Reinforcement Learning-Based Tracking Control for a Three Mecanum Wheeled Mobile Robot

Dianfeng Zhang, Guangcang Wang, and Zhaojing Wu, *Member, IEEE*

***Abstract*— This brief investigates the robust optimal tracking control for a three Mecanum wheeled mobile robot (MWMR) with the external disturbance by the aid of online actor–critic synchronous learning algorithm. The Euler–Lagrange motion equation of MWMR subject to slipping is established by analyzing the structural characteristics of Mecanum wheels. Concatenating the tracking error with the desired trajectory, the tracking control problem is converted into a time-invariant optimal control problem of an augmented system. Then, an approximate optimal tracking controller is obtained by applying online actor–critic synchronous learning algorithm. With the help of Lyapunov-based analysis, the ultimately bounded tracking can be guaranteed. Finally, simulation results show the effectiveness of synchronous learning algorithm and approximate optimal tracking controller.**

***Index Terms*— Hamilton–Jacobi–Bellman (HJB) equation, Mecanum wheeled mobile robot (MWMR), neural networks (NNs), online actor–critic algorithm, optimal tracking control.**

I. Introduction

Wheeled mobile robots (WMRs) have been widely used in various fields, such as industrial, military, aerospace, health sector, scientific, and so on, due to the advantages of good motion flexibility, low mechanical complexity, and energy consumption [1]–[3]. As an important class of WMRs, Mecanum WMRs (MWMRs) have been and still remain the subject of intensive research over the past decades [4]–[10] because of the omnidirectional mobility capability which can make them conveniently transport in a congested or highly dynamic environments.

The trajectory tracking control of omnidirectional WMRs, which is one of the most important motion control problems, has been extensively studied. Sliding-mode control strategies were utilized for robust tracking control of omnidirectional WMRs in the presence of friction, external force disturbance, and uncertainties [7], [9], [11]. To address the control singularity caused by adaptive linearizing control, a smooth switching adaptive robust tracking control algorithm was proposed in [12] for omnidirectional mobile robots with both structured and unstructured uncertainties. Based on the energy shaping plus damping approach, passivity-based tracking control method was proposed for an omnidirectional mobile robot in [13]. Recently, in [14], a model predictive control algorithm was used to deal with the tracking problem of an MWMR. However, most of existing works do not consider slipping. In many real applications, the slipping of wheels is inevitable due to the tire deformation, aging wheels, and slippery ground and other reasons [15], [16]. Hence, it is necessary to consider the modeling and control of MWMR with slipping.

It is worth noting that most of the aforementioned controllers concentrate on applying conventional control methods which do not involve autonomous learning mechanisms. Reinforcement learning (RL), an important branch of machine learning [17], [18], has received significant attention in control engineering (see [19], [20], and the references therein) in view of the capability of solving traditional optimal control problems. RL algorithms are mainly used to find an optimal solution of the Hamilton–Jacobi–Bellman (HJB) equation by iteratively performing policy evaluation and policy improvement. As a consequence, the optimal regulation and optimal tracking of nonlinear systems can be addressed with the help of the implementation of RL algorithm. In [21], an online adaptive algorithm was presented for learning the solution to the optimal regulation of nonlinear systems by using an actor–critic structure consisting of both actor and critic neural networks (NNs). From then on, the optimal regulation problems for nonlinear systems with unknown dynamics have been addressed by employing actor–critic–identifier learning architecture, integral RL algorithm and model-based RL algorithm in [22]–[24]. The robust control problems for nonlinear systems with uncertainties and input constraints were tackled, respectively, in [25] and [26] based on an online policy iteration algorithm using a single critic NN.

The infinite-time optimal tracking control problem for discrete-time nonlinear systems was studied in [27] via transforming the original optimal tracking problem into an optimal regulation problem of an augmented system. Based on system transformation similar to [27], the infinite-horizon optimal tracking control problems for continuous-time nonlinear systems, unknown dynamics nonlinear systems and uncertain nonlinear systems have been developed, respectively, in [28]–[30] by using the actor–critic algorithm related to the desired control, integral RL algorithm and adaptive-critic algorithm. Recently, in [31], a single network-based adaptive dynamic programming (ADP) algorithm was applied to investigate the robust optimal flight control for near space vehicle attitude system with external time-varying disturbance by means of a disturbance observer.

It is noted that the performance index defined in [30] may be infinity because the control depends on the reference trajectory as pointed out in [20] and [27]. This obstacle can be overcome by introducing a discount factor [28] or the error cost between the real control and the desired control [29] in the performance index. However, in the presence of external disturbances, the control error does not go to zero, which might cause the performance index defined in [29] to be infinity. Moreover, although ADP algorithm using a single critic NN [31] can cope with the optimal tracking control of systems with disturbance, the algorithm is implemented offline and intractable to real-time control applications as pointed out in [26].

Motivated by the above discussion, this brief concentrates on using an online actor–critic algorithm to develop a robust tracking control scheme for a three MWMR with external disturbances. To find in real-time a suitable approximate optimal tracking control policy, we introduce an appropriate discounted performance index, which brings together the uncertainty of the disturbance, the tracking error, and the error between the real control and the desired control, and the unbounded of the performance index can be well prevented

Manuscript received 7 September 2021; revised 4 March 2022; accepted 18 June 2022. This work was supported by the National Natural Science Foundation of China under Grant 61803324 and Grant 62073275. *(Corresponding author: Zhaojing Wu.)*

The authors are with the School of Mathematics and Informational Science, Yantai University, Yantai 264005, China (e-mail: d_feng1986@163.com; gcwang20@163.com; wuzhaojing00@188.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3185055.

Digital Object Identifier 10.1109/TNNLS.2022.3185055

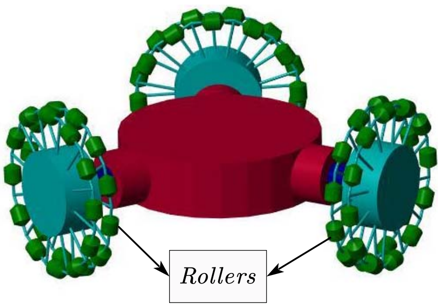

Fig. 1. Three 90° MWMR.

simultaneously. The main work of this brief includes: 1) for an MWMR with three 90° Mecanum wheels, the Euler–Lagrange motion equation of MWMR subject to slipping is presented based on inverse kinematics analysis of Mecanum wheel and Lagrangian mechanics [3] and 2) a discounted performance index containing the bounds of the uncertain term is introduced to improve the robustness of the system. Furthermore, an online actor–critic synchronous learning algorithm is adopted and a modified tuning law is obtained to solve the optimal control problem of the augmented system. Moreover, compared with the method in [23] and [29], the robustness of the system is improved. The ultimately bounded of the closed loop system and robust tracking performance are analyzed by using Lyapunov-based method.

The remainder of this article is organized as follows. In Section II, model description and problem formulation are given. In Section III, system transformation, the implementation of online actor–critic synchronous learning algorithm and stability analysis are provided. Simulations are presented in Section IV. Finally, conclusions are drawn in Section V.

Notations: $\mathbb{R}^n$ is the real n-dimensional space, $\mathbb{R}^{n\times m}$ is the space of all $n\times m$ real matrices. $\|\cdot\|$ denotes the Euclidean norm of a vector or a matrix. For a matrix X, X^{-1} is the inverse matrix, X^{T} represents its transpose, $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote the maximal and minimal eigenvalues of the matrix X, respectively. The bold letter $\boldsymbol{I}$ or $\boldsymbol{I}_n$ denote the identity matrix of appropriate dimensions, $\mathbf{1}_n$ represents a n-dimensions column vector having all of its elements equal to one, and $\mathbf{0}_{n\times m}$ stands for the $n \times m$ zero matrix.

II. Model Description and Problem Formulation

In this brief, we consider an omnidirectional mobile robot which is equipped with three 90° Mecanum wheels, i.e., the angle $\gamma = 0°$ between the wheel plane and the rotation axis of the roller as shown in Fig. 1.

To study the motion of the mobile robot, three coordinate frames are defined as indicated in Fig. 2. The earth-fixed frame $O-\{X, Y\}$ is considered to be an inertial reference frame for the robot motion in the plane. The body-fixed frame $P-\{X_R, Y_R\}$ is attached to the center of the mass (CM) which is located at the center of the robot circle, X_R is along the center of the rotation axis of A_2 wheel. $A_i - \{X_{Wi}, Y_{Wi}\}$ $(i = 1, 2, 3)$ is coordinate frame of the ith wheel attached to the wheel's axis center, X_{Wi} is along the drive direction of the ith wheel. Configuration parameters l denotes the distance of PA_i, α_i is the angle between PA_i and X_R, β_i is the angle between PA_i and Y_{Wi}, θ (the angle between X_R and X) represents orientation of the mobile robot around CM point P, r denotes the radius of the wheel.

Let $\boldsymbol{q} = (x, y, \theta)^{\mathrm{T}}$ be the position (x, y) of CM point P and orientation θ of robot in the inertial frame, $\dot{\boldsymbol{q}} = (\dot{x}, \dot{y}, \dot{\theta})^{\mathrm{T}}$ be the velocity of CM point P in the inertial frame, $\boldsymbol{v} = (v_x, v_y, \omega)^{\mathrm{T}}$ represent generalized velocity of CM point P in the body-fixed frame, $\boldsymbol{\Phi} = (\varphi_1, \varphi_2, \varphi_3)^{\mathrm{T}}$ represent the angular position of all wheels.

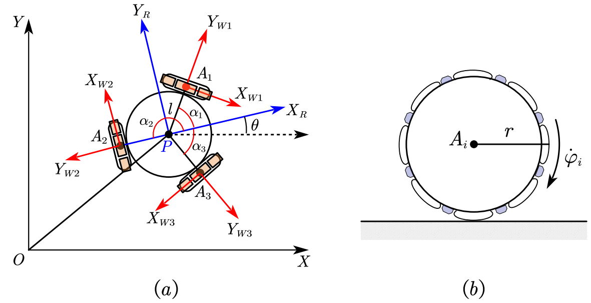

Fig. 2. (a) Coordinate frames and configuration parameters for mobile robot. (b) Parameters for wheels.

The main purpose of this brief is to develop a robust tracking controller for MWMR based on RL so that MWMR can optimally track a time-varying desired trajectory $\boldsymbol{q}_r(t)$.

In the following, the motion equations of MWMR subject to disturbances are constructed based on kinematics and dynamics of mobile robot and Mecanum wheel [3], [32], [33], and the problem formulation is presented.

A. Kinematics and Dynamics of MWMR With Slipping

The velocity relationship between the body-fixed frame and the inertial reference frame is presented as follows:

$$\dot{\boldsymbol{q}} = \boldsymbol{R}(\theta)\boldsymbol{v} \tag{1}$$

where $\boldsymbol{R}(\theta)$ is the transform matrix from the body-fixed frame to the inertial frame

$$\boldsymbol{R}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

By the aid of kinematics for Mecanum wheel [3], in the case of nonslipping, the linear velocity of the wheel is $v_{li} = r\dot{\varphi}_i$. Then the relation between the angular velocity $\dot{\varphi}_i$ of each wheel and the robot velocity $\boldsymbol{v}$ can be expressed by

$$(\sin(\alpha_i + \beta_i), -\cos(\alpha_i + \beta_i), -l\cos(\beta_i))\boldsymbol{v} = v_{li}. \tag{2}$$

In many real applications, the slipping of wheels is inevitable due to the tire deformation, aging wheels and slippery ground. Considering the slipping, the slippage ρ [16] is defined as $\rho = ((r\dot{\varphi}_i - v_{li})/(r\dot{\varphi}_i))$, where $-1 < \rho < 1$. Then, the slip velocity in the direction of the wheel rotation is $\rho r\dot{\varphi}_i = r\dot{\varphi}_i - v_{li}$ and (2) becomes

$$(\sin(\alpha_i + \beta_i), -\cos(\alpha_i + \beta_i), -l\cos(\beta_i))\boldsymbol{v} = (1-\rho)r\dot{\varphi}_i. \tag{3}$$

From the local reference frame $P - \{X_R, Y_R\}$ in Fig. 2 and the configuration parameters $\alpha_i = (((2i-1)\pi)/3)$, $\beta_i = 0$, $i = 1, 2, 3$, a collected expression of (3) is given

$$\begin{pmatrix} \sin\frac{\pi}{3} & -\cos\frac{\pi}{3} & -l \\ \sin\pi & -\cos\pi & -l \\ \sin\left(-\frac{\pi}{3}\right) & -\cos\left(-\frac{\pi}{3}\right) & -l \end{pmatrix}\boldsymbol{v} = (1-\rho)r\begin{pmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \\ \dot{\varphi}_3 \end{pmatrix} = (1-\rho)r\dot{\boldsymbol{\Phi}}. \tag{4}$$

Combining (1) with (4) yields the kinematic equation of MWMR as follows:

$$\dot{\boldsymbol{q}} = (1-\rho)r\boldsymbol{H}(\theta)\dot{\boldsymbol{\Phi}} \tag{5}$$

where

$$\boldsymbol{H}^{-1}(\theta) = \begin{pmatrix} \sin\left(\theta + \frac{\pi}{3}\right) & -\cos\left(\theta + \frac{\pi}{3}\right) & -l \\ -\sin\theta & \cos\theta & -l \\ \sin\left(\theta - \frac{\pi}{3}\right) & -\cos\left(\theta - \frac{\pi}{3}\right) & -l \end{pmatrix}.$$

Applying Lagrangian mechanics [34], the dynamic equations of MWMR can be derived as

$$M\ddot{\boldsymbol{\Phi}} + C\big(\boldsymbol{\Phi}, \dot{\boldsymbol{\Phi}}\big)\dot{\boldsymbol{\Phi}} + \mu\dot{\boldsymbol{\Phi}} = \tau \tag{6}$$

where $M = \begin{pmatrix} m_1 & m_2 & m_2 \\ m_2 & m_1 & m_2 \\ m_2 & m_2 & m_1 \end{pmatrix}, C(\boldsymbol{\Phi}, \dot{\boldsymbol{\Phi}}) = 0$, $m_1 = (4mr^2/9) + (I_\omega r^2/9l^2) + I_\varphi$, $m_2 = (I_\omega r^2/9l^2) - (2mr^2/9)$.

It is easy to verify that the inertia matrix M has the following useful property.

Property 1: The inertia matrix M is symmetric and there exist positive constants $\delta_1 > 0$ and $\delta_2 > 0$, such that $0 < \delta_1 I_{3\times3} \leq M \leq \delta_2 I_{3\times3}$.

Taking the time derivative of (5) yields

$$\ddot{\boldsymbol{q}} = (1-\rho)\dot{\boldsymbol{H}}(\theta)\boldsymbol{H}^{-1}(\theta)\dot{\boldsymbol{q}} + (1-\rho)r\boldsymbol{H}(\theta)\ddot{\boldsymbol{\Phi}}. \tag{7}$$

Substituting (6) into (7), one can obtain the following dynamic equation of MWMR:

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{v}_q \\ \dot{\boldsymbol{v}}_q &= \bar{f}(\boldsymbol{q})\boldsymbol{v}_q + \bar{g}(\boldsymbol{q})\tau + \bar{g}(\boldsymbol{q})d(\boldsymbol{q}, v_q) \end{aligned} \tag{8}$$

where $\bar{f}(\boldsymbol{q}) = \dot{\boldsymbol{H}}(\theta)\boldsymbol{H}^{-1}(\theta) - \mu\boldsymbol{H}(\theta)M^{-1}\boldsymbol{H}^{-1}(\theta)$, $\bar{g}(\boldsymbol{q}) = r\boldsymbol{H}(\theta)M^{-1}$, $\bar{f}_d(\boldsymbol{q}, v_q) = -\rho(\bar{f}(\boldsymbol{q})\boldsymbol{v}_q + \bar{g}(\boldsymbol{q})\tau)$, $d(\boldsymbol{q}, v_q) = \bar{g}^{-1}(\boldsymbol{q})\bar{f}_d(\boldsymbol{q}, v_q)$. The unknown torque $d(\boldsymbol{q}, v_q)$ induced by the slipping enters into the dynamics of the MWMR, which is considered as the external disturbances to the MWMR.

The following assumption is made for the unknown disturbance $d(\boldsymbol{q}, v_q)$ caused by slipping.

Assumption 1: The unknown disturbance $d(\boldsymbol{q}, v_q)$ is bounded by a known function $d_M(\boldsymbol{q})$, i.e., $\|d(\boldsymbol{q}, v_q)\| \leq d_M(\boldsymbol{q})$.

B. Problem Formulation

The objective is to deal with the robust tracking control problem for system (8) such that $\boldsymbol{q}(t)$ can track a desired bounded continuously differentiable trajectory $\boldsymbol{q}_r(t)$.

Let $x = (\boldsymbol{q}^{\mathrm{T}}, \boldsymbol{v}_q^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^6$, $f(x) = (\boldsymbol{v}_q^{\mathrm{T}}, (\bar{f}(\boldsymbol{q})\boldsymbol{v}_q)^{\mathrm{T}})^{\mathrm{T}}$, $g(x) = (\boldsymbol{0}_{3\times3}, \bar{g}(\boldsymbol{q})^{\mathrm{T}})^{\mathrm{T}}$, (8) can be rewritten as

$$\dot{x} = f(x) + g(x)\tau + g(x)d(\boldsymbol{q}, v_q).$$

According to the previous modeling processes, it is not difficulty to verify that one has the following property which coincides with [29, Assumption 1].

Property 2: $f(x)$ and $g(x)$ are Lipschitz, $\|f(x)\| \leq b_f\|x\|$ for some constant b_f, $g(x)$ is bounded by a constant b_g, i.e., $\|g(x)\| \leq b_g$.

Property 3: The matrix $g(x)$ has full column rank for all $x \in \mathbb{R}^6$, and the function $g^+ \triangleq (g^{\mathrm{T}}g)^{-1}g^{\mathrm{T}} : \mathbb{R}^6 \to \mathbb{R}^{3\times6}$ is bounded and locally Lipschitz.

In order to achieve the tracking purpose, we introduce a reference system expressed by

$$\dot{x}_r = h_r(x_r) \tag{9}$$

where $x_r = (\boldsymbol{q}_r^{\mathrm{T}}, \boldsymbol{v}_{q_r}^{\mathrm{T}})^{\mathrm{T}}$ and $\dot{\boldsymbol{q}}_r = \dot{\boldsymbol{v}}_r$, $h_r(x_r) : \mathbb{R}^6 \to \mathbb{R}^6$ is a Lipschitz continuous function. By defining the tracking error as $e \triangleq x - x_r$, the tracking error dynamics is described as

$$\begin{aligned} \dot{e} = f(e + x_r) - h_r(x_r) \\ + g(e + x_r)\tau + g(e + x_r)d(\boldsymbol{q}, v_q). \end{aligned} \tag{10}$$

Introduce the following steady-state control τ_r without disturbance corresponding to the desired trajectory x_r:

$$\tau_r(x_r) = g^+(x_r)(h_r(x_r) - f(x_r)) \tag{11}$$

where $g^+(x_r) = (g^{\mathrm{T}}(x_r)g(x_r))^{-1}g^{\mathrm{T}}(x_r)$.

Define a new concatenated state as $z = (e^{\mathrm{T}}, x_r^{\mathrm{T}})^{\mathrm{T}}$, then the dynamics of z is formulated as

$$\dot{z} = F(z) + G(z)u + \xi \tag{12}$$

where

$$\begin{aligned} F(z) &= \begin{pmatrix} f(e + x_r) - h_r(x_r) + g(e + x_r)\tau_r \\ h_r(x_r) \end{pmatrix} \\ G(z) &= \begin{pmatrix} g(e + x_r) \\ \boldsymbol{0}_{6\times3} \end{pmatrix}, \quad u \triangleq \tau - \tau_r, \ \xi = G(z)d(\boldsymbol{q}, v_q). \end{aligned}$$

Obviously, Properties 2 and 3 imply that F is Lipschitz and G is bounded. Assumption 1 and the definition of ξ imply that the uncertain term ξ is still upper bounded since $\|\xi\| = \|G(z)d(\boldsymbol{q}, v_q)\| \leq d_M(e + x_r)\|g(e + x_r)\| \triangleq \lambda_M(z)$.

By the aid of the above system transformation, the robust tracking control problem of system (8) can be converted into an optimal control problem for the corresponding nominal system of (12) expressed as

$$\dot{z} = F(z) + G(z)u. \tag{13}$$

In what follows, the goal is to design an optimal controller $u^*(z)$ for the nominal system (13) so as to make all signals of the closed-loop system (12) be ultimately bounded by minimizing a predefined performance function.

III. Robust Tracking Design Scheme Based on the Online Actor–Critic Algorithm

A. Optimal Control and the Tracking HJB Equation

For the optimal control problem of the nominal system (13), we consider the following discounted performance function:

$$V(z(t)) = \int_t^\infty e^{-\gamma(s-t)}U(z(s), u(s))ds \tag{14}$$

where $\gamma > 0$, $U(z, u) \triangleq \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + u^{\mathrm{T}}Ru$, $R \in \mathbb{R}^{3\times3}$ is a positive-definite symmetric constant matrix, and $\bar{Q}$ is defined as

$$\bar{Q} = \begin{pmatrix} Q & \boldsymbol{0}_{6\times6} \\ \boldsymbol{0}_{6\times6} & \boldsymbol{0}_{6\times6} \end{pmatrix}$$

where $Q \in \mathbb{R}^{6\times6}$ is a positive-definite matrix.

Remark 1: As pointed out in [28], because of the desired control τ_r and disturbance τ_d, the cost function defined in [29] and [30] might be infinity. Therefore, a discount factor is considered in the cost function $V(z(t))$.

Differentiating $V(z(t))$ along with (13) gives

$$\begin{aligned} \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + u^{\mathrm{T}}Ru - \gamma V(z) \\ + \nabla V^{\mathrm{T}}(z)(F(z) + G(z)u) = 0 \end{aligned}$$

where $\nabla V(z) \triangleq ((\partial V(z))/\partial z)$. The Hamiltonian function is given by

$$\begin{aligned} H(z, u, \nabla V) = \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + u^{\mathrm{T}}(z)Ru(z) - \gamma V(z) \\ + \nabla V^{\mathrm{T}}(z)(F(z) + G(z)u(z)). \end{aligned} \tag{15}$$

The optimal value function $V^*(z)$ is defined as

$$V^*(z) = \min_{u\in\pi(\Omega)} \int_t^{\infty} e^{-\gamma(s-t)}\big(\lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + u^{\mathrm{T}}Ru\big)ds \tag{16}$$

where $\pi(\Omega)$ denotes a set of admissible control laws on a set $\Omega \subset \mathbb{R}^{12}$. Then, the optimal value function $V^*(z)$ satisfies the following HJB equation:

$$\begin{aligned} H^*(z, u^*, \nabla V^*) &= \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + u^{*\mathrm{T}}(z)Ru^*(z) - \gamma V^*(z) \\ &\quad + \nabla V^{*\mathrm{T}}(z)(F(z) + G(z)u^*(z)) = 0. \end{aligned} \tag{17}$$

The optimal control is then given as

$$\begin{aligned} u^*(z) &= \arg\min_{u\in\pi(\Omega)} \big[H(z, u, \nabla V^*(z))\big] \\ &= -\frac{1}{2}R^{-1}G^{\mathrm{T}}(z)\nabla V^*(z). \end{aligned} \tag{18}$$

Taking the optimal control policy (18) into (17), the tracking HJB equation (17) becomes

$$\begin{aligned} &H^*(z, \nabla V^*) \\ &\quad = \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z - \gamma V^*(z) - \frac{1}{4}\nabla V^{*\mathrm{T}}(z)G(z) \\ &\qquad \times R^{-1}G^{\mathrm{T}}(z)\nabla V^*(z) + \nabla V^{*\mathrm{T}}(z)F(z) = 0. \end{aligned} \tag{19}$$

To obtain the optimal control (18), one needs to solve the tracking HJB equation (19) for the optimal value function V^*. However, finding the solution of the HJB equation (19) is extremely difficult or impossible. In Section III-B, an online actor–critic algorithm is applied to solve this equation.

B. Implementation Based on the Online Actor–Critic Algorithm

In this section, an online actor–critic algorithm which involves tuning both critic and actor NNs simultaneously in real-time is used to find an approximate solution of the HJB equation.

According to the Weierstrass high-order approximation theorem [35], the value function $V^*(z)$ can be approximated by using a single-layer NN as

$$V^*(z) = W^{\mathrm{T}}\phi(z) + \varepsilon_v(z) \tag{20}$$

where $W \in \mathbb{R}^N$ is an ideal constant weight vector and bounded by a known positive constant $\bar{W}$ such that $\|W\| \le \bar{W}$, $\phi(z) \in \mathbb{R}^N$ is a suitable activation function vector, N is the number of neurons, $\varepsilon_v(z)$ is the approximation error. The related gradient vector of $V^*(z)$ is

$$\nabla V^*(z) = \nabla\phi^{\mathrm{T}}(z)W + \nabla\varepsilon_v(z). \tag{21}$$

It is known that, on the compact set Ω, the approximation error ε_v and its gradient $\nabla\varepsilon_v$ are bounded so that $\|\varepsilon_v\| \le b_\varepsilon$ and $\|\nabla\varepsilon_v\| \le b_{\varepsilon z}$ [28]. For the activation functions $\phi(z)$, we make the following assumption.

Assumption 2: The activation functions $\phi(z)$ and their gradients are bounded, i.e., $\|\phi(z)\| \le b_\phi$, $\|\nabla\phi(z)\| \le b_{\phi z}$.

Using (20) and (21) in (19), the optimal Hamiltonian can be represented by

$$\begin{aligned} H^*(z, W) &= \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + W^{\mathrm{T}}\nabla\phi F - \gamma W^{\mathrm{T}}\phi \\ &\quad - \frac{1}{4}W^{\mathrm{T}}D_1W + \varepsilon_H = 0 \end{aligned} \tag{22}$$

where $D_1 = \nabla\phi GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}$, $D_2 = GR^{-1}G^{\mathrm{T}}$ and

$$\begin{aligned} \varepsilon_H &= \nabla\varepsilon_v^{\mathrm{T}}F - \frac{1}{2}\nabla\varepsilon_v^{\mathrm{T}}GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}W \\ &\quad - \frac{1}{4}\nabla\varepsilon_v^{\mathrm{T}}D_2\nabla\varepsilon_v - \gamma\varepsilon_v(z). \end{aligned} \tag{23}$$

There exists a constant bound ε_h such that $\sup\|\varepsilon_H\| \le \varepsilon_h$ as the number of hidden layer units N increases [21]. The optimal control (18) becomes

$$u^*(z) = -\frac{1}{2}R^{-1}G^{\mathrm{T}}(z)\big(\nabla\phi^{\mathrm{T}}(z)W + \nabla\varepsilon_v(z)\big). \tag{24}$$

Based on (20) and (24), the critic NN approximation for the optimal value function and the actor NN approximation for the optimal policy are given by

$$\hat{V}\big(z, \hat{W}_c\big) = \hat{W}_c^{\mathrm{T}}\phi(z) \tag{25}$$

$$\hat{u}\big(z, \hat{W}_a\big) = -\frac{1}{2}R^{-1}G^{\mathrm{T}}(z)\nabla\phi^{\mathrm{T}}(z)\hat{W}_a \tag{26}$$

where $\hat{W}_c$ and $\hat{W}_a$ are the estimates of the ideal weights W. From (11), the tracking controller τ of MWMR can be obtained as

$$\tau = -\frac{1}{2}R^{-1}G^{\mathrm{T}}(z)\nabla\phi^{\mathrm{T}}(z)\hat{W}_a + g^{+}(x_r)(h_r(x_r) - f(x_r)).$$

Taking the approximations $\hat{u}$ and $\hat{V}$ into (15), the approximate Hamiltonian is derived as

$$\begin{aligned} &\hat{H}\big(z, \hat{W}_c, \hat{W}_a\big) \\ &\quad = \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + \frac{1}{4}\hat{W}_a^{\mathrm{T}}D_1\hat{W}_a \\ &\qquad - \gamma\hat{W}_c^{\mathrm{T}}\phi + \hat{W}_c^{\mathrm{T}}\nabla\phi\left(F - \frac{1}{2}GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_a\right). \end{aligned} \tag{27}$$

Define the Bellman error as $\delta \triangleq \hat{H} - H^*$, from (22) and (27), one has

$$\begin{aligned} \delta\big(z, \hat{W}_c, \hat{W}_a\big) &= \lambda_M^2(z) + z^{\mathrm{T}}\bar{Q}z + \frac{1}{4}\hat{W}_a^{\mathrm{T}}D_1\hat{W}_a \\ &\quad + \hat{W}_c^{\mathrm{T}}(\nabla\phi(F + G\hat{u}) - \gamma\phi). \end{aligned} \tag{28}$$

To solve the optimal control problem, a normalized least squares tuning law with an exponential forgetting factor [29] is designed for training critic NN to minimize the integral error $E_c = \int_0^t \delta^2(s)ds$ as follows:

$$\begin{aligned} \dot{\hat{W}}_c(t) &= -\eta_c\Gamma(t)\frac{\sigma(t)}{m_\sigma^2(t)}\delta(t) \\ \dot{\Gamma}(t) &= -\eta_c\left(-\beta\Gamma(t) + \Gamma(t)\frac{\sigma(t)\sigma^{\mathrm{T}}(t)}{m_\sigma^2(t)}\Gamma(t)\right) \end{aligned} \tag{29}$$

where $\sigma \triangleq \nabla\phi(F + G\hat{u}) - \gamma\phi$, $m_\sigma \triangleq (1 + \nu\sigma^{\mathrm{T}}\Gamma\sigma)^{1/2}$, η_c, ν are positive constant, $\beta \in (0, 1)$ is the constant forgetting factor, $\Gamma(t)$ is the estimation gain matrix.

Define the critic weight estimation error as $\tilde{W}_c \triangleq W - \hat{W}_c$ and the actor NN estimation error as $\tilde{W}_a \triangleq W - \hat{W}_a$, the Bellman error δ can be rewritten as

$$\delta = -\tilde{W}_c^{\mathrm{T}}\sigma + \frac{1}{4}\tilde{W}_a^{\mathrm{T}}D_1\tilde{W}_a - \varepsilon_H. \tag{30}$$

Let $\bar{\sigma} = (\sigma/m_\sigma)$ and using (30) in (29), the critic weight estimation error dynamics can be expressed as

$$\begin{aligned} \dot{\tilde{W}}_c &= \dot{W} - \dot{\hat{W}}_c \\ &= -\eta_c\Gamma\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_c + \eta_c\Gamma\frac{\bar{\sigma}}{m_\sigma}\left(\frac{1}{4}\tilde{W}_a^{\mathrm{T}}D_1\tilde{W}_a - \varepsilon_H\right). \end{aligned} \tag{31}$$

The following persistently exciting (PE) assumption [21], [29] is imposed on signal $\bar{\sigma}(t)$ to facilitate the convergence of $\hat{W}_c$ to W.

Assumption 3: There exist $T > 0$, $\beta_1 > 0$, $\beta_2 > 0$ such that for all t

$$\beta_1\boldsymbol{I} \le \int_t^{t+T} \bar{\sigma}(s)\bar{\sigma}^{\mathrm{T}}(s)ds \le \beta_2\boldsymbol{I}.$$

From Assumption 2 and [29], there exist $\alpha_2 > \alpha_1 > 0$ such that

$$\alpha_1 \boldsymbol{I} \leq \Gamma(t) \leq \alpha_2 \boldsymbol{I} \quad \forall t \in [0, \infty). \tag{32}$$

Then, based on (32), it can be concluded that

$$\|\bar{\sigma}(t)\| \leq \frac{1}{\sqrt{\nu \alpha_1}} \quad \forall t \in [0, \infty). \tag{33}$$

The tuning law for the actor NN is developed as

$$\dot{\hat{W}}_a = -\eta_{a1}\big(\hat{W}_a - \hat{W}_c\big) - \eta_{a2}\hat{W}_a + \frac{\eta_a}{4} D_1 \hat{W}_a \frac{\bar{\sigma}^{\mathrm{T}}}{m_\sigma}\hat{W}_c \tag{34}$$

where $\eta_{a1} > 0$, $\eta_{a2} > 0$ and $\eta_a > 0$ are tuning parameters.

Remark 2: Since the Bellman error is nonlinear with respect to the policy weight estimates, the least-squares approach cannot be used to update the policy weights [29]. Define the error $\tilde{u} = \hat{u}(z, \hat{W}_a) - \hat{u}_c(z, \hat{W}_c) = -(1/2)R^{-1}G^{\mathrm{T}}(z)\nabla\phi^{\mathrm{T}}(z)(\hat{W}_a - \hat{W}_c)$ and the objective function $E_u = (1/2)\tilde{u}^{\mathrm{T}}R\tilde{u}$, in order to minimize the objective function, we can employ the gradient-descent algorithm to tune the actor NN weights [28], i.e., $\dot{\hat{W}}_a = -\eta_{a1}(E_u/\hat{W}_a) = -\eta_{a1}D_1(\hat{W}_a - \hat{W}_c)$. Therefore, (34) can be seen as a modified tuning law for the actor weights, and the second term on the right-hand side of (34) is to improve robustness, the third term on the right-hand side of (34) is designed to guarantee the stability of the closed-loop system in a Lyapunov sense.

C. Stability Analysis and Robust Trajectory Tracking Property

Theorem 1: Considering the nominal augmented system (13), let the critic NN and the control input be given by (25) and (26). Let the tuning laws for the critic and actor NNs be provided by (29) and (34), respectively. Let Assumptions 1–3 hold and $\bar{\sigma}$ in (31) be PE. Then, the state of the closed-loop system, the critic NN error $\tilde{W}_c$ and the actor NN error $\tilde{W}_a$ are ultimately bounded. Moreover, the critic weight estimation error $\tilde{W}_c$ converges exponentially to a residual set.

Proof: Consider the candidate Lyapunov function

$$\begin{aligned} V(t) = V^*(t) + V_1(t) + V_2(t) = V^*(t)& \\ + \frac{1}{2\eta_c}\tilde{W}_c^{\mathrm{T}}(t)\Gamma^{-1}(t)\tilde{W}_c(t) + \frac{1}{2\eta_a}\tilde{W}_a^{\mathrm{T}}(t)\tilde{W}_a(t)& \end{aligned} \tag{35}$$

where $V^*(t)$ is the optimal value function. The derivative of the first term is

$$\begin{aligned} \dot{V}^*(t) &= \nabla V^{*\mathrm{T}}(F + G\hat{u}) \\ &= \big(\nabla\phi^{\mathrm{T}}W + \nabla\varepsilon_\upsilon\big)^{\mathrm{T}}\left(F - \frac{1}{2}GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_a\right) \\ &= W^{\mathrm{T}}\nabla\phi F - \frac{1}{2}W^{\mathrm{T}}D_1\hat{W}_a \\ &\quad + \nabla\varepsilon_\upsilon^{\mathrm{T}}\left(F - \frac{1}{2}GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_a\right). \end{aligned}$$

From the HJB equation (22), one has

$$\begin{aligned} \dot{V}^*(t) = -\lambda_M^2(z) - z^{\mathrm{T}}\bar{Q}z + \gamma\, W^{\mathrm{T}}\phi + \frac{1}{4}W^{\mathrm{T}}D_1W& \\ - \frac{1}{2}W^{\mathrm{T}}D_1\hat{W}_a - \varepsilon_H + \varepsilon_1& \end{aligned} \tag{36}$$

where $\varepsilon_1 = \dot{\varepsilon}_\upsilon = \nabla\varepsilon_\upsilon^{\mathrm{T}}(F - (1/2)GR^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_a)$, from (23)

$$\varepsilon_1 - \varepsilon_H = \frac{1}{2}\tilde{W}_a^{\mathrm{T}}\nabla\phi D_2\nabla\varepsilon_\upsilon + \frac{1}{4}\nabla\varepsilon_\upsilon^{\mathrm{T}}D_2\nabla\varepsilon_\upsilon + \gamma\,\varepsilon_\upsilon(z).$$

Then, (36) can be rewritten as

$$\begin{aligned} \dot{V}^*(t) &= -\lambda_M^2(z) - z^{\mathrm{T}}\bar{Q}z + \gamma\, W^{\mathrm{T}}\phi - \frac{1}{4}W^{\mathrm{T}}D_1W \\ &\quad + \frac{1}{2}\tilde{W}_a^{\mathrm{T}}D_1W + \frac{1}{2}\tilde{W}_a^{\mathrm{T}}\nabla\phi D_2\nabla\varepsilon_\upsilon + \frac{1}{4}\nabla\varepsilon_\upsilon^{\mathrm{T}}D_2\nabla\varepsilon_\upsilon \\ &\quad + \gamma\,\varepsilon_\upsilon(z). \end{aligned}$$

Substituting (29) and (31) into the time derivative of the second term gives

$$\begin{aligned} \dot{V}_1(t) &= \frac{1}{\eta_c}\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}\dot{\tilde{W}}_c + \frac{1}{2\eta_c}\tilde{W}_c^{\mathrm{T}}\dot{\Gamma}^{-1}\tilde{W}_c \\ &= \tilde{W}_c^{\mathrm{T}}\Gamma^{-1}\left(-\Gamma\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_c + \Gamma\frac{\bar{\sigma}}{m}\left(\frac{1}{4}\tilde{W}_a^{\mathrm{T}}D_1\tilde{W}_a - \varepsilon_H\right)\right) \\ &\quad - \frac{1}{2}\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}\big(\beta\Gamma - \Gamma\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\Gamma\big)\Gamma^{-1}\tilde{W}_c \\ &= -\frac{1}{2}\tilde{W}_c^{\mathrm{T}}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_c - \frac{\beta}{2}\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}\tilde{W}_c \\ &\quad + \tilde{W}_c^{\mathrm{T}}\frac{\bar{\sigma}}{m}\left(\frac{1}{4}\tilde{W}_a^{\mathrm{T}}D_1\tilde{W}_a - \varepsilon_H\right). \end{aligned} \tag{37}$$

For the last term $V_2(t)$, using (34), and $\hat{W}_c = W - \tilde{W}_c$ and $\hat{W}_a = W - \tilde{W}_a$, one has

$$\begin{aligned} \dot{V}_2(t) &= -\frac{1}{\eta_a}\tilde{W}_a^{\mathrm{T}}\dot{\hat{W}}_a = -\frac{(\eta_{a1} + \eta_{a2})}{\eta_a}\tilde{W}_a^{\mathrm{T}}\tilde{W}_a \\ &\quad + \frac{\eta_{a1}}{\eta_a}\tilde{W}_a^{\mathrm{T}}\tilde{W}_c + \frac{\eta_{a2}}{\eta_a}\tilde{W}_a^{\mathrm{T}}W - \frac{1}{4}\tilde{W}_a^{\mathrm{T}}D_1\hat{W}_a\frac{\bar{\sigma}^{\mathrm{T}}}{m_\sigma}\hat{W}_c. \end{aligned} \tag{38}$$

Using (36)–(38) into the derivative of $V(t)$ yields

$$\begin{aligned} \dot{V}(t) &= -\lambda_M^2(z) - e^{\mathrm{T}}Qe - \frac{1}{4}W^{\mathrm{T}}D_1W - \frac{\tilde{W}_c^{\mathrm{T}}\bar{\sigma}}{m_\sigma}\varepsilon_H \\ &\quad - \tilde{W}_c^{\mathrm{T}}A_1\tilde{W}_c - \tilde{W}_a^{\mathrm{T}}A_2\tilde{W}_a + \tilde{W}_a^{\mathrm{T}}A_3\tilde{W}_c \\ &\quad + \tilde{W}_a^{\mathrm{T}}B_1 + B_2 \end{aligned} \tag{39}$$

where

$$\begin{aligned} A_1 &= \frac{1}{2}\bar{\sigma}\bar{\sigma}^{\mathrm{T}} + \frac{\beta}{2}\Gamma^{-1}, \quad A_2 = \frac{(\eta_{a1} + \eta_{a2})}{\eta_a}\boldsymbol{I} - \frac{1}{4}D_1\frac{\bar{\sigma}^{\mathrm{T}}}{m_\sigma}W \\ A_3 &= \frac{1}{4}D_1W\frac{\bar{\sigma}^{\mathrm{T}}}{m_\sigma} + \frac{\eta_{a1}}{\eta_a}\boldsymbol{I} \\ B_1 &= \frac{1}{2}D_1W + \frac{1}{2}\nabla\phi D_2\nabla\varepsilon_\upsilon - \frac{1}{4}D_1W\frac{\bar{\sigma}^{\mathrm{T}}}{m_\sigma}W + \frac{\eta_{a2}}{\eta_a}W \\ B_2 &= \gamma\, W^{\mathrm{T}}\phi + \frac{1}{4}\nabla\varepsilon_\upsilon^{\mathrm{T}}D_2\nabla\varepsilon_\upsilon + \gamma\,\varepsilon_\upsilon(z). \end{aligned}$$

According to Properties 2 and 3, Assumption 2 and the boundness of ε_υ and $\nabla\varepsilon_\upsilon$, there exist positive constants κ_1, κ_2, κ_3 such that $\|B_1\| \leq \kappa_1$, $\|B_2\| \leq \kappa_2$, $\|(1/4)D_1(\bar{\sigma}^{\mathrm{T}}/m_\sigma)W\| \leq \kappa_3$. Using the Young's inequality, (39) becomes

$$\begin{aligned} \dot{V}(t) &\leq -\underline{q}\|e\|^2 - \frac{\beta}{4\alpha_2}\big\|\tilde{W}_c\big\|^2 - \frac{\eta_{a1} + \eta_{a2}}{2\eta_a}\big\|\tilde{W}_a\big\|^2 \\ &\quad - \left(\frac{\beta}{4\alpha_2} - \frac{1}{2\epsilon}\left(\frac{\eta_{a1}}{\eta_a} + \kappa_3\right)\right)\big\|\tilde{W}_c\big\|^2 \\ &\quad - \left(\frac{\eta_{a1} + \eta_{a2}}{2\eta_a} - \kappa_3 - \frac{\epsilon}{2}\left(\frac{\eta_{a1}}{\eta_a} + \kappa_3\right)\right) \\ &\quad \times \big\|\tilde{W}_a\big\|^2 + \frac{\varepsilon_h}{\sqrt{\nu\alpha_1}}\big\|\tilde{W}_c\big\| + \kappa_1\big\|\tilde{W}_a\big\| + \kappa_2 \end{aligned} \tag{40}$$

where $\underline{q} \triangleq \lambda_{\min}(Q)$. Choose the parameters such that $(\beta/4\alpha_2) \geq (1/2\epsilon)((\eta_{a1}/\eta_a) + \kappa_3)$ and $((\eta_{a1} + \eta_{a2})/2\eta_a) \geq \kappa_3 + (\epsilon/2)((\eta_{a1}/\eta_a) + \kappa_3)$ and let $\zeta = (e^{\mathrm{T}}, \tilde{W}_c^{\mathrm{T}}, \tilde{W}_a^{\mathrm{T}})^{\mathrm{T}}$, $\varpi_1 = \min\{\underline{q}, (\beta/4\alpha_2), ((\eta_{a1} + \eta_{a2})/2\eta_a)\}$ and $\varpi_2 = \max\{(\varepsilon_h/\sqrt{\nu\alpha_1}), \kappa_1\}$, (40) becomes

$$\dot{V}(t) \leq -\varpi_1\|\zeta\|^2 + \varpi_2\|\zeta\| + \kappa_2.$$

Therefore, if the inequality $\|\zeta\| > (\varpi_2/2\varpi_1) + ((\varpi_2^2/4\varpi_1^2) + (\kappa_2/\varpi_1))^{1/2}$ holds, it can be derived that the Lyapunov derivative $\dot{V}(t)$ is negative. Using the standard Lyapunov extension theorem imply that the closed-loop system state and the weights estimation errors $\tilde{W}_c$, $\tilde{W}_a$ are ultimately bounded. ■

Applying (18) and (26) gives

$$u^* - \hat{u} = -\frac{1}{2}R^{-1}G^{\mathrm{T}}\big(\nabla\phi^{\mathrm{T}}(z)\tilde{W}_a + \nabla\varepsilon_v\big).$$

According to Theorem 1 and Assumption 2, one can get

$$\|u^* - \hat{u}\| \le \frac{1}{2\lambda_{\min}(R)}b_g\big(b_{\phi z}b_\zeta + b_{\varepsilon z}\big) \triangleq b_u \tag{41}$$

where b_u is a positive constant.

Theorem 2: For the nominal system (13) with the performance function (14), the approximate optimal control law (26) ensures the tracking error dynamics in the closed form of uncertain augmented system (12) is ultimately bounded provided that $\lambda_M^2(z) \ge \tau_d^{\mathrm{T}}R\tau_d$.

Proof: Let the optimal value function $V^*(z)$ be a Lyapunov function candidate. From (18), we have $\nabla V^{*\mathrm{T}}G = -2u^{*\mathrm{T}}R$. It follows from (17) that:

$$\nabla V^{*\mathrm{T}}(F + Gu^*) = -\lambda_M^2(z) - z^{\mathrm{T}}\bar{Q}z - u^{*\mathrm{T}}Ru^* + \gamma V^*.$$

Then, taking the time derivative of $V^*(z)$ along the system (12) with the approximate optimal control law $\hat{u}$, one has

$$\begin{aligned}
&\dot{V}^*(z)\\
&\quad= \nabla V^{*\mathrm{T}}\big(F + G\hat{u} + Gd(\boldsymbol{q}, v_q)\big)\\
&\quad= \nabla V^{*\mathrm{T}}(F + Gu^*) + \nabla V^{*\mathrm{T}}G(\hat{u} - u^*) - 2u^{*\mathrm{T}}Rd(\boldsymbol{q}, v_q)\\
&\quad= -\big(\lambda_M^2(z) - d^{\mathrm{T}}(\boldsymbol{q}, v_q)Rd(\boldsymbol{q}, v_q)\big) + \nabla V^{*\mathrm{T}}G(\hat{u} - u^*)\\
&\qquad - z^{\mathrm{T}}\bar{Q}z - \big(u^* + d(\boldsymbol{q}, v_q)\big)^{\mathrm{T}}R\big(u^* + d(\boldsymbol{q}, v_q)\big) + \gamma V^*.
\end{aligned}$$

By considering (20), (21), (41), and Assumption 2, we have

$$\begin{aligned}
\big\|\nabla V^{*\mathrm{T}}G(\hat{u} - u^*)\big\| &\le \big(b_{\phi z}\bar{W} + b_{\varepsilon z}\big)b_g b_u\\
\|\gamma V^*\| &\le \gamma\,\big(\bar{W}b_\phi + b_\varepsilon\big).
\end{aligned}$$

Combining the above inequalities with the condition $\lambda_M^2(z) \ge d^{\mathrm{T}}(\boldsymbol{q}, v_q)Rd(\boldsymbol{q}, v_q)$, one can derive that

$$\dot{V}^*(z) \le -\underline{q}\|e\|^2 + \lambda_u$$

where $\lambda_u \triangleq (b_{\phi z}\bar{W}+b_{\varepsilon z})b_g b_u+\gamma\,(\bar{W}b_\phi+b_\varepsilon)$. It can be concluded that $\dot{V}^* < 0$ if $\|e\| > (\lambda_u/\underline{q})^{1/2}$. Thus, the tracking error dynamics of the uncertain augmented system (12) is uniformly ultimately bounded, which further shows that the state $\boldsymbol{q}(t)$ can track the desired trajectory $\boldsymbol{q}_r(t)$. This completes the proof. ■

IV. Simulation

In this part, simulation results are presented to evaluate the performance of the algorithm. The reference trajectory is given by $q_r = (0.5\cos(2t)\ \ \cos(t)\ \ \cos(0.5t))^{\mathrm{T}}$. The initial conditions are $x(0) = (1.8\ \ 1.6\ \ 1.4\ \ 0\ \ 0\ \ 0)^{\mathrm{T}}$, $\hat{W}_c(0) = 200 \times \mathbf{1}_{51}$, $\hat{W}_a(0) = 6 \times \mathbf{1}_{51}$ and $\Gamma(0) = 10 \times \boldsymbol{I}_{51}$. The activation function vector is chosen as $\phi(z) = (1/2)[z_1^2\ z_2^2\ z_3^2\ z_1z_4\ z_1z_5\ z_1z_6\ z_2z_4\ z_2z_5\ z_2z_6\ z_3z_4\ z_3z_5\ z_3z_6\ z_1^2z_2^2\ z_1^2z_3^2\ z_2^2z_3^2\ z_1^2z_7^2\ z_1^2z_8^2\ z_1^2z_9^2\ z_1^2z_{10}^2\ z_1^2z_{11}^2\ z_1^2z_{12}^2\ z_2^2z_7^2\ z_2^2z_8^2\ z_2^2z_9^2\ z_2^2z_{10}^2\ z_2^2z_{11}^2\ z_2^2z_{12}^2\ z_3^2z_7^2\ z_3^2z_8^2\ z_3^2z_9^2\ z_3^2z_{10}^2\ z_3^2z_{11}^2\ z_3^2z_{12}^2\ z_4^2z_7^2\ z_4^2z_8^2\ z_4^2z_9^2\ z_4^2z_{10}^2\ z_4^2z_{11}^2\ z_4^2z_{12}^2\ z_5^2z_7^2\ z_5^2z_8^2\ z_5^2z_9^2\ z_5^2z_{10}^2\ z_5^2z_{11}^2\ z_5^2z_{12}^2\ z_6^2z_7^2\ z_6^2z_8^2\ z_6^2z_9^2\ z_6^2z_{10}^2\ z_6^2z_{11}^2\ z_6^2z_{12}^2]^{\mathrm{T}}$. The parameters of MWMR are shown in Table I.

Considering the external disturbances induced by slipping, simulations are performed for different values of ρ, let $\rho = \pm 0.1$. For $\rho = 0.1$, we choose $Q = \boldsymbol{I}_6$ and $R = \boldsymbol{I}_3$, the control gains are $\eta_c = 6.37$, $\eta_{a1} = 7.8$, $\eta_{a2} = 0.001$, $\eta_a = 0.001$, $\beta = 0.001$, $\gamma = 0.0001$, $\nu = 0.002$. For $\rho = -0.1$, the control gains are $\eta_c = 2$, $\eta_{a1} = 1.2$, $\eta_{a2} = 0.001$, $\eta_a = 0.001$, $\beta = 0.001$, $\gamma = 0.0001$, $\nu = 0.002$.

Simulation results are shown in Figs. 3–7. Figs. 3 and 5 show the tracking performance of MWMR under disturbance. Figs. 4 and 6 show NN weights. A probing noise $N(t) = \sin^2(t)\cos(t) + \sin^2(2t)\cos(0.1t) + \sin^2(-1.2t)\cos(0.5t) + \sin^5(t) + \sin^2(1.2t) + \cos(2.4t)\sin^3(2.4t)$ is added to the control input to ensure PE condition and the probing noise affects the system states and the NN weights. The probing noise is turned off at 50 s and the training process lasts 150 s. After 50 s, we can see that the convergence of the NN weights has occurred and the system states are close to reference trajectories. This shows that the PE condition is effectively guaranteed

TABLE I
Parameters of the MWMR

Parameters	Value
Mass (kg)	$m = 10$
Length (m)	$l = 0.5,\ r = 0.05$
Rotational inertia (kg · m^2)	$I_\theta = 5,\ I_\varphi = 0.1$
Friction coefficients (N·m/rad/s)	$\mu = 1$

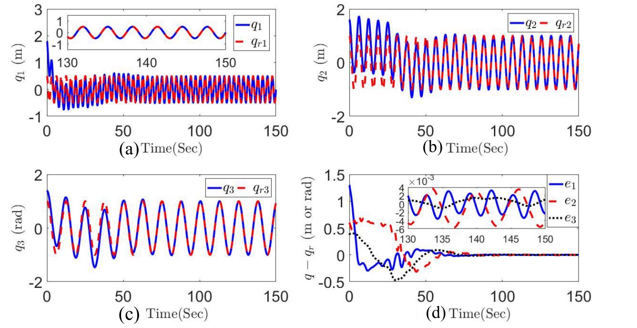

Fig. 3. (a)–(c) Tracking performance of q_1, q_2, and q_3. (d) Time responses of the tracking error $q(t) - q_r(t)$ when $\rho = 0.1$.

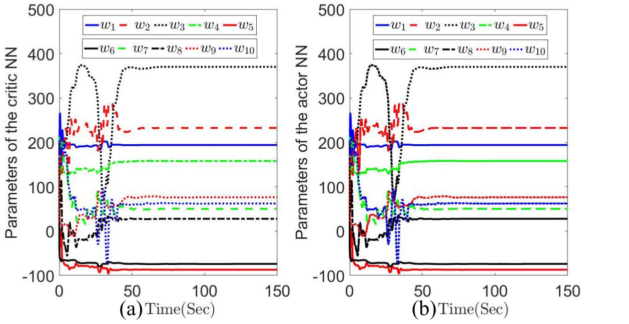

Fig. 4. Weights of the critic and the actor networks when $\rho = 0.1$. (a) Convergence of the critic NN weight $\hat{W}_c$. (b) Convergence of the actor NN weight $\hat{W}_a$.

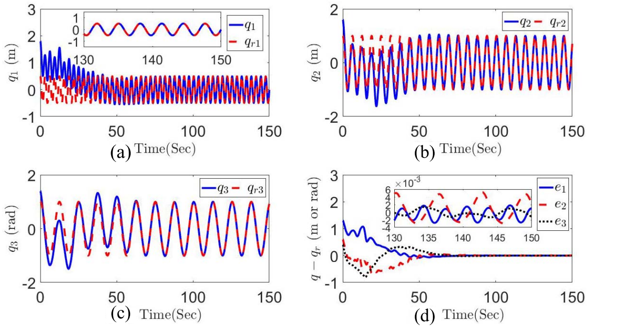

Fig. 5. (a)–(c) Tracking performance of q_1, q_2, and q_3. (d) Time responses of the tracking error $q(t) - q_r(t)$ when $\rho = -0.1$.

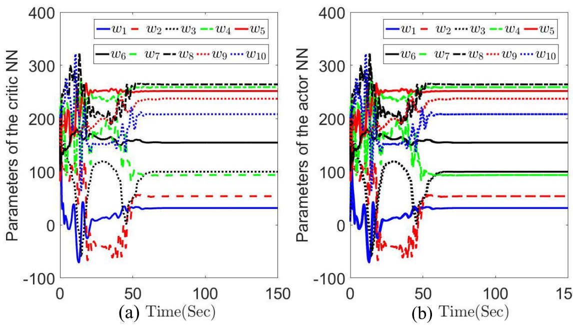

Fig. 6. Weights of the critic and the actor networks when $\rho = -0.1$. (a) Convergence of the critic NN weight $\hat{W}_c$. (b) Convergence of the actor NN weight $\hat{W}_a$.

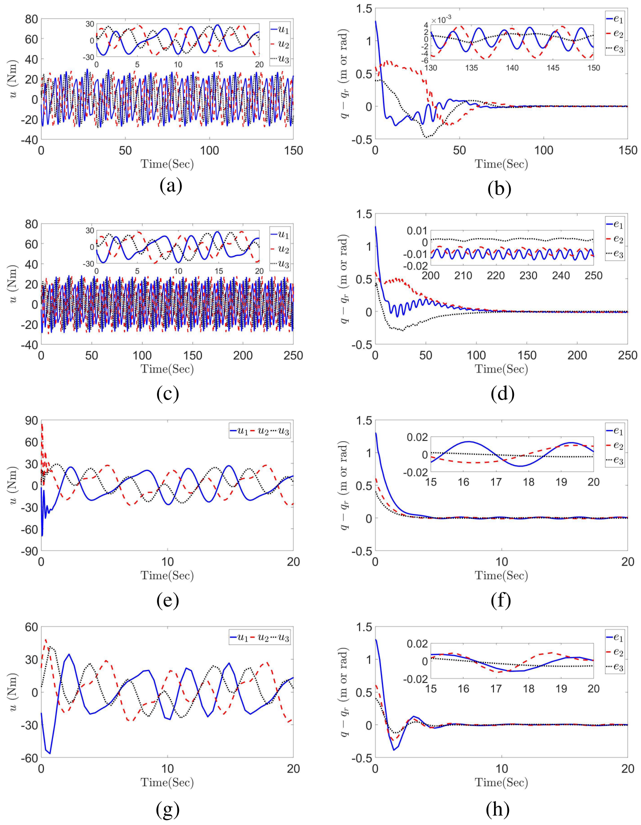

Fig. 7. Comparisons of the tracking error and control input when $\rho = 0.1$. (a) Control input of the designed algorithm. (b) Tracking error of the designed algorithm. (c) Control input of the actor–critic algorithm. (d) Tracking error of the actor–critic algorithm. (e) Control input of robust NN PD algorithm. (f) Tracking error of robust NN PD algorithm. (g) Control input of backstepping. (h) Tracking error of backstepping.

by adding the probing noise. The PE condition is no longer needed on convergence and thus the probing noise was turned off at 50 s. Then, the system state continues to approach reference trajectory.

It can be seen from Figs. 3 and 5 that the MWMR state can track the desired trajectory successfully based on the developed learning algorithm, and the tracking error is within the error range $(-0.006, 0.004)$ and $(-0.004, 0.006)$ when $\rho = 0.1$ and $\rho = -0.1$. In Fig. 4, the first ten critic and actor weights finally converge to $\hat{W}_c = [194.0781\ 232.5053\ 370.2461\ 158.3349\ -87.1293\ -73.9338\ 50.0400\ 27.7158\ 76.4017\ 62.4485]^{\mathrm{T}}$ and $\hat{W}_a = [194.1028\ 232.5361\ 370.2934\ 158.3591\ -87.1408\ -73.9436\ 50.0616\ 27.7184\ 76.4075\ 62.4525]^{\mathrm{T}}$, respectively. In Fig. 6, the first ten critic and actor weights finally converge to $\hat{W}_c = [31.7110\ 53.8465\ 100.0895\ 259.0001\ 251.6835\ 154.7164\ 93.9407\ 264.3532\ 237.1995\ 208.2481]^{\mathrm{T}}$ and $\hat{W}_a = [31.7110\ 53.8470\ 100.0895\ 259.0006\ 251.6833\ 154.7165\ 93.9511\ 264.3536\ 237.1984\ 208.2472]^{\mathrm{T}}$, respectively.

Fig. 7 shows the comparisons of the tracking error and control input u for $\rho = 1$. It can be observed from Fig. 7 that the initial value of input torques u under the designed controller is smaller than that under backstepping controller and robust NN PD controller [36]. This reflects the optimality of the designed method. And the tracking error for the proposed control law is smaller than both the backstepping technique and the robust NN PD control. Compared with the actor–critic algorithm in [29], a fast convergence result is achieved and the tracking error is smaller based on our approach.

V. Conclusion

The robust trajectory tracking control problem for an MWMR subject to external disturbance was addressed by using model-based RL algorithm. Via system transformation, the tracking control problem of the original uncertain system was converted into an optimal control problem of a nominal augmented system. By defining an appropriate performance index, the approximate optimal control law was solved by employing an online actor–critic synchronous learning algorithm, which can be applied to make the MWMR track the desired trajectory. Our future efforts are aimed at investigating the tracking control and obstacle avoidance problems for some robots subject to random disturbances by using RL algorithm.

References

[1] R. Fierro and F. L. Lewis, “Control of a nonholonomic mobile robot using neural networks,” *IEEE Trans. Neural Netw.*, vol. 9, no. 4, pp. 589–600, Jul. 1998.

[2] A. Lazinica, *Mobile Robots: Towards New Applications*. I-Tech Education and Publishing, 2006.

[3] R. Siegwart, I. Nourbakhsh, and D. Scaramuzza, *Introduction to AutonomousMobile Robots* (Intelligent Robotics and Autonomous Agents Series). Cambridge, MA, USA: MIT Press, 2011.

[4] G. Campion, G. Bastin, and B. Dandrea-Novel, “Structural properties and classification of kinematic and dynamic models of wheeled mobile robots,” *IEEE Trans. Robot. Autom.*, vol. 12, no. 1, pp. 47–62, Feb. 1996.

[5] K. L. Moore and N. S. Flann, “A six-wheeled omnidirectional autonomous mobile robot,” *IEEE Control Syst.*, vol. 20, no. 6, pp. 53–66, Dec. 2000.

[6] H.-C. Huang and C.-C. Tsai, “Adaptive trajectory tracking and stabilization for omnidirectional mobile robot with dynamic effect and uncertainties,” *IFAC Proc. Volumes*, vol. 41, no. 2, pp. 5383–5388, 2008.

[7] T. D. Viet, P. T. Doan, N. Hung, H. K. Kim, and S. B. Kim, “Tracking control of a three-wheeled omnidirectional mobile manipulator system with disturbance and friction,” *J. Mech. Sci. Technol.*, vol. 26, no. 7, pp. 2197–2211, Jul. 2012.

[8] L.-C. Lin and H.-Y. Shih, “Modeling and adaptive control of an omni-Mecanum-wheeled robot,” *Intell. Control Autom.*, vol. 4, no. 2, pp. 166–179, 2013.

[9] V. Alakshendra and S. S. Chiddarwar, “Adaptive robust control of Mecanum-wheeled mobile robot with uncertainties,” *Nonlinear Dyn.*, vol. 87, no. 4, pp. 2147–2169, Mar. 2017.

[10] D. Wang *et al.*, “Formation control of multiple Mecanum-wheeled mobile robots with physical constraints and uncertainties,” *Appl. Intell.*, vol. 52, no. 3, pp. 2510–2529, Jun. 2021.

[11] L. Ovalle, H. Ríos, M. Llama, V. Santibáñez, and A. Dzul, “Omnidirectional mobile robot robust tracking: Sliding-mode output-based control approaches,” *Control Eng. Pract.*, vol. 85, pp. 50–58, Apr. 2019.

[12] J.-T. Huang, T. Van Hung, and M.-L. Tseng, “Smooth switching robust adaptive control for omnidirectional mobile robots,” *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1986–1993, Sep. 2015.

[13] C. Ren, Y. Sun, and S. Ma, “Passivity-based control of an omnidirectional mobile robot,” *Robot. Biomimetics*, vol. 3, no. 1, pp. 1–9, Jul. 2016.

[14] C. Wang, X. Liu, X. Yang, F. Hu, A. Jiang, and C. Yang, "Trajectory tracking of an omni-directional wheeled mobile robot using a model predictive control strategy," *Appl. Sci.*, vol. 8, no. 2, p. 231, Feb. 2018.

[15] R. L. Williams, B. E. Carter, and G. Giulio, "Dynamic model with slip for wheeled omnidirectional robots," *IEEE Trans. Robot. Autom.*, vol. 18, no. 3, pp. 285–293, Jun. 2002.

[16] D. Wang and C. B. Low, "Modeling and analysis of skidding and slipping in wheeled mobile robots: Control design perspective," *IEEE Trans. Robot.*, vol. 24, no. 3, pp. 676–687, Jun. 2008.

[17] Y. Anzai, *Pattern Recognition and Machine Learning*. Amsterdam, Netherlands: Elsevier, 2012.

[18] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[19] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control* (Advances in Industrial Control). Berlin, Germany: Springer, 2017.

[20] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

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

[22] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, Jan. 2013.

[23] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[24] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, Feb. 2016.

[25] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

[26] D. Liu, X. Yang, D. Wang, and Q. Wei, "Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1372–1385, Jul. 2015.

[27] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 937–942, Aug. 2008.

[28] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[29] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[30] D. Wang and C. Mu, "Adaptive-critic-based robust trajectory tracking of uncertain dynamics and its application to a spring–mass–damper system," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 654–663, Jan. 2018.

[31] R. Xia, Q. Wu, and S. Shao, "Disturbance observer-based optimal flight control of near space vehicle with external disturbance," *Trans. Inst. Meas. Control*, vol. 42, no. 2, pp. 272–284, Jan. 2020.

[32] A. Gfrerrer, "Geometry and kinematics of the Mecanum wheel," *Comput. Aided Geometric Des.*, vol. 25, no. 9, pp. 784–791, Dec. 2008.

[33] S. G. Tzafestas, *Introduction to Mobile Robot Control*. Amsterdam, The Netherlands: Elsevier, 2014.

[34] R. Ortega, J. A. Perez, P. J. Nicklasson, and H. J. Sira-Ramirez, *Passivity-Based Control of Euler-Lagrange Systems: Mechanical, Electrical and Electromechanical Applications*. London, U.K.: Springer, 2013.

[35] B. A. Finlayson, *The Method of Weighted Residuals and Variational Principles*. New York, NY, USA: Academic Press, 1990.

[36] Z. Hendzel, "Robust neural networks control of omni-Mecanum wheeled robot with Hamilton-Jacobi inequality," *J. Theor. Appl. Mech.*, vol. 56, no. 4, pp. 1193–1204, 2018.