International Journal of Control, Automation, and Systems 21(8) (2023) 2718-2725
http://dx.doi.org/10.1007/s12555-022-0364-5

ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Q-learning Based Adaptive Optimal Control for Linear Quadratic Tracking Problem

Shashi Kant Sharma, Sumit Kumar Jha*, Amit Dhawan, and Manish Tiwari

Abstract: This paper describes a Q-learning based algorithm to design the linear quadratic tracker (LQT) for linear time invariant (LTI) continuous-time systems with partially unknown dynamics. The proposed approach uses a fixed-point equation in terms of Q-function in order to estimate the unknown optimal gain parameters. The fixed-point equation, which is derived by applying the Pontryagin's minimum principle in Q-learning, is based on the modified algebraic Riccati equation (ARE) for LQT problem. The online adaptation of the optimal parameters are achieved by using the gradient descent based parameter update laws by minimizing the Bellman's error term which is derived from fixed-point equation mentioned earlier. A persistence of excitation condition has been used to establish the desired optimal convergence of the estimated control parameters. Simulation results have been shown to validate the efficiency of the proposed Q-learning approach.

Keywords: Adaptive optimal control, algebraic Riccati equation, linear quadratic tracking, Q-learning.

1. INTRODUCTION

Adaptive optimal control (AOC) is a popular method in the modern control system theory. In AOC, the idea is to design the optimized controller in the presence of an uncertain system [1-13]. In some control tasks, the systems to be optimized have uncertainty in parameters at the beginning of the control operation. Such parameters need to converge to some optimum value by using adaptive techniques, otherwise they may produce inaccuracy for the control systems. The controller needs to be redesigned continuously. The main objective of adaptive optimal control is to preserve reliable optimal performance for the system with uncertain dynamics [1]. Since many real world control problems face such uncertainty in the plant parameters, adaptive optimal control based design solution is advantageous in many practical frameworks.

Initial research in AOC domain was concentrated on solving the regulation problem for partially/completely unknown systems [1-4]. In [1-3], the authors have proposed on-policy AOC approaches to find the optimal control for a system with unknown dynamics for linear quadratic regulator (LQR) design. In [4], nonlinear control design for continuous-time (CT) models has been provided. However, since the typical real world control problems require tracking controller designs, the research effort towards the controller design for the tracking AOC problems have also expedited lately [5-7].

In tracking system, the output/states of the system follows a desired output/states in an optimal sense [5,6,8,14-18]. A technique for optimal tracking control have been developed in [5] for discrete-time linear time invariant (LTI) systems with pure-feedback. In [6], reinforcement learning (RL), a popular branch of machine learning, has been used for tracking control of partially unknown continuous time LTI system. The authors in [6] have presented an iterative algorithm based on augmented LQR for linear quadratic tracking (LQT) problem. In [7], the standard solution for the LQT problem have been explained with the help of stored data. The authors in [8] have proposed an integral RL based techniques in which LQT problem is solved by converting LQT design to an augmented LQR design.

The algorithms proposed in [1,4,6,9,11] require initial stabilizing control policy in order to solve AOC problems and establish the convergence of the uncertain optimal control parameters to their respective ideal values. However, it may not be feasible to obtain the desired initial stabilizing control policy for a system with unknown dynamics. Moreover, the iterative policy update in methods proposed in [1,4,6,9] results in an off-policy approach which leads to a controller design whose operation is neither fully continuous-time nor fully discrete-time. This hybrid controller structure may lead to a potentially unstable

Manuscript received April 26, 2022; revised November 16, 2022; accepted January 9, 2023. Recommended by Associate Editor Yajuan Liu under the direction of Senior Editor PooGyeon Park.

Shashi Kant Sharma is with the Department of Electronics and Communication Engineering, MNNIT Allahabad, Prayagraj-211004, India (e-mail: shashish1004@gmail.com). Sumit Kumar Jha, Amit Dhawan, and Manish Tiwari are with the Department of Electronics and Communication Engineering, MNNIT Allahabad, Prayagraj-211004, India (e-mails: {sumit-k, dhawan, mtiwari}@mnnit.ac.in).
* Corresponding author.

system performance.

A typical solution to LQT problem consists of two terms; a feedback term which depends on the system states variable and a feedforward term which depends on the reference trajectory. The concept of Q-learning has been used in this paper to solve LQT optimization [5,9,10]. It is a very popular mathematical tool of RL domain, in which the desired optimal control can be obtained in the absence of sufficient knowledge of system dynamics [19-21]. Q-learning uses Q-function to form a fixed point equation in terms of estimated unknown system parameters. Q-learning was initially introduced for discrete-time systems [5,10]. It is more reliable for discrete-time systems due to its recursive nature. However, it can be used for continuous-time systems as well with some modifications in terms of finding a suitable Q-function and the corresponding iterative relation [2,9,22]. In initial research and developments, Q-learning was suitable for linear quadratic regulator problems for CT systems [2,9,22].

In this paper, a Q-learning based online adaptive optimal controller has been developed based on the continuous updation of the estimated parameters, which converges to their respective ideal optimal values for the given LQT problem without the knowledge of system internal matrix. Based on the developments in [2], an on-policy method has been established to design the proposed controller. The persistence of excitation (PE) condition is required for the gradient-based update laws to achieve the online adaptation of unknown parameters in the Q-function.

The proposed algorithm obviate the need of an initial stabilizing control policy unlike the previous Q-learning based algorithm proposed in the literature for solving LQT problem. Further, unlike the off-policy design approach (which leads to a hybrid controller design) found in [1,4,6,9], the proposed algorithm is an on-policy mechanism (explained in Subsection 4.2) where a completely continuos updation of control parameter estimates eliminates any undesirable discontinuity in controller working and hence leads to desired system performance. Contrary to the algorithms discussed in [1,5-7,23,24], the proposed method does not involve any delayed-window integral terms, which may lead to low memory requirement for controller implementation.

The entire paper is organized as follows: Section 2 briefly talks about the LQT problem and its standard solution. Section 3 gives the concept of Q-learning and Q-function in brief which are to be used in next sections. Section 4 presents the proposed solution of LQT optimization problem using Q-learning by converging the parameter estimates to their optimal values. Section 5 provides a simulation example to verify the effectiveness of the proposed Q-learning algorithm. Finally, Section 6 concludes this paper.

2. CONTINUOUS-TIME LINEAR QUADRATIC TRACKING

Consider a continuous-time LTI system as

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

$$y(t) = Cx(t), \tag{2}$$

where $x(t) \in \mathbb{R}^{n\times 1}$, $y(t) \in \mathbb{R}^{p\times 1}$, and $u(t) \in \mathbb{R}^{m\times 1}$ are the system state, the output and the input vectors, respectively. $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$ are drift dynamics, input matrix and output matrix, respectively. The pair (A,B) are assumed to be controllable.

For a typical LQT problem, the performance index in infinite-horizon is given as [4]

$$J(x,y_d) = \frac{1}{2}\int_{t_0}^{\infty} [(Cx-y_d)^T Z(Cx-y_d) + u^T Ru] d\tau, \tag{3}$$

where $Z \geq 0$ and $R > 0$ are the weight matrices of appropriate dimensions for the system states and control input, respectively. $y_d(t) \in \mathbb{R}^{p\times 1}$ is the desired trajectory to be tracked by the plant output.

The standard solution for LQT problem is given as [14, 15]

$$u = -R^{-1}B^T Sx + R^{-1}B^T v, \tag{4}$$

where S is the solution of ARE given as below

$$A^T S + SA - SBR^{-1}B^T S + W = 0,$$

where W is the matrix defined as $W \triangleq C^T ZC \in \mathbb{R}^{n\times n}$, and the variable v satisfies

$$-\dot{v} = \left(A - BR^{-1}B^T S\right)^T v + C^T Z y_d.$$

The control input in (4) is an affine state feedback. It consists of two terms; the first one is the feedback control part which depends on the system state and second part is the feedforward control element which depends on the desired trajectory. Due to the unknown system internal dynamics and presence of a time varying variable v in the feedforward part, the LQT problem becomes very difficult to solve. Hence, this paper proposes a Q-learning based AOC startegy to achieve desired tracking for a system with partially unknown dynamics.

3. Q-LEARNING

The concept of Q-learning was introduced by Watkins in 1989 [19]. Q-learning is a branch of reinforcement learning which provides a solution for optimal control of linear continuous systems and it does not require the specific model for the system [19-21]. The action value function (known as Q-function) in optimal condition for a continuous-time system is given as [21]

$$Q^*(x,u) = c(x,u) + \gamma(V^*(x^{'})), \tag{5}$$

where $x(t)$ and $u(t)$ refer to the state input and control input and $x'(t)$ denotes next state of the system. The discount factor γ $(0 \leq \gamma \leq 1)$ is used for best estimation for next step. The term $c(x(t),u(t))$ in R.H.S. of (5) refers to the initial cost function/immediate reward given as

$$c(x,u) = x^TZx + u^TRu.$$

The second term in R.H.S of (5) (i.e., $\gamma(V^*(x^{'}))$ denotes the total return/reward for taking an optimal policy thereafter.

In order to obtain the optimal Q-function, the following optimal control policy will be required [2]

$$u^*(x) = \arg\min_u Q^*(x,u) \ . \tag{6}$$

The relation between Hamiltonian using Pontryagin's minimum principle and the Q-function is expressed as [22]

$$Q^*(x,u) = c(x,u) + \frac{\partial V^*(x)}{\partial x}\dot{x}. \tag{7}$$

The relation between Hamiltonian using Pontryagin's minimum principle and the Q-function (7) is expressed in [22], which subsequently paved the path for the research proposed in the current manuscript.

4. ADAPTIVE OPTIMAL CONTROL DESIGN FOR LQT PROBLEM

4.1. Formulation and methodology

Assume the desired trajectory $y_d(t)$ has been produced by the command generator system whose dynamics is given as

$$\dot{y}_d = Gy_d, \tag{8}$$

where $G \in \mathbb{R}^{p\times p}$ denotes a constant output matrix.

Remark 1: It is assumed that the command generator system in (8) generates a bounded trajectory for tracking in order to achieve the tracking behavior from the proposed algorithm. These bounded trajectories can be ensured by keeping the constant matrix G stable or marginally stable. For example, the command generator system can generate a step function or sinusoid function etc.

Define an augmented system state $X(t)$ as

$$X(t) \triangleq [x(t)^T \;\; y_d(t)^T]^T \in \mathbb{R}^{(n+p)\times 1}.$$

Combining both (1) and (8) together, the augmented system is obtained as

$$\begin{aligned}\dot{X} &= \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & G \end{bmatrix} X + \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} u \\ &= A_1X + B_1u, \end{aligned} \tag{9}$$

where $A_1 \triangleq \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & G \end{bmatrix} \in \mathbb{R}^{(n+p)\times(n+p)}$ and $B_1 \triangleq \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+p)\times m}$ are the augmented system and control input matrices of appropriate dimensions, respectively [6]. The new value function can be expressed by using the augmented system states as shown below [6,14]

$$V\,(X(t)) = X(t)^T P\, X(\mathrm{t}). \tag{10}$$

The control input for the augmented system (9) can be expressed in terms of state and desired trajectory as [6]

$$u(t) = Kx(t) + K^{'}y_d(t) = K_aX(t),$$

where $K_a \in \mathbb{R}^{m\times(n+p)}$ is the augmented gain matrix for the augmented system (9) and is defined as $K_a \triangleq [K \;\; K^{'}]$, where $K \in \mathbb{R}^{m\times n}$ and $K^{'} \in \mathbb{R}^{m\times p}$.

Minimization of the value function (10) with respect to $u(t)$ results in the optimized control policy expressed with respect to an unknown matrix $P \in \mathbb{R}^{(n+p)\times(n+p)}$ which is constant and positive definite.

Hence, the optimal control input becomes [6]

$$u^*(t) = K_aX(t), \tag{11}$$

where $K_a = -R^{-1}B_1{}^TP$ is the augmented optimal gain matrix, and P satisfies the augmented linear quadratic tracking Riccati equation (inspired from [6])

$$A_1{}^TP + PA_1 - \gamma P + PB_1R^{-1}B_1{}^TP + W_1 = 0,$$

where $W_1 \triangleq C_1{}^TZC_1 \in \mathbb{R}^{(n+p)\times(n+p)}$ and $C_1 \triangleq [C - I]$.

Solving the above augmented ARE equation without the knowledge of the augmented drift dynamics (i.e., A_1) of the augmented system in (9), is a difficult task. This paper proposes a Q-learning based method that can be used to solve such type of problems.

An optimal tracking system can be extended based on the LQR [7]. In the following, a Q-function based approach is proposed to solve the augmented LQT Riccati equation. Based on (7), the optimal Q-function for the LQT problem can be given as

$$Q^*(X,u) = c(X,u) + \frac{\partial V^*(X)}{\partial X}\dot{X}, \tag{12}$$

where $V^*(X)$ is the optimal value function derived by substituting the optimal policy (11) in (10).

The fixed point equation [22] in terms of the Q-function can be obtained from (12) as

$$\left(\frac{\partial}{\partial X}\min_u (Q^*(X,u))\right)^T \dot{X} = \gamma(Q^*(X,u) - c(X,u)). \tag{13}$$

Solving for (12), we get

$$\begin{aligned} Q^*(X,u) = {} & X^T(t)W_1X(t) + u^T(t)Ru \\ & + 2X^TP(A_1X + B_1u), \end{aligned}$$

which can be written as

$$\begin{aligned} Q^*(X,u) = {} & X^T(t)W_1X(t) + u^T(t)Ru + 2X^TP_{A_1}X \\ & + 2X^TP_{B_1}u, \end{aligned} \tag{14}$$

where $P_{A_1} \triangleq P.A_1$ and $P_{B_1} \triangleq P.B_1$ are two unknown matrices introduced in (14) to reduce the complexity of the expression in (14).

Using (14), the RHS of (13) is given as

$$\gamma(Q^*(X,u) - C(X,u)) = \gamma(2X^TP_{A_1}X + 2X^TP_{B_1}u). \tag{15}$$

For obtaining optimal Q-function, the optimal control policy is substituted in (14)

$$\begin{aligned} Q^*(X) &= \min_u (Q^*(X,u)) \\ &= X^TW_1X + X^TP_{B_1}R^{-1}{P_{B_1}}^TX + 2X^TP_{A_1}X \\ &\quad - 2X^TP_{B_1}R^{-1}{P_{B_1}}^TX) \\ &= X^T(W_1 - P_{B_1}R^{-1}{P_{B_1}}^T + 2P_{A_1})X. \end{aligned} \tag{16}$$

To get the LHS of the fixed point equation (13), taking partial differentiation on both sides of (16)

$$\begin{aligned} \frac{\partial Q^*(X)}{\partial X} = {} & (W_1 - P_{B_1}R^{-1}{P_{B_1}}^T + 2P_{A_1})^TX \\ & + (W_1 - P_{B_1}R^{-1}{P_{B_1}}^T + 2P_{A_1})X. \end{aligned} \tag{17}$$

The fixed point equation, using (15) and (16), becomes

$$\begin{aligned} & X^T\left(W_1 - P_{B_1}R^{-1}{P_{B_1}}^T + {P_{A_1}}^T + P_{A_1}\right)\dot{X} \\ & = \gamma X^TP_{A_1}X + \gamma X^TP_{B_1}u. \end{aligned} \tag{18}$$

Taking the terms of (18) on one side and defining it as a new scalar value $\Phi \in \mathbb{R}$ as

$$\begin{aligned} \Phi \triangleq {} & X^T\left(W_1 - P_{B_1}R^{-1}{P_{B_1}}^T + {P_{A_1}}^T + P_{A_1}\right)\dot{X} \\ & - \gamma X^TP_{A_1}X - \gamma X^TP_{B_1}u = 0. \end{aligned} \tag{19}$$

Since P_{A_1} and $P_{B_1}$are unknown, the estimation of (19) results in $\widehat{\Phi}$, which can be expressed as

$$\begin{aligned} \widehat{\Phi} \triangleq {} & X^T\left(W_1 - \widehat{P_{B_1}}R^{-1}\widehat{P_{B_1}}^T + \widehat{P_{A_1}}^T + \widehat{P_{A_1}}\right)\dot{\hat{X}} \\ & - \gamma X^T(\widehat{P_{A_1}}X + \widehat{P_{B_1}}u), \end{aligned} \tag{20}$$

where $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$ are the estimates of P_{A_1} and P_{B_1}, respectively.

The estimation error between (19) and (20) is given as

$$\begin{aligned} \widetilde{\Phi} &= \widehat{\Phi} - \Phi \\ &= X^T\left(W_1 - \widehat{P_{B_1}}R^{-1}\widehat{P_{B_1}}^T + \widehat{P_{A_1}}^T + \widehat{P_{A_1}}\right)\dot{\hat{X}} \\ &\quad - \gamma X^T(\widehat{P_{A_1}}X + \widehat{P_{B_1}}u). \end{aligned}$$

The parameter estimates are updated using minimization of mean squared error $\varepsilon \in \mathbb{R}$ given below

$$\varepsilon = \frac{1}{2}\widetilde{\Phi}^2.$$

The update laws for the parameter estimates $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$, based on gradient based update laws, mean squared error and further evaluations, are given as

$$\begin{aligned} \dot{\widehat{P_{A_1}}} &= -\eta_{A_1}\frac{\partial \varepsilon}{\partial \widehat{P_{A_1}}} \\ &= proj\big(-\eta_{A_1}\big(\dot{\hat{X}}X^T + X\dot{\hat{X}}^T - \gamma XX^T\big)\widetilde{\Phi}\big), \end{aligned} \tag{21}$$

$$\begin{aligned} \dot{\widehat{P_{B_1}}} &= -\eta_{B_1}\frac{\partial \varepsilon}{\partial \widehat{P_{B_1}}} \\ &= proj\big(-\eta_{B_1}\big(\big(\dot{\hat{X}}X^T + X\dot{\hat{X}}^T\big)\widehat{P_{B_1}}R^{-1} + \gamma Xu^T\big)\widetilde{\Phi}\big). \end{aligned} \tag{22}$$

Smooth projection operator has been used to ensure the boundedness of the parameter estimates of unknown parameters [25].

The estimated value of the optimal control policy in (11) can be obtained in terms the parameter estimates as

$$u(t) = -R^{-1}\widehat{P_{B_1}}^TX(t). \tag{23}$$

In order to compute the parameter estimates $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$ from the update laws (21) and (22), the value of state derivative estimator $\dot{\hat{X}}$ is required. However, since the the augmented system dynamics (i.e., $\dot{X}$) in (9) is partially unknown, it necessary to design a state derivative estimator. This subsection presents a state derivative estimator in terms of state estimation error.

Consider a linear parameterized representation of the augmented system dynamics in (9) as

$$\dot{X} = Y\theta, \tag{24}$$

where $Y(X,u) \in \mathbb{R}^{(n+p)\times((n+p)^2+(n+p)m)}$ is defined as system regressor and $\theta \in \mathbb{R}^{(n+p)^2+(n+p)m\times 1}$ is defined as unknown vector comprises of elements of A_1 and B_1 as shown below

$$\theta = \begin{bmatrix} vec\ A_1^T \\ vec\ B_1^T \end{bmatrix},$$

where $vec(M) \in \mathbb{R}^{ab}$ denotes the vectorization of a matrix $M \in \mathbb{R}^{a\times b}$ and is obtained by stacking columns of the matrix Z.

Hence, a state derivative estimator is designed below based-on (24), the estimate $\widehat{\theta} \in \mathbb{R}^{(n+p)^2+(n+p)m\times 1}$ and the state estimation error $\tilde{X} \triangleq (X - \hat{X}) \in \mathbb{R}^{(n+p)\times 1}$ as

$$\dot{\hat{X}} = Y\widehat{\theta} + \eta\tilde{X}, \tag{25}$$

where $\eta \in \mathbb{R}^{(n+p)\times(n+p)}$ known gain matrix.

Further, the computation of $\hat{X}$ requires $\widehat{\theta}$ which is not available (as θ is unknown), hence an update law for $\widehat{\theta}$ is deigned below

$$\dot{\widehat{\theta}} = \Psi Y^T \widehat{\theta}, \tag{26}$$

where $\Psi \in \mathbb{R}^{((n+p)^2+(n+p)m)\times((n+p)^2+(n+p)m)}$ is a constant gain matrix to facilitate convergence of $\widehat{\theta}$ to θ.

Remark 2: The parameter estimates $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$are continuously updated online using update laws (21), (22), (25) and (26) online and these updated value are used to obtain updated policy by using (23). The same updated policy is then used in the system dynamics (9) at the same time instant and this continuous process makes the current approach as an on-policy mechanism. This continuous on-policy method eventually leads to the optimal control policy which would optimize the performance index.

The steps involved in the above proposed algorithm can be listed as follows:

Step 1: An augmented system dynamics is defined in (9), which comprises of original system states $x(t)$ and command generator variable $y_d(t)$ as augmented system's states.

Step 2: A new value function $V(X)$ is expressed for the augmented system (9) in (10), where $P \in R^{(n+p)\times(n+p)}$ is an unknown, constant and positive definite matrix [6,14].

Step 3: The augmented linear quadratic tracking Riccati equation [6] (expressed below (11)) is expressed in terms of augmented system parameters A_1, B_1, and the augmented Riccati matrix P as

$$A_1{}^T P + PA_1 - \gamma P + PB_1 R^{-1} B_1{}^T P + W_1 = 0.$$

Step 4: A fixed point equation (13) is formed with respect to the Q-function [22] defined in (12).

Step 5: A final expression in (18) is obtained by substituting proper values in (13), which leads to a new scalar quantity Φ, defined in (19).

Step 6: The update laws for the parameter estimates $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$ are given in (21) and (22), respectively, based on the estimation error value $\widetilde{\Phi}$ (derived from (20)).

Step 7: Finally, the estimated control policy is given by (23) in terms of the control parameter estimate obtained from (21), (22), (25) and (26). Steps 6 and 7 will be computed at every time instant to achieve the optimal policy.

Remark 3: The estimated optimal control policy in (23) can be obtained at every time-instant by solving the update laws in (21), (22), (25) and (26). This continuous updation of the control policy eventually leads to the optimal control policy. Hence the proposed AOC algorithm is an on-policy approach for the controller design. Moreover, a PE condition is required for the convergence of the estimated parameters in (21) and (22) to their respective optimal values [2,10,26]. The PE condition can be typically satisfied by adding an exploratory signal to the input to the closed-loop system [2,10]. However, the continuous exposure to the PE signal can introduce oscillations in the system states, hence the PE signal will be removed from the input as soon as desired parameter convergence is achieved.

4.2. Comparative analysis

The proposed result solves adaptive LQT problem by applying a Q-learning algorithm to partially unknown continuous-time linear systems. One of the main features of the proposed method is that it does not require an initial stabilizing policy in order to achieve the convergence of the estimated control parameters towards their respective optimal/ideal values, in contrast to many algorithms proposed in the literature to solve a typical adaptive LQT/LQR design criteria [1,5-7,9,23,24]. Most of these algorithms requires an assumption of an initial stabilizing control policy to achieve the desired convergence of the estimated control parameters. An initial stabilizing control policy is typically required to ensure the initial convergence and stability of the closed-loop system, absence of which may destroy the promise of overall stability of the said system. The assumption of ensuring initially stabilized control gain is not feasible to achieve in real scenarios due to the unavailability of the knowledge of system dynamics. On the other hand, the proposed approach uses the control parameter update laws (21), (22), and adaptive controller (23) to obtain the desired adaptive LQT controller by continuously updation process, while obviating the need of a cumbersome assumption of initial stabilizing control.

Moreover, the iterative policy update in algorithms proposed in [1,4,6,9] lead to a controller which operation is neither fully continuous-time nor fully discrete-time.

Other significant contributions of this paper with respect to some of the existing literatures are highlighted as follows:

The algorithms proposed in [1,5-7,23,24] require evaluation of delayed-window integrals terms (duly defined in [27]) to construct the regressor, and/or "intelligent" data capture/storage mechanism to satisfy an underlying full-rank condition [27]. Computation of such delayed-window integrals terms need previous data storage, which needs significant consumption of memory stacks, especially for higher order systems. In contrast to the above-mentioned algorithms, the proposed Q-learning based result avoids memory expensive data storage mechanism, a precise edge in the case of higher dimensional systems.

Furthermore, the algorithms proposed in [1,5-7,9,23,24,28,29] are elegant approaches to solve complex nonlinear adaptive control problems. However, these algorithms bank on off-policy iterations to achieve the desired controller properties. These methods involve large amount of data collection, processing and storage in order to achieve the ideal adaptation of unknown parameters, which falls in the category of off-policy control algorithm. In an off-policy approach, the controller is updated through the recursive iteration in offline manner and the updated policy is applied to the closed-loop system in the next step. On the contrary, the Q-learning based algorithm proposed in the revised manuscript is an on-policy method in which the estimated policy is continuously updated at every time-instant by using gradient-descent update laws (21) and (22), and the updated policy is applied to the original system in (1) at the same time instant to obtain the state information. A typical on-policy approach largely avoids discontinuous control policy update in terms of iterations and hence bypasses the related computational burden resulting from the iterations.

5. SIMULATION RESULTS

An example of a 3-dimensional state space system is provided in this section to validate the efficiency of the proposed Q-learning method for the solution of LQT problem. The system problem is based on the adaptive LQT controller design for the speed control of a shaft of a DC motor. The considered system is modelled as a third order continuous-time linear system as follows [30]:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 4.438 \\ 0 & -12 & -24 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ 20 \end{bmatrix} u(t),$$

$$y(t) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(t), \; \dot{y_d} = 0,$$

with an initial amplitude $y_d(0) = 2$.
The weight matrices are given as

$$W_1 = I_4, \; R = 1.$$

The simulation is carried out for the discount factor $\gamma = 0$ for the adaptive convergence of parameter estimates to their optimal values. A PE signal is applied to the input as sinusoidal signal with irrational frequencies. After achieving the optimal policy, the PE signal is detached and system output converges to the reference trajectory and states to the origin. Figs. 1 and 2 show the convergence of parameter estimates $\widehat{P_{A_1}}$ and $\widehat{P_{B_1}}$, respectively, to their respective optimal values. Fig. 3 shows the simulation of system output versus reference trajectory. It is observed that the system output efficiently tracks the reference trajectory once the optimal control policy is achieved. Fig. 4 shows the convergence of the opyimal control policy to the desired value.

The simulation results further highlights the novelty of the proposed method. It is clear from the evolution of the parameter estimation curves in Figs. 1 and 2 that the parameter convergence is quicker. This happens due to the uniqeness of the proposed method in terms of applied on-policy approach for the estimation of unknown optimal control parameters and obviation of memory expensive

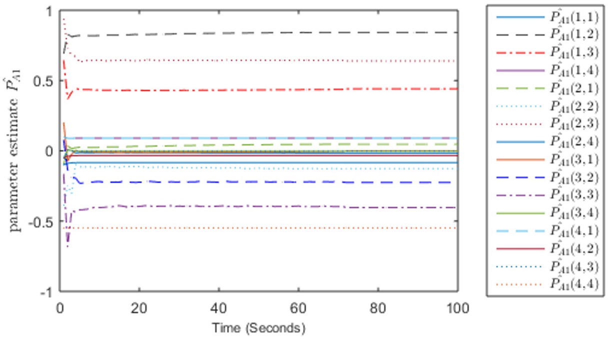

Fig. 1. Convergence of parameter estimate $\widehat{P_{A_1}}$.

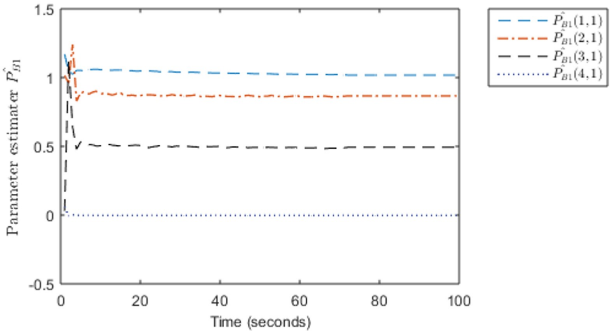

Fig. 2. Convergence of parameter estimate $\widehat{P_{B_1}}$.

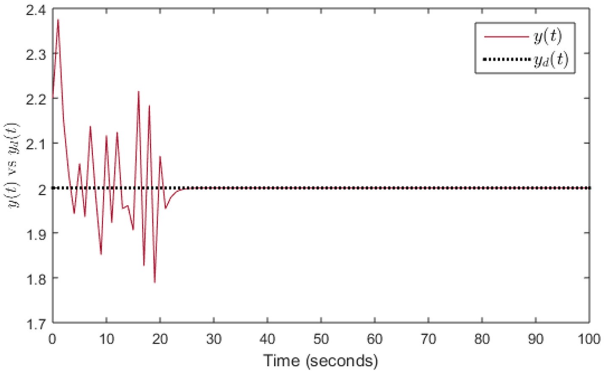

Fig. 3. Output ($y(t)$) vs desired trajectory ($y_d(t)$).

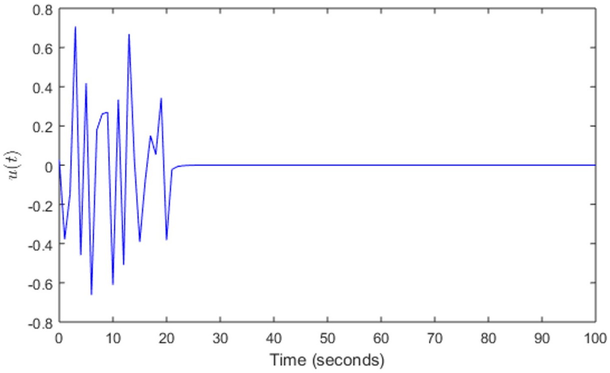

Fig. 4. Convergence of optimal control policy $u(t)$.

delayed-window integrals (explained in Subsection 4.2). Some of the method proposed in the past literature result in discontinuous or hybrid controllers [1,4,6,9], which may lead to unstable performance. In contrast to that, the proposed algorithm results in a LQT controller whose operation is continuous, resulting in a stable closed-loop performance and smooth convergence of graphs in Figs. 1-4. Further, similar argument can be given for the faster tracking of the desired trajectory by the system output. The initial oscillation are present in these graph due the inclusion of the PE signal as mentioned above. The optimal control convergence graph shown in Fig. 4 follows the same pattern.

6. CONCLUSION

In this paper, an online learning algorithm has been presented to solve the standard linear quadratic tracking problem with partially unknown system dynamics. The algorithm uses Q-learning approach to find the solution of the augmented algebraic Riccati equation. The simulation result validates the optimal convergence of the estimates of the unknown optimal, and establishes that the output from the closed-loop system tracks the reference trajectory as desired. This verifies that the proposed algorithm works with adequate tracking behavior.

CONFLICT OF INTEREST

All authors declare that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or contents discussed in this manuscript. No funding/grant was received to assist with the preparation of this submitted manuscript.

REFERENCES

[1] 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.

[2] S. K. Jha and S. Bhasin, "On-policy q-learning for adaptive optimal control," *Proc. of IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, Florida, USA, 2014.

[3] S. K. Jha, S. B. Roy, and S. Bhasin, "Policy iteration-based indirect adaptive optimal control for completely unknown continuous-time LTI systems," *Proc. of IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, Hawaii, USA, pp. 1-7, 2017.

[4] D. Vrabie and F. L. Lewis, "Adaptive optimal control algorithm for continuous-time nonlinear systems based on policy iteration," *Proc. of the IEEE Conference on Decision and Control*, pp. 73 - 79, 2009.

[5] 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.

[6] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3051-3056, 2014.

[7] D. Hou, J. Na, G. Gao, and G. Li, "Data-driven adaptive optimal tracking control for completely unknown systems," *Proc. of IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, Enshi, China, pp. 1039-1044, 2018.

[8] O. Park, H. Shin, and A. Tsourdos, "Linear quadratic tracker with integrator using integral reinforcement learning," *Proc. of Workshop on Research, Education and Development of Unmanned Aerial Systems (RED UAS)*, Cranfield, UK, pp. 31-36, 2019.

[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] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," *Proc. of American Control Conference* , vol. 3, pp. 3475-3479, 1994.

[11] Y. Du, B. Jiang, and Y. Ma, "Policy iteration based online adaptive optimal fault compensation control for spacecraft," *International Journal of Control, Automation, and Systems*, vol. 19, pp. 1607-1617, 2021.

[12] P. Wang, Z. Wang, and Q. Ma, "Adaptive event triggered optimal control for constrained continuous-time nonlinear systems," *International Journal of Control, Autiomation, and Systems*, vol. 20, pp. 857-868, 2022.

[13] Y. Xin, Z. C. Qin, and J. Q. Sun, "Robust experimental study of data-driven optimal control for an underactuated rotary flexible joint," *International Journal of Control, Autiomation, and Systems*, vol. 18, pp. 1202-1214, 2020.

[14] F. L. Lewis, D. Vrabie, and V. Syrmos, *Optimal Control*, 3rd ed., John Wiley & Sons, New Jersey, 2012.

[15] D. Subbaram Naidu, *Optimal Control Systems*, CRC Press, 2002.

[16] Z.-M. Li and J. H. Park, "Dissipative fuzzy tracking control for nonlinear networked systems with quantization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 12, pp. 5130-5141, 2020.

[17] Z.-M. Li, X.-H. Chang, and J. H. Park, "Quantized static output feedback fuzzy tracking control for discrete-time nonlinear networked systems with asynchronous event-triggered constraints," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 3820-3831, 2021.

[18] C. Han and W. Wang, "Optimal LQ tracking control for continuous-time systems with pointwise time-varying input delay," *International Journal of Control, Autiomation, and Systems*, vol. 15, pp. 2243-2252, 2017.

[19] C. J. C. H. Watkins, *Learning from Delayed Rewards*, Ph.D. dissertation, Cambridge University, Cambridge, England, 1989.

[20] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279-292, 1992.

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

[22] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," *Proc. of IEEE Conrerence on Decision and Control*, pp. 3598-3605, 2009.

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

[24] 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.

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

[26] 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.

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

[28] D. Zhang, Z. Ye, G. Feng, and H. Li, "Intelligent event-based fuzzy dynamic positioning control of nonlinear unmanned marine vehicles under DoS attack," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13486-13499, 2022.

[29] Z. Ye, D. Zhang, Z.-G. Wu, and H. Yan, "A3C-based intelligent event-triggering control of networked nonlinear unmanned marine vehicles subject to hybrid attacks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12921-12934, 2022.

[30] C. Edwards and S. Spurgeon, *Sliding Mode Control: Theory and Applications*, CRC Press, London, UK, 1998.

Shashi Kant Sharma received his M.Tech. degree in electronics & communication engineering from Motilal Nehru National Institute of Technology Allahabad, India. His research interests include signal processing, adaptive control, and adaptive optimal control.

Sumit Kumar Jha received his M.Tech. degree in electrical engineering from National Institute of Technology Kurukshetra, India and a Ph.D. degree from IIT Delhi, India. He is currently an Assistant Professor with 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, adaptive control, and signal processing.

Amit Dhawan received his bachelor's degree in electronics and communication engineering from Birla Institute of Technology, Mesra, Ranchi, a master's degree in control & instrumentation, and a Ph.D. degree in electronics and communication engineering from Motilal Nehru National Institute of Technology, Allahabad, India. At present, he is a Professor with the Department of Electronics and Communication Engineering, MNNIT, Allahabad. His current research interests include digital signal processing, robust stability, guaranteed cost control, delayed systems, and multidimensional systems.

Manish Tiwari is an Associate Professor with the Department of Electronics & Communication Engineering, Motilal Nehru National Institute of Technology Allahabad, Uttar Pradesh, India. He has received the first class degrees of B.E. in electronics engineering in 1996 and an M. E. in electronics and controls from Birla Institute of Technology Pilani, India in 1999. He has received a Ph.D. degree from MNNIT Allahabad, India. His research interests include microprocessor based system design, embedded system, digital filter architecture design, multidimensional system design, and signal processing.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.