Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

REGULAR PAPER

Safe adaptive output-feedback optimal control of a class of linear systems

S M Nahid Mahmud[1] | Moad Abudia[2] | Scott A. Nivison[3] | Zachary I. Bell[4] | Rushikesh Kamalapurkar[2]

[1]School of Aeronautics and Astronautics, Purdue University, Indiana, USA
[2]School of Mechanical and Aerospace Engineering, Oklahoma State University, Oklahoma, USA
[3]Johns Hopkins University Applied Physics Laboratory, Florida, USA
[4]Air Force Research Laboratories, Florida, USA

Correspondence
*Rushikesh Kamalapurkar, School of Mechanical and Aerospace Engineering, Oklahoma State University, Oklahoma.
Email: rushikesh.kamalapurkar@okstate.edu

Funding Information
This research was supported, in part, by the Air Force Research Laboratories award number FA8651-23-1-0006, National Science Foundation award number 2027999 and Office of Naval Research award number N00014-21-1-2481. 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.

Abstract

The objective of this research is to enable safety-critical systems to simultaneously learn and execute optimal control policies in a safe manner to achieve complex autonomy. Learning optimal policies via trial and error, i.e., traditional reinforcement learning, is difficult to implement in safety-critical systems, particularly when task restarts are unavailable. Safe model-based reinforcement learning techniques based on a barrier transformation have recently been developed to address this problem. However, these methods rely on full-state feedback, limiting their usability in a real-world environment. In this work, an output-feedback safe model-based reinforcement learning technique based on a novel barrier-aware dynamic state estimator has been designed to address this issue. The developed approach facilitates simultaneous learning and execution of safe control policies for safety-critical linear systems. Simulation results indicate that barrier transformation is an effective approach to achieve online reinforcement learning in safety-critical systems using output feedback.

KEYWORDS:
Output-feedback Control, Optimal Control, Reinforcement Learning, Adaptive Control.

1 | INTRODUCTION

Over the past decade, safe reinforcement learning has gained a lot of attention in the disciplines of robotics and controls. One of the primary reasons for this focus is the increase in the expectation of autonomy in safety-critical systems in real-world tasks. While unmanned autonomous systems have significant advantages, such as repeatability, precision, and lack of physical weariness, over their non-autonomous and biological counterparts, they are often costly to construct and restore. To avoid failures during the learning phase, methods that allow unmanned autonomous agents to learn to perform tasks with safety guarantees are needed.

In the past, reinforcement learning (RL) has been demonstrated to be an effective approach for synthesizing online optimal policies for known and unknown discrete/continuous-time dynamical systems[1,2]. However, due to sample inefficiencies, RL often necessitates a large number of iterations. Model-based reinforcement learning (MBRL) techniques can enhance sample efficiency in RL[3,4,5]. Generally MBRL techniques guarantee stability, not safety. In recent years, significant progress has been made in developing safe model-based reinforcement learning (SMBRL) techniques to learn safe controllers for different classes

of systems [6,7,8,9,10,11,12,13,14,15,16]. While Markov decision process (MDP) based SMBRL methods have been available for discrete time systems with finite state and action spaces [6,7,8,9], synthesizing online controllers for systems in continuous time, under output feedback, while guaranteeing stability and safety is still a challenging problem.

SMBRL techniques that provide probabilistic safety guarantees for continuous-time (CT) stochastic systems have been studied in results such as [10,11,12]; however, not all applications are conducive to probabilistic safety guarantees. Applications such as manned aviation demand deterministic safety guarantees, and as such, online, real-time learning in such systems is challenging. SMBRL techniques for CT deterministic systems have been studied in results such as [13,14,15,16]. In [13], a SMBRL method has been developed to synthesize a real-time safe controller online by incorporating the proximity penalty method developed in [17] with the framework of control barrier functions. While the control barrier function results in safety guarantees, the existence of a smooth value function, in spite of a nonsmooth cost function, needs to be assumed [16].

This paper is inspired by the nonlinear coordinate transformation first introduced in [18]. Leveraging the results of [18], a barrier transformation (BT) to construct an equivalent, unconstrained optimal control problem from a state-constrained optimal control problem was introduced in [14]. The unconstrained problem was then solved using an adaptive optimal control method under persistence of excitation (PE). To soften the restrictive PE requirement, [15] utilized a MBRL formulation to yield a SMBRL technique to synthesize safe controllers. However, the SMBRL method in [15] requires exact model knowledge. To address this limitation, [16] extended the results of [15] to yield a SMBRL solution to the online state-constrained optimal feedback control problem under parametric uncertainty.

While results such as [14,15,16] provide verified safe feedback controllers, they all rely on full-state feedback. Often in unknown or/and adverse environments, the system may not have access to the full state variables. Therefore, development of methods to synthesize safe controllers using BT-based SMBRL that depend on output feedback is needed. The primary challenge in output feedback BT-based SMBRL is that the BT preserves neither the linearity nor the Brunovsky canonical form of the system. As such, observer development in the transformed coordinates is difficult. While state estimation can be done in the original coordinates, due to the nature of the BT, small estimation errors in the original coordinates do not translate to small errors in the transformed coordinates. Since the controllers are designed in transformed coordinates, large state estimation errors in the transformed coordinates can yield unexpected results. To achieve safe learning using output feedback, [19] extended the results in [14,15,16] for nonlinear control-affine systems in Brunovsky canonical form, where the state comprises of the output and its derivatives. This paper aims to extend the results in [19] by developing a controller for partial observable state-constrained linear systems, not necessarily in the Brunovski canonical form, with a more general output equation. To meet such an objective, this paper develops a Luenberger-like BT-based adaptive observer[1], and utilizing the observer, this paper develops an output feedback SMBRL method for linear systems with an output matrix that satisfies Assumption 1 to learn feedback control policies with guaranteed safety and stability during the learning and execution phases.

In the following, Section 2 formalizes the problem statement. Section 3.1 introduces the BT, and Section 3.2 details the developed BT-based state estimator. Section 3.3 describes the SMBRL technique for synthesizing feedback control policies in transformed coordinates. In Section 3.3, a Lypaunov-based analysis, in the transformed coordinates, is utilized to establish practical stability of the closed-loop system resulting from the developed SMBRL technique. Guarantees that the safety requirements are satisfied in the original coordinates are also established. Simulation results in Section 4 show the performance of the developed SMBRL approach.

2 | PROBLEM FORMULATION

We consider the following continuous-time linear dynamical system.

$$\dot{x} = Ax + Bu, \quad y = Cx \tag{1}$$

where $x := [x_1; \dots; x_n] \in \mathbb{R}^n$ is the system state, $A \in \mathbb{R}^{n\times n}$ is the transition matrix, $B \in \mathbb{R}^{n\times m}$ is the control effectiveness matrix, $u \in \mathbb{R}^m$ is the control input, $C \in \mathbb{R}^{q\times n}$ is the output matrix, $y := [y_1; \dots; y_q] \in \mathbb{R}^q$ is the measured output, and the notation $[v; w]$ is used to denote the vector $[v^T \quad w^T]^T$. The following structure is imposed on the problem to make the barrier transformation feasible.

[1]Classical Luenberger observer [20] can not be directly integrated into a BT-based SMBRL framework since the observer becomes infeasible in the transformed coordinate due to the nonlinear structure of the BT.

Assumption 1. For each $j \in \{1, \dots, q\}$, there exists $i \in \{1, \dots, n\}$ such that $y_j = x_i$.

Assumption 1 requires each output variable to be equal to one of the state variables. That is, every row of the matrix C has exactly one element equal to one, and the rest of the elements are zero. Since generally, q is smaller than n, Assumption 1 requires direct measurement of a subset of the state variables. Most autonomous systems commonly use sensors that directly measure one or more state variables. For example, systems such as robot manipulators with encoders that measure angular position, autonomous vehicles with GPS sensors that measure position, gyroscopes that measure angular velocities, and magnetometers that measure yaw angles, electrical circuits equipped with ammeters all have in output matrices that satisfy Assumption 1.

The objective is to design an estimator to estimate the state online, using input-output measurements, and to simultaneously estimate and utilize an output feedback optimal controller, u, such that starting from a given feasible initial condition x^0, the trajectories $x(\cdot)$ decay to a neighborhood of the origin, and satisfies $x_i(t) \in (\underline{z}_i, \overline{z}_i), \forall i = 1, \dots, n$ for user-specified constants $\underline{z}_i < 0 < \overline{z}_i$ define user-specified safety constraints.

The notation $(\cdot)_i$ is used in the rest of the manuscript to denote the ith element of the vector $(\cdot)$, and the notation I_o denotes the identity matrix of size o.

3 | MAIN RESULTS

In this paper, adaptive optimal control techniques developed for unconstrained optimal control problems are adapted to solve the constrained optimal output feedback optimal control problem at hand. To facilitate conversion of a constrained optimal control problem into an unconstrained one, we utilize the BT, first introduced in [14].

3.1 | Barrier Transformation

Given $\underline{\mathscr{E}} < 0 < \overline{\mathscr{E}}$, the function $b_{(\underline{\mathscr{E}}, \overline{\mathscr{E}})} : (\underline{\mathscr{E}}, \overline{\mathscr{E}}) \to \mathbb{R}$, referred to as barrier function (BF), is defined as $b_{(\underline{\mathscr{E}}, \overline{\mathscr{E}})}(\mathscr{Y}) := \log\left(\frac{\overline{\mathscr{E}}(\underline{\mathscr{E}} - \mathscr{Y})}{\underline{\mathscr{E}}(\overline{\mathscr{E}} - \mathscr{Y})}\right)$. The inverse of the BF on the interval $(\underline{\mathscr{E}}, \overline{\mathscr{E}})$ containing the origin is given by $b^{-1}_{(\underline{\mathscr{E}}, \overline{\mathscr{E}})}(\mathscr{Y}) = \left(\frac{\underline{\mathscr{E}}\overline{\mathscr{E}}\, e^{\mathscr{Y}} - 1}{\underline{\mathscr{E}} e^{\mathscr{Y}} - \overline{\mathscr{E}}}\right)$. In this paper, the barrier transformation is a nonlinear coordinate transformation, defined as $s_i := b_{(\underline{z}_i, \overline{z}_i)}(x_i)$, and $x_i = b^{-1}_{(\underline{z}_i, \overline{z}_i)}(s_i)$, where s_i is the ith element of the transformed system state vector, s. Evaluating the derivative of the inverse of the barrier function at s_i yields $\frac{\mathrm{d}b^{-1}_{(\underline{z}_i, \overline{z}_i)}(s_i)}{\mathrm{d}s_i} = \frac{1}{T_i(s_i)}$, where $T_i(s_i) := \frac{\underline{z}_i^2 e^{s_i} - 2\underline{z}_i \overline{z}_i + \overline{z}_i^2 e^{-s_i}}{\overline{z}_i \underline{z}_i^2 - \underline{z}_i \overline{z}_i^2}$. Let $T(s) := [T_1(s_1); \dots; T_n(s_n)]$.

Let $\underline{z} = [\underline{z}_1; \dots; \underline{z}_n]$, and $\overline{z} = [\overline{z}_1; \dots; \overline{z}_n]$. In the following, for any vector $\mathscr{L}$, comprised of components of x, such that $\mathscr{L} = [(x)_p; \dots; (x)_q]$, with $1 \le p \le q \le n$, the notation $b(\mathscr{L})$ is used to denote componentwise application of the barrier function with the appropriate limits selected componentwise from the vectors $\underline{z}$ and $\overline{z}$. That is, $b(\mathscr{L}) := [b_{((\underline{z})_p, (\overline{z})_p)}((x)_p); \dots; b_{((\underline{z})_q, (\overline{z})_q)}((x)_q)]$. Similarly, given any vector $\mathscr{L}$, comprised of components of s, such that $\mathscr{L} = [(s)_p, \cdots, (s)_q]^T$, with $1 \le p \le q \le n$, $b^{-1}(\mathscr{L}) := [b^{-1}_{((\underline{z})_p, (\overline{z})_p)}((s)_p); \dots; b^{-1}_{((\underline{z})_q, (\overline{z})_q)}((s)_q)]^T$.

3.2 | State Estimator

The first technical challenge is designing an estimator to generate safe estimates $\hat{x} \in \mathbb{R}^{2n}$ of the state variables x online, using input-output measurements. In this section, a barrier-based adaptive state estimator inspired by Luenberger observer [20] has been designed to generate estimates of x. The designed estimator is given by

$$\dot{\hat{x}}_i := (Bu)_i + \frac{1}{T_i(b(\hat{x}_i))}\Big(Ab(\hat{x}) + L\left(y_m - Cb(\hat{x})\right)\Big)_i. \tag{2}$$

for $i = 1, \dots, n$, where $y_m = Cb(x)$ and $L \in \mathbb{R}^{n \times q}$ is the gain matrix selected to make $A - LC$ Hurwitz. Note that the design is motivated by the need to obtain the error bound in Lemma 2. Since the estimator needs to deal with the state constraints, BT using the barrier function can be an effective way to address this challenge. Which will transfer the constrained state to the unconstrained state. To transform the dynamics in (1) using the BT, the time derivative of the transformed state, $s \in \mathbb{R}^n$, can be computed as

$$\dot{s} = T(s) \odot (Ax + Bu) = F(s) + G(s)u \tag{3}$$

where, $\odot$ represents Hadamard product[21, Chapter 1], $F(s) := T(s) \odot Ab^{-1}(s)$, $G(s) := T(s) \odot B$.

To transform the dynamics in (2) using the BT, the time derivative of the estimated transformed state, $\hat{s} \in \mathbb{R}^n$, can be computed as

$$\begin{aligned}\dot{\hat{s}} &= T(\hat{s}) \odot (Bu) + A\hat{s} + L\left(y_m - Cb(\hat{x})\right) \\ &= A\hat{s} + G(\hat{s})u + LC\tilde{s}. \end{aligned} \tag{4}$$

Note that the relationship $b(Cx) = Cb(x)$, leveraged in the computation above, is only true under Assumption 1.

In the transformed coordinates, the estimator error can be computed as

$$\dot{\tilde{s}} = F(s) + G(s)u - G(\hat{s})u - As + A\tilde{s} - LC\tilde{s}, \tag{5}$$

with $\tilde{s} = s - \hat{s}, \quad \dot{\tilde{s}} = \dot{s} - \dot{\hat{s}}$.

A detailed analysis of the relationship between the trajectories of the transformed state and state-estimator dynamics and the original state and state estimator dynamics is provided in Section 3.3.

The second technical challenge is designing an optimal output feedback controller such that starting from a given feasible initial condition x^0, the closed-loop trajectory $x(\cdot)$ decays to a neighborhood of the origin and satisfies $x_i(t) \in (\underline{z}_i, \overline{z}_i), \forall t \geq 0$, where $i = 1, 2$. Lemma 1 in Section 3.3 implies that if a feedback controller that practically stabilizes the transformed system in (3) is designed, then the same feedback controller can be adapted to keep the trajectories of the original system within the safe bounds. The following subsection develops a novel OF-SMBRL technique for synthesizing such a feedback controller.

3.3 | Safe Model-based Reinforcement Learning

In the following, a controller that practically stabilizes (3) is designed as an estimate of a controller that minimizes the infinite horizon cost

$$J(u(\cdot)) := \int_0^\infty c(\phi(\tau, s^0, u(\cdot)), u(\tau))d\tau, \tag{6}$$

over the set $\mathcal{U}$ of piecewise continuous functions $t \mapsto u(t)$, subject to (3), where $\phi(\tau, s^0, u(\cdot))$ denotes the trajectory of (3), evaluated at time τ, starting from the state s^0, and under the controller $u(\cdot)$. In (6), $c(s, u) := Q'(s) + u^T Ru$ where $R \in \mathbb{R}^{m\times m}$ is a symmetric positive definite (PD) matrix. For the optimal value function to be a Lyapunov function for the optimal policy, it is assumed that Q' is PD. A state penalty function $x \mapsto E(x)$, given in the original coordinates, can easily be transformed into an equivalent state penalty $Q'(s) = E(b^{-1}(s))$. Since the barrier function is monotonic and $b(0) = 0$, if E is PD, then so is Q'. Furthermore, for applications with bounded control inputs, a non-quadratic penalty function similar to[22, Eq. 17] can be incorporated in (6).

Assuming that an optimal controller exists, let the optimal value function, denoted by $V^* : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}$, be defined as

$$V^*(s) := \min_{u(\cdot)\in\mathcal{U}_{[t,\infty)}} \int_t^\infty c(\phi(\tau, s, u_{[0,\tau)}(\cdot)), u(\cdot))d\tau, \tag{7}$$

where u_I and $\mathcal{U}_I$ are obtained by restricting the domains of u and functions in $\mathcal{U}_I$ to the interval $I \subseteq \mathbb{R}$, respectively. 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[23, Theorem 1.5]

$$\min_{u\in\mathbb{R}^q} \left(V_s\left(F(s) + G(s)u\right) + Q'(s) + u^T Ru\right) = 0, \tag{8}$$

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

$$u^*(s) := -\frac{1}{2}R^{-1}G(s)^T(\nabla_s V^*(s))^T. \tag{9}$$

Considering that analytical solutions to the HJB problem are often infeasible to compute, particularly for nonlinear systems, parametric approximation methods are utilized to estimate the value function V^* and the optimal policy u^*.

The optimal value function can be expressed as

$$V^*(s) = W^T\sigma(s) + \epsilon(s), \tag{10}$$

where $W \in \mathbb{R}^l$ is an unknown vector of bounded weights, $\sigma : \mathbb{R}^n \to \mathbb{R}^l$ is a vector of continuously differentiable nonlinear activation functions[24, Def. 2.1] such that $\sigma(0) = 0$ and $\nabla_s \sigma(0) = 0$, $l \in \mathbb{N}$ is the number of basis functions, and $\epsilon : \mathbb{R}^n \to \mathbb{R}$ is the reconstruction error. Using the universal function approximation property[25, Theorem 1.5] of single layer neural networks (NNs), it can be deduced that given the compact set $\overline{B}(0, \chi) \subset \mathbb{R}^n$ and a positive constant $\overline{\epsilon} \in \mathbb{R}$, there exists a number of basis functions $l \in \mathbb{N}$, and known positive constants $\overline{W}$ and $\overline{\sigma}$ such that $\|W\| \leq \overline{W}$, $\sup_{s \in \overline{B}(0,\chi)} \|\epsilon(s)\| \leq \overline{\epsilon}$, $\sup_{s \in \overline{B}(0,\chi)} \left\|\nabla_s \epsilon(s)\right\| \leq \overline{\epsilon}$, $\sup_{s \in \overline{B}(0,\chi)} \|\sigma(s)\| \leq \overline{\sigma}$, and $\sup_{s \in \overline{B}(0,\chi)} \left\|\nabla_s \sigma(s)\right\| \leq \overline{\sigma}$. Note that $\overline{B}(0, \chi)$ is not assumed to be forward invariant. See Remark 2.

Using (8), a representation of the optimal controller using the same basis as the optimal value function can be derived as

$$u^*(s) = -\frac{1}{2} R^{-1} G^T(s) \left(\nabla_s \sigma^T(s) W + \nabla_s \epsilon^T(s)\right). \tag{11}$$

Since the ideal weights, W, are unknown, an actor-critic technique is utilized in the following to estimate W. Let the NN estimates $\hat{V} : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ and $\hat{u} : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^m$ be defined as

$$\hat{V}\left(\hat{s}, \hat{W}_c\right) := \hat{W}_c^T \sigma(\hat{s}), \tag{12}$$

$$\hat{u}\left(\hat{s}, \hat{W}_a\right) := -\frac{1}{2} R^{-1} G^T(\hat{s}) \nabla_{\hat{s}} \sigma^T(\hat{s}) \hat{W}_a, \tag{13}$$

where the critic weights, $\hat{W}_c \in \mathbb{R}^l$ and actor weights, $\hat{W}_a \in \mathbb{R}^l$ are estimates of the ideal weights, W.

Using the estimate $\hat{s}$ of the transformed state s in (3) and substituting (12) and (13) into (8) for results in a 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), defined as

$$\hat{\delta}(\hat{s}, \hat{W}_c, \hat{W}_a) := \hat{V}_{\hat{s}}(\hat{s}, \hat{W}_c)\left(F(\hat{s}) + G(\hat{s})\hat{u}(\hat{s}, \hat{W}_a)\right) + Q'(\hat{s}) + \hat{u}(\hat{s}, \hat{W}_a)^T R \hat{u}(\hat{s}, \hat{W}_a). \tag{14}$$

To learn the approximation control policy, online RL algorithms traditionally require a PE condition[4,26,27]. It is typically impossible to guarantee PE a priori and verify PE online. While impossible to ensure a priori, by utilizing the model's virtual excitation, stability and convergence of online RL can be established under a PE-like condition that can be validated online (by monitoring the minimum eigenvalue of a matrix in the subsequent Assumption 2)[5]. The BE can be evaluated at any arbitrary point in the state space using the system model. Virtual excitation can then be implemented by selecting a set of points $\left\{r_k \mid k = 1, \cdots, N\right\}$, where $r_k \in \overline{B}(0, \chi)$. The extrapolation state variables r_k are assumed to be constant. However, the approach can be extended in a straightforward manner to time-varying extrapolation state variables confined to a compact neighborhood of the origin. Evaluating the BE at this set of state variables to yield

$$\hat{\delta}_k(r_k, \hat{W}_c, \hat{W}_a) := \hat{V}_{r_k}(r_k, \hat{W}_c)\left(F(r_k) + G(r_k)\hat{u}(r_k, \hat{W}_a)\right) + Q'(r_k) + \hat{u}(r_k, \hat{W}_a)^T R \hat{u}(r_k, \hat{W}_a). \tag{15}$$

To facilitate the analysis, the actor and critic weight estimation errors are defined as $\tilde{W}_c := W - \hat{W}_c$ and $\tilde{W}_a := W - \hat{W}_a$ and substituting the estimates (10) and (11) into (8), and subtracting from (14), the BE that can be expressed in terms of the weight estimation errors as

$$\hat{\delta} = -\omega^T \tilde{W}_c + \frac{1}{4} \tilde{W}_a^T G_\sigma \tilde{W}_a + \Delta, \tag{16}$$

where $\Delta := \frac{1}{2} W^T \nabla_{\hat{s}} \sigma G_R \nabla_{\hat{s}} \epsilon^T + \frac{1}{4} G_\epsilon - \nabla_{\hat{s}} \epsilon F$, $G_R := G R^{-1} G^T$, $G_\epsilon := \nabla_{\hat{s}} \epsilon G_R \nabla_{\hat{s}} \epsilon^T$, $G_\sigma := \nabla_{\hat{s}} \sigma G R^{-1} G^T \nabla_{\hat{s}} \sigma^T$, and $\omega := \nabla_{\hat{s}} \sigma \left(F + G\hat{u}\left(\hat{s}, \hat{W}_a\right)\right)$. Similarly, (15) implies that

$$\hat{\delta}_k = -\omega_k^T \tilde{W}_c + \frac{1}{4} \tilde{W}_a^T G_{\sigma_k} \tilde{W}_a + \Delta_k, \tag{17}$$

where, $\Delta_k := \frac{1}{2} W^T \nabla_{r_k} \sigma_k G_{R_k} \nabla_{r_k} \epsilon_k^T + \frac{1}{4} G_{\epsilon_k} - \nabla_{r_k} \epsilon_k F_k$, $G_{\epsilon_k} := \nabla_{r_k} \epsilon_k G_{R_k} \nabla_{r_k} \epsilon_k^T$, $\omega_k := \nabla_{r_k} \sigma_k \left(F_k + G_k \hat{u}\left(\hat{z}^k, \hat{W}_a\right)\right)$, $G_{\sigma_k} := \nabla_{r_k} \sigma_k G_k R^{-1} G_k^T \nabla_{r_k} \sigma_k^T$, $G_{R_k} := G_k R^{-1} G_k^T$, $F_k := F(r_k)$, $G_k := G(r_k)$, $H_k := H(r_k)$, $\sigma_k := \sigma(r_k)$, and $\epsilon_k := \epsilon(r_k)$.

Note that $\sup_{s \in \overline{B}(0,\chi)} |\Delta| \leq d_a \overline{\epsilon}$ and if $r_k \in \overline{B}(0, \chi)$ then $\left|\Delta_k\right| \leq d_a \overline{\epsilon}_k$, for some constant $d_a > 0$, and the dependence of various functions on the state, $\hat{s}$, are omitted for brevity whenever it is clear from the context.

Using the extrapolated BEs $\hat{\delta}_k$ from (15), the weights are updated according to

$$\dot{\hat{W}}_c = -\frac{k_c}{N} \Gamma \sum_{k=1}^{N} \frac{\omega_k}{\rho_k} \hat{\delta}_k, \tag{18}$$

$$\dot{\Gamma} = \beta \Gamma - \frac{k_c}{N} \Gamma \sum_{k=1}^{N} \frac{\omega_k \omega_k^T}{\rho_k^2} \Gamma, \tag{19}$$

and

$$\dot{\hat{W}}_a = -k_a\left(\hat{W}_a - \hat{W}_c\right) + \sum_{k=1}^{N} \frac{k_c G_{\sigma_k}^T \hat{W}_a \omega_k^T}{4N\rho_k} \hat{W}_c - k_b \hat{W}_a, \tag{20}$$

with $\Gamma\left(t_0\right) = \Gamma_0$, where $\Gamma : \mathbb{R}_{\geq t_0} \to \mathbb{R}^{l\times l}$ is a time-varying least-squares gain matrix, $\rho_k(t) := 1 + \gamma\omega_k^T(t)\,\omega_k(t)$, $\gamma > 0$ is a constant positive normalization gain, $\beta \in \mathbb{R}_+$ is a constant forgetting factor, and $k_c, k_a, k_b \in \mathbb{R}_+$ are constant adaptation gains. The control commands sent to the system are then computed using the actor weights as

$$u(t) = \hat{u}\left(\hat{s}(t), \hat{W}_a(t)\right), \quad t \geq 0. \tag{21}$$

The Lyapunov function needed to analyze the closed loop system defined by (3), (5), (18), (19), and (20) is constructed using stability properties of (3) under the optimal feedback controller (9). To that end, the following section analyzes the optimal closed-loop system.

Using the assumption that $Q'(s)$ is PD,[19, Theorem 1a], and the converse Lyapunov theorem for asymptotic stability[28, Theorem 4.17], the existence of a radially unbounded PD function $\mathcal{V} : \mathbb{R}^n \to \mathbb{R}$ and a PD function $W : \mathbb{R}^n \to \mathbb{R}$ is guaranteed such that

$$\mathcal{V}_s(s)(F(s) + G(s)u^*(s)) \leq -W(s), \tag{22}$$

for all $s \in \mathbb{R}^n$. The functions $\mathcal{V}$ and W are utilized in the following section to analyze the stability of the output feedback approximate optimal controller.

To summarize, the controller in (21) is implemented by selecting a set of basis functions σ, input parameters $k_a > 0$, $k_b > 0$, $k_c > 0$ and $\beta > 0$, and initial conditions $\hat{s}(0), \hat{W}_c(0), \Gamma(0)$ and $\hat{W}_a(0)$, and using (13) where $\hat{s}(t), \hat{W}_c(t), \Gamma(t)$ and $\hat{W}_a(t)$ are computed by solving a system of differential equations, comprised of (4), (18), (19), and (20), respectively.

3.4 | Stability Analysis

In the following lemma, the trajectories of the original system and the transformed system are shown to be related by the barrier transformation provided the trajectories of the transformed system are *complete*[16]. As detailed in Lemma 1, the design of the BT ensures that the trajectories of (1), (3), (2), and (4) are linked by the BT whenever the underlying state trajectories $x(\cdot)$ and $s(\cdot)$ and the initial conditions $\hat{x}^0$ and $\hat{s}^0$ are linked by the BT.

Lemma 1. If $t \mapsto \Phi\left(t; [b(x^0); b(\hat{x}^0)], \zeta\right)$ is a complete Carathéodory solution to (3), and (4), starting from the initial condition $[b(x^0); b(\hat{x}^0)]$, under the feedback policy $(\hat{s}, t) \mapsto \zeta(\hat{s}, t)$, and if $t \mapsto \Lambda\left(t; [x^0; \hat{x}^0], \xi\right)$ is a Carathéodory solution to (1), and (2), starting from the initial condition $[x^0; \hat{x}^0]$, under the feedback policy $(\hat{x}, t) \mapsto \xi(\hat{x}, t)$, defined as $\xi(\hat{x}, t) = \zeta(b(\hat{x}), t)$ then

$$\Lambda\left(t; [x^0; \hat{x}^0], \xi\right) = b^{-1}\left(\Phi\left(t; [b(x^0); b(\hat{x}^0)], \zeta\right)\right)$$

for all $t \in \mathbb{R}_{\geq 0}$.

Proof. See[19, Lemma 1]. □

Remark 1. The feedback ξ is well-defined at $\hat{x}$ only if $b(\hat{x})$ is well-defined, which is the case whenever $\hat{x}$ is inside the barrier. As such, the main conclusion of the lemma also implies that $\Lambda(\cdot, [x^0; \hat{x}^0], \xi)$ remains inside the barrier which indicates the safety of the trajectories by satisfying the user-specified safety constraints. It is thus inferred from Lemma 1 that if the trajectories of (3), and (2) are bounded and decay to a neighborhood of the origin under a feedback policy $(\hat{s}, t) \mapsto \zeta(\hat{s}, t)$, then the feedback policy $(x, t) \mapsto \zeta\left(b(\hat{x}), t\right)$, when applied to the original system in (1), achieves the control objective stated in Section 2.

The following PE-like rank condition is utilized in the stability analysis.

Assumption 2. There exists a constant $\underline{c}_1 > 0$ such that the set of points $\left\{r_k \in \mathbb{R}^n \mid k = 1, \ldots, N\right\}$ satisfies

$$\underline{c}_1 I_l \leq \inf_{t\in\mathbb{R}_{\geq T}} \left(\frac{1}{N}\sum_{k=1}^{N} \frac{\omega_k(t)\,\omega_k^T(t)}{\rho_k^2(t)}\right). \tag{23}$$

Similar to the PE condition traditionally used in online optimal control (see, e.g.,[29]), Assumption 2 cannot be guaranteed a priori as ω_k is a function of the estimates $\hat{s}$ and $\hat{W}_a$. Online verification of the PE condition would require knowledge of a $\delta t > 0$ and computation of the integral $\int_t^{t+\delta t} \frac{\omega(\tau)\omega^T(\tau)}{\rho^2(\tau)} d\tau$ at each t, to check if it has a sufficiently large minimum eigenvalue. Even if the integrals could be efficiently computed, finding a $\delta t > 0$ such that the integrals have a sufficiently large minimum eigenvalue for all t is typically infeasible, making the PE condition impossible to verify online. On the other hand, at each time instant τ, given

the state and parameter estimates, one can easily compute the minimum eigenvalue of $\frac{\omega_k(\tau)\omega_k^T(\tau)}{\rho_k^2(\tau)}$ and check whether it is greater than $\underline{c}_1$. As such, at any given $\tau \geq T$, by keeping track of the smallest minimum eigenvalue recorded over $[T, \tau]$, it is possible to verify the condition $\underline{c}_1 I_l \leq \inf_{t\in[T,\tau]} \left(\frac{1}{N} \sum_{k=1}^{N} \frac{\omega_k(t)\omega_k^T(t)}{\rho_k^2(t)} \right)$.

Furthermore, if the minimum eigenvalue does not satisfy the lower bound, one can either perturb the set of points $\left\{ r_k \in \mathbb{R}^n \mid k = 1, \ldots, N \right\}$ or collect more points to improve the minimum eigenvalue without injecting an excitation signal into the system. Since $\lambda_{\min} \left(\sum_{k=1}^{N} \frac{\omega_k(t)\omega_k^T(t)}{\rho_k^2(t)} \right)$ is non-decreasing in the number of samples, N, Assumption 2 can be met, heuristically, by increasing N. This flexibility makes Assumption 2 much easier to work with than the PE condition.

It is established in [4, Lemma 1] that under Assumption 2 and provided $\lambda_{\min} \left\{ \Gamma_0^{-1} \right\} > 0$, the update law in (19) ensures that the least squares gain matrix satisfies

$$\underline{\Gamma} I_L \leq \Gamma(t) \leq \overline{\Gamma} I_L, \forall t \in \mathbb{R}_{\geq 0} \text{ and some } \overline{\Gamma}, \underline{\Gamma} > 0 \tag{24}$$

Using (3), the orbital derivative of the PD function $\mathcal{V}$ introduced in (22), along the trajectories of (3), under the controller $u = \hat{u}\left(\hat{s}, \hat{W}_a\right)$, is given by $\dot{\mathcal{V}}\left(s, \tilde{s}, \tilde{W}_a\right) = \mathcal{V}_s(s)\left(F(s) + G(s)\hat{u}\left(\hat{s}, \hat{W}_a\right)\right)$, where $\tilde{s} := s - \hat{s}$.

Using (22) and the facts that G is bounded, the basis functions σ are bounded, and the value function approximation error ϵ and its derivative with respect to $s, \hat{s}$ are bounded on compact sets, the time-derivative of $\mathcal{V}$ can be bounded as

$$\dot{\mathcal{V}}\left(s, \tilde{s}, \tilde{W}_a\right) \leq -W(s) + \iota_1 \overline{\epsilon} + \iota_2 \|\tilde{s}\| \left\|\tilde{W}_a\right\| + \iota_3 \left\|\tilde{W}_a\right\| + \iota_4 \|\tilde{s}\|, \tag{25}$$

for all $\hat{W}_a \in \mathbb{R}^l$, for all $s \in \overline{B}(0, \chi)$, and for all $\hat{s} \in \overline{B}(0, \chi)$, where $\iota_1, \cdots, \iota_4$ are positive constants.

Let $\Theta\left(\tilde{W}_c, \tilde{W}_a, t\right) := \frac{1}{2}\tilde{W}_c^T \Gamma^{-1}(t) \tilde{W}_c + \frac{1}{2}\tilde{W}_a^T \tilde{W}_a$. The orbital derivative of Θ along the trajectories of (18) - (20) is given by

$$\dot{\Theta}\left(\tilde{W}_c, \tilde{W}_a, t\right) = \tilde{W}_c^T \Gamma^{-1} \dot{\tilde{W}}_c - \frac{1}{2}\tilde{W}_c^T \Gamma^{-1} \dot{\Gamma} \Gamma^{-1} \tilde{W}_c + \tilde{W}_a^T \dot{\tilde{W}}_a, \tag{26}$$

where $\dot{\tilde{W}}_c = -\dot{\hat{W}}_c$, and $\dot{\tilde{W}}_a = -\dot{\hat{W}}_a$.

Provided the extrapolation state variables are selected such that $r_k \in \overline{B}(0, \chi)$, $\forall k = \{1, \ldots, N\}$, the orbital derivative in (26) can be bounded by

$$\dot{\Theta}\left(\tilde{W}_c, \tilde{W}_a, t\right) \leq -k_c \underline{c} \left\|\tilde{W}_c\right\|^2 - \left(k_a + k_b\right) \left\|\tilde{W}_a\right\|^2 + k_c \iota_8 \overline{\epsilon} \left\|\tilde{W}_c\right\| + k_c \iota_5 \left\|\tilde{W}_a\right\|^2 + \left(k_c \iota_6 + k_a\right) \left\|\tilde{W}_c\right\| \left\|\tilde{W}_a\right\| + \left(k_c \iota_7 + k_b \overline{W}\right) \left\|\tilde{W}_a\right\|, \tag{27}$$

for all $t \geq 0$, where $\iota_5, \ldots, \iota_8$ 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 $\underline{c} = \inf_{t\geq 0} \lambda_{\min} \left\{ \left(\frac{\beta}{2k_c} \Gamma^{-1}(t) + \frac{1}{2N} \sum_{k=1}^{N} \frac{\omega_k \omega_k^T}{\rho_k} \right) \right\}$. Assumption 2 and (24) guarantee that $\underline{c} > 0$. The following Lemma develops a bound on a Lyapunov-like function of the state estimation errors to be utilized in this stability analysis.

Lemma 2. Let $V_{se} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a continuously differentiable candidate Lyapunov function defined as $V_{se}(\tilde{s}, t) := \tilde{s}^T P \tilde{s}$ where P is a PD matrix that satisfies $P(A - LC) + (A - LC)^T P = -\zeta$ for some PD matrix ζ. Provided $\hat{s}(t) \in \overline{B}(0, \chi)$, $s(t) \in \overline{B}(0, \chi)$ for all t, and F, G are locally Lipschitz continuous on $\overline{B}(0, \chi)$ for $\chi > 0$, the orbital derivative of V_{se} along the trajectories of (5), under the controller in (21), can be upper-bounded as $\dot{V}_{se}(\tilde{s}, s, \tilde{W}_a, t) \leq -\left(\lambda_{min}(\zeta) - \varpi_2\right) \|\tilde{s}\|^2 + \varpi_1 \|\tilde{s}\| \|s\| + \varpi_3 \|\tilde{s}\| \|\tilde{W}_a\| + \varpi_4 \|\tilde{s}\|$ where $\varpi_1; \cdots; \varpi_4$ are Lipschitz constants.

Proof. See Appendix 6.1. □

Remark 2. The bound on $\dot{V}_{se}$ established above is only valid if s and $\hat{s}$ remain in $\overline{B}(0, \chi)$. The fact that s and $\hat{s}$ remain in $\overline{B}(0, \chi)$ if trajectories of the closed loop are initialized sufficiently close to the origin is proven rigorously in Theorem 1 using standard local uniform ultimate boundedness results (see, e.g., [28, Theorem 4.18]).

Utilizing the results from (25), (27) and Lemma 2, the following theorem can be obtained

Theorem 1. Provided Assumptions 1 and 2 hold, and F, G are locally Lipschitz continuous on $\overline{B}(0, \chi)$ for $\chi > 0$, the gains are selected large enough to ensure that the sufficient condition (39), introduced in Appendix 6.2 holds, the matrix $M + M^T$, defined in (36) is PD, and the weights $\hat{W}_c$, Γ, and $\hat{W}_a$ are updated according to (18), (19), and (20), respectively, then the estimation errors $\tilde{W}_c$, $\tilde{W}_a$, and the trajectories of the transformed system in (3), under the controller in (21), are locally uniformly ultimately bounded.

Proof. See Appendix 6.2 □

Remark 3. The sufficient condition in (39) and the matrices S and M depend on the upper and lower bounds of the optimal value function. Such bounds can be difficult to obtain in practice. This limitation stems from the use of the optimal value function as a component of the candidate Lyapunov function and is present in all online adaptive optimal control results rely on the optimal value function (see, e.g., [30,31,23,32,33]).

The main conclusion of the Lemma 1 (see Remark 1) implies that if the trajectories of (3) and (4) are bounded and decay to a neighborhood of the origin under a feedback policy $(\hat{s}, t) \mapsto \zeta(\hat{s}, t)$, then the same feedback policy, when applied to the original system in (1), achieves the control objective stated in Section 2. Theorem 1 proves that the trajectories of (3) and (4) under the controller in (21), are locally uniformly ultimately bounded. Therefore, it can be concluded that the controller in (21) can be utilized in the original coordinates to keep the trajectories of the original system within the safe bounds to meet the second technical challenge of this paper.

4 | SIMULATION

This section applies the proposed OF-SMBRL framework to an output feedback controller synthesis problem for a linearized F-16 aircraft longitudinal dynamical system. The linearized F-16 aircraft longitudinal dynamical system[31] is described by (1), where

$$A = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 \\ 0.82225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}.$$

The goal is to learn a stabilizing output feedback control law that drives the state variables to the origin while respecting the state constraints. In this experiment, the state is expressed as $x = [x_1; x_2; x_3] = [\alpha; q; \delta e]$ where α denotes the angle of attack [rad]; q is the pitch rate [rad/sec]; δe is the elevator deflection angle [rad]. The objective is to satisfy the user picked constraints that are $x_1 \in (\underline{z}_1, \overline{z}_1)$, $x_2 \in (\underline{z}_2, \overline{z}_2)$, and $x_3 \in (\underline{z}_3, \overline{z}_3)$ where $\underline{z}_1 = \underline{z}_2 = \underline{z}_3 = -1$, $\overline{z}_1 = \overline{z}_3 = 1$ and $\overline{z}_2 = 0.3$. The system is required to remain in the user-specified bound during the entire duration of the experiment.

To synthesize the controller, the infinite horizon cost in (6) is minimized with $Q'(s) = s^T Q s$ where $Q = 10 I_2$ and $R = 1$. The basis functions for value function approximation are selected as $\sigma(\hat{s}) = [\hat{s}_1\hat{s}_2; \hat{s}_1\hat{s}_3; \hat{s}_2\hat{s}_3; \hat{s}_1^2; \hat{s}_2^2; \hat{s}_3^2]$. The initial conditions for the state variables are selected as $x(0) = [\alpha_0; q_0; \delta e_0] = [0.95; 0; 0.90]$. The initial conditions for the estimated state variables are $\hat{x}(0) = [0.75; -0.75; 0.75]$, and the initial guesses for the weights are $L = [1; 1; 1]$, $\Gamma(0) = I_6$, $\hat{W}_a(0) = [1; 1; 1; 1; 1; 1]$, $\hat{W}_c(0) = [1; 1; 1; 1; 1; 1]$, $K_c = 100$, $K_{a_1} = 100$, $K_{a_2} = 1$, and $\beta = 0.1$. The simulation uses 125 fixed Bellman error extrapolation points in a 0.8×0.6 ×0.8 cube around the origin of the $s-$coordinate system. Since the critic and the actor are trained in the transformed coordinates where the system is nonlinear, the true actor and critic weights are unknown.

Figure 1 (a) and Figure 1 (b) indicate that the system state x remains inside the user-specified safe set under the developed OF-SMBRL scheme, whereas an observer-based LQR solution[34] results in violation of the state constraints. Figure 1 (c) shows that the state estimation errors also converge to zero. As observed from the results in Figure 1 (d), the unknown weights for critic converge to constant values. Since the true actor and critic weights are unknown, the final values that the critic weights converge to in Figure 1 (d) cannot be compared against the true weights. The simulation results indicate that the developed method successfully achieves the stated control objectives.

Remark 4. In this paper, a controller synthesis method for output feedback LQR problems under inequality constraints on the state variables is developed. To the best of our knowledge, online solutions to the state-constrained output feedback LQR problem do not exist in the literature, and as such, we are unable to directly compare the developed method to other output-feedback techniques. However, we have included Figure 1 (a) and Figure 1 (b), where the developed method is compared against a baseline unconstrained output-feedback LQR controller that uses a standard Luenberger observer to generate estimates of the system state.

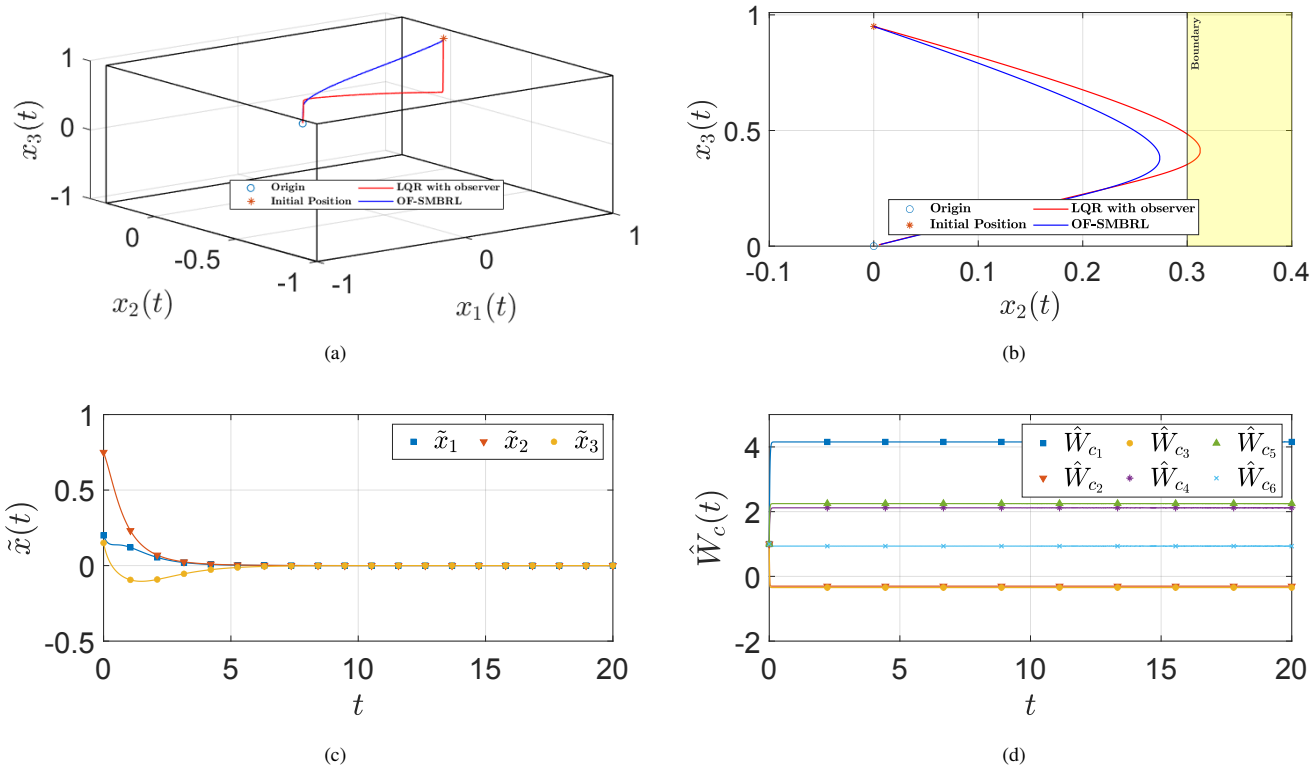

Figure 1 Plot (a) represents a phase portrait of the F-16 aircraft longitudinal dynamical system's state variables in the original coordinates using OF-SMBRL and observer-based LQR solution[34]. The boxed area represents the user-selected safe set. The planes of the box indicate the user-selected safety boundary. Plot (b) shows that the observer-based LQR solution violates the user-selected safety boundary. However, OF-SMBRL satisfies the safety boundary. The yellow region indicates an unsafe region. Plot (c) shows the effectiveness of the developed state estimator. It shows the estimation errors between the original and estimated state variables under selected nominal gains. Plot (d) presents the estimates of the critic weights under selected nominal gains.

5 | CONCLUSION

This paper presents a novel framework that utilizes a new barrier-based adaptive state estimator to yield a safe MBRL-based online, approximate optimal controller synthesizing technique for safety-critical linear systems, under output feedback. BT, a transformation method to transform a constrained optimal control problem into an unconstrained optimal control problem, facilitates existing MBRL techniques to obtain safe optimal controllers in the original coordinate. The newly designed Luenberger-like state estimator enables the SMBRL framework to provide an OF-SMBRL controller that guarantees the state variables of the original system remain within the safety bounds. Regulation of the system state variables to a neighborhood of the origin and convergence of the estimated policy to a neighborhood of the optimal policy is established using a Lyapunov-based stability analysis.

While the simulation results are promising, safety violations are possible due to unmodeled uncertainties in the system dynamics and/or the environment. Furthermore, the simulation study indicates that the technique is sensitive to initial guesses of the unknown policy and the unknown value function, as predicted by the local stability result. Future research targeting these limitations will pave the way for the barrier transformation approach in safety-critical applications such as autonomous driving and manned aviation.

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Sutton RS, Barto AG, Williams RJ. Reinforcement learning is direct adaptive optimal control. *IEEE Control Syst. Mag.* 1992; 12(2): 19–22.
2. Doya K. Reinforcement learning in continuous time and space. *Neural Comput.* 2000; 12(1): 219–245.
3. Wawrzyński P. Real-time reinforcement learning by sequential actor-critics and experience replay. *Neural Netw.* 2009; 22(10): 1484–1497.
4. Kamalapurkar R, Rosenfeld JA, Dixon WE. Efficient model-based reinforcement learning for approximate online optimal control. *Automatica* 2016; 74: 247–258.
5. Kamalapurkar R, Walters P, Dixon WE. Model-based reinforcement learning for approximate optimal regulation. *Automatica* 2016; 64: 94–104.
6. Howard RA. *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press . 1960.
7. Meyn SP, Tweedie RL. Stability of Markovian processes i: criteria for discrete-time chains. *Adv. Appl. Probab.* 1992; 24(3): 542–574.
8. Puterman M. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons . 2014.
9. Wang D, Wang J, Zhao M, Xin P, Qiao J. Adaptive Multi-Step Evaluation Design With Stability Guarantee for Discrete-Time Optimal Learning Control. IEEE/CAA Journal of Automatica Sinica; 2023
10. Fisac JF, Akametalu AK, Zeilinger MN, Kaynama S, Gillula J, Tomlin CJ. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Trans. Autom. Control* 2018; 64(7): 2737–2752.
11. Li Z, Kalabić U, Chu T. Safe Reinforcement Learning: Learning with supervision using a constraint-admissible set. *Proc. Am. Control Conf.* 2018: 6390–6395.
12. Li Y, Li N, Tseng HE, Girard A, Filev D, Kolmanovsky I. Safe reinforcement learning using robust action governor. *Learning for Dynamics and Control* 2021: 1093–1104.
13. Cohen MH, Belta C. Safe exploration in model-based reinforcement learning using control barrier functions. *Automatica* 2023; 147: 110684.
14. Yang Y, Vamvoudakis KG, Modares H, He W, Yin YX, Wunsch D. Safety-Aware Reinforcement Learning Framework with an Actor-Critic-Barrier Structure. *Proc. Am. Control Conf.* 2019: 2352–2358.
15. Greene ML, Deptula P, Nivison S, Dixon WE. Sparse Learning-Based Approximate Dynamic Programming With Barrier Constraints. *IEEE Control Syst. Lett.* 2020; 4(3): 743-748.
16. Mahmud SMN, Nivison SA, Bell ZI, Kamalapurkar R. Safe model-based reinforcement learning for systems with parametric uncertainties. *Front. Robot. AI* 2021; 8(733104): 1–13.

17. Deptula P, Chen H, Licitra RA, Rosenfeld JA, Dixon WE. Approximate Optimal Motion Planning to Avoid Unknown Moving Avoidance Regions. *IEEE Transactions on Robotics* 2020; 36(2): 414-430.

18. Graichen K, Petit N. Incorporating a class of constraints into the dynamics of optimal control problems. *Optim. Control Appl. Methods* 2009; 30(6): 537-561.

19. Mahmud SMN, Abudia M, Nivison SA, Bell ZI, Kamalapurkar R. Safety aware model-based reinforcement learning for optimal control of a class of output-feedback nonlinear systems. arXiv:2110.00271; 2021.

20. Luenberger . Observing the state of a linear system. *IEEE Trans. Mil. Electron.* 1964; 8: 74–80.

21. Horn RA. The Hadamard product. *Proc. Symp. Appl. Math* 1990; 40: 87–169.

22. Yang Y, Ding DW, Xiong H, Yin Y, Wunsch DC. Online barrier-actor-critic learning for H∞ control with full-state constraints and input saturation. *J. Franklin Inst.* 2020; 357(6): 3316 - 3344.

23. Kamalapurkar R, Walters P, Rosenfeld JA, Dixon WE. *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Communications and Control EngineeringSpringer International Publishing . 2018.

24. Sadegh N. A perceptron network for functional identification and control of nonlinear systems. *IEEE Trans. Neural Netw.* 1993; 4(6): 982–988.

25. Sauvigny F. *Partial Differential Equations 1*. Springer . 2012.

26. Modares H, Lewis FL, Naghibi-Sistani MB. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 2013; 24(10): 1513–1525.

27. Kiumarsi B, Lewis FL, Modares H, Karimpour A, Naghibi-Sistani MB. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica* 2014; 50(4): 1167–1175.

28. Khalil HK. *Nonlinear systems*. Upper Saddle River, NJ: Prentice Hall. third ed. 2002.

29. Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica* 2010; 46(5): 878–888.

30. Lewis FL, Vamvoudakis KG. Reinforcement Learning for Partially Observable Dynamic Processes: Adaptive Dynamic Programming Using Measured Output Data. *IEEE Trans. Syst. Man Cybern.* 2011; 41(1): 14-25.

31. Vamvoudakis KG, Vrabie D, Lewis FL. Online adaptive algorithm for optimal control with integral reinforcement learning. *Int. J. Robust Nonlinear Control* 2014; 24(17): 2686–2710.

32. Vamvoudakis KG, Hespanha JP. Cooperative Q-Learning for Rejection of Persistent Adversarial Inputs in Networked Linear Quadratic Systems. *IEEE Trans. Autom. Control* 2018; 63(4): 1018-1031.

33. Kanellopoulos A, Vamvoudakis KG. A Moving Target Defense Control Framework for Cyber-Physical Systems. *IEEE Trans. Autom. Control* 2020; 65(3): 1029-1043.

34. Lewis FL, Vrabie D, Syrmos VL. *Optimal control*. Hoboken, NJ: Wiley. third ed. 2012.

6 | APPENDIX

6.1 | Proof of Lemma 2

Lemma 2. Let $V_{se} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a continuously differentiable candidate Lyapunov function defined as $V_{se}(\tilde{s}, t) := \tilde{s}^T P \tilde{s}$ where P is a PD matrix that satisfies $P(A - LC) + (A - LC)^T P = -\zeta$ for some PD matrix ζ. Provided $\hat{s}(t) \in \overline{B}(0, \chi)$, $s(t) \in \overline{B}(0, \chi)$ for all t, and F, G are locally Lipschitz continuous on $\overline{B}(0, \chi)$ for some $\chi > 0$, the orbital derivative of V_{se} along the trajectories of (5), under the controller in (21), can be upper-bounded as $\dot{V}_{se}(\tilde{s}, s, \tilde{W}_a, t) \leq -\left(\lambda_{min}(\zeta) - \varpi_2\right) \|\tilde{s}\|^2 + \varpi_1 \|\tilde{s}\| \|s\| + \varpi_3 \|\tilde{s}\| \|\tilde{W}_a\| + \varpi_4 \|\tilde{s}\|$ where $\varpi_1; \cdots; \varpi_4$ are Lipschitz constants.

Proof. The estimator error in the transformed coordinate can be computed as

$$\dot{\tilde{s}} = F(s) + G(s)\hat{u} - G(\hat{s})\hat{u} - As + A\tilde{s} - LC\tilde{s}. \tag{28}$$

The orbital derivative can be expressed as

$$\dot{V}_{se} = \tilde{s}^T P\dot{\tilde{s}} + \dot{\tilde{s}}^T P\tilde{s}. \tag{29}$$

Using (28), we can rewrite (29) as

$$\dot{V}_{se} = \tilde{s}^T P\left(F(s) + G(s)\hat{u} - G(\hat{s})\hat{u} - As + A\tilde{s} - LC\tilde{s}\right) + \left(F(s) + G(s)\hat{u} - G(\hat{s})\hat{u} - As + A\tilde{s} - LC\tilde{s}\right)^T P\tilde{s} \tag{30}$$

which yields

$$\dot{V}_{se} = \tilde{s}^T P(A - LC)\tilde{s} + \tilde{s}^T PF(s) + \tilde{s}^T P\tilde{G}(s,\hat{s})\hat{u} - \tilde{s}PAs + \tilde{s}^T(A - LC)^T P\tilde{s} + (\tilde{s}^T PF(s))^T + (\tilde{s}^T P\tilde{G}(s,\hat{s})\hat{u})^T - (\tilde{s}PAs)^T. \tag{31}$$

We can rewrite (31) as

$$\begin{aligned}\dot{V}_{se} = \tilde{s}^T\Big(P(A - LC) + (A - LC)^T P\Big)\tilde{s} + \tilde{s}^T PF(s) + \left(\tilde{s}^T PF(s)\right)^T - \tilde{s}PAs - (\tilde{s}PAs)^T - \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}\left(s,\tilde{W}_a\right)\\ + \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}\left(s,\tilde{W}_a\right) - \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}\left(\hat{s},\tilde{W}_a\right) - \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,W) + \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(\hat{s},W) + \tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,W)\\ - \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,\tilde{W}_a)\right)^T + \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,\tilde{W}_a)\right)^T - \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(\hat{s},\tilde{W}_a)\right)^T - \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,W)\right)^T\\ + \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(\hat{s},W)\right)^T + \left(\tilde{s}^T P\tilde{G}(s,\hat{s})\,\hat{u}(s,W)\right)^T.\end{aligned} \tag{32}$$

Using the Cauchy-Schwarz inequality and the fact that F, G are locally Lipschitz continuous on $\overline{B}(0,\chi)$, we have

$$\dot{V}_{se} \leq -\lambda_{min}(\zeta)\|\tilde{s}\|^2 + \varpi_1\|\tilde{s}\|\|s\| + \varpi_2\|\tilde{s}\|\|\tilde{s}\| + \varpi_3\|\tilde{s}\|\|\tilde{W}_a\| + \varpi_4\|\tilde{s}\| \tag{33}$$

where $\varpi_1; \cdots; \varpi_4$ are Lipschitz constants.

From (33), we obtain the desired bound

$$\dot{V}_{se} \leq -\left(\lambda_{min}(\zeta) - \varpi_2\right)\|\tilde{s}\|^2 + \varpi_1\|\tilde{s}\|\|s\| + \varpi_3\|\tilde{s}\|\|\tilde{W}_a\| + \varpi_4\|\tilde{s}\|. \tag{34}$$

□

6.2 | Proof of Theorem 1

Theorem 1. Provided Assumptions 1 and 2 hold, and F, G are locally Lipschitz continuous on $\overline{B}(0,\chi)$ for $\chi > 0$, the gains are selected large enough to ensure that the sufficient condition (39), introduced in Appendix 6.2 holds, the matrix $M + M^T$, defined in (36) is PD, and the weights $\hat{W}_c$, Γ, and $\hat{W}_a$ are updated according to (18), (19), and (20), respectively, then the estimation errors $\tilde{W}_c$, $\tilde{W}_a$, and the trajectories of the transformed system in (3), under the controller in (21), are locally uniformly ultimately bounded.

Proof. The candidate Lyapunov function for the closed-loop system is selected as

$$V_L(Z,t) := \mathcal{V}(s) + \Theta\left(\tilde{W}_c, \tilde{W}_a, t\right) + V_{se}(\tilde{s}), \tag{35}$$

where $Z := \left[s^T \;\; \tilde{W}_c^T \;\; \tilde{W}_a^T \;\; \tilde{s}^T\right]^T$.

Let $\mathcal{C} \subset \mathbb{R}^{2n}$ be a compact set defined as

$$\mathcal{C} := \left\{(s,\tilde{s}) \in \mathbb{R}^{2n} \mid \|s\| + \|\tilde{s}\| \leq \chi\right\}.$$

Whenever, $(s,\tilde{s}) \in \mathcal{C}$, it can be concluded that $s, \hat{s} \in \overline{B}(0,\chi)$. As a result, (25), (27), and (33) imply that whenever $Z \in \mathcal{C}\times\mathbb{R}^{2l}$, the orbital derivative of the candidate Lyapunov function along the trajectories of (3), (5), (18), (19), (20), under the controller (21), can be bounded as

$$\begin{aligned}\dot{V}_L(Z,t) \leq -W(s) + \iota_1\overline{\epsilon} + \iota_2\|\tilde{s}\|\left\|\tilde{W}_a\right\| + \iota_3\left\|\tilde{W}_a\right\| + \iota_4\|\tilde{s}\| - k_c\underline{c}\left\|\tilde{W}_c\right\|^2 - \left(k_a + k_b\right)\left\|\tilde{W}_a\right\|^2 + k_c\iota_8\overline{\epsilon}\left\|\tilde{W}_c\right\|\\ + k_c\iota_5\left\|\tilde{W}_a\right\|^2 + \left(k_c\iota_6 + k_a\right)\left\|\tilde{W}_c\right\|\left\|\tilde{W}_a\right\| + \left(k_c\iota_7 + k_b\overline{W}\right)\left\|\tilde{W}_a\right\| - \left(\lambda_{min}(\zeta) - \varpi_2\right)\|\tilde{s}\|^2\\ + \left(\varpi_1\right)\|\tilde{s}\|\|s\| + \varpi_3\|\tilde{s}\|\|\tilde{W}_a\| + \varpi_4\|\tilde{s}\|,\end{aligned}$$

which can be re-expressed as,

$$\dot{V}_L(Z,t) \leq -W(s) + \iota_1\overline{\epsilon} + (\iota_2+\varpi_3)\|\tilde{s}\|\left\|\tilde{W}_a\right\| - k_c\underline{c}\left\|\tilde{W}_c\right\|^2 - \left(k_a+k_b-k_c\iota_5\right)\left\|\tilde{W}_a\right\|^2 + k_c\iota_8\overline{\epsilon}\left\|\tilde{W}_c\right\| + \left(k_c\iota_6+k_a\right)\left\|\tilde{W}_c\right\|\left\|\tilde{W}_a\right\| + \left(k_c\iota_7+k_b\overline{W}+\iota_3\right)\left\|\tilde{W}_a\right\| + (\iota_4+\varpi_4)\|\tilde{s}\| - \left(\lambda_{min}(\zeta)-\varpi_2\right)\|\tilde{s}\|^2 + \varpi_1\|\tilde{s}\|\|s\|,$$

using Young's inequality

$$\dot{V}_L(Z,t) \leq -W(s) + \frac{1}{2}(\varpi_1)\|s\|^2 + \iota_1\overline{\epsilon} + (\iota_2+\varpi_3)\|\tilde{s}\|\left\|\tilde{W}_a\right\| + (\iota_4+\varpi_4)\|\tilde{s}\| - k_c\underline{c}\left\|\tilde{W}_c\right\|^2 - \left(k_a+k_b-k_c\iota_5\right)\left\|\tilde{W}_a\right\|^2 + k_c\iota_8\overline{\epsilon}\left\|\tilde{W}_c\right\| + \left(k_c\iota_6+k_a\right)\left\|\tilde{W}_c\right\|\left\|\tilde{W}_a\right\| + \left(k_c\iota_7+k_b\overline{W}+\iota_3\right)\left\|\tilde{W}_a\right\| - \left(\lambda_{min}(\zeta) - \frac{1}{2}\varpi_1 - \varpi_2\right)\|\tilde{s}\|^2,$$

where $z := \left[\left\|\tilde{W}_c\right\| \ \left\|\tilde{W}_a\right\| \ \|\tilde{s}\|\right]^T$. Provided the matrix $M + M^T$ is PD,

$$\dot{V}_L(Z,t) \leq -W(s) + \frac{1}{2}(\varpi_1)\|s\|^2 - \underline{M}\|z\|^2 + \overline{S}\|z\| + \iota_1\overline{\epsilon},$$

where $\underline{M} := \lambda_{\min}\left\{\frac{M+M^T}{2}\right\}$, $\overline{S} = \|S\|_\infty$, and the matrices M and S are defined as

$$S := \begin{bmatrix} k_c\iota_8\overline{\epsilon} \\ \left(k_c\iota_7 + k_b\overline{W} + \iota_3\right) \\ \iota_4 + \varpi_4 \end{bmatrix}^T, \qquad M := \begin{bmatrix} k_c\underline{c} & 0 & 0 \\ -\left(k_c\iota_6+k_a\right) & \left(k_a+k_b-k_c\iota_5\right) & 0 \\ 0 & -(\iota_2+\varpi_3) & \lambda_{min}(\zeta) - \frac{1}{2}\varpi_1 - \varpi_2 \end{bmatrix}^T. \tag{36}$$

Letting $\underline{M} =: \underline{M}_1 + \underline{M}_2$, and letting $\mathcal{W} : \mathbb{R}^{2n+2l} \to \mathbb{R}$ be defined as $\mathcal{W}(Z) = -W(s) + \frac{1}{2}(\varpi_1)\|s\|^2 - \underline{M}_1\|z\|^2$, with $W(s) > \frac{1}{2}(\varpi_1)\|s\|^2$, the orbital derivative can be bounded as

$$\dot{V}_L(Z,t) \leq -\mathcal{W}(Z), \quad \forall \|Z\| \tag{37}$$

such that $\|Z\| > \frac{1}{2}\left(\frac{\overline{S}}{\underline{M}_2} + \sqrt{\frac{\overline{S}^2}{\underline{M}_2^2} + \frac{\iota_1^2\overline{\epsilon}^2}{\underline{M}_2^2}}\right) =: \mu, \forall Z \in \overline{B}\left(0,\overline{\chi}\right)$, for all $t \geq 0$, and some $\overline{\chi}$ such that $\overline{B}(0,\overline{\chi}) \subseteq \mathcal{C} \times \mathbb{R}^{2l}$.

Using the bound in (24) and the fact that the converse Lyapunov function is guaranteed to be time-independent, radially unbounded, and PD,[28, Lemma 4.3] can be applied to conclude that

$$\underline{\upsilon}(\|Z\|) \leq V_L(Z,t) \leq \overline{\upsilon}(\|Z\|), \tag{38}$$

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

Provided the learning gains, the domain radii χ and $\overline{\chi}$, and the basis functions for function approximation are selected such that $M + M^T$ is PD and

$$\mu < \overline{\upsilon}^{-1}\left(\underline{\upsilon}\left(\overline{\chi}\right)\right), \tag{39}$$

then Theorem 4.18 from[28] can be invoked to conclude that Z is locally uniformly ultimately bounded. Since the estimates $\hat{W}_a$ approximate the ideal weights W, the policy $\hat{u}$ approximates the optimal policy u^*. □

□