Output-feedback online optimal control for a class of nonlinear systems

Rushikesh Kamalapurkar

***Abstract*— In this paper an output-feedback model-based reinforcement learning (MBRL) method for a class of second-order nonlinear systems is developed. The control technique uses exact model knowledge and integrates a dynamic state estimator within the model-based reinforcement learning framework to achieve output-feedback MBRL. Simulation results demonstrate the efficacy of the developed method.**

I. Introduction

Over the past decade, online reinforcement learning algorithms that guarantee stability during the learning phase have been developed for deterministic systems [1]–[15]; however, stability and convergence are established under restrictive persistence of excitation (PE) conditions which are difficult, if not impossible, to verify. To soften the PE condition, data-driven methods that employ experience replay have been utilized in results such as [13], [16]–[21]; however, since the data is collected along the system trajectory, added exploration signals are often required to achieve convergence. The need for PE and exploration signals is a result of sample inefficiency, and is a significant drawback of the existing model-free RL-based online optimal control methods.

Model-based reinforcement learning (MBRL) algorithms learn a model of the system from observations using supervised learning and employ the model to learn the policies. Several different MBRL approaches have been developed in the literature over the last few decades. Imaginary roll-outs, i.e., the use of a model as a proxy for the real world to evaluate temporal difference errors (referred to as Bellman errors (BEs) in this paper) are explored in results such as [22] and [23]. While the sample efficiency is of the policy learning algorithms is improved, the performance of the method in [22] decays rapidly with model mismatch, and the method in [23] relies on fitting neural networks to dynamics, which is typically data-intensive, nullifying the sample efficiency gain in the policy learning algorithm.

Policy gradient methods that rely on backpropagation through the model to compute the gradient of the state or the action value function with respect to the policy parameters are developed in results such as [24]–[27]; however, policy gradient methods are often iterative in nature and typically do not study stability during the learning phase, and as a result, are not suitable for real-time simultaneous learning and execution. MBRL methods with provable sample efficiency bounds have been developed in results such as [28]–[34]; however, the theoretical guarantees are obtained under discretization of the continuous state space into finitely many discrete states and a finite action space, and as such, are not directly applicable to systems with continuous state and action spaces.

The MBRL technique developed by the authors in [35]–[40] for continuous time and continuous space systems softens the excitation requirements used in results such as [1]–[15], [41]–[45] by utilizing a model of the system to simulate exploration, where the stability and the performance of the closed-loop system critically depends on the accuracy of the estimated model. A significant drawback of the online optimal control methods mentioned so far is that they require full state measurements.

While model-based and model-free reinforcement learning can be achieved using output feedback instead of state feedback by making use of partially observable Markov decision processes (POMDPs), in general, POMDPs are undecidable if the objective is to find an optimal solution, and finding a near-optimal solution can also be NP-hard [46], [47]. In this paper, the problem is formulated as a state estimation based reinforcement learning problem, and for a specific class of systems, an online solution is obtained that guarantees stability during the learning phase.

A recent result in [48], presents an offline model-free algorithm for linear systems to achieve optimality using output feedback. The objective in this paper is to develop an output-feedback model-based reinforcement learning method for a class of nonlinear systems under exact model knowledge. While the developed results can be extended to systems with uncertain models using model-learning methods such as [49], such extension is not a focus of this work.

The paper is organized as follows. A detailed description of the problem under consideration is provided in Section I. To facilitate the subsequent analysis of the developed technique, section III examines the stability properties of optimal controllers under semidefinite cost functions for the class of systems under consideration. Section IV describes the state estimator used in the design. Section V describes the developed MBRL method. Section VI presents a Lyapunov-based stability analysis, Section VII presents simulation results, and Section VIII concludes the paper.

II. Problem Description

Consider a second order nonlinear system of the form[1]

$$\begin{aligned} \dot{p} &= q, \\ \dot{q} &= f(x) + g(x)\, u, \\ y &= p, \end{aligned} \tag{1}$$

Rushikesh Kamalapurkar is with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. rushikesh.kamalapurkar@okstate.edu.

[1]For $a \in \mathbb{R}$, the notation $\mathbb{R}_{\geq a}$ denotes the interval $[a, \infty)$ and the notation $\mathbb{R}_{>a}$ denotes the interval (a, ∞).

where $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^n$ denote the generalized position states and the generalized velocity states, respectively, $x \coloneqq \begin{bmatrix} p^T & q^T \end{bmatrix}^T$ is the system state, $f : \mathbb{R}^{2n} \to \mathbb{R}^n$ is locally Lipschitz continuous, $f(0) = 0$, and $y \in \mathbb{R}^n$ denotes the output. The drift dynamics f are unknown and the control effectiveness $g : \mathbb{R}^{2n} \to \mathbb{R}^{n \times m}$ is unknown and locally Lipschitz. Systems of the form (1) encompass second-order linear systems and Euler-Lagrange models with known inertial matrices, and hence, represent a wide class of physical plants, including, but not limited to, robotic manipulators and autonomous ground, aerial, and underwater vehicles.

The objective is to design an adaptive estimator to estimate the state x, online, using input-output measurements and to simultaneously estimate and utilize the optimal feedback controller that minimizes the cost functional

$$J(x(\cdot), u(\cdot)) = \int_0^\infty r(x(\tau), u(\tau)) \, \mathrm{d}\tau, \tag{2}$$

while maintaining system stability during the learning phase. The function $r : \mathbb{R}^{n \times m} \to \mathbb{R}$ is defined as $r(x, u) \coloneqq Q(x) + u^T R u$, where $Q : \mathbb{R}^n \to \mathbb{R}$ is continuous, $R \in \mathbb{R}^{m \times m}$ is a constant positive definite matrix, and $\gamma \geq 0$ is the discount factor. It is further assumed that either Q is positive definite or Q is positive semidefinite and the functions $p \mapsto Q(x)$ and $q \mapsto Q(x)$ are positive definite for all nonzero $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^n$, respectively. To facilitate control design, the stability properties of the closed-loop system under optimal feedback are examined.

III. Stability Under Optimal state Feedback

The following theorem establishes global asymptotic stability of the closed-loop system under optimal state feedback.

Theorem 1. *If the optimal state feedback controller $u^* : \mathbb{R}^{2n} \to \mathbb{R}^m$ that minimizes the cost function in* (2) *exists and if the corresponding optimal value function $V : \mathbb{R}^{2n} \to \mathbb{R}$ is continuously differentiable and radially unbounded, then the origin of closed-loop system*

$$\begin{aligned} \dot{y} &= q, \\ \dot{q} &= f(x) + g(x) u^*(x), \end{aligned} \tag{3}$$

is globally asymptotically stable.

Proof. Under the hypothesis of Theorem 1, the optimal value function is the unique solution of the Hamilton-Jacobi-Bellman equation [50, pp. 164]

$$V_y(x) q + V_q(x) (f(x) + g(x) u^*(x)) + r(y, u^*(x)) = 0, \tag{4}$$

with

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) V_q(x), \tag{5}$$

where the notation x_y denotes the partial derivative of x with respect to y. The function V is positive semidefinite by definition. Because Q is positive definite and the solutions of (3) are continuous, if $V\left(\begin{bmatrix} y \\ q \end{bmatrix}\right) = 0$ for some $x \neq 0$, it can be concluded that the output remains zero along the entire trajectory starting from this x. Since the state is comprised of the output and its time-derivative, the state also remains zero along the entire trajectory, contradicting $x \neq 0$. Hence, V cannot be zero for a nonzero x. Furthermore, since $f(0) = 0$, the zero controller is clearly the optimal controller starting from $x = 0$. That is, $V(0) = 0$, and as a result, $V : \mathbb{R}^{2n} \to \mathbb{R}$ is positive definite. Using V as a candidate Lyapunov function and using the HJB equation in (4), it can be concluded that

$$V_y(x) q + V_q(x) (f(x) + g(x) u^*(x)) \leq -Q(x),$$

$\forall x \in \mathbb{R}^{2n}$. If Q is positive definite, the proof is complete using Lyapunov's direct method. If Q is positive semidefinite and $p \mapsto Q(x)$ is positive definite for each nonzero q then, using the fact that if the output is identically zero then so is the state, the invariance principle [51, Corollary 4.2] can be invoked to complete the proof. If Q is positive semidefinite and $q \mapsto Q(x)$ is positive definite for each nonzero p then, using the fact that finiteness of the value function everywhere implies that the origin is the only equilibrium point of the closed-loop system, the invariance principle [51, Corollary 4.2] can be invoked to complete the proof. □

Using Theorem 1 and the converse Lyapunov theorem for asymptotic stability [51, Theorem 4.17], the existence of a radially unbounded positive definite function $\mathcal{V} : \mathbb{R}^{2n} \to \mathbb{R}$ and a positive definite function $W : \mathbb{R}^{2n} \to \mathbb{R}$ is guaranteed such that

$$\mathcal{V}_y(x) q + \mathcal{V}_q(x) (f(x) + g(x) u^*(x)) \leq -W(x), \tag{6}$$

$\forall x \in \mathbb{R}^{2n}$. The functions $\mathcal{V}$ and W are utilized to analyze the stability of the output feedback approximate optimal controller.

IV. Velocity Estimator Design

To generate estimates of the generalized velocity, a velocity estimator inspired by [52] is developed. The estimator is given by

$$\begin{aligned} \dot{\hat{p}} &= \hat{q}, \\ \dot{\hat{q}} &= f(\hat{x}) + g(\hat{x}) u + \nu, \end{aligned} \tag{7}$$

where $\hat{x}$, $\hat{p}$, and $\hat{q}$ are estimates of x, p, and q, respectively, and ν is a feedback term designed in the following.

To facilitate the design of ν, let $\tilde{p} = p - \hat{p}$, $\tilde{q} = q - \hat{q}$, and let

$$r(t) = \dot{\tilde{p}}(t) + \alpha \tilde{p}(t) + \eta(t), \tag{8}$$

where the signal η is added to compensate for the fact that the generalized velocity state, q, is not measurable. Based on the subsequent stability analysis, the signal η is designed as the output of the dynamic filter

$$\dot{\eta}(t) = -\beta \eta(t) - k r(t) - \alpha \tilde{q}(t), \quad \eta(T_0) = 0, \tag{9}$$

where α, k, and β are positive constants and the feedback component ν is designed as

$$\nu(t) = \alpha^2 \tilde{p}(t) - (k + \alpha + \beta) \eta(t). \tag{10}$$

The design of the signals η and ν to estimate the state from output measurements is inspired by the $p-$filter [53]. Using the fact that $\tilde{p}(0)=0$, the signal η can be implemented via the integral form

$$\eta(t) = -\int_{T_0}^{t} (\beta+k)\,\eta(\tau)\,\mathrm{d}\tau - \int_{T_0}^{t} k\alpha\tilde{p}(\tau)\,\mathrm{d}\tau - (k+\alpha)\,\tilde{p}(t). \tag{11}$$

V. Model-based Reinforcement Learning

To estimate the optimal state feedback policy, the optimal value function, defined as

$$V(x) \coloneqq \min_{u(\cdot)} \int_{t}^{\infty} r\left(\phi(\tau, x, u(\cdot)), u(\cdot)\right)\mathrm{d}\tau,$$

where $\phi(\tau, x, u(\cdot))$ denotes the trajectory of (1), evaluated at $t=\tau$, starting from the state x and under the controller $u(\cdot)$, and the optimal policy u^* are approximated using parametric approximators $\hat{V}:\mathbb{R}^{2n}\times\mathbb{R}^{L}\to\mathbb{R}$ and $\hat{u}:\mathbb{R}^{2n}\times\mathbb{R}^{L}\to\mathbb{R}^{m}$ defined as

$$\hat{V}(x, W_c) \coloneqq W_c^T\sigma(x), \text{and} \tag{12}$$

$$\hat{u}(x, W_a) \coloneqq -\frac{1}{2}R^{-1}g^T(x)\,\sigma_x^T(x)\,W_a, \tag{13}$$

where $\sigma \coloneqq [\sigma_1\cdots,\sigma_L]$, $\sigma_i:\mathbb{R}^{2n}\to\mathbb{R}$ for all i is the vector of basis functions and $W_c\in\mathbb{R}^L$ and $W_a\in\mathbb{R}^L$ are estimates of the ideal parameters $W\in\mathbb{R}^L$. The corresponding approximation error $\epsilon:\mathbb{R}^{2n}\to\mathbb{R}$ is defined as $\epsilon(x)\coloneqq V(x)-\hat{V}(x,W)$. Provided the basis functions are selected from an appropriate class of functions, for any given compact ball $\overline{\mathrm{B}}(0,\chi)\subset\mathbb{R}^{2n}$, and any given $\bar{\epsilon}$ there exists $L\in\mathbb{N}$, a set of basis functions $\{\sigma_1,\cdots,\sigma_L\}$, and $W\in\mathbb{R}^L$ such that $\overline{\|\epsilon\|}_{\chi}<\bar{\epsilon}$ and $\overline{\|\epsilon_x\|}_{\chi}<\bar{\epsilon}$, where $\overline{\|\epsilon\|}_{\chi}$ denotes $\sup_{x\in\overline{\mathrm{B}}(0,\chi)}\|\epsilon(x)\|$ (see [54]–[56]).

Substituting the estimates $\hat{V}$, $\hat{u}$, and $\hat{x}$ in (4), the Bellman error $\delta:\mathbb{R}^{2n}\times\mathbb{R}^{L}\times\mathbb{R}^{L}\to\mathbb{R}$ is obtained as

$$\delta(\hat{x}, W_c, W_a) = \hat{V}_q(\hat{x}, W_c)\left(f(\hat{x}) + g(\hat{x})\,\hat{u}(\hat{x}, W_a)\right) + \hat{V}_y(\hat{x}, W_c)\,\hat{q} + r\left(\hat{y}, \hat{u}(\hat{x}, W_a)\right), \tag{14}$$

Similar to [36], the technique developed in this result implements simulation of experience in a model-based RL scheme by using the system model to extrapolate the approximate BE to unexplored areas of the state space. In the following, the trajectories of the state and the weight estimates W_c and W_a, evaluated at time t starting from appropriate initial conditions are denoted by $x(t)$, $W_c(t)$ and $W_a(t)$, respectively. The notation $\delta_t:\mathbb{R}_{\geq 0}\to\mathbb{R}$ denotes the BE in (14), evaluated along the trajectories of the state and the weight estimates as $\delta_t(t)\coloneqq\delta\left(\hat{x}(t),\hat{W}_c(t),\hat{W}_a(t)\right)$ and $\delta_{ti}:\mathbb{R}_{\geq 0}\to\mathbb{R}$ denotes BE extrapolated along the trajectories of the weight estimates and a predefined set of trajectories $\{x_i:\mathbb{R}_{\geq 0}\to\mathbb{R}^n \mid i=1,\cdots,N\}$ as $\delta_{ti}\coloneqq\delta\left(x_i(t),\hat{W}_c(t),\hat{W}_a(t)\right)$. A least-squares update law for the value function weights is designed based on the subsequent stability analysis as

$$\dot{\hat{W}}_c(t) = -\frac{k_c}{N}\Gamma(t)\sum_{i=1}^{N}\frac{\omega_i(t)}{\rho_i(t)}\delta_{ti}(t), \tag{15}$$

$$\dot{\Gamma}(t) = \beta\Gamma(t) - \frac{k_c}{N}\Gamma(t)\sum_{i=1}^{N}\frac{\omega_i(t)\,\omega_i^T(t)}{\rho_i^2(t)}\Gamma(t), \tag{16}$$

$\Gamma(t_0)=\Gamma_0$, where $\Gamma:\mathbb{R}_{\geq t_0}\to\mathbb{R}^{L\times L}$ is a time-varying least-squares gain matrix, $\omega_i(t) \coloneqq \sigma_p(x_i(t))\,q_i(t) + \sigma_q(x_i(t))\left(f(x_i(t)) + g(x_i(t))\,\hat{u}(x_i(t), W_a(t))\right)$, $\rho_i(t)\coloneqq 1+\gamma_1\omega_i^T(t)\,\omega_i(t)$, where $\gamma_1\in\mathbb{R}$ is a constant positive normalization gain, $\beta>0\in\mathbb{R}$ is a constant forgetting factor, and $k_c>0\in\mathbb{R}$ is a constant adaptation gain.

The policy weights are updated to follow the value function weights as

$$\dot{\hat{W}}_a(t) = -k_{a1}\left(W_a(t)-W_c(t)\right) - k_{a2}W_a(t) + \sum_{i=1}^{N}\frac{k_c G_i^T(t)\,W_a(t)\,\omega_i^T(t)}{4N\rho_i(t)}W_c(t), \tag{17}$$

where k_{a1}, $k_{a2}\in\mathbb{R}$ are positive constant adaptation gains, $G_i(t)\coloneqq\sigma_{xi}(t)\,g_i(t)\,R^{-1}g_i^T(t)\,\sigma_{xi}^T(t)\in\mathbb{R}^{L\times L}$, $g_i(t)=g(x_i(t))$ and $\sigma_{xi}(t)=\sigma_x(x_i(t))$. The following rank condition facilitates the subsequent analysis.

Assumption 1. *There exists a finite set of trajectories $\{x_i:\mathbb{R}_{\geq t_0}\to\mathbb{R}^n \mid i=1,\cdots,N\}$ and a constant $T\in\mathbb{R}_{>0}$ such that*

$$\underline{c}_1 I_L \leq \inf_{t\in\mathbb{R}_{\geq t_0}}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\omega_i(t)\,\omega_i^T(t)}{\rho_i^2(t)}\right), \tag{18}$$

$$\underline{c}_2 I_L \leq \frac{1}{N}\int_{t}^{t+T}\left(\sum_{i=1}^{N}\frac{\omega_i(\tau)\,\omega_i^T(\tau)}{\rho_i^2(\tau)}\right)\mathrm{d}\tau,\ \forall t\in\mathbb{R}_{\geq t_0}, \tag{19}$$

where, at least one of the nonnegative constants $\underline{c}_1$ and $\underline{c}_2$ is strictly positive.

The rank conditions in (18) and (19) depend on the estimates W_a; hence, in general, they are impossible to guarantee a priori. However, unlike traditional adaptive dynamic programming literature that assumes that a regressor similar to ω_i evaluated along the system trajectories is PE, Assumption 1 only requires the regressor ω_i to be persistently exciting. When the regressor is evaluated along the system state x excitation in the regressor vanishes as the system states converge. Hence, in general, it is unlikely that a regressor evaluated along the system trajectories will be PE. However, the regressor ω_i depends on x_i, which can be designed independent of the system state x. Hence, $\underline{c}_2$ can be made strictly positive if the signal x_i contains enough frequencies, and $\underline{c}_1$ can be made strictly positive by selecting a sufficient number of extrapolation trajectories, i.e., $N\gg L$. It is established in [38, Lemma 1] that under Assumption 1 and provided $\lambda_{\min}\left\{\Gamma_0^{-1}\right\}>0$, the update law in (16)

ensures that the least squares gain matrix satisfies

$$\underline{\Gamma} I_L \leq \Gamma(t) \leq \overline{\Gamma} I_L, \tag{20}$$

$\forall t \in \mathbb{R}_{\geq 0}$ and for some $\overline{\Gamma}, \underline{\Gamma} > 0$.

VI. Analysis

The approximate BE, evaluated along the selected trajectories $\{x_i \mid i = 1, \cdots, N\}$, can be expressed as

$$\delta_{ti} = -\omega_i^T \tilde{W}_c + \frac{1}{4} \tilde{W}_a^T G_{\sigma i} \tilde{W}_a + \Delta_i, \tag{21}$$

where $\nabla \epsilon_i = \nabla \epsilon(x_i)$, $f_i = f(x_i)$, $G_i \coloneqq g_i R^{-1} g_i^T \in \mathbb{R}^{n \times n}$, $\Delta_i \coloneqq \frac{1}{2} W^T \nabla \sigma_i G_i \nabla \epsilon_i^T + \frac{1}{4} G_{\epsilon i} - \nabla \epsilon_i f_i \in \mathbb{R}$ is a constant, $G_{\epsilon i} \coloneqq \nabla \epsilon_i G_i \nabla \epsilon_i^T \in \mathbb{R}$, and $G_{\sigma i}$ was introduced in (17). Using (21), the time-derivative of the Lyapunov function introduced in (6) along the trajectories of (1) under the controller $u(t) = \hat{u}(\hat{x}, W_a)$ is given by

$$\dot{\mathcal{V}}(x,t) = \mathcal{V}_y(x) q + \mathcal{V}_q(x) (f(x) + g(\hat{x}) \hat{u}(\hat{x}, W_a))$$

Adding and subtracting $\mathcal{V}_q(x)(g(x) u^*(x))$,

$$\dot{\mathcal{V}}(x,t) = \mathcal{V}_y(x) q + \mathcal{V}_q(x) (f(x) + g(x) u^*(x)) + \mathcal{V}_q(x) (g(\hat{x}) \hat{u}(\hat{x}, W_a) - g(x) u^*(x))$$

Using (6), the fact that g is bounded, the basis functions σ are bounded, and that the value function approximation error ϵ and its derivative with respect to x are bounded on compact sets, the time-derivative can be bounded as

$$\dot{\mathcal{V}}(x,t) \leq -W(x) + \iota_1 \overline{\epsilon} + \iota_2 \|\tilde{x}\| \left\|\tilde{W}_a\right\| + \iota_3 \left\|\tilde{W}_a\right\| + \iota_4 \|\tilde{x}\|,$$

for all $t \geq 0$ and for all $x \in \overline{\mathrm{B}}(0, \chi)$ and $\hat{x} \in \mathbb{R}^{2n}$, where $\chi \subset \mathbb{R}^{2n}$ is a compact set, $\iota_1, \cdots, \iota_4$ are positive constants, and $\tilde{x} \coloneqq x - \hat{x}$.

Let $\Theta\left(\tilde{W}_c, \tilde{W}_a, t\right) \coloneqq \frac{1}{2} \tilde{W}_c^T \Gamma^{-1}(t) \tilde{W}_c + \frac{1}{2} \tilde{W}_a^T \tilde{W}_a$ The time-derivative of Θ along the trajectories of (15)-(17) is given by

$$\dot{\Theta}\left(\tilde{W}_c, \tilde{W}_a, t\right) = -\tilde{W}_c^T \Gamma^{-1} \left(-\frac{k_c}{N} \Gamma \sum_{i=1}^{N} \frac{\omega_i}{\rho_i} \delta_{ti}\right) - \frac{1}{2} \tilde{W}_c^T \left(\Gamma^{-1} \beta - \frac{k_c}{N} \sum_{i=1}^{N} \frac{\omega_i \omega_i^T}{\rho_i^2}\right) \tilde{W}_c - \tilde{W}_a^T \left(-k_{a1}(W_a - W_c) - k_{a2} W_a + \sum_{i=1}^{N} \frac{k_c G_i^T W_a \omega_i^T}{4 N \rho_i} W_c\right)$$

Using (14),

$$\dot{\Theta}\left(\tilde{W}_c, \tilde{W}_a, t\right) \leq -k_c \underline{c} \left\|\tilde{W}_c\right\|^2 - (k_{a1} + k_{a2}) \left\|\tilde{W}_a\right\|^2 + k_c \iota_8 \overline{\epsilon} \left\|\tilde{W}_c\right\| + k_c \iota_5 \left\|\tilde{W}_a\right\|^2 + (k_c \iota_6 + k_{a1}) \left\|\tilde{W}_c\right\| \left\|\tilde{W}_a\right\| + \left(k_c \iota_7 + k_{a2} \overline{W}\right) \left\|\tilde{W}_a\right\|,$$

for all $t \geq 0$ and for all $x \in \overline{\mathrm{B}}(0, \chi)$, where $\iota_5, \cdots, \iota_8$ are positive constants that are independent of the learning gains, $\overline{W}$ denotes an upper bound on the norm of the ideal weights W, and $\underline{c} = \min_{t \geq 0} \lambda_{\min} \left\{\left(\frac{\beta}{2 k_c} \Gamma^{-1}(t) + \frac{1}{2N} \sum_{i=1}^{N} \frac{\omega_i \omega_i^T}{\rho_i}\right)\right\}$. Assumption 1 and (20) guarantee that $\underline{c} > 0$.

Let $\Phi(\tilde{p}, r, \eta) \coloneqq \frac{\alpha^2}{2} \tilde{p}^T \tilde{p} + \frac{1}{2} r^T r + \frac{1}{2} \eta^T \eta$. The time-derivative of Φ along the trajectories of (1) and (7)-(10) is given by

$$\dot{\Phi}(\tilde{p}, r, \eta, t) = \alpha^2 \tilde{p}^T (r - \alpha \tilde{p} - \eta) + \eta(-\beta \eta - k r - \alpha \tilde{q}) + r^T \left(\tilde{f}(x, \hat{x}) + \tilde{g}(x, \hat{x}) \hat{u}(\hat{x}, W_a) - \alpha^2 \tilde{p} - k r + k \eta + \alpha \eta\right),$$

where $\tilde{f}(x, \hat{x}) \coloneqq f(x) - f(\hat{x})$ and $\tilde{g}(x, \hat{x}) \coloneqq g(x) - g(\hat{x})$. The time derivative of Φ can be bounded above as

$$\dot{\Phi}(\tilde{p}, r, \eta, t) \leq -\alpha^3 \|\tilde{p}\|^2 - (k - \varpi_1) \|r\|^2 - (\beta - \alpha) \|\eta\|^2 + \varpi_1 (1 + \alpha) \|r\| \|\tilde{p}\| + \varpi_1 \|r\| \|\eta\| + \varpi_3 \|r\| + \varpi_2 \|r\| \left\|\tilde{W}_a\right\|$$

for all $t \geq 0$ and for all $x, \tilde{x} \in \overline{\mathrm{B}}(0, \chi)$, where $\varpi_1, \cdots, \varpi_3$ are positive constants that are independent of the learning gains.

The candidate Lyapunov function for the overall system is then defined as $\mathscr{V}(Z,t) = \mathcal{V}(x) + \Theta\left(\tilde{W}_c, \tilde{W}_a, t\right) + \Phi(\tilde{p}, r, \eta)$, where $Z \coloneqq \begin{bmatrix} x^T & \tilde{p}^T & r^T & \eta^T & \tilde{W}_c^T & \tilde{W}_a^T \end{bmatrix}^T$. The time derivative of the candidate Lyapunov function can be bounded as

$$\dot{\mathscr{V}}(Z,t) \leq -W(x) - z^T \left(\frac{M + M^T}{2}\right) z + P z + \iota_1 \overline{\epsilon},$$

where $z \coloneqq \begin{bmatrix} \left\|\tilde{W}_c\right\| & \left\|\tilde{W}_a\right\| & \|\tilde{p}\| & \|r\| & \|\eta\| \end{bmatrix}^T$, $P = \begin{bmatrix} k_c \iota_8 \overline{\epsilon} & \left(k_c \iota_7 + \iota_3 + k_{a2} \overline{W}\right) & \iota_4 (1 + \alpha) & (\varpi_3 + \iota_4) & \iota_4 \end{bmatrix}$

$M =$

$$\begin{bmatrix} k_c \underline{c} & -(k_c \iota_6 + k_{a1}) & 0 & 0 & 0 \\ 0 & (k_{a1} + k_{a2} - k_c \iota_5) & -\iota_2 (1 + \alpha) & -(\iota_2 + \varpi_2) & -\iota_2 \\ 0 & 0 & \alpha^3 & -\varpi_1 (1 + \alpha) & 0 \\ 0 & 0 & 0 & (k - \varpi_1) & -\varpi_1 \\ 0 & 0 & 0 & 0 & (\beta - \alpha) \end{bmatrix}.$$

Provided the matrix $M + M^T$ is positive definite,

$$\dot{\mathscr{V}}(Z,t) \leq -W(x) - \underline{M} \|z\|^2 + \overline{P} \|z\| + \iota_1 \overline{\epsilon},$$

where $\underline{M} \coloneqq \lambda_{\min} \left\{\frac{M + M^T}{2}\right\}$. Letting $\underline{M} =: \underline{M}_1 + \underline{M}_2$ and letting $\mathcal{W} : \mathbb{R}^{5*n+2*L} \to \mathbb{R}$ be defined as $\mathcal{W}(Z) = -W(x) - \underline{M}_1 \|z\|^2$, the bound

$$\dot{\mathscr{V}}(Z,t) \leq -\mathcal{W}(Z), \forall \|Z\| \geq \mu, Z \in \overline{\mathrm{B}}\left(0, \frac{\chi}{3(1 + \alpha)}\right), \tag{22}$$

for all $t \geq 0$.

Using the bound in (20) and the fact that the converse Lyapunov function is guaranteed to be time-independent, radially unbounded, and positive definite, Lemma 4.3 can be invoked to conclude that

$$\underline{v}(\|Z\|) \leq V_L(Z,t) \leq \overline{v}(\|Z\|), \tag{23}$$

for all $t \in \mathbb{R}_{\geq 0}$ and for all $Z \in \mathbb{R}^{5n+2L}$, where $\underline{v}, \overline{v} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions.

Provided the learning gains, the domain radius χ, and the basis functions for function approximation are selected such

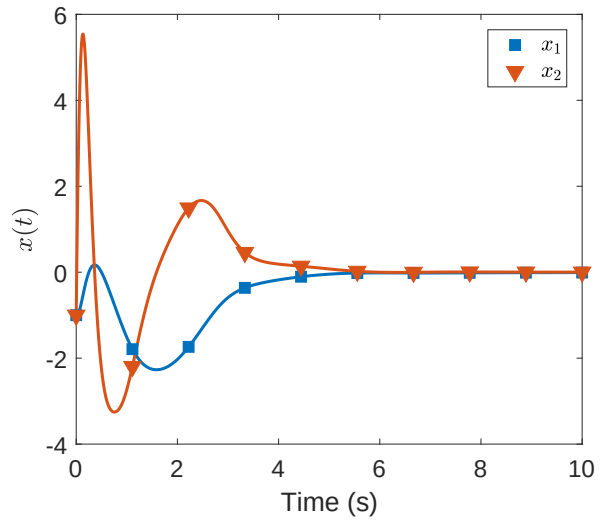

Fig. 1. System state trajectories generated using the developed technique.

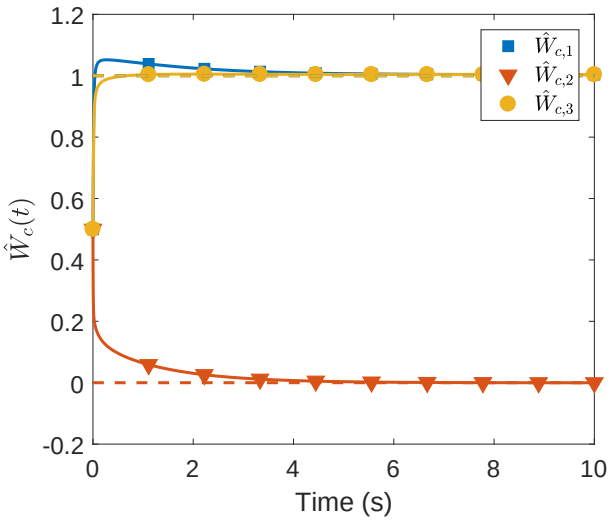

Fig. 3. Critic weight estimates generated using the developed technique, and compared to the ideal values (marked with dashed lines).

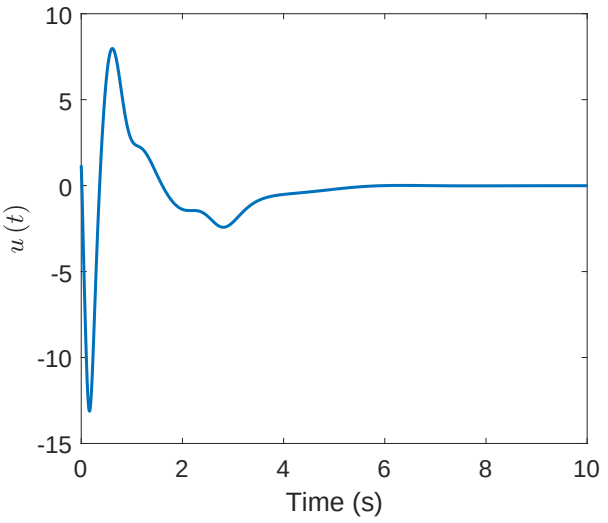

Fig. 2. Control trajectories generated using the developed technique.

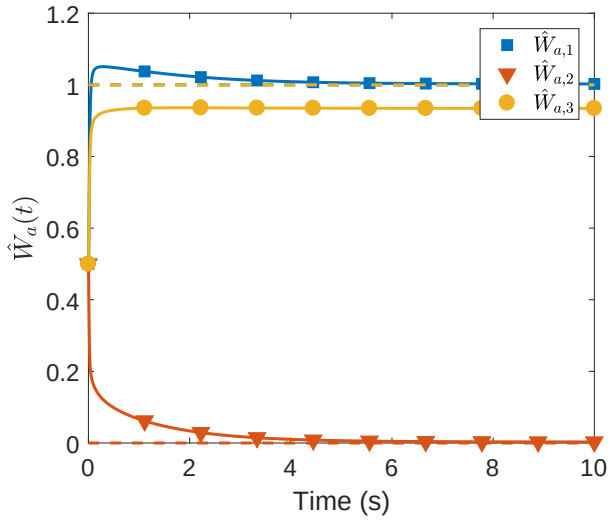

Fig. 4. Actor weight estimates generated using the developed technique, and compared to the ideal values (marked with dashed lines).

that $M+M^T$ is positive definite and $\mu < \overline{v}^{-1}\left(\underline{v}\left(\frac{\chi}{4(1+\alpha)}\right)\right)$, Theorem 4.18 in [51] can be invoked to conclude that Z is uniformly ultimately bounded. Since the estimates W_a approximate the ideal weights W, the policy $\hat{u}$ approximates the optimal policy u^*.

VII. Simulation Results

The performance of the developed controller is demonstrated by simulating a nonlinear, control affine system with a two dimensional state $x = [x_1, x_2]^T$. The system dynamics are described by (1) where

$$f(x) = -x_1 - \frac{1}{2}x_2\left(1 - (\cos(2x_1) + 2)^2\right)$$
$$g(x) = \cos(2x_1) + 2.$$

The origin is an unstable equilibrium point of the unforced system $\dot{x} = f(x)$. The control objective is to minimize the cost in (2), where $Q(x) = q^2$ and $R = 1$. For comparison purposes, the optimal value function for this problem is computed using the converse method in [57] as $V^*(x) = x_1^2 + x_2^2$.

The basis function $\sigma : \mathbb{R}^2 \to \mathbb{R}^3$ for value function approximation is selected as $\sigma = \left[x_1^2, x_1x_2, x_2^2\right]^T$. Based on the analytical solution, the ideal weights are $W = [1,\ 0,\ 1]^T$. The data points for the simulation of experience in the update law (15) are selected to be on a 5×5 grid around the origin. The learning gains are selected as $k_c = 0.2$, $k_{a1} = 100$, $k_{a2} = 0.1$, $\beta_\gamma = 3$, and $\nu = 0.005$. The gains for the state estimator are selected as $k = 5$, $\alpha = 0.2$, and $\beta = 5$. The initial conditions are selected as $x(0) = [1, 1]^T$, $\hat{x}(0) = [-1, -1]^T$, $W_a(0) = W_c(0) = [0.5, 0.5, 0.5]^T$, and $\Gamma(0) = 50\,\mathrm{I}_3$.

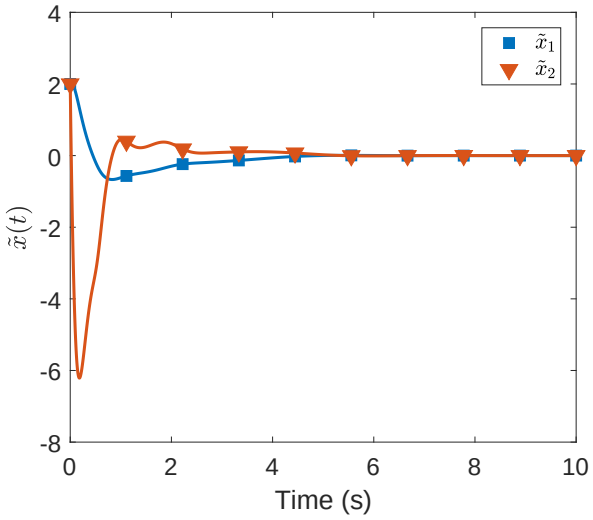

Fig. 5. State estimation error.

Figs. 1-5 demonstrates that the system state is regulated to the origin, the generalized velocities are identified, and the actor and the critic weights converge to their true values. Furthermore, unlike previous results, a probing signal to ensure persistence of excitation is not required.

VIII. Conclusion

An output-feedback MBRL method is developed for a class of second-order nonlinear systems. The control technique uses exact model knowledge and integrates a dynamic state estimator within the model-based reinforcement learning framework to achieve output-feedback MBRL. Simulation results demonstrate the efficacy of the developed method. Integration of simultaneous state and parameter estimation methods such as [49] with the MBRL method to achieve output-feedback MBRL using uncertain models is a topic for future research.

References

[1] 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.

[2] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Conf. Decis. Control*, Dec. 2009, pp. 3598–3605.

[3] D. Vrabie and F. L. Lewis, "Integral reinforcement learning for online computation of feedback nash strategies of nonzero-sum differential games," in *Proc. IEEE Conf. Decis. Control*, 2010, pp. 3066–3071.

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

[5] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*, 3rd ed. Hoboken, NJ: Wiley, 2012.

[6] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, Nov. 2012.

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

[8] 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. 89–92, Jan. 2013.

[9] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming and optimal control of nonlinear nonaffine systems," *Automatica*, vol. 50, no. 10, pp. 2624–2632, 2014.

[10] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, Apr. 2014.

[11] 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, 2014.

[12] T. Bian, Y. Jiang, and Z.-P. Jiang, "Decentralized adaptive optimal control of large-scale systems with application to power systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2439–2447, Apr. 2015.

[13] C. Li, D. Liu, and H. Li, "Finite horizon optimal tracking control of partially unknown linear continuous-time systems using policy iteration," *IET Control Theory Appl.*, vol. 9, no. 12, pp. 1791–1801, 2015.

[14] X. Yang, D. Liu, Q. Wei, and D. Wang, "Direct adaptive control for a class of discrete-time unknown nonaffine nonlinear systems using neural networks," *Int. J. Robust Nonlinear Control*, vol. 25, no. 12, pp. 1844–1861, Apr. 2015.

[15] Q. Zhao, H. Xu, and S. Jagannathan, "Neural network-based finite-horizon optimal control of uncertain affine nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 486–499, 2015.

[16] P. Cichosz, "An analysis of experience replay in temporal difference learning," *Cybern. Syst.*, vol. 30, no. 5, pp. 341–363, 1999.

[17] P. Wawrzyński, "Real-time reinforcement learning by sequential actor-critics and experience replay," *Neural Netw.*, vol. 22, no. 10, pp. 1484–1497, 2009.

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

[19] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 2, pp. 201–212, 2012.

[20] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, 2014.

[21] 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.

[22] R. S. Sutton, "Integrated modeling and control based on reinforcement learning and dynamic programming," in *Advances in Neural Information Processing Systems 3*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds. Morgan-Kaufmann, 1991, pp. 471–478.

[23] T. Lampe and M. Riedmiller, "Approximate model-assisted neural fitted Q-iteration," in *Int. Joint Conf. Neural Netw.*, 2014, pp. 2698–2704.

[24] P. Abbeel, M. Quigley, and A. Y. Ng, "Using inaccurate models in reinforcement learning," in *Proc. Int. Conf. Mach. Learn.* New York, NY, USA: ACM, 2006, pp. 1–8.

[25] M. P. Deisenroth and C. E. Rasmussen, "Pilco: a model-based and data-efficient approach to policy search," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 465–472.

[26] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa, "Learning continuous control policies by stochastic value gradients," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2944–2952.

[27] I. Grondman, "Online model learning algorithms for actor-critic control," Ph.D. dissertation, Technische Universiteit Delft, 2015.

[28] R. I. Brafman and M. Tennenholtz, "R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning," *J. Mach. Learn. Res.*, vol. 3, pp. 213–231, Oct. 2002.

[29] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine Learning*, vol. 49, no. 2, pp. 209–232, Nov. 2002.

[30] S. Kakade, M. J. Kearns, and J. Langford, "Exploration in metric state spaces," in *Proc. Int. Conf. Mach. Learn.*, T. Fawcett and N. Mishra, Eds., 2003, pp. 306–312.

[31] A. Nouri and M. L. Littman, "Multi-resolution exploration in continuous spaces," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1209–1216.

[32] L. Li, M. L. Littman, T. J. Walsh, and A. L. Strehl, "Knows what it knows: a framework for self-aware learning," *Mach. Learn.*, vol. 82, no. 3, pp. 399–443, 2011.

[33] T. Jung and P. Stone, "Gaussian processes for sample efficient reinforcement learning with RMAX-like exploration," in *Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2010*, ser. Lecture Notes in Computer Science, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, vol. 6321, pp. 601–616.

[34] R. Grande, T. Walsh, and J. How, "Sample efficient reinforcement learning with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2, Jun. 2014, pp. 1332–1340.

[35] 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.

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

[37] ——, "Model-based reinforcement learning for approximate optimal regulation," in *Control of Complex Systems: Theory and Applications*, K. Vamvoudakis and S. Jagannathan, Eds. Butterworth-Heinemann, Aug. 2016, pp. 247–273.

[38] 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.

[39] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 753–758, Mar. 2017.

[40] R. Kamalapurkar, J. R. Klotz, P. Walters, and W. E. Dixon, "Model-based reinforcement learning in differential graphical games," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 1, pp. 423–433, Mar. 2018.

[41] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 2009.

[42] D. Vrabie and F. L. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[43] K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations," *Automatica*, vol. 47, pp. 1556–1569, 2011.

[44] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598–1611, 2012.

[45] R. Song, F. L. Lewis, Q. Wei, H.-G. Zhang, Z.-P. Jiang, and D. Levine, "Multiple actor-critic structures for continuous-time optimal control using input-output data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 851–865, Apr. 2015.

[46] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Math. Oper. Res.*, vol. 12, no. 3, pp. 441–450, 1987.

[47] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and related stochastic optimization problems," *Artif. Intell.*, vol. 147, no. 1-2, pp. 5–34, 2003.

[48] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2401–2410, Sep. 2016.

[49] R. Kamalapurkar, "Simultaneous state and parameter estimation for second-order nonlinear systems," in *Proc. IEEE Conf. Decis. Control*, Melbourne, VIC, Australia, Dec. 2017, pp. 2164–2169.

[50] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2012.

[51] H. K. Khalil, *Nonlinear systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[52] H. T. Dinh, R. Kamalapurkar, S. Bhasin, and W. E. Dixon, "Dynamic neural network-based robust observers for uncertain nonlinear systems," *Neural Netw.*, vol. 60, pp. 44–52, Dec. 2014.

[53] B. Xian, M. S. de Queiroz, D. M. Dawson, and M. McIntyre, "A discontinuous output feedback controller and velocity observer for nonlinear mechanical systems," *Automatica*, vol. 40, no. 4, pp. 695–700, 2004.

[54] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Netw.*, vol. 3, no. 5, pp. 551–560, 1990.

[55] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, pp. 251–257, 1991.

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

[57] V. Nevistic and J. A. Primbs, "Constrained nonlinear optimal control: a converse HJB approach," California Institute of Technology, Pasadena, CA 91125, Tech. Rep. CIT-CDS 96-021, 1996.