Received XX Month 2025; revised ; accepted XX Month 2025; Date of publication XX Month 2025; date of current version XX Month 2025.

Adaptive Actor-Critic Based Optimal Regulation for Drift-free Nonlinear Systems

Ashwin P. Dani[1] (Senior Member, IEEE), Shubhendu Bhasin[2] (Member, IEEE)

[1]Electrical and Computer Engineering, University of Connecticut Storrs, CT 06269 USA
[2]Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India

CORRESPONDING AUTHOR: A.P. Dani (e-mail: ashwin.dani@uconn.edu)

This work was supported in part by a Space Technology Research Institutes grant (number 80NSSC19K1076) from NASA Space Technology Research Grants Program and in part by NSF grant no. SMA-2134367.

ABSTRACT In this paper, a continuous-time adaptive actor-critic reinforcement learning (RL) controller is developed for drift-free uncertain nonlinear systems. Practical examples of such systems are image-based visual servoing (IBVS) and wheeled mobile robots (WMR), where the system dynamics include a parametric uncertainty in the control effectiveness matrix with no drift term. The uncertainty in the input term poses a challenge when developing a continuous-time RL controller using existing methods. This paper presents an actor-critic/synchronous policy iteration (PI)-based RL controller with a newly derived constrained concurrent learning (CCL)-based parameter update law for estimating the unknown parameters of the linearly parametrized control effectiveness matrix. The parameter update law ensures that the parameters do not converge to $zero$, avoiding possible loss of stabilization. An infinite-horizon value function minimization objective is achieved by regulating the current states to the desired with near-optimal control efforts. The proposed controller guarantees closed-loop stability, and simulation results in the presence of noise validate the proposed theory using IBVS and WMR examples.

INDEX TERMS Actor-critic policy iteration, drift-free systems, reinforcement learning

I. INTRODUCTION

Drift-free dynamics are commonly found in robotics and other engineering applications. These are systems of the form $\dot{x} = g(x)u$. Some examples of such systems are image-based visual servo (IBVS) control, wheeled mobile robot (WMR) [1], shape servoing control [2], models of kinematic drift effects in space systems [3] etc. Reinforcement learning (RL) has successfully provided a means to design optimal adaptive controllers for various classes of systems [4]–[8]. For the drift-free systems, when there is a parametric uncertainty in the control effectiveness term $g(\cdot)$, existing continuous-time model-based RL solutions cannot be applied to design an RL policy. In this paper, an adaptive actor-critic (AAC) method is developed for a class of drift-free nonlinear systems with uncertainty in the control effectiveness matrix.

RL learns the optimal policy that maximizes a long-term reward. By interacting with the environment, the decision maker gets evaluative feedback about its actions, which is used to improve the control policy [9]. A popular class of iterative RL methods is adaptive dynamic programming (ADP), which Werbos introduced for discrete-time (DT) systems [5], [10], [11], and implemented in actor-critic (AC) framework. Extension of RL algorithms to continuous-time systems is achieved in [12] using the Hamilton-Jacobi-Bellman (HJB) framework with known system dynamics, where a continuous-time version of the temporal difference error is employed. Several offline approaches for solving a generalized HJB equation are developed in [13], [14], using Galerkin's spectral approximation [13] and least-squares successive approximation solution [14] to HJB, which is then used to compute the optimal control.

When the system dynamics are not completely known, among online approaches, an integral reinforcement learning (IRL) method is developed in [15], [16], which requires only partial knowledge of system dynamics. The approach called policy iteration (PI) is designed based on AC structure, where the actor neural network (NN) is learned at a faster time scale than the critic NN. In [17], the IRL approach

is extended to simultaneously learn both the actor and critic NNs, leading to a new method called synchronous PI. Further, in [18], an actor-critic-identifier (ACI) approach is presented, which in addition to actor and critic NNs, uses an identifier network to identify the unknown drift term in the dynamics. A model-based PI algorithm is developed in [19] for unknown drift dynamics where concurrent learning (CL)-based model identification is used to identify the drift part of the dynamics. In [20], the method in [19] is extended to systems with control effectiveness faults. In [21], a robust actor-critic RL policy is developed for a class of nonlinear systems where certain parameterized unknown parts of the dynamics are estimated using adaptive update law. The above-mentioned methods require a complete knowledge of the input gain or control effectiveness matrix.

The method in [22] identifies the complete nonlinear system dynamics using the experience replay technique and learns the actor-critic NN using the PI method for completely unknown dynamics. A data-driven approach to the actor-critic algorithm is developed in [23], where the system dynamics are identified assuming the dynamics can be written as a combination of linear and NN part, and a converged $g(x)$ term is recovered from the identifier before using it in the AC structure. However, in these papers, no constraint on the parameter estimation of $g(x)$ term is used which may lead to $g(x)$ estimates converge to 0 leading to loss of stabilization [24]. In [25] an identifier is used to identify the dynamics in the PI structure, where the state and control input data at sampled time instances is used to estimate the identifier and actor, critic weight parameters which reduces the computational burden of computing the parameter update laws. The control matrix $g(x)$ is estimated away from 0 by using a simple switching rule to a g_{min} value. A backstepping technique is used in [26] to design a finite-time stabilizing controller with unknown dynamics parameterized using NN, which is then proved to be optimal with respect to a performance criterion. The identifier design in these papers is decoupled from the controller stability. However, the problem of loss of stabilization while estimating parameters of control matrix $g(x)$ or $\dot{x} \neq 0$ at non-equilibrium states is not explicitly addressed in these designs. In our paper, a new adaptive parameter update law is designed to estimate control matrix $g(x)$ using Lyapunov stability analysis in the same spirit as an indirect adaptive control in the context of actor-critic PI method while addressing the issue of loss of stabilization. For model parameter estimation, gradient-based adaptive update law is a standard approach in adaptive control, which requires that the regressor be persistently exciting (PE) for parameter convergence. Concurrent learning (CL) or its variant integral concurrent learning (ICL) uses historical data along with relaxed finite excitation conditions. Methods suggested in [27] use initial excitation condition and filtering techniques to derive parameter update law. However, when used for estimating parameters of $g(x)$, these methods may not prevent the parameter estimates, and control matrix $g(x)$ from converging to zero or its neighborhood causing loss of stabilization or $\dot{x} = 0$ for non-equilibrium states. Recently, a metric using Bergman divergence measure is used in [28] to derive parameter update law, which also suggest using barrier type function.

Many model-free approaches for continuous-time systems are developed using Q-learning and its variants where no knowledge of system dynamics is required to obtain an optimal policy [29]–[32]. For linear systems, a completely model-free RL method is developed in [33], which iteratively solves the algebraic Riccati equation using online information of state and input. Initialization with a stabilizing policy is required for this method. Using the IRL framework, an on-policy model-free Q learning approach is developed in [34] for linear systems where no stabilizing policy initialization is required.

The off-policy RL methods have been designed for the regulation and tracking problems for continuous-time dynamics with partially or completely unknown nonlinear system dynamics [35], [36], and for discrete-time linear systems [37]. The off-policy RL methods require two phases, first one is that of data collection where a stabilizing policy collects state, action data. This data is then used to solve a Bellman equation to obtain an optimal control using recursive least squares. Although these methods do not require any knowledge of system dynamics, the data collection phase must be carried out first which can be time consuming before the optimal control action is obtained. Off-policy methods also require a stabilizing behavior policy, the state-action data depends on the choice of behavior policy. Whereas for on-policy methods there is no separate data collection phase required, the policy that is being trained is also used for exploration. Off-policy methods are data-driven, and hence, can be data-intensive for finding optimal control. Whenever some knowledge of the system dynamics is known, e.g., structure of $g(x)$ matrix for drift-free systems, it can be used to derive model-based RL controller instead of purely data-driven approach assuming no knowledge of the system dynamics. Technical development in this paper follows along the lines of model-based RL.

Further, the drift-free systems are in some ways fundamentally different from control-affine systems with drift and, cannot be stabilized using smooth feedback due to Brockett's condition for stabilization of such systems [38]. Owing to the fact that drift-free systems have no natural dynamics and are purely driven by inputs, there is no way for the controller to exploit the natural dynamics for stabilization, necessitating nonlinear, time-varying, and often customized control strategies [38].

Contributions

The contribution of this paper is to design an AAC RL algorithm for a class of uncertain drift-free nonlinear systems. The parametric uncertainty in the control effectiveness matrix complicates the design of the PI

RL algorithm for the drift-free systems. The unknown parameters of the system dynamics (input gain/control effectiveness) are modeled as constants for which a novel constrained concurrent learning (CCL)-based adaptive parameter update law is developed using Barrier function used in optimization literature along with update law for Lagrange multipliers. As long as the parameter estimates are initialized within a set of required bounds, the newly designed inverse Barrier parameter update law ensures that they stay within the bound. The CL method uses a finite excitation condition that can be verified in real time. The technical development in [28] shows that the standard gradient adaptive control parameter update law can also be derived by taking a gradient of an optimization function. The adaptive update laws are termed as a velocity gradient update law in [28], [39], [40]. Following this technical development, if the parameters are constrained then a standard approach in optimization is to solve an unconstrained optimization problem by formulating a Lagrangian function [41]. Thus, to derive a parameter update law that should satisfy constraints, a Lagrangian function is formulated. The stability analysis of the closed-loop system under the constrained adaptive update law remains to be answered. Lyapunov stability analysis with the constrained parameter update law is developed inspired by the stability analysis used for primal-dual saddle point dynamics [42]–[45]. This is one of the key contributions of the paper. The critic and actor NNs approximate the optimal value function and optimal control, respectively. Similar to the development in [18], [19], the critic NN weight update law is derived based on the minimization of the Bellman error computed using optimal and approximate HJB equation. A least-squares weight update law is derived, which uses a PE condition of the critic regressor. Similarly, a gradient-based NN weight update law is derived for actor NN based on the minimization of Bellman error. The parameter and actor-critic weights and the system model are learned simultaneously as new state and control input data becomes available. Lyapunov stability analysis shows an exponential convergence of the state and parameter estimation errors to an ultimate bound, leading to uniformly ultimately bounded (UUB) stability. Even though the actor-critic structure is similar to the one presented in [18], [19], the AAC structure presented in this paper requires a new stability analysis using which the model parameter update law is designed. This is another contribution of this paper. The proposed AAC policy is validated using two simulation examples, IBVS and WMR. Although the controller is designed in a deterministic setting, in both simulation studies, the controller regulates the system state to its desired value in the presence of system state measurement noise. Compared to our prior work in [46], where an RL controller is designed for the IBVS system, this paper derives the RL-policy for a generalized case of vector model parameter using parameter update law derived using constrained CL. Additional simulation example of WMR regulation control is also included.

II. SYSTEM MODEL AND CONTROL OBJECTIVE

A. SYSTEM DYNAMICS

Consider the following system representing the evolution of the states as a function of the control input. The system model can be written in the following form

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

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control input, $\theta \in \mathbb{R}^p$ is an unknown parameter vector. The input gain matrix $g(x, \theta) \in \mathbb{R}^{n\times m}$ is expressed in parametric form as $\text{vec}(g(x, \theta)) = Y(x)\theta$, where $Y \in \mathbb{R}^{nm\times p}$ is a regressor matrix, $\text{vec} : \mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ is a vectorization operator and $\text{vec}^{-1} : \mathbb{R}^{nm} \to \mathbb{R}^{n\times m}$ is an inverse vectorization operator. Using the parametric form of g, and properties of Kronecker product $\otimes$, the dynamics in (1) can be written as linear-in-parameter (LIP) form, given by

$$\dot{x} = \mathcal{Y}\theta \tag{2}$$

where $\mathcal{Y}(x, u) = (u^T \otimes I_{n\times n})Y(x) \in \mathbb{R}^{n\times p}$.

Remark 1:
Many practical system dynamics can be parametrized in LIP form. Two examples, namely, WMR and IBVS system, are provided in the simulation section.

B. CONTROLLER OBJECTIVE

The control objective is to regulate the current state to the desired state, $x_d \in \mathbb{R}^n$ by minimizing an objective function. For the control design, the regulation error $\bar{x}(t) \in \mathbb{R}^n$ is defined as

$$\bar{x}(t) \triangleq x(t) - x_d \tag{3}$$

and the parameter estimation error $\tilde{\theta}(t) \in \mathbb{R}^p$ is defined as

$$\tilde{\theta}(t) \triangleq \theta - \hat{\theta}(t). \tag{4}$$

where $\hat{\theta}(t) \in \mathbb{R}^p$ is the parameter estimate. Since the optimal regulation objective is to bring the state $x(t)$ to a non-zero desired state x_d, the system model (1) is first written in terms of $\bar{x}(t)$

$$\dot{\bar{x}} = g(x, \theta)u = g(\bar{x}, x_d, \theta)u. \tag{5}$$

A continuous adaptive actor-critic controller is designed using the system in (5) with the objective to optimally regulate the state $\bar{x}(t)$ to 0 with the minimum control effort $u(t)$. The following assumption is made on the system dynamics to facilitate the stability analysis.

Assumption 1:
The function $g(x, \theta)$ is continuous and bounded $0 < g(x, \theta) < \bar{g}$ with a known bound on a set $\mathcal{X} \subset \mathbb{R}^n$.

III. OPTIMAL CONTROL DESIGN USING ACTOR-CRITIC STRUCTURE

A. CONTINUOUS RL-BASED CONTROLLER DESIGN

An RL-based controller is designed to achieve the desired control objective given by the optimal value function

$V^*(\bar{x}) \in \mathbb{R}^+$, defined by

$$V^*(\bar{x}(t)) = \min_{u(\tau)\in\Theta(\mathcal{X})} \int_t^\infty r(s)ds \tag{6}$$

where V^* is continuously differentiable, satisfies $V^*(0) = 0$, $\Theta(\mathcal{X})$ is a set of admissible policies, $r(\bar{x}, u) \in \mathbb{R}$ is the local cost

$$r(\bar{x}, u) = Q(\bar{x}) + u^T R u \tag{7}$$

where $Q(\bar{x})$ is a positive definite function and $R = R^T > 0$. Given the dynamics (5) and the value function (6), the optimal control is given by

$$u^* = -\frac{1}{2} R^{-1} g^T(x, \theta) V_{\bar{x}}^{*T} \tag{8}$$

where $V_{\bar{x}}^* = \frac{\partial V^*}{\partial \bar{x}}$.

B. HAMILTONIAN AND BELLMAN ERROR

The Hamiltonian of the system is given by

$$H(\bar{x}, u, V_{\bar{x}}) = V_{\bar{x}} g(x, \theta) u + r(\bar{x}, u) \tag{9}$$

where $V_{\bar{x}} = \frac{\partial V}{\partial \bar{x}}$. The optimal Hamiltonian associated with the optimal cost and optimal control is given by

$$H(\bar{x}, u^*, V_{\bar{x}}^*) = V_{\bar{x}}^* g(x, \theta) u^* + r(\bar{x}, u^*) = 0. \tag{10}$$

Computing the value function $V(\bar{x})$ and the optimal controller requires the solution to the HJB, which is a partial differential equation. It is, in general, hard to find an analytical solution for HJB. The value function is approximated using an NN called a critic NN, and the corresponding optimal control is approximated using an actor NN. Using the approximated cost $\hat{V} \in \mathbb{R}$ and controller $\hat{u} \in \mathbb{R}^m$, the approximated Hamiltonian is computed as

$$H(\bar{x}, \hat{u}, \hat{V}_{\bar{x}}) = \hat{V}_{\bar{x}} g(x, \hat{\theta}) \hat{u} + r(\bar{x}, \hat{u}) \tag{11}$$

Using the optimal and approximated Hamiltonian, a temporal difference or Bellman error $\delta \in \mathbb{R}$ is computed as follows

$$\delta = H(\bar{x}, \hat{u}, \hat{V}_{\bar{x}}) - H(\bar{x}, u^*, V_{\bar{x}}^*) = \hat{V}_{\bar{x}} g(x, \hat{\theta}) \hat{u} + r(\bar{x}, \hat{u}) \tag{12}$$

because the value of the optimal Hamiltonian is 0. Bellman error, δ, in Hamiltonian, is used to learn the critic and actor NN weights. For implementation of the optimal control, the value function and optimal control are approximated using NNs. The following assumptions are made on the NN function approximators.

Assumption 2:
For a given NN, $N(x) = W^T \sigma(V^T x) + \epsilon(x)$, where $x \in \mathcal{X} \subset \mathbb{R}^n$ is a compact set, $\epsilon(x)$ is a function reconstruction error, the ideal NN weights W and V are bounded by known positive constants, i.e., $\|W\|_F \leq \bar{W}$, $\|V\|_F \leq \bar{V}$ [47]. The NN activation function σ and σ' are bounded.

Assumption 3:
Using the universal approximation property of NN, the function reconstruction error and its derivative are bounded, i.e., $\|\epsilon(x)\| \leq \bar{\epsilon}$ and $\|\epsilon'(x)\| \leq \bar{\epsilon}'$ [48].

Assumption 4:
The components θ_i of true parameter vector θ are bounded as $\underline{\theta}_i < \theta_i < \bar{\theta}_i$, where $\bar{\theta}_i$ and $\underline{\theta}_i$ are known.

C. APPROXIMATE OPTIMAL CONTROL

Consider a compact set $\mathcal{X} \subset \mathbb{R}^n$ and the state vector $\bar{x}(t) \in \mathcal{X}$. Using NN representation, the optimal value function and the optimal control are written as

$$\begin{aligned} V^*(\bar{x}(t)) &= W_c^T \phi(\bar{x}) + \epsilon_c(\bar{x}) \\ u^*(\bar{x}) &= -\frac{1}{2} R^{-1} g^T(x, \theta)(\phi'(\bar{x})^T W_c + \epsilon_c'(\bar{x})^T) \end{aligned} \tag{13}$$

where $W_c \in \mathbb{R}^{n_c \times 1}$, $\phi(\bar{x}) : \mathbb{R}^n \to \mathbb{R}^{n_c}$ are the basis functions, $\epsilon(\bar{x}) \in \mathbb{R}$ is the function approximation error and $\epsilon'(\bar{x}) \in \mathbb{R}$ is its derivative with respect to $\bar{x}$. Due to the function approximation error and unknown parameter in $g(x, \theta)$, the value function and optimal control cannot be implemented in practice. Thus, the approximated value function and the optimal control laws are designed as

$$\begin{aligned} \hat{V}(\bar{x}(t)) &= \hat{W}_c^T \phi(\bar{x}) \\ \hat{u}(\bar{x}) &= -\frac{1}{2} R^{-1} g^T(x, \hat{\theta})(\phi'(\bar{x})^T \hat{W}_a) \end{aligned} \tag{14}$$

where $\hat{W}_c \in \mathbb{R}^{n_c \times 1}$ and $\hat{W}_a \in \mathbb{R}^{n_a \times 1}$ are the estimated critic and actor weights.

D. PARAMETER UPDATE LAW

To keep the parameter estimates away from zero or growing too high, which may lead to $\dot{x} = 0$ due to parameter estimation or degrade transient performance before parameter convergence, the following constrained functions are defined on each element $\hat{\theta}_i(t)$ to constrain the parameters

$$c_{1i}(\hat{\theta}) = \frac{-1}{\hat{\theta}_i(t) - \bar{\theta}_i}, \qquad c_{2i}(\hat{\theta}) = \frac{-1}{\underline{\theta}_i - \hat{\theta}_i(t)} \tag{15}$$

where $\bar{\theta}_i \in \mathbb{R}$ are the upper bounds and $\underline{\theta}_i \in \mathbb{R}$ are the lower bounds. The constraints are denoted as $c_j = [c_{j1}, ..., c_{jp}]$ for $j = \{1, 2\}$. A constrained parameter update law is designed using a novel CCL law with an inverse Barrier constrained function as

$$\begin{aligned} \dot{\hat{\theta}} =& PY^T(\hat{u}^T \otimes \hat{W}_c^T \phi')^T + Pk_{cl} \sum_{k=1}^{m} \mathcal{Y}_k^T(\dot{\hat{x}}_k - \mathcal{Y}_k \hat{\theta}(t)) \\ & - \sum_{j=1}^{2} P \mathrm{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j; \end{aligned} \tag{16}$$

$$\dot{\lambda}_j = [-k_j \lambda_j + \Gamma_j^{-1} c_j]_{\lambda_j}^+ \tag{17}$$

where $\mathcal{Y}_k = (\hat{u}_k^T \otimes I_{n\times n}) Y_k$, $P \in \mathbb{R}^{p\times p}$, $\Gamma_j \in \mathbb{R}^{p\times p}$ are PD diagonal learning rate matrices, $\lambda = [\lambda_1^T, \lambda_2^T]^T \in \mathbb{R}^{2p}$ are positive Lagrange multipliers such that $\lambda_j(t_0) > 0_{1\times p}$, $k_j > 0$ and $k_{cl} > 0$ are constant gains. The gradient $\nabla_{\hat{\theta}} c_j$ is computed component-wise, i.e., $\nabla_{\hat{\theta}} c_j = [\frac{\partial c_{j1}}{\partial \hat{\theta}_1}, ..., \frac{\partial c_{jp}}{\partial \hat{\theta}_p}]^T$. The operator $[a]_b^+$ for $b \in \mathbb{R}_{\geq 0}$ is defined as

$$[a]_b^+ = \begin{cases} a, & \text{if } b > 0 \\ \max\{0, a\}, & \text{if } b = 0 \end{cases} \tag{18}$$

To implement the parameter update law in (16), a history stack is collected $\mathcal{H} = \{x_k, \hat{u}_k, \dot{\hat{x}}_k\}_{k=1}^{k=m}$ with m number of data points. An estimate of the state derivative $\dot{\hat{x}}_k$ can be obtained using numerical techniques [19]. By collecting a history stack, information about the constant parameter θ can be obtained. CL-based parameter estimation law uses the finite excitation condition of the system trajectories of (1). In place of inverse Barrier type constrained function presented in (15), log-Barrier type constrained functions can also be used [41].

The design of constrained parameter update law is inspired by the technical development in [28], where the connection is made between standard gradient parameter update law in adaptive control [49] and velocity gradient algorithms in machine learning [39], [40]. It is shown that the standard gradient-based adaptive control law can also be derived as a gradient of an optimization function. Following along this technical development if the parameter has constraints then a standard approach in optimization is to formulate a Lagrangian function and solve the unconstrained optimization problem. To add constraints on the parameters, a Lagrangian function (shown in stability analysis section) is formulated. The gradient of the Lagrangian function yields the constrained parameter update law shown in (16). The first term of the parameter update law is a gradient-like term, the second term is a concurrent learning term and the third term is used to keep the parameter estimates bounded where λ_j are the Lagrange parameters that are updated according to (17). The Lagrangian parameter acts as a scaling factor that controls the weight of the constraint function. The Lagrangian parameters are updated according to the update law in (17). Naturally, the stability analysis of the closed-loop system is a question to be answered for the newly derived constrained parameter update law, which is answered in the stability analysis section.

Consider the parameter estimation error $\tilde{\theta}(t) = \theta - \hat{\theta}(t)$. Substituting $\dot{x}_k$ from (2) in (16), the parameter estimation error dynamics can be written as

$$\dot{\tilde{\theta}} = -PY^T(\hat{u}^T \otimes \hat{W}_c^T \phi')^T - Pk_{cl} \sum_{k=1}^{m} \mathcal{Y}_i^T \mathcal{Y}_i \tilde{\theta} + \sum_{j=1}^{2} P\text{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j \tag{19}$$

Assumption 5:
For the history stack $\mathcal{H} = \{x_k, \hat{u}_k, \dot{\hat{x}}_k\}_{k=1}^{k=m}$ the following condition is satisfied

$$\lambda_{min}(\sum_{k=1}^{m} \mathcal{Y}_k^T \mathcal{Y}_k) = \bar{\sigma}_1 > 0, \tag{20}$$

where $\bar{\sigma}_1 \in \mathbb{R}^+$. The numerically computed derivatives of $x(t)$, $\dot{\hat{x}}_k$ computed at kth data point satisfies $\|\dot{\hat{x}}_k - \dot{x}_k\| \leq \epsilon$ for a small positive number $\epsilon \in \mathbb{R}^+$.

Remark 2:
Assumption 5 is a finite excitation condition that can be verified in real-time [50].

Remark 3:
The initial excitation technique proposed in [27] or ICL proposed in [51] for parameter estimation in adaptive control can also be used for designing parameter update law.

Remark 4:
The newly proposed constrained parameter update law is based on primal-dual dynamics that arise in constrained optimization [44], [45]. It ensures that $\dot{x} \neq 0$ for non-zero control u. The commonly studied adaptive control laws based on gradient descent and recursive least squares do not typically ensure that the parameters stay within the specified bounds [24], [49].

Remark 5:
The parameter estimates should be initialized within specified parameter bounds given in Assumption 4.

Remark 6:
For $\theta \in \mathbb{R}$, the parameter update law reduces to

$$\dot{\hat{\theta}} = P\hat{u}^T Y^T \phi'^T \hat{W}_c + k_{cl} P \sum_{k=1}^{m} u_k^T Y_{rk}^T (\dot{\hat{x}}_k - Y_{rk} u_k \hat{\theta}(t))$$

where $P \in \mathbb{R}$.

E. BELLMAN ERROR

Let the actor-critic NN approximation errors be defined as $\tilde{W}_c(t) = W_c - \hat{W}_c(t)$ and $\tilde{W}_a(t) = W_a - \hat{W}_a(t)$. The actor and critic NN weights are updated using weight update laws that minimize the error between the approximated Hamiltonian and the optimal one, given by Bellman error. The Bellman error in a measurable form in terms of actor and critic NN weights is written as

$$\delta = \hat{W}_c^T \phi'(\bar{x}) g(x, \hat{\theta}) \hat{u} + r(\bar{x}, \hat{u}) \tag{21}$$

For the analysis, another form of Bellman error based on (12) is derived as follows

$$\begin{aligned} \delta = {} & \hat{W}_c^T \phi'(\bar{x}) g(x, \hat{\theta}) \hat{u} + \hat{u}^T R \hat{u} \\ & - W_c^T \phi'(\bar{x}) g(x, \theta) u^* - u^{*T} R u^* - \epsilon_c' g(x, \theta) u^* \end{aligned} \tag{22}$$

which by adding and subtracting $W_c^T \phi'(\bar{x}) g(x, \hat{\theta}) \hat{u}$ can be written as

$$\begin{aligned} \delta = {} & -\tilde{W}_c^T w - W_c^T \phi' \tilde{g} \tilde{u} - \epsilon_c' g u^* + \frac{1}{4} \tilde{W}_a^T g_\phi \tilde{W}_a \\ & - \frac{1}{2} \tilde{W}_a^T g_\phi W_c + \frac{1}{4} \hat{W}_a^T \tilde{g}_\phi \hat{W}_a - \frac{1}{2} \hat{W}_a^T \tilde{g}_{a\phi} \hat{W}_a \\ & - \frac{1}{4} \epsilon_c' g_r \epsilon_c'^T - \frac{1}{2} \epsilon_c'^T g_r \phi'^T W_c \end{aligned} \tag{23}$$

where $\tilde{u} = u^* - \hat{u}$, $\tilde{g} = g - \hat{g}$, $g_\phi = \phi' g R^{-1} g^T \phi'^T$, $g_r = g(x, \theta) R^{-1} g(x, \theta)^T$, $\tilde{g}_{a\phi} = \phi' g R^{-1} \tilde{g}^T \phi'^T$, $\tilde{g}_\phi =$

$\phi'\tilde{g}R^{-1}\tilde{g}^T\phi'^T$, $\hat{u}^TR\hat{u}-u^{*T}Ru^* = \tilde{u}^TR\tilde{u}-2\tilde{u}^TRu^*$ is used and $w \in \mathbb{R}^{n_c}$ is defined as

$$w = \phi'(\bar{x})g(x,\hat{\theta})\hat{u}. \tag{24}$$

F. CRITIC NN WEIGHT UPDATE LAW

Based on the stability analysis and noticing that $\tilde{W}_c$ appears linearly in δ, a recursive least squares update law can be derived as

$$\dot{\hat{W}}_c = \text{proj}\left(-\gamma_c\Lambda\Omega\delta\right) \tag{25}$$

where $\Omega = \frac{w}{1+\nu w^T\Lambda w} \in \mathbb{R}$, $\nu \in \mathbb{R}$ and $\gamma_c \in \mathbb{R}$ are constant gains, $\text{proj}(\cdot)$ is a smooth projection operator, $\Lambda(t) = (\int_0^t w(\tau)w(\tau)^Td\tau)^{-1} \in \mathbb{R}^{n_c\times n_c}$ is a symmetric estimation gain matrix, which is computed using

$$\dot{\Lambda} = -\gamma_c\Lambda\frac{ww^T}{1+\nu w^T\Lambda w}\Lambda, \qquad \Lambda(0) = \beta_2 I \tag{26}$$

where $\beta_2 \in \mathbb{R}^+$. The covariance $\Lambda(t)$ is reset at time instances t^+ when $\Lambda \leq \beta_1 I$, for $\beta_1 \in \mathbb{R}$ by selecting $\Lambda(t^+) = \Lambda(0)$. The resetting ensures that the covariance remains positive definite for all time $t > 0$ [18]. From (26), since $\dot{\Lambda} \leq 0$, the covariance can be upper and lower bounded as $\beta_1 I \leq \Lambda \leq \beta_2 I$.

Assumption 6:
The normalized critic regressor $\xi = \frac{w}{\sqrt{1+\nu w^T\Lambda w}}$ is bounded and is persistently exciting (PE), i.e.,

$$\mu_a I \leq \int_{t_0}^{t_0+T} \xi(\tau)\xi^T(\tau)d\tau \leq \mu_b I, \quad \forall t_0 > 0 \tag{27}$$

where μ_a, μ_b, T are positive constants [18].

Consider the error in the critic weights $\tilde{W}_c = W_c - \hat{W}_c$. Taking derivative of $\tilde{W}_c$ results in

$$\begin{aligned}\dot{\tilde{W}}_c = &-\gamma_c\Lambda\xi\xi^T\tilde{W}_c + \gamma_c\Lambda\Omega\left(-W_c^T\phi'\tilde{g}\tilde{u} + \frac{1}{4}\tilde{W}_a^Tg_\phi\tilde{W}_a\right.\\ &-\frac{1}{2}\tilde{W}_a^Tg_\phi W_c - \epsilon_c'gu^* - \frac{1}{4}\epsilon_c'g_r\epsilon_c'^T - \frac{1}{2}\epsilon_c'^Tg_r\phi'^TW_c\\ &\left.+\frac{1}{4}\hat{W}_a^T\tilde{g}_\phi\hat{W}_a - \frac{1}{2}\hat{W}_a^T\tilde{g}_{a\phi}\hat{W}_a\right)\end{aligned} \tag{28}$$

Under Assumption 6, a nominal system formed using first term of (28) is globally exponentially stable [18], [52], which according to converse Lyapunov Theorem induces a Lyapunov function $V_{wc}(t,\tilde{W}_c)$ with following properties

$$\begin{aligned}\gamma_1\|\tilde{W}_c\|^2 \leq V_{wc}(t,\tilde{W}_c) \leq \gamma_2\|\tilde{W}_c\|^2\\ \frac{\partial V_{wc}}{\partial t} + \frac{\partial V_{wc}}{\partial \tilde{W}_{wc}}(-\gamma_c\Lambda\xi\xi^T\tilde{W}_c) \leq -\eta_1\|\tilde{W}_c\|^2\\ \|\frac{\partial V_{wc}}{\partial \tilde{W}_c}\| \leq \bar{\gamma}\|\tilde{W}_c\|\end{aligned} \tag{29}$$

where $\gamma_1, \gamma_2, \eta_1, \bar{\gamma} \in \mathbb{R}^+$.

G. ACTOR NN WEIGHT UPDATE LAW

The least-squares gradient-based update law for the actor NN is derived using the squared Bellman error $E_a = \delta^2$. Computing the gradient of E_a and setting it to zero, results in the following update law for $\hat{W}_a$

$$\dot{\hat{W}}_a = \text{proj}\left(-\frac{\gamma_a\hat{g}_\phi(\hat{W}_a-\hat{W}_c)\delta}{\sqrt{1+w^Tw}} - \gamma_{a2}(\hat{W}_a-\hat{W}_c)\right) \tag{30}$$

where $\hat{g}_\phi = \phi'g(x,\hat{\theta})R^{-1}g^T(x,\hat{\theta})\phi'^T$, γ_a and γ_{a2} are constant gains. For the stability analysis presented in next section, following bounds are defined

$$\begin{aligned}\|\frac{1}{4}\epsilon_c'g_r\epsilon_c'^T + \frac{1}{4}W_c^Tg_\phi W_c + \frac{1}{2}\epsilon_c'g_r\phi'^TW_c\\ -\frac{1}{2}\epsilon_c'g_{ra}\phi'^T\hat{W}_a - \frac{1}{2}W_c^T\hat{g}_\phi\hat{W}_a + \tilde{W}_c^T\phi'\tilde{g}\hat{u}\| \leq \kappa_1\end{aligned} \tag{31}$$

$$\|\frac{1}{2}W_c^Tg_\phi\| \leq \kappa_2, \tag{32}$$

$$\|\frac{1}{4}\tilde{W}_a^Tg_\phi\tilde{W}_a\| \leq \kappa_3, \quad \overline{w} = \left\|\hat{g}_\phi(\hat{W}_a-\hat{W}_c)\right\| \tag{33}$$

$$\begin{aligned}\|-W_c^T\phi'\tilde{g}\tilde{u} - \epsilon_c'gu^* - \frac{1}{4}\epsilon_c'g_r\epsilon_c'^T - \frac{1}{2}\epsilon_c'^Tg_r\phi'^TW_c\\ +\frac{1}{4}\hat{W}_a^T\tilde{g}_\phi\hat{W}_a - \frac{1}{2}\hat{W}_a^T\tilde{g}_{a\phi}\hat{W}_a\| \leq \kappa_4\end{aligned} \tag{34}$$

where $g_{ra} = gR^{-1}\hat{g}^T$, κ_1, κ_2, κ_3, κ_4, and $\bar{\omega}$ are positive constants.

IV. STABILITY ANALYSIS

Theorem 1:
Given that the Assumptions 1-6 hold and the following sufficient condition is satisfied

$$\gamma_{a2} - \gamma_a\kappa_4\overline{w} > 0,^{1} \tag{35}$$

the actor-critic controller (14) along with the model parameter update law in (16) and critic and actor weight update laws in (25)-(26), (30) guarantee that the signals $\bar{x}(t)$, $\tilde{\theta}(t)$, $\tilde{W}_a(t)$, $\tilde{W}_c(t)$, $\tilde{\lambda}$ are uniformly ultimately bounded.

Proof:
Consider domains $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Z} \subset \mathbb{R}^p$ and a positive definite continuously differentiable Lyapunov function $V : \mathcal{X}\times\mathbb{R}^{n_c}\times\mathbb{R}^{n_a}\times\mathcal{Z}\times\mathbb{R}^{2p}\times[0,\infty) \to \mathbb{R}^+$

$$V_z = V^*(t) + V_{wc}(t,\tilde{W}_c) + \frac{1}{2}\tilde{W}_a^T\tilde{W}_a + \frac{1}{2}\tilde{\theta}^TP^{-1}\tilde{\theta} + \frac{1}{2}\tilde{\lambda}^T\Gamma\tilde{\lambda}$$

where $V^*(t)$ is the optimal value function, V_{wc} is a Lyapunov function defined in (29), $\tilde{\lambda} = \lambda - \lambda^*$ for $\lambda^* > 0$ is the optimal Lagrange parameter, $\Gamma = \text{blkdiag}(\Gamma_1, \Gamma_2)$ such that $\Gamma_{min}I \leq \Gamma \leq \Gamma_{max}I$. Since the optimal value function $V^*(t)$ is continuously differentiable and positive definite, there exists class-$\mathcal{K}$ functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ such that $\alpha_1(\|\bar{x}\|) \leq V^*(t) \leq \alpha_2(\|\bar{x}\|)$, $\forall\bar{x} \in \mathcal{B}_a \subset \mathcal{X}$. Let us define $z(t) = [\bar{x}(t)^T, \tilde{W}_c(t)^T, \tilde{W}_a(t)^T, \tilde{\theta}(t), \tilde{\lambda}(t)]^T \in$

[1] Since γ_{a2} and γ_a are weight update law gains and κ_4 and $\overline{\omega}$ are known constants, the sufficient gain condition in (35) can be easily satisfied.

$\mathcal{X} \times \mathcal{Z} \times \mathbb{R}^{n_c+n_a+2p}$. Based on the bounds on $V^*(t)$, the following bounds can be derived

$$\alpha_3(\|z\|) \leq V_z(\bar{x}, \tilde{W}_c, \tilde{W}_a, \tilde{\theta}, \tilde{\lambda}, t) \leq \alpha_4(\|z\|) \quad \forall \bar{x} \in \mathcal{B}_a \tag{36}$$

where $\alpha_3(\cdot)$ and $\alpha_4(\cdot)$ are class $\mathcal{K}$-functions. Taking the time derivative of the Lyapunov function and using $\dot{\tilde{W}}_a = -\dot{\hat{W}}_a$ results in

$$\dot{V}_z = V_{\bar{x}}^* g(x,\theta)\hat{u} - \tilde{W}_a^T \dot{\hat{W}}_a + \dot{V}_{wc} + \tilde{\theta}^T P^{-1} \dot{\tilde{\theta}} + \tilde{\lambda}^T \Gamma \dot{\tilde{\lambda}} \tag{37}$$

Adding and subtracting $V_{\bar{x}}^* g(x,\theta)u^*$, utilizing $V_{\bar{x}}^* g(x,\theta) = -2u^{*T}R$ and $V_{\bar{x}}^* g(x,\theta)u^* = -Q(\bar{x}) - u^{*T}Ru^*$, and substituting u^* from (13), $\hat{u}$ from (14), the NN form of the V^* from (13) and (19) results in

$$\begin{aligned}
\dot{V}_z = &-Q(\bar{x}) + \frac{1}{4}\epsilon_c' g_r \epsilon_c'^T + \frac{1}{4} W_c^T g_\phi W_c + \frac{1}{2}\epsilon_c' g_r \phi'^T W_c \\
&- \frac{1}{2} W_c^T \phi' g R^{-1} \hat{g}^T \phi'^T \hat{W}_a - \frac{1}{2}\epsilon_c' g R^{-1} \hat{g}^T \phi'^T \hat{W}_a + \dot{V}_{wc} \\
&- \tilde{W}_a^T \dot{\hat{W}}_a - \tilde{\theta}^T Y^T (\hat{u}^T \otimes \hat{W}_c^T \phi')^T - k_{cl}\tilde{\theta} \sum_{k=1}^{m} \mathcal{Y}_k^T \mathcal{Y}_k \tilde{\theta} \\
&+ \tilde{\theta}^T \sum_{j=1}^{2} \text{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j + \sum_{j=1}^{2} \tilde{\lambda}_j^T [-k_1 \Gamma \lambda_j + c_j]_{\lambda_j}^+
\end{aligned} \tag{38}$$

Since $\hat{g}$ term in the first term of second line is not known, and depends on the parameter estimate $\hat{\theta}$, algebraic manipulations as shown next can be performed to obtain a term that is in terms of parameter estimation error $\tilde{\theta}$. This forms the basis of adaptive update law design in the proposed AAC structure, which is different compared to the designs that use identifier to identify the dynamics. Adding and subtracting $W_c^T \phi' \hat{g} \hat{u}$ and $\hat{W}_c^T \phi' \tilde{g} \hat{u}$, and using the fact that $\hat{W}_c^T \phi' \tilde{g} \hat{u} = (\hat{u}^T \otimes \hat{W}_c^T \phi') Y \tilde{\theta}$, the following expression is obtained

$$\begin{aligned}
\dot{V}_z = &-Q(\bar{x}) + \frac{1}{4}\epsilon_c' g_r \epsilon_c'^T + \frac{1}{4} W_c^T g_\phi W_c + \frac{1}{2}\epsilon_c' g_r \phi'^T W_c \\
&- \frac{1}{2}\epsilon_c' g_{ra} \phi'^T \hat{W}_a + \tilde{W}_c^T \phi' \tilde{g} \hat{u} + (\hat{u}^T \otimes \hat{W}_c^T \phi') Y \tilde{\theta} \\
&- \frac{1}{2} W_c^T \hat{g}_\phi \hat{W}_a + \dot{V}_{wc} - \tilde{W}_a^T \dot{\hat{W}}_a - \tilde{\theta}^T Y^T (\hat{u}^T \otimes \hat{W}_c^T \phi')^T \\
&- k_{cl}\tilde{\theta} \sum_{k=1}^{m} \mathcal{Y}_k^T \mathcal{Y}_k \tilde{\theta} + \tilde{\theta}^T \sum_{j=1}^{2} \text{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j \\
&+ \sum_{j=1}^{2} \tilde{\lambda}_j^T (-k_1 \Gamma \lambda_j + c_j) + \sum_{j=1}^{2} \tilde{\lambda}_j^T ([-k_1 \Gamma \lambda_j + c_j]_{\lambda_j}^+ \\
&- (-k_1 \Gamma \lambda_j + c_j)).
\end{aligned} \tag{39}$$

Consider for each $j = 1, 2$, the term $T_i = (\lambda - \lambda^*)_i ([-k_1 \Gamma \lambda + c_j]_\lambda^+ - (-k_1 \Gamma \lambda + c_j))_i \leq 0$ for $i \in \{1, ..., p\}$. To show this, consider if $\lambda_i > 0$, then $T_i = 0$ and if $\lambda_i = 0$, then $(\lambda - \lambda^*)_i \leq 0$ and $([-k_1 \Gamma \lambda + c_j]_\lambda^+ - (-k_1 \Gamma \lambda + c_j))_i \geq 0$ which implies that $T_i \leq 0$. The following bound on $\dot{V}_z$ can now be obtained

$$\begin{aligned}
\dot{V}_z \leq &-Q(\bar{x}) + \frac{1}{4}\epsilon_c' g_r \epsilon_c'^T + \frac{1}{4} W_c^T g_\phi W_c + \frac{1}{2}\epsilon_c' g_r \phi'^T W_c \\
&- \frac{1}{2}\epsilon_c' g_{ra} \phi'^T \hat{W}_a + \tilde{W}_c^T \phi' \tilde{g} \hat{u} + (\hat{u}^T \otimes \hat{W}_c^T \phi') Y \tilde{\theta} \\
&- \frac{1}{2} W_c^T \hat{g}_\phi \hat{W}_a + \dot{V}_{wc} - \tilde{W}_a^T \dot{\hat{W}}_a + \Big[-\tilde{\theta}^T Y^T (\hat{u}^T \otimes \hat{W}_c^T \phi')^T \\
&- k_{cl}\tilde{\theta} \sum_{k=1}^{m} \mathcal{Y}_k^T \mathcal{Y}_k \tilde{\theta} + \tilde{\theta}^T \sum_{j=1}^{2} \text{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j \Big] \\
&- k_1 \Gamma \tilde{\lambda}^T \lambda + \sum_{j=1}^{2} \tilde{\lambda}_j^T c_j
\end{aligned} \tag{40}$$

To further develop the stability analysis, consider a convex-concave Lagrangian function

$$\begin{aligned}
L(\hat{\theta}, \lambda) = &-Q(x) + (\hat{u}^T \otimes \hat{W}_c^T \phi') Y \tilde{\theta} \\
&+ \frac{k_{cl}}{2} \tilde{\theta}^T \sum_{k=1}^{m} \mathcal{Y}_k^T \mathcal{Y}_k \tilde{\theta} + \sum_{j=1}^{2} \lambda_j^T c_j
\end{aligned} \tag{41}$$

Using $-\nabla_{\hat{\theta}} L$ and multiplying by a PD learning rate P yields the parameter update law in (16), which when substituted in $\dot{V}_z$ yields the terms in $[\cdot]$ in (40). The first term in $[\cdot]$ gets canceled with the term $(\hat{u}^T \otimes \hat{W}_c^T \phi') Y \tilde{\theta}$. Next, the third term in $[\cdot]$ and last term of (40) are shown to be ≤ 0 using convex-concave properties of Lagrangian function $L(\hat{\theta}, \lambda)$.

Since L is concave in λ, the following inequalities can be developed using the properties of concave functions (see, [41], [43], [44]). Consider a function $F_{aj}(\hat{\theta}, \lambda) = \sum_{j=1}^{2} \lambda_j^T c_j$ such that $\nabla_\lambda L = \nabla_\lambda F_{aj}$

$$\begin{aligned}
(\lambda_j - \lambda_j^*)^T c_j &= (\lambda_j - \lambda_j^*)^T (\nabla_{\lambda_j} (F_{aj})) \\
&\leq F_{aj}(\hat{\theta}, \lambda) - F_{aj}(\hat{\theta}, \lambda^*) \leq 0
\end{aligned} \tag{42}$$

Consider the term $\tilde{\theta}^T \sum_{j=1}^{2} \text{diag}(\lambda_j) \nabla_{\hat{\theta}} c_j = \sum_{j=1}^{2} \sum_{i=1}^{p} (\theta_i - \hat{\theta}_i) \lambda_{ji} \nabla_{\hat{\theta}} c_{ji}$. Using the fact that c_{ji} is a convex function in $\hat{\theta}(t)$ and using the properties of a convex function, the following bound can be established

$$(\theta_i - \hat{\theta}_i) \lambda_{ji} \nabla_{\hat{\theta}} c_{ji} \leq c_{ji}(\theta) - c_{ji}(\hat{\theta}) \leq 0 \tag{43}$$

To further simplify the $\dot{V}_z$ expression in (40), consider $-\tilde{W}_a^T \dot{\hat{W}}_a$ after substituting actor weight update law from (30)

$$\begin{aligned}
-\tilde{W}_a^T \dot{\hat{W}}_a = &C_1 \tilde{W}_a^T \left(\hat{g}_\phi (\hat{W}_a - \hat{W}_c) \right) \delta \\
&- \gamma_{a2} \tilde{W}_a^T \tilde{W}_a + \gamma_{a2} \tilde{W}_a^T \tilde{W}_c
\end{aligned} \tag{44}$$

where $C_1 = \frac{\gamma_a}{\sqrt{1+w^T w}}$. Substituting (44) and the bounds on Lyapunov function V_{wc} from (29) into (40), utilizing bounds in (42) and (43), adding and subtracting terms $\frac{1}{2} W_c^T g_\phi W_c$, $\dot{V}_z$ can be written as

$$\begin{aligned}
\dot{V}_z \leq &-Q(\bar{x}) - \eta_1 \|\tilde{W}_c\|^2 - \gamma_{a2} \|\tilde{W}_a\|^2 + \frac{1}{4}\epsilon_c' g_r \epsilon_c'^T \\
&+ \frac{1}{4} W_c^T g_\phi W_c + \frac{1}{2}\epsilon_c' g_r \phi'^T W_c - \frac{1}{2}\epsilon_c' g_{ra} \phi'^T \hat{W}_a \\
&- \frac{1}{2} W_c^T \hat{g}_\phi \hat{W}_a + \tilde{W}_c^T \phi' \tilde{g} \hat{u} + \zeta \|\tilde{W}_c\| \Bigg(- W_c^T \phi' \tilde{g} \hat{u}
\end{aligned}$$

$$+\frac{1}{4}\tilde{W}_a^T g_\phi \tilde{W}_a - \frac{1}{2}\tilde{W}_a^T g_\phi W_c - \epsilon_c' g u^* - \frac{1}{4}\epsilon_c' g_r \epsilon_c'^T - \frac{1}{2}\epsilon_c'^T g_r \phi'^T W_c + \frac{1}{4}\hat{W}_a^T \tilde{g}_\phi \hat{W}_a - \frac{1}{2}\hat{W}_a^T \tilde{g}_{a\phi}\hat{W}_a\Big) - k_{cl}\tilde{\theta}\sum_{k=1}^{m}\mathcal{Y}_k^T\mathcal{Y}_k\tilde{\theta} + \gamma_{a2}\tilde{W}_a^T(W_c - \hat{W}_c) + C_1\tilde{W}_a^T\left(\hat{g}_\phi(\hat{W}_a - \hat{W}_c)\right)\delta - k_1\tilde{\lambda}^T\Gamma(\tilde{\lambda}-\lambda^*) \quad (45)$$

where $\zeta = \frac{\bar{\gamma}\beta_2}{\nu\beta_1}$. Utilizing bounds in (31)-(34), using $|C_1| \leq \gamma_a$, and completing the squares, $\dot{V}_z$ can be upper bounded as

$$\begin{aligned}\dot{V}_z \leq& -Q(\bar{x}) - (1-\theta_1)\eta_1\|\tilde{W}_c\|^2 \\ &- (1-\theta_1)\left(\gamma_{a2} - \gamma_a\kappa_3\overline{w}\right)\|\tilde{W}_a\|^2 - \sigma_1\|\tilde{\theta}\|^2 \\ &+ \kappa_1 + (C_1(\kappa_4+\kappa_2)\overline{w})\|\tilde{W}_a\| \\ &+ \zeta\left(\kappa_4+\kappa_3\right)\|\tilde{W}_c\| - k_2\Gamma_{min}\|\tilde{\lambda}\|^2 \\ &- \left(\sqrt{k_3\Gamma_{min}}\|\tilde{\lambda}\| - \frac{k_1\Gamma_{max}\|\lambda^*\|}{2\sqrt{k_3\Gamma_{min}}}\right)^2 + \frac{k_1^2\Gamma_{max}^2\|\lambda\|^{*2}}{4k_2\Gamma_{min}}\end{aligned} \quad (46)$$

where $\eta_2 = \gamma_{a2} - \gamma_a\kappa_3\overline{w}$, $\sigma_1 = k_{cl}\lambda_{\min}(\sum_{k=1}^{m}\mathcal{Y}_k^T\mathcal{Y}_k)$, $1-\theta_1 > 0$, $k_1 = k_2 + k_3$. The final bound on $\dot{V}_z$ can derived as

$$\begin{aligned}\dot{V}_z \leq& -Q(\bar{x}) - (1-\theta_1-\theta_2)\eta_1\|\tilde{W}_c\|^2 \\ &- (1-\theta_1-\theta_2)\eta_2\|\tilde{W}_a\|^2 - \sigma_1\|\tilde{\theta}\|^2 - k_2\Gamma_{min}\|\tilde{\lambda}\|^2 + \kappa_1 \\ &+ \frac{\zeta^2(\kappa_4+\kappa_3)^2}{4(1-\theta_1-\theta_2)\eta_1} + \frac{(C_1(\kappa_4+\kappa_2)\overline{w})^2}{4(1-\theta_1-\theta_2)\eta_2} + \frac{k_1^2\Gamma_{max}^2\|\lambda\|^{*2}}{4k_2\Gamma_{min}}\end{aligned} \quad (47)$$

where $1-\theta_1-\theta_2 > 0$. Since $Q(\bar{x})$ is positive definite, Lemma 4.3 of [53] can be utilized to derive

$$\begin{aligned}\alpha_5(\|z\|) \leq& Q + (1-\theta_1-\theta_2)\eta_1\|\tilde{W}_c\|^2 + \sigma_1\|\tilde{\theta}\|^2 \\ &+ (1-\theta_1-\theta_2)\eta_2\|\tilde{W}_a\|^2 + k_2\Gamma_{min}\|\tilde{\lambda}\|^2 \leq \alpha_6(\|z\|)\end{aligned} \quad (48)$$

where $\alpha_5(\cdot)$ and $\alpha_6(\cdot)$ are class-$\mathcal{K}$ functions. Using (48), the expression (47) can be upper bounded as

$$\begin{aligned}\dot{V}_z \leq& -\alpha_5(\|z\|) + \kappa_1 + \frac{\zeta^2(\kappa_4+\kappa_3)^2}{4(1-\theta_1-\theta_2)\eta_1} \\ &+ \frac{(C_1(\kappa_4+\kappa_2)\overline{w})^2}{4(1-\theta_1-\theta_2)\eta_2} + \frac{k_1^2\Gamma_{max}^2\|\lambda\|^{*2}}{4k_2\Gamma_{min}}\end{aligned} \quad (49)$$

which proves that $\dot{V}_z$ is always negative whenever $z(t)$ is outside the compact set

$$\begin{aligned}\bar{\Omega} = \{z : \|z\| \leq \alpha_5^{-1}(&\kappa_1 + \frac{\zeta^2(\kappa_4+\kappa_3)^2}{4(1-\theta_1-\theta_2)\eta_1} \\ &+ \frac{(C_1(\kappa_4+\kappa_2)\overline{w})^2}{4(1-\theta_1-\theta_2)\eta_2} + \frac{k_1^2\Gamma_{max}^2\|\lambda^*\|^2}{4k_2\Gamma_{min}})\}.\end{aligned} \quad (50)$$

Invoking Theorem 4.18 of [53] $\|z\|$ is uniformly ultimately bounded (UUB). ■

Remark 7:
The ultimate bound on $\|z\|$ can be reduced by appropriately choosing the gains γ_{a2}, γ_c, and increasing the number of neurons in the NN, which reduces the function approximation error of the actor and critic NN.

Remark 8:
When the Assumption 5 is not satisfied, then the term $\sigma_1\|\tilde{\theta}\|^2$ will not be present in $\dot{V}_z$ expression in (47). This is the time period when the history stack data is collected and $\mathcal{N}$ is not full rank. To yield a UUB stability result, a sigma modification term is added to the model parameter update law (16) [24], which yields the parameter update law

$$\begin{aligned}\dot{\hat{\theta}} =& PY^T(\hat{u}^T \otimes \hat{W}_c^T\phi')^T - \sigma_2\hat{\theta} - \sum_{j=1}^{2} P\,\mathrm{diag}(\lambda_j)\nabla_{\hat{\theta}}c_j; \\ \dot{\lambda}_j =& [-k_j\lambda_j + \Gamma_j c_j]_{\lambda_j}^+\end{aligned} \quad (51)$$

Remark 9:
Since the finite excitation condition of Assumption 5 can be verified in real-time, the time of switching between (51) and (16) can be computed.

Remark 10:
Due to switching of the model parameter update law from (51) to (16) based on the finite excitation condition, the error system of $z(t)$ is a switched system. The Lyapunov function in (36) is a common Lyapunov function for the error system.

Remark 11:
The persistence of excitation condition of Assumption 6 may be ensured by adding an exploration noise of distinct frequencies in control input signal $\hat{u}$. With exploration signal of small magnitude, the system signals still remain UUB [18], [22].

Remark 12:
For systems with larger state dimensions, dimensionality reduction methods such as Koopman operators [54] can be used to formulate low dimensional system on which the proposed AAC controller can be used. This can potentially reduce number of unknown parameters of system model, actor and critic NNs.

V. SIMULATION STUDIES

Simulations are carried out to test the performance of the RL-based controller on IBVS and WMR systems.

A. OPTIMAL IBVS CONTROLLER

The dynamics of the IBVS system are given by the system of the form

$$\dot{x} = g(x,\theta)u \quad (52)$$

where $u = [v,\omega]^T$ is the camera velocity consisting of linear velocity $v(t) \in \mathbb{R}^3$ and angular velocity $\omega(t) \in \mathbb{R}^3$. The state $x \in \mathbb{R}^8$ represents the normalized projected coordinates for four points, which can be obtained using image pixels and the internal camera calibration matrix. The Jacobian matrix

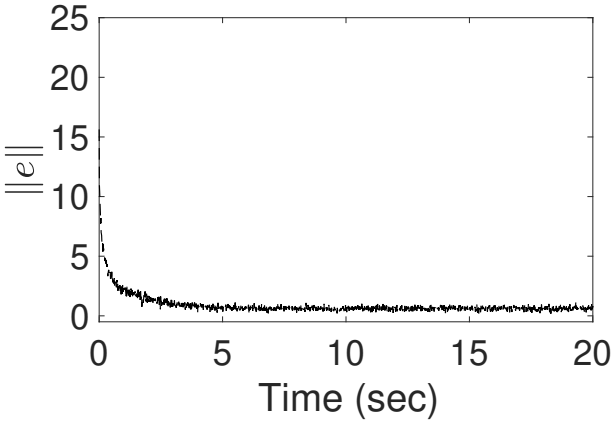

FIGURE 1. Regulation error norm of IBVS-RL controller.

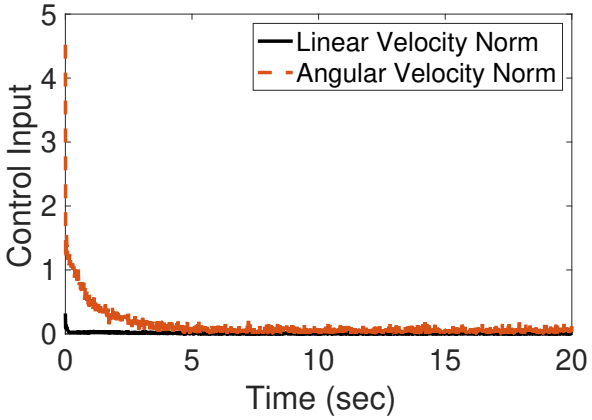

FIGURE 2. Control velocities generated by IBVS-RL controller.

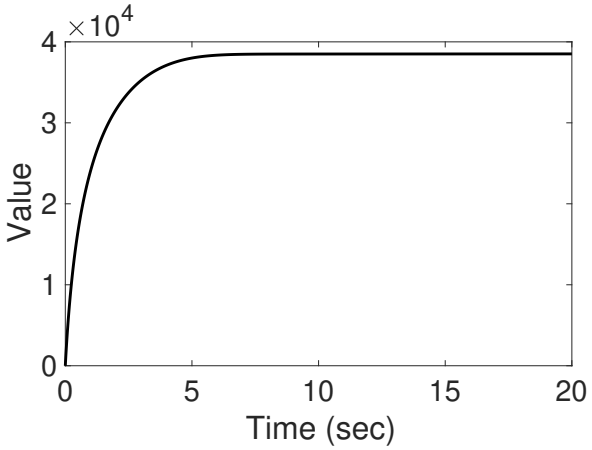

FIGURE 3. Value function of IBVS-RL controller.

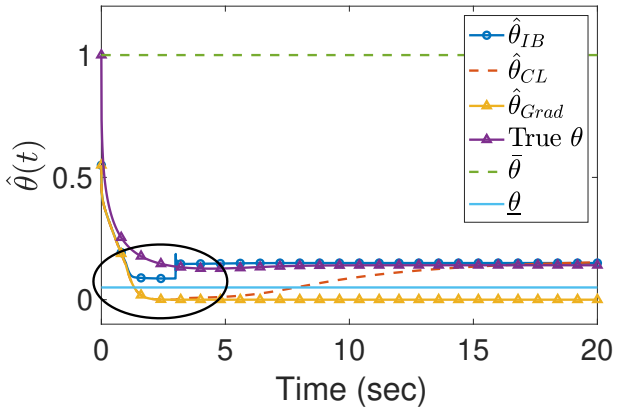

FIGURE 4. Parameter estimates for the proposed RL controller using inverse Barrier (IB), concurrent learning (CL) and gradient update law in [46].

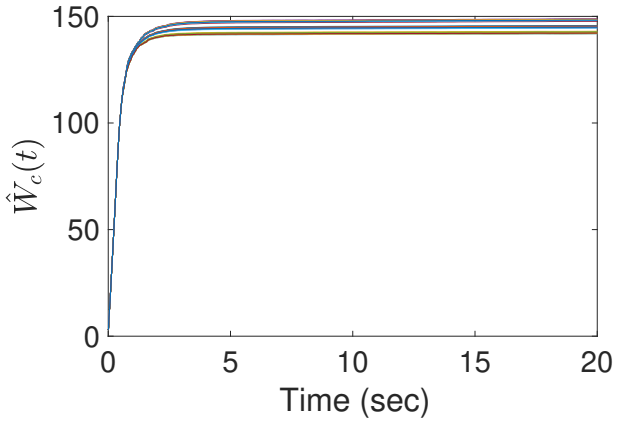

FIGURE 5. Critic weights of IBVS-RL controller.

$g_i \in \mathbb{R}^{2\times 6}$ for a feature point is

$$g_i = \begin{bmatrix} \theta & 0 & -x_1\theta & -x_1x_2 & 1+x_1^2 & -x_2 \\ 0 & \theta & -x_2\theta & -1-x_2^2 & x_1x_2 & x_1 \end{bmatrix} \tag{53}$$

For four feature points the Jacobian matrix is given by $g = [g_1^T\ g_2^T\ g_3^T\ g_4^T]^T \in \mathbb{R}^{8\times 6}$. The parameter θ is an inverse depth of the feature point, which is unknown and varies between $(0, 1]$ with time. In the IBVS implementation, θ is approximated as a constant parameter which is a common practice for IBVS [55]. The presence of the unknown parameter adds uncertainty to the system dynamics. Gaussian noise with zero mean and variance of 0.05 is added to the state measurements.

For testing the performance of the proposed controller, four points were selected on the image plane with the initial pixel values of $[50\ \ 50; 100\ \ 50; 100\ \ 100; 50\ \ 100]^T$ and the desired pixel values of $[825\ 790; 860\ 825; 825\ 860; 790\ 825]^T$. The controller parameters are selected as $Q = 800\mathbb{I}_{8\times 8}$ and $R = 60\text{blkdiag}\{100\mathbb{I}_{3\times 3}, 10\mathbb{I}_{3\times 3}\}$. The basis functions for approximating the value function V are selected to be second-order polynomial combinations of elements of

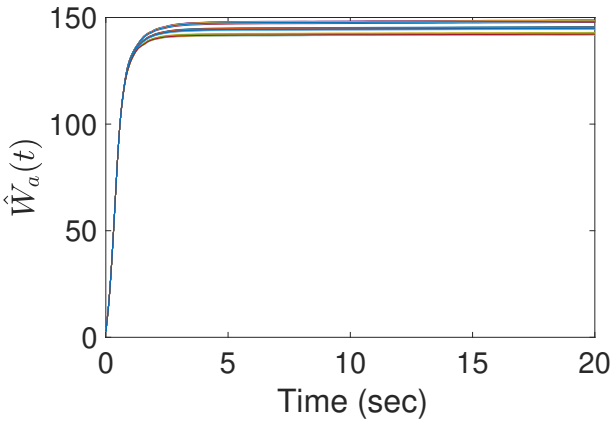

FIGURE 6. Actor weights of IBVS-RL controller.

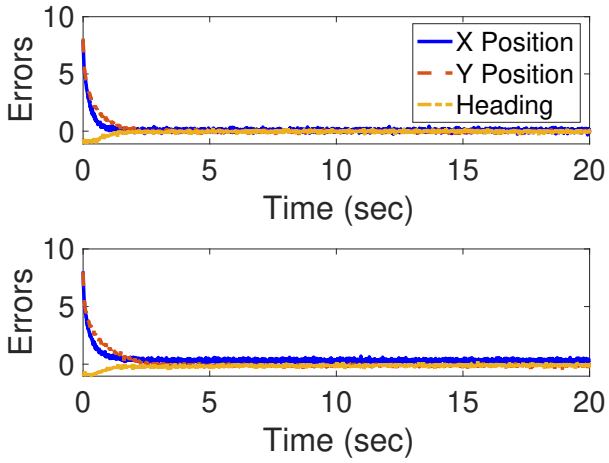

FIGURE 7. Regulation errors of WMR-RL controller in this paper (top plot) and using gradient update law from [46].

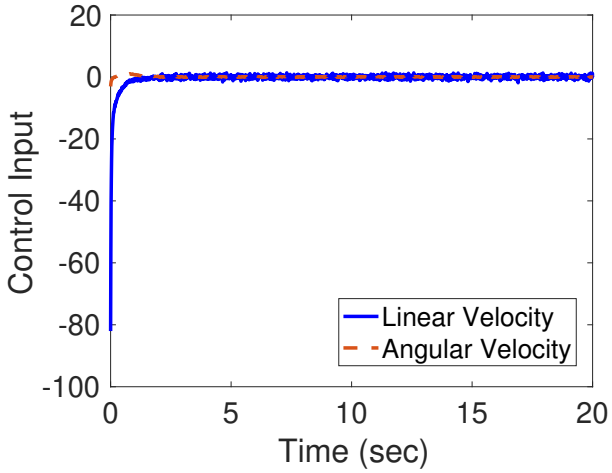

FIGURE 8. Control velocities generated by WMR-RL controller.

$\phi = [\bar{x}_1, \bar{x}_2, ..\bar{x}_n]^T$, the initial weights for the critic NN are set to $W_c(t_0) = 3\mathbb{I}_{36\times1}$. The parameters of critic NN are found by empirical tuning as $\gamma_c = 20$, $\Lambda(t_0) = 1$, and $\nu = 5$. The actor NN weights are initialized to $W_a(t_0) = 3.5\mathbb{I}_{36\times1}$. The parameters of actor NN are selected as $\gamma_a = 0.01$ and $\gamma_{a2} = 10.5$. The model parameter update law gains are selected as $\gamma_\theta = 0.005$ and $k_{cl} = 10000$. Learning rates of $P = 0.0001$, $\Gamma_1 = \Gamma_2 = 0.001$ and $k_1 = k_2 = 0.001$ are used for the parameter estimate and Lagrange multiplier update law. The upper and lower bounds on θ are set to 0.05 and 1, respectively. To ensure the PE condition of Assumption 6, a probing signal is added to the controller of magnitude $0.01e^{-t}(sin^2(\frac{\pi t}{5})cos(\frac{\pi t}{2})+sin^2(\frac{2\pi t}{3})cos(0.1t)+sin^2(-1.2e^{-0.01t})cos(0.5t)+sin^5(e^{-0.1t}))$.

The results of the simulation are shown in Figs. 1-6. The norm of the IBVS regulation errors is shown in Fig. 1. The norm of the linear and angular velocities generated by the proposed controller is shown in Fig. 2. The control velocities are generated in an optimal manner based on the minimization of the value function whose value is shown in Fig. 3. The model parameter weight, which is an inverse depth, in this case, approximated as a constant parameter, is shown in Fig. 4. The parameter is time-varying; hence, the true model parameter is not exactly identified until it is settled nearly at a constant value close to 4 seconds. The parameter estimated by the proposed parameter update law does not go beyond the specified upper and lower bounds, whereas the CL-based and gradient-based parameter update laws violate the prescribed parameter bounds, going very close to 0 between 2-3 seconds, leading to very small values of linear velocities. Both the bounds eventually converge to the constant values, the CL-parameter update law converges to the true parameter eventually. The critic and actor NN weights are shown in Figs. 5-6, which shows that the weights remain bounded and converge to constant values. Moreover, the actor weights converge to the critic weights. For the IBVS controller, since the parameter is time-varying but varies between 0 and 1, exact parameter estimation may not be achieved before it is settled to its steady state value; however, the proposed controller drives the regulation error to a small ball around zero in the presence of uncertainties in the image Jacobian matrix showing robustness against variations in parameter estimation.

B. OPTIMAL WHEELED MOBILE ROBOT REGULATION

The kinematics of WMR can be written in the following form

$$\dot{x} = g(x, \theta)u \tag{54}$$

where $x = [X, Y, \psi]^T \in \mathbb{R}^3$ is the 2D position and orientation state, $u = [v, \omega]^T \in \mathbb{R}^2$ are the linear and angular velocities. The Jacobian matrix g can be written as $g = \begin{bmatrix} acos(\psi) & asin(\psi) & 0 \\ 0 & 0 & b \end{bmatrix}^T$, which contains uncertain parameters a and b related to wheel diameter and distance between the wheels [56]. The proposed RL controller is implemented for this dynamics by first formulating the Jacobian in a

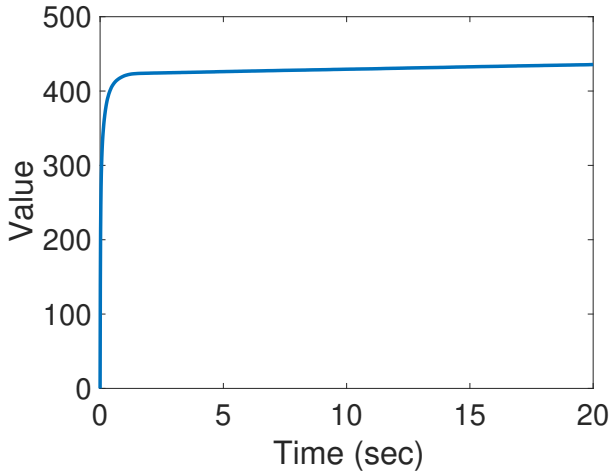

FIGURE 9. Value function for WMR-RL controller.

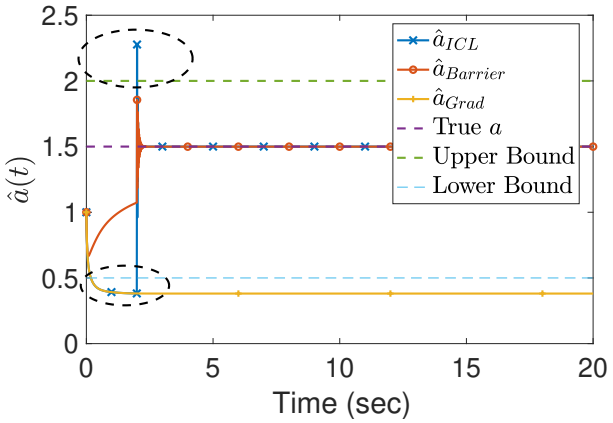

FIGURE 10. Parameter estimate of WMR-RL controller using inverse barrier (IB), concurrent learning (CL), and gradient-based parameter update law used in [46].

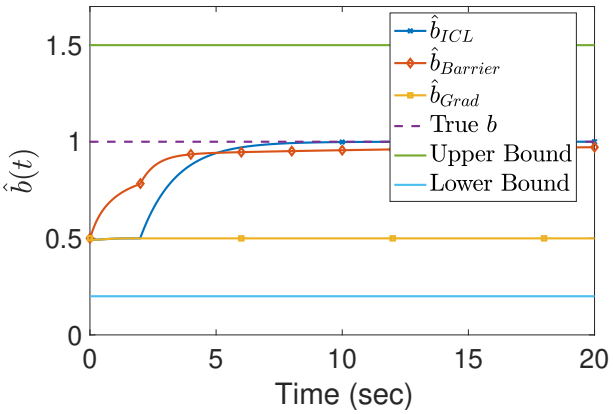

FIGURE 11. Parameter estimate of WMR-RL controller using inverse barrier (IB), concurrent learning (CL), and gradient-based parameter update law in [46].

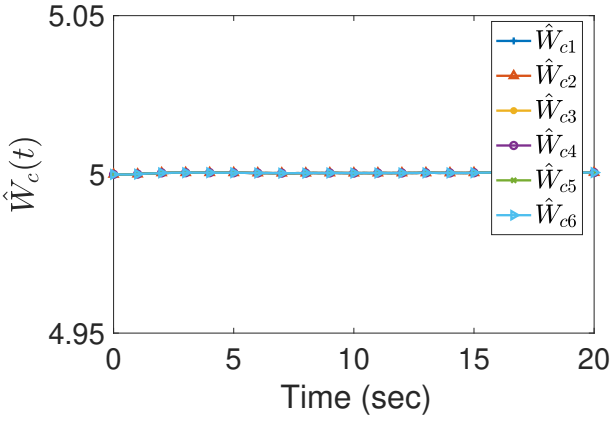

FIGURE 12. Critic weights of WMR-RL controller.

parametric form as

$$\text{vec}(g) = Y\theta, \tag{55}$$

where $\theta = [a, b]^T \in \mathbb{R}^2$ and $Y \in \mathbb{R}^{6\times 2}$ is

$$Y = \begin{bmatrix} cos(\psi) & sin(\psi) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}^T. \tag{56}$$

The control task is to regulate the WMR to a desired state of $x_d = [0, 0, \frac{\pi}{2}]^T$ from an initial state of $x(0) = [2, 2, \frac{\pi}{4}]^T$. Parameter values of the robot model are selected as $a = 1.5$, $b = 1$. Gaussian noise with zero mean and variance of 0.01 is added to the state measurements. The proposed adaptive-actor-critic controller is implemented using following parameters: $Q = 2\mathbb{I}_{3\times 3}$, $R = 1\mathbb{I}_{2\times 2}$. A polynomial basis functions are selected as $\phi = [\bar{x}_1^2, \bar{x}_2^2, \bar{x}_3^2, \bar{x}_1\bar{x}_2, \bar{x}_2\bar{x}_3, \bar{x}_1\bar{x}_3]^T$. The model parameter vector and actor-critic weights are initialized to $\theta(0) = [1, 0.5]^T$, $\hat{W}_c(0) = 5\mathbf{1}_{6\times 1}$ and $\hat{W}_a(0) = 10\mathbf{1}_{6\times 1}$. The controller gain parameters are found by empirical tuning as $\gamma_c = 0.01$, $\Lambda(t_0) = 500$, $\nu = 100$, $\gamma_a = 0.01$ and $\gamma_{a2} = 20$. The parameters of the adaptive update law are selected as $\gamma_\theta = 0.01$, $k_{cl} = 10$. The learning rates are selected as $P = 0.001\mathbb{I}_{2\times 2}$ and $\Gamma_1 = \Gamma_2 = 0.1\mathbb{I}_{2\times 2}$ and $k_1 = k_2 = 0.1$ for the Lagrange multiplier. There are two sets of Lagrange multiplier vectors, each corresponding to upper and lower bounds of θ, which are chosen as 2 and 0.5 for a and 1.5 and 0.2 for b. A probing signal similar to the previous simulation example is added to the control input.

The results are summarized in Figs. 7-13. From Fig. 7 it is seen that the position and heading angle states are regulated to the desired position and heading angle using linear and angular velocities shown in Fig. 8 using the proposed RL control policy. The bottom subplot of Fig. 7 also shows regulation errors when gradient-based parameter update law is used by the controller. It is observed that the position in X-direction converges to a slightly offset value compared to the controller developed using the parameter update law proposed in this paper. The control velocities are bounded and are generated in an optimal manner based on the minimization of the value function, which converges to a constant value in the steady state as seen from Fig.

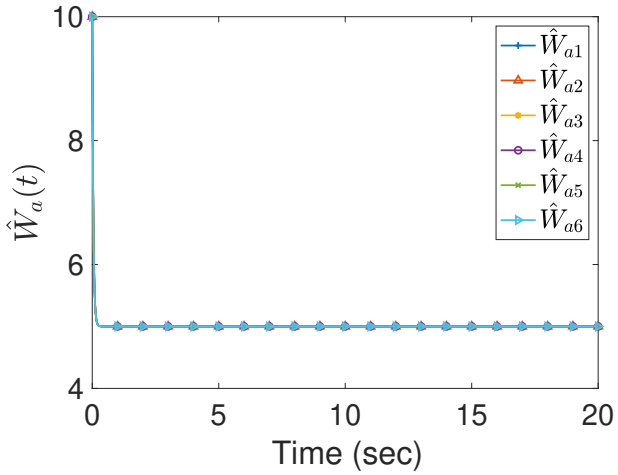

FIGURE 13. Actor weights of WMR-RL controller.

9. The WMR model parameters are estimated using a newly proposed CCL-based parameter estimation law with constraints that achieve the parameter convergence to their true values, as seen from Figs. 10 and 11. The parameter estimates stay within prescribed lower and upper bounds for the proposed model parameter weight update law, whereas for the CCL-based and gradient parameter update laws, the estimated model parameters violate the prescribed bounds as seen from Fig. 10-11. The actor and critic NN weights remain bounded and converge to constant values, as seen from Figs. 12 and 13. The actor weights converge to the critic weights.

C. OPTIMAL CONTROL OF LINEAR DRIFT-FREE SYSTEM

The performance of proposed AAC method is compared with that of LQR controller using a linear systems example. LQR provides an optimal solution for linear systems. The system dynamics is given below

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u \tag{57}$$

The following weight matrices are used to design the cost for both the controllers

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \tag{58}$$

The initial conditions are selected to be $x(t_0) = [1\ 1]^T$. For AAC controller the parameters are initialized to $\theta(t_0) = [1.5\ 0.2]^T$, $\lambda_1(t_0) = \lambda_2(t_0) = [5\ 5]^T$, $\hat{W}_c = 0.5\mathbf{1}_{3\times 1}$, $\hat{W}_a = 0.1\mathbf{1}_{3\times 1}$. The gains are selected as $\gamma_c = 30$, $\Lambda(t_0) = 5$, $\nu = 20$, $\gamma_a = 20$ and $\gamma_{a2} = 2$. For the adaptive parameter update law, the gains are $\gamma_\theta = 0.1$ and $k_{cl} = 10$, $P = 0.1\mathbf{I}_{2\times 2}$, $\Gamma_1 = \Gamma_2 = 0.1\mathbf{I}_{2\times 2}$, $k_1 = k_2 = 0.1$. The performance of our AAC method where the parameters of the B matrix are estimated using the constrained adaptive parameter update law and the optimal LQR controller are shown in Figs. 14-16. It can be seen that our method's performance is very close to the optimal performance obtained by the LQR. The value function is computed as an integral of the local cost $x^TQx + u^TRu$.

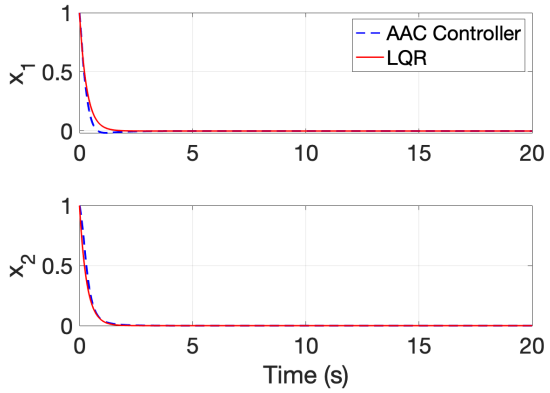

FIGURE 14. LQR state comparison.

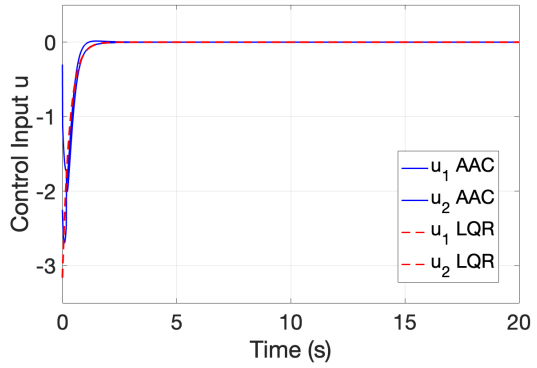

FIGURE 15. Control input comparison.

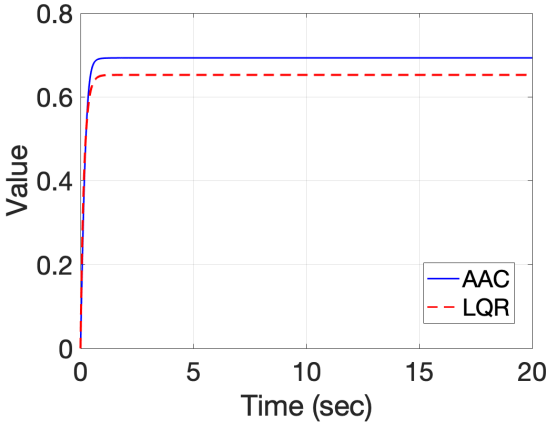

FIGURE 16. Value function comparison.

VI. CONCLUSION AND FUTURE DIRECTIONS

In this paper, a reinforcement learning-based policy is developed based on a continuous-time version of the PI architecture for drift-free nonlinear systems with an uncertain

g matrix. A CCL-based adaptive parameter update law is designed to estimate the model parameters so that they are bounded away from zero and an upper bound. Least squares-based update laws are used for actor and critic NN weights. The CCL-based parameter update law identifies the parameter using the LIP property of the dynamics and history data. Using Lyapunov analysis, it is shown that the signals of the closed-loop system are uniformly ultimately bounded. The simulation results on two examples show that the proposed controller can regulate the state to its desired value. In the case of WMR, the constant parameter vector is also identified by the CCL-based adaptive update law.

The controller developed in this paper considers LIP form of the systems dynamics. When the system dynamics is not in LIP form, e.g., dynamics represented using a 3-layer NN parametrization, a similar stability analysis presented in this paper can be developed to design RL controller for drift-free systems. This will be studied as a part of future work.

REFERENCES

[1] W. E. Dixon, D. M. Dawson, E. Zergeroglu, and A. Behal, *Nonlinear control of wheeled mobile robots*. Springer, 2001, vol. 175.

[2] V. Guthikonda and A. P. Dani, "Shape servoing of deformable objects using model estimation and Barrier Lyapunov function," *IEEE/ASME Transactions on Mechatronics*, p. to appear, 2024.

[3] N. E. Leonard and P. S. Krishnaprasad, "Motion control of drift-free, left-invariant systems on lie groups," *IEEE Transactions on Automatic control*, vol. 40, no. 9, pp. 1539–1554, 1995.

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

[5] Y. Jiang and Z.-P. Jiang, *Robust adaptive dynamic programming*. John Wiley & Sons, 2017.

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[7] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 997–1007, 1997.

[8] K. G. Vamvoudakis, P. J. Antsaklis, W. E. Dixon, J. P. Hespanha, F. L. Lewis, H. Modares, and B. Kiumarsi, "Autonomy and machine intelligence in complex systems: A tutorial," in *2015 American Control Conference (ACC)*, 2015, pp. 5062–5079.

[9] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[10] Z. Chen and S. Jagannathan, "Generalized hamilton–jacobi–bellman formulation-based neural network control of affine nonlinear discrete-time systems," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 90–106, 2008.

[11] P. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of intelligent control*, 1992.

[12] K. Doya, "Reinforcement learning in continuous time and space," *Neural Computation*, vol. 12, no. 1, pp. 219–245, 2000.

[13] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized hamilton-jacobi-bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[14] M. Abu-Khalaf and F. L. 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.

[15] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[16] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

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

[18] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, 2013.

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

[20] P. Deptula, Z. I. Bell, E. A. Doucette, J. W. Curtis, and W. E. Dixon, "Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with control effectiveness faults," *Automatica*, vol. 116, p. 108922, 2020.

[21] Y. Yang, W. Gao, H. Modares, and C.-Z. Xu, "Robust actor–critic learning for continuous-time nonlinear systems with unmodeled dynamics," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 6, pp. 2101–2112, 2021.

[22] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513–1525, 2013.

[23] D. Wang, D. Liu, Q. Zhang, and D. Zhao, "Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 11, pp. 1544–1555, 2015.

[24] P. A. Ioannou and J. Sun, *Robust adaptive control*. PTR Prentice-Hall Upper Saddle River, NJ, 1996, vol. 1.

[25] A. Sahoo, H. Xu, and S. Jagannathan, "Approximate optimal control of affine nonlinear continuous-time systems using event-sampled neurodynamic programming," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 3, pp. 639–652, 2016.

[26] Y. Li, T. Yang, and S. Tong, "Adaptive neural networks finite-time optimal control for a class of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4451–4460, 2019.

[27] S. Basu Roy, S. Bhasin, and I. N. Kar, "Composite adaptive control of uncertain euler-lagrange systems with parameter convergence without pe condition," *Asian Journal of Control*, vol. 22, no. 1, pp. 1–10, 2020.

[28] N. M. Boffi and J.-J. E. Slotine, "Implicit regularization and momentum algorithms in nonlinearly parameterized adaptive control and prediction," *Neural Computation*, vol. 33, no. 3, pp. 590–673, 2021.

[29] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[30] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *IEEE Conference on Decision and Control*, 2009, pp. 3598–3605.

[31] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.

[32] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, vol. 50, no. 12, pp. 3281–3290, 2014.

[33] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[34] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Systems & Control Letters*, vol. 100, pp. 14–20, 2017.

[35] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Transactions on Cybernetics*, vol. 45, no. 1, pp. 65–76, 2014.

[36] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2550–2562, 2015.

[37] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "H_∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[38] M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot modeling and control*. Wiley New York, 2006, vol. 3.

[39] A. Fradkov, "Speed-gradient scheme and its application in adaptive control problems," *Automation and Remote Control*, vol. 9, pp. 90–101, 1979.

[40] A. L. Fradkov, I. V. Miroshnik, and V. O. Nikiforov, *Nonlinear and adaptive control of complex systems*. Springer Science & Business Media, 2013, vol. 491.
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[42] A. A. Adegbege and M. Y. Kim, "Saddle-point convergence of constrained primal-dual dynamics," *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1357–1362, 2020.
[43] A. Cherukuri, E. Mallada, S. Low, and J. Cortés, "The role of convexity in saddle-point dynamics: Lyapunov function and robustness," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2449–2464, 2017.
[44] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal–dual dynamics," *Systems & Control Letters*, vol. 87, pp. 10–15, 2016.
[45] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.
[46] A. P. Dani and S. Bhasin, "Reinforcement learning for image-based visual servo control," in *IEEE Conference on Decision and Control*, 2023, pp. 4358–4363.
[47] F. L. Lewis, J. Campos, and R. Selmic, *Neuro-fuzzy control of industrial systems with actuator nonlinearities*. SIAM, 2002.
[48] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989.
[49] K. S. Narendra and A. M. Annaswamy, *Stable adaptive systems*. Courier Corporation, 2012.
[50] G. Chowdhary, T. Yucelen, M. Muhlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *International Journal of Adaptive Control and Signal Processing*, vol. 27, no. 4, pp. 280–301, 2013.
[51] A. Parikh, R. Kamalapurkar, and W. E. Dixon, "Integral concurrent learning: Adaptive control with parameter convergence using finite excitation," *International Journal of Adaptive Control and Signal Processing*, vol. 33, no. 12, pp. 1775–1787, 2019.
[52] S. Sastry, M. Bodson, and J. F. Bartram, "Adaptive control: stability, convergence, and robustness," 1990.
[53] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.
[54] V. S. Donge, B. Lian, F. L. Lewis, and A. Davoudi, "Data-efficient reinforcement learning for complex nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1391–1402, 2024.
[55] F. Chaumette and S. Hutchinson, "Visual servo control. I. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.
[56] Z.-P. Jiang, "Robust exponential regulation of nonholonomic systems with uncertainties," *Automatica*, vol. 36, no. 2, pp. 189–209, 2000.

Ashwin Dani (Senior Member, IEEE) received his Ph.D. degree in 2011 from the University of Florida, Gainesville, FL. He is currently an Associate Professor in the Department of Electrical and Computer Engineering, University of Connecticut, Storrs. He was a post-doctoral research associate at the University of Illinois at Urbana-Champaign. His research interests are in the areas of estimation and control, learning for control, vision-based estimation and control, and human-robot collaboration. He was an Associate Editor of IEEE Transactions on Mechatronics during 2021-2023. He is an Associate Editor of IEEE Transactions on Automatic Control and a member of the Conference Editorial Board for IEEE Control System Society. He is a co-recipient of the 2020 IFAC Applications Paper Award Finalist, 2016 FUSION student paper award - 1st runner up, 2015 ASME Dynamics, Systems and Controls Conference Robotics paper award and the 2012 Technology Innovator Award from UF.

Shubhendu Bhasin (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2011. He is currently a Professor with the Department of Electrical Engineering, IIT Delhi, New Delhi, India. His research interests include nonlinear, adaptive, and learning control with a current application focus in robotics, biomedical and battery management systems. He is an Associate Editor for IEEE Transactions on Aerospace and Electronic Systems.