Hindawi
Computational Intelligence and Neuroscience
Volume 2023, Article ID 8288527, 9 pages
https://doi.org/10.1155/2023/8288527

Research Article

Adaptive Optimal Control of Hybrid Electric Vehicle Power Battery via Policy Learning

Qinglin Zhu[1] **Huanli Sun,**[2] **Ziliang Zhao**[1] **Yixin Liu,**[2] **and Jun Zhao**[3]

[1] *College of Transportation, Shandong University of Science and Technology, Qingdao 266590, China*
[2] *China FAW Group Corporation, Changchun 130011, China*
[3] *College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao 266590, China*

Correspondence should be addressed to Ziliang Zhao; zhaoziliang1@sdust.edu.cn

Received 27 October 2022; Revised 21 February 2023; Accepted 9 March 2023; Published 29 May 2023

Academic Editor: Radu-Emil Precup

Copyright © 2023 Qinglin Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An online policy learning algorithm is used to solve the optimal control problem of the power battery state of charge (SOC) observer for the first time. The design of adaptive neural network (NN) optimal control is studied for the nonlinear power battery system based on a second-order (RC) equivalent circuit model. First, the unknown uncertainties of the system are approximated by NN, and a time-varying gain nonlinear state observer is designed to address the problem that the resistance capacitance voltage and SOC of the battery cannot be measured. Then, to realize the optimal control, a policy learning-based online algorithm is designed, where only the critic NN is required and the actor NN widely used in most design of the optimal control methods is removed. Finally, the effectiveness of the optimal control theory is verified by simulation.

1. Introduction

Nowadays, electric vehicles are developing at a high speed [1]. The power battery provides the required high power for vehicle start stop, acceleration and deceleration, and other instabilities and greatly improves the service life of fuel cells by controlling the charging and discharging power of the power battery [1, 2]. As an important energy storage part of fuel-cell hybrid vehicles, it has far-reaching significance for the research of power cells. The state of charge (SOC) in the battery is one of the important parameters of the battery management system (BMS), but SOC cannot be directly measured by the on-board sensors. Therefore, SOC estimation is a very important problem in the theory and application. Moreover, the power battery is a highly complex nonlinear system in its working state, which greatly increases the difficulty of estimation [3].

In order to meet the requirements of accurate, fast, and real-time estimation of power battery SOC under different conditions, scholars have carried out a lot of advanced achievements. In [4], the authors proposed an observer-based unilateral Lipschitz conditional nonlinear system control method for a class of nonlinear systems with time-varying parameter uncertainties and norm bounded disturbances. For the state-space equation of the equivalent circuit model, a power battery SOC estimation method based on nonlinear observer is proposed in [5]. The authors in [6] introduced the second-order resistance capacitance (RC) model of the battery pack. Under the unilateral Lipschitz condition, a nonlinear observer based on the H∞ method is designed, but whether the optimal performance of the observer can be guaranteed remains to be verified. For the problem of optimal control design of the observers, the authors proposed an adaptive neural network backstepping recursive optimal control method for nonlinear strict feedback systems with state constraints [7]. The neural network (NN) state identification is used to approximate the unknown nonlinear dynamics, and under the actor-critic structure, the virtual and actual optimal controllers are constructed through the backstepping recursive control algorithm. Because actor-critic structure-based adaptive laws are generated on the basis of the square of Behrman residual error obtained by the gradient descent method, these methods are too complex and difficult to implement. In this regard, the authors in [8] proposed an optimal control method based on reinforcement learning (RL) for a class of nonlinear strict feedback systems with unknown dynamic functions. This method eliminates the persistent excitation assumption necessary for most RL-based adaptive optimal control. On this basis, the adaptive NN output-feedback optimal control problem for a class of strict feedback nonlinear systems with unknown internal dynamics, input saturation, and state constraints is studied in [9]. In [10, 11], the authors proposed the novel optimal control algorithm based on advanced AI techniques, which further promotes the development of the optimal control theory.

Inspired by the abovementioned research results, a nonlinear observer with time-varying gain is designed in this paper. Based on the unilateral Lipschitz condition, the nonlinear dynamic problem contained in the system output is solved. The internal unknown dynamic function is approximated by NN to estimate the SOC and the resistance capacitance voltage of the dynamic battery in the power system. Then, based on estimated system states, we develop a policy learning-based optimal control and the estimated weight error is convergence to zero. Finally, the simulation results show the effectiveness of the proposed method.

The innovations of this paper are summarized as follows:

(1) The optimal control method based on critic NN is used to solve the optimal control problem of the power battery SOC observer for the first time.

(2) Only one critic NN is used to ensure the convergence of the NN weights; thus, the actor NN widely used in most design of optimal control methods [12–14] is removed.

(3) Unlike the existing optimal control with known state, the battery state in this paper is unknown. This leads to a complex optimal control problem.

2. System Modeling

In this paper, we consider the second-order RC equivalent circuit model as shown in Figure 1 [15], where U_{oc} is the open-circuit voltage (OCV) respected to SOC, I_T represents the current, U_T denotes the terminal voltage, R_0 is the ohmic resistance, R_1 and R_2 are the electrochemical polarization resistance and the concentration polarization resistance, respectively, and C_1 and C_2 are the capacitances. U_1 and U_2 show the voltage of the electrochemical capacitor C_1 and concentration polarization capacitor C_2, respectively.

Then, based on the Kirchhoff voltage laws, the state equation of Figure 1 can be given as

$$\begin{cases} \dot{U}_1 = -\dfrac{1}{R_1C_1}U_1 + \dfrac{1}{C_1}I_T, \\[2ex] \dot{U}_2 = -\dfrac{1}{R_2C_2}U_2 + \dfrac{1}{C_2}I_T, \\[2ex] S\dot{O}C = -\dfrac{1}{Q_n}I_T, \end{cases} \tag{1}$$

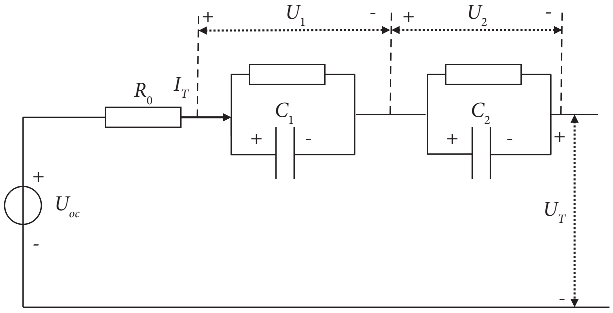


FIGURE 1: The schematic diagram of the second-order RC model.

where Q_n is the nominal capacity of the battery.

Then, its output equation can be defined as

$$U_T = U_{\mathrm{oc}}(\mathrm{SOC}) - R_0I_T - U_1 - U_2, \tag{2}$$

where $0 \leq \mathrm{SOC} \leq 1$, and $U_{\mathrm{oc}}(\mathrm{SOC})$ is the nonlinear monotone increasing function.

Based on (1) and (2), we can obtain state space equation as follows:

$$\begin{cases} \dot{x} = Ax + Bu, x(0) = x_0, \\ y = g(x) + Cx - R_0u, \end{cases} \tag{3}$$

where $x = \begin{bmatrix} U_1 & U_2 & \mathrm{SOC} \end{bmatrix}^T \in \mathbb{R}^3$, $y = U_T \in \mathbb{R}$, $u = I_T \in \mathbb{R}$, $g(x) = U_{oc}(\mathrm{SOC}) \in \mathbb{R}$, and x_0 is the initial state.

$$A = \begin{bmatrix} -\dfrac{1}{R_1C_1} & 0 & 0 \\[2ex] 0 & -\dfrac{1}{R_2C_2} & 0 \\[2ex] 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3}, B = \begin{bmatrix} \dfrac{1}{C_1} & \dfrac{1}{C_2} & -\dfrac{1}{Q_n} \end{bmatrix}^T \in \mathbb{R}^3,$$
$$C = \begin{bmatrix} -1 & -1 & 0 \end{bmatrix} \in \mathbb{R}^{1\times 3}. \tag{4}$$

As the power battery is a highly complex nonlinear system in its working state, there are many unknown uncertainties such as ambient temperature, battery self-discharge, battery life, and cycle interval. Therefore, the state space expression (3) can be expressed as follows:

$$\begin{cases} \dot{x} = \mathrm{Ax} + \mathrm{Bu} + d(x), x(0) = x_0, \\ y = g(x) + Cx - R_0u, \end{cases} \tag{5}$$

where $d(x)$ represents nonlinear characteristics.

Assumption 1. In this paper, we assume that (A, B) is stabilizable and (A, C) is detectable. The nonlinear term $d(x)$ is continuous and bounded.

Control objective: for the second-order RC equivalent model of power battery, based on an adaptive observer a policy learning algorithm-based optimal controller is designed to guarantee all signals of the closed-loop system uniformly ultimately bounded (UUB).

According to the second-order RC model of the power battery, we can derive its state space (3) or (5); then, we should design the control law u for the derived state space equation. Thus, we will use the NN observer and the policy learning algorithm to design the control law u.

3. Optimal Control of Power Battery

3.1. Observer Design via NN. This section will design an observer to estimate the battery voltage and SOC. Thus, we assume

$$d(x) = W_1^T\sigma(x) + \varepsilon(x), \tag{6}$$

where $W_1 \in \mathbb{R}^N$ is the ideal NN weights, $\sigma(x) \in \mathbb{R}^n \longrightarrow \mathbb{R}^N$ is the activation function, and $\varepsilon(x) \in \mathbb{R}$ denotes the NN error.

In this paper, the function $d(x)$ is unknown continuous; hence, the estimated function is

$$\hat{d}(x) = \hat{W}_1^T\sigma(x), \tag{7}$$

where $\hat{W}_1$ is the estimation of W_1.

Then, based on (5) and (7), the observer can be designed as

$$\begin{cases} \dot{\hat{x}} = A\hat{x} + Bu + \hat{W}_1^T\sigma(\hat{x}) + L\left[\dfrac{\partial g}{\partial x}\right]_{x=\hat{x}}^T (y - \hat{y}), \\ \\ \hat{y} = C\hat{x} + g(\hat{x}) - R_0u, \end{cases} \tag{8}$$

where $\hat{x}$ is the estimation of x, $L = P^{-1} \in \mathbb{R}^{3\times 3}$ is the observation matrix, P is the positive matrix, and $\hat{y}$ is the estimation of y.

We define the observation error

$$\tilde{x} = x - \hat{x}. \tag{9}$$

Then, from (5) and (8), we can obtain the observation error dynamic equation as

$$\dot{\tilde{x}} = \left[A - L\left(\frac{\partial g}{\partial x}\right)^T C\right]\tilde{x} - L\left(\frac{\partial g}{\partial x}\right)_{x=\hat{x}}^T \tilde{g} + W_1^T(\sigma(x) - \sigma(\hat{x})) + \tilde{W}_1\sigma(\hat{x}) + \varepsilon, \tag{10}$$

where $\tilde{g} = g(x) - g(\hat{x}) = \partial g/\partial x|x = \xi(x - \hat{x}), \tilde{W}_1 = \hat{W}_1 - W_1$ is the NN weight error.

Lemma 2. *For system (5), if it adopts designed observer (8), the NN weights $\hat{W}_1$ satisfy the adaptive law*

$$\dot{\hat{W}}_1 = -\sigma(\hat{x})\tilde{x}^T P. \tag{11}$$

This can guarantee that errors $\tilde{x}$ and $\tilde{W}_1$ are UUB.

Proof. Consider a Lyapunov function

$$V_1 = \frac{1}{2}\tilde{x}^T P\tilde{x} + \frac{1}{2}tr\left(\tilde{W}_1^T\tilde{W}_1\right). \tag{12}$$

From [15], we have $[\partial g/\partial x]_{x=\hat{x}}^T = [0, 0, \dot{U}_{oc}(\widehat{SOC})]$ with $\alpha_{\min} \le \dot{U}_{oc}(\widehat{SOC}) \le \alpha_{\max}$, where $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum values of the change rate of the $\dot{U}_{oc}$ function, respectively. Then, the derivation of (12) gives

$$\begin{aligned} \dot{V}_1 \le{}& \frac{1}{2}\dot{\tilde{x}}^T\left[PA + A^TP - RMC - C^T(RM)^T\right] - 2Q\Big]\tilde{x} \\ &+ \tilde{x}^T PW_1^T(\sigma(x) - \sigma(\hat{x})) + \tilde{x}^T P\cdot\tilde{W}_1\sigma(\hat{x}) + \tilde{x}^T P\varepsilon + \frac{1}{2}tr\left(\dot{\tilde{W}}_1^T\tilde{W}_1 + \tilde{W}_1^T\dot{\tilde{W}}_1\right), \end{aligned} \tag{13}$$

where $M = [m_1, \ m_2, \ m_3]^T \in \mathbb{R}^3$.

According to the unilateral Lipschitz condition [9], the following inequalities can be obtained:

$$\tilde{x}^T P\varepsilon \le \frac{1}{2}\|\tilde{x}\|^2 + \frac{1}{2}\|P\|^2\sum_{i=1}^{3}\varepsilon_i^{*2}, \tag{14}$$

$$\tilde{x}^T PW_1^{*T}(\sigma(x) - \sigma(\hat{x})) \le \|\tilde{x}\|^2 + \|P\|^2\|W_1\|^2. \tag{15}$$

Taking (14) and (15) into (13), and considering $\mathrm{tr}(\mathrm{ab}^T) = \mathrm{tr}(b^T a) = b^T a$, we have

$$\begin{aligned}\dot{V}_1 \leq & \frac{1}{2}\tilde{x}^T\left[PA + A^T P - RMC - C^T(RM)^T - 2Q\right]\tilde{x} + \|\tilde{x}\|^2 + \|P\|^2\left\|W_1\right\|^2 + \frac{1}{2}\|\tilde{x}\|^2 \\ & + \frac{1}{2}\|P\|^2\sum_{i=1}^{3}\varepsilon_i^{*2} + tr\left(\tilde{W}_1^T\sigma(\hat{x})\tilde{x}^T P + \tilde{W}_1^T\dot{\tilde{W}}_1\right).\end{aligned} \tag{16}$$

Based on [8], let $\mathrm{PA} + A^T P - \mathrm{RMC} - C^T(\mathrm{RM})^T - 2Q = -\Psi$, where $Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \alpha_{\min}^2 \end{bmatrix}$; thus, (16) can be further written as

$$\dot{V}_1 \leq -a_0\|\tilde{x}\|^2 + \frac{1}{2}\|P\|^2\left\|\tilde{W}_1\right\|^2 + D_0, \tag{17}$$

where $a_0 = \lambda_{\min}(\psi) - 3/2$ and $D_0 = \|P\|^2\|W_1\|^2 + 1/2\|P\|^2 \sum_{i=1}^{3}\varepsilon_i^2$.

If $\hat{d}(x) \longrightarrow d(x)$, then the term $1/2\|P\|^2\|\tilde{W}_1\|^2 + D_0$ in (17) can converge to zero. Moreover, by selecting the appropriate matrix ψ, $\lambda_{\min}(\psi)$ can be relatively large. According to (17), the observation error can converge to a small neighborhood containing the origin. □

3.2. Optimal Control Design Based on the Observer

3.2.1. Online Policy Learning Algorithm. In this section, based on critic NN, we construct the policy learning law. Thus, system (8) can be rewritten as

$$\dot{\hat{x}} = F(\hat{x}) + \mathrm{Bu}, \tag{18}$$

where $F(x) = Ax + \hat{W}_1^T\sigma(x) + L[\partial g/\partial x]_{x=\hat{x}}^T(y - \hat{y})$, and L is the Lyapunov function.

To realize the optimal control, we first define the cost function as\

$$V(\hat{x}, u) = \int_0^\infty r(\hat{x}, u)\mathrm{d}s. \tag{19}$$

With $r(\hat{x}, u) = \hat{x}^T Q_s \hat{x} + u^T R_s u$ being the utility function, $Q_s \in \mathbb{R}^{3\times 3}$ and $R_s \in \mathbb{R}$ are the weight matrices of proper dimension.

We define the Hamiltonian function of the optimal control problem and the optimal cost function as

$$H(\hat{x}, u, \nabla V(\hat{x})) = r(\hat{x}, u) + (\nabla V(\hat{x}))^T(F(\hat{x}) + \mathrm{Bu}). \tag{20}$$

$$V^*(\hat{x}) = \min_u \int_0^\infty r(\hat{x}, u)\mathrm{d}s. \tag{21}$$

The optimal cost function $V^*(\hat{x})$ is the solution of the following HJB equation:

$$0 = \min_u H(\hat{x}, u, \nabla V^*(\hat{x})). \tag{22}$$

With $\nabla V^*(x) = \partial V^*(x)/\partial x$, we can obtain this optimal control action as

$$u^* = -\frac{1}{2}R_s^{-1}B^T\nabla V^*(\hat{x}), \tag{23}$$

and the HIB equation in terms of $\nabla V^*(x)$ as

$$\begin{aligned} 0 = & \hat{x}^T Q_s \hat{x} + (\nabla V^*(\hat{x}))^T F(\hat{x}) \\ & - \frac{1}{4}(\nabla V^*(\hat{x}))^T B R_s^{-1} B^T \nabla V^*(\hat{x}), \end{aligned} \tag{24}$$

with $V^*(0) = 0$.

To realize the policy learning, some iteration procedure can be given as follows:

(1) Select the small positive number τ. Set $i = 0$ and $V^{(0)} = 0$, and then give an initial admissible control $u^{(0)}$.

(2) Using the control $u^{(i)}$, resolve

$$0 = r(\hat{x}, u) + \left(\nabla^{i+1}V(\hat{x})\right)^T\left(F(\hat{x}) + Bu^i\right), \tag{25}$$

with $V^{(i+1)}(0) = 0$.

(3) Update the control action using

$$u^{(i+1)} = \frac{1}{2}R_s^{-1}B^T\nabla V^{(i+1)}(\hat{x}). \tag{26}$$

(4) If $\|V^{(i+1)}(\hat{x}) - V^{(i)}(\hat{x})\| \leq \tau$, stop, then apply the optimal control; else, let $i = i + 1$ and go back to (2).

This algorithm will be convergence to the optimal control and optimal cost function when $i \longrightarrow \infty$. The convergence of this algorithm can be referred to [16, 17].

3.2.2. NN Implementation. We assume the cost function $V(\hat{x})$ is continuously differentiable. Then, we can use the NN reconstruct the $V(\hat{x})$ as

$$V(\hat{x}) = W_2^T\sigma_c(\hat{x}) + \varepsilon_c(\hat{x}), \tag{27}$$

where $W_2 \in \mathbb{R}^N$ is the ideal NN weights, $\sigma_c(x) \in \mathbb{R}^n$ is the activation function, and $\varepsilon_c(\hat{x}) \in \mathbb{R}$ denotes the NN error. Then,

$$\nabla V(\hat{x}) = \left(\nabla\sigma_c(\hat{x})\right)^T W_2 + \nabla\varepsilon_c(\hat{x}), \tag{28}$$

where $\nabla\sigma(\hat{x}) = \partial\sigma_c(\hat{x})/\partial\hat{x}$ and $\nabla\varepsilon_c(\hat{x}) = \partial\varepsilon_c(\hat{x})/\partial\hat{x}$ are the gradient of the activation function and NN error, respectively. According to (28), we can obtain the Lyapunov function as

$$0 = r(\hat{x}, u) + \left(W_2^T \nabla\sigma_c(\hat{x}) + (\nabla\varepsilon_c(\hat{x}))^T\right)\dot{\hat{x}}. \tag{29}$$

Assumption 3. (see [12–14, 18]). If the NN weight W_2, the NN error ε_c, the gradient $\nabla\sigma_c$, and derivative $\nabla\varepsilon_c$ are bounded, then we can have $\varepsilon_c \longrightarrow 0$ and $\nabla\varepsilon_c \longrightarrow 0$.

We define the estimation of (27) as

$$\widehat{V}(\hat{x}) = \widehat{W}_2^T \sigma_c(\hat{x}). \tag{30}$$

Then, we have

$$\nabla\widehat{V}(\hat{x}) = (\nabla\sigma_c(\hat{x}))^T \widehat{W}_c. \tag{31}$$

with $\nabla\widehat{V}(\hat{x}) = \partial\widehat{V}(\hat{x})/\partial\hat{x}$. Thus, the estimated Hamiltonian function can be given as

$$H\left(\hat{x}, u, \widehat{W}_2\right) = r(\hat{x}, u) + \widehat{W}_2^T \nabla\sigma_c(\dot{x})\dot{\hat{x}} = e_c. \tag{32}$$

To minimize error (32), we construct the objective function $J = (1/2)e_c^T e_c$, and then the descent algorithm can be designed as

$$\dot{\widehat{W}}_2 = -\alpha_1\left[\frac{\partial J}{\partial W}\right] = -\alpha_1\left[\frac{\partial e_c}{\partial W}\right], \tag{33}$$

with $\alpha_1 > 0$ being the learning gain of the NN.

Based on (29), the Hamiltonian function can be rewritten as

$$H(\hat{x}, u, W_2) = r(\hat{x}, u) + W_2^T \nabla\sigma_c(\hat{x})\dot{\hat{x}} = e_h, \tag{34}$$

where $e_h = -(\nabla\varepsilon_c(\hat{x}))^T \dot{\hat{x}}$ is the residual error.

Define $\phi = \nabla\sigma_c(\hat{x})\dot{\hat{x}}$, if there is a positive constant ϕ_M such that $\|\phi\| \leq \phi_M$, and denote the weight estimation error $\widetilde{W}_2 = W_2 - \widehat{W}_2$, and then based on (32) and (34), we have $e_h - e_c = \widetilde{W}_2^T\phi$; thus, we have the dynamic of the weight estimation error as

$$\dot{\widetilde{W}}_2 = -\dot{\widehat{W}}_2 = \alpha_1\left(e_h - \widetilde{W}_2^T\phi\right)\phi. \tag{35}$$

The persistent excitation (PE) condition is required to tune the NN, guaranteeing $\|\phi\| \geq \phi_m$ with ϕ_m being the positive constant. To this end, a probing noise is inserted into the system to meet the PE.

In this case, the optimal control action can be given as

$$u^* = -\frac{1}{2}R_s^{-1}B^T\left((\nabla\sigma(\hat{x}))^T W_2 + \nabla\varepsilon_c(\hat{x})\right), \tag{36}$$

and its estimation is

$$\hat{u} = -\frac{1}{2}R_s^{-1}B^T(\nabla\sigma(\hat{x}))^T\widehat{W}_2. \tag{37}$$

Equation (37) shows that using the trained critic network, the control policy can be derived directly; thus, the actor NN is removed in this paper. The structural diagram of the algorithm is given in Figure 2.

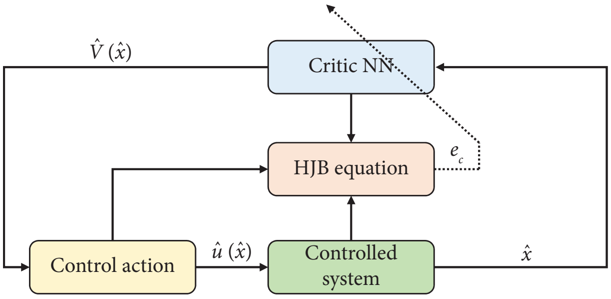


FIGURE 2: The structural diagram of the algorithm.

Lemma 4. *For system (18), the adaptive law for the NN is provided by (33), and then the weight estimation error of NN is UUB.*

Proof. Choose the Lyapunov function as $K(t) = (1/\alpha_1)\mathrm{tr}(\widetilde{W}_2^T\widetilde{W}_2)$. The time derivative of the Lyapunov function along the trajectory of error dynamics (35) is

$$\dot{K}(t) = \frac{2}{\alpha_1}\mathrm{tr}\Bigg(\widetilde{W}_2^T\dot{\widetilde{W}}_2) = \frac{2}{\alpha_1}tr\left(\widetilde{W}_2^T\alpha_1\left(e_h - \widetilde{W}_2^T\phi\right)\phi\right). \tag{38}$$

After doing some basic manipulations, we have

$$\dot{K}(t) \leq -(2-\alpha_1)\left\|\widetilde{W}_2^T\phi\right\|^2 + \frac{1}{\alpha_2}e_h^2. \tag{39}$$

Considering the Cauchy–Schwarz inequality and noticing the assumption $\|\phi\| \leq \phi_M$, we can conclude that $\dot{K}(t) < 0$ as long as $1 < \alpha_1 < 2$ and

$$\left\|\widetilde{W}_2\right\| > \sqrt{\frac{e_h^2}{\alpha_1(2-\alpha_1)\phi_M^2}}. \tag{40}$$

According to the Lyapunov theory, we obtain that the dynamics of the weight estimation error is UUB. The norm of the weight estimation error is bounded as well.

It is noted that the estimated weight $\widehat{W}_2$ is optimal to W_2, and this indicates that the solution $\widehat{V}$ can be extracted from the estimated vector $\widehat{W}_2$ given in (30). Thus, one can derive the actual control $\hat{u} = -1/2R_s^{-1}B^T(\nabla\sigma(\hat{x}))^T\widehat{W}_2$ for system (18) based on $\widehat{W}_2$. As a consequence of Lemma 4, we can conclude that $\hat{u}$ will converge to the optimal control u^*, i.e., $\|\hat{u} - u^*\| \longrightarrow 0$ such that the control system stability can be retained based on Lemma 4. □

Remark 5. In this paper, an observer is designed using NN to online estimate the unknown state (SOC); then, based on the estimated state, we develop a policy learning algorithm to online resolve the optimal control of the battery. The proposed methods are different from our previous work, such as [18], where the system states are assumed to be known, and this limits the application of the optimal control algorithm in practice.

Remark 6. To realize the output-feedback control using the policy learning, the PE condition is required in this paper. As shown in [14, 17], to guarantee the PE condition, an alternative way is to insert an exploration noise into the system for the first two seconds [17].

4. Simulation Results

For the second-order RC equivalent model of power battery, the effectiveness of the optimal control theory in this paper is verified by simulation based on Matlab. The values of resistance, capacitance, and battery capacity in the second-order RC equivalent model (5) are as follows: $R_0 = 10.822\text{m}\Omega$, $R_1 = 3.103\text{m}\Omega$, $R_2 = 2.611\text{m}\Omega$, $C_1 = 8.4379\text{kF}$, $C_2 = 91.401\text{kF}$, and $Q_n = 45\text{A} \cdot \text{h}$.

Let $M = I$, then we can obtain P and L as

$$\begin{aligned} P &= \begin{bmatrix} 14.1250 & 0 & 19.6371 \\ 0 & 128.7451 & 178.9860 \\ 19.6371 & 178.9860 & 0 \end{bmatrix}, \\ L &= \begin{bmatrix} 0.0638 & -0.007 & 0.005 \\ -0.007 & 0.0008 & 0.005 \\ 0.005 & 0.005 & -0.0036 \end{bmatrix}. \end{aligned} \tag{41}$$

Given the design parameters in learning law (33) as $\alpha_1 = 0.1$ and the initial values as $x_1(0) = 0.1, x_2(0) = 0.2, x_3(0) = 1$, $\hat{x}_1(0) = 0.01, \hat{x}_2(0) = 0, \hat{x}_3(0) = 0.99$, and $\widehat{W}_2 = [0.3909\ 0.5812\ 1.0576\ 0.1\ 0.2\ 1]$, we design the regressor of the critic NN as $\sigma(x) = [x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2]^T$.

We aim at obtaining an optimal control policy that can stabilize system (18). For system (18), we need to find a feedback control policy that minimizes the cost function.

$$V(\hat{x}, u) = \int_0^\infty \left(\hat{x}^T Q_s \hat{x} + u^T R_s u\right) \text{d}s, \tag{42}$$

with $Q_s = I$ and $R_s = 2I$. We adopt the online policy iteration algorithm to tackle the optimal control problem, where a critic network is constructed to approximate the cost function. During the implementation process of the policy learning algorithm, we introduce the noise to meet the PE condition. The exponentially decreasing probing noise and sinusoidal signals with different frequencies are used. They are introduced into the control input and thus affect the system states.

The evolution of the state trajectory is depicted in Figure 3, and this can be used to further design the optimal controller for the proposed system. Figure 4 gives the good estimated weights, where we have that the convergence of the weight has occurred after 1000 s. Then, the probing signal is turned off. This good convergence of the NN weights can ensure the stability of the controlled system, which can be found in Figure 5. Figure 5 is the controller system trajectory with the designed optimal controller. We see that the state converge to zero after the probing noise is turned off. Figure 6 shows the cost of the system under which

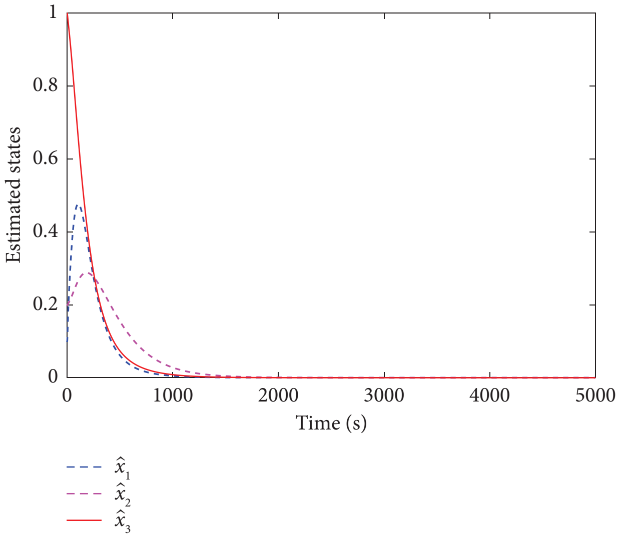


FIGURE 3: Estimated system states.

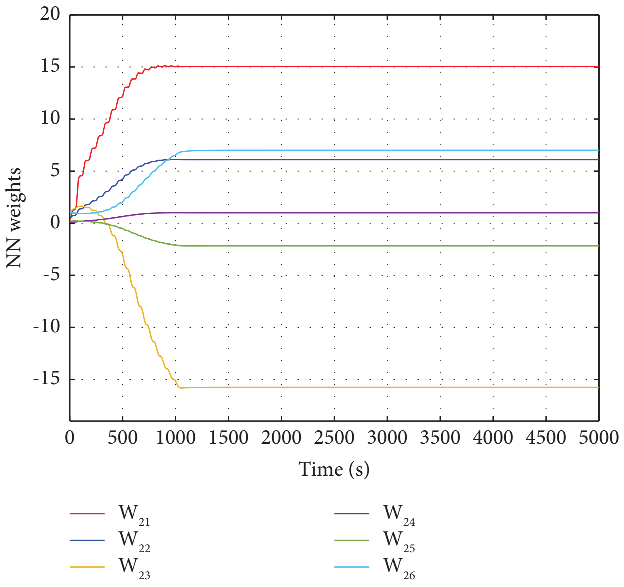


FIGURE 4: NN weights.

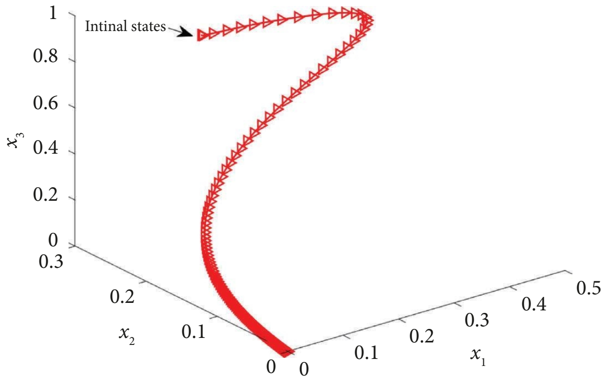


FIGURE 5: System trajectory.

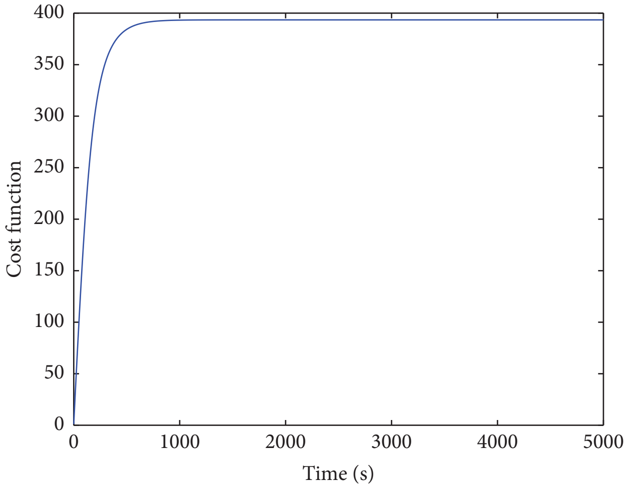


Figure 6: Cost function.

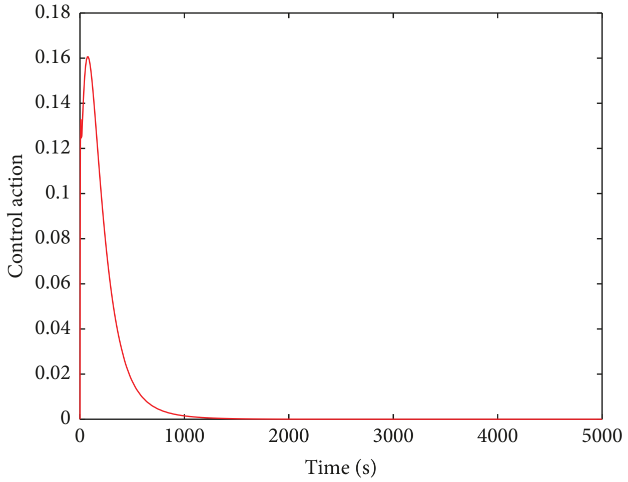


Figure 7: Control action u.

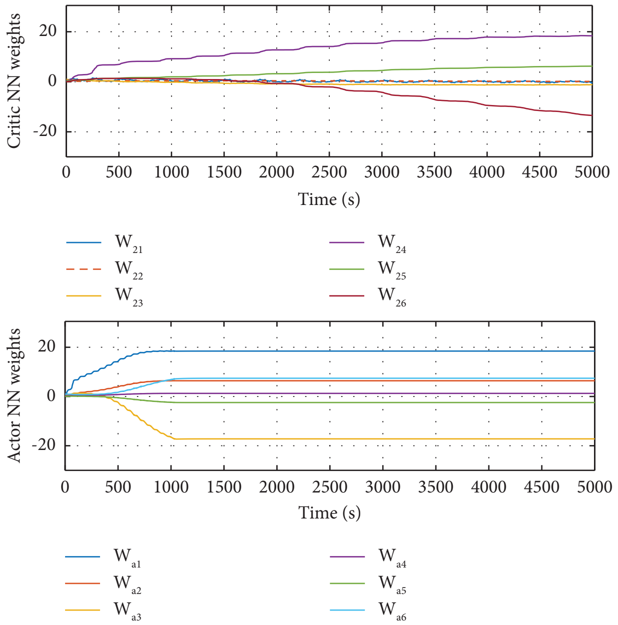
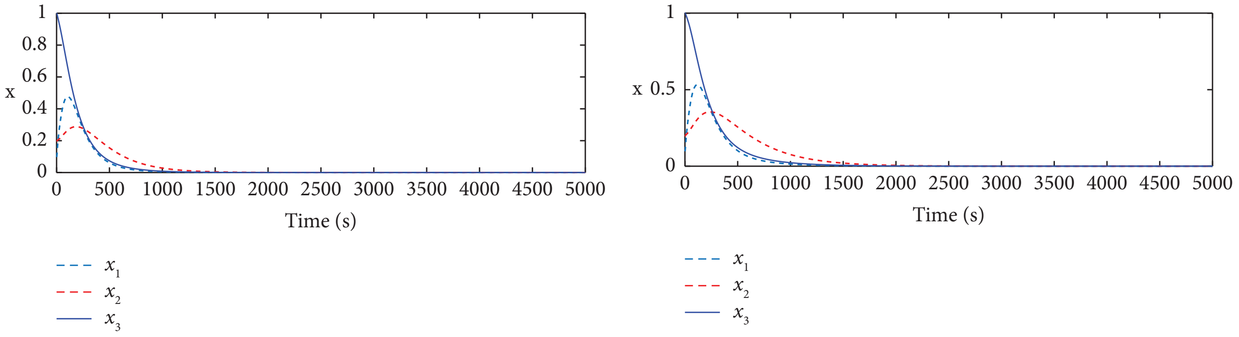


Figure 8: NN weights.



(a)

(b)

Figure 9: System state x (a) using the proposed method and (b) the method proposed in [19, 20].

is smooth, and this indicates that the designed controller is effective. The control action is given in Figure 7, which is bounded. This further shows Lemma 4 is true.

To show the improved performance of the proposed single critic NN-based ADP for solving the derived optimal control problem, a critic-actor NN-based online learning method [19] is also used for comparison. Moreover, in this comparison, we add the robustness verification of the proposed method. To this end, we set the nonlinear term $d(x) = 0.5\sin(x_1)$. The profiles of the critic NN and actor NN weights can be found in Figure 8 and the corresponding control performances are given in Figure 9. Compared with Figures 9(a) and 9(b), it is clear that the proposed single critic NN-based can achieve faster transient state convergence even if there is a nonlinear term.

Generally, the modeling accuracy and control structure will influence the control performance of the closed-loop control systems. In this paper, the main factors affecting the control performance are the modeling uncertainties of the system and the convergence performance of critic NN weights. Moreover, better convergence of critic NN weights, i.e., faster convergence speed can help to achieve better control performance. In this respect, different choices of critic NN parameters and structure will affect the convergence of critic NN weights and the control performance. Hence, proper selection of NN parameters and structure, such as the initial value of weights, learning gain, and regressor structure, is helpful to further improve the control response.

5. Conclusion

For the second-order RC equivalent nonlinear system of power battery, the unknown uncertainty of the system is approximated by NN, and a time-varying gain nonlinear state observer is designed to solve the problem that the resistance capacitance voltage and charge (SOC) of the battery cannot be measured. Then, to realize the optimal control, a policy learning-based online algorithm is designed, where only the critic NN is required, and the actor NN widely used in most design of the optimal control methods is removed. Finally, the effectiveness of the optimal control theory is verified by simulation.

Data Availability

The data used to support the findings of this study are available upon request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qinglin Zhu and Jun Zhao conceptualized the study; Huanli Sun and Ziliang Zhao were responsible for methodology; Ziliang Zhao and Yixin Liu performed formal analysis; Qinglin Zhu wrote the original draft; Qinglin Zhu and Yixin Liu reviewed and edited the manuscript; and Huangli Sun and Ziliang Zhao were responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported in part by Jilin Provincial Major Science and Technology Projects (Grant no.: 20210301020GX).

References

[1] G. Klancar and S. Blazic, "Optimal constant acceleration motion primitives," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8502–8511, 2019.

[2] M. Eshani, Y. Gao, S. E. Gay, and A. Emadi, *Modern Electric, Hybrid Electric, and Fuel Cell Vehicles*, CRC Press, Boca Raton, FL, USA, 2005.

[3] H. He, R. Xiong, X. Zhang, F. Sun, and J. Fan, "State-of-charge estimation of the lithium-ion battery using an adaptive extended kalman filter based on an improved thevenin model," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1461–1469, 2011.

[4] K. W. E. Cheng, B. P. Divakar, H. Wu, K. Ding, and H. F. Ho, "Battery-management system (BMS) and SOC development for electrical vehicles," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 1, pp. 76–88, 2011.

[5] S. Ahmad, M. Rehan, and K. S. Hong, "Observer-based robust control of one-sided Lipschitz nonlinear systems," *ISA Transactions*, vol. 65, pp. 230–240, 2016.

[6] B. Xia, C. Chen, Y. Tian, W. Sun, Z. Xu, and W. Zheng, "A novel method for state of charge estimation of lithium-ion batteries using a nonlinear observer," *Journal of Power Sources*, vol. 270, pp. 359–366, 2014.

[7] Q. Zhu, N. Xiong, M. Yang, R. Huang, and D. Hu, "State of charge estimation for lithium-ion battery based on nonlinear observer: an H∞ method," *Energies*, vol. 10, no. 5, p. 679, 2017.

[8] X. Li and Y. Li, "Neural networks optimized learning control of state constraints systems," *Neurocomputing*, vol. 453, pp. 512–523, 2021.

[9] R.-C. Roman, R.-E. Precup, E.-L. Hedrea et al., "Iterative feedback tuning algorithm for tower crane systems," *Procedia Computer Science*, vol. 199, pp. 157–165, 2022.

[10] T. Chen, A. Babanin, A. Muhannad, B. Chapron, and C. Chen, "Modified evolved bat algorithm of fuzzy optimal control for complex nonlinear systems," *Romanian Journal of Information Science and Technology*, vol. 23, no. T, pp. T28–T40, 2020.

[11] I. A. Zamfirache, R.-E. Precup, R.-C. Roman, and E. M. Petriu, "Policy iteration reinforcement learning-based control using a grey wolf optimizer algorithm," *Information Sciences*, vol. 585, pp. 162–175, 2022.

[12] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[13] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, "Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming," *Automatica*, vol. 48, no. 8, pp. 1825–1832, 2012.

[14] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[15] J. Zhang, K. Li, and Y. Li, "Output feedback based simplified optimized backstepping control for strict-feedback systems with input and state constraints," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 6, pp. 1119–1132, 2021.

[16] Y. Li, X. Pei, and S. Yi, "Adaptive neural network optimal control of hybrid electric vehicle power battery," *Journal of Jilin University Engineering and Technology Edition*, vol. 52, no. 9, pp. 2063–2068, 2022.

[17] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 627–632, 2014.

[18] J. Zhao and Y. Lv, "Output-feedback robust control of systems with uncertain dynamics via data-driven policy learning," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 18, pp. 9791–9807, 2022.

[19] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2226–2236, 2011.

[20] J. Zhao, J. Na, and G. Gao, "Robust tracking control of uncertain nonlinear systems with adaptive dynamic programming," *Neurocomputing*, vol. 471, pp. 21–30, 2022.