Automatica 116 (2020) 108922

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with control effectiveness faults☆,☆☆

Patryk Deptula [a,*], Zachary I. Bell [b], Emily A. Doucette [c], J. Willard Curtis [c], Warren E. Dixon [b]

[a] The Charles Stark Draper Laboratory, Inc., Cambridge, MA, USA
[b] Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, USA
[c] Munitions Directorate, Air Force Research Laboratory, Eglin AFB, FL, USA

ARTICLE INFO

Article history:
Received 19 November 2018
Received in revised form 2 November 2019
Accepted 22 February 2020
Available online xxxx

ABSTRACT

An infinite horizon approximate optimal control problem is developed for a system with unknown drift parameters and control effectiveness faults. A data-based filtered parameter estimator with a novel dynamic gain structure is developed to simultaneously estimate the unknown drift dynamics and control effectiveness fault. A local state-following approximate dynamic programming method is used to approximate the unknown optimal value function for an uncertain system. Using a relaxed persistence of excitation condition, a Lyapunov-based stability analysis shows exponential convergence to a residual error for the parameter estimation and uniformly ultimately bounded convergence for the closed-loop system. Simulation results are presented which demonstrate the effectiveness of the developed method.

1. Introduction

Designing optimal controllers for uncertain nonlinear systems is difficult because the solution to the Hamilton–Jacobi–Bellman (HJB) is generally unknown. Approximate dynamic programming (ADP) provides a viable way of approximating the solution to the HJB via neural network (NN) representations (cf., Jiang & Jiang, 2017; Kamalapurkar, Walters, Rosenfeld, & Dixon, 2018; Lewis & Liu, 2013); however, ADP has two inherent significant challenges.

One challenge for dynamic programming methods is the curse of dimensionality because a large number (e.g., exponential growth) of basis functions is generally required to obtain a sufficient approximation. In this paper, we provide an in-road to address the computational complexity issue by leveraging the recent state-following (StaF) method in Kamalapurkar, Rosenfeld and Dixon (2016) to develop local approximations of the value function with a reduced number (i.e., linear growth) of basis functions.

Another challenge for ADP methods is that, unlike traditional adaptive controllers, the ideal weights of the NN must be exactly learned (i.e., system identification) to approximate the optimal controller. One way to achieve parameter identification is to assume a persistence of excitation (PE) condition is satisfied; yet, for general nonlinear systems, there is no currently known way to ensure the PE condition is satisfied a priori, even with an added disturbance signal or so-called probing noise, and no way to verify if the condition is satisfied online. Motivated by this issue, data-driven techniques such as simulation of experience and experience replay aim to relax the PE assumption by utilizing concurrent learning (CL), where data richness is characterized by the eigenvalues of a history stack, which unlike the PE condition can be verified online (cf., Bhasin et al., 2013; Chowdhary & Johnson, 2011; Fan & Yang, 2016b, 2016c; Kamalapurkar, Rosenfeld et al., 2016; Kamalapurkar, Walters and Dixon, 2016; Modares, Lewis, & Naghibi-Sistani, 2014; Vamvoudakis, Miranda, & Hespanha, 2016; Zhang, Cui, Zhang, & Luo, 2011). Specifically, CL collects pairs of input/output data and stores them in an evolving history stack during task execution. The input/output pairs can then be used, along with methods such as Chowdhary and Johnson (2011) and Kamalapurkar, Reish, Chowdhary, and Dixon (2017) to manage the size and composition of the history stack, to perform system identification (assuming the derivative of the highest order states is available or numerically generated). Integral CL (ICL) removes the need to measure the derivative of the highest order terms by including an integral of the terms in

☆ The material in this paper was partially presented at the 2018 American Control Conference, June 27–29, 2018, Milwaukee, WI, USA. This paper was recommended for publication in revised form by Associate Editor Kyriakos G. Vamvoudakis under the direction of Editor Miroslav Krstic.

☆☆ This work was done prior to Patryk Deptula joining Draper.

* Corresponding author.
E-mail addresses: pdeptula@draper.com (P. Deptula), bellz121@ufl.edu (Z.I. Bell), emily.doucette@us.af.mil (E.A. Doucette), jess.curtis@us.af.mil (J.W. Curtis), wdixon@ufl.edu (W.E. Dixon).

https://doi.org/10.1016/j.automatica.2020.108922
0005-1098/

the history stack. Specifically, in Parikh, Kamalapurkar, and Dixon (2019), ICL is formulated so that the dynamics are integrated over a finite window; hence, the formulation includes both a finite difference and an integrated function. However, such an approach requires numerical techniques to evaluate the integrals, which can cause errors to accumulate if large integration buffers are used or in the presence of measurement noise. Another approach which uses an initial excitation (IE) condition to guarantee parameter identification can be determined via an integral-like update law (Roy, Bhasin, & Kar, 2018).

In addition to the nominal challenges faced with designing and implementing ADP methods, the development of ever more complex systems along with the emerging potential for cyber effects, results in additional challenges. For example, adaptation for fault tolerant control (FTC) has received an increased amount of attention in recent years (cf., Fan & Yang, 2016a, 2016b, 2016c; Jiang, Zhang, Liu, & Han, 2017; Liu, Wang, & Zhang, 2017; Lv, Na, Yang, Wu, & Guo, 2016; Shen, Liu, & Dowell, 2013; Zhao, Liu, & Li, 2017 and references therein). An input observer based approach is considered in Shen et al. (2013) to estimate loss of effectiveness (LoE) faults for linear systems subject to exogenous disturbances while numerical solutions for Linear Matrix Inequalities are used generate control policies. A continuous-time constrained complementary controller using ADP for a partial LoE fault with known nonlinear drift dynamics is developed in Fan and Yang (2016a). In Liu et al. (2017), a high-gain controller is utilized to compensate for an additive fault, while a discrete-time ADP controller provides tracking guarantees. In Jiang et al. (2017) and Zhao et al. (2017), a policy-iteration ADP approach is developed to compensate for biased input faults. An integral sliding-mode approach with a NN estimator in conjunction with an actor–critic design was developed in Fan and Yang (2016b, 2016c) to compensate for time-varying input faults for a class of systems with partially unknown dynamics. The result in Lv et al. (2016) uses a filter-based adaptive NN identifier to estimate the unknown drift dynamics and control effectiveness for a fault-free system. However, the results in Fan and Yang (2016b, 2016c) and Lv et al. (2016) inject a probing signal that is assumed to satisfy the PE condition to ensure the parameter estimates converge to their true value.

Building on our precursory efforts in Deptula, Bell, Doucette, Curtis, and Dixon (2018), a unique filtered ICL estimator is developed in this paper to simultaneously estimate the drift dynamics and unknown state-dependent control effectiveness fault for a class of nonlinear systems. Specifically, we build on our previous ICL development in Parikh et al. (2019) to replace the PE condition with an eigenvalue condition that can be checked each time new data is recorded by computing the minimum eigenvalue of the square of the integral of data-based regression matrix. To account for the potential for numerical integration error accumulation when using large integration windows, the history stack is passed through a low pass filter and a novel dynamic gain structure is developed to increase the eigenvalues and generate less noisy regressors for parameter estimation.

In addition to identifying the unknown drift dynamics and fault, the unknown value function must also be identified. To this end, we build on our previous simulation of experience based approach via Bellman error (BE) extrapolation (cf., Kamalapurkar, Rosenfeld et al., 2016; Kamalapurkar, Walters et al., 2016), where the controller and BE are evaluated at extrapolated trajectories in a neighborhood of the system state. Though the approach seems similar to Fan and Yang (2016a), the developed approach does not assume knowledge of the drift dynamics and considers state-varying faults. Moreover, the result in Fan and Yang (2016a) assumes a robust controller already exists which stabilizes the system and a complementary controller is designed to improve performance. In addition, this result is not limited to constant control effectiveness perturbations as considered in Fan and Yang (2016a). In this result, the overall controller is assumed to be saturated which relates more to real-life systems where torque limits or vehicle velocity and acceleration limits are present. StaF kernel functions are used to estimate the value function and local BE extrapolation is used to relax the PE condition when learning the critic weights. Unlike the preliminary results in Deptula et al. (2018) or the results in Fan and Yang (2016a), this paper considers a wider class of reduced effectiveness faults, by considering state-varying faults. Furthermore, compared to the preliminary result in Deptula et al. (2018), a unique value function representation is included and a more detailed stability analysis is included. Simulation results along with discussions are presented to illustrate the performance of the developed method.

2. Problem formulation

Consider the control affine system with a state-dependent control effectiveness fault

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

with $t \in \mathbb{R}_{\geq t_0}$, where $t_0 \in \mathbb{R}_{\geq 0}$ denotes the initial time, $x : \mathbb{R}_{\geq t_0} \to \mathbb{R}^n$ denotes the system state, $f : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift dynamics, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ denotes the known bounded control effectiveness matrix, $u : \mathbb{R}_{\geq t_0} \to \mathbb{R}^m$ denotes the amplitude limited control input such that $\sup_t |u_i| \leq \alpha \ \forall i = 1, \ldots, m$, where $\alpha \in \mathbb{R}_{>0}$ is the saturation constant, $\mu_a : \mathbb{R}^n \to \mathbb{R}^m$ denotes the unknown state varying actuator perturbation, and $\Lambda : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ is a diagonal operator defined as $\Lambda(\cdot) \triangleq \text{diag}\{\cdot\}$.[1]

Assumption 1. Each element of the actuator control effectiveness perturbation is bounded such that $0 < \mu_{ai}(x) \leq \overline{\mu}_{ai}$, $i = 1, \ldots, m$, where $\overline{\mu}_{ai} \in \mathbb{R}_{>0}$ are constants. Hence, the unknown constant actuator fault is bounded such that $0 < \|\Lambda(\mu_a(x))\| \leq \overline{\mu}_a$ where $\overline{\mu}_a \triangleq \sup\{\overline{\mu}_{a1}, \ldots, \overline{\mu}_{am},\}$ and $\overline{\mu}_{ai} \leq 1$, $i = 1, \ldots, m$. Furthermore, the state varying perturbation is C^1.

Assumption 2. The drift dynamic $f(\cdot)$ is differentiable in its arguments, locally Lipschitz, and $f(0) = 0$. Furthermore, the control effectiveness matrix $g(x)$ is known and bounded, such that $0 < \|g(x)\| \leq \overline{g}$, where $\overline{g} \in \mathbb{R}_{>0}$ (Bhasin et al., 2013; Kamalapurkar et al., 2018; Lewis & Liu, 2013).

To form the optimal control problem, let the cost functional be defined as

$$J(x, u) \triangleq \int_{t_0}^{\infty} r(x(\tau), u(\tau))\, d\tau, \tag{2}$$

where $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ denotes the instantaneous cost $r(x, u) \triangleq Q_x(x) + \Psi(u)$, with $Q_x : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ being continuously differentiable function satisfying $\underline{q}\|x\|^2 \leq Q_x(x) \leq \overline{q}\|x\|^2$ for $\underline{q}, \overline{q} \in \mathbb{R}_{>0}$ and

$$\Psi(u) \triangleq 2\sum_{i=1}^{m} \int_0^{u_i} \alpha r_i \tanh^{-1}\left(\frac{\xi_{u_i}}{\alpha}\right) d\xi_{u_i}. \tag{3}$$

In (3), ξ_{u_i} is an integration variable, and r_i are the diagonal elements which make up the symmetric positive definite weighting matrix $R \in \mathbb{R}^{m\times m}$ where $R \triangleq \text{diag}\{\bar{r}\}$, and $\bar{r} \triangleq [r_1, \ldots, r_m] \in \mathbb{R}^{1\times m}$. The infinite-time optimal scalar value function $V^* : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is $V^*(x(t)) = \inf_{u(\tau)\in U | \tau\in\mathbb{R}_{\geq t}} \int_t^{\infty} r(x(\tau), u(\tau))\, d\tau$, where U denotes the set of admissible controllers.

[1] Although this paper uses a single saturation constant α, different saturation constants can be used for each input such that $\sup_t |u_i| \leq \alpha_i \ \forall i = 1, \ldots, m$, where $\alpha_i \in \mathbb{R}_{>0}$ are the saturation constants.

Assumption 3. A controller $u : \mathbb{R}_{\geq t_0} \to U$ is said to be admissible for $(t_0, x_0) \in \mathbb{R}_{\geq t_0} \times \mathbb{R}^n$ if it is bounded, generates a unique bounded trajectory starting from (t_0, x_0), and results in bounded total cost.

The following assumption facilitates a dynamic programming based solution of the optimal control problem.

Assumption 4. The value function $x \mapsto V^*(x)$ is continuously differentiable.

The objective is to determine the optimal control policy u^* which minimizes the cost functional in (2) subject to the dynamic constraints in (1). To facilitate the subsequent development, let $F(x) \triangleq f(x)$ and $G(x) \triangleq g(x) \Lambda(\mu_a(x))$. Then, the HJB is

$$0 = \nabla V^* \left(F(x) + G(x) u^*\right) + r\left(x, u^*\right), \tag{4}$$

with the boundary condition $V(0) = 0$ (Kirk, 2004, Section 3.11). Using Assumptions 2–4 along with (3)–(4), the optimal control policy is Fan and Yang (2016a) and Modares et al. (2014)

$$u^*(x) = -\alpha \text{Tanh}\Big(\frac{R^{-1}G^T(x)}{2\alpha}\left(\nabla V^*(x)\right)^T\Big). \tag{5}$$

After substituting (5) into (4), the HJB becomes

$$0 = \nabla V^*(x) F(x) + Q_x(x) + \alpha^2 \bar{r} \ln\Big(\underline{1} - \text{Tanh}^2\Big(\frac{R^{-1}G^T(x)}{2\alpha}\left(\nabla V^*(x)\right)^T\Big)\Big), \tag{6}$$

where $\underline{1} \triangleq [1, 1, \ldots, 1]^T \in \mathbb{R}^m$, and $\text{Tanh}(\xi) \triangleq [\tanh(\xi_i), \ldots, \tanh(\xi_m)]^T$ for $\xi \in \mathbb{R}^m$.[2]

The positive definite solution to (6) is the value function V^*, which is an unknown nonlinear function. Furthermore, (6) requires knowledge of the drift dynamics f and the control effectiveness fault μ_a. However, because V^*, f, and μ_a are unknown, approximations are sought.

3. Approximate optimal control

In the following, an adaptive control strategy is presented to approximate the drift dynamics and control effectiveness fault simultaneously. Using a computationally efficient local ADP approach that exploits BE extrapolation is used to estimate V^*.

3.1. System and fault estimation

Since the drift dynamics f and control effectiveness fault μ_a are unknown, using the universal function approximation property, NNs can be used to represent the system in (1) over a compact set $\chi \subset \mathbb{R}^n$ containing the origin as

$$\dot{x} = f^o(x) + \Phi(x, u)\Theta + \varepsilon_{sys}(x, u), \tag{7}$$

where $f^o : \mathbb{R}^n \to \mathbb{R}^n$ denotes the known nominal drift dynamics, $\Theta \triangleq \left[\text{vec}\left(W_1^T\right)^T, \ \text{vec}\left(W_2^T\right)^T\right]^T \in \mathbb{R}^p$, denotes an unknown constant vector, and $\Phi(x, u) \triangleq \left[\left(\phi_1^T(x) \otimes I_n\right) \ , \left(\phi_2^T(x) \otimes (g(x)\Lambda(u))\right)\right] \in \mathbb{R}^{n \times p}$, denotes the known regression matrix, where $p \triangleq p_1 n + p_2 m$, $W_1 \in \mathbb{R}^{p_1 \times n}$ and $W_2 \in \mathbb{R}^{p_2 \times m}$, are unknown constant weight matrices, $\phi_1 : \mathbb{R}^n \to \mathbb{R}^{p_1}$ and $\phi_2 : \mathbb{R}^n \to \mathbb{R}^{p_2}$ are user defined basis functions, and $\varepsilon_{sys} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the total NN function reconstruction error defined as $\varepsilon_{sys}(x, u) \triangleq \varepsilon_1(x) + g(x)\Lambda(u)\varepsilon_2(x)$. The unknown weight, function reconstruction error, and known vector of basis functions are assumed to satisfy the following.

Assumption 5. There exist $\overline{\phi}_1, \overline{\phi}_2, \overline{\varepsilon}_s \in \mathbb{R}_{>0}$ such that $\sup_{x \in \chi} \|\phi_1\| \leq \overline{\phi}_1$, $\sup_{x \in \chi} \|\phi_2\| \leq \overline{\phi}_2$, and $\sup_{x \in \chi} \|\varepsilon_{sys}\| \leq \overline{\varepsilon}_s$, respectively. Furthermore, there exists $\overline{W}_1, \overline{W}_2, \overline{\Theta} \in \mathbb{R}_{>0}$ such that $\|W_1\| \leq \overline{W}_1$ and $\|W_2\| \leq \overline{W}_2$, and based on the definition of Θ, $\|\Theta\| \leq \overline{\Theta}$ follows Kamalapurkar et al. (2018) and Lewis and Liu (2013).

Motivated by the fact that our previous ICL approaches use piecewise constant sample data that are collected online, a filtered representation of (7) is incorporated here to aid in learning the unknown weight Θ. Moreover, implementable filter variables $x_f \in \mathbb{R}^n$, $\Phi_f \in \mathbb{R}^{n \times p}$, and $F_f \in \mathbb{R}^n$ are defined based on the following low-pass filter structures

$$\begin{cases} k\dot{x}_f + x_f = x, & k\dot{\Phi}_f + \Phi_f = \Phi(x, u), \\ k\dot{F}_f + F_f = f^o(x), \end{cases} \tag{8}$$

where $k \in \mathbb{R}_{>0}$. Using (8), (7) can be represented as

$$\dot{x}_f(t) = \frac{x(t) - x_f(t)}{k} = F_f(t) + \Phi_f(t)\Theta + \varepsilon_f(t) \tag{9}$$

where $\varepsilon_f \in \mathbb{R}^n$ is the filtered version of the non-measurable total NN function reconstruction error ε_{sys}, which by following the notation in (8) is generated by $k\dot{\varepsilon}_f + \varepsilon_f = \varepsilon_{sys}(x, u)$. Note that ε_f is not implemented nor measured and is only used in the analysis.

Furthermore, let $P \in \mathbb{R}^{p \times p}$ and $Q \in \mathbb{R}^p$ be dynamic gain matrices updated by the update laws

$$\begin{cases} \dot{P} = -\ell P + k_p \Phi_f^T \Phi_f + k_{CL} \sum_{i=1}^M \mathcal{Y}_i^T \mathcal{Y}_i, \\ \dot{Q} = -\ell Q + k_p \Phi_f^T \left(\frac{x - x_f}{k} - F_f\right) + \\ \quad k_{CL} \sum_{i=1}^M \mathcal{Y}_i^T \left(x(t_i) - x(t_i - T) - \mathcal{F}_i\right), \end{cases} \tag{10}$$

with $P(t_0) = 0$ and $Q(t_0) = 0$, where ℓ, $k_{CL} \in \mathbb{R}_{>0}$, and $k_p \in \mathbb{R}_{\geq 0}$ are constant update gains, $T \in \mathbb{R}_{>0}$ is a user defined integration window, $t_i \in [T, t]$ are points between the initial time and current time, and $M \in \mathbb{N}_{>0}$ is the user-defined number of data points in the history stack.[3] In (10), $\mathcal{Y}_i \triangleq \int_{t_i - T}^{t_i} \Phi(x(\tau), u(\tau))\, d\tau$, $\mathcal{F}_i \triangleq \int_{t_i - T}^{t_i} f^o(x(\tau))\, d\tau$, $x(t_i) - x(t_i - T) = \mathcal{F}_i + \mathcal{Y}_i \Theta + \mathcal{E}_i$, and $\mathcal{E}(t) \triangleq \int_{t-T}^{t} \varepsilon_{sys}(x(\tau))\, d\tau$. Based on the previous development, $\hat{\Theta}$ is designed using a least-squares approach as

$$\dot{\hat{\Theta}}(t) = \text{proj}\left\{-\Gamma(t)\left(P(t)\hat{\Theta}(t) - Q(t)\right)\right\}, \tag{11}$$

$$\dot{\Gamma}(t) = \beta\Gamma(t) - \Gamma(t) P(t) \Gamma(t), \tag{12}$$

where $\beta \in \mathbb{R}_{>0}$ denotes the forgetting factor and $\text{proj}\{\cdot\}$ denotes a smooth projection operator with respect to $\Omega \subset \mathbb{R}^p$, such that $\Theta \in \Omega$, which bounds the weight estimates.[4]

Numerical techniques are required to compute the integrals which can cause errors to accumulate if large integration buffers are used. Moreover, measurements may be noisy which can degrade data when implementing ICL techniques such as Parikh et al. (2019) on hardware. The structure of (10) is a filtering method that also provides a means to increase the minimum eigenvalue of P, and hence the convergence rate of $\hat{\Theta}(t)$.

[2] If different saturation constants α_i are used, then the optimal controller and HJB in (6) become $u^*(x) = -\alpha \text{Tanh}\left(\frac{\alpha^{-1}R^{-1}G^T(x)}{2}\left(\nabla V^*(x)\right)^T\right)$ and $0 = \nabla V^*(x) F(x) + Q_x(x) + \bar{r}\alpha^2 \ln\Big(\underline{1} - \text{Tanh}^2\left(\frac{\alpha^{-1}R^{-1}G^T(x)}{2}\left(\nabla V^*(x)\right)^T\right)\Big)$, respectively, where $\alpha \triangleq \text{diag}\{\alpha_1, \ldots, \alpha_m\} \in \mathbb{R}^{m \times m}$.

[3] As stated in Lv et al. (2016), the regression matrix $\Phi_f^T \Phi_f$ in (10) can aid in increasing $\lambda_{\min}\{P\}$ using the current state and input pairs. However, the use of only $\Phi_f^T \Phi_f$ would require the stringent PE condition to be satisfied; hence, $\sum_{i=1}^M \mathcal{Y}_i^T \mathcal{Y}_i$ is used along with the less restrictive measurable condition in Assumption 6 to show convergence of the parameter estimates in Theorem 1.

[4] See Dixon, Behal, Dawson, and Nagarkatti (2003, Chapter 4) for details of the projection operator.

Remark 1. By Assumption 5, Φ, $\varepsilon_{sys} \in \mathcal{L}_\infty$. Integrating (8), using the definitions for $\mathcal{Y}_i$ and $\mathcal{E}_i$, and then substituting the bounds for Φ and ε_{sys} from Assumption 5, then Φ_f, ε_f, $\mathcal{Y}_i$, $\mathcal{E}_i \in \mathcal{L}_\infty$. Hence, there exist $\overline{\Phi_f}$, $\overline{\varepsilon_f}$, $\overline{\mathcal{Y}}$, $\overline{\mathcal{E}} \in \mathbb{R}_{>0}$ such that $\left\|\Phi_f\right\| \leq \overline{\Phi_f}$, $\left\|\varepsilon_f\right\| \leq \overline{\varepsilon_f}$, $\|\mathcal{Y}_i\| \leq \overline{\mathcal{Y}}$, and $\|\mathcal{E}_i\| \leq \overline{\mathcal{E}}$.

Assumption 6 (*Parikh et al., 2019*)**.**
There exists $T_1 \in \mathbb{R}_{>0}$ such that $T_1 > T$, and strictly positive constants $\lambda_1, \lambda_2 \in \mathbb{R}_{>0}$ where $\lambda_1 I_p \leq \sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i \leq \lambda_2 I_p$, $\forall t \geq T_1$.

Remark 2. In Assumption 6, λ_1 only requires a finite collection of sufficiently exciting $\mathcal{Y}_i$ regressors to have $\lambda_{\min}\left\{\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i\right\} \geq \lambda_1$. Moreover, Assumption 6 says that for all time after $\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i$ becomes full rank, then $\lambda_3 \triangleq \lambda_{\min}\{P(t)\} > 0$, where $\lambda_{\min}\{\cdot\}$ and $\lambda_{\max}\{\cdot\}$ denote the minimum and maximum eigenvalues, respectively, and I_n denotes an $n \times n$ identity matrix.[5,6]

Remark 3. The sufficient condition in Fan and Yang (2016a, Lemma 1) requires the summation of the history stack to be positive for all time (i.e., $\lambda_{\min}\left\{\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i\right\} > 0$ to hold for all $t \geq t_0$). Hence the history stack needs sufficient initial data such that it is full rank from t_0, which is difficult to ensure. Assumption 6 is less restrictive because it does not restrict the history stack to be initially filled with data; and therefore $\lambda_{\min}\left\{\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i\right\} \geq 0$ for $t \in [t_0, T_1)$. However, as input–output data is collected online Assumption 6 indicates that the summation of the history stack becomes positive definite such that $\lambda_{\min}\left\{\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i\right\} \geq \lambda_1 > 0$ for all $t \geq T_1$. Unlike Fan and Yang (2016a), this switch in the minimum eigenvalue of the collected input–output data makes (11)–(12) a switched system. Theorem 1 shows stability during both phases of learning.

Provided $\lambda_{\min}\left\{\Gamma^{-1}(t_0)\right\} > 0$, using similar arguments to Ioannou and Sun (1996, Corollary 4.3.2), Γ can be shown to satisfy $\underline{\Gamma} I_p \leq \Gamma(t) \leq \overline{\Gamma} I_p$, where $\underline{\Gamma}$ and $\overline{\Gamma}$ are positive constants. To facilitate the subsequent stability analysis, a Lyapunov function candidate $V_\theta : \mathbb{R}^p \times \mathbb{R}_{\geq t_0} \to \mathbb{R}_{\geq 0}$ is defined as

$$V_\theta\left(\tilde{\Theta}, t\right) = \frac{1}{2} \tilde{\Theta}^T \Gamma^{-1}(t) \tilde{\Theta}, \tag{13}$$

such that

$$\frac{1}{2\overline{\Gamma}} \left\|\tilde{\Theta}\right\|^2 \leq V_\theta\left(\tilde{\Theta}, t\right) \leq \frac{1}{2\underline{\Gamma}} \left\|\tilde{\Theta}\right\|^2. \tag{14}$$

The following analysis shows that for the time interval $t \in [t_0, T_1)$ the estimation error remains bounded because the matrix P is positive semi-definite. For $t \geq T_1$, after a sufficient collection of input–output pairs $\{\mathcal{Y}_i, \Delta x_i, \mathcal{F}_i\}_{i=1}^{M}$ have been collected to satisfy Assumption 6 and P becomes positive definite, the estimation error $\tilde{\Theta}$ exponentially decays to a smaller bound.

Theorem 1. *The adaptive update laws in (11) and (12) ensure that the estimation error $\tilde{\Theta}$ remains bounded for all $t \geq T_1$ such that*

$$\left\|\tilde{\Theta}(t)\right\| \leq \sqrt{\frac{\overline{\Gamma}}{\underline{\Gamma}}} \sqrt{c_\Theta e^{-\lambda_{\theta 2}(t-T_1)} + \frac{4}{c_{L2}^2} \overline{v_1}^2}, \tag{15}$$

where $c_\Theta \triangleq \left\|\tilde{\Theta}(t_0)\right\|^2 + \frac{4}{c_{L1}^2} \overline{v_1}^2$, $c_{L1} \triangleq \frac{\beta}{\overline{\Gamma}}$, $c_{L2} \triangleq c_{L1} + \lambda_3$, $\lambda_{\theta 2} \triangleq \frac{\underline{\Gamma} c_{L2}}{2}$, *and* $\sup_t \|v_1(t)\| \leq \overline{v_1}$, *where* $\overline{v_1} \triangleq \frac{\left(k_p \overline{\Phi_f} \overline{\varepsilon_f} + k_{CL} M \overline{\mathcal{Y}} \overline{\mathcal{E}}\right)}{\ell}$.

Proof. Substituting (9) into (10), then integrating yields

$$\begin{cases} P(t) = k_p \int_{t_0}^{t} e^{-\ell(t-\tau)} \Phi_f(\tau)^T \Phi_f(\tau)\, d\tau \\ \qquad + k_{CL} \int_{t_0}^{t} e^{-\ell(t-\tau)} \sum_{i=1}^{M} \mathcal{Y}_i^T(\tau) \mathcal{Y}_i(\tau)\, d\tau, \\ Q(t) = P(t)\Theta - v_1(t), \end{cases} \tag{16}$$

where v_1 is defined as

$$v_1(t) \triangleq -\int_{t_0}^{t} e^{-\ell(t-\tau)} k_p \Phi_f(\tau)^T \varepsilon_f(\tau)\, d\tau - \int_{t_0}^{t} e^{-\ell(t-\tau)} k_{CL} \sum_{i=1}^{M} \mathcal{Y}_i^T(\tau) \mathcal{E}_i(\tau)\, d\tau. \tag{17}$$

Substituting (16) into (11) for $Q(t)$ yields

$$\dot{\tilde{\Theta}}(t) = \operatorname{proj}\left\{\Gamma(t)\left(P(t)\tilde{\Theta}(t) - v_1(t)\right)\right\}. \tag{18}$$

Using Remark 1, substituting the bounds in for Φ_f, ε_f, $\mathcal{Y}_i$, $\mathcal{E}_i$ into (17), and integrating yields $\|v_1(t)\| \leq \overline{v_1}$.

Furthermore, let $z(t)$ be a Filippov solution to the differential inclusion $\dot{z}(t) \in K[h](z(t))$ for $t \in \mathbb{R}_{\geq t_0}$, where $K[\cdot]$ is defined in Filippov (1964) and $h : \mathbb{R}^{p+p^2} \to \mathbb{R}^{p+p^2}$ is defined as $h \triangleq \left[\dot{\tilde{\Theta}}^T, \operatorname{vec}\left(\dot{\Gamma}^{-1}\right)^T\right]^T$.[7] Due to discrete collections of data points in (10), the time derivative of (13) exists almost everywhere (a.e.), i.e., for almost all $t \in \mathbb{R}_{t_0}$, and $\dot{V}_\theta(z) \overset{a.e.}{\in} \dot{\tilde{V}}_\theta(z)$, where $\dot{\tilde{V}}_\theta(z)$ is the generalized time-derivative of (13) along the Filippov trajectories of $\dot{z}(t) = K[h](z(t))$ (Paden & Sastry, 1987). Using the calculus of $K[\cdot]$, substituting (12) and (18), using the fact that $P \geq 0$ for $t \in [t_0, T_1)$ and bounding yields $\dot{V}_\theta\left(\tilde{\Theta}, t\right) \overset{a.e.}{\leq} -\frac{c_{L1}\underline{\Gamma}}{2} V_\theta\left(\tilde{\Theta}, t\right) + \frac{\overline{v_1}^2}{c_{L1}}$. Since the set of discontinuities is countable, then $\dot{V}_\theta\left(\tilde{\Theta}, t\right)$ and $V_\theta\left(\tilde{\Theta}, t\right)$ are Lebesgue integrable over $t \in [t_0, T_1)$ such that

$$V_\theta\left(\tilde{\Theta}(t), t\right) \leq V_\theta\left(\tilde{\Theta}(t_0), t_0\right) e^{-\lambda_{\theta 1}(t-t_0)} + \left(1 - e^{-\lambda_{\theta 1}(t-t_0)}\right) \frac{2}{\underline{\Gamma} c_{L1}^2} \overline{v_1}^2, \tag{19}$$

where $\lambda_{\theta 1} \triangleq \frac{c_{L1}\underline{\Gamma}}{2}$. Since $\lambda_{\min}\left\{\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i\right\} \geq 0$ for $t \in [t_0, T_1)$, using (14) in (19), $\forall t \in [t_0, T_1)$ $\left\|\tilde{\Theta}(t)\right\| \leq \sqrt{\frac{\overline{\Gamma}}{\underline{\Gamma}}} \left(\left\|\tilde{\Theta}(t_0)\right\| + 2\frac{\overline{v_1}}{c_{L1}}\right)$.

The matrix P depends on the collection of input–output data and the filtered basis Φ_f; hence, after sufficient data pairs $\{\mathcal{Y}_i, \Delta x_i, \mathcal{F}_i\}_{i=1}^{M}$ have been collected to satisfy Assumption 6, a similar argument as Ioannou and Sun (1996, Corollary 4.3.2) can be used to conclude that $P > 0$ $\forall t \geq T_1$. Using the calculus of K on the time derivative of (13), substituting (12) and (18), using Assumption 6 and bounding yields

$$\dot{V}_\theta\left(\tilde{\Theta}(t), t\right) \overset{a.e.}{\leq} -\frac{c_{L2}}{2} \left\|\tilde{\Theta}\right\|^2 + \left\|\tilde{\Theta}\right\| \overline{v}_1. \tag{20}$$

Completing the squares, using (14), and since $\dot{V}_\theta\left(\tilde{\Theta}, t\right)$ and $V_\theta\left(\tilde{\Theta}, t\right)$ are Lebesgue integrable over $t \in \mathbb{R}_{\geq T_1}$

$$V_\theta\left(\tilde{\Theta}(t), t\right) \leq V_\theta\left(\tilde{\Theta}(T_1), T_1\right) e^{-\lambda_{\theta 2}(t-T_1)} + \left(1 - e^{-\lambda_{\theta 2}(t-T_1)}\right) \frac{2}{\underline{\Gamma} c_{L2}^2} \overline{v_1}^2,$$

[5] It is difficult to ensure the standard PE condition (c.f., Fan & Yang, 2016b; Lv et al., 2016) is satisfied. However, data selection techniques (cf., Chowdhary & Johnson, 2011; Kamalapurkar et al., 2017) can be used to help satisfy Assumption 6, which can be checked with each new data set.

[6] The basis Φ and the history stack $\sum_{i=1}^{M} \mathcal{Y}_i^T \mathcal{Y}_i$ consists of bounded terms, therefore it can be shown that P is upper bounded such that $P \leq \overline{P} I_{p_1}$ $\forall t \in \mathbb{R}_{\geq 0}$ with $\overline{P} \in \mathbb{R}_{>0}$.

[7] The projection algorithm is omitted from the stability analysis for ease of exposition and without loss of generality (cf., Dixon et al., 2003).

where $V_\theta\left(\tilde{\Theta}\left(T_1\right), T_1\right) \leq V_\theta\left(\tilde{\Theta}\left(t_0\right), t_0\right) + \frac{2}{\underline{\Gamma} c_{L1}^2}\overline{v_1}^2$.[8] Using (14) results in (15).[9,10]

3.2. Value function approximation

The solution to (6) is the value function V^*, which is unknown. However, using recent developments in ADP, computationally efficient state-following (StaF) kernels can be used to approximate the value function. Based on the StaF method in Kamalapurkar, Rosenfeld et al. (2016), let $\overline{B_r(x)} \subseteq \mathbb{R}^L$ be a compact set with $y \in \overline{B_r(x)}$ and $c(x) \in \overline{B_r(x)}$, where $c : \chi \to \chi^L$ are centers around the current state $x \in \chi$. Adding and subtracting a known positive definite function $S : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, using StaF kernels centered at x, the optimal value function can be represented as

$$V^*(y) = S(y) + W(x)^T \sigma(y, c(x)) + \varepsilon(x, y), \tag{21}$$

where $\varepsilon : \chi \to \mathbb{R}$ is the continuously differentiable bounded function approximation error, $W : \chi \to \mathbb{R}^L$ is a continuously differentiable ideal StaF weight function, and $\sigma : \chi \times \chi^L \to \mathbb{R}^L$ is a bounded vector of continuously differentiable nonlinear kernels (Kamalapurkar, Rosenfeld et al., 2016).

Property 1. *The selected function S satisfies $S(0) = 0$ and $\nabla S(0) = 0$. Furthermore, there exist known constants $\overline{S}, \overline{\nabla S}, c_s \in \mathbb{R}_{>0}$ such that $\sup_{x\in\chi, y\in\overline{B_r(x)}} |S(y)| \leq \overline{S}$, $\sup_{x\in\chi, y\in\overline{B_r(x)}} \|\nabla S(y)\| \leq \overline{\nabla S}$, and $\|\nabla S(y)\| \leq c_s \|y\|$.*[11]

Using approximations for the ideal weight W, an approximate policy $\hat{u} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^L \to \mathbb{R}^m$ and approximate value function $\hat{V} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$ are expressed as

$$\hat{u}\left(y, x, \hat{W}_a, \hat{W}_2\right) = -\alpha \mathrm{Tanh}\left(\frac{R^{-1}}{2\alpha}\hat{\hat{D}}\left(y, x, \hat{W}_a, \hat{W}_2\right)\right),$$
$$\hat{V}\left(y, x, \hat{W}_c\right) = S(y) + \hat{W}_c^T \sigma(y, c(x)), \tag{22}$$

where $\hat{\hat{D}}\left(y, x, \hat{W}_a, \hat{W}_2\right) \triangleq \hat{G}^T\left(y, \hat{W}_2\right)\left(\nabla\sigma^T(y, c(x))\hat{W}_a + \nabla S^T(y)\right)$, $\hat{G}\left(y, \hat{W}_2\right) \triangleq g(y)\Lambda\left(\hat{W}_2^T\phi_2(y)\right)$, and $\hat{\mu}_a\left(y, \hat{W}_2\right) \triangleq \hat{W}_2^T\phi_2(y)$ denotes the approximated fault.[12] In (22), $\hat{W}_c$, $\hat{W}_a \in \mathbb{R}^L$ denote the critic and actor weight estimates, respectively. Using the system parameter approximations $\hat{\Theta}$, substituting in the approximate value function $\hat{V}$ and control policy $\hat{u}$ into (4) results in a residual error $\delta : \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ called the Bellman Error (BE) computed as

$$\delta\left(y, x, \hat{W}_c, \hat{W}_a, \hat{\Theta}\right) = r\left(y, \hat{u}\left(y, x, \hat{W}_a, \hat{W}_2\right)\right)$$
$$+\hat{W}_c^T(t)\omega\left(y, x, \hat{W}_a, \hat{\Theta}\right) + \omega_S\left(y, x, \hat{W}_a, \hat{\Theta}\right), \tag{23}$$

where $\omega\left(y, x, \hat{W}_a, \hat{\Theta}\right) \triangleq \nabla\sigma(y, c(x))\left(f^o(y) + \Phi\left(y, \hat{u}\left(y, x, \hat{W}_a, \hat{W}_2\right)\right)\hat{\Theta}\right)$, and $\omega_S\left(y, x, \hat{W}_a, \hat{\Theta}\right) \triangleq \nabla S(y)\left(f^o(y) + \Phi\left(y, \hat{u}\left(y, x, \hat{W}_a, \hat{W}_2\right)\right)\hat{\Theta}\right)$.

Since the optimal HJB in (6) is equal to zero for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}_{\geq t_0}$, it is desired to find a set of weights $\hat{W}_c$ and $\hat{W}_a$ such that the BE is driven to zero.

Remark 4. Under the continuity assumptions on the dynamics and the local cost, the admissibility restrictions detailed in Assumption 3, and under the smoothness condition in Assumption 4, the optimal value function can be shown to be the unique positive definite solution of the HJB equation. However, the HJB equation has many other solutions in addition to the unique positive definite solution. Thus, minimization of the BE does not typically guarantee that the resulting weight estimates that approximate the positive definite solution tend to the ideal weights corresponding to this unique positive definite solution, as opposed to tending to the ideal weights corresponding to the other solutions of the HJB that are not positive definite. In this paper, approximation of the positive definite solution is implicitly guaranteed via appropriate selection of initial weight estimates and Lyapunov based update laws that guarantee stability of the closed-loop, see Remark 7.

Remark 5. In the preliminary result in Deptula et al. (2018), a term is only added in the approximate control policy $\hat{u}$, making the policy deviate from the optimal policy. In this work, a positive definite function S is added and subtracted in (21) when approximating the optimal value function; hence the candidate function S serves as a pseudo-value function, and is also contained in the optimal policy u^*.

3.3. Online learning

To implement the approximations online, at each given time instance t, the controller in (22) and BE in (23) are evaluated as $u(t)$ and $\delta_t(t)$, respectively, with $y = x(t)$. Furthermore, leveraging simulation of experience for BE extrapolation (Kamalapurkar et al., 2018, Chapter 7), off-policy trajectories $\left\{x_i : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n\right\}_{i=1}^N$ are selected such that $x_i(x(t), t) \in \overline{B_r(x)}$, then (22) and (23) are evaluated with $y = x_i(x(t), t)$ to give an extrapolated control policy $\hat{u}_i : \mathbb{R}^m \to \mathbb{R}^m$ and BE $\delta_i : \mathbb{R}_{\geq 0} \to \mathbb{R}$ (Kamalapurkar, Rosenfeld et al., 2016).[13] To improve the critic weight estimate, recursive least-squares update laws are designed as

$$\dot{\hat{W}}_c(t) = -\Gamma_c(t)\left(\frac{k_{c1}\omega(t)}{\rho^2(t)}\delta_t(t) + \frac{k_{c2}}{N}\sum_{i=1}^N \frac{\omega_i(t)}{\rho_i^2(t)}\delta_i(t)\right), \tag{24}$$

$$\dot{\Gamma}_c(t) = \beta_c\Gamma_c(t) - \Gamma_c(t)\frac{k_{c1}\omega(t)\omega^T(t)}{\rho^2(t)}\Gamma_c(t)$$
$$-\Gamma_c(t)\frac{k_{c2}}{N}\sum_{i=1}^N \frac{\omega_i(t)\omega_i^T(t)}{\rho_i^2(t)}\Gamma_c(t), \tag{25}$$

with initial conditions $\hat{W}_c(t_0) = \hat{W}_{c0}$ and $\Gamma_c(t_0) = \Gamma_{c0}$. In (24)–(25) $k_{c1}, k_{c2} \in \mathbb{R}_{\geq 0}$ denote learning gains, $\beta_c \in \mathbb{R}_{>0}$ denotes the forgetting factor, $\rho(t) \triangleq 1 + \gamma\omega(t)^T\omega(t)$, and $\gamma \in \mathbb{R}_{>0}$

[8] The bound on V_θ can be made smaller by increasing the number of neurons to decrease $\overline{\varepsilon}_s$, hence decreasing ε_f which decreases v_1. Furthermore, if there is no function approximation error, i.e. $\varepsilon_{sys}(x) = 0$, then it can be shown that $\left\|\tilde{\Theta}(t)\right\| \leq c_E e^{-\frac{\lambda_{\theta 2}}{2}t}\left\|\tilde{\Theta}(t_0)\right\|$, where $c_E \triangleq \sqrt{\frac{\overline{\Gamma}}{\underline{\Gamma}}}e^{\frac{\lambda_{\theta 2}}{2}T_1}$.

[9] Although a switch occurs at the time instance $t = T_1$, the bound is valid for all time before and after T_1, the Lyapunov function in (13) serves as a common Lyapunov function.

[10] Although the parameter estimation and control policy u are coupled in (10), the stability analysis for the update laws in (11) and (12) only require Assumption 5. Moreover, the control policy u is bounded because of the hyperbolic tangent in (5) and (22). Hence, the stability analysis for system identification can be done independently of the stability analysis of the system in (1) and policy u.

[11] Since S is user-defined, it can be selected to satisfy Property 1. An example of a function S which satisfies Property 1 is $\frac{x^Tx}{1+x^Tx}$.

[12] The notation $\nabla W(y, x, \ldots)$ denotes the partial derivative of W with respect to the first argument.

[13] For notational brevity, the notation $C_i(\cdots)$ is defined as $C_i(\cdots) \triangleq C(x_i(x(t), t), \ldots)$, i.e. the function C evaluated at the extrapolated policies $x_i(x(t), t) \in \overline{B_r(x)}$, and the notation $C(t)$ is defined as $C(t) \triangleq C(x(t), \ldots)$.

is a constant normalization gain. The actor weight estimate is improved via the designed update law

$$\dot{\hat{W}}_a(t) = -K_a\left(k_{a1}\left(\hat{W}_a(t) - \hat{W}_c(t)\right) + k_{a2}\hat{W}_a(t)\right) - K_a A^T(t)\hat{W}_c(t), \tag{26}$$

where $A(t) \triangleq \frac{k_{c1}\omega(t)}{\rho^2(t)}G_a^T(t) + \frac{k_{c2}}{N}\sum_{i=1}^{N}\frac{\omega_i(t)}{\rho_i^2(t)}G_{ai}^T(t)$, $G_a(t) \triangleq \alpha\nabla\sigma(x(t))\hat{G}\left(x(t), \hat{W}_2(t)\right)\left(\mathrm{Tanh}\left(\frac{1}{\varepsilon_d}\hat{\bar{D}}(t)\right) - \mathrm{Tanh}\left(\frac{R^{-1}}{2\alpha}\hat{\bar{D}}(t)\right)\right)$, $K_a \in \mathbb{R}^{L\times L}$ is a positive definite gain matrix, $k_{a1}, k_{a2} \in \mathbb{R}_{>0}$ are learning gains, and $\varepsilon_d \in \mathbb{R}_{>0}$ is a small user defined constant.[14]

Furthermore, the states x and x_i, and hence ω and ω_i, are assumed to satisfy the following inequalities.

Assumption 7. There exist constants $T_2 \in \mathbb{R}_{>0}$ and $\underline{c}_1, \underline{c}_2, \underline{c}_3 \in \mathbb{R}_{\geq 0}$ such that

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

$$\underline{c}_2 I_L \leq \int_t^{t+T_2}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\omega_i(\tau)\omega_i^T(\tau)}{\rho_i^2(\tau)}\right)d\tau,\ \forall t\in\mathbb{R}_{\geq t_0},$$

$$\underline{c}_3 I_L \leq \int_t^{t+T_2}\left(\frac{\omega(\tau)\omega^T(\tau)}{\rho^2(\tau)}\right)d\tau,\ \forall t\in\mathbb{R}_{\geq t_0},$$

where at least one of $\underline{c}_1$, $\underline{c}_2$, or $\underline{c}_3$ is strictly positive (Kamalapurkar, Rosenfeld et al., 2016).

Provided $\lambda_{\min}\left\{\Gamma_c^{-1}\right\} > 0$, using a similar argument to Ioannou and Sun (1996, Corollary 4.3.2), the update law in (25) along with Assumption 7 ensure the least squares gain matrix Γ_c satisfies $\underline{\Gamma}_c I_L \leq \Gamma_c \leq \overline{\Gamma}_c I_L$, where $\underline{\Gamma}_c$, $\overline{\Gamma}_c$ are bounds positive bounds (Kamalapurkar, Rosenfeld et al., 2016).

4. Stability analysis

In the following, a Lyapunov-based stability analysis is performed to study the behavior of the closed-loop system. Furthermore, time dependence is suppressed for notational brevity. To facilitate the analysis, the weight estimate errors are defined as $\tilde{W}_c \triangleq W - \hat{W}_c$ and $\tilde{W}_a \triangleq W - \hat{W}_a$, and $\overline{\|(\cdot)\|} \triangleq \sup_{\zeta\in B_\xi}\|(\cdot)\|$ where $B_\xi \subset \chi\times\mathbb{R}^L\times\mathbb{R}^L\times\mathbb{R}^p$ is a compact set containing the origin. Let $Z \triangleq \left[x^T, \tilde{\Theta}^T, \tilde{W}_c^T, \tilde{W}_a^T\right]^T$ and consider the positive definite candidate Lyapunov function $V_L(Z,t) : \mathbb{R}^{n+2L+p}\times\mathbb{R}_{\geq t_0}\to\mathbb{R}$ defined as $V_L(Z,t) \triangleq V^*(x) + \frac{1}{2}\tilde{W}_c^T\Gamma_c^{-1}(t)\tilde{W}_c + \frac{1}{2}\tilde{W}_a^T K_a^{-1}\tilde{W}_a + V_\theta\left(\tilde{\Theta}, t\right)$, such that $\underline{v}_l(\|Z\|) \leq V(Z,t) \leq \overline{v}_l(\|Z\|)$ for class $\mathcal{K}$ functions $\underline{v}_l, \overline{v}_l : \mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$, for all $t\in\mathbb{R}_{\geq t_0}$ and $Z\in\mathbb{R}^{n+2L+p_1}$ (Khalil, 2002, Lemma 4.3).

Theorem 2. *Provided Assumptions 2–7 are satisfied, and* $\lambda_{\min}\{H\} > 0$, $\sqrt{\frac{\iota}{\kappa}} < \underline{v}_l^{-1}\left(\overline{v}_l(\xi)\right)$, *where*

$$H \triangleq \begin{bmatrix} \frac{q}{4} & -\frac{\varphi_{\theta x}}{2} & -\frac{\varphi_{cx}}{2} & -\frac{\varphi_{ac}}{2} \\ -\frac{\varphi_{\theta x}}{2} & \frac{c_{L1}}{16} & -\frac{\varphi_{c\theta}}{2} & 0 \\ -\frac{\varphi_{cx}}{2} & -\frac{\varphi_{c\theta}}{2} & \frac{k_{c2}\underline{c}}{4} & -\frac{\varphi_{ac}}{2} \\ -\frac{\varphi_{ac}}{2} & 0 & -\frac{\varphi_{ac}}{2} & \left(\frac{k_{a1}+k_{a2}}{4} - \varphi_{aa}\right) \end{bmatrix} \text{ with } \iota, \kappa, \varphi_{\theta x}, \varphi_{cx},$$

$\varphi_{ac}, \varphi_{c\theta}, \varphi_{aa} \in \mathbb{R}_{>0}$ *defined in the Appendix and* $\underline{c} \triangleq \frac{\beta_c}{2k_{c2}\overline{\Gamma}_c} + \frac{c_{\underline{1}}}{2}$, *then the approximate policy* $u(t)$, *system state, and weight approximation errors* $\tilde{\Theta}$, $\tilde{W}_c$, *and* $\tilde{W}_a$ *remain uniformly ultimately bounded.*

Proof. Let $z_L(t)$ be a Filippov solution to the differential inclusion $\dot{z}_L(t) \in K[h_L](z_L(t))$ for $t\in\mathbb{R}_{\geq t_0}$ and $h_L : \mathbb{R}^{n+p+2L+p^2+L^2}\to\mathbb{R}^{n+p+2L+p^2+L^2}$ is defined as $h_L \triangleq \left[\dot{Z}^T, \mathrm{vec}\left(\dot{\Gamma}^{-1}\right)^T, \mathrm{vec}\left(\dot{\Gamma}_c^{-1}\right)^T\right]^T$. Due to discrete collections of data points in (10) and (24)–(26), the time derivative of V_L exists almost everywhere and $\dot{V}_L(z_L) \overset{a.e.}{\in} \dot{\tilde{V}}_L(z_L)$, where $\dot{\tilde{V}}_L(z)$ is the generalized time-derivative of (13) along the Filippov trajectories of $\dot{z}_L(t) = K[h_L](z_L(t))$ (Paden & Sastry, 1987). Using the calculus of $K[\cdot]$ and $\dot{V}^* = \nabla V^*(F(x) + G(x)u)$ yields $\dot{\tilde{V}}_L \subseteq \nabla V^* K[F(x) + G(x)u] + \tilde{W}_c^T\Gamma_c^{-1}K\left[\dot{W} - \dot{\hat{W}}_c\right] - \frac{1}{2}\tilde{W}_c^T\Gamma_c^{-1}K\left[\dot{\Gamma}_c\right]\Gamma_c^{-1}\tilde{W}_c + \tilde{W}_a^T K_a^{-1}K\left[\dot{W} - \dot{\hat{W}}_a\right] - \tilde{\Theta}^T\Gamma^{-1}K\left[\dot{\hat{\Theta}}\right] - \frac{1}{2}\tilde{\Theta}^T\Gamma^{-1}K\left[\dot{\Gamma}\right]\Gamma^{-1}\tilde{\Theta}$.

Substituting $\dot{W} = \nabla W(F(x) + G(x)u)$, (4), (11), (12), and (24)–(26) along with the analytical representations of the BEs given by $\delta_t = -\omega^T\tilde{W}_c + G_a^T\tilde{W}_a + \Delta_2$ and $\delta_i = -\omega_i^T\tilde{W}_c + G_{ai}^T\tilde{W}_a + \Delta_{2i}$,[15] then using Young's Inequality and completing with squares, the Lyapunov derivative can be bounded as $\dot{V}_L \overset{a.e.}{\leq} -\kappa\|Z\|^2 - Z_L^T H Z_L - \kappa\|Z\|^2 + \iota$, where $Z_L \triangleq \left[\|x\| \quad \left\|\tilde{\Theta}\right\| \quad \left\|\tilde{W}_c\right\| \quad \left\|\tilde{W}_a\right\|\right]^T \in \mathbb{R}^4$. Provided the sufficient condition $\lambda_{\min}\{H\} > 0$ is met then $\dot{V}_L \overset{a.e.}{\leq} -\kappa\|Z\|^2$, $\forall\|Z\| \geq \sqrt{\frac{\iota}{\kappa}}$, for all $Z\in B_\xi$.

Hence, V_L is a common Lyapunov function, and therefore, Khalil (2002, Theorem 4.18) is invoked to conclude that the concatenated state Z is uniformly ultimately bounded such that $\limsup_{t\to\infty}\|Z(t)\| \leq \underline{v}_l^{-1}\left(\overline{v}_l\left(\sqrt{\frac{\iota}{\kappa}}\right)\right)$. Since, $Z\in\mathcal{L}_\infty$, it follows that $\tilde{\Theta}, \tilde{W}_c, \tilde{W}_a, x \in \mathcal{L}_\infty$, and therefore $u\in\mathcal{L}_\infty$. Furthermore, since W is a continuous function of x, it follows that $W(x)\in\mathcal{L}_\infty$.[16,17]

Remark 6. The sufficient condition $\lambda_{\min}\{H\} > 0$ can be satisfied by increasing the gain k_{a2}, and selecting $Q_x(x)$, the penalty weight R, and update gain K_a with large minimum eigenvalues. For system and fault estimation, increasing the number of neurons p_1 and p_2 can decrease the function approximation errors $\varepsilon_1(x)$ and $\varepsilon_2(x)$, and increasing the size of the history stack can result in a larger minimum eigenvalue, i.e. $\lambda_{\min}\left\{\sum_{i=1}^{M}\mathcal{Y}_i^T\mathcal{Y}_i\right\}$. Therefore, selecting extrapolation points $x_i(x(t), t)$ such that $\underline{c}$ is large, increasing the number of neurons for parameter estimations, i.e. $p_1 \gg n$, $p_2 \gg m$, and selecting $M \gg p$ will ensure the sufficient condition is satisfied.

Remark 7. Since the Lyapunov analysis results in local uniform ultimate boundedness, the convergence results are conditioned on selection of an initial guess for the weights that is close to the ideal weights which correspond to the unique positive definite solution of the HJB as opposed to the ideal weights that correspond to other solutions which are not positive definite. Appropriate selection of an initial guess and the developed

[14] If α is defined as $\alpha \triangleq \mathrm{diag}\{\alpha_1, \ldots, \alpha_m\}$ then $G_a(t)$ can be redefined as $G_a(t) \triangleq \nabla\sigma(x(t))\hat{G}\left(x(t), \hat{W}_2(t)\right)\alpha\left(\mathrm{Tanh}\left(\frac{1}{\varepsilon_d}\hat{\bar{D}}(t)\right) - \mathrm{Tanh}\left(\frac{\alpha^{-1}R^{-1}}{2}\hat{\bar{D}}(t)\right)\right)$.

[15] The terms Δ_2 and Δ_{2i} are defined as $\Delta_2 \triangleq -\Delta_\Phi\Phi_s\tilde{\Theta} + \Delta_a\tilde{W}_a + \nabla S\Delta_S + \Delta_f + \Delta$ and $\Delta_{2i} \triangleq -\Delta_{\Phi i}\Phi_{si}\tilde{\Theta} + \Delta_{ai}\tilde{W}_a + \nabla S_i\Delta_{Si} + \Delta_{fi} + \Delta_i$, where $\Phi_s \triangleq \left[\left(\phi_1^T(x)\otimes I_n\right), \left(\phi_2^T(x)\otimes\left(-\alpha g(x)\Lambda\left(\mathrm{sgn}\left(\hat{\bar{D}}\right)\right)\right)\right)\right]$.

[16] Using Assumptions 2 and 5, the terms Δ_a, Δ_S, and Δ_f are bounded as $\|\Delta_a\| \leq \overline{\Delta}_a$, $\|\Delta_S\| \leq \overline{\Delta}_S$, for $\overline{\Delta}_a$, $\overline{\Delta}_S \in \mathbb{R}_{>0}$, and $\|\Delta_f\| \leq \overline{\left\|\left(\sigma^T\nabla W + \nabla\varepsilon\right)\right\|}L_f\|x\|$, respectively. The functions Δ_Φ, $\Delta_{\Phi i}$, $\Delta, \Delta_i : \mathbb{R}^n\to\mathbb{R}$ are uniformly bounded over χ such that the residual bounds $\|\Delta_\Phi\|$, $\|\Delta_{\Phi i}\|$, $\|\Delta\|$, $\|\Delta_i\|$ decrease with decreasing $\overline{\|\nabla S\|}$, $\overline{\|\nabla W\|}$, $\overline{\|\varepsilon\|}$, and $\overline{\|\varepsilon_i\|}$.

[17] Using Assumption 2 and the Mean Value Theorem between $f(0)$ and $f(x)$ for $x\in\chi$ yields $\|f(x)\| \leq L_f\|x\|$, where $L_f \in \mathbb{R}_{>0}$.

Lyapunov-based stability-preserving update laws implicitly ensure that minimization of the BE results in approximation of the positive definite solution of the HJB equation as desired.

5. Simulation

To demonstrate the effectiveness of the developed method, two simulations for a two-state nonlinear system are provided. The control-affine system in (1) is considered with $f(x) = \left[-x_1 + x_2, \frac{x_2(\cos(2x_1)+2)^2 - x_1 - x_2}{2}\right]^T$ and $g(x) = \left[0, \quad \cos(2x_1) + 2\right]^T$.

In the first simulation, a constant reduced effectiveness fault is considered such that $\mu_a(x) = 0.5$. Moreover, the drift dynamics are assumed to be linear-in-the-parameters (LP) and an exact basis for system identification is provided. The basis to estimate $f(x)$ the basis $\phi_1(x)$ is selected as $\phi_1(x) = \left[x_1, x_2, x_2\left(1 - (\cos(2x_1) + 2)^2\right)\right]^T$ with $f^o(x) = 0_{2\times 1}$, while the basis to estimate $\mu(x)$ is selected as $\phi_2(x) = [1, (\ln(1_{2\times 1} + \exp(x)))^T]^T$. The StaF basis is selected as $\sigma(x, c(x)) = [\sigma_1(x, c_1(x)), \sigma_2(x, c_2(x)), \sigma_3(x, c_3(x))]^T$, where $\sigma_i(x, c_i(x)) = x^T(x + 0.7\upsilon_d(x) d_i(x))$, $i = 1, 2, 3$, $\upsilon_d(x) \triangleq \left(\frac{x^T x + 1\times 10^{-5}}{1 + x^T x}\right)$, and the offsets d_i are selected as $d_1(x) = [0, 1]^T$, $d_2(x) = \left[\frac{\sqrt{3}}{2}, -\frac{1}{2}\right]^T$, and $d_2(x) = \left[-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right]^T$. For BE extrapolation, a single point is selected at random from a uniform distribution over a $\left(\frac{x^T x + 1\times 10^{-4}}{1 + 0.5 x^T x}\right) \times \left(\frac{x^T x + 1\times 10^{-4}}{1 + 0.5 x^T x}\right)$ square centered at the current state $x(t)$. The initial conditions and gains conditions are shown in Table 1.[18]

Fig. 1 shows the developed method regulates the system state and input successfully to the origin for the first simulation. Specifically, Fig. 1a and b show that the states and applied control policy of the system converge to the origin. Fig. 1c shows that the estimates of the unknown parameters, represented by $\hat{W}_1(t)$, converge to the true values $W_1 = \begin{bmatrix} -1, & 1, & 0 \\ -0.5, & 0, & -0.5 \end{bmatrix}^T$, which are known since the drift dynamics $f(x)$ were modeled using a known basis, i.e. $\varepsilon_1(x) = 0_{2\times 1}$. In Fig. 1d, the fault estimate $\hat{\mu}_a$ converges to actual fault $\mu_a = 0.5$; even though the fault is estimated using an uncertain basis. Fig. 1e and f show that the critic and actor weight estimates remain bounded. Since, the optimal weights $W(x)$ are unknown, the estimates cannot be compared to their ideal values.

In the second simulation, a state-varying reduced effectiveness fault is considered as $\mu_a(x) = 0.5 + 0.2\left(1 - \left(\max\left\{\tanh\left(x_1^2\right), \tanh(x_2)\right\}\right)^2\right)\sin(x_1)$. The drift dynamics are assumed to be LP; hence, an exact basis for system identification of $f(x)$ is provided. The basis to estimate $f(x)$ is selected the same as in the first simulation, while to estimate the fault $\mu_a(x)$, the basis $\phi_2(x) = \left[1, \left(\ln\left(1_{4\times 1} + \exp\left(V_\mu^T \bar{x}\right)\right)\right)^T\right]^T$ is selected, where $\bar{x} \triangleq \left[1, x^T\right]^T$ and $V_\mu = U[0.25, 2]\, 1_{3\times 4}$. The rest of the parameters, apart from the dimension of $\hat{\theta}(t)$ changing from a 7×1 to a 11×1 vector, remain the same as in the first simulation.

The results for the second simulation considering a state-varying fault are shown in Fig. 2. Similar to the first simulation, the system state and control policy converge to the origin, as shown in Fig. 2a and b, respectively. Fig. 2c shows that the drift dynamic parameter estimates converge to close to the true values, which are represented by the dashed green lines. Moreover, Fig. 2d shows that the fault estimate converge to the true fault at around 4.5 s. Finally, similar to the first simulation, the actor and critic weight estimates are shown to remain bounded in Fig. 2e and f but cannot be compared to the ideal values since they are not known.

To show the advantage of the developed approach in this paper compared to the preliminary results in Deptula et al. (2018), a comparison of the control policy, system states, and total cost is performed. To provide a fair comparison, the gains are kept the same for each method. The results are shown in Fig. 3 for both the system subject to a constant fault as in the first simulation and a state-dependent fault as in the second simulation. Shown in Fig. 3a and b, the developed result provides a smaller control effort compared to the preliminary result. The developed method also regulates the system to the origin faster than the preliminary approach in Deptula et al. (2018), especially for the constant fault case, as shown by Fig. 3c and d. Moreover, because a smaller control effort is used and the system state is regulated to the origin faster, the total cost is smaller when implementing the developed method compared to the result in Deptula et al. (2018). In Deptula et al. (2018), a robustness term was added to the approximate controller. This term caused the controller deviate from optimality, whereas in the developed method in this paper, the approximate controller in (22) closely resembles the approximate form of the control policy in (5) after substituting in (21). This, and because the developed method in this paper considers state-varying faults, results in the developed method to better handle both the static and state-varying scenarios.

6. Conclusion

An infinite horizon regulation problem for a system with unknown drift dynamics and control effectiveness fault is investigated. An integral data-based system parameter estimator which relaxes the PE condition is developed to simultaneously estimate the drift dynamics and unknown control effectiveness perturbation. Furthermore, the problem is posed as an optimal regulation problem and a local StaF-based ADP method is used to approximate the optimal value function weights. Uniformly ultimately bounded convergence is shown via a Lyapunov stability analysis for the closed-loop system. Simulation results for a two-state nonlinear system are included and compared to preliminary results to illustrate the performance of the developed method.

The developed method focused on a state-varying control effectiveness fault with unknown drift dynamics; however, many systems experience faults that occur at random times and are subject to state-constraints, which can cause potentially unsafe systems. In addition, many systems experience biased faults which degrade controller performance. These different scenarios serve as motivation to possibly investigate the development of a safe ADP approach such as Yang et al. (2019). A topic of future research would be to investigate an ADP approach which considers randomly occurring faults and biased input faults in systems that may be subject to state constraints.

Acknowledgments

This research is supported in part by National Science Foundation, USA award 1509516, Office of Naval Research, USA Grant N00014-13-1-0151, AFOSR, USA award number FA9550-19-1-0169, NEEC, USA award number N00174-18-1-0003, and the OSD Sponsored Autonomy Research Pilot Initiative, USA, and RW Autonomy Initiative, USA. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of sponsoring agencies.

[18] The notation $U[a, b] \times 1_{n\times m}$ denotes a $n \times m-$dimensional matrix with entries selected from a uniform distribution on $[a, b]$, and $1_{n\times m}$ denotes a $n \times m$ matrix of ones.

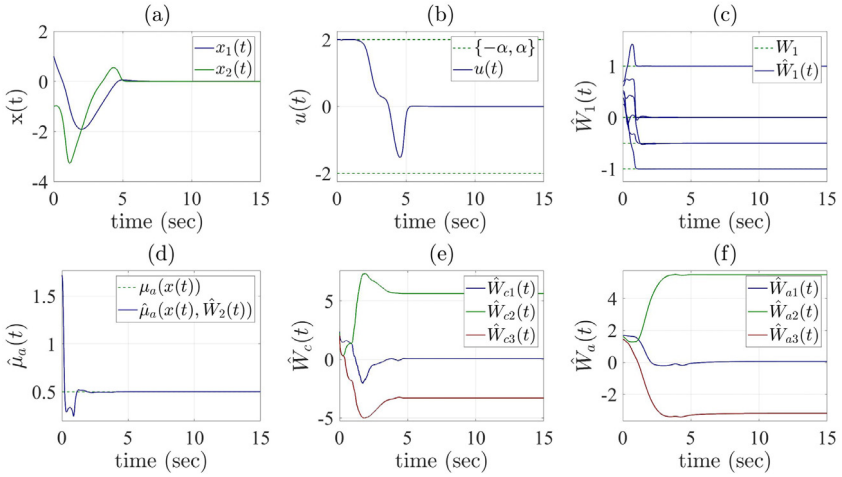


Fig. 1. The system trajectory (a), approximate optimal input (b), estimates of the unknown parameters (c) and fault (d) using the update laws (11) and (12), and estimates of the critic (e) and actor (f) weights using the update laws (24)–(26), respectively. The constant fault, shown in Fig. 1d, is estimated as $\hat{\mu}_a(t) = \hat{W}_2^T(t)\,\phi_2(x(t))$ and converges to the true fault $\mu_a(x)$ shown by the dashed green line.

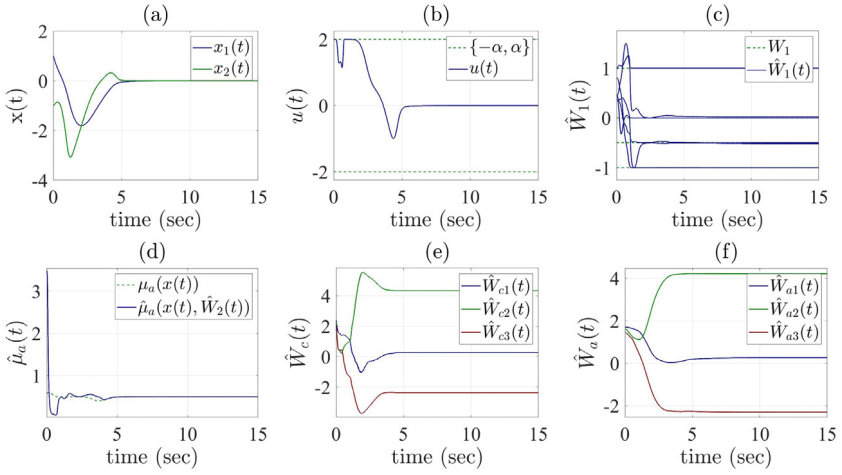


Fig. 2. The system trajectory (a), approximate optimal input (b), estimates of the unknown parameters (c) and fault (d) using the update laws (11) and (12), and estimates of the critic (e) and actor (f) weights using the update laws (24)–(26), respectively. The estimate for state-varying fault, shown in Fig. 2d, is estimated as $\hat{\mu}_a(t) = \hat{W}_2^T(t)\,\phi_2(x(t))$ and converges to the true fault shown by the dashed green line.

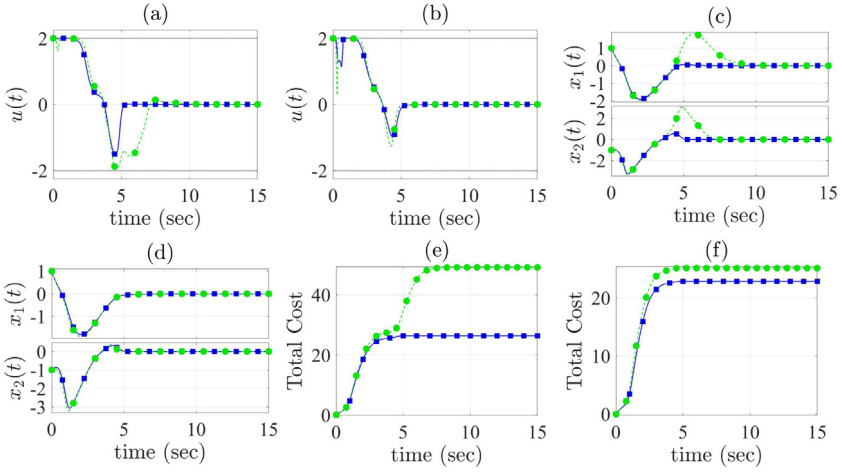


Fig. 3. The comparison of the applied control policy $u(t)$ (Fig. 3a and b), system state $x(t)$ (Fig. 3c and d) and total cost (Fig. 3e and f) between the designed method and the method in Deptula et al. (2018). Fig. 3a, c, and e represent the results for the constant fault while Fig. 3b, d, and f represent the results for the state-varying fault. In each figure, the blue squares represent the developed controller in this paper, while the green circles represent the controller in the preliminary result in Deptula et al. (2018).

Table 1
Initial conditions and parameters for the first simulation.

Initial conditions at $t_0 = 0$
$x(0) = [1.0,\ -1.0]^T$, $\hat{W}_c(0) = U[2,3] \times 1_{3\times 1}$, $\hat{W}_a(0) = 0.7 \times \hat{W}_c(0)$, $\Gamma_c(0) = 10I_3$, $\hat{\theta}(0) = U[0.25,1] \times 1_{9\times 1}$, $\Gamma(0) = 50I_9$.
Penalizing parameters
$Q_x(x) = x^T q_x x$, $q_x = 1.25I_2$, $R = \bar{r} = 0.2$, $\alpha = 2$, $S(x) = \frac{0.1x^T x}{1+x^T x}$.
Gains and parameters for ADP update laws
$k_{c1} = 0.001$, $k_{c2} = 0.9$, $k_{a1} = 0.75$, $k_{a2} = 0.02$, $\gamma_1 = 0.05$, $\beta_c = 0.005$, $K_a = I_5$, $\varepsilon_d = 0.02$, $N = 1$.
Gains and parameters for system identification
$M = 30$, $T = 0.01$, $\ell = 1$, $k_f = 0.1$, $k_p = 0.0001$, $k_{CL} = 10$, $\beta = 10$.

Appendix. Auxiliary terms

To facilitate the analysis in Section 4, $\kappa \in \mathbb{R}_{>0}$ is defined as $\kappa \triangleq \min\left\{\frac{q}{4}, \frac{c_{L1}}{16}, \frac{k_{c2}\underline{c}}{4}, \frac{(k_{a1}+k_{a2})}{4}\right\}$ and the constants $\varphi_{\theta x}, \varphi_{cx}, \varphi_{ac}, \varphi_{c\theta}, \varphi_{aa} \in \mathbb{R}_{>0}$ are defined as $\varphi_{\theta x} \triangleq \overline{\left\|\nabla S + W^T\nabla\sigma + \sigma^T\nabla W + \nabla\varepsilon\right\|}\varphi_g c_s$, $\varphi_{c\theta} \triangleq k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\|\Delta_\Phi\|\|\Phi_s\|} + k_{c2}\frac{1}{N}\sum_{i=1}^{N}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\|\Delta_{\Phi i}\|\|\Phi_{si}\|}$, $\varphi_{ac} \triangleq k_{a1} + k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\|\Delta_a\|} + k_{c2}\frac{1}{N}\sum_{i=1}^{N}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\|\Delta_{ai}\|} + \frac{1}{\underline{\Gamma_c}}\overline{\|\nabla W\|}\varphi_g\overline{W}_2\overline{\|\nabla\sigma\|}$, $\varphi_{cx} \triangleq \frac{1}{\underline{\Gamma_c}}\overline{\|\nabla W\|}L_f + k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\left\|\left(\sigma^T\nabla W + \nabla\varepsilon\right)\right\|}L_f + k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\Delta}_S + \frac{1}{\underline{\Gamma_c}}\overline{\|\nabla W\|}\varphi_g\overline{W}_2$, $\varphi_{ac} \triangleq \frac{1}{\lambda_{\min}(K_a)}\overline{\|\nabla W\|}\left(L_f + \varphi_g\overline{W}_2 c_S + k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\left(\alpha\frac{1}{\varepsilon_d} + \frac{1}{2\lambda_{\min}(R)}\right)\overline{\|W\|}\overline{W}_2^2\overline{\phi}_2^2\overline{g}^2 c_s\right)$, $\varphi_{aa} \triangleq \left(\frac{1}{\lambda_{\min}(K_a)}\overline{\|\nabla W\|}\varphi_g\overline{W}_2\overline{\|\nabla\sigma\|} + \frac{3\sqrt{3}}{16\sqrt{\gamma}}\left(\alpha\frac{1}{\varepsilon_d} + \frac{1}{2\lambda_{\min}(R)}\right)\overline{\|W\|}\overline{W}_2^2\overline{\phi}_2^2\overline{g}^2\left(k_{c1}\overline{\|\nabla\sigma\|}^2 + \frac{k_{c2}}{N}\sum_{i=1}^{N}\overline{\|\nabla\sigma_i\|}^2\right)\right)$, where $\varphi_g \triangleq \overline{g}\overline{\mu}_a\frac{1}{2}\frac{1}{\lambda_{\min}(R)}\overline{\phi}_2\overline{g}$.

Furthermore, the constant $\iota \in \mathbb{R}_{>0}$ is defined as $\iota \triangleq \iota_\Delta + \frac{\iota_x^2}{q} + \frac{4\iota_\theta^2}{c_{L1}} + \frac{\iota_c^2}{k_{c2}\underline{c}} + \frac{\iota_a^2}{(k_{a1}+k_{a2})}$ where $\iota_a \triangleq k_{a2}\overline{\|W\|} + \frac{3\sqrt{3}}{16\sqrt{\gamma}}\left(\alpha\frac{1}{\varepsilon_d} + \frac{1}{2\lambda_{\min}(R)}\right)\overline{W}_2^2\overline{\phi}_2^2\overline{g}^2\left(k_{c1}\overline{\|\nabla\sigma\|}^2\overline{\|W\|} + \frac{k_{c2}}{N}\sum_{i=1}^{N}\left(\overline{\|\nabla\sigma_i\|}^2\overline{\|W\|} + c_s\|x_i\|\right)\right)\overline{\|W\|} + \frac{1}{\lambda_{\min}(K_a)}\overline{\|\nabla W\|}\varphi_g\overline{W}_2\overline{\|\nabla\sigma\|\|W\|} + \overline{\left\|\nabla S + W^T\nabla\sigma + \sigma^T\nabla W + \nabla\varepsilon\right\|}\varphi_g\overline{W}_2\overline{\|\nabla\sigma\|}$, $\iota_c \triangleq \frac{1}{\underline{\Gamma_c}}\overline{\|\nabla W\|}\varphi_g\overline{W}_2\overline{\|\nabla\sigma\|\|W\|} + k_{c1}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\overline{\|\Delta\|} + k_{c2}\frac{1}{N}\sum_{i=1}^{N}\frac{3\sqrt{3}}{16\sqrt{\gamma}}\left(\overline{\nabla S\|\Delta_{Si}\|} + \overline{\|\Delta_{fi}\|} + \overline{\|\Delta_i\|}\right)$, $\iota_\theta \triangleq \overline{\left\|\nabla S + W^T\nabla\sigma + \sigma^T\nabla W + \nabla\varepsilon\right\|}\varphi_g\overline{\|\nabla\sigma\|\|W\|}$, $\iota_x \triangleq \overline{\left\|\nabla S + W^T\nabla\sigma + \sigma^T\nabla W + \nabla\varepsilon\right\|}\overline{g}\overline{\mu}_a\frac{1}{2}\frac{1}{\lambda_{\min}(R)}\overline{\varepsilon}_2\overline{g}c_s$, and $\iota_\Delta \triangleq \frac{\overline{v_1}^2}{c_{L1}} + \frac{\overline{\left\|\nabla S + W^T\nabla\sigma + \sigma^T\nabla W + \nabla\varepsilon\right\|}\overline{g}\overline{\mu}_a}{2\lambda_{\min}(R)} \times \overline{\left\|\Lambda\left(W_2^T\phi_2 + \varepsilon_2\right)g^T\left(\nabla W^T\sigma + \nabla\varepsilon^T\right) + \Lambda(\varepsilon_2)g^T\nabla\sigma^T W\right\|}$.

References

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K. G., Lewis, F. L., & Dixon, W. E. (2013). A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, *49*(1), 89–92.

Chowdhary, G., & Johnson, E. (2011). A singular value maximizing data recording algorithm for concurrent learning. In *Proc. Am. control conf.* (pp. 3547–3552).

Deptula, P., Bell, Z. I., Doucette, E., Curtis, J. W., & Dixon, W. E. (2018). Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with partial loss of control effectiveness. In *Proc. Am. control conf.* (pp. 2521–2526).

Dixon, W. E., Behal, A., Dawson, D. M., & Nagarkatti, S. (2003). *Nonlinear control of engineering systems: A lyapunov-based approach*. Boston: Birkhauser.

Fan, Q.-Y., & Yang, G.-H. (2016a). Active complementary control for affine nonlinear control systems with actuator faults. *IEEE Transactions on Cybernetics*, *47*(11), 3542–3553.

Fan, Q.-Y., & Yang, G.-H. (2016b). Adaptive actor–critic design-based integral sliding-mode control for partially unknown nonlinear systems with input disturbances. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(1), 165–177.

Fan, Q.-Y., & Yang, G.-H. (2016c). Nearly optimal sliding mode fault-tolerant control for affine nonlinear systems with state constraints. *Neurocomputing*, *216*, 78–88.

Filippov, A. F. (1964). Differential equations with discontinuous right-hand side. In *ser. American mathematical society translations - series 2: Vol. 42, Fifteen papers on differential equations* (pp. 199–231). American Mathematical Society.

Ioannou, P., & Sun, J. (1996). *Robust adaptive control*. Prentice Hall.

Jiang, Y., & Jiang, Z.-P. (2017). *Robust adaptive dynamic programming*. John Wiley & Sons.

Jiang, H., Zhang, H., Liu, Y., & Han, J. (2017). Neural-network-based control scheme for a class of nonlinear systems with actuator faults via data-driven reinforcement learning method. *Neurocomputing*, *239*, 1–8.

Kamalapurkar, R., Reish, B., Chowdhary, G., & Dixon, W. E. (2017). Concurrent learning for parameter estimation using dynamic state-derivative estimators. *IEEE Transactions on Automatic Control*, *62*(7), 3594–3601.

Kamalapurkar, R., Rosenfeld, J., & Dixon, W. E. (2016). Efficient model-based reinforcement learning for approximate online optimal control. *Automatica*, *74*, 247–258.

Kamalapurkar, R., Walters, P., & Dixon, W. E. (2016). Model-based reinforcement learning for approximate optimal regulation. *Automatica*, *64*, 94–104.

Kamalapurkar, R., Walters, P. S., Rosenfeld, J. A., & Dixon, W. E. (2018). *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Springer.

Khalil, H. K. (2002). *Nonlinear systems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kirk, D. (2004). *Optimal control theory: An introduction*. Mineola, NY: Dover.

Lewis, F. L., & Liu, D. (2013). *Reinforcement learning and approximate dynamic programming for feedback control* (p. 17). John Wiley & Sons.

Liu, L., Wang, Z., & Zhang, H. (2017). Adaptive fault-tolerant tracking control for mimo discrete-time systems via reinforcement learning algorithm with less learning parameters. *IEEE Transactions on Automatic Sciences*, *14*(1), 299–313.

Lv, Y., Na, J., Yang, Q., Wu, X., & Guo, Y. (2016). Online adaptive optimal control for continuous-time nonlinear systems with completely unknown dynamics. *International Journal of Control*, *89*(1), 99–112.

Modares, H., Lewis, F. L., & Naghibi-Sistani, M.-B. (2014). Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*, *590*(1), 193–202.

Paden, B. E., & Sastry, S. S. (1987). A calculus for computing filippov's differential inclusion with application to the variable structure control of robot manipulators. *IEEE Transactions on Circuits Systems*, *34*(1), 73–82.

Parikh, A., Kamalapurkar, R., & Dixon, W. E. (2019). Integral concurrent learning: Adaptive control with parameter convergence using finite excitation. *International Journal of Adaptive Control and Signal Processing*, *33*(12), 1775–1787.

Roy, S. B., Bhasin, S., & Kar, I. N. (2018). Combined mrac for unknown mimo lti systems with parameter convergence. *IEEE Transactions on Automatic Control*, *63*(1), 283–290.

Shen, Y., Liu, L., & Dowell, E. H. (2013). Adaptive fault-tolerant robust control for a linear system with adaptive fault identification. *IET Control Theory Applications*, *7*(2), 246–252.

Vamvoudakis, K. G., Miranda, M. F., & Hespanha, J. P. (2016). Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(11), 2386–2398.

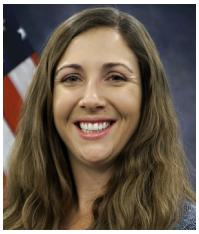
Yang, Y., Yin, Y., He, W., Vamvoudakis, K. G., Modares, H., & Wunsch, D. C. (2019). Safety-aware reinforcement learning framework with an actor-critic-barrier structure. In *Proc. Am. control conf.* (pp. 2352–2358). IEEE.

Zhang, H., Cui, L., Zhang, X., & Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, *22*(12), 2226–2236.

Zhao, B., Liu, D., & Li, Y. (2017). Observer based adaptive dynamic programming for fault tolerant control of a class of nonlinear systems. *Information Sciences*, *384*, 21–33.

Patryk Deptula received his B.Sc. degree in Mechanical Engineering (major) and Mathematics (minor) from Central Connecticut State University in 2014. Dr. Deptula then received his M.S. and Ph.D. degrees in Mechanical Engineering from the Department of Mechanical and Aerospace Engineering from the University of Florida in 2017 and 2019, respectively. In 2019, he joined The Charles Stark Draper Laboratory, Inc. He is a recipient of the Henry Barnard Scholar award (CT), Dean's Citation award (CCSU), NASA CT Space Grant Consortium (CCSU), and Best Graduate Researcher Award (UF). His current research interests include, but are not limited to learning-based and adaptive methods, estimation, guidance, and control; sensor fusion; multi-agent systems; human–machine interaction; vision based navigation and control; biomedical systems; biomechanics; and robotics applied to a variety of fields.

Zachary I. Bell received his M.S. Degree in Mechanical Engineering in 2017 and the B.S. Degree in Mechanical Engineering with a minor in Electrical Engineering in 2015 from the University of Florida. In 2018, he was awarded the Science, Mathematics, and Research for Transformation (SMART) Scholarship, sponsored by the Department of Defense. He received his Ph.D. from the University of Florida in 2019. His research interest is in Lyapunov-based estimation and control theory, visual estimation, simultaneous localization and mapping, and reinforcement learning.



Emily A. Doucette is a senior research engineer at the Air Force Research Laboratory Munitions Directorate where she is the Multi-Domain Networked Weapons Team technical lead. Prior to this post, she has served the Munitions Directorate as the Assistant to the Chief Scientist (2017–2019) and as a research engineer for the Weapon Dynamics and Control Sciences Branch since 2012. She earned a Ph.D. in aerospace engineering from Auburn University and is a recipient of the SMART Scholarship. Her research interests include estimation theory, human–machine teaming, decentralized task assignment, cooperative autonomous engagement, and risk-aware target tracking and interdiction.

J. Willard Curtis is the guidance, navigation and control core technical competency lead at the Air Force Research Laboratory's Munitions Directorate at Eglin AFB Florida. He began his undergraduate studies as a freshman studying engineering physics at Cornell University in 1993 and after a series of adventures and winding roads he received his Ph.D. degree in Electrical and Computer Engineering from Brigham Young University in 2001. His current research interests include swarm control and distributed estimation.

Warren E. Dixon received his Ph.D. in 2000 from the Department of Electrical and Computer Engineering from Clemson University. He worked as a research staff member and Eugene P. Wigner Fellow at Oak Ridge National Laboratory (ORNL) until 2004, when he joined the University of Florida in the Mechanical and Aerospace Engineering Department. His main research interest is the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. He is an ASME Fellow and IEEE Fellow.