Q-Learning for Linear Quadratic Optimal Control with Terminal State Constraint

Juanjuan Xu, Jingmei Liu, Zhaorong Zhang, and Wei Wang

***Abstract*—In this paper, we study the linear quadratic (LQ) optimal control of time-varying difference system with terminal state constraints. The main contribution is to provide the Q-learning algorithm for the optimal controller under the case that the time-varying system matrices and input matrices are both unknown, which consists of learning the solution of the Riccati equation and calculating the specific Lagrange multiplier from the data-driven matrix equation. Different from the existing Q-learning algorithms that mainly focus on unconstrained optimal control problems, the novelty of the proposed algorithm can be applied to handle situations with terminal state constraints. The effectiveness of the proposed Q-learning algorithm is demonstrated through a numerical example.**

***Index Terms*—Linear quadratic optimal control, terminal state constraint, Q-learning, reachability**

I. INTRODUCTION

Linear quadratic (LQ) optimal control has wide applications in fields including aerospace engineering [1], economics [2], and so on, the research of which began in the 1950s [3]. From then on, many important progresses have been made [4, 5]. In Ref. [6], the optimal state feedback solution was provided in terms of the Riccati equation. Ignaciuk and Bartoszewicz [7] gave a closed-form optimal solution for the LQ optimal control with periodic-review perishable inventory systems. Yong [8] proved the existence of equilibrium control for the deterministic LQ time-inconsistent optimal control. Zhang and Xu [9] obtained the necessary and sufficient conditions for the solvability of the LQ optimal control with irregular performance. These results are mainly focused on unconstrained optimal control problems.

Considering the practical demands in production and daily life, such as spacecraft launch and mean-variance portfolio selection, the state/input must satisfy specific constraints, e.g., the terminal state is fixed [10], the control input is bounded [11], etc. This stimulates the study of the LQ optimal control with constraints. Mitze and Mönnigmann [12] derived the optimal controller of an LQ regulator problem with constraint by using dynamic programming. A constraint-compliant feedforward-feedback control update was derived in Ref. [13] for the equality constrained iterative LQ optimal control. Sideris and Rodriguez [14] solved the LQ optimal control with state-only constraints and mixed state-control based on the Riccati equation. Ferrante and Ntogramatzidis [15] provided a closed-form expression that parameterizes all solutions of the Hamiltonian system using a strong non-mixed solution of a continuous algebraic Riccati equation and a solution of the algebraic Lyapunov equation. Sun [16] studied the LQ optimal control with terminal states and integral quadratic constraints. Park et al. [17] gave the optimal controller of LQ tracking control with fixed terminal states in a recursive form. As a brief summary, we note that most results of the constrained LQ optimal control are model-based, that is, the deriving of the optimal controller strongly depends on the exact knowledge of system matrices and input matrices. However, in industrial processes, there are many uncertainties and external disturbances, making it difficult to obtain accurate models [18]. Therefore, studying the constrained LQ optimal control with unknown system matrices and input matrices is of great significance.

As is well known, Q-learning algorithm is one of the methods of reinforcement learning and is very effective for dealing with the optimal control problem with unknown model. Calafiore and Possieri [19] presented a Q-learning algorithm under finite-horizon and provided state-feedback and output-feedback solutions of the standard LQ optimal control problem. Zhang et al. [18] proposed a Q-learning algorithm for the Nash strategy of the nonzero-sum game. Rizvi and Lin [20] provided the Q-learning output-feedback algorithm to give the optimal control of infinite-horizon LQ zero-sum game. In Ref. [21], the Q-learning algorithm was used to solve the consensusability problem of multi-agent systems. It is noted that although the Q-learning algorithm has been widely studied in various problems and many important progresses have been made, the Q-learning algorithm for LQ optimal control problem under terminal state constraints remains to be solved. The main difficulty lies in the learning of a specific Lagrange multiplier in view of the fact that the optimal controller is in the feedback of the state and the specific Lagrange multiplier.

In this paper, we study the LQ optimal control problem of discrete-time time-varying systems with terminal constraints. When the coefficient matrices of state and control inputs in the time-varying system are both unknown, a Q-learning algorithm is proposed for the design of the optimal controller. In particular, the Q-learning algorithm consists of two steps.

Manuscript received: 11 June 2024; revised: 12 July 2024; accepted: 13 August 2024. *(Corresponding author: Juanjuan Xu.)*

Citation: J. Xu, J. Liu, Z. Zhang, and W. Wang, Q-Learning for linear quadratic optimal control with terminal state constraint, *IJICS*, 2024, 29(3), 134–140.

Juanjuan Xu and Wei Wang are with School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: juanjuanxu@sdu.edu.cn; w.wang@sdu.edu.cn).

Jingmei Liu is with School of Automation and Electrical Engineering, Linyi University, Linyi 276000, China (e-mail: jingmei0729@qq.com).

Zhaorong Zhang is with School of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: zhaorong.zhang@uon.edu.au).

Digital Object Identifier 10.62678/IJICS202409.10132

The first step is to learn the solution of the Riccati equation, which thus gives the matrices gains of a specific matrix equation. The second step is to calculate the specific Lagrange multiplier from the data-driven matrix equation. Accordingly, the optimal control is obtained by running the Q-learning algorithm. Finally, a numerical example is shown to verify the effectiveness of the proposed algorithm.

The rest of the paper is organized as follows. Section II states the studied problem. The model-based solution is given in Section III. Section IV presents the specific derivation of the algorithm. Section V shows the numerical example. Some conclusions are given in Section VI.

Notation $\mathbb{R}^n$ and $\mathbb{R}^m$ denote the families of n dimensional and m dimensional vectors. x' means the transpose of vector x. $M > 0\ (\geqslant 0)$ represents that M is a positive-definite (positive semi-definite) matrix. I is an identity matrix with appropriate dimension.

II. Problem Formulation

In this paper, we study the time-varying system which is given as below

$$x(s+1) = A(s)x(s) + B(s)u(s) \tag{1}$$

where $x(s) \in \mathbb{R}^n$ and $u(s) \in \mathbb{R}^m$ are the state and control input, respectively. $A(s) \in \mathbb{R}^{n\times n}$ and $B(s) \in \mathbb{R}^{n\times m}$ are time-varying matrices. $s \in \{0,1,\ldots,S\}$ is the discrete-time variable. The initial value of the state is prescribed as $x(0) = x_0$.

The cost function is given by

$$J = \sum_{s=0}^{S}[x'(s)Qx(s) + u'(s)\mathcal{R}u(s)] + x'(S+1)Hx(S+1) \tag{2}$$

where Q and H are positive-definite matrices, and $\mathcal{R}$ is a positive semi-definite matrix.

The problem studied in this paper is presented as follows.

Problem The aim of the paper is to design a Q-learning algorithm for the optimal solution $u(s)$ which minimizes the cost function in Eq. (2) subject to Eq. (1) and ensures that the terminal state satisfies $x(S+1) = \xi$ with unknown model matrices $A(s)$, $B(s)$, and $s = 0,1,\ldots,S$, where ξ is a given constant vector.

III. Model-Based Solution

In this section, the model-based optimal controller of the problem is firstly presented. To this end, we introduce the following Riccati difference equation

$$\begin{aligned} P(s) = {} & A'(s)P(s+1)A(s) + Q - A'(s)P(s+1)B(s)\times \\ & [\mathcal{R} + B'(s)P(s+1)(s)]^{-1}\times \\ & B'(s)P(s+1)A(s) \end{aligned} \tag{3}$$

with terminal condition $P(S+1) = H$. The following denotations are made

$$\begin{aligned} \Gamma(s) = {} & \mathcal{R} + B'(s)P(s+1)B(s), \\ K(s) = {} & -\Gamma^{-1}(s)B'(s)P(s+1)A(s), \\ A_c(s) = {} & A(s) + B(s)K(s), \\ \Phi(s,S) = {} & \begin{cases} A_c(S)A_c(S-1)\cdots A_c(s), & s \leqslant S; \\ I, & s > S, \end{cases} \\ K_1(s) = {} & -\Gamma^{-1}(s)B'(s)\Phi'(s+1,S), \\ \bar{B}(s) = {} & B(s)\Gamma^{-1}(s)B'(s), \\ G(s) = {} & \sum_{j=s}^{S}\Phi(j+1,S)\bar{B}(j)\Phi'(j+1,S). \end{cases} \end{aligned}$$

Lemma 1 Assuming that the following equation

$$\Phi(0,S)x_0 - G(0)\lambda = \xi \tag{4}$$

has a solution λ^*, where λ is the Lagrange multiplier, then the optimal solution of the problem is given by

$$u(s) = K(s)x(s) + K_1(s)\lambda^* \tag{5}$$

Proof Firstly, considering the terminal state constraint $x(S+1) = \xi$, we introduce a Lagrange multiplier λ and define the cost function as below

$$\begin{aligned} J_1 = {} & \sum_{s=0}^{S}[x'(s)Qx(s) + u'(s)\mathcal{R}u(s)] + \\ & x'(S+1)Hx(S+1) + 2\lambda' x(S+1) \end{aligned} \tag{6}$$

By applying the maximum principle in Ref. [22], there exists a unique optimal controller to the problem $\min\limits_{u(s)} J_1$ subject to Eq. (1), if and only if the responded forward and backward equations have a unique solution, where the forward equation is Eq. (1) and the backward equation is governed by

$$p(s-1) = A'(s)p(s) + Qx(s) \tag{7}$$

with $p(S) = Hx(S+1) + \lambda$ and the optimal controller satisfying

$$0 = \mathcal{R}u(s) + B'(s)p(s) \tag{8}$$

Secondly, we solve the forward and backward equations. The key is to verify that

$$p(s) = P(s+1)x(s+1) + \eta(s) \tag{9}$$

where $P(s+1)$ is defined by Eq. (3) and $\eta(s)$ obeys

$$\eta(s-1) = A_c'(s)\eta(s) \tag{10}$$

with $\eta(S) = \lambda$. In fact, at time S, Eq. (9) holds naturally since $P(S+1) = H$ and $\eta(S) = \lambda$. Assuming that Eq. (9) holds with $s = w$ for any $w \in \{0,1,\ldots,S\}$, that is, $p(w) = P(w+1)x(w+1) + \eta(w)$, we prove that Eq. (9) holds for $s = w-1$. By substituting Eq. (9) into Eq. (8), we have

$$\begin{aligned} 0 = {} & \mathcal{R}u(w) + B'(w)P(w+1)x(w+1) + B'(w)\eta(w) = \\ & [\mathcal{R} + B'(w)P(w+1)B(w)]u(w) + \\ & B'(w)P(w+1)A(w)x(w) + B'(w)\eta(w). \end{aligned}$$

In view of the fact that $\mathcal{R} > 0$ and $P(w+1) \geqslant 0$, it yields that $\Gamma(w) > 0$. This implies that

$$u(w) = K(w)x(w) - \Gamma^{-1}(w)B'(w)\eta(w) \tag{11}$$

By plugging Eq. (11) into Eq. (1), Eq. (12) is obtained

$$x(w+1) = A_c(w)x(w) - \bar{B}(w)\eta(w) \tag{12}$$

Combining with Eq. (7), we have

$$\begin{aligned} p(w-1) = {} & A'(w)P(w+1)A_c(w)x(w) - \\ & A'(w)P(w+1)\bar{B}(w)\eta(w) + \\ & A'(w)\eta(w) + Qx(w), \end{aligned}$$

therefore, at time $s = w-1$, Eq. (9) follows from Eqs. (3) and (10).

Finally, we solve the Lagrange multiplier λ. From Eq. (12), it follows

$$x(S+1) = \Phi(0,S)x(0) - \sum_{s=0}^{S} \Phi(s+1,S)\bar{B}(s)\eta(s) = \Phi(0,S)x(0) - G(0)\lambda \tag{13}$$

where $\eta(s) = \Phi'(s+1,S)\lambda$ is used in the calculation of Eq. (13). Since $x(S+1) = \xi$, Eq. (13) is reduced to Eq. (4). Thus, if Eq. (4) has a solution λ^*, then the optimal controller in Eq. (5) follows from Eq. (11). ■

From Lemma 1, one of the keys to solving the problem is to solve Eq. (4). To further derive the solvability condition of Eq. (4), we introduce the reachability of vector ξ for Eq. (1) as below.

Definition 1 The state ξ is said to be reachable at time $S+1$ from initial value x_0 for Eq. (1), if there exists a sequence of controllers $\{u(s), s = 0,1,\ldots,S\}$ such that $x(0) = x_0$ and $x(S+1) = \xi$.

The criterion to judge the reachability is presented in Lemma 2.

Lemma 2 ξ is reachable at time $S+1$ from the initial value x_0 for Eq. (1) if and only if $\xi - \Phi(0,S)x_0 \in \text{Range}(G_1)$, that is, there exists a vector $\zeta \in \mathbb{R}^n$, such that

$$\xi - \left[\prod_{s=S}^{0} A(s)\right] x_0 = G_1\zeta \tag{14}$$

where $G_1 = \sum_{s=0}^{S}\left[\prod_{i=S}^{s+1} A(i)\right] B(s)B'(s)\left[\prod_{i=S}^{s+1} A(i)\right]'$, and $\prod_{i=S}^{s} A(i) \triangleq A(S)A(S-1)\cdots A(s)$ with $\prod_{i=S}^{S+1} A(i) \triangleq I$.

Proof **Sufficiency:** Assuming that Eq. (14) holds, we prove that ξ is reachable from initial value x_0 for Eq. (1). In this case, we design the following controller

$$u(s) = B'(s)\left[\prod_{i=S}^{s+1} A(i)\right]'\zeta,$$

where ζ satisfies Eq. (14). Together with Eq. (1), it has

$$x(S+1) = \prod_{s=S}^{0} A(s)x_0 + \sum_{s=0}^{S}\prod_{i=S}^{s+1} A(i)B(s)u(s) = \prod_{s=S}^{0} A(s)x_0 + G_1\zeta = \xi.$$

Thus, ξ is reachable at time $S+1$ from the initial value x_0 for Eq. (1).

Necessity: Assuming that ξ is reachable at time $S+1$ from initial value x_0 for Eq. (1), we prove that there exists ζ such that Eq. (14) holds. Otherwise, for any ζ, it holds that $\xi - \left[\prod_{s=S}^{0} A(s)\right] x_0 \neq G_1\zeta$. Then there exists a vector $\alpha \in \mathbb{R}^n$, such that $G_1\alpha = 0$ and $\alpha'\{\xi - \left[\prod_{s=S}^{0} A(s)\right] x_0\} \neq 0$.

On the one hand, from $G_1\alpha = 0$, we have

$$0 = \alpha' G_1\alpha = \sum_{s=0}^{S}\left\|\alpha'\prod_{i=S}^{s+1} A(i)B(s)\right\|^2,$$

which implies that $\alpha'\prod_{i=S}^{s+1} A(i)B(s) = 0$ for $s = 0,1,\ldots,S$.

On the other hand, by denoting the admissible controller such that $x(0) = x_0$ and $x(S+1) = \xi$ by $u(s)$, it is obtained from Eq. (1) that

$$\xi - \prod_{s=S}^{0} A(s)x_0 = \sum_{s=0}^{S}\left[\prod_{i=S}^{s+1} A(i)\right] B(s)u(s).$$

Combining with $\alpha'\{\xi - \left[\prod_{s=S}^{0} A(s)\right] x_0\} \neq 0$, it follows that

$$\sum_{s=0}^{S}\left\{\alpha'\left[\prod_{i=S}^{s+1} A(i)\right] B(s)u(s)\right\} \neq 0,$$

which is a contradiction to the fact that $\alpha'\prod_{i=S}^{s+1} A(i)B(s) = 0$. Thus, there exists ζ such that Eq. (14) holds. ■

We now provide the solvability condition of Eq. (4).

Lemma 3 If ξ is reachable at time $S+1$ from initial value x_0 for Eq. (1), then Eq. (4) has a solution.

Proof We will prove this in two steps. Firstly, a new system where the reachability of ξ is equivalent to that in Eq. (1) is introduced. The second step is to verify that Eq. (4) has a solution under the condition that ξ is reachable for Eq. (1).

Firstly, under the reachability of ξ from the initial value x_0 for Eq. (1), there exists an admissible controller denoted by $u(s)$ such that $x(0) = x_0$ and $x(S+1) = \xi$. We now introduce the following discrete-time system

$$\hat{x}(s+1) = A_c(s)\hat{x}(s) - \hat{B}(s)\hat{u}(s) \tag{15}$$

where the initial value is $\hat{x}(0) = x_0$ and

$$\hat{B}(s) = B(s)[\mathcal{R} + B'(s)P(s+1)B(s)]^{-\frac{1}{2}},$$
$$\hat{u}(s) = -[\mathcal{R} + B'(s)P(s+1)B(s)]^{\frac{1}{2}}u(s) - [\mathcal{R} + B'(s)P(s+1)B(s)]^{-\frac{1}{2}}B'(s)P(s+1)\times A(s)x(s),$$

while $x(s)$ is the solution of Eq. (1) responding to the control $u(s)$.

Rewriting Eq. (15) yields

$$\hat{x}(s+1) = A(s)\hat{x}(s) + B(s)u(s) + B(s)K(s)[\hat{x}(s) - x(s)] \tag{16}$$

It is obvious that $x(s)$ satisfies Eq. (15). Accordingly, under the controller $\hat{u}(s)$, the state $\hat{x}(s)$ in Eq. (15) satisfies $\hat{x}(0) = x_0$ and $\hat{x}(S+1) = \xi$. By using Lemma 2, there exists a vector $\lambda_1 \in \mathbb{R}^n$, such that

$$\xi = \Phi(0,S)x_0 + G(0)\lambda_1 \tag{17}$$

that is, $\xi - \Phi(0,S)x_0 = G(0)\lambda_1$. Thus, Eq. (4) has a solution $\lambda^* = -\lambda_1$. The proof is now completed. ■

Based on Lemmas 1 and 3, we present the optimal solution of the problem as follows.

Theorem 1 If ξ is reachable at time $S+1$ from initial value x_0 for Eq. (1), then Eq. (4) has a solution λ^* and there exists a unique solution to the problem which is given by Eq. (5).

Proof The conclusion can be directly obtained by using Lemmas 1 and 3. The proof is now completed. ■

IV. Q-Learning Algorithm

In this section, we aim to present the Q-learning algorithm for the problem with unknown matrices $A(s)$ and $B(s)$. From Theorem 1, the optimal controller $u(s)$ in Eq. (5) is in the feedback of the state and a specific Lagrange multiplier λ^* which satisfies Eq. (4). To this end, the design of the Q-learning algorithm is divided into two steps. The first step is to learn feedback gains $K(s)$ and $K_1(s)$, the solution $P(s)$ of the Riccati equation, and matrices $\Phi(s,S)$ and $G(s)$. This gives a data-driven representation of matrix gains $\Phi(0,S)$ and $G(0)$ in Eq. (4). Then, the second step is to solve the data-driven matrix equation to derive the parameter λ^*.

A. Q-Function

Firstly, to learn $P(s)$ and feedback gains $K(s)$ and $K_1(s)$, we introduce the Q-function by using the dynamic programming. To this end, we first consider the optimal control problem of minimizing J_1 in Eq. (6) with any parameter λ and define the Q-function for $s=0,1,\ldots,S$ by

$$\begin{aligned} Q(w) = &\sum_{s=w}^{S}[x'(s)Qx(s)+u'(s)\mathcal{R}u(s)]+ \\ &x'(S+1)Hx(S+1)+2\lambda' x(S+1) \end{aligned} \tag{18}$$

By applying the optimality principle, it follows that

$$Q(w) = x'(w)Qx(w)+u'(w)\mathcal{R}u(w)+Q(w+1) \tag{19}$$

where $Q(S+1) = x'(S+1)Hx(S+1)+2\lambda' x(S+1)$. The representation of the Q-function can be given as below.

Lemma 4 The Q-function defined by Eq. (19) is given by

$$Q(w) = x'(w)P(w)x(w)+2x'(w)\Phi'(w,S)\lambda-\lambda' G(w)\lambda \tag{20}$$

The optimal controller to minimize J_1 in Eq. (6) with any parameter λ is

$$u(w) = K(w)x(w)+K_1(w)\lambda \tag{21}$$

Proof The proof is provided by using induction. Firstly, from $w=S$ and using Eq. (1), we obtain

$$\begin{aligned} Q(S) = &x'(S)Qx(S)+u'(S)\mathcal{R}u(S)+ \\ &x'(S+1)Hx(S+1)+2\lambda' x(S+1) = \\ &x'(S)[A'(S)HA'(S)+Q]x(S)+ \\ &u'(S)\Gamma(S)u(S)+ \\ &2x'(S)A'(S)HB(S)u(S)+2\lambda' A(S)x(S)+ \\ &2\lambda' B(S)u(S) = \\ &x'(S)P(S)x(S)+ \\ &[u'(S)-K(S)x(S)-K_1(S)\lambda]'\times \\ &\Gamma(S)[u'(S)-K(S)x(S)-K_1(S)\lambda]+ \\ &2x'(S)A_c'(S)\lambda-\lambda'\bar{B}(S)\lambda. \end{aligned}$$

Thus, the optimal controller is given by Eq. (21) and the corresponding Q-function is given as below

$$Q(S) = x'(S)P(S)x(S)+2x'(S)\Phi'(S,S)\lambda-\lambda' G(S)\lambda.$$

Assuming that the following equation holds

$$\begin{aligned} Q(w+1) = &x'(w+1)P(w+1)x(w+1)+ \\ &2x'(w+1)\Phi'(w+1,S)\lambda-\lambda' G(w+1)\lambda, \end{aligned}$$

we have

$$\begin{aligned} Q(w) = &x'(w)Qx(w)+u'(w)\mathcal{R}u(w)+ \\ &x'(w+1)P(w+1)x(w+1)+ \\ &2x'(w+1)\Phi'(w+1,S)\lambda-\lambda' G(w+1)\lambda = \\ &x'(w)P(w)x(w)+[u(w)-K(w)x(w)- \\ &K_1(w)\lambda]'\Gamma(S)[u(w)- \\ &K(w)x(w)-K_1(w)\lambda]+ \\ &2x'(w)\Phi'(w,S)\lambda-\lambda' G(w)\lambda. \end{aligned}$$

This implies that the optimal controller is given by Eq. (21) and the Q-function in Eq. (20) follows. ■

Based on Lemma 4, we now derive the data-driven representation of the Q-function.

$$\begin{aligned} Q(w) = &x'(w)[Q+A'(w)P(w+1)A(w)]x(w)+ \\ &u'(w)[\mathcal{R}+B'(w)P(w+1)B(w)]u(w)+ \\ &2x'(w)A'(w)P(w+1)B(w)u(w)+ \\ &2x'(w)A'(w)\Phi'(w+1,S)\lambda+ \\ &2u'(w)B'(w)\Phi'(w+1,S)\lambda-\lambda' G(w+1)\lambda = \\ &\begin{bmatrix} x(w) \\ u(w) \\ \lambda \end{bmatrix}' \Lambda(w) \begin{bmatrix} x(w) \\ u(w) \\ \lambda \end{bmatrix} \end{aligned} \tag{22}$$

where

$$\Lambda(w) \triangleq \begin{bmatrix} \Lambda_{11}(w) & \Lambda_{21}'(w) & \Lambda_{31}'(w) \\ \Lambda_{21}(w) & \Lambda_{22}(w) & \Lambda_{32}'(w) \\ \Lambda_{31}(w) & \Lambda_{32}(w) & \Lambda_{33}(w) \end{bmatrix},$$

$$\begin{aligned} &\Lambda_{11}(w) = Q+A'(w)P(w+1)A(w), \\ &\Lambda_{21}(w) = B'(w)P(w+1)A(w), \\ &\Lambda_{22}(w) = \Gamma(w), \\ &\Lambda_{31}(w) = \Phi(w+1,S)A(w), \\ &\Lambda_{32}(w) = \Phi(w+1,S)B(w), \\ &\Lambda_{33}(w) = -G(w+1). \end{aligned}$$

From Eq. (22), the feedback gains in Eq. (21) are reformulated as

$$K(w) = -\Lambda_{22}^{-1}(w)\Lambda_{21}(w) \tag{23}$$

$$K_1(w) = -\Lambda_{22}^{-1}(w)\Lambda_{32}'(w) \tag{24}$$

The Riccati equation is formulated by

$$P(w) = \Lambda_{11}(w)-\Lambda_{21}'(w)\Lambda_{22}^{-1}(w)\Lambda_{21}(w) \tag{25}$$

B. Q-Learning Algorithm

We are now in the position to illustrate the derivation of the Q-learning algorithm. To this end, we take l experiments by entering the inputs $u^i(s)$ and the state $x^i(s)$ where

$i = 1,2,\ldots,l$, and derive the measurements $x^i(s+1)$ of the state for $s = 0,1,\ldots,S$. Moreover, we also enter a parameter vector $\lambda^i(s)$, $i = 1,2,\ldots,l$ to obtain the data-driven Q-function in Eq. (20) for $s = 0,1,\ldots,S$.

Firstly, with the acquired data information $(x^i(S), u^i(S), \lambda^i(S), x^i(S+1))$, we define

$$\begin{aligned}\gamma^i(S) = {} & [x^i(S)]'Qx^i(S) + [u^i(S)]'\mathcal{R}u^i(S) + \\ & [x^i(S+1)]'H[x^i(S+1)] + \\ & 2[x^i(S+1)]'\lambda^i(S)\end{aligned} \tag{26}$$

From Eq. (22), we get

$$\gamma^i(S) = \begin{bmatrix} x^i(S) \\ u^i(S) \\ \lambda^i(S) \end{bmatrix}' \Lambda(S) \begin{bmatrix} x^i(S) \\ u^i(S) \\ \lambda^i(S) \end{bmatrix} \tag{27}$$

This gives the solvability of the matrix $\Lambda(S)$ as follows

$$\Lambda(S) = \arg\min \sum_{i=1}^{l} \|[z^i(S)]'\Lambda(S)z^i(S) - \gamma^i(S)\| \tag{28}$$

where

$$z^i(S) = \begin{bmatrix} x^i(S) \\ u^i(S) \\ \lambda^i(S) \end{bmatrix} \triangleq \begin{bmatrix} z_1^i(S) & z_2^i(S) & \cdots & z_{2n+m}^i(S) \end{bmatrix}.$$

In fact, by denoting the vectorization of the upper triangular part of the matrix $\Lambda(S)$ as $\nu(S)$, the solvability of Eq. (28) can be further reduced to the optimization problem as below

$$\begin{aligned}&\arg\min\|\Upsilon(S)\nu(S) - \gamma(S)\|^2 = \\ &[\Upsilon'(S)\Upsilon(S)]^{-1}\Upsilon'(S)\gamma(S)\end{aligned} \tag{29}$$

where

$$\begin{aligned}\Upsilon(S) &= \begin{bmatrix} \Upsilon^1(S) & \Upsilon^2(S) & \cdots & \Upsilon^l(S) \end{bmatrix}, \\ \gamma(S) &= \begin{bmatrix} \gamma^1(S) & \gamma^2(S) & \cdots & \gamma^l(S) \end{bmatrix}'\end{aligned} \tag{30}$$

while

$$\begin{aligned}\Upsilon^i(S) = \Big[& [z_1^i(S)]^2 \quad z_1^i(S)z_2^i(S) \quad \cdots \quad z_1^i(S)z_{2n+m}^i(S) \\ & [z_2^i(S)]^2 \quad z_2^i(S)z_3^i(S) \quad \cdots \quad z_2^i(S)z_{2n+m}^i(S) \quad \cdots \\ & [z_{2n+m}^i(S)]^2\Big]'.\end{aligned}$$

Accordingly, the feedback gains at time S are

$$K(S) = -\Lambda_{22}^{-1}(S)\Lambda_{21}(S) \tag{31}$$

$$K_1(S) = -\Lambda_{22}^{-1}(S)\Lambda_{32}'(S) \tag{32}$$

$P(S)$ satisfying the Riccati equation is given by

$$P(S) = \Lambda_{11}(S) - \Lambda_{21}'(S)\Lambda_{22}^{-1}(S)\Lambda_{21}(S) \tag{33}$$

and matrices $\Phi(S,S)$ and $G(S)$ are calculated by

$$\Phi(S,S) = \Lambda_{31}(S) - \Lambda_{32}(S)\Lambda_{22}^{-1}(S)\Lambda_{21}(S) \tag{34}$$

$$G(S) = \Lambda_{32}(S)\Lambda_{22}^{-1}(S)\Lambda_{32}'(S) \tag{35}$$

Secondly, by iterating backward from S, using the acquired data information $(x^i(s), u^i(s), \lambda^i(s), x^i(s+1))$ and the data-driven matrix representations in the last step, we define

$$\begin{aligned}\gamma^i(s) = {} & [x^i(s)]'Qx^i(s) + [u^i(s)]'\mathcal{R}u^i(s) + \\ & [x^i(s+1)]'P(s+1)x^i(s+1) + \\ & 2[x^i(s+1)]'\Phi'(s+1,S)\lambda^i(s) - \\ & [\lambda^i(s)]'G(s+1)\lambda^i(s)\end{aligned} \tag{36}$$

which can also be obtained from Eq. (22) that

$$\gamma^i(s) = \begin{bmatrix} x^i(s) \\ u^i(s) \\ \lambda^i(s) \end{bmatrix}' \Lambda(s) \begin{bmatrix} x^i(s) \\ u^i(s) \\ \lambda^i(s) \end{bmatrix} \tag{37}$$

Thus, $\Lambda(s)$ is solved as follows

$$\Lambda(s) = \arg\min \sum_{i=1}^{l} \|[z^i(s)]'\Lambda(s)z^i(s) - \gamma^i(s)\| \tag{38}$$

Similar to the derivation of Eq. (29), we solve Eq. (38) by the following optimization problem

$$\begin{aligned}&\arg\min\|\Upsilon(s)\nu(s) - \gamma(s)\|^2 = \\ &[\Upsilon'(s)\Upsilon(s)]^{-1}\Upsilon'(s)\gamma(s)\end{aligned} \tag{39}$$

Then the feedback gains $K(s)$ and $K_1(s)$ are given by

$$K(s) = -\Lambda_{22}^{-1}(s)\Lambda_{21}(s) \tag{40}$$

$$K_1(s) = -\Lambda_{22}^{-1}(s)\Lambda_{32}'(s) \tag{41}$$

matrix $P(s)$ is provided by

$$P(s) = \Lambda_{11}(s) - \Lambda_{21}'(s)\Lambda_{22}^{-1}(s)\Lambda_{21}(s) \tag{42}$$

and matrices $\Phi(s,S)$ and $G(s)$ are shown by

$$\Phi(s,S) = \Lambda_{31}(s) - \Lambda_{32}(s)\Lambda_{22}^{-1}(s)\Lambda_{21}(s) \tag{43}$$

$$G(s) = G(s+1) + \Lambda_{32}(s)\Lambda_{22}^{-1}(s)\Lambda_{32}'(s) \tag{44}$$

Finally, we give the calculation of the parameter λ^*. From Eq. (4), we have

$$\lambda^* = G^{\dagger}(0)[\Phi(0,S)x_0 - \xi],$$

where $G^{\dagger}$ indicates the Moore-Penrose inverse of matrix G.

Combining with the data-driven representation in Eqs. (43) and (44) at time $s = 0$, it is further obtained that

$$\begin{aligned}\lambda^* = {} & [-\Lambda_{33}(0) + \Lambda_{32}(0)\Lambda_{22}^{-1}(0)\Lambda_{32}'(0)]^{\dagger}\Big[[\Lambda_{31}(0) - \\ & \Lambda_{32}(0)\Lambda_{22}^{-1}(0)\Lambda_{21}(0)]x_0 - \xi\Big]\end{aligned} \tag{45}$$

The detailed algorithm is concluded in Algorithm 1.

Algorithm 1 Q-Learning algorithm

1: Implement l times experiments on Eq. (1), where for the i-th experiment with $i = 1,2,\ldots,l$, obtain the data samples $(x^i(s), u^i(s), \lambda^i(s), x^i(s+1))$ for $s = 0,1,\ldots,S$.

2: Calculate $\gamma^i(S)$ by Eq. (26) and solve the optimization problem in Eq. (29). Calculate the feedback gains $K(S)$ and $K_1(S)$ by Eqs. (31) and (32), $P(S)$ by Eq. (33), and the matrices $\Phi(S,S)$ and $G(S)$ by Eqs. (34) and (35).

3: From $s = S-1$ to $s = 0$, use repeated steps to calculate $\gamma^i(s)$ by Eq. (36) and solve the optimization problem by Eq. (39). Calculate the gains $K(s)$ and $K_1(s)$ by Eqs. (40) and (41), $P(s)$ by Eq. (42), and the matrices $\Phi(s,S)$ and $G(s)$ by Eqs. (43) and (44).

4: Calculate the parameter λ^* by Eq. (45).

5: Derive the optimal controller $u(s)$ by Eq. (5).

The convergence of Algorithm 1 is given as follows.

Theorem 2 Assume that ξ is reachable at time $S+1$ from initial value x_0 for Eq. (1). Collecting $l \geqslant (2n+m)(2n+m+1)/2$ data samples $(x^i(s), u^i(s), \lambda^i(s))$, $i = 1,2,\ldots,l$ for $s = 0,1,\ldots,S$, which are Gaussian with arbitrary mean and arbitrarily positive-definite covariances, then the optimal solution of Eq. (39) exists and the controller in Eq. (5) obtained in Algorithm 1 is the optimal solution of the problem.

Proof By collecting data $x^i(s)$, $u^i(s)$, and $\lambda^i(s)$, $i = 1,2,\ldots,l$ which are Gaussian at time s, we have that the experiments are linearly independent. Considering that the experiment time $l \geqslant (2n+m)(2n+m+1)/2$, it is known that the matrix $\Upsilon(s)$ in Eq. (30) has full rank [23], which implies that Eq. (39) holds. Together with Theorem 1, the derived controller in Algorithm 1 is optimal for the problem. The proof is now completed. ■

V. Numerical Example

In this section, we present a numerical example to show the effectiveness of the proposed algorithm. Consider Eq. (1) with the following constant matrices

$$A(0) = \begin{bmatrix} 1 & 2 \\ -1 & 4 \end{bmatrix}, A(1) = \begin{bmatrix} 5 & 3 \\ -2 & 1 \end{bmatrix},$$
$$A(2) = \begin{bmatrix} -4 & 1 \\ 2 & 5 \end{bmatrix}, B(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$
$$B(1) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, B(2) = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, Q = I,$$
$$\mathcal{R} = 1,\ S = 2,\ x_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \xi = \begin{bmatrix} 6 \\ 7 \end{bmatrix}.$$

By implementing 30 experiments, we have the solution of the optimization problem in Eq. (39) as

$$v(2) = \begin{bmatrix} 21.00 \\ 6.00 \\ -12.00 \\ -4.00 \\ 2.00 \\ 27.00 \\ 14.00 \\ 1.00 \\ 5.00 \\ 21.00 \\ 4.00 \\ 2.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}, v(1) = \begin{bmatrix} 145.24 \\ 162.81 \\ 120.10 \\ -5.24 \\ 8.38 \\ 229.95 \\ 172.52 \\ -6.81 \\ 13.10 \\ 131.24 \\ -5.10 \\ 9.95 \\ -0.76 \\ -0.38 \\ -0.19 \end{bmatrix}, v(0) = \begin{bmatrix} 29.63 \\ 67.93 \\ 28.63 \\ -0.46 \\ -0.74 \\ 271.78 \\ 67.93 \\ -1.60 \\ -1.40 \\ 29.63 \\ -0.46 \\ -0.74 \\ -0.96 \\ 0.01 \\ -0.95 \end{bmatrix}.$$

This accordingly gives $P(s)$ in Eq. (42) as

$$P(3) = I,$$
$$P(2) = \begin{bmatrix} 14.1429 & 14.0000 \\ 14.0000 & 17.6667 \end{bmatrix},$$
$$P(1) = \begin{bmatrix} 35.3396 & 4.9340 \\ 4.9340 & 3.1549 \end{bmatrix},$$
$$P(0) = \begin{bmatrix} 1.9662 & 2.2928 \\ 2.2928 & 116.0380 \end{bmatrix},$$

the feedback gains in Eqs. (40) and (41) as

$$K(0) = \begin{bmatrix} -0.9662 & -2.2928 \end{bmatrix},$$
$$K(1) = \begin{bmatrix} -0.9151 & -1.3146 \end{bmatrix},$$
$$K(2) = \begin{bmatrix} 0.5714 & -0.6667 \end{bmatrix},$$
$$K_1(0) = \begin{bmatrix} 0.0157 & 0.0249 \end{bmatrix},$$
$$K_1(1) = \begin{bmatrix} 0.0388 & -0.0758 \end{bmatrix},$$
$$K_1(2) = \begin{bmatrix} -0.1905 & -0.0952 \end{bmatrix},$$

and the parameter in Eq. (45) as

$$\lambda^* = \begin{bmatrix} -7.2802 \\ -6.6461 \end{bmatrix}.$$

It is easily verified that Eq. (1) under the optimal controller in Eq. (5) achieves the specific terminal state $\xi = [6\ 7]'$.

VI. Conclusion

In this paper, we considered the LQ optimal control problem with terminal state constraint. A Q-learning algorithm was proposed for the optimal controller when the coefficient matrices of state and control input in the time-varying system are both unknown, consisting of learning the solution to the Riccati equation and calculating the specific Lagrange multiplier to a data-driven matrix equation. A numerical example was provided to show the effectiveness of the proposed algorithm. The LQ optimal control problem of time-varying system with terminal constraint and the delays will be studied in our future work.

Acknowledgment

This work was supported by the Natural Science Foundation of Shandong Province (Nos. ZR2021ZD14, ZR2021JQ24, and ZR2024QF198).

References

[1] N. E. Kahveci, P. A. Ioannou, and M. D. Mirmirani, Adaptive LQ control with anti-windup augmentation to optimize UAV performance in autonomous soaring applications, *IEEE Trans. Control Syst. Technol*., 2008, 16(4), 691–707.

[2] Y.-H. Ni, X. Li, J.-F. Zhang, and M. Krstic, Equilibrium solutions of multiperiod mean-variance portfolio selection, *IEEE Trans. Autom. Control*, 2020, 65(4), 1716–1723.

[3] R. E. Bellman, I. Glicksberg, and O. A. Gross, *Some Aspects of the Mathematical Theory of Control Processes*. Santa Monica, CA, USA: RAND Corporation, 1958.

[4] W. M. Wonham, On a matrix Riccati equation of stochastic control, *SIAM J. Control*, 1968, 6(4), 681–697.

[5] M. A. Rami, X. Chen, and X. Y. Zhou, Discrete-time indefinite LQ control with state and control dependent noises, *J. Global Optim*., 2002, 23(3), 245–265.

[6] R. E. Kalman, Contributions to the theory of optimal control, *Bol. Soc. Mat. Mex*., 1960, 5(2), 102–119.

[7] P. Ignaciuk and A. Bartoszewicz, Linear-quadratic optimal control of periodic-review perishable inventory systems, *IEEE Trans. Control Syst. Technol*., 2012, 20(5), 1400–1407.

[8] J. Yong, A deterministic linear quadratic time-inconsistent optimal control problem, *Math. Control Relat. Fields*, 2011, 1(1), 83–118.

[9] H. Zhang and J. Xu, Optimal control with irregular performance, *Sci. China Inf. Sci*., 2019, 62(9), 192203.

[10] Y.-Z. Luo and G.-J. Tang, Spacecraft optimal rendezvous controller design using simulated annealing, *Aerosp. Sci. Technol*., 2005, 9(8), 732–737.

[11] W. Wu, J. Gao, J.-G. Lu, and X. Li, On continuous-time constrained stochastic linear—quadratic control, *Automatica*, 2020, 114, 108809.

[12] R. Mitze and M. Mönnigmann, A dynamic programming approach to solving constrained linear-quadratic optimal control problems, *Automatica*, 2020, 120, 109132.

[13] M. Giftthaler and J. Buchli, A projection approach to equality constrained iterative linear quadratic optimal control, in *Proc. IEEE-RAS 17th International Conference on Humanoid Robotics*, Birmingham, UK, 2017, 61–66.

[14] A. Sideris and L. A. Rodriguez, A Riccati approach to equality constrained linear quadratic optimal control, in *Proc. 2010 American Control Conference*, Baltimore, MD, USA, 2010, 5167–5172.

[15] A. Ferrante and L. Ntogramatzidis, A unified approach to the finite-horizon LQ regulator—Part I: The continuous time, in *Proc. 45th IEEE Conference on Decision and Control*, San Diego, CA, USA, 2006, 5651–5656.

[16] J. Sun, Linear quadratic optimal control problems with fixed terminal states and integral quadratic constraints, *Appl. Math. Optim*., 2021, 83(1), 251–276.

[17] J. H. Park, S. Han, and W. H. Kwon, LQ tracking controls with fixed terminal states and their application to receding horizon controls, *Syst. Control Lett*., 2008, 57(9), 772–777.

[18] Z. Zhang, J. Xu, and M. Fu, Q-learning for feedback Nash strategy of finite-horizon nonzero-sum difference games, *IEEE Trans. Cybern*., 2022, 52(9), 9170–9178.

[19] G. C. Calafiore and C. Possieri, Output feedback Q-learning for linear-quadratic discrete-time finite-horizon control problems, *IEEE Trans. Neural Netw. Learn. Syst*., 2021, 32(7), 3274–3281.

[20] S. A. A. Rizvi and Z. Lin, Output feedback Q-learning for discrete-time linear zero-sum games with application to the H-infinity control, *Automatica*, 2018, 95, 213–221.

[21] T. Feng, J. Zhang, Y. Tong, and H. Zhang, Q-learning algorithm in solving consensusability problem of discrete-time multi-agent systems, *Automatica*, 2021, 128, 109576.

[22] H. Zhang, L. Li, J. Xu, and M. Fu, Linear quadratic regulation and stabilization of discrete-time systems with delay and multiplicative noise, *IEEE Trans. Autom. Control*, 2015, 60(10), 2599–2613.

[23] D. Simon, *Optimal State Estimation*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

Juanjuan Xu received the BS degree in mathematics from Qufu Normal University, China, in 2006, the MS degree in mathematics from Shandong University, China, in 2009, and the PhD degree in control theory from Shandong University, China, in 2013. She is currently a professor at School of Control Science and Engineering, Shandong University, China. Her research interests include distributed control, optimal control, game theory, and reinforcement learning based control.

Jingmei Liu received the BS and MS degrees in mathematics from Qufu Normal University, China, in 2015 and 2018, respectively, and the PhD degree in control theory from Shandong University, China, in 2023. She is currently a lecturer at School of Automation and Electrical Engineering, Linyi University, China. Her research interests include optimal control and game theory.

Zhaorong Zhang received the BS degree in electrical engineering from Nanjing University of Science and Technology, Nanjing, China, in 2016 and the PhD degree from University of Newcastle, Newcastle, Australia, in 2021. She worked as a postdoctoral fellow at The Hong Kong Polytechnic University, China and The Chinese University of Hong Kong, China, from 2021 to 2023 and joined Shandong University, China, in 2024. Her research interests include distributed algorithms, reinforcement learning, and networked control systems.

Wei Wang received the PhD degree in control science and engineering from Shenzhen Graduate School, Harbin Institute of Technology, China, in 2010. He is currently a professor at School of Control Science and Engineering, Shandong University, China. His research interests include optimal control and estimation for delayed systems, distributed control, and estimation.