2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

A Q-Learning Approach to Model-Free Infinite Horizon Control For Linear Time Delay Systems

Zineb Benhmidouch[1], Sadek Belamfedel Alaoui[2], Ahmed Abbou[1], Adnane Saoud[2]

Abstract— **In this paper, an online Q-learning algorithm is proposed to address the infinite-horizon guaranteed cost control problem for linear time delay systems with completely unknown dynamics. The developed approach leverages a Lyapunov-Krasovskii functional as the state value function and integrates guaranteed cost control principles. Specifically, based on Bessel-Legendre integral inequality, a Q-function tailored for handling guaranteed cost control in time delay systems is formulated. Furthermore, an integral reinforcement learning method based on an actor/critic approximator framework is used to dynamically estimate the Q-function parameters. Finally, the proposed approach is successfully applied to an interconnected power system.**

I. Introduction

Time delays are common in various aspects of daily life, particularly in systems requiring processing time, communication, as well as in chemical and power processes. These time delays can result from a variety of effects such as signal propagation, reaction kinetics, and system dynamics as mentioned in [1]. According to [2], these delays disrupt the instantaneous feedback loop resulting in unpredictable behaviors. Mathematically, modeling and analyzing time delay systems involves dealing with delay differential equations, which can be notoriously challenging to solve and analyze. Moreover, these systems are highly sensitive to initial conditions and external disturbances, making them susceptible to unexpected variations in behavior. In many practical systems, it is desirable to develop control systems that ensure stability and guarantee an adequate performance bound. One approach to address this challenge is the guaranteed cost control method.

The concept of guaranteed cost control of uncertain systems was initially proposed in [3] and has been investigated by numerous researchers. This approach aims to develop a controller such that the resulting closed-loop system is asymptotically stable while ensuring an adequate upper bound on the closed-loop value of a quadratic cost function as shown in [4]. The question of guaranteed cost control has also been explored for linear systems with time delay subject to uncertainties. As an example, [5] introduced a delay-dependent memory controller with variable gains, where the gains are tailored based on online estimations of fault parameters using an indirect adaptive method. Moreover, [6] proposed an approach to address the guaranteed cost control problem for continuous-time periodic piecewise linear systems with time delay. However, the design of these control methods mentioned above depends on the knowledge of the system dynamics.

In practical applications, having a complete and accurate system dynamics is not possible. Therefore, to handle this limitation, one may opt for learning-based approaches. Recently, Q-learning algorithms have been employed to deal with the infinite horizon optimal control problems with completely unknown linear or non-linear system dynamics. Notably, in [7], an online algorithm based on Adaptive Dynamic Programming was introduced to learn the continuous time optimal control for a linear system with completely unknown dynamics. Additionally, [8] presented an online model free Q-learning approach for solving the infinite horizon optimal control problem of a linear time invariant system. This approach concurrently estimates the Q-function and overcomes the limitations of the off-policy Q-learning outlined in [9], which depends on a sequential algorithm and requires an initial stabilizing control policy. Moreover, [10] proposed an online Q-learning model free approach in a non-iterative manner for continuous-time nonlinear affine systems. Drawing from the preceding works and related references, there are still gaps to be addressed. An important research line is how to design a model free infinite horizon guaranteed cost control for time delay systems with unknown Dynamics?

This paper presents an online Q-learning algorithm to tackle the challenge of the infinite-horizon guaranteed cost control problem for linear time delay systems with unknown dynamics. First, we revisit the guaranteed cost control problem and how to synthesise feedback gain using Linear Matrix Inequality (LMI). Then, we present our contributions, which can be summarized as follows:

1) We associate the state value function to the augmented Lyapunov functional, while projecting the state over the Bessel-Legendre orthogonal basis.
2) We construct an adequate Q-function for systems with constant delays. This Q-function enables the development of a model free Q-learning algorithm for linear time delay systems with unknown dynamics.
3) We show how to learn the Q-function using integral reinforcement learning based on actor/critic approximator.

This action benefited from the support of the Chair "Sustainable Energy" led by Mohammed VI Polytechnic University, sponsored by OCP.

[1] Zineb Benhmidouch and Ahmed Abbou are with Electrical Engineering Department, Mohammadia School of Engineers, Mohammed V University in Rabat, Rabat, Morocco. benhmidouch@research.emi.ac.ma, abbou@emi.ac.ma

[2] Sadek Belamfedel Alaoui and Adnane Saoud are with College of Computing, Mohammed VI Polytechnic University, Benguerir, Morocco. email (sadek.belamfedel,adnane.saoud)@um6p.ma

Finally, the efficacy of the proposed method is validated through a practical case study conducted on an interconnected power system. Due to space limitations the proofs will be published elsewhere.

Notation: The symbols $\mathbb{N}$, $\mathbb{N}_{>0}$, $\mathbb{R}$, and $\mathbb{R}_{\geqslant 0}$ represent the sets of natural numbers, strictly positive natural numbers, real numbers, and positive real numbers, respectively. The set $\mathbb{S}_n^+ \subseteq \mathbb{R}^{n\times n}$ is the set of symmetric positive definite matrices. For any function $x : [-h, +\infty) \to \mathbb{R}^n$, the notations x and x_h stand for x and $x(t-h)$, for all $t \geq 0$ and all $h \in \mathbb{R}_{\geqslant 0}$, respectively. The notation u stands for $u(t)$, for all $t \geq 0$. The notation $\begin{pmatrix} k \\ l \end{pmatrix}$ is the binomial coefficients given by $\frac{k!}{(k-l)!l!}$. The notation $\sqrt{M}$ represents the square root of a matrix M. The symmetric matrix $\begin{bmatrix} A & B \\ * & C \end{bmatrix}$ stands for $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$. The half-vectorization, denoted $vech(A)$, of a symmetric $n \times n$ matrix A is a column vector of size $\frac{n(n+1)}{2} \times 1$, obtained by stacking the elements from the lower triangular portion of A: $\text{vech}(A) = \begin{bmatrix} A_{11} & A_{21} & A_{22} & A_{31} & A_{32} & A_{33} & \dots & A_{nn} \end{bmatrix}^\intercal$. The notation $\lceil x \rceil$ denotes the ceiling of a real number x, while $\otimes$ denotes the Kronecker product of quadratic polynomial basis vector. The notation $\|.\|_T$ represents the Chebyshev-weighted semi-norm, which is defined by $\|x\|_T = \int_{-1}^{1} \frac{\|\dot{x}(u)\|}{\sqrt{1-u^2}}\, du$, while $\|.\|$ refers to the Euclidean norm.

II. Preliminaries and problem statement

Consider the linear time delay system described by,

$$\begin{cases} \dot{x}(t) = Ax(t) + A_d x(t-h) + Bu(t), & t \geq 0, \\ x(t) = \phi(t), & t \in [-h, 0], \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$ is a state vector, $\phi : [-h, 0] \to \mathbb{R}^n$ is the initial condition function, $u \in \mathbb{R}^m$ is the control input and A, $A_d \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ are the plant and input matrices respectively. Additionally, $h \in \mathbb{R}_{\geqslant 0}$ is the time delay.

Assumption 1. *The following assumptions are made:*

- *The state $x(t)$ is measurable;*
- *$h \in \mathbb{R}_{\geqslant 0}$ is constant and known delay.*

Note that in the case where h is unknown some delay estimation techniques can be used see e.g., [11]. Associated with system (1), we define the following quadratic cost function as follows:

$$J(x(0), u) = \min_u \int_0^\infty (x^\intercal M x + u^\intercal R u)\, dt, \tag{2}$$

where $M \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are user-predefined symmetric positive definite matrices and the pair $(\sqrt{M}, A)$ is detectable. The main objective is to determine the optimal value of the Lyapunov-Krasovskii functional $V^*(x,t)$ given by,

$$V^*(x,t) = \bar{x}^\intercal(t) P \bar{x}(t) + \int_{t-h}^{t} x^\intercal(r) S x(r)\, dr + \int_{t-h}^{t}\int_{\beta}^{t} \dot{x}^\intercal(r) W \dot{x}(r)\, dr\, d\beta. \tag{3}$$

A. Problem definition

Let us first recall the notion of guaranteed cost controller as in [12].

Definition 1. *Consider the time delay system* (1)*, if there exists a control law $u^* : \mathbb{R}^n \to \mathbb{R}^m$ and a positive scalar J^* such that, the closed-loop system is stable and the value of the cost function* (2) *satisfies $J \leqslant J^*$. In this case J^* is said to be a guaranteed cost and $u^* : \mathbb{R}^n \to \mathbb{R}^m$ is said to be a guaranteed cost control law for system* (1).

The problem addressed in this paper is to find a feedback control $u : \mathbb{R}^n \mapsto \mathbb{R}^m$, formally defined by,

$$u^* = Kx, \tag{4}$$

where K is the control gain matrix to be determined, such that system (1) is asymptotically stable. Moreover, among all possible controls satisfying this property, we want to select a control which minimizes the cost function J, but without any prior information on the system dynamics, i.e., the system matrices A, A_d, and B are unknown, such that $J < J^*$, where J^* is a positive scalar.

III. Guaranteed cost control

In this section, a concise description of the synthesis of the guaranteed control problem is provided before proceeding into the actor-critic adaptive structure based model-free Q-learning approach. The following lemma from [13] will be used in the sequel.

Lemma 1. *Let $N \in \mathbb{N}$ and $x : [\alpha, \beta] \to R^n$ be a continuous and differentiable function. For any matrix $Z \in \mathbb{S}_n^+$, the following inequality holds:*

$$-\int_\alpha^\beta \dot{x}^T(u) Z \dot{x}(u) \mathrm{d}u \leq -\tfrac{1}{\beta-\alpha}\zeta_N^T \left[\sum_{k=0}^{N} (2k+1)\pi_N^T(k) Z \pi_N(k)\right] \zeta_N, \tag{5}$$

where

$$\zeta_N = \begin{cases} \begin{bmatrix} x^\intercal(\beta) & x^\intercal(\alpha) \end{bmatrix}^\intercal, & N = 0, \\ \begin{bmatrix} x^\intercal(\beta) & x^\intercal(\alpha) & \frac{1}{\beta-\alpha}\chi_0^\intercal & \cdots & \frac{1}{\beta-\alpha}\chi_{N-1}^\intercal \end{bmatrix}^\intercal, & N > 0, \end{cases}$$

$$\pi_N(k) = \begin{cases} \begin{bmatrix} I & -I \end{bmatrix}, & N = 0, \\ \begin{bmatrix} I & (-1)^{k+1} I & \theta_k^0 I \cdots \theta_k^{N-1} I \end{bmatrix}, & N \geq 1, \end{cases}$$

$$\theta_k^j = \begin{cases} (2j+1)\left((-1)^{k+j} - 1\right), & j \leqslant k, \\ 0, & j > k, \end{cases}$$

$$\mathrm{F}_k(u) = (-1)^k \sum_{i=0}^{k} \left[(-1)^i \begin{pmatrix} k \\ i \end{pmatrix} \begin{pmatrix} k+i \\ i \end{pmatrix}\right] \left(\tfrac{u-\alpha}{\beta-\alpha}\right)^i,$$

$$\chi_k = \int_\alpha^\beta \mathrm{F}_k(u) x(u) du.$$

The following result present a bilinear matrix inequality-based condition allowing to synthesise a guaranteed cost controller.

Theorem 1. *Consider the time delay system (1) with the cost function (2). Given an integer N, if there exist positive definite matrices $P \in \mathbb{S}^+_{(N+1)n}$, and $S, W \in \mathbb{S}_n^+$, an invertible matrix $H \in \mathbb{R}^{n\times n}$ and a matrix $K \in \mathbb{R}^{m\times n}$ such that,*

$$\Lambda - \Gamma_N^\intercal \mathcal{W} \Gamma_N + 2E^\intercal H G + e_1^\intercal (M + K^\intercal R K) e_1 < 0, \tag{6}$$

then, the control law (4) is a guaranteed cost controller. Moreover, the cost function in (2) satisfies,

$$J(t) < J^* = \bar{\phi}^\intercal(0) P \bar{\phi}(0) + \int_{-h}^{0} \phi^\intercal(\xi) S \phi(\xi)\, d\xi + \int_{-h}^{0} \int_{\beta}^{0} \dot{\phi}^\intercal(\xi) W \dot{\phi}(\xi)\, d\xi\, d\beta, \tag{7}$$

where $\bar{\phi}(0) = \begin{bmatrix} \phi^\intercal(0) & \chi_0^\intercal(0) & \dots & \chi_{N-1}^\intercal(0) \end{bmatrix}^\intercal$, *and*

$$e_i = \begin{bmatrix} 0_{n\times(i-1)n} & I_n & 0_{n\times(N+3-i)n} \end{bmatrix}, \quad i = 1, \dots, N+3, \tag{8}$$

$$\mathcal{K} = \begin{bmatrix} h e_3^\intercal & \dots & h e_{N+2}^\intercal \end{bmatrix}^\intercal, \tag{9}$$

$$\mathbb{E}_{N-1} = \begin{cases} \begin{bmatrix} e_1^\intercal & e_2^\intercal \end{bmatrix}, & \text{if } N = 0, \\ \begin{bmatrix} e_1^\intercal & e_2^\intercal & e_\varphi^\intercal \end{bmatrix}, & \text{if } N > 0, \end{cases} \tag{10}$$

$$e_\varphi = \begin{bmatrix} e_3 & \dots & e_{N+2} \end{bmatrix},$$

$$\Gamma_N = \begin{bmatrix} \pi_N(0) & \pi_N(1) & \dots & \pi_N(N) \end{bmatrix}, \tag{11}$$

$$G = e_{N+3} - (A + BK) e_1 - A_d e_2, \tag{12}$$

$$E = e_1 + e_2 + e_{N+3}, \tag{13}$$

$$\Lambda = 2 \left\{ \begin{bmatrix} e_{N+3} \\ \mathbb{E}_{N-1} \end{bmatrix}^\intercal P \begin{bmatrix} e_1 \\ \mathcal{K} \end{bmatrix} + e_1^\intercal S e_1 \right\} - e_2^\intercal S e_2 + h e_{N+3}^\intercal W e_{N+3}, \tag{14}$$

$$\mathcal{W} = diag\{W, 3W, \dots, 2(N+1)W\}, \tag{15}$$

$$\chi_k = \int_{\alpha}^{\beta} \mathrm{F}_k(u) x(u) du. \tag{16}$$

It is worth noticing that Theorem 1 contains the bilinear term HBK, which can not be solved using LMIs-based approaches. For completeness of the design method of a guaranteed cost controller we provide a linear formulation of Theorem 1.

Theorem 2. *Consider the time delay system (1) with the cost function (2). Given an integer N, if there exist positive definite matrices $P \in \mathbb{S}^+_{(N+1)n}$, and $S, W \in \mathbb{S}^+_n$, an invertible matrix $H \in \mathbb{R}^{n\times n}$, and a matrix $V \in \mathbb{R}^{m\times n}$ satisfying,*

$$\begin{bmatrix} \Lambda - \Gamma_N^\intercal \mathcal{W} \Gamma_N + 2E^\intercal \bar{G} + e_1^\intercal M e_1 & e_1^T V^\intercal \\ * & -\mathcal{R} \end{bmatrix} < 0, \tag{17}$$

then, the control law (4) *with $K = VH^{-1}$ is a guaranteed cost controller with the guaranteed cost* (19)*, and*

$$\bar{G} = (AH + BV) e_1 + A_d H e_2 - e_{N+3},$$
$$\mathcal{R} = R^{-1}.$$

Remark 1. *The LMI computation in* (6) *introduces a novel approach, differing from prior methods such as those discussed in [13], notably by incorporating the derivative $\dot{x}(t)$ into the augmented vector $\xi(t)$. This computational technique facilitates LMI linearization through coordinate transformations, as demonstrated in [14]–[16], while employing augmented Lyapunov functionals. Notably, in current literature, knowledge of system matrices A, A_d, and B is essential for solving the LMI in* (17)*.*

IV. Model free formulation

In this section, we explore the Q-learning formulation for model-free control problem of time delay systems as in (1). The Q-function, $Q(x,u) : \mathbb{R}^{n+m} \to \mathbb{R}$, is defined as in [17],

$$Q^*(x,u) := V^*(x) + \frac{\partial V^{*\intercal}}{\partial x} (Ax + A_d x_h + Bu) + x^\intercal M x + u^\intercal R u, \tag{18}$$

which represents the value of taking action u from state x and following the policy (4) afterwards. In the context of time delay systems, expressing the Q-function as in (18) in a compact quadratic form poses a challenge. The main obstacle lies in obtaining a precise finite quadratic representation of the integral terms $\int_{t-h}^{t} x^\intercal(r) S x(r)\, dr$ and $\int_{t-h}^{t} \int_{\beta}^{t} x^\intercal(r) W x(r)\, dr d\beta$. To solve this issue, we relay on Bessel-Legendre based integral inequality [13, Lemma 3], to approximate the Lyapunov–krasovskii functional in (3) by a lower bound, denoted $V(x,t)$, given by,

$$V(x,t) = x^\intercal P x + \frac{1}{h} \sum_{k=0}^{N} (2k+1) \Omega_k^\intercal S \Omega_k, \tag{19}$$

where $\Omega_k = \int_{-h}^{0} L_k(u) x(u) du$, $L_k(u) = (-1)^k \sum_{l=0}^{k} p_l^k (\frac{u+h}{h})^l$, $p_l^k = (-1)^l \begin{pmatrix} k \\ l \end{pmatrix} \begin{pmatrix} k+l \\ l \end{pmatrix}$ and $k = 0, \dots, N$.

Let $\tilde{V}(x,t) = V^*(x,t) - V(x,t)$ be the approximation error. The next results allows us to estimate an order that ensures that $\tilde{V}(x)$ is upper bounded by $\epsilon > 0$.

Lemma 2. *For any $\epsilon > 0$, there exists a corresponding integer $\mathcal{N}(\epsilon)$ such that for all $N \geq \mathcal{N}(\epsilon)$, the following holds:*

$$\left| \tilde{V}(x) \right| \leqslant \epsilon, \tag{20}$$

where $\mathcal{N}(\epsilon) = \left\lceil \frac{2\pi \|x\|_T^2 \left|\bar{\lambda}(S)\right|}{h\epsilon} \right\rceil$.

The following bellman equation is associated to $V(x,t)$ in (19),

$$Q(x,u) := V(x,t) + \tfrac{\partial V^\intercal}{\partial x} (Ax + A_d x_h + Bu) + x^\intercal M x + u^\intercal R u, \tag{21}$$

which from the monotonicity of the Bellman operator, [18] satisfies,

$$Q(x,u) \leqslant \tilde{Q}(x,u) \leqslant Q^*(x,u), \tag{22}$$

where $\tilde{Q}(x,u)$ is formally defined by,

$$\tilde{Q}(x,u) := V(x,t) + \tfrac{\partial V^{*\intercal}}{\partial x} (Ax + A_d x(t-h) + Bu) + x^\intercal M x + u^\intercal R u \tag{23}$$

Note that the controller $u = \arg\min_u \tilde{Q}(x,u)$ in (23) ensures a certain level of performance measured by the cost function J and that the obtained controller is a guaranteed cost controller and not necessarily optimal. It should be noted that the choice of the parameter N significantly affects the optimality of the resulting controller. As N increases, the controller tends to become optimal, and the complexity of

the computation also increases. Thus, for a given $N \in \mathbb{N}$, we express $\tilde{Q}(x,u)$ by the quadratic form,

$$\tilde{Q}(x,u) = \Theta^{\intercal}(t)\Psi\Theta(t), \tag{24}$$

where

$$\Theta(t) = \begin{bmatrix} \xi(t)^{\intercal} & u^{\intercal}(t) \end{bmatrix}^{\intercal},$$
$$\xi(t) = \begin{bmatrix} x^{\intercal}(t) & x_h^{\intercal}(t) & \Omega_0^{\intercal}(t) & \dots & \Omega_N^{\intercal}(t) & \dot{x}^{\intercal}(t) \end{bmatrix}^{\intercal},$$

and $\Psi \in \mathbb{R}^{((N+4)n+m)\times((N+4)n+m)}$ is given by,

$$\Psi = \begin{bmatrix} \Psi_{xx} & \Psi_{xu} \\ * & \Psi_{uu} \end{bmatrix}, \tag{25}$$

with $\Psi_{xu} = e_1^T K^T$, $\Psi_{ux} = -R^{-1}$, and

$$\Psi_{xx} = P + diag\{S, 3S, \dots, 2(N+1)S\} + \Lambda - \Gamma_N^{\intercal}\mathcal{W}\Gamma_N + 2E^{\intercal}HG + e_1^{\intercal}Me_1,$$

where $\Psi_{xx} \in \mathbb{R}^{(N+4)n\times(N+4)n}$, $\Psi_{ux} = \Psi_{xu}^{\intercal} \in \mathbb{R}^{m\times(N+4)n}$, and $\Psi_{uu} \in \mathbb{R}^{m\times m}$. Finding the guaranteed cost controller for the linear time delay system (1) requires computing the Q-function provided in equation (24). This later is based on the computation of the state over the orthogonal basis of the Bessel-Legendre polynomial, see [13].

Lemma 3. *(Guaranteed cost control) Given the Q-function in* (24)*, the value function* $V(x,t)$*, underestimator of* $V^*(x,t)$*, minimizes the cost function in* (2)*, where* $P > 0$*,* $S > 0$ *and* $W > 0$ *satisfying Theorem 1, and guarantees a performance bound,* $\tilde{Q}^*(x,u^*) := \min_u \tilde{Q}(x,u) \leqslant J^*$.

A model-free formulation of the guaranteed cost control (4), can be obtained from $\frac{\tilde{Q}(x,u)}{\partial u} = 0$ as follows,

$$u^*(x) = \arg\min_u \tilde{Q}(x,u) = -\Psi_{uu}^{-1}\Psi_{ux}\Theta. \tag{26}$$

In the computational process, the vector Θ can be derived from a sequence of system data from $x(t-h)$ to $x(t)$. Additionally, the integral terms can be approximated using Riemann numerical method [19]. The following subsection introduce an actor/critic neural network structure to adjust the Q-function based on data instead of a system model.

A. Actor/ Critic neural network structure

In the actor-critic approach, the actor's parameters are updated using gradients derived from the critic's value estimates. The optimal value of the Q-function $\tilde{Q}^*(x,u^*)$ can be expressed as,

$$\tilde{Q}^*(x,u^*) : = \text{vech}(\Psi)^{\intercal}(\Theta \otimes \Theta), \tag{27}$$

where $\text{vech}(\Psi) \in \mathbb{R}^{\frac{1}{2}((N+4)n+m)((N+4)n+m+1)}$ represents a half vectorization of the matrix Ψ. By denoting as $W_c := \text{vech}(\Psi)$, the function (27) can be written in a compact form as, $\tilde{Q}^*(x,u^*) = W_c^{\intercal}(\Theta \otimes \Theta)$, with $W_c \in \mathbb{R}^{\frac{1}{2}((N+4)n+m)((N+4)n+m+1)}$ are the ideal weights of the critic approximator where, $\text{vech}(\Psi_{xx}) := W_c[1 : \frac{l(l+1)}{2}]$, $\text{vech}(\Psi_{xu}) := W_c[\frac{l(l+1)}{2}+1 : \frac{l(l+1)}{2}+lm]$, and $\text{vech}(\Psi_{uu}) := W_c[\frac{l(l+1)}{2}+lm+1 : \frac{1}{2}(l+m)(l+m+1)]$, with $l = (N+4)n$ and the notation $W_c[i:j]$ represents a subset of elements from the vector W_c starting from row i to j. Given the unknown optimal weights for computing $\tilde{Q}^*$ and u^*, it is necessary to consider the following weight approximations. The critic approximator can be written as,

$$\hat{\tilde{Q}}(x,u) = \hat{W}_c^{\intercal}(\Theta \otimes \Theta), \tag{28}$$

where $\hat{W}_c \in \mathbb{R}^{\frac{1}{2}((N+4)n+m)((N+4)n+m+1)}$ are the estimated weights. Similarly, the expression of the actor approximator can be formulated as:

$$\hat{u}(x) = \hat{W}_a^{\intercal}x, \tag{29}$$

where $\hat{W}_a \in \mathbb{R}^{n\times m}$. To find the update law for the critic's weights, we define the following temporal difference error for the critic approximator $e_c \in \mathbb{R}$ as,

$$\begin{aligned} e_c &:= \hat{\tilde{Q}}(x(t),u(t)) - \hat{\tilde{Q}}(x(t-T),u(t-T)) \\ &+ \int_{t-T}^{t} x^{\intercal}(r)Mx(r) + u^{\intercal}(r)Ru(r)\,dr \\ &= \hat{W}_c^{\intercal}(\Theta(t)\otimes\Theta(t)) - \hat{W}_c^{\intercal}(\Theta(t-T)\otimes\Theta(t-T)) \\ &+ \int_{t-T}^{t} x^{\intercal}(r)Mx(r) + u^{\intercal}(r)Ru(r)\,dr. \end{aligned} \tag{30}$$

While the temporal difference error for the actor approximator is given by,

$$e_a := \hat{W}_a^{\intercal}x + \hat{\Psi}_{uu}^{-1}\hat{\Psi}_{ux}\Theta, \tag{31}$$

where the values of $\hat{\Psi}_{uu}$ and $\hat{\Psi}_{ux}$ are going to be derived from the vector $\hat{W}_c$. To learn the optimal weights online, we can define the squared-norm of the critic and actor errors as,

$$\delta_c := \frac{1}{2}\|e_c\|^2, \qquad \delta_a := \frac{1}{2}\|e_a\|^2. \tag{32}$$

B. Weight update mechanism

To guarantee the convergence of e_c to 0 and $\hat{W}_c$ to W_c, the update law for the critic approximator can be formulated using a normalized gradient descent method as follows,

$$\dot{\hat{W}}_c = -\alpha_c \frac{1}{(1+\eta^{\intercal}\eta)^2}\frac{\partial\delta_c}{\partial\hat{W}_c} = -\alpha_c\frac{\eta}{(1+\eta^{\intercal}\eta)^2}e_c^{\intercal}, \tag{33}$$

where $\eta := (\Theta(t)\otimes\Theta(t)) - (\Theta(t-T)\otimes\Theta(t-T))$ and $\alpha_c \in \mathbb{R}^+$ is a user-defined constant critic gain. Similarly, the update law for the actor approximator is given by,

$$\dot{\hat{W}}_a = -\alpha_a\frac{\partial\delta_a}{\partial\hat{W}_a} = -\alpha_a x e_a^{\intercal}, \tag{34}$$

where $\alpha_a \in \mathbb{R}^+$ is a user-defined constant actor gain. We define the critic and actor weight estimation errors as $\tilde{W}_c = W_c - \hat{W}_c$ and $\tilde{W}_a = -\Psi_{xu}\Psi_{uu}^{-1} - \hat{W}_a$, respectively. Based on the update laws in (33) and (34), the weights estimation error dynamics can be expressed as,

$$\dot{\tilde{W}}_c = -\alpha_c\frac{\eta\eta^{\intercal}}{(1+\eta^{\intercal}\eta)^2}\tilde{W}_c, \tag{35}$$

$$\dot{\tilde{W}}_a = -\alpha_a x x^{\intercal}\tilde{W}_a - \alpha_a x\Theta^{\intercal}\tilde{\Psi}_{xu}\Psi_{uu}^{-1}, \tag{36}$$

where $\tilde{\Psi}_{xu} = \text{mat}(\tilde{W}_c[\frac{l(l+1)}{2}+1 : \frac{l(l+1)}{2}+lm]$. Since the precise value of Ψ_{uu} is known, and it is equal to R which

is defined by the predefined performance index provided by the user. Thus, equation (36) can be rewritten as follows,

$$\dot{\tilde{W}}_a = -\alpha_a x x^\intercal \tilde{W}_a - \alpha_a x \Theta^\intercal \tilde{\Psi}_{xu} R^{-1}. \tag{37}$$

C. Stability analysis

In this subsection, it is shown that the estimated weights $\hat{W}_c$ achieve an exponential convergence to the optimal unknown weights W_c for any given control input u under a guaranteed persistence of excitation condition.

Lemma 4. *Consider the update law of the critic approximator given by* (33)*. For any control policy u the critic's weights estimation error dynamics given by* (35) *have an exponentially stable equilibrium point satisfying*

$$\|\tilde{W}_c\| \leqslant \rho_1 e^{-\rho_2 (t-t_0)} \|\tilde{W}_c(t_0)\|, \quad t > t_0 \geq 0, \tag{38}$$

where ρ_1 and ρ_2 are positive constants provided that the signal $\sigma := \frac{\eta}{1+\eta^\intercal \eta}$ is persistently exciting, i.e., $\exists \alpha, \delta \in \mathbb{R}^+$ such that $\forall t \in \mathbb{R}^+$, $\int_t^{t+\delta} \sigma(r)\sigma(r)^\intercal dr > \alpha I$, with I an identity matrix of appropriate dimensions.

The next result provides the main Theorem on stability for the proposed Q-learning method.

Theorem 3. *Consider the linear time delay system given by* (1)*, the critic and actor approximator given by* (28) *and* (29)*, respectively. The tuning law for the weights of the critic and actor are given by* (33) *and* (34)*, respectively. Then the equilibrium point of the closed loop system, characterized by the state $\Phi := \begin{bmatrix} x^\intercal & \tilde{W}_c^\intercal & \tilde{W}_a^\intercal \end{bmatrix}^\intercal$ for all initial conditions $\Phi(\phi)$, is proven to be asymptotically stable provided that the critic gain α_c is significantly larger than the actor gain α_a and the following inequality satisfied:*

$$0 < \alpha_a < \frac{1}{\delta \bar{\lambda}(R^{-1})} \left(2\underline{\lambda}(M + \Psi_{xu} R^{-1} \Psi_{xu}^\intercal) - \bar{\lambda}(\Psi_{xu} \Psi_{xu}^\intercal) \right), \tag{39}$$

where δ is a constant of unity order.

V. Simulation results

In this section, a two-area interconnected power system with time delay is considered to check the effectiveness of the proposed model free Q-learning scheme. The system experiences frequency deviations due to active power load changes, leading to network instability. Although a local governor can adjust generator output to counteract these load changes, it typically causes frequency fluctuations. To maintain stable system frequency at the standard 50 Hz, especially during load fluctuations, an additional Load Frequency Control (LFC) is required. The dynamics of the LFC are described by specific governing equations.

$$\begin{cases} \Delta P_{ij} & = -\Delta P_{ji}, \quad i,j = 1,2 \quad i \neq j \\ \Delta \dot{P}_{m_i} & = \frac{1}{T_{t_i}} (\Delta P_{g_i} - \Delta P_{m_i}), \\ \Delta \dot{P}_{12} & = 2\pi T_{12} (\Delta f_1 - \Delta f_2), \\ \Delta \dot{P}_{c_i} & = K_i \Delta P_{12} + K_i \beta_i \Delta f_i, \\ \Delta \dot{f}_i & = -\frac{K_{p_i}}{T_{p_i}} (\Delta P_{l_i} + \Delta P_{ij} - \Delta P_{m_i}) - \frac{1}{T_{p_i}} \Delta f_i, \\ \Delta \dot{P}_{g_i} & = -\frac{1}{T_{g_i}} (R_i^{-1} \Delta f_i + \Delta P_{g_i} + \Delta P_{c_i}(t-h) - u_i), \end{cases} \tag{40}$$

where Δf_i, ΔP_{m_i}, ΔP_{g_i}, ΔP_{c_i}, and ΔP_{12} denote the frequency deviation, generator mechanical power output, power output of turbine generator, area control error signals, and tie-line power flow from area 1 to area 2, respectively. While K_{p_i}, R_i, K_i and β_i represent power system gain, speed regulation coefficient, integral control gain and frequency bias parameter, respectively. T_{p_i}, T_{t_i}, and T_{g_i} are power system, turbine, and governor time constants, and T_{12} is stiffness coefficient. In the considered simulation, the load disturbances ΔP_{l_i} are assumed to be zero and the time delay in the two control areas is given by $h = 0.5s$. Table I shows the model parameters, which vary due to real-world factors and aging. Let the state vector be expressed as

TABLE I: List of simulation parameters with assigned values

Parameters	Values	Parameters	Values
T_{t_1}, T_{t_2}	0.3, 0.17	K_1, K_2	0.5, 0.6
T_{g_1}, T_{g_2}	0.1, 0.4	D_1, D_2	1, 1.5
R_1, R_2	0.05, 0.05	H_1, H_2	10, 12

$x = \begin{bmatrix} \Delta f_1 & \Delta P_{m_1} & \Delta P_{g_1} & \Delta P_{c_1} & \Delta P_{12} & \Delta f_2 & \Delta P_{m_2} & \Delta P_{g_2} & \Delta P_{c_2} \end{bmatrix}^\intercal$. Based on this, the LFC system parameter matrices are defined as follows:

$$A = \begin{bmatrix} -\frac{1}{T_{p_1}} & \frac{K_{p_1}}{T_{p_1}} & 0 & 0 & -\frac{K_{p_1}}{T_{p_1}} & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{T_{t_1}} & \frac{1}{T_{t_1}} & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{R_1 T_{g_1}} & 0 & -\frac{1}{T_{g_1}} & 0 & 0 & 0 & 0 & 0 & 0 \\ K_1\beta_1 & 0 & 0 & 0 & K_1 & 0 & 0 & 0 & 0 \\ 2\pi T_{12} & 0 & 0 & 0 & 0 & -2\pi T_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{K_{p_2}}{T_{p_2}} & -\frac{1}{T_{p_2}} & \frac{K_{p_2}}{T_{p_2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{T_{t_2}} & \frac{1}{T_{t_2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{R_2 T_{g_2}} & 0 & -\frac{1}{T_{g_2}} & 0 \\ 0 & 0 & 0 & 0 & K_2 & K_2\beta_2 & 0 & 0 & 0 \end{bmatrix},$$

$$A_d = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{T_{g_1}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{T_{g_2}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{1}{T_{g_1}} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & \frac{1}{T_{g_2}} \\ 0 & 0 \end{bmatrix}.$$

The matrices for the performance index (2) are chosen as $M := I$ and $R := 0.5I$, where I is the identity matrix with appropriate dimensions. The algorithm from Theorem 3 is applied with parameters $\alpha_c = 30$, $\alpha_a = 1.1$, $T = 0.1$ s, $N = 4$, and a 1.5 s delay for the integral terms, while the matrices A, A_d, and B are considered unknown. It is necessary to note that the delay used for Bessel integrals must be significantly larger than the system delay. Initial weights for both actor and critic networks are randomly assigned within the ranges of $[0, 3]$ and $[0, 1]$, respectively. Fig. 1 shows the time evolution of the state trajectories of the system. As seen in Fig. 2, one can observe that the responses of frequency deviations Δf_1 and Δf_2 of the closed loop system, subject to a transient variation in load demands prior to $t = 0$, quickly converge to zero after a few seconds of settling time. This convergence stems from the swift adjustment of the control input signal, as illustrated in Fig. 4. Therefore, the model-free Q-learning approach can help reduce frequency deviations between power areas. On the

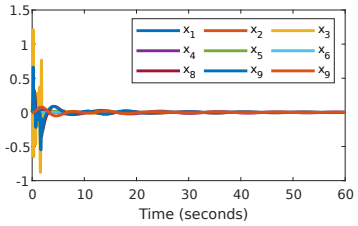

Fig. 1: Time evolution of the system states.

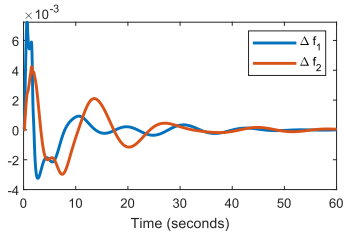

Fig. 2: The frequency deviation of LFC system under the control input.

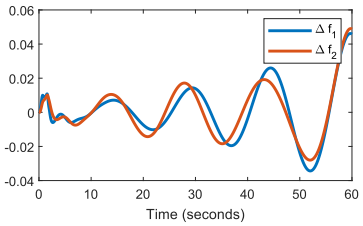

Fig. 3: The frequency deviation without control input.

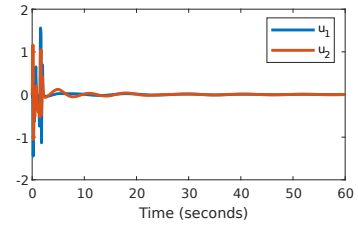

Fig. 4: The control input during the learning process.

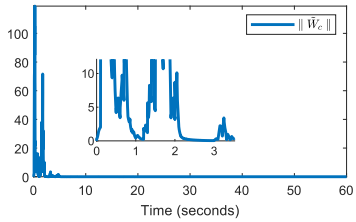

Fig. 5: The norm of the critic network weights error.

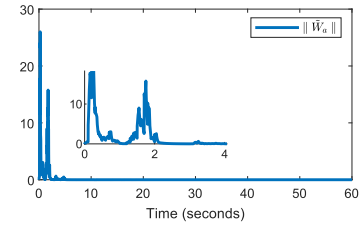

Fig. 6: The norm of the actor network weights error.

contrary, the frequency deviations of the open-loop system without the control input diverge, as can be seen in Fig. 3, which demonstrates the algorithm's ability to dynamically dampen frequency and power oscillations. The convergence to zero of the norm of the critic weights error, depicted in Fig. 5, demonstrates that the proposed method effectively achieves a guaranteed cost control, minimizing the cost function defined in (2). While, Fig. 6 shows the convergence of the actor weights' norm to the optimal value. It is worth noting that when employing model-free Q-learning control for unstable systems, selecting appropriate initial weight ranges is important. The chattering phenomenon observed in several figures, results from exploration noise added to meet the PE condition in Lemma 4.

VI. Conclusion

In the present work, a model free approach is developed to address the infinite horizon guaranteed cost control for linear time delay systems with unknown dynamics. Based on a sufficient condition obtained from LMI approach, the appropriate Q-function to drive the control policy is formulated in terms of the state and control variables using Bessel-Legendre integral inequality. Furthermore, an actor critic structure is developed to approximate simultaneously and in an adaptive manner both the Q-function and the control policy. Afterwards, the stability of the closed-loop system and its convergence to the guaranteed cost control are verified without any prior knowledge of the system dynamics.

References

[1] K. Gu, J. Chen, and V. L. Kharitonov, *Stability of time-delay systems*. Springer Science & Business Media, 2003.

[2] F. M. Atay, *Complex time-delay systems: theory and applications*. Springer, 2010.

[3] S. Chang and T. Peng, "Adaptive guaranteed cost control of systems with uncertain parameters," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 474–483, 1972.

[4] G.-H. Yang, J. L. Wang, and Y. C. Soh, "Reliable guaranteed cost control for uncertain nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 11, pp. 2188–2192, 2000.

[5] D. Ye and G.-H. Yang, "Reliable guaranteed cost control for linear state delayed systems with adaptive memory state feedback controllers," *Asian Journal of Control*, vol. 10, no. 6, pp. 678–686, 2008.

[6] X. Xie and J. Lam, "Guaranteed cost control of periodic piecewise linear time-delay systems," *Automatica*, vol. 94, pp. 274–282, 2018.

[7] 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.

[8] 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.

[9] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, 2014.

[10] A. S. Chen and G. Herrmann, "Adaptive optimal control via continuous-time q-learning for unknown nonlinear affine systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 1007–1012, IEEE, 2019.

[11] V. Léchappé, S. Rouquet, A. Gonzalez, F. Plestan, J. De León, E. Moulay, and A. Glumineau, "Delay estimation and predictive control of uncertain systems with input delay: Application to a dc motor," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 9, pp. 5849–5857, 2016.

[12] L. Yu and J. Chu, "An lmi approach to guaranteed cost control of linear uncertain time-delay systems," *Automatica*, vol. 35, no. 6, pp. 1155–1159, 1999.

[13] A. Seuret and F. Gouaisbaut, "Hierarchy of lmi conditions for the stability analysis of time-delay systems," *Systems & Control Letters*, vol. 81, pp. 1–7, 2015.

[14] B. A. Sadek, T. E. Houssaine, and C. Noreddine, "Congestion control with aqm and dynamic quantisers," *IET Control Theory & Applications*, vol. 14, no. 20, pp. 3601–3609, 2020.

[15] B. A. Sadek, T. El Houssaine, and C. Noreddine, "On designing lyapunov-krasovskii functional for time-varying delay t–s fuzzy systems," *Journal of the Franklin Institute*, vol. 359, no. 5, pp. 2192–2205, 2022.

[16] B. A. Sadek, T. El Houssaine, K. A. Barbosa, A. J. Rojas, *et al.*, "Consensus congestion control for ad hoc networks: time-delay and saturation," *IEEE Transactions on Network Science and Engineering*, 2023.

[17] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proceedings of 1994 American Control Conference-ACC'94*, vol. 3, pp. 3475–3479, IEEE, 1994.

[18] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 4. Athena scientific, 2012.

[19] R. M. McLeod, *The generalized Riemann integral*, vol. 20. American Mathematical Soc., 1980.