Data-Driven Reinforcement Learning–Based Real-Time Energy Management System for Plug-In Hybrid Electric Vehicles

Xuewei Qi, Guoyuan Wu, Kanok Boriboonsomsin, Matthew J. Barth,
and Jeffrey Gonder

Plug-in hybrid electric vehicles (PHEVs) show great promise in reducing transportation-related fossil fuel consumption and greenhouse gas emissions. Designing an efficient energy management system (EMS) for PHEVs to achieve better fuel economy has been an active research topic for decades. Most of the advanced systems rely either on a priori knowledge of future driving conditions to achieve the optimal but not real-time solution (e.g., using a dynamic programming strategy) or on only current driving situations to achieve a real-time but nonoptimal solution (e.g., rule-based strategy). This paper proposes a reinforcement learning–based real-time EMS for PHEVs to address the trade-off between real-time performance and optimal energy savings. The proposed model can optimize the power-split control in real time while learning the optimal decisions from historical driving cycles. A case study on a real-world commute trip shows that about a 12% fuel saving can be achieved without considering charging opportunities; further, an 8% fuel saving can be achieved when charging opportunities are considered, compared with the standard binary mode control strategy.

Reducing transportation-related energy consumption and greenhouse gas (GHG) emissions has been a common goal of public agencies and research institutes for years. In 2013, the total energy consumed by the transportation sector in the United States was as high as 24.90 quadrillion BTUs (*1*). The U.S. Environmental Protection Agency reported that nearly 27% of GHG emissions resulted from fossil fuel combustion for transportation activities in 2013 (*2*). From a vehicle perspective, innovative power train technologies, such as hybrid electric vehicles (HEVs), are very promising in improving fossil fuel efficiency and reducing exhaust emissions. Plug-in hybrid electric vehicles (PHEVs) attracted most of the attention because of their ability to also use energy off the electricity grid through charging their batteries, thereby achieving even higher overall energy efficiency (*3*).

The energy management system (EMS) is at the heart of PHEV fuel economy; its functionality is to control the power streams from the internal combustion engine (ICE) and the battery pack, on the basis of vehicle and engine operating conditions. In the past decade, a large variety of EMS implementations have been developed for PHEVs, whose control strategies may be well categorized into two major classes as shown in Figure 1: (*a*) rule-based strategies that rely on a set of simple rules without a priori knowledge of driving conditions (*4–7*); such strategies make control decisions on the basis of instant conditions only and are easily implemented, but their solutions are often far from being optimal because of the lack of consideration of variations in trip characteristics and prevailing traffic conditions; and (*b*) optimization-based strategies that are aimed at optimizing some predefined cost function according to the driving conditions and the vehicle's dynamics (*3, 8–18*). The selected cost function is usually related to fuel consumption or tailpipe emissions. According to the way the optimization is implemented, such strategies can be further divided into two groups: (*a*) offline optimization, which requires full knowledge of the entire trip to achieve the global optimal solution and (*b*) short-term prediction-based optimization, which takes into account the predicted driving conditions in the near future and achieves local optimal solutions segment-by-segment in an entire trip. However, major drawbacks of these strategies include (*a*) heavy dependence on a priori knowledge of future driving conditions and (*b*) high computation costs that are difficult to implement in real time.

As discussed above, there is a trade-off between the real-time performance and optimality in the energy management for PHEVs. Specifically, rule-based methods can easily be implemented in real time but are far from being optimal while optimization-based methods are able to achieve optimal solutions but are difficult to implement in real time. To achieve a good balance in between, reinforcement learning (RL) has recently attracted researchers' attention. Liu et al. proposed the first and only existing RL-based EMS for PHEVs; it outperforms the rule-based controller with respect to the defined reward function but is worse in regard to fuel consumption without considering charging opportunity in the model (*19*).

In this study, a novel model-free RL-based real-time EMS of PHEVs is proposed and evaluated; it is capable of simultaneously controlling and learning the optimal power-split operations in real time. The proposed model is theoretically derived from dynamic programming (DP) formulations and compared with DP in the computational complexity perspective. Three major features distinguish it from existing methods: (*a*) the proposed model can be implemented in real time without any prediction efforts since the control decisions are made only on the current system state; the control decisions are also considered for the entire trip information

X. Qi and M. J. Barth, Department of Electrical and Computer Engineering, G. Wu and K. Boriboonsomsin, CE-CERT, University of California, Riverside, 1084 Columbia Avenue, Riverside, CA 92507. J. Gonder, National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401. Corresponding author: X. Qi, xqi001@ucr.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2572, Transportation Research Board, Washington, D.C., 2016, pp. 1–8.
DOI: 10.3141/2572-01

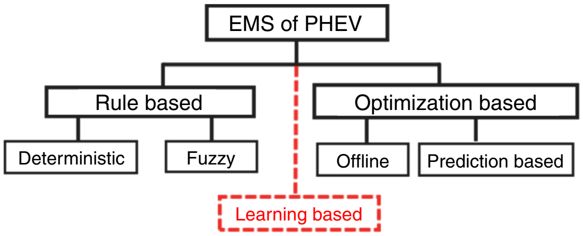


FIGURE 1 Taxonomy of current EMS.

by learning the optimal or near-optimal control decisions from historical driving behavior; therefore, a good balance between real-time performance and energy saving optimality is achieved; (*b*) the proposed model is data driven and does not need any PHEV model information once it is well trained since all the decision variables can be observed and are not calculated with the use of any vehicle power train models (these details are described in the following sections); and (*c*) compared with existing RL-based EMS implementations, the proposed strategy considers charging opportunities along the way (a key distinguishing feature of PHEVs as compared with HEVs) (*19*). This proposed method represents a new class of models that could be a good supplement to the current method taxonomy as shown in Figure 1.

BACKGROUND

PHEV Power Train and Optimal Energy Management Formulation

There are three types of PHEV power train architectures: (*a*) series, (*b*) parallel, and (*c*) power split (series parallel) (*1*). In this study the focus is on the power-split architecture. The decision making on the power-split ratio between internal combustion engine (ICE) and battery pack is called the power-split control problem (*3*). Mathematically, the optimal energy management (i.e., power-split control) for PHEVs can be defined as a nonlinear constrained optimization problem (*3*). In this study, the ICE power supply is discretized into different levels, and the optimal PHEV power-split control problem therefore can be formulated as follows:

$$\min \sum_{t=1}^{M} \sum_{i=1}^{N} x(t,i) \frac{P_i^{\text{eng}}}{\eta_i^{\text{eng}}} \tag{1}$$

subject to

$$\sum_{t=1}^{j} f\left(P_t - \sum_{i=1}^{N} x(t,i) P_i^{\text{eng}} \right) \leq C \qquad \forall j = 1, \ldots, T \tag{2}$$

$$\sum_{i=1}^{N} x(t,i) = 1 \qquad \forall t \tag{3}$$

$$x(t,i) = \{0,1\} \qquad \forall t, 1 \tag{4}$$

where

M = time span of entire trip,
N = discretized power level value for engine,
t = time step index,
i = ICE power level index,
C = gap of battery pack's state of charge (SOC) between initial and minimum,
P_i^{eng} = ith discretized level for engine power,
η_i^{eng} = associated engine efficiency,
P_t = driving demand power at time step t, and
j = any time step between first and last time step.

The objective of the energy management problem is to find the optimal action (i.e., selection of the optimal ICE power level) for each time step to achieve the best fuel efficiency along the entire trip.

Dynamic Programming

The optimization problem represented by Equations 1 to 4 can be solved by DP since it satisfies Bellman's Principle of Optimality (*20*). Let $s \in S$ be the state vector of the system and $a \in A$ the decision variable. The optimization problem can be converted into the following single equation given the initial state s_0 and the decisions a_t for each time step t:

$$\min_{a_t \in A} E\left\{ \sum_{t=0}^{T-1} \beta^t g(s_t, s_{t+1}) \middle| s_0 = s \right\} \tag{5}$$

where β is discount factor, $\beta \in (0, 1)$, and s_t is the current system state, and it can be solved by recursively calculating

$$J(s_t) = \min_{a_t \in A} E\left\{ \sum_{t=0}^{T-1} g(s_t, s_{t+1}) + \beta J(s_{t+1}) \middle| s_t = s \right\}$$
$$\text{for } t = T-1, T-2, \ldots, 0 \tag{6}$$

where

T = time duration,
$g(.)$ = one-step cost function, and
$J(s)$ = true value function associated with state s.

Equation 6 is also often noted as Bellman's equation. In the case of PHEV energy management, s_t can be defined as a combination of vehicle states, such as the current SOC level and the remaining time to the destination, which is discussed in the following sections. a_t can be defined as the ICE power supply at each time step.

It is well known that the high computation cost of Equation 6 is always the barrier that impedes its real-world application although it is a very simple and descriptive definition. It could be computationally intractable even for a small-scale problem (in regard to state space and time span). The major reason is that the algorithm has to loop over the entire state space to evaluate the optimal decision for every single step. Another obvious drawback in the real-world application of DP is that it requires the availability of the full information of the optimization problem. In the present case, it means the power demand along the entire trip should be known before the trip, which is always impossible in practice.

Approximate Dynamic Programming and Reinforcement Learning

To address the above issues, approximate dynamic programming (ADP) has been proposed (*21*). The major contribution of ADP is that it significantly reduces the state space by introducing an approximate value function $\hat{J}(s_t, p_t)$, where p_t is a parameter vector. With the replacement of this approximate value function, Equation 6 can be reformulated as

$$\hat{J}(s_t) = \min_{a_t \in A} E\left\{\sum_{t=0}^{T-1} g(s_t, s_{t+1}) + \beta\hat{J}(s_{t+1}, p_t)\right\} \quad \text{for } t = 0, 1, \ldots, T-1 \qquad (7)$$

Now the optimal decision can be calculated at each time step t by

$$a_t = \arg\min_{a_t \in A} E\left\{\sum_{t=0}^{T-1} g(s_t, s_{t+1}) + \beta\hat{J}(s_{t+1}, p_t)\right\} \qquad (8)$$

The calculation of Equation 8 now relies only on the current system state s_t, which substantially reduces the computational requirement of Equation 6 to polynomial time in regard to the number of state variables, rather than being exponential to the size of state space (*22*). In addition, the value iteration that solves the DP problem becomes forward into time, rather than being backward in Equation 6. In the case of PHEV energy management, this fact is actually a bonus since the predicted state (e.g., power demand) at the end of the time horizon is much less reliable compared with that at the beginning of the time horizon.

In principle, the value function approximation can be learned by tuning and updating the parameter vector p_t on the addition of each observation on state transitions (*22*). RL is an effective tool for that purpose. The specific learning technique used in this study is temporal-difference learning, which was originally proposed by Sutton and Barto to approximate the long-term future cost as a function of current states (*23*). The details on the implementation of the algorithm are presented in the following sections.

REINFORCEMENT LEARNING–BASED EMS

In this study, a temporal-difference-learning strategy is adopted for the RL problem. An action-value function, $Q(s, a)$, is defined as the expected total reward for the future receipt starting from that state. This function is to estimate how good it is to perform a given action in a given state in regard to the expected return. More specifically, $Q^\pi(s, a)$ is defined as the value of taking action a in state s under a control policy π (i.e., a map that maps the optimal action to a system state), which is also the expected return starting from s, taking the action a, and thereafter following policy π:

$$Q^\pi(s, a) = E_\pi\left\{\sum_{k=1}^{\infty} \gamma^k * r(s_{t+k}, a_{t+k}) | s_t = s, a_t = a\right\} \qquad (9)$$

where

s_t = state at time step t,
γ = discount factor in (0, 1) to guarantee convergence (*26*),
k = time steps elapsed after time step t, and
$r(s_{t+k}, a_{t+k})$ = immediate reward based on state s and action a at given time step $(t + k)$.

The ultimate goal of RL is to identify the optimal control policy that maximizes the above action-value function for all state-action pairs. Compared with the formulations defined by Equations 6 and 7, the policy π is the ultimate decision for each time step along the entire time horizon. The reward function $r(s_{t+k}, a_{t+k})$ here is $g(.)$ in Equation 6. The action-value function [i.e., $Q(s, a)$] is actually an instantiation of the approximate value function $\hat{J}(s_t)$. So, it is not difficult to understand that the DP formulas are the basis for an RL problem.

Conceptually, an RL system consists of two basic components: a learning agent and an environment. The learning agent interacts continuously with the environment in the following manner: at each time step, the learning agent receives an observation on the environment state. The learning agent then chooses an action that is subsequently input into the environment. The environment then moves to a new state as a result of the action, and the reward associated with the transition is calculated and fed back to the learning agent. Along with each state transition, the agent receives an immediate reward, and these rewards are used to form a control policy that maps the current state to the best control action on that state. At each time step, the agent makes the decision on the basis of its control policy. Ultimately, the optimal policy can guide the learning agent to take the best series of actions to maximize the cumulated reward over time that can be learned after sufficient training. A graphical illustration of the learning system is given in Figure 2. The definition of the environmental states, actions, and reward are provided next.

Action and Environmental States

In this study, the discretized ICE power supply level (i.e., P_i^{eng} in Equation 1) is defined as the only action the learning agent can take. The environment states include any other system parameters that could influence the decision of engine power supply. Here a definition is given for a five-dimensional state space for the environment, including the vehicle velocity (v_{veh}), road grade (g_{road}), percentage of remaining time to destination (t_{togo}), battery pack's state-of-charge (b_{soc}), and available charging gain (c_g) of the selected charging station:

$$S = \left\{ s = \left[v_{\text{veh}}, g_{\text{road}}, t_{\text{togo}}, b_{\text{soc}}, c_g\right]^T \,\middle|\, \begin{array}{l} v_{\text{veh}} \in V_{\text{veh}}, g_{\text{road}} \in G_{\text{road}}, \\ t_{\text{togo}} \in T_{\text{togo}}, b_{\text{soc}} \in B_{\text{soc}}, c_g \in C_g \end{array} \right\}$$

where V_{veh} is the set of discretized vehicle speed levels and G_{road} is the set of discretized road grade levels. The minimum and maximum value of vehicle velocity and road grade can be estimated from the historical data of commuting trips, which will be elaborated in the

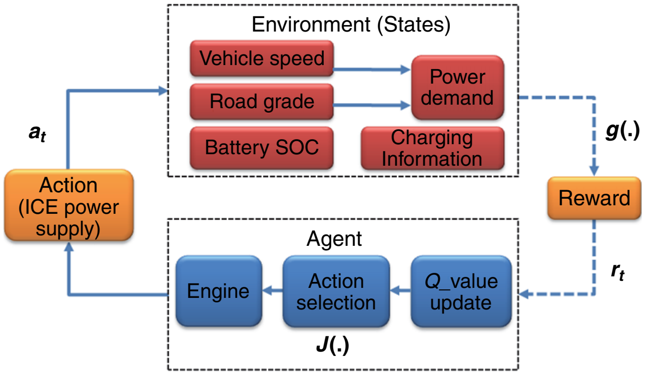


FIGURE 2 Graphical illustration of RL system.

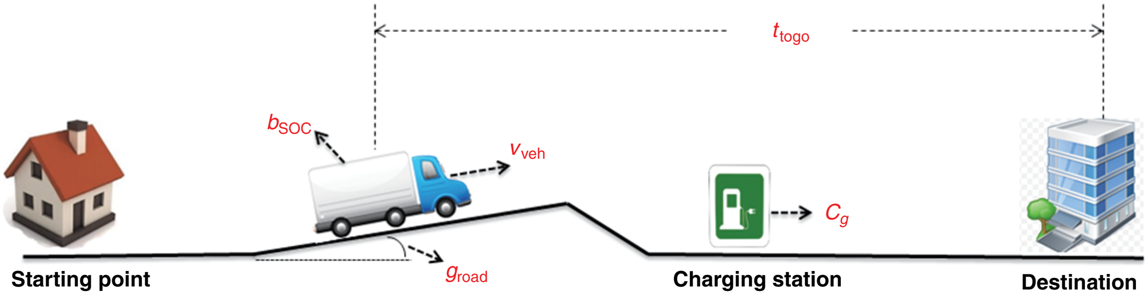


FIGURE 3 Illustration of environment states along trip.

following section. B_{soc} is the set of battery SOC levels between the lower bound (e.g., 20%) and upper bound (e.g., 80%); T_{togo} is the percentage (10%~90%) of remaining time in the entire trip duration, which is calculated on the basis of the remaining distance to the destination. C_g is the set of discretized charging gain percentages (e.g., 30% and 60%) of the selected charger. This charging gain represents the availability of the charger, which may be a function of the charging time and charging rate and is assumed to be known beforehand. All of the states can be measured and updated in real time as the vehicle is running. Figure 3 shows all the real-time environmental states.

Reward Initialization with Optimal Results from Simulation

The definition of reward depends on the control objective, which is to minimize the fuel cost while satisfying the power demand requirement. Hence, the reciprocal of the resultant ICE power consumption at each time step is defined as the immediate reward. A penalty term is also included to penalize the situation in which the SOC is beyond the predefined SOC boundaries. Immediate reward is calculated by the following equations:

$$r_{ss'}^{a}=\begin{cases}\dfrac{1}{P_{\text{ICE}}} & \text{if } P_{\text{ICE}}\neq 0 \text{ and } 0.2\leq \text{SOC}\leq 0.8\\[2ex] \dfrac{1}{P_{\text{ICE}}+P} & \text{if } P_{\text{ICE}}\neq 0 \text{ and } \left(\text{SOC}\leq 0.2 \text{ or } \text{SOC}\geq 0.8\right)\\[2ex] \dfrac{2}{\min_{P_{\text{ICE}}}} & \text{if } P_{\text{ICE}}=0 \text{ and } 0.2\leq \text{SOC}\leq 0.8\\[2ex] \dfrac{1}{2*P} & \text{if } P_{\text{ICE}}=0 \text{ and } \left(\text{SOC}\leq 0.2 \text{ or } \text{SOC}\geq 0.8\right)\end{cases}\tag{10}$$

where

$r_{ss'}^{a}$ = immediate reward when state changes from s to s' by taking action a,
P_{ICE} = ICE power supply,
P = penalty value and is set as maximum power supply from ICE in this study, and
min_P_{ICE} = minimum nonzero value of ICE power supply.

This definition guarantees that the minimum ICE power supply (action) that satisfies the power demand as well as the SOC constraints can have the largest numerical reward. A good initialization of reward is also critical for the quick convergence of the proposed algorithm. In this case, the optimal or near-optimal results of typical trips obtained from simulation are used as the initial seeds. These optimal or near-optimal results are deemed as the control decisions made by good drivers from historical driving. To obtain a large number of such good results for algorithm training, an evolutionary algorithm is adopted for the offline full-trip optimization since an evolutionary algorithm can provide multiple solutions for one single run. The quality of these solutions is different and, as a result, different levels of driving proficiency in the real-world situation may well be represented.

Q-Value Update and Action Selection

In the algorithm, a Q value, denoted by $Q(s, a)$, is associated with each possible state-action pair (s, a). Hence there is a Q table that is kept updated during the learning process and can be interpreted as the optimal control policy that the learning agent is trying to learn. At each time step, the action is selected by using this table after it is updated. The details of the RL-based PHEV EMS algorithmic process are given in the following pseudocode.

Inputs. Initialization 6-D $Q(s, a)$ table; discount factor $\gamma = 0.5$; learning rate $\alpha = 0.5$; exploration probability $\varepsilon \in (0, 1)$; vehicle power demand profile P_d (N time steps).

Outputs. $Q(s, a)$ array; control decisions P_d (T time steps).

```
Initialize $Q(s, a)$ arbitrarily.
For each time step $t = 1{:}T$
    observe current $s_t$ ($v_{\text{veh}}$, $g_{\text{road}}$, $t_{\text{togo}}$, $b_{\text{soc}}$, $C_g$).
    Choose action $a_t$ for current state $s_t$:
        temp = random (0, 1);
        if temp <= $\varepsilon$
            $a_t = \arg\max_{a\in A}\{Q(s_t, a)\}$
        else
            $a_t$ = randomly choose an action;
        end.
    Take action $a_t$, observe next state $s_{t+1}$ ($P_{t+1}$, $\text{SOC}_{t+1}$).
    If $\text{SOC}_{t+1} < 0.2$
        switch into charging–sustaining mode;
        give big penalty to $r_t$ according to Equation 10.
    else
        calculate reward $r_t$ according to Equation 10
    end.
    Update $Q(s_t, a_t)$ with following value:
    $Q(s_t, a_t) + \alpha\{r_t + \gamma * \max_{a_{t+1}}\{Q(s_{t+1}, a_{t+1})\} - Q(s_t, a_t)\}$
end.
```

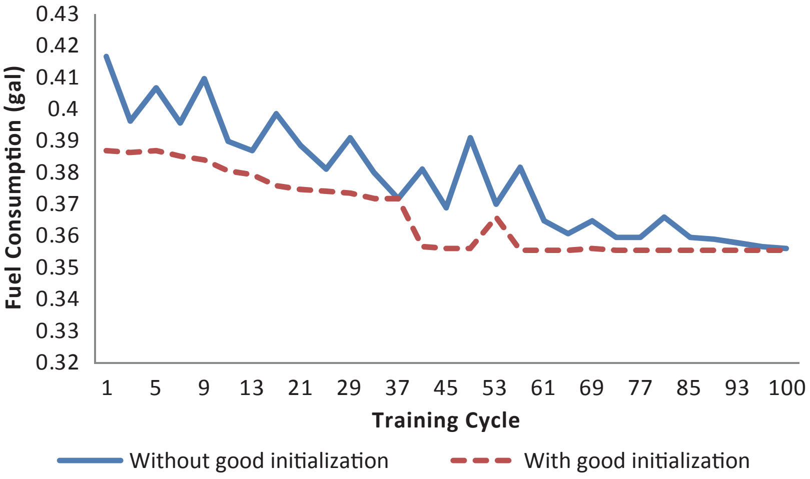


FIGURE 4 Convergence analysis (ε = 0.7, γ = 0.5, and α = 0.5).

CASE STUDY

The proposed model is then evaluated with real-world data in two scenarios: one without charging opportunities and the other with charging opportunities.

Data Description

To obtain a series of real trip data (second-by-second velocity trajectories), the trajectory synthesis technique proposed in previous work (*3*) is applied to the inductive loop detector data archived in the California Freeway Performance Measurement System (*24*). The trajectory synthesis is a two-step process: (*a*) estimating average velocity by applying a two-dimensional interpolation method to real-world traffic data (e.g., volumes and occupancy) collected from inductive loop detectors and (*b*) generating random velocity disturbance based on representative driving cycles from the Motor Vehicle Emission Simulator database. Real traffic data were collected at the I-210 freeway segment between I-605 and Day Creek Boulevard in Southern California; data were collected on traffic starting at 8:00 a.m. in the morning (westbound) and returning at 4:00 p.m. in the afternoon every weekday during the period between January 9, 2012, and January 17, 2012. Twelve trips (including eastbound and westbound) were generated in total. The road grade information was also synchronized with the trip data to estimate the second-by-second power demands. For more detailed information on the trajectory synthesis and power demand profile generation, see Wu et al. (*3*).

Model Without Charging Opportunity: Trip Level

To validate the proposed strategy, the model without charging opportunity being considered is first trained and tested with trips for which there is no charging opportunity in the trip. Data for multiple westbound trips described in Wu et al. are used for training (*3*). Although it has been proved that *Q* learning is guaranteed to converge mathematically, an experimental analysis of convergence is conducted in this study (*19*). In the experiment, the trip data for the first 6 days are concatenated one by one to form a single training cycle. The proposed model is trained with repeated training cycles. At the end of each training cycle, the trained model is tested with the seventh day trip; the fuel consumption is recorded in Figure 4. In addition, the training of the model with good initialization using the simulated optimal or near-optimal solution and the training of the model without good initialization are compared. As can be seen in the figure, there is a clear convergence in fuel consumption for the two cases. However, the initialization with simulated optimal or near-optimal solutions helps achieve a faster convergence.

As previously described, the selected state space is five-dimensional and the action space has one dimension. Therefore the $Q(s, a)$ table is six-dimensional. Figure 5 shows the 4-D slice diagram of the learned

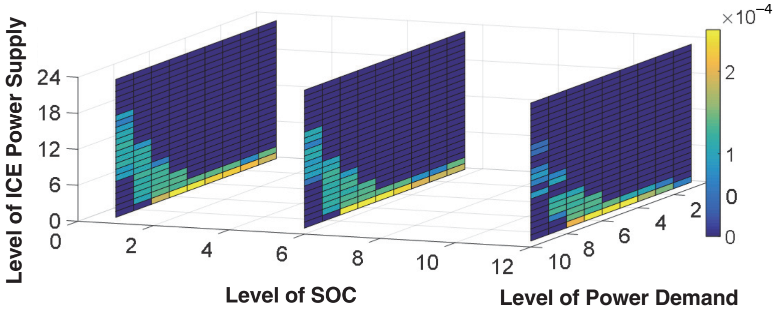


FIGURE 5 Four-dimensional slice diagram of learned *Q* table.

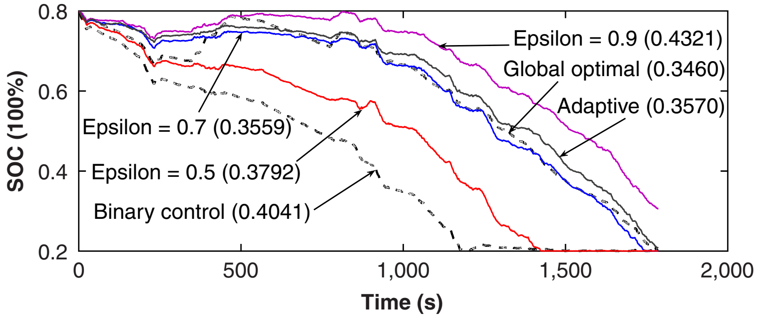


FIGURE 6 Fuel consumption (values in parentheses are gallons) and SOC curves by different exploration probabilities.

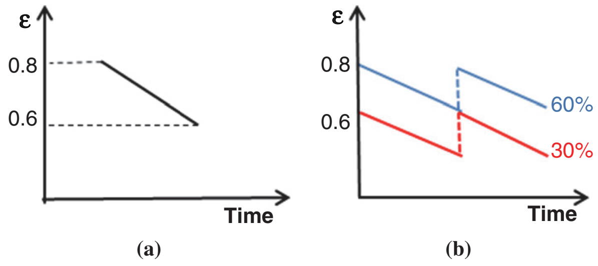
$Q(s, a)$ table in which different color grids represent different numerical reward values (e.g., blue means lower values) and three slices on the ICE power supply and power demand space are given at three SOC levels: 1, 6, and 12 (i.e., 20%, 50%, and 80%). The road grade and vehicle speed are implicitly aggregated into power demand. The dimension of remaining time is not indicated in the figure. As can be observed in each slice, when the power demand is not so high (e.g., below Level 5), Action Level 1 or 2 is usually the most appropriate because the least ICE power is consumed. When the power demand becomes higher, the range of the feasible action levels widens also. In such cases, lower levels of ICE power supply may not be enough to satisfy the power demand and the resultant SOC level could be lower than 0.2, resulting in a penalty defined in Equation 10. When the SOC level is high, it is less likely that the higher ICE power supply level would be chosen to satisfy the same power demand. The reason is that when the vehicle battery SOC is high, the ICE power is not likely to be used aggressively.

As discussed in the previous sections, an exploration–exploitation strategy is adopted for the action selection process to avoid premature convergence. The action with the biggest Q value has a probability of 1-ε to be selected. Hence the value of ε may significantly affect the performance of the proposed method. To evaluate such effects, a sensitivity analysis of ε is carried out and is illustrated in Figure 6. It can be observed that the fuel consumption and the resultant SOC curves are very close to those of the binary mode control if the value of ε is small. A possible explanation is that a small ε value indicates a large probability to select the most aggressive action with the biggest Q value (or the lowest levels of ICE power supply). Therefore, the battery power is consumed drastically as it is with the binary mode control. However, if the value of ε is too large (e.g., >0.8), the battery power is used too conservatively so that the final SOC is far away from the lower bound, resulting in much greater fuel consumption. It is found that the best value of ε in this study is about 0.7, which ensures that the SOC curve is quite close to the global optimal solution obtained by the offline DP strategy. With this best ε value, the fuel consumption is 0.3559 gal, which is 11.9% less than that of the binary mode control and only 2.86% more than that of the DP strategy as shown in Figure 6. The implication is that an adaptive strategy for determining the exploration rate along the trip could be useful. Figure 7*a* shows a linearly decreasing control of ε along the trip. A smaller ε is preferred at the later stage of the trip because SOC is low and the battery power should be consumed more conservatively. With this adaptive strategy for ε, the proposed mode could also achieve a good solution with a 0.3570-gal fuel consumption, which is 11.7% less than by binary control shown in Figure 6.

Model with Charging Opportunity: Tour Level

The most distinctive characteristic of PHEVs from HEVs is that PHEV can be externally charged whenever a charging opportunity is available. To further evaluate the effects resulting from charging availability, this information is included in the proposed model as a decision variable. For simplicity, the charging opportunity is quantified by the gain in the battery's SOC, which may be a function of available charging time and charging rate. Data for a typical tour are constructed by combining a round-trip between the origin and destination (*3*). It is assumed that there is a charger in the workplace (westmost point on the map) and the available charging gain has only two levels: 30% and 60%. In this case, a corresponding adaptive strategy of ε is also used as shown in Figure 7*b.* The rationale behind this adaptive strategy is that battery power should be used less conservatively (i.e., higher ε value) after it has been charged, when C_g is higher, or in both cases.

The obtained optimal results are shown in Figures 8 and 9. As can be seen in the two figures, the resultant SOC curves are much closer to the global optimal solutions obtained by DP than by binary control. To obtain the statistical significance of the performance, the proposed model is tested with 30 trips by randomly combining two trips and assuming a charging station in between with a random C_g (randomly choose from 30% and 60%). With binary control taken as the baseline, the reduced fuel consumption is given in Figure 10. As can be seen in the figure, the RL model achieves an average of 7.9% fuel savings. It seems that having more information results in lower fuel savings, which is a counterintuitive result. The reason is that the inclusion of additional information or state variable to the model, exponentially increases the search space of the problem, which thereby increases the difficulty of learning the optimal solution. And also more uncertainty is introduced into the learning process



(a) (b)

FIGURE 7 Linear adaptive control of ε: (*a*) without charging opportunity and (*b*) with charging opportunity.

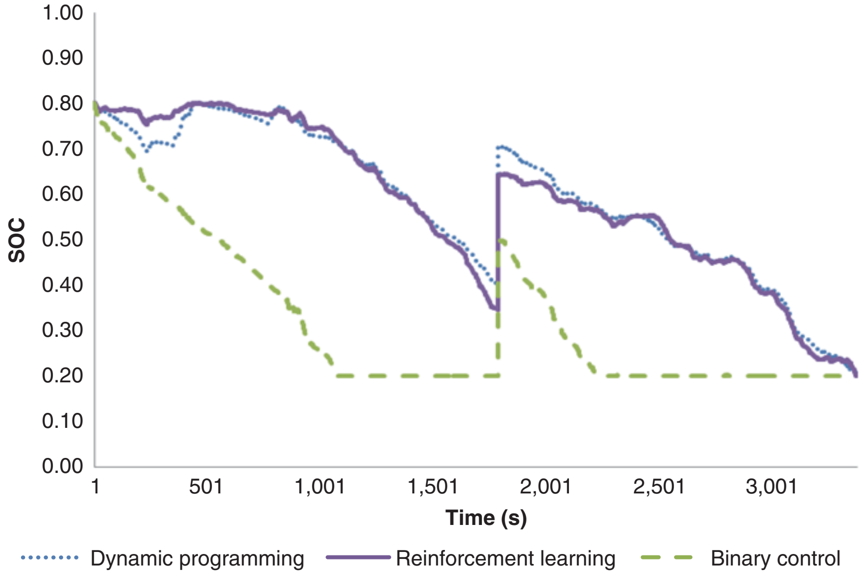


FIGURE 8 Optimal results when available charging gain is 0.3 (C_g = 0.3).

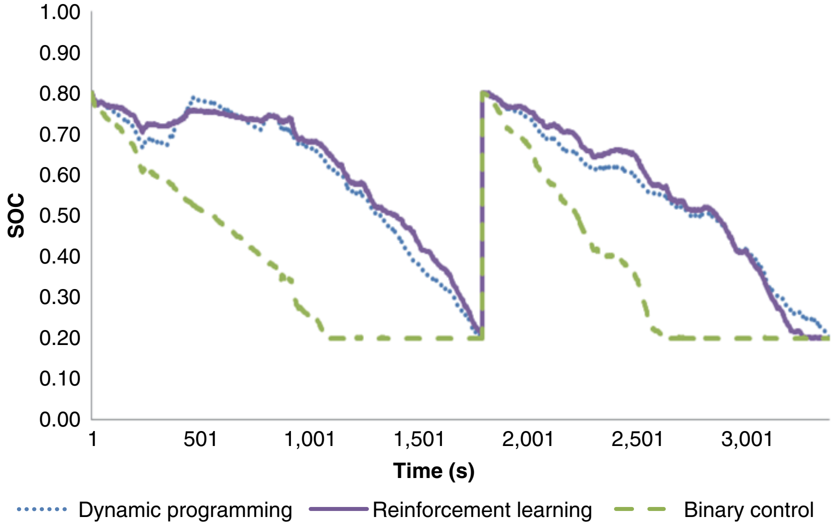


FIGURE 9 Optimal results when available charging gain is 0.6 (C_g = 0.6).

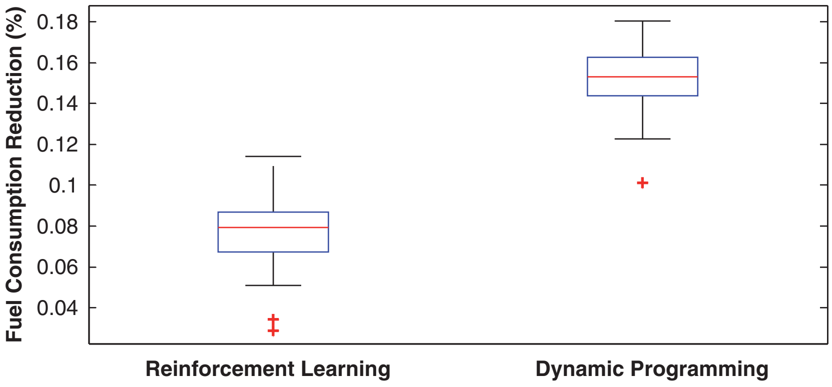


FIGURE 10 Fuel consumption reduction compared with binary control.

because of the errors in the added information, which degrades the quality of the best solution the model can achieve.

CONCLUSIONS AND FUTURE WORK

This paper proposes a data-driven reinforcement learning–based real-time energy management system for PHEVs that is capable of simultaneously controlling and learning the optimal power-split operation. The proposed EMS model is tested with trip data (i.e., multiple speed profiles) synthesized from real-world traffic measurements. Numerical analyses show that a near-optimal solution can be obtained in real time when the model is well trained with historical driving cycles. For the study cases, the proposed EMS model can achieve better fuel economy than the binary mode strategy by about 12% and 8% at the trip level and tour level (with charging opportunity), respectively. The possible topics for future work are (*a*) propose a self-adaptive tuning strategy for exploration–exploitation (ε); (*b*) test the proposed model with more real-world trip data, which could include other environmental states, such as the position of charging stations; and (*c*) conduct a robustness analysis to evaluate the performance of the proposed EMS model when there is error present in the measurement of environment states.

ACKNOWLEDGMENT

This work was supported in part by the U.S. Department of Transportation.

REFERENCES

1. Bureau of Transportation Statistics (BTS). http://www.bts.gov/publications/national_transportation_statistic.
2. *DRAFT Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2013.* Final report. U.S. Environmental Protection Agency, Feb. 2015.
3. Wu, G., K. Boriboonsomsin, and M. Barth. Development and Evaluation of an Intelligent Energy-Management Strategy for Plug-In Hybrid Electric Vehicles. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 15, No. 3, June 2014, pp. 1091–1100.
4. Tribioli, L., M. Barbielri, R. Capata, E. Sciubba, E. Jannelli, and G. Bella. A Real Time Energy Management Strategy for Plug-In Hybrid Electric Vehicles Based on Optimal Control Theory. *Energy Procedia,* Vol. 45, 2014, pp. 949–958.
5. Denis, N., M.R. Dubois, and A. Desrochers. Fuzzy-Based Blended Control for the Energy Management of a Parallel Plug-In Hybrid Electric Vehicle. *Intelligent Transport Systems,* Vol. 9, No. 1, 2015, pp. 30–37.
6. Wang, X., H. He, F. Sun, X. Sun, and H. Tang. Comparative Study on Different Energy Management Strategies for Plug-In Hybrid Electric Vehicles. *Energies,* Vol. 6, 2013, pp. 5656–5675.
7. Wu, J. Fuzzy Energy Management Strategy for Plug-In HEV Based on Driving Cycle Modeling. *2014 33rd Chinese Control Conference,* July 28–30, 2014, pp. 4472–4476.
8. Tribioli, L., and S. Onori. Analysis of Energy Management Strategies in Plug-In Hybrid Electric Vehicles: Application to the GM Chevrolet Volt. *American Control Conference,* June 17–19, 2013, pp. 5966–5971.
9. Yu, H., M. Kuang, and R. McGee. Trip-Oriented Energy Management Control Strategy for Plug-In Hybrid Electric Vehicle*s. IEEE Transactions on Control Systems Technology,* Vol. 22, No. 4, July 2014, pp. 1323–1336.
10. Gong, Q., Y. Li, and Z.-R. Peng. Trip Based Optimal Power Management of Plug-In Hybrid Electric Vehicles Using Gas-Kinetic Traffic Flow Model. *American Control Conference,* June 11–13, 2008, pp. 3225–3230.
11. Feng, T., L. Yang, Q. Gu, Y. Hu, T. Yan, and B. Yan. A Supervisory Control Strategy for PHEVs Based on Energy Demand Prediction and Route Preview. *IEEE Transactions on Vehicular Technology,* Vol. 64, No. 5, May 2015, pp. 1691–1700.
12. Larsson, V., L. Johannesson Mårdh, B. Egardt, and S. Karlsson. Commuter Route Optimized Energy Management of Hybrid Electric Vehicles. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 15, No. 3, June 2014, pp. 1145–1154.
13. Qi, X., G. Wu, K. Boriboonsomsin, and M.J. Barth. An On-Line Energy Management Strategy for Plug-In Hybrid Electric Vehicles Using an Estimation Distribution Algorithm. *2014 IEEE 17th International Conference on Intelligent Transportation Systems,* Oct. 8–11, 2014, pp. 2480–2485.
14. O'Keefe, M.P., and T. Markel. *Dynamic Programming Applied to Investigate Energy Management Strategies for a Plug-In HEV.* Report NREL/CP-540-40376. National Renewable Energy Laboratory, Golden, Colo., 2006.
15. Chen, Z., C.C. Mi, R. Xiong, J. Xu, and C. You. Energy Management of a Power-Split Plug-In Hybrid Electric Vehicle Based on Genetic Algorithm and Quadratic Programming. *Journal of Power Sources,* Vol. 248, Feb. 15, 2014, pp. 416–426.
16. Lin, X., H. Banvait, S. Anwar, and Y. Chen. Optimal Energy Management for a Plug-In Hybrid Electric Vehicle: Real-Time Controller. *American Control Conference,* June 30–July 2, 2010, pp. 5037–5042.
17. Hou, C., L. Xu, H. Wang, Mi. Ouyang, and H. Peng. Energy Management of Plug-In Hybrid Electric Vehicles with Unknown Trip Length. *Journal of the Franklin Institute,* Vol. 352, No. 2, Feb. 2015, pp. 500–518.
18. Qi, X., G. Wu, K. Boriboonsomsin, and M.J. Barth. Evolutionary Algorithm-Based On-Line PHEV Energy Management System with Self-Adaptive SOC Control. *Intelligent Vehicles Symposium,* June 28, 2015–July 1, 2015, pp. 425–430.
19. Liu, J., and H. Peng. Modeling and Control of a Power-Split Hybrid Vehicle. *IEEE Transactions on Control Systems Technology,* Vol. 16, No. 6, 2008, pp. 1242–1251.
20. Bellman, R.E. *Dynamic Programming.* Princeton University Press, Princeton, N.J. Republished 2003: Dover, N.J.
21. Powell, W.B. *Approximate Dynamic Programming: Solving the Curses of Dimensionality,* 2nd ed., John Wiley & Sons, Inc., Hoboken, N.J., 2011.
22. Cai, C., C.K. Wong, and B.G. Heydecker. Adaptive Traffic Signal Control Using Approximate Dynamic Programming. *Transportation Research Part C: Emerging Technologies,* Vol. 17, No. 5, Oct. 2009, pp. 456–474.
23. Sutton, R.S., and A.G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, Mass., 1998.
24. California Performance Measurement System. http://pems.dot.ca.gov/. Accessed July 7, 2015.

The Standing Committee on Alternative Transportation Fuels and Technologies peer-reviewed this paper.