Automatica 118 (2020) 109035

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems[☆]

Bo Pang[a,*], Zhong-Ping Jiang[a], Iven Mareels[b]

[a] *Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 370 Jay Street, Brooklyn,, NY 11201, USA*
[b] *Iven Mareels, IBM Research - Australia, Melbourne, Vic 3006, Australia*

ARTICLE INFO

Article history:
Received 5 March 2019
Received in revised form 9 December 2019
Accepted 25 April 2020
Available online xxxx

Keywords:
Optimal control
Reinforcement learning (RL)
Policy iteration (PI)
Adaptive dynamic programming (ADP)

ABSTRACT

This paper studies the infinite-horizon adaptive optimal control of continuous-time linear periodic (CTLP) systems, using reinforcement learning techniques. By means of policy iteration (PI) for CTLP systems, both on-policy and off-policy adaptive dynamic programming (ADP) algorithms are derived, such that the solution of the optimal control problem can be found without the exact knowledge of the system dynamics. Starting with initial stabilizing controllers, the proposed PI-based ADP algorithms converge to the optimal solutions under mild conditions. Application to the adaptive optimal control of the lossy Mathieu equation demonstrates the efficacy of the proposed learning-based adaptive optimal control algorithm.

1. Introduction

Tremendous research efforts have been put into the analysis and control of continuous-time linear periodic (CTLP) systems. There are generally two kinds of motivations. Firstly, many problems in engineering applications can be described and solved in the setting of CTLP systems, such as vibration attenuation in helicopters (Camino & Santos, 2019), controlling robot manipulators (Oh, Bien, & Suh, 1988) and programmatic advertising (Karlsson, 2018). Secondly, CTLP systems play an important role in the study of adaptive control of linear time-varying systems (Xu, 2004; Zhang & Serrani, 2009). It is widely recognized that finding a universal solution to adaptive control of general linear time-varying systems is extremely difficult (Mareels & Polderman, 2012). Thus it is more realistic to classify linear time-varying systems into different categories and study them independently (Narendra & Esfandiari, 2019). CTLP system is one of such categories. Besides adaptive control design, ensuring certain optimality properties for the closed-loop adaptive systems is another major challenge. With this in mind, the optimal control problem of CTLP systems has received considerable attention; see, for instance, Bittanti, Colaneri, and De Nicolao (1991), Shayman (1985), Varga and Stefan (1998) and the references therein. However, even for this special class of CTLP systems, the adaptive control and the optimal control problems have been studied as two separate problems. That is, the adaptive control results presented in Narendra and Esfandiari (2019), Xu (2004) and Zhang and Serrani (2009) do not guarantee optimization of any prescribed cost function, while the optimal control solutions presented in Bittanti et al. (1991), Shayman (1985) and Varga and Stefan (1998) require the precise knowledge of the system dynamics.

In this paper, we aim to invoke reinforcement learning (RL) techniques to address the adaptive optimal control problem for CTLP systems. The objective is to come up with a method which solves the infinite-horizon optimal control problem of CTLP systems without the exact knowledge of the system dynamics. Bellman's dynamic programming (Bellman, 1957) is a powerful method to investigate complex optimal control problems (Li, Yu, Teo, & Duan, 2011; Yang et al., 2016). But the original dynamic programming is haunted by the "curse of dimensionality" (Bellman, 1957). It is also haunted by the "curse of modeling" (Bertsekas & Tsitsiklis, 1996), that is, a mathematical model must be precisely known *a priori*. RL overcomes these two curses by solving Bellman equations through successive approximations using the data generated from interactions between the controller and the plant (Sutton & Barto, 2018). Although the most general array of RL algorithms is provided by researchers within the artificial intelligence community, algorithms with stability and robustness guarantees are not available until recently, mainly by the efforts of researchers within the control systems community (Buşoniu, de Bruin, Tolić, Kober, & Palunko, 2018). Adaptive dynamic

[☆] The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Kok Lay Teo under the direction of Editor Ian R. Petersen.
[*] Corresponding author.
E-mail addresses: bo.pang@nyu.edu (B. Pang), zjiang@nyu.edu (Z.-P. Jiang), i.mareels@unimelb.edu.au (I. Mareels).

https://doi.org/10.1016/j.automatica.2020.109035
0005-1098/© 2020 Elsevier Ltd. All rights reserved.

programming (ADP) (Bertsekas & Tsitsiklis, 1996; Jiang & Jiang, 2017; Lewis & Liu, 2013; Werbos, 2007), as a control-theoretic RL subfield, is devoted to this class of RL algorithms with stability and robustness guarantees. In the past decade, a good number of ADP algorithms have been proposed for systems described by linear or nonlinear, difference or differential equations (Deptula, Rosenfeld, Kamalapurkar, & Dixon, 2018; Jiang & Jiang, 2017; Kamalapurkar, Rosenfeld, & Dixon, 2016; Kamalapurkar, Walters, Rosenfeld, & Dixon, 2018; Lewis & Liu, 2013), with many applications in, e.g., power and energy systems (Wei, Liu, Lewis, Liu, & Zhang, 2017), autonomous systems (Pane, Nageshrao, & Babuška, 2016), to name a few. Nevertheless, most of the existing ADP algorithms assume time-invariant systems, and relatively less results are known for time-varying systems. Adaptive optimal control of linear time-varying systems is studied in Fong, Tan, Crocher, Oetomo, and Mareels (2018) and Pang, Bian, and Jiang (2019), for the continuous-time case and the discrete-time case, respectively. But the optimal control problems considered in Fong et al. (2018) and Pang et al. (2019) are finite-horizon, where no stability issue arises. Thus it is of interest to investigate how to derive ADP algorithms and optimal solutions for time-varying systems in the infinite-horizon optimal control setting.

Inspired by the time-invariant results in Jiang and Jiang (2017), ADP algorithms for CTLP systems are proposed in this paper by using policy iteration (PI). For certain classes of optimal control problems (Kleinman, 1968, Theorem, Saridis & Lee, 1979, Theorem 4), starting with an initial stabilizing controller, PI yields successively stabilizing controllers with improved performance in each iteration, that will converge to the optimal solution as the iteration step goes to infinity. For the infinite-horizon periodic linear quadratic optimal control problem of CTLP systems (defined in next section), Theorem 6.2 in Bittanti et al. (1991) can be viewed as the prototype of the model-based PI. However, the controllers given by this prototypical PI converge pointwise to the controllers that are not necessarily stabilizing and optimal for the corresponding periodic linear quadratic optimal control problem. We firstly add one additional assumption to the prototypical PI, such that it will converge pointwise to the unique stabilizing optimal controller. Then we prove that this pointwise convergence can be further strengthened into uniform convergence. Next, based on the modified model-based PI, we derive two novel PI-based on-policy and off-policy ADP algorithms, to find stabilizing approximate optimal controllers directly from the input/state data, without the exact knowledge of system dynamics. The Fourier basis functions are utilized to approximate different periodic nonlinear functions involved in the algorithms. Rigorous convergence analysis is presented to guarantee the convergence of the proposed algorithms to the optimal solutions, under mild conditions. Finally, the proposed algorithms are applied to the periodic linear quadratic optimal control of the well-known lossy Mathieu equation, which shows the effectiveness and feasibility of our methods.

The rest of this paper is organized as follows: Section 2 introduces the problem formulation and reviews the mathematical preliminaries. Section 3 contains the main results of this paper, i.e., the model-based PI, the PI-based on-policy and off-policy ADP algorithms and their convergence proofs. Section 4 presents the simulation results. Section 5 concludes the whole paper.

Notations: $\mathbb{R}$ is the set of real numbers. $\mathbb{Z}_+$ is the set of nonnegative integers. $\otimes$ is the Kronecker product operator. $\mathbb{S}^n$ denotes the vector space of all n-by-n real symmetric matrices. $|\cdot|_p$ and $\|\cdot\|_p$, $p \in [1,\infty]$ denote the p-norm for vectors and the induced p-norm for matrices, respectively. When subscript p is omitted, $|\cdot|$ and $\|\cdot\|$ represent the Euclidean norm for vectors and the Frobenius norm for matrices, respectively. $[x]_j$ denotes the jth element of vector $x \in \mathbb{R}^n$. $[X]_{i,\cdot}$ ($[X]_{\cdot,j}$) denotes the ith row (jth column) of matrix $X \in \mathbb{R}^{m\times n}$. $[X]_{i,j}$ denotes the element in the ith row and jth column of matrix $X \in \mathbb{R}^{m\times n}$. For real symmetric matrices A and B, $A > B$ ($A \geq B$) means that matrix $A - B$ is positive definite (positive semidefinite).

2. Problem formulation and preliminaries

Consider the following class of continuous-time linear periodic systems

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the system state, $u(t) \in \mathbb{R}^m$ is the control input, $A(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times n}$, $B(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times m}$ are continuous and T-periodic matrix-valued functions, i.e.,

$$A(t+T) = A(t), \qquad B(t+T) = B(t), \qquad \forall t \in \mathbb{R}.$$

$B(\cdot)$ is piecewise continuously differentiable. Let $\Phi(t,\tau)$ be the state transition matrix of the unforced system (1), i.e., $u \equiv 0$. Then $\Phi(t,\tau)$ satisfies

$$\dot{\Phi}(t,\tau) = A(t)\Phi(t,\tau), \quad \Phi(\tau,\tau) = I,$$

and $\Phi(t+T,\tau+T) = \Phi(t,\tau)$. In the setting of CTLP systems, $\Phi(t+T,t)$ is known as the monodromy matrix at time t, whose eigenvalues (known as characteristic multipliers) are independent of t. By Floquet theory (DaCunha & Davis, 2011), we obtain the following lemma about the stability of CTLP systems.

Lemma 1. *For the unforced system (1), the following are equivalent:*

(i) It is globally uniformly asymptotically stable.
(ii) It is globally uniformly exponentially stable.
(iii) The characteristic multipliers associated with $A(\cdot)$ belong to the open unit disk.

By Bittanti et al. (1991, Section 6.5.1.1), the periodic linear quadratic optimal control problem consists of finding a linear stabilizing control policy $u(\cdot)$ that minimizes the quadratic cost functional

$$J(t_0,\xi,u(\cdot)) = \int_{t_0}^{\infty} |C(t)x(t)|^2 + u^T(t)R(t)u(t)dt, \tag{2}$$

where $u(t) = -K(t)x(t)$, $K(\cdot) : \mathbb{R} \to \mathbb{R}^{m\times n}$, $C(\cdot) : \mathbb{R} \to \mathbb{R}^{r\times n}$ are continuous and T-periodic; $R(\cdot) : \mathbb{R} \to \mathbb{R}^{m\times m}$ is continuous, T-periodic, positive definite and piecewise continuously differentiable; $x(t)$ is the solution of Eq. (1) with initial state $x(t_0) = \xi$, $\xi \in \mathbb{R}^n$.

Remark 2. For convenience, in this paper, we just use "stable" or "stabilizing" to refer to the type of stability in Lemma 1. For example, we say that a control gain $K(\cdot)$ is stabilizing, if system $\dot{x}(t) = (A(t) - B(t)K(t))x(t)$ is stable in the sense of Lemma 1.

Associated with the optimal control problem is the well-known periodic Riccati equation (PRE)

$$\begin{aligned}-\dot{P}(t) &= A^T(t)P(t) + P(t)A(t)\\ &\quad - P(t)B(t)R^{-1}(t)B^T(t)P(t) + C^T(t)C(t).\end{aligned}$$

Under certain conditions, the optimal solution to the periodic linear quadratic control problem exists and is unique (Bittanti et al., 1991, Theorem 6.5 and Theorem 6.12).

Lemma 3. *There exists a unique symmetric, periodic and positive semidefinite (SPPS) solution $P^*(\cdot)$ of the PRE, and the corresponding closed-loop system is stable, if and only if $(A(\cdot),B(\cdot))$ is stabilizable and $(A(\cdot),C(\cdot))$ is detectable (Bittanti, 1986, Theorem 4). The cost function (2) is minimized by the optimal control gain $K^*(t) = R^{-1}(t)B^T(t)P^*(t)$, and the corresponding minimum cost is $J^*(t_0,\xi) = J(t_0,\xi,u^*(\cdot)) = \xi^T P^*(t_0)\xi$.*

Note that $P^*(t)$ is a nonlinear matrix-valued function of time t, whose analytic expression is generally difficult to be obtained. In this paper, Fourier basis functions are adopted to approximate different periodic functions. Suppose $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is a periodic function with a period T. Then, define the partial sums of Fourier series of $f(\cdot)$ as

$$f_N(t) = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos(\omega n t) + b_n \sin(\omega n t)\right),$$

where $\omega = 2\pi/T$, a_n and b_n are the Fourier coefficients.

Lemma 4 (Anton, 2005, Theorem 1.5.1)**.** *Iff is T-periodic, continuous and piecewise continuously differentiable, then $f_N \to f$ uniformly, as $N \to \infty$.*

When matrices $A(\cdot)$ and $B(\cdot)$ are known, the optimal solution $P^*(\cdot)$ can be approximately solved using existing numerical methods (see, e.g. Varga (2008)). When matrices $A(\cdot)$ and $B(\cdot)$ are unknown, those numerical methods can hardly be applied directly due to the nonlinearity of the PRE. By reinforcement learning techniques, in the sequel, two PI-based ADP algorithms are proposed to find approximate optimal controllers directly from the collected data.

Definition 5. For matrices $X \in \mathbb{R}^{n\times m}$, $Y \in \mathbb{S}^m$, and vector $v \in \mathbb{R}^n$, define

$$\begin{aligned}\mathrm{vec}(X) &= [x_1^T, x_2^T, \ldots, x_m^T]^T,\\ \mathrm{vecs}(Y) &= [y_{11}, \sqrt{2}y_{12}, \ldots, \sqrt{2}y_{1m}, y_{22}, \sqrt{2}y_{23},\\ &\quad \ldots, \sqrt{2}y_{m-1,m}, y_{m,m}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)},\\ \tilde{v} &= \mathrm{vecs}(vv^T),\end{aligned}$$

where x_i is the ith column of X. In addition, $\mathrm{vec}^{-1}(\cdot)$ and $\mathrm{vecs}^{-1}(\cdot)$ denote the operators such that $X = \mathrm{vec}^{-1}(\mathrm{vec}(X))$ and $Y = \mathrm{vecs}^{-1}(\mathrm{vecs}(Y))$, respectively.

As it can be directly checked, we have

Lemma 6. *For $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{S}^n$, $|\mathrm{vec}(X)| = \|X\|$, $|\mathrm{vecs}(Y)| = \|Y\|$.*

3. Policy iteration based adaptive dynamic programming for continuous-time linear periodic systems

In this section, the model-based PI algorithm is firstly presented. Then the corresponding on-policy and off-policy PI-based ADP algorithms are derived in two subsequent subsections, respectively.

3.1. Model-based PI for CTLP systems with known dynamics

Before proceeding, the following lemma is useful.

Lemma 7. *Suppose $L(\cdot)$ is a continuous, T-periodic and stabilizing control gain. The cost induced by control policy $u_L(t) = -L(t)x(t)$ is $J(t_0, \xi, u_L(\cdot)) = \xi^T P_L(t_0)\xi$, where $P_L(\cdot)$ is the unique SPPS solution of the periodic Lyapunov equation (PLE)*

$$\begin{aligned}-\dot{P}_L(t) &= A_L^T(t)P_L(t) + P_L(t)A_L(t)\\ &\quad + C^T(t)C(t) + L^T(t)R(t)L(t),\end{aligned} \tag{3}$$

and

$$\begin{aligned}P_L(t) &= \int_t^\infty \left[\Phi_L^T(\tau,t)\left(C(\tau)^T C(\tau)\right.\right.\\ &\quad \left.\left.+L^T(\tau)R(\tau)L(\tau)\right)\Phi_L(\tau,t)\right]\mathrm{d}\tau,\end{aligned} \tag{4}$$

where $\Phi_L(\tau, t)$ is the state transition matrix corresponding to $A_L(t) = A(t) - B(t)L(t)$.

Proof. By Lemma 2 in Bittanti, Bolzern, and Colaneri (1984), $P_L(\cdot)$ in (4) is the unique T-periodic solution of PLE (3). Since $R(\cdot) > 0$, $P_L(\cdot)$ is the unique SPPS solution.

The model-based policy iteration for CTLP systems is presented in the following theorem.

Theorem 8. *Suppose that $(A(\cdot), B(\cdot))$ is stabilizable and $(A(\cdot), C(\cdot))$ is detectable. Let $K_0(\cdot)$ be a continuous, T-periodic stabilizing control gain. Set $i = 0$, and consider the following stepwise procedure:*

(1) (Policy Evaluation) Solve the unique SPPS solution $P_i(\cdot)$ from the PLE

$$\begin{aligned}-\dot{P}_i(t) &= A_i^T(t)P_i(t) + P_i(t)A_i(t)\\ &\quad + C^T(t)C(t) + K_i^T(t)R(t)K_i(t),\end{aligned} \tag{5}$$

where $A_i(t) = A(t) - B(t)K_i(t)$.

(2) (Policy Improvement) Obtain improved control policy using

$$K_{i+1}(t) = R^{-1}(t)B^T(t)P_i(t). \tag{6}$$

(3) Let $i = i + 1$, and return to Step (1).

Then for all $i \in \mathbb{Z}_+$:

(i) $A_i(\cdot)$ is stable.
(ii) $0 \le P^*(t) \le P_{i+1}(t) \le P_i(t)$, $\forall t \in \mathbb{R}$.
(iii) $P_i(\cdot)$ and $K_i(\cdot)$ converge pointwise to $P^*(\cdot)$ and $K^*(\cdot)$, respectively.

Proof. In Bittanti et al. (1991, Theorem 6.2), if $(A(\cdot), B(\cdot))$ is stabilizable, it is shown that $P_i(\cdot)$ converges pointwise and monotonically to the maximal solution (Bittanti et al., 1991, Section 6.3.1.1) of the PRE. By Lemma 3, if further $(A(\cdot), C(\cdot))$ is detectable, $P^*(\cdot)$ is equal to the maximal solution of PRE. Thus Theorem 8 follows readily.

Next, we show that the pointwise convergence presented in Theorem 8 can be actually strengthened into uniform convergence. In the rest of this paper, we omit the dependence of variables on time t when there is no ambiguity.

Corollary 9. *Under the conditions of Theorem 8, $\lim_{i\to\infty} P_i(t) = P^*(t)$ uniformly, $\lim_{i\to\infty} K_i(t) = K^*(t)$ uniformly, on $\mathbb{R}$.*

Proof. Firstly, we show that $[P_i(t)]_{j,k}$ converges to $[P^*(t)]_{j,k}$ uniformly on $\mathbb{R}$, for $j = 1, \ldots, n$, $k = 1, \ldots, n$. For $i > 0$, substituting (6) into (5) yields

$$\begin{aligned}\|\dot{P}_i\| \le \bar{U}_i &:= \eta_1\|P_i\| + \eta_2\\ &\quad + \eta_3\left(\|P_{i-1}\|^2 + 2\|P_{i-1}\|\|P_i\|\right),\end{aligned} \tag{7}$$

where η_1, η_2, η_3 are positive constants. From Theorem 8, $P_{i-1} \ge P_i \ge P_{i+1}$, the monotonicity of Frobenius norm (Ciarlet, Miara, & Thomas, 1989, 2.2-10) implies $\|P_{i-1}\| \ge \|P_i\| \ge \|P_{i+1}\|$. In view of (7), this implies that $\{\bar{U}_i\}_{i=1}^{\infty}$ is nonincreasing. Thus we have

$$\|\dot{P}_i\| \le \bar{U}^* := \max\left\{\max_t \bar{U}_1(t), \max_t \|\dot{P}_0(t)\|\right\},$$

for all $i \in \mathbb{Z}_+$. Consider function $[P_i(\cdot)]_{j,k}$ on compact set $[0, T]$. For $0 \le t_1 < t_2 \le T$, the mean value theorem yields $\left|[P_i(t_1)]_{j,k} - [P_i(t_2)]_{j,k}\right| \le \bar{U}^*|t_1 - t_2|$. Since $\bar{U}^*$ is independent of iteration index i, it follows from the above inequality that the sequence of functions $\{[P_i(\cdot)]_{j,k}\}_{i=0}^{\infty}$ is equicontinuous (Rudin, 1976, Definition 7.22) on $[0, T]$. Furthermore, through Theorem 8, we know that $\{[P_i(\cdot)]_{j,k}\}_{i=0}^{\infty}$ converges pointwise to $[P^*(\cdot)]_{j,k}$ on $[0, T]$. Then by a corollary (Rudin, 1976, Exercise 16 on Page 168) of the Arzelà–Ascoli theorem (Rudin, 1976, Theorem 7.25), $[P_i(\cdot)]_{j,k}$ converges uniformly to $[P^*(\cdot)]_{j,k}$ on $[0, T]$. As a result of the periodicity, $[P_i(\cdot)]_{j,k}$ converges uniformly on $\mathbb{R}$. This implies that $P_i(\cdot)$ itself converges uniformly to $P^*(\cdot)$ on $\mathbb{R}$. Due to the boundedness of $R(\cdot)$ and $B(\cdot)$, by (6) the uniform convergence of $K_i(\cdot)$ to $K^*(\cdot)$ on $\mathbb{R}$ follows. This completes the proof of Corollary 9.

In the following subsections, using Corollary 9, we propose two algorithms to solve iteratively (5) and (6) in the absence of the exact knowledge of matrices $(A(\cdot), B(\cdot))$.

3.2. PI-based on-policy ADP algorithm for the unknown dynamics case

Assume the following policy is applied to system (1) to collect data

$$v_L(t) = -L(t)x(t) + u_e(t), \tag{8}$$

where $L(t)$ is known and defined in Lemma 7, $u_e(t)$ is the exploration noise. Then the evolution of the closed-loop system states is

$$\dot{x}(t) = A_L(t)x(t) + B(t)u_e(t). \tag{9}$$

Since $L(\cdot)$ is stabilizing, we can construct an improved control gain based on $P_L(\cdot)$

$$\bar{L}(t) = R^{-1}(t)B^T(t)P_L(t). \tag{10}$$

Then by (3), (9) and (10), differentiating x^TP_Lx with respect to time t yields

$$\frac{\mathrm{d}x^TP_Lx}{\mathrm{d}t} = -x^T(C^TC + L^TRL)x + 2u_e^TR\bar{L}x. \tag{11}$$

Define $t_j = t_0 + j\Delta t$, where $j \in \mathbb{Z}_+$, $\Delta t > 0$ is the sampling interval. By integrating both sides of (11) from t_j to t_{j+1} and rearranging the terms, we have

$$\begin{aligned} &-\int_{t_j}^{t_{j+1}} [x^T(C^TC + L^TRL)x]dt = \\ &\tilde{x}^T(t_{j+1})\mathrm{vecs}(P_L(t_{j+1})) - \tilde{x}^T(t_j)\mathrm{vecs}(P_L(t_j)) \\ &-\int_{t_j}^{t_{j+1}} (x^T \otimes 2u_e^TR)\mathrm{vec}(\bar{L})dt. \end{aligned} \tag{12}$$

Note that $P_L(\cdot)$ and $\bar{L}(\cdot)$ are periodic matrix-valued functions. Thus, we can express $P_L(\cdot)$ and $\bar{L}(\cdot)$ using the linear combination of $(2N+1)$ Fourier basis functions as follows

$$\begin{aligned} \mathrm{vecs}(P_L(t)) &= \hat{X}^{(1)}_{L,N}F_N(t) + \hat{e}^{(1)}_{L,N}(t), \\ \mathrm{vec}(\bar{L}(t)) &= \hat{X}^{(2)}_{L,N}F_N(t) + \hat{e}^{(2)}_{L,N}(t), \end{aligned} \tag{13}$$

where $\hat{X}^{(1)}_{L,N} \in \mathbb{R}^{n_1\times(2N+1)}$ and $\hat{X}^{(2)}_{L,N} \in \mathbb{R}^{n_2\times(2N+1)}$ are weight matrices, $n_1 = \frac{n(n+1)}{2}$, $n_2 = mn$; $\hat{e}^{(1)}_{L,N}(t) \in \mathbb{R}^{n_1}$ and $\hat{e}^{(2)}_{L,N}(t) \in \mathbb{R}^{n_2}$ are approximation errors; and

$$\begin{aligned} F_N(t) = [&1, \cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t), \\ &\cdots, \cos(N\omega t), \sin(N\omega t)]^T. \end{aligned}$$

By inserting (13) into (12) and rearranging the terms, we obtain

$$d_{L,j,N}\begin{bmatrix}\mathrm{vec}(\hat{X}^{(1)}_{L,N}) \\ \mathrm{vec}(\hat{X}^{(2)}_{L,N})\end{bmatrix} = -r_{L,j} + \hat{e}_{L,j,N}, \tag{14}$$

where error term $\hat{e}_{L,j,N}$ summarizes the effect of errors $\hat{e}^{(1)}_{L,N}(\cdot)$ and $\hat{e}^{(2)}_{L,N}(\cdot)$, and

$$\begin{aligned} d_{L,j,N} &= \left[F_{x,N}(t_{j+1}) - F_{x,N}(t_j), -F_{xu,j,N}\right], \\ F_{x,N}(t) &= F_N^T(t) \otimes \tilde{x}^T(t), \\ F_{xu,j,N} &= \int_{t_j}^{t_{j+1}} F_N^T \otimes x^T \otimes 2u_e^TRdt \\ r_{L,j} &= \int_{t_j}^{t_{j+1}} [x^T(C^TC + L^TRL)x]dt. \end{aligned}$$

The subscript L of $d_{L,j,N}$ is used to emphasize that the state trajectory $x(t)$ involved in (14) is generated by control policy $v_L(\cdot)$ in (8). Letting $j = 0, 1, 2, \ldots, M-1$ in (14), where $M \in \mathbb{Z}_+\backslash\{0\}$, we can reorganize the resulting equations into a single linear matrix equation

$$\Theta_{L,N}\begin{bmatrix}\mathrm{vec}(\hat{X}^{(1)}_{L,N}) \\ \mathrm{vec}(\hat{X}^{(2)}_{L,N})\end{bmatrix} = \Psi_{L,N} + \hat{E}_{L,N} \tag{15}$$

where $[\Theta_{L,N}]_{j,\cdot} = d_{L,j,N}$, $[\Psi_{L,N}]_j = -r_{L,j}$, $[\hat{E}_{L,N}]_j = \hat{e}_{L,j,N}$. In (15), $\Theta_{L,N}$ and $\Psi_{L,N}$ are known data matrices. Thus it is possible to apply the least square regression to determine the weight matrices $\hat{X}^{(1)}_{L,N}$ and $\hat{X}^{(2)}_{L,N}$.

Now, we are ready to derive the PI-based on-policy ADP algorithm to solve (5) and (6) directly from the collected input/state data. Define the control gains

$$\hat{K}_{i,N}(t) = \begin{cases} K_0(t), & i = 0, \\ \mathrm{vec}^{-1}\left(\hat{X}^{(2)}_{i-1,N}F_N(t)\right), & i = 1, 2, \ldots, \end{cases}$$

and the control law

$$\hat{v}_{i,N}(t) = -\hat{K}_{i,N}(t)x(t) + u_e(t), \tag{16}$$

where $\hat{X}^{(2)}_{i-1,N}$ is the weight matrix used to approximate the improved control gain at the $(i-1)$th iteration. Let $\check{P}_{i,N}(\cdot)$ be the unique SPPS solution of PLE

$$\begin{aligned} -\dot{\check{P}}_{i,N}(t) = &\hat{A}^T_{i,N}(t)\check{P}_{i,N}(t) + \check{P}_{i,N}(t)\hat{A}_{i,N}(t) \\ &+ C^T(t)C(t) + \hat{K}^T_{i,N}(t)R(t)\hat{K}_{i,N}(t) \end{aligned} \tag{17}$$

with $\hat{A}^T_{i,N} = A(t) - B(t)\hat{K}_{i,N}(t)$. By Lemma 7, if $\hat{K}_{i,N}(\cdot)$ is stabilizing, such a $\check{P}_{i,N}(\cdot)$ exists. Then the policy improvement step is

$$\check{K}_{i+1,N}(t) = R^{-1}(t)B^T(t)\check{P}_{i,N}(t). \tag{18}$$

By using (17) and (18), and replacing $L(\cdot)$ and $v_L(\cdot)$ with $\hat{K}_{i,N}(\cdot)$ and $\hat{v}_{i,N}(\cdot)$ in the derivations from (11) to (15), we obtain

$$\hat{\Theta}_{i,N}\begin{bmatrix}\mathrm{vec}(\hat{X}^{(1)}_{i,N}) \\ \mathrm{vec}(\hat{X}^{(2)}_{i,N})\end{bmatrix} = \hat{\Psi}_{i,N} + \hat{E}_{i,N}, \tag{19}$$

where $\hat{X}^{(1)}_{i,N} \in \mathbb{R}^{n_1\times(2N+1)}$, $\hat{X}^{(2)}_{i,N} \in \mathbb{R}^{n_2\times(2N+1)}$ are unknown weight matrices to be determined. To ensure that the least square regression problem represented by (19) is feasible, we make an assumption in spirit of persistent excitation (PE) condition in adaptive control (Jiang & Jiang, 2017; Mareels & Polderman, 2012).

Assumption 10. For all $i \in \mathbb{Z}_+$, there exist $\bar{M} \geq (n_1+n_2)(2N+1)$ and $\alpha > 0$ (independent of N), such that for all $M > \bar{M}$, $M \in \mathbb{Z}_+$, we have

$$\frac{1}{M}\hat{\Theta}^T_{i,N}\hat{\Theta}_{i,N} \geq \alpha I_{(n_1+n_2)(2N+1)}.$$

Remark 11. Analogous assumptions appeared in the past literature of ADP (Bian, Jiang, & Jiang, 2014; Jiang & Jiang, 2017; Lewis & Liu, 2013). The exploration noise $u_e(t)$ can be chosen as, e.g., sinusoidal signals or random noise, to satisfy this kind of assumptions. As long as $L(t)$ is stabilizing, by Lemma 1 and Khalil (2002, Lemma 4.6), system (9) is input-to-state stable, which means that the states of system (9) are bounded for any bounded exploration noise $u_e(t)$.

Under Assumption 10, the weight matrices which achieve the minimum approximation error are given by

$$\begin{bmatrix} \mathrm{vec}(\hat{X}^{(1)}_{i,N}) \\ \mathrm{vec}(\hat{X}^{(2)}_{i,N}) \end{bmatrix} = (\hat{\Theta}^T_{i,N}\hat{\Theta}_{i,N})^{-1}\hat{\Theta}^T_{i,N}\hat{\Psi}_{i,N}. \tag{20}$$

The PI-based on-policy ADP algorithm is summarized in Algorithm 1.

Algorithm 1 PI-based on-policy ADP

```
Choose a stabilizing initial control gain $K_0(\cdot)$, threshold $\epsilon > 0$, $N \in \mathbb{Z}_+$, $M \in \mathbb{Z}_+\backslash\{0\}$ and $\Delta t > 0$.
Set $\hat{K}_{0,N}(t) = K_0(t)$, and let $i \leftarrow 0$.
repeat
    Apply (16) to the system (1) and construct the data matrices $\hat{\Theta}_{i,N}$, $\hat{\Psi}_{i,N}$ in (19).
    Compute $\hat{X}^{(1)}_{i,N}$, $\hat{X}^{(2)}_{i,N}$ by (20).
    $\hat{P}_{i,N}(t) \leftarrow \mathrm{vecs}^{-1}\left(\hat{X}^{(1)}_{i,N}F_N(t)\right)$
    $\hat{K}_{i+1,N}(t) \leftarrow \mathrm{vec}^{-1}\left(\hat{X}^{(2)}_{i,N}F_N(t)\right)$
    if $i > 0$ then
        $\gamma \leftarrow \|\hat{X}^{(1)}_{i,N} - \hat{X}^{(1)}_{i-1,N}\| + \|\hat{X}^{(2)}_{i,N} - \hat{X}^{(2)}_{i-1,N}\|$
    else
        $\gamma \leftarrow 2\epsilon$
    end if
    $i \leftarrow i + 1$
until $\gamma < \epsilon$
Use $\hat{u}_i(t) = -\hat{K}_{i,N}(t)x(t)$ as the approximate optimal control.
```

Lemma 12. *For each $i \in \mathbb{Z}_+$, if $\lim_{N\to\infty}\hat{K}_{i,N}(t) = K_i(t)$ uniformly on $\mathbb{R}$, then*

(i) When N is large enough, $\hat{K}_{i,N}(\cdot)$ is stabilizing.
(ii) $\lim_{N\to\infty}\check{P}_{i,N}(t) = P_i(t)$ uniformly on $\mathbb{R}$.

Proof. See Appendix A.

Lemma 13. *For each $i \in \mathbb{Z}_+$, if $\lim_{N\to\infty}\hat{K}_{i,N}(t) = K_i(t)$ uniformly on $\mathbb{R}$, and Assumption 10 is satisfied, then $\forall\epsilon > 0$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}, N \in \mathbb{Z}_+$,*

$$\|\hat{P}_{i,N}(t) - \check{P}_{i,N}(t)\| < \epsilon, \|\hat{K}_{i+1,N}(t) - \check{K}_{i+1,N}(t)\| < \epsilon,$$

for all $t \in \mathbb{R}$.

Proof. See Appendix B.

The convergence analysis of Algorithm 1 is given in the following theorem and corollary.

Theorem 14. *Under Assumption 10, given $\bar{i} \in \mathbb{Z}_+$, for any $\epsilon > 0$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}$, $N \in \mathbb{Z}_+$,*

$$\|\hat{P}_{i,N}(t) - P_i(t)\| < \epsilon, \quad \|\hat{K}_{i+1,N}(t) - K_{i+1}(t)\| < \epsilon,$$

for all $t \in \mathbb{R}$, $i = 1, 2, \dots, \bar{i}$. In addition, $\hat{A}_{i,N}(\cdot)$ in (17) is uniformly asymptotically stable, for $i = 1, 2, \dots, \bar{i} + 1$.

Proof. Since $\hat{K}_{0,N}(t) = K_0(t)$, there are $\check{P}_{0,N}(t) = P_0(t)$, $\check{K}_{1,N}(t) = K_1(t)$. By Lemma 13, $\exists\bar{N}_0 > 0$, such that $\forall N > \bar{N}_0$, $\forall t \in \mathbb{R}$, $\|\hat{P}_{0,N}(t) - P_0(t)\| < \epsilon$, $\|\hat{K}_{1,N}(t) - K_1(t)\| < \epsilon$. Then by Lemmas 12, 13 and (18), $\exists\bar{N}_1 \geq \bar{N}_0$, such that $\forall N > \bar{N}_1$, $\hat{A}_{1,N}(t)$ is uniformly asymptotically stable, and $\forall t \in \mathbb{R}$, $\|\hat{P}_{1,N}(t) - P_1(t)\| < \epsilon$, $\|\hat{K}_{2,N}(t) - K_2(t)\| < \epsilon$. Through similar derivations, $\exists\bar{N}_{\bar{i}} > 0$, such that $\forall N > \bar{N}_{\bar{i}}$, $\hat{A}_{i,N}(\cdot)$ is uniformly asymptotically stable, and $\forall t \in \mathbb{R}$, $\|\hat{P}_{i,N}(t) - P_i(t)\| < \epsilon$, $\|\hat{K}_{i+1,N}(t) - K_{i+1}(t)\| < \epsilon$, for $i = 1, 2, \dots, \bar{i}$. Finally, one could choose $\bar{N}_{\bar{i}+1} \geq \bar{N}_{\bar{i}}$, such that $\forall N > \bar{N}_{\bar{i}+1}$, $\hat{A}_{\bar{i}+1,N}(\cdot)$ is uniformly asymptotically stable. The proof is completed by setting $\bar{N} = \bar{N}_{\bar{i}+1}$.

A direct combination of Corollary 9, Theorem 14 and the triangle inequality yields the next corollary.

Corollary 15. *Under Assumption 10 and the conditions of Theorem 8, $\forall\epsilon > 0$, $\exists\bar{i} \in \mathbb{Z}_+$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}$, $N \in \mathbb{Z}_+$, $\forall t \in \mathbb{R}$,*

$$\|\hat{P}_{\bar{i},N}(t) - P^*(t)\| < \epsilon, \quad \|\hat{K}_{\bar{i}+1,N}(t) - K^*(t)\| < \epsilon.$$

3.3. PI-based off-policy ADP algorithm for the unknown dynamics case

In order to find the near-optimal policy, in Algorithm 1 one needs to collect new data in every iteration of the inner loop, which may be costly and inconvenient. In engineering applications, one is often more interested in finding an approximate optimal solution by utilizing less data. To this end, PI-based off-policy ADP algorithm is proposed in this section.

Define the control gains

$$\mathring{K}_{i,N}(t) = \mathrm{vec}^{-1}\left(\mathring{X}^{(2)}_{i-1,N}F_N(t)\right), \quad i \in \mathbb{Z}_+, \tag{21}$$

where $\mathring{X}^{(2)}_{-1,N}$ is chosen so that $\mathring{K}_{0,N}(t)$ is stabilizing, $\mathring{X}^{(2)}_{i,N}$, $i \in \mathbb{Z}_+$ is the weight matrix used to approximate the improved control gain in ith iteration. Take an arbitrary control policy u_0 which, when applied to (1), yields the boundedness of the solutions of the closed-loop system, i.e.,

$$\dot{x}(t) = \mathring{A}_{i,N}(t)x(t) + B(t)(\mathring{K}_{i,N}(t)x(t) + u_0(t)), \tag{22}$$

where $\mathring{A}^T_{i,N} = A(t) - B(t)\mathring{K}_{i,N}(t)$. Let $\check{P}_{i,N}(t)$ denote the unique SPPS solution of PLE,

$$\begin{aligned} -\dot{\check{P}}_{i,N}(t) = {} & \mathring{A}^T_{i,N}(t)\check{P}_{i,N}(t) + \check{P}_{i,N}(t)\mathring{A}^T_{i,N} \\ & + C^T(t)C(t) + \mathring{K}^T_{i,N}(t)R(t)\mathring{K}_{i,N}(t). \end{aligned} \tag{23}$$

If $\mathring{K}_{i,N}(t)$ is stabilizing, such a $\check{P}_{i,N}(t)$ always exists by Lemma 7. Then an improved control gain can be obtained

$$\check{K}_{i+1,N}(t) = R^{-1}(t)B^T(t)\check{P}_{i,N}(t). \tag{24}$$

Similar to (11), by (22), (23) and (24), differentiating $x^T\check{P}_{i,N}x$ with respect to t yields

$$\begin{aligned} \frac{\mathrm{d}x^T\check{P}_{i,N}x}{\mathrm{d}t} = {} & x^T(-C^TC - \mathring{K}_{i,N}R\check{K}_{i,N})x \\ & + 2(u_0 + \mathring{K}_{i,N}x)^TR\check{K}_{i+1,N}x. \end{aligned} \tag{25}$$

By integrating both sides of (25) from t_j to t_{j+1} and rearranging the terms, we have

$$\begin{aligned} & \tilde{x}^T(t_{j+1})\mathrm{vecs}(\check{P}_{i,N}(t_{j+1})) - \tilde{x}^T(t_j)\mathrm{vecs}(\check{P}_{i,N}(t_j)) \\ & - \int_{t_j}^{t_{j+1}}\left(x^T \otimes (2(u_0 + \mathring{K}_{i,N}x)^TR)\right)\mathrm{vec}(\check{K}_{i+1,N})\mathrm{d}t \\ & = -\int_{t_j}^{t_{j+1}} x^TC^TCx\mathrm{d}t - \int_{t_j}^{t_{j+1}}\mathrm{vec}(\mathring{K}_{i,N})^T \\ & (x^T \otimes I_m)^TR(x^T \otimes I_m)\mathrm{vec}(\mathring{K}_{i,N})\mathrm{d}t. \end{aligned} \tag{26}$$

Substituting (21) and the following approximations into (26)

$$\begin{aligned} \mathrm{vecs}(\check{P}_{i,N}(t)) &= \mathring{X}^{(1)}_{i,N}F_N(t) + \mathring{e}^{(1)}_{i,N}(t), \\ \mathrm{vec}(\check{K}_{i+1,N}(t)) &= \mathring{X}^{(2)}_{i,N}F_N(t) + \mathring{e}^{(2)}_{i,N}(t), \end{aligned} \tag{27}$$

we obtain

$$\mathring{d}_{i,j,N}\begin{bmatrix}\mathrm{vec}(\mathring{X}^{(1)}_{i,N})\\ \mathrm{vec}(\mathring{X}^{(2)}_{i,N})\end{bmatrix} = -c_{i,j,N} + \mathring{e}_{i,j,N}, \tag{28}$$

where

$$\begin{aligned}\mathring{d}_{i,j,N} &= \Big[F_{x,N}(t_{j+1}) - F_{x,N}(t_j),\\ &\quad -F_{xu_0,j,N} - \mathrm{vec}^T(\mathring{X}^{(2)}_{i-1,N})\Delta_{1,j,N}\Big],\\ \Delta_{1,j,N} &= \int_{t_j}^{t_{j+1}} F_N^T \otimes x^T \otimes 2F_N \otimes x \otimes R\mathrm{d}t,\\ F_{xu_0,j,N} &= \int_{t_j}^{t_{j+1}} F_N^T \otimes x^T \otimes (2u_0^T R)\mathrm{d}t,\\ c_{i,j,N} &= \int_{t_j}^{t_{j+1}} x^T C^T Cx\mathrm{d}t\\ &\quad - \mathrm{vec}^T(\mathring{X}^{(2)}_{i-1,N})\Delta_{2,j,N}\mathrm{vec}(\mathring{X}^{(2)}_{i-1,N}),\\ \Delta_{2,j,N} &= \int_{t_j}^{t_{j+1}} (F_N \otimes x \otimes I_m)\, R\left(F_N^T \otimes x^T \otimes I_m\right)\mathrm{d}t,\end{aligned}$$

and $\mathring{e}_{i,j,N}$ is the approximation error as that in (14). Analogous to last subsection, we can reorganize Eqs. (28) with $j = 1, 2, \ldots, M-1$ into a single linear matrix equation,

$$\mathring{\Theta}_{i,N}\begin{bmatrix}\mathrm{vec}(\mathring{X}^{(1)}_{i,N})\\ \mathrm{vec}(\mathring{X}^{(2)}_{i,N})\end{bmatrix} = \mathring{\Psi}_{i,N} + \mathring{E}_{i,N}, \tag{29}$$

where $[\mathring{\Theta}_{i,N}]_{j,\cdot} = \mathring{d}_{i,j,N}$, $[\mathring{\Psi}_{i,N}]_j = -c_{i,j,N}$, $[\mathring{E}_{i,N}]_j = \mathring{e}_{i,j,N}$. Again, we need to impose an assumption on the data matrix in the spirit of PE condition.

Assumption 16. Assumption 10 holds with $\hat{\Theta}_{i,N}$ replaced by $\mathring{\Theta}_{i,N}$.

Under Assumption 16, the weighting matrices that achieve the least squares approximation error can be computed by

$$\begin{bmatrix}\mathrm{vec}(\mathring{X}^{(1)}_{i,N})\\ \mathrm{vec}(\mathring{X}^{(2)}_{i,N})\end{bmatrix} = (\mathring{\Theta}^T_{i,N}\mathring{\Theta}_{i,N})^{-1}\mathring{\Theta}^T_{i,N}\mathring{\Psi}_{i,N}. \tag{30}$$

From (29) and (30), the PI-based off-policy ADP algorithm is presented in Algorithm 2.

Algorithm 2 PI-based off-policy ADP

```
Choose threshold $\epsilon > 0$, $N \in \mathbb{Z}_+$, $M \in \mathbb{Z}_+\backslash\{0\}$, $\Delta t > 0$ and $\mathring{X}^{(2)}_{-1,N}$ such that $\mathring{K}_{0,N}(\cdot)$ is stabilizing.
Apply $u_0$ (with exploration noise) to system (1), collect the system state and control input data.
Set $i = 0$, $\mathring{X}^{(1)}_{-1,N} = 0$.
repeat
    Construct data matrices $\mathring{\Theta}_{i,N}$ and $\mathring{\Psi}_{i,N}$ in (29).
    Compute $\mathring{X}^{(1)}_{i,N}$, $\mathring{X}^{(2)}_{i,N}$ by (30).
    $\mathring{P}_{i,N}(t) \leftarrow \mathrm{vecs}^{-1}\left(\mathring{X}^{(1)}_{i,N}F_N(t)\right)$
    $\mathring{K}_{i+1,N}(t) \leftarrow \mathrm{vec}^{-1}\left(\mathring{X}^{(2)}_{i,N}F_N(t)\right)$
    $\gamma \leftarrow \|\mathring{X}^{(1)}_{i,N} - \mathring{X}^{(1)}_{i-1,N}\| + \|\mathring{X}^{(2)}_{i,N} - \mathring{X}^{(2)}_{i-1,N}\|$
    $i \leftarrow i+1$
until $\gamma < \epsilon$
Use $\mathring{u}_i(t) = -\mathring{K}_{i,N}(t)x(t)$ as the approximate optimal control.
```

The convergence of Algorithm 2 to the optimal solutions is given in the following corollary, whose derivation and proof are similar to Corollary 15, thus omitted.

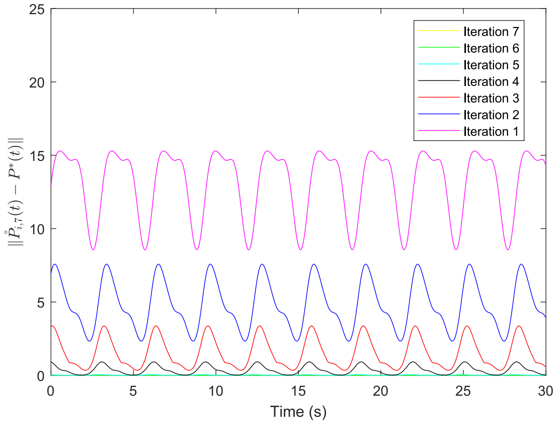

Fig. 1. Differences between $\mathring{P}_{i,7}(\cdot)$ and $P^*(\cdot)$.

Corollary 17. *Under Assumption 16 and the conditions of Theorem 8, $\forall\epsilon > 0$, $\exists\bar{i} \in \mathbb{Z}_+$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}$, $N \in \mathbb{Z}_+$, $\forall t \in \mathbb{R}$,*

$$\|\mathring{P}_{\bar{i},N}(t) - P^*(t)\| < \epsilon, \quad \|\mathring{K}_{\bar{i}+1,N}(t) - K^*(t)\| < \epsilon.$$

4. An example

In this section, the proposed algorithms in last section are applied to the periodic linear quadratic optimal control of the well-known lossy Mathieu equation, without the exact knowledge of system dynamics. The lossy Mathieu equation is a classic example, both in the theoretic study of linear periodic systems (Jovanović & Fardad, 2008; Zhang & Serrani, 2009; Zhou, Hagiwara, & Araki, 2002), and in the modeling of many engineering applications (Wereley, 1990).

Consider the following second-order linear periodic system (Wereley, 1990, Section 1.1)

$$\dot{x}(t) = \begin{bmatrix}0 & 1\\ -(a-2q\cos(\omega_p t)) & -2\zeta\end{bmatrix}x(t) + \begin{bmatrix}0\\1\end{bmatrix}u(t), \tag{31}$$

where the parameters a, q and ζ are not required to be known for the application of our PI-based learning algorithms. When some bounds on these parameters are known, we can apply robust control techniques to find a stabilizing, not necessarily optimal, controller for (31). For example, under the following condition,

$$|a| < 5, \quad |q| < 5, \quad |\zeta| < 5, \tag{32}$$

by Khalil (2002, Theorem 4.9), a choice of initial controller gain $K_0 = [15, 10]$ stabilizes the system (31). Here we are interested in finding a desired suboptimal controller (that is close to the optimal controller) without the exact system dynamics, starting from the robustly stabilizing (but not optimal) control gain K_0. In the simulation, we set parameters $a = 1$, $q = 2$, $\zeta = 0.2$, which satisfy condition (32). The exploration noise is chosen as $u_e(t) = 0.2\sum_{j=1}^{10}\sin(\omega_j t)$, where ω_j is sampled from the uniform distribution over $[-10, 10]$. Other parameters are chosen as $C = I_2$, $R = 1$, $\epsilon = 0.01$, $N = 7$, $M = 100$, $\Delta t = 0.1$. Both algorithms stopped after only 7 iterations. The norm of the difference between $\mathring{P}_{i,7}(\cdot)$ and $P^*(\cdot)$ is shown in Fig. 1. The norm of the difference between $\mathring{K}_{i+1,7}(\cdot)$ and $K^*(\cdot)$ is shown in Fig. 2. The convergence can be easily identified from Figs. 1 and 2, aligned with the conclusions of Corollary 17. Since two algorithms start with the same initial control gain, they generate almost the same learning processes, thus the learning process of Algorithm 1 is omitted.

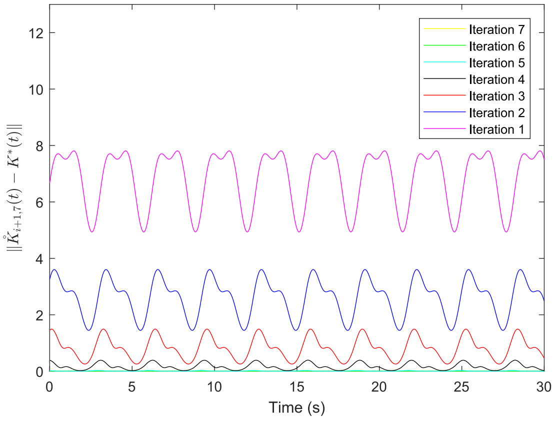

Fig. 2. Differences between $\mathring{K}_{i+1,7}(\cdot)$ and $K^*(\cdot)$.

5. Conclusions

In this paper, reinforcement learning techniques have been used to solve the infinite-horizon adaptive optimal control problem for linear periodic systems with unknown dynamics. Specifically, by means of policy iteration, both on-policy and off-policy ADP algorithms are proposed to solve the infinite-horizon adaptive periodic linear quadratic optimal control problem, using the input/state data collected along the system trajectories. Under mild conditions, it is shown that, starting with an initial stabilizing controller, either the on-policy or off-policy ADP algorithm generates a sequence of suboptimal controllers whose convergence to the optimal solution is guaranteed. The numerical example of the well-known lossy Mathieu equation illustrates the effectiveness of the proposed approaches.

Acknowledgments

This work has been supported in part by the National Science Foundation, USA under Awards ECCS-1501044 and EPCN-1903781.

Appendix A. Proof of Lemma 12

Since $A_i(\cdot)$ is stable, by Lemma 1, we have

$$\|\Phi_i(t,t_0)\| \leq \gamma_i e^{-\alpha_i(t-t_0)}, \quad t \geq t_0, \tag{A.1}$$

where $\Phi_i(t,t_0)$ is the state transition matrix associated with $A_i(\cdot)$, γ_i and α_i are positive constants associated with $A_i(\cdot)$, and are independent of t_0. Define

$$\begin{aligned} \dot{x}(t) &= \hat{A}_{i,N}(t)x(t) \\ &= A_i(t)x(t) + B(t)(K_i(t) - \hat{K}_{i,N}(t))x(t). \end{aligned} \tag{A.2}$$

Since $\lim_{N\to\infty} \hat{K}_{i,N}(t) = K_i(t)$ uniformly on $\mathbb{R}$, there exists $\bar{N}_1$, such that $\forall N > \bar{N}_1$,

$$\|B(t)(K_i(t) - \hat{K}_{i,N}(t))\| < \epsilon_1, \quad \forall t \in \mathbb{R}.$$

where $\epsilon_1\gamma_i < \alpha_i$. Then by Lemma (Teschl, 2012, Theorem 3.20),

$$\|\hat{\Phi}_{i,N}(t,t_0)\| \leq \sqrt{n}\gamma_i e^{-(\alpha_i - \gamma_i\epsilon_1)(t-t_0)}, \quad t \geq t_0, \tag{A.3}$$

where $\hat{\Phi}_{i,N}(t,t_0)$ is the state transition matrix corresponding to $\hat{A}_{i,N}(\cdot)$. This implies that $\hat{A}_{i,N}(\cdot)$ is globally uniformly exponentially stable, i.e., (i) is proved.

By (4) in Lemma 7, we have

$$\check{P}_{i,N}(t) - P_i(t) = \int_t^\infty \delta_{i,N}(\tau,t)\mathrm{d}\tau, \tag{A.4}$$

where

$$\begin{aligned} \delta_{i,N}(\tau,t) &= \hat{\Phi}_{i,N}^T(\tau,t)C^TC\left(\hat{\Phi}_{i,N}(\tau,t) - \Phi_i(\tau,t)\right) \\ &+ \left(\hat{\Phi}_{i,N}(\tau,t) - \Phi_i(\tau,t)\right)^T C^TC\Phi_i(\tau,t) \\ &+ \hat{\Phi}_{i,N}^T(\tau,t)\hat{K}_{i,N}^T R\left(\hat{K}_{i,N}\hat{\Phi}_{i,N}(\tau,t) - K_i\Phi_i(\tau,t)\right) \\ &+ \left(\hat{K}_{i,N}\hat{\Phi}_{i,N}(\tau,t) - K_i\Phi_i(\tau,t)\right)^T RK_i\Phi_i(\tau,t). \end{aligned}$$

Let

$$t_{i,N}^* = \operatorname*{argmax}_{t\in[t_0,t_0+T]} \|\check{P}_{i,N}(t) - P_i(t)\|.$$

Due to the continuity and periodicity of $\check{P}_{i,N}(\cdot)$ and $P_i(\cdot)$, such a $t_{i,N}^*$ always exists.

On one hand, for any $\epsilon > 0$, there exists $\bar{t} > t_0 + T$, such that $\forall N > \bar{N}_1$,

$$\begin{aligned} &\left\| \int_{\bar{t}}^\infty \delta_{i,N}(\tau,t_{i,N}^*)\mathrm{d}\tau \right\| \\ &\leq c_0 \int_{\bar{t}}^\infty \|\hat{\Phi}_{i,N}(\tau,t_{i,N}^*)\|^2 + \|\Phi_i(\tau,t_{i,N}^*)\|^2\mathrm{d}\tau \\ &\leq c_1 e^{-2(\alpha_i - \gamma_i\epsilon_1)\bar{t}} < \frac{\epsilon}{2}, \end{aligned} \tag{A.5}$$

where c_0, c_1 are constants, the first inequality comes from (4) and Rudin (1976, Theorem 6.25), the second inequality is obtained by (A.1) and (A.3).

On the other hand, from (A.3), for $\forall N > \bar{N}_1$, any solution of (A.2) with initial state $|x(t_{i,N}^*)| = 1$ will stay in the ball $\mathcal{B}_0 = \{x \in \mathbb{R}^n | |x| \leq \sqrt{n}\gamma_i\}$. Therefore for any $\mu > 0$, there exists $\bar{N}_2 \geq \bar{N}_1$, such that $\forall N > \bar{N}_2$, $\forall x \in \mathcal{B}_0$,

$$\|B(t)(K_i(t) - \hat{K}_{i,N}(t))x(t)\| < \mu, \quad \forall t \geq t_{i,N}^*. \tag{A.6}$$

Then by Khalil (2002, Theorem 3.4), for $\forall \tau \in [t_{i,N}^*, \bar{t}]$, $\forall N > \bar{N}_2$,

$$\|\Phi_i(\tau,t_{i,N}^*) - \hat{\Phi}_{i,N}(\tau,t_{i,N}^*)\| \leq \frac{\sqrt{n}\mu}{c_2}\left(e^{c_2(\bar{t}-t_0)} - 1\right)$$

where $c_2 = \max_t \|A(t)\|$. Above inequality combined with (A.6) implies that $\delta_{i,N}(\tau,t_{i,N}^*)$ converges uniformly to 0 on $\tau \in [t_{i,N}^*, \bar{t}]$, as $N \to \infty$. Then, $\exists \bar{N}_3 \geq \bar{N}_1$, such that $\forall N > \bar{N}_3$,

$$\left\| \int_{t_{i,N}^*}^{\bar{t}} \delta_{i,N}(\tau,t_{i,N}^*)\mathrm{d}\tau \right\| < \frac{\epsilon}{2}. \tag{A.7}$$

Thus from (A.4), (A.7) and (A.5), we have for $N > \bar{N}_3$,

$$\|\check{P}_{i,N}(t) - P_i(t)\| \leq \|\check{P}_{i,N}(t_{i,N}^*) - P_i(t_{i,N}^*)\| < \epsilon, \forall t \in \mathbb{R}.$$

This completes the proof of (ii).

Appendix B. Proof of Lemma 13

By Lemma 12, $\exists \bar{N}_0 > 0$, such that $\forall N > \bar{N}_0$, $\hat{K}_{i,N}(\cdot)$ is stabilizing. Using the same $x(t)$ generated by (16), and (17), (18), by the similar derivation for (15), we have

$$\hat{\Theta}_{i,N}\begin{bmatrix} \mathrm{vec}(\check{X}_{i,N}^{(1)}) \\ \mathrm{vec}(\check{X}_{i,N}^{(2)}) \end{bmatrix} = \hat{\Psi}_{i,N} + \check{E}_{i,N}, \tag{B.1}$$

where $\check{X}_{i,N}^{(1)}$ and $\check{X}_{i,N}^{(2)}$ are Fourier coefficients satisfying,

$$\begin{aligned} \mathrm{vecs}(\check{P}_{i,N}(t)) &= \check{X}_{i,N}^{(1)}F_N(t) + \check{e}_{i,N}^{(1)}(t), \\ \mathrm{vec}(\check{K}_{i+1,N}(t)) &= \check{X}_{i,N}^{(2)}F_N(t) + \check{e}_{i,N}^{(2)}(t), \end{aligned}$$

and

$$\check{e}_{i,j,N} = \tilde{x}^T(t_j)\check{e}^{(1)}_{i,N}(t_j) - \tilde{x}^T(t_{j+1})\check{e}^{(1)}_{i,N}(t_{j+1}) + \int_{t_j}^{t_{j+1}} (x^T \otimes 2u_e^T R)\check{e}^{(2)}_{i,N} dt.$$

Subtracting (19) from (B.1) yields

$$\hat{\Theta}_{i,N}(\check{Z}_{i,N} - \hat{Z}_{i,N}) = \check{E}_{i,N} - \hat{E}_{i,N},$$

where $\hat{Z}_{i,N} = \left[\mathrm{vec}^T(\hat{X}^{(1)}_{i,N}), \quad \mathrm{vec}^T(\hat{X}^{(2)}_{i,N})\right]^T$, $\check{Z}_{i,N} = \left[\mathrm{vec}^T(\check{X}^{(1)}_{i,N}), \quad \mathrm{vec}^T(\check{X}^{(2)}_{i,N})\right]^T$. By Assumption 10, there is

$$|\check{Z}_{i,N} - \hat{Z}_{i,N}|^2 \leq \frac{1}{M\alpha}|\check{E}_{i,N} - \hat{E}_{i,N}|^2 \leq \frac{4}{M\alpha}|\check{E}_{i,N}|^2,$$

where the last inequality holds because by the least square regression, $|\hat{E}_{i,N}| \leq |\check{E}_{i,N}|$. Therefore,

$$|\check{Z}_{i,N} - \hat{Z}_{i,N}|^2 \leq \frac{4}{M\alpha}\sum_{j=0}^{M-1} \check{e}^2_{i,j,N} \leq \frac{4}{\alpha}(\max_j |\check{e}_{i,j,N}|)^2.$$

By Lemma 12, $\lim_{N\to\infty} \check{P}_{i,N}(t) = P_i(t)$ uniformly on $\mathbb{R}$, so the Fourier coefficients of $\check{P}_{i,N}(\cdot)$ also converge to those of $P_i(\cdot)$. Then by Lemma 4, errors $|\check{e}^{(1)}_{i,N}(t)|$ and $|\check{e}^{(2)}_{i,N}(t)|$ converge uniformly to 0, as $N \to \infty$. Thus, for each $\epsilon > 0$, there exists some positive integer $\bar{N}_1 > \bar{N}_0$ such that $\forall N > \bar{N}_1$, $N \in \mathbb{Z}_+$, $\max_j |\check{e}_{i,j,N}| < \sqrt{\alpha\epsilon}/2$, which leads to $|\check{Z}_{i,N} - \hat{Z}_{i,N}|^2 < \epsilon$.

By Lemma 4 and Hölder's inequality, $\forall \epsilon > 0$, $\exists \bar{N}_{a,j} > \bar{N}_0$, such that $\forall N > \bar{N}_{a,j}$, $\forall t \in \mathbb{R}$,

$$\left|\left[\mathrm{vecs}(\hat{P}_{i,N}(t) - \check{P}_{i,N}(t))\right]_j\right| \leq \left|\left[\hat{X}^{(1)}_{i,N} - \check{X}^{(1)}_{i,N}\right]_{j,\cdot}\right|_1 |F_N(t)|_\infty + \left|\left[\check{e}^{(1)}_{i,N}(t)\right]_j\right| < \frac{\epsilon}{\sqrt{n_1}}.$$

Thus, setting $\bar{N}_a = \max_j \bar{N}_{a,j}$, $j = 1, 2, \ldots, n_1$, and using Lemma 6, we obtain $\forall \epsilon > 0$, $\exists \bar{N}_a > \bar{N}_0$, such that $\forall N > \bar{N}_a$, $\forall t \in \mathbb{R}$, $\|\hat{P}_{i,N}(t) - \check{P}_{i,N}(t)\| < \epsilon$. Similarly, we have $\forall \epsilon > 0$, $\exists \bar{N}_b > \bar{N}_0$, such that $\forall N > \bar{N}_b$, $\forall t \in \mathbb{R}$, $\|\hat{K}_{i+1,N}(t) - \check{K}_{i+1,N}(t)\| < \epsilon$. By choosing $\bar{N} = \max\{\bar{N}_a, \bar{N}_b\}$, Lemma 13 is thus proved.

References

Anton, Deitmar (2005). *A first course in harmonic analysis* (2nd ed.). New York: Springer.

Bellman, Richard E. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Bertsekas, Dimitri P., & Tsitsiklis, John N. (1996). *Neuro-dynamic programming*. Belmont: Athena Scientific.

Bian, Tao, Jiang, Yu, & Jiang, Zhong-Ping (2014). Adaptive dynamic programming and optimal control of nonlinear nonaffine systems. *Automatica*, *50*(10), 2624–2632.

Bittanti, Sergio (1986). Deterministic and stochastic linear periodic systems. In Sergio Bittanti (Ed.), *Time series and linear systems* (pp. 141–182). Berlin, DE: Springer, chapter 5.

Bittanti, Sergio, Bolzern, Paolo, & Colaneri, Patrizio (1984). Stability analysis of linear periodic systems via the Lyapunov equation. In *9th IFAC World congress, Budapest, Hungary* (pp. 213–216).

Bittanti, Sergio, Colaneri, Patrizio, & De Nicolao, Giuseppe (1991). The periodic Riccati equation. In Sergio Bittanti, Alan J. Laub, & Jan C. Willems (Eds.), *The Riccati equation* (pp. 127–162). Berlin, DE: Springer, chapter 6.

Buşoniu, Lucian, de Bruin, Tim, Tolić, Domagoj, Kober, Jens, & Palunko, Ivana (2018). Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, *46*, 8–28.

Camino, J. F., & Santos, I. F. (2019). A periodic linear–quadratic controller for suppressing rotor-blade vibration. *Journal of Vibration and Control*, *25*(17), 2351–2364.

Ciarlet, Philippe G., Miara, Bernadette, & Thomas, Jean-Marie (1989). *Introduction to numerical linear algebra and optimisation*. Cambridge: Cambridge University Press.

DaCunha, Jeffrey J., & Davis, John M. (2011). A unified Floquet theory for discrete, continuous, and hybrid periodic linear systems. *Journal of Differential Equations*, *251*(11), 2987–3027.

Deptula, Patryk, Rosenfeld, Joel A., Kamalapurkar, Rushikesh, & Dixon, Warren E. (2018). Approximate dynamic programming: Combining regional and local state following approximations. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2154–2166.

Fong, Justin, Tan, Ying, Crocher, Vincent, Oetomo, Denny, & Mareels, Iven (2018). Dual-loop iterative optimal control for the finite horizon LQR problem with unknown dynamics. *Systems & Control Letters*, *111*, 49–57.

Jiang, Yu, & Jiang, Zhong-Ping (2017). *Robust adaptive dynamic programming*. Hoboken: Wiley-IEEE Press.

Jovanović, Mihailo R., & Fardad, Makan (2008). *H*2 Norm of linear time-periodic systems: A perturbation analysis. *Automatica*, *44*(8), 2090–2098.

Kamalapurkar, Rushikesh, Rosenfeld, Joel A., & Dixon, Warren E. (2016). Efficient model-based reinforcement learning for approximate online optimal control. *Automatica*, *74*, 247–258.

Kamalapurkar, Rushikesh, Walters, Patrick, Rosenfeld, Joel A., & Dixon, Warren E. (2018). *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Springer.

Karlsson, Niklas (2018). Control of periodic systems in online advertising. In *IEEE 57th annual conference on decision and control (CDC)*, Miami, FL, USA (pp. 5928–5933).

Khalil, Hassan K. (2002). *Nonlinear systems* (3rd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Kleinman, David L. (1968). On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, *13*(1), 114–115.

Lewis, Frank L., & Liu, Derong (Eds.), (2013). *Reinforcement learning and approximate dynamic programming for feedback control*. Hoboken, New Jersey: Wiley-IEEE Press.

Li, Bin, Yu, Changjun, Teo, Kok Lay, & Duan, Guangren (2011). An exact penalty function method for continuous inequality constrained optimal control problem. *Journal of Optimization Theory and Applications*, *151*(2), 260.

Mareels, Iven, & Polderman, Jan Willem (2012). *Adaptive systems: An introduction*. Boston: Birkhauser.

Narendra, Kumpati S., & Esfandiari, Kasra (2019). Adaptive identification and control of linear periodic systems using second-level adaptation. *International Journal of Adaptive Control and Signal Processing*, *33*(6), 956–971.

Oh, Sang-Rok, Bien, Zeungnam, & Suh, Il Hong (1988). An iterative learning control method with application to robot manipulators. *IEEE Journal of Robotics and Automation*, *4*(5), 508–514.

Pane, Yudha P., Nageshrao, Subramanya P., & Babuška, Robert (2016). Actor-critic reinforcement learning for tracking control in robotics. In *2016 IEEE 55th conference on decision and control (CDC)*, Las Vegas, USA (pp. 5819–5826).

Pang, Bo, Bian, Tao, & Jiang, Zhong-Ping (2019). Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems. *Control Theory and Technology*, *17*(1), 18–29.

Rudin, Walter (1976). *Principles of mathematical analysis*. New York: McGraw-hill.

Saridis, George N., & Lee, Chun-Sing G. (1979). An approximation theory of optimal control for trainable manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(3), 152–159.

Shayman, Mark A. (1985). On the phase portrait of the matrix Riccati equation arising from the periodic control problem. *SIAM Journal on Control and Optimization*, *23*(5), 717–751.

Sutton, Richard S., & Barto, Andrew G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, Massachusetts: MIT Press.

Teschl, Gerald (2012). *Ordinary differential equations and dynamical systems*. Providence, Rhode Island: American Mathematical Society.

Varga, Andras (2008). On solving periodic Riccati equations. *Numerical Linear Algebra with Applications*, *15*(9), 809–835.

Varga, Andras, & Stefan, Pieters (1998). Gradient-based approach to solve optimal periodic output feedback control problems. *Automatica*, *34*(4), 477–481.

Wei, Qinglai, Liu, Derong, Lewis, Frank L., Liu, Yu, & Zhang, Jie (2017). Mixed iterative adaptive dynamic programming for optimal battery energy control in smart residential microgrids. *IEEE Transactions on Industrial Electronics*, *64*(5), 4110–4120.

Werbos, Paul J. (2007). Using ADP to understand and replicate brain intelligence: The next level design? In Leonid I. Perlovsky, & Robert Kozma (Eds.), *Neurodynamics of cognition and consciousness* (pp. 109–123). Springer.

Wereley, Norman M. (1990). *Analysis and control of linear periodically time varying systems* (PhD thesis), Cambridge, Massachusetts: Massachusetts Institute of Technology.

Xu, Jian-Xin (2004). A new periodic adaptive control approach for time-varying parameters with known periodicity. *IEEE Transactions on Automatic Control*, *49*(4), 579–583.

Yang, Feng, Teo, Kok Lay, Loxton, Ryan, Rehbock, Volker, Li, Bin, Yu, Changjun, et al. (2016). Visual MISER: An efficient user-friendly visual program for solving optimal control problems. *Journal of Industrial and Management Optimization*, *12*, 781–810.

Zhang, Zhen, & Serrani, Andrea (2009). Adaptive robust output regulation of uncertain linear periodic systems. *IEEE Transactions on Automatic Control*, *54*(2), 266–278.

Zhou, Jun, Hagiwara, Tomomichi, & Araki, Mituhiko (2002). Stability analysis of continuous-time periodic systems via the harmonic analysis. *IEEE Transactions on Automatic Control*, *47*(2), 292–298.

Bo Pang received the B.Sc. Degree in Automation from the Beihang University, Beijing, China, in 2014, and the M.Sc. degree in Control Science and Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2017. He is currently working toward the Ph.D. degree in Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, U.S.A. His research interests include optimal/stochastic control, approximate/adaptive dynamic programming, and reinforcement learning.

Zhong-Ping JIANG received the M.Sc. degree in statistics from the University of Paris XI, France, in 1989, and the Ph.D. degree in automatic control and mathematics from the Ecole des Mines de Paris (now, called ParisTech-Mines), France, in 1993, under the direction of Prof. Laurent Praly.

Currently, he is a Professor of Electrical and Computer Engineering at the Tandon School of Engineering, New York University. His main research interests include stability theory, robust/adaptive/distributed nonlinear control, robust adaptive dynamic programming, learning based control and their applications to information, mechanical and biological systems. In these fields, he has written five books and is author/co-author of over 450 peer-reviewed journal and conference papers.

Dr. Jiang has served as Senior Editor and Associate Editor for numerous journals. Prof. Jiang is a Fellow of the IEEE, a Fellow of the IFAC and a Clarivate Analytics Highly Cited Researcher.

Prof. Iven Mareels is the director of the IBM Research Laboratory in Australia. He joined IBM Research in February 2018 following a 20-year career at the University of Melbourne, where he spent the last 10 years as the Dean of the Melbourne School of Engineering.

Prof Mareels received the Masters of Electromechanical Engineering in 1982 summa cum laude from the University of Gent, Belgium and the Ph.D. from the Australian National University in 1987 with a thesis on dynamics of adaptive or learning systems. He has contributed to more than 500 refereed publications, and he has co-supervised more than 50 Ph.D. students. He is a coinventor of 39 internationally granted patents that focus on the management of large scale, open channel, water distribution networks.

Prof Mareels is a Fellow, and a Director, of The Academy of Technological Sciences and Engineering (Australia); The Institute of Electrical and Electronics Engineers (USA), the International Federation of Automatic Control (IFAC) and Engineers Australia. He is a Foreign Member of the Royal Flemish Academy of Belgium for Science and the Arts. He is internationally registered as a professional engineer. He is a Commander in the Order of the Crown (Belgium).