Output Feedback Adaptive Optimal Control of Affine Nonlinear systems with a Linear Measurement Model

Tochukwu Elijah Ogri[1] S M Nahid Mahmud[2] Zachary I. Bell[3] Rushikesh Kamalapurkar[1]

***Abstract*— Real-world control applications in complex and uncertain environments require adaptability to handle model uncertainties and robustness against disturbances. This paper presents an online, output-feedback, critic-only, model-based reinforcement learning architecture that simultaneously learns and implements an optimal controller while maintaining stability during the learning phase. Using multiplier matrices, a convenient way to search for observer gains is designed along with a controller that learns from simulated experience to ensure stability and convergence of trajectories of the closed-loop system to a neighborhood of the origin. Local uniform ultimate boundedness of the trajectories is established using a Lyapunov-based analysis and demonstrated through simulation results, under mild excitation conditions.**

I. Introduction

Reinforcement learning (RL) has proven to be robust to modeling errors in dynamic systems, ensuring a fast convergence to the optimal solution while maintaining stability regardless of disturbances to the system [1]–[4]. In the absence of full state measurement information, model-based reinforcement learning (MBRL) controllers in [5]–[8], tend to perform poorly since the excitation conditions require the accuracy of the estimated model to guarantee the closed-loop stability of the system.

Motivated by the performance of the observer developed in [9] which augments the extended Luenberger observer in [10] by introducing a third observer gain to cancel non-convex terms in the semi-definite condition, this paper offers a modification to that observer structure with fewer restrictions on the class of nonlinear systems. An observer for real-time state estimation using semi-definite programming (SDP) to search for the extended Luenberger observer gains is developed for continuous-time nonlinear systems. Using multiplier matrix approach which involves placing bounds on the derivatives of the drift and control effectiveness functions of the system, sufficient conditions developed using Lyapunov analysis are used to guarantee the stability of the state estimation error dynamics [11], [12]. The state estimates are then used in a MBRL framework to design a adaptive dynamic programming (ADP) based controller that optimizes a given performance objective while ensuring the stability of the closed-loop system during learning.

MBRL learns an optimal controller that approximates the value function, and subsequently, the optimal policy for the nonlinear system. While adaptive optimal control methods have been extensively studied in the literature to solve the online optimal control problem, [4]–[8], [13]–[16], most existing results require full state feedback. In this paper, an output feedback problem is solved for systems with a linear measurement model. Unlike actor-critic MBRL methods popular in the literature [17], [18], this paper presents a critic-only structure to provide an approximate solution of the Hamilton–Jacobi–Bellman (HJB) equation that requires the identification of fewer free parameters. Lyapunov methods are used to show that the states of the system, the state estimation error, and the critic weights are locally uniformly ultimately bounded (UUB) for all time starting from any initial condition.

This novel architecture is different from existing NN network observers in literature like [15], [19]–[23] whose convergence analysis relies solely on negative terms that result from a $\sigma-$modification-like term added to the weight update laws. As a result, similar to adaptive control, the convergence of the observer weights to their true values cannot be expected, and convergence of state estimates to the true states is not robust to disturbances and approximation errors. In addition, the observer technique in this paper does not require restrictions on the form and rank of the C matrix unlike NN based observers in [15], [21], [22]. A drawback of existing state feedback control methods, such as [21], is that the substitution $x = C^{+}y$ implicitly restricts the technique to systems where the number of outputs is larger than the number of states, which is typically not the case in output feedback control.

The rest of the paper is organized as follows: Section II contains the problem formulation, Section III introduces the state estimator/observer, Section IV presents the Multiplier matrices and sector Conditions, Section V contains control design using MBRL methods, Section VI contains stability analysis of the developed architecture, and Section VII concludes the paper.

*This research was supported in part by Office of Naval Research under award number N00014-21-1-2481 and the Air Force Research Laboratories under contract numbers FA8651-19-2-0009 and AFRL-FA8651-23-1-0006. Any opinions, findings, or recommendations in this article are those of the author(s), and do not necessarily reflect the views of the sponsoring agencies.

[1]School of Mechanical and Aerospace Engineering, Oklahoma State University, email: {tochukwu.ogri, rushikesh.kamalapurkar}@okstate.edu.

[2] School of Aeronautics and Astronautics, Purdue University, West Lafayette, 47907, USA, e-mail: {mahmud7}@purdue.edu.

[3] Air Force Research Laboratories, Florida, USA, email: zachary.bell.10@us.af.mil.

II. Problem Formulation

This paper considers nonlinear dynamical systems of the form

$$\dot{x} = f(x) + g(x)u, \quad y = Cx, \tag{1}$$

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^m$ is the control input, $C \in \mathbb{R}^{q\times n}$ is the output matrix, and $y \in \mathbb{R}^q$ is the measured output. The functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, denote the drift and the control effectiveness matrix, respectively.

The objective is to design an observer to estimate the states online, using input-output measurements, and to simultaneously synthesize and utilize a controller that minimizes the cost functional defined in (24), under the saturation constraint $|(u)_i| \leq \lambda > 0$ for $i = 1, \ldots, m$, while ensuring local uniform ultimate boundedness of the trajectories of the system in (1).

In order to facilitate the development and analysis of the method presented in this paper, the following assumption is necessary.

Assumption 1. The functions f and g are known, their derivatives exist on a compact set $\mathcal{C} \subset \mathbb{R}^n$, and satisfy the element-wise bounds

$$(M_{f_1})_{i,j} \leq \frac{\mathrm{d}(f(x))_i}{\mathrm{d}(x)_j} \leq (M_{f_2})_{i,j}, \tag{2}$$

$$(M_{g_1})_{i,j} \leq \left(\frac{\mathrm{d}(g(x))_{i,k}}{\mathrm{d}(x)_j}\right)(u)_k \leq (M_{g_2})_{i,j}, \tag{3}$$

for all $x \in \mathcal{C}$, $|(u)_k| \leq \lambda$, $i, j = 1, \ldots, n$ and $k = 1, \ldots, m$, where $(\cdot)_i$ and $(\cdot)_{i,j}$ denote the element of the array $(\cdot)$ at the indices indicated by the subscript.

Remark 1. Conditions similar to those in Assumption 1 are commonly required in several observer design schemes (see, e.g., [9], [24]–[26].

In the following section, sufficient conditions involving multiplier matrices that characterize the affine system will be presented, along with the design of the state estimator.

III. State Estimator

In this section, a state estimator inspired by the extended Luenberger observer is developed to generate estimates of x. Let the nonlinear dynamics in (1) be expressed in the form

$$\dot{x} = M_{f_1}x + M_{g_1}x + \overline{f}(x) + \overline{g}_u(x, u), \tag{4}$$

where $\overline{f}(x) = -M_{f_1}x + f(x)$, and $\overline{g}_u(x, u) = -M_{g_1}x + \sum_{i=1}^{m} g_i(x)(u)_i$. Under Assumption 1, the derivatives of $\overline{f}$ and $\overline{g}$ satisfy the element-wise inequalities

$$0 \leq \frac{\mathrm{d}(\overline{f}(x))_i}{\mathrm{d}(x)_j} \leq (M_{f_2})_{i,j} - (M_{f_1})_{i,j}, \text{ and} \tag{5}$$

$$0 \leq \frac{\mathrm{d}(\overline{g}_u(x, u))_i}{\mathrm{d}(x)_j} \leq (M_{g_2})_{i,j} - (M_{g_1})_{i,j}, \tag{6}$$

where $i, j = 1, \ldots, n$. Let $\overline{M}_{f_1} \coloneqq 0_{n\times n}$, $\overline{M}_{f_2} \coloneqq M_{f_2} - M_{f_1}$, $\overline{M}_{g_1} \coloneqq 0_{n\times n}$ and $\overline{M}_{g_2} \coloneqq M_{g_2} - M_{g_1}$. Using the derivative bounds, a state estimator with three correction terms is designed as

$$\begin{aligned}\dot{\hat{x}} = {} & M_{f_1}\hat{x} + M_{g_1}\hat{x} + \overline{f}[\hat{x} + H\,(y - C\hat{x})] \\ & + \overline{g}_u[\hat{x} + K\,(y - C\hat{x})\,, u] + L\,(y - C\hat{x})\,,\end{aligned} \tag{7}$$

where $\hat{x} \in \mathbb{R}^n$ is the estimate of x, $H \in \mathbb{R}^{n\times q}$, $K \in \mathbb{R}^{n\times q}$ and $L \in \mathbb{R}^{n\times q}$ are observer gains, $H\,(y - C\hat{x})$ and $K\,(y - C\hat{x})$ are nonlinear injection terms, and $L\,(y - C\hat{x})$ is a linear correction term.

The estimation error is defined as $e \coloneqq x - \hat{x}$, and the estimation error dynamics are given by

$$\begin{aligned}\dot{e} = {} & (M_{f_1} + M_{g_1} - LC)\,e + \overline{f}(x) + \overline{g}_u(x, u) \\ & - [\hat{x} + H\,(y - C\hat{x})] - \overline{g}_u[\hat{x} + K\,(y - C\hat{x})\,, u].\end{aligned} \tag{8}$$

IV. Multiplier Formulation and Sector Conditions

In this section, conditions sufficient for Lyapunov stability are derived by designing multiplier matrices that characterize the nonlinear functions f and g (cf. [11]).

For convenience of notation, let $\phi_f(t, e) \coloneqq \overline{f}(x) - \overline{f}[\hat{x} + H\,(y - C\hat{x})]$ and $\phi_g(t, e, u) \coloneqq \overline{g}_u(x, u) - \overline{g}_u[\hat{x} + K\,(y - C\hat{x})\,, u]$. The differential mean value theorem (DMVT) [27, Theorem 2.1] guarantees that the difference function ϕ_f can be expressed as $\phi_f(t, e) = \overline{M}_f(I - HC)(x - \hat{x})$ where $\overline{M}_f$ is a time-varying matrix that is constrained in a compact set defined by $\overline{M}_{f_1}$ and $\overline{M}_{f_2}$ in (5). Similarly, $\phi_g(t, e, u) = \overline{M}_g(I - KC)(x - \hat{x})$, where $\overline{M}_g$ is a time-varying matrix that is constrained in a compact set defined by $\overline{M}_{g_1}$ and $\overline{M}_{g_2}$ in (6).

The DMVT implies that the difference functions $\phi_f(t, e)$ and $\phi_g(t, e, u)$ are bounded as

$$\overline{M}_{f_1}(I - HC)e \leq \phi_f(t, e) \leq \overline{M}_{f_2}(I - HC)e, \text{ and} \tag{9}$$

$$\overline{M}_{g_1}(I - KC)e \leq \phi_g(t, e, u) \leq \overline{M}_{g_2}(I - KC)e. \tag{10}$$

The stability of the state estimation error dynamics can now be shown using only the sector information about $\phi_f(t, e)$ and $\phi_g(t, e, u)$ constrained on a compact set $\mathcal{C}$, where the Jacobian bounds in (5) and (6) hold. The bounds in (9) and (10) can be used to obtain the inequalities

$$[\phi_f(t, e)]^{\mathrm{T}}\,[\phi_f(t, e) - \overline{M}_{f_2}(I - HC)e] \leq 0 \text{ and} \tag{11}$$

$$[\phi_g(t, e, u)]^{\mathrm{T}}\,[\phi_g(t, e, u) - \overline{M}_{g_2}(I - KC)e] \leq 0. \tag{12}$$

Rewriting the inequalities in (11) and (12) into their quadratic form yields

$$\begin{bmatrix} e \\ \phi_f \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I - HC & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}^{\mathrm{T}} J_f \begin{bmatrix} I - HC & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} e \\ \phi_f \end{bmatrix} \leq 0, \text{ and} \tag{13}$$

$$\begin{bmatrix} e \\ \phi_g \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I - KC & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}^{\mathrm{T}} J_g \begin{bmatrix} I - KC & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} e \\ \phi_g \end{bmatrix} \leq 0, \tag{14}$$

with

$$J_f = \begin{bmatrix} \mathbf{0} & -\frac{M_{f_2}^{\mathrm{T}} - M_{f_1}^{\mathrm{T}}}{2} \\ -\frac{M_{f_2} - M_{f_1}}{2} & I \end{bmatrix} \text{ and} \tag{15}$$

$$J_g = \begin{bmatrix} \mathbf{0} & -\frac{M_{g_2}^{\mathrm{T}} - M_{g_1}^{\mathrm{T}}}{2} \\ -\frac{M_{g_2} - M_{g_1}}{2} & I \end{bmatrix}, \tag{16}$$

where $\mathbf{0}$ denotes an $n \times n$ matrix of zeros and I is an $n \times n$ identity matrix. The observer error dynamics in (8) can now be expressed as

$$\dot{e} = (A - LC)\, e + \phi_f(t, e) + \phi_g(t, e, u). \tag{17}$$

where $A \coloneqq M_{f_1} + M_{g_1}$. The following theorem establishes convergence of the estimator, provided the control input remains bounded and the system trajectories remain within the compact set $\mathcal{C}$.

Remark 2. Note that the assumption that the trajectory and control signals are bounded applies only to the following theorem, and not to the controller designed in Section V. The controller designed in Section V ensures that provided the initial condition is close enough to the origin, the trajectories stay within the compact set $\mathcal{C}$. As such, provided the initial condition is close enough to the origin the bounds on the derivatives of $\overline{f}$ and $\overline{g}$ are guaranteed to hold along the trajectories of the closed-loop system.

Theorem 1. Given a system satisfying Assumption 1, provided the control input remains bounded and the system trajectories remain within the compact set $\mathcal{C}$, if there exists a symmetric positive definite matrix P, and observer gains L, H, K that satisfy the matrix inequality

$$\begin{bmatrix} \begin{pmatrix} (A - LC)^{\mathrm{T}} P \\ +P\,(A - LC) \end{pmatrix} + 2\alpha P & P - {J_{fh}}_{21}^{\mathrm{T}} & P - {J_{gk}}_{21}^{\mathrm{T}} \\ P - {J_{fh}}_{21} & -(J_f)_{22} & \mathbf{0} \\ P - {J_{gk}}_{21} & \mathbf{0} & -(J_g)_{22} \end{bmatrix} \leq 0, \tag{18}$$

where ${J_{fh}}_{21} \coloneqq (J_f)_{21}\,(I - HC)$ and ${J_{gk}}_{21} \coloneqq (J_g)_{21}\,(I - KC)$, then the observer error system in (17) is locally uniformly asymptotically stable.

Proof. Let $\mathcal{D}$ be an open subset of the set $\{e \in \mathbb{R}^n : x, \hat{x} \in \mathcal{C}\}$ and consider the continuously differentiable candidate Lyapunov function, $V_e : \mathcal{D} \to \mathbb{R}$ defined as $V_e(e) \coloneqq e^{\mathrm{T}} P e$, which satisfies the inequality $\lambda_{\min}(P)\|e\|^2 \leq V_e(e) \leq \lambda_{\max}(P)\|e\|^2$ where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a matrix, respectively. Since P is a positive definite matrix, both eigenvalues are positive. On the set $\mathcal{D}$, the orbital derivative of the Lyapunov function along the trajectories of (8) can be expressed as

$$\dot{V}_e(e) = \begin{bmatrix} e \\ \phi_f \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \begin{pmatrix} \left(M_{f_1} - \frac{LC}{2}\right)^{\mathrm{T}} P \\ +P\left(M_{f_1} - \frac{LC}{2}\right) \end{pmatrix} & P \\ P & \mathbf{0} \end{bmatrix} \begin{bmatrix} e \\ \phi_f \end{bmatrix} + \begin{bmatrix} e \\ \phi_g \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \begin{pmatrix} \left(M_{g_1} - \frac{LC}{2}\right)^{\mathrm{T}} P \\ +P\left(M_{g_1} - \frac{LC}{2}\right) \end{pmatrix} & P \\ P & \mathbf{0} \end{bmatrix} \begin{bmatrix} e \\ \phi_g \end{bmatrix}. \tag{19}$$

Substituting (13) and (14) in (19) yields

$$\dot{V}_e(e) = \begin{bmatrix} e \\ \phi_f \\ \phi_g \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} (A-LC)^{\mathrm{T}}P + P(A-LC) & P & P \\ P & \mathbf{0} & \mathbf{0} \\ P & \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & (I-HC)^{\mathrm{T}}(J_f)_{12} & \mathbf{0} \\ (J_f)_{21}(I-HC) & (J_f)_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} 0 & \mathbf{0} & (I-KC)^{\mathrm{T}}(J_g)_{12} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ (J_g)_{21}(I-KC) & (J_g)_{22} & \mathbf{0} \end{bmatrix} \begin{bmatrix} e \\ \phi_f \\ \phi_g \end{bmatrix} \leq 0 \tag{20}$$

Provided the LMI in (18) is satisfied for some constant $\alpha > 0$, the multiplier matrices and sector conditions formulated in (13) and (14), and the S-Procedure Lemma [28] can be used to guarantee that the orbital derivative is bounded as (cf. [11])

$$\dot{V}_e(e) \leq -2\alpha V_e(e), \forall e \in \mathcal{D}. \tag{21}$$

Using the bound in (21), it can be concluded that the origin of the error system, $e = 0$, is locally uniformly asymptotically stable. In particular, let $r \in \mathbb{R}_{>0}$ be a constant such that, $B_r \coloneqq \{e \in \mathbb{R}^n \mid \|e\| \leq r\} \subset \mathcal{D}$ and, $W_2(\|e\|) \coloneqq \lambda_{\min}(P)\|e\|^2$. Select $c > 0$ such that $c < \frac{r^2 \lambda_{\min}(P)}{2}$, Theorem 4.9 in [29] can then be invoked to conclude that every trajectory starting in $\{e \in B_r \mid W_2(\|e\|) \leq c\}$ stays within $\mathcal{D}$ for all $t \geq 0$ and satisfies

$$\|e(t)\| \leq \beta(\|e(t_0)\|, t - t_0), \forall t \geq t_0 \geq 0 \tag{22}$$

where β is a class $\mathcal{KL}$ function. □

Remark 3. The matrix inequality can be reformulated as a linear matrix inequality (LMI) using the typical variable substitution method. Indeed, substituting $L = P^{-1}R$ in (18), the matrix P and the observer gains L, H, and K can be obtained by solving the LMI

$$\begin{bmatrix} \begin{pmatrix} A^{\mathrm{T}}P + PA \\ -C^{\mathrm{T}}R^{\mathrm{T}} - RC \end{pmatrix} + 2\alpha P & P - {J_{fh}}_{21}^{\mathrm{T}} & P - {J_{gk}}_{21}^{\mathrm{T}} \\ P - {J_{fh}}_{21} & -(J_f)_{22} & \mathbf{0} \\ P - {J_{gk}}_{21} & \mathbf{0} & -(J_g)_{22} \end{bmatrix} \leq 0, \tag{23}$$

for P, R, H, and K.

Remark 4. The observer design is only valid if the control input remains bounded and the system trajectories remain within the compact set $\mathcal{C}$ where the bounds on the derivatives of $\overline{f}$ and $\overline{g}$, in (5) and (6), respectively, are valid. In the theorem above, the derivative bounds are local, and as a result the observer error is locally uniformly asymptotically stable. If the derivative bounds hold globally, then a similar analysis can be used to show that the observer error is globally uniformly asymptotically stable. The controller designed in Section V ensures that provided the initial condition is close enough to the origin, the trajectories stay within the desired compact set $\mathcal{C}$.

V. Control Design

To achieve the control objective stated above while satisfying all constraints of the system, the cost functional to be minimized is given as

$$J(x, u(\cdot)) \coloneqq \int_0^\infty Q(\phi(\tau, x, u_{[t,\tau)}(\cdot))) + U(u(\tau))d\tau, \quad (24)$$

over the set $\mathcal{U}$ piecewise continuous functions $t \to u(t)$, where $\phi(t, x, u(\cdot))$ is a solution of (1) under control signal $u(\cdot)$ starting from $x(0)$, $Q : \mathbb{R}^n \to \mathbb{R}$ is a continuous, positive definite function and $U : \mathbb{R}^m \to \mathbb{R}$, introduced to address the saturation constraint on the control, is defined as

$$U(u) \coloneqq 2\int_0^u (\lambda \tanh^{-1}(\upsilon/\lambda))^{\mathrm{T}} R d\upsilon, \quad (25)$$

where $R \coloneqq \mathrm{diag}(r_1, \ldots, r_m)$ and u_I and $\mathcal{U}_I$ are obtained by restricting the domains of u and functions in $\mathcal{U}_I$ to the interval $I \subseteq R$, respectively. Assuming the optimal controller exists, then let the optimal value function, $V^* : \mathbb{R}^n \to \mathbb{R}$, be expressed as

$$V^*(x) \coloneqq \min_{u(\cdot)\in\mathcal{U}_{[t,\infty)}} \int_t^\infty Q(\phi(\tau, x, u_{[t,\tau)}(\cdot))) + U(u(\tau))d\tau. \quad (26)$$

Assuming that the optimal value function is continuously differentiable, it can be shown to be the unique PD solution of the Hamilton-Jacobi-Bellman (HJB) equation, [30, Theorem 1.5],

$$\min_{u\in\mathbb{R}^m} \left(\nabla_x V\left(f(x) + g(x)u\right) + Q(x) + U(u)\right) = 0, \quad (27)$$

where $\nabla_{(\cdot)} \coloneqq \frac{\partial}{\partial(\cdot)}$. Therefore, the optimal controller is given by the feedback policy $u(t) = u^*(\phi(t, x, u_{[0,t)}))$ where $u^* : \mathbb{R}^n \to \mathbb{R}^m$ defined as

$$u^*(x) \coloneqq -\lambda \tanh(D^*), \quad (28)$$

where $D^* = (1/2\lambda)R^{-1}g(x)^{\mathrm{T}}\nabla_x V^*(x) \in \mathbb{R}^m$. Substituting equation (28) in (25), the function U is given as

$$U(u^*) = \lambda \nabla_x V^*(x)^{\mathrm{T}} g(x) \tanh(D^*) + \lambda^2 \overline{R} \ln(\mathbf{1} - \tanh^2(D^*)), \quad (29)$$

where $\overline{R} \coloneqq [r_1, \ldots, r_m] \in \mathbb{R}^{1\times m}$ and $\mathbf{1} \in \mathbb{R}^m$ denotes a column vector having all of its elements equal to one. Substituting optimal control input, (28) in the HJB equation in (27) yields, $\nabla_x V^*\left(f(x) + g(x)u^*(x)\right) + Q(x) + U(u^*) = 0$.

A. Value Function Approximation

Solving the above HJB equation is generally infeasible due to its inherent non-linearity, hence to find an approximate solution, estimates of the value function and the control policy are introduced. The value function and its gradient can be expressed as

$$V^*(x) = W^{\mathrm{T}}\sigma(x) + \epsilon(x), \quad (30)$$

$$\nabla_x V^*(x) = \nabla_x \sigma^{\mathrm{T}}(x) W + \nabla_x \epsilon(x), \quad (31)$$

respectively. $W \in \mathbb{R}^L$ is an unknown vector of bounded weights, $\sigma : \mathbb{R}^n \to \mathbb{R}^L$ is a vector of continuously differentiable nonlinear activation functions such that $\sigma(0) = 0$ and $\nabla_x \sigma(0) = 0$, $L \in \mathbb{N}$ is the number of basis functions, and $\epsilon : \mathbb{R}^n \to \mathbb{R}$ is the reconstruction error. Using the Stone-Weierstrass Theorem [31, Theorem 1.5], given a compact set $\mathcal{C}$, the activation functions σ can be selected so that the weights and the approximation errors satisfy $\sup_{x\in\mathcal{C}} \|W\| \leq \overline{W}$, $\sup_{x\in\mathcal{C}} \|\sigma(\cdot)\| \leq \overline{\|\sigma\|}$, $\sup_{x\in\mathcal{C}} \|\nabla_{(\cdot)}\sigma(\cdot)\| \leq \overline{\|\nabla\sigma\|}$, $\sup_{x\in\mathcal{C}} \|\epsilon(\cdot)\| \leq \overline{\|\epsilon\|}$ and $\sup_{x\in\mathcal{C}} \|\nabla_{(\cdot)}\epsilon(\cdot)\| \leq \overline{\|\nabla\epsilon\|}$, where $\overline{\|(\cdot)\|}$ denotes a positive constant.

Since the ideal weights, W, are unknown, estimates $\hat{V} : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$ and $\hat{u} : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}^m$ are defined as

$$\hat{V}\left(\hat{x}, \hat{W}_c\right) \coloneqq \hat{W}_c^{\mathrm{T}}\sigma(\hat{x}), \quad (32)$$

$$\hat{u}\left(\hat{x}, \hat{W}_c\right) \coloneqq -\lambda \tanh(\hat{D}), \quad (33)$$

where $\hat{D} = \frac{1}{2\lambda}R^{-1}g(\hat{x})^{\mathrm{T}}\nabla_{\hat{x}}\sigma(\hat{x})^{\mathrm{T}}\hat{W}_c$. The critic weights, $\hat{W}_c \in \mathbb{R}^L$ are an estimate of the ideal weights W. Substituting (32) and (33) into (27), the residual term, $\hat{\delta} : \mathbb{R}^n \times \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, referred to as the Bellman error (BE), is obtained as

$$\hat{\delta}\left(\hat{x}, \hat{W}_c\right) = \nabla_x \hat{V}(\hat{x}, \hat{W}_c)\left(f(\hat{x}) + g(\hat{x})\hat{u}(\hat{x}, \hat{W}_c)\right) + U(\hat{u}) + Q(\hat{x}). \quad (34)$$

By simplifying (34), the BE can be expressed as

$$\hat{\delta}\left(\hat{x}, \hat{W}_c\right) = -\omega\left(\hat{x}, \hat{W}_c\right)^{\mathrm{T}}\tilde{W}_c + \Delta\left(\hat{x}, \hat{W}_c\right), \quad (35)$$

where $\omega \coloneqq \nabla\sigma\left(\hat{f} + \hat{g}\hat{u}\right)$ and $\Delta \coloneqq -\nabla\epsilon\left(\hat{f} + \hat{g}u^*\right) + \lambda W^{\mathrm{T}}\nabla\sigma\hat{g}\left(\tanh(D^*) - \tanh(\hat{D})\right) + 2\lambda^2\overline{R}(C_{D^*} - C_{\hat{D}}) + \lambda^2\overline{R}(\varepsilon_{\hat{D}} - \varepsilon_{D^*})$.

To accurately approximate the value function, online RL methods require persistence of excitation (PE) condition [17], [32], which is difficult to guarantee in practice. However, through BE extrapolation for excitation via simulation, stability and convergence of online RL can be established using Assumption 2 [17]. To simulate experience using BE extrapolation, select a set of trajectories $\{x_i : \mathbb{R}_{\geq t_0} \to \mathbb{R}^n \mid i = 1, \cdots, N\}$ and extrapolate the BE along these trajectories to yield the BEs, $\hat{\delta}_i : \mathbb{R}^n \times \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, given by

$$\hat{\delta}_i\left(x_i, \hat{W}_c\right) \coloneqq \nabla_x \hat{V}(x_i, \hat{W}_c)\left(f(x_i) + g(x_i)\hat{u}(x_i, \hat{W}_c)\right) + U(\hat{u}) + Q(x_i). \quad (36)$$

Given the critic weight estimation error $\tilde{W}_c \coloneqq W - \hat{W}_c$ and substituting (32) and (33) into (27), and subtracting from (34), the BE can be expressed as

$$\hat{\delta}_i\left(x_i, \hat{W}_c\right) \coloneqq -\omega_i\left(x_i, \hat{W}_c\right)^{\mathrm{T}}\tilde{W}_c + \Delta_i\left(x_i, \hat{W}_c\right), \quad (37)$$

where $\hat{f}_i \coloneqq f(x_i)$, $\hat{g}_i \coloneqq g(x_i)$, $\sigma_i \coloneqq \sigma(x_i)$, $\omega_i \coloneqq$

$\nabla\sigma_i\left(\hat{f}_i+\hat{g}_i\hat{u}(x_i,\hat{W}_c)\right)$, $\Delta_i := -\nabla\epsilon_i\left(\hat{f}_i+\hat{g}_i u^*(x_i)\right)+\lambda W^{\mathrm{T}}\nabla\sigma_i\hat{g}_i\left(\tanh(D_i{}^*)-\tanh(\hat{D_i})\right)+2\lambda^2\overline{R}(C_{D_i{}^*}-C_{\hat{D}_i})+\lambda^2\overline{R}(\varepsilon_{\hat{D}_i}-\varepsilon_{D_i{}^*})$, $\nabla\epsilon_i=\nabla\epsilon(x_i)$. To simplify notation, the function arguments are being suppressed.

B. Update laws for Critic weights

To guarantee that the estimated value function weights, $\hat{W}_c$, converge to their ideal weights in (30), the estimated value function weights are updated based on the result of the stability analysis in Section VI as

$$\dot{\hat{W}}_c=-\frac{k_c}{N}\Gamma\sum_{i=1}^{N}\frac{\omega_i}{\rho_i}\hat{\delta}_i,\quad \dot{\Gamma}=\beta\Gamma-\frac{k_c}{N}\Gamma\sum_{i=1}^{N}\frac{\omega_i\omega_i^{\mathrm{T}}}{\rho_i^2}\Gamma, \tag{38}$$

with $\Gamma(t_0)=\Gamma_0$, where $\Gamma:\mathbb{R}_{\geq t_0}\to\mathbb{R}^{L\times L}$ is a time-varying least-squares gain matrix, $\rho_i(t):=1+\gamma\omega_i^{\mathrm{T}}(t)\omega_i(t)$, $\gamma\in\mathbb{R}_{>0}$ is a constant normalization gain, $\beta\in\mathbb{R}_{>0}$ is a constant forgetting factor, and $k_c\in\mathbb{R}_{>0}$ is a constant adaptation gain.

VI. Stability Analysis

In this section, stability analysis of the observer-controller RL architecture will be carried out using Lyapunov methods. To facilitate the stability analysis, the following rank condition is utilized in the stability analysis

Assumption 2. There exists a constant $\underline{c}_1$ such that the finite set of trajectories $\{x_i:\mathbb{R}_{\geq t_0}\mid i=1,\ldots,N\}$ satisfies

$$0<\underline{c}_1\leq\inf_{t\in\mathbb{R}_{\geq T}}\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\omega_i(t)\omega_i^{\mathrm{T}}(t)}{\rho_i^2(t)}\right). \tag{39}$$

As described in [4], since ω_i is a function of x_i and $\hat{W}_c$, Assumption 2 cannot be guaranteed a priori. However, unlike the PE condition utilized in [33], Assumption 2 can be verified online. Furthermore, since $\lambda_{\min}\left(\sum_{i=1}^{N}\frac{\omega_i(t)\omega_i^{\mathrm{T}}(t)}{\rho_i^2(t)}\right)$ is non-decreasing in the number of samples, N, Assumption 2 can be met, heuristically, by increasing the number of extrapolation trajectories. The calculation of a precise bound on the number of extrapolation trajectories is out of the scope of this paper.

Let $Z:=[x^{\mathrm{T}},e^{\mathrm{T}},\tilde{W}_c^{\mathrm{T}}]^{\mathrm{T}}$ represent the concatenated state of the closed-loop system and let a continuously differentiable candidate Lyapunov function, $V_L:R^{2n+L}\times\mathbb{R}_{\geq 0}\to\mathbb{R}$, be defined as,

$$V_L(Z,t):=V^*(x)+\frac{1}{2}\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}(t)\tilde{W}_c+V_e(e), \tag{40}$$

where V^* represent the optimal value function and the lyapunov function V_e is introduced in Section III. To facilitate the stability analysis, let $\chi\subset\mathcal{C}\times\mathcal{D}\times\mathbb{R}^L$ be an open set, let $\underline{c}\in\mathbb{R}_{>0}$ be a constant defined as $\underline{c}:=\frac{\beta}{2\Gamma k_c}+\frac{\underline{c}_1}{2}$, and let $\iota\in\mathbb{R}$ be a positive constant defined as

$$\iota:=\frac{L_{g\sigma}^2\overline{W}^2}{2\lambda^2\lambda_{\min}(P)}+\frac{3\overline{\|G_{r\sigma}\|}^2}{4\lambda^2k_c\underline{c}}+(1/2\lambda)\overline{\|G_r\|}\overline{\|\nabla\epsilon\|}+\lambda L_g\overline{\|\nabla\epsilon\|}+\frac{3k_c}{4\underline{c}}\|\frac{\omega_i}{\rho_i}\|^2\|\Delta_i\|^2, \tag{41}$$

where $G_{r\sigma}(x):=R^{-1}g(x)^{\mathrm{T}}\nabla_x\sigma^{\mathrm{T}}(x)$, $G_{r\sigma}(\hat{x}):=R^{-1}g(\hat{x})^{\mathrm{T}}\nabla_{\hat{x}}\sigma(\hat{x})^{\mathrm{T}}$, $G_r(x):=R^{-1}g(x)^{\mathrm{T}}$, and $L_{g\sigma}$ denotes the Lipschitz constant of $G_{r\sigma}$ over the set χ. As shown in [17, Lemma 1], provided (2) holds and $\lambda_{\min}\{\Gamma_0^{-1}\}>0$, the update law in (38) ensures that the least squares update law satisfies

$$\underline{\Gamma}I_L\leq\Gamma(t)\leq\overline{\Gamma}I_L, \tag{42}$$

$\forall t\in\mathbb{R}_{\geq 0}$ and some $\overline{\Gamma},\underline{\Gamma}>0$. Since the candidate Lyapunov function is positive definite, [29, Lemma 4.3] and the bound in (42) can be used to conclude that it is bounded as

$$\underline{\upsilon}(\|Z\|)\leq V_L(Z,t)\leq\overline{\upsilon}(\|Z\|), \tag{43}$$

for all $t\in\mathbb{R}_{\geq 0}$ and for all $Z\in\mathbb{R}^{2n+L}$, where $\underline{\upsilon},\overline{\upsilon}:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions. Let $\upsilon_l:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ be a class $\mathcal{K}$ function such that $\upsilon_l(\|Z\|)\leq\frac{\lambda_{\min}(Q)}{2}\|x\|^2+\frac{k_c\underline{c}}{6}\|\tilde{W}_c\|^2+\frac{\lambda_{\min}(P)}{4}\|e\|^2$.

Theorem 2. Provided Assumptions 1 and 2 hold, there exists a symmetric positive definite matrix P, and observer gains L, H, K that satisfy the matrix inequality in (18), the control gains are selected large enough based on the sufficient condition [1]

$$\upsilon_l{}^{-1}(\iota)\leq\overline{\upsilon}^{-1}(\underline{\upsilon}(\zeta)), \tag{44}$$

and the weights $\hat{W}_c$ and Γ are updated according to (38), then the concatenated state, Z, is locally uniformly ultimately bounded under the controller designed in (33).

Proof. The orbital derivative of the candidate Lyapunov function, V_L, along the trajectories of (1), (8), (38) is given by

$$\dot{V}_L(Z,t)=\nabla_xV^*(x)\dot{x}-\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}(t)\dot{\hat{W}}_c-\frac{1}{2}\tilde{W}_c^{\mathrm{T}}\Gamma^{-1}\dot{\Gamma}\Gamma^{-1}(t)\tilde{W}_c+\dot{V}_e(e). \tag{45}$$

Substituting (1), (21), (27), (28), (33), and (38) in (45), using the fact that $\frac{\omega_i\omega_i{}^{\mathrm{T}}}{\rho_i^2}\leq\frac{\omega_i\omega_i{}^{\mathrm{T}}}{\rho_i}$, applying completing of squares, triangle inequality and Cauchy Schwartz inequality, the orbital derivative is bounded, on the set $\chi\times\mathbb{R}_{\geq 0}$, as

$$\dot{V}_L(Z,t)\leq-\lambda_{\min}(Q)\|x\|^2-\frac{k_c\underline{c}}{3}\|\tilde{W}_c\|^2-\frac{\lambda_{\min}(P)}{2}\|e\|^2+\iota. \tag{46}$$

Let ζ be a constant such that $B_\zeta\subset\chi$. Based on the conditions stated in (44) and (21), the orbital derivative can be bounded as

$$\dot{V}_L(Z,t)\leq-\upsilon_l(\|Z\|),\forall\upsilon_l{}^{-1}(\iota)<\|Z\|<\zeta,\forall t\geq 0. \tag{47}$$

Using the sufficient condition stated in (44), [29, Theorem 4.18] can be invoked to conclude that Z is locally uniformly ultimately bounded. In particular, all trajectories starting from initial conditions bounded by $\|Z(0)\|\leq$

[1] Despite the fact that ι generally increases with increasing ζ, the condition in (44) can be satisfied provided the points for BE extrapolation are selected such that $\underline{c}$, introduced in (VI) and control gain, k_c is large enough, and the basis for the value function approximation are selected such that $\overline{\|\epsilon\|}$ and $\overline{\|\nabla\epsilon\|}$ are sufficiently small.

$\overline{v}^{-1}\left(\underline{v_l}\left(\zeta\right)\right)$ remain with χ for all $t \geq 0$ and satisfy $\limsup_{t\to\infty}\|Z(t)\| \leq \overline{v}^{-1}\left(\underline{v_l}\left(\iota\right)\right)$. Therefore, provided $\|Z(0)\| \leq \overline{v}^{-1}\left(\underline{v_l}\left(\zeta\right)\right)$, the state and the state estimates, under the controller in (33) and the observer in (7), remain within the compact set $\mathcal{C}$ where the Jacobian bounds and the Lipschitz constants are valid. □

VII. Conclusion

An observer-controller framework for output feedback RL in input-constrained nonlinear systems is developed. LMIs are formulated to obtain observer gain matrices and an MBRL-based controller is developed that maintains stability while finding an approximate solution to the optimal control problem. Simulation results demonstrate the effectiveness of the developed method and local uniform ultimately boundedness of the system states is guaranteed using a Lyapunov-based stability analysis.

If the LMI is poorly conditioned, this can lead to rank deficiency in certain regions of the state space. To address these numerical issues, the current LMI architecture can be augmented with techniques such as [34] which uses a delta operator formulation of the LMI. Future research will also involve introducing a system identifier into the observer RL architecture that learns the system's dynamics for systems where the parameters of the system model are uncertain.

References

[1] R. Kamalapurkar, "Model-based reinforcement learning for online approximate optimal control," Ph.D. dissertation, University of Florida, 2014.

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

[3] R. V. Self, M. Harlan, and R. Kamalapurkar, "Model-based reinforcement learning for output-feedback optimal control of a class of nonlinear systems," in *Proc. Am. Control Conf.*, Philadelphia, PA, USA, Jul. 2019, pp. 2378–2383.

[4] S. M. N. Mahmud, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, "Safe model-based reinforcement learning for systems with parametric uncertainties," *Front. Robot. AI*, vol. 8, no. 733104, pp. 1–13, Dec. 2021.

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

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

[7] 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.

[8] 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.

[9] Y. Wang, R. Rajamani, and D. M. Bevly, "Observer Design for Parameter Varying Differentiable Nonlinear Systems, With Application toSlip Angle Estimation," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1940–1945, 2017.

[10] M. Arcak and P. Kokotović, "Nonlinear observers: a circle criterion design and robustness analysis," *Automatica*, vol. 37, no. 12, pp. 1923–1930, 2001.

[11] B. Açıkmeşe and M. Corless, "Stability analysis with quadratic Lyapunov functions: Some necessary and sufficient multiplier conditions," *Syst. Control Lett.*, vol. 57, no. 1, pp. 78–94, 2008.

[12] D. Quintana, V. Estrada-Manzo, and M. Bernal, "An exact handling of the gradient for overcoming persistent problems in nonlinear observer design via convex optimization techniques," *Fuzzy Sets Syst.*, vol. 416, pp. 125–140, 2021, systems Engineering.

[13] H. Modares and F. L. Lewis, "Online solution to the linear quadratic tracking problem of continuous-time systems using reinforcement learning," in *Proc. IEEE Conf. Decis. Control*, Florence, IT, Dec. 2013, pp. 3851–3856.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] Y. H. Kim, F. L. Lewis, and C. T. Abdallah, "A dynamic recurrent neural-network-based adaptive observer for a class of nonlinear systems," *Automatica*, vol. 33, pp. 1539–1543, 1997.

[20] F. Abdollahi, H. A. Talebi, and R. V. Patel, "A stable neural network-based observer with application to flexible-joint manipulators," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 118–129, 2006.

[21] X. Yang, D. Liu, and Q. Wei, "Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming," *IET Control Theory Appl.*, vol. 8, no. 16, pp. 1676–1688, 2014.

[22] Y. Huang and H. Jiang, "Neural network observer-based optimal control for unknown nonlinear systems with control constraints," in *Int. Joint Conf. Neural Netw.*, 2015, pp. 1–7.

[23] M. Farza, A. Sboui, E. Cherrier, and M. M'Saad, "High-gain observer for a class of time-delay nonlinear systems," *Int. J. Control*, vol. 83, no. 2, pp. 273–280, 2010.

[24] A. Zemouche, M. Boutayeb, and G. Bara, "Observer Design for Nonlinear systems: An Approach Based on the Differential Mean Value Theorem." in *Proc. IEEE Conf. Decis. Control*, 2005, pp. 6353–6358.

[25] Y. Wang, R. Rajamani, and D. M. Bevly, "Observer design for differentiable Lipschitz nonlinear systems with time-varying parameters," in *Proc. IEEE Conf. Decis. Control*, 2014, pp. 145–152.

[26] R. Rajamani, W. Jeon, H. Movahedi, and A. Zemouche, "On the need for switched-gain observers for non-monotonic nonlinear systems," *Automatica*, vol. 114, p. 108814, 2020.

[27] A. Zemouche, M. Boutayeb, and G. Bara, "Observer Design for Nonlinear Systems: An Approach Based on the Differential Mean Value Theorem." in *Proc. IEEE Conf. Decis. Control*, 2005, pp. 6353–6358.

[28] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994.

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

[30] R. Kamalapurkar, P. Walters, J. A. Rosenfeld, and W. E. Dixon, *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*, ser. Communications and Control Engineering. Springer International Publishing, 2018.

[31] F. Sauvigny, *Partial Differential Equations 1*. Springer, 2012.

[32] 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.

[33] 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.

[34] B. Lennartson and R. Middleton, "Numerical sensitivity of Linear Matrix Inequalities for shorter sampling periods," in *Proc. IEEE Conf. Decis. Control*, 2012, pp. 4247–4252.