Adaptive Linear Quadratic Regulator for Continuous-Time Systems With Uncertain Dynamics

Sumit Kumar Jha, *Member, IEEE,* and Shubhendu Bhasin, *Member, IEEE*

***Abstract*—In this paper, adaptive linear quadratic regulator (LQR) is proposed for continuous-time systems with uncertain dynamics. The dynamic state-feedback controller uses input-output data along the system trajectory to continuously adapt and converge to the optimal controller. The result differs from previous results in that the adaptive optimal controller is designed without the knowledge of the system dynamics and an initial stabilizing policy. Further, the controller is updated continuously using input-output data, as opposed to the commonly used switched/intermittent updates which can potentially lead to stability issues. An online state derivative estimator facilitates the design of a model-free controller. Gradient-based update laws are developed for online estimation of the optimal gain. Uniform exponential stability of the closed-loop system is established using the Lyapunov-based analysis, and a simulation example is provided to validate the theoretical contribution.**

***Index Terms*—Adaptive optimal control, continuous policy update, linear quadratic regulator, uncertain system, dynamics.**

I. Introduction

THE development of the infinite-horizon linear quadratic regulator (LQR) [1] has been one of the most important contributions in linear optimal control theory. The optimal control law for the LQR problem is expressed in state-feedback form, where the optimal gain is obtained from the solution of the nonlinear matrix equation – the algebraic Riccati equation (ARE). The solution of the ARE requires exact knowledge of the system matrices and is typically found offline, a major impediment to online real-time control.

Recent research has focused on solving the optimal control problem using iterative, data-driven algorithms which can be implemented online and require minimal knowledge of the system dynamics [2]−[15]. In [2], Kleinman proposed a computationally efficient procedure for solving the ARE by iterating on the solution of the linear Lyapunov equation, with proven convergence to the optimal policy for any initial condition. The Newton-Kleinman algorithm [2], although offline and model-based, paved the way for a class of reinforcement learning (RL)/approximate dynamic programming (ADP)-based algorithms which utilize data along the system trajectory to learn the optimal policy [4], [7], [10], [16]−[18]. Strong connections between RL/ADP and optimal control have been established [19]−[23] and several RL algorithms including policy iteration (PI), value iteration (VI) and Q-learning have been adapted for optimal control problems [4], [7]−[9], [13], [22], [24]. Initial research on adaptive optimal control was mostly concentrated in the discrete-time domain due to the recursive nature of RL/ADP algorithms. An important contribution in [4] is the development of a model-free PI algorithm using Q-functions for discrete-time adaptive linear quadratic control. The iterative RL/ADP algorithms have since been applied to various discrete-time optimal control problems [25]−[27].

Extension to continuous-time systems entails challenges in controller development and convergence/stability proofs. One of the first adaptive optimal controllers for continuous-time systems is proposed in [17], where a model-based algorithm is designed using a continuous-time version of the temporal difference (TD) error. Model-free RL algorithms for continuous-time systems are proposed in [22], which require measurement of the state derivatives. In chapter 7 of [3], an indirect adaptive optimal linear quadratic (ALQ) controller is proposed, where the unknown system parameters are identified using an online adaptive update law, and the ARE is solved at every time instant using the current parameter estimates. However, the algorithm may become computationally prohibitive for higher dimensional systems, owing to the need for solving the ARE at every time instant. More recently, partially model-free PI algorithms are developed in [7], [24] for linear systems with unknown internal dynamics. In [9], [10], the idea in [7] is extended to adaptive optimal control of linear systems with completely unknown dynamics. In another significant contribution [6], the connections between Q-learning and the Pontryagin's minimum principle are established, based on which an off policy control algorithm is proposed.

A common feature of RL algorithms adapted for continuous-time systems is the requirement of an initial stabilizing policy [7], [9], [10], [18], [24], and a batch least square estimation algorithm leading to intermittent updates of the control policy [7], [9]. Finding an initial stabilizing policy for systems with unknown dynamics may not always possible. Further, the intermittent control policy updates in [7], [9], [18] render the control law discontinuous, potentially leading

Manuscript received April 2, 2018; accepted June 12, 2018. Recommended by Associate Editor Qinglai Wei. *(Corresponding author: Sumit Kumar Jha.)*

Citation: S. K. Jha and S. Bhasin, "Adaptive linear quadratic regulator for continuous-time systems with uncertain dynamics," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 3, pp. 833−841, May 2020.

S. K. Jha is with the Department of Electronics and Communication Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj-211004, India (e-mail: sumitjha54@gmail.com).

S. Bhasin is with the Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi-110016, India (e-mail: sbhasin@ee.iitd.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2019.1911438

to challenges in proving stability. Moreover, many adaptive optimal control algorithms require to implement delayed-window integrals to construct the regressor/design update laws [5], [7], [9], [14], and "intelligent" data storage mechanism (procedure for populating independent set of data) [5], [7], [9], [10] to satisfy an underlying full-rank condition. The computation of delayed-window integrals of functions of states requires past data storage for the time interval $[t-T,\ t]$, $\forall t > 0$, where t and T are the current time instant and the window length, respectively, which demands significant memory consumption, especially for large scale systems.

Recent works in [8], [11], [13] have cast the continuous-time RL problem in an adaptive control framework with continuous policy updates, without the need for an initial stabilizing policy. However, for continuous-time RL, it is not straight forward to develop a fixed-point equation for parameter updation, which is independent of the knowledge of system dynamics and state derivatives. A synchronous PI algorithm for known system dynamics is developed in [8], which is extended to a partially model-free method using a novel actor-critic-identifier architecture [11]. For input-constrained systems with completely unknown dynamics, a PI and neural network (NN) based adaptive control algorithm is proposed in [13]. However, the work in [13] utilizes past stored data along with the current data for identifier design, while guaranteeing bounded convergence of critic weight estimation error for bounded NN reconstruction error.

The contribution of this paper is the design of a continuous-time adaptive LQR with a time-varying state-feedback gain, which is shown to exponentially converge to the optimal gain. The novelty of the proposed result lies in the computational/memory efficient algorithm used to solve the optimal control problem for uncertain dynamics, without requiring an initial stabilizing control policy, unlike previous results which either use an initial stabilizing control policy and a switched policy update [5], [7], [9], [10] or past data storage [5], [7], [9], [10], [28], [29] or memory-intensive delayed-window integrals [5], [7], [9], [14]. The result in this paper is facilitated by the development of a fixed point equation which is independent of system matrices, and the design of a state derivative estimator. A gradient-based update law is devised for online adaptation of the state-feedback gain and convergence to the optimal gain is shown, provided a uniform persistence of excitation (u-PE) condition [30], [31] on the state-dependent regressor is satisfied. The u-PE condition, although restrictive in its verification and implementation, establishes the theoretical requirements for convergence of adaptive linear quadratic controller proposed in the paper. The Lyapunov analysis is used to prove uniform exponential stability of the overall system.

This paper is organized as follows. Section II discusses the primary concepts of linear optimal control, problem formulation, and subsequently the general methodology. The proposed model-free adaptive optimal control design along with the state derivative estimator is described in Section III. Convergence and exponential stability of the proposed result is shown in Section IV. Finally, an illustrative example is given in Section V.

Notations: Throughout this paper, $\mathbb{R}$ is used to denote the set of real numbers. The operator $||\,.\,||$ designates the Euclidean norm for vectors and induced matrix norm for matrices. The symbol $\otimes$ denotes the Kronecker product operator and $vec(Z) \in \mathbb{R}^{qr}$ denotes the vectorization of the argument matrix $Z \in \mathbb{R}^{q\times r}$ and is obtained by stacking columns of the argument matrix on top of one another. The operators $\lambda_{\min}(.)$ and $\lambda_{\max}(.)$ denote the minimum and maximum eigenvalues of the argument matrix, respectively. The symbol B_d denotes the open ball $B_d = \{z \in \mathbb{R}^{n(n+m)} : ||z|| < d\}$. The following standard properties of vec and Kronecker product have been used for the matrices D, E and F of appropriate dimension

1) $vec(DEF) = (F^T \otimes D)vec(E)$, where matrix multiplication (DEF) is defined.

2) $vec(D+E+F) = vec(D) + vec(E) + vec(F)$, where matrix summation $(D+E+F)$ is defined.

The partial derivative formula, $\frac{\partial(a^T D b)}{\partial D} = ab^T$, where a, b are vectors, D is a matrix and the multiplication $(a^T D b)$ is defined, has also been used.

II. Preliminaries and Problem Formulation

Consider a continuous-time deterministic LTI system given as

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

where $x(t) \in \mathbb{R}^n$ denotes the state and $u(t) \in \mathbb{R}^m$ denotes the control input. $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant unknown matrices and (A, B) are assumed to be controllable.

The infinite horizon quadratic value function can be defined as the total cost starting from state $x(t)$ and following a fixed control action $u(t)$ from time t onwards as

$$V(x(t)) = \int_t^\infty (x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau))d\tau \tag{2}$$

where $Q \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite with (Q, A) being observable and $R \in \mathbb{R}^{m\times m}$ is a positive definite matrix.

When A and B are accurately known, the standard LQR problem is to find the optimal policy by minimizing the value function (2) with respect to the policy u.

$$u^*(t) = -K^* x(t) \tag{3}$$

where $K^* = R^{-1}B^T P^* \in \mathbb{R}^{m\times n}$ is the optimal control gain matrix and $P^* \in \mathbb{R}^{n\times n}$ is the constant positive definite matrix solution of ARE [32]

$$A^T P^* + P^* A + Q - P^* B R^{-1} B^T P^* = 0. \tag{4}$$

Remark 1: It is obvious that solving the ARE for P^* requires knowledge of the system matrices A and B, however, in the case where information about A and B is unavailable, it is challenging to determine P^* and K^* online.

The following assumptions are required to facilitate the subsequent design.

Assumption 1: The optimal Riccati matrix P^* is upper bounded as $\|P^*\| \le \alpha_1$, where α_1 is a known positive scalar constant.

Assumption 2: The optimal gain matrix K^* is upper bounded as $\|K^*\| \leq \alpha_2$, where α_2 is a known positive scalar constant.

For the linear system in (1), the optimal value function can be written as a quadratic function [33]

$$V^*(x) = x^T P^* x. \tag{5}$$

To facilitate the development of the model-free LQR, differentiate (5) with respect to time and use system dynamics (1) to obtain

$$\dot{V}^*(x) = x^T(P^*A + A^T P^*)x + 2x^T P^* B u. \tag{6}$$

Using (4), (6) reduces to

$$\dot{V}^*(x) = x^T(-Q + P^* B R^{-1} B^T P^*)x + 2x^T P^* B u. \tag{7}$$

The LHS of (7) can be written as $\dot{V}^*(x) = \nabla V^*(x)\dot{x} = 2x^T P^* \dot{x}$ by considering (5), which is then substituted in (7) as

$$2x^T P^* \dot{x} = x^T(-Q + {K^*}^T R K^*)x + 2x^T {K^*}^T R u. \tag{8}$$

The expression in (8) acts as the fixed point equation used to define $\mathcal{D} \in \mathbb{R}$ as the difference between LHS and RHS of (8)

$$\mathcal{D} = 2x^T P^* \dot{x} - x^T(-Q + {K^*}^T R K^*)x - 2x^T {K^*}^T R u = 0. \tag{9}$$

Remark 2: The motivation behind the formulation of (9) is to represent the fixed point equation in a model-free way without using memory-intensive delayed-window integrals and subsequently design a parameter estimation algorithm to learn P^* and K^* without knowledge of system matrices A and B.

III. Optimal Control Design for Completely Unknown LTI Systems

In (9), P^* and K^* are unknown parameter matrices and the objective is to estimate these parameters using gradient-based update laws.

Let $\hat{\mathcal{D}} \in \mathbb{R}$ denote the estimate of $\mathcal{D}$ as

$$\hat{\mathcal{D}} = 2x^T \hat{P} \dot{\hat{x}} - x^T(-Q + \hat{K}^T R \hat{K})x - 2x^T \hat{K}^T R u \tag{10}$$

where $\hat{P}(t) \in \mathbb{R}^{n\times n}$, $\hat{K}(t) \in \mathbb{R}^{m\times n}$ and $\dot{\hat{x}} \in \mathbb{R}^n$ are the subsequently defined estimates of P^*, K^* and $\dot{x}$, respectively. The TD-like estimation error $\mathcal{E} \in \mathbb{R}$, from (9) and (10), can be defined as

$$\begin{aligned}\mathcal{E} &= \hat{\mathcal{D}} - \mathcal{D} \\ &= 2x^T \hat{P} \dot{\hat{x}} - x^T(-Q + \hat{K}^T R \hat{K})x - 2x^T \hat{K}^T R u.\end{aligned} \tag{11}$$

The gradient-based update laws are developed which minimize the squared error $\Xi \in \mathbb{R}$ defined as $\Xi = \mathcal{E}^2/2$. The update laws for the parameters to be estimated are given by

$$\dot{\hat{P}} = -\nu \frac{\partial \Xi}{\partial \hat{P}}$$
$$\dot{\hat{K}} = -\nu_k \frac{\partial \Xi}{\partial \hat{K}}$$

where $\nu \in \mathbb{R}^+$ and $\nu_k \in \mathbb{R}^+$ are adaptation gains. Substituting the values of gradients of Ξ with respect to $\hat{P}(t)$ and $\hat{K}(t)$, the normalized update laws are given as

$$\dot{\hat{P}} = proj\left(-2\nu x \dot{\hat{x}}^T\right)\mathcal{E} \tag{12}$$

$$\dot{\hat{K}} = \frac{2\nu_k(R\hat{K}xx^T + Rux^T)\mathcal{E}}{1 + \eta_k \omega_k^T \omega_k} \tag{13}$$

where $1/(1 + \eta_k \omega_k^T \omega_k)$ is the normalization term, $\eta_k \in \mathbb{R}^+$ is a constant gain and $\omega_k = x \otimes Ru \in \mathbb{R}^{nm}$. Further, $proj(.)$ is a smooth projection operator which ensures boundedness of the parameter estimate $\hat{P}(t)$ within a compact region in the parameter space [34], [35]. Referring to Definition 5 and Lemma 6 of [35], and using Assumptions 1 and 2, the negative semi-definite term in the $proj(.)$ always keeps the parameter estimates inside the bounded region whenever boundary condition is reached. In (12), the convex and compact region for parameter estimation is chosen as $\left\|\hat{P}\right\| \leq \alpha_1$, which is in line with Assumption 1.

The continuous policy update is given as

$$u = -\hat{K}x. \tag{14}$$

The design of the state derivative estimator $\dot{\hat{x}}(t)$, mentioned in (11) and (12), is facilitated by expressing the system dynamics (1) as linear-in-the-parameters (LIP)

$$\dot{x} = Y\theta \tag{15}$$

where $Y(x,u) \in \mathbb{R}^{n\times n(n+m)}$ is the regressor matrix and $\theta \in \mathbb{R}^{n(n+m)}$ is the unknown vector defined as

$$\theta = \begin{bmatrix} vec(A^T) \\ vec(B^T) \end{bmatrix}. \tag{16}$$

Assumption 3: The system parameter vector θ in (16) is upper bounded as $\|\theta\| \leq a_1$, where a_1 is a known positive constant.

The state derivative estimator is designed as

$$\dot{\hat{x}} = Y\hat{\theta} + L\tilde{x} \tag{17}$$

where $\hat{\theta}(t) \in \mathbb{R}^{n(n+m)}$ is the estimate of θ, $\tilde{x}(t) = x - \hat{x} \in \mathbb{R}^n$ is the state estimation error and $L \in \mathbb{R}^{n\times n}$ is the symmetric positive definite high gain matrix. Using (15) and (17), the state derivative estimation error is given as

$$\dot{\tilde{x}} = \dot{x} - \dot{\hat{x}} = Y\tilde{\theta} - L\tilde{x} \tag{18}$$

where $\tilde{\theta}(t) = \theta - \hat{\theta} \in \mathbb{R}^{n(n+m)}$ is the system parameter estimation error.

The update law for $\hat{\theta}(t)$, which minimizes the state derivative estimation error, is designed as

$$\dot{\hat{\theta}} = \Gamma Y^T \tilde{x} \tag{19}$$

where $\Gamma \in \mathbb{R}^{n(n+m)\times n(n+m)}$ is the constant positive definite gain matrix.

Lemma 1: The update laws in (17) and (19) ensure that the state estimation and the system parameter estimation error dynamics are Lyapunov stable $\forall t \geq 0$.

Proof: Consider a positive-definite Lyapunov function candidate as

$$U(\tilde{x}, \tilde{\theta}) = \frac{1}{2}\tilde{x}^T\tilde{x} + \frac{1}{2}\tilde{\theta}^T\Gamma^{-1}\tilde{\theta}. \tag{20}$$

Taking time derivative of (20) and substituting the value of $\dot{\tilde{x}}(t)$ from (18), the following expression is obtained

$$\dot{U}(\tilde{x},\tilde{\theta}) = \tilde{x}^T Y\tilde{\theta} - \tilde{x}^T L\tilde{x} - \tilde{\theta}^T\Gamma^{-1}\dot{\hat{\theta}}.$$

Substituting $\dot{\hat{\theta}}(t)$ from (19), the following expression is obtained

$$\begin{aligned}\dot{U}(\tilde{x},\tilde{\theta}) &= -\tilde{x}^T L\tilde{x}\\ &\leq -\underline{L}\left\|\tilde{x}\right\|^2 \leq 0, \quad \forall t \geq 0\end{aligned} \tag{21}$$

where $\underline{L} = \lambda_{\min}(L)$. ∎

Since $U(\tilde{x},\tilde{\theta}) > 0$ and $\dot{U}(\tilde{x},\tilde{\theta}) \leq 0$, $U(\tilde{x},\tilde{\theta})$ is bounded which implies that $\tilde{x}(t)$, $\tilde{\theta}(t)$, $\hat{\theta}(t) \in \mathcal{L}_\infty$.

Remark 3: Assumptions 1 and 2 are standard assumptions required for projection based adaptive algorithms, frequently used in robust adaptive control literature ([3], Chapter 11 of [36], Chapter 3 of [37], [38]). In fact, in the context of adaptive optimal control, analogous to Assumptions 1 and 2, many existing results [8], [11], [13], [14], [29] assume a known upper bound of the unknown parameters associated with the value function, an essential requirement for proving stability of the closed-loop system. Although the true system parameters (A and B) are unknown, a range of operating values (a compact set containing the true values of the elements of A and B) may be known in many cases from the particular domain knowledge of the plant. Performing a uniform sampling over the known compact set and solving the ARE offline with those samples, a set of Riccati matrices can be obtained, and hence, the upper bounds (α_1 and α_2), assumed in Assumptions 1 and 2, can be conservatively estimated using this set. Moreover, the proposed algorithm serves as an effective approach for the case where it is hard to obtain the initial stabilizing policy for uncertain systems.

IV. Convergence and Stability

A. Development of Controller Parameter Estimation Error Dynamics

The controller parameter estimation error dynamics for $\tilde{K}(t) = \hat{K}(t) - K^* \in \mathbb{R}^{m\times n}$ can be obtained using (11) and (13) as

$$\begin{aligned}\dot{\tilde{K}} = \frac{2\nu_k(R\hat{K}xx^T + Rux^T)}{1+\eta_k\omega_k^T\omega_k}\Big(&2x^T\hat{P}\dot{\hat{x}} - W - 2x^T\tilde{K}Ru\\ &- 2x^TK^*Ru\Big)\end{aligned} \tag{22}$$

where $W = x^T(-Q + \hat{K}^TR\hat{K})x \in \mathbb{R}$.

Using the vec operator in (22), the following expression is obtained

$$\begin{aligned}vec(\dot{\tilde{K}}) = &-4\nu_k\varphi_k\varphi_k^T vec(\tilde{K}) - 4\nu_k\varphi_k\varphi_k^T vec(K^*)\\ &+ \frac{2\nu_k\varphi_k}{\gamma}\Big(2(\dot{\hat{x}}\otimes x)^T vec(\hat{P}) - vec(W)\Big)\\ &+ \frac{2\nu_k vec(R\hat{K}xx^T)}{\gamma^2}\Big(2(\dot{\hat{x}}\otimes x)^T vec(\hat{P})\\ &- vec(W)\Big) - \frac{4\nu_k vec(R\hat{K}xx^T)\varphi_k^T}{\gamma}\\ &\times\Big(vec(\tilde{K}) + vec(K^*)\Big)\end{aligned} \tag{23}$$

where $\varphi_k(z,\,t) = (\omega_k/\gamma) \in \mathbb{R}^{nm}$ is the normalized parameter regressor vector, $\omega_k(z,t) = x\otimes Ru$, where u is substituted from (14) and $\gamma(z,\,t) = \sqrt{1+\eta_k\omega_k^T(z,\,t)\omega_k(z,\,t)} \in \mathbb{R}$ is the normalization term with $z \in \mathbb{R}^{n(1+m)}$ defined as

$$z = [x^T\ (vec(\tilde{K}))^T]^T \tag{24}$$

and

$$\|\varphi_k\| \leq \frac{1}{\sqrt{\eta_k}}. \tag{25}$$

Using (15) and (23), the system dynamics in terms of the error state $z(t)$ can be expressed as

$$\dot{z}(t) = \mathcal{F}(z,\,t)$$

where $\mathcal{F} \in \mathbb{R}^{n(1+m)}$ is a vector valued function containing the right hand sides of (15) and (23).

Assumption 4: The pair $(\varphi_k,\,\mathcal{F})$ is u-PE, i.e., PE uniformly in the initial conditions $(z_0,\,t_0)$, if for each $d > 0$, $\exists\varepsilon,\delta > 0$ such that, $\forall(z_0,\,t_0) \in B_d\times[0,\infty)$, all corresponding solutions satisfy

$$\int_t^{t+\delta}\varphi_k(z(\tau,\,t_0,\,z_0),\,\tau)\varphi_k(z(\tau,\,t_0,\,z_0),\,\tau)^T d\tau \geq \varepsilon I \tag{26}$$

$\forall t \geq t_0$ [30].

Remark 4: Since the regressor $\varphi_k(z,\,t)$ in (23) is state dependent, the u-PE condition in (26), which is uniform in initial condition, is used instead of the classical PE condition, where the regressor is only function of time and not the states, e.g., where the objective is identification (Section 2.5 of [39]).

Remark 5: In adaptive control, convergence of system and control parameter error vectors are dependent on the excitation of the system regressors. This excitation property, typically known as persistence of excitation (PE), is necessary to achieve perfect identification and adaptation. The PE condition, although restrictive in its verification and implementation, is typically imposed by using a reference input with as many spectral lines as the number of unknown parameters [40]. The u-PE condition mentioned in Assumption 4 may be satisfied by adding a probing exploratory signal to the control input [4], [8], [11], [13], [41]. This signal can be removed once the parameter estimate $\hat{K}(t)$ converges to optimal control policy and subsequently, exact regulation of the system states will be achieved. Exact regulation of the system states in presence of persistently exciting signal can also be achieved by following the method given in [42], in which the PE property is generated in a finite time interval by an asymptotically decaying "rich" feedback law.

The expression in (23) can be represented using a perturbed system as

$$vec(\dot{\tilde{K}}) = \Sigma_{\text{nom}} + \Sigma_{\text{per}} \tag{27}$$

where $\Sigma_{\text{nom}}(vec(\tilde{K}),\, t) = -4\nu_k\varphi_k\varphi_k^T vec(\tilde{K})$, represents the nominal system, and the perturbation is represented by

$$\begin{aligned}\Sigma_{\text{per}} = &-4\nu_k\varphi_k\varphi_k^T vec(K^*) + \frac{2\nu_k\varphi_k}{\gamma}\Big(2(\dot{\hat{x}}\otimes x)^T vec(\hat{P}) \\ &- vec(W)\Big) + \frac{2\nu_k vec(R\hat{K}xx^T)}{\gamma^2}\Big(2(\dot{\hat{x}}\otimes x)^T vec(\hat{P}) \\ &- vec(W)\Big) - \frac{4\nu_k vec(R\hat{K}xx^T)\varphi_k^T}{\gamma} \\ &\times \big(vec(\tilde{K}) + vec(K^*)\big).\end{aligned}$$

For each $d > 0$, the dynamics of the nominal system

$$vec(\dot{\tilde{K}}) = -4\nu_k\varphi_k\varphi_k^T vec(\tilde{K}) \tag{28}$$

can be shown to be uniformly exponentially stable $\forall(z_0,\, t_0) \in B_d \times [0, \infty)$ by using Assumption 4, (25) and Lemma 5 of [31].

Since $\Sigma_{\text{nom}}(vec(\tilde{K}),\, t)$ is continuously differentiable and the Jacobian $\frac{\partial \Sigma_{\text{nom}}}{\partial vec(\tilde{K})}$ is bounded for the nominal system (28), it can be shown, by referring to the converse Lyapunov Theorem 4.14 in [43] and definitions and results in [31], [44], that there exists a Lyapunov function $V_c(vec(\tilde{K}), t)$, which satisfies following inequalities.

$$\begin{aligned}&d_1\left\|vec(\tilde{K})\right\|^2 \le V_c(vec(\tilde{K}),\, t) \le d_2\left\|vec(\tilde{K})\right\|^2 \\ &\frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial vec(\tilde{K})}\Sigma_{\text{nom}} \le -d_3\left\|vec(\tilde{K})\right\|^2 \\ &\left\|\frac{\partial V_c}{\partial vec(\tilde{K})}\right\| \le d_4\left\|vec(\tilde{K})\right\|\end{aligned} \tag{29}$$

for some positive constants $d_1,\, d_2,\, d_3,\, d_4 \in \mathbb{R}$**.**

B. Lyapunov Stability Analysis

Theorem 1: If Assumption 4 holds, the adaptive optimal controller (14) along with the parameter update laws (12) and (13) and the state derivative estimators (17) and (19) guarantees that the system states and the controller parameter estimation errors $z(t)$ are uniformly exponentially stable $\forall t \ge 0$, provided $z(0) \in \varrho$, where the set ϱ is defined as[1]

$$\varrho = \left\{z(t) \in \mathbb{R}^{n(1+m)} \mid \|z\| < \rho^{-1}(\beta)\right\} \tag{30}$$

where $\beta = \min\left(\underline{Q},\, d_3/2\right) \in \mathbb{R}^+$, $\underline{Q} = \lambda_{\min}(Q)$ and the known function $\rho(\|z\|) : \mathbb{R} \to \mathbb{R}$, defined subsequently, is positive, globally invertible and non-decreasing.

Proof: A positive-definite, continuously differentiable Lyapunov function candidate $V_L : B_d \times [0, \infty) \to \mathbb{R}$ is defined for each $d > 0$ as

$$V_L(z,\, t) = V^*(x) + V_c(vec(\tilde{K}),\, t) \tag{31}$$

where $V^*(x)$ is the optimal value function defined in (5) which is positive definite and continuously differentiable and V_c is defined in (29). Taking the time derivative of V_L, along the trajectories of (1) and (27), the following expression is obtained

$$\dot{V}_L = \frac{\partial V^*}{\partial x}\dot{x} + \frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial vec(\tilde{K})}\Sigma_{\text{nom}} + \frac{\partial V_c}{\partial vec(\tilde{K})}\Sigma_{\text{per}}.$$

Using (6), (29) and the Rayleigh-Ritz theorem, $\dot{V}_L$ can be upper bounded as

$$\begin{aligned}\dot{V}_L \le &-\underline{Q}\left\|x\right\|^2 - u^{*T}Ru^* - 2x^TK^{*T}R(u^* - u) \\ &- d_3\left\|vec(\tilde{K})\right\|^2 + d_4\left\|vec(\tilde{K})\right\|\Sigma_{\text{per}}\end{aligned} \tag{32}$$

where $\underline{Q} = \lambda_{\min}(Q)$.

Substituting the bounds on the term $d_4\left\|vec(\tilde{K})\right\|\Sigma_{\text{per}}$ from the Appendix and expressing $d_3 = d_3/2 + d_3/2$, (32) is written as

$$\begin{aligned}\dot{V}_L \le &-\underline{Q}\left\|x\right\|^2 - \frac{d_3}{2}\left\|vec(\tilde{K})\right\|^2 + \Big[l\left\|x\right\|^2\left\|vec(\tilde{K})\right\| \\ &- \frac{d_3}{2}\left\|vec(\tilde{K})\right\|^2\Big] + \nu_k\rho_1(\|z\|)\left\|z\right\|^2\end{aligned} \tag{33}$$

where bounds are applied on the third term of the RHS of (32) as $\left\|-2x^TK^{*T}R\tilde{K}x\right\| \le l\left\|x\right\|^2\left\|vec(\tilde{K})\right\|$ using (3) and (14) with $l \in \mathbb{R}$ as the positive constant and ρ_1 is defined in (42). By completing the square on the square bracketed terms, (33) can be written as

$$\begin{aligned}\dot{V}_L \le &-\underline{Q}\left\|x\right\|^2 - \frac{d_3}{2}\left\|vec(\tilde{K})\right\|^2 + \rho_2(\|z\|)\left\|z\right\|^2 \\ &+ \frac{\rho_1(\|z\|)\left\|z\right\|^2}{\bar{\nu}}\end{aligned} \tag{34}$$

where the known function $\rho_2(\|z\|) : \mathbb{R} \to \mathbb{R}$, defined as $\rho_2(\|z\|) = 2l^2\left\|x\right\|^2/d_3$, is positive, globally invertible and non-decreasing and $\bar{\nu} = 1/\nu_k \in \mathbb{R}$. By using (24), (34) can be further expressed as

$$\dot{V}_L \le -\left(\beta - \rho(\|z\|)\right)\left\|z\right\|^2 \tag{35}$$

where $\beta = \min\left(\underline{Q},\, d_3/2\right) \in \mathbb{R}^+$ and $\rho(\|z\|) = \rho_1(\|z\|)/\bar{\nu} + \rho_2(\|z\|)$.

Using (5), (24) and (29), the Lyapunov function candidate V_L can be bounded as

$$\sigma_1\left\|z\right\|^2 \le V_L(z,\, t) \le \sigma_2\left\|z\right\|^2 \tag{36}$$

where σ_1 and σ_2 are positive constants.

Using (36), (35) can be expressed as

$$\dot{V}_L \le -\left(\beta - \rho(\|z\|)\right)\frac{V_L}{\sigma_2}. \tag{37}$$

The expression in (37) can be further upper bounded by

$$\dot{V}_L \le -\frac{\bar{\beta}V_L}{\sigma_2} \tag{38}$$

where $\bar{\beta} \in \mathbb{R}^+$ is given as

$$\bar{\beta} \le \left(\beta - \rho(\|z\|)\right), \qquad \forall z(t) \in \varrho$$

where the set ϱ is defined as

$$\varrho = \left\{z(t) \in \mathbb{R}^{n(2+2m+n)} \mid \|z\| < \rho^{-1}(\beta)\right\}.$$

[1]The initial condition region ϱ can be increased by appropriately choosing user defined matrices Q, R, and by tuning design parameters ν, ν_k and η_k.

If $z(0) \in \varrho$, then by looking at the solution of (38),

$$V_L \leq V_L(0)e^{-\frac{\bar{\beta}}{\sigma_2}t}, \qquad \forall t > 0$$

it can be said that system states and the parameter estimation errors uniformly exponentially converge to the origin. ■

Remark 6: The positive constants d_1, d_2, d_4 in (29) do not appear in the design of the control law (14) or the parameter update law (13) and are only utilized for the stability analysis purpose. As a result, knowing the exact values of these constants is not required in general. However, the quantity d_3, which appears in Theorem 1, can be determined by following the procedure given in [43] (for details see proof of Theorem 4.14 in [43]).

Remark 7: Traditionally, the parameter update laws in adaptive control have user defined design parameters termed as adaptation gains (in this paper ν and ν_k defined in (12) and (13), respectively). Typically, these gains are responsible for the convergence rate of the estimation of the unknown parameters. Hence, a careful selection of gains govern the performance of the designed estimators. However, a large value of adaptation gain may result in an unstable adaptive system, which can be overcome by introducing "normalization" in the update laws [45]. The normalized estimator in the update law (13) involves constant tunable gain η_k, which can be chosen in such a way that maintains the system stability in presence of high adaptation gain ν_k.

Remark 8: The estimates of the system matrices A and B, given by (19), are not guaranteed to converge to the optimal parameters, since Lemma 1 only proves that the parameter estimation error $\tilde{\theta}(t)$ is bounded. Therefore, solving ARE in (4) using the estimates of A and B may not yield the optimal parameter P^* and K^*. Moreover, solving P^* directly from the ARE, which is nonlinear in P^*, can be challenging, especially for large scale systems. However, the proposed method utilizes the estimates of A and B in the estimator design of the controller parameters P^* and K^*. The adaptive update laws for $\hat{P}$ and $\hat{K}$, in (12) and (13), include the identifier $\dot{\hat{x}}(t)$, which is designed in (17), and uses $\hat{\theta}(t)$ (estimates of A and B). The proposed design is architecturally analogous to [11], [13], [29], where a system identifier is utilized in controller parameter estimation. Also, note that although the system parameter estimates $\hat{A}$ and $\hat{B}$ are only guaranteed to be bounded, the controller parameter estimates $\hat{P}$ and $\hat{K}$ are proved to be exponentially convergent to the optimal parameters, as proved in Theorem 1.

C. Comparison With Existing Literature

One of the main contributions of the result is that the initial stabilizing policy assumption is not required, unlike the iterative algorithms in [5], [7], [9], [10], where an initial stabilizing policy is assumed to ensure that the subsequent policies remain stabilizing. On the other hand, an adaptive control framework is considered in the proposed approach where the control policies are continuously updated until convergence to the optimal policy. The design of the controller, the parameter update laws and the state derivative estimator ensure exponential stability of the closed-loop system which is proved using a rigorous Lyapunov-based stability analysis, irrespective of the initial control policy (stabilizing or destabilizing) chosen.

Moreover, other significant contributions of this paper with respect to some of the existing literatures are highlighted as follows.

The algorithms proposed in [5], [7], [9], [10] require computation of delayed-window integrals to construct the regressor, and/or "intelligent" data storage mechanism to satisfy an underlying full-rank condition. Computation of delayed-window integrals require past data storage for the time interval $[t - T,\ t]$, $\forall t > 0$, where t and T are the current time instant and the window length, respectively, which demands significant consumption of memory stacks, especially for large scale systems. Unlike [5], [7], [9], [10], the proposed work strategically obviates the requirement of memory intensive delayed-window integrals and "intelligent" data storage, a definite advantage in the case of large scale systems implemented on embedded hardware.

Although the result in [14] designs an actor-critic architecture based adaptive optimal controller for uncertain LTI systems, it uses memory-intensive delayed-window integral based Bellman error (see the error expression for "e" defined below (17) in [14]) to tune the critic weight estimates $\hat{W}_c$. Unlike [14], the proposed algorithm uses an online state derivative estimator to obviate the need of past data storage for control parameter estimation by strategically formulating Bellman error "$\mathcal{E}$" (11) to be independent of delayed-window integrals. Further, an exponential stability result is obtained using the proposed algorithm as compared to the asymptotic result achieved in [14].

Recent results in [28], [29] relax the PE condition by concurrently applying past stored data along with the current parameter estimates, however, unlike [28], [29], the proposed result is established for completely uncertain systems without requiring past data storage. Moreover, a stronger exponential regulation result is obtained using the proposed controller, while obviating the need of past data storage, as compared to [28], [29].

The proposed result also differs from the ALQ algorithm [3] in that it avoids the computational burden of solving the ARE (with the estimates of A and B) at every iteration, thus also avoiding the restrictive condition on stabilizability of estimates of A and B, at every iteration.

V. Simulation

To verify the effectiveness of the proposed result, the problem of controlling the angular position of the shaft in a DC motor is considered [12]. The plant is modeled as a third order continuous-time LTI system and its system matrices are given as

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 4.438 \\ 0 & -12 & -24 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 20 \end{bmatrix}.$$

The objective is to find the optimal control policy for the infinite horizon value function (2), where the state and input penalties are taken as $Q = I_3$ and $R = 1$, respectively. Solving

ARE (4) for the given system dynamics, the optimal control gain K^* is obtained as $K^* = [1.0 \quad 0.8549 \quad 0.4791]$. The gains for parameter update laws (12) and (13) are chosen as $\nu = 35$, $\nu_k = 55$ and $\eta_k = 5$. The gain matrix of the state derivative estimator is selected as $L = I_3$. An exploration signal, comprising of a sum of sinusoids with irrational frequencies, is added to the control input in (14) which subsequently leads to the convergence of control gain to its optimal values (depicted by ☆) as shown in Fig. 1.

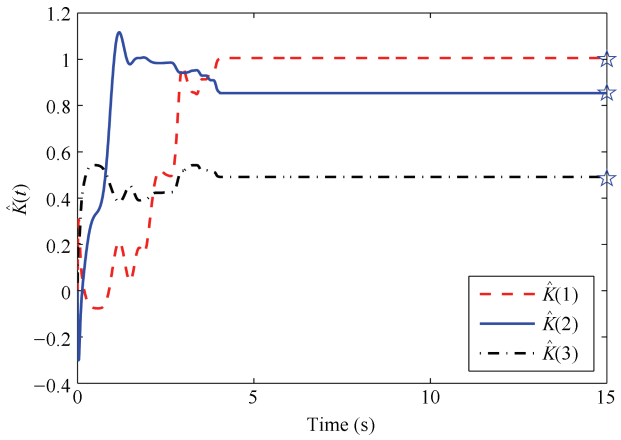

Fig. 1. The evolution of parameter estimate $\hat{K}(t)$ for the proposed method.

The proposed method is compared with the recently published work in [14]. The Q-learning algorithm proposed in [14] solves adaptive optimal control problem for completely uncertain linear time invariant (LTI) systems. The norms of the control gain estimation error $||\tilde{K}(t)||$ (used in the proposed work) and the actor weight estimation error $||\tilde{W}_a(t)||$ (as discussed in [14] and analogous to the $||\tilde{K}(t)||$) are depicted in Fig. 2.

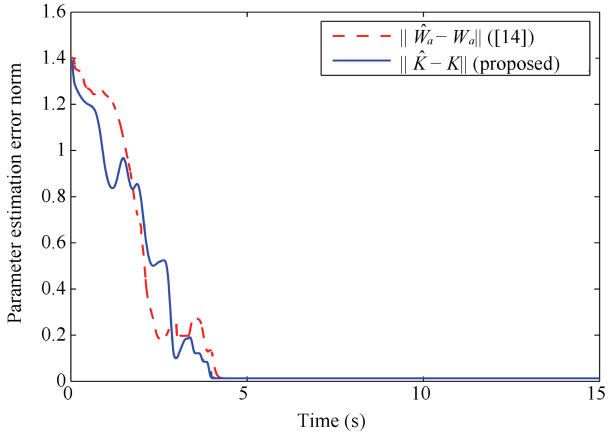

Fig. 2. Comparison of the parameter estimation error norms between [14] and the proposed method.

The initial conditions are chosen as $\hat{W}_a^T(0) = \hat{K}(0) = [0\ 0\ 0]$ and $x(0) = [-0.2 \quad 0.2 \quad -0.2]^T$, and the gains for the update laws of the approach in [14] are chosen as $\alpha_a = 6$ and $\alpha_c = 50$. To ensure sufficient excitation, an exploration noise is added to the control input up to $t = 4$ s in both cases.

From the Fig. 3, it can be observed that for similar control inputs, the convergence rates for both the methods (as shown in Fig. 2) are comparable. However, as opposed to the memory-intensive delayed-window integration for the calculation of the regressor in [14], the proposed result does not use past-stored data and hence is more memory efficient. Further, an exponential stability result is obtained using the proposed controller as compared to the asymptotic result obtained in [14]. As seen from Figs. 4 and 5, the state trajectories for both the methods initially have bounded perturbation around origin due to the presence of the exploration signal. However, once this signal is removed after $t = 4$ s, the trajectories converge to the origin.

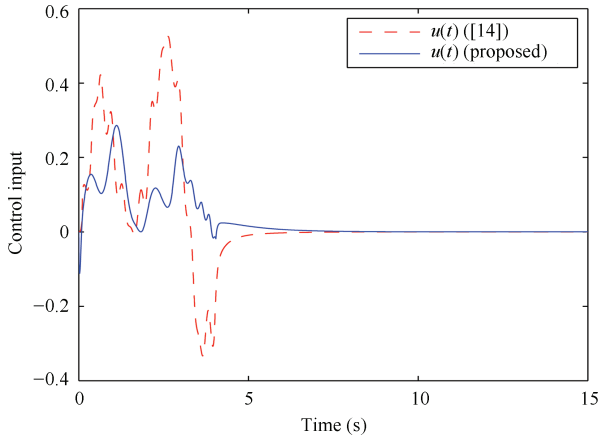

Fig. 3. Comparison of the control inputs between [14] and the proposed method.

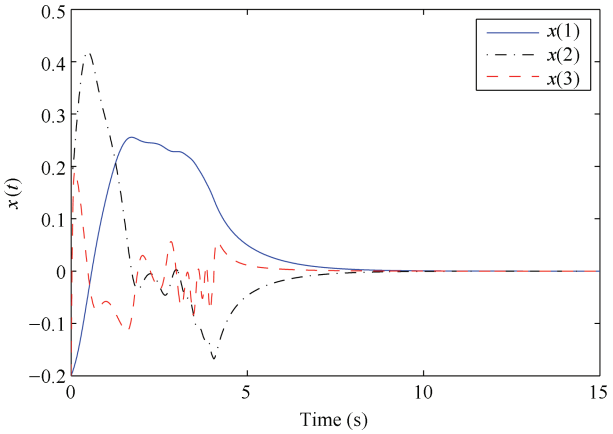

Fig. 4. System state trajectories for the proposed method.

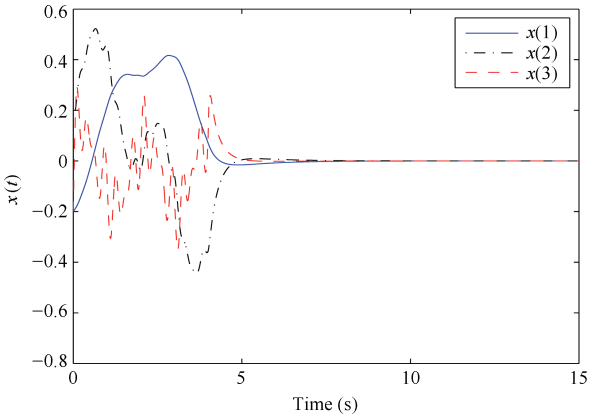

Fig. 5. System state trajectories for [14].

VI. Conclusion

An adaptive LQR is developed for continuous-time LTI systems with uncertain dynamics. Unlike previous results on adaptive optimal control which use RL/ADP methods, the proposed adaptive controller is memory/computationally efficient and does not require an initial stabilizing policy. The result hinges on a u-PE condition on the regressor vector, which is shown to be critical for proving convergence to the optimal controller. Future work will be focused on relaxing the restrictive u-PE condition without compromising the merits of the proposed result. The Lyapunov analysis is used to prove uniform exponential stability of the tracking error and parameter estimation error dynamics. Simulation results validate the efficacy of the proposed algorithm.

Appendix
Evaluation of Bound for $d_4 \left\|vec(\tilde{K})\right\| \Sigma_{\text{per}}$

This section presents bounds on different terms encountered at different stages of the proof for Theorem 1. These bounds, comprising of norms of the elements of the vector $z(t)$ defined in (24), are developed by using (13), (15), (18), (19), Lemma 1 and considering standard vec operator and Kronecker product properties.

Substituting for Σ_{per} in $d_4 \left\|vec(\tilde{K})\right\| \Sigma_{\text{per}}$ from (27), the simplified expression is given as

$$\begin{aligned} d_4 \left\|vec(\tilde{K})\right\| \Sigma_{\text{per}} = d_4 \left\|vec(\tilde{K})\right\| \Bigg[& -4\nu_k \varphi_k \varphi_k^T vec(K^*) \\ & + \frac{2\nu_k \varphi_k}{\gamma} \Big\{ 2(\dot{\hat{x}} \otimes x)^T vec(\hat{P}) \\ & - vec(W) \Big\} + \frac{2\nu_k vec(R\hat{K}xx^T)}{\gamma^2} \\ & \times \Big\{ 2(\dot{\hat{x}} \otimes x)^T vec(\hat{P}) - vec(W) \Big\} \\ & - \frac{4\nu_k vec(R\hat{K}xx^T)\varphi_k^T}{\gamma} \\ & \times \Big\{ vec(\tilde{K}) + vec(K^*) \Big\} \Bigg]. \end{aligned} \tag{39}$$

The following inequality results from the use of projection operator in (12) [35].

$$\left\|\hat{P}(t)\right\| \leq 2\alpha_1, \qquad \forall t \geq 0. \tag{40}$$

The expression in (39) is upper bounded, by using Assumptions 1 and 2, Lemma 1, (40) and the following supporting bounds

$$\|Y\| \leq \|x\| \left(1 + h_1 \left\|vec(\tilde{K})\right\|\right) \tag{41a}$$

$$\|\varphi_k\| \leq \|\omega_k\| = \left\|vec(R\hat{K}xx^T)\right\| \tag{41b}$$

$$\left\|vec(R\hat{K}xx^T)\right\| \leq \|x\|^2 \left(h_2 + h_3 \left\|vec(\tilde{K})\right\|\right) \tag{41c}$$

$$\begin{aligned} \left\|2(\dot{\hat{x}} \otimes x)^T\right\| \leq\ & h_4 \|x\|^2 \left\|vec(\tilde{K})\right\| + h_5 \|x\|^2 \\ & + h_6 \|x\| + h_7 \|x\|^2 \\ & + h_8 \|x\|^2 \left\|vec(\tilde{K})\right\| \end{aligned} \tag{41d}$$

$$\begin{aligned} \|vec(W)\| \leq\ & \|x\|^2 \left(h_9 + h_{10} \left\|vec(\tilde{K})\right\|^2 \right. \\ & \left. + h_{11} \left\|vec(\tilde{K})\right\| \right) \end{aligned} \tag{41e}$$

where $h_i \in \mathbb{R}$ for $i = 1, 2, \ldots, 11$ are positive constants and in (41b), equality expression $\omega_k = x \otimes Ru = vec(R\hat{K}xx^T)$ is used , as

$$d_4 \left\|vec(\tilde{K})\right\| \Sigma_{\text{per}} \leq \nu_k \rho_1(\|z\|) \|z\|^2 \tag{42}$$

where the known function $\rho_1(\|z\|) : \mathbb{R} \to \mathbb{R}$ is a positive, globally invertible and non decreasing and $z \in \mathbb{R}^{n(n+m)}$ is defined in (24).

References

[1] R. E. Kalman, "Contributions to the theory of optimal control," *Bol. Soc. Mat. Mexicana*, vol. 5, no. 2, pp. 102−119, 1960.

[2] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Automatic Control*, vol. 13, no. 1, pp. 114−115, 1968.

[3] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[4] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. Amer. Control Conf.*, vol. 3, 1994, pp. 3475−3479.

[5] D. Vrabie, M. Abu-Khalaf, F. L. Lewis, and Y. Wang, "Continuous-time ADP for linear systems with partially unknown dynamics," in *Proc. IEEE Int. Symp. Approximate Dynamic Programming and Reinforcement Learning*, 2007, pp. 247−253.

[6] P. Mehta and S. Meyn, "Q-learning and Pontryagin's minimum principle," in *Proc. IEEE Conf. Decision and Control*, 2009, pp. 3598−3605.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] S. Bhasin, R. Kamalapurkar, M. Johnson, and 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.

[12] S. K. Jha, S. B. Roy, and S. Bhasin, "Direct adaptive optimal control for uncertain continuous-time LTI systems without persistence of excitation," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 65, no. 12, pp. 1993−1997, 2018.

[13] 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 Trans. Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513−1525, 2013.

[14] 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.

[15] S. K. Jha, S. B. Roy, and S. Bhasin, "Data-driven adaptive LQR for completely unknown LTI systems," in *Proc. World Congr. IFAC*, 2017, pp. 4224−4229.

[16] 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.

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

[18] S. K. Jha, S. B. Roy, and S. Bhasin, "Policy iteration-based indirect adaptive optimal control for completely unknown continuous-time LTI systems," in *Proc. IEEE Symp. Adaptive Dynamic Programming and Reinforcement Learning*, 2017, pp. 1−7.

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

[20] P. J. Werbos, "Neural networks for control and system identification," in *Proc. 28th IEEE Conf. Decision and Control*, 1989, pp. 260−265.

[21] L. C. Baird, "Reinforcement learning in continuous time: advantage updating," in *Proc. IEEE World Congr. Computational Intelligence Int. Conf. Neural Networks*, vol. 4, 1994, pp. 2448−2453.

[22] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 32, no. 2, pp. 140−153, 2002.

[23] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA, USA: Athena Scientific, 2007.

[24] S. K. Jha, S. B. Roy, and S. Bhasin, "Memory-efficient filter based novel policy iteration technique for adaptive LQR," in *Proc. 2018 American Control Conf.*, 2018, pp. 4963−4968.

[25] T. Dierks and S. Jagannathan, "Online optimal control of nonlinear discrete-time systems using approximate dynamic programming," *J. Control Theory and Applications*, vol. 9, no. 3, pp. 361−369, 2011.

[26] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167−1175, 2014.

[27] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 4, pp. 943−949, 2008.

[28] K. G. Vamvoudakis, M. F. Miranda, and J. Hespanha, "Asymptotically stable adaptive-optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Trans. Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2386−2398, 2016.

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

[30] E. Panteley, A. Loría, and A. Teel, "Relaxed persistency of excitation for uniform asymptotic stability," *IEEE Trans. Automatic Control*, vol. 46, no. 12, pp. 1874−1886, 2001.

[31] A. Loría and E. Panteley, "Uniform exponential stability of linear timevarying systems: revisited," *Systems & Control Letters*, vol. 47, no. 1, pp. 13−24, 2002.

[32] F. Lewis and V. Syrmos, *Optimal Control*, 2nd ed. John Wiley & sons, INC., 1995.

[33] D. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, 1987.

[34] P. Kokotovic, M. Krstic, and I. Kanellakopoulos, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, 1995.

[35] L. Eugene, W. Kevin, and D. Howe, *Robust and Adaptive Control With Aerospace Applications*, Springer London, 2013.

[36] E. Lavretsky and K. Wise, *Robust and Adaptive Control: With Aerospace Applications*, Springer, 2013.

[37] P. Ioannou and B. Fidan, *Adaptive Control Tutorial*, SIAM, 2006.

[38] E. Lavretsky, T. E. Gibson, and A. M. Annaswamy, "Projection operator in adaptive systems," *arXiv preprint arXiv:1112.4232v6*, 2012.

[39] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*, Engle-wood Cliffs, NJ: Prentice Hall, 1989.

[40] S. Boyd and S. S. Sastry, "Necessary and sufficient conditions for parameter convergence in adaptive control," *Automatica*, vol. 22, no. 6, pp. 629−639, 1986.

[41] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32−50, 2009.

[42] G. Kreisselmeier and G. Rietze-Augst, "Richness and excitation on an interval-with application to continuous-time adaptive control," *IEEE Trans. Automatic Control*, vol. 35, no. 2, pp. 165−171, 1990.

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

[44] M. Corless and L. Glielmo, "New converse lyapunov theorems and related results on exponential stability," *Mathematics of Control, Signals and Systems*, vol. 11, no. 1, pp. 79−100, 1998.

[45] K. J. A ström and B. Wittenmark, *Adaptive Control*, Courier Corporation, 2013.

Sumit Kumar Jha (M'20) received the B. Tech. degree in electronics and communication engineering from Maulana Abul Kalam Azad University of Technology, India in 2009 and the M. Tech. degree in control systems from National Institute of Technology Kurukshetra, India in 2011. He received the Ph.D. degree in control and automation from the Department of Electrical Engineering, Indian Institute of Technology Delhi, India.

He is currently an Assistant Professor in the Department of Electronics and Communication Engineering at the Motilal Nehru National Institute of Technology Allahabad, India. His research interests include adaptive optimal control, reinforcement learning, and adaptive control.

Shubhendu Bhasin (M'08) received the Ph.D. degree in 2011 from the Department of Mechanical and Aerospace Engineering at the University of Florida, USA. He is currently an Associate Professor in the Department of Electrical Engineering at the Indian Institute of Technology Delhi, India. His research interests include reinforcement learning-based feedback control, approximate dynamic programming, neural network-based control, nonlinear system identification and parameter estimation, robust and adaptive control of uncertain nonlinear systems.