CHAPTER 12

AN ACTOR-CRITIC-IDENTIFIER ARCHITECTURE FOR ADAPTIVE APPROXIMATE OPTIMAL CONTROL

S. Bhasin[1], R. Kamalapurkar[2], M. Johnson[2], K. G. Vamvoudakis[3], F. L. Lewis[3] and W. E. Dixon[2]

[1]Department of Electrical Engineering, Indian Institute of Technology, Delhi, India

[2]Department of Mechanical and Aerospace Engineering, University of Florida, USA

[3]Automation and Robotics Research Institute, The University of Texas at Arlington, USA

12.1 INTRODUCTION

Reinforcement Learning (RL) provides a way for learning agents to optimally interact with uncertain complex environments, and hence, can address problems from a variety of domains, including artificial intelligence, controls, economics, operations research, etc. Actor-critic (AC) architectures have been proposed as models of RL [43, 51]. Since AC methods are amenable to online implementation, they have become an important subject of research, particularly in the controls community [13, 16, 19, 28, 30, 32, 35, 44, 46, 49]. Typically, the AC architecture consists of two neural networks (NNs) – an actor NN and a critic NN. The critic NN approximates

the evaluation function, mapping states to an estimated measure of the value function, while the action NN approximates an optimal control law and generates actions or control signals. Following the works of Sutton [41], Barto [4], Watkins [47], and Werbos [48], current research focuses on the relationship between RL and dynamic programming (DP) [6] methods for solving optimal control problems. Due to the *curse of dimensionality* associated with using DP [6], Werbos [49] introduced an alternative Approximate Dynamic Programming (ADP) approach which gives an approximate solution to the DP problem, or the *Hamiltonian-Jacobi-Bellman* (HJB) equation for optimal control. A detailed review of ADP designs can be found in [35]. Various modifications to ADP algorithms have since been proposed [16, 38, 39].

Convergence of ADP algorithms for RL-based control is studied in [13, 16, 26, 28, 30]. A policy iteration (PI) algorithm is proposed in [9] using Q-functions for the discrete-time LQR problem and convergence to the state feedback optimal solution is proven. In [2], model-free Q-learning is proposed for linear discrete-time systems with guaranteed convergence to the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ state feedback control solution. Most of the previous work on ADP has focused on either finite state Markovian systems or discrete-time systems [7, 50]. The inherently iterative nature of the ADP algorithm has impeded the development of closed-loop controllers for continuous-time uncertain nonlinear systems. Extensions of ADP-based controllers to continuous-time systems entails challenges in proving stability, convergence, and ensuring the algorithm is online and model-free. Early solutions to the problem consisted of using a discrete-time formulation of time and state, and then applying an RL algorithm on the discretized system. Discretizing the state space for high dimensional systems requires a large memory space and a computationally prohibitive learning process. Convergence of PI for continuous-time LQR was first proved in [24]. Baird [3] proposed *Advantage Updating*, an extension of the Q-learning algorithm which could be implemented in continuous-time and provided faster convergence. Doya [15] used an HJB framework to derive algorithms for value function approximation and policy improvement, based on a continuous-time version of the temporal difference (TD) error. Murray et al. [30] also used the HJB framework to develop a *stepwise stable* iterative ADP algorithm for continuous-time input-affine systems with an in-

put quadratic performance measure. In Beard et al. [5], Galerkin's spectral method is used to approximate the solution to the generalized HJB (GHJB), using which a stabilizing feedback controller was computed offline. Similar to [5], Abu-Khalaf and Lewis [1] proposed a least-squares successive approximation solution to the GHJB, where an NN is trained offline to learn the GHJB solution. Another continuous-time formulation of adaptive critic is proposed in Hanselman [19].

All of the aforementioned approaches for continuous-time nonlinear systems require complete knowledge of system dynamics. The fact that continuous-time ADP requires knowledge of the system dynamics has hampered the development of continuous-time extensions to ADP-based controllers for nonlinear systems. Recent result by [46] has made new inroads by addressing the problem for partially unknown nonlinear systems. A PI-based hybrid continuous-time/discrete-time sampled data controller is designed in [45, 46], where the feedback control operation of the actor occurs at faster time scale than the learning process of the critic. Vamvoudakis and Lewis [44] extended the idea by designing a model-based online algorithm called *synchronous PI* which involved synchronous continuous-time adaptation of both actor and critic NNs. A contribution of this work is the use of a novel actor-critic-identifier architecture, which obviates the need to know the system drift dynamics, and where the learning of the actor, critic and identifier is continuous and simultaneous. Moreover, the actor-critic-identifier method utilizes an identification-based online learning scheme, and hence is the first ever indirect adaptive control approach to RL.

In the developed method, the actor and critic NNs use gradient and least squares-based update laws, respectively, to minimize the Bellman error, which is the difference between the exact and the approximate HJB equation. The identifier is a combination of a Hopfield-type [20] dynamic NN (DNN), in parallel configuration with the system [34], and a novel RISE (Robust Integral of Sign of the Error) component [33]. The identifier asymptotically estimates the state derivative, allowing the actor-critic-identifier architecture to be implemented without knowledge of system drift dynamics; however, knowledge of the input gain matrix is required to implement the control policy. Convergence of the actor-critic-identifier-based algorithm and stability of the closed-loop system are analyzed using Lyapunov-based adaptive control

methods, and a *persistence of excitation* (PE) condition is used to guarantee exponential convergence to a bounded region in the neighborhood of the optimal control and uniformly ultimately bounded (UUB) stability of the closed-loop system. The PE condition is equivalent to the exploration paradigm in RL [42] and ensures adequate sampling of the system's dynamics, required for convergence to the optimal policy.

12.2 ACTOR-CRITIC-IDENTIFIER ARCHITECTURE FOR HJB APPROXIMATION

Consider a continuous-time nonlinear system

$$\dot{x} = F(x, u),$$

where $x(t) \in \mathcal{X} \subseteq \mathbb{R}^n$, $u(t) \in \mathcal{U} \subseteq \mathbb{R}^m$ is the control input, $F : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^n$ is Lipschitz continuous on $\mathcal{X} \times \mathcal{U}$ containing the origin, such that the solution $x(t)$ of the system is unique for any finite initial condition x_0 and control $u \in \mathcal{U}$. The optimal value function can be defined as

$$V^*(x(t)) = \min_{\substack{u(\tau)\in\Psi(\mathcal{X}) \\ t\leq\tau<\infty}} \int_t^\infty r(x(s), u(x(s)))\, ds, \tag{12.1}$$

where $\Psi(\mathcal{X})$ is a set of admissible policies, and $r(x, u) \in \mathbb{R}$ is the immediate or local cost, defined as

$$r(x, u) = Q(x) + u^T R u, \tag{12.2}$$

where $Q(x) \in \mathbb{R}$ is continuously differentiable and positive definite, and $R \in \mathbb{R}^{m\times m}$ is a positive-definite symmetric matrix. For the local cost in (12.2), which is convex in the control, and control-affine dynamics of the form

$$\dot{x} = f(x) + g(x)u, \tag{12.3}$$

where $f(x) \in \mathbb{R}^n$ and $g(x) \in \mathbb{R}^{n\times m}$, the closed-form expression for optimal control is derived as [23]

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \frac{\partial V^*(x)}{\partial x}^T, \tag{12.4}$$

where it is assumed that the value function $V^*(x)$ is continuously differentiable and satisfies $V^*(0) = 0$.

The Hamiltonian of the system in 12.3 is given by

$$H(x, u, V_x^*) \triangleq V_x^* F_u + r_u,$$

where $V_x^* \triangleq \frac{\partial V^*}{\partial x} \in \mathbb{R}^{1\times n}$ denotes the gradient of the optimal value function $V^*(x)$, $F_u(x,u) \triangleq f(x) + g(x)u \in \mathbb{R}^n$ denotes the system dynamics with control $u(x)$, and $r_u \triangleq r(x,u)$ denotes the local cost with control $u(x)$. The optimal value function $V^*(x)$ in 12.1 and the associated optimal policy $u^*(x)$ in 12.4 satisfy the HJB equation

$$H^*(x, u^*, V_x^*) = V_x^* F_{u^*} + r_{u^*} = 0. \tag{12.5}$$

Replacing $u^*(x)$, $V_x^*(x)$, and $F_{u^*}(x,u^*)$ in 12.5 by their approximations, $\hat{u}(x)$ (actor), $\hat{V}(x)$(critic), and $\hat{F}_{\hat{u}}(x,\hat{x},\hat{u})$ (identifier), respectively, the approximate HJB equation is given by

$$\hat{H}^*(x, \hat{x}, \hat{u}, \hat{V}_x) = \hat{V}_x \hat{F}_{\hat{u}} + r_{\hat{u}}, \tag{12.6}$$

where $\hat{x}(t)$ denotes the state of the identifier. Using 12.5 and 12.6, the error between the actual and the approximate HJB equation is given by the Bellman residual error $\delta_{hjb}(x, \hat{x}, \hat{u}, \hat{V}_x)$, defined as

$$\delta_{hjb} \triangleq \hat{H}^*(x, \hat{x}, \hat{u}, \hat{V}_x) - H^*(x, u^*, V_x^*). \tag{12.7}$$

Since $H^*(x, u^*, V_x^*) \equiv 0$, the Bellman error can be written in a measurable form as

$$\delta_{hjb} = \hat{H}^*(x, \hat{x}, \hat{u}, \hat{V}_x) = \hat{V}_x \hat{F}_{\hat{u}} + r(x, \hat{u}). \tag{12.8}$$

The actor and critic learn based on the Bellman error $\delta_{hjb}(\cdot)$, whereas the identifier estimates the system dynamics online using the identification error $\tilde{x}(t) \triangleq x(t) - \hat{x}(t)$, and hence is decoupled from the actor-critic design. The block diagram of the ACI architecture is shown in Fig. 12.1.

The following assumptions are made about the control-affine system in (12.3). **Assumption 1:** The functions $f(x)$ and $g(x)$ are second-order differentiable. **Assumption 2:** The input gain matrix $g(x)$ is known and bounded i.e. $0 < \|g(x)\| \leq \bar{g}$, where $\bar{g}$ is a known positive constant.

Assuming the optimal control, the optimal value function and the system dynamics are continuous and defined on compact sets, NNs can be used to approximate them

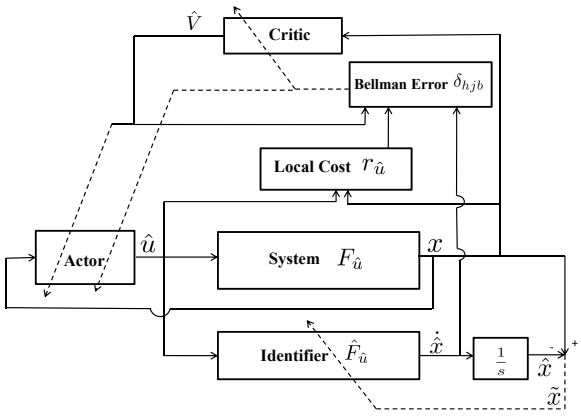

Figure 12.1: Actor-critic-identifier architecture to approximate the HJB.

[12, 21]. Some standard NN assumptions which will be used throughout the chapter are: **Assumption 3:** Given a continuous function $\Upsilon : \mathbb{S} \to \mathbb{R}^n$, where $\mathbb{S}$ is a compact simply connected set, there exists ideal weights $W,\ V$ such that the function can be represented by a NN as

$$\Upsilon(x) = W^T \sigma(V^T x) + \varepsilon(x),$$

where $\sigma(\cdot)$ is the nonlinear activation function, and $\varepsilon(x)$ is the function reconstruction error. **Assumption 4:** The ideal NN weights are bounded by known positive constants i.e. $\|W\| \leq \bar{W}$, $\|V\| \leq \bar{V}$ [27]. **Assumption 5:** The NN activation function $\sigma(\cdot)$ and its derivative with respect to its arguments, $\sigma'(\cdot)$, are bounded. **Assumption 6:** Using the NN universal approximation property [12, 21], the function reconstruction errors and its derivative with respect to its arguments are bounded [27] as $\|\varepsilon(\cdot)\| \leq \bar{\varepsilon}$, $\|\varepsilon'(\cdot)\| \leq \bar{\varepsilon'}$.

12.3 ACTOR-CRITIC DESIGN

Using Assumption 3 and (12.4), the optimal value function and the optimal control can be represented by NNs as

$$V^*(x) = W^T\phi(x) + \varepsilon_v(x),$$
$$u^*(x) = -\frac{1}{2}R^{-1}g^T(x)(\phi'(x)^T W + \varepsilon_v'(x)^T), \tag{12.9}$$

where $W \in \mathbb{R}^N$ are unknown ideal NN weights, N is the number of neurons, $\phi(x) = [\phi_1(x)\ \ \phi_2(x). . . \phi_N(x)]^T \in \mathbb{R}^N$ is a smooth NN activation function such that $\phi_i(0) = 0$ and $\phi_i'(0) = 0 \forall i = 1...N$, and $\varepsilon_v(\cdot) \in \mathbb{R}$ is the function reconstruction error. **Assumption 7:** The NN activation functions $\{\phi_i(x) : i = 1...N\}$ are selected so that as $N \to \infty$, $\phi(x)$ provides a complete independent basis for $V^*(x)$.

Using Assumption 7 and the Weierstrass higher-order approximation Theorem, both $V^*(x)$ and $\frac{\partial V^*(x)}{\partial x}$ can be uniformly approximated by NNs in (12.9), i.e. as $N \to \infty$, the approximation errors $\varepsilon_v(x)$, $\varepsilon_v'(x) \to 0$ [1]. The critic $\hat{V}(x)$ and the actor $\hat{u}(x)$ approximate the optimal value function and the optimal control in (12.9), and are given by

$$\hat{V}(x) = \hat{W}_c^T\phi(x); \qquad \hat{u}(x) = -\frac{1}{2}R^{-1}g^T(x)\phi'^T(x)\hat{W}_a, \tag{12.10}$$

where $\hat{W}_c(t) \in \mathbb{R}^N$ and $\hat{W}_a(t) \in \mathbb{R}^N$ are estimates of the ideal weights of the critic and actor NNs, respectively. The weight estimation errors for the critic and actor NNs are defined as $\tilde{W}_c(t) \triangleq W - \hat{W}_c(t) \in \mathbb{R}^N$ and $\tilde{W}_a(t) \triangleq W - \hat{W}_a(t) \in \mathbb{R}^N$, respectively.

Remark 12.1 *Since the optimal control is determined using the gradient of the optimal value function in (12.9), the critic NN in (12.10) may be used to determine the actor without using another NN for the actor. However, for ease in deriving weight update laws and subsequent stability analysis, separate NNs are used for the actor and the critic [44].*

The actor and critic NN weights are both updated based on the minimization of the Bellman error $\delta_{hjb}(\cdot)$ in (12.8), which can be rewritten by substituting $\hat{V}(x)$ from

(12.10) as

$$\delta_{hjb} = \hat{W}_c^T \omega + r(x, \hat{u}), \tag{12.11}$$

where $\omega(x, \hat{x}, \hat{u}) \triangleq \phi'(x)\hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \in \mathbb{R}^N$ is the critic NN regressor vector. Let $E_c(\delta_{hjb}) = \int_0^t \delta_{hjb}^2(\tau)\, d\tau \in \mathbb{R}^+$ denote the integral squared Bellman error for the critic. The normalized recursive least squares (LS) update law for the critic is given by [36]

$$\dot{\hat{W}}_c = -\eta_c \Gamma \frac{\omega}{1 + \nu \omega^T \Gamma \omega} \delta_{hjb}, \tag{12.12}$$

where $\nu, \eta_c \in \mathbb{R}$ are constant positive gains, and $\Gamma(t) \in \mathbb{R}^{N \times N}$ is a symmetric estimation gain matrix generated as

$$\dot{\Gamma} = -\eta_c \Gamma \frac{\omega \omega^T}{1 + \nu \omega^T \Gamma \omega} \Gamma; \quad \Gamma(t_r^+) = \Gamma(0) = \varphi_0 I, \tag{12.13}$$

where t_r^+ is the resetting time at which $\lambda_{min}\{\Gamma(t)\} \leq \varphi_1$, $\varphi_0 > \varphi_1 > 0$. The covariance resetting ensures that $\Gamma(t)$ is positive-definite for all time and prevents its value from becoming arbitrarily small in some directions, thus avoiding slow adaptation in some directions (also called the covariance wind-up problem) [36]. From (12.13), it is clear that $\dot{\Gamma} \leq 0$, which means that the covariance matrix $\Gamma(t)$ can be bounded as

$$\varphi_1 I \leq \Gamma(t) \leq \varphi_0 I. \tag{12.14}$$

Unlike the critic weights, the actor weights appear nonlinearly in $\delta_{hjb}(\cdot)$, making it problematic to develop a LS update law. Hence, a gradient update law is developed for the actor which minimizes the squared Bellman error $E_a(t) \triangleq \delta_{hjb}^2$, whose gradient is given by

$$\begin{aligned} \dot{\hat{W}}_a = & proj \left[-\frac{\eta_{a1}}{\sqrt{1 + \omega^T \omega}} \left(\hat{W}_c^T \phi' \frac{\partial \hat{F}_{\hat{u}}}{\partial \hat{u}} \frac{\partial \hat{u}}{\partial \hat{W}_a} + \hat{W}_a^T \phi' G \phi'^T \right)^T \delta_{hjb} \right. \\ & \left. - \eta_{a2}(\hat{W}_a - \hat{W}_c) \right], \end{aligned} \tag{12.15}$$

where $proj\{\cdot\}$ is a projection operator used to bound the weight estimates [14], [25], $G(x) \triangleq g(x)R^{-1}g(x)^T \in \mathbb{R}^{n \times n}$, $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are positive adaptation gains, $\frac{1}{\sqrt{1+\omega^T\omega}}$ is the normalization term, and the last term in (12.15) is added for stability (based on the subsequent stability analysis).

12.4 IDENTIFIER DESIGN

The following assumption is made for the identifier design: **Assumption 8:** The control input is bounded i.e. $u(t) \in \mathcal{L}_\infty$. Using Assumptions 2, 5 and the projection algorithm in (12.15), this assumption holds for the control design $u(t) = \hat{u}(x)$ in (12.10). Using Assumption 3, the dynamic system in (12.3), with control $\hat{u}(x)$, can be represented using a multi-layer NN as

$$\dot{x} = F_{\hat{u}}(x, \hat{u}) = W_f^T \sigma(V_f^T x) + \varepsilon_f(x) + g(x)\hat{u}, \tag{12.16}$$

where $W_f \in \mathbb{R}^{L_f+1 \times n}$, $V_f \in \mathbb{R}^{n \times L_f}$ are the unknown ideal NN weights, $\sigma_f \triangleq \sigma(V_f^T x) \in \mathbb{R}^{L_f+1}$ is the NN activation function, and $\varepsilon_f(x) \in \mathbb{R}^n$ is the function reconstruction error. The following multi-layer dynamic neural network (MLDNN) identifier is used to approximate the system in (12.16)

$$\dot{\hat{x}} = \hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) = \hat{W}_f^T \hat{\sigma}_f + g(x)\hat{u} + \mu, \tag{12.17}$$

where $\hat{x}(t) \in \mathbb{R}^n$ is the DNN state, $\hat{\sigma}_f \triangleq \sigma(\hat{V}_f^T \hat{x}) \in \mathbb{R}^{L_f+1}$, $\hat{W}_f(t) \in \mathbb{R}^{L_f+1 \times n}$ and $\hat{V}_f(t) \in \mathbb{R}^{n \times L_f}$ are weight estimates, and $\mu(t) \in \mathbb{R}^n$ denotes the RISE feedback term defined as [33, 52]

$$\mu \triangleq k\tilde{x}(t) - k\tilde{x}(0) + v, \tag{12.18}$$

where $\tilde{x}(t) \triangleq x(t) - \hat{x}(t) \in \mathbb{R}^n$ is the identification error, and $v(t) \in \mathbb{R}^n$ is the generalized solution (in Filippov's sense [18]) to

$$\dot{v} = (k\alpha + \gamma)\tilde{x} + \beta_1 sgn(\tilde{x}); \quad v(0) = 0,$$

where $k,\ \alpha,\ \gamma,\ \beta_1 \in \mathbb{R}$ are positive constant control gains, and $sgn(\cdot)$ denotes a vector signum function. The identification error dynamics can be written as

$$\dot{\tilde{x}} = \tilde{F}_u(x, \hat{x}, u) = W_f^T \sigma_f - \hat{W}_f^T \hat{\sigma}_f + \varepsilon_f(x) - \mu, \tag{12.19}$$

where $\tilde{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \triangleq F_{\hat{u}}(x, \hat{u}) - \hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \in \mathbb{R}^n$. A filtered identification error is defined as

$$r \triangleq \dot{\tilde{x}} + \alpha\tilde{x}. \tag{12.20}$$

Taking the time derivative of (12.20) and using (12.19) yields

$$\begin{aligned}\dot{r} &= W_f^T\sigma_f' V_f^T\dot{x} - \dot{\hat{W}}_f^T\hat{\sigma}_f - \hat{W}_f^T\hat{\sigma}_f'\dot{\hat{V}}_f^T\hat{x} - \hat{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}} + \dot{\varepsilon}_f(x) - kr \\ &\quad -\gamma\tilde{x} - \beta_1 sgn(\tilde{x}) + \alpha\dot{\tilde{x}}. \end{aligned} \tag{12.21}$$

Based on (12.21) and the subsequent stability analysis, the weight update laws for the DNN are designed as

$$\dot{\hat{W}}_f = proj(\Gamma_{wf}\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}}\tilde{x}^T), \quad \dot{\hat{V}}_f = proj(\Gamma_{vf}\dot{\hat{x}}\tilde{x}^T\hat{W}_f^T\hat{\sigma}_f'), \tag{12.22}$$

where $proj(\cdot)$ is a smooth projection operator [14], [25], and $\Gamma_{wf} \in \mathbb{R}^{L_f+1\times L_f+1}$, $\Gamma_{vf} \in \mathbb{R}^{n\times n}$ are positive constant adaptation gain matrices. The expression in (12.21) can be rewritten as

$$\dot{r} = \tilde{N} + N_{B1} + \hat{N}_{B2} - kr - \gamma\tilde{x} - \beta_1 sgn(\tilde{x}), \tag{12.23}$$

where the auxiliary signals, $\tilde{N}(x,\tilde{x},r,\hat{W}_f,\hat{V}_f,t)$, $N_{B1}(x,\hat{x},\hat{W}_f,\hat{V}_f,t)$, and $\hat{N}_{B2}(\hat{x},\dot{\hat{x}},\hat{W}_f,\hat{V}_f,t) \in \mathbb{R}^n$ are defined as

$$\tilde{N} \triangleq \alpha\dot{\tilde{x}} - \dot{\hat{W}}_f^T\hat{\sigma}_f - \hat{W}_f^T\hat{\sigma}_f'\dot{\hat{V}}_f^T\hat{x} + \frac{1}{2}W_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\tilde{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'V_f^T\dot{\tilde{x}}, \tag{12.24}$$

$$N_{B1} \triangleq W_f^T\sigma_f'V_f^T\dot{x} - \frac{1}{2}W_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{x} - \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'V_f^T\dot{x} + \dot{\varepsilon}_f(x), \tag{12.25}$$

$$\hat{N}_{B2} \triangleq \frac{1}{2}\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'\tilde{V}_f^T\dot{\hat{x}}, \tag{12.26}$$

where $\tilde{W}_f \triangleq W_f - \hat{W}_f(t) \in \mathbb{R}^{L_f+1\times n}$ and $\tilde{V}_f \triangleq V_f - \hat{V}_f(t) \in \mathbb{R}^{n\times L_f}$. To facilitate the subsequent stability analysis, an auxiliary term $N_{B2}(\hat{x},\dot{x},\hat{W}_f,\hat{V}_f,t) \in \mathbb{R}^n$ is defined by replacing $\dot{\hat{x}}(t)$ in $\hat{N}_{B2}(\cdot)$ by $\dot{x}(t)$, and $\tilde{N}_{B2}(\hat{x},\dot{\tilde{x}},\hat{W}_f,\hat{V}_f,t) \triangleq \hat{N}_{B2}(\cdot) - N_{B2}(\cdot)$. The terms $N_{B1}(\cdot)$ and $N_{B2}(\cdot)$ are grouped as $N_B \triangleq N_{B1} + N_{B2}$. Using Assumptions 2, 4-6, and (12.20), (12.22), (12.25) and (12.26), the following bounds can be obtained

$$\left\|\tilde{N}\right\| \leq \rho_1(\|z\|)\|z\|, \tag{12.27}$$

$$\|N_{B1}\| \leq \zeta_1, \quad \|N_{B2}\| \leq \zeta_2, \quad \left\|\dot{N}_B\right\| \leq \zeta_3 + \zeta_4\rho_2(\|z\|)\|z\|, \tag{12.28}$$

$$\left\|\dot{\tilde{x}}^T\tilde{N}_{B2}\right\| \leq \zeta_5\|\tilde{x}\|^2 + \zeta_6\|r\|^2, \tag{12.29}$$

where $z \triangleq \left[\tilde{x}^T \; r^T\right]^T \in \mathbb{R}^{2n}$, $\rho_1(\cdot), \rho_2(\cdot) \in \mathbb{R}$ are positive, globally invertible, non-decreasing functions, and $\zeta_i \in \mathbb{R}$, $i = 1, ..., 6$ are computable positive constants. To facilitate the subsequent stability analysis, let $\mathcal{D} \subset \mathbb{R}^{2n+2}$ be a domain containing $y(t) = 0$, where $y(t) \in \mathbb{R}^{2n+2}$ is defined as

$$y \triangleq \left[\tilde{x}^T \;\; r^T \;\; \sqrt{P} \;\; \sqrt{Q}\right]^T, \tag{12.30}$$

where the auxiliary function $P(z,t) \in \mathbb{R}$ is the generalized solution to the differential equation

$$\dot{P} = -L, \qquad P(0) = \beta_1 \sum_{i=1}^{n} |\tilde{x}_i(0)| - \tilde{x}^T(0)\, N_B(0), \tag{12.31}$$

where the auxiliary function $L(z,t) \in \mathbb{R}$ is defined as

$$L \triangleq r^T(N_{B1} - \beta_1 sgn(\tilde{x})) + \dot{\tilde{x}}^T N_{B2} - \beta_2 \rho_2(\|z\|)\, \|z\|\, \|\tilde{x}\|, \tag{12.32}$$

where $\beta_1, \beta_2 \in \mathbb{R}$ are chosen according to the following sufficient conditions to ensure $P(t) \geq 0$ [33]

$$\beta_1 > \max(\zeta_1 + \zeta_2,\, \zeta_1 + \frac{\zeta_3}{\alpha}), \qquad \beta_2 > \zeta_4. \tag{12.33}$$

The auxiliary function $Q(\tilde{W}_f, \tilde{V}_f) \in \mathbb{R}$ in (12.30) is defined as

$$Q \triangleq \frac{1}{4}\alpha \left[tr(\tilde{W}_f^T \Gamma_{wf}^{-1} \tilde{W}_f) + tr(\tilde{V}_f^T \Gamma_{vf}^{-1} \tilde{V}_f)\right],$$

where $tr(\cdot)$ denotes the trace of a matrix.

Theorem 12.1 *For the system in (12.3), the identifier developed in (12.17) along with the weight update laws in (12.22) ensures asymptotic identification of the state and its derivative, in the sense that*

$$\lim_{t\to\infty} \|\tilde{x}(t)\| = 0 \qquad \textit{and} \qquad \lim_{t\to\infty} \left\|\dot{\tilde{x}}(t)\right\| = 0,$$

provided the control gains k and γ are chosen sufficiently large based on the initial conditions of the states[1]*, and satisfy the following sufficient conditions*

$$\gamma > \frac{\zeta_5}{\alpha}, \qquad k > \zeta_6, \tag{12.34}$$

[1]See subsequent semi-global stability analysis.

where ζ_5 and ζ_6 are introduced in (12.29), and $\beta_1,\ \beta_2$ introduced in (12.32), are chosen according to the sufficient conditions in (12.33).

Proof: Let $V_I(y) : \mathcal{D} \to \mathbb{R}$ be a Lipschitz continuous regular positive definite function defined as

$$V_I \triangleq \frac{1}{2} r^T r + \frac{1}{2}\gamma \tilde{x}^T \tilde{x} + P + Q, \tag{12.35}$$

which satisfies the following inequalities:

$$U_1(y) \leq V_I(y) \leq U_2(y), \tag{12.36}$$

where $U_1(y), U_2(y) \in \mathbb{R}$ are continuous positive definite functions defined as

$$U_1 \triangleq \frac{1}{2} min(1,\gamma) \left\|y\right\|^2 \qquad U_2 \triangleq max(1,\gamma) \left\|y\right\|^2.$$

From (12.19), (12.22), (12.23), and (12.31), the differential equations of the closed-loop system are continuous except in the set $\{y | \tilde{x} = 0\}$. Using Filippov's differential inclusion [17, 40], the existence of solutions can be established for $\dot{y} = f(y)$, where $f(y) \in \mathbb{R}^{2n+2}$ denotes the right-hand side of the the closed-loop error signals. Under Filippov's framework, a generalized Lyapunov stability theory can be used (see [11, 31, 37] for further details) to establish strong stability of the closed-loop system. The generalized time derivative of (12.35) exists almost everywhere (a.e.), and $\dot{V}_I(y) \overset{a.e.}{\in} \dot{\tilde{V}}_I(y)$ where

$$\dot{\tilde{V}}_I = \bigcap_{\xi \in \partial V_I(y)} \xi^T K \left[\dot{r}^T \ \dot{\tilde{x}}^T \ \frac{1}{2} P^{-\frac{1}{2}} \dot{P} \ \frac{1}{2} Q^{-\frac{1}{2}} \dot{Q} \right]^T,$$

where ∂V_I is the generalized gradient of V_I [11], and $K[\cdot]$ is defined as [31, 37]

$$K[f](y) \triangleq \bigcap_{\delta > 0} \bigcap_{\mu M = 0} \overline{co} f(B(y,\delta) - M),$$

where $\bigcap_{\mu M = 0}$ denotes the intersection of all sets M of Lebesgue measure zero, $\overline{co}$ denotes convex closure, and $B(y,\delta) = \{x \in \mathbb{R}^{2n+2} |\ \|y - x\| < \delta\}$. Since $V_I(y)$ is a Lipschitz continuous regular function,

$$\begin{aligned} \dot{\tilde{V}}_I &= \nabla V_I^T K \left[\dot{r}^T \ \dot{\tilde{x}}^T \ \frac{1}{2} P^{-\frac{1}{2}} \dot{P} \ \frac{1}{2} Q^{-\frac{1}{2}} \dot{Q} \right]^T \\ &= \left[r^T \ \gamma \tilde{x}^T \ 2P^{\frac{1}{2}} \ 2Q^{\frac{1}{2}} \right] K \left[\dot{r}^T \ \dot{\tilde{x}}^T \ \frac{1}{2} P^{-\frac{1}{2}} \dot{P} \ \frac{1}{2} Q^{-\frac{1}{2}} \dot{Q} \right]^T. \end{aligned}$$

Using the calculus for $K[\cdot]$ from [31], and substituting the dynamics from (12.23) and (12.31), yields

$$\begin{aligned}\dot{\tilde{V}}_I \subset\ & r^T(\tilde{N} + N_{B1} + \hat{N}_{B2} - kr - \beta_1 K[sgn(\tilde{x})] - \gamma\tilde{x}) + \gamma\tilde{x}^T(r - \alpha\tilde{x}) \\ & -r^T(N_{B1} - \beta_1 K[sgn(\tilde{x})]) - \dot{\tilde{x}}^T N_{B2} + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\| \\ & -\frac{1}{2}\alpha\left[tr(\tilde{W}_f^T\Gamma_{wf}^{-1}\dot{\hat{W}}_f) + tr(\tilde{V}_f^T\Gamma_{vf}^{-1}\dot{\hat{V}}_f)\right] \\ =\ & -\alpha\gamma\tilde{x}^T\tilde{x} - kr^Tr + r^T\tilde{N} + \frac{1}{2}\alpha\tilde{x}^T\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}} + \dot{\tilde{x}}^T(\hat{N}_{B2} - N_{B2}) \\ & +\frac{1}{2}\alpha\tilde{x}^T\hat{W}_f^T\hat{\sigma}_f'\tilde{V}_f^T\dot{\hat{x}} + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\| \\ & -\frac{1}{2}\alpha tr(\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}}\tilde{x}^T) - \frac{1}{2}\alpha tr(\tilde{V}_f^T\dot{\hat{x}}\tilde{x}^T\hat{W}_f^T\hat{\sigma}_f'),\end{aligned} \tag{12.37}$$

where (12.22) and the fact that $(r^T - r^T)_i SGN(\tilde{x}_i) = 0$ is used (the subscript i denotes the i^{th} element), where $K[sgn(\tilde{x})] = SGN(\tilde{x})$ [31], such that $SGN(\tilde{x}_i) = 1$ if $\tilde{x}_i > 0$, $[-1, 1]$ if $\tilde{x}_i = 0$, and -1 if $\tilde{x}_i < 0$. Canceling common terms, substituting for $k \triangleq k_1 + k_2$ and $\gamma \triangleq \gamma_1 + \gamma_2$, using (12.27), (12.29), and completing the squares, the expression in (12.37) can be upper bounded as

$$\dot{\tilde{V}}_I \leq -(\alpha\gamma_1 - \zeta_5)\|\tilde{x}\|^2 - (k_1 - \zeta_6)\|r\|^2 + \frac{\rho_1(\|z\|)^2}{4k_2}\|z\|^2 + \frac{\beta_2^2\rho_2(\|z\|)^2}{4\alpha\gamma_2}\|z\|^2. \tag{12.38}$$

Provided the sufficient conditions in (12.34) are satisfied, the expression in (12.38) can be rewritten as

$$\dot{\tilde{V}}_I \leq -\lambda\|z\|^2 + \frac{\rho(\|z\|)^2}{4\eta}\|z\|^2 \leq -U(y) \qquad \forall y \in \mathcal{D}, \tag{12.39}$$

where $\lambda \triangleq min\{\alpha\gamma_1 - \zeta_5,\ k_1 - \zeta_6\}$, $\rho(\|z\|)^2 \triangleq \rho_1(\|z\|)^2 + \rho_2(\|z\|)^2$, $\eta \triangleq min\{k_2,\ \frac{\alpha\gamma_2}{\beta_2^2}\}$, and $U(y) = c\|z\|^2$, for some positive constant c, is a continuous, positive semi-definite function defined on the domain $\mathcal{D} \triangleq \left\{y(t) \in \mathbb{R}^{2n+2} \mid \|y\| \leq \rho^{-1}\left(2\sqrt{\lambda\eta}\right)\right\}$. The size of the domain $\mathcal{D}$ can be increased by increasing the gains k and γ. The result in (12.39) indicates that $\dot{V}_I(y) \leq -U(y)\ \forall\ \dot{V}_I(y) \in^{a.e.} \dot{\tilde{V}}_I(y)\ \forall y \in \mathcal{D}$. The inequalities in (12.36) and (12.39) can be used to show that $V_I(y) \in \mathcal{L}_\infty$ in $\mathcal{D}$; hence, $\tilde{x}(t), r(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Using (12.20), standard linear analysis can be used to show that $\dot{\tilde{x}}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$, and since $\dot{x}(t) \in \mathcal{L}_\infty$, $\dot{\hat{x}}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Since $\hat{W}_f(t) \in \mathcal{L}_\infty$ from the use

of projection in (12.22), $\hat{\sigma}_f(t) \in \mathcal{L}_\infty$ from Assumption 5, and $\hat{u}(t) \in \mathcal{L}_\infty$ from Assumption 8, $\mu(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$ from (12.17). Using the above bounds and the fact that $\hat{\sigma}'_f(t), \dot{\varepsilon}_f(t) \in \mathcal{L}_\infty$, it can be shown from (12.21) that $\dot{r}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Since $\tilde{x}(t)$,$r(t) \in \mathcal{L}_\infty$, the definition of $U(y)$ can be used to show that it is uniformly continuous in $\mathcal{D}$. Let $\mathcal{S} \subset \mathcal{D}$ denote a set defined as

$$\mathcal{S} \triangleq \left\{ y(t) \subset \mathcal{D} \mid U_2(y(t)) < \frac{1}{2} \left(\rho^{-1} \left(2\sqrt{\lambda \eta} \right) \right)^2 \right\}. \tag{12.40}$$

The region of attraction in (12.40) can be made arbitrarily large to include any initial conditions by increasing the control gain η (i.e. a semi-global type of stability result), and hence

$$c \left\| z \right\|^2 \to 0 \qquad \text{as} \qquad t \to \infty \qquad \forall y(0) \in \mathcal{S},$$

and using the definition of $z(t)$

$$\left\| \tilde{x}(t) \right\|, \left\| \dot{\tilde{x}}(t) \right\|, \left\| r \right\| \to 0 \qquad \text{as} \qquad t \to \infty \qquad \forall y(0) \in \mathcal{S}.$$

■

Using the developed identifier in (12.17), the actor weight update law can now be simplified using (12.15) as

$$\dot{\hat{W}}_a = proj \left\{ -\frac{\eta_{a1}}{\sqrt{1+\omega^T \omega}} \phi' G \phi'^T \left(\hat{W}_a - \hat{W}_c \right) \delta_{hjb} - \eta_{a2} (\hat{W}_a - \hat{W}_c) \right\}. \tag{12.41}$$

12.5 CONVERGENCE AND STABILITY ANALYSIS

The unmeasurable form of the Bellman error can be written using (12.5)-(12.8) and (12.11), as

$$\begin{aligned} \delta_{hjb} &= \hat{W}_c^T \omega - W_c^T \phi' F_{u^*} + \hat{u}^T R \hat{u} - u^{*^T} R u^* - \varepsilon'_v F_{u^*}. \\ &= -\tilde{W}_c^T \omega - W^T \phi' \tilde{F}_{\hat{u}} + \frac{1}{4} \tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a - \frac{1}{4} \varepsilon'_v G \varepsilon'^T_v - \varepsilon'_v F_{u^*}, \end{aligned} \tag{12.42}$$

where (12.9) and (12.10) are used. The dynamics of the critic weight estimation error $\tilde{W}_c(t)$ can now be developed by substituting (12.42) in (12.12), as

$$\begin{aligned} \dot{\tilde{W}}_c &= -\eta_c \Gamma \psi \psi^T \tilde{W}_c + \eta_c \Gamma \frac{\omega}{1+\nu \omega^T \Gamma \omega} \left[-W^T \phi' \tilde{F}_{\hat{u}} + \frac{1}{4} \tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a \right. \\ &\quad \left. -\frac{1}{4} \varepsilon'_v G \varepsilon'^T_v - \varepsilon'_v F_{u^*} \right], \end{aligned} \tag{12.43}$$

where $\psi(t) \triangleq \frac{\omega(t)}{\sqrt{1+\nu\omega(t)^T\Gamma(t)\omega(t)}} \in \mathbb{R}^N$ is the normalized critic regressor vector, bounded as

$$\|\psi\| \leq \frac{1}{\sqrt{\nu\varphi_1}}, \tag{12.44}$$

where φ_1 is introduced in (12.14). The error system in (12.43) can be represented by the following perturbed system

$$\dot{\tilde{W}}_c = \Omega_{nom} + \Delta_{per}, \tag{12.45}$$

where $\Omega_{nom}(\tilde{W}_c, t) \triangleq -\eta_c\Gamma\psi\psi^T\tilde{W}_c \in \mathbb{R}^N$, denotes the nominal system, and $\Delta_{per}(t) \triangleq \eta_c\Gamma\frac{\omega}{1+\nu\omega^T\Gamma\omega}\left[-W^T\phi'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T\phi' G\phi'^T\tilde{W}_a - \frac{1}{4}\varepsilon_v' G\varepsilon_v'^T - \varepsilon_v' F_{u^*}\right] \in \mathbb{R}^N$ denotes the perturbation. Using Theorem 2.5.1 in [36], the nominal system

$$\dot{\tilde{W}}_c = -\eta_c\Gamma\psi\psi^T\tilde{W}_c \tag{12.46}$$

is globally exponentially stable, if the bounded signal $\psi(t)$ is PE, i.e.

$$\mu_2 I \geq \int_{t_0}^{t_0+\delta} \psi(\tau)\psi(\tau)^T d\tau \geq \mu_1 I \qquad \forall t_0 \geq 0,$$

for some positive constants $\mu_1, \mu_2, \delta \in \mathbb{R}$. Since $\Omega_{nom}(\tilde{W}_c, t)$ is continuously differentiable and the Jacobian $\frac{\partial\Omega_{nom}}{\partial\tilde{W}_c} = -\eta_c\Gamma\psi\psi^T$ is bounded for the exponentially stable system in (12.46), the converse Lyapunov Theorem 4.14 in [22] can be used to show that there exists a function $V_c : \mathbb{R}^N \times [0, \infty) \to \mathbb{R}$, which satisfies the following inequalities

$$\begin{aligned} c_1\left\|\tilde{W}_c\right\|^2 \leq V_c(\tilde{W}_c, t) \leq c_2\left\|\tilde{W}_c\right\|^2 \\ \frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial\tilde{W}_c}\Omega_{nom}(\tilde{W}_c, t) \leq -c_3\left\|\tilde{W}_c\right\|^2 \\ \left\|\frac{\partial V_c}{\partial\tilde{W}_c}\right\| \leq c_4\left\|\tilde{W}_c\right\|, \end{aligned} \tag{12.47}$$

for some positive constants $c_1,\ c_2,\ c_3,\ c_4 \in \mathbb{R}$. Using Assumptions 2, 4-6, 8, the projection bounds in (12.15), the fact that $F_{u^*} \in \mathcal{L}_\infty$ (using (12.4), Assumptions 2-6, and (12.9)), and provided the conditions of Theorem 12.1 hold (required to prove

that $\tilde{F}_{\hat{u}} \in \mathcal{L}_\infty$), the following bounds can be developed:

$$\left\|\tilde{W}_a\right\| \leq \kappa_1, \qquad \left\|\phi' G \phi'^T\right\| \leq \kappa_2,$$
$$\left\|\frac{1}{4}\tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a - \frac{1}{4}\varepsilon_v' G \varepsilon_v'^T - W^T \phi' \tilde{F}_{\hat{u}} - \varepsilon_v' F_{u^*}\right\| \leq \kappa_3,$$
$$\left\|\frac{1}{2}W^T \phi' G \varepsilon_v'^T + \frac{1}{2}\varepsilon_v' G \varepsilon_v'^T + \frac{1}{2}W^T \phi' G \phi'^T \tilde{W}_a + \frac{1}{2}\varepsilon_v' G \phi'^T\right\| \leq \kappa_4, \qquad (12.48)$$

where $\kappa_1, \kappa_2, \kappa_3, \kappa_4 \in \mathbb{R}$ are computable positive constants.

Theorem 12.2 *If Assumptions 1-8 hold, the regressor* $\psi(t) \triangleq \frac{\omega}{\sqrt{1+\omega^T \Gamma \omega}}$ *is PE (persistently exciting), and provided (12.33), (12.34) and the following sufficient gain condition is satisfied*[2]

$$\frac{c_3}{\eta_{a1}} > \kappa_1 \kappa_2, \qquad (12.49)$$

where η_{a1}, c_3, κ_1, κ_2 *are introduced in (12.15), (12.47), and (12.48), then the controller in (12.10), the actor-critic weight update laws in (12.12)-(12.13) and (12.41), and the identifier in (12.17) and (12.22), guarantee that the state of the system* $x(t)$*, and the actor-critic weight estimation errors* $\tilde{W}_a(t)$ *and* $\tilde{W}_c(t)$ *are UUB.*

Proof: To investigate the stability of (12.3) with control $\hat{u}(x)$, and the perturbed system in (12.45), consider $V_L : \mathcal{X} \times \mathbb{R}^N \times \mathbb{R}^N \times [0, \infty) \to \mathbb{R}$ as the continuously differentiable, positive-definite Lyapunov function candidate defined as

$$V_L(x, \tilde{W}_c, \tilde{W}_a, t) \triangleq V^*(x) + V_c(\tilde{W}_c, t) + \frac{1}{2}\tilde{W}_a^T \tilde{W}_a,$$

where $V^*(x)$ (the optimal value function), is the Lyapunov function for (12.3), and $V_c(\tilde{W}_c, t)$ is the Lyapunov function for the exponentially stable system in (12.46). Since $V^*(x)$ is continuously differentiable and positive-definite from (12.1) and (12.2), there exist class $\mathcal{K}$ functions α_1 and α_2 defined on $[0,\ r]$, where $B_r \subset \mathcal{X}$ (see Lemma 4.3 in [22]), such that

$$\alpha_1(\|x\|) \leq V^*(x) \leq \alpha_2(\|x\|) \qquad \forall x \in B_r. \qquad (12.50)$$

[2]Since c_3 is a function of the critic adaptation gain η_c, η_{a1} is the actor adaptation gain, and κ_1, κ_2 are known constants, the sufficient gain condition in (12.49) can be easily satisfied.

Using (12.47) and (12.50), $V_L(x, \tilde{W}_c, \tilde{W}_a, t)$ can be bounded as

$$\alpha_1(\|x\|) + c_1 \left\|\tilde{W}_c\right\|^2 + \frac{1}{2}\left\|\tilde{W}_a\right\|^2 \leq V_L(x, \tilde{W}_c, \tilde{W}_a, t) \leq \alpha_2(\|x\|) + c_2 \left\|\tilde{W}_c\right\|^2 + \frac{1}{2}\left\|\tilde{W}_a\right\|^2,$$

which can be written as

$$\alpha_3(\|\tilde{z}\|) \leq V_L(x, \tilde{W}_c, \tilde{W}_a, t) \leq \alpha_4(\|\tilde{z}\|) \qquad \forall \tilde{z} \in B_s,$$

where $\tilde{z}(t) \triangleq [x(t)^T \ \tilde{W}_c(t)^T \ \tilde{W}_a(t)^T]^T \in \mathbb{R}^{n+2N}$, α_3 and α_4 are class $\mathcal{K}$ functions defined on $[0, \ s]$, where $B_s \subset \mathcal{X} \times \mathbb{R}^N \times \mathbb{R}^N$. Taking the time derivative of $V_L(\cdot)$ yields

$$\dot{V}_L = \frac{\partial V^*}{\partial x} f + \frac{\partial V^*}{\partial x} g\hat{u} + \frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial \tilde{W}_c}\Omega_{nom} + \frac{\partial V_c}{\partial \tilde{W}_c}\Delta_{per} - \tilde{W}_a^T \dot{\hat{W}}_a, \quad (12.51)$$

where the time derivative of $V^*(\cdot)$ is taken along the the trajectories of the system (12.3) with control $\hat{u}(\cdot)$ and the time derivative of $V_c(\cdot)$ is taken along the along the trajectories of the perturbed system (12.45). To facilitate the subsequent analysis, the HJB in (12.5) is rewritten as $\frac{\partial V^*}{\partial x} f = -\frac{\partial V^*}{\partial x} g u^* - Q(x) - {u^*}^T R u^*$. Substituting for $\frac{\partial V^*}{\partial x} f$ in (12.51), using the fact that $\frac{\partial V^*}{\partial x} g = -2{u^*}^T R$ from (12.4), and using (12.15) and (12.47), (12.51) can be upper bounded as

$$\dot{V}_L \leq -Q - {u^*}^T R u^* - c_3 \left\|\tilde{W}_c\right\|^2 + c_4 \left\|\tilde{W}_c\right\| \|\Delta_{per}\| + 2{u^*}^T R(u^* - \hat{u}) + \eta_{a2}\tilde{W}_a^T(\hat{W}_a - \hat{W}_c) + \frac{\eta_{a1}}{\sqrt{1+\omega^T\omega}}\tilde{W}_a^T \phi' G {\phi'}^T (\hat{W}_a - \hat{W}_c)\delta_{hjb}. \quad (12.52)$$

Substituting for $u^*, \ \hat{u}, \ \delta_{hjb}$, and Δ_{per} using (12.4), (12.10), (12.42), and (12.45), respectively, and using (12.14) and (12.44) in (12.52), yields

$$\begin{aligned}\dot{V}_L \leq & -Q - c_3 \left\|\tilde{W}_c\right\|^2 - \eta_{a2}\left\|\tilde{W}_a\right\|^2 + \frac{1}{2}W^T\phi' G {\varepsilon'_v}^T + \frac{1}{2}W^T\phi' G {\phi'}^T \tilde{W}_a \\ & + c_4 \frac{\eta_c \varphi_0}{2\sqrt{\nu\varphi_1}} \left\| -W^T\phi'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T\phi' G {\phi'}^T \tilde{W}_a - \frac{1}{4}\varepsilon'_v G {\varepsilon'_v}^T - \varepsilon'_v F_{u^*} \right\| \left\|\tilde{W}_c\right\| \\ & + \frac{\eta_{a1}}{\sqrt{1+\omega^T\omega}}\tilde{W}_a^T\phi' G {\phi'}^T(\tilde{W}_c - \tilde{W}_a)\left(-\tilde{W}_c^T\omega - W^T\phi'\tilde{F}_{\hat{u}} - \frac{1}{4}\varepsilon'_v G {\varepsilon'_v}^T \right. \\ & \left. + \frac{1}{4}\tilde{W}_a^T\phi' G {\phi'}^T \tilde{W}_a - \varepsilon'_v F_{u^*}\right) + \frac{1}{2}\varepsilon'_v G {\phi'}^T \tilde{W}_a + \eta_{a2}\left\|\tilde{W}_a\right\|\left\|\tilde{W}_c\right\| + \frac{1}{2}\varepsilon'_v G {\varepsilon'_v}^T. \end{aligned} \quad (12.53)$$

Using the bounds developed in (12.48), (12.53) can be further upper bounded as

$$\begin{aligned}\dot{V}_L &\leq -Q - (c_3 - \eta_{a1}\kappa_1\kappa_2)\left\|\tilde{W}_c\right\|^2 - \eta_{a2}\left\|\tilde{W}_a\right\|^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4 \\ &+ \left(\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right)\left\|\tilde{W}_c\right\|.\end{aligned}$$

Provided $c_3 > \eta_{a1}\kappa_1\kappa_2$, and completing the square yields

$$\begin{aligned}\dot{V}_L \leq -Q - (1-\theta)(c_3 - \eta_{a1}\kappa_1\kappa_2)\left\|\tilde{W}_c\right\|^2 - \eta_{a2}\left\|\tilde{W}_a\right\|^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4 \\ + \frac{1}{4\theta(c_3 - \eta_{a1}\kappa_1\kappa_2)}\left[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right]^2\end{aligned} \tag{12.54}$$

where $0 < \theta < 1$. Since $Q(x)$ is positive definite, Lemma 4.3 in [22] indicates that there exist class $\mathcal{K}$ functions α_5 and α_6 such that

$$\alpha_5(\|\tilde{z}\|) \leq Q + (1-\theta)(c_3 - \eta_{a1}\kappa_1\kappa_2)\left\|\tilde{W}_c\right\|^2 + \eta_{a2}\left\|\tilde{W}_a\right\|^2 \leq \alpha_6(\|\tilde{z}\|) \quad \forall v \in B_s,$$

which can be used to further upper bound the expression in (12.54) as

$$\begin{aligned}\dot{V}_L \leq -\alpha_5(\|\tilde{z}\|) + \frac{1}{4\theta(c_3 - \eta_{a1}\kappa_1\kappa_2)}\left[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right]^2 \\ + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4,\end{aligned}$$

which proves that $\dot{V}_L(\cdot)$ is negative whenever $\tilde{z}(t)$ lies outside the compact set $\Omega_{\tilde{z}} \triangleq \left\{\tilde{z} : \|\tilde{z}\| \leq \alpha_5^{-1}\left(\frac{1}{4\theta(c_3-\eta_{a1}\kappa_1\kappa_2)}\left[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right]^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4\right)\right\}$, and hence, $\|\tilde{z}(t)\|$ is UUB (see Theorem 4.18 in [22]). The bounds in (12.48) depend on the actor NN approximation error ε'_v, which can be reduced by increasing the number of neurons N, thereby reducing the size of the residual set $\Omega_{\tilde{z}}$. From Assumption 7, as the number of neurons of the actor and critic NNs $N \to \infty$, the reconstruction error $\varepsilon'_v \to 0$. ■

Remark 12.2 *Since the actor, critic and identifier are continuously updated, the developed RL algorithm can be compared to fully optimistic PI in machine learning literature [8], where policy evaluation and policy improvement are done after every state transition, unlike traditional PI, where policy improvement is done after convergence of the policy evaluation step. Proving convergence of optimistic PI*

is complicated and is an active area of research in machine learning [8, 10]. By considering an adaptive control framework, this result investigates the convergence and stability behavior of fully optimistic PI in continuous-time.

Remark 12.3 *The PE condition in Theorem 12.2 is equivalent to the exploration paradigm in RL which ensures sufficient sampling of the state space and convergence to the optimal policy [42].*

12.6 SIMULATION

The following nonlinear system is considered [44]

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (cos(2x_1) + 2)^2) \end{bmatrix} + \begin{bmatrix} 0 \\ cos(2x_1) + 2 \end{bmatrix} u, \quad (12.55)$$

where $x(t) \triangleq [x_1(t) \;\; x_2(t)]^T \in \mathbb{R}^2$ and $u(t) \in \mathbb{R}$. The state and control penalties are chosen as

$$Q(x) = x^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x; \qquad R = 1.$$

The optimal value function and optimal control for the system in (12.55) are known, and given by [44]

$$V^*(x) = \frac{1}{2}x_1^2 + x_2^2; \qquad u^*(x) = -(cos(2x_1) + 2)x_2.$$

The activation function for the critic NN is selected with $N = 3$ neurons as

$$\phi(x) = [x_1^2 \quad x_1x_2 \quad x_2^2]^T,$$

while the activation function for the identifier DNN is selected as a symmetric sigmoid with $L_f = 5$ neurons in the hidden layer. The identifier gains are selected as

$$k = 800, \quad \alpha = 300, \quad \gamma = 5, \quad \beta_1 = 0.2, \quad \Gamma_{wf} = 0.1\mathbb{I}_{6\times 6}, \quad \Gamma_{vf} = 0.1\mathbb{I}_{2\times 2},$$

and the gains for the actor-critic learning laws are selected as

$$\eta_{a1} = 10, \quad \eta_{a2} = 50, \quad \eta_c = 20, \quad \nu = 0.005.$$

The covariance matrix is initialized to $\Gamma(0) = 5000$, all the NN weights are randomly initialized in $[-1,\ 1]$, and the states are initialized to $x(0) = [3,\ -1]$. An implementation issue in using the developed algorithm is to ensure PE of the critic regressor vector. Unlike linear systems, where PE of the regressor translates to sufficient richness of the external input, no verifiable method exists to ensure PE in nonlinear regulation problems. To ensure PE qualitatively, a small exploratory signal consisting of sinusoids of varying frequencies, $n(t) = sin^2(t)cos(t) + sin^2(2t)cos(0.1t) + sin^2(-1.2t)cos(0.5t) + sin^5(t)$, is added to the control $u(t)$ for the first 3 seconds [44]. The evolution of states is shown in Fig. 12.2. The identifier approximates the system dynamics, and the state derivative estimation error is shown in Fig. 12.3. Persistence of excitation ensures that the weights converge to their optimal values of $W = [0.5\ \ 0\ \ 1]^T$ in approximately 2 seconds, as seen from the evolution of actor-critic weights in Figs. 12.4 and 12.5. The errors in approximating the optimal value function and optimal control at steady state ($t = 10\,sec.$) are plotted against the states in Figs. 12.6 and 12.7, respectively.

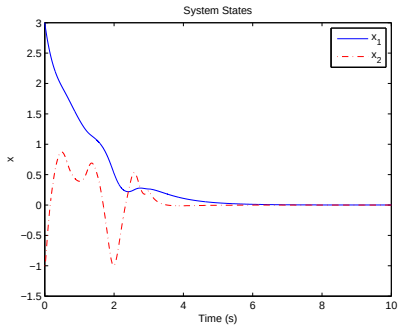

Figure 12.2: System states $x(t)$ with persistently excited input for the first 3 seconds.

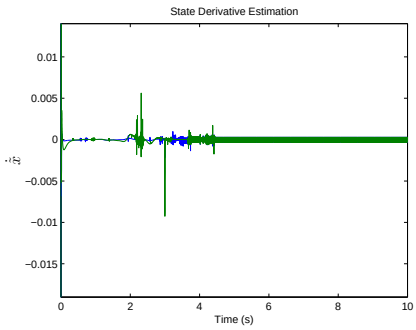

Figure 12.3: Error in estimating the state derivative $\dot{\hat{x}}(t)$ by the identifier.

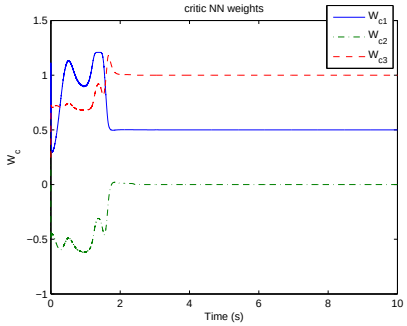

Figure 12.4: Convergence of critic weights $\hat{W}_c(t)$.

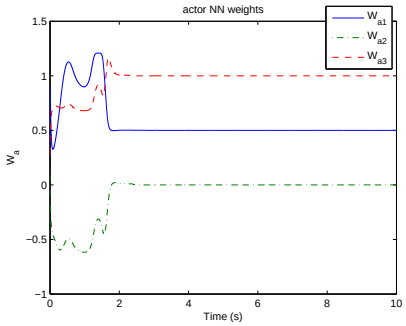

Figure 12.5: Convergence of actor weights $\hat{W}_a(t)$.

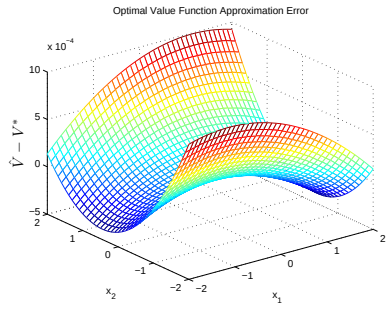

Figure 12.6: Error in approximating the optimal value function by the critic at steady state.

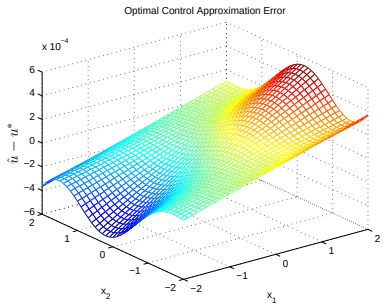

Figure 12.7: Error in approximating the optimal control by the actor at steady state.

12.7 CONCLUSION

An actor-critic-identifier architecture is proposed to learn the approximate solution to the HJB equation for infinite-horizon optimal control of uncertain nonlinear systems. The online method is the first ever indirect adaptive control approach to continuous-time RL. The learning by the actor, critic and identifier is continuous and simultaneous, and the novel addition of the identifier to the traditional actor-critic architecture eliminates the need to know the system drift dynamics. The actor and critic minimize the Bellman error using gradient and least-squares update laws, respectively, and provide online approximations to the optimal control and the optimal value function, respectively. The identifier estimates the system dynamics online and asymptotically converges to the system state and its derivative. A PE condition is required to ensure exponential convergence to a bounded region in the neighborhood of the optimal control and UUB stability of the closed-loop system. Simulation results demonstrate the performance of the actor-critic-identifier-based method. A limitation of the method is the requirement of the knowledge of the input gain matrix. Future efforts will investigate ways to overcome this limitation, e.g., using methods similar to the model-free Q-learning methods [9, 29, 47].

REFERENCES

1. M. Abu-Khalaf and F.L. Lewis. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 41(5):779–791, 2005.

2. A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf. Model-free q-learning designs for linear discrete-time zero-sum games with application to h-[infinity] control. *Automatica*, 43:473–481, 2007.

3. L.C. Baird. Advantage updating. Technical report, Wright Lab, Wright-Patterson Air Force Base, OH, 1993.

4. A.G. Barto, R.S. Sutton, and C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybern.*, 13(5):834–846, 1983.

5. R.W. Beard, G.N. Saridis, and J.T. Wen. Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation. *Automatica*, 33:2159–2178, 1997.

6. R. Bellman. *Dynamic Programming*. Dover Publications, Inc., 2003.

7. D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.

8. D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

9. S.J. Bradtke, B.E. Ydstie, and A.G. Barto. Adaptive linear quadratic control using policy iteration. In *Proc. Am. Control Conf.*, pages 3475–3479. IEEE, 1994.

10. L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.

11. F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

12. G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2:303–314, 1989.

13. T. Dierks, B.T. Thumati, and S. Jagannathan. Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence. *Neural Networks*, 22(5-6):851–860, 2009.

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

15. K. Doya. Reinforcement learning in continuous time and space. *Neural Comput.*, 12(1):219–245, 2000.

16. S. Ferrari and RF Stengel. An adaptive critic global controller. In *Proc. Am. Control Conf.*, volume 4, 2002.

17. A. Filippov. Differential equations with discontinuous right-hand side. *Am. Math. Soc. Transl.*, 42 no. 2:199–231, 1964.

18. A. F. Fillipov. *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, 1988, pp. 48-122.

19. T. Hanselmann, L. Noakes, and A. Zaknich. Continuous-time adaptive critics. *IEEE Trans. Neural Networks*, 18(3):631–647, 2007.

20. J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci. U.S.A.*, 81(10):3088, 1984.

21. K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1985.

22. H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002.

23. D. Kirk. *Optimal Control Theory: An Introduction*. Dover Pubns, 2004.

24. D. Kleinman. On an iterative technique for riccati equation computations. *IEEE Trans. Autom. Contr.*, 13(1):114–115, 1968.

25. Miroslav Krstic, Petar V. Kokotovic, and Ioannis Kanellakopoulos. *Nonlinear and Adaptive Control Design*. John Wiley & Sons, 1995.

26. T. Landelius. *Reinforcement learning and Distributed Local Model Synthesis*. PhD thesis, Linkoping University, Sweden, 1997.

27. F. L. Lewis, R. Selmic, and J. Campos. *Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

28. X. Liu and S. N. Balakrishnan. Convergence analysis of adaptive critic based optimal control. In *Proc. Am. Control Conf.*, volume 3, 2000.

29. P. Mehta and S. Meyn. Q-learning and Pontryagin's minimum principle. In *Proc. IEEE Conf. Decis. Control*, pages 3598–3605, 2009.

30. J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks. Adaptive dynamic programming. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 32(2):140–153, 2002.

31. B. Paden and S. Sastry. A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators. *IEEE Trans. Circuits Syst.*, 34 no. 1:73–82, 1987.

32. R. Padhi, S.N. Balakrishnan, and T. Randolph. Adaptive-critic based optimal neuro control synthesis for distributed parameter systems. *Automatica*, 37(8):1223–1234, 2001.

33. P. M. Patre, W. MacKunis, K. Kaiser, and W. E. Dixon. Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure. *IEEE Trans. Autom. Control*, 53(9):2180–2185, 2008.

34. A. S. Poznyak, E. N. Sanchez, and W. Yu. *Differential neural networks for robust nonlinear control: identification, state estimation and trajectory tracking*. World Scientific Pub Co Inc, 2001.

35. D. V. Prokhorov and II Wunsch, D. C. Adaptive critic designs. *IEEE Trans. Neural Networks*, 8:997–1007, 1997.

36. S. Sastry and M. Bodson. *Adaptive Control: Stability, Convergence, and Robustness*. Prentice-Hall, Upper Saddle River, NJ, 1989.

37. D. Shevitz and B. Paden. Lyapunov stability theory of nonsmooth systems. *IEEE Trans. Autom. Control*, 39 no. 9:1910–1914, 1994.

38. J. Si, A. Barto, W. Powell, and D. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*. Wiley-IEEE Press, 2004.

39. J. Si and Y. T. Wang. On-line learning control by association and reinforcement. *IEEE Trans. Neural Networks*, 12(2):264–276, 2001.

40. G. V. Smirnov. *Introduction to the theory of differential inclusions*. American Mathematical Society, 2002.

41. R. S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44, 1988.

42. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

43. R.S. Sutton and A.G. Barto. *Introduction to reinforcement learning*. MIT Press Cambridge, MA, USA, 1998.

44. K.G. Vamvoudakis and F.L. Lewis. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46:878–888, 2010.

45. D. Vrabie, M. Abu-Khalaf, FL Lewis, and Y. Wang. Continuous-time ADP for linear systems with partially unknown dynamics. In *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinf. Learn.*, pages 247–253, 2007.

46. D. Vrabie and F. L. Lewis. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 22(3):237 – 246, 2009.

47. C. J. C. H. Watkins and P. Dayan. Q-learning. *Mach. Learn.*, 8(3):279–292, 1992.

48. P. J. Werbos. Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Trans. Syst. Man Cybern.*, 17(1):7–20, 1987.

49. P. J. Werbos. Approximate dynamic programming for real-time control and neural modeling. In David A. White and Donald A. Sofge, editors, *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*. Van Nostrand Reinhold, New York, 1992.

50. D. A. White and D. A. Sofge. *Handbook of intelligent control: neural, fuzzy, and adaptive approaches*. Van Nostrand Reinhold Company, 1992.

51. B. Widrow, N. K. Gupta, and S. Maitra. Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Trans. Syst. Man Cybern.*, 3(5):455–465, 1973.

52. B. Xian, D. M. Dawson, M. S. de Queiroz, and J. Chen. A continuous asymptotic tracking control strategy for uncertain nonlinear systems. *IEEE Trans. Autom. Control*, 49:1206–1211, 2004.