Reinforcement Learning Based Power Management for Hybrid Electric Vehicles

Xue Lin[1], Yanzhi Wang[1], Paul Bogdan[1], Naehyuck Chang[2], and Massoud Pedram[1]
[1]University of Southern California, Los Angeles, CA, USA
[2]Korea Advanced Institute of Science and Technology, Daejeon, Korea
[1]{xuelin, yanzhiwa, pbogdan, pedram}@usc.edu, [2]naehyuck@cad4x.kaist.ac.kr

ABSTRACT
Compared to conventional internal combustion engine (ICE) propelled vehicles, hybrid electric vehicles (HEVs) can achieve both higher fuel economy and lower pollution emissions. The HEV consists of a hybrid propulsion system containing one ICE and one or more electric motors (EMs). The use of both ICE and EM increases the complexity of HEV power management, and therefore requires advanced power management policies to achieve higher performance and lower fuel consumption. Towards this end, our work aims at minimizing the HEV fuel consumption over any driving cycle (without prior knowledge of the cycle) by using a reinforcement learning technique. This is in clear contrast to prior work, which requires deterministic or stochastic knowledge of the driving cycles. In addition, the proposed reinforcement learning technique enables us to (partially) avoid reliance on complex HEV modeling while coping with driver specific behaviors. To our knowledge, this is the first work that applies the reinforcement learning technique to the HEV power management problem. Simulation results over real-world and testing driving cycles demonstrate the proposed HEV power management policy can improve fuel economy by 42%.

Categories and Subject Descriptors
B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms
Algorithms, Management, Performance, Design.

Keywords
Hybrid electric vehicle (HEV), power management, reinforcement learning.

1. INTRODUCTION
Automobiles have contributed significantly to the development of modern society by satisfying many of the requirements for mobility in everyday life. However, large amounts of fuel consumption and pollution emissions resulting from the increasing number of automobiles in use around the world have drawn attention of researchers and developers towards more energy efficient and environmentally friendly automobiles. Hybrid electric vehicles (HEVs) represent a promising approach towards sustainable mobility. In contrast to conventional internal combustion engine (ICE) propelled vehicles, HEVs can simultaneously achieve higher fuel economy and lower pollution emissions [1], [2], [3].

The HEV features a hybrid propulsion system comprised of an ICE with an associated fuel tank and an electric motor (EM) with an associated electrical energy storage system (e.g., batteries), both of which may be coupled directly to the drivetrain. The ICE consumes fuel to provide the primary propulsion, whereas the EM converts the stored electrical energy to the secondary propulsion when extra torque is needed. Besides assisting the ICE with extra torque, the EM also serves as a generator for recovering kinetic energy during braking (known as regenerative braking) and collecting excess energy from the ICE during coasting. The introduction of the secondary propulsion by the EM allows for a smaller ICE design and makes HEVs more efficient than conventional ICE vehicles in terms of acceleration, hill climbing, and braking energy utilization [4], [5].

On the other hand, the use of both ICE and EM increases the complexity of HEV power management and advanced power management policy is required for achieving higher performance and lower fuel consumption. A power management policy for HEVs determines the power split between the ICE and EM to satisfy the speed and torque requirements and, meanwhile, to ensure safe and smooth operation of the involved power components (e.g., ICE, EM and batteries). Furthermore, a "good" power management policy should result in reduced fuel consumption and lower pollution emissions. Rule-based power management approaches have been designed based on heuristics, intuition, and human expertise [6], [7]. Although rule-based approaches are effective for real-time supervisory control, they may be far from being optimal. Dynamic programming (DP) techniques have been applied to the power management of various types of HEVs [8], [9], [10]. DP techniques can derive a globally optimal solution that minimizes the total fuel consumption during a whole driving cycle, which is given as a vehicle speed versus time profile for a specific trip. Unfortunately, the DP techniques require *a priori* knowledge of the driving cycles as well as detailed and accurate HEV modeling; therefore they are not applicable for real-time implementation.

The equivalent consumption minimization strategy (ECMS) approach has been proposed to reduce the global optimization problem (as in DP techniques) to an instantaneous optimization problem [11]. However, the ECMS approach strongly depends on the equivalence factors, which convert the electrical energy consumption of EM into the equivalent fuel consumption of ICE. The equivalence factors are quite sensitive to the driving cycles. For instance, the equivalence factors that are suitable for a driving cycle may lead to poor performance for other driving cycles. To overcome this challenge, the adaptive-ECMS (A-ECMS) approach has been applied for HEV power management based on driving cycle prediction within a finite horizon [12]. Although the A-ECMS approach has good performance, the detailed driving cycle prediction method has been omitted. Gong *et al.* has provided a trip modeling method using a combination of geographical information systems (GISs), global positioning systems (GPSs), and intelligent transportation systems (ITSs) [13]. However, the driving cycle constructed by this trip modeling method is synthetic and not accurate enough to capture the real driving scenarios, such as the effect of traffic lights and some unforeseen circumstances. In [14] and [15], the authors proposed the stochastic control method for HEVs based on a Markov chain model of the driving cycles. This method does not rely on *a priori* knowledge of the driving cycles, but it is not adaptive to the dynamical driving conditions.

Towards this end, our work aims at minimizing the HEV fuel consumption over any driving cycles. We propose to use the

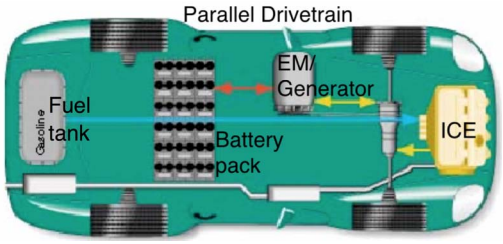

Figure 1. The parallel hybrid drivetrain configuration [16].

reinforcement learning technique for deriving the optimal HEV power management policy. Unlike some previous approaches, which require complete or stochastic information of the driving cycles, in our method the HEV controller does not require any prior information about the driving cycles and uses only partial information about the HEV modeling. Consequently, we carefully define the state space, action space, and reward in the reinforcement learning technique such that the objective of the reinforcement learning agent coincides with our goal of minimizing the HEV overall fuel consumption. We employ the TD(λ)-learning algorithm to derive the optimal HEV power management policy, due to its relatively higher convergence rate and higher performance in non-Markovian environment. To the best of our knowledge, this is the first work that applies the reinforcement learning technique to the HEV power management problem. Simulation results over real-world and testing driving cycles demonstrate that the proposed HEV power management policy can improve fuel economy by 42%.

2. HEV SYSTEM DESCRIPTION

By way of an example and without loss of generality, our proposed power management policy is designed exemplarily for (but not limited to) the parallel hybrid drivetrain configuration displayed in Figure 1. In an HEV with the parallel hybrid drivetrain, i.e., a parallel HEV, the ICE and EM can deliver power in parallel to drive the wheels. There are five different operation modes in a parallel HEV, depending on the flow of energy:

1) ICE only mode: wheels are driven only by the ICE.
2) EM only mode: wheels are driven only by the EM.
3) Power assist mode: wheels are driven by both the ICE and EM.
4) Battery charging mode: a part of the ICE power drives the EM as a generator to charge the battery pack, while the other part of the ICE power drives the wheels.
5) Regenerative braking mode: the wheels drive the EM as a generator to charge the battery pack when the vehicle is braking.

2.1 Internal Combustion Engine (ICE)

We describe a quasi-static ICE model [17] as follows. The fuel consumption rate $\dot{m}_f$ (in $\mathrm{g}\cdot\mathrm{s}^{-1}$) of an ICE is a nonlinear function of the ICE speed ω_{ICE} (in $\mathrm{rad}\cdot\mathrm{s}^{-1}$) and torque T_{ICE} (in $\mathrm{N}\cdot\mathrm{m}$). The fuel efficiency of an ICE is calculated by

$$\eta_{ICE}(\omega_{ICE}, T_{ICE}) = (\omega_{ICE}\cdot T_{ICE})/(\dot{m}_f \cdot D_f), \quad (1)$$

where D_f is the fuel energy density (in $\mathrm{J}\cdot\mathrm{g}^{-1}$).

Figure 2 shows a contour map of the fuel efficiency of an ICE in the *speed-torque* plane. The number labeled with each contour is the corresponding ICE efficiency. It is a 1.0L VTEC-E SI ICE modeled by the advanced vehicle simulator ADVISOR [16]. The ICE has a peak power of 50 kW and a peak efficiency of 40%. A "good" power management policy should avoid ICE operation point (ω_{ICE}, T_{ICE}) in the low efficiency region. Superimposed on the contour map is the maximum ICE torque $T_{ICE}^{max}(\omega_{ICE})$ (the dashed line.) To ensure safe and smooth operation of an ICE, the following constraints should be satisfied:

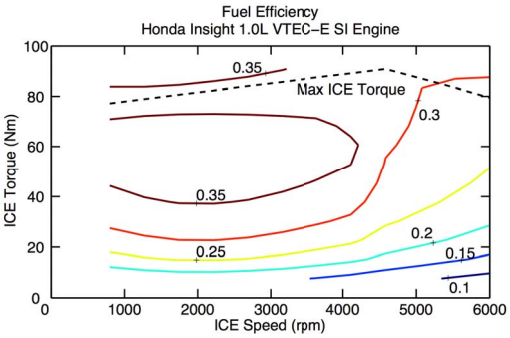

Figure 2. Fuel efficiency map of an ICE.

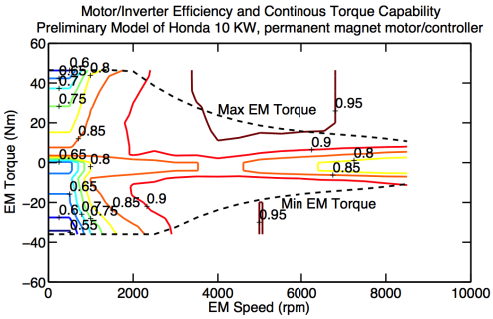

Figure 3. Efficiency map of an EM.

$$\omega_{ICE}^{min} \le \omega_{ICE} \le \omega_{ICE}^{max}, \quad (2)$$

$$0 \le T_{ICE} \le T_{ICE}^{max}(\omega_{ICE}). \quad (3)$$

2.2 Electric Motor (EM)

Figure 3 presents a contour map of the efficiency of an EM also in the speed-torque plane. It is a permanent magnet EM modeled by ADVISOR. The EM has a peak power of 10 kW and a peak efficiency of 96%. Let ω_{EM} and T_{EM}, respectively, denote the speed and torque of the EM. When $T_{EM} \ge 0$, the EM operates as a motor; when $T_{EM} < 0$, the EM operates as a generator. The efficiency of the EM is defined by

$$\eta_{EM}(\omega_{EM}, T_{EM}) = \begin{cases} (\omega_{EM}\cdot T_{EM})/P_{batt}, & T_{EM} \ge 0 \\ P_{batt}/(\omega_{EM}\cdot T_{EM}), & T_{EM} < 0 \end{cases}, \quad (4)$$

where P_{batt} is the output power of the battery pack. When $T_{EM} \ge 0$, the battery pack is discharging and P_{batt} is a positive value; when $T_{EM} < 0$, the battery pack is charging and P_{batt} is a negative value. Superimposed on the contour map are the maximum and minimum EM torques (the dashed lines) i.e., $T_{EM}^{max}(\omega_{EM})$ and $T_{EM}^{min}(\omega_{EM})$, respectively. To ensure safe and smooth operation of an EM, the following constraints should be satisfied:

$$0 \le \omega_{EM} \le \omega_{EM}^{max}, \quad (5)$$

$$T_{EM}^{min}(\omega_{EM}) \le T_{EM} \le T_{EM}^{max}(\omega_{EM}). \quad (6)$$

2.3 Drivetrain Mechanics

In what follows, we discuss a simplified but sufficiently accurate drivetrain model as in [18], [19]. The following equations describe the drivetrain mechanics, showing the mechanical coupling between different components and the vehicle.

- Speed relation

$$\omega_{wh} = \frac{\omega_{ICE}}{R(k)} = \frac{\omega_{EM}}{R(k)\cdot\rho_{reg}}. \quad (7)$$

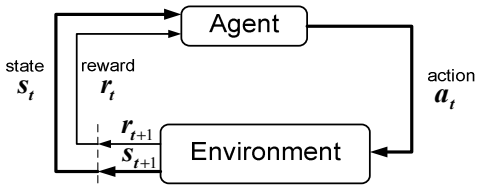

Figure 4. The agent-environment interaction.

- Torque relation
$$T_{wh} = R(k) \cdot (T_{ICE} + \rho_{reg} \cdot T_{EM} \cdot (\eta_{reg})^{\alpha}) \cdot (\eta_{gb})^{\beta}. \quad (8)$$

ω_{wh} and T_{wh} are the wheel speed and torque, respectively. $R(k)$ is the gear ratio of the k-th gear. The ρ_{reg} is the reduction gear ratio. The η_{reg} and η_{gb} are the reduction gear efficiency and gear box efficiency, respectively. α equals $+1$ if $T_{EM} \geq 0$, and -1 otherwise. β equals $+1$ if $T_{ICE} + \rho_{reg} \cdot T_{EM} \cdot (\eta_{reg})^{\alpha} \geq 0$, and -1 otherwise.

2.4 Vehicle Dynamics

The vehicle is considered as a rigid body with four wheels and the vehicle mass is assumed to be concentrated in a single point. The following force balance equation describes the vehicle dynamics:

$$m \cdot a = F_{TR} - F_g - F_R - F_{AD}. \quad (9)$$

m is the vehicle mass, a is the vehicle acceleration, and F_{TR} is the total tractive force. The force due to road slope is given by

$$F_g = m \cdot g \cdot \sin\theta, \quad (10)$$

where θ is the road slope angle. The rolling friction force is given by

$$F_R = m \cdot g \cdot \cos\theta \cdot C_R, \quad (11)$$

where C_R is rolling friction coefficient. The air drag force is give by

$$F_{AD} = 0.5 \cdot \rho \cdot C_D \cdot A_F \cdot v^2, \quad (12)$$

where ρ is air density, C_D is air drag coefficient, A_F is the vehicle frontal area, and v is the vehicle speed. Given v, a, and θ, the total tractive force F_{TR} can be derived using (9)~(12). Then, the wheel speed and torque are related to F_{TR}, v, and wheel radius r_{wh} by

$$\omega_{wh} = v/r_{wh}, \quad (13)$$

$$T_{wh} = F_{TR} \cdot r_{wh}. \quad (14)$$

2.5 Backward-Looking Optimization

In this work, the *backward-looking* optimization approach [8]~[15] is adopted, which implies that the HEV controller determines the operation of ICE and EM, so that the vehicle meets the target performance (speed v and acceleration a) specified in benchmark driving cycles [21]. In reality, the drivers determine the speed v and power demand $p_{dem} = \omega_{wh} \cdot T_{wh}$ profiles for propelling the HEV (through pressing the acceleration or brake pedal.) The backward-looking optimization is equivalent to actual HEV management because p_{dem} and a satisfy a relationship specified in Section 2.4.

With given values of vehicle speed v and acceleration a (or power demand p_{dem}), the required wheel speed ω_{wh} and torque T_{wh} satisfy (9)~(14). In addition, the five variables, i.e., the ICE speed ω_{ICE} and torque T_{ICE}, the EM speed ω_{EM} and torque T_{EM}, and the gear ratio $R(k)$, should satisfy (7) and (8) to support the required wheel speed and torque. The HEV controller chooses the battery output power P_{batt} (or equivalently, battery charging/discharging current) and the gear ratio $R(k)$ as the control variables. Then, the rest of variables (i.e., ω_{ICE}, T_{ICE}, ω_{EM}, and T_{EM}) become dependent (associate) variables, the values of which are determined by P_{batt} and $R(k)$. The results of the HEV power management policy are the fuel consumption rate of the ICE.

3. REINFORCEMENT LEARNING BACKGROUND

Reinforcement learning provides a mathematical framework for discovering or learning strategies that map situations onto actions with the goal of maximizing a reward function [20]. The learner and decision-maker is called the *agent*. The thing it interacts with, comprising everything outside the agent, is called the *environment*. The agent and environment interact continually, the agent selecting actions and the environment responding to those actions and presenting new situations to the agent. The environment also gives rise to rewards, which are special numerical values that the agent tries to maximize over time.

Figure 4 illustrates the agent-environment interaction in reinforcement learning. Specifically, the agent and environment interact at each of a sequence of discrete time steps, i.e., $t = 0, 1, 2, 3, \ldots$. At each time step t, the agent receives some representation of the environment's *state*, i.e., $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible states, and on that basis selects an *action*, i.e., $a_t \in \mathcal{A}(s_t) \subseteq \mathcal{A}$, where $\mathcal{A}(s_t)$ is the set of actions available in state s_t and $\mathcal{A}$ is the set of all possible actions. One time step later, in part as a consequence of its action, the agent receives a numerical *reward*, i.e., $r_{t+1} \in \mathcal{R}$, and finds itself in a new state, i.e., s_{t+1}.

A policy, denoted by π, of the agent is a mapping from each state $s \in \mathcal{S}$ to an action $a \in \mathcal{A}$ that specifies the action $a = \pi(s)$ that the agent will choose when the environment is in state s. The ultimate goal of an agent is to find the optimal policy, such that

$$V^{\pi}(s) = E\{\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} | s_t = s\} \quad (15)$$

is maximized for each state $s \in \mathcal{S}$. The *value function* $V^{\pi}(s)$ is the expected return when the environment starts in state s at time step t and follows policy π thereafter. γ is a parameter, $0 < \gamma < 1$, called the *discount rate* that ensures the infinite sum (i.e., $\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1}$) converges to a finite value. More importantly, γ reflects the uncertainty in the future. r_{t+k+1} is the reward received at time step $t + k + 1$.

4. REINFORCEMENT LEARNING BASED HEV POWER MANAGEMENT

4.1 Motivations

Reinforcement learning provides a powerful solution to the problems in which (i) different actions should be taken according to the change of system states, and the future state depends on both the current state and the selected action; (ii) an expected cumulative return instead of an immediate reward will be optimized; (iii) the agent only needs knowledge of the current state and the reward it receives, while it needs not have knowledge of the system input in prior or the detailed system modeling; and (iv) the system might be non-stationary to some extent. The second, third, and fourth properties differentiate reinforcement learning from other machine learning techniques, model-based optimal control and dynamic programming, and Markov decision process-based approach, respectively.

The HEV power management problem, on the other hand, possesses all of the four above-mentioned properties. (i) During a driving cycle, the change of vehicle speed, power demand, and

battery charge level necessitates different operation modes and actions as discussed in Section 2, and also the future battery charge level depends on the battery charging/discharging current. (ii) The HEV power management aims at minimizing the total fuel consumption during a whole driving cycle rather than the fuel consumption rate at a certain time step. (iii) The HEV power management agent does not have *a priori* knowledge of a whole driving cycle, while it has only the knowledge of the current vehicle speed and power demand values and the current fuel consumption rate as a result of the action taken. (iv) The actual driving cycles are non-stationary [21]. Therefore, the reinforcement learning technique better suits the HEV power management problem than other optimization methods.

4.2 State, Action and Reward of HEV Power Management

4.2.1 State Space

We define the state space of the HEV power management problem as a finite number of states, each represented by the power demand, vehicle speed, and battery pack stored charge levels:

$$\mathcal{S} = \{s = [p_{dem}, v, q]^T | p_{dem} \in \mathcal{P}_{dem}, v \in \mathcal{V}, q \in \mathcal{Q}\}, \quad (16)$$

where p_{dem} is the power demand for propelling the HEV[1], which can be interpreted from the positions of the acceleration pedal and the brake pedal; q is the battery pack stored charge; $\mathcal{P}_{dem}$, $\mathcal{V}$, and $\mathcal{Q}$ are respectively the finite sets of power demand levels, vehicle speed levels, and battery pack stored charge levels. Discretization is required when defining these finite sets. In particular, $\mathcal{Q}$ is defined by discretizing the range of the battery pack stored charge i.e., $[q_{min}, q_{max}]$ into a finite number of charge levels:

$$\mathcal{Q} = \{q_1, q_2, \dots, q_N\}, \quad (17)$$

where $q_{min} \le q_1 < q_2 < \cdots < q_N \le q_{max}$. q_{min} and q_{max} are 40% and 80% of the battery pack capacity, respectively, in the SOC-sustaining power management for ordinary HEVs [8]~[10]. On the other hand, q_{min} and q_{max} are 0% and 80% of the battery pack capacity, respectively, in the SOC-depletion power management for plug-in HEVs (PHEVs) [13], in which the battery pack can be recharged from the power grid during parking time.

4.2.2 Action Space

We define the action space of the HEV power management problem as a finite number of actions, each represented by the discharging current of the battery pack and gear ratio values:

$$\mathcal{A} = \{a = [i, R(k)]^T | i \in I, R(k) \in R\}, \quad (18)$$

where an action $a = [i, R(k)]^T$ taken by the agent is to discharge the battery pack with a current value of i and choose the k-th gear ratio[2]. The set I contains within it a finite number of current values in the range of $[-I_{max}, I_{max}]$. Please note that $i > 0$ denotes discharging the battery pack; $i < 0$ denotes charging the battery pack; and $i = 0$ denotes idle. The set R contains the allowable gear ratio values, which depend on the drivetrain design. Usually, there are four or five gear ratio values in total [7], [14].

The above definition of the action space enables that the reinforcement learning agent does not require detailed HEV modeling (we will elaborate this in Section 4.4). The complexity and convergence speed of reinforcement learning algorithms are proportional to the number of state-action pairs [20]. In order to reduce computation complexity and accelerate convergence, we modify the action space to reduce the number of actions based on the HEV modeling. The reduced action space only contains charging/discharging current values of the battery pack:

$$\mathcal{A} = \{a = [i] | i \in I\}. \quad (19)$$

The inherent principle of reducing the action space is: with the selected action $a = [i]$, we can derive the best-suited gear ratio analytically when we have the knowledge of the HEV modeling. More precisely, we derive the best-suited gear ratio by solving the following ***fuel optimization*** (FO) problem:

Given the values of the current state $s = [p_{dem}, v, q]^T$ and the current action $a = [i]$, ***find*** the gear ratio $R(k)$ to ***minimize*** the fuel consumption rate $\dot{m}_f$ ***subject to*** (2)~(8).

Based on the current state $s = [p_{dem}, v, q]^T$ and the current action $a = [i]$, ω_{wh}, T_{wh}, and battery output power P_{batt} are calculated according to

$$\omega_{wh} = v / r_{wh}, \quad (20)$$

$$T_{wh} = p_{dem} \cdot r_{wh} / v, \quad (21)$$

and

$$P_{batt} = V^{OC} \cdot i - R_{batt} \cdot i^2, \quad (22)$$

where V^{OC} is the open-circuit voltage of the battery pack and R_{batt} is the internal resistance of the battery pack.

To solve the FO problem, for each of the possible $R(k)$ values, we first calculate ω_{ICE} and ω_{EM} using (7), next calculate T_{EM} using (4) while satisfying (5)~(6), and then calculate T_{ICE} using (8) while satisfying (2)~(3). With ω_{ICE} and T_{ICE}, the fuel consumption rate $\dot{m}_f$ is obtained based on the ICE model. We pick the $R(k)$ value that results in the minimum $\dot{m}_f$ i.e., $\dot{m}_f^{opt}$.

We will refer to the action space shown in (18) and (19) as the *original action space* and the *reduced action space*, respectively, in the following discussions.

4.2.3 Reward

We define the reward r that the agent receives after taking action a while in state s as the negative of the fuel consumption in that time step i.e., $-\dot{m}_f^{opt} \cdot \Delta T$, where ΔT is the length of a time step. Remember from Section 3 that the agent in reinforcement learning aims at maximizing the expected return i.e., the discounted sum of rewards. Therefore, by using the negative of the fuel consumption in a time step as the reward, the total fuel consumption will be minimized while maximizing the expected return.

4.3 TD(λ)-Learning Algorithm for HEV Power Management

To derive the optimal HEV power management policy, we employ a specific type of reinforcement learning algorithm, namely the TD(λ)-learning algorithm [22], due to its relatively higher convergence rate and higher performance in non-Markovian environment. In the TD(λ)-learning algorithm, a Q value, denoted by $Q(s, a)$, is associated with each state-action pair (s, a), which approximates the expected discounted cumulative

[1] The power demand p_{dem} instead of vehicle acceleration is selected as a state variable because: (i) the power demand can be interpreted from positions of acceleration and brake pedals, and (ii) experiments show that the power demand has higher correlation with actions in the system.

[2] According to the discussions in Section 2.5, the selected action will be sufficient to determine the values of all dependent variables in HEV control.

reward of taking action a at state s. There are two basic steps in the TD(λ)-learning algorithm: *action selection* and *Q-value update*.

4.3.1 Action Selection

A straightforward approach for action selection is to always choose the action with the highest Q value. If we do so, however, we are at the risk of getting stuck in a sub-optimal solution. A judicious reinforcement learning agent should exploit the best action known so far to gain rewards while in the meantime explore all possible actions to find a potentially better choice. We address this *exploration versus exploitation* issue by breaking the learning procedure into two phases: In the exploration phase, *ε-greedy-policy* is adopted, i.e., the current best action is chosen only with probability of $1-\varepsilon$. In the exploitation phase, the action with the highest Q value is always chosen.

4.3.2 Q-Value Update

Suppose that action a_t is taken in state s_t at time step t, and reward r_{t+1} and new state s_{t+1} are observed at time step $t+1$. Then at time step $t+1$, the TD(λ)-learning algorithm updates the Q value for each state-action pair (s,a) as:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot e(s,a) \cdot \delta, \tag{23}$$

where α is a coefficient controlling the *learning rate*, $e(s,a)$ is the *eligibility* of the state-action pair (s,a), and δ is calculated as

$$\delta \leftarrow r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t). \tag{24}$$

In (24), γ is the discount rate.

At time step $t+1$, the eligibility $e(s,a)$ of each state-action pair is updated by

$$e(s,a) \leftarrow \begin{cases} \gamma \cdot \lambda \cdot e(s,a) + 1, & s = s_t \cap a = a_t \\ \gamma \cdot \lambda \cdot e(s,a), & otherwise \end{cases} \tag{25}$$

to reflect the degree to which the particular state-action pair has been chosen in the recent past, where λ is a constant between 0 and 1. In practice, we do not have to update Q values and eligibility e of all state-action pairs. We only keep a list of M most recent state-action pairs since the eligibility of all other state-action pairs is at most λ^M, which is negligible when M is large enough.

4.3.3 Algorithm Description

The pseudo code of the TD(λ)-learning algorithm for HEV power management is summarized as follows.

TD(λ)-Learning Algorithm for HEV Management:

Initialize $Q(s,a)$ arbitrarily for all the state-action pairs.

For each time step t:

  Choose action a_t for state s_t using the exploration-exploitation policy discussed in Section 4.3.1.

  Take action a_t, observe reward r_{t+1} and the next state s_{t+1}.

  $\delta \leftarrow r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$.

  $e(s_t, a_t) \leftarrow e(s_t, a_t) + 1$.

  For all state-action pair (s,a):

    $Q(s,a) \leftarrow Q(s,a) + \alpha \cdot e(s,a) \cdot \delta$.

    $e(s,a) \leftarrow \gamma \cdot \lambda \cdot e(s,a)$.

  End

End

4.4 Model-Free Property Analysis

Theoretically, the reinforcement learning technique could be *model-free*, i.e., the agent does not require detailed system model to choose actions as long as it can observe the current state and reward as a result of an action previously taken by it. For the HEV power management problem, model-free reinforcement learning means that the controller (agent) should be able to observe the current state (i.e., power demand, vehicle speed, and battery pack charge levels) and the reward (i.e., the negative of fuel consumption in a time step) as a result of an action (i.e., battery pack discharging current and gear ratio selection), while the detailed HEV models are not needed by the controller. Now let us carefully examine whether the proposed reinforcement learning technique could be exactly model-free (or to which extent it could be model-free) in practical implementations.

For the reinforcement learning technique using the original action space: To observe the current state, the agent can use sensors to measure power demand level and the vehicle speed. And also, the reward can be obtained by measuring the fuel consumption. However, the battery pack charge level cannot be obtained directly from online measurement during HEV driving, since the battery pack terminal voltage changes with the charging/discharging current and therefore it could not be an accurate indicator of the battery pack stored charge level [23]. To address this problem, a battery pack model together with the Coulomb counting method [24] is needed by the agent. In summary, the reinforcement learning technique with the original action space is mostly model-free, i.e., only the battery pack model is needed.

For the reinforcement learning technique using the reduced action space: Given the current state and the action (charging/discharging current) taken, the agent should decide the gear ratio by solving the FO problem, where the ICE, the EM, the drivetrain mechanics and the battery pack models are needed. On the other hand, the vehicle dynamics model (discussed in Section 2.4) is not needed by the agent. In summary, the reinforcement learning technique with the reduced action space is partially model-free.

Table 1 summarizes the models needed by the agent for reinforcement learning technique with the original and reduced action spaces.

Table 1. Models needed for the original and reduced action spaces.

	Original action space	Reduced action space
ICE model	*no*	*needed*
EM model	*no*	*needed*
Drivetrain mechanics model	*no*	*needed*
Vehicle dynamics model	*no*	*no*
Battery pack model	*needed*	*needed*
Future driving cycle profile	*no*	*no*

4.5 Complexity and Convergence Analysis

The time complexity of the TD(λ)-learning algorithm in a time step is $O(|\mathcal{A}| + M)$, where $|\mathcal{A}|$ is the number of actions and M is the number of the most recent state-action pairs kept in memory. Generally, $|\mathcal{A}|$ and M are set to be less than 100. Therefore, the algorithm has negligible computation overhead when implementing in the state-of-the-art micro-controller/processors.

As for the convergence speed, normally, the TD(λ)-learning algorithm can converge within L time steps, where L is approximately three to five times of the number of state-action pairs. The total number of states could be as large as $|\mathcal{P}_{dem}| \cdot |\mathcal{V}| \cdot |\mathcal{Q}|$. However, some of the states do not have any physical meanings and will never be encountered by the system. And only 10% of the states are valid in the simulation. In summary, the

TD(λ)-learning algorithm can converge after two or three-hour driving, which is much shorter than the total lifespan of an HEV. To further speed up the convergence, the Q values can also be initialized by the manufacturers with optimized values.

4.6 Application-Specific Implementations

The actual implementation of the TD(λ)-learning algorithm for HEV power management can be application-specific. For example, the range of the battery pack stored charge level in the state space for PHEVs (SoC-depletion mode) is different from that for ordinary HEVs (SoC-sustaining mode). In the former case, it is more desirable to use up the energy stored in the battery pack by the end of a trip since the battery can be recharged from the power grid. Also, the parameters (e.g., α, γ, and λ) used in the TD(λ)-learning algorithm can be modified for different types of trips. For instances, the HEV controller can use different sets of parameters for urban trips from those for highway trips. Of course, the controller does not need the knowledge of detailed driving cycle profiles in prior.

5. EXPERIMENTAL RESULTS

We simulate the operation of an HEV based on Honda Insight Hybrid, the model of which is developed in ADVISOR [16]. Key parameters are summarized in Table 2. We compare our proposed optimal power management policy derived by reinforcement learning with the rule-based power management policy described in [7] using both real-world and testing driving cycles. A driving cycle is given as a vehicle speed versus time profile for a specific trip. The driving cycles may come from real measurements or from specialized generation for testing purposes. In this work, we use the real-world and testing driving cycles provided by different organizations and projects such as U.S. EPA (Environmental Protection Agency), E.U. MODEM (Modeling of Emissions and Fuel Consumption in Urban Areas project) and E.U. ARTEMIS (Assessment and Reliability of Transport Emission Models and Inventory Systems project).

Table 2. Honda Insight Hybrid key parameters.

Vehicle C_D	0.32	ICE Max power (kW)	50
Vehicle A_F (m^2)	1.48	ICE Max Torque (Nm)	89.5
Vehicle r_{wh} (m)	0.3	EM Max power (kW)	10
Vehicle m (kg)	1000	Battery capacity (Ah)	6.5
Reduction gear ratio ρ_{reg}	1.4	Battery voltage (V)	144

We improve the battery pack model used in ADVISOR to take into account the *rate capacity effect* and the *recovery effect*. Specifically, the majority of literature on HEV power management adopts a simple battery pack model as follows [3]:

$$q_t = q_{ini} - \sum_{k=0}^{t} I_k \cdot \Delta T, \quad (26)$$

where q_t is the amount of charge stored in the battery pack at the end of time step t, q_{ini} is the amount of charge stored in the battery pack at the beginning of time step 0, I_t is the discharging current of the battery pack at time step t ($I_t < 0$ means battery charging), and ΔT is the length of a time step. However, this model ignores the rate capacity effect, which causes the most significant power loss when the battery pack charging/discharging current is high [23]. We know that the battery pack charging/discharging current is high during deceleration and acceleration, and therefore the rate capacity effect should be considered carefully. The rate capacity effect specifies that if the battery pack is discharging ($I > 0$), the actual charge decreasing rate inside the battery pack is higher than I; and if the battery pack is charging ($I < 0$), the actual charge increasing rate inside the battery pack is lower than $|I|$. In addition, the battery model (26) also ignores the recovery effect, which specifies that the battery pack can partially recover the charge loss in previous discharges if relaxation time is allowed in between discharges [23].

Table 3. Fuel consumption of an ordinary HEV using proposed and rule-based policies.

Driving Cycle	Proposed Policy	Rule-based Policy	Reduction
IM240	68.5 g	92.2 g	25.7 %
LA92	426.6 g	585.3 g	27.1 %
NEDC	229.4 g	319.8 g	28.3 %
NYCC	38.8 g	86.1 g	54.9 %
HWFET	223.7 g	364.0 g	38.5 %
MODEM_1	151.7 g	228.6 g	33.6 %
MODEM_2	246.5 g	344.9 g	28.5 %
MODEM_3	75.8 g	137.1 g	44.7 %
Artemis_urban	128.9 g	220.5 g	41.5 %
Artemis_rural	460.3 g	499.7 g	7.9 %
total	2050.2 g	2878.2 g	28.8 %

First, we test the fuel consumption of an ordinary HEV in the battery SOC-sustaining mode using the proposed power management policy and the rule-based policy. The fuel consumption over some driving cycles is summarized in Table 3. We can observe that the proposed policy always results in lower fuel consumption and the maximum reduction in fuel consumption is as high as 54.9%. The last row in Table 3 shows that the proposed policy can reduce the fuel consumption by 28.8% on average. We also compare the overall fuel economy of the proposed policy and the rule-based policy over the 10 real-world and testing driving cycles in Table 3. The rule-based policy achieves an MPG value of 48 and the proposed policy achieves an MPG value of 67. Therefore, the proposed policy improves the

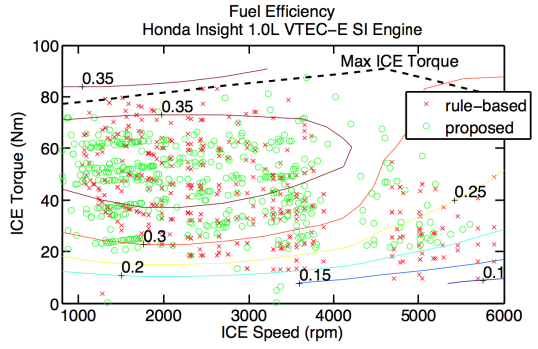

Figure 5. ICE operation points of an ordinary HEV from the proposed and rule-based policies.

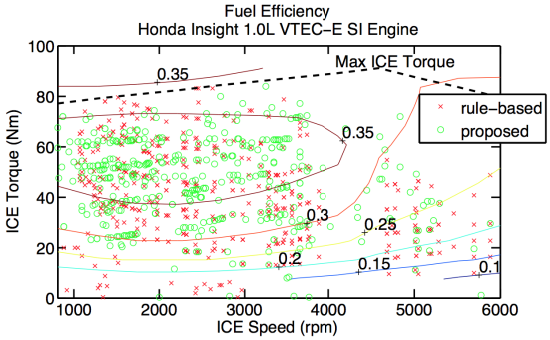

Figure 6. ICE operation points of a PHEV from the proposed and rule-based policies.

fuel economy by 39% compared to the rule-based policy in the ordinary HEV.

We plot the ICE operation points over a driving cycle on the ICE fuel efficiency map in Figure 5. The "x" points are from rule-based policy and the "o" points are from our proposed policy. We can observe that the operation points from the proposed policy are more concentrated on the high efficiency region of the ICE, validating the effectiveness of the proposed policy.

We also test the fuel consumption of a PHEV in the battery SOC-depletion mode using the proposed power management policy and the rule-based policy. Again, the proposed policy always results in lower fuel consumption. The proposed policy can reduce the fuel consumption by 60.8% in maximum and 30.4% on average. The MPG value of the rule-based policy over the 10 driving cycles is 55 and the MPG value of the proposed policy over the 10 driving cycles is 78. Therefore, the proposed policy improves the fuel economy by 42% compared to the rule-based policy in the PHEV. In addition, comparing Table 4 with Table 3, we can observe that the PHEV usually has higher fuel economy than the ordinary HEV. We also plot the ICE operation points over a driving cycle on the ICE fuel efficiency map in Figure 6. We can observe that the operation points from the proposed policy are more concentrated on the high efficiency region of the ICE, again validating the effectiveness of the proposed policy.

Table 4. Fuel consumption of a PHEV using proposed and rule-based policies.

Driving Cycle	Proposed Policy	Rule-based Policy	Reduction
IM240	42.4 g	52.8 g	19.7 %
LA92	408.1 g	544.6 g	25.1 %
NEDC	214.1 g	270.2 g	20.8 %
NYCC	24.4 g	62.3 g	60.8 %
HWFET	193.1 g	323.0 g	40.2 %
MODEM_1	110.6 g	192.7 g	42.6 %
MODEM_2	191.0 g	318.0 g	39.9 %
MODEM_3	50.0 g	108.0 g	53.7 %
Artemis_urban	100.3 g	200.8 g	50.0 %
Artemis_rural	422.1 g	451.8 g	6.6 %
total	1756.1 g	2524.2 g	30.4 %

6. CONCLUSION

The HEV features a hybrid propulsion system consisting of one ICE and one or more EMs. The use of both ICE and EM improves the performance and fuel economy of HEVs but also increases the complexity of HEV power management. Our proposed approach minimizes the HEV fuel consumption over any driving cycles. Different from previous works, our strategy derives the optimal HEV power management policy using reinforcement learning, which requires neither complete nor stochastic information of the driving cycles in prior, and only partial information of detailed HEV modeling. Simulation results over real-world and testing driving cycles demonstrate the effectiveness of the proposed HEV power management policy.

7. ACKNOWLEDGMENTS

This work is supported in part by a grant from the Directorate for Computer & Information Science & Engineering of the NSF, and the Mid-Career Researcher Program and the International Research & Development Program of the NRF of Korea funded by the MSIP (NRF-2014-023320).

8. REFERENCES

[1] C. C. Chan, "The state of the art of electric, hybrid, and fuel cell vehicles," *Proceedings of the IEEE*, vol. 95, pp. 704-718, Apr. 2007.

[2] F. R. Salmasi, "Control strategies for hybrid electric vehicles: evolution, classification, comparison, and future trends," *IEEE Trans. Vehicular Technology*, vol. 56, pp. 2393-2404, Sep. 2007.

[3] M. Ehsani, Y. Gao, and A. Emadi, *Modern electric, hybrid electric, and fuel cell vehicles: fundamentals, theory, and design*, CRC press, 2009.

[4] M. Ahman, "Assessing the future competitiveness of alternative powertrains," *Int. J. of Vehicle Design*, vol. 33, no. 4, pp. 309-331, 2003.

[5] A. Emadi, K. Rajashekara, S. S. Williamson, and S. M. Lukic, "Topological overview of hybrid electric and fuel cell vehicular power system architectures and configurations," *IEEE Trans. Vehicular Technology*, vol. 54, pp. 763-770, May 2005.

[6] H. D. Lee, E. S. Koo, S. K. Sul, and J. S. Kim, "Torque control strategy for a parallel-hybrid vehicle using fuzzy logic," *IEEE Industry Applications Magazine*, vol. 6, pp. 33-38, Nov. 2000.

[7] N. J. Schouten, M. A. Salman, and N. A. Kheir, "Fuzzy logic control for parallel hybrid vehicles," *IEEE Trans. Control Systems Technology*, vol. 10, pp. 460-468, May 2002.

[8] A. Brahma, Y. Guezennec, and G. Rizzoni, "Optimal energy management in series hybrid electric vehicles," in *Proc. American Control Conf.*, 2000, pp. 60-64.

[9] C. C. Lin, H. Peng, J. W. Grizzle, and J. M. Kang, "Power management strategy for a parallel hybrid electric truck," *IEEE Trans. Control Systems Technology*, vol. 11, pp. 839-849, Nov. 2003.

[10] L. V. Perez, G. R. Bossio, D. Moitre, and G. O. Garcia, "Optimization of power management in an hybrid electric vehicle using dynamic programming," *Mathematics and Computers in Simulation*, vol. 73, pp. 244-254, Nov. 2006.

[11] G. Paganelli, M. Tateno, A. Brahma, G. Rizzoni, and Y. Guezennec, "Control development for a hybrid-electric sport-utility vehicle: strategy, implementation and field test results," in *Proc. American Control Conf.*, 2001, pp. 5064-5069.

[12] P. Pisu, and G. Rizzoni, "A comparative study of supervisory control strategies for hybrid electric vehicles," *IEEE Trans. Control Systems Technology*, vol. 15, pp. 506-518, May 2007.

[13] Q. Gong, Y. Li, and Z. R. Peng, "Trip-based optimal power management of plug-in hybrid electric vehicles," *IEEE Trans. Vehicular Technology*, vol. 57, pp. 3393-3401, Nov. 2008.

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

[15] 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 Systems Technology*, vol. 19, pp. 545-555, May 2011.

[16] National Renewable Energy Lab. ADVISOR 2003 documentation. http://bigladdersoftware.com/advisor/docs/advisor_doc.html.

[17] J. M. Kang, I. Kolmanovsky, and J. W. Grizzle, "Dynamic optimization of lean burn engine aftertreatment," *J. Dyn. Sys., Meas., Control*, vol. 123, pp. 153-160, Jun. 2001.

[18] S. Delprat, J. Lauber, T. M. Guerra, and J. Rimaux, "Control of a parallel hybrid powertrain: optimal control," *IEEE Trans. Vehicular Technology*, vol. 53, pp. 872-881, May 2004.

[19] S. Kermanil, S. Delprat, T.M. Guerra, and R. Trigui, "Predictive control for HEV energy management: experimental results," in *Proc. Vehicle Power and Propulsion Conf.*, 2009, pp. 364-369.

[20] R. S. Sutton, and A. G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, Cambridge, MA, 1998.

[21] Dynamometer Drive Schedules. http://www.epa.gov/nvfel/testing/dynamometer.htm.

[22] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9-44, 1988.

[23] D. Linden, and T. B. Reddy. *Handbook of Batteries*. McGrew-Hill Professional, 2001.

[24] G. L. Plett, "Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part 1. Background," *Journal of Power Sources*, vol. 134, pp. 252-261, 2004.