IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 19, 2020, accepted July 3, 2020, date of publication August 11, 2020, date of current version August 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015801

Optimal Control of Iron-Removal Systems Based on Off-Policy Reinforcement Learning

NING CHEN, SHUHAN LUO, JIAYANG DAI, BIAO LUO, (Senior Member, IEEE), AND WEIHUA GUI, (Member, IEEE)
School of Automation, Central South University, Changsha 410083, China

Corresponding author: Ning Chen (ningchen@csu.edu.cn)

This work was supported in part by the Program of National Natural Science Foundation of China under Grant 61673399, and in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621062.

ABSTRACT The goethite iron-removal process is an important procedure to remove the iron ions from the zinc hydrometallurgy. However, as a coherent system with complex reaction mechanism, associated uncertainties, and interconnected adjacent reactors, it is difficult for the process to accurately control the ion concentration. Because a large amount of historical data can be obtained during the process, an optimal control algorithm based on off-policy reinforcement learning is proposed in this paper to overcome these difficulties. According to the historical data, the weights of neural network are learned offline, and the optimal control strategy is solved online. Firstly, a bounded function is introduced to define the maximum effect of the coherent system on the subsystem cost function and to extend the cost function of the nominal system, so that the decentralized guaranteed cost control problem can be expressed as the optimal control problem of the nominal system. Then, an approximate iterative control algorithm based on actor-critic structure is proposed. The actor and critic neural networks are used to approximate control strategies and cost functions respectively. To achieve complete off-line, a new neural network is added to the actor-critic structure to approximate a part of the unknown system structure, and the three neural network parameters are optimized by the state transition algorithm. Finally, the strategy update and strategy iteration operations are performed alternately to learn optimal control strategies. The effectiveness and flexibility of the proposed off-policy optimal control method is validated by data from a real industrial goethite iron-removal process.

INDEX TERMS Goethite iron-removal process, optimal control, off-policy, reinforcement learning.

I. INTRODUCTION

Zinc is an important non-ferrous raw material, which plays an important role in various fields. It is widely used in non-ferrous metallurgy, batteries, machinery, automobile manufacturing and other industries. At present, most zinc smelting enterprises adopt the atmospheric pressure oxygen enriched direct leaching zinc smelting method with high iron zinc sulfide concentrate as raw material [1], [2], which can effectively reduce sulfur dioxide emissions and improve the recovery rates of valuable metals in leaching solution. Because the zinc sulfide concentrate is rich in iron, the leaching solution will contain high concentration of iron ions. If the iron ion concentration in the leaching solution exceeds the range of the technical requirement, it will lead to impurity of the zinc product, and it will increase power consumption in the electrolytic process. Therefore, the goethite iron-removal process is an important link in the leaching process. In the goethite iron-removal process, oxygen is added to oxidize the ferrous iron into ferric iron. Then, the ferric iron is hydrolyzed to goethite precipitate for iron-removal. In the process, excessive oxygen is generally replenished into the rector for the required range of iron ion concentration. This approach will cause waste of raw materials and sharp fluctuation of pH value, as well as low grade of goethite precipitate or even no goethite precipitate. Therefore, in order to ensure the quality of iron-removal and save raw materials, controlling the oxygen addition rate is a critical step.

Goethite iron-removal process consists of four continuous reactors arranged in descending order, overflowing zinc

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

leach solution from one reactor to the next. It is a non-linear process involving a series of complex chemical reactions such as oxidation, hydrolysis, and neutralization. For the non-linear problems in the industrial production process, a lot of effective control methods have been proposed [3]–[10]. According to the law of mass conservation and reaction kinetics, Chen *et al.* [4] established a single-tank continuous stirred reactor model based on the law of conservation of mass and reaction kinetics for the first time. Based on the single-reactor model, considering the existence of unreacted oxygen in the leaching solution, a cascade reactor coupled control distributed model is established. The distributed model predictive control strategy is adopted to solve the optimization control problem of iron-removal process. Xie *et al.* [5] established a weighted coupled CSTR model for goethite iron-removal process by introducing weighting parameters. A parameter identification method to determine unknown parameters is proposed. Then, a model predictive control scheme is designed to achieve process performance objectives and minimize process costs. Han *et al.* [6] and others transformed the dynamic optimization problem of the iron-removal process into a nonlinear mathematical programming problem and proposed a multi-objective optimization method based on the state transition algorithm and constrained non-dominated sorting to find the optimal solution of the oxygen concentration and zinc oxide addition. Sun *et al.* [7] proposed a steady-state multiple reactors gradient optimization, unsteady-state operational pattern adjustment strategy, and a process evaluation strategy based on the oxidation-reduction potential and proved the effectiveness of this study in industrial experiments. Yang *et al.* [8] proposed a model-based hybrid adaptive dynamic programming (ADP) framework consisting of continuous feedback-based policy evaluation and policy improvement steps as well as an intermittent policy implementation procedure. Shahlavi *et al.* [9] developed a novel fully distributed controller based on backstepping technique and neuro-adaptive update mechanism. The simulation results are carried out to demonstrate the effectiveness of the proposed approach. And literature [10] presents a distributed solution for consensus control of a network of single-integrator incommensurate fractional-order systems with nonlinear and uncertain dynamics. However, most of the industrial processes have complex environment, and the mechanism model cannot reflect the system dynamics completely and truly, and there must be modeling errors, unmodeled dynamics and various uncertainties, which makes the model-based control theory and method defective.

In order to solve a series of problems caused by imprecise model, many researches put forward data-based control methods [11]–[13]. Different from other data-based control methods, a significant advantage of reinforcement learning based control method [14]–[23] is that it can achieve performance optimization control in unknown environment, which is undoubtedly of great significance to practical engineering applications. Because of this, many researchers introduce reinforcement learning into optimal control problems. Li *et al.* [18] proposed a novel off-policy interleaved Q-learning algorithm for solving optimal control problem of affine nonlinear discrete-time systems, using only the measured data along the system trajectories. Lewis and Vamvoudakis [20] proposed an online strategy iteration algorithm based on reinforcement learning. The algorithm uses an actor-critic network architecture and adjusts the weights of the actor network and the critic network synchronously. Yang *et al.* [21] proposed a novel barrier-actor-critic algorithm that is presented for adaptive optimal learning while guaranteeing the full-state constraints and input saturation. Luo *et al.* [22] proposed a data-based approximation strategy iteration algorithm, which updates the weights using the weighted residual method based on the least square. Yan *et al.* [23] proposed a Q-learning algorithm based on policy iteration, and proved that under the given bounded condition, the approximate q-function would converge to the finite neighborhood of the optimal Q-function. Zhu *et al.* [25] transformed the H∞ optimal control problem into the zero-sum game problem and obtained the H∞ optimal control law of the perturbed system by using the strategy iteration algorithm. Dong [28] studied the event triggered iterative ADP method and applied it to the optimal control of grinding process [29]. Maldonado *et al.* [30] *et al.* studied the optimal control of flotation cell by using adaptive dynamic programming method. This method can learn from the operation data of flotation cell and improve the controller iteratively. In recent years, some scholars also extended RL method to decentralized control. Liu *et al.* [31] used an online learning optimal control method based on neural network, put forward a decentralized control strategy to stabilize a class of continuous time nonlinear interconnected system, and designed the optimal controller isolated from the system by using the cost function reflecting the interconnect boundary. Wang *et al.* [32] proposed a learning based optimal control method for the optimal control of interconnected systems. By combining the robust decentralized control formula with adaptive critical learning technology, the decentralized guaranteed cost [33], [34] controller was designed. This method is still implemented under the condition where the model is known. At present, the modeless RL method for nonlinear continuous time decentralized control is still an open problem, which also promotes the research of this paper.

In this paper, the decentralized control problem of continuous time nonlinear system with unknown model is considered. Inspired by Wang *et al.* [32], this paper introduces a bounded function to transform the decentralized guaranteed cost control problem into the nominal system optimal control problem. Based on the literature [32], a reinforcement learning based optimal control algorithm is proposed. Three neural networks are used to approach the critic network, the actor network and a part of the unknown system structure respectively. For the problem that the traditional solution method, such as the minimum residual method [22] doesn't work when the linear relationship between the residual and

the parameters is not satisfied, the three neural network parameters are optimized by the state transition algorithm [35], and the optimal control strategy is learned when the system model is unknown.

The rest of this paper is arranged as follows. In the second section, the decentralized optimal control problem is described. In the third section, the iterative algorithm of decentralized control based on data is derived. In the fourth section, through the simulation experiment of real industrial data, the applicability of the decentralized control strategy is verified. A brief conclusion is given in Section 5.

II. DESCRIPTION OF THE OPTIMAL CONTROL OF IRON-REMOVAL SYSTEM

The object of goethite process is zinc sulphate solution obtained by direct leaching of zinc concentrate. Usually, the iron-removal needs to be performed in the temperature range of 65°C to 82 °C.

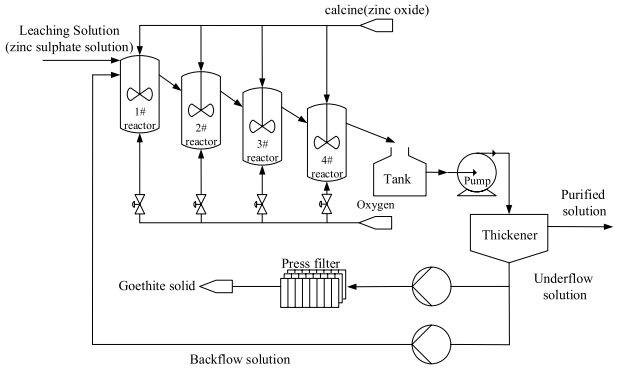

FIGURE 1. Process flow chart of goethite process.

The goethite process in a representative Chinese zinc smelting plant is taken to investigate, the simplified flow chart is shown in Fig.1. The iron-removal is completed slowly in the continuous stirred reactors. The zinc sulphate solution from the previous procedure, which enters the 1# reactor and can be called as inlet solution, would flow out of reactor 1# as the outlet solution and flow into reactor 2# afterwards. Similarly, the outlet solution of the previous reactor is the inlet solution of the next one. Moreover, the outlet concertation of Fe^{3+}, Fe^{2+} and pH in each reactor should be controlled with in the set range, making the excessive iron-ion less than $1\text{g}\cdot\text{L}^{-1}$, and the pH value between 3.0 and 4.0. Therefore, the zinc sulphate solution leaves the 4# reactor is qualified for the next procedures. In addition, partial zinc sulphate solution exiting in the last reactor is sent back to 1# reactor as the backflow solution, since the crystal nucleus of goethite in it can promote the proceed of iron-removal.

In the process, each reactor has a series of complex chemical reactions among gas, liquid and solid. From the aspect of the influence on iron-removal, the main chemical reactions are presented as follows:

Oxidation reaction:

$$4Fe^{2+} + 4H^{+} + O_2 \rightarrow 4Fe^{3+} + 2H_2O \tag{1}$$

Neutralization reaction:

$$2H^{+} + ZnO \rightarrow Zn^{2+} + H_2O \tag{2}$$

Hydrolysis reaction:

$$Fe^{3+} + 2H_2O \leftrightarrow FeOOH + 3H^{+} \tag{3}$$

In the oxidation reaction, Fe^{2+} is oxidized to Fe^{3+} that hydrolyzes to form goethite precipitation, which can be removed through filtration. And the neutralization reaction ensures certain reaction conditions, i.e. The pH value of the solution is kept within a certain range.

In the actual goethite process, data sampling can only be carried out every two hours because of the sealed reactors. Under this condition, those process data obtained by periodic sampling cannot be directly used in continuous time optimal control. Therefore, it is necessary to establish the mechanism model of the iron-removal process for optimal control.

Assume the temperature in the reactor remains unchanged and the solution is mechanically stirred uniformly. Then the rates of the oxidation, hydrolysis, and neutralization reaction in the solution can be obtained from the chemical reaction kinetics:

$$r_{Fe^{2+}} = k_1 c^{\alpha}_{Fe^{2+}} c^{\beta}_{H^{+}} c^{\gamma}_{O_2} \tag{4}$$

$$r_{Fe^{3+}} = k_2 c_{Fe^{3+}} \tag{5}$$

$$r_{H^{+}} = \frac{3m}{\rho R_s} k_3 c_{H^{+}} \tag{6}$$

where k_1, k_2, k_3 are the reaction rate constants, α, β, γ are the reaction orders. C_{O_2} is the concentration of dissolved oxygen in the solution and an important variable that affects the oxidation rate of the ferrous ions. Parameter m is the mass of zinc oxide, ρ represents the density of zinc oxide particles and R_s is the radius of zinc oxide particles.

Simulation based on the CSTR model of goethite process is proposed in reference [5]. In the research, the control variable available to the controller is oxygen. According to the reaction rate equation, the dynamic equation of dissolved oxygen and the law of mass conservation, the ion concentration and dissolved oxygen in the solution at the outlet of the reactor are taken as states. The model of a single reactor can be described as follows:

$$\begin{cases} \dfrac{dc_{Fe^{2+}}}{dt} = \dfrac{F}{V}(C_{Fe^{2+},in} - C_{Fe^{2+}}) - k_1 C^{\alpha}_{Fe^{2+}} C^{\beta}_{H^{+}} C^{\gamma}_{o_2} \\ \dfrac{dc_{Fe^{3+}}}{dt} = \dfrac{F}{V}(C_{Fe^{3+},in} - C_{Fe^{3+}}) + k_1 C^{\alpha}_{Fe^{2+}} C^{\beta}_{H^{+}} C^{\gamma}_{o_2} - k_2 C_{Fe^{3+}} \\ \dfrac{dc_{H^{+}}}{dt} = \dfrac{F}{V}(C_{H^{+},in} - C_{H^{+}}) - k_1 C^{\alpha}_{Fe^{2+}} C^{\beta}_{H^{+}} C^{\gamma}_{o_2} \\ \qquad + k_2 C_{Fe^{3+}} - \dfrac{3m_{zno}}{\rho R_s} k_3 C_{H^{+}} \end{cases} \tag{7}$$

where F is the flow rate, V is the reactor volume. Parameters $C_{Fe^{2+},in}$, $C_{Fe^{3+},in}$ and $C_{H^+,in}$ are the inlet concentrations of Fe^{3+}, Fe^{2+} and H^+, respectively. Parameter m_{zno} represents the mass of zinc oxide, and C_{O_2} is the dissolved oxygen concentration. The dissolved oxygen concentration is selected as a new state variable, where ρ, R_S, V and ρ_{O_2} are constants, and $F, C_{Fe^{2+},in}, C_{Fe^{3+},in}, C_{H^+,in}, C_{Fe^{2+}}, C_{Fe^{3+}}, C_{H^+}$ are obtained from the sampled data. The reaction rate constants k_1, k_2, k_3 and the reaction orders α, β, γ are the parameters to be obtained.

In the actual goethite process, the ion concentration range at the inlet of the reactor is shown in the table below:

TABLE 1. Concentration ranges of three ion at the reactor inlet.

	Ion concentration range
Fe^{3+}	1.43g/L-1.81 g/L
Fe^{2+}	10.45 g/L -14.24 g/L
H^+	3.37-3.79

And reaction rate constant range is shown as the table below:

TABLE 2. Reaction rate constant range.

	Reaction rate constant range
α	0.2-5.0
β	0.2-5.0
γ	0.2-5.0

According to literature [4], normally the value of reaction rate constants α, β, γ can be 1. In this paper, therefore, the value of α, β, γ is set 1.

However, even if only the key variable set of the goethite process is considered, the interaction between these variables makes the solution of the parameters still a considerable challenge. Therefore, a certain degree of modelling accuracy is often sacrificed in practice, which directly affects control accuracy.

Moreover, the goethite system is a coherent system. The solution in the former reactor will flow into the latter reactor, hence the ion concentration at the outlet of the former reactor is equal to the ion concentration at the inlet of the subsequent reactor. Consider the j# reactor as subsystem j, define the state of subsystem j as $x_j(t) = [c_{j,Fe^{2+}}, c_{j,Fe^{3+}}, c_{j,H^+}]^T$, and $u_j = c_{j,o_2}$ as the control variable of subsystem j. To obtain the optimal control when the parameters are difficult to solve, it is assumed that the goethite process in any reactor can be expressed as the state space in equation (8) referring to equation (7).

$$\dot{x}_j(t) = f_j(x_j) + g_j(x_j)u_j + h_j(x), \quad j = 1, \ldots, 4 \tag{8}$$

The system functions $f_j(\cdot)$ and $g_j(\cdot)$ are both differentiable, and $h_j(x(t))$ represents the concentrations between subsystem j and other subsystems. System functions and concentrations have two forms in the iron-removal process. For the 1# reactor:

$$f_1(x_1) = \begin{pmatrix} \frac{F}{V}C_{1,in} - \frac{F}{V}x_{11} \\ \frac{F}{V}C_{2,in} - x_{12}\frac{F}{V} - k_{12}x_{12} \\ \frac{F}{V}C_{3,in} - x_{13}\frac{F}{V} + k_{12}x_{12} - \frac{3m_{i,zno}}{\rho R_s}k_{13}x_{13} \end{pmatrix},$$
$$g_1(x_1) = \begin{pmatrix} -k_{11}x_{11}x_{13} \\ k_{11}x_{11}x_{13} \\ -k_{11}x_{11}x_{13} \end{pmatrix}$$
$$h_1(x) = \left[(\frac{F_b}{V}x_{41})^T (\frac{F_b}{V}x_{42})^T (\frac{F_b}{V}x_{43})^T \right]^T \tag{9}$$

where $C_{1,in}$, $C_{2,in}$ and $C_{3,in}$ represent the Fe^{2+}, Fe^{3+} and H^+ concentration of the solution at the inlet of the 1# reactor, respectively. F_b represents the backflow rate of the 4# reactor. For 2#-4# reactors:

$$f_j(x_j) = \begin{pmatrix} -\frac{F}{V}x_{j1} \\ -x_{j2}\frac{F}{V} - k_{j2}x_{j2} \\ -x_{j3}\frac{F}{V} + k_{j2}x_{j2} - \frac{3m_{j,zno}}{\rho R_s}k_{j3}x_{j3} \end{pmatrix},$$
$$g_j(x_j) = \begin{pmatrix} -k_{j1}x_{j1}x_{j3} \\ k_{j1}x_{j1}x_{j3} \\ -k_{j1}x_{j1}x_{j3} \end{pmatrix}$$
$$h_j(x) = \left[(\frac{F}{V}x_{(j-1)1})^T (\frac{F}{V}x_{(j-1)2})^T (\frac{F}{V}x_{(j-1)3})^T \right]^T, \quad j=2, 3, 4. \tag{10}$$

For further study, if the concentrations are not considered, the nominal subsystem of subsystem (8) can be defined as follows:

$$\dot{x}_j(t) = f_j(x_j(t)) + g_j(x_j(t))u_j(t), \quad j = 1, \ldots, 4. \tag{11}$$

Hypothesis 1 [32]: Assume the concentrations of three ions Fe^{2+}, Fe^{3+}, and H^+ in subsystem j are within the boundary given in TABLE 1, and satisfy the following structure:

$$h_j(x) = D_j(x)c_j(\zeta(x)) \tag{12}$$
$$c_j^T(\zeta(x))c_j(\zeta(x)) \le d_j^T(\zeta(x))d_j(\zeta(x)) \tag{13}$$

where $D_j(\cdot) \in \mathbb{R}^{n_j \times r_j}$ and ζ are function structures of concentrations, and there is $\zeta(0) = 0$. $c_j \in R^{r_j}$ is the uncertainty function of the concentrations, with $c_j(0) = 0$. $d_j(\cdot) \in \mathbb{R}^{r_j}$ is a known bounded function with $d_j(0) = 0$, j $= 1, \ldots, 4$.

For the nominal system (11), the cost function of subsystem j can be expressed as:

$$\bar{V}_j(x_j(t)) = \int_0^\infty (Q_j(x_j) + u_j^T R_j u_j)d\tau \tag{14}$$

where $Q_j(x_j)$ is a positive definite function and $R_j = R_j^T > 0$ is a square matrix.

The objective of optimal control for goethite process is: give an initial state x_{j0}, design an approximate optimal control strategy $u_j(t) = u_j^*(x)$ to make the local subsystem j

asymptotically stable and minimize the cost function (14). The optimization control problem can be described as:

$$u_j(t) \triangleq u_j^*(x) \triangleq \arg\min_{u_j} \bar{V}_j(x_{j0}) \tag{15}$$

III. DECENTRALIZED OPTIMAL CONTROL OF IRON-REMOVAL SYSTEM BASED ON REINFORCEMENT LEARNING

A. APPROXIMATE ITERATIVE ALGORITHM

From the former section, it is known that the cost function (14) cannot directly evaluate the coherent system. To solve that problem, the optimal guaranteed cost control problem of the original coherent system (8) is converted into the optimal feedback control problem of the nominal system (11), according to the idea in [17].

Lemma 1 [32]: Assume that there are a cost function $V_j(x)$, a bounded function $B_j(x)$ with $B_j(x) > 0$, and a control law $u_j(x)$, hence:

$$[\nabla V_j(x)]^T h_j(x) \le B_j(x) \tag{16}$$

$$Q_j(x_j) + u_j^T R_j u_j + (\nabla V_j(x))^T \times (f_j(x_j) + g_j(x_j)u_j) + B_j(x) = 0 \tag{17}$$

where $\nabla V_j(x)$ is the partial derivative of the cost function V of the subsystem j to the system state x, $\nabla V_j(x) \triangleq \partial V_j/\partial x$. Then a neighbourhood of the origin system (8) is locally asymptotically stable. Also, $\bar{V}_j(x_{j0}, u_j) \le V_j(x_0, u_j)$, where $V_j(x_0, u_j)$ is the modified cost function of nominal system (11) described as:

$$V_j(x_0, u_j) = \int_0^\infty (Q_j(x_j) + u_j^T R_j u_j + B_j(x))d\tau \tag{18}$$

To deal with interconnections, $B_j(x)$ is set to a specific form as:

$$B_j(x) = \frac{1}{4}(\nabla V_j^{(i+1)}(x))^T D_j(x) D_j^T(x) \nabla V_j^{(i+1)}(x) + d_j^T(\xi(x)) d_j(\xi(x)) \tag{19}$$

For the new cost function (18), the Hamiltonian equation of the nominal system (11) can be defined as:

$$H_j(x_j, u_j, \nabla V_j) \triangleq Q_j(x_j) + u_j^T R_j u_j + B_j(x) + (\nabla V_j)^T (f_j(x_j) + g_j(x_j)u_j) \tag{20}$$

Set the Hamilton function as $H_j(x_j, u_j, \nabla V_j^*) = 0$, the optimal control of the HJB equation can be obtained:

$$u_j^*(x) = -\frac{1}{2} R_j^{-1} g_j^T(x_j) \nabla V_j^*(x) \tag{21}$$

The modified HJB equation can be written as:

$$Q_j(x_j) + (u_j^*)^T R_j u_j^* + B_j(x) + (\nabla V_j)^T (f_j(x_j) + g_j(x_j)u_j^*) \frac{1}{4}(\nabla V_j^*(x))^T D_j(x) D_j^T(x) \nabla V_j^*(x) + d_j^T(\xi(x)) d_j(\xi(x)) = 0 \tag{22}$$

Therefore, the optimal guaranteed cost control problem of the original coherent system is transformed into the optimal feedback control problem of the nominal system. The optimal control strategy (21) depends on the solution of the HJB equation (22), and the equation can be successively approximated by the GHJB sequence as follow:

$$\left[\nabla V_j^{(i+1)}\right]^T \left[f_j(x_j) + g_j(x_j)u_j^{(i)}\right] + Q_j(x_j) + (u_j^{(i)})^T R_j u_j^{(i)} + \mathrm{d}_j^T(\xi(x))\mathrm{d}_j(\xi(x)) + \frac{1}{4}(\nabla V_j^{(i+1)}(x))^T D_j(x) D_j^T(x) \nabla V_j^{(i+1)}(x) = 0 \tag{23}$$

$$u_j^{(i)} = -\frac{1}{2} R_j^{-1} g_j^T(x_j) \nabla V_j^{(i)}(x) \tag{24}$$

Since the model of the actual goethite process is not completely accurate, R_j and $g_j(x_j)$ in equation (24) are often not precisely obtained. In order to obtain the optimal control strategy when the model is not accurate enough or the model is unknown, an approximation strategy iterative algorithm is proposed. In the strategy, the actual system data is used to learn the solution of the HJB equation through neural network learning. For that purpose, system (11) can be rewritten as:

$$\dot{x}_j = f_j(x_j) + g_j u_j^{(i)} + g_j(x_j)\left[u_j - u_j^{(i)}\right] \tag{25}$$

For system (25), the derivative of $V^{(i+1)}(x)$ with respect to time can be found as:

$$\frac{dV_j^{(i+1)}(x)}{dt} = \left[\nabla V_j^{(i+1)}\right]^T \left[f_j(x_j) + g_j(x_j)u_j^{(i)}\right] + \left[\nabla V_j^{(i+1)}\right]^T g_j(x_j)\left[u_j - u_j^{(i)}\right] \tag{26}$$

Using equations (23) and (24), equation (25) can be written as follows:

$$\begin{aligned}
&\frac{dV_j^{(i+1)}(x)}{dt} \\
&= -(u_j^{(i)})^T R_j u_j^{(i)} + 2\left[u_j^{(i+1)}\right]^T R_j \left[u_j^{(i)} - u_j\right] \\
&\quad -Q_j(x_j) - \frac{1}{4}(\nabla V_j^{(i+1)}(x))^T D_j(x) D_j^T(x) \nabla V_j^{(i+1)}(x) \\
&\quad - \mathrm{d}_j^T(\xi(x))\mathrm{d}_j(\xi(x)) \\
&= -Q_j(x_j) + 2\left[u_j^{(i+1)}\right]^T R_j \left[u_j^{(i)} - u_j\right] \\
&\quad - \mathrm{d}_j^T(\xi(x))\mathrm{d}_j(\xi(x)) - (u_j^{(i)})^T R_j u_j^{(i)} \\
&\quad -g_j^{-1}(x_j) R_j^T u_j^{(i+1)} D_j(x) D_j^T(x) g_j^{-1}(x_j) R_j^T u_j^{(i+1)} \\
&= 0
\end{aligned} \tag{27}$$

Both sides of (27) on the interval $[t, t + \Delta t]$ can be integrated as:

$$\begin{aligned}
&V_j^{(i+1)}(t) - V_j^{(i+1)}(t+\Delta t) + 2\int_t^{t+\Delta t} \left(u_j^{(i+1)}\right)^T R_j \left[u_j^{(i)} - u_j\right] d\tau \\
&= \int_t^{t+\Delta t} [g_j^{-1}(x_j) R_j^T u_j^{(i+1)} D_j(x) D_j^T(x) g_j^{-1}(x_j) R_j^T u_j^{(i+1)}] d\tau \\
&\quad + \int_t^{t+\Delta t} \left[\mathrm{d}_j^T(\xi(x))\mathrm{d}_j(\xi(x)) + Q_j(x_j) + (u_j^{(i)})^T R_j u_j^{(i)}\right] d\tau
\end{aligned} \tag{28}$$

where $V_j^{(i+1)}(x)$ and $u_j^{(i+1)}(x_j)$ are unknown functions and unknown vectors of subsystem j, respectively. The problem of solving the GHJB equation (23) for $V_j^{(i+1)}(x)$ is transformed to the problem of solving equation (28).

Lemma 2: Note $\lambda(x) \in R^m$, $b(x) \in R$ and $c \in R^m$, in which c is the variable. If $\forall c \neq 0$, there is $\lambda^T(x)c = b(x)$, then $\lambda(x) = 0$ and $b(x) = 0$. When c is given a fixed value that satisfies $c_0 \neq 0$, then there is $\lambda^T(x)c_0 = b(x)$. Summing up the above analysis, the equation in (29) can be obtained.

$$\lambda^T(x)(c - c_0) = 0 \quad \forall x, \forall c \neq 0. \tag{29}$$

Let $F(x, c) = \lambda^T(x)(c - c_0)$, then:

$$F(x, c) = 0\forall \mathrm{x}, \quad \forall \mathrm{c} \neq 0 \tag{30}$$

The partial derivative to c for equation (30) is:

$$\frac{\partial F(x, c)}{\partial c} = \frac{\partial(\lambda^T(x)(c - c_0))}{\partial c} = \lambda(x) = 0 \tag{31}$$

Hence, $b(x) = \lambda^T(x)c = 0$.

Theorem 1: Equation (28) is the equivalent of equations (23) and (24).

Proof: Rewrite equation (28) as:

$$\begin{aligned}
&\left[\nabla V_j^{(i+1)}\right]^T \left[f_j(x_j) + g_j(x_j)u_j^{(i)}\right] \\
&\quad + (\mathrm{d}_j^T(\xi(x))\mathrm{d}_j(\xi(x)) + Q_j(x_j) + (u_j^{(i)})^T R_j u_j^{(i)}) \\
&\quad + \frac{1}{4}(\nabla V_j^{(i+1)}(x))^T D_j(x) D_j^T(x) \nabla V_j^{(i+1)}(x) \\
&= -\left\{\left[\nabla V_j^{(i+1)}\right]^T g_j(x_j) + 2\left[u_j^{(i+1)}\right]^T R_j\right\}\left[u_j - u_j^{(i)}\right]
\end{aligned} \tag{32}$$

According to Lemma 2, there are:

$$\begin{aligned}
&\left[\nabla V_j^{(i+1)}\right]^T \left[f_j(x_j) + g_j(x_j)u_j^{(i)}\right] \\
&\quad + (d_j^T(\xi(x))d_j(\xi(x)) + Q_j(x_j) + (u_j^{(i)})^T R_j u_j^{(i)}) \\
&\quad + \frac{1}{4}(\nabla V_j^{(i+1)}(x))^T D_j(x) D_j^T(x) \nabla V_j^{(i+1)}(x) = 0
\end{aligned} \tag{33}$$

$$\left[\nabla V_j^{(i+1)}\right]^T g_j(x_j) + 2\left[u_j^{(i+1)}\right]^T R_j = 0 \tag{34}$$

By observing equations (33) and (34), it is easy to find that they are exactly the same as (23) and (24), respectively. Accordingly, the proof is completed.

B. ACTOR-CRITIC NEURAL NETWORK AND ITS PARAMETER SOLUTION

In order to solve Eq.(28) for $V_j^{(i+1)}(x)$ and $u_j^{(i+1)}(x)$, a method based on actor-critic neural network (NN) structure is adopted. Combining the advantages of both actor-only and critic-only, actor-critic neural network has low variance and continuous action. The actor neural network is used to approximate the cost function $V_j^{(i+1)}(x)$, and the critic neural network is used to approximate the control strategy $u_j^{(i+1)}(x)$. $\varphi_j(x) \triangleq [\varphi_{j,1}(x) \dots \varphi_{j,L_V}(x)]^T$ is the vector of linearly independent activation functions for critic NN, where $j_k = 1, \dots L_V$ is the number of hide layer neuron for critic NN. $\psi_j^l(x) \triangleq [\psi_{j,1}^l(x) \dots \psi_{j,L_u}^l(x)]^T$ is the vector of linearly independent activation functions for actor NN, where $k = 1, \dots L_u$ is the number of hide layer neuron for actor NN.

The outputs of the critic NN and the actor NN are given by:

$$\hat{V}_j^{(i)}(x) = \sum_{jk=1}^{L_V} \theta_{j,V,jk}^{(i)} \varphi_{j,jk}(x) = \varphi_j^T(x)\theta_{j,V}^{(i)} \tag{35}$$

$$\hat{u}_{j,l}^{(i)}(x) = \sum_{k=1}^{L_u} \theta_{j,u_l,k}^{(i)} \psi_{j,k}^l(x) = (\psi_j^T(x))^T \theta_{j,u_l}^{(i)} \tag{36}$$

$\forall i = 0, 1, 2, \dots, \theta_{j,V}^{(i)} \triangleq [\theta_{j,V,1}^{(i)} \dots \theta_{j,V,LV}^{(i)}]^T$ and $\theta_{j,u_l}^{(i)} \triangleq [\theta_{j,u_l,1}^{(i)} \dots \theta_{j,u_l,L_u}^{(i)}]^T$ are the weight vectors of critic and actor neural network, respectively.

According to [22], equation (36) can be rewritten as:

$$\begin{aligned}
\hat{u}_j^{(i)}(x) &= [\hat{u}_{j,1}^{(i)}(x) \dots \hat{u}_{j,m}^{(i)}(x)]^T \\
&= [(\psi_j^1(x))^T \theta_{j,u_1}^{(i)} \dots (\psi_j^m(x))^T \theta_{j,u_m}^{(i)}]^T
\end{aligned} \tag{37}$$

Define residuals as:

$$\begin{aligned}
&\sigma_j^{(i)}(x(t), u_j(t), x(t + \Delta t)) \\
&\triangleq [\varphi_j(x(t)) - \varphi_j(x(t + \Delta t))]^T \theta_{j,V}^{(i+1)} \\
&\quad + 2\int_t^{t+\Delta t} [u_j^{(i)}(x(\tau)) - u_j(\tau)]^T R_j u_j^{(i+1)}(x(\tau)) d\tau \\
&\quad - \int_t^{t+\Delta t} Q_j(x_j(\tau)) d\tau - \int_t^{t+\Delta t} \left[\mathrm{d}_j^T(\xi(x(\tau)))\mathrm{d}_j(\xi(x(\tau)))\right] d\tau \\
&\quad - \int_t^{t+\Delta t} u_j^{(i)} R_j u_j d\tau - \frac{1}{4}\int_t^{t+\Delta t} [(\nabla V_j^{(i+1)})^T O_j(x) \nabla V_j^{(i+1)}] d\tau \\
&= [\varphi_j(x(t)) - \varphi_j(x(t + \Delta t))]^T \theta_{j,V}^{(i+1)} - \int_t^{t+\Delta t} Q_j(x_j(\tau)) d\tau \\
&\quad - \int_t^{t+\Delta t} \left[\mathrm{d}_j^T(\xi(x(\tau)))\mathrm{d}_j(\xi(x(\tau)))\right] d\tau + 2\sum_{l_1=1}^{m}\sum_{l_2=1}^{m} r_{l_1,l_2} \\
&\quad \times \int_t^{t+\Delta t} [\psi_j^{l_1}(x(\tau))^T \theta_{j,u_{l_1}}^{(i)} - u_{j,l_1}(\tau)]^T (\psi_j^{l_2}(x(\tau)))^T \theta_{j,u_{l_2}}^{(i+1)} d\tau \\
&\quad - \int_t^{t+\Delta t} [g_j^{-1}(x_j) R_j^T u_j^{(i+1)} D_j(x) D_j^T(x) g_j^{-1}(x_j) R_j^T u_j^{(i+1)}] d\tau \\
&\quad - \sum_{l_1=1}^{m}\sum_{l_2=1}^{m} r_{l_1,l_2} \times \int_t^{t+\Delta t} \theta_{j,u_{l_1}}^{(i)} \psi_j^{l_1}(x(\tau))(\psi_j^{l_2}(x(\tau)))^T \theta_{j,u_{l_2}}^{(i)} d\tau
\end{aligned} \tag{38}$$

In order to eliminate the dependence of the interconnecttions on $g_j(x_j)$, based on the actor-critic neural network, a structural neural network is constructed as the follows:

$$\begin{aligned}
\hat{q}_j^{(i)}(x) &= g_j^{-1}(x_j) R_j^T u_j^{(i)} \\
&= \sum_{s=1}^{L_A} \theta_{j,q,s}^{(i)} \phi_{j,s}(x) = \phi_j^T(x) \theta_{j,q}^{(i)}
\end{aligned} \tag{39}$$

where $\forall i = 0, 1, 2 \dots, \theta_{j,q}^{(i)} \triangleq [\theta_{j,q,1}^{(i)} \dots \theta_{j,q,LA}^{(i)}]^T$ is the vector of linearly independent activation functions for the structural

neural network, then the residual can be expressed as:

$$\begin{aligned}
&\sigma_j^{(i)}(x(t), u_j(t), x(t+\Delta t)) \\
&\triangleq -\int_t^{t+\Delta t} Q_j(x_j(\tau))d\tau \\
&\quad + [\varphi_j(x(t)) - \varphi_j(x(t+\Delta t))]^T \theta_{j,V}^{(i+1)} + 2\sum_{l_1=1}^{m}\sum_{l_2=1}^{m} r_{l_1,l_2} \\
&\quad \times \int_t^{t+\Delta t} [\psi_j^{l_1}(x(\tau))^T \theta_{j,u_{l_1}}^{(i)} - u_{j,l_1}(\tau)]^T (\psi_j^{l_2}(x(\tau)))^T \theta_{j,u_{l_2}}^{(i+1)} d\tau \\
&\quad - \sum_{l_1=1}^{m}\sum_{l_1=1}^{m} r_{l_1,l_2} \times \int_t^{t+\Delta t} \theta_{j,u_{l_1}}^{(i)} \psi_j^{l_1}(x(\tau))(\psi_j^{l_2}(x(\tau)))^T \theta_{j,u_{l_2}}^{(i)} d\tau \\
&\quad - \int_t^{t+\Delta t} \left[\mathrm{d}^T(\xi(x(\tau)))\mathrm{d}(\xi(x(\tau))) \right] d\tau - \sum_{s_1=1}^{w}\sum_{s_2=1}^{w} O_{j,s_1,s_2}(x) \\
&\quad \times \int_t^{t+\Delta t} [(\theta_{j,q_{s_1}}^{(i+1)})^T \phi_j^{s_1}(x(t))(\phi_j^{s_2}(x(t)))^T \theta_{j,q_{s_2}}^{(i+1)}] d\tau \qquad (40)
\end{aligned}$$

where $O_j(x) = D_j(x)D_j^T(x)$.

In order to solve the unknown parameters in equation (38), the parameters of the neural network are first obtained by solving the following objective function:

$$\min J(\theta_j) = \frac{1}{2}(\sigma_j^{(i)})^2 \qquad (41)$$

where $\sigma_j^{(i)}$ is the residual defined in equation (40), and $\theta_j = [\theta_{j,V}, \theta_{j,u}, \theta_{j,\rho}]$ are the weight vector to be identified by the critic neural network and the actor neural network.

It is necessary for the commonly used methods such as the minimum residual method [22] to satisfy the linear relationship between the residuals and parameters defined by the HJB equation, when it comes to solving optimization problems. However, there are many optimization variables in the optimal control problem of goethite process, and it is difficult to meet the constraint where the residuals and parameters must have a linear relationship. In considering the above problems, an intelligent global optimization, algorithm-State Transition Algorithm [35] (STA), is used to optimize the solution parameters.

The parameter to be identified in equation (41) is encoded as the state $\tilde{x}$, and the process of parameter optimization by the state transition algorithm can be expressed as follows:

$$\begin{cases} \tilde{x}_{k+1} = \tilde{A}_k \tilde{x}_k + \tilde{B}_k \tilde{u}_k \\ \tilde{y}_{k+1} = \tilde{f}(\tilde{x}_{k+1}) \end{cases} \qquad (42)$$

where $\tilde{x} \in \boldsymbol{R}^{n'}$ represents the state of the parametric solution, k is the number of iteration steps, and $\tilde{y}_k$ represents the fitness of the state $\tilde{x}_k$. $\tilde{A}_k$ and $\tilde{B}_k$ indicate the state transition matrices at each update of the solution state. $\tilde{u}_k$ is a function related to the current state $\tilde{x}_k$ and historical state $\tilde{x}_{k-1}$, while $\tilde{f}(\cdot)$ is regarded as the fitness function corresponding to the state $\tilde{x}_k$. The state transition algorithm generates random iterative solutions through four operators: a rotation transformation operator, a translation transformation operator, a telescopic transformation operator, and an axis search operator.

1) ROTATION TRANSFORM OPERATOR

$$\tilde{x}(k+1) = \left(I_{n'} + \alpha' \frac{1}{n' ||\tilde{x}(k)||_2} R_r \right) \tilde{x}(k) \qquad (43)$$

where α' is the rotation transformation operator of the STA, and it usually takes a positive integer; n' is the dimension of the solution state. $R_r \in \boldsymbol{R}^{n' \times n'}$ obeys the uniform distribution of [-1,1]. The rotation transformation of the state $\tilde{x}(k)$ is performed in the hypersphere with its current value as the center and the rotation operator α' as the radius.

2) TRANSLATION OPERATOR

$$\tilde{x}(k+1) = \tilde{x}(k) + \beta' R_t \frac{\tilde{x}(k) - \tilde{x}(k-1)}{\|\tilde{x}(k) - \tilde{x}(k-1)\|_2} \qquad (44)$$

where β' is a positive integer, which is the translation operator of STA. $R_t \in \boldsymbol{R}^{n' \times n'}$ obeys the uniform distribution among [0,1]. The translation of state $\tilde{x}(k)$ is performed in a gradient direction of $\tilde{x}(k)$ to $\tilde{x}(k-1)$ with a maximum step size of β'.

3) SCALING OPERATOR

$$\tilde{x}(k+1) = \tilde{x}(k) + \gamma' R_e \tilde{x}(k) \qquad (45)$$

where γ' is a normal number, which is a scaling operator of STA; $R_e \in R^{n' \times n'}$ is a diagonal matrix obeying a Gaussian distribution. The scaling operator can be optimized across the entire search space.

4) AXIS SEARCH OPERATOR

$$\tilde{x}(k+1) = \tilde{x}(k) + \delta' R_a \tilde{x}(k) \qquad (46)$$

where δ' is a coordinate search operator of the STA, and its value is a positive integer. $R_a \in \boldsymbol{R}^{n' \times n'}$ is a diagonal sparse matrix, which has only non-zero elements at a random position, and the elements obey Gaussian distribution.

After the parameters of the critic and actor of the neural network are obtained, the final control strategy u can be solved according to the weights, basis functions and the current state of each subsystem to minimize the cost function. Combining the approximate iterative algorithm and STA algorithm proposed in this paper, the steps to solve the control strategy of the associated iron-removal system are as follows:

Step 1: Under the given initial stable controller and initial state, collect sample data of ion concentration at the reactor outlet for a period of time;

Step 2: Select the basis functions for critic NN, actor NN, and the structure NN, then encode the weights to be identified as the states in the STA algorithm;

Step 3: Select the state of a set of solutions that make the fitness function $\tilde{f}(\cdot)$ (that is, the objective function (41)) reach the minimum value from the current population. Record it as *best* and the corresponding fitness is f_{best}, then copy *best* as the number of individuals with *SE*. The population is recorded as $\tilde{x}(k)$, and a new population is obtained by performing a scaling transformation according to equation (45).

The optimal individual in the population after the scaling transformation is new_{best}, and the corresponding fitness is g_{best}. If g_{best} is less than f_{best}, then use equation (44). Perform a translation transformation on the individual new_{best}, and update the *best* and f_{best} after the translation transformation.

Step 4: Copy the best into a group with S individuals, and then perform rotation transformation according to equation (43) to obtain a new population. Select the best individual new_{best} in the population after the rotation transformation, and the corresponding fitness is g_{best}; if g_{best} is less than f_{best}, perform translation transformation according to equation (44), and update the *best* and f_{best} after the translation transformation.

Step 5: Copy *best* into a group, and *SE* is the number of group. Then perform coordinate search and transformation according to equation (46). Select the solution state of the optimal solution among all individuals after transformation as new_{best}, and the corresponding fitness as g_{best}; if g_{best} is less than f_{best}, perform translation transformation according to equation (44), and update the *best* and f_{best} after the translation transformation.

Step 6: Repeat steps 3-5. When the given termination condition $\left\|\theta_j^{(i)} - \theta_j^{(i+1)}\right\| \leq \zeta$ is satisfied or the number of iterations is greater than the given number of times, find a set of parameter vectors that minimizes the objective function as parameters for critic NN and actor NN;

Step 7: According to the weights and basis functions of the neural network obtained by optimization and the current status of each subsystem collected by the system, the real-time optimization control strategy u_j for each subsystem is solved according to equation (41).

IV. SIMULATION

Assume that the goethite iron-removal system satisfies the affine nonlinear structure of formula (25). And to verify the proposed decentralized optimization control method of the coherent iron-removal system, simulation experiments are carried out by actual data of the goethite iron removal process. According to these data, the flow rates of the reactors are $F_b \in [110\text{m}^3/\text{h},\ 120\text{m}^3/\text{h}]$, $F \in [120\text{m}^3/\text{h},\ 150\text{m}^3/\text{h}]$, and the effective volume of the reactor is $V = 300\text{m}^3$. Therefore, for the 1# reactor, the parameters of the concentrations can be set as:

$$D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad d_1(\xi(x)) = \begin{pmatrix} 0.4x_{41} \\ 0.4x_{42} \\ 0.4x_{43} \end{pmatrix}$$

Similarly, the parameters of 2#-4# reactor concentrations can be set as:

$$D_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad D_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad D_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$d_2 = \begin{pmatrix} 0.5x_{11} \\ 0.5x_{12} \\ 0.5x_{13} \end{pmatrix} \quad d_3 = \begin{pmatrix} 0.5x_{21} \\ 0.5x_{22} \\ 0.5x_{23} \end{pmatrix} \quad d_4 = \begin{pmatrix} 0.5x_{31} \\ 0.5x_{32} \\ 0.5x_{33} \end{pmatrix}$$

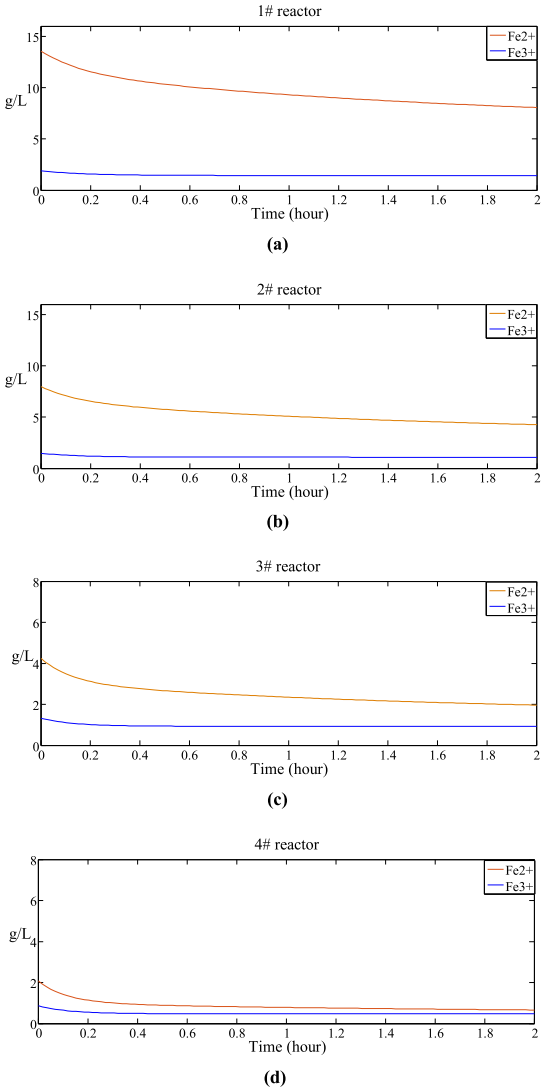

FIGURE 2. Fe^{2+} and Fe^{3+} concentrations in 1#-4# reactor.

Taking the 1# reactor as an example, according to the initial value of the actual system setting state $x_{10} = [14\ 1.7\ 3.6]$ and the initial controller $u_{10} = 51.6164$ according to the established model (10). The parameters $k_{j1} = 1.4623$, $k_{j2} = 1.6693$, $k_{j3} = 0.2802$ in the model are identified by the least squares method. With the choice of $Q_1 = x_1^T x_1$ and $R_1 = I$ in the cost function (18), the following functions are selected as the basis functions of the critic network: $\varphi_1(x) = [x_{1,1}, x_{1,2}, x_{1,3}, x_{1,1}x_{1,2}, x_{1,1}x_{1,3}, x_{1,2}x_{1,3}, x_{1,1}^2, x_{1,2}^2, x_{1,3}^2, x_{4,1}, x_{4,2}, x_{4,3}]$. The selection rules of $\varphi_2(x)$, $\varphi_3(x)$, $\varphi_4(x)$ are the same as $\varphi_1(x)$, and the number of hidden m^3/h layer nodes $L_V = 12$. Similarly, the following function is selected as the basis function for actor network:

$\psi_1(x) = [x_{1,1}, x_{1,2}, x_{1,3}, x_{1,1}^2, x_{1,2}^2, x_{1,3}^2, x_{1,1}^3, x_{1,2}^3, x_{1,3}^3, x_{1,1}^4, x_{1,2}^4, x_{1,3}^4]$. And $\psi_2(x)$, $\psi_3(x)$, $\psi_4(x)$ is the same as $\psi_1(x)$. The number of hidden layer nodes is $L_u = 12$. Similarly, the following is selected as the basis function of the structural neural network: $\phi_1(x) = [x_{1,1}, x_{1,2}, x_{1,3}, x_{1,1}^2,$

TABLE 3. Oxygen consumptions in iron-removal process 1#-4# reactor.

	1#	2#	3#	4#
Total oxygen reduction in two hours (m^3/h)	13.73	10.86	8.12	8.91
Oxygen reduction percentage	13.93%	11.47%	8.83%	9.55%

TABLE 4. pH change in iron-removal process 1#-4# reactor.

	1#	2#	3#	4#
Mean pH by the initial control	3.5142	3.4464	3.4714	3.4155
Mean pH by the optimal control	3.4077	3.4006	3.3887	3.3861
Variance of pH by the initial control	0.2280	0.2694	0.2911	0.2863
Variance of pH by the optimal control	0.0938	0.0806	0.0825	0.0624

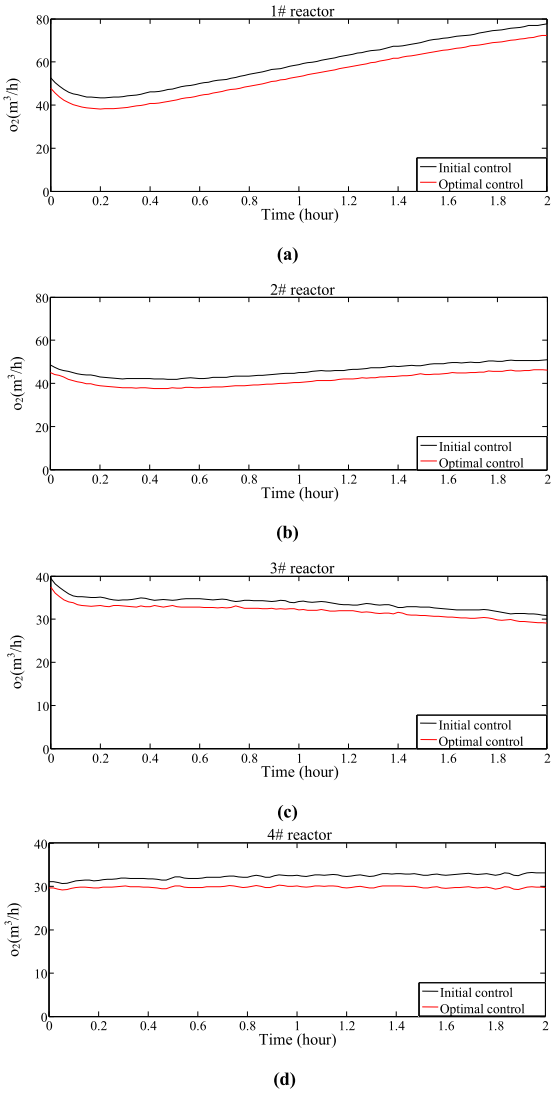

FIGURE 3. Oxygen consumptions in 1#-4# reactor.

$x^2_{1,2}, x^2_{1,3}, x^3_{1,1}, x^3_{1,2}, x^3_{1,3}, x^4_{1,1}, x^4_{1,2}, x^4_{1,3}]$. The selection rules of $\phi_2(x)$, $\phi_3(x)$, $\phi_4(x)$ are the same as $\phi_1(x)$. The initial controller u_0 is obtained from actual experience. Set an initial weight vector according to the initial state and the initial control u_0: $\theta^{(0)}_{1,u} = [-11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91\ -11.91]$.

Setting the value of the convergence criterion $\xi = 10^{-4}$, it is found that the critic and actor NN weight vectors converge respectively to θ^*_V, θ^*_u and θ^*_q. For the first reactor critic network, the weight vector converges to $\theta_{1,\mathrm{v}} = [-14.9446\ -0.1209\ 14.9999\ -2.4667\ 1.7839\ -3.9240\ 0.8326\ -7.3507\ -9.6520\ 0.0006\ 14.9481\ 2.3172]$, the actor network's weight vector converges to $\theta_{1,u} = [-1.4835\ 9.3696\ -0.1576\ -8.2513\ 5.7104\ 14.8071\ 1.5260\ 4.1283\ 11.2396\ -14.9668\ -12.8070\ 14.9220]$ and the structural neural network's weight vector converges to $\theta_{1,q} = [1.3457\ 15.0001\ 0.8082\ 15.0021\ 0.0107\ -0.1543\ -7.6880\ 8.3908 - 0.0680\ 14.9888\ 3.3743]$.

Similarly, in 2#-4# reactors, the initial values $x_{20} = [7.92\ 1.63\ 3.37]$, $x_{30} = [4.21\ 1.59\ 3.41]$ and $x_{30} = [2.03\ 1.52\ 3.28]$ are set according to the exit value of the previous reactor. The initial controller $u_{20} = 43.8154$, $u_{20} = 39.9451$, $u_{20} = 31.6164$.

Taking the 1# reactor as an example, the control strategy (33) is used for closed-loop simulation. Figures 2 (a), (b), (c) (d) are the Fe^{2+}, Fe^{3+} concentrations in 1#-4# reactor, respectively. Figures 3 (a), (b), (c) and (d) are the compared results of oxygen consumption between the proposed optimal control and initial control in the 1#-4# reactor, respectively. Figures 4 (a), (b), (c) and (d) are the changes of the pH value in the 1#-4# reactor within two hours, respectively. It can be seen from Figure 2 that the Fe^{2+} ion concentration in the solution to be treated is reduced from 14g/L to 0.4g/L, while the Fe^{3+} ions are reduced from 1.7g/L to 0.8g/L, Changes in the concentration of these ions all meet the process technical requirements. The fluctuations of the Fe^{2+} and Fe^{3+} concentrations in the reactors are small, which avoids the formation of some by-products and ensures the smoothness of the goethite process. It can be seen from Figure 3 that the oxygen consumptions of the proposed optimal control have been significantly reduced compared to the initial control. The results of the oxygen consumption comparisons are shown in Table 1. Compared with the initial controller, the proposed optimal control reduced oxygen consumptions of the 1#-4# reactors by 13.73m^3/h, 10.86m^3/h, 8.12m^3/h, 8.91m^3/h respectively in two hours. The results show that the proposed control method is resources-saving. Table 2 shows the detailed comparison results of pH values, which indicates that compared to the initial control, the proposed optimal

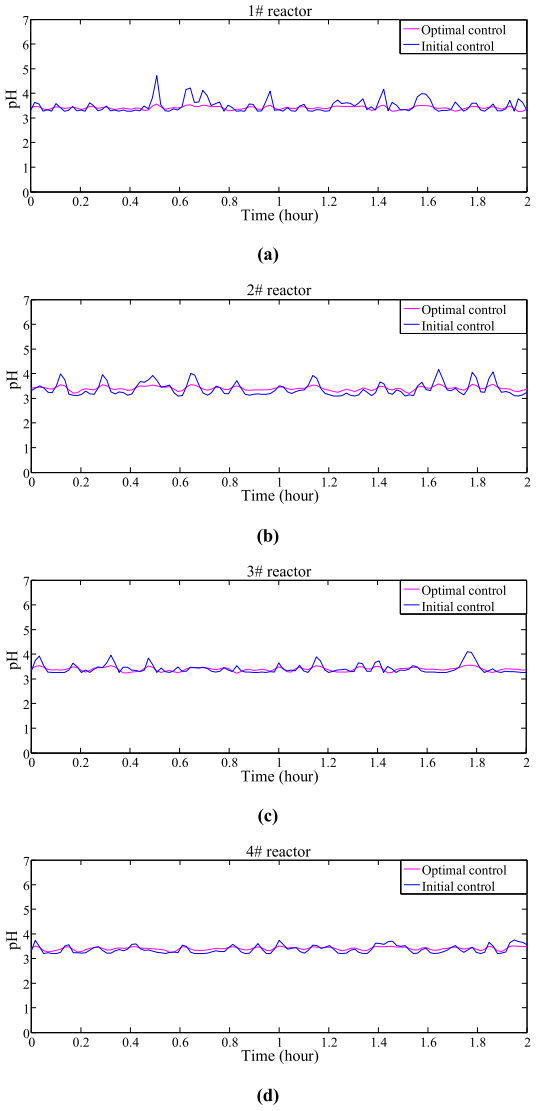

FIGURE 4. pH value of the solution in 1#-4# reactor.

control leads to smaller pH value fluctuations of the solution in reactor.

V. CONCLUSION

This paper proposes an off-policy optimal control method based on reinforcement learning for the associated Iron-Removal system. A bounded function is introduced to define the maximum impact of the associated system on the subsystem cost function. The bound function extends the cost function of the nominal system, and optimizes the new cost function to ensure that the cost function of the associated system is not higher than the nominal system cost function, thereby obtaining approximately optimal control. Taking advantages of the large amount of data obtained in goethite iron-removal process, the weight of neural network learned offline, and the strategy of solving optimal control online, this method provides convenience for practical operation of industries. Based on the actor-critic structure, a new neural network is introduced to approximate a part of the unknown system structure. In this way, we extended the optimal control method in [22] to the coherent system. And this method relaxes the constraints between parameters and residuals. According to the actual industrial data from the simulation experiment, the two ion concentrations and pH values in the goethite iron-removal process are strictly controlled within the range required by the technological requirement, and the ion fluctuation is less than that under the initial control, which proves the effectiveness of the proposed off-policy optimal control method.

REFERENCES

[1] J. O. Claassen, E. H. O. Meyer, J. Rennie, and R. F. Sandenbergh, "Iron precipitation from zinc-rich solutions: Defining the zincor process," *Hydrometallurgy*, vol. 67, nos. 1–3, pp. 87–108, Dec. 2002.

[2] M. R. C. Ismael and J. M. R. Carvalho, "Iron recovery from sulphate leach liquors in zinc hydrometallurgy," *Minerals Eng.*, vol. 16, no. 1, pp. 31–39, Jan. 2003.

[3] M. Shahvali, N. Pariz, and M. Akbariyan, "Distributed finite-time control for arbitrary switched nonlinear multi-agent systems: An observer-based approach," *Nonlinear Dyn.*, vol. 94, no. 3, pp. 2127–2142, Nov. 2018.

[4] N. Chen, J. Dai, X. Zhou, Q. Yang, and W. Gui, "Distributed model predictive control of iron precipitation process by goethite based on dual iterative method," *Int. J. Control, Autom. Syst.*, vol. 17, no. 5, pp. 1233–1245, May 2019.

[5] S. Xie, Y. Xie, H. Ying, W. Gui, and C. Yang, "A hybrid control strategy for real-time control of the iron removal process of the zinc hydrometallurgy plants," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5278–5288, Dec. 2018.

[6] J. Han, C. Yang, X. Zhou, and W. Gui, "Dynamic multi-objective optimization arising in iron precipitation of zinc hydrometallurgy," *Hydrometallurgy*, vol. 173, pp. 134–148, Nov. 2017.

[7] B. Sun, C. Yang, H. Zhu, Y. Li, and W. Gui, "Modeling, optimization, and control of solution purification process in zinc hydrometallurgy," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 2, pp. 564–576, Mar. 2018.

[8] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Hamiltonian-driven hybrid adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Jan. 13, 2020, doi: 10.1109/TSMC.2019.2962103.

[9] M. Shahvali, A. Azarbahram, M.-B. Naghibi-Sistani, and J. Askari, "Bipartite consensus control for fractional-order nonlinear multi-agent systems: An output constraint approach," *Neurocomputing*, vol. 397, pp. 212–223, Jul. 2020.

[10] M. Shahvali, M.-B. Naghibi-Sistani, and H. Modares, "Distributed consensus control for a network of incommensurate fractional-order systems," *IEEE Control Syst. Lett.*, vol. 3, no. 2, pp. 481–486, Apr. 2019.

[11] 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, May 2005.

[12] G. N. Saridis and C.-S.-G. Lee, "An approximation theory of optimal control for trainable manipulators," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 3, pp. 152–159, Mar. 1979.

[13] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, Dec. 1997.

[14] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[15] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Safe intermittent reinforcement learning with static and dynamic event generators," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 10, 2020, doi: 10.1109/TNNLS.2020.2967871.

[16] D. Wang, "Robust policy learning control of nonlinear plants with case studies for a power system application," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1733–1741, Mar. 2020.

[17] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 65–76, Jan. 2015.

[18] J. Li, T. Chai, F. L. Lewis, Z. Ding, and Y. Jiang, "Off-policy interleaved *Q*-learning: Optimal control for affine nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1308–1320, May 2019.
[19] J. Li, J. Ding, T. Chai, and F. L. Lewis, "Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes," *IEEE Trans. Cybern.*, early access, Nov. 19, 2019, doi: 10.1109/TCYB.2019.2950262.
[20] F. L. Lewis and K. G. Vamvoudakis, "Optimal adaptive control for unknown systems using output feedback by reinforcement learning methods," in *Proc. IEEE Int. Conf. Control Autom.*, Jun. 2010, pp. 9–11.
[21] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for H_∞ control with full-state constraints and input saturation," *J. Franklin Inst.*, vol. 357, no. 6, pp. 3316–3344, Apr. 2020.
[22] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, vol. 50, no. 12, pp. 3281–3290, Dec. 2014.
[23] P. Yan, D. Wang, H. Li, and D. Liu, "Error bound analysis of *Q*-function for discounted optimal control problems with policy iteration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1207–1216, Jul. 2017.
[24] B. Luo, H.-N. Wu, and H.-X. Li, "Adaptive optimal control of highly dissipative nonlinear spatially distributed processes with neuro-dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 684–696, Apr. 2015.
[25] Y. Zhu, D. Zhao, X. Yang, and Q. Zhang, "Policy iteration for H_∞ optimal control of polynomial nonlinear systems via sum of squares programming," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 500–509, Feb. 2018.
[26] S. He, H. Fang, M. Zhang, F. Liu, X. Luan, and Z. Ding, "Online policy iterative-based H_∞ optimization algorithm for a class of nonlinear systems," *Inf. Sci.*, vol. 495, pp. 1–13, Aug. 2019.
[27] S. He, H. Fang, M. Zhang, F. Liu, and Z. Ding, "Adaptive optimal control for a class of nonlinear systems: The online policy iteration approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 549–558, Feb. 2020.
[28] L. Dong, X. Zhong, C. Sun, and H. He, "Event-triggered adaptive dynamic programming for continuous-time systems with control constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1941–1952, Aug. 2017.
[29] X. Lu, B. Kiumarsi, T. Chai, Y. Jiang, and F. L. Lewis, "Operational control of mineral grinding processes using adaptive dynamic programming and reference governor," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2210–2221, Apr. 2019.
[30] M. Maldonado, D. Sbarbaro, and E. Lizama, "Optimal control of a rougher flotation process based on dynamic programming," *Minerals Eng.*, vol. 20, no. 3, pp. 221–232, Mar. 2007.
[31] D. Liu, D. Wang, and H. Li, "Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 418–428, Feb. 2014.
[32] D. Wang, D. Liu, C. Mu, and H. Ma, "Decentralized guaranteed cost control of interconnected systems with uncertainties: A learning-based optimal control strategy," *Neurocomputing*, vol. 35, no. 6, pp. 297–301, 2016.
[33] L. Yu and J. Chu, "An LMI approach to guaranteed cost control of linear uncertain time-delay systems," *Automatica*, vol. 35, no. 6, pp. 1155–1159, Jun. 1999.
[34] L. Yu, J.-M. Xu, and Q.-L. Han, "Optimal guaranteed cost control of linear uncertain systems with input constraints," in *Proc. 5th World Congr. Intell. Control Autom. (WCICA)*, 2004, pp. 15–19.
[35] X. Zhou, C. Yang, and W. Gui, "State transition algorithm," *J. Ind. Manage. Optim.*, vol. 8, no. 4, pp. 1039–1056, 2012.