Hierarchical Reinforcement Learning and Gain Scheduling-based Control of a Hypersonic Vehicle

Wanjiku A. Makumi*,
University of Florida, Gainesville, Florida 32611
Max L. Greene†,
Aurora Flight Sciences, A Boeing Company, Massachusetts 02142
Zachary I. Bell‡,
U.S. Air Force Research Laboratory, Eglin Air Force Base, Florida 32542
Brendan J. Bialy§,
U.S. Air Force Research Laboratory, Eglin Air Force Base, Florida 32542
Rushikesh Kamalapurkar¶
Oklahoma State University, Stillwater, Oklahoma 74078
and
Warren E. Dixon‖
University of Florida, Gainesville, Florida 32611

A hierarchical reinforcement learning-based control strategy is introduced to facilitate state regulation for a hypersonic vehicle. To account for time-varying aerothermoelastic parameters in real-time, a hierarchical switching policy selects a subsystem from a larger set of potential subsystems. The selection depends on an approximation of the optimal value function of each subsystem. Integral concurrent learning is used to approximate the parametric uncertainties in each dynamical system. The approximate optimal control policy is proven to converge to a neighborhood of the optimal control policy. Uniformly ultimately bounded stability of each subsystem and stability of the overall switched system are proven using a Lyapunov-based stability analysis and dwell-time analysis.

I. Introduction

Flight control design for hypersonic vehicles (HSVs) has several challenges. An extremely large operating envelope, various multi-objective constrained phases of flight, and complex interactions between the fluid, thermal, and structural dynamics of an HSV jeopardize traditional autopilot design methods [1]. Furthermore, the expense and difficulty associated with characterizing these interactions underscore the importance of autopilot design methods capable of maintaining strict performance metrics while operating in complex and highly uncertain environments. Results such as [2] and [3] use linear-parameter-varying (LPV) HSV models where the HSV drift dynamics and control effectiveness are represented by linear nominal matrices. Disturbance terms are added to represent the time-varying aerothermoelastic parameters that can result in vehicle instability.

Optimal control is often used for aircraft control to synthesize a stabilizing controller. Optimal control problems are used for complex systems by implementing the control across a large flight envelope using gain scheduling [4–6]. Linear gain-scheduled optimal control based on linear quadratic (LQ) methods is well-established and meets the performance and robustness design requirements on a point by point basis [7]. However, for large flight envelopes, questions may arise as to whether the resulting set of gain tables will be computationally tractable and sufficiently dense to maintain closed-loop performance and stability.

The optimal control policy depends on the optimal value function, which is the solution to the Hamilton Jacobi Bellman (HJB) equation. For linear quadratic regulator (LQR) problems, the HJB equation is reduced to the algebraic Riccati equation (ARE) [6]. Gain scheduling is based on flight envelope conditions, and, due to time-varying parameters

*Graduate Research Assistant, Department of Mechanical and Aerospace Engineering; makumiw@ufl.edu
†Aerospace Controls Researcher, Aurora Flight Sciences, A Boeing Company; greene.max@aurora.aero
‡Research Engineer, Munitions Directorate, Eglin Air Force Base; zachary.bell.10@us.af.mil
§Research Aerospace Engineer, Munitions Directorate, Eglin Air Force Base; brendan.bialy@us.af.mil.
¶Assistant Professor, Department of Mechanical and Aerospace Engineering; rushikesh.kamalapurkar@okstate.edu.
‖Professor, Department of Mechanical and Aerospace Engineering; wdixon@ufl.edu.

in the flight envelope, research has focused on switched control systems [8–10]. Switching has been investigated in the context of LQR problems in [11–13]. However, in many applications, most systems are considered to be unknown, i.e., the structure of the dynamics is known but contains parametric uncertainties. It is difficult to evaluate the optimal value function offline for unknown systems, thus motivating online and suboptimal adaptive methods.

This work provides a method for switching between subsystems online using the infinite horizon cost-to-go as a metric to measure performance. Each subsystem consists of its own unique control policy, cost function, and set of dynamics. Since the change in the dynamics resulting from aerothermoelastic effects can destabilize HSV controllers, the nominal plant must be updated to accurately reflect the time-varying parameters in the plant dynamics. Due to parameter uncertainties in the plant dynamics, the cost function that generates the most desirable state response for its corresponding nominal model is unknown *a priori*. Motivated by the fact that there will be different instantaneous cost functions in different regions of the flight envelope, and the fact that there is model uncertainty in the dynamics, a hierarchical reinforcement learning (HRL) framework is used to select the appropriate approximately optimal controller that minimizes each unknown system cost for the flight envelope. Since the infinite-horizon value function is the cost-to-go from using an optimal controller, a hierarchical agent is tasked with selecting the approximate optimal value function with the smallest cost-to-go. In the HRL architecture, the optimal value function approximations of several subsystems are compared, and the hierarchical agent selects the subsystem, consisting of the control policy and model dynamics, associated the lowest-valued cost-to-go to use in the control loop. Therefore, the control policy that results in the lowest-valued approximated cost-to-go is used as the applied control policy for its corresponding model at each time instance.

Due to the challenges of finding an analytical solution to the HJB equation, classical optimal control techniques are of limited use on complex systems. Reinforcement learning (RL) tools are commonly used to approximate solutions to optimal control problems. Approximate dynamic programming (ADP) is a RL method that uses an actor-critic framework to approximate the solution to the HJB equation, i.e., the optimal value function, online [14]. ADP is suitable for adaptive flight control applications that include system dynamics that contain parametric uncertainties. Actor and critic neural networks (NNs) are used to approximate the optimal control policy and optimal value function, respectively, in real-time.

The Bellman error (BE) is a metric that indirectly measures the quality of the optimal value function approximation. Continuous-time update laws update the weights to minimize the BE. By using a model-based formulation of ADP, the BE can be calculated at user-defined off-trajectory states to perform simulation of experience via BE extrapolation [14]. BE extrapolation yields simultaneous exploration and exploitation to facilitate improved online-learning. For unknown systems, an online data-driven system identifier can be used to estimate the system model. An integral concurrent learning (ICL)-based parameter identifier is used to approximate the unknown drift dynamics online. Previous results in [15] and [16] have used concurrent learning to approximate the uncertain parameters, but these methods require knowledge of the highest order state derivative which may not always be available. ICL eliminates the need to estimate the unmeasurable state derivatives [17].

A switched systems approach provides a framework for modeling changing aerothermoelastic parameters. Switched systems can be difficult to analyze because of discontinuities and instantaneous growth of the Lyapunov function(s) [18]. Switching between stable subsystems can lead to instability of the overall switched system; therefore, the stability of the switched system must be analyzed [19]. However, since each optimal value function, generally, is distinct between subsystems, and the optimal value function is contained within each Lyapunov function, a multiple Lyapunov function approach is necessary. One way to ensure stability in multiple Lyapunov function-based problems is via a dwell-time analysis [20, Ch. 3]. A previous result that addresses switched ADP using a minimum dwell-time analysis [21] contains constraints and assumptions that are unnecessarily restrictive. Therefore, a generalized Lyapunov-based dwell-time analysis is developed in this paper that can be applied to any switched uniformly ultimately bounded (UUB) stable subsystem that relaxes the constraints and assumptions in [21].

The idea of switching to account for the time varying parameters in HSVs was introduced for ADP in [22]. However, in [22], the dynamics were known, it was assumed that the linear model was updated at every switch without a method to facilitate the switching, and the analysis contained a restrictive dwell-time analysis. Motivated by the previous result in [22], this paper introduces a method for switching between unknown dynamics and uses a generalized dwell-time analysis. In this paper, an HRL agent is tasked with selecting the subsystem that facilitates the most desirable state trajectory for a given flight condition. The hierarchical framework allows for the value function approximations of multiple subsystems to be evaluated, and selects the subsystem which corresponds to the lowest-valued approximated cost-to-go, resulting in the most desirable switching pattern. A Lyapunov-based switched subsystem stability analysis proves UUB stability of the subsystems using multiple Lyapunov-like functions.

Notation

For notational brevity, time-dependence is omitted while denoting trajectories of the dynamical systems. For example, given the trajectories $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ and $y : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, the equation $f + h\left(y, t\right) = g\left(x\right)$ should be interpreted as $f\left(t\right) + h\left(\left(y\left(t\right)\right), t\right) = g\left(x\left(t\right)\right)$. The gradient $\left[\frac{\partial f(x,y)}{\partial x_1}^T, \ldots, \frac{\partial f(x,y)}{\partial x_n}^T\right]^T$ is denoted by $\nabla_x f\left(x, y\right)$. Unless otherwise specified, let $\nabla \triangleq \nabla_y$. Both the Euclidean norm for vectors and the Frobenius norm for matrices are denoted by $\|\cdot\|$. The cardinality of a set A is denoted by $|A|$. Let the subscript p define the quantity or function belonging to the p^{th} subsystem of the overall system. Let $p \in \mathcal{P}$, where $\mathcal{P} \subset \mathbb{N}$ and $|\mathcal{P}| < \infty$ represent a family of switched subsystems.

II. HSV Dynamics

Consider the nonlinear equations of motion for an HSV including aerothermoelastic effects and structural dynamics given in [23] as

$$\dot{V} = \frac{T\cos\left(\alpha\right) - D}{m} - g\sin\left(\theta - \alpha\right) \tag{1}$$

$$\dot{\alpha} = \frac{L + T\sin\left(\alpha\right)}{mV} + q + \frac{g}{V}\cos\left(\theta - \alpha\right) \tag{2}$$

$$\dot{q} = \frac{M}{I_{yy}} \tag{3}$$

$$\dot{h} = V\sin\left(\theta - \alpha\right) \tag{4}$$

$$\dot{\theta} = q \tag{5}$$

$$\ddot{\eta}_{s,i} = -2\zeta_{s,i}\omega_{s,i}\dot{\eta}_{s,i} - \omega_{s,i}^2\eta_{s,i} + N_i, \tag{6}$$

where $V \in \mathbb{R}$ denotes the forward velocity of the HSV, $T \in \mathbb{R}_{\geq 0}$ denotes the thrust, $\alpha \in \mathbb{R}$ denotes the angle of attack, $D \in \mathbb{R}_{\geq 0}$ denotes the drag, $h \in \mathbb{R}_{\geq 0}$ denotes the altitude, $m \in \mathbb{R}_{>0}$ denotes the HSV mass, $g \in \mathbb{R}_{>0}$ denotes the gravitational constant, $\theta \in \mathbb{R}$ denotes the pitch angle, $L \in \mathbb{R}_{\geq 0}$ denotes the lift, $q \in \mathbb{R}$ denotes the pitch rate, $M \in \mathbb{R}$ denotes the pitching moment about the HSV body y-axis, $I_{yy} \in \mathbb{R}_{>0}$ denotes the moment of inertia about the body y-axis, $\eta_{s,i} \in \mathbb{R}$ denotes the i^{th} flexible structural mode displacement for $i \in \{1, 2, 3\}$, $\zeta_{s,i}$, $\omega_{s,i}$, $N_{s,i} \in \mathbb{R}_{\geq 0}$ denote the damping factor, natural frequency, and generalized elastic forces of the i^{th} structural mode, respectively. The concatenated state vector $x \in \mathbb{R}^{11}$ includes the flight dynamic and structural dynamic states

$$x \triangleq \begin{bmatrix} \Delta V & \Delta\alpha & q & \Delta h & \Delta\theta & \eta_{s,1} & \dot{\eta}_{s,1} & \eta_{s,2} & \dot{\eta}_{s,2} & \eta_{s,3} & \dot{\eta}_{s,3} \end{bmatrix}^T,$$

where Δ denotes the difference between the state and its respective trim condition. The HSV aerodynamic and structural modes are coupled such that T, L, and D, depend on the structural modes $\eta_{s,i}$. The modulus of elasticity linearly decreases as the HSV temperature increases [24]. The change in the modulus of elasticity alters the structural damping ratio $\zeta_{s,i}$ and natural frequency $\omega_{s,i}$, which, in turn, significantly alters the structural dynamic responses (i.e., $\ddot{\eta}_{s,i}$) [24].

A. LPV Model

The aforementioned HSV dynamics are modeled as a controllable LPV system including uncertainty from unmodeled effects, given in [2] and [24] as

$$\dot{x} = A\left(\rho\left(t\right)\right)x + B\left(\rho\left(t\right)\right)u + d\left(t\right) \qquad y = Cx \tag{7}$$

where $A\left(\rho\left(t\right)\right) \in \mathbb{R}^{11\times 11}$ is the unknown LPV state matrix, $\rho\left(t\right) \in \mathbb{R}_{\geq 0}$ is the unknown time-dependent temperature profile of the HSV, $B\left(\rho\left(t\right)\right) \in \mathbb{R}^{11\times 2}$ is the LPV control effectiveness matrix, $u \in \mathbb{R}^2$ is the subsequently defined control

input, $d(t) \in \mathbb{R}^{11}$ is a time-varying uncertainty and disturbance term, $y \in \mathbb{R}^5$ is the measurable states, and $C \in \mathbb{R}^{5\times 11}$ is the output matrix. The control input $u \triangleq \begin{bmatrix} \delta_e & \delta_c \end{bmatrix}^T$ consists of the deflection angle of the elevators and canards from their trim condition. The HSV fuel equivalence ratio and diffuser area ratio are fixed at their operational trim condition. The outputs for this example are selected as $y \triangleq \begin{bmatrix} \Delta V & \Delta\alpha & q & \Delta h & \Delta\theta \end{bmatrix}^T$.

Let $p \in \mathcal{P} \subset \mathbb{N}$ represent the total number of nominal dynamic models. The state matrix and control effectiveness matrix are represented from [2] as

$$A(\rho(t)) = A_p + w_p(\rho(t)) \tag{8}$$

$$B(\rho(t)) = B_p + v_p(\rho(t)), \tag{9}$$

where $A_p \in \mathbb{R}^{11\times 11}$ is the p^{th} nominal state matrix, $B_p \in \mathbb{R}^{11\times 2}$ is the p^{th} nominal control effectiveness matrix, and $w_p(\rho(t)) \in \mathbb{R}^{11\times 11}$ is an unknown parameter-varying disturbance term associated with the p^{th} nominal state matrix, and $v_p(\rho(t)) \in \mathbb{R}^{11\times 2}$ is an unknown parameter-varying disturbance term associated with the p^{th} nominal control effectiveness matrix. The matrices $A(\rho(t))$ and $B(\rho(t))$ are time-varying since the temperature profile $\rho(t)$ in (8) and (9) is time-varying.

B. Measurable States

The nominal HSV dynamics are represented by A_p and B_p; however, some states in x, such as the structural dynamic modes $\eta_{s,i}$ and $\dot{\eta}_{s,i}$ for all i, may not be measurable. To facilitate the subsequent analysis, the structural dynamic modes are upperbounded, and thus collected into the disturbances $d(t)$. To reflect this change, the matrices $A_{p,y} \in \mathbb{R}^{5\times 5}$, $B_{p,y} \in \mathbb{R}^{5\times 2}$, and $C_y \in \mathbb{R}^{5\times 5}$ represent the nominal HSV dynamics that correspond only to the measurable states. Recall, the output is defined as $y \triangleq \begin{bmatrix} \Delta V & \Delta\alpha & q & \Delta h & \Delta\theta \end{bmatrix}^T$. The disturbances are assumed to be negligible such that $\|d(t)\| = 0$, $\left\|w_p(\rho(t))\right\| = 0$, $\left\|v_p(\rho(t))\right\| = 0$ for all $p \in \mathcal{P}$. The reduced-order model used for the control development is

$$\dot{y} = C_y A_{p,y} y + C_y B_{p,y} u. \tag{10}$$

While this paper neglects disturbances; in practice these disturbances will be nonzero. Results such as [25] have investigated different update laws to account for the disturbances, but not in the context of switched ADP. Future work aims to explicitly compensate for nonzero disturbances in the switched ADP framework.

III. Control Development

A. Control Objective

The control objective is to solve the infinite-horizon optimal regulation problem online, i.e. find an optimal control policy u that minimizes the cost functional for the p^{th} subsystem

$$J_p(y,u) = \int_{t_0}^{\infty} y^T Q_p y + u^T R_p u \, d\tau, \tag{11}$$

while regulating the system states to the origin and to switch to the subsystem with the lowest-valued cost-to-go. In (11), $Q_p \in \mathbb{R}^{5\times 5}$ and $R_p \in \mathbb{R}^{2\times 2}$ are user-defined constant positive definite (PD) symmetric cost matrices. The infinite horizon value function (i.e. the cost-to-go) for the p^{th} mode $V_p^* : \mathbb{R}^5 \to \mathbb{R}_{\geq 0}$ is defined as

$$V_p^*(y) \triangleq \min_{u(\cdot)\in\mathcal{U}} \int_t^{\infty} y^T Q_p y + u^T R_p u \, d\tau, \tag{12}$$

where $\mathcal{U} \subseteq \mathbb{R}^2$ is the set of admissible controllers [14].

Remark 1. Each subsystem, generally has a different cost function made up of a state penalty function Q_p and a cost penalty matrix R_p, a different developed optimal control policy, and a different dynamic model. Due to the parametric uncertainty in (10), the developed method switches between multiple user-defined cost functions to assess which controller generates the most desirable behavior for the unknown parameters by selecting V_p^* with the lowest value in real time. The selected controller is applied to its corresponding set of dynamics, thus facilitating switching between dynamic models as the time-varying parameters change.

Assuming that the optimal value function is continuously differentiable for each $p \in \mathcal{P}$, then the optimal control policy $u_p^* : \mathbb{R}^5 \to \mathbb{R}^2$ is defined as

$$u_p^*(y) = -\frac{1}{2} R_p^{-1} \left(C_y B_{p,y}\right)^T \left(\nabla V_p^*(y)\right)^T. \tag{13}$$

The optimal value function and the optimal control policy satisfy the HJB equation

$$0 = \nabla V_p^*(y) \left(C_y A_{p,y} y + C_y B_{p,y} u_p^*\right) + y^T Q_p y + u_p^{*T} R_p u_p^*, \tag{14}$$

which has the boundary condition $V_p^*(0) = 0$.

B. Value Function Approximation

The solution to the HJB in 14, i.e., the optimal value function is difficult to find for LPV systems. Parametric methods, specifically NNs, can be used to approximate the optimal value function in real-time. Let $\Omega \subset \mathbb{R}^5$ be a compact set.* The optimal value function can be approximated with an NN in Ω by invoking the Stone-Weierstrass Theorem [26] to obtain

$$V_p^*(y) = W_p^T \phi_p(y) + \epsilon_p(y) \ \forall y \in \Omega, \tag{15}$$

where $W_p \in \mathbb{R}^{15}$ is a vector of unknown weights, $\phi_p : \mathbb{R}^5 \to \mathbb{R}^{15}$ is a user-defined vector of basis functions,† and $\epsilon_p : \mathbb{R}^5 \to \mathbb{R}$ is the bounded function reconstruction error. By substituting (15) into (13), the NN representation of the p^{th} mode optimal control policy in (13) becomes

$$u_p^*(y) = -\frac{1}{2} R_p^{-1} C_y B_{p,y} \left(\nabla \phi_p(y)^T W_p + \nabla \epsilon_p(y)\right)^T. \tag{16}$$

Assumption 1. There exists a set of known positive constants $\overline{W}, \overline{\phi}, \overline{\nabla\phi}, \overline{\epsilon}, \overline{\nabla\epsilon} \in \mathbb{R}_{>0}$ such that $\sup_{p\in\mathcal{P}} \|W_p\| \leq \overline{W}$, $\sup_{y\in\Omega,\, p\in\mathcal{P}} \|\phi_p(y)\| \leq \overline{\phi}$, $\sup_{y\in\Omega,\, p\in\mathcal{P}} \|\nabla\phi_p(y)\| \leq \overline{\nabla\phi}$, $\sup_{y\in\Omega,\, p\in\mathcal{P}} \|\epsilon_p(y)\| \leq \overline{\epsilon}$, and $\sup_{y\in\Omega,\, p\in\mathcal{P}} \|\nabla\epsilon_p(y)\| \leq \overline{\nabla\epsilon}$ for all p [27, Assumptions 9.1.c-e].‡

Assumption 2. The ideal weights W_p in (15) and (16) are unknown *a priori*; hence, an approximation of W is sought using actor and critic weight estimates.

The critic weight estimate vector $\hat{W}_{c,p} \in \mathbb{R}^{15}$ is used to derive the approximate optimal value function $\hat{V}_p : \mathbb{R}^5 \times \mathbb{R}^{15} \to \mathbb{R}$, defined as

$$\hat{V}_p\left(y, \hat{W}_{c,p}\right) \triangleq \hat{W}_{c,p}^T \phi_p(y). \tag{17}$$

The actor weight estimate vector $\hat{W}_{a,p} \in \mathbb{R}^{15}$ is used to derive the approximate optimal control policy $\hat{u}_p : \mathbb{R}^5 \times \mathbb{R}^{15} \to \mathbb{R}$, defined as

$$\hat{u}_p\left(y, \hat{W}_{a,p}\right) \triangleq -\frac{1}{2} R_p^{-1} \left(C_y B_{p,y}\right)^T \left(\nabla \phi_p(y)^T \hat{W}_{a,p}\right). \tag{18}$$

IV. Hierarchical Agent

A. Switching Algorithm

The hierarchical agent uses the value function approximations of several suboptimal lower-level controllers as a metric to select the applied feedback control policy and the active nominal model. The hierarchical agent selects the

*The subsequent stability analysis guarantees that if y is initialized in an appropriately-sized subset of Ω, then it will stay in Ω.

†For brevity, each subsystem has the same number of elements in the basis function vector L.

‡The assumption can be met, for example, by selecting polynomials as basis functions (see [28, Theroem 1.5]).

subsystem associated with the lowest-valued approximated cost-to-go at each instance in time using the switching signal

$$\sigma \triangleq \text{argmin}_{p\in\mathcal{P}} \left\{\hat{V}_p\left(y, \hat{W}_{c,p}\right)\right\}. \tag{19}$$

The switching signal in (19) facilitates switching in real time and outputs the number of the subsystem that corresponds to the lowest-valued approximated cost-to-go. The optimal value function approximations for all the individual ADP subsystems $p \in \mathcal{P}$ are quantitatively compared in real-time, and the applied control input u is selected as

$$u = \hat{u}_\sigma\left(y, \hat{W}_{a,p}\right). \tag{20}$$

The selected control policy will be implemented as the controller in the active nominal model at each instance in time, as seen in Figure 1.

B. System Identification

System identification is used to approximate the uncertain parameters in the drift dynamics. To facilitate online system identification, let $C_y A_{p,y} y = Y_p(y)\,\theta_p$ where $Y_p : \mathbb{R}^5 \to \mathbb{R}^{5\times 4}$ is the known regression matrix, and $\theta_p \in \mathbb{R}^4$ is a vector of constant unknown parameters. Let the approximation of the uncertain drift dynamics $C_y A_{p,y} y$ be denoted as $C_y \hat{A}_p y$ and be defined as $C_y \hat{A}_{p,y} y \triangleq Y_p(y)\,\hat{\theta}_p$, where $\hat{\theta}_p \in \mathbb{R}^4$ is an approximation of the unknown parameter vector θ_p. The parameter estimate $\hat{\theta}_p$ is updated using the following ICL-based update law [17]

$$\dot{\hat{\theta}}_p \triangleq k_{p,ICL}\Gamma_{p,\theta}\sum_{j=1}^{M} \mathcal{Y}_{p,j}^T\left(y\left(t_j\right) - y\left(t_j - \Delta t\right) - \mathcal{U}_{pj} - \mathcal{Y}_{p,j}\hat{\theta}_p\right), \tag{21}$$

where $k_{p,ICL} \in \mathbb{R}_{>0}$ and $\Gamma_{p,\theta} \in \mathbb{R}^{4\times 4}$ are user-selected PD constants, $\mathcal{Y}_{p,j} \triangleq \mathcal{Y}_p\left(t_j\right)$, $\mathcal{U}_{p,j} \triangleq \mathcal{U}_p\left(t_j\right)$, $\mathcal{Y}_p(t) \triangleq \int_{\max[t-\Delta t,0]}^{t} Y_p\left(y(\tau)\right) d\tau$, and $\mathcal{U}_p(t) \triangleq \int_{\max[t-\Delta t,0]}^{t} C_y B_{p,y} u(\tau)\, d\tau$. The parameter update law in (21) can be rewritten in an analytical form as

$$\dot{\hat{\theta}}_p = k_{p,ICL}\Gamma_{p,\theta}\sum_{j=1}^{M} \mathcal{Y}_{p,j}^T \mathcal{Y}_{p,j}\tilde{\theta}_p, \tag{22}$$

where $\tilde{\theta}_p \triangleq \theta_p - \hat{\theta}_p$ is the parametric error.

Assumption 3. *A history stack of recorded state and control inputs $\left\{y\left(t_j\right), u\left(t_j\right)\right\}_{j=1}^{M_p}$ is available that satisfies $\underline{\mathcal{Y}}_p \triangleq \lambda_{\min}\left\{\sum_{j=1}^{M} \mathcal{Y}_{p,j}^T \mathcal{Y}_{p,j}\right\} > 0$ and ensures the finite excitation condition in [17] is satisfied a priori for all subsystems $p \in \mathcal{P}$.*

V. Bellman Error

The BE is a measure of suboptimality representing how close the actor and critic weight estimates are to their ideal weight values. By substituting (17), (18), and the approximated drift dynamics $C_y \hat{A}_p y$ into (14), the BE $\hat{\delta}_p : \mathbb{R}^5 \times \mathbb{R}^{15} \times \mathbb{R}^{15} \times \mathbb{R}^4 \to \mathbb{R}$ is defined as

$$\hat{\delta}_p\left(y, \hat{W}_{c,p}, \hat{W}_{a,p}, \hat{\theta}\right) \triangleq y^T Q_p y + \hat{u}_p\left(y, \hat{W}_{a,p}\right)^T R_p \hat{u}_p\left(y, \hat{W}_{a,p}\right) + \nabla\hat{V}_p\left(y, \hat{W}_{c,p}\right)\left(Y_p(y)\,\hat{\theta}_p + C_y B_{p,y}\hat{u}_p\left(y, \hat{W}_{a,p}\right)\right). \tag{23}$$

To facilitate the subsequent stability analysis, the BE can also be expressed in terms of the mismatch between the estimates and the ideal values defined as $\tilde{W}_{c,p} \triangleq W_p - \hat{W}_{c,p}$ and $\tilde{W}_{a,p} \triangleq W_p - \hat{W}_{a,p}$. Subtracting (14) from (23) and substituting (15)-(18), the analytical form of the BE in (23) can be expressed as

$$\hat{\delta}_p\left(y, \hat{W}_{c,p}, \hat{W}_{a,p}, \hat{\theta}\right) = -\omega_p^T \tilde{W}_{c,p} - W_p^T \nabla\phi_p Y_p(y)\,\tilde{\theta}_p + \frac{1}{4}\tilde{W}_{a,p}^T G_{\phi,p}(y)\,\tilde{W}_{a,p} + O_p(y), \tag{24}$$

where $\omega_p\left(y, \hat{W}_{a,p}, \hat{\theta}\right) \triangleq \nabla\phi_p(y)\left(Y_p(y)\hat{\theta}_p + C_y B_{p,y}\hat{u}_p\left(y, \hat{W}_{a,p}\right)\right)$ and $\Theta_p(y) \triangleq \frac{1}{2}\nabla\epsilon_p(y) G_{R,p}\nabla\phi_p(y)^T W_p + \frac{1}{4}G_{\epsilon,p} - \nabla\epsilon_p(y) C_y A_{p,y} y$. The functions $G_{R,p}$, $G_{\phi,p}$, and $G_{\epsilon,p}$ are defined as $G_{R,p}(y) \triangleq C_y B_{p,y} R_p^{-1}\left(C_y B_{p,y}\right)^T$, $G_{\phi,p}(y) \triangleq \nabla\phi_p(y) G_{R,p}(y) \nabla\phi_p(y)^T$, and $G_{\epsilon,p}(y) \triangleq \nabla\epsilon_p(y) G_{R,p}(y) \nabla\epsilon_p(y)^T$ respectively.

As described in [15], the BE in (23) can be calculated along any set of off-trajectory points in the state space using a user-selected state x_i, the critic weight estimate $\hat{W}_{c,p}$, the actor weight estimate $\hat{W}_{a,p}$, and the estimate of the system model from the aforementioned online system identifier via BE extrapolation. BE extrapolation yields simultaneous exploration and exploitation, enabling faster policy learning. To facilitate sufficient exploration, the BE is extrapolated from the user-defined off-trajectory points $\{y_i : y_i \in \Omega\}_{i=1}^{N_p}$, where $N_p \in \mathbb{N}$ denotes a user-specified total number of extrapolation points in the compact set Ω. Each subsystem p has its own distinct set of gain values, data, and update laws.

Assumption 4. On the compact set, Ω, a finite set of user-selected, off-trajectory points $\{y_i : y_i \in \Omega\}_{i=1}^{N_p}$ exists such that $0 < \underline{c}_p \triangleq \inf_{t\in\mathbb{R}_{\geq 0}} \lambda_{\min}\left\{\frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}\omega_{i,p}^T}{\rho_{i,p}^2}\right\}$ for all $p \in \mathcal{P}$, where $\rho_{i,p} = 1 + \nu_p \omega_{i,p}^T \Gamma_p \omega_{i,p}$, $\nu_p \in \mathbb{R}_{>0}$ is a user-defined gain, $\Gamma_p : \mathbb{R}^{15\times 15}$ is a time-varying least-squares gain matrix, and $\underline{c}_p$ is a constant scalar lower bound of the value of each input-output data pair's minimum eigenvalues for the p^{th} subsystem [15].

VI. Update Laws for Actor and Critic Weights

The actor and critic weights for each subsystem are updated simultaneously via BE error extrapolation. Using the extrapolated BE error, the actor and critic weight estimates for the p^{th}subsystem are updated according to the critic update law for the p^{th} mode $\dot{\hat{W}}_{c,p} \in \mathbb{R}^{15}$,

$$\dot{\hat{W}}_{c,p} \triangleq -\eta_{c,p}\Gamma\frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}}{\rho_{i,p}}\delta_{i,p}, \tag{25}$$

the actor update law for the p^{th} mode $\dot{\hat{W}}_{a,p} \in \mathbb{R}^{15}$,

$$\dot{\hat{W}}_{a,p} \triangleq -\eta_{a1,p}\left(\hat{W}_{a,p} - \hat{W}_{c,p}\right) - \eta_{a2,p}\hat{W}_{a,p} + \eta_{c,p}\frac{1}{N_p}\sum_{i=1}^{N_p}\frac{G_{\phi i,p}^T \hat{W}_{a,p}\omega_{i,p}^T}{4\rho_{i,p}}\hat{W}_{c,p}, \tag{26}$$

and the least-squares gain matrix update law of the p^{th} mode $\dot{\Gamma}_p \in \mathbb{R}^{15\times 15}$,

$$\dot{\Gamma}_p \triangleq \left(\lambda_p\Gamma_p - \frac{\eta_{c,p}\Gamma_p}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}\omega_{i,p}^T\Gamma_p}{\rho_{i,p}^2}\right)\cdot \mathbf{1}_{\left\{\underline{\Gamma}_p \leq \|\Gamma_p\| \leq \overline{\Gamma}_p\right\}}, \tag{27}$$

where $\eta_{c,p}$, $\eta_{a1,p}$, $\eta_{a2,p}$, $\lambda_p \in \mathbb{R}_{>0}$ are positive constant adaptation gains, $\underline{\Gamma}_p$, $\overline{\Gamma}_p \in \mathbb{R}_{>0}$ denote lower and upper bounds for Γ_p, and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function.

The update laws in (25)-(27) are always active for each subsystem regardless of whether a subsystem is active or inactive. Therefore, the update laws will update each subsystem p's weight estimates and least-squares gain matrix even if subsystem p is not active. Convergence of the states of each subsystem can be proven concurrently since the update laws are simultaneously learning for each subsystem.

VII. Stability Analysis

In a switched system, the stability of the individual subsystems does not guarantee stability of the overall switched system [19]. Hence, the switching signal must be properly designed to ensure the stability of the overall switched system. In addition to the stability of the individual subsystems, the stability of the switched system must be analyzed. The following development analyzes the dynamics in (10) using the control policy in (20) and the update laws in (22), (25)-(27).

A. Subsystem Stability Analysis

To facilitate the stability analysis, let $z \triangleq \left[y^T, \tilde{W}_{c,1}^T, \ldots, \tilde{W}_{c,p}^T, \tilde{W}_{a,1}^T, \ldots, \tilde{W}_{a,p}^T, \tilde{\theta}_1^T, \ldots, \tilde{\theta}_p^T\right]^T$ denote a concatenated state, and let $V_{L,p} : \mathbb{R}^{5+34|\mathcal{P}|} \to \mathbb{R}_{\geq 0}$ be a candidate Lyapunov function for the p^{th} mode defined as

$$V_{L,p}(z) \triangleq V_p^*(y) + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{W}_{c,p}^T \Gamma_p^{-1} \tilde{W}_{c,p} + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{W}_{a,p}^T \tilde{W}_{a,p} + \frac{1}{2}\sum_{p\in\mathcal{P}} \tilde{\theta}_p^T \Gamma_{\theta,p}^{-1} \tilde{\theta}_p. \tag{28}$$

According to [14, Lemma 4.3], (28) can be bounded as $\alpha_{1,p}(\|z\|) \leq V_{L,p}(z) \leq \alpha_{2,p}(\|z\|)$ using class $\mathcal{K}$ functions $\alpha_{1,p}, \alpha_{2,p} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. The normalized regressors $\frac{\omega_p}{\rho_p}$ and $\frac{\omega_{i,p}}{\rho_{i,p}}$ are 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 $\sup_{t\in\mathbb{R}_{\geq 0}} \left\|\frac{\omega_{i,p}}{\rho_{i,p}}\right\| \leq \frac{1}{2\sqrt{\nu_p \underline{\Gamma}_p}}$ for all $y \in \Omega$ and $y_i \in \Omega$, respectively. The function $G_{R,p}$ is bounded as $\sup_{y\in\Omega} \|G_{R,p}\| \leq \overline{G}_p^2 \lambda_{\max}\left\{R_p^{-1}\right\}$, $G_{\phi,p}$ is bounded as $\sup_{y\in\Omega} \|G_{\phi,p}\| \leq \left(\overline{\nabla\phi G}_p\right)^2 \lambda_{\max}\left\{R_p^{-1}\right\}$, and $Y(y)$ is bounded as $\sup_{y\in\Omega} \|Y(y)\| \leq \overline{Y}$. To facilitate the subsequent analysis, define $r \in \mathbb{R}_{>0}$ to be the radius of a compact ball $\mathcal{B}_r \in \mathbb{R}^{5+34|\mathcal{P}|}$ centered at the origin.

Theorem 1. *The state y, every critic weight estimate error $\tilde{W}_{c,p}$ $\forall p \in \mathcal{P}$, every actor weight estimate error $\tilde{W}_{a,p}$ $\forall p \in \mathcal{P}$, and every parameter estimation error $\tilde{\theta}_p$ $\forall p \in \mathcal{P}$ are UUB while each subsystem is active, provided the control policy in (18) is used, the weight update laws in (25)-(27) are implemented, Assumptions (1)-(4) hold, and the conditions*

$$\eta_{a1,p} + \eta_{a2,p} \geq \frac{5}{4\sqrt{\nu_p \underline{\Gamma}_p}} \eta_{c,p} \overline{W G_{\phi,p}} \tag{29}$$

$$\underline{c}_p \geq 3\frac{\eta_{a1,p}}{\eta_{c,p}} + \frac{3\eta_{c,p}^2 \overline{W}^2}{4\eta_{c,p}\nu_p \underline{\Gamma}_p}\left(\frac{\overline{\nabla\phi}^2 \overline{Y}^2}{k_{ICL}\underline{\mathcal{Y}}} + \frac{5\overline{G_{\phi,p}}^2}{16\left(\eta_{a1,p} + \eta_{a2,p}\right)}\right) \tag{30}$$

$$v_{L,p}^{-1}\left(L_p\right) < \alpha_{2,p}^{-1}\left(\alpha_{1,p}(r)\right) \tag{31}$$

are satisfied for each individual subsystem, where L_p is a positive constant that depends on the NN bounding constants in Assumption 1. Therefore, each control policy $\hat{u}_p$ converges to a neighborhood of its respective optimal control policy u_p^*.

Proof. Using the HJB equation in (14), the BE in (24), the gain conditions in (29) and (30), and the weight update laws in (25)-(27), the time derivative of (28) can be bounded as

$$\dot{V}_{L,p} \leq -v_{L,p}(\|z\|) \quad \forall \|z\| \geq v_{L,p}^{-1}\left(L_p\right) \tag{32}$$

for all $p \in \mathcal{P}$ and $t \in \mathbb{R}_{>0}$, where

$$v_{L,p} \triangleq \frac{1}{2}\underline{q}_p \|y\|^2 + \sum_{p\in\mathcal{P}}\left[\frac{1}{12}\eta_{c,p}\underline{c}_p \left\|\tilde{W}_{c,p}\right\|^2 + \frac{1}{20}\left(\eta_{a1,p} + \eta_{a2,p}\right)\left\|\tilde{W}_{a,p}\right\|^2 + \frac{1}{4}k_{ICL,p}\underline{\mathcal{Y}_p}\left\|\tilde{\theta}_p\right\|^2\right]. \tag{33}$$

Using (32), $v_{L,p}(\|z\|)$, and (31), [29, Theorem 4.18] can be invoked to conclude that z is UUB such that $\limsup_{t\to\infty} \|z\| \leq \alpha_{1,p}^{-1}\left(\alpha_{2,p}\left(v_{L,p}^{-1}\left(L_p\right)\right)\right)$ and the control policy $\hat{u}_p$ converges to a neighborhood of the optimal control policy u_p^*. Since $z \in \mathcal{L}_\infty$, it follows that $y, \tilde{W}_{c,1}, ..., \tilde{W}_{c,|\mathcal{P}|}, \tilde{W}_{a,1}, ..., \tilde{W}_{a,|\mathcal{P}|}, \tilde{\theta}_1, ..., \tilde{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$; hence, $y, \hat{W}_{c,1}, \ldots, \hat{W}_{c,|\mathcal{P}|}, \hat{W}_{a,1}, \ldots, \hat{W}_{a,|\mathcal{P}|}, \hat{\theta}_1, ..., \hat{\theta}_{|\mathcal{P}|} \in \mathcal{L}_\infty$ and $u \in \mathcal{L}_\infty$. Additionally, every trajectory z that is initialized in the ball $\mathcal{B}_r$ is bounded such that $z \in \mathcal{B}_r$, $\forall t \in \mathbb{R}_{\geq 0}$, $\forall p \in \mathcal{P}$. Since $z \in \mathcal{B}_r$, the states $y, \tilde{W}_{c,1}, ..., \tilde{W}_{c,|\mathcal{P}|}, \tilde{W}_{a,1}, ..., \tilde{W}_{a,|\mathcal{P}|}, \tilde{\theta}_1, ..., \tilde{\theta}_{|\mathcal{P}|}$ similarly lie in a compact set. □

Remark 2. See [15] for insight into satisfying the gain conditions in (29) and (30). See [15, Algorithm 1] for insight into selecting the size of the compact set Ω.

B. Switched UUB Stability Analysis

Theorem 1 proves the stability of the individual subsystems, but does not guarantee stability of the overall switched system. The Lyapunov function for the switched system may instantaneously increase due to changes in the optimal value function and real-time updates of the weights. Since the unknown optimal value function $V_p^*(y)$ in (28) is different for each subsystem, the aforementioned UUB subsystems contain multiple Lyapunov functions, thus preventing (28) from being a common Lyapunov function. Due to switching between multiple Lyapunov functions, a dwell-time analysis is necessary to prove convergence of the overall system [20, Ch. 3]. The proof is available upon request.

VIII. Conclusion

This paper investigates the application of an HRL-based switched ADP control framework applied to a HSV vehicle to account for time-varying aerothermoelastic effects in the flight envelope. An HRL-based switching law is used to switch between subsystems by selecting the controller with the least approximated cost-to-go at each time instance. Each individual subsystem is proven to be UUB via a Lyapunov-based stability analysis, and the stability of the overall switched system is proven via a dwell-time analysis.

IX. Acknowledgments

This research is supported in part by AFOSR grant FA9550-19-1-0169, AFRL grant FA8651-21-F-1027, Office of Naval Research grant N00014-21-1-2481, and AFRL grant FA8651-21-F-1025. 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.

References

[1] Bolender, M. and Doman, D., "A Non-linear Model for the Longitudinal Dynamics of a Hypersonic Air-Breathing Vehicle," *Proc. AIAA Guid. Navig. Control Conf.*, San Francisco, CA, Aug. 2005.

[2] Lind, R., "Linear Parameter-Varying Modeling and Control of Structural Dynamics with Aerothermoelastic Effects," *J. Guid. Control Dynam.*, Vol. 25, No. 4, July 2002, pp. 733–739.

[3] Zhang, L., Nie, L., Cai, B., Yuan, S., and Wang, D., "Switched linear parameter-varying modeling and tracking control for flexible hypersonic vehicle," *Aerosp. Science and Technology*, Vol. 95, 2019, pp. 105445.

[4] Siranosian, A. A., Krstic, M., Smyshlyaev, A., and Bement, M., "Gain Scheduling-Inspired Boundary Control for Nonlinear Partial Differential Equations," *J. Dyn. Syst. Meas. Control*, Vol. 133, 2011, pp. 051007.

[5] Rugh, W. J. and Shamma, J. S., "Research on gain scheduling," *Automatica*, Vol. 36, No. 10, Oct. 2000, pp. 1401–1425.

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

[7] Lewis, F. L., Vrabie, D., and Syrmos, V. L., *Optimal Control*, Wiley, Hoboken, NJ, 3rd ed., 2012.

[8] Lu, B., Wu, F., and Kim, S., "Switching LPV control of an F-16 aircraft via controller state reset," *IEEE Trans Control Syst Technol*, Vol. 14, No. 2, 2006, pp. 267–277.

[9] Huang, B., Lu, B., Li, Q., and Tong, Y., "Average dwell time based smooth switching linear parameter-varying proportional-integral-derivative control for an f-16 aircraft," *IEEE Access*, Vol. 9, 2021, pp. 30979–30992.

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

[11] Zhang, W., Hu, J., and Abate, A., "On the Value Functions of the Discrete-Time Switched LQR Problem," *IEEE Trans. Autom. Control*, Vol. 54, No. 11, 2009, pp. 2669–2674.

[12] Balandat, M., Zhang, W., and Abate, A., "On infinite horizon switched LQR problems with state and control constraints," *Systems & Control Letters*, Vol. 61, No. 4, 2012, pp. 464–471.

[13] Zhang, W., Abate, A., and Hu, J., "Efficient suboptimal solutions of switched LQR problems," *Proc. Am. Control Conf.*, IEEE, 2009, pp. 1084–1091.

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

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

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

[17] Parikh, A., Kamalapurkar, R., and Dixon, W. E., "Integral Concurrent Learning: Adaptive Control with Parameter Convergence using Finite Excitation," *Int J Adapt Control Signal Process*, Vol. 33, No. 12, Dec. 2019, pp. 1775–1787.

[18] Parikh, A., Cheng, T.-H., Licitra, R., and Dixon, W. E., "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, 2018, pp. 2149–2156.

[19] Branicky, M., "Multiple Lyapunov functions and other analysis tools for switched and hybrid systems," *IEEE Trans. Autom. Control*, Vol. 43, 1998, pp. 475–482.

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

[21] Greene, M., Abudia, M., Kamalapurkar, R., and Dixon, W. E., "Model-based Reinforcement Learning for Optimal Feedback Control of Switched Systems," *Proc. IEEE Conf. Decis. Control*, 2020, pp. 162–167.

[22] Greene, M. L., Deptula, P., Bialy, B., and Dixon, W. E., "Model-Based Approximate Optimal Feedback Control of a Hypersonic Vehicle," *AIAA SCITECH*, Jan. 2022, AIAA 2022-0613.

[23] Williams, T., Bolender, M. A., Doman, D. B., and Morataya, O., "An Aerothermal Flexible Mode Analysis of a Hypersonic Vehicle," *AIAA Paper 2006-6647*, Aug. 2006.

[24] Wilcox, Z. D., MacKunis, W., Bhat, S., Lind, R., and Dixon, W. E., "Lyapunov-Based Exponential Tracking Control of a Hypersonic Aircraft with Aerothermoelastic Effects," *AIAA J. Guid. Control Dyn.*, Vol. 33, No. 4, 2010, pp. 1213–1224.

[25] Huang, Y., Wang, D., and Liu, D., "Bounded robust control design for uncertain nonlinear systems using single-network adaptive dynamic programming," *Neurocomputing*, Vol. 266, 2017, pp. 128–140.

[26] Stone, M. H., "The Generalized Weierstrass Approximation Theorem," *Math. Mag.*, Vol. 21, No. 4, 1948, pp. 167–184.

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

[28] Sauvigny, F., *Partial Differential Equations 1: Foundations and Integral Representations*, Springer Science & Business Media, 2012.

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

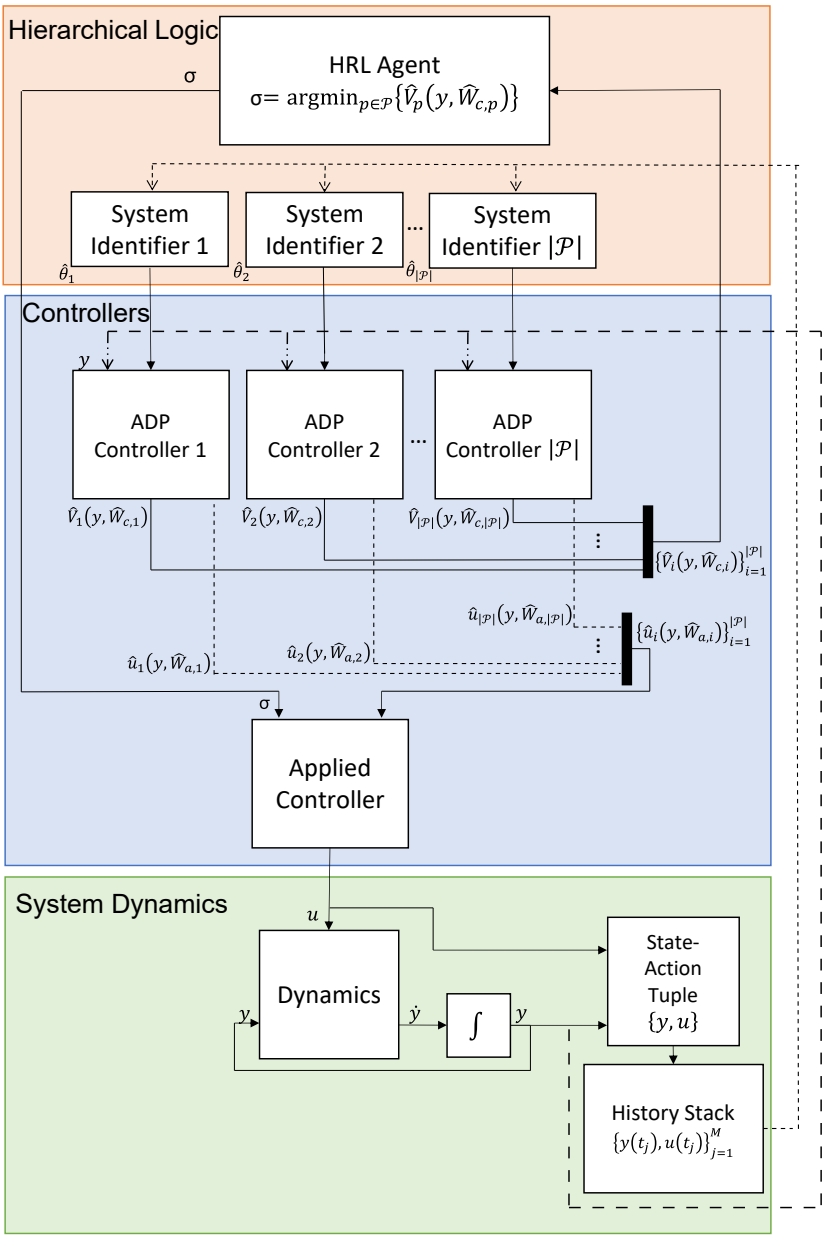

Figure 1 The hierarchical logic in the control loop contains the HRL agent and the system identifiers. The HRL agent compares the approximate optimal value function $\hat{V}_p$ for each control policy and returns the number of the subsystem with the lowest value function approximation. The system identifiers approximate the unknown dynamics of each subsystem to be used in the actor and critic weight estimates. The control execution level contains a family of ADP controllers, each containing a unique cost function, with the goal of minimizing each subsystem's respective cost-to-go. The control input with the lowest approximated cost-to-go is selected at the hierarchical level and applied to the corresponding dynamical system in (10). The history stack data is then provided to the system identifiers, the new state is provided to the ADP controllers, and the control policy in (20) is evaluated again.