AFRL-RW-EG-TR-2023-150

MODEL AWARE REINFORCEMENT LEARNING

Rushikesh Kamalapurkar

Oklahoma State University
401 Whitehurst Hall
Stillwater, OK 74078

May 2023

Final Report

AIR FORCE RESEARCH LABORATORY, MUNITIONS DIRECTORATE
Air Force Materiel Command • United States Air Force • Eglin Air Force Base

PUBLIC AFFAIRS SECURITY AND POLICY REVIEW WORKSHEET

(See page 2 for instructions)

1. DATE NEEDED	2. SUBMITTER REFERENCE NO.
Jul 27, 2023	

NOTE: Applica ion to clear information for Public Release. Public release clearance is NOT required for material presented in a closed meeting and which will not be made available to the general public, on the Internet, in print or electronic media. **Items marked with an asterisk (*) and Blocks 13-15 are required.**

3. SUBMITTER

*NAME: Zachary Bell
*PHONE: 850-512-7203 *ORG/OFC SYM: RWTAD
*EMAIL: zachary.bell.10@us.af.mil
*ORG. EMAIL: zachary.bell.10@us.af.mil

4. PRIMARY AUTHOR

*NAME: Zachary Bell
*PHONE: 850-512-7203
*ORG/OFC SYM: RWTAD
*EMAIL: zachary.bell.10@us.af.mil

***5. DOCUMENT TITLE**

Model Aware Reinforcement Learning

***6. CONFERENCE/EVENT/PUBLICATION/WEBSITE/PUBLIC WEB URL**

DTIC

***7. EVENT/PUBLICATION DATE**

Jul 27, 2023

***8. THE SPONSOR/OWNER of this conference/event/publication/website is:**

☐ *Industry* ☐ *Professional Association* ☐ *Academia* ☒ *Other:*

9. THIS DOCUMENT IS WRITTEN in my official/work** ☒ **or in a** ☐ **personal/private capacity.** ***(If personal/private, you must include a disclaimer as noted in AFI 35-101, para. 9.4)

***10. DOCUMENT TYPE**

Technical Report

OTHER

***11. BUDGET CATEGORY (Choose N/A if not applicable)**

6.1

OTHER

***12.a. NATIONAL SECURITY STATUTES/TECHNOLOGY ISSUES:** Are any aspects of this technology included in: U.S. Munitions List; ITAR 22, CFR Part 121; CCL; CIIL, S&T Protection Plan or Security Classification Guide? (If YES, explain rationale for release in Block 14)

☐ YES ☒ NO

***b.** Does this information meet the criteria for Public Release – unclassified, unlimited distribution?

☒ YES ☐ NO

c.** Are any references classified or subject to distribution limitations? (If YES***, *explain rationale for release in Block 14)*

☐ YES ☒ NO

d. STINFO.** Does this document contain technical information relating to research, development, engineering, testing, evaluation, production, operation or maintenance of any military/space equipment or technology? (If YES***, *the STINFO officer must sign in Block 18 per AFI 61-201, para. 3.4.5*)

☐ YES ☒ NO

***e. S&T.** S&T. Does this document contain any information about Critical Technology Elements (CTE) as defined in AFRL| 61-113; or is the effort covered by a Science and Technology Protection Plan? *(**If YES**, the S&T Protection Lead must sign in Block 19)*

☐ YES ☒ NO

***f.** Is this information identified as a topic of potential elevation for SAF review in AFI 35-101?

☐ YES ☒ NO *(**If YES,** explain in Block 14)*

***g.** If this material results from an international agreement, is the DoD authorized to release program information? *(**If NO**, identify release authority organization in Block 14)*

☐ YES ☐ NO ☒ N/A

***h.** If a joint program, does your organization maintain primary management responsibility and authority to release all information? *(If NO, provide name of lead organization / POC [i.e. DARPA, NASA, Army, Navy, etc.] in Block 14)*

☐ YES ☐ NO ☒ N/A

***13. PROJECT TYPE.** ☐ **PIA** ☐ **AFOSR** ☒ **6.x Funded** ☐ **Other**

If a PIA, enter the contract # below and the Approval Officer must sign in Block 18 or 19. **For AFOSR,** provide the LRIR and name of program officer below. **For 6.x** include the Work Unit # (WU) and work unit manager's name (WUM) below. **For Other** specify below.

Work Unit # W1AS, Manager: Zachary Bell

14. EXPLANATION *(Additional comments, previous related cases [include case number], additional coordination accomplished/required - continued on next page.)*

Technical report containing review of publicly released work under the program.

CERTIFICATION AND COORDINATION SIGNATURES. ***SIGNATURES MAY NOT BE REPEATED IN MULTIPLE BLOCKS.***
PER REGULATORY GUIDANCE, CONTRACTORS MAY NOT SIGN IN BLOCKS 15-20
NOTE: Once the first signature is applied, all blocks above except 1, 2 3 (emails) and 14 are locked and cannot be modified or changed.

15. DoD ORIGINATOR/PROGRAM MANAGER** ***(Required)
I certify the attached material is unclassified, technically accurate, contains no critical military technology, is not subject to export controls, has no foreign disclosure issues as outlined in para. 5.7 of AFMAN 16-201 and is suitable for public release.

NAME: Zachary Bell
ORG: RW OFFICE SYMBOL: RWTAD
SIGNATURE: BELL.ZACHARY.IAN.[redacted]
DUTY TITLE: Research Engineer DATE: Jun 27, 2023

***16. TECHNICAL REVIEW AND CERTIFICATION (Required)**
The information contained in the attached document is technically accurate; does not disclose classified, sensitive, or militarily critical technology; does not violate proprietary rights or copyright restrictions, is not subject to export control regulations and is suitable for public release.

NAME: Emily Doucette
ORG: AFRL OFFICE SYMBOL: RWTA
SIGNATURE: [redacted]
DUTY TITLE: Technical Advisor DATE: Jul 27, 2023

17. OPSEC MANAGER REVIEW** ***(Required)
I certify that the information has been reviewed and contains no Operational Security disclosure issues, is not prohibited by the CIIL, or otherwise prohibited by AFI 110-701.

NAME: burgess
ORG: OFFICE SYMBOL:
SIGNATURE: BURGESS.KURT.D.[redacted]
DUTY TITLE: DATE:

18. STINFO REVIEWER ONLY** ***(Required IF 12.d above is checked YES)
I certify that the information has been reviewed and may be released IAW DoDI 5230.24 and AFI 61-201.

NAME: Wilbur McDew
ORG: AFRL OFFICE SYMBOL: RW
SIGNATURE: MCDEW.WILBUR.1[redacted]
DUTY TITLE: STINFO DATE: Jul 31, 2023

19. S&T PROTECTION LEAD REVIEW ONLY ***(Required IF 12.e above is checked YES)***
I certify that the information has been reviewed and may be released IAW AFRL| 61-113.

NAME:
ORG: OFFICE SYMBOL:
SIGNATURE:
DUTY TITLE: DATE:

20. Additional review, if required by the unit or requested by PA
I certify that this information is suitable for public release.

NAME:
ORG: OFFICE SYMBOL:
SIGNATURE:
DUTY TITLE: DATE:

21. PA USE ONLY

NOTES:

☐ CLEARED ☐ AS AMENDED ☐ w/RECOMMENDATION ☐ OTHER (Annotate in notes)
☐ NO OBJECTION ☐ RETURN - NO ACTION ☐ NOT CLEARED ☐ OBJECTION

PUBLIC AFFAIRS OFFICER

CASE NUMBER

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
31-05-2023	Final Technical Report	16 August 2019 – 31 May 2023

4. TITLE AND SUBTITLE
Model Aware Reinforcement Learning

5a. CONTRACT NUMBER: N/A

5b. GRANT NUMBER
FA8651-19-2-0009

5c. PROGRAM ELEMENT NUMBER: N/A

6. AUTHOR(S)
Rushikesh Kamalapurkar

5d. PROJECT NUMBER: N/A

5e. TASK NUMBER: N/A

5f. WORK UNIT NUMBER
W1AS

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Oklahoma State University
401 Whitehurst Hall
Stillwater, OK 74078

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Munitions Directorate
101 West Eglin Blvd
Eglin AFB, FL 32542-6810

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RWTA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RW-EG-TR-2023-150

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION A – Approved for public release: Distribution Unlimited. AFRL/PA-2023-4147

13. SUPPLEMENTARY NOTES SUBJECT TO EXPORT CONTROL LAWS
DISTRIBUTION STATEMENT INDICATING AUTHORIZED ACCESS IS ON THE COVER PAGE AND BLOCK 12 OF THIS FORM. DATA RIGHTS RESTRICTIONS AND AVAILABILITY OF THIS REPORT ARE SHOWN ON THE NOTICE AND SIGNATURE PAGE.

14. ABSTRACT
The overall goal in this proposal was to develop a novel model-aware RL (MARL) framework for nonlinear systems in continuous time and space specifically focused on mitigating large modeling errors and maintaining closed-loop stability during the learning phase. The original scope of work was to focus on the development of model-aware RL methods that utilize parametric models of the environment, are robust to modeling errors, and can adapt online to changing models and objectives. Data-driven adaptive estimation techniques were proposed to achieve online model estimation. Novel real-time model validation methods were proposed to gauge the quality of the estimated models. Development of fall-back policies was proposed to achieve robust learning in the presence of inaccurate models. In addition, the development model-aware RL methods that utilize non-parametric models such as Gaussian Processes (GPs) was proposed along with the use of the confidence bounds obtained for the GPs to guide model-aware virtual exploration. The development of model-aware RL techniques that utilize local parametric and non-parametric models was also proposed to synthesize locally optimal policies.

15. SUBJECT TERMS
Model-based reinforcement learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Zachary I. Bell
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	19	**19b. TELEPHONE NUMBER** (850) 882-5326

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

Section **Page**

1.0. Goals and objectives

The overall goal in this proposal was to develop a novel model-aware RL (MARL) framework for nonlinear systems in continuous time and space specifically focused on mitigating large modeling errors and maintaining closed-loop stability during the learning phase. The original scope of work was to focus on the development of model-aware RL methods that utilize parametric models of the environment, are robust to modeling errors, and can adapt online to changing models and objectives. Data-driven adaptive estimation techniques were proposed to achieve online model estimation. Novel real-time model validation methods were proposed to gauge the quality of the estimated models. Development of fall-back policies was proposed to achieve robust learning in the presence of inaccurate models. In addition, the development model-aware RL methods that utilize non-parametric models such as Gaussian Processes (GPs) was proposed along with the use of the confidence bounds obtained for the GPs to guide model-aware virtual exploration. The development of model-aware RL techniques that utilize local parametric and non-parametric models was also proposed to synthesize locally optimal policies.

2.0. Findings of the investigators

Over the course of this project, the investigators made progress along three principal research directions, safety-aware model-based reinforcement learning, system identification, and learning-based control using output feedback. Published and in-review research articles that were supported, in part, by this grant are summarized below, followed by a brief description of the progress in each of the three areas above.

3.0. Published and accepted articles partially supported by this grant

[1] J. A. Rosenfeld, R. Kamalapurkar, B. Russo, and T. T. Johnson, "Occupation kernels and densely defined Liouville operators for system identification," in *Proc. IEEE Conf. Decis. Control*, 2019, pp. 6455–6460.

[2] M. Abudia, M. Harlan, R. V. Self, and R. Kamalapurkar, "Switched optimal control and dwell time constraints: A preliminary study," in *Proc. IEEE Conf. Decis. Control*, 2020, pp. 3261–3266.

[3] M. L. Greene, M. Abudia, R. Kamalapurkar, and W. E. Dixon, "Model-based reinforcement learning for optimal feedback control of switched systems," in *Proc. IEEE Conf. Decis. Control*, 2020, pp. 162–167.

[4] R. Kamalapurkar, W. E. Dixon, and A. R. Teel, "On reduction of differential inclusions and Lyapunov stability," *ESAIM Control Optim. Calc. Var.*, vol. 26, 2020.

[5] S. M. N. Mahmud, K. Hareland, S. Nivison, Z. I. Bell, and R. Kamalapurkar, "A safety aware model-based reinforcement learning framework for systems with uncertainties," in *Proc. Am. Control Conf.*, 2021, pp. 1979–1984.

[6] 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, 2021.

[7] J. A. Rosenfeld and R. Kamalapurkar, "Dynamic mode decomposition with control Liouville operators," in *IFAC-PapersOnLine*, vol. 54, 2021, pp. 707–712.

[8] G. Rotithor, D. Trombetta, R. Kamalapurkar, and A. P. Dani, "Full and reduced order observers for image-based depth estimation using concurrent learning," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 6, pp. 2647–2653, 2021.

[9] J. A. Rosenfeld, R. Kamalapurkar, L. F. Gruss, and T. T. Johnson, "Dynamic mode decomposition for continuous time systems with the Liouville operator," *J. Nonlinear Sci.*, vol. 32, no. 1, pp. 1–30, 2022.

[10] R. V. Self, M. Abudia, S. M. N. Mahmud, and R. Kamalapurkar, "Model-based inverse reinforcement learning for deterministic systems," *Automatica*, vol. 140, no. 110242, pp. 1–13, 2022.

[11] T. E. Ogri, S. M. N. Mahmud, Z. I. Bell, and R. Kamalapurkar, "Output feedback adaptive optimal control of affine nonlinear systems with a linear measurement model," in *Proc. IEEE Conf. Control Technol. Appl.*, 2023, to appear.

3.1. Under-review papers partially supported by this grant

[1] M. L. Greene, M. Seyed Sakha, R. Kamalapurkar, and W. E. Dixon, *Approximate dynamic programming for practical stabilization of switched systems*, Submitted to IEEE Transactions on Automatic Control.

[2] S. M. N. Mahmud, M. Abudia, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, *Safe adaptive output-feedback optimal control of a class of linear systems*, submitted to International Journal of Robust and Nonlinaer Control.

[3] S. M. N. Mahmud, M. Abudia, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, *Safe adaptive output-feedback optimal control of second order nonlinear deterministic systems*, submitted to IEEE Transactions on Automatic Control.

[4] T. E. Ogri, Z. I. Bell, and R. Kamalapurkar, *State and parameter estimation for affine nonlinear systems*, submitted to IEEE Conference on Decision and Control.

[5] J. A. Rosenfeld, B. Russo, R. Kamalapurkar, and T. Johnson, *The occupation kernel method for nonlinear system identification*, arXiv:1909.11792, Submitted to SIAM Journal on Control and Optimization.

4.0. Safety aware model-based reinforcement learning

Safety awareness is critical in reinforcement learning when task restarts are not available and/or when the system is safety critical. Safety requirements are often expressed in terms of state and/or control constraints. In the past, model-based reinforcement learning (MBRL) approaches combined with barrier transformations have been used as an effective tool to learn the optimal control policy under state constraints for systems with fully known models. In this project, the investigators primarily focused on the development reinforcement learning techniques that utilize novel filtered concurrent learning methods to realize simultaneous learning and control in the presence of model uncertainties and using partial state feedback for safety critical systems.

A MBRL approach to address the state-constrained optimal control problem appeared in [1], where the results in [2] are extended to soften the restrictive persistence of excitation requirement. However, the methods developed in [2] and [1] require fully known models, which are often difficult to obtain. In this paper, a MBRL technique is developed for barrier transformed, safety critical systems, which utilizes a novel filtered concurrent learning method to realize simultaneous learning and control in the presence of parametric uncertainties. The inclusion of filtered concurrent learning makes the feedback controller robust to modeling errors and guarantees closed-loop stability under a *finite* (as opposed to *persistent*) excitation condition. A Lypaunov-based analysis proves the developed MBRL technique is stable and guarantees safety requirements are satisfied. Simulation results are provided to demonstrate the performance of the developed MBRL approach compared to an existing optimal control method. The idea, published in [3] is summarized below.

Consider a nonlinear dynamical system of the form

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

where $x = [x_1; \ldots; x_n] \in \mathbb{R}^n$ with $i = 1, 2, \ldots, n$ is the system state, $\theta \in \mathbb{R}^p$ are the unknown parameters, $u \in \mathbb{R}^q$ is the control input, and the functions $f : \mathbb{R}^n \to \mathbb{R}^{n\times p}$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times q}$ are known, locally Lipschitz functions with $f(x) = [f_1(x); \cdots; f_n(x)]$ and $g(x) = [g_1(x); \cdots; g_n(x)]$. The notation $[a; b]$ denotes the vector $[a\ b]^T$.

The objective is to design a controller u for the system in (16) such that starting from a given feasible initial condition x^0, the trajectories $x(\cdot)$ decay to the origin and satisfy $x_i(t) \in (a_i, A_i), \forall t \geq 0$, where $a_i < 0 < A_i$.

Using a barrier transformation of the form

$$s_i \coloneqq b(x_i, a_i, A_i), \quad x_i = b^{-1}(s_i, a_i, A_i), \quad b(z, a, A) \coloneqq \log \frac{A(a-z)}{a(A-z)}, \quad z \in \mathbb{R}, \tag{2}$$

it can be shown that the system in (16) can be transformed into

$$\dot{s} = F(s) + G(s)u = y(s)\theta + G(s)u, \tag{3}$$

where

$$y(s) \coloneqq [F_1; \ldots; F_n] \in \mathbb{R}^{n\times p}, \quad G(s) \coloneqq [G_1; \ldots; G_n] \in \mathbb{R}^{n\times q}.$$

The idea is then to formulate and solve an unconstrained optimal control problem to design an adaptive feedback controller $u(t) = \phi(t, s(t))$ for the transformed system, and use the barrier transform to get the adaptive feedback controller $u(t) = \psi(t, x(t)) = \phi(t, b(x(t)))$ for the original system.

To cope with the uncertain parameters, θ, a novel system identifier was developed. Estimates of the unknown parameters, $\hat{\theta} \in \mathbb{R}^p$ are generated using the filter

$$\dot{Y} = \begin{cases} y(s), & \|Y\| \leq \overline{Y} \\ 0, & \text{otherwise} \end{cases}, \quad Y(0) = 0, \tag{4}$$

$$\dot{Y}_f = \begin{cases} Y^T Y, & \|Y_f\| \leq \overline{Y_f} \\ 0, & \text{otherwise} \end{cases}, \quad Y_f(0) = 0, \tag{5}$$

$$\dot{G}_f = \begin{cases} G(s)u, & \|Y_f\| \leq \overline{Y_f} \\ 0, & \text{otherwise} \end{cases}, \quad G_f(0) = 0, \tag{6}$$

$$\dot{X}_f = \begin{cases} Y^T(s - s^0 - G_f), & \|Y_f\| \leq \overline{Y_f} \\ 0, & \text{otherwise} \end{cases}, \quad X_f(0) = 0, \tag{7}$$

where $s^0 = [b(x_1^0); \ldots; b(x_n^0)]$, and the update law

$$\dot{\hat{\theta}} = \beta_1 Y_f^T(t)(X_f(t) - Y_f(t)\hat{\theta}), \quad \hat{\theta}(0) = \theta^0, \tag{8}$$

where β_1 is a symmetric positive definite gain matrix and $\overline{Y_f}$ is a tunable upper bound on the filtered regressor Y_f. Provided there exists a time instance $T > 0$ such that $Y_f(T)$ is full rank, the estimates, $\hat{\theta}$ can be shown to converge to their true values, θ.

Provided the optimal control policy exists, the value function of the optimal control problem, defined as

$$V^*(s) := \min_{u(\cdot)} \int_t^\infty r(\phi(\tau, s, u(\cdot)), u(\cdot)) d\tau, \tag{9}$$

where $\phi(\tau, s, u(\cdot))$ denotes the trajectory of (17), evaluated at time τ, starting from the state s and under the controller $u(\cdot)$, is characterized by the corresponding Hamilton-Jacobi-Bellman (HJB) equation

$$0 = \min_{u \in U} \left(\nabla_s V(s) \left(F(s) + G(s)u \right) + s^T Q s + u^T R u \right), \tag{10}$$

where $\nabla_s := \frac{\partial}{\partial s}$. If the value function is continuously differentiable, then it can be shown to be the unique positive definite solution of the HJB equation in (42), and provides the optimal closed-loop policy $u^* : \mathbb{R}^n \to \mathbb{R}^q$ defined as $u^*(s) := -\frac{1}{2} R^{-1} G(s)^T (\nabla_s V^*(s))^T$. The development in REF then shows that by approximating the optimal value function and the optimal control policy using a basis σ as

$$\hat{V}\left(s, \hat{W}_c\right) := \hat{W}_c^T \sigma(s), \tag{11}$$

$$\hat{u}\left(s, \hat{W}_a\right) := -\frac{1}{2} R^{-1} G^T(s) \nabla_s \sigma^T(s) \hat{W}_a, \tag{12}$$

the HJB equation can be exploited to drive the weight estimates $\hat{W}_c$ and $\hat{W}_a$ to their ideal values while keeping the system in (17) stable, thereby estimating the optimal controller online and in real time.

5.0. System identification

In addition to the work on reinforcement learning above, the investigators have also made inroads into an alternative method for system identification that could potentially be integrated into the learning framework. The new technique is developed in the framework of reproducing kernel Hilbert spaces and allows for accurate identification of unknown parameters in the system model.

Definition 1. A RKHS, H, over a set X is a Hilbert space of real valued functions over the set X such that for all $x \in X$ the evaluation functional, $E_x : H \to \mathbb{R}$, given as $E_x g := g(x)$ is bounded.

The Riesz representation theorem guarantees, for all $x \in X$, the existence of a function $k_x \in H$ such that $\langle g, k_x \rangle_H = g(x)$, where $\langle \cdot, \cdot \rangle_H$ is the inner product for H [4, Chapter 1]. The function k_x is called the reproducing kernel function at x, and the function $K(x, y) = \langle k_y, k_x \rangle_H$ is called the kernel function corresponding to H.

Each kernel function has an associated feature mapping, $\Psi : X \to \ell^2(\mathbb{N})$, such that $K(x, y) = \langle \Psi(x), \Psi(y) \rangle_{\ell^2(\mathbb{N})}$. The feature map can be obtained by using an orthonormal basis for H, and if $K : X \times X \to \mathbb{R}$ can be represented through a feature mapping, then there is a unique RKHS for which K is its kernel function [5].

To establish a connection between RKHSs and nonlinear dynamical systems, the following operator is introduced, which is inspired by the study of occupation measures [6].

Definition 2. Let $\dot{x} = f(x)$ be a dynamical system with the dynamics, $f : \mathbb{R}^n \to \mathbb{R}^n$, locally Lipschitz continuous, and suppose that H is a RKHS over a set X, where $X \subset \mathbb{R}^n$ is compact. The *Liouville operator with symbol* f, $A_f : \mathcal{D}(A_f) \to H$, is given as

$$A_f g := \nabla_x g \cdot f,$$

where

$$\mathcal{D}(A_f) := \{ g \in H : \nabla_x g \cdot f \in H \}.$$

Associated with Liouville operators in particular are a special class of functions within the domain of the Liouville operators' adjoints, and these functions are also the main object of study of this part of the project.

Definition 3. Let $X \subset \mathbb{R}^n$ be compact, H be a RKHS of continuous functions over X, and $\gamma : [0, T] \to X$ be a continuous trajectory. The functional $g \mapsto \int_0^T g(\gamma(\tau)) d\tau$ is bounded over H, and may be respresented as $\int_0^T g(\gamma(\tau)) d\tau = \langle g, \Gamma_\gamma \rangle_H$, for some $\Gamma_\gamma \in H$ by the Riesz representation theorem. The function Γ_γ is called the occupation kernel corresponding to γ in H.

The following relationship between the adjoint of the Liouville operators and the occupation kernels is exploited below for system identification.

Proposition 4. *Let H be a RKHS of continuously differentiable functions over a compact set X, and suppose that $f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous. If $\gamma : [0, T] \to X$ is a trajectory as in Definition 3 that satisfies $\dot{\gamma} = f(\gamma)$, then $\Gamma_\gamma \in \mathcal{D}(A_f^*)$, and $A_f^* \Gamma_\gamma = K(\cdot, \gamma(T)) - K(\cdot, \gamma(0))$.*

Proposition 4 completes the integration of nonlinear dynamical systems with RKHSs. In particular, valid trajectories for the dynamical system appear as occupation kernels within the domain of the adjoint of the Liouville operator corresponding to the dynamics. This intertwining allows for the expression of finite dimensional nonlinear dynamics as linear systems in infinite dimensions.

Moreover, the relation

$$\langle A_f g, \Gamma_\gamma \rangle_H = g(\gamma(T)) - g(\gamma(0)) \text{ for all } g \in \mathcal{D}(A_f)$$

uniquely determines Γ_γ. Consequently, this relation will be used subsequently to establish constraints for parameter identification in a system identification setting.

The occupation kernels themselves can be expressed as an integral against the kernel function in a RKHS as demonstrated in Proposition 5.

Proposition 5. *Let H be a RKHS over a compact set X consisting of continuous functions and let $\gamma : [0, T] \to X$ be a continuous trajectory as in Definition 3. The occupation kernel corresponding to γ in H, Γ_γ, may be expressed as*

$$\Gamma_\gamma(x) = \int_0^T K(x, \gamma(t))dt. \tag{13}$$

Thus, to compute occupation kernels and inner products of functions against occuption kernels, one simply needs to integrate numerically along the trajectories of the system leveraging, for example, quadrature techniques for integration.

In a gray box system identification setting, the system dynamics, $f : \mathbb{R}^n \to \mathbb{R}^n$, is parameterized in terms of a collection of basis functions, $Y_i : \mathbb{R}^n \to \mathbb{R}^n$ for $i = 1, \ldots, M$, as

$$f(x) = \sum_{i=1}^{M} \theta_i Y_i(x). \tag{14}$$

The goal of the system identification problem given a collection of trajectories, $\{\gamma_j\}_{j=1}^N$, satisfying the dynamics as in Definition 3, is to determine the values of the parameters, θ_i for $i = 1, \ldots, M$, such that (14) may be used to reproduce the trajectories.

For a compact set $X \subset \mathbb{R}^n$, let $\{\gamma_j : [0, T] \to X\}_{j=1}^N$ be a collection of trajectories satisfying the dynamics $\dot{x} = f(x) = \sum_{i=1}^M \theta_i Y_i(x)$, and let Γ_{γ_j} be the corresponding occupation kernels inside a RKHS, H of continuously differentiable functions over X. Suppose that $\{c_s\}_{s=1}^\infty \subset X$ is dense. Constraints on θ_i are then established as

$$\langle A_f K(\cdot, c_s), \Gamma_{\gamma_j} \rangle_H = \sum_{i=1}^{M} \theta_i \langle A_{Y_i} K(\cdot, c_s), \Gamma_{\gamma_j} \rangle_H = K(\gamma_j(T), c_s) - K(\gamma_j(0), c_s),$$

for each $s = 1, \ldots, \infty$ and $j = 1, \ldots, N$.

After the selection of a finite and representative collection of centers, $\{c_s\}_{s=1}^S$, () may be expressed as a matrix equation. Let $\{n_i\}_{i=1}^{S \cdot N}$ be an enumeration of $\{(s, j)\}_{s=1,j=1}^{S,N}$, then the matrix equation in (15) holds.

$$\mathbf{A}\theta = \mathbf{K}(T) - \mathbf{K}(0), \text{ where} \tag{15}$$
$$\mathbf{A} = \left(\langle A_{Y_i} K(\cdot, c_{n_{j,1}}), \Gamma_{\gamma_{n_{j,2}}} \rangle_H\right)_{j=1,i=1}^{j=SN,i=M} \in \mathbb{R}^{SN\times M}, \theta = \begin{pmatrix}\theta_1 & \cdots & \theta_M\end{pmatrix}^T \in \mathbb{R}^M, \text{ and}$$
$$\mathbf{K}(t) = \begin{pmatrix} K(\gamma_{n_{1,2}}(t), c_{n_{1,1}}) \\ \vdots \\ K(\gamma_{n_{SN,2}}(t), c_{n_{SN,1}}) \end{pmatrix} \in \mathbb{R}^{SN}.$$

Since the matrix $\mathbf{A}$ must be numerically estimated, written as $\hat{\mathbf{A}}$, the parameter values obtained using this method are approximate, and will be represented as $\hat{\theta}$, obtained via

$$\hat{\theta} := (\hat{\mathbf{A}}^T \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T (\mathbf{K}(T) - \mathbf{K}(0)).$$

For further details, see the published article [7]. The investigators have also extended the ideas presented here to nonparametric system identification through dynamic mode decomposition (DMD), see [8] for further details.

Safe output feedback adaptive optimal control

The ability to learn and execute control policies safely is critical to realization of complex autonomy, especially when task restarts are not available and/or the systems are safety-critical. Safety requirements are often expressed in terms of state and/or control constraints. Methods such as barrier transformation and control barrier functions have been successfully used in conjunction with model-based reinforcement learning, for safe learning in systems under state constraints. However, existing barrier-based safe learning methods rely on full-state feedback. In this line of enquiry, we develop output-feedback safe model-based reinforcement learning techniques using novel dynamic state estimators to implement simultaneous learning and control for a class of safety-critical systems with partially observable state.

We started the enquiry by developing adaptive observers for specific classes nonlinear systems. The first class considered was systems in a Brunovsky form where the state is comprised of the output and its derivatives. While adaptive observers for this class of systems were already available in results such as [9, 10], integration of such observers with the barrier transformation framework described above presents several technical challenges related to discontinuity of the transformation at the boundary of the barrier. To address these challenges, we developed a novel barrier-aware adaptive observer as follows.

We consider the following continuous-time affine nonlinear dynamical system in Brunovsky canonical form.

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = f(x) + g(x)u, \tag{16}$$

where $x \coloneqq [x_1; x_2] \in \mathbb{R}^{2n}$ is the system state, $u \in \mathbb{R}^m$ is the control input, and $x_1 \in \mathbb{R}^n$ is the output. The drift dynamics, $f : \mathbb{R}^{2n} \to \mathbb{R}^n$, and control effectiveness, $g : \mathbb{R}^{2n} \to \mathbb{R}^{n\times m}$, are known, locally Lipschitz continuous functions. Let $\hat{x}_1$ and $\hat{x}_2$ be the estimates of x_1 and x_2

respectively and $\hat{x} \coloneqq [\hat{x}_1; \hat{x}_2]$. The notation $[v; w]$ is used to denote the vector $[v^T \quad w^T]^T$, and the notation z_{i_j} is used to denote the jth element of the vector z_i. The notation I_o denotes the identity matrix of size o.

The objective is to design an adaptive 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 satisfy $x_{i_j}(t) \in (a_{i_j}, A_{i_j})$, $\forall t \geq 0$, where $i = 1, 2$, $j = 1, \ldots, n$, and the constants $a_{i_j} < 0 < A_{i_j}$ define user-specified safety constraints.

To transform the dynamics in (16) using the BT, the time derivative of the transformed state, $s \coloneqq [s_1; s_2] \in \mathbb{R}^{2n}$, can be computed as

$$\dot{s}_1 = H(s), \quad \dot{s}_2 = F(s) + G(s)u, \tag{17}$$

where $(H(s))_j \coloneqq B_{1_j}(s_{1_j})b^{-1}(s_{2_j})$, $(F(s))_j \coloneqq B_{2_j}(s_{2_j})\left(f(b^{-1}(s))\right)_j$, and $(G(s))_j \coloneqq B_{2_j}(s_{2_j})\left(g(b^{-1}(s))\right)_j$.

The estimator is given by

$$\dot{\hat{x}}_1 = \hat{x}_2, \quad \dot{\hat{x}}_2 = f(\hat{x}) + g(\hat{x})u + \nu_1, \tag{18}$$

where, $\nu_1 = [\nu_{1_1}; \ldots; \nu_{1_n}] \in \mathbb{R}^n$ is a feedback term designed in the following. The design of ν_1 is motivated by the need to establish bounds on the state estimation errors in the transformed coordinates. To facilitate the design of ν_1, let the state estimation errors be defined as $\tilde{x}_1 = x_1 - \hat{x}_1$, and $\tilde{x}_2 = x_2 - \hat{x}_2$. The feedback component ν_{1_j} where $j \in \{1, \ldots, n\}$ is designed as

$$\nu_{1_j} = \frac{\alpha^2(b(x_{1_j}) - b(\hat{x}_{1_j})) - (k + \alpha + \beta_1)\varsigma_j}{B_{1_j}\left(b\left(\hat{x}_{1_j}\right)\right)}, \tag{19}$$

where the signal ς_j is added to compensate for the fact that x_{2_j} is not measurable, and $B_{i_j}(s_{i_j}) \coloneqq \frac{a_{i_j}^2 e^{s_{i_j}} - 2a_{i_j}A_{i_j} + A_{i_j}^2 e^{-s_{i_j}}}{A_{i_j}a_{i_j}^2 - a_{i_j}A_{i_j}^2}$ is the reciprocal of the derivative of the barrier function.

Based on the stability analysis, the signal ς_j is designed as the output of the dynamic filter

$$\dot{\varsigma}_j = -(k + \beta_1)\varsigma_j - (k + \alpha)\frac{\mathrm{d}}{\mathrm{d}t}\left(b(x_{1_j}) - b(\hat{x}_{1_j})\right) - k\alpha\left(b(x_{1_j}) - b(\hat{x}_{1_j})\right), \quad \varsigma_j(0) = 0, \tag{20}$$

where α, k, and β_1 are positive constants. The signal ς_j can be implemented, without numerically differentiating $b(x_{1_j})$, via the filter,

$$\begin{aligned} \dot{\overline{\varsigma_j}} &= -(k + \beta_1)\varsigma_j - k\alpha\left(b\left(x_{1_j}\right) - b\left(\hat{x}_{1_j}\right)\right), \quad \overline{\varsigma_j}(0) = 0, \\ \varsigma_j(t) &= \overline{\varsigma_j}(t) - (k + \alpha)\Big(b\left(x_{1_j}(t)\right) - b\left(\hat{x}_{1_j}(t)\right) - b\left(x_{1_j}(0)\right) + b\left(\hat{x}_{1_j}(0)\right)\Big), \end{aligned} \tag{21}$$

where, $\overline{\varsigma_j}$ is an auxiliary signal.

To facilitate the analysis, which is done in transformed coordinates, an equivalent expression of the state estimator in the transformed coordinates is needed. To transform the state

estimator using the BT, let $\hat{s}_{i_j} \coloneqq b(\hat{x}_{i_j})$, and $\tilde{s}_{i_j} \coloneqq s_{i_j} - \hat{s}_{i_j}$. The state estimator can then be expressed in transformed coordinates $\hat{s} \coloneqq [\hat{s}_1; \hat{s}_2] \in \mathbb{R}^{2n}$, as

$$\dot{\hat{s}}_1 = H(\hat{s}), \quad \dot{\hat{s}}_2 = F(\hat{s}) + G(\hat{s})u + \nu_2, \tag{22}$$

where $\nu_2 = [\nu_{2_1}; \ldots; \nu_{2_n}] \in \mathbb{R}^n$ is given by $\nu_{2_j} = \frac{B_{2_j}(\hat{s}_{2_j})(\tilde{s}_{1_j} - (k+\alpha+\beta_1)\eta_j)}{B_{1_j}(\hat{s}_{1_j})}$, where $\eta = [\eta_1; \ldots; \eta_n]$ is the output of the dynamic filter

$$\dot{\eta} = -\beta_1 \eta - kr - \alpha \dot{\tilde{s}}_1, \quad \eta(0) = 0, \tag{23}$$

and the error signal $r = [r_1; \ldots; r_n]$ is given by

$$r = \dot{\tilde{s}}_1 + \alpha \tilde{s}_1 + \eta. \tag{24}$$

Using (23), the time derivative of r is given by

$$\dot{r} = \tilde{F}_2(s, \hat{s}) + \tilde{F}_3(s, \hat{s}) + \tilde{G}_1(s, \hat{s})u - \alpha^2 \tilde{s}_1 - kr + (k+\alpha)\eta, \tag{25}$$

where $\tilde{F}_2(s, \hat{s}) \coloneqq F_2(s) - F_2(\hat{s})$, $\tilde{F}_3(s, \hat{s}) \coloneqq F_3(s) - F_3(\hat{s})$, $\tilde{G}_1(s, \hat{s}) \coloneqq G_1(s) - G_1(\hat{s})$. The design of the estimator is motivated by the need to get the following bound.

Lemma 6. *Let $V_{se} : \mathbb{R}^{3n} \to \mathbb{R}_{\geq 0}$ be a continuously differentiable candidate Lyapunov function defined as $V_{se}(Z_1) \coloneqq \frac{\alpha^2}{2}\tilde{s}_1^T \tilde{s}_1 + \frac{1}{2} r^T r + \frac{1}{2}\eta^T \eta$, where $Z_1 \coloneqq [\tilde{s}_1^T, r^T, \eta^T]$. Provided s, $\hat{s} \in \overline{B}(0, \chi)$, where $\overline{B}(0, \chi)$ is the closed ball of radius $\chi > 0$ centered at the origin, the orbital derivative of V_{se}, along the trajectories of (17), (22), (23), and (25), under the approximate optimal controller, defined as $\dot{V}_{se}(Z_1, s, \tilde{s}, \tilde{W}_a) \coloneqq \frac{\partial V_{se}(Z_1)}{\partial \tilde{s}_1}(H(s) - H(\hat{s})) + \frac{\partial V_{se}(Z_1)}{\partial r}\dot{r} + \frac{\partial V_{se}(Z_1)}{\partial \eta}\dot{\eta}$, can be bounded as*

$$\dot{V}_{se}(Z_1, s, \tilde{s}, \tilde{W}_a) \leq -\alpha^3 \|\tilde{s}_1\|^2 - (k - \varpi_1\varpi_4)\|r\|^2 - (\beta_1 - \alpha)\|\eta\|^2 + \varpi_1 (1 + \varpi_4 + \varpi_4 \alpha\|) \|r\| \|\tilde{s}_1\| + \varpi_1 \varpi_4 \|r\| \|\eta\| + \varpi_2 \|r\| \|\tilde{W}_a\| + \varpi_3 \|r\|$$

The general idea behind the developed technique is the fact that (see [11, Lemma 1]) if a feedback controller that practically stabilizes the transformed system in (17) is designed, then the same feedback controller, applied to the original system by inverting the BT, also achieves the control objective as stated above.

In the following, a controller that practically stabilizes (17) is designed as an estimate of a controller that minimizes the infinite horizon cost[1]

$$J(u(\cdot)) \coloneqq \int_0^\infty c(\phi(\tau, s^0, u(\cdot)), u(\tau)) d\tau, \tag{26}$$

over the set $\mathcal{U}$ of piecewise continuous functions $t \mapsto u(t)$, subject to (17), where $\phi(\tau, s^0, u(\cdot))$ denotes the trajectory of (17), evaluated at time τ, starting from the state s^0, and under the controller $u(\cdot)$. In (26), $c(s, u) \coloneqq Q'(s) + u^T R u$, with $Q'(s) : \mathbb{R}^{2n} \mapsto \mathbb{R}$, and $R \in \mathbb{R}^{m \times m}$ is a symmetric PD matrix. For the optimal value function to be a Lyapunov function for the optimal policy, the following assumption is needed [13].

[1] 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 positive definite (PD), then so is Q. Furthermore, for applications with bounded control inputs, a non-quadratic penalty function similar to Eq. 17 of [12] can be incorporated in (26).

Assumption 7. One of the following is true:

1. Q' is PD.
2. Q' is positive semidefinite (PSD), and $s_1 \mapsto Q'(s)$ is PD for all nonzero $s_2 \in \mathbb{R}^n$.
3. Q' is PSD, $s_2 \mapsto Q'(s)$ is PD for all nonzero $s_1 \in \mathbb{R}^n$ and $F(s) \neq 0$ whenever $s_1 \neq 0$.

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

$$V^*(s) := \min_{u(\cdot) \in \mathcal{U}_{[t,\infty)}} \int_t^\infty c(\phi(\tau, s, u_{[0,\tau)}(\cdot)), u(\cdot)) d\tau, \tag{27}$$

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 [14, Theorem 1.5]

$$\min_{u \in \mathbb{R}^q} \Big(V_{s_1}(H(s)) \quad + \quad V_{s_2}(F(s) + G(s)u) \quad + \quad Q'(s) \quad + \quad u^T R u \Big) \qquad = \qquad 0, \tag{28}$$

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

$$u^*(s) := -\frac{1}{2} R^{-1} G(s)^T (\nabla_{s_2} V^*(s))^T. \tag{29}$$

The following theorem establishes global asymptotic stability of the closed-loop system under optimal state feedback (see [11]).

Theorem 8. *If the optimal state feedback controller* (43) *that minimizes the cost function in* (26) *exists and if the corresponding optimal value function is continuously differentiable and radially unbounded, then the origin of closed-loop system* $\dot{s}_1 = H(s)$ *and* $\dot{s}_2 = F(s) + G(s)u^*(s)$ *is globally asymptotically stable.*

The design of the controller is similar to the safe controller described above except that it uses state estimates instead of state measurements. The analysis utilizes the bound in Lemma 6 along with a Lyapunov function that guarantees stability of the optimal state feedback controller. Such a Lyapunov function is guaranteed to exist by the converse Lyapunov theorem for asymptotic stability. See [11] for details.

A similar modified barrier-aware observer was developed for linear systems, with a more general linear measurement model, based on the Luenberger observer, and used in a similar fashion to develop safe output feedback adaptive optimal controllers in [15].

To enable output feedback adaptive optimal control in nonlinear systems with linear measurement models, we integrated the bounded Jacobian observer technique with a critic-only saturated adaptive optimal controller for nonlinear dynamical systems of the form,

$$\dot{x} = f(x) + g(x)u, \quad y = Cx, \tag{30}$$

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. In the bounded Jacobian technique, we assume the following.

Assumption 9. 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

$$(M_{f_1})_{i,j} \leq \frac{\mathrm{d}(f(x))_i}{\mathrm{d}(x)_j} \leq (M_{f_2})_{i,j}, \tag{31}$$

$$(M_{g_1})_{i,j,k} \leq \frac{\mathrm{d}(g(x))_{i,k}}{\mathrm{d}(x)_j} \leq (M_{g_2})_{i,j,k}, \tag{32}$$

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

The bounded Jacobian assumption is used to express the system model in the form

$$\dot{x} = M_{f_1} x + M_{g_1} u x + \bar{f}(x) + \bar{g}_u(x, u), \tag{33}$$

where

$$\bar{f}(x) = -M_{f_1} x + f(x), \text{ and} \tag{34}$$

$$\bar{g}_u(x, u) = -M_{g_1} u x + \sum_{i=1}^{m} g_i(x)(u)_i, \tag{35}$$

and the derivatives of $\bar{f}$ and $\bar{g}$ satisfy the element-wise inequalities

$$0 \leq \frac{\mathrm{d}(\bar{f}(x))_i}{\mathrm{d}(x)_j} \leq (M_{f_2})_{i,j} - (M_{f_1})_{i,j} \text{ and} \tag{36}$$

$$0 \leq \frac{\mathrm{d}(\bar{g}_u(x, u))_{i,k}}{\mathrm{d}(x)_j} \leq \left[(M_{g_2})_{i,j,k} - (M_{g_1})_{i,j,k}\right](u)_k, \tag{37}$$

where $i, j \coloneqq 1, \ldots, n$, and $k \coloneqq 1, \ldots, m$. Thus, $\bar{M}_{f_1} = 0_{n \times n}$, $\bar{M}_{f_2} = M_{f_2} - M_{f_1}$, $\bar{M}_{g_1} = 0_{n \times n \times m}$ and $\bar{M}_{g_2} = M_{g_2} - M_{g_1}$.

Using the derivative bounds, a state estimator with three correction terms is designed as

$$\dot{\hat{x}} = M_{f_1} \hat{x} + M_{g_1} u \hat{x} + \bar{f}[\hat{x} + H(y - C\hat{x})] + \bar{g}_u[\hat{x} + K(y - C\hat{x}), u] + L(y - C\hat{x}), \tag{38}$$

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. Standard Lyapunov methods are then used to conclude that if the system state and the control input remain bounded, then the estimator results in asymptotic convergence of the state estimates to the true state (see [16].

To ensure boundedness of the control input, the controller is designed by minimizing the non-quadratic cost functional

$$J(x, u(\cdot)) \coloneqq \int_0^\infty Q(\phi(\tau, x, u_{[t,\infty)}(\cdot))) + U(u(\tau))d\tau, \tag{39}$$

over the set $\mathcal{U}$ piecewise continuous functions $t \to u(t)$, $\forall t \in [0, \infty)$ where $\phi(t, x, u(\cdot))$ is a solution of (30) under control signal $u(\cdot)$ starting from x, $Q : \mathbb{R}^n \to \mathbb{R}$ 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

$$U(u) \coloneqq 2 \int_0^u (\bar{\lambda} \tanh^{-1}(\upsilon/\bar{\lambda}))^{\mathrm{T}} R d\upsilon, \tag{40}$$

where $R \coloneqq \mathrm{diag}(r_1, \ldots, r_m)$. Assuming the optimal controller exists, then let the optimal value function, $V^* : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, be expressed as

$$V^*(x) := \min_{u(\cdot) \in \mathcal{U}_{[t,\infty)}} \int_t^\infty Q(\phi(\tau, x, u_{[t,\tau)}(\cdot))) + U(u(\tau))d\tau, \tag{41}$$

where u_I and $\mathcal{U}_I$ are obtained by restricting the domains of u and functions in $\mathcal{U}_I$ to the interval $I \subseteq R$, 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, [14, Theorem 1.5],

$$\min_{u \in \mathbb{R}^m} \Big(\nabla_x V \left(f(x) + g(x)u \right) + Q(x) + U(u) \Big) = 0, \tag{42}$$

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

$$u^*(x) := -\bar{\lambda} \tanh(D^*), \tag{43}$$

where $D^* = (1/2\bar{\lambda}) R^{-1} g(x)^{\mathrm{T}} \nabla_x V^*(x) \in \mathbb{R}^m$. Substituting equation (43) in (40), the function U is given as

$$U(u^*) \coloneqq \bar{\lambda} \nabla_x V^{*\mathrm{T}}(x) g(x) \tanh(D^*) + \bar{\lambda}^2 \bar{R} \ln(\mathbf{1} - \tanh^2(D^*)), \tag{44}$$

where $\bar{R} \coloneqq [r_1, \ldots, r_m] \in \mathbb{R}^{1 \times m}$ and $\mathbf{1}$ denotes a column vector having all of its elements equal to one. Substituting optimal control input, (43) in (42), the following equation is obtained,

$$\nabla_x V^* \left(f(x) + g(x)u^*(x) \right) + Q(x) + U(u^*) = 0 \tag{45}$$

The design of the approximate optimal controller is then similar to the designs detailed above, and uses Bellman error extrapolation and minimization. The controller saturation obviates the need for the actor network, and as a result, stability guarantees can be obtained using a single critic network that approximates the optimal value function. For further details, see [16].

We are currently in the process of extending the above result to systems with uncertain dynamic models. To that end, we have developed an adaptive observer using the bounded Jacobian technique in a paper under review for possible presentation at the 2023 IEEE Conference on Decision and Control (see [17]).

6.0. Reasons why established goals were not met

One of the goals in year 1 was to develop fallback controllers and switching based methods that utilize the fallback controllers to achieve safe reinforcement learning. The authors surmised that the barrier transformation approach presented above was a more effective way to achieve safe reinforcement learning. Furthermore, the fallback controller approach has been extensively investigated by other researchers using interval reachability and Hamilton-Jacobi reachability. Similarly, results are now available in the literature on model-based reinforcement learning using Gaussian processes.

On the other hand, the output feedback adaptive optimal control problem has not seen much attention in the literature, and has potential applications in multi-agent systems. As a result, the fallback controller approach was not investigated in this effort a nd w e f ocused i nstead on the output feedback problem. However, since the barrier transformation approach requires a well-parameterized system, the PI believes that there is still value in a fallback-based approach with online monitoring for model mismatch. The PI will continue to investigate this approach in upcoming years.

7.0. References Cited

[1] M. L. Greene, P. Deptula, S. Nivison, and W. E. Dixon, "Sparse learning-based approximate dynamic programming with barrier constraints," *IEEE Control Syst. Lett.*, vol. 4, no. 3, pp. 743–748, 2020.

[2] Y. Yang, K. G. Vamvoudakis, H. Modares, W. He, Y.-X. Yin, and D. Wunsch, "Safety-aware reinforcement learning framework with an actor-critic-barrier structure," in *Proc. Am. Control Conf.*, 2019, pp. 2352–2358.

[3] 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, 2021.

[4] V. I. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*. Cambridge University Press, 2016, vol. 152.

[5] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[6] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat, "Nonlinear optimal control via occupation measures and LMI-relaxations," *SIAM J. Control Optim.*, vol. 47, no. 4, pp. 1643–1666, 2008.

[7] J. A. Rosenfeld, R. Kamalapurkar, B. Russo, and T. T. Johnson, "Occupation kernels and densely defined Liouville operators for system identification," in *Proc. IEEE Conf. Decis. Control*, 2019, pp. 6455–6460.

[8] J. A. Rosenfeld, R. Kamalapurkar, L. F. Gruss, and T. T. Johnson, "Dynamic mode decomposition for continuous time systems with the Liouville operator," *J. Nonlinear Sci.*, vol. 32, no. 1, pp. 1–30, 2022.

[9] R. Kamalapurkar, "Online output-feedback parameter and state estimation for second order linear systems," in *Proc. Am. Control Conf.*, 2017, pp. 5672–5677.

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

[11] S. M. N. Mahmud, M. Abudia, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, *Safe adaptive output-feedback optimal control of second order nonlinear deterministic systems*, submitted to IEEE Transactions on Automatic Control.

[12] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for H∞ control with full-state constraints and input saturation," *J. Franklin Inst.*, vol. 357, no. 6, pp. 3316–3344, 2020.

[13] 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.*, 2019, pp. 2378–2383.

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

[15] S. M. N. Mahmud, M. Abudia, S. A. Nivison, Z. I. Bell, and R. Kamalapurkar, *Safe adaptive output-feedback optimal control of a class of linear systems*, submitted to International Journal of Robust and Nonlinaer Control.

[16] T. E. Ogri, S. M. N. Mahmud, Z. I. Bell, and R. Kamalapurkar, "Output feedback adaptive optimal control of affine nonlinear systems with a linear measurement model," in *Proc. IEEE Conf. Control Technol. Appl.*, 2023, to appear.

[17] T. E. Ogri, Z. I. Bell, and R. Kamalapurkar, *State and parameter estimation for affine nonlinear systems*, submitted to IEEE Conference on Decision and Control.