Machine Learning-Based Energy Management in a Hybrid Electric Vehicle to Minimize Total Operating Cost

Xue Lin, Paul Bogdan
University of Southern California
Los Angeles, California 90089
Email: {xuelin, pbogdan}@usc.edu

Naehyuck Chang
Korea Advanced Institute of Science and Technology
Daejeon, 305-701, Korea
Email: naehyuck@cad4x.kaist.ac.kr

Massoud Pedram
University of Southern California
Los Angeles, California 90089
Email: pedram@usc.edu

***Abstract*—This paper investigates the energy management problem in hybrid electric vehicles (HEVs) focusing on the minimization of the operating cost of an HEV, including both fuel and battery replacement cost. More precisely, the paper presents a nested learning framework in which both the optimal actions (which include the gear ratio selection and the use of internal combustion engine versus the electric motor to drive the vehicle) and limits on the range of the state-of-charge of the battery are learned on the fly. The inner-loop learning process is the key to minimization of the fuel usage whereas the outer-loop learning process is critical to minimization of the amortized battery replacement cost. Experimental results demonstrate a maximum of 48% operating cost reduction by the proposed HEV energy management policy.**

I. Introduction

Electric vehicles (EVs) and hybrid electric vehicles (HEVs) have been gaining market share nowadays in the automotive market due to the concerns about large amounts of fuel consumption and pollution resulted from the conventional internal combustion engine (ICE) vehicles [21]. By integrating electric motors (EMs) into the vehicle propulsion system, EVs and HEVs achieve higher energy efficiency and lower pollution emission compared with the conventional vehicles [9].

HEVs, which represent a transition from conventional ICE vehicles to full electric vehicles, have higher fuel efficiency than conventional vehicles and fewer battery-related problems than EVs. However, due to the hybrid structure of the propulsion system, advanced HEV energy management techniques are needed to fully explore the advantages of HEVs [24]. The hybrid propulsion system of an HEV consists of an ICE and one or more EMs. The ICE converts chemical energy of the fuel into mechanical energy to propel the vehicle. The EM converts electrical energy stored in the battery pack to propel the vehicle, and it can also operate as a generator collecting kinetic energy during braking to charge the battery pack, which is called the regenerative braking, a mechanism improving the energy efficiency of EVs and HEVs [10]. HEV energy management techniques coordinate the operation of ICE and EM to improve the energy efficiency of HEVs.

The fuel cost is one major operating cost component of the HEV. Therefore, the majority of previous work on HEV energy management aimed at improving the fuel economy. The rule-based strategies for HEV energy management interpret the driver controlled pedal motion into the required propulsion power, and determine the power split between the ICE and the EM based on intuition, human expertise or fuzzy logic [5], [7]. The optimization-based control strategies either minimize the fuel consumption during a trip with given, predicted or stochastic future driving profile [22], [8], [19], or perform control by converting battery charge into equivalent fuel consumption (ECMS and adaptive-ECMS approaches) [11], [20].

The state-of-health (SoH) of the HEV battery pack is degrading with the operation of an HEV due to the frequent charging/discharging of the battery pack by the EM. The work [14], [18] studied the SoH degradation model for the EV/HEV battery pack as a function of the state-of-charge (SoC) swing, the number of charging/discharging cycles, etc. The battery pack will reach its end-of-life when its SoH degrades to 80% or 70% [17] and the battery pack replacement will result in additional operating cost of the HEV. Enlarging the battery pack energy capacity within size, weight and cost constraints is preferred for exploring the energy storage capability of the battery pack to improve the HEV fuel economy, and especially the plug-in HEV (PHEV) employs a higher-capacity battery pack [27], which is charged using the grid power. The battery replacement cost increases significantly with enlarged battery capacity, and therefore the amortized battery replacement cost must not be neglected in the HEV. There are some work taking into account battery SoH degradation when optimizing the fuel efficiency [12], [26], [25]. However, these work have one or more of the following shortcomings: (i) The HEV energy management policies they use are based on ECMS or adaptive-ECMS approaches [11], [20], which rely on the knowledge of the future driving profile. If the prediction of the future driving profile is not accurate, the effectiveness of these ECMS and adapative-ECMS based approaches can be degraded. (ii) They do not use accurate analytical battery SoH degradation model in the optimization and evaluation, instead, they use Ah-throughput or battery output power as the equivalent of the battery SoH degradation during charging and discharging processes.

Machine learning provides a powerful tool for the agent (i.e., decision-maker) to "learn" how to "act" optimally when the explicit and accurate system modeling is difficult or even impossible to obtain [4]. The agent can observe the environment's *state* and take an *action* according to the observed state. A *reward* will be given to the agent as a result of the action taken. Stimulated by the reward, the agent aims to derive a policy, which is a mapping from each possible state to an action, by "learning" from its past experience. The reinforcement learning has been applied to the HEV energy management

problem [16], such that the HEV energy management policy does not rely on any knowledge of the future driving profile. An inverse reinforcement learning technique [29] has been applied for learning the driver behavior, however, it is out of our focus.

In this proposed work, we investigate the HEV energy management problem focusing on the minimization of the operating cost of an HEV, including both fuel cost and amortized battery replacement cost (i.e., battery purchase plus installation cost). We present a nested learning framework in which both the optimal actions (which include the gear ratio selection and the use of ICE versus EM to drive the vehicle) and limits on the range of battery SoC are learned on the fly. More precisely, the inner-loop learning process determines the operation modes of the HEV components whereas the outer-loop learning process modulates the battery SoH degradation from a global point of view. Due to the usage of the machine learning techniques, the proposed HEV energy management does not rely on perfect and accurate system modeling (i.e., HEV component modeling and driving profile modeling.) The proposed nested learning framework for HEV energy management differs from the reinforcement learning-based framework [16] in that (i) the amortized battery replacement cost is incorporated into the HEV energy management; and (ii) two nested learning processes are used in which the inner-loop learning process is the key to minimization of the fuel usage and the outer-loop learning process is critical to minimization of the amortized battery replacement cost. Experimental results demonstrate a maximum of 48% operating cost reduction by the proposed HEV management policy.

II. System Description

Although this work aims to design a smart HEV controller that discovers the energy management policy by learning from its experience, it is still necessary to understand the fundamentals of HEV operation. By way of an example and without loss of generality, we discuss the parallel HEV configuration as in most of the literature work on HEV energy management [10]. There are five operation modes of a parallel HEV, depending on the flow of energy: (i) only the ICE propels the vehicle, (ii) only the EM propels the vehicle, (iii) the ICE and EM propel the vehicle in parallel, (iv) the ICE propels the vehicle and at the same time drives the EM to charge the battery pack, and (v) the EM charges the battery pack when the vehicle is braking (i.e., regenerative braking mode.)

A. HEV Component Analysis

1) Internal Combustion Engine (ICE): According to the quasi-static ICE model [15], the fuel efficiency of an ICE is calculated as

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

In (1), T_{ICE} and ω_{ICE} are the torque (in N·m) and speed (in rad/s) of the ICE, respectively, which represent the operation point of the ICE. $\dot{m}_f$ is the fuel consumption rate (in g/s) of the ICE, depending on the ICE operation point. And D_f is the fuel energy density (in J/g). Figure 1 (a) represents the contour map of the fuel consumption rate of an example ICE in the ICE speed-torque plane. Figure 1 (b) shows the corresponding fuel efficiency contour map. To ensure safe and smooth operation

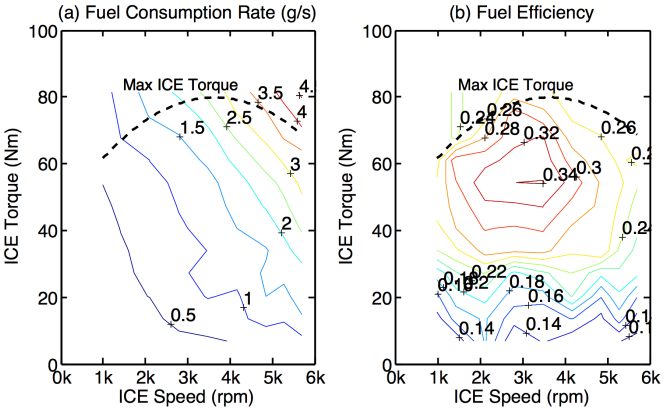

Fig. 1. The (a) fuel consumption rate map and (b) fuel efficiency map of an ICE.

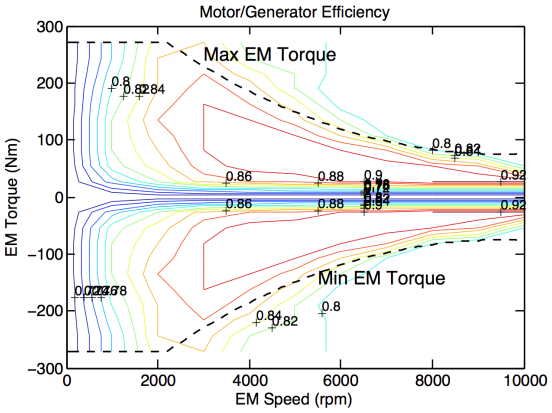

Fig. 2. The efficiency map of an EM.

of an ICE, the following constraints should be satisfied:

$$\begin{aligned} \omega_{ICE}^{min} \leq \quad & \omega_{ICE} \quad \leq \omega_{ICE}^{max}, \\ 0 \leq \quad & T_{ICE} \quad \leq T_{ICE}^{max}(\omega_{ICE}). \end{aligned} \quad (2)$$

2) Electric Motor (EM): The EM operates either as a motor to propel the vehicle or as a generator to charge the battery pack. The efficiency of the EM is

$$\eta_{EM}(T_{EM}, \omega_{EM}) = \begin{cases} (T_{EM} \cdot \omega_{EM})/P_{batt} & T_{EM} \geq 0 \\ P_{batt}/(T_{EM} \cdot \omega_{EM}) & T_{EM} < 0 \end{cases} \quad (3)$$

where T_{EM} and ω_{EM} are respectively the torque and speed of the EM, and P_{batt} is the output power of the battery pack. When the EM operates as a motor, T_{EM} is positive and the battery pack is discharging i.e., $P_{batt} > 0$; when the EM operates as a generator, T_{EM} is negtive and the battery pack is charging i.e., $P_{batt} < 0$. Figure 2 represents the efficiency contour map of the EM as a motor or a generator. To ensure safe and smooth operation of an EM, the following constraints should be satisfied:

$$\begin{aligned} 0 \leq \quad & \omega_{EM} \quad \leq \omega_{EM}^{max}, \\ T_{EM}^{min}(\omega_{EM}) \leq \quad & T_{EM} \quad \leq T_{EM}^{max}(\omega_{EM}). \end{aligned} \quad (4)$$

3) Vehicle Tractive Force: The vehicle tractive force F_{TR} to support the vehicle speed and acceleration (which are

determined by the driver through pressing the braking or acceleration pedal) is derived by

$$\begin{aligned} F_{TR} &= m\cdot a + F_g + F_R + F_{AD} \\ F_g &= m\cdot g\cdot \sin\theta \\ F_R &= m\cdot g\cdot \cos\theta\cdot C_R \\ F_{AD} &= 0.5\cdot \rho\cdot C_D\cdot A_F\cdot v^2 \end{aligned} \tag{5}$$

where m is the vehicle mass, a is the vehicle acceleration, F_g is the force due to road slope, F_R is the rolling friction force, F_{AD} is the air drag force, θ is the road slope angle, C_R is the rolling friction coefficient, ρ is the air density, C_D is the air drag coefficient, A_F is the vehicle frontal area, and v is the vehicle speed. Given v, a and θ, the tractive force F_{TR} can be derived using (5). Then, the vehicle wheel torque T_{wh} and wheel speed ω_{wh} are related to F_{TR}, v, and wheel radius r_{wh} by

$$\begin{aligned} T_{wh} &= F_{TR}\cdot r_{wh}, \\ \omega_{wh} &= v/r_{wh}. \end{aligned} \tag{6}$$

The demanded power for propelling the vehicle i.e., p_{dem} satisfies

$$p_{dem} = F_{TR}\cdot v = T_{wh}\cdot \omega_{wh}. \tag{7}$$

4) Drivetrain Coupling: The ICE and EM are coupled together through the drivetrain to propel the vehicle cooperatively. The speed and torque of the ICE, the EM, and the vehicle wheel obey the following speed and torque relation:

$$\begin{aligned} \omega_{wh} &= \frac{\omega_{ICE}}{R(k)} = \frac{\omega_{EM}}{R(k)\cdot\rho_{reg}} \\ T_{wh} &= R(k)\cdot(T_{ICE} + \rho_{reg}\cdot T_{EM}\cdot(\eta_{reg})^{\alpha})\cdot(\eta_{gb})^{\beta} \end{aligned} \tag{8}$$

where

$$\alpha = \begin{cases} +1 & T_{EM}\geq 0 \\ -1 & T_{EM}<0 \end{cases} \tag{9}$$

$$\beta = \begin{cases} +1 & T_{ICE} + \rho_{reg}\cdot T_{EM}\cdot(\eta_{reg})^{\alpha}\geq 0 \\ -1 & T_{ICE} + \rho_{reg}\cdot T_{EM}\cdot(\eta_{reg})^{\alpha}<0 \end{cases} \tag{10}$$

In (8), $R(k)$ is the k-th gear ratio (there are a total number of four or five gear ratios), ρ_{reg} is the reduction gear ratio, and η_{reg} and η_{gb} are the reduction gear efficiency and the gear box efficiency, respectively.

B. HEV Control Flow

During the actual HEV control process, it is the driver that determines the speed v and power demand p_{dem} (or equivalently, the speed v and acceleration a) for propelling the vehicle on the fly through pressing the acceleration or brake pedal. Then the HEV controller controls the operation of the ICE, EM and drivetrain such that the vehicle meets the target performance (speed v and acceleration a.) Generally, the HEV controller chooses a couple of control variables, such as the battery output power P_{batt} (or equivalently, the battery output current i) and the gear ratio $R(k)$, etc., and then the rest of the variables (i.e., the ICE torque T_{ICE} and speed ω_{ICE}, the EM torque T_{EM} and speed ω_{EM}) become the dependent (associate) variables, the values of which are determined by P_{batt} and $R(k)$ according to the operational principle of HEV components as discussed previously.

There are HEV control strategies that rely on accurate HEV system modeling, such as the dynamic programming-based strategy [22], the model predictive control strategy [8], and the equivalent consumption minimization strategy (ECMS) [11]. And also, there are model-free or partially model-free HEV control strategies that do not rely on detailed HEV system modeling or only need partial HEV modeling. For example, the rule-based strategies [5], [7] only need the battery modeling. The model-free HEV control strategies are preferred due to their flexibility and feasibility. The reinforcement learning-based strategy [16] generally can be a model-free or partially model-free HEV control framework.

C. Battery SoH Estimation

With the operation of an HEV, the SoH of the HEV battery is degrading i.e., the battery gradually loses its capacity. We say a battery reaches its end-of-life when its SoH degrades to 80% or 70% i.e., the battery capacity fading reaches 20% or 30% [17]. The battery capacity fading C_{fade} is formally defined as

$$C_{fade} = (1 - C_{full}/C_{full}^{nom})\times 100\%, \tag{11}$$

where C_{full} is the battery full charge capacity and C_{full}^{nom} is the nominal value of C_{full} i.e., the full charge capacity of a brand new battery. The battery capacity fading results from long-term electrochemical reactions involving the carrier concentration loss and internal impedance growth. We will discuss in the following the battery capacity fading model in [18], which shows a good match with real data but can only be applied for cycled charging/discharging pattern.

The state-of-charge (SoC) of a battery is given by

$$SoC = C_{batt}/C_{full}\times 100\%, \tag{12}$$

where C_{batt} is the amount of charge stored in the battery. A battery charging/discharging *cycle* is defined as a charging process of the battery from SoC_{low} to SoC_{high} and a subsequent discharging process from SoC_{high} to SoC_{low}. Then the average SoC and SoC swing in a cycle are calculated as

$$\begin{aligned} SoC_{avg} &= (SoC_{low} + SoC_{high})/2, \\ SoC_{swing} &= SoC_{high} - SoC_{low}. \end{aligned} \tag{13}$$

Reference [18] estimates the capacity fading of a battery in a charging/discharging cycle i.e., $C_{fade,cycle}$ as

$$\begin{aligned} D_1 &= K_{CO}\cdot\exp[(SoC_{swing}-1)\frac{T_{ref}}{K_{ex}T_B}] + 0.2\frac{\tau}{\tau_{life}} \\ D_2 &= D_1\cdot\exp[4K_{SoC}(SoC_{avg}-0.5)](1-C_{fade}) \\ C_{fade,cycle} &= D_2\cdot\exp[K_T(T_B - T_{ref})\frac{T_{ref}}{T_B}] \end{aligned} \tag{14}$$

where K_{CO}, K_{ex}, K_{SoC}, and K_T are battery specific parameters; T_B and T_{ref} are the battery temperature and reference temperature, respectively; τ is the duration of this charging/discharging cycle; τ_{life} is the calendar life of the battery. Please note that $C_{fade,cycle}$ is a function of SoC_{avg} and SoC_{swing}. The total capacity fading after M charging/discharging cycles is calculated by

$$C_{fade} = \Sigma_{m=1}^{M} C_{fade,cycle}(m), \tag{15}$$

where $C_{fade,cycle}(m)$ denotes the battery capacity fading in the m-th cycle. The battery capacity fading C_{fade} will increase over the battery lifetime from 0 (brand new) to 100% (no capacity

left.) Generally, $C_{fade} = 20\%$ or 30% is used to indicate end-of-life of the battery.

The battery capacity fading model in [18] can only be applied to the cycled charging/discharging pattern i.e., the battery experiences the charging/discharging cycles with the same SoC swing and the same average SoC. However, a battery may not follow a cycled charging/discharging pattern. A cycle-decoupling method [30] was proposed which can identify and decouple cycles from arbitrary battery charging/discharging patterns. Then, the battery capacity fading in a cycle can be calculated using (14) and the total capacity fading is derived using (15). Moreover, the battery internal resistance grows with increased C_{fade} value and thereby reducing the output power rating of the battery. This is called the battery power fading effect [18]. Therefore, the battery end-of-life criterion i.e., $C_{fade} = 20\%$ or $C_{fade} = 30\%$ also indicates significant degradation in the battery output power during battery aging process [1].

III. A Nested Learning Framework for HEV Energy Management

In this work, we aim to minimize the operating cost of an HEV including both fuel cost and amortized battery replacement cost. To achieve this goal, we propose a nested learning framework for HEV energy management, in which the optimal actions to propel the vehicle and the limits on the change in the SoC of the battery are learned on the fly by the inner-loop reinforcement learning and the outer-loop adaptive learning, respectively. The inner-loop reinforcement learning process is the key to minimization of the fuel usage, whereas the outer-loop adaptive learning process is critical to minimization of the amortized battery replacement cost.

A. Motivation

We use reinforcement learning in the inner loop due to the following reasons. (i) The inner-loop HEV energy management aims to minimize the total fuel consumption during a driving trip rather than the instantaneous fuel consumption rate at each time step; the reinforcement learning also aims to optimize an expected cumulative return (16) rather than an immediate reward. (ii) During a driving trip, the changes of vehicle speed, power demand and battery charge level require different HEV operation modes; the reinforcement learning agent takes different actions depending on the current state. (iii) The inner-loop HEV energy management does not have *a priori* knowledge of a whole driving trip, 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 an action taken previously; the reinforcement learning agent only needs the knowledge of the current state and the current reward in order to learn the optimal policy, while it needs not have knowledge of the system input in prior or the detailed system modeling. The inner loop is the key to minimization of the fuel usage, however, we also consider battery SoH degradation in the inner loop by incorporating the battery capacity fading term into the reward of the reinforcement learning, such that the inner loop itself can be used as an independent HEV energy management framework for minimizing the total operating cost.

[1] In addition, the battery calendar life also affects the battery SoH degradation, but it is out of the focus of this work, since we only focus on the HEV energy management.

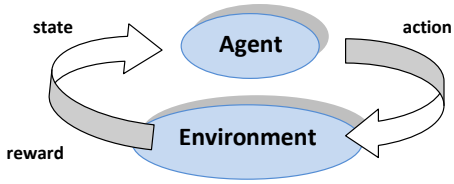

Fig. 3. The interactions between agent and environment.

In the previous work on HEV energy management, a fixed battery SoC range is used i.e., the battery pack SoC is clamped by fixed lower bound and upper bound. Then, the resultant HEV energy management strategies may tend to use up the available battery energy to improve fuel economy even for some very short urban trips, which may harm battery SoH seriously. The battery can obtain significant amount of energy by regenerative braking in urban trips. It is not always necessary to use up the available battery energy. Simulation results demonstrate that the battery SoC swing of $<15\%$ is enough to improve fuel economy in short urban trips due to the regenerative braking. Therefore, we use adaptive learning in the outer loop to learn the optimal SoC range for different types of trips, because the SoC range is an important factor determining the battery SoH degradation. The outer loop is critical to minimization of the amortized battery replacement cost, since it modulates the battery SoH degradation globally. The outer loop performs better than the inner loop in terms of reducing the amortized battery replacement cost if it has prior knowledge of the driving trips such as trip length and average speed, which are given as input by the driver at the beginning of each trip. In case such kind of knowledge is inaccurate or not available, we can rely on the inner loop for reducing the total operating cost.

B. Inner-Loop Reinforcement Learning Process

1) Reinforcement Learning Background: In reinforcement learning, the decision-maker is called the *agent* and everything outside the agent is called the *environment*. Figure 3 illustrates the agent-environment interaction at each of a sequence of discrete time steps $t = 0, 1, 2, ...$ At each time step t, the agent observes the environment's *state* $s_t \in \mathcal{S}$ and on that basis takes an *action* $a_t \in \mathcal{A}$, where $\mathcal{S}$ and $\mathcal{A}$ are the sets of possible states and actions, respectively. One time step later, in part as a consequence of the action taken, the agent receives a numerical *reward* r_{t+1} and finds the environment in a new state s_{t+1}.

A policy 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

$$V^{\pi}(s) = E\left\{\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \mid s_t = s\right\} \tag{16}$$

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 $t+k+1$.

2) State Space: We define the state space of the inner-loop reinforcement learning

$$\mathcal{S} = \left\{ s = [p_{dem}, v, q]^T | p_{dem} \in \mathcal{P}_{dem}, v \in \mathcal{V}, q \in Q \right\}, \quad (17)$$

where p_{dem} is the power demand for propelling the HEV, v is the vehicle speed, and q is the charge stored in the battery pack. Different actions should be taken under different states. For example, if the power demand is negative i.e., the vehicle is braking, the action taken by the agent (i.e., HEV controller) should be charging the battery by using the EM as a generator. On the other hand, if the power demand is a very large positive value, the action should be discharging battery to power the EM, which propels the vehicle in assistance with ICE.

A reinforcement learning agent should be able to observe a state. In the actual implementation of the inner-loop reinforcement learning, the current power demand level p_{dem} and vehicle speed level v can be obtained by using sensors to measure the driver controlled pedal motion. However, the charge level q cannot be obtained from online measurement of terminal voltage, since the battery pack terminal voltage changes with the charging/discharging current and therefore it cannot be an accurate indicator of q [17]. To observe the charge level q, the Coulomb counting method [23] is needed by the agent, which is typically realized using a dedicated circuit [2].

$\mathcal{P}_{dem}$, $\mathcal{V}$, and Q in (17) are respectively the finite sets of power demand levels, vehicle speed levels, and levels of charge stored in the battery pack. Discretization is required when defining these finite sets. In particular, Q is defined by discretizing the range of charge stored in the battery pack i.e., $[q_{min}, q_{max}]$ into a finite number of charge levels:

$$Q = \{q_1, q_2, ..., q_N\}, \quad (18)$$

where $q_{min} = q_1 < q_2 < ... < q_N = q_{max}$. Generally, q_{min} and q_{max} are 40% and 80% of the battery pack nominal capacity, respectively, in the charge-sustaining energy management for ordinary HEVs [22]; 0% and 80%, respectively, in the charge-depleting energy management for PHEVs [13]. In the outer-loop adaptive learning process, we will optimize q_{min} value to modulate the battery SoH degradation and q_{max} is usually fixed in the HEV control.

3) Action Space: We define the action space of the inner-loop reinforcement learning as a finite number of actions, each represented by the discharging current of the battery pack and the gear ratio value:

$$\mathcal{A} = \left\{ a = [i, R(k)]^T | i \in I, R(k) \in R \right\}, \quad (19)$$

where an action $a = [i, R(k)]^T$ taken by the agent is to discharge the battery pack with current i and choose the k-th gear ratio. 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; and $i < 0$ denotes charging the battery pack. 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 [8].

Alternatively, we can define a reduced action space $\mathcal{A}_{re}$, in which an action $a_{re} = [i]$ is to discharge the battery pack with current i (and the gear ratio $R(k)$ is selected by solving an optimization problem such that the resultant fuel consumption rate is minimized.) The complexity and convergence speed of reinforcement learning algorithms are proportional to the number of state-action pairs [6]. Therefore, the reduced action space $\mathcal{A}_{re}$ helps to reduce the complexity and increase convergence speed by a factor of four or five. However, this reduced action space relies on HEV component modeling when solving the optimization problem. In summary, we can either use the original action space (19) for model-free control or use the reduced action space $\mathcal{A}_{re}$ for reduced complexity and increased convergence rate.

4) Reward: The objective of the inner-loop reinforcement learning is to minimize the HEV operating cost including both fuel cost and amortized battery replacement cost. Therefore, we define the reward r that the agent receives after taking action a in state s as the negative of the weighted sum of the fuel consumption and battery capacity fading in that time step i.e., $-\dot{m}_f \cdot \Delta T - w \cdot \Delta C_{fade}$, where ΔT is the length of a time step, w is the weight of battery capacity fading (w is determined by the ratio of the fuel cost to the amortized battery cost), and $\dot{m}_f$ and ΔC_{fade} are respectively the fuel consumption rate and battery capacity fading in that time step. The reinforcement learning agent aims to maximize the expected return (16), which is a discounted sum of rewards. Therefore, by using the negative of the weighed sum of the fuel consumption and battery capacity fading in a time step as the reward, the fuel consumption and battery capacity fading will be minimized while maximizing the expected return.

For the implementation of the inner-loop reinforcement learning, the agent (HEV controller) should be aware of the reward it receives after taking an action, since the observation of reward is critical in deriving an optimal policy. In the above-mentioned reward definition, the $\dot{m}_f \cdot \Delta T$ part can be obtained by measuring the fuel consumption directly. The ΔC_{fade} part cannot be obtained by online measurement[2]. A battery SoH estimation method is needed. The HEV battery does not follow a cycled charging/discharging pattern and it could be an arbitrary charging/discharging pattern. Therefore, the cycle-decoupling method [30] can be used for the battery SoH estimation. However, the time complexity of the cycle-decoupling method is high. We can use the following *equivalent cycle method* to derive ΔC_{fade}.

The reinforcement learning agent keeps a record of the battery charging/discharging profile i.e., $i(t)$ and therefore the battery SoC profile $SoC(t)$ from the beginning of a trip. Based on the discussion in Section II-C, the battery capacity fading in one charing/discharging cycle is a function of the average SoC and SoC swing, i.e., $C_{fade,cycle}(SoC_{avg}, SoC_{swing})$. The SoC_{avg} and SoC_{swing} values can be approximated by

$$\begin{aligned} SoC_{high} &= \max_t SoC(t), \\ SoC_{low} &= \min_t SoC(t), \\ SoC_{avg} &= \frac{SoC_{high} + SoC_{low}}{2}, \\ SoC_{swing} &= SoC_{high} - SoC_{low}. \end{aligned} \quad (20)$$

Then, we can calculate the battery capacity fading in a cycle $C_{fade,cycle}$ by (14). The total number of cycles that the battery pack has experienced so far in the trip is approximated as

$$N_C = \sum_t \frac{-i(t) \cdot \Delta T \cdot \boldsymbol{I}[i(t) < 0]}{C_{full} \cdot SoC_{swing}}, \quad (21)$$

[2]The battery SoH can only be measured offline by depleting and recharging the battery.

where the indicator function $\boldsymbol{I}[x]$=1 when x is true. The total battery capacity fading after taking action a is then calculated by

$$C_{fade} = C_{fade,cycle}(SoC_{avg}, SoC_{swing}) \cdot N_C. \quad (22)$$

Therefore, the ΔC_{fade} value can be calculated as

$$\Delta C_{fade} = C_{fade} - C'_{fade}, \quad (23)$$

where C'_{fade} and C_{fade} are the battery total capacity fading before and after taking action a.

The equivalent cycle method can be further improved for reduced complexity. The reinforcement learning agent does not need to keep a record of the battery charging/discharging profile from the beginning of a trip. Instead, only the latest battery current value $i(t)$ and the observed maximum and minimum SoC i.e., SoC_{high} and SoC_{low} are updated and kept in record, and then SoC_{swing} is updated. The ΔC_{fade} value can be calculated as

$$\Delta C_{fade} = \frac{-i(t) \cdot \Delta T \cdot \boldsymbol{I}[i(t) < 0]}{C_{full} \cdot SoC_{swing}} \cdot C_{fade,cycle}. \quad (24)$$

In this way, we reduce a global calculation into a local one with $O(1)$ complexity.

5) TD(λ)-Learning Algorithm: We adopt the TD(λ)-learning algorithm [28] for deriving the optimal policy of the inner-loop reinforcement learning, due to its relatively higher convergence rate and higher performance in non-Markovian environment. In this algorithm, a Q value, denoted by $Q(s,a)$, is associated with each state-action pair (s,a), where a state s is represented by the power demand p_{dem}, the vehicle speed v, and the battery charge level q, and an action a is to discharge the battery with current i and choose the k-th gear ratio. The $Q(s,a)$ value approximates the expected discounted cumulative reward of taking action a in state s. The TD(λ)-learning algorithm is summarized as follows.

In the TD(λ)-learning algorithm, the Q values are initialized arbitrarily at first. At each time step t, the agent first selects an action a_t for the current state s_t based on the $Q(s,a)$ values. To avoid the risk of getting stuck in a sub-optimal solution, the exploration-exploitation policy [28] is employed for the action selection, i.e., the agent does not always select the action a that results in the maximum $Q(s_t,a)$ value for the current state s_t. After taking the selected action a_t, the agent observes a new state s_{t+1} and receives reward r_{t+1}. Then, based on the observed s_{t+1} and r_{t+1}, the agent updates the $Q(s,a)$ values for all the state-action pairs, in which the *eligibility* $e(s,a)$ of each state-action pair is updated and utilized during the Q value update. The eligibility $e(s,a)$ of a state-action pair reflects the degree to which the particular state-action pair has been encountered in the recent past and λ is a constant between 0 and 1. Due to the use of the eligibility of the state-action pairs, we do not need 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.

6) Application Specific Improvement of the TD(λ)-Learning Algorithm: We modify the TD(λ)-learning algorithm to improve its performance and convergence rate in the HEV control scenario by accommodating different operation modes of an HEV. Specifically, when selecting an action for the current state, the agent takes into account the actual HEV operation mode besides the stored Q values. For example, if the power demand is negative i.e., the regenerative braking mode, the agent will definitely choose the maximum allowable charging current for the battery pack to harvest the kinetic energy as much as possible. If the battery charge level is very high, the agent will use EM power with higher likelihood to propel the vehicle. And if the battery charge level is very low, the agent is likely to use more ICE power to propel the vehicle and at the same time charge the battery. In summary, these application specific modifications significantly improve performance and convergence rate of TD(λ)-learning algorithm.

Algorithm 1 TD(λ)-Learning Algorithm for the Inner Loop

```
Initialize Q(s,a) arbitrarily for all the state-action pairs.
for each time step t do
    Choose action a_t for state s_t using the exploration-exploitation policy.
    Take action a_t, observe reward r_{t+1} and next state s_{t+1}.
    δ ← r_{t+1} + γ·max_{a'} Q(s_{t+1}, a') − Q(s_t, a_t).
    e(s_t, a_t) ← e(s_t, a_t) + 1.
    for all state-action pair (s,a) do
        Q(s,a) ← Q(s,a) + α·e(s,a)·δ.
        e(s,a) ← γ·λ·e(s,a).
    end for
end for
```

7) Complexity and Model-Free Analysis: The time complexity of the TD(λ)-learning algorithm at a time step is $O(|\mathcal{A}| + M)$, where $|\mathcal{A}|$ is the total number of actions and M is the number of the most recent state-action pairs kept in memory. Usually, $|\mathcal{A}| + M$ is within a few hundred, and therefore, the algorithm has negligible computation overhead. In terms of convergence rate, 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. In simulation, due to the application specific improvement, the TD(λ)-learning algorithm can converge within one hour driving, which is much shorter than the total lifespan of an HEV. To further speed up the convergence rate, the Q values can be initialized by the manufacturers with optimized values.

In theory, the reinforcement learning technique can be model-free. As long as the agent can observe the current state and be aware of the reward as a result of an action taken previously, the agent can find the optimal action selection policy during this observation-and-estimation process, whereas the detailed system modeling is not required. Based on the previous analysis, if the original action space is used, the HEV component modeling is not required by the agent, whereas the battery SoH estimation method is needed. If the reduced action space is used, the inner-loop reinforcement learning needs partial HEV component modeling (the ICE model, the EM model and the drivetrain model are needed, whereas the vehicle tractive force model and driving profile are not needed) besides the battery SoH estimation method.

C. Outer-Loop Adaptive Learning Process

The battery SoH degradation together with the fuel consumption have been taken into account at each time step in the driving by the inner-loop reinforcement learning. In this outer-loop adaptive learning process, the learning agent modulates the battery SoH degradation from a global point of view by tuning the maximum SoC range. More specifically, when we

define the state space of the inner loop, we have actually limited the battery SoC range within $[\frac{q_{min}}{C_{full}}, \frac{q_{max}}{C_{full}}]$. We know from Section II-C that the SoC range (from which SoC swing and average SoC can be derived) strongly affects the battery capacity fading. Therefore, in this outer loop the agent tunes the q_{min} value for different driving trip types such that battery capacity fading can be reduced.

1) State Space: We define the state s of the outer loop by using the trip characteristics including the trip length, average speed, and road condition (urban or highway). In the real implementation in an HEV, the outer-loop agent can obtain such trip information at the beginning of a trip from driver input. The battery usage (i.e., charging/discharging profile) should be different for different driving trip types, and therefore we use the trip characteristics as the state. For example, a smaller SoC range is enough for urban trips, whereas a larger SoC range should be used for highway trips. The state in the supervised machine learning technique is also called the feature.

2) Action Space: The action taken by the outer-loop agent is to choose the q_{min} value as discussed in Section III-B2, whereas q_{max} is usually fixed in the HEV control. Therefore, the SoC range during a trip can be clamped between the selected q_{min} value and the q_{max}. The action in the supervised machine learning technique is also called the target.

3) $C(s,a)$ Cost Function: If the system is in a state s (i.e., a specific trip type) and an action a (i.e., a q_{min} value) is taken by the learning agent, the agent will observe a cost value C, which is associated with each state-action pair by the cost function $C(s,a)$. The learning agent aims at minimizing the cost value when choosing an action for a state. In order to both improve the fuel economy and reduce the SoH degradation during a trip, we use the weighted sum of the fuel consumption and the SoH degradation during that trip as the cost function. The agent should be able observe the cost value after taking an action in a state. In the implementation of the outer-loop adaptive learning, the fuel consumption is obtained by online measurement whereas the battery SoH degradation during the trip is obtained using the SoH estimation method in Section II-C.

4) Adaptive Learning Algorithm: The outer-loop learning agent can choose the optimal action for the current state based on its past experience. When the system is in state s, the agent chooses the action that results in the minimum cost value,

$$a \leftarrow \arg\min_{a'} C(s,a'). \tag{25}$$

After taking action a, the agent observes the new cost value and updates the $C(s,a)$ value accordingly.

However, the outer-loop learning agent does not have the knowledge of the $C(s,a)$ values and therefore could not make decision on the action selection for a brand new HEV. To address this issue, the manufacturer can pre-set the $C(s,a)$ values by performing driving tests on the same type of HEV for different state and action combinations. This initialization of the $C(s,a)$ values is called regulation in the supervised machine learning technique.

The time complexity of the adaptive learning algorithm, which is performed for each driving trip, is $O(|\mathcal{A}|)$, where $|\mathcal{A}|$ is the number of actions in the action space of the outer-loop adaptive learning. Normally, we choose the q_{min} value from a finite set consisting of up to ten allowable q_{min} levels. Therefore, the adaptive learning algorithm has negligible computation overhead. In addition, the outer-loop adaptive learning does not rely on accurate HEV modeling and only the battery SoH estimation method are needed.

TABLE I. PHEV KEY PARAMETERS.

Vehicle	**Transmission**	**ICE**
$m = 1254$ kg	$\rho_{reg} = 1.75$	peak power 41kW
$C_R = 0.009$	$\eta_{reg} = 0.98$	peak eff. 34%
$C_D = 0.335$	$\eta_{gb} = 0.98$	**EM**
$A_F = 2$ m^2	$R(k) = [13.5; 7.6;$	peak power 56kW
$r_{wh} = 0.282$ m	$5.0; 3.8; 2.8]$	peak eff. 92%
battery		
Capacity 25A·h Voltage 240V		

IV. EXPERIMENTAL RESULTS

We simulate the operation of a PHEV, the model of which is developed in the vehicle simulator ADVISOR [1]. The key parameters of the PHEV are summarized in Table I. We test our proposed policy and compare with the reinforcement learning (RL) policy [16] and the rule-based policy [5]. We use both real-world and testing driving trip profiles, which are developed and provided by different organizations and projects such as U.S. EPA (Environmental Protection Agency) and E.U. MODEM (Modeling of Emissions and Fuel Consumption in Urban Areas project).

Table II presents the simulation results of the operating cost of the PHEV during different driving trips when the proposed, the RL, and the rule-based policies are adopted. For example, as shown in Table II, the proposed policy results in 0.0028% battery capacity fading and 344.17g fuel consumption in the MODEM5713 driving trip, which correspond to \$0.76 amortized battery replacement cost and \$0.37 fuel consumption cost, and the total operating cost is \$1.13. When calculating the operating cost, we use the America average gasoline price of \$3/gal and the total battery replacement cost of \$8,000 for the PHEV. Generally, the battery replacement cost of a PHEV is in the range \$10,000∼\$12,000 [3] for battery pack with average capacity of 10kWh. We use the battery replacement cost of \$8,000 for the 6kWh battery. We use the complete cycle-decoupling method [30] to evaluate the battery capacity fading during each trip. From Table II we can observe that the proposed policy consistently achieves the lowest operating cost comparing with the RL and rule-based policies. The proposed policy achieves a maximum of 47% operating cost reduction comparing with the rule-based policy, and a maximum of 48% reduction comparing with the RL policy.

Based on Table II, we also have the following observations: (i) For a PHEV, the amortized battery replacement cost is a large portion of the total operating cost and is even higher than the fuel cost for some driving trips. (ii) The relative amortized battery replacement cost is more significant for shorter driving trip. (iii) Our proposed policy can prolong the battery life significantly besides reducing the operating cost. (iv) Although the RL policy can reduce the fuel consumption comparing with the rule-based policy, in some case the operating cost from the RL policy is even higher because the RL policy does not take into account the battery cost when optimizing the fuel consumption. (v) The amortized battery replacement cost is non-negligible when optimizing the total operating cost.

Furthermore, we also simulate an HEV (without the plug-in feature) using the Honda Insight Hybrid model from AD-

TABLE II. OPERATING COST OF THE PHEV IN DIFFERENT TRIPS USING THE PROPOSED, RL, AND RULE-BASED POLICIES.

Trip	Proposed	RL	Rule
MODEM 5713	0.0028%($0.76) +344.17g($0.37)	0.0045%($1.22) +310.56g($0.33)	0.0044%($1.18) +383.30g($0.41)
cost	=($1.13)	=($1.55)	=($1.59)
Hyzem motorway	0.0018%($0.50) +1991.9g($2.16)	0.0048%($1.28) +2001.9g($2.17)	0.0050%($1.36) +2093.6g($2.27)
cost	=($2.66)	=($3.45)	=($3.63)
FTP75	0.0027%($0.73) +311.40g($0.33)	0.0043%($1.16) +295.97g($0.32)	0.0048%($1.30) +623.73g($0.67)
cost	=($1.06)	=($1.48)	=($1.97)
US06	0.0028%($0.74) +414.17g($0.45)	0.0043%($1.17) +354.34g($0.38)	0.0036%($0.98) +321.02g($0.34)
cost	=($1.19)	=($1.55)	=($1.32)
UDDS	0.0032%($0.85) +298.48g($0.32)	0.0044%($1.19) +355.85g($0.38)	0.0048%($1.30) +630.22g($0.68)
cost	=($1.17)	=($1.57)	=($1.98)
OSCAR	0.0021%($0.57) +149.51g($0.16)	0.0043%($1.16) +222.75g($0.24)	0.0042%($1.12) +242.54g($0.26)
cost	=($0.73)	=($1.40)	=($1.38)

TABLE III. OPERATING COST OF THE HEV IN DIFFERENT TRIPS BY THE PROPOSED, RL, AND RULE-BASED POLICIES.

Trip	Proposed	RL	Rule
LA92	0.0010%($0.07) +474.83g($0.51)	0.0039%($0.26) +460.03g($0.50)	0.0067%($0.44) +568.43g($0.61)
cost	=($0.58)	=($0.76)	=($1.05)
Artemis urban	0.0015%($0.10) +110.34g($0.12)	0.0028%($0.18) +109.61g($0.11)	0.0051%($0.34) +209.20g($0.22)
cost	=($0.22)	=($0.29)	=($0.56)
Modem1	0.0009%($0.06) +143.25g($0.15)	0.0029%($0.19) +138.33g($0.15)	0.0058%($0.39) +215.48g($0.23)
cost	=($0.21)	=($0.34)	=($0.62)
Modem2	0.0012%($0.08) +221.62g($0.24)	0.0029%($0.19) +229.32g($0.24)	0.0056%($0.37) +330.26g($0.35)
cost	=($0.32)	=($0.43)	=($0.72)
Modem3	0.0015%($0.10) +66.00g($0.07)	0.0026%($0.18) +58.30g($0.06)	0.0044%($0.29) +121.21g($0.13)
cost	=($0.17)	=($0.24)	=($0.42)

VISOR. The battery pack replacement of an HEV is $2,000 [3]. Table III presents the operating cost of an HEV. We can observe that the proposed policy achieves the lowest operating cost comparing with the RL, and the rule-based policies. We also find that the amortized battery replacement cost is less significant for an HEV than for a PHEV.

V. CONCLUSIONS

This paper investigates the HEV energy management problem for the minimization of the operating cost of an HEV by using a nested learning framework. The inner loop determines the operation modes of the HEV components and is the key to minimization of the fuel usage, whereas the outer loop modulates the battery SoH degradation globally. Experimental results demonstrate a maximum of 48% operating cost reduction by the proposed HEV energy management policy.

ACKNOWLEDGMENT

This research is supported in part by a grant from the Software and Hardware Foundations program of the National Science Foundation, and by the Center for Integrated Smart Sensors funded by Science, ICT & Future Planning as Global Frontier Project (CISS-2011-0031863).

REFERENCES

[1] *ADVISOR 2003 documentation*. National Renewable Energy Lab.

[2] *High-performance battery monitor IC with coulomb counter, voltage and, temperature measurement*. Texas Instruments.

[3] *http://batteryuniversity.com/learn/article/hybrid_electric_vehicle*.

[4] E. Alpaydin. *Introduction to machine learning*. MIT press, 2004.

[5] H. Banvait and et al. A rule-based energy management strategy for plug-in hybrid electric vehicle (phev). In *ACC'09*.

[6] A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[7] B. M. Baumann and et al. Mechatronic design and control of hybrid electric vehicles. *Mechatronics, IEEE/ASME Trans*, 2000.

[8] H. Borhan and et al. Mpc-based energy management of a power-split hybrid electric vehicle. *Control Systems Technology, IEEE Trans*, 2012.

[9] C. Chan. The state of the art of electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, 2007.

[10] C.-C. Chan and et al. Electric, hybrid, and fuel-cell vehicles: Architectures and modeling. *Vehicular Technology, IEEE Trans*, 2010.

[11] S. Delprat and et al. Optimal control of a parallel powertrain: from global optimization to real time control strategy. In *Vehicular Technology Conference, 2002*.

[12] S. Ebbesen, P. Elbert, and L. Guzzella. Battery state-of-health perceptive energy management for hybrid electric vehicles. *Vehicular Technology, IEEE Transactions on*, 61(7):2893–2900, 2012.

[13] Q. Gong and et al. Trip-based optimal power management of plug-in hybrid electric vehicles. *Vehicular Technology, IEEE Trans*, 2008.

[14] D. Haifeng and et al. A new soh prediction concept for the power lithium-ion battery used on hevs. In *VPPC'09*.

[15] J.-M. Kang, I. Kolmanovsky, and J. Grizzle. Dynamic optimization of lean burn engine aftertreatment. *Journal of Dynamic Systems, Measurement, and Control*, 2001.

[16] X. Lin and et al. Reinforcement learning based power management for hybrid electric vehicles. In *ICCAD*, 2014.

[17] D. Linden and T. Reddy. Handbook of batteries, 2002.

[18] A. Millner. Modeling lithium ion battery degradation in electric vehicles. In *CITRES, 2010*.

[19] S. J. Moura and et al. A stochastic optimal control approach for power management in plug-in hybrid electric vehicles. *Control Systems Technology, IEEE Trans*, 2011.

[20] C. Musardo, G. Rizzoni, Y. Guezennec, and B. Staccia. A-ecms: An adaptive algorithm for hybrid electric vehicle energy management. *European Journal of Control*, 11(4):509–524, 2005.

[21] S. Pelletier and et al. Battery electric vehicles for goods distribution: A survey of vehicle technology, market penetration, incentives and practices. 2014.

[22] L. V. Pérez and et al. Optimization of power management in an hybrid electric vehicle using dynamic programming. *Mathematics and Computers in Simulation*, 2006.

[23] G. L. Plett. Extended kalman filtering for battery management systems of lipb-based hev battery packs: Part 1. background. *JPS*, 2004.

[24] F. R. Salmasi. Control strategies for hybrid electric vehicles: Evolution, classification, comparison, and future trends. *Vehicular Technology, IEEE Transactions on*, 2007.

[25] A. Sciarretta, D. di Domenico, P. Pognant-Gros, and G. Zito. Optimal energy management of automotive battery systems including thermal dynamics and aging. In *Optimization and Optimal Control in Automotive Systems*, pages 219–236. Springer, 2014.

[26] L. Serrao, S. Onori, A. Sciarretta, Y. Guezennec, and G. Rizzoni. Optimal energy management of hybrid electric vehicles including battery aging. In *American Control Conference (ACC), 2011*, pages 2125–2130. IEEE, 2011.

[27] S. Shao and et al. Challenges of phev penetration to the residential distribution network. In *PES'09*.

[28] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 1988.

[29] A. Vogel, D. Ramachandran, R. Gupta, and A. Raux. Improving hybrid vehicle fuel efficiency using inverse reinforcement learning. In *AAAI*, 2012.

[30] Y. Wang and et al. Minimizing state-of-health degradation in hybrid electrical energy storage systems with arbitrary source and load profiles. In *DATE2014*.