Adaptive dynamic programming for terminally constrained finite-horizon optimal control problems

L. Andrews, J. R. Klotz, R. Kamalapurkar, and W. E. Dixon

***Abstract*—Adaptive dynamic programming is applied to control-affine nonlinear systems with uncertain drift dynamics to obtain a near-optimal solution to a finite-horizon optimal control problem with hard terminal constraints. A reinforcement learning-based actor-critic framework is used to approximately solve the Hamilton-Jacobi-Bellman equation, wherein critic and actor neural networks (NN) are used for approximate learning of the optimal value function and control policy, while enforcing the optimality condition resulting from the hard terminal constraint. Concurrent learning-based update laws relax the restrictive persistence of excitation requirement. A Lyapunov-based stability analysis guarantees uniformly ultimately bounded convergence of the enacted control policy to the optimal control policy.**

I. Introduction

Many practical problems are best described as finite-horizon optimal control problems with hard terminal constraints. For example, missile interception problems seek to hit a target within a finite amount of time and with a specified terminal angle; an aircraft seeks to arrive at a specified destination and also minimize fuel consumption along the way; a spacecraft must obtain an orbital position with a specific velocity at a given final time. An abundance of applications can be formulated into this type of optimal control problem. However, an analytical solution is not generally feasible because it requires the solution to a time-varying Hamilton-Jacobi-Bellman (HJB) equation, which is a nonlinear partial differential equation. This motivates the development of an approximate optimal solution.

Approximate optimal control solutions are well-established for unconstrained infinite-horizon problems. One type of suboptimal control method utilizes the State-Dependent Riccati Equation (SDRE) technique, which minimizes the cost functional by restructuring the nonlinear system into a linear form with state-dependent coefficients. Then the corresponding Algebraic Riccati equation is solved at each time step. However, SDRE techniques have a high computing cost, and therefore require advanced numerical methods for implementation [1]. One alternative to the computationally expensive SDRE method is the $\theta - \mathrm{D}$ technique [2]. Rather than solve the Algebraic Riccati Equation at each step, the $\theta - \mathrm{D}$ method uses a power series expansion of the costate to solve for a closed form approximate optimal control solution.

Alternatively, adaptive dynamic programming (ADP) can be used with an actor-critic architecture to learn the optimal control solution. In ADP methods, reinforcement learning is used with an actor-critic framework to obtain an approximate solution to the HJB equation and learn the optimal control policy. The actor-critic framework consists of actor and critic neural networks (NNs), which approximate the optimal control policy and the optimal value function. The actor NN interacts with the environment through the control, and the critic NN evaluates the response to update the NNs. A strong foundation has been established for terminally unconstrained, infinite-horizon problems solved by either offline or online ADP techniques [3]–[7]. ADP methods have also been applied to the unconstrained, finite-horizon problem [8]–[11]. However, less attention has been given to the finite-horizon problem with hard terminal constraints.

Current approaches to the finite-horizon problem with hard terminal constraints fall into two main categories. The first category involves numerically solving for the optimal control and then using the open loop optimal control solution with another technique, such as neighboring optimal control [12]. Alternatively, an approximate optimal feedback control law can be developed. For example, a closed-form series solution has been developed in [13] to exactly satisfy the hard terminal constraint; however, convergence is not guaranteed for a highly nonlinear system. Constrained optimal control problems have been approached with offline ADP methods in [14]–[16]. However, the offline ADP solution uses exact model knowledge to train the weights. Consequently, the solution does not have the flexibility to react online to uncertainty. An online ADP solution can provide this flexibility.

The objective of this paper is to develop an online ADP solution for terminally constrained, finite-horizon optimal control problems with linear-in-the-parameters (LP) uncertainty in the drift dynamics. Online ADP methods typically employ a persistence of excitation (PE) condition, which requires sufficient exploration in the observed data to guarantee approximation convergence [4]. However, in this work, concurrent learning (CL) is used to eliminate the need for the PE condition. In CL, the approximate Bellman error is evaluated at a pre-sampled set of data points. As a result, the approximate Bellman error can be evaluated at any point in state space rather than the limited set of observed data points

L. Andrews, J. R. Klotz, R. Kamalapurkar, and W. E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. Email: {landr010, jklotz, rkamalapurkar, wdixon}@ufl.edu.

This research is supported in part by NSF award numbers 1161260, 1217908, ONR grant number N00014-13-1-0151, and a contract with the AFRL Mathematical Modeling and Optimization Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agency.

[17]. A Lyapunov-based stability analysis is presented to establish uniformly ultimately bounded (UUB) convergence of the enacted control policy to the optimal control policy.

II. Problem Formulation

Consider the nonlinear control affine system

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

where $x \in \mathbb{R}^n$ represents the states, $f : \mathbb{R}^n \to \mathbb{R}^n$ represents the drift dynamics, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ is the control effectiveness matrix, and $\hat{u} \in \mathbb{R}^m$ is the control input. The class of systems to be considered satisfies the following:

Assumption 1. The drift dynamics $f(x)$ are unknown with LP uncertainty, are locally Lipschitz, and $f(0) = 0$. The control effectiveness matrix $g(x)$ is known, bounded, and locally Lipschitz.

The objective is to solve a constrained optimal control problem by minimizing the cost functional $J \in \mathbb{R}$, defined as

$$J \triangleq \int_{t_0}^{t_f} \left(x^T Q x + \hat{u}^T R \hat{u}\right) dt, \tag{2}$$

subject to the dynamics from (1) and the hard terminal constraint $\psi(x(t_f)) = \psi_f \in \mathbb{R}^p$, where $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are constant, positive definite symmetric weighting matrices. The hard constraint on the terminal state is defined as $\psi_f \triangleq 0_p$, where $0_p \in \mathbb{R}^p$ denotes a p-dimensional vector of zeros. The optimal value function, $V : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, is defined using the minimum cost functional within the set of admissible control policies $\mathcal{U}$ [5], given by

$$V(x,t) \triangleq \int_t^{t_f} \left(x^T Q x + u(x,\tau)^T R u(x,\tau)\right) d\tau + \nu^T \left(\psi(x(t_f)) - \psi_f\right), \tag{3}$$

subject to the dynamic constraints, where $\nu \in \mathbb{R}^p$ is the optimal constant vector of Lagrange multipliers, and $u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ is the optimal control policy. If the hard terminal constraint is satisfied, then $(\psi(x(t_f)) - \psi_f) = 0_p$. As a result, the optimal value function will have no sensitivity to ν. Consequently, the optimal value function must satisfy the necessary condition [13]

$$\frac{\partial V(x,t)}{\partial \nu} = 0_p,\ \forall x \in \mathbb{R}^n,\ t \in \mathbb{R}. \tag{4}$$

The HJB equation is given by [18]

$$H(x,t) + \frac{\partial V(x,t)}{\partial t} = 0, \tag{5}$$

where the Hamiltonian, $H : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, is defined as

$$H(x,t) \triangleq x^T Q x + u(x,t)^T R u(x,t) + \frac{\partial V(x,t)}{\partial x}\left(f(x) + g(x)u(x,t)\right). \tag{6}$$

Using the stationary condition [18], the optimal control policy is defined as

$$u(x,t) = -\frac{1}{2} R^{-1} g(x)^T \left(\frac{\partial V(x,t)}{\partial x}\right)^T, \tag{7}$$

where it is assumed that a minimizing policy exists and that the value function is continuously differentiable. Solving for the optimal control policy is often analytically infeasible, since it requires the solution to a nonlinear partial differential equation. Consequently, an approximate optimal control solution is desired. To facilitate the development of an approximate control solution, an identifier is designed in the following section for identification of the uncertain parameters in the dynamics.

III. System Identification

The drift dynamics $f(x)$ can be linearly parametrized as $f(x) = Y(x)\theta$, where $Y : \mathbb{R}^n \to \mathbb{R}^{n\times s}$ is the known regression matrix, and $\theta \in \mathbb{R}^s$ is the vector of constant unknown parameters. The drift dynamics are estimated with $\hat{f}\left(x,\hat{\theta}\right) : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}^n$, defined as $\hat{f}\left(x,\hat{\theta}\right) \triangleq Y(x)\hat{\theta}$, where $\hat{\theta} \in \mathbb{R}^s$ is a vector of the estimated uncertain parameters. The estimated dynamics are defined as

$$\dot{\hat{x}} = \hat{f}\left(x,\hat{\theta}\right) + g(x)\hat{u} + k_x\tilde{x}, \tag{8}$$

where $\tilde{x} \triangleq x - \hat{x}$ is the state estimation error, and $k_x \in \mathbb{R}^{n\times n}$ is a constant, diagonal, positive definite gain matrix. Using (1) and (8), the state estimation error dynamics are given by

$$\dot{\tilde{x}} = Y(x)\tilde{\theta} - k_x\tilde{x}, \tag{9}$$

where the parameter estimation error is defined as $\tilde{\theta} \triangleq \theta - \hat{\theta}$.

A. Parameter Update Law

Traditional parameter update laws rely on the PE condition to guarantee parameter identification. Rather than assume that the system states are exciting over all time, the PE condition is relaxed by using a CL approach. Contrary to the PE condition, CL only requires that the system states be exciting for a finite amount of time to provide a rich history of data points. The estimation of the uncertain parameters is updated with a CL-based gradient descent law given by

$$\dot{\hat{\theta}} \triangleq \Gamma_\theta Y(x)^T \tilde{x} + \Gamma_\theta k_\theta \sum_{i=1}^{M} Y_i^T \left(\dot{x}_i - g_i\hat{u}_i - Y_i\hat{\theta}\right), \tag{10}$$

where $\Gamma_\theta \in \mathbb{R}^{s\times s}$ is a constant positive definite gain matrix, $k_\theta \in \mathbb{R}$ is a constant positive gain, and $(\cdot)_i$ represents evaluation at a recorded data point. The update law depends on the derivative $\dot{x}_i$, which is unknown. However, this can be numerically obtained with a smoothing filter based on past and current data [19]. To incorporate new information, the history stack of recorded points is updated using a

singular value maximizing algorithm [20]. By substituting $\dot{x}_i = Y_i\theta + g_i\hat{u}_i$, (10) can be expressed as

$$\dot{\hat{\theta}} = \Gamma_\theta Y(x)^T \tilde{x} + \Gamma_\theta k_\theta \sum_{i=1}^{M} \left(Y_i^T Y_i\right) \tilde{\theta}. \tag{11}$$

Assumption 2. The CL-based matrix: $\sum_{i=1}^{M} Y_i^T Y_i \in \mathbb{R}^{s\times s}$ is full rank $\forall t \in [t_0, t_f]$, with a minimum eigenvalue denoted by $\underline{y_1} > 0$.

To initially satisfy assumption 2, the data stack is initialized with pre-obtained experimental data. The assumption is weaker than the usual PE assumption because it only requires the states to be exciting for a finite amount of time. Furthermore, satisfaction of Assumption 2 can be verified online unlike the PE condition [17].

B. Convergence Analysis

Consider the following positive definite continuously differentiable Lyapunov function candidate

$$V_0(z) \triangleq \frac{1}{2}\tilde{x}^T\tilde{x} + \frac{1}{2}\tilde{\theta}^T\Gamma_\theta^{-1}\tilde{\theta}, \tag{12}$$

which is bounded by

$$\underline{v_0}\|z\|^2 \leq V_0(z) \leq \overline{v_0}\|z\|^2, \tag{13}$$

where $\underline{v_0} \triangleq \frac{1}{2}\min\left(1, \underline{\gamma_\theta}\right)$, $\overline{v_0} \triangleq \frac{1}{2}\max\left(1, \overline{\gamma_\theta}\right)$, $z \triangleq \begin{bmatrix} \tilde{x}^T & \tilde{\theta}^T \end{bmatrix}^T \in \mathbb{R}^{n+s}$, and $\underline{\gamma_\theta}$, $\overline{\gamma_\theta} \in \mathbb{R}$ are the minimum and maximum eigenvalues of Γ_θ^{-1}.

The time derivative of V_0 is given by

$$\dot{V}_0 = -\tilde{x}^T k_x \tilde{x} - \tilde{\theta}^T k_\theta \sum_{i=1}^{M} \left(Y_i^T Y_i\right) \tilde{\theta}. \tag{14}$$

Using (13), (14) can be upperbounded as

$$\dot{V}_0 \leq -\min\left(\underline{k_x}, k_\theta \underline{y_1}\right) \frac{V_0}{\overline{v_0}}, \tag{15}$$

where $\underline{k_x}$ is the minimum eigenvalue of k_x. From (15), $\left\|\tilde{\theta}(t)\right\| \to 0$ and $\|\tilde{x}(t)\| \to 0$ exponentially fast. Furthermore, it can be shown that $\left\|\dot{\tilde{x}}(t)\right\| \to 0$ exponentially fast, resulting in exponential regulation of the parameter and state derivative estimation errors [17]. The parameter estimates are used within the development of an approximate optimal control solution in the following section. Note that some function arguments are suppressed hereafter for brevity.

IV. Approximate Solution

An approximate optimal control policy is developed with an actor-critic NN framework based on reinforcement learning [4]. Typically in ADP, the value function is represented with one NN. However, due to the hard constraint, the value function is augmented with a second NN, which represents the effect of the hard constraint. A temporary assumption is made that the state x lies on a compact set $\chi \subset \mathbb{R}^n$. This assumption is common in NN-based control and will be relaxed in the stability analysis, as long as the initial condition $x(0)$ is bounded.

Motivated by the development in [14], the optimal value function can be approximated as

$$V = W^T\sigma(x,t) + \Gamma^T\phi(x,t,\nu) - \nu^T\psi_f + \varepsilon(x,t,\nu), \tag{16}$$

where $W \in \mathbb{R}^N$ and $\Gamma \in \mathbb{R}^L$ are the unknown ideal NN weights, $\sigma : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^N$ and $\phi : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^L$ are the continuous basis functions, and $\varepsilon : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$ is the unknown reconstruction error. The ideal NN weights are bounded above by positive constants, i.e. $\|W\| \leq \bar{W} \in \mathbb{R}$ and $\|\Gamma\| \leq \bar{\Gamma} \in \mathbb{R}$. The selected basis functions σ and ϕ as well as the partial derivatives with respect to state and time are bounded. The reconstruction error and its partial derivatives with respect to state and time are upperbounded, i.e. $\Lambda|\varepsilon| \leq \bar{\varepsilon} \in \mathbb{R}_{\geq 0}$, $\Lambda\left\|\frac{\partial \varepsilon}{\partial x}\right\| \leq \bar{\varepsilon}_x \in \mathbb{R}_{\geq 0}$, and $\Lambda\left|\frac{\partial \varepsilon}{\partial t}\right| \leq \bar{\varepsilon}_t \in \mathbb{R}_{\geq 0}$, where the operator $\Lambda(\cdot) \triangleq \sup_{x\in\chi,\, t\in[t_0,t_f]} (\cdot)$. The upperbounds $\bar{W}$, $\bar{\Gamma}$, $\bar{\varepsilon}$, $\bar{\varepsilon}_x$, $\bar{\varepsilon}_t$ are assumed to be known.

Enforcing the optimality condition from (4), the optimal Lagrange multiplier is given by

$$\left(\frac{\partial \phi(x,t,\nu)}{\partial \nu}\right)^T \Gamma - \psi_f + \left(\frac{\partial \varepsilon(x,t,\nu)}{\partial \nu}\right)^T = 0_p. \tag{17}$$

Since the value function in (16) contains the unknown ideal weights W and Γ, and the unknown optimal Lagrange multipliers ν, the value function is approximated as

$$\hat{V} \triangleq \hat{W}_c^T\sigma(x,t) + \hat{\Gamma}_c^T\phi(x,t,\hat{\nu}) - \hat{\nu}^T\psi_f, \tag{18}$$

where $\hat{W}_c \in \mathbb{R}^N$ and $\hat{\Gamma}_c \in \mathbb{R}^L$ are approximations of W and Γ, and where $\hat{\nu} \in \mathbb{R}^p$ is an estimate of ν. The update policies for these estimates are given in the following section.

Using the optimal control definition from (7) and the value function in (16), the optimal control can be represented as

$$u(x,t,\nu) = -\frac{1}{2}R^{-1}g(x)^T\Big(\sigma_x(x,t)^T W + \phi_x(x,t,\nu)^T\Gamma + \varepsilon_x(x,t,\nu)^T\Big), \tag{19}$$

where $\sigma_x \triangleq \frac{\partial \sigma}{\partial x} \in \mathbb{R}^{N\times n}$, $\phi_x \triangleq \frac{\partial \phi}{\partial x} \in \mathbb{R}^{L\times n}$, and $\varepsilon_x \triangleq \frac{\partial \varepsilon}{\partial x} \in \mathbb{R}^{1\times n}$. The approximate optimal control solution is given by

$$\hat{u} = -\frac{1}{2}R^{-1}g(x)^T\left(\sigma_x(x,t)^T\hat{W}_a + \phi_x(x,t,\hat{\nu})^T\hat{\Gamma}_a\right), \tag{20}$$

where $\hat{W}_a \in \mathbb{R}^N$ and $\hat{\Gamma}_a \in \mathbb{R}^L$ are approximations of the ideal NN weights W and Γ.

To develop an estimate for the optimal Lagrange multipliers, the optimality condition in (4) is applied to the approximated value function, resulting in the following condition:

$$\left(\frac{\partial \phi(x,t,\hat{\nu})}{\partial \hat{\nu}}\right)^T \hat{\Gamma}_c - \psi_f = 0_p. \tag{21}$$

The approximate Hamiltonian $\hat{H} \in \mathbb{R}$, rewritten with the NN representation, is given by

$$\hat{H} = x^T Q x + \hat{u}^T R \hat{u} + \left(\hat{W}_c^T \sigma_x + \hat{\Gamma}_c^T \phi_x \right) (f + g\hat{u}). \quad (22)$$

The Bellman error (BE) is defined as the difference between the approximated HJB and the optimal HJB. Using (5) and the approximated drift dynamics $\hat{f}\left(x, \hat{\theta}\right)$, an estimate of the BE, $\hat{\delta} \in \mathbb{R}$, is expressed as

$$\hat{\delta} \triangleq \hat{W}_c^T \left(\sigma_t + \sigma_x \left(\hat{f} + g\hat{u} \right) \right) + \hat{\Gamma}_c^T \left(\phi_t + \phi_x \left(\hat{f} + g\hat{u} \right) \right) + x^T Q x + \hat{u}^T R \hat{u}, \quad (23)$$

where $\sigma_t \triangleq \frac{\partial \sigma}{\partial t} \in \mathbb{R}^N$, and $\phi_t \triangleq \frac{\partial \phi}{\partial t} \in \mathbb{R}^L$. The measurable estimate of the BE in (23) can be computed to reveal the proximity of the approximations to the actual values. Consequently, the estimated BE can be used to learn the ideal NN weights by incorporating it into the weight update laws, which are developed in the following section.

V. NN Weight Update Laws

In this section, weight update laws are defined so that the approximated NN weights converge to the ideal NN weights. In the NN representation, the integrated cost is approximated by $\hat{W}_c^T \sigma$. Therefore, the weights $\hat{W}_c$ are trained based on the BE estimate corresponding to the optimal control problem without the hard constraint, denoted by $\hat{\delta}_1 \in \mathbb{R}$. The hard terminal constraint is approximated by $\hat{\Gamma}_c^T \phi$. Consequently, the weights $\hat{\Gamma}_c$ are trained based on $\hat{\delta}_{HC} \triangleq \left(\hat{\delta} - \hat{\delta}_1 \right) \in \mathbb{R}$. By taking this difference, the contribution to the BE from the hard terminal constraint is isolated.

A. Integrated Cost Weight Estimates

Since $\hat{W}_c$ and $\hat{W}_a$ represent estimates associated with the integrated cost, the weights are trained based on the BE estimate that corresponds to the terminally unconstrained problem, $\hat{\delta}_1$, defined as

$$\hat{\delta}_1 \triangleq \hat{W}_c^T \left(\sigma_t + \sigma_x \left(\hat{f} + g\hat{u}_1 \right) \right) + x^T Q x + \hat{u}_1^T R \hat{u}_1, \quad (24)$$

where the approximate optimal control $\hat{u}_1$ is given by

$$\hat{u}_1 \left(x, t, \hat{W}_a \right) = -\frac{1}{2} R^{-1} g(x)^T \sigma_x(x, t) \hat{W}_a, \quad (25)$$

where $\varepsilon_1 : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is the reconstruction error and where $\varepsilon_{1x}(x, t) \triangleq \frac{\partial \varepsilon_1(x,t)}{\partial x} \in \mathbb{R}^{1 \times n}$, which is upperbounded by a constant, i.e. $\Lambda(\|\varepsilon_{1x}\|) \leq \bar{\varepsilon}_{1x} \in \mathbb{R}_{\geq 0}$. The constant $\bar{\varepsilon}_{1x}$ is assumed to be known.

The critic weights $\hat{W}_c$ are updated based on the BE estimate $\hat{\delta}_1$. To provide sufficient data richness for learning, the BE estimate is also evaluated over a set of k user-selected sample points, given by $d_j = \{x_j \mid x_j \in \chi\} \cup \{t_j \mid t_j \in [t_0, t_f]\}$ for $j = 1, \cdots, k$. The critic NN update law, $\dot{\hat{W}}_c \in \mathbb{R}^N$, is defined by the CL-based gradient descent of $\hat{\delta}_1$, given by

$$\dot{\hat{W}}_c \triangleq \text{proj} \left(-\eta_{c1} \frac{\omega_1}{\Omega_1} \hat{\delta}_1 - \eta_{c2} \sum_{j=1}^{k} \left(\frac{\omega_{1j}}{\Omega_{1j}} \hat{\delta}_{1j} \right) \right), \quad (26)$$

where $(\cdot)_j$ represents evaluation at d_j, the adaptation gains $\eta_{c1}, \eta_{c2} \in \mathbb{R}$ are positive constants, the regressor vectors $\omega_1, \omega_{1j} \in \mathbb{R}^N$ are defined as $\omega_1 \triangleq \sigma_t + \sigma_x \left(\hat{f} + g\hat{u}_1 \right)$ and $\omega_{1j} \triangleq \sigma_{tj} + \sigma_{xj} \left(\hat{f}_j + g_j \hat{u}_{1j} \right)$, the normalization terms $\Omega_1, \Omega_{1j} \in \mathbb{R}$ are defined as $\Omega_1 \triangleq \sqrt{1 + \omega_1^T \omega_1}$ and $\Omega_{1j} \triangleq \sqrt{1 + \omega_{1j}^T \omega_{1j}}$, and $\text{proj}\{\cdot\}$ is a smooth orthogonal projection operator used to bound the weight estimates [21].

The actor NN update law, $\dot{\hat{W}}_a \in \mathbb{R}^N$, is given by

$$\dot{\hat{W}}_a \triangleq \text{proj} \left(-\eta_{a1} \left(\hat{W}_a - \hat{W}_c \right) \right), \quad (27)$$

where $\eta_{a1} \in \mathbb{R}$ is a positive constant gain.

B. Terminal Constraint Weight Estimates

Since $\hat{\Gamma}_c$ and $\hat{\Gamma}_a$ represent the hard terminal constraint, they are updated based on $\hat{\delta}_{HC}$, which represents the contributions to the estimated BE that come from the hard terminal constraint.

The critic NN update law for $\hat{\Gamma}_c$ is defined by the CL-based gradient descent of $\hat{\delta}_{HC}$, given by

$$\dot{\hat{\Gamma}}_c \triangleq -\eta_{c3} \frac{\omega_2}{\Omega_2} \hat{\delta}_{HC} - \eta_{c4} \sum_{j=1}^{k} \left(\frac{\omega_{2j}}{\Omega_{2j}} \hat{\delta}_{HC_j} \right), \quad (28)$$

where $\eta_{c3}, \eta_{c4} \in \mathbb{R}$ are positive constant adaptation gains, $\omega_2, \omega_{2j} \in \mathbb{R}^L$ are the regressor vectors defined as $\omega_2 \triangleq \phi_t + \phi_x \left(\hat{f} + g\hat{u} \right)$ and $\omega_{2j} \triangleq \phi_{tj} + \phi_{xj} \left(\hat{f}_j + g_j \hat{u}_j \right)$, and $\Omega_2, \Omega_{2j} \in \mathbb{R}$ are normalization terms given by $\Omega_2 = \sqrt{1 + \omega_2^T \omega_2}$ and $\Omega_{2j} = \sqrt{1 + \omega_{2j}^T \omega_{2j}}$.

The actor NN update law for $\hat{\Gamma}_a$ is given by

$$\dot{\hat{\Gamma}}_a \triangleq \text{proj} \left(-\eta_{a2} \left(\hat{\Gamma}_a - \hat{\Gamma}_c \right) \right), \quad (29)$$

where $\eta_{a2} \in \mathbb{R}$ is a positive constant gain.

VI. Stability Analysis

To facilitate the stability analysis, let $G \triangleq g R^{-1} g^T \in \mathbb{R}^{n \times n}$, $G_\sigma \triangleq \sigma_x G \sigma_x^T \in \mathbb{R}^{N \times N}$, and let the minimum eigenvalue of Q be denoted as $\underline{q}$. The drift dynamics and regression matrix can be upperbounded on the compact set χ as $\|f\| \leq L_f \|x\|$, $\|Y\| \leq \overline{Y}$, and $\|Y_i\| \leq \overline{Y}_i$ where $L_f, \overline{Y}, \overline{Y}_i \in \mathbb{R}$ are positive constants. The upperbounds on $\|\sigma_x\|$ and $\|\phi_x\|$ are defined as $\bar{\sigma}_x$ and $\bar{\phi}_x$. For brevity, let $\overline{Y}_{12}, \overline{Y}_{34} \in \mathbb{R}$ be defined as $\overline{Y}_{12} \triangleq \left(\eta_{c1} \overline{Y} + \eta_{c2} \max_{i \in [1,M]} \left(\overline{Y}_i \right) \right)$ and $\overline{Y}_{34} \triangleq \left(\eta_{c3} \overline{Y} + \eta_{c4} k \max_{i \in [1,M]} \left(\overline{Y}_i \right) \right)$. Let the operators Λ_t and Λ_x

be defined as $\Lambda_t(\cdot) \triangleq \sup_{t\in[t_0,t_f]}(\cdot)$ and $\Lambda_x(\cdot) \triangleq \sup_{x\in\chi}(\cdot)$. Lastly, let the operator $\Upsilon_{a,b}(\cdot)$ be defined as $\Upsilon_{a,b}(\cdot) \triangleq a(\cdot) + b\sum_{j=1}^{k}(\cdot)_j$ for any $a,\, b \in \mathbb{R}$.

Throughout the stability analysis, unmeasurable forms of the BE estimates are used, which contain the weight estimation errors $\tilde{W}_c$, $\tilde{W}_a$, $\tilde{\Gamma}_c$, $\tilde{\Gamma}_a$ defined as $\tilde{W}_c \triangleq W - \hat{W}_c$, $\tilde{W}_a \triangleq W - \hat{W}_a$, $\tilde{\Gamma}_c \triangleq \Gamma - \hat{\Gamma}_c$, and $\tilde{\Gamma}_a \triangleq \Gamma - \hat{\Gamma}_a$. Expressing the BE estimates in an unmeasurable form allows the weight estimation errors to enter the stability analysis through the weight update laws.

Assumption 3. The CL-based matrices: $c_1 = \sum_{j=1}^{k}\frac{\omega_{1j}\omega_{1j}^T}{\Omega_{1j}} \in \mathbb{R}^{N\times N}$ and $c_2 = \sum_{j=1}^{k}\frac{\omega_{2j}\omega_{2j}^T}{\Omega_{2j}} \in \mathbb{R}^{L\times L}$ are full rank $\forall t \in [t_0, t_f]$, with minimum eigenvalues of $\underline{c_1} \triangleq \inf_{t\in[t_0,t_f]}(\lambda_{\min}(c_1)) > 0$ and $\underline{c_2} \triangleq \inf_{t\in[t_0,t_f]}(\lambda_{\min}(c_2)) > 0$.

Although Assumption 3 cannot be guaranteed a priori, it can be verified online and, in general, can be satisfied by selecting many sample points to create a rich database [17].

Theorem 1. *Provided that Assumptions 1-3 hold and the gains k_θ, η_{c1}, η_{c2}, η_{c3}, η_{c4} are designed such that the following sufficient conditions are met*

$$\underline{q} > \frac{\eta_{c1}}{2}\bar{\varepsilon}_{1x}L_f + \frac{\eta_{c3}}{2}L_f\left|(\bar{\varepsilon}_x - \bar{\varepsilon}_{1x})\right|,$$

$$k_\theta > \frac{\bar{W}\bar{\sigma}_x}{2\underline{y_1}}\overline{Y}_{12} + \frac{\bar{\Gamma}\bar{\phi}_x}{2\underline{y_1}}\overline{Y}_{34},$$

$$\eta_{c2} > \frac{1}{2\underline{c_1}}\left(\eta_{a1} + \eta_{c1}\bar{\varepsilon}_{1x}L_f + \bar{W}\bar{\sigma}_x\overline{Y}_{12}\right),$$

$$\eta_{c4} > \frac{1}{2\underline{c_2}}\left(\eta_{a2} + \eta_{c3}L_f\left|(\bar{\varepsilon}_x - \bar{\varepsilon}_{1x})\right| + \bar{\Gamma}\bar{\phi}_x\overline{Y}_{34}\right), \quad (30)$$

then the state x, the weight estimation errors $\tilde{W}_a$, $\tilde{W}_c$, $\tilde{\Gamma}_a$, $\tilde{\Gamma}_c$, the state estimation error $\tilde{x}$, and the parameter estimation error $\tilde{\theta}$ are UUB, resulting in UUB convergence of the approximate control policy to the optimal control policy.

Proof: Consider the following continuously differentiable, positive definite Lyapunov function candidate

$$V_L(Z,t) \triangleq V(x,t) + \frac{1}{2}\tilde{W}_a^T\tilde{W}_a + \frac{1}{2}\tilde{\Gamma}_a^T\tilde{\Gamma}_a + \frac{1}{2}\tilde{W}_c^T\tilde{W}_c + \frac{1}{2}\tilde{\Gamma}_c^T\tilde{\Gamma}_c + V_0(z), \quad (31)$$

where $Z \triangleq \begin{bmatrix} x^T & \tilde{W}_a^T & \tilde{\Gamma}_a^T & \tilde{W}_c^T & \tilde{\Gamma}_c^T & \tilde{x}^T & \tilde{\theta}^T \end{bmatrix}^T$. From Lemma 4.3 of [22], there exist class $\mathcal{K}$ functions, α_1 and α_2, such that

$$\alpha_1(\|Z\|) \le V_L(Z,t) \le \alpha_2(\|Z\|), \quad (32)$$

for all $t \in [0,\infty)$ and for all $Z \in \mathbb{R}^{2n+2N+2L+s}$. The time derivative of the candidate Lyapunov function is given by

$$\dot{V}_L = \frac{\partial V}{\partial t} + \left(\frac{\partial V}{\partial x}\right)^T(f + g\hat{u}) - \tilde{W}_a^T\left(\dot{\hat{W}}_a\right) - \tilde{\Gamma}_a^T\left(\dot{\hat{\Gamma}}_a\right) - \tilde{W}_c^T\left(\dot{\hat{W}}_c\right) - \tilde{\Gamma}_c^T\left(\dot{\hat{\Gamma}}_c\right) + \dot{V}_0. \quad (33)$$

Substituting $\frac{\partial V}{\partial t}$ from (5) and then using (6), (14), (16), (19), (20), (26), (27), (28), (29), the unmeasurable forms of the BE estimates, Young's Inequality, and completion of the squares, $\dot{V}_L$ can be upperbounded as

$$\dot{V}_L \le -\psi_x\|x\|^2 - \underline{k_x}\|\tilde{x}\|^2 - \frac{\psi_{a1}}{2}\left\|\tilde{W}_a\right\|^2 - \frac{\psi_{a2}}{2}\left\|\tilde{\Gamma}_a\right\|^2 - \psi_\theta\left\|\tilde{\theta}\right\|^2 - \frac{\psi_{c1}}{2}\left\|\tilde{W}_c\right\|^2 - \frac{\psi_{c2}}{2}\left\|\tilde{\Gamma}_c\right\|^2 + \frac{k_{a1}^2}{2\psi_{a1}} + \frac{k_{a2}^2}{2\psi_{a2}} + \frac{k_{c1}^2}{2\psi_{c1}} + \frac{k_{c2}^2}{2\psi_{c2}} + k, \quad (34)$$

where

$$\psi_x \triangleq \underline{q} - \frac{\eta_{c1}}{2}\bar{\varepsilon}_{1x}L_f - \frac{\eta_{c3}}{2}L_f\left|(\bar{\varepsilon}_x - \bar{\varepsilon}_{1x})\right|,\ \psi_{a1} \triangleq \frac{\eta_{a1}}{2},$$

$$\psi_{a2} \triangleq \frac{\eta_{a2}}{2},\quad \psi_\theta \triangleq k_\theta\underline{y_1} - \frac{\bar{W}\bar{\sigma}_x}{2}\overline{Y}_{12} - \frac{\bar{\Gamma}\bar{\phi}_x}{2}\overline{Y}_{34},$$

$$\psi_{c1} \triangleq \eta_{c2}\underline{c_1} - \frac{\eta_{a1}}{2} - \frac{\eta_{c1}}{2}\bar{\varepsilon}_{1x}L_f - \frac{\bar{W}\bar{\sigma}_x}{2}\overline{Y}_{12},$$

$$\psi_{c2} \triangleq \eta_{c4}\underline{c_2} - \frac{\eta_{a2}}{2} - \frac{\eta_{c3}}{2}L_f\left|(\bar{\varepsilon}_x - \bar{\varepsilon}_{1x})\right| - \frac{\bar{\Gamma}\bar{\phi}_x}{2}\overline{Y}_{34},$$

$$k_{a1} \triangleq \Lambda\left(\left\|\frac{1}{2}\left(G_\sigma W + \sigma_x G\varepsilon_x^T\right)\right\|\right) + \frac{1}{2}\bar{\sigma}_x\bar{\phi}_x\bar{\Gamma}\Lambda_x(\|G\|),$$

$$k_{a2} \triangleq \frac{1}{2}\Lambda_x(\|G\|)\bar{\phi}_x\left(\bar{\phi}_x\bar{\Gamma} + \bar{\sigma}_x\bar{W}\right) + \frac{1}{2}\bar{\phi}_x\Lambda\left(\left\|G\varepsilon_x^T\right\|\right),$$

$$k_{c1} \triangleq \Lambda\left\|\frac{1}{2}\Upsilon_{\eta_{c1},\eta_{c2}}\left(\frac{1}{2}\varepsilon_{1x}G\varepsilon_{1x}^T + W^T\sigma_x G\varepsilon_{1x}^T\right) + \Upsilon_{\eta_{c1},\eta_{c2}}\left(\frac{1}{4}\tilde{W}_a^T G_\sigma\tilde{W}_a - \varepsilon_{1t}\right) - \eta_{c2}\sum_{j=1}^{k}\varepsilon_{1xj}f_j\right\|,$$

$$k_{c2} \triangleq \Lambda\left\|\frac{1}{2}\Upsilon_{\eta_{c3},\eta_{c4}}\left(W^T\sigma_x G\left(\varepsilon_x^T - \varepsilon_{1x}^T\right)\right) + \Upsilon_{\eta_{c3},\eta_{c4}}\left(\frac{1}{4}\varepsilon_x G\varepsilon_x^T - \frac{1}{4}\varepsilon_{1x}G\varepsilon_{1x}^T + \varepsilon_{1t} - \varepsilon_t\right) - \eta_{c4}\sum_{j=1}^{k}\left(\varepsilon_{xj} - \varepsilon_{1xj}\right)f_j\right\| + \Lambda_t\left(\left\|\tilde{\Gamma}_a\right\|\left\|\tilde{W}_a\right\|\right)\bar{\sigma}_x\bar{\phi}_x + \frac{1}{2}\left(\eta_{c3} + \eta_{c4}k\right)\Lambda_x(\|G\|)\left(\frac{1}{2}\Lambda_t\left(\left\|\tilde{\Gamma}_a\right\|^2\right)\bar{\phi}_x^2\right.$$

$$+\Lambda_t\left(\left\|\tilde{W}_c\right\|\right)\bar{\sigma}_x\bar{\phi}_x\bar{\Gamma}+\bar{\phi}_x\bar{\Gamma}\bar{\varepsilon}_x$$
$$+\Lambda_t\left(\left\|\tilde{\Gamma}_a\right\|\left\|\tilde{W}_c\right\|\right)\bar{\sigma}_x\bar{\phi}_x\Big),$$

$$k\triangleq\Lambda\left(\frac{1}{4}\varepsilon_x G\varepsilon_x^T\right). \tag{35}$$

A further upperbound for (34) is

$$\dot{V}_L\leq-\frac{\psi_z}{2}\left\|Z\right\|^2,\ \forall\left\|Z\right\|\geq\mu>0, \tag{36}$$

where $\psi_z\triangleq\min\left(\psi_x,\underline{k_x},\psi_\theta,\frac{\psi_{a1}}{2},\frac{\psi_{a2}}{2},\frac{\psi_{c1}}{2},\frac{\psi_{c2}}{2}\right)$, $\mu\triangleq\sqrt{\frac{2k_z}{\psi_z}}$, and $k_z\triangleq\frac{k_{a1}}{2\psi_{a1}}+\frac{k_{a2}^2}{2\psi_{a2}}+\frac{k_{c1}^2}{2\psi_{c1}}+\frac{k_{c2}^2}{2\psi_{c2}}+k$. The upperbound can be made smaller by increasing the number of neurons N and L or by adjusting the adaptation gains and CL sample points. Based on (36), Z is UUB by invoking a finite time version of Theorem 4.18 in [22]. ■

Remark: If $\left\|Z\left(t_0\right)\right\|\geq\mu$, then $\dot{V}_L\left(Z\left(t_0\right)\right)<0$. There exists an $\epsilon_0\in\mathbb{R}^+$ such that $V_L\left(Z\left(t_0+\epsilon_0\right)\right)<V_L\left(Z\left(t_0\right)\right)$. Then, $\alpha_1\left(\left\|Z\left(t_0+\epsilon_0\right)\right\|\right)\leq V_L\left(Z\left(t_0+\epsilon_0\right)\right)<\alpha_2\left(\left\|Z\left(t_0\right)\right\|\right)$. Rearranging terms, $\left\|Z\left(t_0+\epsilon_0\right)\right\|<\alpha_1^{-1}\left(\alpha_2\left(\left\|Z(t_0)\right\|\right)\right)$. Hence, $Z\left(t_0+\epsilon_0\right)\in\mathcal{L}_\infty$. It can be shown by induction that $Z\left(t\right)\in\mathcal{L}_\infty$ and $\left\|Z(t)\right\|<\alpha_1^{-1}\left(\alpha_2\left(\left\|Z(t_0)\right\|\right)\right)\ \forall t\in\mathbb{R}^+$ when $\left\|Z\left(t_0\right)\right\|\geq\mu$. If instead, $\left\|Z\left(t_0\right)\right\|<\mu$, then $\left\|Z\left(t\right)\right\|<\alpha_1^{-1}\left(\alpha_2\left(\sqrt{\frac{\mu}{\alpha_5}}\right)\right)$. Therefore, $\left\|Z(t)\right\|\in\mathcal{L}_\infty$, $\forall t\in\mathbb{R}^+$ when $\left\|Z\left(t_0\right)\right\|<\mu$. Since $Z\left(t\right)\in\mathcal{L}_\infty$ $\forall t\in\mathbb{R}^+$, then x lies on the compact set χ, where $\chi\triangleq\left\{x\in\mathbb{R}^n\left|\left\|x\right\|\leq\alpha_1^{-1}\left(\alpha_2\left(\max\left(\left\|Z\left(t_0\right)\right\|,\mu\right)\right)\right)\right.\right\}$. This validates the compactness assumption and also implies that if the gain conditions initially satisfy (30), then the gain conditions are sufficient for all time.

VII. Conclusion

An approximately optimal controller is developed for finite-horizon optimal control problems with hard terminal constraints and LP uncertainty in the drift dynamics. The solution is determined online by learning the optimal value function and optimal control policy with an actor-critic framework guided by reinforcement learning, while also enforcing the optimality condition that results from the hard terminal constraint. Actor and critic update laws were based on the minimization of the BE estimate. Additionally, CL is used in place of the more restrictive PE requirement. A Lyapunov-based stability analysis proves UUB convergence of the enacted control policy to the optimal policy.

One advantage of using an online ADP solution is the ability to handle uncertain drift dynamics. Offline ADP results from [14] and [16] show exact satisfaction of the hard constraint; however, the NN weights are trained based on exact model knowledge. Therefore, if there is uncertainty in the drift dynamics, then the hard constraint will not be satisfied. Furthermore, the stability of the controller even becomes questionable. Although the online ADP result developed in this paper only proves UUB convergence, the advantage of using online ADP is the ability to compensate for uncertain parameters in the drift dynamics while simultaneously learning the optimal solution. In practice, the NN weights for the online ADP controller could be initialized according to offline training based on the best knowledge of the dynamics by leveraging results such as [14] and [16].

References

[1] P. K. Menon, T. Lam, L. S. Crawford, and V. H. L. Cheng, "Real-time computational methods for sdre nonlinear control of missiles," in *Proc. Am. Control Conf.*, May 2002, pp. 232–237.

[2] M. Xin, S. N. Balakrishnan, D. T. Stansbery, and E. J. Ohlmeyer, "Nonlinear missile autopilot design with theta-d technique," *J. Guid. Control Dynam.*, vol. 27, pp. 406–417, 2004.

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

[4] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. L. Lewis, and W. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 89–92, 2013.

[5] M. Abu-Khalaf and F. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[6] S. Ferrari and R. Stengel, "Online adaptive critic flight control," *J. Guid. Control Dynam.*, vol. 27, pp. 777–786, 2004.

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

[8] D. Wang, D. Liu, and Q. Wei, "Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach," *Neurocomputing*, vol. 78, no. 1, pp. 14 – 22, 2012.

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

[10] T. Cheng, F. Lewis, and M. Abu-Khalaf, "A neural network solution for fixed-final time optimal control of nonlinear systems," *Automatica*, vol. 43, no. 3, pp. 482–490, 2007.

[11] F. Wang, N. Jin, D. Liu, and Q. Wei, "Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with epsilon-error bound," *IEEE Trans. Neural Netw.*, vol. 22, pp. 24–36, 2011.

[12] A. E. Bryson and Y. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. Hemisphere Publishing Corporation, 1975.

[13] S. R. Vadali and R. Sharma, "Optimal finite-time feedback controllers for nonlinear systems with terminal constraints," *J. Guid. Control Dynam.*, vol. 29, pp. 921–928, 2006.

[14] A. Heydari and S. N. Balakrishnan, "Fixed-final-time optimal control of nonlinear systems with terminal constraints," *Neural Netw.*, vol. 48, pp. 61–71, 2013.

[15] D. Han and S. Balakrishnan, "State-constrained agile missile control with adaptive-critic-based neural networks," *IEEE Trans. Control Syst. Technol.*, vol. 10, no. 4, pp. 481–489, 2002.

[16] A. Heydari and S. N. Balakrishnan, "Adaptive critic-based solution to an orbital rendezvous problem," *J. Guid. Control Dynam.*, vol. 37, pp. 344–350, 2014.

[17] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Concurrent learning-based approximate optimal regulation," in *Proc. IEEE Conf. Decis. Control*, Florence, IT, Dec. 2013, pp. 6256–6261.

[18] F. L. Lewis and V. L. Syrmos, *Optimal Control*, 2nd ed. Wiley, 1995.

[19] G. V. Chowdhary and E. N. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *J. Guid. Control Dynam.*, vol. 34, no. 2, pp. 592–607, March 2011.

[20] G. Chowdhary and E. Johnson, "A singular value maximizing data recording algorithm for concurrent learning," in *Proc. American Control Conf.*, 2011, pp. 3547–3552.

[21] W. E. Dixon, A. Behal, D. M. Dawson, and S. Nagarkatti, *Nonlinear Control of Engineering Systems: A Lyapunov-Based Approach*. Birkhauser: Boston, 2003.

[22] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.