Integration of adaptive control and reinforcement learning for real-time control and learning

Anuradha M. Annaswamy, Anubhav Guha, Yingnan Cui, Sunbochen Tang, Peter A. Fisher, and Joseph E. Gaudio

Abstract— This paper considers the problem of real-time control and learning in dynamic systems subjected to parametric uncertainties. We propose a combination of a Reinforcement Learning (RL) based policy in the outer loop suitably chosen to ensure stability and optimality for the nominal dynamics, together with Adaptive Control (AC) in the inner loop so that in real-time AC contracts the closed-loop dynamics towards a stable trajectory traced out by RL. Two classes of nonlinear dynamic systems are considered, both of which are control-affine. The first class of dynamic systems utilizes equilibrium points and a Lyapunov approach while second class of nonlinear systems uses contraction theory. AC-RL controllers are proposed for both classes of systems and shown to lead to online policies that guarantee stability using a high-order tuner and accommodate parametric uncertainties and magnitude limits on the input. In addition to establishing a stability guarantee with real-time control, the AC-RL controller is also shown to lead to parameter learning with persistent excitation for the first class of systems. Numerical validations of all algorithms are carried out using a quadrotor landing task on a moving platform.

I. INTRODUCTION

This paper considers the problem of real-time control and learning of a class of dynamic systems with parametric uncertainties using a combination of Adaptive Control (AC) and Reinforcement Learning (RL). Over the past seven decades, the field of AC has focused on ensuring global boundedness with asymptotic tracking, parameter learning with persistent excitation, and robustness to nonparametric uncertainties such as disturbances and unmodeled dynamics for a large class of linear and nonlinear dynamic systems [1]–[11]. The field of RL has focused on the determination of a sequence of inputs that drives a dynamical system to minimize a suitable objective with minimal knowledge of the system model. The central structure of the RL solution revolves around a policy that is learned so as to maximize a desired reward [12]–[15].

Both AC and RL-based control methods have addressed the problem of control in the presence of uncertainty with a component of learning, but with entirely different approaches. AC methods have been proven to be effective in the enforcement of

objectives for specific classes of problems, such as regulation and tracking, in real-time, with analytical guarantees. In addition to accommodating parametric uncertainties, constraints on the control input magnitude [16], [17] and rate [18] can also be incorporated within an analytical framework. They are however unable to directly guarantee the realization of long-term optimality-based objectives. RL- trained policies, on the other hand, can handle a broad range of objectives [15], where the control policies are learned offline in simulation, allowing for a near infinite number of agent-environment interactions to allow the policy to become near-optimal with guarantees of local stability [14], [19]–[23]. In practice, however, offline policies trained in simulation often exhibit degenerate performance when used for real-time control due to modeling errors that can occur online [24]–[28]. This paper takes a first step towards an integration of the AC and RL approaches so as to reduce this "sim-to-real" gap by realizing a combined set of advantages of both approaches.

The AC-RL approach that we propose in this paper uses AC-based components in the inner-loop and RL-based components in the outer-loop. The AC-RL controller is designed such that the inner-loop AC accommodates the effect of parametric uncertainties in real-time and contracts the closed-loop dynamics towards a reference system through a nonlinear adaptive law. The RL policy in the outerloop, which is used to determine the control input to the reference system, is designed to lead to an almost-optimal control policy for the reference system, through training in simulation. An important point to note is that this policy is chosen so that local stability is ensured, which can be accomplished by introducing additional components in the underlying loss function (ex. [23]). Two classes of nonlinear dynamic systems are considered, denoted as Problem 1 and Problem 2. The first class of dynamic systems utilizes equilibrium points and expansion forms around these points and employs a Lyapunov approach to establish convergence. The second class of nonlinear systems uses contraction theory as the underlying framework [29]. In each class, an AC- RL controller is proposed for a class of parametric uncertainties, and the resulting closed-loop system is proved to be globally bounded with asymptotic properties of tracking and cost. In both problems, the states of the dynamic system are assumed to be accessible for measurement.

In the first class of systems, the RL component in the outer loop is chosen so as to lead to near-optimality of a desired cost function with analytical guarantees of local stability for a nonlinear reference system [23]. With this RL, and an

AC-based design that employs a high-order tuner (HT) in the inner-loop for parameter estimation, we show that for a class of parametric uncertainties, the closed-loop system, with the integrated AC-RL controller, can be guaranteed to be stable for all initial conditions of the true system. We also quantify the asymptotic behavior of its state, input, and overall cost. The advantages of the AC-RL controller are illustrated using numerical experiments of a quadrotor required to land on a moving platform using a nonlinear high-fidelity model, illustrating the reduction in the sim-to-real gap. Various extensions of the AC-RL approach are also proposed including a relaxation of the parametric uncertainties in the underlying nonlinearities, inclusion of multiple equilibrium points, non-affine systems, and incorporation of magnitude limits in the control input. In all cases except the last extension, the AC-RL ensures global boundedness; in the last case, all initial conditions of the state and parameter estimate that lie inside a bounded domain are shown to lead to boundedness for an arbitrary nonlinear dynamic system.

In the second class of dynamic systems, denoted as Problem 2, we introduce parametric uncertainties similar to the first class and employ contraction theory to establish convergence which utilizes tools in Riemannian geometry to analyze the convergence of differential dynamics by constructing differential Lyapunov functions [29], [30]. In contrast to prior work [31], we introduce multiplicative parametric uncertainties in the control input, employ a HT for parameter estimation, and an RL-based outer-loop as before, with analytical guarantees of local stability to synthesize the reference system.

In both classes, the central component of the AC algorithm uses a HT which was first proposed by [32], and later expanded in [33], [34] in an effort to develop stable low-order adaptive controllers. Independently, HT has also been sought after in the ML community, in an effort to obtain accelerated convergence of an underlying cost function and the associated accelerated learning of the minimizer of this function (see for example, [35]–[39]). The basic idea is to include momentum-based updates so as to get a faster convergence of the performance error. The HT algorithms in [32] have seen widespread applications in machine learning [40]. They were shown in [41]–[43] to lead to stability with time-varying regressors, and in [42] to have accelerated convergence with constant regressors in discrete-time. The first contribution of this paper is the demonstration of these properties of global boundedness and asymptotic behavior of the overall AC-RL controller in problems 1 and 2.

The second contribution of this paper is the demonstration of parameter learning in uncertain dynamic systems. We focus our attention primarily on the first class of dynamic systems with a scalar input. We show that under conditions of persistent excitation [1], the AC-RL controller with the HT guarantees that the parameter estimates converge to their true values, thereby allowing parameter learning. Since the HT inserts additional filtering actions, the proof of parameter convergence is a nontrivial result. Earlier results related to parameter convergence with HT required additional processing and time-scale transformations on the underlying regressors [33], which are avoided in this paper. The main benefits

of parameter learning are robustness [10] and avoiding the bursting phenomenon [1], [44].

Efforts to combine AC and RL approaches have been highlighted in several recent works, which include [45]–[48] in discrete-time and [49]–[52] in continuous-time. Reference [50] proposes the use of adaptive control for nonlinear systems in a data-driven manner, but requires offline trajectories from a target system and does not address accelerated learning or magnitude saturation. References [51], [52] study the linear-quadratic-regulator problem and its adaptive control variants from an optimization and machine learning perspective. In [47] a reinforcement learning approach is used to determine an adaptive controller for an unknown system, while in [48] principles from adaptive control and Lyapunov analysis are used to adjust and train a deep neural network. These approaches, including our earlier work in [49], have not addressed a comprehensive treatment of an integrated AC-RL approach with analytical guarantees. While the use of an AC in the inner-loop and RL in the outer-loop is a common feature, all other elements in this paper including high-order tuner, magnitude saturation, multiple equilibrium points, non-affine systems, and contraction-theory based adaptive control design have been considered for the first time. None of these prior papers have addressed the problem of parameter convergence using the AC-RL controller, and this is also a key contribution of this paper.

In summary, the contributions of the paper are the following, which are applicable to two classes of nonlinear dynamic systems with certain types of parametric uncertainties:

- An AC-RL controller approach that ensures global stability with the AC in the inner loop employing a high-order tuner and RL component in the outer loop, with asymptotic quantification of its state, input, and overall cost with respect to a reference system. The RL-based control policies are designed so that local stability is guaranteed and are near-optimal for the reference system.
- A guarantee of parameter learning in real-time using the AC-RL controller with persistent excitation.

In Section II, we describe the problem statements. Section III includes overviews of a range of underlying tools related to the RL and AC approaches. Sections IV and V describe the AC-RL controller for Problem 1 and Section VI outlines the controller for Problem 2. Section VII includes numerical studies of the AC-RL controllers for Problems 1 and 2. Proofs of Theorems 1 and 4-8 as well as all lemmas can be found in the Appendix, with additional details in [53]. Proofs of Theorems 2 and 3 can be found in [30] and [54] respectively.

II. PROBLEM STATEMENT

We consider two classes of problems in this paper, for which we propose control and learning solutions in real-time. In all problems, states are assumed to be accessible.

A. Problem 1: Real-time control and learning for systems with equilibrium points

Consider a continuous-time, deterministic nonlinear system described by the following dynamics:

$$\dot{X}(t) = F(X(t), U(t)) \tag{1}$$

where $X(t) \in \mathbb{R}^n$, and $U(t) \in \mathbb{R}^m$. We define $x = X - X_0$ and $u = U - U_0$, where (X_0, U_0) is an equilibrium point[1], i.e., $F(X_0, U_0) = 0$. Using a Taylor series expansion on (1) yields

$$\dot{x} = Ax + Bu + f(x, u) \qquad (2)$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, and (A, B) is controllable. It should be noted that (1) can always be written in the form of (2) for any analytic function F. The following assumption is made about f:

Assumption 1: The higher order effects represented by the nonlinearity $f(x, u)$ in (2) (a) lies in the span of B and (b) are solely a function of the state x, i.e. the system in (2) is control affine and $g(x)$ exists such that $Bg(x) = f(x, u)$:

$$\dot{x} = Ax + B[u + g(x)] \qquad (3)$$

The goal in Problem 1 is the determination of the control input $u(t)$ in real-time, when parametric uncertainties are present in the system dynamics in (3), so as to minimize the cost function

$$\min_{u(t) \in \mathcal{U}, \, \forall t \in [t_0, t_0 + T]} \int_{t_0}^{t_0 + T} c(x(t), u(t)) dt \qquad (4)$$

subject to the dynamics in (3), where $\mathcal{U}$ represents the set of all allowable inputs, and c is a bounded function that is piecewise continuous in x and continuous in u.

B. Problem 2: Real-time control for a class of nonlinear systems

Problem 2 corresponds to a class of control-affine nonlinear systems with parametric uncertainties which does not necessarily utilize the presence of an equilibrium point. This class is assumed to be of the form

$$\dot{X}(t) = f(X) + B(X)[U(t)] \qquad (5)$$

where $X(t) \in \mathbb{R}^n$ and $U(t) \in \mathbb{R}^m$. We note that (5) differs from the expansion form of (2) and therefore allows both $f(X)$ and $B(X)$ to be nonlinear. The goal once again is to determine $U(t)$ in real-time which minimizes a cost function of $X(t)$ and $U(t)$, in the same form as (4), when parameter uncertainties are present in $f(X)$ and $B(X)$.

III. UNDERLYING TOOLS

A. An Offline Approach Based on Reinforcement Learning

We define a reference system

$$\dot{x}_r = Ax_r + Bu_r + f_r(x_r, u_r) \qquad (6)$$

where the subscript r is used to denote signals and parameters of the reference system. We briefly describe the offline training procedure in an RL: First, it is assumed that the continuous-time dynamics in (6) are sampled with sufficient accuracy, resulting in the discrete time dynamics

$$x_{r,k+1} = h(x_{r,k}, u_{r,k}) \qquad (7)$$

An appropriate numerical integration scheme ensures that this discrete-time formulation closely approximates the dynamics.

---

[1] In most of what follows, we suppress the argument t for ease of exposition.

The goal of RL is to learn a feedback policy $u_{r,k} = \pi(x_{r,k})$, with repeated training of $\pi(\cdot)$, so as to achieve the control objective of (4) [55].

In order to facilitate such a training, a simulation environment is assumed to be available so that each timestep, an observation $x_{r,k}$ is received, a control $u_{r,k}$ is chosen, and the resulting cost $c_k = c(x_{r,k}, u_{r,k})$ is computed. Repeating this process, a set of input-state-cost tuples $\mathcal{D} = [(x_{r,1}, u_{r,1}, c_1), \ldots, (x_{r,N}, u_{r,N}, c_N)]$ is formed. This data is used to train and update the policy π [55]. Often the policy is parameterized using neural networks so that $u_k = \pi_\theta(x_k)$, where θ denotes the weights of the neural network. A learning algorithm, often based on a gradient-descent approach, is used to adjust θ so that the expected accumulated cost is minimized [56]. In stochastic settings, the RL goal can be summarized as

$$\min_{\theta} \quad J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{k=0}^{T} c_k \right] \qquad (8)$$

1) Analytical guarantees: While majority of the RL literature is focused on extensive simulation studies, a few papers have developed specific RL methods with analytical properties, which can be used for control of nonlinear systems (see for example, [14], [19]–[23]). Reference [14] shows convergence of Q-learning algorithms can be achieved using action-replay process, Ref. [19] develops convergence analysis based on stochastic approximation techniques that apply to both Q-learning and TD(λ) algorithms, and Ref. [20] discusses local convergence results for actor-critic algorithms. Convergence analyses in [14], [19], [20] are established under a tabular setting where state and action spaces are both finite, which may be restrictive for control applications since most physical systems have continuous state space. Results in papers such as [21]–[23] avoid this restriction. Ref. [21] proposes an RL algorithm provide stability guarantee for linear quadratic regulator (LQR) problems; results in [22] provide a proof of convergence of global optimality, and hence stable policy for linear MDPs and cost functions that are a linear combination of a set of known feature functions. In [23], RL is explored for general nonlinear systems, with the requirement that the initial policy is stable and a modification of the cost function that enforces Lyapunov stability based constraints. Overall, the approach in [23] can be viewed as a policy design with a trade-off of optimality for closed-loop stability. We propose the use of such RL-based policies in the RL-part of our AC-RL controller.

B. Classical Adaptive Control Methods

We start with an error model description of the form

$$\dot{e}(t) = A_m e(t) + B\Lambda \left[\widetilde{\Theta}(t) \Phi(t) \right] \qquad (9)$$

where $e(t) \in \mathbb{R}^n$ is a performance error that is required to be brought to zero, such as tracking error, identification error, and state estimation error. $\widetilde{\Theta}(t) \in \mathbb{R}^{m \times l}$ is a matrix of parameter errors that quantify the learning error. If the true parameter in the system dynamics is $\Theta^* \in \mathbb{R}^{m \times l}$, and it is estimated as $\widehat{\Theta}(t)$ at time t, then $\widetilde{\Theta} = \widehat{\Theta} - \Theta^*$. A_m is a Hurwitz matrix, A_m and

B are known with (A_m, B) controllable, and $\Lambda \in \mathbb{R}^{m \times m}$ is an unknown parameter matrix that is positive definite. Finally $\Phi(t) \in \mathbb{R}^l$ is a vector of regressors that correspond to all real-time information measured or computed from the system dynamics and controllers at time t. Such an error model is ubiquitous in adaptive control of a large class of nonlinear dynamic systems with parametric uncertainties [1] including (3). The following theorem summarizes the standard global stability result in AC:

Theorem 1: Let $\Gamma \in \mathbb{R}^{m \times m}$ and Q_a be symmetric positive-definite matrices with P corresponding to the solution of the Lyapunov equation

$$A_m^T P + P A_m = -Q_a. \tag{10}$$

An adaptive law that adjusts the parameter estimate in real time as

$$\dot{\widehat{\Theta}} = -\Gamma B^T P e \Phi^T \tag{11}$$

guarantees that $e(t)$ and $\widetilde{\Theta}(t)$ are bounded for any initial conditions $e(0)$ and $\widetilde{\Theta}(0)$. If in addition $\Phi(t)$ is bounded for all t, then $\lim_{t \to \infty} e(t) = 0$.

C. High-order Tuners

The core of the adaptive components proposed in this paper is based on high-order tuners (HT) [32]–[34]. The idea behind HT is summarized below.

Parameter identification in a linear regression problem of the form $y^*(t) = \theta^{*T} \phi(t)$ where $\theta^* \in \mathbb{R}^n$ represents an unknown parameter, and $\phi(t) \in \mathbb{R}^n$ is an underlying regressor that can be measured at each t can be formulated as a minimization problem $L_t(\theta) = e^2(t)$ where $e = y - y^*$ and $y(t) = \theta^T \phi(t)$. A first-order tuner that can be used to solve the minimization problem is of the form $\dot{\theta} = -\gamma \nabla_\theta L_t(\theta)$. In [41] [42], a HT of the form

$$\ddot{\theta} + \beta \dot{\theta} = -\frac{\gamma \beta}{\mathcal{N}_t} \nabla_\theta L_t(\theta) \tag{12}$$

was proposed and shown to correspond to the Lagrangian

$$\mathcal{L}(\theta, \dot{\theta}, t) = e^{\beta(t - t_0)} \left(\frac{1}{2} \|\dot{\theta}\|^2 - \frac{\gamma \beta}{\mathcal{N}_t} L_t(\theta) \right). \tag{13}$$

The benefits of the HT in (12) are that (a) it can be guaranteed to be stable even with time-varying regressors for a dynamic error model with a scalar control input [41], and (b) a particular discretization was shown in [42] to lead to an accelerated algorithm which reaches an ϵ sub-optimal point in $\widetilde{\mathcal{O}}(1/\sqrt{\epsilon})$ iterations for a linear regression-type convex loss function with constant regressors, as compared to the $\widetilde{\mathcal{O}}(1/\epsilon)$ guaranteed rate for the standard gradient descent algorithm. We employ elements of HT in [41] in the AC part of our AC-RL controller.

D. Contraction Theory

Contraction theory provides tools to analyze the stability of nonlinear systems beyond attraction to an equilibrium point. Based on results in Riemannian geometry, control contraction metrics [30] have been developed to design a feedback controller that stabilizes a class of control-affine systems to a reference trajectory. A brief review of the salient features of this approach follows.

For an arbitrary pair of points $X_1, X_2 \in \mathbb{R}^n$, let $\zeta(X_1, X_2)$ denote the set of all smooth curves connecting X_1 and X_2, in which each path $c \in \zeta(X_1, X_2)$ is parameterized by $s \in [0, 1]$, i.e. $c(s) : [0, 1] \to \mathbb{R}^n$, $c(0) = X_1, c(1) = X_2$. Let $\mathcal{S}_n^+$ be the set of $n \times n$ positive-definite matrices. A Riemannian metric $M(X) : \mathbb{R}^n \to \mathcal{S}_n^+$ is a function that defines a smoothly varying inner product $<\cdot, \cdot>$ on the tangent space of $\mathbb{R}^n$.

A Riemannian metric $M(X)$ is uniformly bounded if there exists $\alpha_2 \geq \alpha_1 > 0$ such that $\alpha_1 I \leq M(X) \leq \alpha_2 I$. For a smooth path $c(s, t)$ between the pair $(X_1(t), X_2(t))$, which is a time-varying smooth curve parameterized by s at any given $t \geq 0$, its Riemannian length and energy are defined as follows,

$$L(c) := \int_0^1 \|c_s\|_c ds, \quad E(c) := \int_0^1 \|c_s\|_c^2 ds$$

where $c_s := \frac{\partial c(s, t)}{\partial s}$ and $\|c_s\|_c = \sqrt{c_s^T M(c(s, t)) c_s}$. A smooth curve is regular if $\frac{\partial c(s, t)}{\partial s} \neq 0$ for all $s \in [0, 1]$. By the Hopf-Rinow theorem, a smooth regular minimum-length curve (geodesic) $\gamma(s, t)$ exists and connects any pair $(X_1(t), X_2(t))$ [30]. Let the minimum Riemannian length of all paths connecting $X_1(t)$ and $X_2(t)$ be a Riemannian distance function, i.e. $d(X_1(t), X_2(t)) := \inf_{c \in \Gamma(X_1(t), X_2(t))} L(c)$. The Riemannian energy of $\gamma(s, t)$ satisfies $d(X_1(t), X_2(t))^2 = E(\gamma) = L(\gamma)^2 \leq E(c)$ for all $c \in \Gamma(X_1(t), X_2(t))$, where the second equality follows from the fact that $\|\gamma_s\|$ is constant [57].

We now consider the dynamic system in (5), let b_i be the ith row of $B(X)$ and U_i be the ith element of control input vector $U(t)$. We can write the differential dynamics of (5) as

$$\dot{\delta}_X = A(X, U) \delta_x + B(X) \delta_U, \tag{14}$$

where $A(X, U) = \frac{\partial f}{\partial X} + \sum_{i=1}^m \frac{\partial b_i}{\partial X} U_i$, and $\delta_x = \gamma_s$ is a differential element of the geodesic connecting a pair of points on two trajectories $X_1(t)$ and $X_2(t)$ in the state space $\mathbb{R}^n$. With these preliminaries established, we can now introduce the control contraction metric.

Theorem 2 ([30]): If there exists a uniformly bounded metric $M(X)$ and $\lambda > 0$ such that

$$\delta_X^T M B = 0 \implies \delta_X^T (A^T M + M A + \dot{M} + 2\lambda M) \delta_X \leq 0$$

holds for all $\delta_X \neq 0$, then system (5) is universally exponentially stabilizable with rate λ, and $M(X)$ is the control contraction metric.

Remark 1: The above theorem states that under the metric $M(X)$, every tangent vector δ_X orthogonal to the span of $B(X)$ is naturally contracting with rate λ [30]. Note that the differential dynamics in (14) is linear time-varying, which reduces the complexity in finding a suitable δ_U that stablizes (14) when $\delta_X^T M B \neq 0$. By computing path integrals of such a δ_U, control contraction metrics $M(X)$ allows construction of stablizing controllers $U(t)$.

E. Persistent Excitation and Parameter Learning

Definition 1 (Persistent Excitation): [1] A bounded function $\Phi : [t_0, \infty) \to \mathbb{R}^N$ is persistently exciting (PE) if there

exists $T > 0$ and $\alpha > 0$ such that

$$\int_t^{t+T} \Phi(\tau)\Phi^T(\tau)d\tau \geq \alpha I, \quad \forall t \geq t_0. \tag{15}$$

The definition of PE in (15) is equivalent to the following:

$$\frac{1}{T} \int_t^{t+T} \left| \Phi(\tau)^T w \right| d\tau \geq \epsilon_0, \forall \text{ unit vectors } w \in \mathbb{R}^N, \forall t \geq t_0 \tag{16}$$

for some $\epsilon_0 > 0$. In what follows, we will utilize an alternate definition, which is equivalent to (15) and (16) if $\|\dot{\Phi}(t)\|$ is bounded for all t [58]:

Definition 2: Φ is PE if there exists an $\epsilon > 0$ a t_2 and a sub-interval $[t_2, t_2 + \delta_0] \subset [t, t+T]$ with

$$\frac{1}{T} \left| \int_{t_2}^{t_2+\delta_0} \Phi(\tau)^T w d\tau \right| \geq \epsilon_0, \forall \text{ unit vectors } w \in \mathbb{R}^N, \forall t \geq t_0. \tag{17}$$

Adaptive control systems enable parameter learning by imposing properties of persistent excitation defined in Section III-E. We briefly summarize the classical result related to this topic, established in [54], [58]. The starting point is the same error model as in (9), which contains two errors, $e(t) \in \mathbb{R}^n$ is a performance error that can be measured, and $\widetilde{\theta}^T(t) = \widetilde{\Theta}(t) \in \mathbb{R}^{1 \times l}$ is a parameter learning error. As before we assume that A_m is known and Hurwitz, and B is known. The following theorem summarizes this result:

Theorem 3 ([54]): The solutions of the error dynamics in (9) together with the adaptive law in (11) lead to $\lim_{t \to \infty} \widetilde{\theta}(t) = 0$ if the regressor Φ satisfies the PE condition in Definition 2.

IV. INTEGRATED AC-RL SOLUTIONS FOR REAL-TIME CONTROL AND LEARNING: PROBLEM 1

We address Problem 1 in this section, and propose an integrated AC-RL controller that combines adaptive control (AC) and reinforcement learning (RL) approaches.

A. A Restatement of Problem 1

We shall denote (3) when there are no parametric uncertainties as the reference system, and express it as

$$\dot{x}_r = Ax_r + B[u_r + g(x_r)] \tag{18}$$

We assume that u_r is designed using RL:

$$u_r = \pi(x_r) \tag{19}$$

It should be noted that $\pi(x_r)$ in (19) does not necessarily cancel $g(x_r)$, but accommodates it so that the closed-loop system behaves in a satisfactory manner for the requisite control task. This is quantified in Assumption 2, a desirable property of RL controllers.

Assumption 2: RL is used to train a feedback policy $\pi : x \to u$ such that for a given positive constant R_2, a positive constant R_1 exists such that $||x_r(0)|| \leq R_1$ implies $||x_r(t)|| \leq R_2$ $\forall t$.

Remark 2: Assumption 2 encapsulates the notion that, given a sufficiently long training period, a performant RL algorithm, and appropriate choices of hyperparameters, the learned policy π will be "good" enough to maintain boundedness of the reference system states. As mentioned in Section III-A, for tabular settings where the action and state are finite [14], [19], and for linear dynamics and cost functions that are linear combinations of known feature functions [22], the learned policy is optimal and therefore stable, which validates Assumption 2. For nonlinear dynamics, [23] shows that a stable controller can be learned by optimizing a soft loss function that augments cost functions in (4) with Lyapunov stability based constraints. Relaxing the soft loss constraints and expanding scope of RL for general nonlinear systems is one of the objectives in the area of *safe reinforcement-learning*, a topic of intense research activity. We note that any of aforementioned frameworks including [23] can be used to compute a stable RL policy u_r in (19). In order to avoid the associated optimality gap induced by these soft loss constraints, methods such as PPO, shown to guarantee local optimality in [59], engineer the underlying cost function through the addition of high magnitude penalties only on out-of-set excursions.

1) Parametric uncertainties: We now introduce two parametric uncertainties, one in the form of control effectiveness, i.e., u gets perturbed as Λu, and the second as a perturbation in $g(x)$. Loss of control effectiveness is ubiquitous in practical problems, due to unforeseen anomalies that may occur in real-time, such as accidents or aging in system components, especially in actuators. Parametric uncertainties in the nonlinearity $g(x)$ may be due to modeling errors. The following assumptions are made about these two uncertainties.

Assumption 3: The nonlinearity $g(x)$ in (3) is parameterized linearly, i.e., $g(x) = \Theta_{n,r}\Phi_n(x)$ where $\Theta_{n,r} \in \mathbb{R}^{m \times l}$, and $\Phi_n(x) \in \mathbb{R}^l$.

Assumption 4: Λ is symmetric and positive definite, with $\|\Lambda\| \leq 1$.

Remark 3: Assumption 3 implies that $g(x)$ in (3) can be approximated accurately using l basis functions. The use of basis functions to approximate a nonlinearity is commonplace in adaptive control: see e.g. [60]–[62]. For simplicity, we assume that the approximation is exact in Assumption 3, which will be relaxed in Section V-A. Assumption 4 introduces a structure in the loss of control effectiveness.

With Assumption 3 in place, a second parametric uncertainty is introduced where the nonlinear component $g(x)$ is perturbed as $\Theta_{n,r}\Phi_n(x)$ to $\Theta_n'\Phi_n(x)$. With these parametric uncertainties, the plant equation in (3) then becomes

$$\dot{x} = Ax + B\Lambda \left[u + \Lambda^{-1}\Theta_n'\Phi_n(x) \right] \tag{20}$$

where $B \in \mathbb{R}^{n \times m}, \Lambda \in \mathbb{R}^{m \times m}$, and $\Theta_n' \in \mathbb{R}^{m \times l}$. Altogether we note that Problem 1 is restated as the control of (20) where A, B, and $\Phi_n(x)$ are known, but Λ and Θ_n' are unknown parameters.

B. The AC-RL controller

The goal of the AC-RL controller is to design a control input u in (20) such that the true system state x converges to the reference system state x_r in the presence of the aforementioned perturbations in Λ and Θ_n'. Subtracting (18) from (20), we obtain the error dynamics

$$\dot{e} = A_H e + B\Lambda[u - \Theta\Phi] \tag{21}$$

where

$$e := x - x_r, \Theta := \Lambda^{-1}\left[I, \; -\Theta_n'\right], \Phi := \begin{bmatrix} \Phi_r(u_r, x_r, x) \\ \Phi_n(x) \end{bmatrix} \tag{22}$$

$$\Phi_r(u_r, x_r, x) = u_r + g(x_r) + \Theta_{l,r}e$$

and $\Theta_{l,r}$ is such that $A_H := A + B\Theta_{l,r}$ is Hurwitz. In (21), $\Theta \in \mathbb{R}^{m \times (m+l)}$ corresponds to an unknown parameter matrix and $\Phi(t) \in \mathbb{R}^{m+l}$ is the regressor vector used for adaptation. The error equation (21) is central to the development of the AC-RL algorithms derived in the following subsections. We now propose the AC-RL controller.

$$u = \widehat{\Theta}(t)\Phi(t) \tag{23}$$

$$\dot{\Xi} = -\gamma B^T P e \Phi^T \tag{24}$$

$$\dot{\widehat{\Theta}} = -\beta(\widehat{\Theta} - \Xi)\mathcal{N}_t, \tag{25}$$

where

$$\mathcal{N}_t = 1 + \mu\Phi^T\Phi \tag{26}$$

$$\mu \geq \frac{2\gamma}{\beta}\|PB\|_F^2\|\Omega\|_2^2 \tag{27}$$

Ω is such that $\Omega^T\Omega = \Lambda$, $\|\cdot\|_F$ in (27) denotes the Frobenius matrix norm, and $P = P^T \in \mathbb{R}^{n \times n}$ is a positive definite matrix that solves the equation $A_H^T P + P A_H = -Q_H$, where Q_H is a positive-definite matrix and $Q_H \geq 2I$. The following theorem is the first main result of this paper:

Theorem 4: Under Assumptions 1-4, the closed-loop adaptive system specified by the plant in (20), the reference system in (18), and the adaptive controller in (23)-(25) leads to bounded solutions for any initial conditions $x(t_0)$, with $\lim_{t\to\infty} e(t) = 0$.

Using the result in Theorem 4, we quantify the asymptotic property of AC-RL controller in Corollary 1 where $o(\cdot)$ is defined as in [63]. For this purpose, we define the ideal control input as $u^* := \Theta\Phi$. From (21), it is easy to see that if $u = u^*$, then $x(t) \to x_r(t)$ as $t \to \infty$.

Corollary 1: If $\dot{\Phi}$ is bounded, the AC-RL controller ensures that (i) $\lim_{t\to\infty} |u(t) - u^*(t)| = 0$ and

$$\text{(ii)} \int_{t_0}^{t_0+T} \left|c(x(t), u(t)) - c(x_r(t), u^*(t))\right| dt = o(T) \tag{28}$$

We make several important observations about the AC-RL controller:

1) The parameter update in (24) - (25) is a second-order tuner, and an extension of our earlier results in [41] to the multivariable case. Two different regressors are employed in the adaptive control input, Φ_r in (23) and Φ_n, which are utilized to address the two different sources of parametric uncertainties Λ and Θ_n. The first regressor component, Φ_r, comes predominantly from the RL-component, where $g(x_r)$ can be computed as $(B^T B)^{-1} B^T (\dot{x}_r - Ax_r) - u_r$ using measurements from the reference system and numerical approximations. Alternately, $g(x_r)$ can be computed using comparisons

of simulations of the reference system for infinitesimal and large inputs. It should be noted that these computations are offline, and therefore do not add to the computational burden in the AC-RL. The second regressor accommodates the uncertainty Θ_n' in $g(x)$ (and employs assumption 3). The additional feedback from e in (22) is essential in guaranteeing global stability.

2) The RL-policy is required to satisfy Assumption 2, such as the one in [23].

3) In comparison to a purely AC approach, in the AC-RL controller the RL plays the role of a reference model that is nonlinear with a controller that is nonlinear as well and trained so as to elicit desirable properties from the reference system. It should also be noted that a purely AC approach is agnostic to u_r, relegates the design of u_r to an outer-loop, and essentially focuses only on the innerloop stability and tracking of $x_r(t)$.

4) If there are no parametric uncertainties, and if the initial conditions of (20) are identical to those of (18), then the choices of $\widehat{\Theta}(0) = [I, \Theta_{n,r}^T]^T$ and $\Xi(0) = 0$ ensures that the AC-RL control $u(t) \to u^*(t) \to u_r(t)$, thereby accomplishing (4). That is, the AC-RL policy coincides with the near-optimal RL-based policy. When there are parametric uncertainties, the AC-RL policy is necessarily modified; $u(t) \to u^*(t)$ but $u^*(t) \not\to u_r(t)$. Theorem 4 guarantees that with this modified policy, the closed-loop system state x is globally bounded, sans any online training, and converges to x_r. Corollary 1 precisely quantifies the corresponding cost of the AC-RL controller.

Remark 4: Alternate methods rather than RL for determining the feedback policy may be used as well provided they satisfy Assumption 2. Methods based on MPC may be used as well, but may require additional structures on the cost function in the form of terminal costs [64]. The same comment is applicable for other classical nonlinear control methods, but these may involve additional assumptions or restrictions. Furthermore, MPC generally requires online computation, which might be challenging especially when system dynamics is nonlinear, while RL can be trained offline and deployed online with reduced computational burden.

C. A comparison of AC-RL and RL-based controllers

We make a few comments regarding Assumptions 1-4 and the global properties of this AC-RL controller below.

1) Global properties of the AC-RL controller vs. local properties of the RL-controller: The focus is on nonlinear dynamic systems with parametric uncertainties that satisfy Assumption 1, 3, and 4 and an RL-component that satisfies 2. Suppose we start with $x(t_0)$ with $\|x(t_0)\| \leq R_1$. Suppose that the parametric uncertainties are such that $x(t)$ departs from $\mathcal{X}$ for some $t > t_0$. In such a case, an RL controller alone will not suffice. In contrast, the AC-RL controller will accommodate these departures, modify the policy from $u_r(t)$ to $u^*(t)$, and ensure that

$$\|x(t)\| \leq (1/\lambda(P_{\min}))V_0 + R_1 \qquad \forall t \geq t_0$$

where $V_0 = V((e(t_0), \Xi(t_0) - \Theta, \Theta(t_0) - \Theta)$, where V is defined as in (59). This accommodation was possible because of the change in the control input from u_r to u in (23). It should be noted that the AC-RL controller relaxes the restriction of the RL-approach that the state always remains in a compact set, allow parametric uncertainties to be introduced at any t_0, provided they are in the form specified in Assumptions 1, 3, and 4, and provided the RL approach satisfies Assumption 2. As this is a nontrivial class of nonlinear problems that can occur in practice, this extension from a local result with RL-approaches to a global result with AC-RL approach proposed is an important contribution, and bridges the sim-real gap for this class of nonlinear problems. The benefit of using RL rather than other control approaches is that we can address tasks that correspond to arbitrary cost functions. The cost of using RL is the requirement that Assumption 2 is satisfied.

2) Assumptions 1 to 4: Assumption 2 was needed for the RL part of the controller, while the other three assumptions were needed for the AC part of the controller. As the adaptive controller is designed to function in real time, certain structures needed to be imposed on the type of uncertainties that the AC can deal with, which led to Assumption 1 (that required uncertainties to lie in the span of the control matrix B), Assumption 3 (that required the nonlinearities to be represented using a linear network with a bounded approximation error), and Assumption 4 (which required certain analytic properties in the linear parametric uncertainty Λ).

3) Optimality of the control policy u*: While we have established the asymptotic nature of u in the form of u^*, we note that $u^*(t) \neq u_{opt}(t)$, where

$$u_{opt}(t) = argmin_{u(t) \in \mathcal{U}, \, \forall t \in [t_0, t_0 + T]} \int_{t_0}^{t_0 + T} c(x(t), u(t)) dt$$

for the plant (20). The reason for this difference can be argued thus: The input u^* is designed to make the closed-loop system stable with uncertainties with the reference system as the target response. But such a u^* is optimal with respect to the cost function in (4) subject to the reference system in (18) (i.e. no uncertainties), but not optimal with respect to the perturbed system in (20) (with uncertainties). While choosing the reference system in (18) as a target response provides tractability in the form of global stability, it may not be the right choice for the compromised true system in (20) from the point of view of optimality. A tractable design of a reference system that assures both stability (using AC methods) and optimality (using RL methods) is a topic for future research.

D. Learning in AC-RL controllers with persistent excitation

Section IV-B focused on the control solution, and showed global boundedness with a cost as in (28). In particular we showed in corollary 1 that $\widetilde{\Theta}(t)\Phi(t) \to 0$ asymptotically, where $\widetilde{\Theta} = \widehat{\Theta} - \Theta$. As the vector Φ can be arbitrary, $\widetilde{\Theta}(t)$ can remain orthogonal to $\Phi(t)$, and need not go to zero asymptotically, i.e., there is no guarantee that parameter learning will take place. In this section, we establish learning with the AC-RL algorithm by imposing additional conditions

of persistent excitation on Φ. We limit our discussion to the case when $u(t)$ is a scalar.

The starting point is the error equation in (21) and the AC-RL controller in (23)-(27). With a scalar input, $m = 1$, we obtain that $\Lambda \in \mathbb{R}^+$ (using Assumption 4), $\Theta \in \mathbb{R}^{1 \times (l+1)}$, $\Phi_r(t) \in \mathbb{R}$, and $\Phi_n(t) \in \mathbb{R}^l$ in (22). Defining $\tilde{\theta} = (\widehat{\Theta} - \Theta)^T$, $\tilde{\vartheta} = (\Xi - \Theta)^T$ and $B_0 = B\Lambda$, we rewrite (21), (24) and (25) as

$$\dot{e} = A_H e + B_0 \tilde{\theta}^T \Phi \tag{29}$$

and

$$\dot{\tilde{\vartheta}} = -\gamma \Phi e^T P B \tag{30}$$

$$\dot{\tilde{\theta}} = -\beta(\tilde{\theta} - \tilde{\vartheta}) \mathcal{N}_t \tag{31}$$

where $\mathcal{N}_t = 1 + \mu \Phi^T \Phi$, $\mu \geq 2\gamma \|PB\|^2 / \beta$. As before, the matrix P solves $A_H^T P + P A_H = -Q$ and $Q \geq 2I$ is a symmetric, positive-definite matrix. The constants γ and β are positive. Let $x_1(t) = [e(t)^T, (\tilde{\theta}(t) - \tilde{\vartheta}(t))^T]^T$ and $z(t) = [x_1(t)^T, \tilde{\vartheta}(t)^T]^T$.

We note that the stability result related to the AC-RL controller stated in Theorem 4 guarantees that $z(t)$ is uniformly bounded, which follows from a Lyapunov function of the form

$$V = \frac{\Lambda}{\gamma} \tilde{\vartheta}^T \tilde{\vartheta} + \frac{\Lambda}{\gamma} \left[(\tilde{\vartheta} - \tilde{\theta})^T (\tilde{\vartheta} - \tilde{\theta}) \right] + e^T P e \tag{32}$$

whose derivative is bounded by

$$\dot{V} \leq -\frac{2\beta\Lambda}{\gamma} \|\tilde{\theta} - \tilde{\vartheta}\|^2 - \Lambda \|e\|^2 \tag{33}$$

Noting that our goal is to show that $\lim_{t \to \infty} \tilde{\theta}(t) = 0$, it suffices to show that $V(t) \to 0$ as $t \to \infty$. As the time-derivative $\dot{V}$ in (33) is only negative semi-definite, additional conditions of persistent excitation are utilized to show parameter learning, and corresponds to the second contribution of this paper.

The following lemmas are useful in proving the second contribution of this paper, which is stated in Theorem 5. We note that Theorem 4 guarantees that $z(t)$ and $\Phi(x, e, u_r)$ are bounded, making the properties of persistent excitation applicable for what follows.

Lemma 1: Let $\epsilon_1 > \epsilon_2 > 0$, then there is an $n = n(\epsilon_1, \epsilon_2)$ such that if $z(t) = \left[x_1(t)^T, \tilde{\vartheta}(t)^T \right]^T$ is a solution with $\|z(t_1)\| \leq \epsilon_1$ and $S = \{t \in [t_1, \infty) | \|x_1(t)\| > \epsilon_2\}$, then $\mu(S) \leq n$ where μ denotes Lebesgue measure.

Lemma 2: Let $\delta > 0$ and $\epsilon_1 > 0$ be given. Then there exist positive numbers ϵ and T such that if $z(t)$ is a solution with $\|z(t_1)\| \leq \epsilon_1$ and if $\left\|\tilde{\vartheta}(t)\right\| \geq \delta$ for $t \in [t_1, t_1 + T]$, then there exists a $t_2 \in [t_1, t_1 + T]$ such that $\|x_1(t_2)\| \geq \epsilon$.

Lemma 3: Let ϵ_1 and δ be given positive numbers. Then there is a $T = T(\epsilon_1, \delta)$ such that if $z(t)$ is a solution and $\left\|z(t_1)\right\| \leq \epsilon_1$, then there exist some $t_2 \in [t_1, t_1 + T]$ such that $\left\|\tilde{\vartheta}(t_2)\right\| \leq \delta$.

Theorem 5: If $\Phi(t)$ satisfies the persistent excitation property in (17), then the origin $(e = 0, \tilde{\vartheta} = 0, \tilde{\theta} = 0)$ in (29)-(31) is uniformly asymptotically stable.

The proof of Theorem 5 stems from the three lemmas listed above, which represent the three main steps. Lemmas 1 and 2 establish that $x_1(t)$ cannot remain small over the entire period

of persistent excitation. Lemma 3 then leverages this fact to show that this leads to the parameter error $\tilde{\vartheta}(t)$ to decrease. Together this allows the conclusion of u.a.s. of the origin (29)-(31) and therefore that $\lim_{t\to\infty}\tilde{\theta}(t)=0$. As is apparent from the details of the proof provided in the Appendix, first principles-based arguments had to be employed in order to derive this result. No standard observability properties or time-scale transformations as in [33] have been employed; these are inadequate as the error model structure in (29) includes system dynamics and no filtering techniques are used to convert the error model to a static linear regression model. Extensions to multivariable inputs are a topic for future work. The results of this section also indicate that the RL-based policy u_r should be such that various components of $\Phi(t)$ satisfy the PE condition.

V. EXTENSIONS OF THE AC-RL CONTROLLER FOR PROBLEM 1

A. Relaxation of Assumption 3

Assumption 5: The higher order term in (3) is parameterized linearly, i.e., $g(x)=\Theta_{n,r}\Phi_n(x)+d(t)$ where $\Theta_{n,r}\in\mathbb{R}^{m\times l}$, $\Phi_n(x)\in\mathbb{R}^l$, and $d(t)$ denotes an approximation error with $\|d(t)\|\le d_{max}$ for all t.

Remark 5: While Assumption 5 is a relaxation of Assumption 3, the more general case of interest is when d to be dependent on the state x, i.e. not known to be bounded *a priori*. This extension is relegated to future work.

We note that with Assumption 3 replaced by Assumption 5, Eq. (20) is modified as

$$\dot{x}=Ax+B\Lambda\left[u+\Lambda^{-1}\Theta_n^{'}\Phi_n(x)+\Lambda^{-1}d\right] \quad (34)$$

with the resulting error equation modified from (21) as

$$\dot{e}=A_He+B\Lambda[u-\Theta\Phi+\overline{d}] \quad (35)$$

where $\overline{d}=\Lambda^{-1}d$ with $|\overline{d}|\le\overline{d}_{max}$.

The AC-RL controller in this case remains as in (23) and (25), but the tuner in (24) is modified as

$$\dot{\Xi}=-\gamma B^TPe\Phi^T-f_1(\Xi,\widehat{\Theta}) \quad (36)$$

$$f_1=\left(2\Xi-\widehat{\Theta}\right) \quad (37)$$

with all other quantities defined as in Section IV. The analytical guarantee for this AC-RL controller is summarized in the following theorem:

Theorem 6: Under Assumptions 1, 2, 4, and 5, the closed-loop adaptive system specified by the plant in (20), the reference system in (18), and the adaptive controller in (23)-(27), and (24) replaced by (36)-(37), leads to globally bounded solutions, with $e(t)=O(D)$, where $D=\max(\overline{d}_{max},\|\Theta\|)$.

B. MSAC-RL Controller

As control inputs are often subject to magnitude saturation, we propose another extension, Magnitude Saturated Adaptive Controller (MSAC)-RL, which builds on the ideas introduced in [16]. The saturated control input into the true plant is calculated as:

$$u_i(t)=u_{i,\max}\text{sat}\left(\frac{u_{i,c}(t)}{u_{i,\max}}\right) \quad (38)$$

where $u_c(t)$ denotes the output of the controller and $u_{i,\max}$ is the allowable magnitude limit on u_i, chosen as $\|\mathcal{U}\|$ where $\mathcal{U}$ is the set of all allowable inputs. This induces a saturation-triggered disturbance Δu vector defined by

$$\Delta u(t)=u_c(t)-u(t) \quad (39)$$

It is easy to see that $\Delta u(t)=0$ when the desired control $u_c(t)$ does not saturate. The output of the controller, u_c, is given by:

$$u_c=\widehat{\Theta}(t)\Phi(t) \quad (40)$$

The presence of the disturbance Δu causes the error equation to vary from (21) to

$$\dot{e}=A_He+B\Lambda[u_c-\Delta u-\Theta\Phi] \quad (41)$$

We introduce a new performance error e_a in order to accommodate the disturbance Δu as follows:

$$\dot{e}_a=A_He_a+BK_a(t)\Delta u \quad (42)$$

where K_a is an estimate for Λ which is adjusted according to (44). The error e_a leads to a new augmented error $e_u=e-e_a$:

$$\dot{e}_u=A_He_u+B\Lambda\widetilde{\Theta}(t)\Phi(t)-B(\Lambda+K_a)\Delta u \quad (43)$$

This suggests a different set of adaptive laws,

$$\dot{\Xi}=-\gamma B^TPe_u\overline{\Phi}^T \quad (44)$$

$$\dot{\overline{\Theta}}=-\beta(\overline{\Theta}-\Xi)\mathcal{N}_t, \quad (45)$$

where $\overline{\Theta}=[\widehat{\Theta},-K_a]$ and $\overline{\Phi}=[\Phi^T,\Delta u^T]^T$, γ and β are positive constants, and $\mathcal{N}_t$ defined as in (26) with Φ replaced by $\overline{\Phi}$ and μ as in (27). The following theorem provides the analytical guarantees for this MSAC-RL controller.

Theorem 7: Under Assumptions 1-4, the closed-loop adaptive system specified by the plant in (20), the reference system in (18), the magnitude constraint in (38) and the MSAC-RL controller given by (40), (44) and (45) leads to

(i) globally bounded solutions, with $\lim_{t\to\infty}\|e(t)\|=0$ if the target system in (20) is open-loop stable.
(ii) bounded solutions for all initial conditions $x(0)$, $\Xi(0)$, $\widehat{\Theta}(0)$ and $K_a(0)$ in a bounded domain, with $\|e(t)\|=O[\int_0^t\|\Delta u(\tau)\|d\tau]$ if the target system in (18) is not open-loop stable.

Remark 6: As in Section V-A, if we relax Assumption 3 in the form of Assumption 5, then the asymptotic result in Theorem 7(i) is replaced by convergence to a compact set as in Theorem 6. It should be noted that in both Section V-A and V-B, the proposed controller still remains global.

C. Multiple Equilibrium Points

Suppose the system in (1) has p equilibrum points $(X_1,U_1),\ldots,(X_p,U_p)$, so that $F(X_i,U_i)=0$ for $i=1,\ldots,p$. Define $x_i=x-X_i$, $u_i=u-U_i$. Denoting the state and action spaces as $\mathcal{X}$ and $\mathcal{U}$, respectively, we partition the composite domain $\mathbb{D}=\mathcal{X}\times\mathcal{U}$ into p disjoint subsets $S_1,\ldots,S_p$, such that:

$$\mathbb{D}=S_1\cup S_2\cup\ldots S_p$$
$$S_i\cap S_j=\emptyset, i\ne j$$

One can then express (1) in the presence of parametric uncertainties as

$$\dot{x} = \begin{cases} A_1 x_1 + B_1[\Lambda u_1 + g(x_1)], & x_1, u_1 \in S_1 \\ \quad\quad\quad\quad \vdots \\ A_p x_p + B_p[\Lambda u_p + g(x_p)], & x_p, u_p \in S_p. \end{cases} \quad (46)$$

If we now consider the effect of parametric uncertainties in u_i and $g(x_i)$ as in Section III-A, it is easy to have a switching set of controllers as in (23)-(25) that are invoked when the trajectories enter the set S_i. A corresponding stability result to Theorem 4 can be derived, provided the command signal is such that the dwell time in each set S_i exceeds a certain threshold, which is not discussed here. We summarize the overall AC-RL controller in Algorithm 1, which specifies the discrete time implementation of the overall AC-RL Controller. Δt denotes the integration timestep and the AdaptiveControl function corresponds to the adaptive control input; in the single equilibrium case, this function corresponds to equations (23)-(27). F and F_r represent the velocity vector fields of the target and reference dynamical systems, respectively. That is, F and F_r are the right hand sides of equations (2) and (6), respectively.

Algorithm 1 Multiple Equilibrum AC-RL

Require: F, F_r, π, x_r, x
  while running **do**
    $u_r = \pi(x_r)$
    $x_i \leftarrow x - X_i$
    $u_i \leftarrow u - U_i$
    $e \leftarrow x_i - (x_r - X_i)$
    $\Phi \leftarrow [x_i, u_i]$
    $u \leftarrow$ AdaptiveControl$(\Phi, e, \hat{\Theta}_i, P_i, B_i) + U_i$
    $x_r \leftarrow x_r + F_r(x_r, u_r)\Delta t$
    $x \leftarrow x + F(x, u)\Delta t$
    $i \leftarrow j : x, u \in S_j$
  end while

D. Nonaffine Systems

We once again start with (1), expand the dynamics around (X_0, U_0), which together with Assumption 1 leads to a class of nonlinear systems

$$\dot{x} = Ax + B[u + h(x, u)] \quad (47)$$

Noting that u is required to minimize (4), we assume that u can be expressed as an analytic function of the state x, which in turn leads to the assumption that

$$h(x, u) = g(x) \quad (48)$$

The goal is to determine u in real time when parametric uncertainties are present in (47)-(48) so that (4) is accomplished for initial condition x_0. As before, the reference system dynamics with no parametric uncertainties is chosen as in (18). Noting that (48) makes the plant in (47) identical to (3), the same AC-RL controller in Section IV-B leads to global boundedness under Assumptions 1-4, with valid extensions as in Sections V-A and V-B. We skip a formal statement of the result due to space limitations.

VI. INTEGRATED AC-RL SOLUTIONS: PROBLEM 2

We now consider Problem 2, which pertains to the class of dynamic system (5), and introduce parametric uncertainties $\theta \in \mathbb{R}^{n_\theta}$ and $\nu \in \mathbb{R}^{n_\nu}$ in $f(X)$, and $\Lambda \in \mathcal{S}_m^+$ in U which may be due to loss of effectiveness, leading to the system

$$\dot{X}(t) = f(X) - \rho(X)\nu + B(X)\left[\Lambda U(t) - \phi(X)\theta\right], \quad (49)$$

where $\rho(X) \in \mathbb{R}^{n \times n_\nu}$ and $\phi(X) \in \mathbb{R}^{m \times n_\theta}$ are known nonlinear dynamics of the parametric uncertainties θ and ν, respectively. It is clear that $\phi(X)\theta$ denotes a matched uncertainty similar to that introduced by Assumptions 3, while $\rho(X)\nu$ denotes an unmatched uncertainty. We make the following assumption regarding the latter, which is similar to an extended matching condition in [5]:

Assumption 6: The unmatched uncertainty satisfies $\rho(X)\nu \in span\{ad_f B\}$, where $ad_f B = [f, B] = \frac{\partial f}{\partial X} B - \frac{\partial B}{\partial X} f$ denotes the Lie bracket of f, B. $f(X)$ and $B(X)$ are nonlinear functions such that B and $ad_f B$ are linearly independent.

For the system in (49) with Assumption 6, we now propose an AC-RL controller. We first consider a control-affine reference system, similar to (5), of the form,

$$\dot{X}_r(t) = f(X_r) + B(X_r)U_r(t). \quad (50)$$

Given an initial condition X_0, $U_r(t)$ minimizes the cost function given in (4) where $x(t)$ and $u(t)$ are replaced by $X(t)$ and $U(t)$, respectively. The reference input $U_r(t)$ can be constructed using the offline RL-based approach in Section III-A. In addition to Assumption 2 and 4, we introduce the following assumption:

Assumption 7: A uniformly bounded parameter-dependent contraction metric $M(X, \hat{\nu})$ can be computed for each parameter estimate $\hat{\nu} \in \mathbb{R}^{p_\nu}$.

Under Assumption 7, let $\gamma(t, s)$ be the geodesic that connects $X_r(t)$ to $X(t)$, i.e., $\gamma(t, 0) = X_r(t)$ and $\gamma(t, 1) = X(t)$. The Riemannian energy of the geodesic is defined as $E = \int_0^1 \gamma_s^T M(X, \hat{\nu})\gamma_s ds$. We now describe the AC-RL controller in equations (51)-(53) and the stability result:

$$U = \Psi\left(U_c + U_\nu + \phi(X)\hat{\theta}\right), \quad (51a)$$

$$U_\nu = \sum_{i=1}^{p_\nu} \Omega_i \dot{\hat{\nu}}_i \int_0^1 \left[\frac{\partial \rho_i^T(\gamma(s))}{\partial X_1} \; \cdots \; \frac{\partial \rho_i^T(\gamma(s))}{\partial X_{n-1}} \; 0 \right] \gamma_s(s) ds, \quad (51b)$$

$$\dot{\hat{\nu}} = \beta_\nu(\nu_a - \hat{\nu})\mathcal{N}_\nu, \quad (51c)$$

$$\dot{\nu}_a = -\gamma_\nu \rho(X)^T M \gamma_s(1) E, \quad (51d)$$

$$\dot{\hat{\theta}} = \beta_\theta(\theta_a - \hat{\theta})\mathcal{N}_\theta, \quad (51e)$$

$$\dot{\theta}_a = -\gamma_\theta \phi(X)^T B(X)^T M \gamma_s(1) E, \quad (51f)$$

$$\dot{\Psi} = \beta_\Psi(\Psi_a - \Psi)\mathcal{N}_\Psi, \quad (51g)$$

$$\dot{\Psi}_a = -\gamma_\Psi B(X)^T M \gamma_s(1) E U^T, \quad (51h)$$

where

$$\mathcal{N}_\nu = 1 + \omega_\nu^T \omega_\nu, \quad \omega_\nu = \rho(X)^T M \gamma_s(1), \quad (52a)$$

$$\mathcal{N}_\theta = 1 + \omega_\theta^T \omega_\theta, \quad \omega_\theta = \phi(X)^T B(X)^T M \gamma_s(1), \quad (52b)$$

$$\mathcal{N}_\Psi = 1 + \omega_\Psi^T \omega_\Psi, \quad \omega_\Psi = U, \quad (52c)$$

ρ_i is the ith column of $\rho(X,t)$, $\Omega_i \in \mathbb{R}^m$ extracts the jth column vector of $B(x)$ such that $\rho_i(x)\nu_i \in span\{ad_f b_j\}$, and U_c is a nominal controller that leverages the RL-based input U_r and satisfies

$$\frac{\partial E}{\partial t} + 2\gamma_s(1)^T M(X,\hat{\nu})(f(X) - \rho(X)^T \hat{\nu} + B(X)U_c)$$
$$- 2\gamma_s(0)^T M(X_r, \hat{\nu}) \left[f(X_r) + B(X_r)U_r \right] \leq -2\lambda E. \quad (53)$$

where $\lambda > 0$. The adaptive gain parameters β_ν, γ_ν, β_θ, γ_θ, β_Ψ, γ_Ψ are chosen as positive numbers that satisfy

$$\frac{\beta_\nu}{\gamma_\nu} \geq \frac{3}{\lambda}, \quad \frac{\beta_\theta}{\gamma_\theta} \geq \frac{3}{\lambda}, \quad \frac{\beta_\Psi}{\gamma_\Psi} \geq \frac{3\|\gamma_s(1)\|\|M\|\|B\|}{\lambda} \quad (54)$$

Theorem 8: Under Assumption 2, 4, 6, and 7, the closed-loop system specified by the plant in (49), and the adaptive controller in (51)-(54) has bounded solutions and ensures that $\lim_{t\to\infty} \|X(t) - X_r(t)\| = 0$.

Remark 7: Theorem 8 proposes a solution to Problem 2 and leverages contraction theory rather than Lyapunov theory. The new contributions here, in comparison with [31], are that the controller is based on AC-RL, it utilizes a high-order tuner, and incorporates larger uncertainties such as loss of effectiveness.

VII. NUMERICAL STUDIES

A. Problem 1

We consider a control task of a quadrotor moving in 3-D to land on a moving platform. The adaptive control design was based on a linearized quadrotor model

$$\ddot{x} = g\theta \quad \ddot{\theta} = \frac{L}{I_y}\tau_\theta$$
$$\ddot{y} = -g\phi \quad \ddot{\phi} = \frac{L}{I_x}\tau_\phi \quad (55)$$
$$\ddot{z} = \frac{f_z - mg}{m} \quad \ddot{\psi} = \frac{1}{I_z}\tau_\psi$$

which defines the matrices A_r, B. $\Theta_{l,r}$ was chosen via $\Theta_{l,r} = B^\dagger(A_r - A_H)$ such that all eigenvalues of A_H were at -6, and we also chose $Q_H = I$. The generalized forces were generated via the control input as:

$$\begin{bmatrix} f_z \\ \tau_\phi \\ \tau_\theta \\ \tau_\psi \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ L & 0 & -L & 0 \\ 0 & L & 0 & -L \\ \nu & -\nu & \nu & -\nu \end{bmatrix} \Lambda u \quad (56)$$

In (55), $g = 9.81, m = 1.2, L = 0.3, I_x = I_y = 0.22, I_z = 0.44, \nu = 1.0$ (in SI units). Λ was chosen to be diagonal with entries decreased from the nominal case of unity to represent loss of effectiveness of each actuator. A linearized quadrotor model was used in the AC-RL control design, and therefore $g(\cdot) = 0$ in (18). A nonlinear quadrotor model as in [65], however, was used for the training of the RL as well as for the evaluation of the AC-RL controllers.

The cost function in (4) was chosen as

$$c(\vec{x},t) = \begin{cases} -1 & \texttt{box} \\ 1 & \neg\texttt{box} \wedge (\Delta z \leq 0 \vee t \geq T_{max}) \\ 0 & \texttt{else} \end{cases} \quad (57)$$

where the boolean variable $\texttt{box}$ is assumed to be True if ALL of the following simultaneously hold: $|z| \leq 5\text{cm}$, $|\Delta xy| \leq 25\text{cm}$, $|\phi| \leq 10°$, $|\theta| \leq 10°$, $|v_{xy}| \leq 50\text{cm/s}$, $|\dot{z}| \leq 10\text{cm/s}$, and $T_{\max} = 10\text{sec}$, where $|\Delta xy|^2 = ((x - X_{\text{plat}})^2 + (y - Y_{\text{plat}})^2)$, $v_{xy} = \|[\dot{x} \ \dot{y}]^T\|$, and $(X_{\text{plat}}, Y_{\text{plat}})$ denote the platform position in the (x,y)-plane [53]. $\vec{x}$ is the whole quadrotor state, and the integral in (4) is computed with a discretization of $\Delta t = 0.1$. $z = 0$ was assumed to correspond to the plane of motion of the moving platform. Overall, (57) implies that the cost increases if the quadrotor leaves a compact set $\mathcal{D}(\vec{x})$, if its altitude falls below the platform altitude, or if the termination time exceeds 10sec.

We use the popular PPO method [66] to determine the RL component even though PPO has been shown only to have properties of local optimality [59], as it provides a powerful methodology for training. Any departures of the state from the compact set are accommodated through suitable reward engineering, i.e. through the addition of high magnitude penalties in the cost function only on out-of-set excursions. The following hyperparameters are used in the training of the RL algorithm, PPO (see [53], [66] for details), in the outerloop: $\gamma = .99, \Delta t = .1, T = 1000, N = 64$. The fourth-order Runge-Kutta method is used to integrate the dynamics, with a step-size of $h = .001$. The learning rate used for SGD algorithms is .0001, and both the value and policy networks were chosen to have two hidden layers, each with 128 neurons and ReLU activations. The inner-loop controller uses the MSAC algorithm described in V-B, with $\gamma = 50, \beta = 25, \mu = 1$.

We test two types of parametric uncertainties: 1) the mass, length and inertia properties of the quadrotor varied between $\pm25\%$ of their nominal (reference) values (Table I) and 2) an abrupt loss of effectiveness (LOE) in the fourth propeller (Table II) and therefore in Λ. Such a LOE may occur if the propeller blades are broken midflight, as demonstrated in [65]. Both types of parametric uncertainties correspond to the target-system structure as in (20), with the symmetric part of Λ being positive definite. We observed that in all cases, the AC-RL ensured boundedness of all signals in the closed-loop system and that the tracking error converged to zero, which validates Theorem 4.

As our goal extends beyond Theorem 4 to minimizing the cost in (57), we evaluate AC-RL using additional performance metrics, success rate (SR) and success time (ST), where SR denotes the percentage of all trials where the quadrotor entered the compact set $\mathcal{D}(\vec{x})$ within 10 seconds and ST is the mean time required to complete this task over all successful tests. The results are reported in Tables I and II, with Table I also including a comparison of the average cost in (57) averaged over 1000 trials. It should be noted that Theorem 4 provides no guarantees for SR to be at 100% or for ST to be less than a specified amount, as both correspond to the transient performance of a nonlinear controller, which is very difficult to quantify. Nevertheless, the results reported in Tables I and II show that AC-RL outperforms its competitors.

We compare four control methods: 1) RL, which just uses the trained π to dictate control; 2) AC-RL, which utilizes both π and the adaptive laws in V-B; 3) a meta-learning approach

(ME-RL) which employs a deep-learning based policy update in addition to a network that learns all hyperparameters [67]; and 4) a domain-randomized approach (DR-RL) [68] which uses domain randomization of the plant parameters during policy training. The following observations follow:

Algorithm	Results		
	SR	**ST**	**Cost** (57)
RL	48%	7.5s	1.2
AC-RL	82%	3.5s	-0.3
DR-RL	75%	7.1s	0.11
ME-RL	88%	3.4s	-0.41

TABLE I: $\pm 25\%$ parametric uncertainty results

A. AC-RL performs better than pure RL or DR-RL (Table I).
B. ME-RL outperforms AC-RL on all three metrics. This comes with two qualifiers: 1) the meta-learner in ME-RL is trained using the same distributional shift on which it was tested (the $\pm 25\%$ parameter perturbations), and 2) the meta-learner is a DNN and thus has no guarantees of convergence, unlike AC-RL.
C. The relevance of point 1) becomes apparent from the ME-RL results in Table II, where it can be see that ME-RL greatly underperforms AC-RL. This is because the specific type of uncertainty (asymmetric LOE) studied in Table II was not incorporated into the meta-learner's training regimen. The LOE column represents the degree of propeller thrust lost (with 0% being no loss). For a 75% LOE there is no data on the RL or ME-RL success time because there were no successful tests. See [53] for results with noise.
D. Table I also shows that for a 25% parametric uncertainty, the average cost was -0.3 and 1.2 for AC-RL and RL, respectively. In comparison, the average cost for the case when there were no uncertainties, was observed to be -0.51 for a pure RL approach, which is the best achievable cost, i.e. AC-RL has an optimality gap of 40%.
E. We observed that pure AC led to success only with a smaller set of initial conditions for the state, primarily due to the stringent metric in (57). This led to an optimality gap of 40% under AC-RL, in contrast to 115% under AC alone (see [53] for further details).

AC-RL SR	ME-RL SR	RL SR	LOE
97%	98%	97%	0%
88%	78%	74%	10%
73%	41%	34%	25%
54%	8%	14%	50%
27%	0%	0%	75%
AC-RL ST	**ME-RL ST**	**RL ST**	**LOE**
2.6s	2.8s	2.6s	0%
3.1s	3.9s	3.1s	10%
2.9s	5.1s	4.7s	25%
3.3s	7.5s	5.4s	50%
3.3s	——	——	75%

TABLE II: Results from the simulated quadrotor experiments.
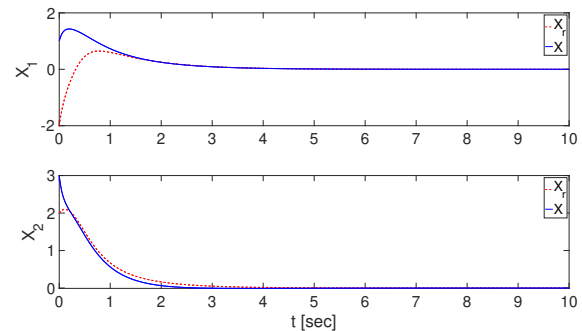

Fig. 1: State trajectories of reference and true systems.

B. Problem 2

To demonstrate the effectiveness of our proposed AC-RL controller in (51)-(54), we provide numerical simulation results for a system described in (49) with

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, f(X) = \begin{bmatrix} -X_1 - X_1^3 + X_2^2 \\ 0 \end{bmatrix}, B(X) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$\phi(X) = X^T, \rho(X) = \begin{bmatrix} 0 & 0 \end{bmatrix}^T, \Lambda = I, \theta = \begin{bmatrix} 1 & 1 \end{bmatrix}^T.$$

Note that unmatched uncertainty $\rho(X)\nu$ in (49) is removed for simplicity in the numerical example. As a result, the contraction metric $M(X)$ does not depend on parameter estimate $\hat{\nu}$. The initial estimates $(\hat{\theta}, \theta_a)$ and (Ψ, Ψ_a) for unknown parameters θ and Λ, respectively, are 20% off of true values. As shown in Figure 1, our proposed controller provides bounded closed-loop solution $X(t)$ that converges to reference state trajectory $X_r(t)$ generated using RL-based controller U_r asymptotically, which validates theoretical results stated in Theorem 8.

VIII. SUMMARY AND CONCLUSIONS

This paper proposed solutions for real time control and learning in dynamic systems with parametric uncertainties using a combination of adaptive control and reinforcement learning. Two classes of nonlinear dynamic systems are considered, both of which are control-affine. The first class of dynamic systems utilizes equilibrium points and expansion forms around these points and employs a Lyapunov approach. The second class of nonlinear systems uses contraction theory as the underlying framework. For both classes of systems, the AC-RL controller is shown to lead to online policies that guarantee stability, and leverage accelerated convergence properties using a high-order tuner. In both cases, when there are no uncertainties, the RL component is assumed to be capable of generating policies that are guaranteed to be stable and near-optimal. Additionally, for the first class of systems, the AC-RL controller is shown to lead to parameter learning with persistent excitation. Together, this paper takes a first step towards real-time control using machine learning with provable guarantees, by drawing upon key insights, tools, and approaches developed in these two disparate and powerful methodologies of adaptive control and reinforcement learning. Several more steps need to be carried out in this direction and address the gap between stability and optimality of general nonlinear systems with arbitrary cost functions.

APPENDIX

A. Proof of Theorem 1:

A Lyapunov function candidate is chosen as

$$V(e, \widetilde{\Theta}) = e^T P e + \text{Tr}\left(\widetilde{\Theta}^T \left(\Lambda^T S\right) \widetilde{\Theta}\right) \tag{58}$$

where $S = \Gamma^{-1}$ in (11) with the symmetric part of ΛS positive-definite. Using arguments as in [69], we can show that $\dot{V} = -e^T Q_a e$ by using (9) and (11). Thus, $e(t)$ and the parameter estimates $\widehat{\Theta}$ are uniformly bounded. If in addition $\Phi(t)$ is bounded, we note that $\dot{e}(t)$ is bounded, and that $e \in \mathcal{L}_2$. It follows from Barbalat's lemma that $\lim_{t \to \infty} \|e(t)\| = 0$.

B. Proof of Theorem 4:

The closed-loop system dynamics with the AC-RL controller is represented by the error equation (21), the adaptive control input (23), and the adaptive laws in (24)-(27). A Lyapunov function candidate

$$V = \frac{1}{\gamma}\text{Tr}\left[(\Xi - \Theta)^T \Lambda^T (\Xi - \Theta)\right] \\ + \frac{1}{\gamma}\text{Tr}\left[(\widehat{\Theta} - \Xi)^T \Lambda^T (\widehat{\Theta} - \Xi)\right] + e^T P e \tag{59}$$

yields a time-derivative

$$\begin{aligned}\dot{V} = & -\text{Tr}\left[(\Xi - \Theta)^T (\Lambda + \Lambda^T) B^T P e \Phi^T\right] \\ & -\frac{\beta}{\gamma}\text{Tr}\left[(\widehat{\Theta} - \Xi)^T (\Lambda + \Lambda^T)(\widehat{\Theta} - \Xi)\right] \mathcal{N}_t \\ & +\text{Tr}\left[(\widehat{\Theta} - \Xi)^T (\Lambda + \Lambda^T) B^T P e \Phi^T\right] \\ & +e^T\left(A_H^T P + P A_H\right) e + 2 e^T P B \Lambda \widetilde{\Theta} \Phi\end{aligned} \tag{60}$$

Through algebraic manipulations, it can be shown that

$$\begin{aligned}\dot{V} \leq & -\frac{2\beta}{\gamma}\text{Tr}\left[(\widehat{\Theta} - \Xi)^T \Omega^T \Omega(\widehat{\Theta} - \Xi)\right] - 2\|e\|^2 \\ & -4\|PB\|_2^2\,\text{Tr}\left[(\widehat{\Theta} - \Xi)^T \Omega^T \Omega(\widehat{\Theta} - \Xi)\right]\|\Phi\|^2 \\ & +4\|e\|\|PB\|_2\|\Lambda(\widehat{\Theta} - \Xi)\|_F\|\Phi\| \\ \leq & -\|e\|^2 - \frac{2\beta}{\gamma}\left\|\Omega(\widehat{\Theta} - \Xi)\right\|_F^2 \\ & -\left[\|e\| - 2\|PB\|_2\|\Omega\|_2\|\Omega(\widehat{\Theta} - \Xi)\|_F\|\Phi\|\right]^2 \leq 0 \tag{61}\end{aligned}$$

where we have expressed the positive definite matrix Λ using Ω with $2\Omega^T \Omega = \Lambda + \Lambda^T$, $\|\cdot\|_F$ is the Frobenius matrix norm and $\|\cdot\|_2$ is the matrix 2-norm. In the above derivation we have used (26), (27), Assumption 4, and the inequality $\|AB\|_F \leq \|A\|_2\|B\|_F$. It can be concluded that $e, \widehat{\Theta}, \Xi \in \mathcal{L}_\infty$. From Assumption 2 and (22), we conclude that Φ is bounded. Using Barbalat's lemma and arguments as in Appendix A, we conclude that $\lim_{t \to \infty} e(t) = 0$.

C. Proof of Corollary 1

From Theorem 4, it follows that $\lim_{t \to \infty} e(t) \to 0$. Using (21), it follows that

$$\lim_{t \to \infty} \int_{t_0}^t \exp(A_H(t - \tau)) B \Lambda[u(\tau) - \Theta\Phi(\tau)] d\tau = 0. \tag{62}$$

It also follows that $\Phi \in \mathcal{L}_\infty$. Additionally, from the proof of Theorem 4 we know that $\widehat{\Theta}, \Xi \in \mathcal{L}_\infty$. Using (23) and (25) and the supposition that $\dot{\Phi} \in \mathcal{L}_\infty$, we have $u, \dot{u} \in \mathcal{L}_\infty$. From here, with (62) it can be shown using first principles of real analysis that $\lim_{t \to \infty}[u(t) - \Theta\Phi(t)] = 0$, implying conclusion (i). Due to the continuity arguments on $c(x, u)$ as noted in Section II-A, we have that $\lim_{t \to \infty}[c(x, u) - c(x_r, u^*)] = 0$.

D. Proof of Lemma 1

Note: In the proofs of Lemma 1 - 3 and Theorem 5, we assume that $\|B_0\Phi(t)^T\| \leq c_1$. Such a c_1 exists as Φ is bounded following Theorem 4.

Consider the following candidate Lyapunov function

$$V = \frac{\Lambda}{\gamma}\|\tilde{\vartheta}\|^2 + \frac{\Lambda}{\gamma}\|\tilde{\theta} - \tilde{\vartheta}\|^2 + e^T P e$$

Similar to what has been shown in the proof of Theorem 4, the time derivative of V may be bounded by

$$\dot{V} \leq -\frac{2\beta\Lambda}{\gamma}\|\tilde{\theta} - \tilde{\vartheta}\|^2 - \Lambda\|e\|^2 \leq -c_2\|x_1\|^2$$

where $c_2 = \Lambda \min\left\{1, \frac{2\beta}{\gamma}\right\}$. Noting that P is a positive definite matrix, it follows that for any vector $v \in \mathbb{R}^n$, $\alpha, \rho > 0$ exist such that

$$\alpha v^T v \leq v^T P v \leq \rho v^T v \tag{63}$$

We prove Lemma 1 by using contradiction. Assume $\mu(S) > n$ and denote $\overline{S} = \{t \in [t_1, \infty)\}$. Integrating $\dot{V}$, we have

$$\begin{aligned}\int_{t_1}^\infty \dot{V}(\tau)d\tau = & \int_S \dot{V}(\tau)d\tau + \int_{\overline{S}-S} \dot{V}(\tau)d\tau \\ \leq & \int_S -c_2\|x_1(\tau)\|^2 d\tau + \int_{\overline{S}-S} \dot{V}(\tau)d\tau \leq -c_2 n \epsilon_2^2\end{aligned}$$

Choose $n(\epsilon_1, \epsilon_2) = c_3 \epsilon_1^2/(c_2 \epsilon_2^2)$. This leads to a contradiction since $\|V(t_1)\| \leq c_3 \epsilon_1^2$, where $c_3 = \max\left\{\frac{\Lambda}{\gamma}, \rho\right\}$.

E. Proof of Lemma 2

Let $z(t)$ be a solution with initial condition $\|z(t_1)\| \leq \epsilon_1$. Suppose that $\left\|\tilde{\vartheta}(t)\right\| \geq \delta$ for all $t \in [t_1, t_1 + T]$, where $T = T_0 + \delta_0$.

From the error model in (29), for any $t \geq t_1$,

$$e(t + \delta_0) = e(t) + \int_t^{t+\delta_0} A_H e(\tau) + B_0 \Phi(\tau)^T \tilde{\theta}(\tau) d\tau \tag{64}$$

from which we have

$$\begin{aligned}& \|e(t + \delta_0)\| \\ & \geq \left\|\int_t^{t+\delta_0} B_0 \Phi(\tau)^T \tilde{\theta}(\tau) d\tau\right\| - \left\|e(t) + \int_t^{t+\delta_0} A_H e(\tau) d\tau\right\| \\ & \tag{65}\end{aligned}$$

Given $\epsilon' = \epsilon_0 \delta/(2c_2\delta_0)$ and $T = T_0 + \delta_0$, from the adaptive law in (31), there is an $\epsilon_2 > 0$ such that if $z(\tau)$ is a solution to (30)-(31) with $\|x_1(\tau)\| \leq \epsilon_2$ for all $\tau \in [t_1, t_1 + T]$, then $\|\tilde{\theta}(\tau) - \tilde{\theta}(t_1)\| \leq \epsilon'$. Define $\epsilon = \min\left\{\frac{\delta\epsilon_0}{8}, \frac{\delta\epsilon_0}{8c_1\delta_0}, \epsilon_2\right\}$. Now we show the lemma holds for this choice of T and ϵ.

If $\|x_1(t_2)\| \geq \epsilon$ for some $t_2 \in [t_1, t_1 + T]$, then we are done. Assume $\|x_1(t)\| \leq \epsilon$ for all $t \in [t_1, t_1 + T]$, then

$$\left\| e(t) + \int_t^{t+\delta_0} A_H e(\tau) d\tau \right\| \leq \epsilon + c_1 \epsilon \delta_0 \leq \frac{\epsilon_0 \delta}{4}$$

for all $t \in [t_1, t_1 + T]$. By hypothesis, there exist a $t' \in [t_1, t_1 + T_0]$ such that

$$\left\| \int_{t'}^{t'+\delta_0} B_0 \Phi(\tau)^T w d\tau \right\| \geq \epsilon_0$$

where $w = \tilde{\theta}(t_1) / \left\| \tilde{\theta}(t_1) \right\|$ is a unit vector.

We have

$$\left\| \int_{t'}^{t'+\delta_0} B_0 \Phi(\tau)^T \left[w \left\| \tilde{\theta}(t_1) \right\| - \tilde{\theta}(\tau) \right] d\tau \right\|$$
$$\leq c_2 \int_{t'}^{t'+\delta_0} \left\| \tilde{\theta}(t_1) - \tilde{\theta}(\tau) \right\| d\tau \leq c_2 \delta_0 \epsilon' = \frac{\epsilon_0 \delta}{2}$$

Since $\|e(\tau)\| \leq \|x_1(\tau)\| \leq \epsilon \leq \epsilon_2$ for $\tau \in [t_1, t_1 + T]$,

$$\left\| \tilde{\theta}(t_1) \right\| \left\| \int_{t'}^{t'+\delta_0} B_0 \Phi(\tau)^T w d\tau \right\|$$
$$- \left\| \int_{t'}^{t'+\delta_0} B_0 \Phi(\tau)^T \tilde{\theta}(\tau) d\tau \right\| \leq \frac{\epsilon_0 \delta}{2}$$

which implies

$$\left\| \int_{t'}^{t'+\delta_0} B_0 \Phi(\tau)^T \tilde{\theta}(\tau) d\tau \right\| \geq \epsilon_0 \delta - \frac{\epsilon_0 \delta}{2} = \frac{\epsilon_0 \delta}{2}$$

Thus

$$\|x_1(t' + \delta_0)\| \geq \|e(t' + \delta_0)\| \geq \frac{\epsilon_0 \delta}{2} - \frac{\epsilon_0 \delta}{4} = \frac{\epsilon_0 \delta}{4} > \epsilon$$

which is a contradiction.

F. Proof of Lemma 3

By Lemma 2, the assumption that $\|z(t_1)\| \leq \epsilon_1$ and $\left\| \tilde{\vartheta}(t_2) \right\| \geq \delta$ implies that there exist an ϵ such that $\|x_1(t)\|$ is periodically both less than $\epsilon/2$ and greater than ϵ. This leads to a contradiction with Lemma 1 if we choose $\epsilon_1 = \|z(t_1)\|$ and $\epsilon_2 = \epsilon/2$. We thus conclude that $\left\| \tilde{\vartheta}(t_2) \right\| \leq \delta$.

G. Proof of Theorem 5

Consider the candidate

$$V = \frac{\Lambda}{\gamma} \|\tilde{\vartheta}\|^2 + \frac{\Lambda}{\gamma} \|\tilde{\theta} - \tilde{\vartheta}\|^2 + e^T P e \tag{66}$$

With $\mu \geq 2\gamma \|PB\|^2 / \beta$ and $Q_H \geq 2I$ which solves $A_H^T P + P A_H = -Q_H$, the time derivative of (66) may be bounded by

$$\dot{V} \leq -\frac{2\beta\Lambda}{\gamma} \|\tilde{\theta} - \tilde{\vartheta}\|^2 - \Lambda \|e\|^2 \leq 0 \tag{67}$$

Since P is positive-definite, (63) holds for some $\alpha, \rho > 0$. Now we show that given $\epsilon_1 > \epsilon_2 > 0$, there is a η with $0 < \eta < 1$ and $\Delta T_1 > 0$ such that if $z(t)$ is a solution with

$$\epsilon_2 \leq V(t) \leq \epsilon_1, \quad \text{for } t \in [t_1, t_1 + \Delta T_1],$$

then there is a $t_2 \in [t_1, t_1 + \Delta T_1]$ such that $V(t_2) \leq \eta V(t_1)$. Choose $0 < \nu < 1$, $\nu \leq \sigma < 1$ and $\Delta T_2 > 0$ so that $\rho\sqrt{1-\sigma} - \Delta T_2 \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) > 0$, $\sqrt{\gamma(1-\nu)} - \Delta T_2 \left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}} \right) > 0$, $0 < \Delta T_2 \left[\rho\sqrt{1-\sigma} - \Delta T_2 \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) \right]^2 < 1$ and $0 < \frac{2\beta\Delta T_2}{\gamma} \left[\sqrt{\gamma(1-\nu)} - \Delta T_2 \left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}} \right) \right]^2 < 1$. From Lemma 3, we can obtain a T when $\epsilon = \epsilon_1$ and $\delta = \sqrt{\epsilon_2\nu}$. Define $\eta = 1 - \min \left\{ \Delta T_2 \left[\rho\sqrt{1-\sigma} - \Delta T_2 \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) \right]^2, \frac{2\beta\Delta T_2}{\gamma} \left[\sqrt{\gamma(1-\nu)} - \Delta T_2 \left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}} \right) \right]^2 \right\}$ and $\Delta T_1 = T + \Delta T_2$. Next we show that for this η and ΔT_1 the results hold.

Let $t_2' \in [t_1, t_1 + T]$ be such that $\left\| \tilde{\vartheta}(t_2') \right\| \leq \delta\sqrt{\gamma}$. If $V(t_2') \leq \epsilon_2$, we are done. If $V(t_2') \geq \epsilon_2$, then

$$V(t_2') = \frac{\Lambda}{\gamma} \left\| \tilde{\vartheta}(t_2') \right\|^2 + \frac{\Lambda}{\gamma} \left\| \tilde{\theta}(t_2') - \tilde{\vartheta}(t_2') \right\|^2 + e(t_2')^T P e(t_2') \tag{68}$$

implies

$$(1 - \nu)V(t_2') \leq V(t_2') - \delta^2 \tag{69}$$
$$\leq \frac{\Lambda}{\gamma} \left\| \tilde{\theta}(t_2') - \tilde{\vartheta}(t_2') \right\|^2 + e(t_2')^T P e(t_2')$$
$$\leq \frac{\Lambda}{\gamma} \left\| \tilde{\theta}(t_2') - \tilde{\vartheta}(t_2') \right\|^2 + \rho e(t_2')^2$$

Case 1: $\frac{\Lambda}{\gamma} \|\tilde{\theta}(t_2') - \tilde{\vartheta}(t_2')\|^2 < (1-\nu)V(t_2')$. From (69),

$$e(t_2')^2 \geq \frac{1}{\rho}(1-\sigma)V(t_2'), \tag{70}$$

where $0 < \nu \leq \sigma < 1$. From (29), for any $t \geq t_2'$,

$$\|e(t_2')\| - \|e(t)\| \leq \int_{t_2'}^t \left\| A_H e(\tau) + B_0 \Phi(\tau)^T \tilde{\theta}(\tau) \right\| d\tau$$
$$\leq \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) (t - t_2') \sqrt{V(t_2')}$$

where the last inequality is due to the assumption that $\|A_H\| \leq c_1$ and $\|B_0 \Phi(\tau)^T\| \leq c_2$ for all τ. If $t_2 = t_2' + \Delta T_2$, we obtain

$$\|e(t)\| \geq \|e(t_2')\| - \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) (t_2 - t_2')\|z(t_2')\|$$
$$\geq \rho\sqrt{1-\sigma}\sqrt{V(t_2')} - \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) \Delta T_2 \sqrt{V(t_2')}$$
$$= \left[\rho\sqrt{1-\sigma} - \Delta T_2 \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) \right] \sqrt{V(t_2')}$$

Integrating $\dot{V}$, we obtain

$$V(t_2') - V(t_2) = \int_{t_2'}^{t_2} -\dot{V}(\tau) d\tau$$
$$\geq \int_{t_2'}^{t_2} \left(\|e(\tau)\|^2 + \frac{2\beta}{\gamma} \left\| \tilde{\theta}(\tau) - \tilde{\vartheta}(\tau) \right\|^2 \right) d\tau$$
$$\geq \Delta T_2 \left[\rho\sqrt{1-\sigma} - \Delta T_2 \left(\frac{c_1}{\sqrt{\alpha}} + 2c_2\sqrt{\gamma} \right) \right]^2 V(t_2')$$

Therefore, $V(t_2) \leq \eta V(t'_2)$ and uniform asymptotic stability holds.

Case 2: $\frac{1}{\gamma}\left\|\tilde{\theta}(t'_2) - \tilde{\vartheta}(t'_2)\right\|^2 \geq (1-\nu)V(t'_2)$. For any $t \geq t'_2$, following the process in case 1, we can show that

$$\left\|\tilde{\theta}(t'_2) - \tilde{\vartheta}(t'_2)\right\| - \left\|\tilde{\theta}(t) - \tilde{\vartheta}(t)\right\|$$
$$\leq (t - t'_2)\left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}}\right)\sqrt{V(t'_2)}$$

If we let $t_2 = t'_2 + \Delta T_2$, then

$$\left\|\tilde{\theta}(t) - \tilde{\vartheta}(t)\right\|$$
$$\geq \left\|\tilde{\theta}(t'_2) - \tilde{\vartheta}(t'_2)\right\| - (t_2 - t'_2)\left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}}\right)\sqrt{V(t'_2)}$$
$$\geq \left[\sqrt{\gamma(1-\nu)} - \Delta T_2\left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}}\right)\right]\sqrt{V(t'_2)}$$

Integrating $\dot{V}$, we have

$$V(t'_2) - V(t_2)$$
$$\geq \frac{2\beta\Delta T_2}{\gamma}\left[\sqrt{\gamma(1-\nu)} - \Delta T_2\left(\beta\sqrt{\gamma} + \frac{c_2\gamma\rho}{\sqrt{\alpha}}\right)\right]^2 V(t'_2)$$

Therefore, $V(t_2) \leq \eta V(t'_2)$ which proves the theorem.

H. Proof of Theorem 6:

When $d(t) \neq 0$, with f_1 as in (37), we obtain additional terms in $\dot{V}$, in the right-hand-side of (60) of the form (for ease of exposition, γ is set to unity)

$$2e^T P B \Lambda d - 2Tr\left(\left(2\widetilde{\Xi} - \tilde{\theta}\right)^T (\Lambda + \Lambda^T)(2\Xi - \Theta)\right)$$

where $\widetilde{\Xi} = \Xi - \Theta$ and $\widetilde{\Theta} = \hat{\Theta} - \Theta$. Together with these terms, $\dot{V}$ can be shown to be negative outside an ellipsoid $Z^T \overline{A} Z = D_0$ where $Z = \left[\|e\|, \left\|\widetilde{\Xi} - \widetilde{\Theta}\right\|, \left\|\widetilde{\Xi}\right\|\right]^T$, $\overline{A}$ is a positive definite matrix in $\mathbb{R}^{3\times3}$ and D_0 is a constant that depends on $\bar{d}_0$ and $\|\Theta\|$ (the constant unknown parameter) We refer the reader to [53] for all technical details of the derivation.

I. Proof of Theorem 7

As in Theorem 4, we consider a candidate

$$V = \frac{1}{\gamma}\text{Tr}\left[(\Xi - \Theta)^T \Lambda^T (\Xi - \Theta)\right]$$
$$+ \frac{1}{\gamma}\text{Tr}\left[(\overline{\Theta} - \Xi)^T \Lambda^T (\overline{\Theta} - \Xi)\right] + e_u^T P e_u \qquad (71)$$

Using (43) and (44), we obtain that $\dot{V} \leq -e_u^T e_u$ since $Q_H \geq 2I$. This in turn allows us to conclude that $e_u, \Xi, \overline{\Theta} \in \mathcal{L}_\infty$. For case (i), since all parameters are bounded, together with magnitude saturation, we have that the input u is bounded. For this case, it implies that the state x is bounded. Therefore $e_u \in \mathcal{L}_\infty$. Proceeding in a similar fashion to the proof of Theorem 1, Barbalat's lemma is applied to find that $\lim_{t\to\infty}\|e_u(t)\| = 0$. It is easy to show then that $\|e(t)\| = O[\int_0^t \|\Delta u(\tau)\| d\tau]$.

For case (ii), as the structure of the error model in (43) is identical to that considered in Theorem 1 in [70], the same arguments can be used to establish boundedness of the state.

J. Proof of Theorem 8:

The first variation of the Riemannien energy [57] is

$$\dot{E} = \frac{\partial E}{\partial t} + \frac{\partial E}{\partial \hat{\nu}}\dot{\hat{\nu}} + 2\langle\gamma_s(s), \dot{\gamma}(s)\rangle|_{s=0}^{s=1} - 2\int_0^1 \langle\frac{D}{ds}\gamma_s, \dot{\gamma}\rangle ds$$

where $\frac{D(\cdot)}{ds}$ is the covariant derivative. Since $\gamma(s)$ is the geodesic, $\frac{D\gamma_s}{ds} = 0$.

Under Assumption 6, Lemma 2 in [31] implies

$$\frac{\partial E}{\partial \hat{\nu}}\dot{\hat{\nu}} + 2\gamma_s(1)^T M(X, \hat{\nu}, t) B(X, t) U_\nu = 0$$

Therefore, using (51) and (53), we have

$$\dot{E} = \frac{\partial E}{\partial t} + \frac{\partial E}{\partial \hat{\nu}}\dot{\hat{\nu}} + 2\gamma_s(1)^T M(X, \hat{\nu}, t)\dot{X}$$
$$- 2\gamma_s(0)^T M(X_r, \hat{\nu}, t)\dot{X}_r$$
$$= \frac{\partial E}{\partial t} + \left[\frac{\partial E}{\partial \hat{\nu}}\dot{\hat{\nu}} + 2\gamma_s(1)^T M(X, \hat{\nu}, t) B(X, t) U_\nu\right]$$
$$+ 2\gamma_s(1)^T M(X, \hat{\nu}, t) B(X, t)(\Lambda\Psi - I)U_\nu$$
$$+ 2\gamma_s(1)^T M(X, \hat{\nu}, t)\left[f(X, t) - \rho(X, t)\nu\right.$$
$$+ B(X, t)(\Lambda\Psi - I)(U_c + \phi(X, t)\hat{\theta})\right]$$
$$+ 2\gamma_s(1)^T M(X, \hat{\nu}, t) B(X, t)(U_c + \phi(X, t)\tilde{\theta})$$
$$- 2\gamma_s(0)^T M(X_r, \hat{\nu}, t)\dot{X}_r$$
$$\leq -2\lambda E + 2\gamma_s(1)^T M(X, \hat{\nu}, t)\left[\rho(X, t)\tilde{\nu}\right.$$
$$+ B(X, t)((\Lambda\Psi - I)U + \phi(X, t)\tilde{\theta})\right]$$

Consider a Lyapunov candidate

$$V = \frac{1}{\gamma_\nu}(\nu_a - \hat{\nu})^T(\nu_a - \hat{\nu}) + \frac{1}{\gamma_\nu}(\nu_a - \nu)^T(\nu_a - \nu)$$
$$+ \frac{1}{\gamma_\theta}(\theta_a - \hat{\theta})^T(\theta_a - \hat{\theta}) + \frac{1}{\gamma_\theta}(\theta_a - \theta)^T(\theta_a - \theta)$$
$$+ \frac{1}{\gamma_\Psi}\text{Tr}\left[(\Lambda\Psi_a - \Lambda\Psi)^T|\Lambda|^{-1}(\Lambda\Psi_a - \Lambda\Psi)\right]$$
$$+ \frac{1}{\gamma_\Psi}\text{Tr}\left[(\Lambda\Psi_a - I)^T|\Lambda|^{-1}(\Lambda\Psi_a - I)\right] + \frac{1}{2}E(\gamma, \hat{\nu}, t)^2$$

Define $e = \gamma_s(1)ME$, and use short-hand notation $\rho = \rho(X, t), \phi = \phi(X, t)$, and $B = B(X, t)$. Since the Riemmanien energy is always non-negative, i.e. $E \geq 0$, we have

$$\frac{1}{2}\dot{V} \leq -\lambda E^2 + e^T \rho(\nu_a - \nu) + e^T B\left[(\Lambda\Psi - I)U\right.$$
$$+ \phi(\theta_a - \theta)\right] - (\nu_a - \hat{\nu})^T \rho^T e - \frac{\beta_\nu}{\gamma_\nu}(\nu_a - \hat{\nu})^T(\nu_a - \hat{\nu})\mathcal{N}_\nu$$
$$- (\nu_a - \nu)^T \rho^T e - (\theta_a - \hat{\theta})^T \phi^T B^T e - \frac{\beta_\theta}{\gamma_\theta}(\theta_a - \hat{\theta})^T$$
$$(\theta_a - \hat{\theta})\mathcal{N}_\theta - (\theta_a - \theta)^T \phi^T B^T e + \text{Tr}\left[-ue^T B(\Lambda\Psi_a - I)\right]$$
$$+ \text{Tr}\left[-Ue^T B\Lambda(\Psi_a - \Psi) - \frac{\beta_\Psi}{\gamma_\Psi}\mathcal{N}_\Psi^T(\Psi_a - \Psi)^T\Lambda(\Psi_a - \Psi)\right]$$

which can be simplified as $\dot{V} \leq 2(V_1 + V_2 + V_3)$, where

$$V_1 = -\frac{\lambda}{3}E^2 + 2e^T\rho(\hat{\nu} - \nu_a) - \frac{\beta_\nu}{\gamma_\nu}(\nu_a - \hat{\nu})^T(\nu_a - \hat{\nu})\mathcal{N}_\nu$$

$$V_2 = -\frac{\lambda}{3}E^2 + 2e^T B\phi(\hat{\theta} - \theta_a) - \frac{\beta_\theta}{\gamma_\theta}(\theta_a - \hat{\theta})^T(\theta_a - \hat{\theta})\mathcal{N}_\theta$$

$$V_3 = -\frac{\lambda}{3}E^2 + 2e^T B\Lambda(\Psi - \Psi_a)U$$
$$\quad - \frac{\beta_\Psi}{\gamma_\Psi}\mathrm{Tr}\left[(\Psi_a - \Psi)^T\Lambda(\Psi_a - \Psi)\mathcal{N}_\Psi\right]$$

It follows from (52) and (54) that

$$V_1 \leq -\frac{\beta_\nu}{\gamma_\nu}\|\nu_a - \hat{\nu}\|^2 - \lambda_1(E - 2\|\omega_\nu\|\|\nu_a - \hat{\nu}\|)^2$$

$$V_2 \leq -\frac{\beta_\theta}{\gamma_\theta}\|\theta_a - \hat{\theta}\|^2 - \lambda_2(E - \|\omega_\theta\|\|\theta_a - \hat{\theta}\|)^2 \leq 0$$

Using similar manipulations as in the proof of Theorem 4,

$$V_3 \leq -\frac{\beta_\Psi}{\gamma_\Psi}\|(\Psi_a - \Psi)\Omega\|_F^2 - \lambda_3(E$$
$$\quad - \|\gamma_s(1)^T MB\|E\|(\Psi_a - \Psi)\Omega\|_F\|U\|)^2 \leq 0$$

Thus we have $\dot{V} \leq 2(V_1 + V_2 + V_3) \leq 0$, which implies that $E, \hat{\nu}, \hat{\theta}, \Psi \in \mathcal{L}_\infty$, and that $E \in \mathcal{L}_2$. Since θ is a constant vector and $\tilde{\theta}$ is bounded, $\hat{\theta}$ is bounded. Similarly, $\tilde{\nu}$ and $(\Lambda\Psi - I)$ are also bounded. By Assumption 7, $M(x, \hat{\nu}, t)$ is bounded. Since geodesics have constant speed and $E = \int_0^1 \gamma_s(s)^T M(x, \hat{\nu}, t)\gamma_s(s)ds = \langle\gamma_s, \gamma_s\rangle$ is bounded, γ_s is bounded. Therefore, $\dot{E} \in \mathcal{L}_\infty$. By Barbalat's Lemma, $x(t) \to x_d(t)$ as $t \to \infty$.

REFERENCES

[1] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. NJ: Dover Publications, 2005, (original publication by Prentice-Hall Inc., 1989).

[2] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice-Hall, 1996.

[3] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*. Prentice-Hall, 1989.

[4] J.-J. E. Slotine and W. Li, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991.

[5] M. Krstic, P. V. Kokotovic, and I. Kanellakopoulos, *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.

[6] K. J. Åström and B. Wittenmark, *Adaptive control*. Courier Corporation, 2013.

[7] A. Astolfi, D. Karagiannis, and R. Ortega, *Nonlinear and adaptive control with applications*. Springer Science & Business Media, 2007.

[8] M. Guay and T. Zhang, "Adaptive extremum seeking control of nonlinear dynamic systems with parametric uncertainties," *Automatica*, vol. 39, no. 7, pp. 1283–1293, 2003.

[9] N. Hovakimyan and C. Cao, *L1 adaptive control theory: Guaranteed robustness with fast adaptation*. SIAM, 2010.

[10] K. Narendra and A. Annaswamy, "Robust adaptive control in the presence of bounded disturbances," *IEEE Transactions on Automatic Control*, vol. 31, no. 4, pp. 306–315, 1986.

[11] H. S. Hussain, Y. Yildiz, M. Matsutani, A. M. Annaswamy, and E. Lavretsky, "Computable delay margins for adaptive systems with state variables accessible," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5039–5054, 2017.

[12] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[13] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.

[14] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[15] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[16] S. P. Karason and A. M. Annaswamy, "Adaptive control in the presence of input constraints," *IEEE Transactions on Automatic Control*, vol. 39, no. 11, pp. 2325–2330, 1994.

[17] E. Lavretsky and N. Hovakimyan, "Positive μ-modification for stable adaptation in the presence of input constraints," in *Proceedings of the 2004 American Control Conference*, vol. 3. IEEE, 2004.

[18] J. E. Gaudio, A. M. Annaswamy, and E. Lavretsky, "Adaptive control of hypersonic vehicles in the presence of rate limits," in *2018 AIAA Guidance, Navigation, and Control Conference*, 2018, p. 0846.

[19] T. Jaakkola, M. Jordan, and S. Singh, "Convergence of stochastic iterative dynamic programming algorithms," *Advances in neural information processing systems*, vol. 6, 1993.

[20] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, vol. 12, 1999.

[21] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Reinforcement learning with fast stabilization in linear dynamical systems," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 5354–5390.

[22] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*. PMLR, 2020, pp. 2137–2143.

[23] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in neural information processing systems*, 2017, pp. 908–918.

[24] S. Koos, J.-B. Mouret, and S. Doncieux, "Crossing the reality gap in evolutionary robotics by promoting transferable controllers," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, 2010, pp. 119–126.

[25] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, "Sim-to-real: Learning agile locomotion for quadruped robots," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[26] N. Fulton and A. Platzer, "Safe reinforcement learning via formal methods: Toward safe control through proof and learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[27] A. Roy, H. Xu, and S. Pokutta, "Reinforcement learning under model mismatch," in *Advances in neural information processing systems*, 2017.

[28] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade, "Towards generalization and simplicity in continuous control," in *Advances in Neural Information Processing Systems*, 2017, pp. 6550–6561.

[29] W. Lohmiller and J.-J. E. Slotine, "On contraction analysis for non-linear systems," *Automatica*, vol. 34, no. 6, pp. 683–696, 1998.

[30] I. R. Manchester and J.-J. E. Slotine, "Control contraction metrics: Convex and intrinsic criteria for nonlinear feedback design," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 3046–3053, 2017.

[31] B. T. Lopez and J.-J. E. Slotine, "Adaptive nonlinear control with contraction metrics," *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 205–210, 2020.

[32] A. S. Morse, "High-order parameter tuners for the adaptive control of linear and nonlinear systems," in *Systems, Models and Feedback: Theory and Applications*. Birkhauser Boston, 1992, pp. 339–364.

[33] R. Ortega, "On morse's new adaptive controller: parameter convergence and transient performance," *IEEE transactions on Automatic Control*, vol. 38, no. 8, pp. 1191–1202, 1993.

[34] S. Evesque, A. M. Annaswamy, S. Niculescu, and A. P. Dowling, "Adaptive control of a class of time-delay systems," *Journal of Dynamic Systems, Measurement, and Control*, vol. 125, no. 2, p. 186, 2003.

[35] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.

[36] ——, *Lectures on Convex Optimization*. Springer International Publishing, 2018.

[37] W. Su, S. Boyd, and E. J. Candès, "A differential equation for modeling nesterov's accelerated gradient method: Theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[38] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7351–E7358, 2016.

[39] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie, "Direct runge-kutta discretization achieves acceleration," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, 2017.

[41] J. E. Gaudio, A. M. Annaswamy, M. A. Bolender, E. Lavretsky, and T. E. Gibson, "A class of high order tuners for adaptive systems," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 391–396, apr 2021.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2023.3290037

16

[42] J. E. Gaudio, A. M. Annaswamy, J. M. Moreu, M. A. Bolender, and T. E. Gibson, "Accelerated learning with robustness to adversarial regressors," *Proceedings of the 3rd Conference on Learning for Dynamics and Control, PMLR 144:636-650*, 2020.

[43] J. M. Moreu and A. M. Annaswamy, "A stable high-order tuner for general convex functions," *IEEE Control Systems Letters*, vol. 6, pp. 566–571, 2022.

[44] A. M. Annaswamy, "Adaptive control and intersections with reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, 2023.

[45] J. E. Gaudio, T. E. Gibson, A. M. Annaswamy, M. A. Bolender, and E. Lavretsky, "Connections between adaptive control and optimization in machine learning," *58th IEEE Conference on Decision and Control (CDC)*, 2019.

[46] N. Matni, A. Proutiere, A. Rantzer, and S. Tu, "From self-tuning regulators to reinforcement learning and back again," in *IEEE 58th Conference on Decision and Control (CDC)*, 2019.

[47] T. Westenbroek, E. Mazumdar, D. Fridovich-Keil, V. Prabhu, C. J. Tomlin, and S. S. Sastry, "Adaptive control for linearizable systems using on-policy reinforcement learning," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 118–125.

[48] R. Sun, M. L. Greene, D. M. Le, Z. I. Bell, G. Chowdhary, and W. E. Dixon, "Lyapunov-based real-time and iterative adjustment of deep neural networks," *IEEE Control Systems Letters*, 2021.

[49] A. Guha and A. Annaswamy, "Online policies for real-time control using mrac-rl," in *60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 1806–1811.

[50] S. M. Richards, N. Azizan, J.-J. Slotine, and M. Pavone, "Adaptive-Control-Oriented Meta-Learning for Nonlinear Systems," in *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[51] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 4192–4201.

[52] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1467–1476.

[53] A. M. Annaswamy, A. Guha, Y. Cui, S. Tang, P. A. Fisher, and J. E. Gaudio, "Online algorithms and policies using adaptive and machine learning approaches," *arXiv preprint arXiv:2105.06577*, 2023.

[54] A. Morgan and K. Narendra, "On the stability of nonautonomous differential equations $\dot{x} = [A + B(t)]x$, with skew symmetric matrix $B(t)$," *SIAM Journal on Control and Optimization*, vol. 15, no. 1, pp. 163–176, 1977.

[55] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[56] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[57] M. P. Do Carmo, *Riemannian geometry*. Springer Science & Business Media, 2013.

[58] A. P. Morgan and K. S. Narendra, "On the uniform asymptotic stability of certain linear nonautonomous differential equations," *SIAM Journal on Control and Optimization*, vol. 15, no. 1, pp. 5–24, 1977.

[59] M. Holzleitner, L. Gruber, J. Arjona-Medina, J. Brandstetter, and S. Hochreiter, "Convergence proof for actor-critic methods applied to ppo and rudder," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLVIII*. Springer, 2021, pp. 105–130.

[60] R. M. Sanner and J.-J. E. Slotine, "Gaussian networks for direct adaptive control," *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 837–863, 1992.

[61] H. Palanthandalam-Madapusi, J. Hoagg, and D. Bernstein, "Basis-function optimization for subspace-based nonlinear identification of systems with measured-input nonlinearities," in *Proceedings of the 2004 American Control Conference*, vol. 5, 2004, pp. 4788–4793 vol.5.

[62] E. Lavretsky and K. A. Wise, *Robust and Adaptive Control with Aerospace Applications*. Springer London, 2013.

[63] G. H. Hardy, E. M. Wright *et al.*, *An introduction to the theory of numbers*. Oxford university press, 1979.

[64] B. Kouvaritakis and M. Cannon, "Model predictive control," *Switzerland: Springer International Publishing*, vol. 38, 2016.

[65] Z. T. Dydek, A. M. Annaswamy, and E. Lavretsky, "Adaptive control of quadrotor UAVs: a design trade study with flight evaluations," *IEEE Transactions on control systems technology*, vol. 21, no. 4, pp. 1400–1406, 2012.

[66] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[67] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[68] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, "Deep drone racing: From simulation to reality with domain randomization," *IEEE Transactions on Robotics*, vol. 36, no. 1, pp. 1–14, 2019.

[69] A. Somanath and A. Annaswamy, "Adaptive control of hypersonic vehicles in presence of aerodynamic and center of gravity uncertainties," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 4661–4666.

[70] M. Schwager and A. M. Annaswamy, "Direct adaptive control of multi-input plants with magnitude saturation constraints," in *Proceedings of the 44th IEEE Conference on Decision and Control*. IEEE, 2005.

Anuradha Annaswamy is Founder and Director of the Active-Adaptive Control Laboratory (AACL) in the Department of Mechanical Engineering at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. Her research interests span adaptive control theory and its applications to aerospace, automotive, propulsion, and energy systems. She has received best paper awards (Axelby; CSM), Distinguished Member, and Distinguished Lecturer awards from the IEEE Control Systems Society (CSS) and a Presidential Young Investigator award from NSF. She is the author of a graduate textbook on adaptive control, co-editor of two vision documents on smart grids and the Impact of Control Technology report, and a coauthor of the National Academy of Sciences Committee report on the Future of Electric Power in the United States in 2021. She is a Fellow of IEEE and IFAC. She served as the President of CSS in 2020.

Anubhav Guha is a Masters student in the AACL at MIT. His research interests and focus include adaptive control, reinforcement and machine learning, and robotics.

Yingnan Cui is a PhD student in the AACL in the Department of Mechanical Engineering at MIT. His research interests include adaptive control and optimization methods for machine learning.

Sunbochen Tang is a PhD student in the AACL at MIT. His research interests include adaptive control, optimization methods, and machine learning.

Peter Fisher is a Masters student in the AACL at MIT. His research interests include adaptive control, optimization methods, and statistical learning.

Joseph Gaudio received his PhD degree in Mechanical Engineering from MIT in 2020. He is currently a Staff Engineer at Aurora Flight Sciences, a Boeing Company. His research interests are in adaptive control and machine learning.