International Journal of Control, Automation and Systems 17(12) (2019) 3140-3148
http://dx.doi.org/10.1007/s12555-019-0083-8
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Adaptive Dynamic Programming for Minimal Energy Control with Guaranteed Convergence Rate of Linear Systems

Kai Zhang, Suoliang Ge*, and Yuling Ge

Abstract: The traditional linear quadratic optimal control can be summarized as finding the state feedback controller, so that the closed-loop system is stable and the performance index is minimum. And it is well known that the solution of the linear quadratic optimal control problem can be obtained by algebraic Riccati equation (ARE) with the standard assumptions. However, results developed for the traditional linear quadratic optimal control problem cannot be directly applied to solve the problem of minimal energy control with guaranteed convergence rate (MECGCR), because the standard assumptions cannot be satisfied in the MECGCR problem. In this paper, we mainly consider the problem of MECGCR and prove that ARE can be applied to solve the MECGCR problem under some conditions. Furthermore, with the assumption that the system dynamics is unknown, we propose a policy iteration (PI) based adaptive dynamic programming (ADP) algorithm to iteratively solve the ARE using the online information of state and input, without requiring the a priori knowledge of the system matrices. Finally, a numerical example is worked out to show the effectiveness of the proposed approach.

Keywords: Adaptive dynamic programming, guaranteed convergence rate, minimal energy control, policy iteration.

1. INTRODUCTION

The linear quadratic regulator (LQR) has been widely used in various industrial applications [1–4] due to its high practicability, high efficiency and simple structure. Such a controller design method can be described as follows. Consider a continuous-time linear system described by

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

where $x \in \mathbb{R}^n$ is the system state vector; $u \in \mathbb{R}^m$ is the control input; $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant matrices. Finding a state feedback controller in the form of

$$u = -Kx, \tag{2}$$

which minimizes the following performance index

$$J(x_0, u) = \int_0^\infty (x^T Qx + u^T Ru)dt, \tag{3}$$

where x_0 is the initial state vector of system; $Q = Q^T \geq 0$, $R = R^T > 0$. The solution of the LQR can be obtained by the following ARE [5]

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

Equation (4) has a unique symmetric positive definite solution P^*_{LQR}, if the following standard assumptions are satisfied: 1) The pair (A, B) is stabilizable; 2) The pair $(A, Q^{1/2})$ has no unobservable modes on the imaginary axis. Then the optimal feedback gain matrix in (2) can thus be determined by

$$K^*_{LQR} = R^{-1}B^T P^*_{LQR}. \tag{5}$$

Since (4) is nonlinear in P, it is usually difficult to directly solve it, especially for high-order systems. Nevertheless, many efficient algorithms have been proposed to numerically approximate the solution to (4). One of such algorithms is the famous Kleinman algorithm [6]. According to the Kleinman algorithm, the solution of the ARE can be numerically approximated by iteratively solving the Lyapunov equation. However, the information of the system is needed in the Kleinman algorithm. When the information of the system is unknown, the general approach to design an adaptive optimal control law can be pursued by first identifying the system parameters, and then solving the related ARE [7, 8]. However, this algorithm responds slowly to parameter variations of the system.

Inspired by the learning behavior from biological systems, reinforcement learning (RL) theories have been broadly applied for solving optimal control problems for unknown systems in recent years [9–13]. With the help of RL, an adaptive optimal control scheme for linear systems with unknown internal system dynamics was studied

Manuscript received February 9, 2019; revised May 27, 2019; accepted July 1, 2019. Recommended by Associate Editor Hyun Myung under the direction of Editor PooGyeon Park.

Kai Zhang, Suoliang Ge, and Yuling Ge are with the School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, Anhui 230009, China (e-mails: zhangkaistd@mail.hfut.edu.cn, gesuol@163.com, 2018110326@mail.hfut.edu.cn).
* Corresponding author.

in [14–16]. However, this method is based on the case that partial knowledge of the system dynamics is exactly known. In order to remove the assumption on partial knowledge of the system dynamics, a computational model-free ADP methodology was proposed in [17]. This approach can serve as a computational tool to study ADP related problems for linear systems, such as multiagent systems [18, 19], adaptive optimal output regulation problem [20], markov jump linear systems [21] and zero-sum games [22, 23] and so on.

These ADP algorithms mentioned above can be employed to online numerically approximate the solution of the optimal control problems in the case that the standard assumptions are satisfied. However, the purpose of the minimal energy control (MEC) is to find the state feedback controller, so that the closed-loop system is stable and the following performance index is minimum [24–26].

$$J(x_0,u)=\int_0^\infty u^T R u dt. \tag{6}$$

Compared with (3), the performance index (6) shows that the special case where $Q=0$. That implies the standard assumptions cannot be satisfied in the MEC problem and these ADP algorithms cannot be directly applied to find the numerical approximate solution of the MEC. On the other hand, a common feature of the above ADP-based results is that the convergence rate is not pre-specified. This may result in the phenomena of slow convergence. From the engineering point of view, we hope that the eigenvalues of the optimal control system are located in a certain region of the s-plane (i.e., the asymptotically stable optimal control systems have a given dynamic performance). However, there are a few studies related to optimal control with guaranteed convergence rate and these studies are based on the system models, such as linear-quadratic optimal observers with guaranteed convergence rate [27] and so on.

It can be seen from the above mentioned that the application of ADP algorithm in both minimum energy control and optimal control with guaranteed convergence rate is still an open question. In order to solve this question, in this paper, we first consider the problem of MECGCR and prove that ARE can be applied to this problem under some conditions. Futhermore, a PI based ADP algorithm is proposed for finding the approximate solution of the problem of MECGCR without using the system information. Our technical contributions from this paper can be briefly summarized as follows:

1) To the best of our knowledge, this note is the first attempt to apply the PI based ADP algorithm for the the problem of MECGCR. By using the online information of state and input, the PI based ADP algorithm can find the approximate solution of the MECGCR without using the system information A or B, or both.
2) The stability and convergence of the proposed PI based ADP algorithm are analyzed, and the conditions that ensure the stability of closed-loop system have been formulated.
3) The approach proposed in this paper can serve as a computational tool to study the MECGCR related problems, such as multiagent systems, tracking control and zero-sum games and so on.

The rest of this paper is organized as follows: In Section 2, we first give some useful lemmas, and then consider the problem of MECGCR. In Section 3, a computational adaptive optimal control method is developed and its convergence is proved. In Section 4, simulation study on a three-order linear system is provided. In Section 5, concluding remarks as well as potential future work are contained.

Notations: Throughout this paper, $\otimes$ is used to indicate the Kronecker product, and $vec(\mathrm{A})$ is defined to be the mn-vector formed by stacking the columns of $\mathrm{A}\in\mathbb{R}^{n\times m}$on top of one another, i.e., $vec(\mathrm{A})=\begin{bmatrix} a_1^T & a_2^T & \cdots & a_m^T \end{bmatrix}^T$, where $a_i\in\mathbb{R}^n$ are the columns of A. $\mathrm{Re}(\lambda(\mathrm{A}))$ denotes the set of the real parts of the eigenvalues of A. If A and B are positive semidefinite, $\mathrm{A}>\mathrm{B}$ $(\mathrm{A}\geq\mathrm{B})$ is used to denote the matric $(\mathrm{A}-\mathrm{B})$ is positive (semi-positive) definite.

2. PRELIMINARIES AND PROBLEM STATEMENT

2.1. Some useful lemmas

Lemma 1 [28]**:** If the n-dimensional pair (A,B) is controllable and all eigenvalues of A have negative real parts, then the unique solution W_c of

$$AW_c+W_cA^T=-BB^T \tag{7}$$

is positive definite. The solution is called the controllability Gramian and can be expressed as

$$W_c=\int_0^\infty e^{A\tau}BB^Te^{A^T\tau}d\tau. \tag{8}$$

Lemma 2 [30]**:** With the standard assumptions, the state feedback controller $\mathrm{u}=-R^{-1}B^TP^*x$, where P^* is the unique symmetric positive definite solution of the following ARE

$$(A+aI)^TP+P(A+aI)+Q-PBR^{-1}B^TP=0 \tag{9}$$

can minimize the following performance index

$$J(x_0,u)=\int_0^\infty e^{2at}(x^TQx+u^TRu)dt, \tag{10}$$

where $a\geq 0$, $Q=Q^T\geq 0$, $R=R^T>0$; and let the system (1) be globally exponentially stable with $\lim_{t\to\infty}x(t)e^{at}=0$.

Remark 1: The problem in Lemma 2 can be summarized as the problem of linear quadratic optimal control with guaranteed convergence rate. It can be converted to the problem of traditional linear quadratic optimal control by linear transformation and the relevant proof can be found in [29, 30]

2.2. Problem statement

In this note, we are interested in using a model-free method to iteratively solve the ARE of the problem of MECGCR. In what follows, we will introduce the problem of MECGCR and prove that ARE can be applied to this problem under some conditions.

Consider a continuous-time linear system described by (1). The design objective of MECGCR is to find a state feedback controller with the form (2), which minimizes the following performance index

$$J(x_0,u)=\int_0^\infty e^{2at}u^TRudt. \tag{11}$$

With Lemma 2, it can be proved that the following ARE can be used to solve the problem of MECGCR

$$(A+aI)^TP+P(A+aI)-PBR^{-1}B^TP=0. \tag{12}$$

Due to the special case where $Q=0$, the standard assumptions cannot be satisfied. Then the solution of (12) cannot be guaranteed to be positive definite and the closed-loop system cannot be guaranteed to be stable. Fortunately, with the help of Lemma 1, the conditions that ensure the stability of closed-loop system will be formulated without using the standard assumptions.

Theorem 1 [31]**:** Let (A,B) is controllable and let $a\geq 0$ be such that

$$a>-\min\{Re(\lambda(A))\}. \tag{13}$$

Then

1) ARE (12) has a unique positive definite solution P^*, where $P^*=W^{-1}$ and W is the unique positive definite solution to the following matrix equation

$$W(-A-aI)^T+(-A-aI)W=-BR^{-1}B^T. \tag{14}$$

2) The closed-loop system $\dot{x}=(A-BR^{-1}B^TP^*)x$ be globally exponentially stable with $\lim_{t\to\infty}x(t)e^{at}=0$ (i.e., $\max\left\{\mathrm{Re}\left(A-BR^{-1}B^TP^*\right)\right\}<-a$).

Proof:

1) With $P^*=W^{-1}$, (12) can be rewritten as (14). Then, (A,B) is controllable, hence $(-A-aI,BR^{-1/2})$ is controllable. Since a satisfies (13), all eigenvalues of $(-A-aI)$ have negative real parts. Finally, based on Lemma 1, It can be proved that $W$is the unique positive definite solution of (14), i.e., P^* is the unique positive definite solution of (12).

2) (12) can be also rewritten as the following form

$$A-BR^{-1}B^TP=P^{-1}(-A^T-2aI)P. \tag{15}$$

Hence, P^* is also the unique positive definite solution of (15). That implies that the closed-loop system matric $A-BR^{-1}B^TP^*$ and A^T-2aI are similar to each other. That is

$$Re(\lambda(A-BR^{-1}B^TP^*))=-Re((A^T))-2a, \tag{16}$$

which indicates that the eigenvalues of $A-\mathrm{B}R^{-1}B^TP^*$ are symmetric to those of A with respect to the line $\mathrm{s}=-a$ on the s-plane. Hence, $\max\left\{\mathrm{Re}\left(A-BR^{-1}B^TP^*\right)\right\}<-a$ and Properties (2) in Theorem 1 can be proved. □

Remark 2: If $a=0$, in order to ensure that ARE (12) has a unique positive definite solutionP^*, all eigenvalues of A must satisfy the following condition: $\min\{\mathrm{Re}(\lambda(A))\}>0$. This implies that most systems do not meet this requirement. Therefore, the function of parameter a can be summarized as expanding the scope of controlled objects. In addition, the parameter a also has the advantage of specifying the convergence rate of the closed-loop system.

Remark 3: Theorem 1 developed in [31] can also be applied to low gain feedback control and some relevant properties can be found in [31–33].

3. PI BASED ADP ALGORITHM FOR MECGCR

In Section 2, it is proved that if the conditions in Theorem 1 are satisfied, ARE (12) can be used to solve the MECGCR problem. However, (12) is nonlinear in P, it is usually difficult to directly solve it, especially for high-order systems. In this section, in order to solve this question, we first propose an offline PI based ADP algorithm to obtain the numerical approximate solution, which is difficult to be solved in ARE (12). Then based on the offline PI based ADP algorithm, a model-free online PI based ADP algorithm is developed.

3.1. Offline PI based ADP algorithm for MECGCR

Theorem 2: Let $K_0\in\mathbb{R}^{m\times n}$ be any stabilizing feedback gain matrix with the convergence rate which is faster than e^{-at} (i.e., $A+aI-BK_0$ is Hurwitz or $\max\{\mathrm{Re}[\lambda(A-BK_0)]\}<-a$), and repeat the following steps for $k=0,1,\ldots,$

1) Solve for $P_k=P_k^T>0$ of the Lyapunov equation

$$\begin{aligned}&(A+aI-BK_k)^TP_k+P_k(A+aI-BK_k)\\&\quad+K_k^TRK_k=0.\end{aligned} \tag{17}$$

2) Update K_k by

$$K_{k+1}=R^{-1}B^TP_k. \tag{18}$$

Then, the following properties holds:

1) $A+aI-BK_k$ is Hurwitz (i.e., $\max\{\mathrm{Re}[\lambda(A-BK_k)]\} < -a$),
2) $P^* \le P_{k+1} \le P_k$,
3) $\lim_{k\to\infty} K_k = K^*$, $\lim_{k\to\infty} P_k = P^*$.

Proof: The proof is provided in Appendix A. □

Remark 4: By solving the Lyapunov equation (17) iteratively, which is linear in P_k, and updating K_k by (18), the solution to the nonlinear equation (12) is numerically approximated by using the offline PI based ADP algorithm. In addition, this algorithm also provides theoretical support for the model-free online PI based ADP algorithm

3.2. Online PI based ADP algorithm for MECGCR

To begin with, let us consider a virtual system described by

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}\bar{u}, \tag{19}$$

where $\bar{x} = xe^{at}$, $\bar{u} = ue^{at}$, $\bar{B}$=B and $\bar{A} = A+aI$ with the diagonal matrix $I \in \mathbb{R}^{n\times n}$, and propose the following control law

$$\bar{u} = -K_k\bar{x} + \varepsilon, \tag{20}$$

where ε denotes a time-varying artificial noise, known as the exploration noise, added for the purpose of online learning. Then, the closed-loop system (19) and (20) can be written as

$$\dot{\bar{x}} = \bar{A}_k\bar{x} + \bar{B}\varepsilon, \tag{21}$$

where $\bar{A}_k = \bar{A} - \bar{B}K_k = A + aI - BK_k$. Taking the time derivative of $\bar{x}^T P_k \bar{x}$ along the system (21), it follows that

$$\frac{d}{dx}(\bar{x}^T P_k \bar{x}) = \bar{x}^T(\bar{A}_k{}^T P_k + P_k\bar{A}_k)\bar{x} + 2\varepsilon^T\bar{B}^T P_k\bar{x}. \tag{22}$$

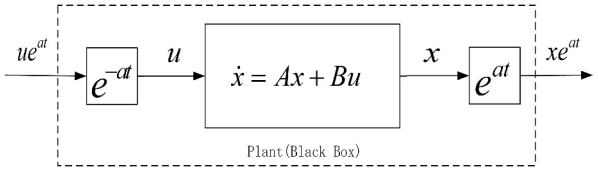
Finally, we use the term $-K_k^T RK_k$ in (17) to replace the term $\bar{A}_k{}^T P_k + P_k\bar{A}_k$ in (22) which depends on $\bar{A}$ and $\bar{B}$. Also, we use the term RK_{k+1} in (18) to replace the term $\bar{B}^T P_k$ in (22) which depends on $\bar{B}$. Therefore, (22) can be rewritten as

$$\frac{d}{dx}(\bar{x}^T P_k \bar{x}) = -\bar{x}^T(K_k^T RK_k)\bar{x} + 2\varepsilon^T RK_{k+1}\bar{x}. \tag{23}$$

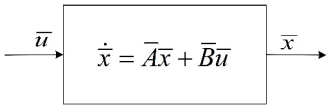
By integrating both sides of (23) on any given interval $[t, t+\delta t]$ and rearranging the terms, we have

$$\begin{aligned}&\bar{x}^T(t+\delta t)P_k\bar{x}(t+\delta t) - \bar{x}^T(t)P_k\bar{x}(t)\\ &-2\int_t^{t+\delta t}\varepsilon^T RK_{k+1}\bar{x}d\tau = -\int_t^{t+\delta t}\bar{x}^T(K_k^T RK_k)\bar{x}d\tau.\end{aligned} \tag{24}$$

It can be clearly seen that (24) relies on the knowledge of state measurements $\bar{x}(t)$, instead of the system matrices



(a) System consisting of plant (1) and the exponential signals.



(b) A virtual linear system.

Fig. 1. Two equivalent system models.

$\bar{A}$ and $\bar{B}$. If the state information $\bar{x}(t)$ can be obtained, it becomes possible to solve for P_k and K_{k+1} using the online measurements $\bar{x}$ (i.e., xe^{at}). However, in this paper, we assume that the system (1) is a black box(i.e., the matrices A and B are unknown), the virtual system (19) thus cannot be built directly and the state information $\bar{x}(t)$ cannot be measured. Fortunately, it can be proved that the system in Fig. 1(a), which consists of system (1) and the exponential signals is similar to the system in Fig. 1(b). With the same initial conditions and control law, the state measurements xe^{at} in Fig. 1(a) are similar to the state measurements $\bar{x}(t)$ in Fig. 1(b).

Hence P_k and K_{k+1} can be solved by the measurements xe^{at} (i.e., $\bar{x}$). For convenience, we still use $\bar{x}$ below.

In order to iteratively solve P_k and K_{k+1} by using (24), the specific steps are as follows. First, $\bar{x}^T P_k\bar{x}$ and $\varepsilon^T RK_{k+1}\bar{x}$ can be rewritten as the following forms by Kronecker product, respectively [34].

$$\begin{aligned}&\bar{x}^T P_k\bar{x} = \bar{x}^T \otimes \bar{x}^T vec(P_k),\\ &\varepsilon^T RK_{k+1}\bar{x} = \{\bar{x}^T \otimes (\varepsilon^T R)\}vec(K_{k+1}).\end{aligned} \tag{25}$$

With (25), (24) can be rewritten as

$$\begin{aligned}&\left[\bar{x}^T \otimes \bar{x}^T|_t^{t+\delta t} - 2\int_t^{t+\delta t}\{\bar{x}^T \otimes (\varepsilon^T R)\}dt\right]\begin{bmatrix}vec(P_k)\\ vec(K_{k+1})\end{bmatrix}\\ &= -\int_t^{t+\delta t}\bar{x}^T(K_k^T RK_k)\bar{x}dt.\end{aligned} \tag{26}$$

By specifying $t = t_{k,1}, t_{k,2}, \ldots, t_{k,l_k}$ with $0 \le t_{k,i} + \delta t \le t_{k,i+1}$ and $t_{k,i} + \delta t \le t_{k+1,1}$ for all $k = 0, 1, \ldots$ and $i = 1, 2, \ldots, l_k$, (26) can be used to generate a series of equations as follows:

$$\Theta_k\begin{bmatrix}vec(P_k)\\ vec(K_{k+1})\end{bmatrix} = \Xi_k, \tag{27}$$

where

$$\Theta_k = \begin{bmatrix} \bar{x}^T \otimes \bar{x}^T \big|_{t_{k,1}}^{t_{k,1}+\delta t} & -2\int_{t_{k,1}}^{t_{k,1}+\delta t} \bar{x}^T \otimes (\varepsilon^T R)dt \\ \bar{x}^T \otimes \bar{x}^T \big|_{t_{k,2}}^{t_{k,2}+\delta t} & -2\int_{t_{k,2}}^{t_{k,2}+\delta t} \bar{x}^T \otimes (\varepsilon^T R)dt \\ \vdots & \vdots \\ \bar{x}^T \otimes \bar{x}^T \big|_{t_{k,l_k}}^{t_{k,l_k}+\delta t} & -2\int_{t_{k,l_k}}^{t_{k,l_k}+\delta t} \bar{x}^T \otimes (\varepsilon^T R)dt \end{bmatrix},$$

$$\Xi_k = \begin{bmatrix} -\int_{t_{k,1}}^{t_{k,1}+\delta t} (\bar{x}^T K_k^T R K_k \bar{x})dt \\ -\int_{t_{k,2}}^{t_{k,2}+\delta t} (\bar{x}^T K_k^T R K_k \bar{x})dt \\ \vdots \\ -\int_{t_{k,l_k}}^{t_{k,l_k}+\delta t} (\bar{x}^T K_k^T R K_k \bar{x})dt \end{bmatrix}.$$

Assumption 1: For each $k = 0, 1, \ldots$, there exists a sufficiently large integer $l_k > 0$ to ensure the following rank condition hold.

$$rank(\Theta_k) = \frac{n(n+1)}{2} + mn. \tag{28}$$

Remark 5: The rank condition (28) is essentially inspired from the persistent excitation condition in adaptive control [35, 36]. To satisfy the rank condition in (28), the choice of the exploration noise plays an important role.

Lemma 3: Under Assumption 1, there is a unique pair $(P_k, K_{k+1}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ satisfying (27) with $P_k = P_k^T$, i.e.,

$$\begin{bmatrix} vec(P_k) \\ vec(K_{k+1}) \end{bmatrix} = (\Theta_k^T \Theta_k)^{-1} \Theta_k^T \Xi_k. \tag{29}$$

The proof of Lemma 3 is similar to the proof of Lemma 2.3.3 in [37], thus omitted.

Theorem 3: If (28) holds, let $K_0 \in \mathbb{R}^{m\times n}$ be any stabilizing feedback gain matrix with the convergence rate which is faster than e^{-at} (i.e., $A + aI - BK_0$ is Hurwitz or $\max\{\mathrm{Re}[\lambda(A - BK_0)]\} < -a$), the sequences $\{P_k\}_0^\infty$, $\{K_k\}_0^\infty$ obtained from solving (29) converge to P^* and K^*, respectively.

Proof: It can be seen that the pair(P_k, K_{k+1}) obtained from (29) must satisfy (17) and (18). In addition, with Lemma 3, such a pair (P_k, K_{k+1}) obtained from (29) is unique. Therefore, the solution to (29) is the same as the solution to (17) and (18) for all $k = 0, 1, \ldots$. The proof is thus completed by Theorem 2. □

Finally, we obtain the online PI based ADP algorithm (Algorithm 1) for MECGCR.

Algorithm 1 (Online PI Based ADP Algorithm for MECGCR)**:**

1) Initialization: Find K_0 such that $A + aI - BK_0$ is Hurwitz. Let $k = 0$ and $t_{0,1} = 0$.
2) Policy Evaluation and Improvement: Apply $\bar{u} = K_k \bar{x} + \varepsilon$ to the system (1) from $t = t_{k,1}$ and construct each row of the data matrices Θ_k and Ξ_k. If the rank condition (28) is satisfied, solve for P_k and K_{k+1} from (29).
3) Stopping criterion: If $\|P_k - P_{k-1}\| \le \eta$ (η is a prescribed small positive threshold), stop and output P_k; else, set $k = k + 1$and go to 2).

4. SIMULATIONS

Consider the following linear constant system

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 & a_2 & a_3 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} u, \tag{30}$$

with the initial state vector $x_0 = [10, -10, -5]^T$. In the linear system (30), a_1, a_2, a_3, b are the uncertain parameters. Only for simulation purpose, we set $a_1 = 1$, $a_2 = 5$, $a_3 = 7$, $b = 1$. The design objective is to find a linear optimal control law to minimize the following performance index

$$J(x_0, u) = \int_0^\infty e^{2t} u^2 dt. \tag{31}$$

For comparison purposes, the optimal cost matrix P^* and the optimal gain matrix K^*, which can be directly obtained by ARE (12), are given below:

$$P^* = \begin{bmatrix} 364.00 & 288.00 & 28.00 \\ 288.00 & 312.00 & 40.00 \\ 28.00 & 40.00 & 20.00 \end{bmatrix},$$
$$K^* = \begin{bmatrix} 28.00 & 40.00 & 20.00 \end{bmatrix}. \tag{32}$$

Then, we use the developed online model-free PI based ADP algorithm to solve this problem. All the relevant parameters designed in this paper are set as follows: $K_0 = [50,\ 50,\ 20]$; $\delta t = 0.1$; $t_{k+1,1} - t_{k,l_k} = t_{k,i+1} - t_{k,i} = 0.1$; $\eta = 0.5 \times 10^{-2}$; $\varepsilon = \sin(10t)$. The simulation results are given as follows:

1) After five iterations, the approximate optimal matrices P_5 and K_5 are obtained as follows:

$$P_5 = \begin{bmatrix} 364.00 & 288.00 & 28.00 \\ 288.00 & 312.00 & 40.00 \\ 28.00 & 40.00 & 20.00 \end{bmatrix},$$
$$K_5 = \begin{bmatrix} 28.00 & 40.00 & 20.00 \end{bmatrix}. \tag{33}$$

2) The states trajectories are plotted in Fig. 2.
3) The convergence of P_k to its optimal values is illustrated in Fig. 3.
4) The convergence of K_k to its optimal values is illustrated in Fig. 4.

As shown in Fig. 2, $\lim_{t\to\infty} xe^t = 0$. This implies that the closed-loop system is exponentially stable with a given dynamic performance in the learning process. The convergence of the cost function matrix is shown in Fig. 3, and its final estimate can be found in (33). Parallel to the cost function matrix, the convergence of the feedback matrix is shown in Fig. 4, and its final estimate can also be found in (33).

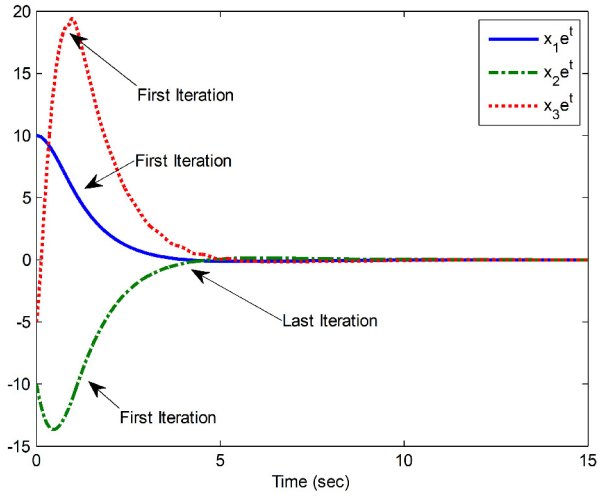


Fig. 2. The system states trajectories.

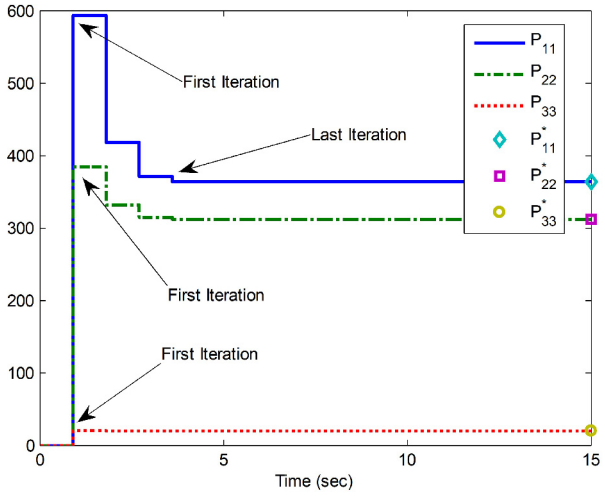


Fig. 3. The convergence of the cost matrix P_k.

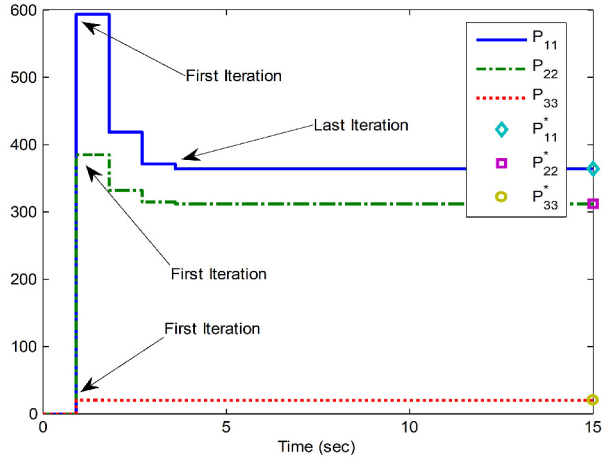


Fig. 4. The convergence of the feedback matrix K_k.

5. CONCLUSION

In this paper, we have proved that ARE can be applied to MECGCR under some conditions without using the standard assumptions. Then, with the help of reinforcing learning, a policy iteration based adaptive dynamic programming algorithm has been developed for MECGCR. In this algorithm, by iteratively solving the algebraic Riccati equation with system state and input information collected online, the numerically approximate solution of MECGCR can be obtained without knowing the system matrices. The methodology developed in this paper may be employed to study the related questions of MECGCR, such as tracing control [38], H_∞ state feedback control [39] and so on.

APPENDIX A

A.1. Proof of Theorem 2

The analytical solution of P in ARE(12) can be first given

$$P=\int_0^\infty e^{(A+aI-BK)^T t}K^T RKe^{(A+aI-BK)t}dt, \tag{A.1}$$

i.e.,

$$P=\int_0^\infty e^{(\bar{A}-BK)^T t}K^T RKe^{(\bar{A}-BK)t}dt.$$

As we all know that R is positive, P is hence non-negative. And P is non-negative and finite if and only if $A+aI-BK$ is Hurwitz. Then, with (A.1), the proof of Theorem 2 will be given.

Proof of Properties 1) and 2) in Theorem 2: Let P_0 be the cost function matrix determined by K_0. From (17), we have

$$\bar{A}_0^T P_0+P_0\bar{A}_0+K_0^T RK_0=0, \tag{A.2}$$

where $\bar{A}_0=\bar{A}-BK_0$. Similarly, we have

$$\bar{A}_1^T P_1+P_1\bar{A}_1+K_1^T RK_1=0, \tag{A.3}$$

where

$$\bar{A}_1=\bar{A}-BK_1, \tag{A.4}$$
$$K_1=R^{-1}B^T P_0. \tag{A.5}$$

Let $\bar{A}_0=\bar{A}_1-B(K_0-K_1)$, (A.2) can be rewritten as

$$\bar{A}_1^T P_0-(K_0-K_1)^T B^T P_0+P_0\bar{A}_1-P_0B(K_0-K_1)+K_0^T RK_0=0. \tag{A.6}$$

By subtracting (A.3) from (A.6), we have

$$\bar{A}_1^T(P_0-P_1)+(P_0-P_1)\bar{A}_1-(K_0-K_1)^T B^T P_0$$

$$-P_0B(K_0-K_1)+K_0^TRK_0-K_1^TRK_1=0. \quad (A.7)$$

By adding the following term to both sides of (A.7)

$$K_0-K_1)^TRK_1+K_1^TR(K_0-K_1), \quad (A.8)$$

and rearranging the terms, we have

$$\begin{aligned}&\bar{A}_1^T(P_0-P_1)+(P_0-P_1)\bar{A}_1+(K_0-K_1)^TR(K_0-K_1)\\&\quad+(K_0-K_1)^T(RKs_1-B^TP_0)\\&\quad+(K_1^TR-P_0B)(K_0-K_1)=0. \quad (A.9)\end{aligned}$$

From (A.5), we know that $RK_1=B^TP_0$. Therefore, (A.9) can be rewritten as the following form:

$$\bar{A}_1^T(P_0-P_1)+(P_0-P_1)\bar{A}_1+(K_0-K_1)^TR(K_0-K_1). \quad (A.10)$$

Due to the same form between (12) and (A.10), along with (A.1), the analytical solution of (P_0-P_1) can be written as

$$P_0-P_1=\int_0^\infty e^{\bar{A}_1^Tt}((K_0-K_1)^TR(K_0-K_1))e^{\bar{A}_1t}dt. \quad (A.11)$$

Therefore, $P_0-P_1\geq 0$ (i.e., $P_1\leq P_0$). Similarly, we have

$$P_1-P^*=\int_0^\infty e^{\bar{A}_1^Tt}((K_1-K^*)^TR(K_1-K^*))e^{\bar{A}_1t}dt\geq 0. \quad (A.12)$$

So that $P^*\leq P_1\leq P_0$ can be obtained. In addition, since both P_0 and P^* are finite, P_1 must be finite. This implies that $\bar{A}_1=A+aI-BK_1$ is Hurwitz. Repeating the above analysis for $k=1,2,\ldots$, Properties (1) and (2) in Theorem 2 can be proved.

Finally, since sequence $\{P_k\}$ is monotonically decreasing and lower bounded by P^*, $\lim_{k\to\infty}P_k=P_\infty$ exists. And by taking the limit of (17) as $k\to\infty$, we have

$$\bar{A}^TP_\infty+P_\infty\bar{A}-P_\infty\bar{B}R^{-1}\bar{B}^TP_\infty=0. \quad (A.13)$$

that satisfies (12). Therefore, P^* is the unique positive definite solution of (17) and $P_\infty=P^*$ can be proved.

REFERENCES

[1] B. Zhou and Z. Lin, "Truncated predictor feedback stabilization of polynomially unstable linear systems with multiple time-varying input delays," *IEEE Transactions on Automatic Control*, vol. 59, no. 8, pp. 2157-2163, August 2014.

[2] C. Xia, N. Liu, Z. Zhou, Y. Yan, and T. Shi, "Steady-state performance improvement for LQR-based PMSM drives," *IEEE Transactions on Power Electronics*, vol. 33, no. 12, pp. 10622-10632, December 2018.

[3] B. Zhou and Z. Li, "Truncated predictor feedback for periodic linear systems with input delays with applications to the elliptical spacecraft rendezvous," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 6, pp. 2238-2250, November 2015.

[4] H. Sun, Y. Liu, F. Li, and X. Niu, "Distributed LQR optimal protocol for leader-following consensus," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3532-3546, September 2019.

[5] F. L. Lewis, D. Vrabie, *Optimal Control*, 3rd Edition, John Wiley & Sons, Inc., 2013.

[6] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114-115, February 1968.

[7] G. Tao, *Adaptive Control Design and Analysis*, Wiley-IEEE Press, 2003.

[8] I. Mareels, J. Polderman, "Adaptive systems," *Systems & Control Foundations & Applications*, vol. 12, no. 1, pp. 1-26, 1996.

[9] Y. Jiang, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Data-driven flotation industrial process operational optimal control based on reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1974-1989, May 2018.

[10] H. Zhang, Y. Liu, G. Xiao, and H. Jiang, "Data-based adaptive dynamic programming for a class of discrete-time systems with multiple delays," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017. DOI: 10.1109/TSMC.2017.2758849

[11] C. Li, D. Liu, and D. Wang, "Data-based optimal control for weakly coupled nonlinear systems using policy iteration," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 4, pp. 511-521, April 2018.

[12] B. Luo, Y. Yang, and D. Liu, "Adaptive Q learning for data-based optimal output regulation with experience replay," *IEEE Transactions on Cybernetics*, vol. 48, no. 12, pp. 3337-3348, December 2018.

[13] S. Zuo, Y. Song, F. L. Lewis, and A. Davoudi, "Optimal robust output containment of unknown heterogeneous multiagent system using off-policy reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3197-3207, November 2018.

[14] 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, February 2009.

[15] 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, November 2014.

[16] H. Wu and B. Luo, "Simultaneous policy update algorithms for learning the solution of linear continuous-time H_∞ state feedback control," *Information Sciences*, vol. 222, no. 10, pp. 472–485, February 2013.

[17] Y. Jiang and Z. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699-2704, October 2012.

[18] H. Zhang, J. Zhang, G. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 1, pp. 152-163, February 2015.

[19] W. Gao, Z. Jiang, F. L. Lewis, and Y. Wang, "Leader-to-formation stability of multiagent systems: an adaptive optimal control approach," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3581-3587, October 2018.

[20] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning," *IEEE Transactions on Cybernetics*, 2019. DOI: 10.1109/TCYB.2018.2890046

[21] S. He, J. Song, Z. Ding, and F. Liu, "Online adaptive optimal control for continuous-time Markov jump linear systems using a novel policy iteration algorithm," *IET Control Theory & Applications*, vol. 9, no. 10, pp. 1536-1543, 2015.

[22] Y. Fu, J. Fu, and T. Chai, "Robust adaptive dynamic programming of two-player zero-sum games for continuous-time linear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3314-3319, December 2015.

[23] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 706-714, July 2014.

[24] T. Kaczorek, "Minimum energy control of positive fractional descriptor continuous-time linear systems," *IET Control Theory & Applications*, vol. 8, no. 4, pp. 219-225, Mar 2014.

[25] T. Kaczorek, "Minimum energy control of fractional positive electrical circuits with bounded inputs," *Circuits Systems & Signal Processing*, vol. 65, no. 2, pp. 191-201, Mar 2014.

[26] J. L. Willems, "Minimum energy and maximum accuracy optimal control of linear stochastic systems," *International Journal of Control*, vol. 22, no.1 pp. 103-112, 1975.

[27] M. Stocks and A. Medvedev, "Guaranteed convergence rate for linear-quadratic optimal time-varying observers," *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 1653-1658, 2006.

[28] C. T. Chen, *Linear System Theory and Design*, Holt, Rinehart, and Winston, 1984.

[29] B. Anderson and J. Moore, "Linear system optimization with prescribed degree of stability," *Proceedings of the Institution of Electrical Engineers*, vol. 116, no.12, pp. 2083-2087, 1969.

[30] K. Zhang and S. Ge, "Adaptive optimal control with guaranteed convergence rate for continuous-time linear systems with completely unknown dynamics," *IEEE Access*, vol. 7, pp. 11526-11532, 2019.

[31] B. Zhou, G. Duan, and Z. Lin, "A parametric Lyapunov equation approach to the design of low gain feedback," *IEEE Transactions on Automatic Control*, vol. 53, no. 6, pp. 1548-1554, July 2008.

[32] B. Zhou and G. Duan, "Periodic Lyapunov equation based approaches to the stabilization of continuous-time periodic linear systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 8, pp. 2139-2146, August 2012.

[33] B. Zhou, G. Duan, and Z. Lin, "Approximation and monotonicity of the maximal invariant ellipsoid for discrete-time systems by bounded controls," *IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 440-446, February 2010.

[34] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, Society for Industrial and Applied Mathematics, 2004.

[35] P. A. Ioannou and J. Sun, *Robust Adaptive Control*, Prentice-Hall, Inc., 1995.

[36] I. Mareels and J. W. Polderman, *Adaptive Systems: An Introduction*, DBLP, 1996.

[37] Y. Jiang and Z. Jiang, *Robust Adaptive Dynamic Programming. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, John Wiley & Sons, Inc., 2013.

[38] W. Gao and Z. Jiang, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2614-2624, June 2018.

[39] J. Fan, Z. Li, Y. Jiang, T. Chai, and F. L. Lewis, "Model-free linear discrete-time system H_∞ control using input-output data," *Proc. of International Conference on Advanced Mechatronic Systems*, pp. 207-212, 2018.

Kai Zhang received his B.S. degree in automation in 2016 from Hefei University of Technology, Anhui, China, where he is currently pursuing an M.S. degree in control theory and control engineering. His research interests include reinforcement learning, adaptive dynamic programming, optimization and game theory.

Suoliang Ge received his M.S. degree in control theory and control engineering from Hefei University of Technology, Anhui, China. His research interests include adaptive dynamic programming, optimization and game theory.

Yuling Ge received her B.S. degree in automation in 2018 from Hefei University, Anhui, China. Now she is pursuing an M.S. degree in control theory and control engineering from Hefei University of Technology, Anhui, China. Her research interests include neural networks, adaptive dynamic programming, deep learning.