Hierarchical Reinforcement Learning-based Supervisory Control of Unknown Nonlinear Systems ⋆

Wanjiku A. Makumi [*] **Max L. Greene** [**] **Zachary I. Bell** [***]
Scott Nivison [****] **Rushikesh Kamalapurkar** [†]
Warren E. Dixon [*]

[*] *Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA Email: makumiw, wdixon@ufl.edu*
[**] *Aurora Flight Sciences, A Boeing Company, Cambridge, MA, USA. Email: greene.max@aurora.aero*
[***] *U.S. Air Force Research Laboratory, Eglin Air Force Base, Florida, USA. Email: zachary.bell.10@us.af.mil*
[****] *Johns Hopkins University Applied Physics Laboratory, Fort Walton Beach, FL, USA Email: scott.nivison@jhuapl.edu*
[†] *Department of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA Email: rushikesh.kamalapurkar@okstate.edu*

Abstract: A supervisory control approach using hierarchical reinforcement learning (HRL) is developed to approximate the solution to optimal regulation problems for a control-affine, continuous-time nonlinear system with unknown drift dynamics. This result contains two objectives. The first objective is to approximate the optimal control policy that minimizes the infinite horizon cost function of each approximate dynamic programming (ADP) sub-controller. The second objective is to design a switching rule, by comparing the approximated optimal value functions of the ADP sub-controllers, to ensure that switching between subsystems yields a lower cost than using one subsystem. An integral concurrent learning-based parameter identifier approximates the unknown drift dynamics. Uniformly ultimately bounded regulation of the system states to a neighborhood of the origin, and convergence of the approximate control policy to a neighborhood of the optimal control policy, are proven using a Lyapunov-based stability and dwell-time analysis.

1. INTRODUCTION

Supervisory control methods provide alternatives to traditional continuously-tuned adaptive control laws and are useful when traditional control methodologies based on a single continuous controller do not provide satisfactory performance Battistelli et al. (2012). Switching between multiple controllers is orchestrated by a supervisory agent that uses data obtained to dictate the active control policy at each instance of time Battistelli et al. (2012). The key difference between supervisory switching control and standard adaptive algorithms based on continuous tuning is the use of higher-level logic to control the lower-level performance Hespanha (2001). Some of the first supervisory control results were developed in Morse (1996) and Morse (1997). Since then, the field of supervisory control has expanded Vu and Liberzon (2010); Chong et al. (2015); Leonessa et al. (2001).

⋆ 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 the sponsoring agency.

Supervised switching can be used in the context of optimality by using a hierarchy to optimize a certain performance index. The infinite horizon value function is a valuable metric to observe because it provides the cost-to-go of implementing its respective optimal controller Anderson and Moore (1971). Supervisory control approaches have been used to obtain optimality in Jing et al. (2021) and Pantelic and Lawford (2012). Many results in this field do not consider nonlinear systems because it is challenging to solve optimal control problems for nonlinear systems. However, recent advancements in Kamalapurkar et al. (2018); Jiang and Jiang (2017); Lewis and Liu (2013) have created a framework for approximating optimal control policies online, and these methods can be integrated into a supervisory control problem.

For unknown systems, i.e., the structure of the dynamics is known, but it contains unknown parametric uncertainties, the optimal value function cannot be determined offline; hence, there is a need to approximate it online. Due to the parametric uncertainties, it is difficult to know which controller yields the lowest cost for the system. The focus of this work is to develop a hierarchical agent that uses the value function approximation of several approximately

optimal lower-level controllers as a metric to select which controller should be active in the feedback loop; i.e, the hierarchical agent selects the controller associated with the least approximated cost-to-go at each instance in time.

For nonlinear systems, the Hamilton-Jacobi-Bellman (HJB) equation can be used to determine the optimal value function. However, there is not a general closed-form solution to the HJB for nonlinear systems. Therefore, approximate dynamic programming (ADP) has been developed as a method to approximate the solution to the HJB and has yet to be used in the context of supervisory control. ADP uses a reinforcement learning-based actor-critic framework to approximate the optimal value function (and hence, the optimal controller) in real-time Lewis and Liu (2013).

ADP uses a critic neural network (NN) to approximate the optimal value function and an actor NN to approximate the optimal control policy. The weights of the NN are adjusted online using the Bellman error (BE) as a performance metric. To facilitate improved online learning, the BE can be evaluated at user-defined, off-trajectory states within a compact set via BE extrapolation. BE extrapolation can provide simulation of experience by selecting an appropriate number of off-trajectory points using the system model. However, if the system model contains parametric uncertainties, then an estimate of the system model can be used. An integral concurrent learning (ICL)-based parameter identifier, as in Deptula et al. (2020), is used in the feedback loop of the hierarchical reinforcement learning (HRL) structure with the supervisory agent to identify the unknown drift dynamics online.

At each instance in time, the HRL closed-loop system switches between control policies, resulting in a switched system. In general, switched systems are challenging to analyze due to discontinuities and instantaneous growth of the Lyapunov function(s) Liberzon (2003). Switching between multiple stable subsystems can result in an unstable switched system; hence, a switched systems stability analysis is motivated Branicky (1998). Since optimal value functions are generally distinct between subsystems, and are a part of the candidate Lyapunov function for each subsystem, a common Lyapunov function cannot be constructed. Hence, a multiple Lyapunov function-based approach is motivated. One way to ensure stability using multiple Lyapunov functions is to ensure that each subsystem remains active for a minimum amount of time (i.e., a minimum dwell-time analysis Liberzon (2003, Ch. 3)) or to establish an upper bound on the number of switches in any given time interval (i.e., an average dwell-time analysis Liberzon (2003, Ch. 3)). While a switched ADP technique that uses a minimum dwell-time analysis is available in Greene et al. (2020), the analysis therein assumes that the optimal value function is upper and lower bounded by quadratic functions. In Greene et al. (2020, Assumption 6), a very restrictive bound on the optimal value function in the Lyapunov function is used to facilitate an exponential result; however, for general nonlinear systems, the assumption cannot be verified. The subsequent Lyapunov-based switched system stability analysis relaxes that previous assumption.

In this paper, an HRL-based framework is developed that uses a hierarchical supervisory control strategy to determine the policy corresponding to the lowest cost-to-go between ADP controllers and optimizes the corresponding subsystem. The hierarchical framework identifies which controller should be active at a given time and generates a switching signal indicating the most desirable switching pattern based on comparing multiple value function estimates. A Lyapunov-based dwell-time analysis is used to establish stability while relaxing the constraints and assumptions in Greene et al. (2020). The dwell-time analysis uses a novel Lyapunov-based stability theorem that is generally applicable to switched systems where all subsystems can be shown to be uniformly ultimately bounded (UUB) using multiple Lyapunov-like functions.

2. PROBLEM FORMULATION

Consider a continuous-time, control-affine nonlinear dynamical system

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

where $x \in \mathbb{R}^n$ denotes the system state trajectory, $u \in \mathbb{R}^m$ denotes the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift dynamics, and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ denotes the control effectiveness.

Assumption 1. The function f is an unknown locally Lipschitz function and $f(0) = 0$. Furthermore, $\nabla f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is continuous.

Assumption 2. The function g is a known locally Lipschitz function, bounded such that $0 < \|g(x)\| \leq \overline{g}\ \forall x \in \mathbb{R}^n$, where $\overline{g} \in \mathbb{R}_{>0}$ is the supremum over all x of the maximum singular values of $g(x)$.

2.1 Control Objective

Let $\mathcal{P} \subset \mathbb{N}$ with $\mathcal{P} < \infty$ represent a family of subsystems, and 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. The cost function

$$J_p(x, u_p) = \int_{t_0}^{\infty} Q_p(x) + u_p^T R_p u_p \; d\tau, \tag{2}$$

denotes the cost of running subsystem p the entire time. The cost function

$$J_{p(t)}(x, u) = \int_{t_0}^{\infty} Q_{p(t)}(x) + u_{p(t)}^T R_{p(t)} u_{p(t)} \; d\tau, \tag{3}$$

denotes the cost of switching between subsystems. 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 and to design the switching rule so that the cost in (3) is smaller than the cost in (2).

In (2), $Q_p : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a positive definite (PD) cost function where Q_p satisfies $\underline{q_p}(\|x\|) \leq Q_p(x) \leq \overline{q_p}(\|x\|)$ for $\underline{q_p}, \overline{q_p} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and $R_p \in \mathbb{R}^{m\times m}$ is a user-defined constant PD symmetric cost matrix.

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

$$V_p^*(x) \triangleq \min_{u \in U} \int_t^\infty Q_p(x) + u_p^T R_p u_p \, d\tau, \tag{4}$$

where $U \subseteq \mathbb{R}$ is the action space for u_p.

Remark 1. While each subsystem has the same set of dynamics, each has a different state penalty function Q_p, a different cost penalty matrix R_p and, thereby, a different respective controller. There are a user-defined number of cost functions that yield different desirable behavior, but since (1) is unknown a priori, supervised switching between the cost functions with different parameters will result in different expressions for (4), which motivates selecting the V_p^* with the lowest value for the specific unknown system.

Assumption 3. The optimal value function V_p^* is continuously differentiable for all $p \in \mathcal{P}$ Kamalapurkar et al. (2016).

The optimal value function is the solution to the corresponding HJB equation

$$0 = \nabla V_p^*(x) \left(f(x) + g(x) u_p^*\right) + Q_p(x) + u_p^{*T} R_p u_p^*, \tag{5}$$

where $u_p^* : \mathbb{R}^n \to \mathbb{R}^m$ is the optimal control policy for the p^{th} mode. The HJB equation in (5) has the boundary condition $V_p^*(0) = 0$. The optimal control policy u_p^* is defined as

$$u_p^*(x) = -\frac{1}{2} R_p^{-1} g(x)^T \left(\nabla V_p^*(x)\right)^T. \tag{6}$$

Remark 2. Under Assumptions 1-3, the optimal value function is the unique PD solution of the HJB equation for each system. The approximation of the PD solution to the HJB is guaranteed by the appropriate selection of Lyapunov-based update laws and initial weight estimates Deptula et al. (2020).

2.2 Value Function Approximation

The optimal control policy in (6) requires knowledge of the optimal value function, which is generally unknown for nonlinear systems. Let $\Omega \subset \mathbb{R}^n$ be a compact set. [1] Using the Universal Function Approximation Theorem, the optimal value function can be approximated with an NN in Ω as

$$V_p^*(x) = W_p^T \phi_p(x) + \epsilon_p(x) \; \forall x \in \Omega, \tag{7}$$

where $W_p \in \mathbb{R}^L$ is a vector of unknown weights, $\phi_p : \mathbb{R}^n \to \mathbb{R}^L$ is a user-defined vector of basis functions, [2] and $\epsilon_p : \mathbb{R}^n \to \mathbb{R}$ is the bounded function reconstruction error. Substituting (7) into (6), the NN representation of the p^{th} mode optimal control policy in (6) is

$$u_p^*(x) = -\frac{1}{2} R_p^{-1} g(x) \left(\nabla \phi_p(x) W_p + \nabla \epsilon_p(x)\right)^T. \tag{8}$$

Assumption 4. 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 $\sup_{p \in \mathcal{P}} \|W_p\| \leq \overline{W}$, $\sup_{x \in \Omega, \, p \in \mathcal{P}} \|\phi_p(x)\| \leq \overline{\phi}$, $\sup_{x \in \Omega, \, p \in \mathcal{P}} \|\nabla \phi_p(x)\| \leq \overline{\nabla\phi}$, $\sup_{x \in \Omega, \, p \in \mathcal{P}} \|\epsilon_p(x)\| \leq \overline{\epsilon}$, and $\sup_{x \in \Omega, \, p \in \mathcal{P}} \|\nabla \epsilon_p(x)\| \leq \overline{\nabla\epsilon}$ for all p Vrabie et al. (2013, Assumptions 9.1.c-e).

[1] The subsequent stability analysis guarantees that if x is initialized in an appropriately-sized subset of Ω, then it will stay in Ω.

[2] For brevity, each subsystem uses the same number of elements in the basis function vector L.

The critic weight estimate vector $\hat{W}_{c,p} \in \mathbb{R}^L$ is used to approximate (7), resulting in the optimal value function estimate $\hat{V}_p : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}$, defined as

$$\hat{V}_p\left(x, \hat{W}_{c,p}\right) \triangleq \hat{W}_{c,p}^T \phi_p(x). \tag{9}$$

The actor weight estimate vector $\hat{W}_{a,p} \in \mathbb{R}^L$ is used to approximate (8), resulting in the optimal control policy estimate $\hat{u}_p : \mathbb{R}^n \times \mathbb{R}^L \to \mathbb{R}^m$, defined as

$$\hat{u}_p\left(x, \hat{W}_{a,p}\right) \triangleq -\frac{1}{2} R_p^{-1} g(x)^T \left(\nabla \phi_p(x)^T \hat{W}_{a,p}\right). \tag{10}$$

3. HIERARCHICAL AGENT

3.1 Switching Rule

The hierarchical agent is tasked with identifying which policy minimizes the infinite horizon cost functional based on a switching policy. The supervisory algorithm

$$\sigma \triangleq \text{argmin}_{p \in \mathcal{P}} \left\{\hat{V}_p\left(x, \hat{W}_{c,p}\right)\right\} \tag{11}$$

returns the number of the subsystem associated with the smallest approximated cost-to-go, computed using estimates of the optimal value function corresponding to each subsystem. The switched signal in (11) will switch in real-time; therefore, to guarantee closed-loop stability of the overall system, a subsequently defined dwell-time condition must be satisfied. The optimal value function approximations are used to quantitatively compare all individual ADP controllers $p \in \mathcal{P}$ in real-time. The switching rule in (11) evaluates all of the approximated costs-to-go and selects the applied control input u in (1) as

$$u = \hat{u}_\sigma\left(x, \hat{W}_{a,p}\right), \tag{12}$$

that corresponds to the smallest optimal value function approximation at a given time. The goal is to determine which control policy provides the least approximate cost-to-go for the system.

3.2 System Identification

In addition to approximating the optimal value function for each subsystem, there is also uncertainty in the drift dynamics, and those uncertain parameters are approximated using system identification. To facilitate the online system identification, assume the drift dynamics f are linearly parameterizable such that $f(x) = Y(x)\theta$, where $Y : \mathbb{R}^n \to \mathbb{R}^{n \times s}$ is the known regression matrix and $\theta \in \mathbb{R}^s$ is a vector of constant unknown parameters. Let $\hat{\theta} \in \mathbb{R}^s$ be an approximation of the unknown parameter vector θ, which is updated according to the subsequently defined parameter update policy. The uncertain drift dynamics f are approximated by $\hat{f} : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$ which is defined as $\hat{f}\left(x, \hat{\theta}\right) \triangleq Y(x)\hat{\theta}$. [3] The parameter estimate $\hat{\theta}$ is updated with the ICL-based update policy Parikh et al. (2019)

[3] All subsystem controllers have the same dynamical system. The system parameters are being identified strictly in one drift dynamics model.

$$\dot{\hat{\theta}}(t) \triangleq k_{ICL}\Gamma_\theta \sum_{j=1}^{M} \mathcal{Y}_j^T \left(x(t_j) - x(t_j - \Delta t) - \mathcal{U}_j - \mathcal{Y}_j \hat{\theta} \right), \tag{13}$$

where $k_{ICL} \in \mathbb{R}_{>0}$ and $\Gamma_\theta \in \mathbb{R}^{s\times s}$ are user-selected PD constants, $\mathcal{Y}_j \triangleq \mathcal{Y}(t_j)$, $\mathcal{U}_j \triangleq \mathcal{U}(t_j)$, $\mathcal{Y}(t) \triangleq \int_{\max[t-\Delta t,0]}^{t} Y(x(\tau))\, d\tau$, and $\mathcal{U}(t) \triangleq \int_{\max[t-\Delta t,0]}^{t} g(x(\tau))\, u(\tau)\, d\tau$. The parameter update law in (13) can be rewritten in an analytical form as

$$\dot{\hat{\theta}} = k_{ICL}\Gamma_\theta \sum_{j=1}^{M} \mathcal{Y}_j^T \mathcal{Y}_j \tilde{\theta}, \tag{14}$$

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

Assumption 5. A history stack of recorded state and control inputs $\{x(t_j), u(t_j)\}_{j=1}^{M}$ is available that satisfies $\underline{\mathcal{Y}} \triangleq \lambda_{\min}\left\{\sum_{j=1}^{M} \mathcal{Y}^T{}_j \mathcal{Y}_j\right\} > 0$ and ensures the finite excitation condition in Parikh et al. (2019) is satisfied a priori. [4]

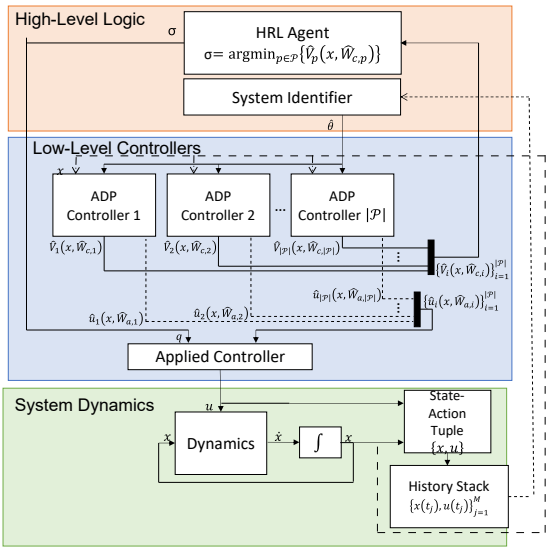

Fig. 1. The high-level logic in the hierarchical supervisory control architecture contains the HRL agent and the system identifier. The HRL agent evaluates the family of $\hat{V}_p$s and outputs the number of the subsystem with the lowest value function approximation. The system identifier approximates the uncertain model parameters that are used to update the actor and critic weight estimates. Each ADP controller contains a different cost function, and the objective is to minimize each subsystem's respective cost-to-go. The selected controller in (12) is applied to the dynamical system in (1). Then history stack data is provided to the high-level system identifier, the new state is provided to the low-level ADP controllers, and the policy in (12) is evaluated again.

[4] The *a priori* availability of the history stack is used for ease of exposition but is not necessary Kamalapurkar et al. (2016).

4. BELLMAN ERROR

The BE indicates how close the actor and critic weight estimates are to their ideal weight values. By substituting the approximate optimal value function $\hat{V}_p\left(x, \hat{W}_{c,p}\right)$ and approximate optimal control policy $\hat{u}_p\left(x, \hat{W}_{a,p}\right)$ into (5), the BE $\hat{\delta}_p : \mathbb{R}^n \times \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^s \to \mathbb{R}$ is defined as

$$\begin{aligned}\hat{\delta}_p\left(x, \hat{W}_{c,p}, \hat{W}_{a,p}, \hat{\theta}\right) &\triangleq Q_p(x) \\ &+ \hat{u}_p\left(x, \hat{W}_{a,p}\right)^T R_p \hat{u}_p\left(x, \hat{W}_{a,p}\right) \\ &+ \nabla \hat{V}_p\left(x, \hat{W}_{c,p}\right)\left(Y(x)\hat{\theta} + g(x)\hat{u}_p\left(x, \hat{W}_{a,p}\right)\right).\end{aligned} \tag{15}$$

While (15) is used for implementation, to facilitate the subsequent stability analysis, the BE can be expressed in terms of the weight approximation errors $\tilde{W}_{c,p} \triangleq W_p - \hat{W}_{c,p}$ and $\tilde{W}_{a,p} \triangleq W_p - \hat{W}_{a,p}$. Subtracting (5) from (15) and substituting (7)-(10), the analytical form of the BE in (15) can be expressed as

$$\begin{aligned}\hat{\delta}_p\left(x, \hat{W}_{c,p}, \hat{W}_{a,p}, \hat{\theta}\right) &= -\omega_p^T \tilde{W}_{c,p} - W_p^T \nabla\phi_p Y(x)\tilde{\theta} \\ &+ \frac{1}{4}\tilde{W}_{a,p}^T G_{\phi,p}(x)\tilde{W}_{a,p} + O_p(x),\end{aligned} \tag{16}$$

where $\omega_p : \mathbb{R}^n \times \mathbb{R}^L \times \mathbb{R}^s \to \mathbb{R}^n$ is $\omega_p\left(x, \hat{W}_{a,p}, \hat{\theta}\right) \triangleq \nabla\phi_p(x)\left(\hat{f}\left(x,\hat{\theta}\right) + g(x)\hat{u}_p\left(x,\hat{W}_{a,p}\right)\right)$ and $O_p(x) \triangleq \frac{1}{2}\nabla\epsilon_p(x) G_{R,p} \nabla\phi_p(x)^T W_p + \frac{1}{4} G_{\epsilon,p} - \nabla\epsilon_p(x) f(x)$. The functions $G_{R,p}$, $G_{\phi,p}$, and $G_{\epsilon,p}$ are defined as $G_{R,p}(x) \triangleq g_p(x) R_p^{-1} g_p(x)^T$, $G_{\phi,p}(x) \triangleq \nabla\phi_p(x) G_{R,p}(x) \nabla\phi_p(x)^T$, and $G_{\epsilon,p}(x) \triangleq \nabla\epsilon_p(x) G_{R,p}(x) \nabla\epsilon_p(x)^T$ respectively.

Bellman Error Extrapolation

As described in Kamalapurkar et al. (2016), the BE in (15) can be calculated at any user-defined point in the state space using a user-selected state x_i, the critic weight estimate $\hat{W}_{c,p}$, and the actor weight estimate $\hat{W}_{a,p}$. To estimate the value function over the compact set, the estimate of the system model from the aforementioned online system identifier is used to evaluate the BE along a set of off-trajectory points via BE extrapolation. BE extrapolation yields simultaneous exploration and exploitation, and can provide simulation of experience, enabling faster policy learning.

To facilitate sufficient exploration, the BE is extrapolated from the user-defined off-trajectory points $\{x_i : x_i \in \Omega\}_{i=1}^{N_p}$, where $N_p \in \mathbb{N}$ denotes a user-specified number of total extrapolation trajectories in the compact set Ω. Each subsystem p has its own distinct set of gain values, data, and update laws.

Assumption 6. On the compact set, Ω, a finite set of off-trajectory points $\{x_i : x_i \in \Omega\}_{i=1}^{N_p}$ are user-selected 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, $\Gamma_p : \mathbb{R}^{L\times L}$ 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 Kamalapurkar et al. (2016).

5. UPDATE LAWS FOR ACTOR AND CRITIC WEIGHTS

The actor and critic weights for each subsystem are updated simultaneously via BE error extrapolation. In the subsequent weight update laws, $\eta_{c,p}$, $\eta_{a1,p}$, $\eta_{a2,p}$, $\lambda_p \in \mathbb{R}_{>0}$ are positive constant adaptation gains, and $\underline{\Gamma}_p$, $\overline{\Gamma}_p \in \mathbb{R}_{>0}$ denote lower and upper bounds for Γ_p. The critic update law for the p^{th} mode $\dot{\hat{W}}_{c,p} \in \mathbb{R}^L$ is defined as

$$\dot{\hat{W}}_{c,p} \triangleq -\eta_{c,p}\Gamma \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}}{\rho_{i,p}}\delta_{i,p}. \tag{17}$$

The actor update law for the p^{th} mode $\dot{\hat{W}}_{a,p} \in \mathbb{R}^L$ is defined as

$$\dot{\hat{W}}_{a,p} \triangleq -\eta_{a1,p}\left(\hat{W}_{a,p} - \hat{W}_{c,p}\right) - \eta_{a2,p}\hat{W}_{a,p} + \eta_{c,p}\frac{1}{N_p}\sum_{i=1}^{N_p}\frac{G_{\phi i,p}^T \hat{W}_{a,p}\omega_{i,p}^T}{4\rho_{i,p}}\hat{W}_{c,p}. \tag{18}$$

The least-squares gain matrix update law of the p^{th} mode $\dot{\Gamma}_p \in \mathbb{R}^{L\times L}$ is defined as

$$\dot{\Gamma}_p \triangleq \left(\lambda_p\Gamma_p - \frac{\eta_{c,p}\Gamma_p}{N_p}\sum_{i=1}^{N_p}\frac{\omega_{i,p}\omega_{i,p}^T\Gamma_p}{\rho_{i,p}^2}\right)\cdot \mathbf{1}_{\left\{\underline{\Gamma}_p \le \|\Gamma_p\| \le \overline{\Gamma}_p\right\}}, \tag{19}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. [5] , [6]

The update laws in (17)-(19) are always active for each subsystem regardless of a subsystem's activity or inactivity. Hence, the update laws will update each subsystem p's weight estimates and least-squares gain matrix even if subsystem p is not active. Since the update laws are always learning for each subsystem, convergence of the states of each subsystem can be proven concurrently.

6. STABILITY ANALYSIS

6.1 Subsystem Stability Analysis

To facilitate the stability analysis, a concatenated state $z \in \mathbb{R}^{n+2L|\mathcal{P}|+s}$ is defined as $z \triangleq \left[x^T, \tilde{W}_{c,1}^T, \ldots, \tilde{W}_{c,p}^T, \tilde{W}_{a,1}^T, \ldots, \tilde{W}_{a,p}^T, \tilde{\theta}^T\right]^T$, and the candidate Lyapunov function $V_{L,p} : \mathbb{R}^{n+2L|\mathcal{P}|+s} \to \mathbb{R}_{\ge 0}$ is defined as

$$V_{L,p}(z) \triangleq V_p^*(x) + \frac{1}{2}\sum_{p\in\mathcal{P}}\tilde{W}_{c,p}^T\Gamma_p^{-1}\tilde{W}_{c,p} + \frac{1}{2}\sum_{p\in\mathcal{P}}\tilde{W}_{a,p}^T\tilde{W}_{a,p} + \frac{|\mathcal{P}|}{2}\tilde{\theta}^T\Gamma_\theta^{-1}\tilde{\theta}. \tag{20}$$

According to Kamalapurkar et al. (2018, Lemma 4.3), (20) can generally be bounded as $\alpha_{1,p}(\|z\|) \le V_{L,p}(z) \le \alpha_{2,p}(\|z\|)$ using class $\mathcal{K}$ functions $\alpha_{1,p}, \alpha_{2,p} : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$. The normalized regressors $\frac{\omega_p}{\rho_p}$ and $\frac{\omega_{i,p}}{\rho_{i,p}}$ are bounded as $\sup_{t\in\mathbb{R}_{\ge 0}}\left\|\frac{\omega_p}{\rho_p}\right\| \le \frac{1}{2\sqrt{\nu_p\underline{\Gamma}_p}}$ and $\sup_{t\in\mathbb{R}_{\ge 0}}\left\|\frac{\omega_{i,p}}{\rho_{i,p}}\right\| \le \frac{1}{2\sqrt{\nu_p\underline{\Gamma}_p}}$ for all $x \in \Omega$ and $x_i \in \Omega$, respectively. The function $G_{R,p}$ is bounded as $\sup_{x\in\Omega}\|G_{R,p}\| \le \overline{G}_p^2\lambda_{\max}\left\{R_p^{-1}\right\}$, $G_{\phi,p}$ is bounded as $\sup_{x\in\Omega}\|G_{\phi,p}\| \le \left(\overline{\nabla\phi G}_p\right)^2\lambda_{\max}\left\{R_p^{-1}\right\}$, and $Y(x)$ is bounded as $\sup_{x\in\Omega}\|Y(x)\| \le \overline{Y}$. To facilitate the subsequent analysis, define $r \in \mathbb{R}_{>0}$ to be the radius of a compact ball $\mathcal{B}_r \in \mathbb{R}^{n+2L|\mathcal{P}|+s}$ centered at the origin.

Theorem 1. Let $x(\cdot)$ denote the trajectory of the p^{th} subsystem for a fixed p. Provided the control policy in (10) is used, the weight update laws in (17)-(19) are implemented, Assumptions (1)-(6) hold, and the conditions

$$\eta_{a1,p} + \eta_{a2,p} \ge \frac{5}{4\sqrt{\nu_p\underline{\Gamma}_p}}\eta_{c2,p}\overline{WG_{\phi,p}} \tag{21}$$

$$\underline{c}_p \ge 3\frac{\eta_{a1,p}}{\eta_{c2,p}} + \frac{3\eta_{c,p}^2\overline{W}^2}{4\eta_{c,p}\nu_p\underline{\Gamma}_p}\left(\frac{\overline{\nabla\phi}^2\overline{Y}^2}{k_{ICL}\underline{\mathcal{Y}}} + \frac{5\overline{G_{\phi,p}}^2}{16(\eta_{a1,p}+\eta_{a2,p})}\right) \tag{22}$$

$$v_{L,p}^{-1}(L_p) < \alpha_{2,p}^{-1}(\alpha_{1,p}(r)) \tag{23}$$

are satisfied for each individual subsystem, where L_p is a positive constant that depends on the NN bounding constants in Assumption 4, then the state x, 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 the parameter estimation error $\tilde{\theta}$ are UUB. Hence, each control policy $\hat{u}_p$ converges to a neighborhood of its respective optimal control policy u_p^*.

The proof is available upon request.

Remark 3. See Kamalapurkar et al. (2016) for insight into satisfying the gain conditions in (21) and (22). See Kamalapurkar et al. (2016, Algorithm 1) for insight into selecting the size of the compact set Ω.

6.2 Switched UUB Stability Analysis

Since the unknown optimal value function $V_p^*(x)$ in (20) is different for each subsystem, (20) is not a common Lyapunov function. The previous theorem proves stability of the individual subsystems, but not stability of the overall switched system. The Lyapunov function for the switched system may instantaneously increase due to the increase in the optimal value function and the real-time updates of the weights. The HRL strategy includes switching between individually UUB subsystems with multiple Lyapunov functions; hence, a dwell-time analysis is necessary to prove the convergence of the overall system Liberzon (2003, Ch. 3).

Theorem 2. Let $\dot{x} = f_p(x,t)$ be a finite family of UUB subsystems and $V_p : \mathbb{R}^n \times \mathbb{R}_{\ge 0} \to \mathbb{R}$ be a family of corresponding Lyapunov-like functions that satisfy

$$\alpha_{1,p}(\|x\|) \le V_p(x,t) \le \alpha_{2,p}(\|x\|), \tag{24}$$

[5] The on-trajectory points can be included in the weight update laws, such as in Kamalapurkar et al. (2016), but to focus the Lyapunov-based analysis, only off-trajectory BE extrapolation is performed.

[6] Using (19) ensures that each $\underline{\Gamma}_p \le \|\Gamma_p\| \le \overline{\Gamma}_p$ for all $t \in \mathbb{R}_{>0}$.

$$\frac{\partial V_p}{\partial t} + \frac{\partial V_p}{\partial x}\left(f\left(x,t\right) + g\left(x,t\right)u_p\left(x,t\right)\right) \leq -W_p\left(x\right), \quad (25)$$

and

$$\max_{p\in\mathcal{P}} \alpha_{2,p}\left(\mu_p\right) < \min_{p\in\mathcal{P}} \alpha_{1,p}\left(r\right) \quad (26)$$

for all $x \in \Lambda_p$, $p \in \mathcal{P}$, and $t \geq 0$, where $\Lambda_p \triangleq \{x \,|\, 0 \leq \mu_p \leq \|x\| \leq r\}$, $\alpha_{1,p}$, $\alpha_{2,p} : [0,r] \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions, r is the radius of a compact ball $\mathcal{B}_r$, μ_p is the radius of a compact ball $\mathcal{B}_{\mu_p}$, and $W_p : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a continuous PD function. Let $c_1\left(t\right) \triangleq \frac{\alpha_{2,\sigma(t_i)}(\|x(t_i)\|)-\alpha_{1,\sigma\left(t_i^-\right)}(\|x(t_i)\|)}{\kappa}$. If the conditions in (24)-(26) and the minimum dwell-time condition

$$\tau\left(t_i\right) \quad \geq \quad \begin{cases} c_1\left(t\right) & \forall V_{\sigma(t_i)}\left(x\left(t_i\right),t_i\right) > \overline{\alpha} \\ > 0 & \forall V_{\sigma(t_i)}\left(x\left(t_i\right),t_i\right) \leq \overline{\alpha} \end{cases} \quad (27)$$

are satisfied for all $p \in \mathcal{P}$ and for every switching instant $t_i \in t_\sigma$, where V_j represents the Lyapunov function of the j^{th} subsystem, κ is a subsequently defined positive constant, and $t_i \in t_\sigma$ represents a general switching instance, then the trajectories of the switched system $\dot{x} = f\left(x,t\right) + g\left(x,t\right)u_p\left(x,t\right)$ initialized in the set $\left\{x \,|\, \|x\| \leq \min_{p,q\in\mathcal{P}} \alpha_{2,p}^{-1}\left(\alpha_{1,q}\left(r\right)\right)\right\}$ converge to a bounded region such that $\lim_{t\to\infty}\|x\left(t\right)\| \leq \max_{p,q\in\mathcal{P}} \alpha_{1,p}^{-1}\left(\alpha_{2,q}\left(\mu_q\right)\right)$.

The proof is available upon request.

6.3 Application to Switched ADP

As proved in Theorem 1, each individual subsystem is UUB; i.e., each subsystem satisfies (24) and (25). In addition, to apply Theorem 2, (26) and (27) must also be satisfied. Hence, following the switching policy in (11) and given that the dwell-time condition in (27) is satisfied, then Theorem 2 can be applied to show that $\limsup_{t\to\infty}\|z\left(t\right)\| \leq \max_{p,q\in\mathcal{P}} \alpha_{1,p}^{-1}\left(\alpha_{2,q}\left(\mu_q\right)\right)$. Moreover, since $z \in \mathcal{L}_\infty$, it follows that $x, \tilde{W}_{c,1}, \ldots, \tilde{W}_{c,|\mathcal{P}|}, \tilde{W}_{a,1}, \ldots, \tilde{W}_{a,|\mathcal{P}|}, \tilde{\theta} \in \mathcal{L}_\infty$; hence, $x, \hat{W}_{c,1}, \ldots, \hat{W}_{c,|\mathcal{P}|}, \hat{W}_{a,1}, \ldots, \hat{W}_{a,|\mathcal{P}|}, \hat{\theta} \in \mathcal{L}_\infty$ and $u \in \mathcal{L}_\infty$.

7. CONCLUSION

An HRL-based supervisory control strategy is developed to approximate solutions to multiple infinite horizon regulation problems for nonlinear continuous-time control-affine systems online. A supervisory switching policy is used to switch between the control policy with the least approximated cost-to-go in real-time. Stability of each subsystem is proven via a Lyapunov-based stability analysis. The overall switched system is proven to be stable in the sense that the system states converge to a neighborhood of the origin and the applied policy converges to a neighborhood of the selected optimal policy.

REFERENCES

Anderson, B.D. and Moore, J.B. (1971). *Optimal control: linear quadratic methods*. Courier Corp.

Battistelli, G., Hespanha, J., and Tesi, P. (2012). Supervisory control of switched nonlinear systems. *Int. J. Adapt. Control Signal Process.*, 26(8), 723–738.

Branicky, M. (1998). Multiple Lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE Trans. Autom. Control*, 43, 475–482.

Chong, M.S., Nesic, D., Postoyan, R., and Kuhlmann, L. (2015). Parameter and state estimation of nonlinear systems using a multi-observer under the supervisory framework. *IEEE Trans. Autom. Control*, 60(9), 2336–2349.

Deptula, P., Bell, Z., Doucette, E., Curtis, W.J., and Dixon, W.E. (2020). Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with control effectiveness faults. *Automatica*, 116, 1–10.

Greene, M., Abudia, M., Kamalapurkar, R., and Dixon, W.E. (2020). Model-based reinforcement learning for optimal feedback control of switched systems. In *Proc. IEEE Conf. Decis. Control*, 162–167.

Hespanha, J. (2001). Tutorial on supervisory control. *Lecture Notes for the workshop Control using Logic and Switching for the 40th Conf. on Decision and Contr.*

Jiang, Y. and Jiang, Z.P. (2017). *Robust Adaptive Dynamic Programming*. John Wiley & Sons.

Jing, G., Bai, H., George, J., and Chakrabortty, A. (2021). Model-free optimal control of linear multiagent systems via decomposition and hierarchical approximation. *IEEE Control Netw. Syst.*, 8(3), 1069–1081.

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

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

Leonessa, A., Haddad, W.M., and Chellaboina, V. (2001). Nonlinear system stabilization via hierarchical switching control. *IEEE Trans. Autom. Control*, 46(1), 17–28.

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

Liberzon, D. (2003). *Switching in Systems and Control*. Birkhauser.

Morse, A.S. (1996). Supervisory control of families of linear set-point controllers part I. exact matching. *IEEE Trans. Autom. Control*, 41(10), 1413–1431.

Morse, A.S. (1997). Supervisory control of families of linear set-point controllers part II. robustness. *IEEE Trans Autom. Contol*, 42(11), 1500–1515.

Pantelic, V. and Lawford, M. (2012). Optimal supervisory control of probabilistic discrete event systems. *IEEE Trans. Autom. Control*, 57(5), 1110–1124. doi: 10.1109/TAC.2011.2173420.

Parikh, A., Kamalapurkar, R., and Dixon, W.E. (2019). Integral concurrent learning: Adaptive control with parameter convergence using finite excitation. *Int J Adapt Control Signal Process*, 33(12), 1775–1787.

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

Vu, L. and Liberzon, D. (2010). Supervisory control of uncertain linear time-varying systems. *IEEE Trans. Autom. Control*, 56(1), 27–42.