IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2021, accepted October 12, 2021, date of publication October 15, 2021, date of current version October 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120835

Adaptive Neural Network Optimized Control Using Reinforcement Learning of Critic-Actor Architecture for a Class of Non-Affine Nonlinear Systems

XUE YANG[1], BIN LI[1], AND GUOXING WEN[1,2]
[1]School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[2]College of Science, Binzhou University, Binzhou, Shandong 256600, China

Corresponding author: Bin Li (ribbenlee@126.com)

This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2020MF097, in part by the National Natural Science Foundation of China under Grant 62073045 and Grant 61973185, and in part by the Development Plan of Young Innovation Team in the colleges and universities of Shandong Province under Grant 2019KJN011.

ABSTRACT In this article, an optimized tracking control using critic-actor reinforcement learning (RL) strategy is investigated for a class of non-affine nonlinear continuous-time systems. Since the non-affine system is with the implicit control input in dynamic equation, it is a more general modeling form than the affine case, hence this also makes the optimized control more challenging and rewarding. However, most existing RL-based optimal controllers are very complex in algorithm because their actor and critic training laws obtained by implementing gradient descent on the square of Bellman residual error, which equals to the approximation of Hamilton-Jacobi-Bellman (HJB) equation, hence these methods are difficult to be extended to non-affine systems. In this optimized control, the RL algorithm is produced from implementing gradient descent to a simple positive-definite function, which is derived from HJB equation's partial derivative. As a result, the proposed control algorithm can be significantly simple so as to alleviate the computational burden. Finally, a typical numerical simulation is carried out, and the results also further confirm effectiveness of the proposed control scheme.

INDEX TERMS Non-affine nonlinear system, optimal control, reinforcement learning (RL), neural network (NN), Lyapunov function.

I. INTRODUCTION

In control community, non-affine system control has always played an important and key role because most practical engineering must be modeled in non-affine dynamic form. Unlike affine system that has the explicit control input, non-affine system's control is implicit, so that it has not the concept of control gain and control direction. As a result, studying non-affine control system becomes very challenging and rewarding [1]. Several mathematical tools are available to help find an equivalent affine system, such as mean value theorem, implicit function theorem, and Taylor serious expansion, they can be used to convert the system into affine form [2]–[4]. Especially, in [4], the adaptive NN controller is derived by first transforming the non-affine single-input single-output (SISO) nonlinear system to affine form via Taylor series expansion.

In the recent decades, optimal control has always been a hot and attractive academical topic in control community, especially to optimal control of nonlinear systems [5]–[8]. In general, nonlinear optimal control problem refers to solve a nonlinear partial differential equation, regarded as Hamilton-Jacobi-Bellman (HJB) equation [9]. Due to inherent non-linearity of the equation, its analytical solution is obtained difficultly or even impossibly. To address this challenge, Bellman proposed a technique that is the famous dynamic programming (DP) method [10]. However, the technique has an inevitable disadvantage, of which the calculation amount will increase exponentially as the increasing of system dimension, and thus it will result difficult

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

application in practice. In order to address the difficulty of DP method so that nonlinear optimal control can be effectively achieved, Werbos developed an adaptive algorithm by taking the advantage of NN approximators, which was called approximate/adaptive dynamic programming (ADP) [11]. Up to now, ADP method has received the increasing attention and spawned many other schemes, such as adaptive critic design [12], [13], neural dynamic programming [14] and the like. Additionally, in the reference [15], Liu *et al.* proposed a strategic of iterative ADP to address the infinite level optimal control problem of nonlinear systems. In the reference [16], a complex ADP approach was developed to solve the infinite-horizon optimal problem of complex-valued nonlinear systems.

In fact, ADP can be regarded as a class of reinforcement learning (RL) [17]. RL is a machine learning strategy that modifies agent behavior based on the response from environment [18]. In general, a common structure of RL is the critic-actor architecture, in which the critic is to evaluate the control performance according to the interaction with their environment and return the feedback for the actor, and the actor is to execute these continuous improving control operations. Since RL enables an agent to learn autonomously according to their own experience [19]–[21], it is an universal strategy in the nonlinear optimal control [22]–[24]. In [22], to solve the infinite-horizon optimal control problem, Yang *et al.* developed an adaptive optimal control strategy by using the RL of identifier-critic architecture. In [23], by using the NN-approximator-based RL, Wen *et al.* proposed a decentralized optimized formation control for a class of nonlinear multi-agent systems, and a significant breakthrough in the work is that two common requirements of known dynamic and persistence excitation are removed. In [24], the RL optimized control was extended to stochastic nonlinear system.

Because neural network (NN) and fuzzy logic system (FLS) are the effective approximators [25]–[27], some adaptive nonlinear approaches based on FLS or NN were proposed in recent years [28]–[32]. By using NN to estimate the solution of HJB equation, the optimal control based on RL of nonlinear systems was further developed, and many outstanding achievements have been made recently [33]–[35]. In [33] and [34], to optimal control of nonlinear strict feedback system, the new technique, optimized backstepping (OB), was proposed first time. Its basic thought is to design the actual control and all virtual controls as the optimal solution of the corresponding backstepping step, so that the overall system control can be optimized. In [35], OB technique was applied for surface vessel control. But the above optimization control methods requires complete system knowledge in the RL training. In fact, some systems are often with unknown dynamics. To solve this problem, many highlighted approaches have been presented, such as [6], [23], [36]. In [36], an observer-based optimal control scheme was developed, and thus unknown dynamics can be compensated by the adaptive observer. In the references [6], [23], an optimal formation control of nonlinear multi-agent system was addressed, the identifier technique was employed to overcome the difficulty of unknown dynamic.

Inspired by the above discussions, an optimal control using RL strategy is presented for a class of non-affine continuous-time nonlinear systems in this article. The primary contributions in the work can be summarized as follows.

1) The optimized control approach is developed for a class of non-affine nonlinear systems, and it is a significant extension in optimal control area.

2) The optimized control is significantly simple compared with the existing methods, so that it can be well performed for engineering.

3) The optimized control is easy to be implemented and applied, because it can release the condition of persistence excitation required for most existing optimal control.

II. PROBLEM FORMULATION

Consider the following non-affine nonlinear continuous-time system, which is a stabilizable system [33]:

$$\dot{x}(t) = F(u, x) \tag{1}$$

where $x(t) \in R^m$ and $u \in R^m$ are, respectively, the system state and control input, $F(u, x) \in R^m$ with $F(0, 0) = 0_m$ is the unknown nonlinear vector-value function. The term $F(u, x)$ is assumed to be Lipschitz continuous on the set Ω containing origin so that the solution of system (1) is unique for any control u and bounded initial value $x(0)$.

Since the control u is implicitly contained in the dynamic function $F(u, x)$, the control cannot be constructed via direct seeking for help from the system (1). In order to overcome the difficulty, Taylor series expansion is implemented so that the relation between control and dynamic can become explicit [37]:

$$F(u, x) = F(u^0, x) + \left.\frac{\partial F(u, x)}{\partial u^T}\right|_{u=u^0} (u - u^0) + \Delta \tag{2}$$

where $\Delta \in R^m$ denotes the infinitesimal term, which can be limited by a constant μ as $0 \leq \|\Delta\| \leq \mu$, and $u^0(x) \in R^m$ is an unknown smooth function. Furthermore, by choosing $u^0 = 0$, equation (2) is expressed as

$$F(u, x) = F(0, x) + \left.\frac{\partial F(u, x)}{\partial u^T}\right|_{u=0} u + \Delta. \tag{3}$$

Insert (3) into system dynamics (1), it results in

$$\dot{x}(t) = f(x) + g(x)u + \Delta \tag{4}$$

where $g(x) = \left.\frac{\partial F(u, x)}{\partial u^T}\right|_{u=0} \in R^{m\times m}$ and $f(x) = F(0, x) \in R^m$.

Assumption 1 ([37], [38]): The matrix $g(x)$ in system (4) is non-singular and bounded, i.e., it is an invertible matrix and there exist two constants $\bar{\xi} > \underline{\xi} > 0$ such that $\bar{\xi} > \|g(x)\| > \underline{\xi}$. As a result, it implies that the matrix $g(x)$ is either strictly positive or strictly negative. Without losing of generality, we assume $\bar{\xi} > g(x) > \underline{\xi}$.

The desired tracking trajectory is denoted by $y(t) \in R^m$, then define the tracking error as $z(t) = x(t) - y(t)$. From (4), we obtain the following equation:

$$\dot{z}(t) = f(x) + g(x)u + \Delta - \dot{y}(t). \tag{5}$$

Assumption 2 ([3], [39]): The reference tracking trajectory $y(t)$ and its derivative $\dot{y}(t)$ are assumed to be bounded.

Let us introduce the performance index as follows

$$J(z(0)) = \int_0^\infty r(z(s), u(z))ds, \tag{6}$$

where $r(z, u) = z^T(t)z(t) + u^T u$ is the local cost function.

Definition 1 [9]: A control policy u associated with (1) is admissible on Ω, that is denoted by $u \in \Phi(\Omega)$, if u is continuous, and $u(0) = 0$, and stabilizes (1), and makes the performance index (6) finite on Ω.

Optimal Control: An admissible control $u \in \Phi(\Omega)$ for the system (1) is said to be optimal one if it can minimize the performance index (6).

According to (6), define the performance index function as

$$J(z(t)) = \int_t^\infty r(z(s), u(z))ds. \tag{7}$$

Represent the optimal control via u^*, the optimal value function is generated as

$$\begin{aligned} J^*(z) &= \min_{u\in\Phi(\Omega)} \left(\int_t^\infty r(z(s), u(z))ds \right) \\ &= \int_t^\infty r(z(s), u^*(z))ds. \end{aligned} \tag{8}$$

Taking the time derivation on both sides of the optimal value function (8), the HJB equation is got as follows:

$$\begin{aligned} H(z, u^*, J_z^*) &= r(z, u^*) + J_z^* \dot{z}(t) \\ &= z^T(t)z(t) + u^{*T}u^* + J_z^{*T}(z)\Big(f(x) \\ &\quad + g(x)u^* + \Delta - \dot{y}(t)\Big) = 0, \end{aligned} \tag{9}$$

where $J_z^*(z) = \dfrac{dJ^*(z)}{dz} \in R^m$.

Assuming the solution of (9) to be existing and unique, then the optimal control u^* can be got by solving the equation $\dfrac{\partial H(z, u^*, J_z^*)}{\partial u^*} = 0$ as

$$u^* = -\frac{1}{2}g^T(x)J_z^*(z). \tag{10}$$

Define a function $K^*(z, x)$ as

$$K^*(z, x) = g^T(x)J_z^*(z), \tag{11}$$

then the optimal control described in (10) can be rewritten as

$$u^* = -\frac{1}{2}K^*(z, x). \tag{12}$$

Substituting (12) into (9), we get

$$\begin{aligned} H(z, u^*, J_z^*) &= z^T(t)z(t) + J_z^* f(x) \\ &\quad - J_z^*\dot{y}(t) + J_z^*\Delta \\ &\quad - \frac{1}{4}K^{*T}(z, x)K^*(z, x) = 0. \end{aligned} \tag{13}$$

Since the optimal control (10) contains the uncertain term $J_z^*(z)$, it is unavailable for the non-affine system (1). For the sake of deriving available optimal control, the gradient term $J_z^*(z)$ is expected to obtain by solving the HJB equation (13). But solving the equation is difficult or even impossible because the equation has strong nonlinearity. In order to solve this difficulty, the critic-actor RL algorithm based on NN approximation be usually considered.

III. MAIN RESULTS

A. REINFORCEMENT LEARNING DESIGN

To construct the critic-actor architecture RL, rewrite the term $J_z^*(z)$ as

$$K^*(z, x) = 2kz(t) + K^0(z, x) \tag{14}$$

where $k > 0$ is a design parameter, $K^0(z, x) = -2kz(t) + K^*(z, x)$.

Substituting (14) into (10), the optimal control becomes

$$u^* = -kz(t) - \frac{1}{2}K^0(z, x). \tag{15}$$

In adaptive control field, NN has become the popular tool for solving the unknown dynamic problem because of its universal function approximation ability, they can approximate a continuous function to desired accuracy over a compact set (the detailed introduction refers to [26], [31]). Since the term $K^0(z, x)$ is unknown but continuous, NN can approximate it over a compact set Ω_K in the following form

$$K^0(z, x) = \omega_K^{*T}\Pi_K(z, x) + \varepsilon_K(z, x) \tag{16}$$

where $\omega_K^* \in R^{n\times m}$ is the ideal NN weight, $\Pi_K(z, x) \in R^n$ is the basis function vector, and $\varepsilon_K \in R^m$ is the NN approximation error to satisfy $\|\varepsilon_K\| \le \tau$, where τ is a constant.

Inserting (16) into (14) and (15), we get

$$K_z^*(z) = 2kz(t) + \omega_K^{*T}\Pi_K(z, x) + \varepsilon_K(z, x), \tag{17}$$

$$u^* = -kz(t) - \frac{1}{2}\omega_K^{*T}\Pi_K(z, x) - \frac{1}{2}\varepsilon_K(z, x). \tag{18}$$

It should be noted that the NN weight ω_K^* is an unknown constant weight just for analytical purpose, therefore the optimal control (18) cannot be directly adopted for system (1). For obtaining the valid control, the critic and actor NNs for implementing RL are constructed in accordance with (17) and (18).

Critic NN is designed to evaluate the control performance as

$$\hat{K}_z(z) = 2kz(t) + \hat{\omega}_c^T(t)\Pi_K(z, x) \tag{19}$$

where $\hat{K}_z(z)$ is the estimation of $K_z^*(z)$, $\hat{\omega}_c(t) \in R^{n\times m}$ is the critic NN weight matrix.

The tuning law for critic NN weight is

$$\dot{\hat{\omega}}_c(t) = -\gamma_c \Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t) \tag{20}$$

where $\gamma_c > 0$ is the critic designed parameter.

The actor NN is designed to perform the control behavior as

$$u = -kz(t) - \frac{1}{2}\hat{\omega}_a^T(t)\Pi_K(z,x) \tag{21}$$

where u is the estimation of u^*, $\hat{\omega}_a(t) \in R^{n\times m}$ is the actor NN weight matrix.

The tuning law for actor NN weight is

$$\dot{\hat{\omega}}_a(t) = -\Pi_K(z,x)\Pi_K^T(z,x) \times\Big(\gamma_a\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big) + \gamma_c\hat{\omega}_c(t)\Big) \tag{22}$$

where $\gamma_a > 0$ is actor designed parameter.

Remark 1: The critic and actor updating laws (20) and (22) are analyzed below.

Substituting the (20) and (22) into (9), the approximated HJB equation is generated in the following

$$\begin{aligned} H(z,u,\hat{J}_z) = {} & z^T(t)z(t) + \| - kz(t) - \frac{1}{2}\hat{\omega}_a^T\Pi_K(z,x)\|^2 \\ & + \Big(\Big(g^T(x)\Big)^{-1}\Big(2kz(t) \\ & + \hat{\omega}_c^T(t)\Pi_K(z,x)\Big)\Big)^T \\ & \times\Big(f(x) - kg(x)z(t) - \dot{y}(t) + \Delta \\ & - \frac{1}{2}g(x)\hat{\omega}_a^T(t)\Pi_K(z,x)\Big). \end{aligned} \tag{23}$$

Use the HJB equation (13) and its approximation (23) to define the Bellman residual error $e(t)$ as

$$e(t) = H(z,u,\hat{J}_z) - H\Big(z,u^*,J_z^*\Big) = H(z,u,\hat{J}_z). \tag{24}$$

Based on the previous analysis, the optimized solution $u(z)$ will be required to satisfy $e(t) = H(z,u,\hat{J}_z) \to 0$. If $H(z,u,\hat{J}_z) = 0$ is true and has a unique solution, then the following is true,

$$\frac{\partial H(z,u,\hat{J}_z)}{\partial\hat{\omega}_a(t)} = \frac{1}{2}\Pi_K(z,x)\Pi_K^T(z,x)\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big) = 0 \tag{25}$$

Defined a positive function as follows:

$$P(t) = Tr\Big(\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big)^T\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big)\Big). \tag{26}$$

Obviously, $P(t) = 0$ is equivalent to the equation (25). Then the updating laws (20) and (22) are designed from the following fact.

On the basis of $\frac{\partial P(t)}{\partial\hat{\omega}_a(t)} = -\frac{\partial P(t)}{\partial\hat{\omega}_c(t)} = 2\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big)$, the time derivative of $P(t)$ is computed along (20) and (22) as

$$\begin{aligned} \frac{dP(t)}{dt} = {} & Tr\Big(\frac{\partial P(t)}{\partial\hat{\omega}_c(t)}\dot{\hat{\omega}}_c(t) + \frac{\partial P(t)}{\partial\hat{\omega}_a(t)}\dot{\hat{\omega}}_a(t)\Big) \\ = {} & Tr\Big(-\gamma_c\frac{\partial P(t)}{\partial\hat{\omega}_c(t)}\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t) \\ & - \frac{\partial P(t)}{\partial\hat{\omega}_a(t)}\Pi_K(z,x)\Pi_K^T(z,x) \\ & \times\Big(\gamma_a\Big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\Big) + \gamma_c\hat{\omega}_c(t)\Big)\Big) \\ = {} & -\frac{\gamma_a}{2}Tr\Big(\frac{\partial P(t)}{\partial\hat{\omega}_a(t)}\Pi_K(z,x)\Pi_K^T(z,x)\frac{\partial P(t)}{\partial\hat{\omega}_a(t)}\Big) \le 0. \end{aligned} \tag{27}$$

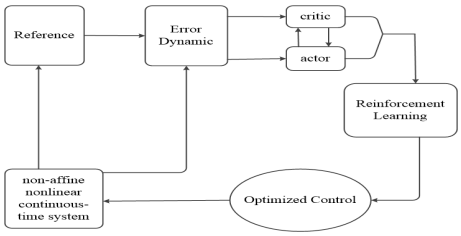

FIGURE 1. Block diagram of the proposed optimized control design.

The inequality (27) indicates that using the RL updating laws (20) and (22) can achieve $P(t) = 0$ finally, therefore the (25) can be established.

The main advantages are that: 1) in contrast to the existing methods, this optimized control algorithm is greatly simple; 2) it can remove the persistent excitation condition.

Remark 2: In this paper, RL for optimal control is adopted (as is shown in Fig.1), which is an iterative process that synchronously trains both critic and actor. Therefore, the challenge of control design is mainly focused on the derivation of the critic and actor updating laws. In the existing optimal methods, critic and actor updating laws are designed based on the square of Bellman residual. Because the equation is a complex nonlinear equation, the complexity of control design is inevitably increased. In this paper, RL algorithm is designed based on a simple positive function, which is equivalent to the HJB equation, as a result, it is of great significance to reduce the complexity of control design.

B. STABILITY ANALYSIS

Lemma 1 [40]: For a positive definite continuous function $G(t) \in R$ meets $\dot{G}(t) \le -pG(t) + q$, where $p > 0$ and $q > 0$ are two constants, then a following inequality is true:

$$G(t) \le e^{-pt}G(0) + \frac{q}{p}(1 - e^{-pt}). \tag{28}$$

Theorem 1: Consider the non-affine nonlinear system (1) under bounded initial condition. If the proposed RL optimized tracking control is performed by the critic and actor NNs (19) and (21) with the training laws (20) and (22), and these designed constants, k, γ_a and γ_c, are chosen to satisfy

$$k > \frac{\bar{\xi}^2}{4\underline{\xi}} + \frac{3}{2\underline{\xi}}, \gamma_a > \frac{1}{2}, \gamma_a > \gamma_c > \frac{1}{2}\gamma_a. \tag{29}$$

Then the proposed optimized approach can guarantee the following objectives:

1) all the errors can be guaranteed to be semi-globally uniformly ultimately bounded(SGUUB);
2) the system state $x(t)$ can track the trajectory $y(t)$ in desired accuracy.

Proof: Choose a Lyapunov function candidate as

$$L(t) = \frac{1}{2}z^T(t)z(t) + \frac{1}{2}Tr\left\{\tilde{\omega}_c^T(t)\tilde{\omega}_c(t)\right\} + \frac{1}{2}Tr\left\{\tilde{\omega}_a^T(t)\tilde{\omega}_a(t)\right\}, \quad (30)$$

where $\tilde{\omega}_c(t) = \hat{\omega}_c(t) - \omega_K^*$ and $\tilde{\omega}_a(t) = \hat{\omega}_a(t) - \omega_K^*$ are the critic and actor NN weight errors, respectively.

Taking the time derivative of $L(t)$ along (5), (20) and (22) is

$$\begin{aligned}\dot{L}(t) = {} & z^T(t)\Big(f(x) + g(x)u - \dot{y}(t) + \Delta\Big) \\ & -\gamma_c Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & -Tr\Big\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x) \\ & \times\Big(\gamma_a\big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\big) + \gamma_c\hat{\omega}_c(t)\Big)\Big\}. \end{aligned} \quad (31)$$

Adding (21) into (31) gets

$$\begin{aligned}\dot{L}(t) = {} & -kz^T(t)g(x)z(t) - \frac{1}{2}z^T(t)g(x)\hat{\omega}_a^T(t)\Pi_K(z,x) \\ & +z^T(t)f(x) - z^T(t)\dot{y}(t) + z^T(t)\Delta \\ & -\gamma_c Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & -Tr\Big\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x) \\ & \times\Big(\gamma_a\big(\hat{\omega}_a(t) - \hat{\omega}_c(t)\big) + \gamma_c\hat{\omega}_c(t)\Big)\Big\}. \end{aligned} \quad (32)$$

On the basis of Cauchy-Schwartz and Young's inequalities, it follows

$$\begin{aligned} z^T(t)f(x) &\le \frac{1}{2}\|z(t)\|^2 + \frac{1}{2}\|f(x)\|^2, \\ -z^T(t)\dot{y}(t) &\le \frac{1}{2}\|z(t)\|^2 + \frac{1}{2}\|\dot{y}(t)\|^2, \\ z^T(t)\Delta &\le \frac{1}{2}\|z(t)\|^2 + \frac{1}{2}\|\Delta\|^2, \\ & -\frac{1}{2}z^T(t)g(x)\hat{\omega}_a^T(t)\Pi_K(z,x) \\ & \le \frac{1}{4}z^T(t)g(x)g^T(x)z(t) \\ & +\frac{1}{4}Tr\left\{\hat{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\}. \end{aligned} \quad (33)$$

Substituting the inequalities (33) into (32) becomes

$$\begin{aligned}\dot{L}(t) \le {} & -kz^T(t)g(x)z(t) + \frac{3}{2}\|z(t)\|^2 \\ & +\frac{1}{4}z^T(t)g(x)g^T(x)z(t) \\ & +\frac{1}{2}\|f(x)\|^2 + \frac{1}{2}\|\dot{y}(t)\|^2 + \frac{1}{2}\|\Delta\|^2 \\ & +\frac{1}{4}Tr\left\{\hat{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\} \\ & -\gamma_c Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & -\gamma_a Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\} \\ & +(\gamma_a - \gamma_c)Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\}. \end{aligned} \quad (34)$$

By using the facts $\tilde{\omega}_c(t) = \hat{\omega}_c(t) - \omega_K^*$ and $\tilde{\omega}_a(t) = \hat{\omega}_a(t) - \omega_K^*$, the following equations can be gained:

$$\begin{aligned} & Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & = \frac{1}{2}Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_c(t)\right\} \\ & \quad +\frac{1}{2}Tr\left\{\hat{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & \quad -\frac{1}{2}Tr\left\{\omega_K^{*T}\Pi_K(z,x)\Pi_K^T(z,x)\omega_K^*\right\}, \\ & Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\} \\ & = \frac{1}{2}Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_a(t)\right\} \\ & \quad +\frac{1}{2}Tr\left\{\hat{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\} \\ & \quad -\frac{1}{2}Tr\left\{\omega_K^{*T}\Pi_K(z,x)\Pi_K^T(z,x)\omega_K^*\right\}. \end{aligned} \quad (35)$$

Applying the above results to the inequality (34) has

$$\begin{aligned}\dot{L}(t) \le {} & -kz^T(t)g(x)z(t) + \frac{3}{2}\|z(t)\|^2 \\ & +\frac{1}{4}z^T(t)g(x)g^T(x)z(t) \\ & +\frac{1}{2}\|f(x)\|^2 + \frac{1}{2}\|\dot{y}(t)\|^2 + \frac{1}{2}\|\Delta\|^2 \\ & -\frac{\gamma_c}{2}Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_c(t)\right\} \\ & -\frac{\gamma_a}{2}Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_a(t)\right\} \\ & +(\gamma_a - \gamma_c)Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & -\frac{\gamma_c}{2}Tr\left\{\hat{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & -\Big(\frac{\gamma_a}{2} - \frac{1}{4}\Big)Tr\left\{\hat{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\} \\ & +\Big(\frac{\gamma_a}{2} + \frac{\gamma_c}{2}\Big)Tr\left\{\omega_K^{*T}\Pi_K(z,x)\Pi_K^T(z,x)\omega_K^*\right\}. \end{aligned} \quad (36)$$

According to the Young's inequality and (29), there is the following one that

$$\begin{aligned} & (\gamma_a - \gamma_c)Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\} \\ & \le \frac{\gamma_a - \gamma_c}{2}Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_a(t)\right\} \\ & \quad +\frac{\gamma_a - \gamma_c}{2}Tr\left\{\hat{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\}. \end{aligned} \quad (37)$$

Then, (36) can be expressed as

$$\dot{L}(t) \le -kz^T(t)g(x)z(t) + \frac{3}{2}\|z(t)\|^2$$

$$
\begin{aligned}
&+\frac{1}{4}z^T(t)g(x)g^T(x)z(t)\\
&-\frac{\gamma_c}{2}Tr\left\{\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_c(t)\right\}\\
&-\frac{\gamma_c}{2}Tr\left\{\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_a(t)\right\}\\
&-\left(\gamma_c-\frac{\gamma_a}{2}\right)Tr\left\{\hat{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_c(t)\right\}\\
&-\left(\frac{\gamma_a}{2}-\frac{1}{4}\right)Tr\left\{\hat{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\hat{\omega}_a(t)\right\}\\
&+S(t), \qquad (38)
\end{aligned}
$$

where $S(t) = \frac{1}{2}\|f(x)\|^2 + \frac{1}{2}\|\dot{y}(t)\|^2 + \frac{1}{2}\mu^2 + \left(\frac{\gamma_a}{2} + \frac{\gamma_c}{2}\right)\left(\omega_K^{*T}(t)\Pi_K(z,x)\right)^2$, which is bounded by a constant s, that is, $S \le s$, because all the terms are bounded.

According to Assumption 1 and condition (29), the inequality (38) can be reorganized as

$$
\begin{aligned}
\dot{L}(t) \le &-\left(k\underline{\xi} - \frac{1}{4}\bar{\xi}^2 - \frac{3}{2}\right)z^T(t)z(t)\\
&-\frac{\gamma_c}{2}\tilde{\omega}_c^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_c(t)\\
&-\frac{\gamma_c}{2}\tilde{\omega}_a^T(t)\Pi_K(z,x)\Pi_K^T(z,x)\tilde{\omega}_a(t) + s. \qquad (39)
\end{aligned}
$$

Letting $h = min\{2k\underline{\xi} - \frac{1}{2}\bar{\xi}^2 - 3, \gamma_c\lambda_{min}\}$, λ_{min} is the minimum eigenvalue of $\Pi_K(z,x)\Pi_K^T(z,x)$, then the inequality (39) can be described as

$$\dot{L}(t) \le -hL(t) + s. \qquad (40)$$

Applying Lemma 1 into (40) becomes

$$L(t) \le e^{-ht}L(0) + \frac{s}{h}(1 - e^{-ht}). \qquad (41)$$

According to the above inequality, all error signals are SGUUB, and when the design parameter k is large enough, the tracking error can convergent to the desired accuracy.

IV. SIMULATION EXAMPLE

Consider the following numerical example of non-affine nonlinear systems

$$\dot{x}(t) = \begin{bmatrix} -0.7\sin x_1 + 0.5x_2 \\ \sin x_1 cos x_2 \end{bmatrix} + 0.8u + 0.2\begin{bmatrix} sin(u) \\ cos(u) \end{bmatrix} \qquad (42)$$

where $x(t) = [x_1, x_2]^T \in R^2$ and $x(0) = [0.3, 0.3]^T$, $u \in R^2$. The expected reference signal is $y(t) = [4sin(0.8t), 5cos(t)]^T$ and its initial state is $y(t) = [0, 0]^T$.

Corresponding to the control protocol (21), the parameter is chosen as $\beta = 18$. The NN with 12 neurons are employed for the NN approximation (16). The basis function vector is designed as $\Pi(x) = [\Pi_1(x), \ldots, \Pi_{12}(x)]^T$ with $\Pi_i(x) = exp[-(x-\eta_i)^T(x-\eta_i)/2]$. And the NN center $\eta_i \in R^2$, $i = 1, 2, \ldots, 12$, equally spaced in an interval of -6 to 6.

Corresponding to the two updating laws (20) and (22), the parameter $\gamma_c = 16$ represents the critic updating and the parameter $\gamma_a = 14$ for actor updating. And the initial weights are written as $\hat{\omega}_c = [0.2, \ldots, 0.2]^T \in R^{12\times 1}$, $\hat{\omega}_a = [0.2, \ldots, 0.2]^T \in R^{12\times 1}$.

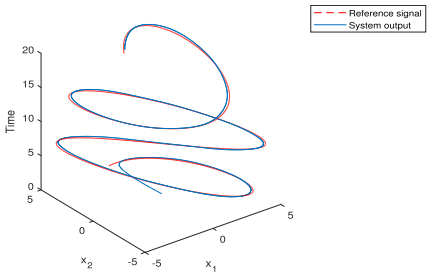

FIGURE 2. Tracking performance.

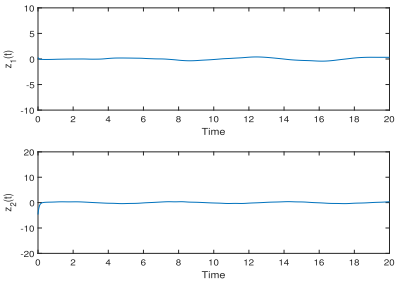

FIGURE 3. Tracking error.

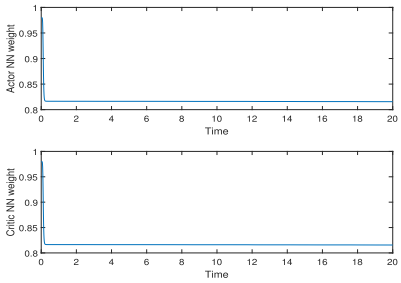

FIGURE 4. Norm of actor and critic NN weights.

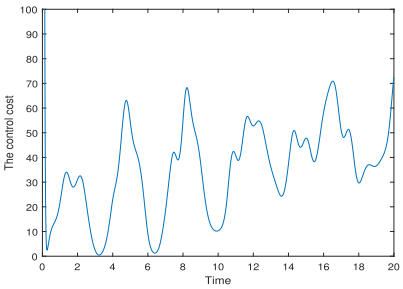

FIGURE 5. Cost function $r(z, u)$.

The simulation results are shown in Figs. 2-5. Fig. 2 describes the tracking performance. Fig. 3 presents the tracking error to be convergent. From Figs. 2-3, the system states can follow to the desired reference trajectory. In the Fig. 4, the boundedness of critic and actor NN weights are shown, respectively. The cost function is displayed in the Fig. 5. The tracking capability of controller is demonstrated through the simulation results.

We make a comparison with the method of reference [9] in the computation time by using the ''tic'' function of MATLAB. Their averaged times are $0.1132s$ for the proposed method and $0.3779s$ for the method of [9] respectively. It is obvious that the proposed method is with the less computational time.

V. CONCLUSION

In this article, an optimized control method is developed for a class of continuous-time non-affine nonlinear systems. Since the system is with the implicit control, it needs to transform the system to the affine-like form for revealing the control. According to the transformed system, the optimal control is derived by employing the NN-based RL. Since the RL updating laws is derived from negative gradient of a simple positive function, which is designed based on the partial derivatives of the HJB equation, the proposed optimized control can be significantly simple to compare with the existing RL optimal methods. Moreover, it can remove the condition of persistence excitation. Finally, it is proven that the control targets with the desired control performance are achieved. The effectiveness of the proposed optimizing method is demonstrated by the theory proof and simulation.

The disadvantages of the methods are mainly involved two aspects: 1) the control scheme is designed for an abstract mathematics model, it is not for a specific practical dynamic system. Hence we will extend this method to the practical engineering systems; 2) the optimal control of non-affine nonlinear system is focused on the first-order case, we will consider to develop the method to the second-order non-affine nonlinear systems.

REFERENCES

[1] S. Ge and J. Zhang, ''Neural-network control of nonaffine nonlinear system with zero dynamics by state and output feedback,'' *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 900–918, Jul. 2003.

[2] Y. Li, S. Tong, and T. Li, ''Adaptive fuzzy backstepping control design for a class of pure-feedback switched nonlinear systems,'' *Nonlinear Anal., Hybrid Syst.*, vol. 16, pp. 72–80, May 2015.

[3] A. Boubakir, S. Labiod, F. Boudjema, and F. Plestan, ''Linear adaptive control of a class of SISO nonaffine nonlinear systems,'' *Int. J. Syst. Sci.*, vol. 45, no. 12, pp. 2490–2498, Dec. 2014.

[4] S. S. Ge, T. H. Lee, and J. Wang, ''Adaptive control of non-affine nonlinear systems using neural networks,'' in *Proc. IEEE Int. Symp. Intell. Control*, Oct. 2000, pp. 13–18.

[5] Z. Chen and S. Jagannathan, ''Generalized Hamilton–Jacobi–Bellman formulation-based neural network control of affine nonlinear discrete-time systems,'' *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 90–106, Jan. 2008.

[6] G. Wen, C. L. P. Chen, J. Feng, and N. Zhou, ''Optimized multi-agent formation control based on an identifier-actor-critic reinforcement learning algorithm,'' *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2719–2731, Oct. 2018.

[7] T. Dierks and S. Jagannathan, ''Online optimal control of nonlinear discrete-time systems using approximate dynamic programming,'' *J. Control Theory Appl.*, vol. 9, no. 3, pp. 361–369, 2011.

[8] T. Dierks, B. T. Thumati, and S. Jagannathan, ''Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence,'' *Neural Netw.*, vol. 22, nos. 5–6, pp. 851–860, 2009.

[9] G. Wen, C. L. P. Chen, S. S. Ge, H. Yang, and X. Liu, ''Optimized adaptive nonlinear tracking control using actor–critic reinforcement learning strategy,'' *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 4969–4977, Sep. 2019.

[10] R. Bellman, *Dynamic Programming*, vol. 1, no. 2. Princeton, NJ, USA: Princeton Univ. Press, 1957, p. 3.

[11] P. Werbos, ''Approximate dynamic programming for realtime control and neural modelling,'' in *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. New York, NY, USA: Van Nostrand, 1992, pp. 493–525.

[12] D. Wang, M. Zhao, M. Ha, and L. Hu, ''Adaptive-critic-based hybrid intelligent optimal tracking for a class of nonlinear discrete-time systems,'' *Eng. Appl. Artif. Intell.*, vol. 105, Oct. 2021, Art. no. 104443.

[13] P. Ghanooni, A. M. Yazdani, A. Mahmoudi, S. M. Zadeh, M. A. Movahed, and M. Fathi, ''Robust precise trajectory tracking of hybrid stepper motor using adaptive critic-based neuro-fuzzy controller,'' *Comput. Electr. Eng.*, vol. 81, Jan. 2020, Art. no. 106535.

[14] M. Abu-Khalaf, F. L. Lewis, and J. Huang, ''Neuro dynamic programming and zero-sum games for constrained control systems,'' *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1243–1252, Jul. 2008.

[15] D. Liu and Q. Wei, ''Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[16] R. Song, W. Xiao, H. Zhang, and C. Sun, ''Adaptive dynamic programming for a class of complex-valued nonlinear systems,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1733–1739, Sep. 2014.

[17] A. Perrusquía and W. Yu, ''Identification and optimal control of nonlinear systems using recurrent neural networks and reinforcement learning: An overview,'' *Neurocomputing*, vol. 438, pp. 145–154, May 2021.

[18] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, ''Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers,'' *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[19] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, ''Reinforcement learning for optimal control of low exergy buildings,'' *Appl. Energy*, vol. 156, pp. 577–586, Oct. 2015.

[20] R. Kamalapurkar, J. A. Rosenfeld, and W. E. Dixon, ''Efficient model-based reinforcement learning for approximate online optimal control,'' *Automatica*, vol. 74, pp. 247–258, Dec. 2016.

[21] X. Yang, D. Liu, and D. Wang, ''Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints,'' *Int. J. Control*, vol. 87, no. 3, pp. 553–566, Oct. 2013.

[22] X. Yang, D. Liu, and Y. Huang, ''Neural-network-based online optimal control for uncertain non-linear continuous-time systems with control constraints,'' *IET Control Theory Appl.*, vol. 7, no. 17, pp. 2037–2047, Nov. 2013.

[23] G. Wen, C. L. P. Chen, and B. Li, ''Optimized formation control using simplified reinforcement learning for a class of multiagent systems with unknown dynamics,'' *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 7879–7888, Sep. 2020.

[24] G. Wen, C. L. Philip Chen, and W. N. Li, ''Simplified optimized control using reinforcement learning algorithm for a class of stochastic nonlinear systems,'' *Inf. Sci.*, vol. 517, pp. 230–243, May 2020.

[25] K. Hornik, M. Stinchcombe, and H. White, ''Multilayer feedforward networks are universal approximators,'' *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[26] G. Wen, C. Zhang, P. Hu, and Y. Cui, ''Adaptive neural network leader-follower formation control for a class of second-order nonlinear multi-agent systems with unknown dynamics,'' *IEEE Access*, vol. 8, pp. 148149–148156, 2020.

[27] L. X. Wang and J. M. Mendel, ''Fuzzy basis functions, universal approximation, and orthogonal least-squares learning,'' *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 807–814, Sep. 1992.

[28] G. Wen, S. S. Ge, F. Tu, and Y. S. Choo, ''Artificial potential-based adaptive H_∞ synchronized tracking control for accommodation vessel,'' *IEEE Trans. Ind. Electron.*, vol. 64, no. 7, pp. 5640–5647, Mar. 2017.

[29] F. C. Chen and H. K. Khalil, ''Adaptive-control of nonlinear-systems using neural networks,'' *Int. J. Control*, vol. 55, no. 6, pp. 1299–1317, 1992.

[30] W.-Y. Wang, M.-L. Chan, C.-C. J. Hsu, and T.-T. Lee, ''H_∞ tracking-based sliding mode control for uncertain nonlinear systems via an adaptive fuzzy-neural approach,'' *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 32, no. 4, pp. 483–492, Aug. 2002.

[31] G. Wen, C. L. P. Chen, Y.-J. Liu, and Z. Liu, ''Neural network-based adaptive leader-following consensus control for a class of nonlinear multiagent state-delay systems,'' *IEEE Trans. Cybern.*, vol. 47, no. 8, pp. 2151–2160, Aug. 2017.

[32] S. S. Ge and C. Wang, "Adaptive neural control of uncertain MIMO nonlinear systems," *IEEE Trans. Neural Netw.*, vol. 15, no. 3, pp. 674–692, May 2004.
[33] G. Wen, S. S. Ge, and F. Tu, "Optimized backstepping for tracking control of strict-feedback systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3850–3862, Aug. 2018.
[34] G. Wen, C. L. P. Chen, and S. S. Ge, "Simplified optimized backstepping control for a class of nonlinear strict-feedback systems with unknown dynamic functions," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4567–4580, Sep. 2020.
[35] G. Wen, S. S. Ge, C. L. P. Chen, F. Tu, and S. Wang, "Adaptive tracking control of surface vessel using optimized backstepping technique," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3420–3431, Sep. 2019.
[36] D. Liu, Y. Huang, D. Wang, and Q. Wei, "Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming," *Int. J. Control*, vol. 86, no. 9, pp. 1554–1566, Sep. 2013.
[37] S. Doudou and F. Khaber, "Adaptive fuzzy sliding mode control for a class of uncertain nonaffine nonlinear strict-feedback systems," *Iranian J. Sci. Technol., Trans. Electr. Eng.*, vol. 43, no. 1, pp. 33–45, Oct. 2018.
[38] A. Boulkroune, M. M'Saad, and M. Farza, "Adaptive fuzzy tracking control for a class of MIMO nonaffine uncertain systems," *Neurocomputing*, vol. 93, pp. 48–55, Sep. 2012.
[39] S. Labiod and T. M. Guerra, "Indirect adaptive fuzzy control for a class of nonaffine nonlinear systems with unknown control directions," *Int. J. Control, Autom. Syst.*, vol. 8, no. 4, pp. 903–907, Aug. 2010.
[40] S. S. Ge, C. C. Hang, and T. Zhang, "Adaptive neural network control of nonlinear systems by state and output feedback," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 29, no. 6, pp. 818–828, Dec. 1999.

XUE YANG received the B.S. degree from the Qilu University of Technology (Shandong Academy of Sciences), in 2015, where she is currently pursuing the master's degree. Her main research interests include optimal control and adaptive control.

BIN LI received the B.S., M.S., and Ph.D. degrees from Shandong University, China, in 2002, 2005, and 2012, respectively. He is currently a Professor with the School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), China. His research interests include algorithms for neural networks, gait planning, and adaptive control of legged robots.

GUOXING WEN received the M.S. degree in applied mathematics from the Liaoning University of Technology, Jinzhou, China, in 2011, and the Ph.D. degree in computer and information science from Macau University, Macau, China, in 2014.

He was a Research Fellow with the Department of Electrical and Computer Engineering, Faculty of Engineering, National University of Singapore, Singapore, from September 2015 to September 2016. He is currently an Associate Professor with the College of Science, Binzhou University, Shandong, China. His research interests include adaptive control, optimal control, multi-agent control, nonlinear systems, reinforcement learning, neural networks, and fuzzy logic systems.