. The advantage of the proposed algorithm is that it separates the vehicle model from the environment differently from the existing Q-learning based energy management strategy. In the case of the vehicle model, future vehicle state (fuel consumption and battery SOC) can be derived when certain control inputs are given with current vehicle state information. However, in the case of driving cycle information, it is not easy to accurately predict the change of the vehicle speed and the required power demand. In the case of vehicle powertrain, modeling is inserted to the control algorithm through modeling, but in the case of the vehicle driving cycle, the model is configured to learn based on the interaction of the agent and the environment as in the existing Q-learning based energy management strategy.

The biggest difference between the proposed algorithm and the existing SDP algorithm is whether to use TPM. In SDP, driving cycle information is expressed as TPM, which is defined in the optimal control problem as driving cycle information. However, in the proposed algorithm, instantaneous driving cycle information is updated using the Q function value and stored based on the Bellman equation as if the TPM is updated every moment. On the other hand, using the vehicle model, it is possible to derive the optimum control value by examining the vehicle state change and the compensation value according to all possible control inputs as in the SDP. Therefore,

it is possible to control the vehicle in real time by taking all the advantages of SDP and Q−learning.

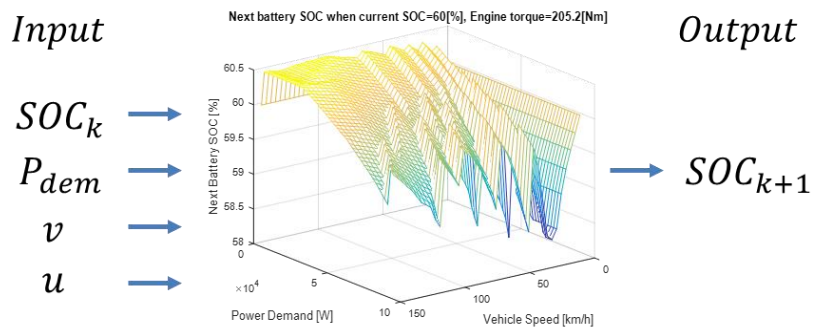


Figure 4.7 Battery SOC model

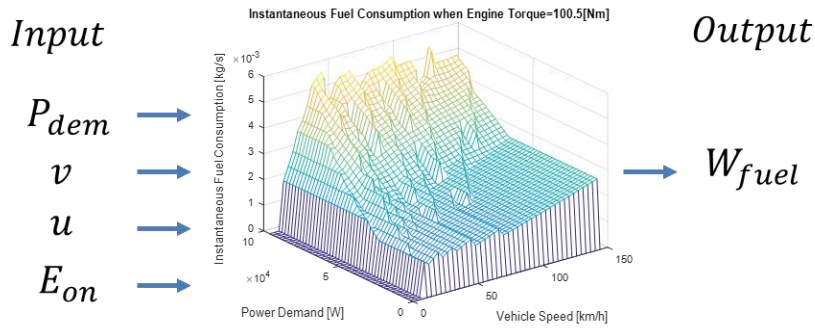


Figure 4.8 Fuel consumption model

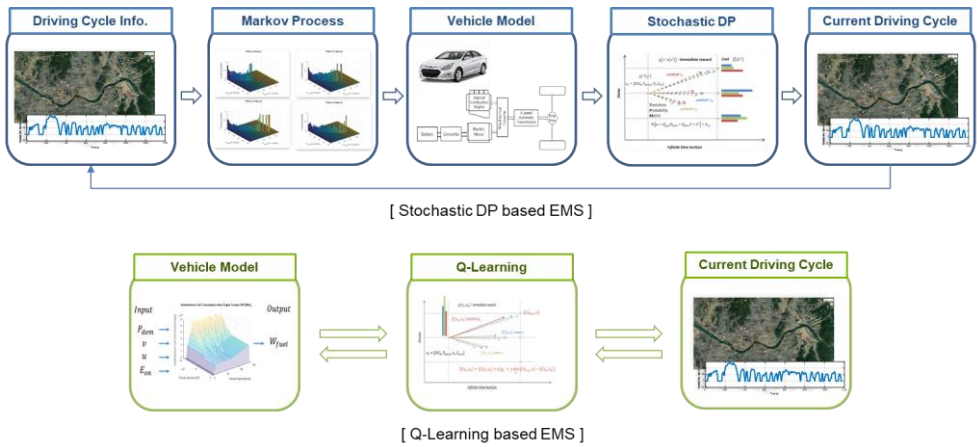


Figure 4.9 Comparison between SDP and Q-learning based energy management strategy

Figure 4.9 presents comparison between SDP and Q-learning based EMS. In addition, it is possible to use the initial value calculated by SDP for the initial values of Q-learning based EMS of HEVs, that initialization can be replaced by SDP when the Q-learning based algorithm is used for controlling the real-world vehicle.

CHAPTER 5 SIMULATION ANALYSIS

In this chapter, the newly developed EMS in this paper are simulated and tested for different driving cycles. Backward-looking vehicle simulator are used to verify the effectiveness of the proposed EMS, and simulation conducted using standard driving cycle and real-world driving cycle. For the comparison study, DDP and rule-based strategy are used. In case of the rule-based strategy, power follower strategy is used, in which engine produce required power according to battery SOC [80],[81]. Fuel economy result for the simulation modified to compensate electric energy use based on relation between ΔSOC and $\Delta \dot{m}$ [82], since the initial and final SOC has different value except DDP, in which final SOC value is constrained to be same as initial SOC.

5.1 Vehicle Simulation based on Stochastic Dynamic Programming based Energy Management Strategy

Firstly, SDP based EMS is simulated on standard driving cycle. In order to make TPM, the real-world driving cycle is used. The real-world driving cycle data is obtained from digital tachographs(DTGs) of Taxi in the city of Seoul. Figure 5.1 presents driving cycle data A,

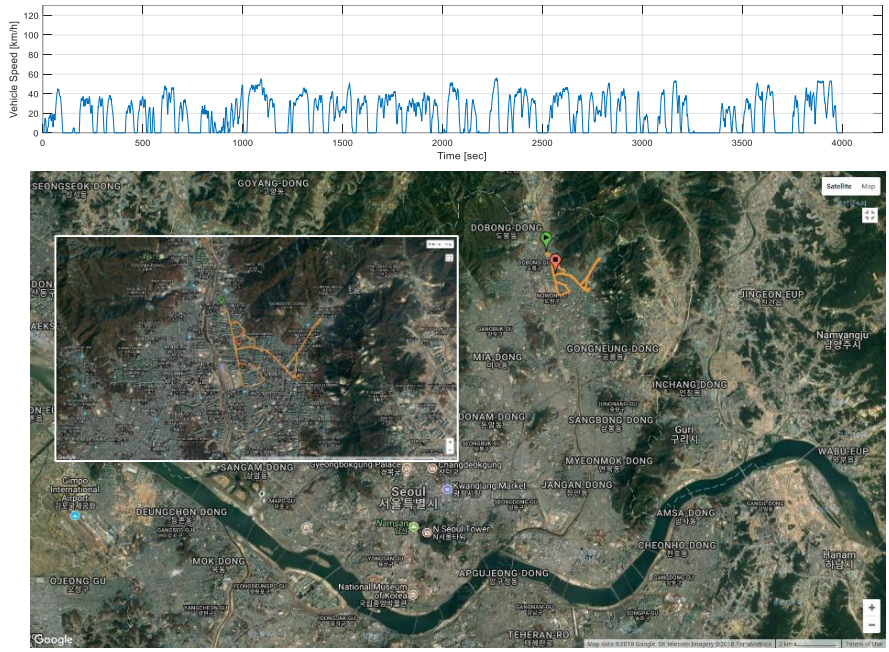


Figure 5.1 Real-world driving cycle A

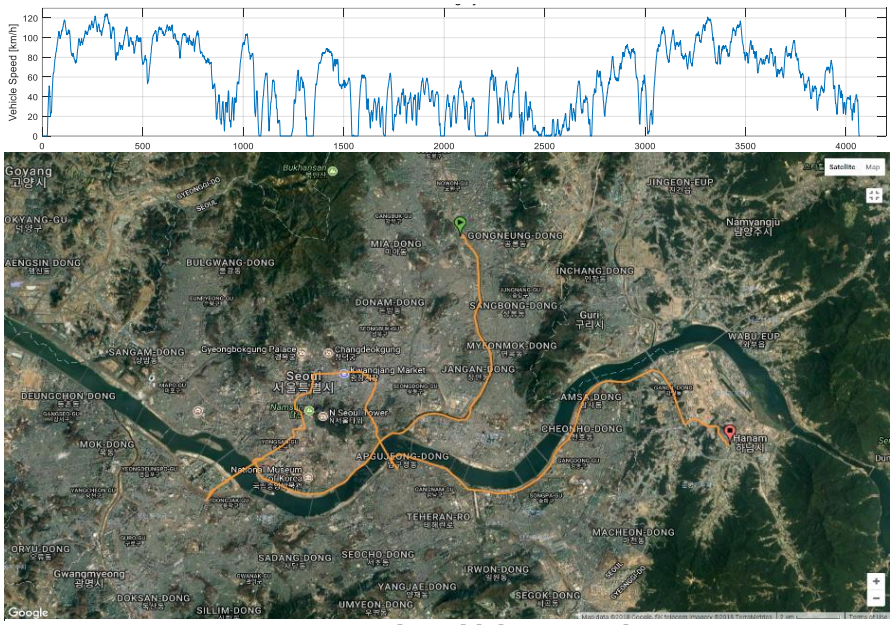


Figure 5.2 Real-world driving cycle B

which is a speed profile obtained from taxi driving cycle data in Seoul. The maximum speed of the driving cycle is less than 60 km/h, and length of the driving cycle is about 4000 seconds. Figure 5.2 presents driving cycle data B, that is also real-world driving cycle with maximum vehicle speed more than 120 km/h, obtained from the road along with the river. The length of the driving cycle is also about 4000 seconds. For SDP, it is important to have a lot of driving data to construct TPM, therefore real-world driving cycle data is used in this study. Based on this TPM, iterative calculation is conducted to get optimal control policy. In this study, each state and control is discretized as shown in Table 5.1. Note that a parallel hybrid powertrain is used, thus the control of engine torque, T_{eng} and engine power, P_e have same meaning. The acquired control policy as term of PSR map is applied in the backward-looking vehicle simulator and tested. Engine on/off penalty coefficient β is 0.0004, SOC penalty coefficient μ is 0.0001 , and $C_{Penalty}$ is 10. For target SOC value SOC_{ref} , 0.60 is used and discount factor γ is 0.99995 in this simulation.

Table 5.1 Calculation assumption for SDP

	Minimum value	Maximum value	Interval
Vehicle speed, v $[m/s]$	0	40	1
Battery SOC, SOC [%]	45	75	0.01
Power demand, P_{dem} $[W]$	0	96000	2000
Engine on/off, E_{on}	1 (off)	2 (on)	-
Engine Torque T_{eng} $[Nm]$	0	205.2	4.1

Table 5.2 Equivalent fuel economy [km/l] result for SDP simulation on various driving cycle (% compared to DDP result)

Algorithm / TPM Cycle	Driving Cycle					Average
	UDDS	HWFET	JN1015	WLTC	NEDC	
DDP	26.1	26.2	26.3	24.6	24.6	25.6
SDP / DTGs A +DTGs B	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 (83.4)	22.8 (87.0)	21.8 (82.9)	21.0 (85.4)	20.5 (83.3)	21.5 (84.1)

Table 5.2 presents simulation results for DDP, SDP and rule-based strategy. The results show that DDP shows optimal fuel efficiency, average 25.6 km/l. In the case of SDP, average fuel economy is 24.1 km/l, which is 94.1 % of DDP result. The SDP fuel economy results are less than the DDP optimum fuel economy results, but improved compared to the existing rule-based ones, which is average 21.5 km/l. Figure 5.3-5.4 present simulation results of DDP, SDP and rule-based strategy for UDDS driving cycle and figure 5.5-5.6 for HWFET driving cycle. As shown in figure 5.3, and 5.5, the result presents that DDP operates engine at Optimal Operating Line (OOL) mostly, which is a lowest brake specific fuel consumption(BSFC) line with respect to engine speed. SDP also utilize engine near OOL mostly, but also together with other area for change sustenance.

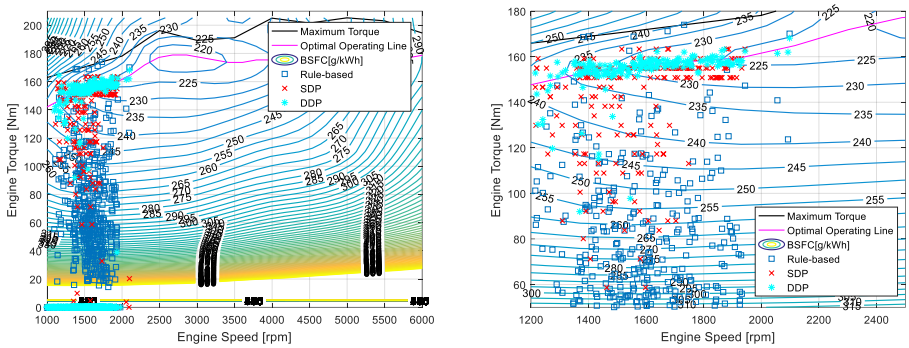


Figure 5.3 Simulation results of engine operating point for UDDS using SDP

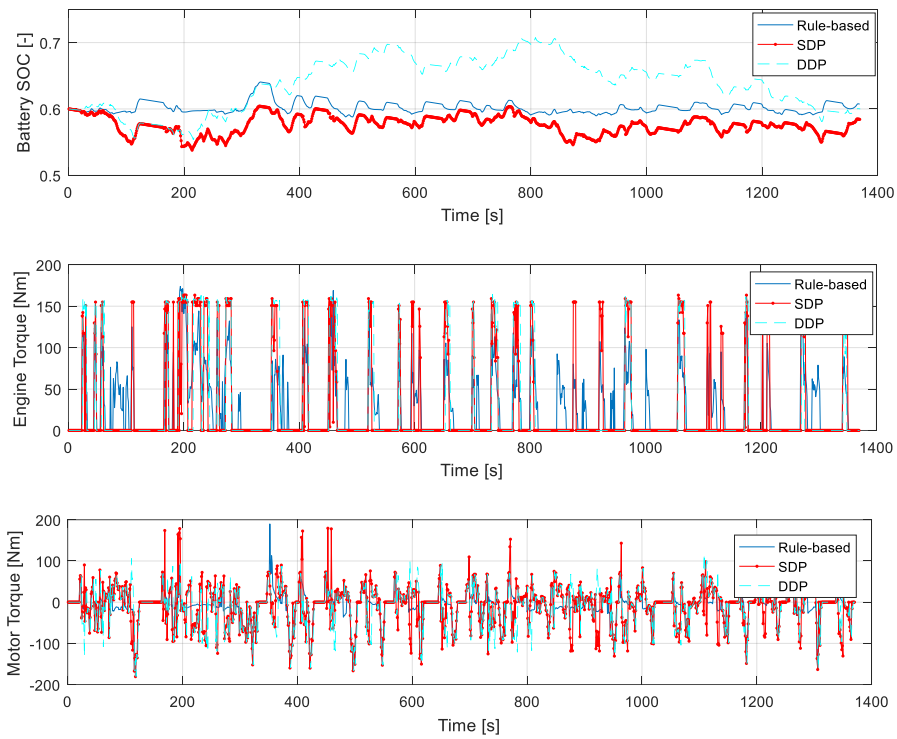


Figure 5.4 Simulation results for UDDS using SDP

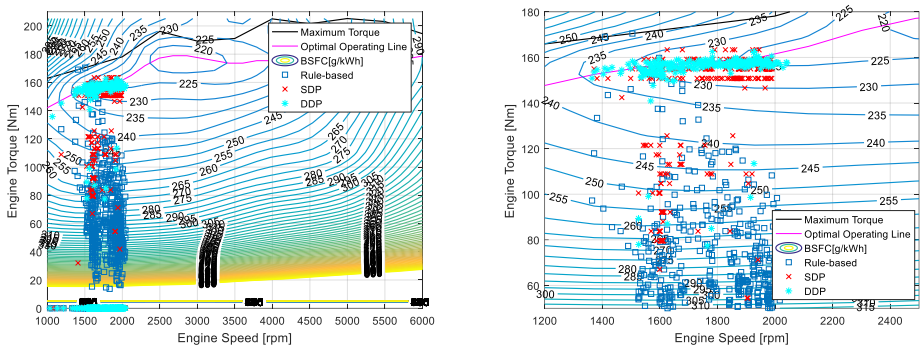


Figure 5.5 Simulation results of engine operating point for HWFET using SDP

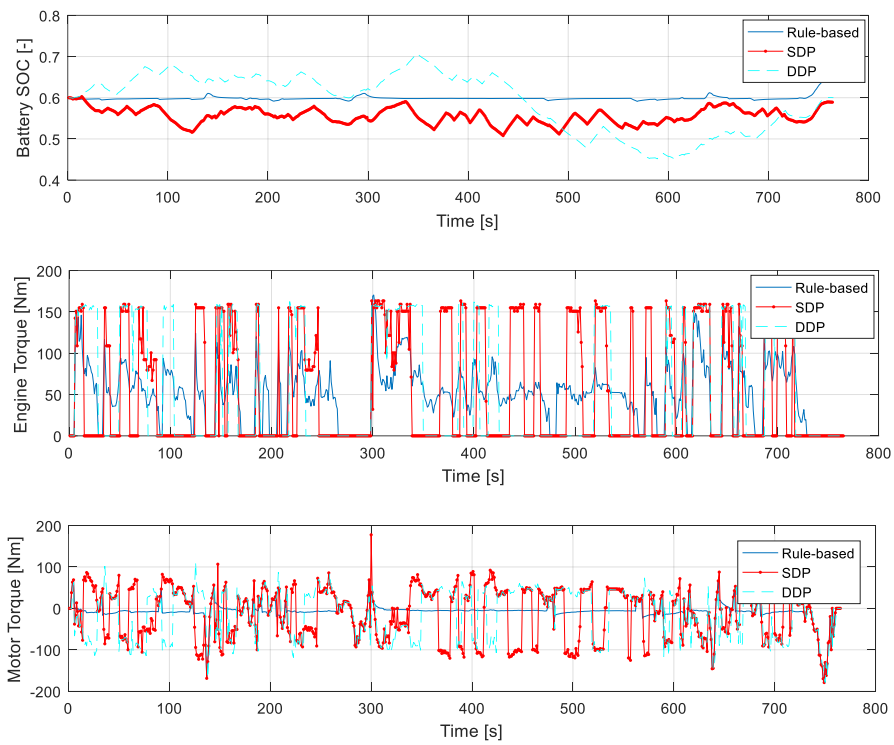


Figure 5.6 Simulation result for HWFET using SDP

DDP suggests optimal solution, but SOC deviation is not considered and only boundary value and final SOC value is considered as the optimal problem is defined, thus battery is extremely utilized to maximize fuel economy performance according to driving cycle profile. On the other hand, for SDP and rule-based strategy, the entire driving cycle information is not given, thus control strategy is designed as battery SOC deviation is penalized, which result in narrow range of battery SOC use as shown figure 5.4 and figure 5.6.

Secondly, another simulation is conducted to show the relationship between characteristic of TPM and the driving cycle to be driven in the future in terms of fuel economy performance of the vehicle. In this simulation, we examined how the fuel economy performance changes according to each driving cycle, when the driving cycle used to construct the TPM for optimization process in SDP is different with the driving cycle being driven. First, SDP was used to optimize the control policy using DTGs A driving cycle data and DTGs B driving cycle data, respectively, and an optimal control strategy was derived for each case. After that the simulations were conducted on DTGs A and DTGs B driving cycles, respectively, and the tendency of vehicle fuel consumption results was observed. Table 5.3 shows the results of fuel economy for the simulation. It was confirmed that the fuel economy was the highest when the control policy was extracted using the TPM based on the same driving cycle in both driving cycles. In this case, the fuel efficiency of the vehicle reached 94.9% (DTGs A) and 97.2% (DTGs B), of optimal fuel economy acquired from DDP, respectively. It is confirmed that the fuel consumption is reduced in the case of using the TPM obtained in the other driving cycle.

Table 5.3 Equivalent fuel economy [km/l] result for different TPM use (% compared to DDP result)

Algorithm / TPM Cycle		Driving Cycle	
		DTGs A	**DTGs B**
DDP		23.5	24.6
SDP	DTGs A	22.3 (94.9)	22.3 (90.7)
	DTGs B	22.0 (93.6)	23.9 (97.2)
Rule-based		19.1 (81.3)	21.7 (88.2)

The simulation results indicate that the SDP needs to know the cycle information to get higher fuel economy results. Control policy obtained from SDP could be easily implemented as real-time control, but it essentially shows the characteristics of SDP, which is an offline line policy. Therefore, in order to derive the control rule having the cycle dependent characteristic, it is necessary to recognize the pattern of the driving cycle and update the control rule accordingly. In this study, we developed an algorithm to update the control rules by reflecting characteristics of driving cycle being driven in real-time using RL. In the next chapter, the performance of the newly proposed RL-based strategy is confirmed through various simulations.

5.2 Vehicle Simulation using Reinforcement Learning based Energy Management Strategy

Vehicle simulation using RL-based strategy is conducted. Backward-looking vehicle simulation are used to verify the effectiveness of the proposed EMS and different driving cycle are used for the vehicle simulation. For comparison study, deterministic dynamic programming and rule-based strategy are used. The state variables and control are discretized as they as in SDP, and same parameter values are used together with learning rate 0.95.

Firstly, UDDS driving cycle is used for learning and UDDS driving cycle is used for the simulation also. Figure 5.7 presents learning curve, in which cumulative reward decreased rapidly as iteration is repeated. As the iterative learning continues, the cumulative reward value becomes smaller and convergence can be confirmed. Figure 5. 8 present battery SOC result for each simulation for UDDS. Firstly, there is nothing previously learned, thus battery SOC is decreased since controller will select control to minimize immediate fuel consumption and SOC deviation penalty only without considering discounted cost of next state, thus battery SOC value becomes smaller to reach minimum SOC value 0.45. However, as learning process is repeated battery SOC is sustained near target battery SOC value.

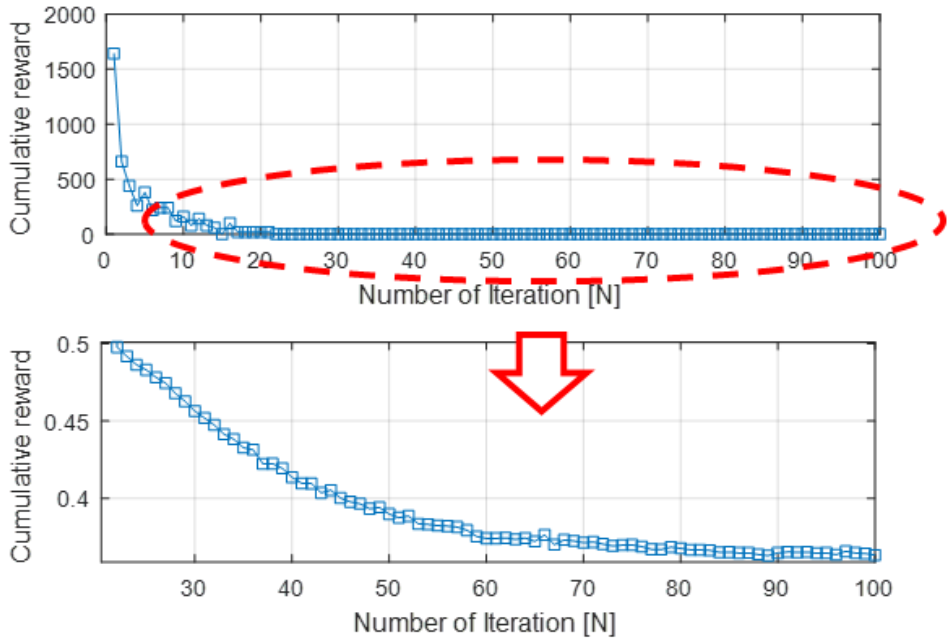


Figure 5.7 Learning-curve for UDDS driving cycle

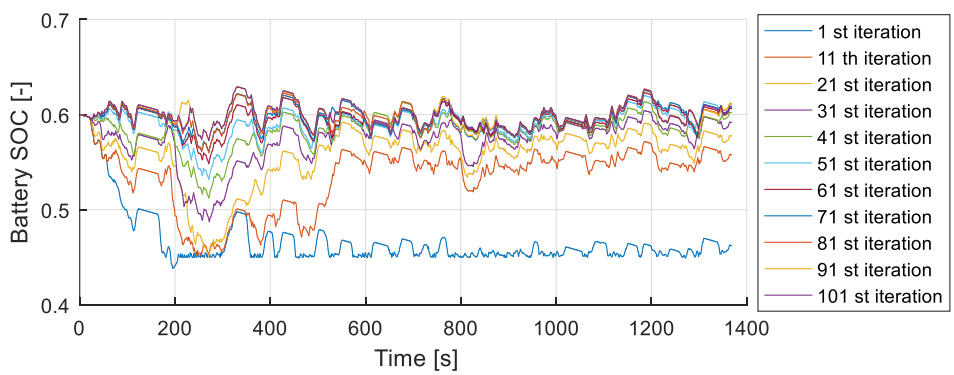


Figure 5.8 Battery SOC trajectory change according to learning

In the same way, the simulation for HWFET is conducted. The strategy is simulated on HWFET driving cycle repeatedly for learning and fuel efficiency performance is measured. Table 5.4 presents fuel economy performance of the strategy which is trained for each cycle separately. Simulation result shows that in case of UDDS driving cycle, RL-based strategy present fuel economy performance of 24.9 km/l which is 95.4 % of optimal fuel efficiency acquired in DDP. In case of HWFET driving cycle, RL-based strategy presents 25.7% which is 98.1% of DDP result. In both cases, we can confirm that the results of the RL-based strategy outweigh the results of the rule-based strategy. However, fuel economy result of RL-based strategy cannot reach to that of DDP even though it is trained using the driving cycle information repeatedly. This is because of the optimization problem is defined as infinite time horizon rather than finite driving cycle, thus derived optimal control is not for deterministic case as same as SDP.

Table 5.4 Equivalent fuel economy [km/l] result for RL-based strategy for UDDS and HWFET (% compared to DDP result)

Algorithm	**Driving Cycle**	
	UDDS	**HWFET**
DDP	26.1	26.2
RL-based (Trained for each cycle)	24.9 (95.4)	25.7 (98.1)
Rule-based	21.5 (82.4)	22.8 (87.0)

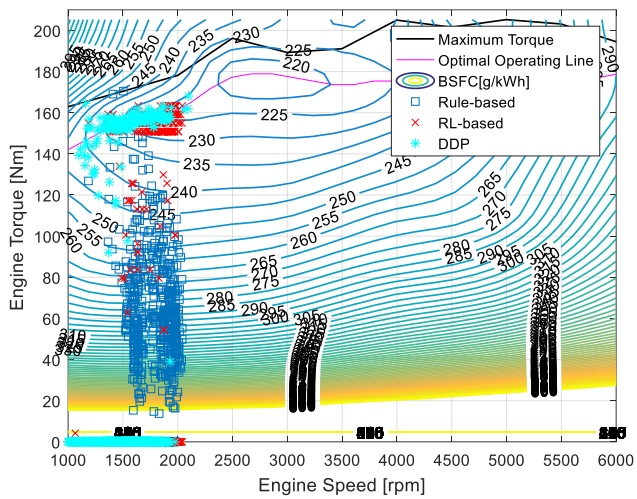


Figure 5.9 Simulation results of engine operating point for UDDS using RL-based strategy

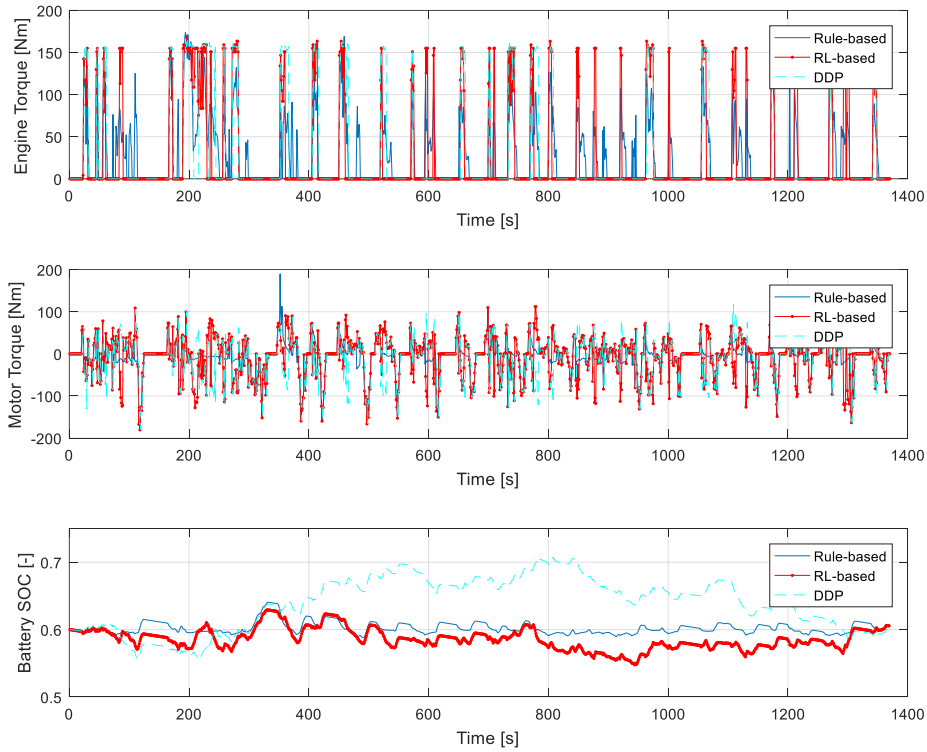


Figure 5.10 Simulation results for UDDS using RL-based strategy

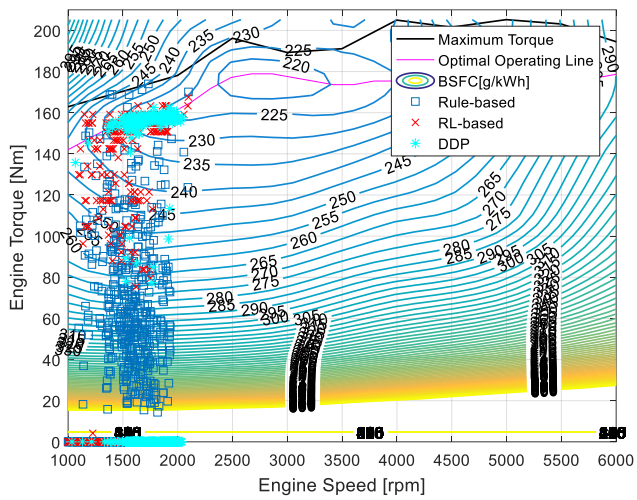


Figure 5.11 Simulation results of engine operating point for HWFET using RL-based strategy

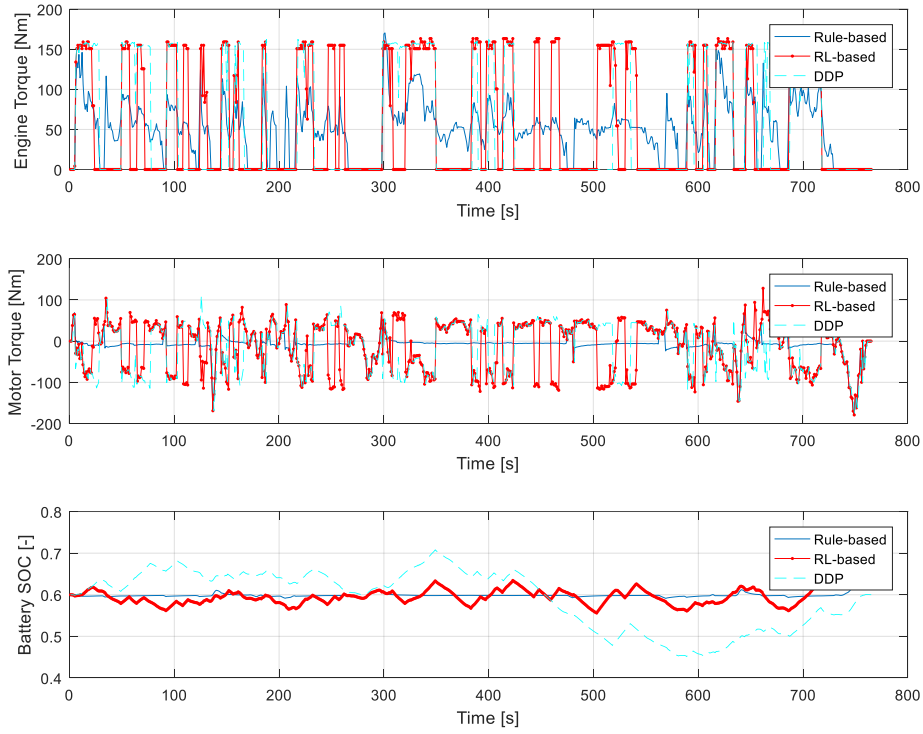


Figure 5.12 Simulation results for HWFET using RL-based strategy

Figure 5.9-10 and figure 5.11-12 present simulation results for UDDS and HWFET driving cycle separately. It is shown that for both driving cycle, RL-based strategy's engine operating points are very similar to that of DDP, while battery SOC range is relatively narrow compared to that of DDP as result of change sustaining constraints.

On the other hand, the learning ability of the proposed strategy was also tested through simulation. In this simulation, UDDS and HWFET driving cycle are used for learning, and then it is learned again in different driving cycles (HWFET and UDDS) to determine whether new learning occurs well on existing learned data. Figure 5.13 presents equivalent fuel economy results as learning occurs for HWFET driving cycle using pre-learned data with UDDS driving cycle. It is shown that equivalent fuel economy is increased as iteration is repeated. Figure 5.14 presents battery SOC trajectory results as learning is repeated in which the first iteration, the battery SOC is decreased to reach the minimum value and the charge sustaining is not performed well. However, as the learning is repeated, it can be confirmed that the battery SOC is maintained at the target SOC value, which is 0.6. Similarly, figure 5.15 and 5.16 show the process of re-learning UDDS driving cycle on data learned in the HWFET driving cycle. Also, it can be confirmed that the fuel consumption value increases as the learning is repeated and converges to a constant value. Also, the battery SOC can be confirmed that the charge sustaining fails at first and then it is improved as learning is repeated.

Table 5.5 shows the fuel efficiency performance for the UDDS, HWFET cycle of the learned strategy with different cycles. In case

of UDDS driving cycle, the case of learning in only UDDS shows the best fuel economy and also for the case of learning in HWFET only shows the best fuel economy for the HWFET driving cycle. However, the performance of fuel efficiency in case of simulation in different cycle presents reduced performance. However, when two cycles are learned both, that is, when the strategy learned in the HWFET is learned again in the UDDS, or when the strategy learned in the UDDS and learned in HWFET again shows similar or close fuel economy performance compared to best fuel efficiency.

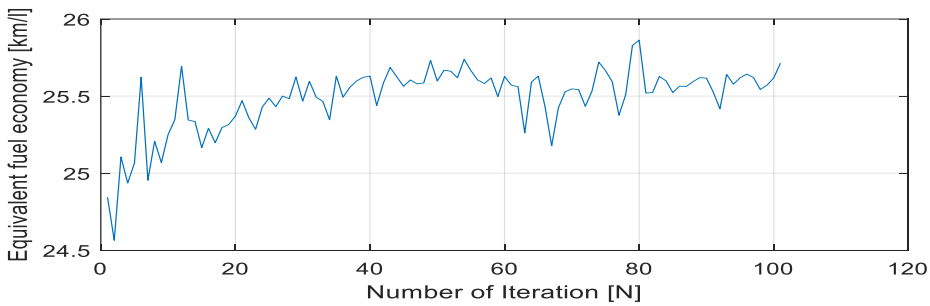


Figure 5.13 Equivalent fuel economy results for re-learning/ HWFET

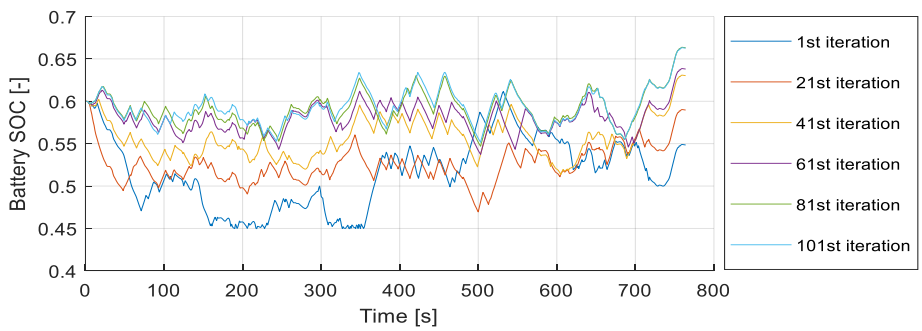


Figure 5.14 Battery SOC trajectory results for re-learning/ HWFET

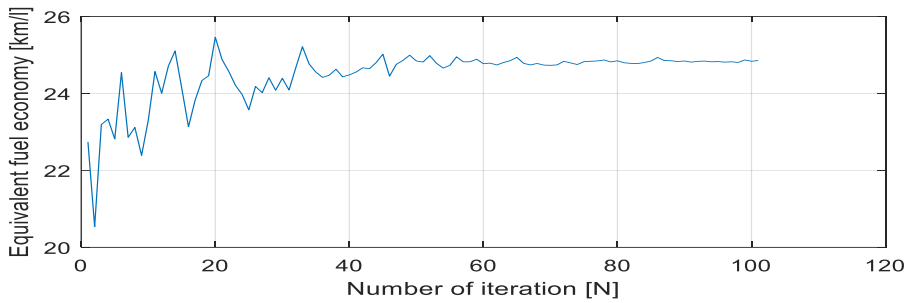


Figure 5.15 Equivalent fuel economy results for re-learning/ UDDS

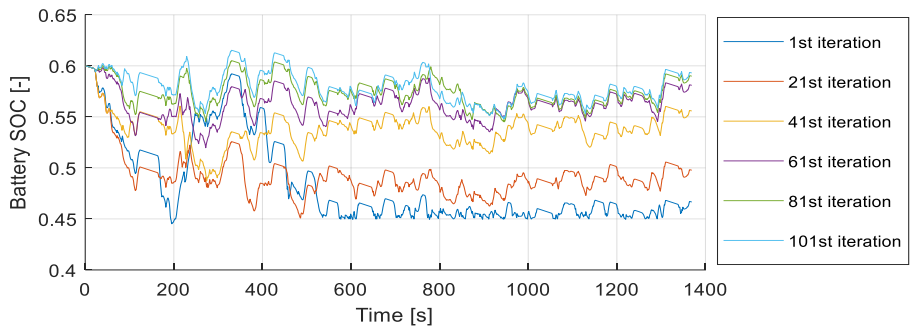


Figure 5.16 Battery SOC trajectory results for re-learning/ HWFET

Table 5.5 Equivalent fuel economy result for RL-based strategy [km/l] for re-learning (% compared to DDP result)

Algorithm/Trained Cycle		**Driving Cycle**	
		UDDS	**HWFET**
DDP		26.1	26.2
RL-based	UDDS	24.9 (95.4)	24.6 (93.9)
	HWFET	23.9 (91.6)	25.7 (98.0)
	HWFET + UDDS	24.9 (95.4)	25.4 (96.9)
	UDDS + HWFET	24.9 (95.4)	25.7 (98.1)
Rule-based		21.5 (82.4)	22.8 (87.0)

In the case of the above-mentioned simulation, the control strategy is tested that a certain driving cycle is learned and it is simulated in the corresponding driving cycle. Therefore, it is difficult to say that they show the fuel efficiency performance of the vehicle when the speed profile of the vehicle is unknown. In the simulation below, the control strategy is learned in the DTGs A and DTGs B driving cycle, and the performance is verified by simulation on the UDDS, HWFET, JN1015, WLTC, and NEDC driving cycle which are assumed as unknown driving cycle. In this case, since it is difficult to reflect the information about the speed profile to be driven in future on the control strategy, general control rules obtained through RL according DTGs A and DTGs B driving cycle are derived, and it result in the fuel efficiency of the vehicle.

Table 5.6 Equivalent fuel economy result [km/l] for RL-based strategy for various driving cycle (% compared to DDP result)

Algorithm /Trained Cycle	**Driving Cycle**					**Average**
	UDDS	**HWFET**	**JN1015**	**WLTC**	**NEDC**	
DDP	26.1	26.2	26.3	24.6	24.6	25.6
RL-based /DTGs A +DTGs B	24.6 (94.3)	25.5 (97.3)	24.7 (93.9)	23.0 (93.5)	23.2 (94.3)	24.2 (94.5)
SDP /DTGs A +DTGs B	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)

Simulation result shows that the average fuel economy of the new strategy is 24.2 km/l which is 94.5% of the fuel economy of DDP. Compared to rule-based strategy, the new strategy presents increased fuel economy performance for all driving cycle. Compared to SDP, RL-based strategy shows very close fuel economy result with SDP, that is obvious considering two strategy have same framework based on Bellman optimality equation.

On the other hand, another advantage of the RL-based strategy is that the learning ability of the strategy could be used for diagnosing vehicle system performance and adaptation of it into the control strategy. The vehicle is exposed to various driving environments and performance deteriorates naturally. For example, the aging of the engine or the performance degradation of the power train over time can happen in real-world vehicles, thus the adaptation of the controller according to the change of the vehicle component performance is a necessary factor to minimize fuel efficiency reduction of the vehicle. One of the advantages of the proposed strategy is that the algorithms can learn by themselves and find the optimal control according to these environmental changes. In this case study, we deliberately changed the fuel consumption map of the engine model with assumption that engine consumes more fuel in high torque area according to performance reduction and verified that the proposed algorithm can work properly to derive the optimal control rules according to this change. The fuel consumption map is intentionally modified as shown in figure 5.17, in which faint part of high engine torque indicates the fuel consumption rise.

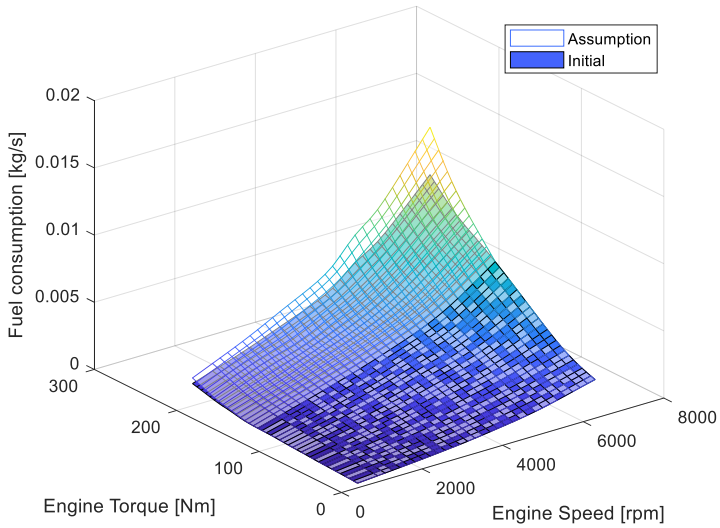


Figure 5.17 Assumption for fuel consumption map change

With this modified engine model, the RL-based EMS is implemented on HWFET driving cycle. As a result, the strategy dynamically changes the existing set of parameters according to the change of element and to find the optimum control rule. Figure 5.18 present vehicle fuel consumption model in RL-based energy strategy before and after fuel consumption map change. Figure 5.20(a) present BSFC map and engine operating point before fuel consumption map change and figure 5.20(b)-(d) present BSFC map and engine operating point after map change. It is shown that in figure 5.19(a) BSFC map is changed as fuel consumption map is modified intentionally, however engine operating point still remains in the same area. However, after a few iterations, the engine operating point moves to the most efficient area of the BSFC and fuel economy performance is also increased as given in table 5.7.

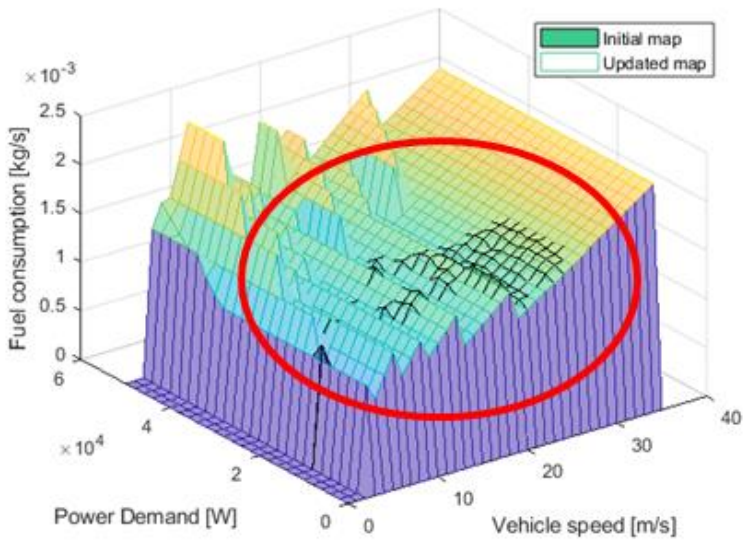
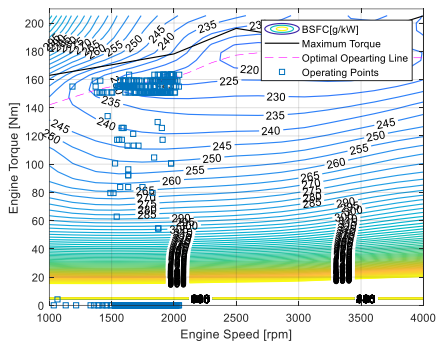


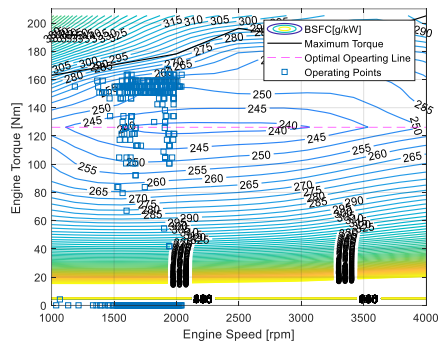
Figure 5.18 Vehicle model update according to fuel consumption map change

Table 5.7 Simulation result for the vehicle model update

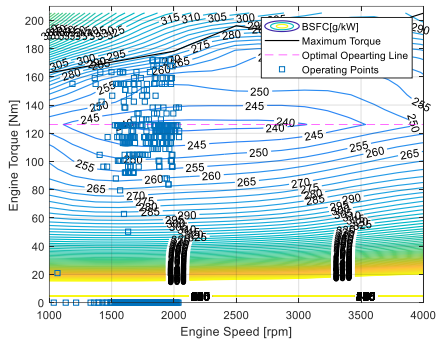
Algorithm	**Number of iteration**		
	1	**10**	**20**
Equivalent Fuel economy [km/l]	23.6	24.1	24.6



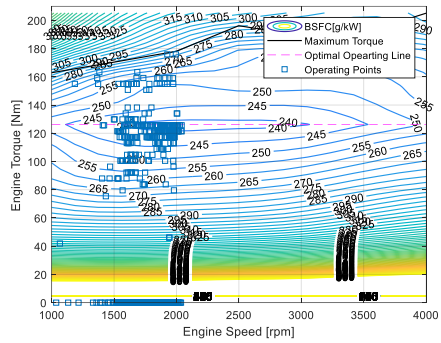
(a) Initial engine operating point without fuel consumption map change



(b) 1st iteration after fuel consumption map changed



(c) 10th iteration after fuel consumption map changed



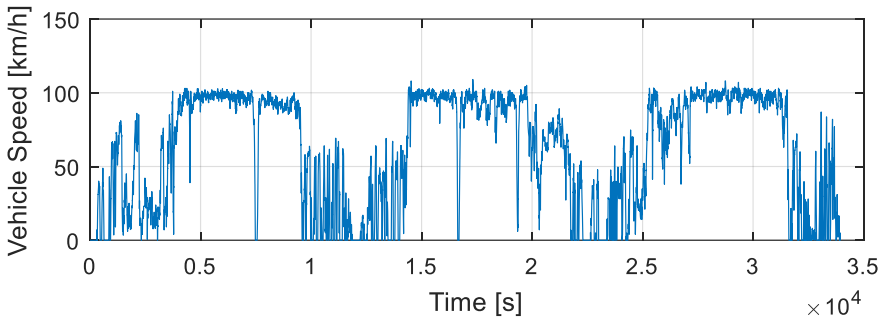
(d) 20th iteration after fuel consumption map changed

Figure 5.19 Vehicle model update according to change

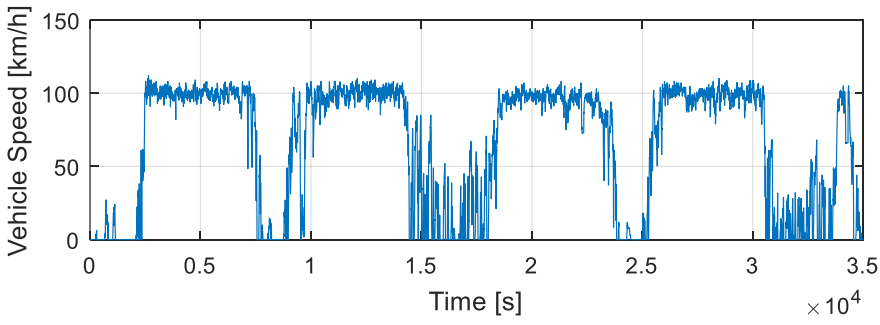
In the next simulation, the control strategy is learned through the vehicle speed profiles with high similarity, and it is confirmed how the fuel efficiency of the control strategy varies when each vehicle speed profile is used for the simulation. In other words, the effectiveness of the control strategy can be verified by confirming the fuel efficiency performance when the control strategy learned through the speed profile, that has a similar characteristic with the speed profile used for learning. To obtain vehicle speed profiles with similar probability characteristics, we used DTGs data obtained from the speed profile of express bus running between cities. Figure 5.20 shows the speed profiles of the express bus. The control strategy was learned in bus DTGs A cycle and bus DTGs B cycle, respectively, and then simulated in each driving cycle.

Simulation results is given table 5.8. There is almost no difference in the fuel consumption results when the DTGs A and DTGs B cycles are simulated using the control strategy learned using the bus DTGs A cycle and the bus DTGs B cycle, respectively. The implication of these simulation results is that the fuel economy of the control strategy can be maintained if the characteristics of the driving cycle do not change significantly even if the driving cycle changes, therefore, it is possible to obtain high fuel economy performance by using the proposed control strategy without the future speed prediction of the vehicle or the control strategy based on the prediction with the assumption that the characteristic of the driving cycle does not change abruptly, which is true for the actual vehicle driving environment.

.



(a) Bus DTGs driving cycle A



(b) Bus DTGs driving cycle B

Figure 5.20 Bus DTGs speed profiles

Table 5.8 Equivalent fuel economy result for RL-based strategy [km/l] for re-learning (% compared to DDP result)

Algorithm/Trained Cycle		**Driving Cycle**	
		BUS DTGs A	**BUS DTGs B**
DDP		25.0	26.8
RL-based	BUS DTGs A	24.4 (97.6)	25.9 (96.8)
	BUS DTGs B	24.4 (97.6)	26.0 (97.0)
Rule-based		20.8 (83.2)	21.4 (79.9)

On the other hand, the initial Q value of this RL-based control strategy can be defined using the SDP result as shown figure 5.21. Therefore, for example, even when the driving cycle characteristic of the vehicle suddenly changes, high fuel efficiency performance can be still obtained by using the Q value derived from the SDP in advance. Figure 5.22 shows the results of the control strategy using SDP initialization and the control strategy not using SDP initialization. In the case of SDP initialization, the result of SDP using DTGs A driving cycle and DTGs B driving cycle, which were used in the previous simulation, is defined as the initial value of Q value for the control strategy. In the case without SDP initializing, the Q value was learned only 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 figure 5.22.

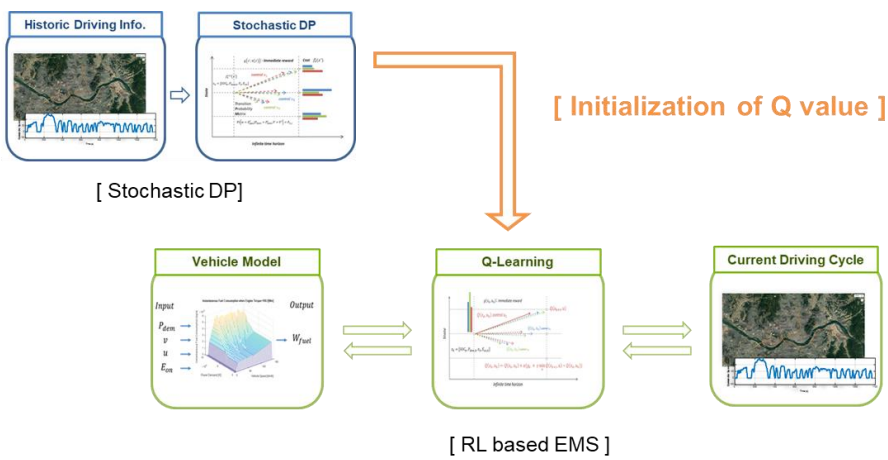


Figure 5.21 Initialization of Q value using SDP

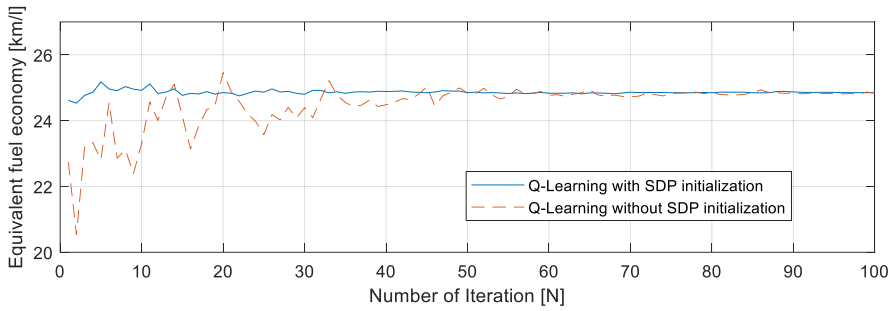


Figure 5.22 Equivalent fuel economy result using SDP initialization

The simulation results show that even if the driving cycle change suddenly, it is possible to improve the fuel efficiency of the proposed strategy by using the information optimized through the SDP and it is possible to reduce the learning time for a specific driving cycle using pre-calculated SDP result.

CHAPTER 6 CONCLUDING REMARKS

6.1 Conclusion

In this dissertation, Stochastic Dynamic Programming(SDP) algorithm and Reinforcement Learning(RL) algorithm are introduced, and applied into optimal control problem for Energy Management Strategy (EMS) of Hybrid Electric Vehicles (HEVs). SDP and RL, especially Q-learning are based on Bellman's optimality equation like Deterministic Dynamic Programming (DDP), but has an advantage that it can be used for a real-time implementable supervisory control of HEVs directly unlike DDP.

Firstly, to apply SDP algorithm into the control problem of the powertrain control for HEVs, driving cycle information is modeled using Markov process. Based on Transition Probability Matrix (TPM) of vehicle speed and power demand, optimization control problem on infinite-horizon is defined, and control policy is derived using SDP as function of vehicle speed, power demand, battery SOC, and engine on/off status to minimize the expected total fuel consumption of the vehicle, while penalizes the battery SOC deviation and frequent engine on/off. In this study, engine on/off cost is considered especially for more practical application and pre-calculated map based on vehicle dynamic equation is used to reduce computation burdensome of SDP.

However, SDP is an offline control policy in nature that TPM of the existing driving cycle is utilized for optimization, and the control strategy obtained is only optimized for the specific driving cycle that TPM is derived, or it is a general control policy in average sense. To overcome this cycle dependent problem, it is necessary to change the TPM according to the driving cycle of the vehicle to be driven and to extract an optimized control strategy using it.

In this study, RL-based control strategy was developed to overcome the shortcomings of SDP. In the newly proposed RL-based control strategy, the transition probability of the vehicle's driving speed profile is learned in real time by using the existing SDP framework as it is, and control strategy is optimized based on Q-learning technique, which is one of RL algorithm.

To verify the achievement of the SDP and RL-based strategies, vehicle simulation is conducted based on backward-looking vehicle simulator. As a result of simulation, both control strategies show improved fuel efficiency compared to existing rule-based control strategy. In the case of the control strategy based on RL, the characteristics of the driving cycle are learned without the TPM, and the control strategy is modified to improve the fuel efficiency as the driving is repeated.

In order to obtain improved fuel economy results in HEV, it is necessary not only to increase the efficiency of the vehicle powertrain but also to characterize the speed profile of the vehicle and to reflect it in the control strategy. Recently, with rapid development of intelligent transportation system, and global positioning system, the type of information and quantity of the driving

cycle information which can be achieved in the vehicle level is increased. Thus, how to utilize this information in order to increased fuel economy performance of the vehicle is a very important problem. The proposed control strategies in this paper have a powerful mathematical framework to model the driving cycle information in stochastic view, and to solve HEVs supervisory control problem based on optimization using Q−learning. Therefore, it is essential and has many advantages to use these EMS in the power distribution of HEVs.

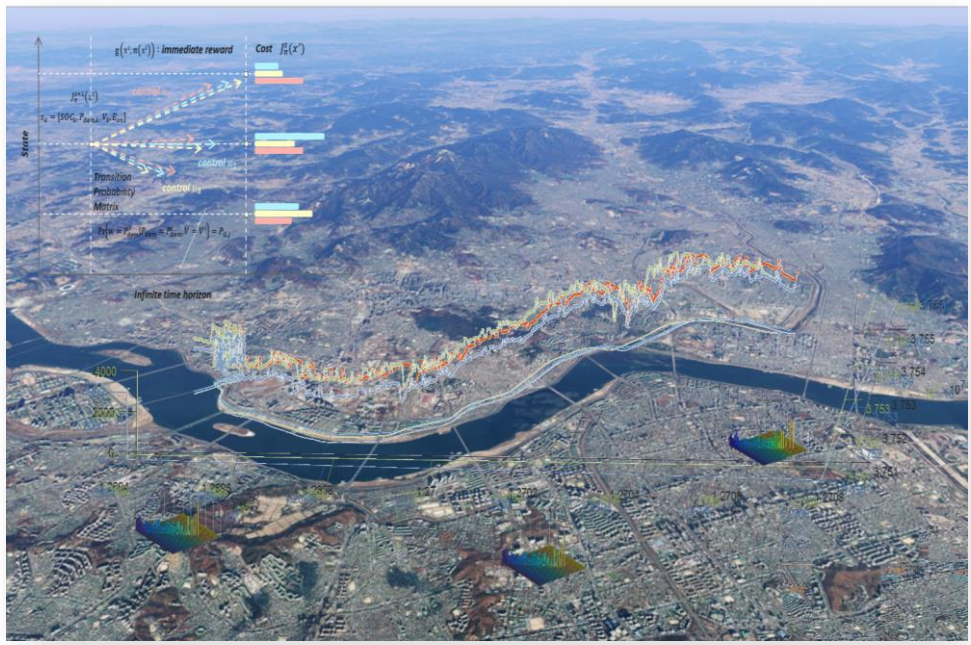


Figure 6.1 Stochastic optimal control concept

6.2 Future Work

The dissertation proposed a new approach to design supervisory controller for HEV. There exist several opportunities to improve the work presented in this paper and the proposed EMS.

First, experimental validation of the newly proposed control strategy is needed. Since the control strategy is verified based on the simulation, it is necessary to verify the strategy based on the experiment. In particular, it is necessary to check the fuel efficiency of the control strategy through various case studies while validating the learning ability of the strategy upon diverse driving situations of the vehicle. The state variables in this study are discretized, thus tradeoff relationship of computational burdensome and fuel economy performance of the strategy should be investigated based on experiment. Also, combined with other practical issue such as emission or drivability, it is possible to advance the proposed strategy more practical and realistic.

Second, it is necessary to further improve the fuel economy of the proposed algorithm. The simulation results show that the proposed algorithm is inferior to the optimal fuel efficiency of DDP. This is because the future velocity profile information of the vehicle is not used directly but in the form of a TPM, and optimal control problem is defined in infinite time horizon. Therefore, if the future driving speed profile information can be predicted or parameters based on prediction could be used for the learning process, then it could be used for improve the fuel efficiency performance of the proposed

EMS. Also, driver's behavior such as driving pattern information could be used for setting learning rate of the proposed strategy and could improve fuel economy performance.

REFERENCE

[1] "[Online]. Available: https://www.afdc.energy.gov/data/10305."

[2] "[Online]. Available: https://www.afdc.energy.gov/data/10562."

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

[4] X. Zhang, H. Peng, and J. Sun, "A Near – Optimal Power Management Strategy for Rapid Component Sizing of Power Split Hybrid Vehicles with Multiple Operating Modes," *Am. Control Conf.*, no. 1, pp. 5972–5977, 2013.

[5] D. Karbowski, N. Kim, and A. Rousseau, "Route–based online energy management of a PHEV and sensitivity to trip prediction," *2014 IEEE Veh. Power Propuls. Conf. VPPC 2014*, 2014.

[6] P. Pisu and G. Rizzoni, "A comparative study of supervisory control strategies for hybrid electric vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 3, pp. 506–518, 2007.

[7] J. Lars, A. Mattias, and E. Bo, "Assessing the Potential of Predictive Control for Hybrid Vehicle Powertrains Using Stochastic Dynamic Programming," *Intell. Transp. Syst. IEEE Trans.*, vol. 8, no. 1, pp. 71–83, 2007.

[8] H. Yang, B. Kim, Y. Park, W. Lim, and S. Cha, "ANALYSIS OF PLANETARY GEAR HYBRID POWERTRAIN SYSTEM PART 2 : OUTPUT SPLIT SYSTEM," *Int.J Automot. Technol.,* vol. 10, no. 3, 2009.

[9] Y. Zou, Z. Kong, T. Liu, and D. Liu, "A real–time Markov chain

driver model for tracked vehicles and its validation: Its adaptability via stochastic dynamic programming," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3571–3582, 2017.

[10] D. Karbowski, J. Kwon, A. Rousseau, K. Pechmann, "'Fair' Comparison of Powertrain Configurations for Plug-In Hybrid Operation Using Global Optimization," *SAE Technical Paper* 2009-01-1334, 2009 .

[11] R. Fellini, N. Michelena, P. Papalambros, and M. Sasena, "Optimal design of automotive hybrid powertrain systems," *Proceedings First International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, pp. 400-405, 1999.

[12] X. Zhang, C. T. Li, D. Kum, and H. Peng, "Prius+and volt-: Configuration analysis of power-split hybrid vehicles with a single planetary gear," *IEEE Trans. Veh. Technol.*, vol. 61, no. 8, pp. 3544–3552, 2012.

[13] C. Li and H. Peng, "Optimal Configuration Design for Hydraulic Split Hybrid Vehicles," *Proc. Am. Control Conf.*, pp. 5812–5817, 2010.

[14] H. Lee, C. Kang, J. Kim, S. Won, and C. Y. Park, "Component Sizing for Development of Novel PHEV System," *Transactions of KSAE*, vol. 24, no. 3, pp. 330–337, 2016.

[15] A. Sciarretta and L. Guzzella, "Control of hybrid electric vehicles," *IEEE Control Syst.*, vol. 27, no. 2, pp. 60–70, 2007.

[16] D. F. Opila, X. Wang, R. McGee, R. B. Gillespie, J. A. Cook, and J. W. Grizzle, "An Energy Management Controller to Optimally Trade Off Fuel Economy and Drivability for Hybrid Vehicles,"

IEEE Trans. Control Syst. Technol., vol. 20, no. 99, pp. 1–16, 2011.

[17] B. Skugor, D. Pavkovic, and J. Deur, "A series-parallel hybrid electric vehicle control strategy including instantaneous optimization of equivalent fuel consumption," *Proc. IEEE Int. Conf. Control Appl.*, no. 1, pp. 310–316, 2012.

[18] F. R. Salmasi, "Control Strategies for Hybrid Electric Vehicles: Evolution, Classification, Comparison, and Future Trends," *Veh. Technol. IEEE Trans.*, vol. 56, no. 5, pp. 2393–2404, 2007.

[19] S. G. Wirasingha and A. Emadi, "Classification and review of control strategies for plug-in hybrid electric vehicles," *IEEE Trans. Veh. Technol.*, vol. 60, no. 1, pp. 111–122, 2011.

[20] T. Hofman, M. Steinbuch, R. Van Druten, and A. Serrarens, "Rule-based energy management strategies for hybrid vehicles," *Int. J. Electr. Hybrid Veh.*, vol. 1, no. 1, pp. 71–94, 2007.

[21] X. Lin, A. Ivanco, and Z. Filipi, "Optimization of Rule-Based Control Strategy for a Hydraulic- Electric Hybrid Light Urban Vehicle Based on Dynamic Programming," *SAE Int. J. Alt. Power.,* pp. 249–259, 2012.

[22] H. Banvait, S. Anwar, and C. Yaobin, "A rule-based energy management strategy for Plug-in Hybrid Electric Vehicle (PHEV)," *Am. Control Conf. 2009. ACC '09.*, pp. 3938–3943, 2009.

[23] C. J. Mansour, "Trip-based optimization methodology for a rule-based energy management strategy using a global optimization routine: The case of the Prius plug-in hybrid

electric vehicle," *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.*, vol. 230, no. 11, pp. 1529–1545, 2016.

[24] N. Jalil, N. a Kheir, and M. Salman, "A rule-based energy management strategy for a series hybrid vehicle," *Am. Control Conf. 1997. Proc. 1997*, vol. 1, no. June, pp. 689–693 vol.1, 1997.

[25] H.D. Lee and S.K. Sul, "0-Fuzzy-logic-based torque control strategy for parallel-type hybrid electric vehicle," *Ind. Electron. IEEE Trans.*, vol. 45, no. 4, pp. 625–632, 1998.

[26] M. R. Dubois, A. Desrochers, and N. Denis, "Fuzzy-based blended control for the energy management of a parallel plug-in hybrid electric vehicle," *IET Intell. Transp. Syst.*, vol. 9, no. 1, pp. 30–37, 2015.

[27] R. Zhang and J. Tao, "GA based fuzzy energy management system for FC/SC powered HEV considering H2 consumption and load variation," *IEEE Trans. Fuzzy Syst.*, vol. 6706, no. c, 2017.

[28] D. Zhao, R. Stobart, G. Dong, and E. Winward, "Real-Time Energy Management for Diesel Heavy Duty Hybrid Electric Vehicles," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 829–841, 2015.

[29] A. Ravey, B. Blunier, and A. Miraoui, "Control strategies for fuel-cell-based hybrid electric vehicles: From offline to online and experimental results," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2452–2457, 2012.

[30] N. W. Kim, D. H. Lee, C. Zheng, C. Shin, H. Seo, and S. W. Cha, "REALIZATION OF PMP-BASED CONTROL FOR HYBRID

ELECTRIC VEHICLES IN A BACKWARD-LOOKING SIMULATION," *Int.J Automot. Technol.,* vol. 15, no. 4, pp. 625–635, 2014.

[31] L. Serrao, S. Onori, and G. Rizzoni, "ECMS as a realization of pontryagin's minimum principle for HEV control," *Proc. Am. Control Conf.*, pp. 3964–3969, 2009.

[32] M. Sivertsson, "Adaptive control using map-based ECMS for a PHEV", *IFAC Proceedings volumes*, vol. 45, no. 30., 2012.

[33] C. Zhang and A. Vahidi, "Real-time optimal control of plug-in hybrid vehicles with trip preview," *2010 Am. Control Conf.*, pp. 6917–6922, 2010.

[34] P. Rodatz, G. Paganelli, A. Sciarretta, and L. Guzzella, "Optimal power management of an experimental fuel cell / supercapacitor-powered hybrid vehicle," *Control Engineering Practice*, vol. 13, pp. 41–53, 2005.

[35] C. Zheng, Y. P. Suk, and W. Cha, "Fuel Economy Evaluation Methods of Fuel Cell Hybrid Vehicles," *Int J Automot Technol.*, vol. 13, pp. 2–5, 2012

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

[37] C. Zheng, G. Xu, and Y. Zhou, "Realization of PMP-based Power Management Strategy for Hybrid Vehicles Based on MPC Scheme," *4th IEEE Int. Conf. Inf. Sci. Technol.*, pp. 682–685, 2014.

[38] A. Rezaei, J. B. Burl, and B. Zhou, "Estimation of the ECMS Equivalent Factor Bounds for Hybrid Electric Vehicles," *IEEE*

Trans. Control Syst. Technol., pp. 1–8, 2017.

[39] R. Wang, “Dynamic Programming Technique in Hybrid Electric Vehicle Optimization,” *IEEE Int. Electr. Veh. Conf.*, pp. 1–8, 2012.

[40] C. Lin, H. Peng, J. W. Grizzle, and J. Kang, “Power Management Strategy for a Parallel Hybrid Electric Truck,” *IEEE Transactions on Control Systems Technology,* vol. 11, no. 6, pp. 839–849, 2003.

[41] L. Johannesson and B. S. Egardt, "Approximate dynamic programming applied to parallel hybrid powertrains", *IFAC world congress*, vol. 41, Issue2, 2008.

[42] H.-G. Wahl and F. Gauterin, “An iterative dynamic programming approach for the global optimal control of hybrid electric vehicles under real-time constraints,” *2013 IEEE Intell. Veh. Symp.*, no. Iv, pp. 592–597, 2013.

[43] X. Wang, H. He, F. Sun, and J. Zhang, “Application study on the dynamic programming algorithm for energy management of plug-in hybrid electric vehicles,” *Energies*, vol. 8, no. 4, pp. 3225–3244, 2015.

[44] Q. Gong, Y. Li, and Z. R. Peng, “Trip Based Power Management of Plug-in Hybrid Electric Vehicle with Two-Scale Dynamic Programming,” *2007 IEEE Veh. Power Propuls. Conf.*, pp. 12–19, 2007.

[45] H. Lee, Y. Park, and S. W. Cha, “Power Management Strategy of Hybrid Electric Vehicle using Power Split Ratio Line Control Strategy based on Dynamic Programming,” *International Conference on Control, Automation and Systems*, pp. 1739–

1742, 2015.

[46] L. Lai and M. Ehsani, “Dynamic programming optimized constrained engine on and off control strategy for parallel hev,” *2013 9th IEEE Veh. Power Propuls. Conf. IEEE VPPC 2013*, pp. 422–426, 2013.

[47] I. Kolmanovsky, I. Siverguina, and B. Lygoe, “Optimization of powertrain operating policy for feasibility assesment and calibration: stochastic dynamic programming approach,” *Proc. Am. Contr. Conf.*, vol. 8, no. i, pp. 1425–1430, 2002.

[48] J. Liu and H. Peng, “Modeling and Control of a Power-Split,” *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 6, pp. 1242–1251, 2008.

[49] S. J. Moura, H. K. Fathy, D. S. Callaway, and J. L. Stein, “A Stochastic Optimal Control Approach for Power Management in Plug-In Hybrid Electric Vehicles,” *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 3, pp. 545–555, 2011.

[50] C.-C. Lin, M.-J. Kim, H. Peng, and J. W. Grizzle, “System-Level Model and Stochastic Optimal Control for a PEM Fuel Cell Hybrid Vehicle,” *J. Dyn. Syst. Meas. Control*, vol. 128, no. 4, p. 878, 2006.

[51] K. McDonough, I. Kolmanovsky, D. Filev, D. Yanakiev, S. Szwabowski, and J. Michelini, “Stochastic dynamic programming control policies for fuel efficient in-traffic driving,” *2012 Am. Control Conf.*, no. 734, pp. 3986–3991, 2012.

[52] D. F. Opila, X. Wang, R. McGee, and J. W. Grizzle, “Real-Time Implementation and Hardware Testing of a Hybrid Vehicle Energy Management Controller Based on Stochastic Dynamic

Programming," *J. Dyn. Syst. Meas. Control*, vol. 135, no. 2, p. 021002, 2012.

[53] T. Leroy, F. Vidal-Naquet, P. Tona, and F. Rueil-Malmaison, "Stochastic Dynamic Programming based Energy Management of HEV's: an Experimental Validation," *IFAC World Congr.*, pp. 4813–4818, 2014.

[54] 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, 2016.

[55] D. Bardini and T. Leroy, "Impact of driveability constraints on local optimal energy management strategies for hybrid powertrains", *IFAC Symposimum on Advances in Autonotive Control*, vol. 7, no. PART 1. pp 23-28, 2013.

[56] E. D. Tate, J. W. Grizzle, and H. Peng, "Shortest path stochastic control for hybrid electric vehicles," *Int. J. robust nonlinear Control*, no. December 2007, pp. 1409–1429, 207AD.

[57] T. Leroy, J. Malaize, and G. Corde, "Towards Real-Time Optimal Energy Management of HEV Powertrains Using Stochastic Dynamic Programming," *Veh. Power Propuls.*, pp. 383–388, 2012.

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

[59] C. Liu and Y. L. Murphey, "Power management for Plug-in Hybrid Electric Vehicles using Reinforcement Learning with

trip information," *2014 IEEE Transp. Electrif. Conf. Expo*, pp. 1–6, 2014.

[60] 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, no. 5, pp. 538–548, 2018.

[61] 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.

[62] 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, 2017.

[63] 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.

[64] "[Online]. Available: https://www.autonomie.net/."

[65] H.G. Wahl, K.L. Bauer, F. Gauterin, and M. Holzapfel, "A real-time capable enhanced dynamic programming approach for predictive optimal cruise control in hybrid electric vehicles," *2013 IEEE 16th Int. Conf. Intell. Transp. Syst.*, no. Itsc, pp. 1662–1667, 2013.

[66] A. Kreutzmann, D. Wolter, F. Dylla, and J. H. Lee, "Towards Safe Navigation by Formalizing Navigation Rules," *TransNav, Int. J. Mar. Navig. Saf. Sea Transp.*, vol. 7, no. 2, pp. 161–168, 2013.

[67] C. H. Zheng, N. W. Kim, and S. W. Cha, “Optimal control in the power management of fuel cell hybrid vehicles,” *Int. J. Hydrogen Energy*, vol. 37, no. 1, pp. 655–663, 2011.

[68] G. Qiuming, L. Yaoyu, and P. Zhong-Ren, “Computationally efficient optimal power management for plug-in hybrid electric vehicles based on spatial-domain two-scale dynamic programming,” *Veh. Electron. Safety, 2008. ICVES 2008. IEEE Int. Conf.*, pp. 90–95, 2008.

[69] P. R. Montague, “Reinforcement Learning: An Introduction, by Sutton, R.S. and Barto, A.G.,” *Trends Cogn. Sci.*, vol. 3, no. 9, p. 360, 1999.

[70] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*. New York: Academic, 1976.

[71] R. J. Leake and R.-W. Liu, “Construction of Suboptimal Control Sequences,” *J. SIAM Contr.*, vol. 5, no. 1, pp. 54–63, 1967.

[72] D. P. Bertsekas and J.N. Tsitsiklis, “Neuro-Dynamic Programming ,” Belmont, MA: Athena Scientific, 1996.

[73] P. Mehta and S. Meyn, “Q-learning and Pontryagin’s minimum principle,” *Proc. IEEE Conf. Decis. Control*, no. 1, pp. 3598–3605, 2009.

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

[75] F. Lewis, D. Vrabie, and K. Vamvoudakis, “Reinforcement Learning and Feedback Control: Using Natural Decision Methods to Design Optimal Adaptive Controllers,” *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, 2012.

[76] C. Watkins, “Learning from Delayed Reward,” Ph.D. dissertation, Cambridge University, Cambridge, U.K., 1989.

[77] C. J. C. H. Watkins and P. Dayan, “Q−learning,” *Mach. Learn.*, vol. 8, no. 3–4, pp. 279–292, 1992.

[78] P. J. Werbos, “Neural networks for control and system identification,” *Proc. 28th IEEE Conf. Decis. Control.*, pp. 260–265, 1989.

[79] R. S. Sutton, “Learning to Predict by the Method of Temporal Differences,” *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.

[80] H. He, J. Gao, and Y. Zhang, “Fuel cell output power−oriented control for a fuel cell hybrid electric vehicle,” *2008 Am. Control Conf.*, vol. 1, pp. 605–610, 2008.

[81] S. Barsali, C. Miulli, and a. Possenti, “A Control Strategy to Minimize Fuel Consumption of Series Hybrid Electric Vehicles,” *IEEE Trans. Energy Convers.*, vol. 19, no. 1, pp. 187–195, 2004.

[82] C. H. Zheng, C. E. Oh, Y. I. Park, and S. W. Cha, “Fuel economy evaluation of fuel cell hybrid vehicles based on equivalent fuel consumption,” *Int. J. Hydrog. Energy*, vol. 7, pp. 2–8, 2011.

국 문 초 록

본 논문에서는 하이브리드 자동차(Hybrid Electric Vehicles)의 연비향상을 위해 확률론적 최적화 제어 이론을 적용한 동력 분배 제어 전략(Energy management Strategy) 개발에 대한 연구를 진행하였다.

하이브리드 자동차는 내연기관과 전기모터 두가지의 동력원을 사용하여 차량을 주행하는 자동차로 기존 내연기관 차량에 비해 연비가 증대되고 친환경적이라는 장점을 가지고 있다. 이러한 하이브리드의 연비 성능을 향상시키기 위해서는, 엔진과 모터의 출력을 결정하는 상위 제어기인 동력분배 제어 전략이 매우 중요하며, 이를 위해 여러가지 최적화 이론 기반 제어전략들이 개발되고 있다. 이중 다이나믹 프로그래밍(Dynamic Programming)은 시스템의 비선형성이나 구속 조건에 상관없이 전역 최적 해를 얻을 수 있다는 장점을 가지고 있지만, 차량의 미래 속도 프로파일을 주행 전에 미리 알아야 하므로 차량의 실시간 제어에는 적용이 어렵다는 단점을 가지고 있다. 본 연구에서는 기존 다이나믹 프로그래밍 이론을 기반으로 하되 실시간으로 활용이 가능한 확률론적 다이나믹 프로그래밍(Stochastic Dynamic Programming) 및 강화 학습(Reinforcement Learning)의 기법을 이용하여 차량 동력 분배 제어 전략을 개발하였다.

확률론적 다이나믹 프로그래밍을 차량 제어에 적용하기 위해, 과거의 차량 주행 속도 프로파일의 특성은 속도 프로파일의 속도와 요구 파워량이 이산화 되어 마르코프 프로세스(Markov Process) 를 통해 확률밀도 함수(Transition Probability Matrix)로 표현되었다. 최적화 문제의 목적 함수로는 차량 연료소모량의 기대 값과 배터리 충전량(State of Charge)의 편차, 그리고 빈번한 엔진 on/off 를 최소화하도록 문제가 정의되었으며, 결과로 얻어진 최적 제어 규칙은 차량의 속도, 요구 파워량, 배터리 충전량, 엔진 on/off 상태에 따른 엔진과 모터 간의 동력분배

비로 주어진다.

그러나 이러한 확률론적 다이나믹 프로그래밍의 경우에도 현재 주행하고자 하는 차량의 속도프로파일이 최적화에서 사용된 속도프로파일의 확률 밀도 함수와 다를 경우 차량의 연비 성능이 저감 될 수 있다. 따라서 본 연구에서는 이를 보완하기 위해 기계 학습의 한 분야인 강화 학습(Reinforcement Learning)의 기법을 이용하여 기존 확률론적 다이나믹 프로그래밍의 프레임워크 안에서 차량의 주행상황에 따라 스스로 학습하여 최적화가 진행될 수 있는 제어 전략을 개발하였다. 새롭게 제안된 제어전략에서는 Q-learning 알고리즘을 기반으로 주행속도 프로파일의 확률 정보가 차량의 각 상태 변수 및 제어 입력에 대한 가치를 나타내는 Q value에 업데이트 되며, 확률론적 다이나믹 프로그래밍과 마찬가지로 가능한 모든 제어 입력에 대한 비용 함수를 차량 모델을 통해 계산함으로써 최적 제어 규칙을 도출한다.

차량 시뮬레이션 검증을 위하여 후 방향 차량 시뮬레이터가 개발되었으며 병렬형 하이브리드 차량 구조에 대해 검증을 수행하였다. 시뮬레이션 결과로써 기존 규칙기반 제어 전략에 비해 연비가 향상되었음을 확인하였으며, 차량의 성능 변화에도 학습을 통해 최적 제어 규칙이 도출되는 것을 확인하였다.

본 논문에서 연구된 제어 전략의 경우 차량의 주행속도 프로파일의 특성을 반영하여 연비 성능을 향상 시킬 수 있는 적응형 알고리즘으로, 실제 차량에 적용되어 차량의 여러가지 다양한 주행환경에서 연비를 향상시킬 수 있는 제어 기술이 될 수 있을 것으로 기대된다.

주요어: 하이브리드 자동차, 동력분배 제어 전략, 최적 제어, 확률론적 다이나믹 프로그래밍, 강화학습

학 번: 2013-20707