Model-based Reinforcement Learning for Optimal Feedback Control of Switched Systems

Max L. Greene, Moad Abudia, Rushikesh Kamalapurkar, Warren E. Dixon

***Abstract*—This paper examines the use of reinforcement learning-based controllers to approximate multiple value functions of specific classes of subsystems while following an arbitrarily switching sequence. Each subsystem may have varying characteristics, such as different cost function or system dynamics. Stability of the overall switching sequence is proven using Lyapunov-based analysis techniques. Specifically, Lyapunov-based methods are developed to prove boundedness of individual subsystems and to determine a minimum dwell-time condition to ensure stability of the overall switching sequence. Uniformly ultimately bounded regulation of the states, approximation of the value function, and approximation of the optimal control policy is achieved for arbitrary switching sequences provided the minimum dwell-time condition is satisfied.**

I. Introduction

Gain scheduling has been shown to be a beneficial tool in the control of nonlinear systems [1]–[3]. However, gain scheduling control techniques often use a divide and conquer strategy to decompose nonlinear control design tasks into multiple linear design problems [1].

A motivating example is the work in [4], in which active magnetic bearing motors (e.g., hard disk drive motors) perform a tracking objective. In [4], three motor controllers are synthesized and switched between depending on the system state (e.g., motor revolutions per minute (RPM)). The dynamics differ between RPM ranges since vibratory disturbances are activated at varying RPMs (see [4]). In the aforementioned design, each region has a distinct controller with different gains to attenuate disturbances within each respective region.

Reinforcement learning (RL)-based methods such as [5]–[15] have been used to obtain online approximate solutions to optimal control problems for systems with finite state-spaces and stationary environments. Approximate dynamic programming (ADP) uses RL to approximate the value function (i.e., the solution) corresponding to optimal control problems for deterministic autonomous control-affine systems (see [7], [8], [16]–[18]). The optimal control policy is derived from the Hamilton-Jacobi-Bellman (HJB) equation and depends on the optimal value function [19], [20]. However, obtaining an analytical solution of the HJB (the optimal value function) is, generally, not possible; hence, an approximate value function is sought. ADP techniques use universal function approximators, such as certain neural networks (NNs) (see [21]–[23]), to approximate the value function by using the state as the inputs. Obtaining a more accurate approximate value function leads to a more accurate approximate optimal control policy. ADP-based controllers do not use stabilizing feedback gains and are defined by the estimated parameters of the value function [15]. Because of the lack of traditional feedback, gain tuning cannot be performed in the sense of increasing and decreasing feedback gains.

Unlike gain scheduling, in which feedback gains are varied, in ADP-based controllers costs are assigned to the control input and states to achieve different tracking performance. Altering the weights of the cost function in an ADP-based controller affects system performance by modifying the reward gained from system's states or control input. Drawing inspiration from the gain tuning's strategy, this paper proposes a method by which the weights of the cost function and system dynamics can be varied. That is, different controller properties can be achieved by varying the weights of the cost of the states.

Considering the hard disk drive example again, suppose that the speed error is weighted more than the position error during startup. This cost would reward a faster initial dynamic response. Once the speed error is within a certain tolerance, then the cost function could change to more heavily weight the position error over the speed error. Using each controller independently may provide undesirable performance (i.e., high overshoot or slow rise time, respectively). Intuitively, switching between the two control policies at appropriate times could provide improved overall controller performance (i.e., fast rise time for speed tracking and accurate position tracking).

Previous ADP results consider fixed state cost and control cost matrices within the cost function, such as [14], [15], [24], [25]. Works such as [26] and [27] examine the use of ADP-based methods for switched discrete-time nonlinear systems. Previous results such as [28]–[32] use optimal control methods to minimize cost function(s) of a switched system. These methods use a fixed mode sequence (see [28], [31]–[33]) or fixed switching instances (see [30]). In comparison, the developed method uses an arbitrary switching sequence that satisfies a dwell-time condition to approximate

Max L. Greene and Warren E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, 32611-6250 USA. Email: {maxgreene12, wdixon}@ufl.edu. Moad Abudia and Rushikesh Kamalpurkar are with the Department of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. Email: {abudia, rushikesh.kamalapurkar}@okstate.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 sponsoring agencies.

the value functions of a finite number of continuous-time subsystems. Unlike the aforementioned methods, this method proposes a Lyapunov-based framework to prove convergence of a control policy to the neighborhood of an optimal policy. While this paper focuses on a framework for switching between multiple ADP-based controllers and modifying control system performance by using different weighting matrices and dynamical models, it does not address optimality of the overall developed trajectory.

A complication in Lyapunov-based analyses for switched systems is the growth and discontinuity of Lyapunov functions at switching instances [34]. To overcome this issue, a dwell-time analysis is performed to prove stability of the overall switching sequence. The included dwell-time analysis accounts for the worst-case growth and discontinuity between Lyapunov functions during switching instances by explicitly determining the minimum time required before the system can switch to a different subsystem. In doing so, overall stability of the system for an arbitrary switching sequence is established.

This paper develops a continuous-time ADP-based controller that follows an arbitrary switching sequence between multiple dynamical systems and cost functions based on environmental conditions or at the user's discretion, given that certain conditions are satisfied. Section II describes the framework of a general ADP-based controller. Section III presents the stability analysis for one subsystem and behavior of the overall system for an arbitrary switching sequence by developing a dwell-time condition. Simulation results for a three state dynamical system are presented in Section IV to demonstrate the performance of the developed technique.

II. Approximate Optimal Controller Development

Let $k \in \mathbb{S}$, where $\mathbb{S} \subset \mathbb{N}$ and $|\mathbb{S}| < \infty$, represent a family of switched subsystems. Consider the continuous-time control-affine nonlinear dynamical subsystem of the k^{th} mode,

$$\dot{x} = f_k\left(x\left(t\right)\right) + g_k\left(x\left(t\right)\right)u\left(t\right), \tag{1}$$

with initial condition $x\left(0\right) = x_0 \in \mathbb{R}^n$, where $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ denotes the system state, $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ denotes the control input, $f_k : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift dynamics, and $g_k : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ is the control effectiveness. The considered class of functions satisfy the following assumptions.[1]

Assumption 1. The drift dynamics f_k are locally Lipschitz, $f_k' : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is continuous,[2] and $f_k\left(0\right) = 0$.

Assumption 2. The control effectiveness g_k is a locally Lipschitz function and bounded such that $0 < \left\|g_k\left(x\right)\right\| \leq \overline{g}_k$ where $\overline{g}_k \in \mathbb{R}_{>0}$.

[1]Throughout this paper, the subscript k defines the quantity or function belonging to the k^{th} mode of the overall system.

[2]*The notation $(\cdot)'$ denotes the partial derivative with respect to the first argument.*

The control objective is to solve the infinite-horizon optimal regulation problem for each subsystem, i.e., determine a control policy u that minimizes the infinite horizon cost function, $J_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$, defined as

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

subject to (1) while regulating the system states of the k^{th} mode to the origin (i.e., $x = 0$), where $r_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is the instantaneous cost defined as $r_k\left(x, u\right) \triangleq x^T Q_k x + u^T R_k u$, $Q_k \in \mathbb{R}^{n\times n}$ is a constant user-defined symmetric positive definite (PD) matrix, and $R_k \in \mathbb{R}^{m\times m}$ is a constant PD symmetric matrix.

Property 1. The state cost matrix Q_k satisfies $\underline{q}_k I_n \leq Q_k \leq \overline{q}_k I_n$ where $\underline{q}_k, \overline{q}_k \in \mathbb{R}_{>0}$, and I_n represents the $n \times n$ identify matrix.

The infinite horizon value function (i.e., the cost to go) for the optimal solution of the k^{th} mode is denoted by $V_k^* : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and given by

$$V_k^*\left(x\right) = \min_{u(\tau)\in U,\, \tau\in\mathbb{R}_{\geq t}} \int_t^{\infty} r_k\left(x\left(\tau\right), u\left(\tau\right)\right) d\tau, \tag{3}$$

where $U \subseteq \mathbb{R}^m$ denotes the action space. Provided an optimal control policy exists, the value function is characterized by the corresponding HJB

$$0 = \min_{u(\tau)\in U}\Big(V_k^{*\prime}\left(x\right)\left(f_k\left(x\right) + g_k\left(x\right)u\right) + x^T Q_k x + u^T R_k u\Big), \tag{4}$$

with the boundary condition $V_k^*\left(0\right) = 0$.

Assumption 3. The value function V_k^* is continuously differentiable.

Provided the HJB in (4) admits a continuously differentiable PD solution, then the optimal closed-loop control policy $u_k^* : \mathbb{R}^n \to \mathbb{R}^m$ is

$$u_k^*\left(x\right) = -\frac{1}{2} R_k^{-1} g_k\left(x\right)^T \left(V_k^{*\prime}\left(x\right)\right)^T. \tag{5}$$

A. Value Function Approximation

The HJB in (4) requires knowledge of the optimal value function, which, generally, is an unknown function for nonlinear systems. Parametric methods can be used to approximate the value function over a compact domain. To facilitate the solution of (4), let $\Omega \subset \mathbb{R}^n$ be a compact set containing the origin with $x \in \Omega$. The universal function approximation property of single-layer NNs is used to represent the value function of the k^{th} mode V_k^* as

$$V_k^*\left(x\right) = W_k^T \phi\left(x\right) + \epsilon_k\left(x\right), \tag{6}$$

where $W_k \in \mathbb{R}^L$ is an unknown bounded vector of weights, $\phi : \mathbb{R}^n \to \mathbb{R}^L$ is a user-defined vector of basis functions,[3]

[3]There is no subscript k for the basis function because each mode uses the same basis function.

and $\epsilon_k : \mathbb{R}^n \to \mathbb{R}$ is the bounded function approximation error. Substituting (6) into (5), the optimal control policy of the k^{th} mode, u_k^*, can be expressed in terms of the gradient of the value function V_k^* as

$$u_k^*(x) = -\frac{1}{2} R_k^{-1} g_k(x) \left(\phi'(x)^T W_k + \epsilon_k'(x)^T \right). \quad (7)$$

Assumption 4. [10], [35], [36] There exists a set of constants that bound the unknown weight vector W_k, the user-defined basis vector ϕ, and approximation error ϵ_k, from above such that $\|W_k\| \leq \bar{W}_k$, $\sup_{x\in\Omega} \|\phi(x)\| \leq \bar{\phi}$, $\sup_{x\in\Omega} \|\phi'(x)\| \leq \bar{\phi}'$, $\sup_{x\in\Omega} \|\epsilon_k(x)\| \leq \bar{\epsilon}_k$, $\sup_{x\in\Omega} \|\epsilon_k'(x)\| \leq \bar{\epsilon}_k'$ for all k, where $\bar{W}_k$, $\bar{\phi}$, $\bar{\phi}'$, $\bar{\epsilon}_k$, $\bar{\epsilon}_k' \in \mathbb{R}_{>0}$.

Since the ideal weights are unknown, a parametric estimate, called a critic weight vector $\hat{W}_{c,k} \in \mathbb{R}^L$, is substituted to calculate the optimal value function estimate $\hat{V}_k : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$, where

$$\hat{V}_k\left(x, \hat{W}_{c,k}\right) = \hat{W}_{c,k}^T \phi(x). \quad (8)$$

An actor weight vector $\hat{W}_{a,k} \in \mathbb{R}^L$, is used to provide an approximate version of (7), the approximate optimal control policy $\hat{u}_k : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$ is given by

$$\hat{u}_k\left(x, \hat{W}_{a,k}\right) = -\frac{1}{2} R_k^{-1} g_k(x)^T \left(\phi'(x)^T \hat{W}_{a,k} \right). \quad (9)$$

B. Bellman Error

The HJB in (4) is equal to zero under optimal conditions; however, substituting (8) and (9) into (4) results in a residual term $\hat{\delta}_k : \mathbb{R}^n \times \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, which is referred to as the Bellman Error (BE), defined as

$$\hat{\delta}_k\left(x, \hat{W}_{c,k}, \hat{W}_{a,k}\right) \triangleq \hat{V}_k'\left(x, \hat{W}_{c,k}\right)\left(f_k(x) + g_k(x)\hat{u}_k\left(x, \hat{W}_{a,k}\right)\right) + \hat{u}_k\left(x, \hat{W}_{a,k}\right)^T R_k \hat{u}_k\left(x, \hat{W}_{a,k}\right) + x^T Q_k x, \quad (10)$$

where $\hat{V}_k'\left(x, \hat{W}_{c,k}\right) = \hat{W}_{c,k}^T \phi'(x)$ denotes the gradient of the value function estimate. The BE is indicative of how close the actor and critic weight estimates are to the ideal weights. By defining the mismatch between the estimates and the ideal values as $\tilde{W}_{c,k} \triangleq W_k - \hat{W}_{c,k}$ and $\tilde{W}_{a,k} \triangleq W_k - \hat{W}_{a,k}$, substituting (6) and (9) in (4), and subtracting from (10) yields

$$\hat{\delta}_k = \frac{1}{4} \tilde{W}_{a,k}^T G_{\phi,k} \tilde{W}_{a,k} - \omega_k^T \tilde{W}_{c,k} + O_k(x), \quad (11)$$

where $\omega_k : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}^n$ is defined as

$$\omega_k\left(x, \hat{W}_{a,k}\right) \triangleq \phi'(x)\left(f_k(x) + g_k(x)\hat{u}\left(x, \hat{W}_{a,k}\right)\right),$$

and $O_k(x) \triangleq \frac{1}{2} W_k^T \phi'(x) G_k \epsilon_k'^T + \frac{1}{4} G_{\epsilon,k} - \epsilon_k' f_k(x)$.[4]

Remark 1. The expressions in (10) and (11) are equivalent for the BE. However, (10) is used in implementation, while (11) is used in the stability analysis in Section III.

[4] G_k, $G_{\phi,k}$, and $G_{\epsilon,k}$ are defined as $G_k = G_k(x) \triangleq g_k(x) R_k^{-1} g_k(x)^T$, $G_{\phi,k} = G_{\phi,k}(x) \triangleq \phi'(x) G_k(x) \phi'(x)^T$, and $G_{\epsilon,k} = G_{\epsilon,k}(x) \triangleq \epsilon_k'(x) g_k(x) \epsilon_k'(x)^T$, respectively.

C. Switched Subsystems

Let the switching signal $\sigma(t) : \mathbb{R}_{\geq 0} \to \{k\}$ indicate the active subsystem. Let $t_k^{\text{ON}} \in [0, t]$ denote the time instant when the k^{th} subsystem in the switching sequence is activated. Similarly, let $t_k^{\text{OFF}} \in [0, t]$ denote the time instant when the k^{th} subsystem in the switching sequence is deactivated. The dwell-time in any active mode of a subsystem is denoted by $\tau \in \mathbb{R}_{\geq 0}$. Similarly, let $\tau^* \in \mathbb{R}_{\geq 0}$ denote the minimum dwell-time for any active mode of a subsystem.

D. Bellman Error Extrapolation

At each time instant, the BE in (10) is calculated using the control policy given by (9) evaluated using the current system state, critic weight estimates, and actor weight estimates to obtain the instantaneous BE denoted by $\hat{\delta}_k\left(x(t), \hat{W}_{c,k}(t), \hat{W}_{a,k}(t)\right)$.

A classical problem in learning-based control is exploration versus exploitation. Results such as [12], [37], [38] add an exploration signal to sufficiently explore the operating domain. However, no analytical methods exist to compute the appropriate exploration signal. Alternatively, results such as [14] evaluate the BE along the system trajectory and at any desired point in the state space (i.e., so-called BE extrapolation). The BE extrapolation technique provides simulation of experience to avoid using an exploration signal.

Specifically, BE is extrapolated from a user-specified number and location of off-trajectory points $\{x_{i,k} : x_{i,k} \in \Omega\}_{i=1}^{N_k}$, where $N_k \in \mathbb{N}$ denotes a user-specified number of points in the compact set Ω. The data is represented by the tuple $(\Sigma_{c,k}, \Sigma_{a,k}, \Sigma_{\Gamma,k})$, defined as $\Sigma_{c,k} \triangleq \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\omega_{i,k}(t)}{\rho_{i,k}(t)} \hat{\delta}_{i,k}(t)$, $\Sigma_{a,k} \triangleq \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{G_{\sigma i,k}^T \hat{W}_{a,k}(t) \omega_{i,k}^T(t)}{4\rho_{i,k}(t)}$, $\Sigma_{\Gamma,k} \triangleq \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\omega_{i,k}(t)\omega_{i,k}^T(t)}{\rho_{i,k}(t)}$, where $\hat{\delta}_{i,k}(t) \triangleq \hat{\delta}_k\left(x_{i,k}(t), \hat{W}_{c,k}(t), \hat{W}_{a,k}(t)\right)$, $\omega_{i,k}(t) \triangleq \omega_k\left(x_{i,k}(t), \hat{W}_{a,k}(t)\right)$, and $\rho_{i,k}(t) = 1 + \nu_k \omega_{i,k}^T(t) \Gamma_k(t) \omega_{i,k}(t)$, $\nu_k \in \mathbb{R}_{>0}$ is a user-defined gain, and $\Gamma_k : \mathbb{R} \to \mathbb{R}^{L\times L}$ is a time-varying least-squares gain matrix. Each subsystem, k, must have distinct sets of data, gain values, and update laws.

Assumption 5. Over the compact set, Ω, a finite set of off-trajectory points $\{x_{i,k} : x_{i,k} \in \Omega\}_{i=1}^{N_k}$ exists such that $0 < \underline{c}_k \triangleq \inf_{t\in\mathbb{R}_{\geq 0}} \lambda_{\min}\{\Sigma_{\Gamma,k}(t)\}$ for all $t \in \mathbb{R}_{\geq 0}$, where $\lambda_{\min}\{\cdot\}$ is the minimum eigenvalue.

Remark 2. The constant $\underline{c}_k$ is a scalar lower bound of the value of each input-output data pair's minimum eigenvalues.

E. Update Laws for Actor and Critic Weights

Using the instantaneous BE $\hat{\delta}_k(t)$, policy $u(t)$, and extrapolated BEs $\hat{\delta}_{i,k}(t)$, the critic and actor weights are updated according to the following policies while $t \in \left[t_k^{\text{ON}}, t_k^{\text{OFF}}\right)$. In the following definitions, $\eta_{c1,k}$, $\eta_{c2,k}$, $\eta_{a1,k}$, $\eta_{a2,k}$, $\lambda_k \in \mathbb{R}$ are positive constant

learning gains, and $\underline{\hat{W}}_{c,k}$, $\overline{\hat{W}}_{c,k}$, $\underline{\hat{W}}_{a,k}$, $\overline{\hat{W}}_{a,k}$, $\underline{\Gamma}_k$, $\overline{\Gamma}_k \in \mathbb{R}$ are upper and lower bound constants of subsystem k.[5] For the development of the weight update laws, define the following convex sets as

$$\Pi_{c,k} \triangleq \left\{ \hat{W}_{c,k} \in \left[\underline{\hat{W}}_{c,k}, \overline{\hat{W}}_{c,k} \right] \mid h_{c,k}\left(\hat{W}_{c,k} \right) \leq \xi_{c,k} \right\},$$
$$\Pi_{a,k} \triangleq \left\{ \hat{W}_{a,k} \in \left[\underline{\hat{W}}_{a,k}, \overline{\hat{W}}_{a,k} \right] \mid h_{a,k}\left(\hat{W}_{a,k} \right) \leq \xi_{a,k} \right\},$$

where $h_{c,k} : \mathbb{R}^L \to \mathbb{R}$ and $h_{a,k} : \mathbb{R}^L \to \mathbb{R}$ are smooth functions and $\xi_{c,k}$, $\xi_{a,k} > 0$. Denote the interior of a set Π by $\mathring{\Pi}$ and the boundary of Π by $\partial\Pi$. Observe that $h'_{c,k}$ and $h'_{a,k}$ represent outward normal vectors at $\partial\Pi_{c,k}$ and $\partial\Pi_{a,k}$, respectively. The critic update law of the k^{th} mode, $\dot{\hat{W}}_{c,k} : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$, is defined as

$$\dot{\hat{W}}_{c,k}(t) \triangleq \operatorname{proj}\left\{ \Phi_{c,k}(t) \right\} = \begin{cases} \Phi_{c,k}, & \hat{W}_{c,k} \in \mathring{\Pi}_{c,k} \text{ or } h'^T_{c,k} \Phi_{c,k} \leq 0 \\ \mathcal{C}_{c,k} \Phi_{c,k}, & \hat{W}_{c,k} \in \partial\Pi_{c,k} \text{ and } h'^T_{c,k} \Phi_{c,k} > 0, \end{cases} \tag{12}$$

where $\mathcal{C}_{c,k} : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$ is defined as $\mathcal{C}_{c,k} \triangleq 1 - \min\left(1, \frac{h_{c,k}}{\xi_{c,k}}\right) \frac{h'_{c,k} h'^T_{c,k}}{\left\| h'_{c,k} \right\|^2}$, and $\Phi_{c,k}(t) \triangleq -\eta_{c1,k} \Gamma_k \frac{\omega_k}{\rho_k} \hat{\delta}_k - \eta_{c2,k} \Sigma_{c,k}$. The actor update law of the k^{th} mode, $\dot{\hat{W}}_{a,k} : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$, is defined as

$$\dot{\hat{W}}_{a,k}(t) \triangleq \operatorname{proj}\left\{ \Phi_{a,k}(t) \right\} = \begin{cases} \Phi_{a,k}, & \hat{W}_{a,k} \in \mathring{\Pi}_{a,k} \text{ or } h'^T_{a,k} \Phi_{a,k} \leq 0 \\ \mathcal{C}_{a,k} \Phi_{a,k}, & \hat{W}_{a,k} \in \partial\Pi_{a,k} \text{ and } h'^T_{a,k} \Phi_{a,k} > 0, \end{cases} \tag{13}$$

where $\mathcal{C}_{a,k} : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$ is defined as $\mathcal{C}_{a,k} \triangleq 1 - \min\left(1, \frac{h_{a,k}}{\xi_{a,k}}\right) \frac{h'_{a,k} h'^T_{a,k}}{\left\| h'_{a,k} \right\|^2}$, and $\Phi_{a,k} \triangleq -\eta_{a1,k}\left(\hat{W}_{a,k} - \hat{W}_{c,k} \right) - \eta_{a2,k} \hat{W}_{a,k} + \frac{\eta_{c1,k} G^T_{\sigma,k} \hat{W}_{a,k} \omega^T_k}{4\rho_k} \hat{W}_{c,k} + \eta_{c2,k} \Sigma_{a,k} \hat{W}_{c,k}$. The least-squares gain matrix update law of the k^{th} mode, $\dot{\Gamma}_k : \mathbb{R}_{\geq 0} \to \mathbb{R}^{L \times L}$, is expressed as

$$\dot{\Gamma}_k(t) \triangleq \left(\lambda_k \Gamma_k - \eta_{c1,k} \frac{\Gamma_k \omega_k \omega_k^T \Gamma_k}{\rho_k^2} - \eta_{c2,k} \Gamma_k \Sigma_{\Gamma,k} \Gamma_k \right) \cdot \mathbf{1}_{\left\{ \underline{\Gamma}_k \leq \|\Gamma_k\| \leq \overline{\Gamma}_k \right\}}, \tag{14}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function.[6] While the k^{th} mode is inactive, (i.e., $t \notin \left[t_k^{\text{ON}}, t_k^{\text{OFF}} \right)$) : $\dot{\hat{W}}_{c,k}(t) = 0_{L \times 1}$ $\dot{\Gamma}_k(t) = 0_{L \times L}$, and $\dot{\hat{W}}_{a,k}(t) = 0_{L \times 1}$.[7]

[5] The arguments of each function have been omitted for notional brevity.

[6] Each $\|\Gamma_k(t)\|$ is bounded from above and below by some user-defined saturation gains, $\overline{\Gamma}_k$ and $\underline{\Gamma}_k$, respectively. Using (14) ensures that $\underline{\Gamma}_k \leq \|\Gamma_k(t)\| \leq \overline{\Gamma}_k$ for all $t \in \mathbb{R}_{>0}$ and $k \in \mathbb{S}$, where $\underline{\Gamma}_k \in \mathbb{R}_{>0}$. $\hat{W}_{c,k}$ and $\hat{W}_{a,k}$ are updated according to an orthogonal projection operator.

[7] The update laws will not update a subsystem k's weight estimates or least-squares matrix unless subsystem k is active.

III. Stability Analysis

Generally, the trajectory of a switched system can diverge even when all the subsystems that compose the switched system are stable. Hence, the switching signal must be properly designed to keep the overall system stable. Before the switching signal is designed, the stability of each subsystem must be analyzed. In the following development, k subsystems, each with a class of dynamics in (1), will be analyzed with the control policy and update laws outlined in (9), (12), (13), and (14).

A. Subsystem Stability Analysis

To facilitate the analysis, let $z_k \triangleq \left[x^T, \tilde{W}^T_{c,k}, \tilde{W}^T_{a,k} \right]^T$ denote a concatenated state, and let $V_{L,k} : \mathbb{R}^{n+2L} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a candidate Lyapunov function for the k^{th} mode be defined as

$$V_{L,k}(z_k, t) = V_k^*(x) + \frac{1}{2} \tilde{W}^T_{c,k} \Gamma_k^{-1}(t) \tilde{W}_{c,k} + \frac{1}{2} \tilde{W}^T_{a,k} \tilde{W}_{a,k}, \tag{15}$$

where k represents the active subsystem mode. Define the sequence of times instants at which a switching event occurs as $\{t_{N_\sigma}\}$, such that $0 < t_1 < t_2 < \cdots < t_{N_\sigma} < t < t_{N_\sigma + 1}$ and $N_\sigma \in \mathbb{N}_{>0}$ denotes the number of switching events. Using the positive definiteness of V_k^* and [39, Lemma 4.3], (15) can generally be bounded as $\underline{v}_{l,k}(\|z_k\|) \leq V_{L,k}(z_k, t) \leq \overline{v}_{l,k}(\|z_k\|)$ using class $\mathcal{K}$ functions $\underline{v}_{l,k}, \overline{v}_{l,k} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. For the subsequent analysis, the following more restrictive assumption is required.

Assumption 6. The optimal value function $V_k^*(x)$ can be bounded by the square of the norm of its argument times a positive constant, i.e.,

$$\beta_{1,k} \|x\|^2 \leq V_k^*(x) \leq \beta_{2,k} \|x\|^2, \ \forall k \in \mathbb{S}, \ \beta_{1,k}, \ \beta_{2,k} \in \mathbb{R}_{\geq 0}. \tag{16}$$

Remark 3. It is known that the value function of a classical linear-quadratic-regulator problem is quadratic and can be bounded from above and below by a quadratic function, as in (16) [40]. Assumption 6 may not be valid for some nonlinear systems. Future efforts will seek to generalize the following development without requiring this assumption.

Using Assumption 6, (15) can be bounded as $\alpha_{1,k} \|z_k\|^2 \leq V_{L,k}(z_k, t) \leq \alpha_{2,k} \|z_k\|^2$, where $\alpha_{1,k}$, $\alpha_{2,k} \in \mathbb{R}_{\geq 0}$ are positive constants. To facilitate the analysis, the notation $\overline{(\cdot)}$ is defined as $\overline{(\cdot)} \triangleq \sup_{x \in \Omega}(\cdot)$. Using (14), the normalized regressors $\frac{\omega_k}{\rho_k}$ and $\frac{\omega_{i,k}}{\rho_{i,k}}$ can be bounded as $\sup_{t \in \mathbb{R}_{\geq 0}} \left\| \frac{\omega_k}{\rho_k} \right\| \leq \frac{1}{2\sqrt{\nu_k \underline{\Gamma}_k}}$ for all $x \in \Omega$ and $\sup_{t \in \mathbb{R}_{\geq 0}} \left\| \frac{\omega_{i,k}}{\rho_{i,k}} \right\| \leq \frac{1}{2\sqrt{\nu_k \underline{\Gamma}_k}}$ for all $x_i \in \Omega$ and $k \in \mathbb{S}$. The matrices G_k and $G_{\phi,k}$ can be bounded as $\sup_{x \in \Omega} \|G_k\| \leq \lambda_{\max}\left\{ R_k^{-1} \right\} \overline{g}_k^2$ and $\sup_{x \in \Omega} \|G_{\phi,k}\| \leq \left(\overline{\phi}' \overline{g}_k \right)^2 \lambda_{\max}\left\{ R_k^{-1} \right\}$, respectively, for all $k \in \mathbb{S}$, where $\lambda_{\max}\{\cdot\}$ denotes the maximum eigenvalue.

Remark 4. Using the projection operator from the critic update law in (12) and [41, Lemma E.1], $-\tilde{W}^T_{c,k}(t) \Gamma_k^{-1}(t) \dot{\hat{W}}_{c,k}(t)$ is bounded from above as

$$
\begin{aligned}
&- \tilde{W}_{c,k}^{T}(t)\,\Gamma_k^{-1}(t)\,\dot{\hat{W}}_{c,k}(t)\\
&= -\tilde{W}_{c,k}^{T}(t)\,\Gamma_k^{-1}(t)\,\mathrm{proj}\left\{\Phi_{c,k}(t)\right\}\\
&\leq -\tilde{W}_{c,k}^{T}(t)\,\Gamma_k^{-1}(t)\,\Phi_{c,k}(t).
\end{aligned}
$$

Similarly, from the actor update law in (13) and [41, Lemma E.1], $-\tilde{W}_{a,k}^{T}(t)\,\dot{\hat{W}}_{a,k}(t)$ is bounded from above as

$$
\begin{aligned}
-\tilde{W}_{a,k}^{T}(t)\,\dot{\hat{W}}_{a,k}(t) &= -\tilde{W}_{a,k}^{T}(t)\,\mathrm{proj}\left\{\Phi_{a,k}(t)\right\}\\
&\leq -\tilde{W}_{a,k}^{T}(t)\,\Phi_{a,k}(t).
\end{aligned}
$$

To facilitate the subsequent analysis, let $l_k \in \mathbb{R}_{>0}$ be defined as $l_k \triangleq \frac{2\bar{a}_k^2}{\eta_{a1,k}+\eta_{a2,k}} + \frac{3(\eta_{c1,k}+\eta_{c2,k})^2}{8\nu_k\underline{\Gamma}_k\eta_{c2,k}\underline{c}_k}\overline{\|O_k(x)\|}^2 + \frac{1}{4}\overline{\|G_{\epsilon,k}\|} + \frac{1}{2}\eta_{a2,k}\overline{\|W_k\|}^2$, where $\bar{a}_k \triangleq \frac{1}{2}\lambda_{\max}\left\{R_k^{-1}\right\}\overline{\|\phi'\|\|G_k\|\|W_k\|}\left(\overline{\|\phi'\|\|G_k\|} + \overline{\|\epsilon_k'\|} + \frac{\eta_{c1,k}+\eta_{c2,k}}{4\sqrt{\nu_k\underline{\Gamma}_k}}\overline{\|\phi'\|\|G_k\|\|W_k\|}\right)$, and $\Lambda_k \triangleq \inf_{k\in\mathbb{S}}\left\{\frac{1}{2}\bar{q}_k,\ \frac{1}{16}\left(\eta_{a1,k}+\eta_{a2,k}\right),\ \frac{1}{12}\eta_{c2,k}\underline{c}_k\right\}$, where $\bar{q}_k$ is defined in Property 1. Furthermore, define $\mathcal{R} \in \mathbb{R}_{>0}$ as the radius of a ball $\mathcal{B}_{\mathcal{R}}$ centered at the origin, where $\mathcal{B}_{\mathcal{R}} \subset \Omega$.

Theorem 1. *Provided Assumptions 1-6 hold, the weight update laws in (12)-(14) are used, and the gain conditions*

$$
\eta_{a1,k}+\eta_{a2,k} > \frac{1}{\sqrt{\nu_k\underline{\Gamma}_k}}\left(\eta_{c1,k}+\eta_{c2,k}\right)\|W_k\|\,\overline{\|G_{\phi,k}\|}, \tag{17}
$$

$$
\underline{c}_k > \frac{3\left(\eta_{c1,k}+\eta_{c2,k}\right)^2\|W_k\|^2\,\overline{\|G_{\phi,k}\|}^2}{16\nu_k\underline{\Gamma}_k\left(\eta_{a1,k}+\eta_{a2,k}\right)\eta_{c2,k}} + \frac{3\eta_{a1,k}}{\eta_{c2,k}}, \tag{18}
$$

$$
\frac{\alpha_{2,k}}{\alpha_{1,k}}\sqrt{\frac{2l_k}{\Lambda_k}} < \mathcal{R}, \tag{19}
$$

are satisfied, the system state $x(t)$, the value function weight estimate error $\tilde{W}_{c,k}(t)$, and the control policy weight estimate error $\tilde{W}_{a,k}(t)$, are uniformly ultimately bounded (UUB). Hence, the error between the stabilizing control policy for each mode $\hat{u}_k(t)$ in (9) and its respective optimal control policy $u_k^*(t)$ in (5) is UUB.

Proof: Taking the time derivative of the Lyapunov function in (15) yields

$$
\begin{aligned}
\dot{V}_{L,k}\left(z_k(t),t\right) = {} & V_k^{*\prime}\dot{x} - \tilde{W}_{c,k}^{T}\Gamma_k^{-1}\dot{\hat{W}}_{c,k}\\
& - \tilde{W}_{a,k}^{T}\dot{\hat{W}}_{a,k} - \frac{1}{2}\tilde{W}_{c,k}^{T}\Gamma_k^{-1}\dot{\Gamma}_k\Gamma_k^{-1}\tilde{W}_{c,k},
\end{aligned}
$$

where the fact $\frac{d}{dt}\Gamma^{-1} = \Gamma^{-1}\dot{\Gamma}\Gamma^{-1}$ is used. Using the given class of dynamics, update laws, Assumptions 1-6 and the sufficient conditions in (17)-(19) yields

$$
\dot{V}_{L,k}\left(z_k(t),t\right) \leq -\frac{\Lambda_k}{\alpha_{2,k}}V_{L,k}\left(z_k(t),t\right) + l_k,
$$

for all $k \in \mathbb{S}$ and $t \in \left[t_k^{\mathrm{ON}}, t_k^{\mathrm{OFF}}\right)$. ∎

Remark 5. The condition in (17) can be satisfied by increasing $\eta_{a2,k}$ and ν_k, and selecting a penalty weight matrix R_k such that $\lambda_{\max}\left\{R_k^{-1}\right\}$ is small. Selecting a R_k with a large minimum eigenvalue and a large gain ν_k will also help satisfy the gain condition in (18). The condition in (18) can be satisfied by selecting off-trajectory points that increase the minimum eigenvalue of each $\underline{c}_k \triangleq \inf_{t\in\mathbb{R}_{\geq 0}}\left\{\Sigma_{\Gamma,k}(t)\right\}$.[8]

B. Dwell-Time Analysis

Theorem 1 indicates that each subsystem is UUB. However, this does not account for switching between subsystems. To ensure that the system is stable, a dwell-time must be designed to switch between subsystems. Furthermore, switching between control policies may result in instantaneous growth when switching between Lyapunov functions.[9] Hence, continuity is not guaranteed between Lyapunov functions $V_{L,k}$, across all subsystems.

Theorem 2. *The system consisting of a family of subsystems with the dynamics in (1) with a properly designed dwell-time, $\tau \in \mathbb{R}$ ensures that the states, critic estimate errors, and actor estimate errors will converge to a neighborhood of the origin in the sense that $V_{L,\sigma(t)}\left(z_{\sigma(t)}(t),t\right) \leq V_{L,B}$ for all $t \geq T$, where $V_{L,B} \in \mathbb{R}$ is the maximum ultimate bound for all subsystems, and $T \in \mathbb{R}_{\geq 0}$ is the time required to reach the ultimate bound $V_{L,B}$, provided a minimum dwell-time τ^* is satisfied.*

Proof: From Theorem 1, the derivative of the Lyapunov function of the k^{th} subsystem can be bounded from above by

$$
\dot{V}_{L,k}\left(z_k(t),t\right) \leq -\frac{\Lambda_k}{\alpha_{2,k}}V_{L,k}\left(z_k(t),t\right) + l_k,\ \forall k \in \mathbb{S},\ t \geq 0. \tag{20}
$$

Based on (20), the region $\mathbb{D}_k$, which represents the region within the ultimate bound of the Lyapunov function of the k^{th} subsystem, is defined as $\mathbb{D}_k \triangleq \left\{z_k : \|z_k\| \leq \frac{\alpha_{2,k}}{\alpha_{1,k}}\sqrt{\frac{2l_k}{\Lambda_k}}\right\}$. The union of the individual regions, $\mathbb{D}_k$, is denoted as $\overline{\mathbb{D}}_k \triangleq \bigcup_{k\in\mathbb{S}}\mathbb{D}_k$.The value of $V_{L,k}\left(z_k(t),t\right)$ due to switching inside of the region $\overline{\mathbb{D}}_k$ is bounded from above by

$$
V_{L,k}\left(z_k(t),t\right) \leq V_{L,B} \triangleq \sup_{k\in\mathbb{S}}\left\{\frac{2l_k\alpha_{2,k}^3}{\Lambda_k\alpha_{1,k}^2}\right\}.
$$

Generally, switching between control policies may result in instantaneous growth of the Lyapunov function. In this case, there is instantaneous growth between Lyapunov functions at the switching instances. Following [34] and [42], the scalar multiple that defines the maximum ratio of the discontinuities in the Lyapunov function is defined as $\mu \triangleq \left\{\frac{\sup_{k\in\mathbb{S}}\{\beta_{2,k}\}\|x(0)\|^2+\sup_{k\in\mathbb{S}}\{\Delta_k\}}{\inf_{k\in\mathbb{S}}\{\Delta_k\}}\right\}$, where $\Delta_k \triangleq$

[8] The minimum eigenvalue of each $\Sigma_{\Gamma,k}(t)$ can be increased by collecting redundant data, i.e., selecting $N \gg L$ for each subsystem.

[9] $V_{k+1}^*(x)$, corresponding to mode $k+1$, may be larger than $V_k^*(x)$ corresponding to mode k. Similarly, the actor and critic weight errors could be larger in magnitude while in mode $k+1$ than in mode k.

$\underline{\Gamma}_k^{-1}\left\|\overline{\hat{W}}_{c,k}-\underline{\hat{W}}_{c,k}\right\|^2+\left\|\overline{\hat{W}}_{a,k}-\underline{\hat{W}}_{a,k}\right\|^2$, such that the inequalities $V_{L,i}\leq\mu V_{L,j}$ $V_{L,j}\leq\mu V_{L,i}$, $\forall i\neq j$ hold, where $i,j\in\mathbb{S}$ index any two subsystems. Note that Δ_k can be calculated for each mode since $\overline{\hat{W}}_{c,k}$, $\underline{\hat{W}}_{c,k}$, $\overline{\hat{W}}_{a,k}$, $\underline{\hat{W}}_{a,k}$, and $\underline{\Gamma}_k$ are user-selected. From Assumption 6, $\beta_{2,k}$ can be selected sufficiently large to satisfy Assumption 6. Since the initial condition of the state $x(0)$ is known, μ can be calculated.

Starting with (20) and following the development from [42] and [43], the magnitude of $V_{L,k}\left(z_k(t),t\right)$ for any k can be expressed as

$$\begin{aligned}&V_{L,k}\left(z_k\left(t-t_k^{\mathrm{ON}}\right),t-t_k^{\mathrm{ON}}\right)\\&\leq\max\left\{V_{L,k}\left(z_k\left(t_k^{\mathrm{ON}}\right),t_k^{\mathrm{ON}}\right)e^{-\frac{\Lambda_k\alpha_{1,k}}{2\alpha_{2,k}}\left(t-t_k^{\mathrm{ON}}\right)},V_{L,B}\right\},\end{aligned}\tag{21}$$

for all $t\geq t_k^{\mathrm{ON}}$. Accounting for switching, by induction, (21) can be rewritten as

$$V_{L,k}\left(z_k(t),t\right)\leq\max\left\{V_{L,1}\left(z_1(0),0\right)\mu^{N_\sigma}e^{-\zeta_0 t},V_{L,B}\right\},\tag{22}$$

where $N_\sigma\in\mathbb{N}_{>0}$ is the total number of switches during $[0,t)$, and $\zeta_0\triangleq\inf_{k\in\mathbb{S}}\left\{\frac{\Lambda_k\alpha_{1,k}}{2\alpha_{2,k}}\right\}\in\mathbb{R}_{>0}$ is a constant. The inequality in (22) is true for an arbitrary sequence of switches provided that the subsequently defined minimum dwell-time condition is satisfied. A desired decay rate, ζ^*, can be determined such that

$$V_{L,k}\left(z_k(t),t\right)\leq\max\left\{V_{L,1}\left(z_1(0),0\right)e^{-\zeta^* t},V_{L,B}\right\},$$

where $\zeta^*\in(0,\zeta_0)$ is an arbitrarily selected decay rate that satisfies the inequality

$$\mu^{N_\sigma}e^{-\zeta_0 t}\leq e^{-\zeta^* t}.\tag{23}$$

A minimum dwell-time (i.e., the minimum amount of time required to converge low enough that the subsequent potential jump in the Lyapunov functions at the next mode will decrease) can be determined from (23) as

$$\tau^*=\frac{\ln\left(\mu^{N_\sigma}\right)}{\zeta_0-\zeta^*}.\tag{24}$$

Since $\zeta^*\in(0,\zeta_0)$, then $\zeta_0-\zeta^*>0$. Since $\mu\geq1$, $\zeta_0-\zeta^*>0$, $N_\sigma\in\mathbb{N}_{>0}$, and $N_\sigma<\infty$ then $\tau^*>0$, i.e., the dwell-time will always be positive. Since the number of switches is finite, the number of switches is bounded from above by $N_\sigma\leq\frac{t}{\tau^*}$, $t\in\left[t_{N_\sigma}^{ON},t_{N_\sigma+1}^{ON}\right)$, hence, τ^* is a minimum dwell-time. The time, $T\in\mathbb{R}_{\geq0}$, required to reach the region $\overline{\mathbb{D}}_k$ for the initial condition $V_L(z_1(0),0)$ is

$$T=\begin{cases}T\geq\frac{\ln\left(\frac{V_{L,B}}{V_{L,1}\left(z_k(0),0\right)}\right)}{\zeta^*} & \text{if } V_{L,1}\left(z_1(0),0\right)>V_{L,B}\\ T=0 & \text{if } V_{L,1}\left(z_1(0),0\right)\leq V_{L,B}.\end{cases}$$

Hence, the system state, actor weight estimates, and critic weight estimates will converge to a neighborhood of the origin in the sense that $V_{L,\sigma(t)}\left(z_k(t),t\right)\leq V_{L,B}$ for all $t\geq T$ provided that the minimum dwell-time condition in (24) is met. ■

Remark 6. Under Assumptions 1-3, the optimal value function can be shown to be the unique positive definite solution of the HJB equations. In this paper, the approximation of the positive definite solution to the HJB is guaranteed by appropriately selecting initial weight estimates and Lyapunov-based update laws [44].

IV. Simulation

To demonstrate the performance of the developed method, the ADP controller is applied to a family of dynamical systems. The simulation is performed on the control-affine systems in (25)-(27). The dynamic models are based on the continuous-time F-16 longitudinal dynamics from [45]. Since the dynamics are linear, the value functions are quadratic, hence Assumption 6 is satisfied. The dynamics of the first mode are

$$\dot{x}=\begin{bmatrix}-1&0.9&-0.002\\0.8&-1.1&-0.2\\0&0&-1\end{bmatrix}x+\begin{bmatrix}0\\0\\1\end{bmatrix}u,\tag{25}$$

the dynamics of the second mode are

$$\dot{x}=\begin{bmatrix}-0.8&0.2&-0.01\\0.6&-1.3&-0.1\\0&0&-1\end{bmatrix}x+\begin{bmatrix}0\\0\\0.5\end{bmatrix}u,\tag{26}$$

and the dynamics of the third mode are

$$\dot{x}=\begin{bmatrix}-1&0.5&-0.02\\0.9&-0.8&-0.4\\0&0&-1\end{bmatrix}x+\begin{bmatrix}0\\0\\1\end{bmatrix}u,\tag{27}$$

where $x=\begin{bmatrix}x_1, & x_2, & x_3\end{bmatrix}^T$, $x\in\mathbb{R}$ is measured in radians, and $u\in\mathbb{R}$. The initial condition is $x(0)=\begin{bmatrix}0.35 & 0.26 & -0.35\end{bmatrix}^T$.

The mode described by (25) is the closest to the dynamic model given in [45]. (26) and (27) vary from (25). A different mode was arbitrarily selected every 5 second as the active subsystem to highlight this method's ability to switch between different dynamical systems. A switching time of 5 seconds was chosen because it is larger than the calculated minimum required dwell-time for each system. The simulated switching sequence is $\{1,2,3,1,3,2\}$. The basis function is $\phi(x)=\begin{bmatrix}x_1^2, & x_1x_2, & x_1x_3, & x_2^2, & x_2x_3, & x_3^2\end{bmatrix}^T$.

Modes 1-3 have different cost matrices and gains, which alters V_k^*,and hence, the performance. The simulation parameters for each mode are listed in Table I.

Table I
SIMULATION PARAMETERS

Parameter	Mode 1	Mode 2	Mode 3
Q	diag([1, 1, 1])	diag([5, 5, 5])	diag([3, 3, 3])
R	0.5	2	1
$\overline{\Gamma}$	10^3	10^3	10^3
$\underline{\Gamma}$	500	500	50
λ	0.4	0.5	0.5
ν	0.005	0.005	0.005
η_{c1}	3	1	1
η_{c2}	5	2.5	5
η_{a1}	20	10	5
η_{a2}	1	0.75	1
N	10	10	10

Figure 1 illustrates that the system states are driven to the origin with an arbitrary switching sequence and sufficiently long dwell-time.

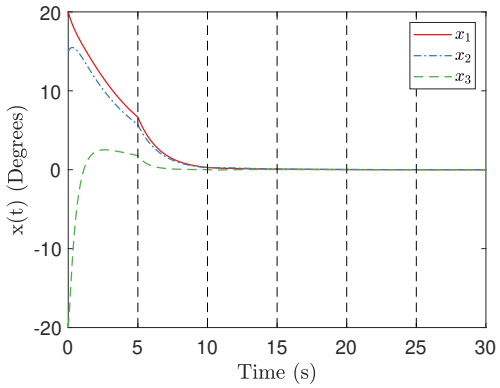

Figure 1. System states. The vertical dashed lines represent the time instances at which the mode was switched.

Since the dynamic systems are linear, the analytical value function can be determined by solving the Algebraic Riccati Equation (ARE). Solving the ARE provides a matrix which corresponds to the value function weights W_k, and, hence, the value functions $V_k^*(x)$. Figure 2 compares the value of the approximate value function to analytical value function while switching between modes.

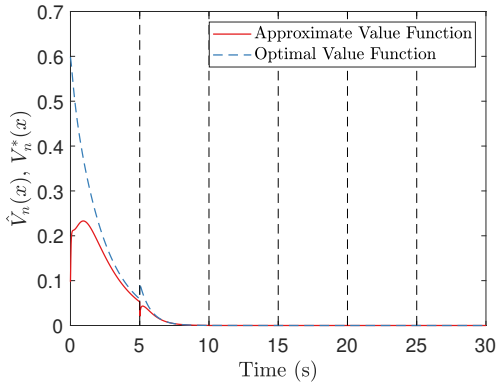

Figure 2. Comparison of the analytical value functions, $V_k^*(x)$, and the approximate value functions, $\hat{V}_k\left(x, \hat{W}_{c,k}\right)$. The vertical dashed lines represent the time instances at which the mode was switched.

Figure 3 presents the evolution of the critic weights $\hat{W}_{c,k}$, while switching. A mode's weights only update while that mode is active. If the mode is not active, the weight values do not change. Note that the weights of mode 1 and 2 converge before switching to another mode for the first time, while mode 3 is switched before it finishes learning. This illustrates that the weights do not need to be learned before the switching occurs.

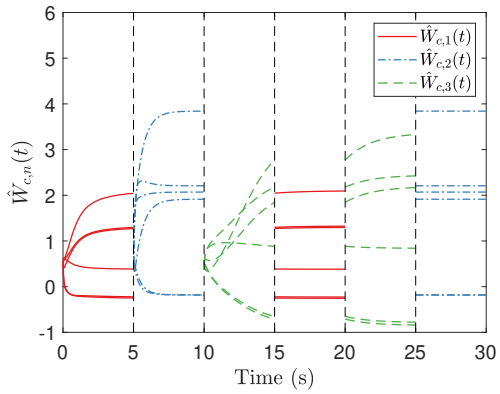

Figure 3. Critic weight estimates of each mode, $\hat{W}_{c,k}$. The vertical dashed lines represent the time instances at which the mode was switched.

V. CONCLUSION

A set of online approximate optimal controllers are developed for an arbitrary sequence of subsystems. Each controller is proven to regulate the state to within a neighborhood of the origin. Furthermore, the control policies are shown to converge to the neighborhood of the optimal policy using a Lyapunov-based analysis, while switching between different dynamic models and cost matrices. Simulation results show that switching according to an arbitrary sequence yields different performance. Future research will focus on generalizing the result to broader classes of systems.

REFERENCES

[1] 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.

[2] N. Stefanovic, M. Ding, and L. Pavel, "An application of L2 nonlinear control and gain scheduling to erbium doped fiber amplifiers," *Control Eng. Pract.*, vol. 15, pp. 1107–1117, 2007.

[3] A. A. Siranosian, M. Krstic, A. Smyshlyaev, and M. Bement, "Gain scheduling-inspired boundary control for nonlinear partial differential equations," *J. Dyn. Syst. Meas. Control*, vol. 133, p. 051007, 2011.

[4] H. Balini, J. Witte, and C. W. Scherer, "Synthesis and implementation of gain-scheduling and lpv controllers for an amb system," *Automatica*, vol. 48, no. 3, pp. 521–527, 2012.

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

[6] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Policy iterations on the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation," *IEEE Trans. Autom. Control*, vol. 51, pp. 1989–1995, Dec. 2006.

[7] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 38, pp. 943–949, 2008.

[8] T. Dierks, B. Thumati, and S. Jagannathan, "Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence," *Neural Netw.*, vol. 22, no. 5-6, pp. 851–860, 2009.

[9] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[10] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, pp. 89–92, Jan. 2013.

[11] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, pp. 621–634, Mar. 2014.

[12] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[13] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *Int. J. Control*, vol. 87, no. 3, pp. 553–566, 2014.

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

[15] R. Kamalapurkar, J. Rosenfeld, and W. E. Dixon, "Efficient model-based reinforcement learning for approximate online optimal control," *Automatica*, vol. 74, pp. 247–258, Dec. 2016.

[16] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 2009.

[17] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Conf. Decis. Control*, pp. 3598–3605, Dec. 2009.

[18] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control Algorithms and Stability*. Communications and Control Engineering, London: Springer-Verlag, 2013.

[19] D. Kirk, *Optimal Control Theory: An Introduction*. Mineola, NY: Dover, 2004.

[20] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2012.

[21] A. Heydari and S. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 145–157, 2013.

[22] D. Liu, Y. Huang, D. Wang, and Q. Wei, "Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming," *Int. J. Control*, vol. 86, no. 9, pp. 1554–1566, 2013.

[23] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, pp. 1167–1175, Apr. 2014.

[24] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, ch. An Actor-Critic-Identifier Architecture for Adaptive Approximate Optimal Control, pp. 258–278. IEEE Press Series on Computational Intelligence, Wiley and IEEE Press, 2012.

[25] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2017.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

[36] D. Wang and C. Mu, *Adaptive Critic Control with Robust Stabilization for Uncertain Nonlinear Systems*. Springer, 2019.

[37] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Adaptive critic designs for discrete-time zero-sum games with application to h-[infinity] control," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 37, pp. 240–247, 2007.

[38] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, 2013.

[39] H. K. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice Hall, 3 ed., 2002.

[40] J. Blot and P. Michel, "The value-function of an infinite-horizon linear-quadratic problem," *Appl. Math. Letters*, vol. 16, no. 1, pp. 71–78, 2003.

[41] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*. New York, NY, USA: John Wiley & Sons, 1995.

[42] T.-H. Cheng, *Lyapunov-Based Switched Systems Control*. PhD thesis, University of Florida, 2015.

[43] H. Lin and P. J. Antsaklis, "Asymptotic disturbance attenuation properties for uncertain switched linear systems," *Nonlin. Anal.: Hybrid Syst.*, vol. 4, pp. 279–290, 2010.

[44] 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*, to appear.

[45] 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.