Approximate Dynamic Programming for Trajectory Tracking of Switched Systems

Max L. Greene, Masoud S. Sakha, Rushikesh Kamalapurkar, Warren E. Dixon

***Abstract*—This paper develops a technique for online approximate optimization of tracking control policies for a family of switched nonlinear dynamical systems. Optimization is realized via approximate dynamic programming, and integral concurrent learning is used for robustness to parametric uncertainties. The family of switched systems is composed of finitely many subsystems, which may have differing characteristics, such as dynamics and cost functions. This paper develops a new result on the analysis of switched systems comprised of locally practically stable subsystems using multiple Lyapunov-like functions. Local practical stability of the overall switched system and convergence of the applied tracking control policies to a neighborhood of the optimal tracking control policies is then proven for an arbitrary switching sequence provided that a set of sufficient gain conditions and a minimum dwell-time condition are satisfied. Simulation results are presented for optimal control of an autonomous underwater vehicle in the presence of a set of discretely varying irrotational currents to show the efficacy of the developed technique.**

I. Introduction

Switched systems are dynamical systems that can operate in various modes of operation (also referred to as subsystems) in response to internal and external stimuli [1]. Switching behaviors can result from changes in control objectives, system parameters, actuator limitations including saturation and on/off control, modeling choices where a complex system is composed of several simpler models, and design choices including gain scheduling, where a complex control design problem is separated into multiple control design problems [2]–[6]. Solutions to optimal control problems provide a stabilizing control policy, which can be used to facilitate a regulation or tracking objective [7, Ch. 5]. This paper considers the optimal control of a switched nonlinear system. The control objective is to optimize the performance of each subsystem and to schedule switching behaviors to ensure stability of the overall switched system.

The performance objective for each subsystem is encoded in terms of the minimization of a cost functional, which results in a nonlinear optimal control problem (NOCP) corresponding to each subsystem. Solutions to a large classes of NOCPs can be characterized using the Hamilton-Jacobi-Bellman (HJB) equation, which is generally difficult to solve [7, Ch. 2]. HJB equations can be solved numerically (e.g,. [8] and [9]) given a dynamic model of the system; however, the resulting feedback controllers may be rendered ineffective if the model includes uncertainty.

Reinforcement learning (RL) has been used to approximate solutions of optimal control problems [10] and [11]. Approximate dynamic programming (ADP) utilizes a RL-based actor-critic framework to solve NOCPs in the presence of modeling uncertainty via value function approximation [12]–[16]. For a class of NOCPs that includes affine-quadratic NOCPs, once the optimal value function is successfully approximated, a stabilizing optimal control policy can be determined. The optimal value function of each subsystem is approximated with a separate single-layer linear-in-the-parameters neural network (NN); the weights of the NN are updated according to the Bellman error (BE), which is a performance metric that indirectly measures the quality of the value function approximation. If a system model is available, it can be used to improve learning efficiency via simulation of experience [17]–[19]. If the system model includes uncertain model parameters, they may be approximated in real-time using techniques such as integral concurrent learning (ICL) under finite excitation (FE) conditions [20]. Unlike standard concurrent learning results [21], ICL circumvents the need to compute state derivatives. In a switched systems context, since ICL relies on historical input-output data, ICL facilitates exponential identification of a subsystem's uncertain parameters regardless of the subsystem's (in)activity. The identified parameters enable simulation of experience (i.e., to calculate the BE at off-trajectory locations in the state space) to facilitate value function approximation. While ADP with simulation of experience has proven to be an effective tool for approximate optimal control of the subsystems (e.g., [15], [19], [22]–[26]), optimal control of switched nonlinear systems remains a challenge.

Numerical optimal control of switched systems has been investigated in results such as [27]–[31]. Early results on switched optimal control rely on a simplifying assumption such as a fixed switching sequence [27], [31], [32] or a fixed switching surface [33] and [34]. Methods to optimize the switching sequence, under the assumption that the switching times are fixed, are also developed in [29]. More recently, methods based on mode insertion [30], dynamic programming [31], and embedding [35] and [36] have been addressed

Max L. Greene is with Aurora Flight Sciences, a Boeing Company, Cambridge, MA, 02142 USA. Email: greene.max@aurora.aero. Masoud Sakha, Rushikesh Kamalpurkar, and Warren E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, 32611 USA. Email: {masoud.sakha,rkamalapurkar,wdixon}@ufl.edu.

This research is supported in part by Office of Naval Research grant N00014-13-1-0151, NEEC award N00174-18-1-0003, AFOSR award FA9550-18-1-0109, AFOSR award FA9550-19-1-0169, AFRL award number FA8651-19-2-0009, and NSF award 1762829 . 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 Aurora Flight Sciences or the sponsoring agencies.

simultaneous offline optimization of switching sequences and switching times. Results such as [37] and [38] also examine the use of dynamic programming for offline optimization of switched discrete-time nonlinear systems. Unlike the aforementioned methods, the objective in this paper is online optimization, where a Lyapunov-based framework establishes convergence of the subsystem control policies to a neighborhood of their respective optimal policies while maintaining stability of the overall switched system. Optimization of the switching logic relative to a system-wide performance metric is not considered in this work (cf. [30], [31], [35], and [36]) and is a topic for future research.

The Lyapunov-like function for each subsystem includes its respective optimal value function. Since the optimal value function is generally different for each subsystem [39, Ch. 3], the switched system is analyzed using multiple Lyapunov-like functions [1, Ch. 3.1]. A complication in Lyapunov-based analyses of switched systems is the growth and discontinuity of Lyapunov-like functions at the switching instances [40]. To overcome this complication, a minimum dwell-time analysis is utilized. A minimum dwell-time is a lower bound on the time required before the scheduler can switch to a different subsystem. While the value of one Lyapunov-like function may decrease while the corresponding subsystem is active, its value may increase when the subsystem becomes inactive. The minimum dwell-time accounts for the worst-case growth between multiple Lyapunov-like functions at switching instances [1, Sec. 3.1]. In doing so, overall system stability, when following an arbitrary switching sequence of subsystems (i.e., the sequence of active subsystems is arbitrary, but the timing between switching instances is not arbitrary), can be established provided the switching time instances satisfy the minimum dwell-time condition. This paper develops a continuous-time ADP-based tracking controller for an arbitrary switching sequence between multiple dynamical subsystems with a time-varying switching signal, which may be based on environmental conditions or the user's discretion.

Online optimal regulation of uncertain linear and nonlinear switched systems is studied in [41]–[43]. In [42], algorithms for learning the optimal feedback gains in each subsystem are developed, but stability of the switched system is analyzed assuming fixed learned feedback gains, and as such, stability during the learning phase is not analyzed. The result in [43] concerns safe ADP-based control within a hybrid systems framework; however, it does not analyze multiple subsystems that are intermittently activated. In [41], the trajectory tracking problem is not considered, quadratic bounds on the optimal value function are assumed, projection operators are required, and the adaptive parameters (solely the actor-critic weights) of a subsystem are only updated while that subsystem is active. As such, the state of each subsystem only includes its respective state. This approach results in a concatenated state vector that is unique to each subsystem. Since dwell-time analysis typically relies on continuity of the state vector at the switching instance, the absence of a common state vector between subsystems makes the analysis challenging, resulting in overly conservative assumptions on the Lyapunov-like functions and overly conservative dwell-time bounds. Another issue noticed in [41] is that the actor-critic weights only update while their respective subsystem is active, resulting in piecewise continuous trajectories. Similar sample-and-hold strategies have been analyzed using a hybrid systems framework [44, Ch. 3] but they do not consider subsystems that are locally practically stable or have uniformly ultimately bounded (UUB) or locally practically stable trajectories.

Unlike [41], the method developed in this paper relaxes the assumptions on the optimal value function, removes the projection operators on the update policies, considers the optimal trajectory tracking problem, performs online system identification of each subsystem simultaneously, and simultaneously updates every subsystem's value function weight estimates. Furthermore, this paper uses model-based ADP to continuously learn the value functions and system parameters of one subsystem while another subsystem is active. Model-based ADP relies on simulation of experience, which enables policy updates while a subsystem is inactive. The advantage of this simultaneous online learning technique is that it results in a common state that remains continuous at the switching instance (i.e., no reset map is needed [1, Ch. 1.1.1]). Having a common and continuous state enables quantification of the convergence rate and the ultimate bound of the active Lyapunov-like function and also the change in value of the active Lyapunov-like functions before and after a switch. Such quantification enables computation of the minimum dwell-time needed to maintain local practical stability (see [45] and [46]) of the switched system. This paper develops a new result on the relationship between local practical stability of the subsystems and local practical stability of the switched system, which can be applied to general switching problems with locally practically stable subsystems that must be analyzed using multiple Lyapunov-like functions.

The remainder of this paper is structured as follows. Section II formulates the tracking ADP problem and objectives. Section III outlines the ICL-based system identifier, which is used to estimate the system dynamics online. Section IV introduces BE extrapolation, which requires the parametric model from Section III. Section V defines the update laws that facilitate the ADP algorithm and subsequent stability analysis. Section VI outlines the simultaneous learning that facilitates the subsequent stability analysis. Section VII presents a Lyapunov-based stability analysis and dwell-time analysis. Section VIII presents a simulation result to illustrate the effectiveness of the developed technique.

Notation

For notational brevity, time-dependence is omitted while denoting trajectories of the dynamical systems. For example, the trajectory $x(t)$, where $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, is denoted as $x \in \mathbb{R}^n$ and referred to as x instead of $x(t)$. For example, an equation of the form $f + h(y,t) = g(x)$ should be interpreted as $f(t) + h(y(t),t) = g(x(t)) \; \forall t \in \mathbb{R}_{\geq 0}$. The gradient $\left[\frac{\partial f(x,y)}{\partial x_1}, \ldots, \frac{\partial f(x,y)}{\partial x_n}\right]^T$, where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$,

$f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ and $\frac{\partial f(x,y)}{\partial x_k} \in \mathbb{R}^m$ is denoted by $\nabla_x f(x,y)$. $\|\cdot\|$ denotes both the Euclidean norm for vectors and Frobenius norm for matrices. The cardinality of a set A is denoted by $|A|$. $\mathbf{1}_{n\times m}$ and $\mathbf{0}_{n\times m}$ denote matrices of ones and zeros with n rows and m columns, respectively. $I_{n\times n}$ denotes an $n \times n$ identity matrix. Generally, the subscript p defines the quantity or function belonging to the p^{th} mode of the overall system.

II. Problem Formulation

Let $\dot{x} = f_p(x) + g_p(x)\,u$ denote a family of finitely many dynamical systems, indexed by $p \in \mathcal{P} \subset \mathbb{N}$ with $|\mathcal{P}| < \infty$, where $x \in \mathbb{R}^n$ denotes the system state and $u \in \mathbb{R}^m$ denotes the control input. The function $f_p : \mathbb{R}^n \to \mathbb{R}^n$ models the drift dynamics, and the function $g_p : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ models the control effectiveness of the p^{th} subsystem. The control objective is to track a time-varying piecewise continuously differentiable signal $x_d : \mathbb{R}_{\geq t_0} \to \mathbb{R}^n$. To quantify the tracking objective, the tracking error is defined as $e \triangleq x - x_d$. Using the technique in [19] to transform the time-varying tracking problem into an infinite horizon regulation problem, the control-affine dynamics are rewritten as

$$\dot{\zeta} = F_p(\zeta) + G_p(\zeta)\,\mu_p, \tag{1}$$

where $\zeta \in \mathbb{R}^{2n}$ is the concatenated state vector $\zeta \triangleq \left[e^T, x_d^T\right]^T$, $\mu_p \triangleq u - u_{d,p}(x_d)$ is the feedback portion of the controller, $u_{d,p} : \mathbb{R}^n \to \mathbb{R}^m$ is the subsequently-defined feedforward component of the controller that facilitates the trajectory tracking objective, $F_p : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is defined as

$$F_p(\zeta) \triangleq \begin{bmatrix} f_p(e + x_d) - h_{d,p}(x_d) + g_p(e + x_d)\,u_{d,p}(x_d) \\ h_{d,p}(x_d) \end{bmatrix}, \tag{2}$$

where $h_{d,p} : \mathbb{R}^n \to \mathbb{R}^n$ is a user-selected desired trajectory generation function and subsequently-defined in Assumption 4, and $G_p : \mathbb{R}^{2n} \to \mathbb{R}^{2n\times m}$ is defined as

$$G_p(\zeta) \triangleq \left[g_p(x)^T, \mathbf{0}_{m\times n}\right]^T. \tag{3}$$

The following assumptions facilitate the development of the approximate optimal tracking controller (see [19]).

Assumption 1. The function f_p is continuously differentiable and $f_p(0) = 0$ for all $p \in \mathcal{P}$.

Assumption 2. The function g_p is locally Lipschitz, $g_p(x)$ has full column rank for all $(x,p) \in \mathbb{R}^n \times \mathcal{P}$, and there exists a known constant $\overline{g_p} \in \mathbb{R}_{>0}$ such that $\|g_p(x)\| \leq \overline{g_p}$, for all $(x,p) \in \mathbb{R}^n \times \mathcal{P}$. It follows that there exists a known constant $\overline{G_p} \in \mathbb{R}_{>0}$ such that $0 < \|G_p(\zeta)\| \leq \overline{G_p}$ for all $(\zeta,p) \in \mathbb{R}^{2n} \times \mathcal{P}$.

Assumption 3. The desired trajectory is bounded from above by a positive constant $\overline{x_d} \in \mathbb{R}_{\geq t_0}$ such that $\sup_{t\in\mathbb{R}_{\geq 0}} \|x_d(t)\| \leq \overline{x_d}$.

Assumption 4. The desired trajectory of the p^{th} system is a solution of $\dot{x}_d = h_{d,p}(x_d)$, starting from $x_d(0)$, where $h_{d,p} : \mathbb{R}^n \to \mathbb{R}^n$ are user-selected locally Lipschitz continuous trajectory generation functions that satisfy $h_{d,p}(0) = 0$ and $g_p(x_d)\,g_p^+(x_d)\,(h_{d,p}(x_d) - f_p(x_d)) = h_{d,p}(x_d) - f_p(x_d)$ for all $x_d \in \mathbb{R}^n$, and $p \in \mathcal{P}$, where $g_p^+ : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is defined as $g_p^+(x) \triangleq \left(g_p^T(x)\,g_p(x)\right)^{-1} g_p^T(x)$.

Based on Assumptions 2-4, the controller u can be separated into two components: a feedback component μ_p, which is subsequently defined, and a feedforward component $u_{d,p}(x_d)$, defined as $u_{d,p}(x_d) \triangleq g_p^+(x_d)\,(h_{d,p}(x_d) - f(x_d))$.

A. Control Objectives

Solutions to optimal control problems provide control polices that facilitate tracking objectives [7, Ch. 11]. Tracking objectives prescribe the system states to follow a specific, often user-defined, function of time. The control design problem under consideration has three objectives. The first objective is to solve the infinite-horizon optimal tracking problem for each subsystem online and in real time. That is, for each fixed $p \in \mathcal{P}$, the aim is to determine a feedback control policy μ_p that minimizes the infinite horizon cost functional, J_p, defined as

$$J_p(\zeta(\cdot), \mu_p(\cdot)) \triangleq \int_{t_0}^{\infty} \overline{Q}_p(\zeta(\tau)) + \mu_p^T(\tau)\,R_p\mu_p(\tau)\;d\tau, \tag{4}$$

subject to (1) while tracking the desired trajectory output by the p^{th} subsystem's desired trajectory generator, where $\overline{Q}_p \in \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ is a positive semidefinite (PSD) user-selected state cost function of the p^{th} subsystem, and $R_p \in \mathbb{R}^{m\times m}$ is a user-selected positive definite (PD) symmetric input cost matrix for the p^{th} subsystem. The second objective is to generate an online estimate of the feedforward controllers $u_{d,p}(x_d)$ by learning the uncertain parameters in the drift models of the subsystems. The third objective is to characterize the class of allowable switching signals $\sigma : \mathbb{R}_{\geq 0} \to \mathcal{P}$ and estimate a set of initial conditions such that the dynamics of the error between the trajectories of the switched desired trajectory generator $\dot{x}_d = h_{d,\sigma(t)}(x_d)$ and the switched system $\dot{x} = f_{\sigma(t)}(x) + g_{\sigma(t)}(x)\,u$, under the developed controller, are practically stable.[1]

To ensure that the optimal controllers in each subsystem are stabilizing, it is assumed that the state cost function is of the form $\overline{Q}_p(\zeta) \triangleq Q_p(e)$, where $Q_p : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a PD function that is independent of x_d. By [47, Lemma 4.3], $\overline{Q}_p$ satisfies $\underline{q}_p(\|e\|) \leq \overline{Q}_p(\zeta) \leq \overline{q}_p(\|e\|)$ for all $(\zeta,p) \in \mathbb{R}^{2n} \times \mathcal{P}$, where $\underline{q}_p, \overline{q}_p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions. For example, let $\overline{Q}_p(\zeta) = e^T e + x_d^T \mathbf{0}_{n\times n} x_d$.

[1] In this problem formulation we consider the case in which x_d is continuous through the switching instances, but owing to different desired behaviors in different modes of operation, modeled by the functions $h_{d,p}$, the time derivative $\dot{x}_d$ may be piecewise continuous at the switching instances.

B. Exact Solution of Subsystem Optimal Control Problems

The infinite horizon value function (i.e., the cost-to-go) for the p^{th} subsystem $V_p^* : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ is defined as

$$V_p^*(\zeta_0) \triangleq \min_{\mu_p} \int_t^\infty \overline{Q}_p(\zeta(\tau)) + \mu_p^T(\zeta(\tau)) R_p \mu_p(\zeta(\tau)) \, d\tau, \tag{5}$$

where $\zeta(\cdot)$ is the trajectory of (1) starting from some initial state ζ_0 under the feedback policy μ_p and the minimization is over the set of admissible feedback policies [15, Chapter 1]. If the optimal value function V_p^* is continuously differentiable and $(e,t) \mapsto V_p^*\left(\begin{bmatrix} e \\ x_d(t) \end{bmatrix}\right)$ is PD, then V_p^* is the unique stabilizing solution of the corresponding HJB equation (e.g., [48, Chapter 5])

$$\begin{aligned} 0 = \overline{Q}_p(\zeta) + \mu_p^{*T}(\zeta) R_p \mu_p^*(\zeta) \\ + \nabla_\zeta V_p^*(\zeta) \left(F_p(\zeta) + G_p(\zeta) \mu_p^*(\zeta)\right), \end{aligned} \tag{6}$$

where the optimal feedback policy $\mu_p^* : \mathbb{R}^{2n} \to \mathbb{R}^m$ is given by

$$\mu_p^*(\zeta) = -\frac{1}{2} R_p^{-1} G_p(\zeta)^T \left(\nabla_\zeta V_p^*(\zeta)\right)^T. \tag{7}$$

C. Value Function Approximation

Parametric NN-based methods can be used to approximate the optimal value function over a compact domain. To facilitate solving the HJB equation in (6), let $\Omega_p \subset \mathbb{R}^{2n}$ be a compact set and consider the approximation of the optimal value function in (5) over the set Ω_p given by[2]

$$V_p^*(\zeta) = W_p^T \phi_p(\zeta) + \epsilon_p(\zeta), \tag{8}$$

where $W_p \in \mathbb{R}^L$ is an unknown bounded vector of weights, $\phi_p : \mathbb{R}^{2n} \to \mathbb{R}^L$ is a user-selected vector of basis functions, and $\epsilon_p : \mathbb{R}^{2n} \to \mathbb{R}$ is the bounded function approximation error.[3] Substituting (8) into (7), the transient optimal control policy of the p^{th} subsystem μ_p^* can be expressed as

$$\mu_p^*(\zeta) = -\frac{1}{2} R_p^{-1} G_p(\zeta)^T \left(W_p \nabla_\zeta \phi_p(\zeta) + \nabla_\zeta \epsilon_p(\zeta)\right)^T. \tag{9}$$

Assumption 5. There exist constants $\overline{W_p}, \overline{\phi_p}, \overline{\nabla_\zeta \phi_p}, \overline{\epsilon_p}, \overline{\nabla_\zeta \epsilon_p} \in \mathbb{R}_{>0}$ such that the unknown weights W, user-defined vector of basis functions $\phi_p(\cdot)$, and function approximation error ϵ_p, can be bounded such that $\|W_p\| \leq \overline{W_p}$, $\sup_{\zeta \in \Omega_p} \|\phi_p(\zeta)\| \leq \overline{\phi_p}$, $\sup_{\zeta \in \Omega_p} \|\nabla_\zeta \phi_p(\zeta)\| \leq \overline{\nabla_\zeta \phi_p}$, $\sup_{\zeta \in \Omega_p} \|\epsilon_p(\zeta)\| \leq \overline{\epsilon_p}$, and $\sup_{\zeta \in \Omega_p} \|\nabla_\zeta \epsilon_p(\zeta)\| \leq \overline{\nabla_\zeta \epsilon_p}$ [13, Assumptions 9.1.c-e].[4]

The ideal weights W_p in (8) and (9) are unknown; hence, an approximation of W_p is sought. Specifically, the critic weight estimate, $\hat{W}_{c,p} \in \mathbb{R}^L$ is substituted to approximate the optimal value function $\hat{V}_p : \mathbb{R}^{2n} \times \mathbb{R}^L \to \mathbb{R}$ denoted as

$$\hat{V}_p\left(\zeta, \hat{W}_{c,p}\right) \triangleq \hat{W}_{c,p}^T \phi_p(\zeta). \tag{10}$$

Similarly, another estimate for W_p, called the actor weight estimate $\hat{W}_{a,p} \in \mathbb{R}^L$, is used to provide an approximate version of (9), the approximate optimal control policy $\hat{\mu}_p : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$ is

$$\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right) = -\frac{1}{2} R_p^{-1} G_p(\zeta)^T \left(\nabla_\zeta \phi_p(\zeta)^T \hat{W}_{a,p}\right). \tag{11}$$

III. Identification of the Feedforward Component

If the drift dynamics contain parametric uncertainties then online system identification is needed to learn the unknown feedforward component $u_{d,p}$ of the controller and for model-based learning of the unknown feedback component μ_p of the controller. To facilitate system identification objective, let $\hat{f}_p : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ be a parametric estimate of the drift dynamics f_p.[5] Assume that the drift dynamics are linearly parameterizable, i.e., $f_p(x) \triangleq Y_p(x)\theta_p$, where $Y_p : \mathbb{R}^n \to \mathbb{R}^{n \times s}$ is a known regression matrix and $\theta_p \in \mathbb{R}^s$ denotes the unknown constant parameters of f_p.[6] Using an approximation of the unknown parameter vector $\hat{\theta}_p \in \mathbb{R}^s$, an approximation of the p^{th} uncertain drift dynamics $\hat{f}_p : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ is defined as $\hat{f}_p\left(x, \hat{\theta}_p\right) \triangleq Y_p(x)\hat{\theta}_p$. The parameter estimates are updated using the ICL-based update policy

$$\begin{aligned} \dot{\hat{\theta}}_p \triangleq -k_{p,\theta} \Gamma_{p,\theta} \\ \cdot \sum_{j=1}^{M_p} \mathcal{Y}_{p,j}^T \left(x(t_j) - x(t_j - \Delta t) - \mathcal{U}_{p,j} - \mathcal{Y}_{p,j} \hat{\theta}_p\right) \end{aligned} \tag{12}$$

based on the result in [20], where $k_{p,\theta} \in \mathbb{R}_{>0}$ and $\Gamma_{p,\theta} \in \mathbb{R}^{s \times s}$ are PD learning gains, $M_p \in \mathbb{Z}^+$ is the user-defined number of elements in the subsequently-defined history stacks, $\mathcal{Y}_{p,j} \triangleq \mathcal{Y}_p(t_j)$, $\mathcal{U}_{p,j} \triangleq \mathcal{U}_p(t_j)$, $\mathcal{Y}_p(t) \triangleq \int_{\max\{t-\Delta t, t_0\}}^t Y_p(x(\tau)) \, \mathrm{d}\tau$, $\mathcal{U}_p(t) \triangleq \int_{\max\{t-\Delta t, t_0\}}^t g_p(x(\tau)) u(\tau) \, \mathrm{d}\tau$, where τ is the integration variable, and $t_j \in \mathbb{R}_{\geq t_0}$ is the time when the state-input pairs are recorded.

Assumption 6. History stacks containing recorded values of state and control signals $\{x(t_j), x(t_j - \Delta t), u(t_j)\}_{j=1}^{M_p}$ that satisfy $\underline{\mathcal{Y}}_p \triangleq \lambda_{\min}\left\{\sum_{j=1}^{M_p} \mathcal{Y}_{p,j}^T \mathcal{Y}_{p,j}\right\} > 0$ are available *a priori* for all subsystems $p \in \mathcal{P}$ [20, Assumption 1].

[2] The subsequent stability analysis in Theorem 1 proves that if ζ is initialized within an appropriately-sized subset of Ω_p, then it will remain in Ω_p for all $t \in \mathbb{R}_{\geq 0}$.

[3] Each subsystem p can have finitely many neurons. The number of neurons in each subsystem's ϕ_p can be different (e.g., L for subsystem 2 need not be the same as L in subsystem 1). However, to focus the subject of this manuscript and to minimize the amount of notation, generally, L represents the number of neurons in ϕ_p for all $p \in \mathcal{P}$.

[4] Assumption 5 can be satisfied by selecting ϕ_p to be a polynomial basis function [49, Theorem 1.5].

[5] $s \in \mathbb{N}$ represents the number of uncertain parameters for each subsystem p. Each subsystem p may have a different number of uncertainties and, therefore, different value of s. However, to focus the subject of this manuscript and to simplify the notation, let $s = s_p$ represent the number of parametric uncertainties for each subsystem.

[6] Linear parameterizations of the drift dynamics f_p require partial knowledge of a system's dynamics. The developed technique can be extended to include a larger class of nonlinear systems by using NNs to approximate the drift dynamics. To focus the scope of this manuscript on switched systems, the drift dynamics are assumed to be linear-in-the-parameters.

Remark 1. For systems without finite escape behaviors and tuning of the initial values of $\hat{W}_{c,p}$ and $\hat{W}_{a,p}$, the availability of the history stack *a priori* is not necessary [19]. Assumption 6 is used to focus the scope of this manuscript and simplify the subsequent stability analysis.

Remark 2. To relax the common persistence of excitation (PE) condition [50, Def. 4.3.1], the update law in (12) uses a history stack comprised of recorded state and input data. Assumption 6, also known as the finite excitation (FE) condition, facilitates parameter convergence in the subsequent stability analysis. Assumption 6 requires excitation of the system states and is significantly less restrictive than the typical PE condition. The advantage of FE over PE is not due to which mechanism yields the excitation, but rather the interval of time concerned (finite versus persistent/infinite) and the verifiability of the condition: FE can be verified online and PE cannot generally be verified *a priori* or online for general nonlinear systems. Unlike the typical PE condition, which assumes that $\alpha_1 I \geq \frac{1}{T_0}\int_t^{t+T_0} \phi(\tau)\,\phi^\top(\tau)\,d\tau \geq \alpha_0 I\ \forall t \geq t_0$ over any time interval $[t, t+T_0]$, the FE condition uses data that is collected online the provide excitation [50, Def. 4.3.1].

With the parameter estimation error defined as $\tilde{\theta}_p \triangleq \theta_p - \hat{\theta}_p$, the update law in (12) can be rewritten in an analytical form as $\dot{\tilde{\theta}}_p = -k_{p,\theta}\Gamma_{p,\theta}\sum_{j=1}^{M_p}\mathcal{Y}_{p,j}^T\mathcal{Y}_{p,j}\tilde{\theta}_p$. If f_p is unknown, then the feedforward control component is approximated using $\hat{u}_{d,p} : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^m$, defined as $\hat{u}_{d,p}\left(x_d, \hat{\theta}\right) \triangleq g_p^+(x_d)\left(h_{d,p}(x_d) - \hat{f}_p\left(x, \hat{\theta}\right)\right)$. Hence, the applied control policy is given by

$$u \triangleq \hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right) + \hat{u}_{d,p}\left(x_d, \hat{\theta}_p\right). \tag{13}$$

IV. Bellman Error

The right-hand side of (6) is equal to zero under optimal conditions; however, substituting (10) and (11) into (6) results in a residual term $\hat{\delta}_p : \mathbb{R}^{2n} \times \mathbb{R}^s \times \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, which is referred to as the Bellman Error (BE), defined as

$$\begin{aligned}\hat{\delta}_p\left(\zeta, \hat{\theta}_p, \hat{W}_{c,p}, \hat{W}_{a,p}\right) &\triangleq \hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)^T R_p \hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right) \\ &+ \overline{Q}_p(\zeta) + \nabla_\zeta \hat{V}_p\left(\zeta, \hat{W}_{c,p}\right)\left(F_{\theta,p}\left(\zeta, \hat{\theta}_p\right)\right. \\ &\left. + F_{1,p}(\zeta) + G_p(\zeta)\,\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)\right),\end{aligned} \tag{14}$$

where

$$F_{\theta,p}\left(\zeta, \hat{\theta}_p\right) \triangleq \begin{bmatrix} \hat{f}_p\left(x, \hat{\theta}_p\right) - g_p(x)\,g_p^+(x_d)\,\hat{f}_p\left(x_d, \hat{\theta}_p\right) \\ \mathbf{0}_{n\times 1}\end{bmatrix}, \tag{15}$$

and

$$F_{1,p}(\zeta) \triangleq \begin{bmatrix} -h_{d,p}(x_d) + g_p(x)\,g_p^+(x_d)\,h_{d,p}(x_d) \\ h_{d,p}(x_d)\end{bmatrix}. \tag{16}$$

The BE is an indirect measure of the proximity of the actor and critic weight estimates to the ideal weights. By defining the mismatch between the estimates and their ideal values as $\tilde{W}_{c,p} \triangleq W_p - \hat{W}_{c,p}$ and $\tilde{W}_{a,p} \triangleq W_p - \hat{W}_{a,p}$, substituting (10) and (11) in (6), and subtracting from (14) yields the analytical form of the BE, which is used in the subsequent stability analysis,

$$\begin{aligned}\hat{\delta}_p\left(\zeta, \hat{\theta}_p, \hat{W}_{c,p}, \hat{W}_{a,p}\right) &= -W_p^T\nabla_\zeta\phi_p(\zeta) \\ &\cdot\left(F_{\theta,p}(\zeta, \theta_p) - F_{\theta,p}\left(\zeta, \hat{\theta}_p\right)\right) - \omega_p^T\tilde{W}_{c,p} \\ &+\frac{1}{4}\tilde{W}_{a,p}^T G_{\phi,p}\tilde{W}_{a,p} + O_p(\zeta),\end{aligned} \tag{17}$$

where $\omega_p : \mathbb{R}^{2n} \times \mathbb{R}^L \times \mathbb{R}^s \to \mathbb{R}^{2n}$ is defined as $\omega_p\left(\zeta, \hat{W}_{a,p}, \hat{\theta}_p\right) \triangleq \nabla_\zeta\phi_p(\zeta)\left(F_{p,\theta}\left(\zeta, \hat{\theta}_p\right) + F_{1,p}(\zeta) + G_p(\zeta)\,\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)\right)$, $O_p(\zeta) \triangleq \frac{1}{2}\nabla_\zeta\epsilon_p(\zeta)\,G_{R,p}\nabla_\zeta\phi_p(\zeta)^T W_p + \frac{1}{4}G_{\epsilon,p} - \nabla_\zeta\epsilon_p(\zeta)\,F_{p,\theta}(\zeta, \theta_p) - \nabla_\zeta\epsilon_p(\zeta)\,F_{1,p}(\zeta)$, $G_{R,p} = G_{R,p}(\zeta) \triangleq G_p(\zeta)\,R_p^{-1}G_p(\zeta)^T$, $G_{\phi,p} = G_{\phi,p}(\zeta) \triangleq \nabla_\zeta\phi_p(\zeta)\,G_{R,p}(\zeta)\,\nabla_\zeta\phi_p(\zeta)^T$, and $G_{\epsilon,p} = G_{\epsilon,p}(\zeta) \triangleq \nabla_\zeta\epsilon_p(\zeta)\,G_p(\zeta)\,\nabla_\zeta\epsilon_p(\zeta)^T$. Unlike BE definitions in typical tracking model-based ADP results [19], the definitions in (14)-(17) differ because the BE $\hat{\delta}_p$ is specific to each subsystem. The definitions in (14)-(17) are switched-system analogs of BE definitions in typical tracking model-based ADP results such as [19].

A. Bellman Error Extrapolation

In this section, the concept of BE extrapolation, developed for model-based reinforcement learning results (see [22]) is adapted to the switched ADP problem. At each time instant, the BE in (14) is calculated using the current system state, critic weight estimates, and actor weight estimates. A classical problem in learning-based control is exploration versus exploitation. Results such as [51] require an exploration signal to sufficiently explore the operating domain. However, no analytical methods exist to compute the appropriate exploration signal. Alternatively, results such as [22] evaluate the BE along the system trajectory and other desired points in the state space to avoid using an exploration signal. Specifically, the BE is computed at a user-specified number and location of off-trajectory points $\{\zeta_i : \zeta_i \in \Omega_p\}_{i=1}^{N_p}$, where $N_p \in \mathbb{N}$ denotes a user-specified number of points in the compact set Ω_p. The BE extrapolation data is represented by the tuple $(\Sigma_{c,p}, \Sigma_{a,p}, \Sigma_{\Gamma,p})$, defined as $\Sigma_{c,p} \triangleq \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}}{\rho_{i,p}}\hat{\delta}_{i,p}$, $\Sigma_{a,p} \triangleq \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{G_{\sigma i,p}^T\hat{W}_{a,p}\omega_{i,p}^T}{4\rho_{i,p}}$, $\Sigma_{\Gamma,p} \triangleq \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}\omega_{i,p}^T}{\rho_{i,p}}$, where $\hat{\delta}_{i,p} \triangleq \hat{\delta}_p\left(\zeta_i, \hat{\theta}_p, \hat{W}_{c,p}, \hat{W}_{a,p}\right)$, $\omega_{i,p} \triangleq \omega_p\left(\zeta_i, \hat{W}_{a,p}, \hat{\theta}_p\right)$, and $\rho_{i,p} = 1 + \nu_p\omega_{i,p}^T\Gamma_p\omega_{i,p}$, $\nu_p \in \mathbb{R}_{>0}$ is a user-defined gain, and $\Gamma_p : \mathbb{R} \to \mathbb{R}^{L\times L}$ is a time-varying least-squares gain matrix. Generally, each subsystem, p, has distinct sets of data, history stacks, gain values, and update laws.[7]

Assumption 7. A finite set of points $\{\zeta_{i,p}\}_{i=1}^{N_p} \subset \Omega_p$ exists along with known constants $\underline{c}_p$ such that $0 < \underline{c}_p \triangleq$

[7]Since each subsystem has respective data and parameters, they are treated as unique to that subsystem, which is similar to [52].

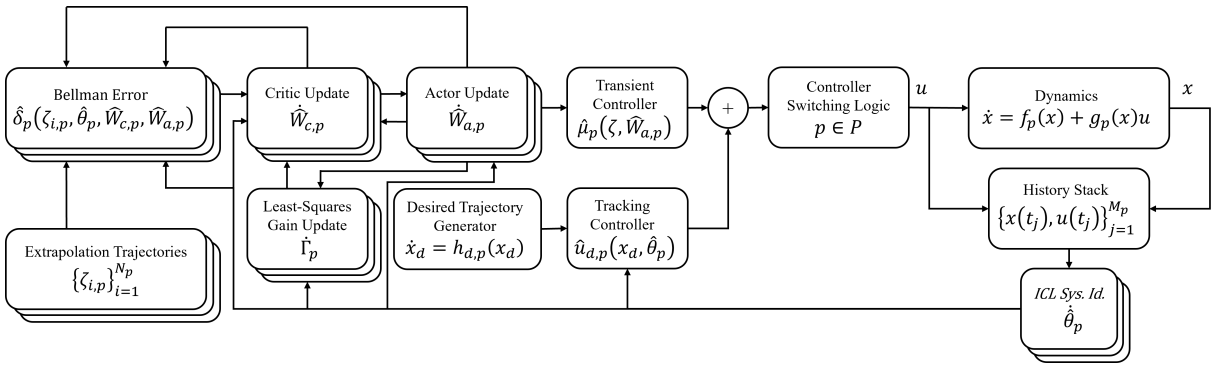


Figure 1. This diagram describes the overall switched actor-critic control architecture. The stacked boxes represent the subsystems that are simultaneously online (i.e., always active). The ICL-based system identification update policy $\dot{\hat{\theta}}_p$, BE extrapolation trajectories, BE evaluation, critic update policy $\dot{\hat{W}}_{c,p}$, least-squares gain update policy $\dot{\Gamma}_p$, and actor update policy $\dot{\hat{W}}_{a,p}$ are simultaneously active for all $p \in \mathcal{P}$.

$\inf_{t \in \mathbb{R}_{\geq 0}} \lambda_{\min}\{\Sigma_{\Gamma,p}\}$ for all $t \in \mathbb{R}_{\geq 0}$ and each $p \in \mathcal{P}$ [22, Assumption 3].

Remark 3. Assumption 7 states that BE extrapolation must provide a sufficiently exciting data. This assumption facilitates convergence of the weight approximation error term in the subsequent Lyapunov-based analysis. In practice, Assumption 7 is satisfied by selecting a large number of BE extrapolation trajectories (i.e., $N_p \gg L$ [22, Assumption 3]). Assumption 7 can also be verified online by evaluating $\lambda_{\min}\{\Sigma_{\Gamma,p}\}$.[8]

V. Update Laws for Actor and Critic Weights

Using the extrapolated BEs $\hat{\delta}_{i,p}$, the critic and actor weights are updated according to the following continuous-time update policies. These update policies are derived from the Lyapunov-based analysis in Section VII. In the following definitions, $\eta_{c,p}$, $\eta_{a1,p}$, $\eta_{a2,p}$, $\lambda_p \in \mathbb{R}_{>0}$ are user-selected constant learning gains, and $\underline{\Gamma}_p$, $\overline{\Gamma}_p \in \mathbb{R}_{>0}$ are upper and lower bounds of the least-squares learning gains of subsystem p. The critic update law of the p^{th} subsystem is

$$\dot{\hat{W}}_{c,p} \triangleq -\eta_{c,p}\Gamma_p\Sigma_{c,p}. \tag{18}$$

The actor update law of the p^{th} subsystem is

$$\dot{\hat{W}}_{a,p} \triangleq -\eta_{a1,p}\left(\hat{W}_{a,p} - \hat{W}_{c,p}\right) - \eta_{a2,p}\hat{W}_{a,p} + \eta_{c2,p}\Sigma_{a,p}\hat{W}_{c,p}. \tag{19}$$

The time-varying least-squares gain matrix update law [53, Sec. 8.7.4] of the p^{th} subsystem is

$$\dot{\Gamma}_p \triangleq \left(\lambda_p\Gamma_p - \eta_{c,p}\Gamma_p\Sigma_{\Gamma,p}\Gamma_p\right) \cdot \mathbf{1}_{\left\{\underline{\Gamma}_p \leq \|\Gamma_p\| \leq \overline{\Gamma}_p\right\}}, \tag{20}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, and $\overline{\Gamma}_p, \underline{\Gamma}_p \in \mathbb{R}_{>0}$ are user-selected saturation gains that bound $\|\Gamma_p\|$ such that $\underline{\Gamma}_p \leq \|\Gamma_p(t)\| \leq \overline{\Gamma}_p$ for all $t \in \mathbb{R}_{>0}$ and $p \in \mathcal{P}$.

VI. Simultaneous Online Learning

In [41], each subsystem's weights and least-squares gain matrix ($\hat{W}_{c,p}$, $\hat{W}_{a,p}$, and Γ_p) are updated strictly while that subsystem is active. For example, the $(p+1)^{\text{th}}$ subsystem's parameters are held constant as the p^{th} subsystem's parameters are updated. The previous approach introduces a problem in the switched systems analysis because the system states are not continuous between switching instances. For example, in [41] the active state changes from $\hat{W}_{c,p}$ to $\hat{W}_{c,p+1}$ as the p^{th} subsystem switches to the $(p+1)^{\text{th}}$ subsystem. To account for the state discontinuity, the weight update policies in [41] include smooth projection operators. Due to the projection operators, the bounds on the instantaneous changes in the values of the Lyapunov-like functions at the switching instances become overly conservative. As a result, the minimum dwell-time condition to prove stability is overly conservative for practical implementation.

In this paper, a dwell-time condition that is less conservative than [41] is obtained by keeping the parameter update policies in (12) and (18)-(20) simultaneously active for all $p \in \mathcal{P}$. Having each subsystem active simultaneously enables the creation of a concatenated state that contains $\tilde{W}_{c,p}$, $\tilde{W}_{a,p}$, $\tilde{\theta}_p$ $\forall p \in \mathcal{P}$. Hence, when the system switches from the p^{th} subsystem to the $(p+1)^{\text{th}}$ subsystem, the concatenated state is continuous at the switching instance.

The family of update laws in (18)-(20) is different from that of the typical model-based ADP [19]. The result in [19] includes on-trajectory BE data in the critic update law and an additional on-trajectory term in the actor update law to compensate for the on-trajectory BE data. Instead, the update laws in (18)-(20) omit the terms related to on-trajectory BE. Unlike the update laws in [41], the update laws in (18)-(20) only use model-based evaluation of the BE at user-

[8] Both BE extrapolation and CL-based system identification (see Assumption 6) are different techniques that relax the PE condition, which is required for parameter convergence. While the PE condition is traditionally studied within the context of system identification, ADP-based controllers require similarly exciting signals for convergence of the ADP controller to the optimal controller [13, Ch. 6]. CL system identification provides excitation from stored input-output data pairs, and BE extrapolation in ADP provides excitation from simulation of experience (i.e., simulating the learned system model at user-selection regions of the state space to simulate policy excitation).

selected points in the state space.[9] The omission of on-trajectory data is motivated by the need to have a continuous state between all subsystems. Similarly, the parameter update law in (12) is implemented using only the history stack $\{x(t_j), x(t_j - \Delta t), u(t_j)\}_{j=1}^{M_p}$ for each subsystem, which contains data that was recorded while that subsystem was active. As shown in the subsequent Lyapunov-based stability analysis, each subsystem is practically stable and the state trajectory of the switched system is continuous at the switching instances. The control system architecture, which leverages simultaneous online learning, is detailed in Figure 1.

VII. Stability Analysis

First, the ADP-based controllers in each subsystem are analyzed using a Lyapunov-based approach. In comparison to the analysis in [41, Thm. 1], the state vector in the following analysis includes parameter estimates from every subsystem. The resulting dwell-time analysis of the overall switched system is less restrictive that the result in [41, Thm. 1].

A. Subsystem Analysis

To motivate the overall switched systems stability analysis in Section VII-B, the stability of the system while p^{th} subsystem is active must first be analyzed. The development in this subsection does not address switching from the p^{th} to the $(p+1)^{\text{th}}$ subsystem, as switching between a family of stable subsystems may not result in a stable switched system [1].

Since the function $\overline{Q}$ and, therefore, the optimal value function V_p^* in (8) is PSD, V_p^* is not a valid candidate Lyapunov function. It is shown in [55] that a nonautonomous form of V_p^*, denoted as $V_{na,p}^* : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ and defined as $V_{na,p}^*(e,t) \triangleq V_p^*(\zeta)$, is positive definite and decrescent. Hence, $V_{na,p}^*(0,t) = 0$ and there exist class $\mathcal{K}_\infty$ functions $\underline{v}, \overline{v} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that bound $\underline{v}_p(\|e\|) \leq V_{na,p}^*(e,t) \leq \overline{v}_p(\|e\|)$, for all $e \in \mathbb{R}^n$ and $t \in \mathbb{R}_{\geq 0}$. Hence, $V_{na,p}^*(e,t)$ is a valid candidate Lyapunov function.

Let $Z \in \mathbb{R}^{n+|\mathcal{P}|(2L+s)}$ denote a concatenated state defined as $Z \triangleq \left[e^T, \tilde{W}_{c,1}^T, \ldots, \tilde{W}_{c,k}^T, \tilde{W}_{a,1}^T, \ldots, \tilde{W}_{a,k}^T, \tilde{\theta}_1^T, \ldots, \tilde{\theta}_k^T\right]^T$.[10] Let $V_{L,p} : Z \in \mathbb{R}^{n+|\mathcal{P}|(2L+s)} \times \mathbb{R}_{\geq_0} \to \mathbb{R}$ be a candidate Lyapunov function defined as

$$V_{L,p}(Z,t) \triangleq V_{na,p}^*(e,t) + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{W}_{c,p}^T \Gamma_p(t)^{-1} \tilde{W}_{c,p} + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{W}_{a,p}^T \tilde{W}_{a,p} + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{\theta}_p^T \Gamma_{\theta,p}^{-1} \tilde{\theta}_p. \quad (21)$$

[9] Depending on the state dimension n, the dimension of the basis functions L, and number of BE extrapolation points N_p, it may be computationally expensive or intractable to compute (18)-(20) in parallel for each subsystem in real-time. Sparse NN BE extrapolation methods in [26] and [54] can be leveraged to reduce the computational cost associated with (18)-(20) for each subsystem in parallel.

[10] The inclusion of $\tilde{\theta}_p$ terms in (21) complicate the Lyapunov-based analysis, cf. [41, Thm. 1]. Recall from Section IV-A that $\hat{\delta}_{i,p} \triangleq \hat{\delta}_p\left(\zeta_i, \hat{\theta}_p, \hat{W}_{c,p}, \hat{W}_{a,p}\right)$, which includes the parameter estimate $\hat{\theta}_p$. The coupling between the actor-critic and system identification update laws motivates their respective designs and inclusion in (21).

Using the properties of $V_{na,p}^*(e,t)$ and the fact that Γ_p is bounded, (21) can be bounded as $\alpha_{1,p}(\|Z\|) \leq V_{L,p}(Z,t) \leq \alpha_{2,p}(\|Z\|)$ using class $\mathcal{K}_\infty$ functions $\alpha_{1,p}, \alpha_{2,p} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Using (20), the normalized regressors $\frac{\omega_p}{\rho_p}$ and $\frac{\omega_{e,p}}{\rho_{e,p}}$ can be bounded as $\sup_{t\in\mathbb{R}_{\geq 0}} \left\|\frac{\omega_p}{\rho_p}\right\| \leq \frac{1}{2\sqrt{\nu_p \underline{\Gamma}_p}}$ and $\sup_{t\in\mathbb{R}_{\geq 0}} \left\|\frac{\omega_{e,p}}{\rho_{e,p}}\right\| \leq \frac{1}{2\sqrt{\nu_p \underline{\Gamma}_p}}$. The matrices $G_{R,p}$ and $G_{\sigma,p}$ can be bounded as $\sup_{\zeta_p\in\Omega_p} \|G_R\| \leq \lambda_{\max}\left(R_p^{-1}\right)\overline{G_p}^2 \triangleq \overline{G_{R,p}}$ and $\sup_{\zeta_p\in\Omega_p} \|G_{\sigma,p}\| \leq \left(\overline{\nabla_\zeta \phi_p}\,\overline{G_p}\right)^2 \lambda_{\max}\left(R_p^{-1}\right) \triangleq \overline{G_{\sigma,p}}$, respectively. These facts are leveraged in the subsequent Lyapunov-based analysis in Theorem 1. The following theorem provides sufficient conditions for the ADP controller with continuous-time update policies in (18)-(20) to ensure that the closed-loop subsystems are locally practically stable, uniformly in t_0, in the sense of the definition below.

Definition 1. A system $\dot{x} = f(x,t)$ is locally practically stable, uniformly over $t_0 \in \mathbb{R}_{\geq 0}$, if there exist constants $0 \leq \nu < r$ and $\beta \in \mathcal{KL}$ such that for all $t_0 \geq 0$ and $\|x(t_0)\| \leq r$, the trajectory $x(\cdot)$ of the system starting from (t_0, x_0) satisfies $\|x(t)\| \leq \beta(\|x_0\|, t - t_0) + \nu$ (see [45, Def. 2.2]).

Subsequently, Theorem 2 provides a minimum dwell time condition to ensure that the switched system is also locally practically stable.

Theorem 1. *Provided Assumptions 1-7 hold, the weight update laws in (18)-(20) are used, and the conditions*

$$\eta_{a1,p} + \eta_{a2,p} > \frac{\eta_{c1,p} + \eta_{c2,p}}{\sqrt{\nu_p \underline{\Gamma}_p}} \overline{W_p^* G_{\phi,p}}, \quad (22)$$

$$\underline{c}_p > \frac{3\eta_{c,p}^2 \overline{W_p^*}^2 \overline{G_{\phi,p}}^2}{16\nu_p \underline{\Gamma}_p (\eta_{a1,p} + \eta_{a2,p})\eta_{c2,p}} + \frac{3\eta_{a1,p}}{\eta_{c,p}}, \quad (23)$$

$$L_p < \alpha_{2,p}^{-1}(\alpha_{1,p}(\mathcal{R}_p)), \quad (24)$$

are satisfied[11] *for all $p \in \mathcal{P}$, where $\mathcal{R}_p$ is the radius of a ball contained in Ω_p and L_p is a positive constant introduced below, then the closed loop system defined by (1), (12), (18), and (19), with state Z, is locally practically stable, uniformly over $t_0 \in \mathbb{R}_{\geq 0}$.*

Proof: Taking the time derivative of the Lyapunov-like function in (21), the fact $\frac{d}{dt}\Gamma^{-1} = \Gamma^{-1}\dot{\Gamma}\Gamma^{-1}$, along with Assumptions 1-7 and the sufficient conditions in (22)-(24) yields $\dot{V}_{L,p} \leq -v_{L,p}(\|Z\|)$, $\forall v_{L,p}^{-1}(L_p) \leq \|Z\| \leq \alpha_{2,p}^{-1}(\alpha_{1,p}(\mathcal{R}_p))$, where

$$v_{L,p}(\|Z\|) \triangleq \frac{1}{2}\underline{q}_p(\|e\|) + \sum_{p=1}^{|\mathcal{P}|}\left[\frac{\eta_{c,p}\underline{c}_p}{12}\left\|\tilde{W}_{c,p}\right\|^2 + \frac{\eta_{a1,p} + \eta_{a2,p}}{20}\left\|\tilde{W}_{a,p}\right\|^2 + \frac{k_{ICL,p}\underline{\mathcal{Y}}_p}{6}\left\|\tilde{\theta}_p\right\|^2\right], \quad (25)$$

[11] See [22] for insight into satisfying the condition in (22)-(24).

and L_p is a positive constant. While each individual subsystem is active, [47, Thm. 4.18] can be invoked to infer the existence of a class $\mathcal{KL}$ function β_p such that for all $t_0 \in \mathbb{R}_{\geq 0}$, if $\|Z(t_0)\| \leq \alpha_{2,p}^{-1}(\alpha_{1,p}(\mathcal{R}_p))$ then $\|Z(t)\| \leq \max\left\{\beta_p(\|Z(t_0)\|, t-t_0), \alpha_{1,p}^{-1}\left(\alpha_{2,p}\left(v_{L,p}^{-1}(L_p)\right)\right)\right\}$ and that the subsystem trajectories are UUB. Using Definition 1, the subsystem is also locally practically stable. Furthermore, $\hat{\mu}_p$ converges to a neighborhood of the optimal policy μ_p^*. Furthermore, since $Z \in \mathcal{L}_\infty$, it follows that e, $\tilde{W}_{c,1} \ldots, \tilde{W}_{c,|\mathcal{P}|}$, $\tilde{W}_{a,1}, \ldots, \tilde{W}_{a,|\mathcal{P}|}$, $\tilde{\theta}_1, \ldots, \tilde{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$, hence x, $\hat{W}_{c,1}, \ldots, \hat{W}_{c,|\mathcal{P}|}$, $\hat{W}_{a,1}, \ldots, \hat{W}_{a,|\mathcal{P}|}$, $\hat{\theta}_1, \ldots, \hat{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$ and $u \in \mathcal{L}_\infty$. ■

Remark 4. Under Assumptions 1-7, the optimal value function can be shown to be the unique positive definite solution of the HJB equation. Convergence to the positive definite solution of the HJB equation is guaranteed by appropriately selecting initial weight estimate values [56].

Remark 5. The result in Theorem 1 improves upon the result in [41, Thm. 1] by relaxing the quadratic bounds on (21), addressing the trajectory tracking problem, including an online system identification term, performing simultaneous online learning of all subsystems simultaneously, and providing a stability result of a state that is common between all subsystems Z (cf. subsystem-specific states in [41, Thm. 1]).

In addition to establishing that the subsystem trajectories are UUB, Theorem 1 establishes local practical stability of the individual subsystems. Switching between a family of such subsystems may not result in an overall locally practically stable switched system [1]. The candidate Lyapunov-like function for the p^{th} subsystem in (21) contains the optimal value function V_p^*, which is generally unique to each subsystem; hence, the switched system does not generally admit a common Lyapunov-like function. As a result, the use of multiple Lyapunov-like functions is motivated. To the best of the authors' knowledge, a general result that provides conditions under which local practical stability of the switched system may be inferred from local practical stability of individual subsystems is not available in the literature. The development of such a result, in the context of general nonlinear nonautonomous systems, is the focus of the following section.

B. Switching between locally practically stable subsystems

Consider a family of finitely many nonlinear subsystems of the form

$$\dot{x} = f_p(x,t),\ p \in \mathcal{P}, \tag{26}$$

where the functions $f_p : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ are locally Lipschitz continuous in x, for all t, and piecewise continuous in t, for all x. Given a piecewise constant right-continuous (i.e., $\sigma(t) = \lim_{\tau \downarrow t^+} \sigma(\tau)$) switching signal $\sigma : \mathbb{R}_{\geq t_0} \to \mathcal{P}$ where $\sigma(t)$ indicates the active subsystem at time t, the corresponding switched system can be expressed as

$$\dot{x} = f_{\sigma(t)}(x,t). \tag{27}$$

The objective of the following theoretical development is to provide sufficient conditions such that the switched system is locally practically stable, uniformly over the initial time and over a suitable set Σ of switching signals, as defined below.

Definition 2. A switched system $\dot{x} = f_{\sigma(t)}(x,t)$ is locally practically stable, uniformly over $t_0 \in \mathbb{R}_{\geq 0}$ and $\sigma \in \Sigma$, if there exist constants $0 \leq \nu < r$ and $\beta \in \mathcal{KL}$ such that for all $t_0 \geq 0$, $\sigma \in \Sigma$, and $\|x_0\| \leq r$, the trajectory $x(\cdot)$ of the system starting from (t_0, x_0) satisfies $\|x(t)\| \leq \beta(\|x_0\|, t-t_0) + \nu$ (see [45, Def. 2.2]).

In this paper, the sufficient conditions are derived using a minimum dwell-time approach [1, Ch. 3.2.1].

Definition 3. Given a switching signal σ and the corresponding sequence of switching times $t_\sigma \triangleq \{t_1, \ldots, t_i, t_j, \ldots\}$, the dwell time $\tau \in \mathbb{R}_{>0}$ is defined as the time between switching instances. Specifically, $\tau(t_i, t_j) \triangleq t_j - t_i$ such that $\sigma(t_i) \neq \sigma(t_j)$ [57].

Another objective of the analysis is to infer the size of the ultimate bound of the trajectories of the switched system from the ultimate bounds of the trajectories of the subsystems.

The subsequent stability analysis relies on multiple Lyapunov-like functions V_p, where each V_p, for $p \in \mathcal{P}$, establishes local practical stability of the p^{th} subsystem, uniformly in t_0. Multiple Lyapunov-like functions are a tool used for proving stability of switched systems [1, Sec. 3.1]; for the subsequent stability analysis, each subsystem has a respective Lyapunov-like function that is used to determine the behavior of that system while active. While each V_p is continuous, the function $t \mapsto V_{\sigma(t)}(x(t))$, evaluated along the trajectories of the switched system in (27) is generally discontinuous (i.e., V_p may instantaneously change its value at the switching instances). Furthermore, while the p^{th} system is active (i.e., $\sigma(t) = p$), the corresponding V_p, evaluated along the trajectories of the switched system in (27), decreases or is bounded within an ultimate bound. However, the functions V_q, corresponding to all inactive subsystems may increase when evaluated along the trajectories of the switched system in (27) (see [1, Ch. 3.1]).

The following theorem provides sufficient conditions on the switching signal and the initial conditions to ensure that the Lyapunov-like functions corresponding to all subsystems decrease to an ultimate bound. Furthermore, if the sufficient conditions are satisfied, then Theorem 2 can also be used to show that each $\hat{\mu}_p$ converges to a neighborhood of the respective optimal policy μ_p^* for all $p \in \mathcal{P}$.

Theorem 2. *If $D \subset \mathbb{R}^n$ is an open and connected set containing the origin, $r > 0$ is such that $B_r \triangleq \{x \in \mathbb{R}^n \mid \|x\| \leq r\} \subset D$, $\dot{x} = f_p(x,t)$ is a finite family of subsystems, there exist continuously differentiable functions $V_p : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, continuous PD functions $W_p : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, class $\mathcal{K}$ functions $\alpha_{1,p} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and class $\mathcal{K}_\infty$ functions $\alpha_{2,p} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, such that*

$$\alpha_{1,p}(\|x\|) \leq V_p(x,t) \leq \alpha_{2,p}(\|x\|), \tag{28}$$

for all $(p,x,t) \in \mathcal{P} \times D \times \mathbb{R}_{\geq 0}$,

$$\frac{\partial V_p}{\partial t} + \frac{\partial V_p}{\partial x} f_p(x,t) \leq -W_p(x), \tag{29}$$

for all $(p,x,t) \in \mathcal{P} \times \Lambda_p \times \mathbb{R}_{\geq 0}$, *and*

$$\max_{p,q \in \mathcal{P}} \left\{ \alpha_{2,q}\left(\alpha_{1,p}^{-1}\left(\alpha_{2,p}(\nu_p)\right)\right)\right\} \leq \min_{p \in \mathcal{P}} \left\{\alpha_{1,p}(r)\right\}, \tag{30}$$

where $\Lambda_p \triangleq \{x \,|\, 0 \leq \nu_p \leq \|x\| \leq r\}$, $\alpha_{1,p}^{-1} : \text{range}(\alpha_{1,p}) \to \mathbb{R}_{\geq 0}$ *is the inverse of* $\alpha_{1,p}$, *and for a given* $\lambda \in (0,1)$, *if* Σ_λ *is the set of all switching signals* σ *such that the resulting sequence of switching times* $t_\sigma = \{t_1, t_2, \ldots\}$ *satisfies the minimum dwell-time condition*

$$\tau(t_{i-1}, t_i) > \begin{cases} 0 & \|x(t_i)\| \leq \nu_i \\ \tau_\sigma(t_i) & \textit{otherwise} \end{cases} \tag{31}$$

for all switching instants $t_i \in t_\sigma$, *where*

$$\tau_\sigma(t_i) \triangleq \max\Bigg\{0, \frac{\begin{pmatrix}\alpha_{2,\sigma(t_{i-1})}\left(\|x(t_{i-1})\|\right) \\ -\alpha_{1,\sigma(t_{i-1})}\left(\alpha_{2,\sigma(t_i)}^{-1}\left(\lambda\alpha_{1,\sigma(t_{i-1})}\left(\|x(t_{i-1})\|\right)\right)\right)\end{pmatrix}}{\kappa_{\sigma(t_{i-1})}}, \frac{\begin{pmatrix}\alpha_{2,\sigma(t_{i-1})}\left(\|x(t_{i-1})\|\right) \\ -\alpha_{1,\sigma(t_{i-1})}\left(\alpha_{2,\sigma(t_i)}^{-1}\left(\alpha_{1,\sigma(t_i)}(r)\right)\right)\end{pmatrix}}{\kappa_{\sigma(t_{i-1})}}\Bigg\}, \tag{32}$$

then the switched system $\dot{x} = f_{\sigma(t)}(x,t)$ *is locally practically stable, uniformly over* $t_0 \in \mathbb{R}_{\geq 0}$ *and* $\sigma \in \Sigma_\lambda$. *In particular, the trajectories of* $\dot{x} = f_{\sigma(t)}(x,t)$, *with*

$$\max_{p \in \mathcal{P}} \left\{\alpha_{2,p}\left(\|x(t_0)\|\right)\right\} \leq \min_{q \in \mathcal{P}} \left\{\alpha_{1,q}(r)\right\} \tag{33}$$

satisfy

$$\limsup_{t \to \infty} \|x(t)\| \leq \max_{p \in \mathcal{P}} \left\{ \alpha_{1,p}^{-1}\left(\max_{q,s \in \mathcal{P}} \left\{\alpha_{2,s}\left(\alpha_{1,q}^{-1}\left(\alpha_{2,q}(\nu_q)\right)\right)\right\}\right)\right\}. \tag{34}$$

Proof: The proof relies on the observation that if (30) holds, then for any $p \in \mathcal{P}$, if $\alpha_{2,p}(\nu_p) < V_p(x,t) \leq \alpha_{1,p}(r)$ then $x \in \Lambda_p$, and from (29), $\dot{V}_p(x,t) < 0$. As a result, the $\alpha_{2,p}(\nu_p)-$ and $\alpha_{1,p}(r)-$ sublevel sets of V_p are forward invariant whenever the p^{th} subsystem is active.

Let $o,p,q \in \mathcal{P}$ represent the first three active subsystems i.e., $\sigma(t_0) = o$, $\sigma(t_1) = p$, and $\sigma(t_2) = q$. From (29), whenever $x \in \Lambda_o$, then $\dot{V}_o(x,t) \leq -W_o(x) \leq -\kappa_o$, where generally $\kappa_o = \min_{x \in \Lambda_o} W_o(x) > 0$. Using forward invariance of $\alpha_{1,o}(r)-$ and $\alpha_{2,o}(\nu_o)-$sublevel sets of V_o and (33), it can be concluded that $\alpha_{2,o}(\|x(t_0)\|) \leq \alpha_{1,o}(r)$, which implies that $V_o(x(t_0),t_0) \leq \alpha_{1,o}(r)$, and as a result

$$V_o(x(t),t) \leq \max\left\{V_o(x(t_0),t_0) - \kappa_o(t-t_0), \alpha_{2,o}(\nu_o)\right\} \tag{35}$$

for all $t \in [t_0, t_1]$. From (30), ν_o satisfies $\alpha_{2,p}\left(\alpha_{1,o}^{-1}\left(\alpha_{2,o}(\nu_o)\right)\right) \leq \alpha_{1,p}(r)$. From (30) and (31), either $x(t_1)$ and ν_p satisfy $\|x(t_1)\| \leq \nu_p$ and $\alpha_{2,p}(\nu_p) \leq \alpha_{1,p}(r)$ or the dwell-time satisfies $\alpha_{2,o}(\|x(t_0)\|) - \kappa_o(t_1 - t_0) \leq \alpha_{1,o}\left(\alpha_{2,p}^{-1}\left(\alpha_{1,p}(r)\right)\right)$. In either case, $\alpha_{2,p}(\|x(t_1)\|) \leq \alpha_{1,p}(r)$, which implies that $V_p(x(t_1),t_1) \leq \alpha_{1,p}(r)$, and as a result

$$V_p(x(t),t) \leq \max\left\{V_p(x(t_1),t_1) - \kappa_p(t-t_1), \alpha_{2,p}(\nu_p)\right\} \tag{36}$$

for all $t \in [t_1, t_2]$. Similarly, from (30) and (31), either $x(t_1)$ and ν_p satisfy $\|x(t_1)\| \leq \nu_p$ and $\alpha_{2,p}(\nu_p) \leq \alpha_{1,p}(r)$ or the dwell-time $\tau(t_0,t_1)$ satisfies $\alpha_{2,o}(\|x(t_0)\|) - \kappa_o(t_1 - t_0) \leq \alpha_{1,o}\left(\alpha_{2,p}^{-1}\left(\lambda\alpha_{1,o}(\|x(t_0)\|)\right)\right)$ for some $\lambda \in (0,1)$. In either case, (35) implies that

$$V_p(x(t_1),t_1) \leq \max\left\{\lambda V_o(x(t_0),t_0), \alpha_{2,p}\left(\alpha_{1,o}^{-1}\left(\alpha_{2,o}(\nu_o)\right)\right)\right\}, \tag{37}$$

Let $\tilde{\alpha} \triangleq \max_{p,q \in \mathcal{P}} \alpha_{2,p}\left(\alpha_{1,q}^{-1}\left(\alpha_{2,q}(\nu_q)\right)\right)$. Observe that if $V_o(x(t_o),t_o) \leq \tilde{\alpha}$, then from (35) $V_o(x(t),t) \leq \tilde{\alpha}$ for all $t \in [t_0,t_1)$, and from (37), $V_p(x(t_1),t_1) \leq \tilde{\alpha}$. An inductive argument then shows that $V_{\sigma(t_i)}(x(t_i),t_i) \leq \tilde{\alpha}$ for some $t_i \in t_\sigma$ implies that $V_{\sigma(t)}(x(t),t) \leq \tilde{\alpha}$ for all $t \geq t_i$.

Furthermore, given (33), for all switching instances $t_i \in t_\sigma$, the initial condition satisfies

$$\alpha_{2,\sigma(t_0)}\left(\|x(t_0)\|\right) \leq \alpha_{1,\sigma(t_0)}(r), \tag{38}$$

given (30), the residuals $\nu_{\sigma(\cdot)}$ satisfy

$$\alpha_{2,\sigma(t_i)}\left(\alpha_{1,\sigma(t_{i-1})}^{-1}\left(\alpha_{2,\sigma(t_{i-1})}\left(\nu_{\sigma(t_{i-1})}\right)\right)\right) \leq \alpha_{1,\sigma(t_i)}(r), \tag{39}$$

and given (31), either the dwell-time must satisfy

$$t_i - t_{i-1} \geq \tau_\sigma(t_i) \tag{40}$$

or the state at the time of switch must satisfy $\|x(t_i)\| \leq \nu_{\sigma(t_i)}$. As a result,

$$V_{\sigma(t_i)}(x(t_i),t_i) \leq \max\left\{\lambda V_{\sigma(t_{i-1})}(x(t_{i-1}),t_{i-1}), \tilde{\alpha}\right\}, \tag{41}$$

and for all $t \in [t_i, t_{i+1})$,

$$V_{\sigma(t)}(x(t),t) \leq \max\left\{V_{\sigma(t_i)}(x(t_i),t_i) - \kappa_{\sigma(t_i)}(t-t_i), \alpha_{2,\sigma(t_i)}\left(\nu_{\sigma(t_i)}\right)\right\}. \tag{42}$$

Let $i_\sigma \triangleq \min\left\{i \geq 0 \mid V_{\sigma(t_{i+1})}(x(t_{i+1}),t_{i+1}) \leq \tilde{\alpha}\right\}$ denote the number of switches for which V_σ remains larger than $\tilde{\alpha}$, let $t_{i_\sigma} \in t_\sigma$ denote the corresponding switching time, and let $N_\sigma(t) = \max\{i \mid 0 \leq i \leq i_\sigma \wedge t_i \leq t\}$ denote the number of switches up to and including $t \leq t_{i_\sigma}$. Since (30) implies that

$$\alpha_{2,p}\left(\alpha_{1,o}^{-1}\left(\alpha_{2,o}(\nu_o)\right)\right) \leq \alpha_{1,q}(r), \forall p,q,o \in \mathcal{P}, \tag{43}$$

(41) and (42) can be combined to conclude that for all $t \geq t_0$,

$$V_{\sigma(t)}(x(t),t) \leq \max\left\{\begin{pmatrix}\lambda^{N_\sigma(t)} V_{\sigma(t_0)}(x(t_0),t_0) \\ -\kappa_{\sigma(t)}\left(t - t_{N_\sigma(t)}\right)\end{pmatrix}, \tilde{\alpha}\right\}. \tag{44}$$

The bound in (44) can be used to establish local practical stability of the switched system for a given fixed switching signal σ. The purpose of the following arguments is to compute a decay bound on $V_{\sigma(t)}$ that holds for all $\sigma \in \Sigma_\lambda$. For brevity of notation, let $V_0 \triangleq V_{\sigma(t_0)}(x(t_0), t_0)$. If the initial condition satisfies (33), then $V_0 \leq r^* \triangleq \min_p \alpha_{1,p}(r)$. Under the dwell-time restriction, (44) results in $V_{\sigma(t_{i+1})}(x(t_{i+1}), t_{i+1}) \leq \max\left\{\lambda^{i+1} V_0, \tilde{\alpha}\right\}$. Therefore, the number of possible switches over the interval $[t_0, t_{i_\sigma}]$ is bounded, uniformly over $\sigma \in \Sigma_\lambda$ by i^*. That is, for all $\sigma \in \Sigma_\lambda$, $i_\sigma \leq \min\left\{i \mid \lambda^{i+1} V_0 \leq \tilde{\alpha}\right\} \leq i^*$, where

$$i^* \triangleq \min\left\{i \mid \lambda^{i+1} r^* \leq \tilde{\alpha}\right\}. \tag{45}$$

Similarly, over the interval $[t_0, t_{i_\sigma}]$, the time between any two switches can also shown to be bounded, uniformly over $\sigma \in \Sigma_\lambda$. Indeed, since $\alpha_{2,p}(\nu_p)-$sublevel sets are invariant whenever the p^{th} subsystem is active, $t \in [t_0, t_{i_\sigma}]$ implies that $V_{\sigma(t)}(x(t), t) > \alpha_{2,\sigma(t)}\left(\nu_{\sigma(t)}\right)$. Therefore, if t_i and t_{i+1} are two switching instances in $[t_0, t_{i_\sigma}]$, then $\lambda^i V_0 - \kappa_{\sigma(t_i)}(t_{i+1} - t_i) \geq \alpha_{2,\sigma(t_i)}\left(\nu_{\sigma(t_i)}\right)$, which results in the bound $t_{i+1} - t_i \leq \frac{V_0 - \min_p\{\alpha_{2,p}(\nu_p)\}}{\kappa} \leq \tau^*$ for all $\sigma \in \Sigma_\lambda$, where $\kappa \triangleq \min_{p\in\mathcal{P}}\{\kappa_p\} = \min_{p\in\mathcal{P}}\left\{\min_{x\in\Lambda_p}\{W_p(x)\}\right\}$ and

$$\tau^* \triangleq \frac{r^* - \min_p\{\alpha_{2,p}(\nu_p)\}}{\kappa}. \tag{46}$$

As a result, the last switching time for which V_σ remains larger than $\tilde{\alpha}$ also admits a bound that is uniform over $\sigma \in \Sigma_\lambda$. Specifically, $t_{i_\sigma} \leq t^* \triangleq \tau^* i^*$.

Consider a closed interval $\mathcal{I}_i \triangleq [t_0 + i\tau^*, t_0 + (i+1)\tau^*]$, for $0 \leq i \leq i^*$. Note that since there are at least i switches over $[t_0, t_0 + i\tau^*]$, at the start of the interval, V_σ satisfies $V_{\sigma(t_0+i\tau^*)}(x(t_0 + i\tau^*), t_0 + i\tau^*) \leq \lambda^i V_0$. If there are $j \geq 1$ switches over $\mathcal{I}_i$, then at the end of the interval, V_σ satisfies $V_{\sigma(t_0+(i+1)\tau^*)}(x(t_0 + (i+1)\tau^*), t_0 + (i+1)\tau^*) \leq \lambda^{i+j} V_0 \leq \lambda^{i+1} V_0$. Furthermore, V_σ is bounded by an affine function with slope $-\kappa$ between switches. As a result, on $\mathcal{I}_i$, V_σ satisfies an affine decay bound with slope $-\min\left\{\kappa, \frac{V_0 \lambda^i (1-\lambda)}{\tau^*}\right\}$.

In particular, letting $N^*(t) \triangleq \left\lfloor \frac{t}{\tau^*} \right\rfloor$, $\kappa^*(s,t) \triangleq \min\left\{\kappa, \frac{s\lambda^{N^*(t)}(1-\lambda)}{\tau^*}\right\}$, and

$$\beta^*(s,t) \triangleq \begin{cases} s\lambda^{N^*(t)} - \kappa^*(s,t)\left(t - t_{N^*(t)}\right) & t < t^* \\ s\lambda^{i^*} \mathrm{e}^{\frac{-\kappa^*(s,t)}{s\lambda^{i^*}}(t-t^*)} & t \geq t^* \end{cases}, \tag{47}$$

it can be concluded that for all $\sigma \in \Sigma_\lambda$, V_σ satisfies the bound $V_{\sigma(t)}(x(t), t) \leq \max\{\beta^*(\max_{p\in\mathcal{P}}\{V_p(x(t_0), t_0)\}, t), \tilde{\alpha}\}$. As a result, for all $\sigma \in \Sigma_\lambda$, the system state is bounded by[12]

$$\|x(t)\| \leq \max\left\{\max_{q\in\mathcal{P}}\left\{\alpha_{1,q}^{-1}(\tilde{\alpha})\right\}, \max_{q\in\mathcal{P}}\left\{\alpha_{1,q}^{-1}\left(\beta^*\left(\max_{p\in\mathcal{P}}\{\alpha_{2,p}(\|x(t_0)\|)\}, t\right)\right)\right\}\right\}. \tag{48}$$

Since $\beta^* \in \mathcal{KL}$, $\max_{q\in\mathcal{P}}\left\{\alpha_{1,q}^{-1}\right\} \in \mathcal{K}$, and $\max_{p\in\mathcal{P}}\{\alpha_{2,p}\} \in \mathcal{K}$ [47, Lemma 4.2] can be invoked to conclude that $(s,t) \mapsto \max_{q\in\mathcal{P}}\left\{\alpha_{1,p}^{-1}(\beta^*(\max_{p\in\mathcal{P}}\{\alpha_{2,p}(s)\}, t))\right\} \in \mathcal{KL}$ and local practical stability of the closed loop system is established, uniformly over $t_0 \in \mathbb{R}_{\geq 0}$ and $\sigma \in \Sigma_\lambda$. ■

C. Application to Switched ADP

From Theorem 1, every individual subsystem is locally practically stable, uniformly over $t_0 \in \mathbb{R}^n$; i.e., each subsystem satisfies (28)-(30). Hence, Theorem 2 can be invoked to show that provided the concatenated state is initialized in the set defined by (33) with $r = \min_p\{\mathcal{R}_p\}$, then the closed-loop switched system is locally practically stable, uniformly over $t_0 \in \mathbb{R}^n$ and over switching signals that satisfy $\sigma \in \Sigma_\lambda$. In particular, the resulting trajectory of the closed-loop switched system satisfies the ultimate bound in (34) with $\nu_q = v_{L,q}^{-1}(L_q)$. Furthermore, each $\hat{\mu}_p$ converges to a neighborhood of the respective optimal policy μ_p^* for all $p \in \mathcal{P}$. Furthermore, since $Z \in \mathcal{L}_\infty$, it follows that $e, \tilde{W}_{c,1}, \ldots, \tilde{W}_{c,|\mathcal{P}|}, \tilde{W}_{a,1}, \ldots, \tilde{W}_{a,|\mathcal{P}|}, \tilde{\theta}_1, \ldots, \tilde{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$; hence $e, \hat{W}_{c,1}, \ldots, \hat{W}_{c,|\mathcal{P}|}, \hat{W}_{a,1}, \ldots, \hat{W}_{a,|\mathcal{P}|}, \hat{\theta}_1, \ldots, \hat{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$ and $u \in \mathcal{L}_\infty$.

VIII. Simulation

The developed switched ADP controller is applied to a fully actuated autonomous undersea vehicle (AUV), to complete an Earth-fixed position tracking objective. Specifically, the simulation is based on the SubjuGator AUV detailed in [23] and [58]. To focus the scope of this simulation section, it is assumed that the AUV is neutrally buoyant if submerged, the center of gravity is located vertically below the center of buoyancy on the z-axis, and the vehicle model accounts for small roll and pitch angles. The nonlinear equations of motion for an AUV under the effects of an irrotational current are given in [59, Sec. 7.5] as

$$\dot{\eta}_{AUV} = J_E(\eta_{AUV})\,\nu_{AUV} \tag{49}$$

$$M_{RB}\dot{\nu}_{AUV} + C_{RB}(\nu_{AUV})\,\nu_{AUV} + M_A\dot{\nu}_r + C_A(\nu_r)\,\nu_r + D_A(\nu_r)\,\nu_r + G(\eta_{AUV}) = \tau_b, \tag{50}$$

where $\nu_{AUV} \in \mathbb{R}^3$ is the body-fixed translational and angular velocity vector, $\nu_c \in \mathbb{R}^3$ is the body-fixed irrotational current velocity vector, $\nu_r \triangleq \nu_{AUV} - \nu_c$ is the relative body-fixed translational and angular fluid velocity vector, $\eta_{AUV} \in \mathbb{R}^3$ is the Earth-fixed position and orientation vector, $J_E : \mathbb{R}^3 \to \mathbb{R}^{3\times3}$ is the coordinate transformation between the body-fixed and Earth-fixed coordinates, $M_{RB} \in \mathbb{R}^{3\times3}$ is the constant rigid body inertia matrix, $C_{RB} : \mathbb{R}^3 \to \mathbb{R}^{3\times3}$ is the rigid body centripetal and Coriolis matrix, $M_A \in \mathbb{R}^{3\times3}$ is the constant hydrodynamic added mass matrix, $C_A : \mathbb{R}^3 \to \mathbb{R}^{3\times3}$ is the unknown hydrodynamic centripetal and Coriolis matrix, $D_A : \mathbb{R}^3 \to \mathbb{R}^{3\times3}$ is the unknown hydrodynamic damping and friction matrix, $G : \mathbb{R}^3 \to \mathbb{R}^3$ is the gravitational and buoyancy force and moment vector, and $\tau_b \in \mathbb{R}^3$ is the body-fixed force

[12] Note that condition (30) and (33) imply that $\tilde{\alpha}$ and $\max_{p\in\mathcal{P}}\{\alpha_{2,p}(\|x(t_0)\|)\}$, respectively, are in the codomain of $\alpha_{1,q}$ for all $q \in \mathcal{P}$. Furthermore, since $\beta^*(s,t) \leq s$, $\beta^*\left(\max_{p\in\mathcal{P}}\{\alpha_{2,p}(\|x(t_0)\|)\}, t\right)$ is also in the codomain of $\alpha_{1,q}$ for all $q \in \mathcal{P}$.

and moment control input. Further define $\eta_{AUV} \triangleq [x,\, y,\, \psi]^T$ and $\nu_{AUV} \triangleq [u_b,\, v_b,\, r_b]^T$, where $x,\, y \in \mathbb{R}$ are the Earth-fixed position vector components of the center of mass, $\psi \in [0, 2\pi]$ represents the yaw angle, $u_b,\, v_b \in \mathbb{R}$ are the body-fixed translational velocities, and $r_b \in \mathbb{R}$ is the body-fixed angular velocity. The constant irrotational current vector is generally defined as $\nu_c \triangleq \begin{bmatrix} u_c & v_c & 0 \end{bmatrix}$, where $u_c, v_c \in \mathbb{R}$ are the body-fixed translational velocities. The coordinate transformation $J_E : \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ is

$$J_E(\eta_{AUV}) = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{51}$$

Given the previous definitions, the control affine form of the AUV dynamics is

$$\dot{\xi} = Y(\xi, \nu_c)\,\theta + f_0(\xi, \dot{\nu}_c) + g\tau_b, \tag{52}$$

where $\xi \triangleq \left[\eta_{AUV}^T,\, \nu_{AUV}^T\right]^T \in \mathbb{R}^6$ is the concatenated state vector, $f_0 : \mathbb{R}^6 \times \mathbb{R}^3 \to \mathbb{R}^6$ is the known rigid body drift dynamics, $Y : \mathbb{R}^6 \times \mathbb{R}^3 \to \mathbb{R}^{6\times 5}$ is the known regression matrix, and $\theta \in \mathbb{R}^5$ is a vector of unknown hydrodynamic parameters. Furthermore, let $e \triangleq \xi - \xi_d$ and $\zeta = \left[e^T,\, \xi_d^T\right]^T$.

Each mode of the controller corresponds to a different irrotational current vector. The subsystem that each quantity belongs to is marked with an appropriate subscript; if there is no subscript, then that quantity can be identically applied across all subsystems. Three irrotational currents, which correspond to the three different subsystems are $\nu_{c1} = [-0.1, 0.1, 0]^T$, $\nu_{c2} = [0.05, -0.2, 0]^T$, and $\nu_{c3} = [-0.15, -0.1, 0]^T$. The current direction and magnitude are switched every 20 seconds resulting in the switching signal[13]

$$\sigma(t) = \begin{cases} 1, & 60\left\lfloor \frac{t}{60} \right\rfloor \le t < 60\left\lfloor \frac{t}{60} \right\rfloor + 20, \\ 2, & 60\left\lfloor \frac{t}{60} \right\rfloor + 20 \le t < 60\left\lfloor \frac{t}{60} \right\rfloor + 40, \\ 3, & 60\left\lfloor \frac{t}{60} \right\rfloor + 40 \le t < 60\left\lfloor \frac{t}{60} \right\rfloor + 60, \end{cases} \tag{53}$$

where $\lfloor \cdot \rfloor$ denotes the floor operator. The initial state is $\xi(0) = \left[-1, 1.5, \frac{3\pi}{4}, 0, 0, 0\right]^T$. The initial parameter estimate is $\hat{\theta}(0) = \mathbf{0}_{6\times 1}$. The desired trajectory is generated by

$$\xi_d(t) = \left[\cos\left(\frac{\pi}{20}t\right), \cos\left(\frac{\pi}{30}t\right), 0 \right. \\ \left. -\frac{\pi \sin\left(\frac{\pi}{20}t\right)}{20}, -\frac{\pi \sin\left(\frac{\pi}{30}t\right)}{30}, 0\right]^T \tag{54}$$

and, hence, is initialized as $\xi_d(0) = [1, 1, 0, 0, 0, 0]^T$.

The cost function for each subsystem is selected as $r(\zeta, \mu) = \zeta^T Q \zeta + \mu^T R \mu$, where $Q = \operatorname{diag}(100, 100, 200, 10, 10, 50, 0, 0, 0, 0, 0, 0)$, and $R = I_{3\times 3}$, $I_{n\times n}$ denotes the $n \times n$ identity matrix, and $\operatorname{diag}(v)$ for a vector v denotes a diagonal matrix with entries of the vector on the diagonal. The learning parameters are selected as $\eta_c = 0.5$, $\eta_{a1} = 10$, $\eta_{a2} = 0.1$, $\nu = 0.025$, $\lambda = 0.025$, $\overline{\Gamma} = 5000$, $\underline{\Gamma} = 100$, and $\Gamma(0) = 5000 \cdot I_{27\times 27}$.

[13] Future work will investigate detecting and accounting for a sudden change in the system model.

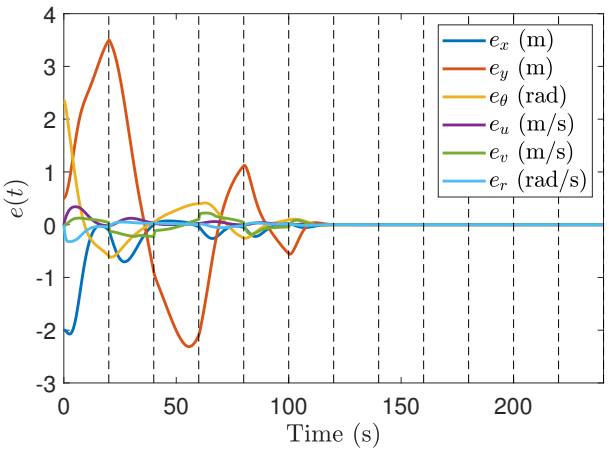


Figure 2. Error trajectories e of the AUV. The vertical dashed lines denote the time at which a switching instance occurred.

The actor and critic weights $\hat{W}_a(0)$ and $\hat{W}_c(0)$ were initialized by solving the ARE for the linearized rigid body AUV dynamics about the position $\xi = \mathbf{0}_{6\times 1}$. For each subsystem the same BE extrapolation trajectories ζ_i were used, where each element of ζ_i was selected from a uniform distribution on the interval $[-1, 1]$. The polynomial basis function $\phi = \phi_p$ for all $p \in \{1, 2, 3\}$ used for value function approximation is

$$\phi(\zeta) = [\zeta_1\zeta_2,\, \zeta_1\zeta_3,\, \zeta_1\zeta_4,\, \zeta_1\zeta_5,\, \zeta_1\zeta_6,\, \zeta_2\zeta_3,\, \zeta_2\zeta_4, \\ \zeta_2\zeta_5,\, \zeta_2\zeta_6,\, \zeta_3\zeta_4,\, \zeta_3\zeta_5,\, \zeta_3\zeta_6,\, \zeta_4\zeta_5,\, \zeta_4\zeta_6,\, \zeta_5\zeta_6,\, \zeta_1^2,\, \zeta_2^2, \\ \zeta_3^2,\, \zeta_4^2,\, \zeta_5^2,\, \zeta_6^2,\, \zeta_3\zeta_7,\, \zeta_3\zeta_8,\, \zeta_3^2,\, \zeta_3\zeta_{10},\, \zeta_3\zeta_{11},\, \zeta_3\zeta_{12}]^T. \tag{55}$$

To facilitate ICL, a maximum of $M_p = 100$ state-action pairs for all $p \in \{1, 2, 3\}$ are recorded and replaced according to the singular value maximization algorithm defined in [21, Algorithm 1]. The state-action pairs are not populated *a priori*; all data needed for ICL is generated online. The ICL learning parameters are $\Gamma_\theta = \operatorname{diag}(50,\, 30,\, 10,\, 2.5,\, 1)$, $k_\theta = 5 \cdot 10^5$, and $\Delta t = 0.25$.

Figure 2 shows the error trajectories of the AUV while switching between the multiple subsystems every 20 seconds. Despite the three distinct currents acting on the AUV, each subsystem's control policy approximates the dynamic effect of its respective current and appropriately compensates for it with a feedforward control term. The approximation of this feedforward term facilitates convergence at approximately 120 seconds.

Figure 3 shows the parameter estimation error $\tilde{\theta}$ for the active subsystem. The parameters are estimated to within a neighborhood of the actual values at approximately 100 seconds. Parameter estimation is facilitated by the fact that each subsystem, even when inactive, continues to learn the uncertain parameters via the ICL history stack. The error convergence in Figure 2 occurs approximately at 120 seconds, highlighting the fact that error convergence occurs after the uncertain parameters are identified. Hence, as predicted by

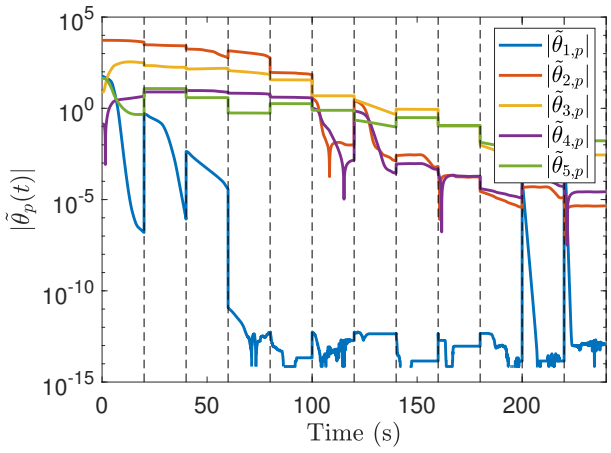


Figure 3. System identification errors $\tilde{\theta}_p$ for the hydrodynamic parameters of the AUV. The vertical dashed lines denote the time at which a switching instance occurred.

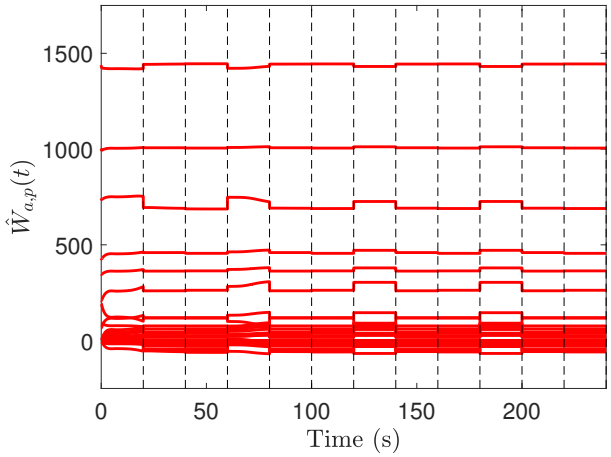


Figure 4. The actor weight estimates $\hat{W}_{a,p}$. The vertical dashed lines denote the time at which a switching instance occurred.

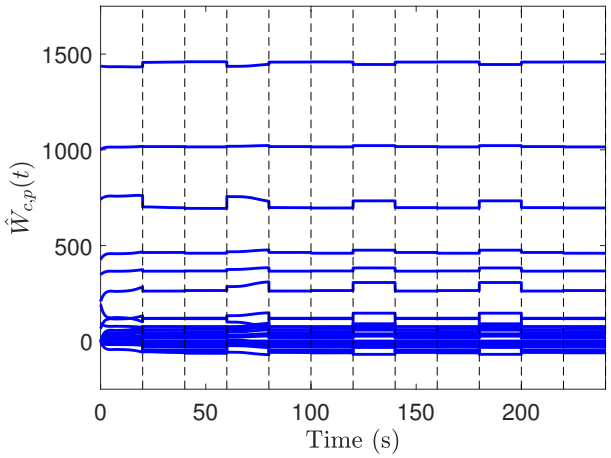


Figure 5. The critic weight estimates $\hat{W}_{c,p}$. The vertical dashed lines denote the time at which a switching instance occurred.

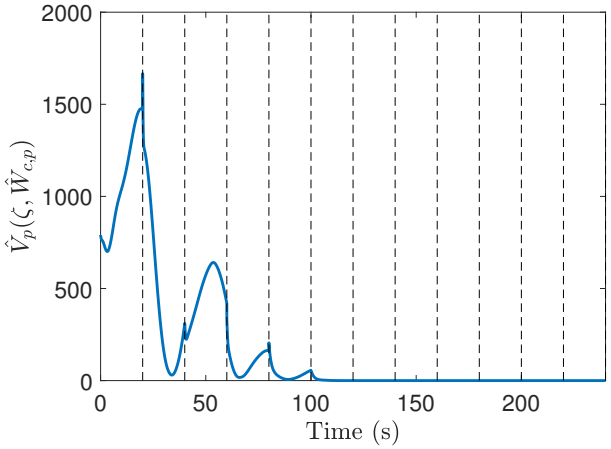


Figure 6. Value function approximation value $\hat{V}\left(\zeta, \hat{W}_{c,p}\right)$. The vertical dashed lines denote the time at which a switching instance occurred.

the theoretical analysis, once the parameters, and therefore the correct feedforward control term for trajectory tracking, is identified, then the controller is able to drive the tracking error to zero.

Figures 4 and 5 show the actor and critic weight approximations, respectively, for the active subsystem. The weights initially change to reflect the parameters updated parameter estimates (see Figure 3). It may appear that the weights do not change significantly. This behavior is due to the large magnitude of the weights. Hence, a small change in the actor and critic weight approximation in Figures 4 and 5 result in a large magnitude change in the applied control input.

Figure 6 shows each approximated optimal value function $\hat{V}_p\left(\zeta, \hat{W}_{c,p}\right)$ for the active subsystem. Based on the construction of the cost function $r_p\left(\zeta, \mu_p\right)$, the convergence of the approximation optimal value function corresponds to the error convergence in Figure 2.

Figure 7 shows each approximated optimal transient control policy $\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)$ for the active subsystem. $\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)$ converges to 0 at 120 seconds because the transient component has been eliminated. At that point, only the trajectory tracking component of the controller $\hat{u}_{d,p}\left(\zeta, \hat{\theta}_p\right)$ is nonzero (i.e., active).

IX. Conclusion

In this paper, a new Lyapunov-based theorem is developed to analyze convergence properties of switched systems in the case where each subsystem has UUB or locally practically stable trajectories. Sufficient conditions that relate the minimum dwell time, the initial conditions, the convergence rates, and the ultimate bounds of the subsystem to those of the switched system are developed.

This new theorem aids in the design of an ADP-based controller to optimize the performance of a switched system while achieving a tracking objective and compensating for parametric uncertainties in the system's drift dynamics. Local practical stability of individual subsystems, along with local practical stability of the overall switched system are proven via two Lyapunov-based stability analyses. Simulations results are

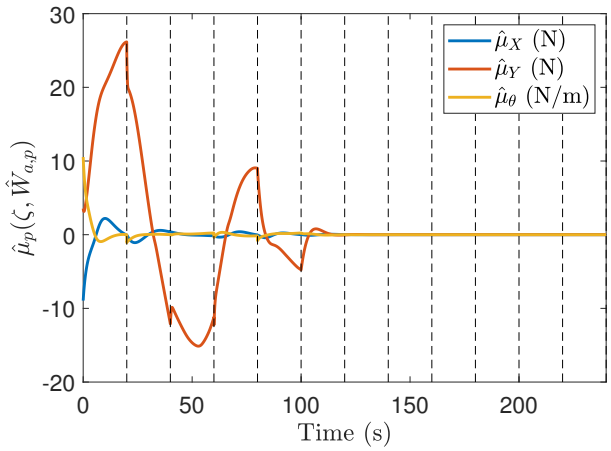


Figure 7. Transient controller term $\hat{\mu}_p\left(\zeta, \hat{W}_{a,p}\right)$. The vertical dashed lines denote the time at which a switching instance occurred.

presented for optimal control of an AUV in the presence of a discretely varying set of irrotational currents to show the efficacy of the developed technique. Future research will expand on the results in this paper by compensating for uncertainty in the control effectiveness matrix and investigating stronger subsystem stability results.

When applied to the ADP-based design, the sufficient conditions of the developed theorem provide qualitative intuition as to which parameters affect the needed minimum dwell time. Since the sufficient conditions require knowledge of bounds on the optimal value function, the bounds are difficult to compute in applications where estimation of bounds on the optimal value function is not feasible. The need to estimate bounds on the optimal value function, while limiting, is typical in ADP-based designs for computation of control gains, ultimate bounds, and regions of attraction (see [12]–[16]).

In this paper, the objective is to optimize the performance of each subsystem, with respect to given subsystem-specific performance metrics, while maintaining stability of the switched system. Optimization of the switched system relative to a system-wide performance metric is out of the scope of this work and is also a topic for future research.

REFERENCES

[1] D. Liberzon, *Switching in Systems and Control*. Birkhauser, 2003.

[2] D. J. Leith and W. E. Leithead, "Survey of gain-scheduling analysis and design," *Int. J. Control*, vol. 73, no. 11, pp. 1001–1025, 2000.

[3] W. J. Rugh and J. S. Shamma, "Research on gain scheduling," *Automatica*, vol. 36, pp. 1401–1425, Oct. 2000.

[4] J. Shamma and M. Athans, "Gain scheduling: Potential hazards and possible remedies," *IEEE Control System Magazine*, vol. vol. 12, no. no. 3, pp. 101–107, 1992.

[5] B. Stevens and F. Lewis, *Aircraft Control and Simulation*. Hoboken, NJ: John Wiley and Sons, 2003.

[6] L. Eugene, W. Kevin, and D. Howe, "Robust and adaptive control with aerospace applications," 2013.

[7] B. D. Anderson and J. B. Moore, *Optimal control: linear quadratic methods*. Courier Corp., 1971.

[8] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge university press, 1999.

[9] D. Bertsekas, "Approximate policy iteration: a survey and some new methods," *J. Control Theory Appl.*, vol. 9, pp. 310–335, 2011.

[10] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[12] J. Si, A. Barto, W. Powell, and D. Wunsch, eds., *Handbook of Learning and Approximate Dynamic Programming*. Wiley-IEEE Press, 2004.

[13] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. The Institution of Engineering and Technology, 2013.

[14] F. L. Lewis and D. Liu, *Reinforcement learning and approximate dynamic programming for feedback control*, vol. 17. John Wiley & Sons, 2013.

[15] R. Kamalapurkar, P. S. Walters, J. A. Rosenfeld, and W. E. Dixon, *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Springer, 2018.

[16] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. John Wiley & Sons, 2017.

[17] D. Mitrovic, S. Klanke, and S. Vijayakumar, "Adaptive optimal feedback control with learned internal dynamics models," in *From Motor Learning to Interaction Learning in Robots* (O. Sigaud and J. Peters, eds.), vol. 264 of *Studies in Computational Intelligence*, pp. 65–84, Springer Berlin Heidelberg, 2010.

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

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

[20] A. Parikh, R. Kamalapurkar, and W. E. Dixon, "Integral concurrent learning: Adaptive control with parameter convergence using finite excitation," *Int J Adapt Control Signal Process*, vol. 33, pp. 1775–1787, Dec. 2019.

[21] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, 2013.

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

[23] P. Walters, R. Kamalapurkar, F. Voight, E. Schwartz, and W. E. Dixon, "Online approximate optimal station keeping of a marine craft in the presence of an irrotational current," *IEEE Trans. Robot.*, vol. 34, pp. 486–496, April 2018.

[24] P. Deptula, J. Rosenfeld, R. Kamalapurkar, and W. E. Dixon, "Approximate dynamic programming: Combining regional and local state following approximations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, pp. 2154–2166, June 2018.

[25] 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, pp. 743–748, July 2020.

[26] M. L. Greene, P. Deptula, R. Kamalapurkar, and W. E. Dixon, *Handbook of Reinforcement Learning and Control*, ch. Mixed Density Methods for Approximate Dynamic Programming, pp. 139–172. Cham: Springer International Publishing, 2021.

[27] X. Xu and P. J. Antsaklis, "Optimal control of switched systems based on parameterization of the switching instants," *IEEE Trans. Autom. Control*, vol. 49, no. 1, pp. 2–16, 2004.

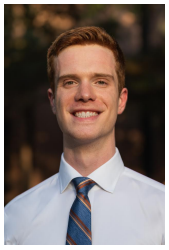
[28] W. Zhang, J. Hu, and A. Abate, "On the value functions of the discrete-time switched lqr problem," *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2669–2674, 2009.

[29] A. Heydari and S. N. Balakrishnan, "Optimal switching and control of nonlinear switching systems using approximate dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1106–1117, 2013.

[30] Y. Wardi, M. Egerstedt, and M. Hale, "Switched-mode systems: gradient-descent algorithms with Armijo step sizes," *Discrete Event Dyn. Syst.*, vol. 25, no. 4, pp. 571–599, 2015.

[31] A. Heydari, "Optimal switching with minimum dwell time constraint," *Journal of the Franklin Institute*, vol. 354, no. 11, pp. 4498–4518, 2017.

[32] M. Kamgarpour and C. Tomlin, "On optimal control of non-autonomous switched systems with a fixed mode sequence," *Automatica*, vol. 48, no. 6, pp. 1177–1181, 2012.
[33] S. Dharmatti and M. Ramaswamy, "Hybrid control systems and viscosity solutions," *SIAM J. on Control and Optim.*, vol. 44, no. 4, pp. 1259–1288, 2005.
[34] H. Axelsson, M. Boccadoro, M. Egerstedt, P. Valigi, and Y. Wardi, "Optimal mode-switching for hybrid systems with varying initial states," *Nonlinear Anal. Hybrid Syst.*, vol. 2, no. 3, pp. 765–772, 2008.
[35] M. Abudia, M. Harlan, R. Self, and R. Kamalapurkar, "Switched optimal control and dwell time constraints: A preliminary study," in *IEEE Conf. Decis. Control*, pp. 3261–3266, IEEE, 2020.
[36] S. C. Bengea and R. A. DeCarlo, "Optimal control of switching systems," *Automatica*, vol. 41, no. 1, pp. 11–27, 2005.
[37] H. Zhang, C. Qin, and Y. Luo, "Neural-network-based constrained optimal control scheme for discrete-time switched nonlinear system using dual heuristic programming," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 839–849, 2014.
[38] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *Int. J. Control*, vol. 87, no. 5, pp. 1000–1009, 2014.
[39] B. D. Anderson and J. B. Moore, *Linear Optimal Control*. Prentice-Hall, 1971.
[40] A. Parikh, T.-H. Cheng, R. Licitra, and W. E. Dixon, "A switched systems approach to image-based localization of targets that temporarily leave the camera field of view," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 6, pp. 2149–2156, 2018.
[41] M. 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*, pp. 162–167, 2020.
[42] L. Zhai and K. G. Vamvoudakis, "Data-based and secure switched cyber–physical systems," *Sys. & Control Lett.*, vol. 148, p. 104826, 2021.
[43] M. H. Cohen, Z. Serlin, K. Leahy, and C. Belta, "Temporal logic guided safe model-based reinforcement learning: A hybrid systems approach," *Nonlinear Anal.: Hybrid Sys.*, vol. 47, p. 101295, 2023.
[44] R. Goebel, R. G. Sanfelice, and A. R. Teel, *Hybrid Dynamical Systems*. Princeton University Press, 2012.
[45] A. Chaillet and A. Loría, "Necessary and sufficient conditions for uniform semiglobal practical asymptotic stability: Application to cascaded systems," *Automatica*, vol. 42, no. 11, pp. 1899–1906, 2006.
[46] A. Chaillet and A. Loría, "Uniform semiglobal practical asymptotic stability for non-autonomous cascaded systems and applications," *Automatica*, vol. 44, no. 2, pp. 337–347, 2008.
[47] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 3 ed., 2002.
[48] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2012.
[49] F. Sauvigny, *Partial Differential Equations 1: Foundations and Integral Representations*. Springer Science & Business Media, 2012.
[50] P. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice Hall, 1996.
[51] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, "Online adaptive algorithm for optimal control with integral reinforcement learning," *Int. J. of Robust and Nonlinear Control*, vol. 24, no. 17, pp. 2686–2710, 2014.
[52] C. Wu, J. Li, B. Niu, and X. Huang, "Switched concurrent learning adaptive control of switched systems with nonlinear matched uncertainties," *IEEE Access*, vol. 8, pp. 33560–33573, 2020.
[53] J.-J. E. Slotine and W. Li, "On the adaptive control of robot manipulators," *Int. J. Robot. Res.*, vol. 6, no. 3, pp. 49–59, 1987.
[54] M. L. Greene, P. Deptula, B. Bialy, and W. E. Dixon, "Model-based approximate optimal feedback control of a hypersonic vehicle," in *AIAA SCITECH*, Jan. 2022. AIAA 2022-0613.
[55] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.
[56] P. Deptula, Z. Bell, E. Doucette, W. J. Curtis, and W. E. Dixon, "Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with control effectiveness faults," *Automatica*, vol. 116, pp. 1–10, June 2020.
[57] A. S. Morse, "Supervisory control of families of linear set-point controllers part I. exact matching," *IEEE Trans. Autom. Control*, vol. 41, no. 10, pp. 1413–1431, 1996.
[58] N. Fischer, D. Hughes, P. Walters, E. Schwartz, and W. E. Dixon, "Nonlinear RISE-based control of an autonomous underwater vehicle," *IEEE Trans. Robot.*, vol. 30, pp. 845–852, Aug. 2014.
[59] T. I. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control*. Wiley, 2011.



Max L. Greene received his Ph.D. in mechanical engineering from the University of Florida, Gainesville, FL, USA in 2022. His research interests include Lyapunov-based, adaptive, and reinforcement learning-based control of dynamical systems. In 2022 he joined Aurora Flight Sciences as an Aerospace Controls Researcher.

Masoud S. Sakha received his bachelor's degree in the field of electrical engineering and master's degree (first rank) in the field of Control engineering, both from the University of Tabriz, Iran. He is currently pursuing a Ph.D. under the guidance of Dr. Kamalapurkar in the Systems, Cognition, and Control Laboratory. His main research interests are Nonlinear and Optimal Control Theory.

Rushikesh Kamalapurkar received his M. S. and his Ph. D. degree in 2011 and 2014, respectively, from the Mechanical and Aerospace Engineering Department at the University of Florida. In 2024 he joined the Department of Mechanical and Aerospace Engineering at the University of Florida as an associate professor. His primary research interest has been intelligent, learning-based control of uncertain nonlinear dynamic systems.

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 has been the development and application of Lyapunov-based control techniques for uncertain nonlinear systems.