Concurrent learning-based approximate optimal regulation

Rushikesh Kamalapurkar, Patrick Walters, and Warren Dixon

Abstract—In deterministic systems, reinforcement learning-based online approximate optimal control methods typically require a restrictive persistence of excitation (PE) condition for convergence. This paper presents a concurrent learning-based solution to the online approximate optimal regulation problem that eliminates the need for PE. The development is based on the observation that given a model of the system, the Bellman error, which quantifies the deviation of the system Hamiltonian from the optimal Hamiltonian, can be evaluated at any point in the state space. Further, a concurrent learning-based parameter identifier is developed to compensate for parametric uncertainty in the plant dynamics. Uniformly ultimately bounded (UUB) convergence of the system states to the origin, and UUB convergence of the developed policy to an approximate optimal policy are established using a Lyapunov-based analysis, and simulations are performed to demonstrate the performance of the developed controller.

I. INTRODUCTION

Reinforcement learning (RL) enables a cognitive agent to learn desirable behavior from interactions with its environment. In control theory, the desirable behavior is typically quantified using a cost function, and the control problem is formulated as the desire to find the optimal policy that minimizes the cumulative cost. Recently, various RL-based techniques have been developed to approximately solve optimal control problems for continuous-time and discrete-time deterministic systems [1]–[13]. The approximate solution is facilitated via value function approximation, where the value function is approximated using a linear-in-the-parameters (LP) approximation, and the optimal policy is computed based on the estimated value function.

Methods that seek an online solution to the optimal control problem, (cf., [1], [4]) are structurally similar to adaptive control schemes. In adaptive control, the estimates for the uncertain parameters in the plant model are updated using the current tracking error as the performance metric, whereas, in online RL-based techniques, estimates for the uncertain parameters in the value function are updated using the Bellman error (BE) as the performance metric. Convergence of online RL-based techniques to the optimal solution is analogous to parameter convergence in adaptive control.

Parameter convergence has been a focus of research in adaptive control for several decades. It is common knowledge that the least squares and gradient descent-based update laws generally require persistence of excitation (PE) in the system state for convergence of the parameter estimates. Modification schemes such as projection algorithms, $\sigma-$modification, and $e-$modification are used to guarantee boundedness of parameter estimates and overall system stability. However, these modification schemes do not guarantee parameter convergence unless the PE condition, which is often impossible to verify online, is satisfied [14], [15].

As recently shown in results such as [16] and [17], concurrent learning methods can be used to guarantee parameter convergence in adaptive control without relying on the PE condition. Concurrent learning relies on recorded state information along with current state measurements to update the parameter estimates. Learning from recorded data is effective since it is based on the model error, which is closely related to the parameter estimation error. The key concept that enables the computation of the model error from past recorded data is that the model error can be computed if the state derivative is known, and the state derivative can be accurately computed at a past recorded data point using numerical smoothing techniques [16], [17].

In RL-based approximate online optimal control, parameter estimates are updated based on the BE along the state trajectories. Such weight update strategies create two challenges for analyzing convergence. The system states need to be PE for parameter convergence, and the policy, which is based on the estimated weights, needs to regulate the system states to a neighborhood around the desired goal so the information around the desired trajectory can be used to learn the value function. For example, in an infinite horizon regulation problem, if the policy does not regulate the system states to a neighborhood around the origin, the optimal value function (and hence, the optimal policy) near the origin can not be learned, defeating one of the control objectives. These challenges are typically addressed by adding an exploration signal to the control input (cf. [4], [13], [18]) to ensure sufficient exploration in the desired region of the state space. However, no analytical methods exist to compute the appropriate exploration signal for nonlinear systems.

In this paper, the aforementioned challenges are addressed for an infinite horizon optimal regulation problem on a nonlinear, control affine plant with LP uncertainties in the drift dynamics by observing that if the system dynamics are known, the BE can be computed at any desired point in the

state space. Unknown parameters in the value function can therefore be adjusted based on least square minimization of the BE evaluated at any number of desired points in the state space. For example, in an infinite horizon regulation problem, the BE can be computed at sampled points uniformly distributed in a neighborhood around the origin of the state space. The results of this paper indicate that convergence of the unknown parameters in the value function is guaranteed provided the selected points satisfy a rank condition that is weaker than the PE condition. Since the BE can be evaluated at any desired point in the state space, sufficient exploration can be achieved by appropriately selecting the points in a desired neighborhood.

If the system dynamics are partially unknown, an approximation to the BE can be evaluated at any desired point in the state space based on an estimate of the system dynamics. In this paper, a concurrent learning-based parameter estimator is developed to exponentially identify the unknown parameters in the system model, and the parameter estimates are used to compute an approximation to the BE. The unknown parameters in the value function are updated based on the approximate BE, and uniformly ultimately bounded (UUB) convergence of the system states to the origin and UUB convergence of the parameter estimates (and hence, UUB convergence of the developed policy to the optimal policy) is established using a Lyapunov-based analysis. Simulations are performed on a nonlinear system with a four dimensional state to demonstrate the performance of the developed controller.

II. Problem Formulation

Consider a control affine nonlinear dynamic system

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

where $x \in \mathbb{R}^n$ denotes the system state, $\hat{u} \in \mathbb{R}^m$ denotes the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift dynamics, and $g : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ denotes the control effectiveness matrix. The objective is to solve the infinite horizon optimal regulation problem online, i.e., to find the optimal policy $u^* : \mathbb{R}^n \to \mathbb{R}^m$ defined as

$$u^* \triangleq \underset{\substack{\hat{u}:\mathbb{R}^n \to \mathbb{R}^m \\ \hat{u} \in U}}{arg\,min} \int_{t_0}^{\infty} r(x(\tau), \hat{u}(x(\tau)))\,d\tau, \tag{2}$$

while regulating the system states to the origin. In (2), U denotes the set of admissible state feedback policies, and $r : \mathbb{R}^n \times \mathbb{R}^m \to [0, \infty)$ denotes the instantaneous cost defined as

$$r(x, \hat{u}) \triangleq x^T Q x + \hat{u}^T R \hat{u},$$

where $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{m \times m}$ are constant, positive definite symmetric matrices. The class of nonlinear systems considered in this paper is characterized by the following assumption.

Assumption 1. The drift dynamics f is an unknown, LP locally Lipschitz function such that $f(0) = 0$, and the control

effectiveness matrix g is a known, bounded, locally Lipschitz function.

A closed-form solution to the optimal control problem is formulated in terms of the optimal value function $V^* : \mathbb{R}^n \to [0, \infty)$ defined as

$$V^*(x_0) \triangleq \underset{\substack{\hat{u}:\mathbb{R}^n \to \mathbb{R}^m \\ \hat{u} \in U}}{min} \int_{t_0}^{\infty} r(x(\tau), \hat{u}(x(\tau)))\,d\tau, \; \forall x_0 \in \mathbb{R}^n, \tag{3}$$

where $x(\tau)$, for $\tau \in [t_0, \infty)$ denotes the trajectory of (1) with the feedback control law $\hat{u}(x(\tau))$ and the initial condition $x(t_0) = x_0$. Assuming that V^* is continuously differentiable, and $V^*(0) = 0$, the optimal control law can be determined as

$$u^* = -\frac{1}{2}R^{-1}g^T (\nabla_x V^*)^T,$$

where ∇_x denotes the partial derivative with respect to x.

The optimal value function can be obtained by solving the corresponding Hamilton-Jacobi-Bellman (HJB) equation

$$\nabla_x V^* (f + gu^*) + x^T Q x + u^{*T} R u^* = 0. \tag{4}$$

An analytical solution of the HJB equation is generally infeasible to obtain; hence, an approximate solution is sought. An approximate solution based on minimizing the BE is facilitated by replacing V^* and u^* in (4) by their respective subsequently defined estimates $\hat{V}$ and $\hat{u}$ to compute the BE $\delta\left(\hat{V}, x, \hat{u}\right) \in \mathbb{R}$ as

$$\delta = \nabla_x \hat{V} (f + g\hat{u}) + x^T Q x + \hat{u}^T R \hat{u}. \tag{5}$$

The control objective is achieved by simultaneously adjusting the estimates $\hat{V}$ and $\hat{u}$ to minimize the BE evaluated along the trajectory x. The BE depends on the drift dynamics f. Since the drift dynamics are unknown, an adaptive system identifier is developed in the following section.

III. System Identification

Let $f(x) = Y(x)\theta^*$ be the linear parametrization of the function f, where $Y : \mathbb{R}^n \to \mathbb{R}^{n \times p}$ is the regression matrix and $\theta^* \in \mathbb{R}^p$ is the vector of constant unknown parameters. Let $\hat{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ be an estimate of the unknown function f defined as $\hat{f}\left(x, \hat{\theta}\right) \triangleq Y(x)\hat{\theta}$, where $\hat{\theta} \in \mathbb{R}^n$ is the vector of parameter estimates. To estimate the drift dynamics, an identifier is designed as

$$\dot{\hat{x}} = \hat{f} + g\hat{u} + k_x \tilde{x}, \tag{6}$$

where the state estimation error $\tilde{x}$ is defined as $\tilde{x} \triangleq x - \hat{x}$ and $k_x \in \mathbb{R}^{n \times n}$ is a positive definite, constant diagonal observer gain matrix. From (1) and (6) the identification error dynamics can be derived as

$$\dot{\tilde{x}} = Y\tilde{\theta} - k_x \tilde{x}, \tag{7}$$

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

A. Concurrent learning-based parameter update

In traditional adaptive control, convergence of the estimates $\hat{\theta}$ to their true values θ^* is ensured by using a PE condition [14], [15]. To ensure convergence without the PE condition, a concurrent learning-based approach can be employed [16], [17]. The following observability assumption relaxes the PE condition that is required for parameter convergence in adaptive control.

Assumption 2. [16], [17] There exists a finite set of time instances $\{t_j \mid j = 1, \cdots, M\}$ such that

$$rank\left(\sum_{j=1}^{M} Y_j^T Y_j\right) = p, \qquad (8)$$

where $Y_j = Y\left(x\left(t_j\right)\right)$.

Let $\underline{y} \in \mathbb{R}$ be the minimum eigenvalue of $\left(\sum_{j=1}^{M} Y_j^T Y_j\right)$. Based on (8), $\left(\sum_{j=1}^{M} Y_j^T Y_j\right)$ is symmetric and positive definite, and hence, $\underline{y} > 0$. The condition in (8) is satisfied as long as the system states are exciting over a finite period of time, and hence, is weaker than the PE condition. Furthermore, unlike the PE condition, the rank condition in (8) can be verified online since it is a function of past states. To design the concurrent learning-based parameter update law, time instances $\{t_j \mid j = 1, \cdots, M\}$ are selected such that the condition in (8) holds, and the states $\left\{x_j \triangleq x\left(t_j\right) \mid j = 1, \cdots, M\right\}$ and the corresponding control values $\left\{\hat{u}_j \triangleq \hat{u}\left(t_j\right) \mid j = 1, \cdots, M\right\}$ are recorded in a history stack. The update law is then designed as

$$\dot{\hat{\theta}} = \Gamma_\theta Y^T \tilde{x} + \Gamma_\theta k_\theta \sum_{j=1}^{M} Y_j^T \left(\dot{x}_j - g_j \hat{u}_j - Y_j \hat{\theta}\right), \qquad (9)$$

where $g_j = g(x_j)$, $\Gamma_\theta \in \mathbb{R}^{p \times p}$ is a constant positive definite adaptation gain matrix, and k_θ is a constant positive concurrent learning gain. The update law in (9) depends on the unknown state derivative $\dot{x}_j = \dot{x}\left(t_j\right)$. However, since the state derivative is from recorded data, numerical smoothing techniques based on past and future data can be used to obtain good estimates of the derivative. In the presence of derivative estimation errors, the parameter estimation errors can be shown to be UUB, where the size of the ultimate bound depends on the error in the derivative estimate [17]. From (1) and the definitions of $\tilde{\theta}$ and $\hat{f}$, the bracketed term in (9), can be expressed as $\dot{x}_j - g_j \hat{u}_j - Y_j \hat{\theta} = Y_j \tilde{\theta}$ and the parameter update law in (9) can be expressed as

$$\dot{\hat{\theta}} = \Gamma_\theta Y^T \tilde{x} + \Gamma_\theta k_\theta \left(\sum_{j=1}^{M} Y_j^T Y_j\right) \tilde{\theta}. \qquad (10)$$

IV. APPROXIMATE OPTIMAL CONTROL

Based on the system identifier developed in Section III, the BE in (5) can be approximated as

$$\hat{\delta}\left(x, \hat{u}, \hat{V}, \hat{\theta}\right) = \nabla_x \hat{V}\left(Y\hat{\theta} + g\hat{u}\right) + x^T Q x + \hat{u}^T R \hat{u}. \qquad (11)$$

In the following, the approximate BE in (11) is used to obtain an approximate solution to the HJB equation in (4).

A. Value function approximation

Approximations to the optimal value function V^* and the optimal policy u^* are designed based on neural network (NN)-based representations. The NN-based representation is facilitated by a temporary assumption that the state trajectory x evolves on a compact subspace $\chi \subset \mathbb{R}^n$. The compactness assumption is common in neural network-based adaptive control (cf. [19], [20]), and it is shown in the subsequent stability analysis that the states evolve on a compact set provided the initial condition $x(t_0)$ is bounded (see Remark 1 in the subsequent stability analysis). The following standard NN assumption describes a NN-based representation of the optimal value function.

Assumption 3. On the compact set χ, the optimal value function V^* can be represented using a NN as

$$V^* = W^{*T}\sigma + \epsilon, \qquad (12)$$

where $W^* \in \mathbb{R}^L$ is the constant ideal weight matrix, which is bounded above by a known positive constant $\bar{W}$ in the sense that $\|W^*\| \leq \bar{W}$, $\sigma : \chi \to \mathbb{R}^L = \begin{bmatrix} \sigma_1 & \cdots & \sigma_L \end{bmatrix}^T$ is a bounded continuously differentiable nonlinear activation function such that $\sigma(0) = 0$ and $\sigma'(0) = 0$, $L \in \mathbb{N}$ is the number of neurons, and $\epsilon : \chi \to \mathbb{R}$ is the function reconstruction error such that $\sup_{x \in \chi} |\epsilon(x)| \leq \bar{\epsilon}$ and $\sup_{x \in \chi} |\epsilon'(x)| \leq \bar{\epsilon}'$, where $\bar{\epsilon}, \bar{\epsilon}' \in \mathbb{R}$ are known positive constants.

Based on (12) a NN-based representation of the optimal controller is derived as

$$u^* = -\frac{1}{2}R^{-1}g^T\left(\sigma'^T W^* + \epsilon'^T\right). \qquad (13)$$

The NN-based approximations to the optimal value function in (12) and the optimal policy in (13) are defined as

$$\hat{V} \triangleq \hat{W}_c^T \sigma, \quad \hat{u} \triangleq -\frac{1}{2}R^{-1}g^T\sigma'^T \hat{W}_a, \qquad (14)$$

where $\hat{W}_c \in \mathbb{R}^L$ and $\hat{W}_a \in \mathbb{R}^L$ are estimates of the ideal weights W^*. The use of two sets of weights to estimate the same set of ideal weights is motivated by the stability analysis and fact that it enables a formulation of the BE that is linear in the value function weight estimates $\hat{W}_c$, enabling a least squares-based adaptive update law.

Based on (14), the approximate BE in (11) can be expressed as

$$\hat{\delta}\left(x, \hat{W}_a, \hat{W}_c, \hat{\theta}\right) = \omega^T \hat{W}_c + x^T Q x + \hat{u}^T R \hat{u}, \qquad (15)$$

where $\omega\left(x, \hat{\theta}, \hat{W}_a\right) \triangleq \sigma'(x)\left(\hat{f}\left(x, \hat{\theta}\right) + g(x)\hat{u}\left(x, \hat{W}_a\right)\right)$ is the regressor vector.

B. Learning based on desired behavior

In traditional RL-based algorithms, the value function estimate and the policy estimate are updated based on observed data, and hence require sufficient exploration. In stochastic systems, this is achieved using a randomized stationary policy (cf. [13], [18], [21]), whereas in deterministic systems, a probing noise is added to the derived control law (cf. [1]–[3], [7], [22]). The technique developed in this result is based on the observation that if an estimate of the system dynamics is available, an approximation to the BE can be evaluated at any desired point in the state space. The following condition, similar to the condition in (8), enables the use of an approximate BE evaluated at a pre-sampled set of points $\{x_i \in \chi \mid i = 1, \cdots, N\}$ in the state space.

Assumption 4. There exists a set of points $\{x_i \in \chi \mid i = 1, \cdots, N\}$ such that $\forall t \in [0, \infty)$,

$$rank\left(\sum_{i=1}^{N} \frac{\omega_i \omega_i^T}{\rho_i}\right) = L. \quad (16)$$

In (16), $\rho_i \triangleq 1 + \nu \omega_i^T \Gamma \omega_i \in \mathbb{R}$ are the normalization terms, where $\nu \in \mathbb{R}$ is a constant positive normalization gain, $\Gamma \in \mathbb{R}^{L \times L}$ is the least-squares gain matrix, and

$$\omega_i\left(x_i, \hat{\theta}, \hat{W}_a\right) \triangleq \sigma'(x_i)\left(\hat{f}\left(x_i, \hat{\theta}\right) + g(x_i)\hat{u}\left(x_i, \hat{W}_a\right)\right).$$

To facilitate the stability analysis, let

$$\underline{c} = \frac{1}{N}\left(\inf_{t \in \mathbb{R}_{\geq 0}}\left(\lambda_{min}\left\{\sum_{i=1}^{N} \frac{\omega_i \omega_i^T}{\rho_i}\right\}\right)\right), \quad (17)$$

where $\lambda_{min}\{\cdot\}$ denotes the minimum eigenvalue. Provided Assumption 4 holds, the infimum in (17) is positive. The condition in (8) is weaker than the PE condition in previous results such as [1]–[3], [7], [22], and unlike the PE condition, (8) can be verified online. Since the rank condition in (16) depends on the estimates $\hat{\theta}$ and $\hat{W}_a$, it is in general, impossible to guarantee a priori. However, heuristically, the condition in (16) can be met by collecting redundant data, i.e., by selecting more points than the number of neurons by choosing $N \gg L$.

The approximate BE can be evaluated at the sampled points $\{x_i \mid i = 1, \cdots, N\}$ as

$$\hat{\delta}_i\left(x_i, \hat{W}_a, \hat{W}_c, \hat{\theta}\right) = \omega_i^T \hat{W}_c + x_i^T Q x_i + \hat{u}_i^T R \hat{u}_i, \quad (18)$$

where $\hat{u}_i\left(x_i, \hat{W}_a\right) \triangleq -\frac{1}{2}R^{-1}g(x_i)^T \sigma(x_i)'^T \hat{W}_a$. A concurrent learning-based least-squares update law for the value function weights is designed based on the subsequent stability analysis as

$$\dot{\hat{W}}_c = -\eta_{c1}\Gamma\frac{\omega}{\rho}\hat{\delta} - \frac{\eta_{c2}}{N}\Gamma\sum_{i=1}^{N} \frac{\omega_i}{\rho_i}\hat{\delta}_i,$$

$$\dot{\Gamma} = \left(\beta\Gamma - \eta_{c1}\Gamma\frac{\omega\omega^T}{\rho}\Gamma\right)\mathbf{1}_{\{\|\Gamma\| \leq \overline{\Gamma}\}}, \ \|\Gamma(t_0)\| \leq \overline{\Gamma}, \quad (19)$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, $\overline{\Gamma} > 0 \in \mathbb{R}$ is the saturation constant, $\beta > 0 \in \mathbb{R}$ is the forgetting factor, and $\eta_{c1}, \eta_{c2} > 0 \in \mathbb{R}$ are constant adaptation gains. The update law in (19) ensures that the adaptation gain matrix is bounded such that

$$\underline{\Gamma} \leq \|\Gamma(t)\| \leq \overline{\Gamma}, \ \forall t \in \mathbb{R}_{\geq 0}. \quad (20)$$

The policy weights are then updated to follow the value function weights as

$$\dot{\hat{W}}_a = -\eta_{a1}\left(\hat{W}_a - \hat{W}_c\right) - \eta_{a2}\hat{W}_a$$
$$+ \left(\frac{\eta_{c1}G_\sigma^T \hat{W}_a \omega^T}{4\rho} + \sum_{i=1}^{N} \frac{\eta_{c2}G_{\sigma i}^T \hat{W}_a \omega_i^T}{4N\rho_i}\right)\hat{W}_c, \quad (21)$$

where $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are positive constant adaptation gains and $G_\sigma \triangleq \sigma' g R^{-1} g^T \sigma'^T \in \mathbb{R}^{L \times L}$.

The update law in (19) is fundamentally different from the concurrent learning adaptive update in results such as [16], [17], in the sense that the points $\{x_i \in \chi \mid i = 1, \cdots, N\}$ are selected a priori based on prior information about the desired behavior of the system. For example, in the present case, since the objective is to regulate the system states to the origin and the system is deterministic, it is natural to select a bounded set of points uniformly distributed around the origin of the state space. This difference is a result of the fact that the developed RL-based technique uses the approximate BE as the metric to update the weight estimates. Given the system dynamics, or an estimate of the system dynamics, the approximate BE can be evaluated at any desired point in the state space, whereas in adaptive control, the prediction error is used as a metric which can only be evaluated at observed data points along the state trajectory.

V. STABILITY ANALYSIS

To facilitate the subsequent stability analysis, the approximate BE is expressed in terms of the weight estimation errors $\tilde{W}_c \triangleq W^* - \hat{W}_c$ and $\tilde{W}_a \triangleq W^* - \hat{W}_a$. Subtracting (4) from (15), the unmeasurable form of the instantaneous BE can be expressed as

$$\hat{\delta} = \nabla_x \hat{V}\left(\hat{f} + g\hat{u}\right) + \hat{u}^T R \hat{u} - \nabla_x V^*(f + gu^*) - u^{*T}Ru^*,$$
$$= -\omega^T \tilde{W}_c - W^{*T}\sigma'Y\tilde{\theta} + \frac{1}{4}\tilde{W}_a^T G_\sigma \tilde{W}_a + \frac{1}{4}G_\epsilon - \epsilon'f$$
$$+ \frac{1}{2}W^{*T}\sigma'G\epsilon'^T, \quad (22)$$

where $G \triangleq gR^{-1}g^T \in \mathbb{R}^{n \times n}$ and $G_\epsilon \triangleq \epsilon'G\epsilon'^T \in \mathbb{R}$. Similarly, the approximate BE evaluated at the sampled states $\{x_i \mid i = 1, \cdots, N\}$ can be expressed as

$$\hat{\delta}_i = -\omega_i^T \tilde{W}_c + \frac{1}{4}\tilde{W}_a^T G_{\sigma i}\tilde{W}_a - W^{*T}\sigma'Y_i\tilde{\theta} + \Delta_i, \quad (23)$$

where $\Delta_i \triangleq \frac{1}{2}W^{*T}\sigma'_i G_i \epsilon'^T_i + \frac{1}{4}G_{\epsilon i} - \epsilon'_i f_i \in \mathbb{R}$ is a constant.

On the compact set χ the functions f and Y are Lipschitz continuous, and hence, there exist positive constants L_f,

$L_Y \in \mathbb{R}$ such that[1]

$$\|f(x)\| \le L_f \|x\|, \quad \|Y(x)\| \le L_Y \|x\|, \forall x \in \chi. \quad (24)$$

Using (20), the normalized regressor $\frac{\omega}{\rho}$ can be bounded as

$$\overline{\left\|\frac{\omega(x)}{\rho(x)}\right\|} \le \frac{1}{2\sqrt{\nu \underline{\Gamma}}}. \quad (25)$$

where the operator $\overline{(\cdot)} : [0, \infty) \to [0, \infty)$ is defined as $\overline{(\cdot)} \triangleq \sup_{x \in \mathbb{R}^n} (\cdot)$. The following positive constants are defined to facilitate the subsequent analysis

$$\vartheta_1 \triangleq \frac{\eta_{c1} L_f \overline{\epsilon'}}{4\sqrt{\nu \underline{\Gamma}}}, \ \vartheta_2 \triangleq \sum_{i=1}^{N} \left(\frac{\eta_{c2} \|\sigma_i' Y_i\| \overline{W}}{4N\sqrt{\nu \underline{\Gamma}}} \right), \ \vartheta_4 \triangleq \overline{\left\|\frac{1}{4} G_\epsilon\right\|},$$

$$\vartheta_7 \triangleq \frac{(\eta_{c1} + \eta_{c2}) \overline{\|G_\sigma\|}}{8\sqrt{\nu \underline{\Gamma}}}, \ \vartheta_3 \triangleq \frac{L_Y \eta_{c1} \overline{W} \|\sigma'\|}{4\sqrt{\nu \underline{\Gamma}}},$$

$$\vartheta_5 \triangleq \overline{\left\|\frac{\eta_{c1} \omega (2W^T \sigma' G \epsilon'^T + G_\epsilon)}{4\rho} + \sum_{i=1}^{N} \frac{\eta_{c2} \omega_i \Delta_i}{N \rho_i}\right\|},$$

$$\vartheta_6 \triangleq \overline{\left\|\frac{1}{2} W^T G_\sigma + \frac{1}{2} \epsilon' G^T \sigma'^T\right\|} + \vartheta_7 \overline{W}^2 + \eta_{a2} \overline{W}.$$

The main result of this paper can now be stated as follows.

Theorem 1. *Provided Assumptions (1) - (4) hold and gains $\underline{q}$, η_{c2}, η_{a2}, and k_θ are selected large enough based on the following sufficient conditions*

$$\eta_{a2} > -\frac{\eta_{a1}}{2} + \vartheta_7 \overline{W} \left(\frac{2\zeta_2 + 1}{2\zeta_2} \right), \quad k_\theta > \frac{\vartheta_2 + \zeta_1 \vartheta_3 \overline{Z}}{\underline{y} \zeta_1},$$

$$\underline{q} > \vartheta_1, \ \eta_{c2} > \frac{\zeta_2 \vartheta_7 \overline{W} + \eta_{a1} + 2\left(\vartheta_1 + \zeta_1 \vartheta_2 + \vartheta_3 \overline{Z}\right)}{2\underline{c}}, \quad (26)$$

where $\zeta_1, \zeta_2 \in \mathbb{R}$ are constant adjustable parameters and $\overline{Z}(t_0) \in \mathbb{R}$ is a positive constant that depends on the initial condition of the system, the observer in (6) along with the adaptive update law in (9) and the controller in (14) along with the adaptive update laws in (19) and (21) ensure that the state x, the state estimation error $\tilde{x}$, the parameter estimation error $\tilde{\theta}$, the value function weight estimation error $\tilde{W}_c$ and the policy weight estimation error $\tilde{W}_a$ are UUB, resulting in UUB convergence of the policy $\hat{u}$ to the optimal policy u^*.

Proof: Let $V_L : \mathbb{R}^{2n+2L+p} \times [0, \infty) \to [0, \infty)$ be a continuously differentiable positive definite candidate Lyapunov function defined as

$$V_L \triangleq V^* + \frac{1}{2} \tilde{W}_c^T \Gamma^{-1} \tilde{W}_c + \frac{1}{2} \tilde{W}_a^T \tilde{W}_a + \frac{1}{2} \tilde{x}^T \tilde{x} + \frac{1}{2} \tilde{\theta}^T \Gamma_\theta^{-1} \tilde{\theta} \quad (27)$$

where V^* is the optimal value function. Using the fact that V^* is positive definite, (20) and Lemma 4.3 from [23] yield

$$\underline{v_l}(\|Z\|) \le V_L(Z, t) \le \overline{v_l}(\|Z\|), \quad (28)$$

[1]The Lipschitz property is exploited here for clarity of exposition. The bounds in (24) can be easily generalized to $\|f(x)\| \le L_f(\|x\|)\|x\|$, $\|Y(x)\| \le L_Y(\|x\|)\|x\|$, where $L_f, L_Y : \mathbb{R} \to \mathbb{R}$ are positive, non decreasing functions.

for all $t \in [0, \infty)$ and for all $Z \in \mathbb{R}^{2n+2L+p}$. In (28), $\underline{v_l}, \overline{v_l} : [0, \infty] \to [0, \infty)$ are class $\mathcal{K}$ functions and

$$Z \triangleq \left[x^T, \tilde{W}_c^T, \tilde{W}_a^T, \tilde{x}^T, \tilde{\theta}^T \right]^T.$$

Substituting for the approximate BEs from (22) and (23), using the bounds in (24) and (25), and using Young's inequality, the time derivative of the Lyapunov function along the trajectories of (1), (7), (10), (19), and (21) can be upper-bounded as

$$\dot{V}_L \le -\iota_x \|x\|^2 - \iota_c \left\|\tilde{W}_c\right\|^2 - \iota_a \left\|\tilde{W}_a\right\|^2 - \underline{k_x} \|\tilde{x}\|^2$$
$$- \iota_\theta \left\|\tilde{\theta}\right\|^2 + \vartheta_5 \left\|\tilde{W}_c\right\| + \vartheta_6 \left\|\tilde{W}_a\right\| + \vartheta_4,$$

where the constants $\iota_x, \iota_a \in \mathbb{R}$ and the coefficients $\iota_c, \iota_\theta \in \mathbb{R}$ are defined as

$$\iota_x \triangleq \underline{q} - \vartheta_1, \quad \iota_a \triangleq \frac{\eta_{a1}}{2} + \eta_{a2} - \vartheta_7 \overline{W} \left(\frac{2\zeta_2 + 1}{2\zeta_2} \right),$$

$$\iota_c \triangleq \eta_{c2} \underline{c} - \vartheta_1 - \zeta_1 \vartheta_2 - \frac{\zeta_2 \vartheta_7 \overline{W} + \eta_{a1}}{2} - \vartheta_3 \|x\|,$$

$$\iota_\theta \triangleq k_\theta \underline{y} - \frac{\vartheta_2}{\zeta_1} - \vartheta_3 \|x\|.$$

Provided the conditions in (26) are satisfied, the coefficients ι_x, ι_a, ι_c, and ι_θ are positive for all $t \in [0, \infty)$(see Remark 1). Completing the squares, the Lyapunov derivative can be expressed as

$$\dot{V}_L \le -\iota_x \|x\|^2 - \frac{\underline{\iota_c}}{2} \left\|\tilde{W}_c\right\|^2 - \frac{\iota_a}{2} \left\|\tilde{W}_a\right\|^2 - \underline{k_x} \|\tilde{x}\|^2$$
$$- \underline{\iota_\theta} \left\|\tilde{\theta}\right\|^2 + \iota,$$
$$\le -v_l \|Z\|^2, \quad \forall \|Z\| \ge \iota > 0, \quad (29)$$

where $v_l = \min\left(\iota_x, \frac{\underline{\iota_c}}{2}, \frac{\iota_a}{2}, \underline{k_x}, \underline{\iota_\theta} \right) \in \mathbb{R}$, $\iota \triangleq \sqrt{\frac{\vartheta_5^2}{2v_l \underline{\iota_c}} + \frac{\vartheta_6^2}{2v_l \iota_a} + \frac{\vartheta_4}{v_l}} \in \mathbb{R}$, and $\underline{\iota_c}, \underline{\iota_\theta} \in \mathbb{R}$ are the lower bounds on ι_c and ι_θ, respectively. Theorem 4.18 in [23] can now be invoked to conclude that Z is UUB. ∎

Remark 1. If $\|Z(0)\| \ge \iota$ then $\dot{V}_L(Z(0)) < 0$. Thus, $V_L(Z)$ is decreasing at $t = 0$, and hence, $Z \in \mathcal{L}_\infty$ at $t = 0^+$. Thus all the conditions of Theorem 1 are satisfied at $t = 0^+$. As a result, $V_L(Z)$ is decreasing at $t = 0^+$. By induction, $\|Z(0)\| \ge \iota \implies V_L(Z(t)) \le V_L(Z(0)), \forall t \in \mathbb{R}^+$. Thus, from (28), $\|Z(t)\| \le \underline{v_l}^{-1}(\overline{v_l}(\|Z(0)\|)), \forall t \in [0, \infty)$. If $\|Z(0)\| < \iota$ then (28) and (29) can be used to determine that $\underline{v_l}(\|Z(t)\|) \le V_L(Z(t)) \le \overline{v_l}(\|\iota\|), \forall t \in [0, \infty)$. As a result, $\|Z(t)\| \le \underline{v_l}^{-1}(\overline{v_l}(\iota)), \forall t \in [0, \infty)$. Let the positive constant $\overline{Z} \in \mathbb{R}$ be defined as $\overline{Z} \triangleq \underline{v_l}^{-1}(\overline{v_l}(\max(\|Z(0)\|, \iota)))$. This relieves the compactness assumption in the sense that the compact set $\chi \subset \mathbb{R}^n$ that contains the system trajectories $x(t), \forall t \in [0, \infty)$ is given by $\chi \triangleq \left\{ \xi \in \mathbb{R}^n \mid \|\xi\| \le \overline{Z} \right\}$. Furthermore, $\|Z(t)\| \le \overline{Z}, \forall t \in [0, \infty)$ implies that the gain conditions in (26) are sufficient for the the coefficients ι_x, ι_a, ι_c, and ι_θ to be positive for all $t \in [0, \infty)$.

VI. SIMULATION

To demonstrate the performance of the developed technique, simulations are performed to solve the two dimensional nonlinear optimal control problem in [3]. The dynamics are given by

$$f\left(x\right) = \left[\begin{array}{c} ax_1 + bx_2 \\ cx_1 + dx_2\left(1 - \left(cos\left(2x_1\right) + 2\right)^2\right) \end{array}\right],$$

$$g\left(x\right) = \left[\begin{array}{c} 0 \\ cos\left(2x_1\right) + 2 \end{array}\right],$$

where $a, b, c, d \in \mathbb{R}$ are unknown parameters with true values $a = -1$, $b = 1$, $c = -0.5$, and $d = -0.5$. The optimal value function for this problem is known and is given by $V^* = \frac{1}{2}x_1^2 + x_2^2$. The basis for value function approximation is $\sigma = \left[\begin{array}{ccc} x_1^2, & x_1 x_2, & x_2^2 \end{array}\right]$, and the points for concurrent learning are selected to lie on a 5×5 grid around the origin.

Figure 1 shows that the value function weights and the unknown parameters in the drift dynamics converge to their true values. The existing techniques rely on an ad-hoc probing signal added to the controller to ensure PE and parameter convergence. The state trajectory and the control trajectory plots demonstrate that unlike the existing techniques, parameter convergence is obtained without the use of an ad-hoc probing signal.
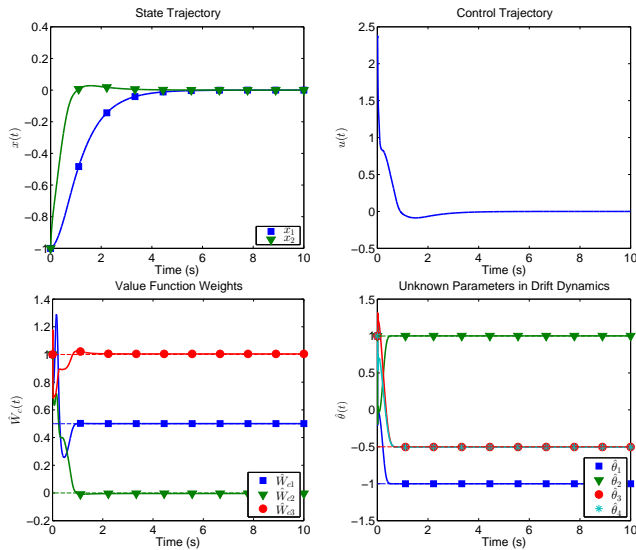

Figure 1. System trajectories generated using the developed method.

VII. CONCLUSION

A RL-based online approximate optimal controller is developed that does not require PE for convergence. The PE condition is replaced by a weaker rank condition that can be verified online from recorded data. An approximation to the BE computed at pre-sampled desired values of the system state using an estimate of the system dynamics is used to improve the value function approximation, and UUB convergence of the system states to the origin, and UUB convergence of the policy to the optimal policy are established using a Lyapunov-based analysis. Simulations are performed to demonstrate the performance of the developed controller.

REFERENCES

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

[2] K. Vamvoudakis and F. Lewis, "Online synchronous policy iteration method for optimal control," in *Recent Advances in Intelligent Control Systems*. Springer, 2009, pp. 357–374.

[3] ——, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, pp. 878–888, 2010.

[4] D. Vrabie, M. Abu-Khalaf, F. Lewis, and Y. Wang, "Continuous-time ADP for linear systems with partially unknown dynamics," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinf. Learn.*, 2007, pp. 247–253.

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

[6] T. Dierks, B. Thumati, and S. Jagannathan, "Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence," *Neural Netw.*, vol. 22, no. 5-6, pp. 851–860, 2009.

[7] T. Dierks and S. Jagannathan, "Optimal tracking control of affine nonlinear discrete-time systems with unknown internal dynamics," in *Proc. IEEE Conf. Decis. Control*, 2009, pp. 6750–6755.

[8] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2226–2236, 2011.

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

[10] 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_∞ control," *Automatica*, vol. 43, pp. 473–481, 2007.

[11] ——, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 38, pp. 943–949, 2008.

[12] R. Padhi, N. Unnikrishnan, X. Wang, and S. Balakrishnan, "A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems," *Neural Netw.*, vol. 19, no. 10, pp. 1648–1660, 2006.

[13] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Conf. Decis. Control*, Dec. 2009, pp. 3598 –3605.

[14] K. Narendra and A. Annaswamy, *Stable Adaptive Systems*. Prentice-Hall, Inc., 1989.

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

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

[17] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt Control Signal Process.*, 2012.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[19] F. Lewis, K. Liu, and A. Yesildirek, "Neural net robot controller with guaranteed tracking performance," in *Proc. IEEE Int. Symp. Intell. Control*, Chicago, Illinois, 1993, pp. 225–231.

[20] K. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.

[21] V. Konda and J. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Contr. Optim.*, vol. 42, no. 4, pp. 1143–1166, 2004.

[22] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems," in *Proc. Am. Control Conf.*, 2010, pp. 1568–1573.

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