2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Reinforcement Learning for Adaptive Periodic Linear Quadratic Control

Bo Pang, Zhong-Ping Jiang and Iven Mareels

Abstract— **This paper presents a first solution to the problem of adaptive LQR for continuous-time linear periodic systems. Specifically, reinforcement learning and adaptive dynamic programming (ADP) techniques are used to develop two algorithms to obtain near-optimal controllers. Firstly, the policy iteration (PI) and value iteration (VI) methods are proposed when the model is known. Then, PI-based and VI-based off-policy ADP algorithms are derived to find near-optimal solutions directly from input/state data collected along the system trajectories, without the exact knowledge of system dynamics. The effectiveness of the derived algorithms is validated using the well-known lossy Mathieu equation.**

I. INTRODUCTION

Recently, with significant progress in reinforcement learning and adaptive dynamic programming (ADP), innovative solutions have been obtained for various kinds of problems ranging from engineering to neuroscience, (see [1], [2] and the references therein). In ADP, data generated through the interactions between the plant and the controller is exploited to find approximate optimal controllers without explicitly identifying the plant model. Despite these significant progresses over the past decade [3], [4], [5], [6], little research has been devoted to the adaptive optimal control problem for time-varying systems. It is widely recognized that, for general uncertain time-varying systems, the problem is fundamentally challenging [7], [8]. Thus it is more suitable to focus on special classes of time-varying systems. In this paper, we will study a class of continuous-time linear periodic (CTLP) systems with unknown dynamics.

Indeed, the analysis and control of CTLP systems has been studied by many authors (see [9], [10] and the references therein), due to its important role in modeling practical problems including spacecraft attitude control [11], vibration attenuation [12], online advertising [13] and so on. Optimal control problems of CTLP systems are addressed in [14], [15], [16], assuming the complete knowledge of the system dynamics. Adaptive control of CTLP systems can be found in [7], [8], [17], where optimality was not addressed.

In this paper, two reinforcement learning based ADP algorithms are proposed, such that the adaptive optimal control problem of CTLP systems is solved in the absence of the exact knowledge of system dynamics. Firstly, the model-based policy iteration (PI) and value iteration (VI) for CTLP systems are proposed, based on [14, Theorem 6.2] and [16, Corollary], respectively. Then, we develop the PI-based and VI-based off-policy ADP algorithms, to find approximate optimal controllers directly from input/state data, when the system dynamics are unknown. Under mild conditions, both algorithms converge uniformly to the optimal solutions. With appropriate choice of an initial stabilizing policy, the PI-based ADP algorithm enjoys quadratic convergence rate, while the VI-based ADP algorithm does not require the knowledge of any initial stabilizing controller. The feasibility and effectiveness of both algorithms are demonstrated by the optimal control design for the well-known lossy Mathieu equation.

Notations. $\mathbb{Z}_+$ is the set of nonnegative integers. $\mathbb{S}^n$ denotes the vector space of all n-by-n real symmetric matrices. $|\cdot|$ and $\|\cdot\|$ represent the Euclidean norm for vectors and the Frobenius norm for matrices, respectively. $[x]_j$ denotes the jth element of vector $x \in \mathbb{R}^n$. $[X]_{i,j}$ denotes the element in the ith row and jth column of matrix $X \in \mathbb{R}^{m\times n}$. $\lfloor x \rfloor$ represents the largest integer less than or equal to $x \in \mathbb{R}$.

II. 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, $T \in \mathbb{R}_+$.

By [14, 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 function

$$J(t_0, \xi, u(\cdot)) = \int_{t_0}^{\infty} r(x(t), u(t))dt, \tag{2}$$

where $u(t) = -K(t)x(t)$, $r(x(t), u(t)) = |C(t)x(t)|^2 + u^T(t)R(t)u(t)$, $K(\cdot) : \mathbb{R} \to \mathbb{R}^{m\times n}$, $C(\cdot) : \mathbb{R} \to \mathbb{R}^{l\times n}$, $R(\cdot) : \mathbb{R} \to \mathbb{R}^{m\times m}$ are continuous and T-periodic, with $R(\cdot) > 0$; $x(t)$ is the solution of equation (1) with initial state $x(t_0) = \xi$. Associated with the optimal control problem is the well-known periodic Riccati equation (PRE)

$$\begin{aligned} -\dot{P}(t) = {} & A^T(t)P(t) + P(t)A(t) \\ & - P(t)B(t)R^{-1}(t)B^T(t)P(t) + C^T(t)C(t). \end{aligned} \tag{3}$$

*This work has been supported in part by the U.S. National Science Foundation grant ECCS-1501044.

B. Pang and Z.-P. Jiang are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Six Metrotech Center, Brooklyn, NY 11201, USA (bo.pang@nyu.edu; zjiang@nyu.edu).

I. Mareels is with the IBM Research - Australia, Melbourne, Vic 3006, Australia (i.mareels@unimelb.edu.au).

978-1-7281-1398-2/19/$31.00 ©2019 IEEE

Under certain conditions, the optimal solution to the periodic linear quadratic control problem exists and is unique.

Assumption 1 ([18, Theorem 4]). *$(A(\cdot), B(\cdot))$ is stabilizable and $(A(\cdot), C(\cdot))$ is detectable.*

Lemma 1 ([14, Theorems 6.5 and 6.12]). *There exists a unique symmetric, periodic and positive semidefinite (SPPS) solution of the PRE, and the corresponding closed-loop system is stable, if and only if Assumption 1 is satisfied. Denote the unique SPPS solution as $P^*(\cdot)$. Then 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 $J^*(t_0, \xi) = J(t_0, \xi, u^*(\cdot)) = \xi^T P^*(t_0)\xi$.*

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 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. The following lemma gives the asymptotic property of using f_N to approximate f.

Lemma 2 ([19, Theorem 1.5.1]). *If f is T-periodic, continuous and piecewise continuously differentiable, then $f_N \to f$ uniformly on $\mathbb{R}$, as $N \to \infty$.*

When the matrices $A(\cdot)$ and $B(\cdot)$ are unknown, the PRE can hardly be directly solved. By reinforcement learning techniques, PI-based and VI-based ADP algorithms are proposed to find approximate optimal controllers directly from the collected data in the next two sections, respectively.

Definition 1. *For matrices $X \in \mathbb{R}^{n\times m}$, $Y \in \mathbb{S}^m$, and vector $v \in \mathbb{R}^n$, define*

$$\begin{aligned}
\mathrm{vec}(X) &= [x_1^T, x_2^T, \cdots, x_m^T]^T,\\
\mathrm{vecs}(Y) &= [y_{11}, \sqrt{2}y_{12}, \cdots, \sqrt{2}y_{1m}, y_{22}, \sqrt{2}y_{23},\\
&\quad \cdots, \sqrt{2}y_{m-1,m}, y_{m,m}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)},\\
\tilde{v} &= [v_1^2, \sqrt{2}v_1v_2, \cdots, \sqrt{2}v_1v_n, v_2^2, \sqrt{2}v_2v_3,\\
&\quad \cdots, \sqrt{2}v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)},
\end{aligned}$$

where x_i is the ith column of X. Let $\mathrm{vecs}^{-1}(\cdot)$ and $\mathrm{vec}^{-1}(\cdot)$ denote the inverse functions of $\mathrm{vecs}(\cdot)$ and $\mathrm{vec}(\cdot)$, resp..

As it can be directly checked, we have

Lemma 3. *For $X \in \mathbb{S}^n$, $Y \in \mathbb{R}^{m\times n}$, $|\mathrm{vecs}(X)| = \|X\|$, $|\mathrm{vec}(Y)| = \|Y\|$.*

In the rest of this paper, we omit the dependence of variables on time t when there is no ambiguity.

III. PI-BASED ADP ALGORITHM FOR CTLP SYSTEMS

A. Model-based PI for known CTLP Systems

Theorem 1. *Under Assumption 1, 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 periodic Lyapunov equation*

$$-\dot{P}_i = A_i^T P_i + P_i A_i + C^T C + K_i^T R K_i, \quad (4)$$

where $A_i = A - BK_i$.

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

$$K_{i+1} = R^{-1}B^T P_i. \quad (5)$$

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 uniformly to $P^*(\cdot)$ and $K^*(\cdot)$, respectively.

Sketch of Proof: In [14, Theorem 6.2], if $(A(\cdot), B(\cdot))$ is stabilizable, it is shown that $P_i(\cdot)$ converge pointwise and monotonically to the maximal solution [14, Section 6.3.1.1] of PRE (3) . By Lemma 1, if further $(A(\cdot), C(\cdot))$ is detectable, $P^*(\cdot)$ is equal to the maximal solution. Then it is established that $\{P_i(\cdot)\}_{i=0}^{\infty}$ is equicontinuous. The uniform convergence follows by the Arzelà-Ascoli theorem. □

B. PI-based ADP Algorithm for Unknown CTLP Systems

Define the periodic control gains

$$\mathring{K}_i(t) = \mathrm{vec}^{-1}\left(\mathring{X}_{i-1}^{(2)} F_N(t)\right), \quad i \in \mathbb{Z}_+, \quad (6)$$

where

$$\begin{aligned}
F_N(t) = [&1, \cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t),\\
&\cdots, \cos(N\omega t), \sin(N\omega t)]^T, \quad N \in \mathbb{Z}_+,
\end{aligned}$$

$\mathring{X}_i^{(2)} \in \mathbb{R}^{mn\times(2N+1)}$, $i \in \mathbb{Z}_+$ is the weight matrix to be determined in ith iteration, $\mathring{X}_{-1}^{(2)}$ is chosen to make $\mathring{K}_0(t)$ stabilizing. Let $\breve{P}_i(\cdot)$ denote the unique SPPS solution of periodic Lyapunov equation,

$$-\dot{\breve{P}}_i = \mathring{A}_i^T \breve{P}_i + \breve{P}_i \mathring{A}_i^T + C^T C + \mathring{K}_i^T R \mathring{K}_i, \quad (7)$$

where $\mathring{A}_i^T = A(t) - B(t)\mathring{K}_i(t)$. If $\mathring{K}_i(t)$ is stabilizing, such a $\breve{P}_i(t)$ always exists by [20, Lemma 2]. Then an improved control gain can be obtained

$$\breve{K}_{i+1} = R^{-1}B^T \breve{P}_i. \quad (8)$$

Apply a feedback control policy u_0 to system (1), which yields the boundedness of the solutions of the closed-loop system. We obtain

$$\dot{x} = \mathring{A}_i x + B(\mathring{K}_i x + u_0). \quad (9)$$

By (7), (8) and (9), differentiating $x^T \breve{P}_i x$ with respect to t yields

$$\begin{aligned}
\frac{\mathrm{d}x^T \breve{P}_i x}{\mathrm{d}t} &= x^T(-C^T C - \mathring{K}_i R \mathring{K}_i)x\\
&\quad + 2(u_0 + \mathring{K}_i x)^T R \breve{K}_{i+1} x.
\end{aligned} \quad (10)$$

By integrating both sides of (10) from t_j to t_{j+1}, where $t_j = j\Delta t$, $j \in \mathbb{Z}_+$, Δt is the sampling interval, and rearranging the terms, we have

$$\begin{aligned}&\tilde{x}^T(t_{j+1})\mathrm{vecs}(\breve{P}_i(t_{j+1})) - \tilde{x}^T(t_j)\mathrm{vecs}(\breve{P}_i(t_j))\\&-\int_{t_j}^{t_{j+1}}\left(x^T\otimes(2(u_0+\mathring{K}_i x)^T R)\right)\mathrm{vec}(\breve{K}_{i+1})\mathrm{d}t\\&=-\int_{t_j}^{t_{j+1}} x^T C^T C x\mathrm{d}t-\int_{t_j}^{t_{j+1}}\mathrm{vec}(\mathring{K}_i)^T\\&(x^T\otimes I_m)^T R(x^T\otimes I_m)\mathrm{vec}(\mathring{K}_i)\mathrm{d}t.\end{aligned}\tag{11}$$

If $j = 1, 2\cdots, M$, $M \in \mathbb{Z}_+\backslash\{0\}$, substituting (6) and the following approximations into (11)

$$\begin{aligned}\mathrm{vecs}(\breve{P}_i(t)) &= \breve{X}_i^{(1)}F_N(t)+\breve{e}_{i,N}^{(1)}(t),\\ \mathrm{vec}(\breve{K}_{i+1}(t)) &= \breve{X}_i^{(2)}F_N(t)+\breve{e}_{i,N}^{(2)}(t),\end{aligned}\tag{12}$$

where $\mathring{e}_{i,N}^{(1)}$ and $\mathring{e}_{i,N}^{(1)}$ are approximation errors, we obtain

$$\mathring{\Theta}_i\begin{bmatrix}\mathrm{vec}(\mathring{X}_i^{(1)})\\ \mathrm{vec}(\mathring{X}_i^{(2)})\end{bmatrix} = \mathring{\Psi}_i+\mathring{E}_{i,N},\tag{13}$$

where $\mathring{E}_{i,N}$ summarizes the effects of the approximation errors $\mathring{e}_{i,N}^{(1)}$ and $\mathring{e}_{i,N}^{(1)}$,

$$\mathring{\Theta}_i=\begin{bmatrix}F_x(t_1)-F_x(t_0), & -F_{xu_0,0}-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{1,0}\\ F_x(t_2)-F_x(t_1), & -F_{xu_0,1}-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{1,1}\\ \vdots & \vdots\\ F_x(t_M)-F_x(t_{M-1}), & -F_{xu_0,M-1}-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{1,M-1}\end{bmatrix},$$

$$\mathring{\Psi}_i=\begin{bmatrix}-c_0-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{2,0}\mathrm{vec}(\mathring{X}_{i-1}^{(2)})\\ -c_1-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{2,1}\mathrm{vec}(\mathring{X}_{i-1}^{(2)})\\ \vdots\\ -c_{M-1}-\mathrm{vec}^T(\mathring{X}_{i-1}^{(2)})\Delta_{2,M-1}\mathrm{vec}(\mathring{X}_{i-1}^{(2)})\end{bmatrix},$$

$$\begin{aligned}\Delta_{1,j}&=\int_{t_j}^{t_{j+1}}F_n^T\otimes x^T\otimes 2F_N\otimes x\otimes R\mathrm{d}t,\\ c_j&=\int_{t_j}^{t_{j+1}}x^T C^T C x\mathrm{d}t,\\ \Delta_{2,j}&=\int_{t_j}^{t_{j+1}}(F_N\otimes x\otimes I_m)\,R\left(F_N^T\otimes x^T\otimes I_m\right)\mathrm{d}t,\\ F_{xu_0,j}&=\int_{t_j}^{t_{j+1}}F_N^T\otimes x^T\otimes(2u_0^T R)\mathrm{d}t,\end{aligned}$$

Note that we hope the approximation errors can be made as small as possible in (12) and (13). To this end, the following assumption is imposed in the spirit of persistent excitation (PE) in adaptive control [21].

Assumption 2. *For all $i \in \mathbb{Z}_+$, there exist $\bar{M} > 0$ and $\alpha > 0$, such that for all $M > \bar{M}$, $M \in \mathbb{Z}_+$, we have*

$$\frac{1}{M}\mathring{\Theta}_i^T\mathring{\Theta}_i \geq \alpha I_{(n_1+n_2)(2N+1)},$$

where $n_1 = n(n+1)/2$, $n_2 = mn$.

The convergence of the PI-based off-policy ADP algorithm to the optimal solutions is true under Assumption 2 and the following assumption.

Assumption 3. *Matrix-valued functions $B(\cdot)$ and $R(\cdot)$ are piecewise continuously differentiable.*

Lemma 4. *Under Assumptions 2 and 3, if $\mathring{K}_i(\cdot)$ is stabilizing for all $i \in \mathbb{Z}_+$, then $\forall\epsilon > 0$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}, N \in \mathbb{Z}_+$, $\forall t \in \mathbb{R}$,*

$$\|\mathring{P}_i(t)-\breve{P}_i(t)\|<\epsilon,\quad \|\mathring{K}_{i+1}(t)-\breve{K}_{i+1}(t)\|<\epsilon.$$

Sketch of Proof: Firstly $\breve{P}_i(t)$ is expanded to

$$\mathrm{vecs}(\breve{P}_i(t)) = \breve{X}_i^{(1)}F_N(t)+\breve{e}_{i,N}^{(1)}(t),$$

where $\breve{X}_i^{(1)} \in \mathbb{R}^{n_1\times(2N+1)}$ is Fourier coefficients, $\breve{e}_{i,N}$ is truncation error. Secondly by Lemma 2, Assumption 2 and the property of least square regression, it is shown that $\mathring{X}_i^{(1)}$ can be made arbitrarily close to $\breve{X}_i^{(1)}$ by choosing large N. Then $\lim_{N\to\infty}\mathring{P}_i(t) = \breve{P}_i(t)$ follows from Lemma 2. The proof for $\mathring{K}_i(t)$ and $\breve{K}_i(t)$ is similar. □

Lemma 5. *For all $i \in \mathbb{Z}_+$, if $\lim_{N\to\infty}\mathring{K}_i(t) = K_i(t)$ uniformly on $\mathbb{R}$, then*

(i) When N is large enough, $\mathring{K}_i(\cdot)$ is stabilizing.

(ii) $\lim_{N\to\infty}\breve{P}_i(t) = P_i(t)$ uniformly on $\mathbb{R}$.

Proof: By continuity, [22, Theorem 3.4] and [20, Lemma 2], this lemma is not hard to derive, thus omitted. □

Theorem 2. *Under Assumptions 2 and 3, given $\bar{i} \in \mathbb{Z}_+$, for any $\epsilon > 0$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}$, $N \in \mathbb{Z}_+$, $\mathring{K}_i(\cdot)$ is stabilizing, and*

$$\|\mathring{P}_i(t)-P_i(t)\|<\epsilon,\quad \|\mathring{K}_{i+1}(t)-K_{i+1}(t)\|<\epsilon,\quad \forall t\in\mathbb{R},$$

for $i = 1, 2, \cdots, \bar{i}$.

Sketch of Proof: By suitable applications of Lemmas 4 and 5, we can prove that $\exists\bar{N}_i > 0$, such that $\forall N > \bar{N}_i$,

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

for $i = 1, 2, \cdots, \bar{i}$. The proof is completed by setting $\bar{N} = \max_i \bar{N}_i$. □

Corollary 1. *Under Assumptions 1, 2, 3 and the conditions of Theorem 1, $\forall\epsilon > 0$, $\exists\bar{i} \in \mathbb{Z}_+$, $\exists\bar{N} > 0$, such that $\forall N > \bar{N}$, $N \in \mathbb{Z}_+$,*

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

Proof: It follows directly from Theorem 1 and Theorem 2, by the well-known triangle inequality. □

IV. VI-BASED ADP ALGORITHM FOR CTLP SYSTEMS

As Newton's method, PI enjoys quadratic convergence rate [14, Section 6.3.1.5]. However, it may be not easy to find an initial stabilizing controller required in Theorem 1. In this section, VI is adopted to find the approximate optimal controller without the knowledge of any stabilizing controller.

A. Model-based VI for known CTLP Systems

The value iteration method is based on the asymptotic property of the solution to the finite-horizon periodic linear quadratic optimal control problem [5]. For any $t_0 < t_f$ and a measurable locally essentially bounded input $u : [t_0, t_f) \to \mathbb{R}^m$, define mapping

$$\mathcal{T}^u_{[t_0,t_f)}(F)(\xi) = x(t_f)^T F x(t_f) + \int_{t_0}^{t_f} r(x(t), u(t))dt,$$

for all $\xi \in \mathbb{R}^n$ and $F \geq 0$, where $x(t_0) = \xi$. Starting with $P(t_f) = F$ at t_f, the corresponding solution of the PRE (3) at time $t < t_f$, denoted by $P(t; t_f, F)$, satisfies

$$\xi^T P(t; t_f, F)\xi = \min_u \mathcal{T}^u_{[t,t_f)}(F)(\xi).$$

The model-based VI is presented in the following theorem.

Theorem 3. *Under Assumption 1, if $F \geq 0$, then*

$$\lim_{t\to-\infty} (P(t; t_f, F) - P^*(t)) = 0. \tag{14}$$

Sketch of Proof: If $F > 0$, [16, Corollary] and Assumption 1 implies (14). If $F = 0$, (14) follows by finding that $P(t; t_f, 0)$ is nondecreasing and bounded from the above. The remaining case is proved by the application of the Squeeze theorem to the first and second cases. □

B. VI-based ADP Algorithm for Unknown CTLP Systems

Assumption 4. *$A(\cdot)$, $B(\cdot)$ and $R(\cdot)$ are T-periodic and continuously differentiable on $\mathbb{R}$.*

To avoid confusion, next we use $s \in \mathbb{R}$ for the algorithmic time, and $t \in \mathbb{R}$ is reserved for the system evolution time. By Theorem 3, we are interested in solving following final value problem on $[0, s_f]$ for large $s_f > 0$

$$\begin{aligned} -\dot{P}(s) &= A^T(s)P(s) + P(s)A(s) + C^T(s)C(s) \\ &- P(s)B(s)R^{-1}(s)B^T(s)P(s), \quad P(s_f) = F, \end{aligned} \tag{15}$$

where $F \geq 0$. Notice that $P(\cdot)$ exists and is bounded on $[0, s_f]$.

Define matrix-valued functions

$$\begin{aligned} H(s,t) &= A^T(t)P(s) + P(s)A(t), \\ K^t(s) &= R(t)^{-1}B^T(t)P(s). \end{aligned} \tag{16}$$

Assuming a measurable locally essentially bounded input u_0 is applied to system (1) to collect input/state data for learning, we have

$$\begin{aligned} &\frac{\mathrm{d}x^T(t)P(s)x(t)}{\mathrm{d}t} = \dot{x}^T(t)P(s)x(t) + x^T(t)P(s)\dot{x}(t) \\ &= x^T(t)H(s,t)x(t) + 2u_0^T(t)R(t)K^t(s)x(t). \end{aligned} \tag{17}$$

Integrating both sides of (17) from t_j to t_{j+1}, and rearranging the terms, we obtain

$$\begin{aligned} &(\tilde{x}(t_{j+1}) - \tilde{x}(t_j))^T \mathrm{vecs}(P(s)) = \\ &\int_{t_j}^{t_{j+1}} \tilde{x}^T(t)\mathrm{vecs}(H(s,t))\mathrm{d}t + \\ &\int_{t_j}^{t_{j+1}} \left(x^T(t) \otimes 2u_0^T(t)R(t)\right) \mathrm{vec}(K^t(s))\mathrm{d}t. \end{aligned} \tag{18}$$

Note that for fixed $s \in [0, s_f]$, $H(s,t)$ and $K^t(s)$ are periodic with respect to time $t \in \mathbb{R}$. Thus we can express $\mathrm{vecs}(H(s,t))$ and $\mathrm{vec}(K^t(s))$ by their Fourier series

$$\begin{aligned} \mathrm{vecs}(H(s,t)) &= X^{(1)}(s)F_N(t) + e_N^{(1)}(s,t), \\ \mathrm{vec}(K^t(s)) &= X^{(2)}(s)F_N(t) + e_N^{(2)}(s,t), \end{aligned} \tag{19}$$

where $X^{(1)}(s) \in \mathbb{R}^{n_1 \times (2N+1)}$ and $X^{(2)}(s) \in \mathbb{R}^{n_2 \times (2N+1)}$ are Fourier coefficients at algorithmic time s. Analogous to (13), we can construct a linear matrix equation from (18)

$$\Theta \begin{bmatrix} \mathrm{vec}(X^{(1)}(s)) \\ \mathrm{vec}(X^{(2)}(s)) \end{bmatrix} + E_N(s) = \Gamma_{\tilde{x}}\mathrm{vecs}(P(s)), \tag{20}$$

where

$$\Theta = \begin{bmatrix} \int_{t_0}^{t_1} F_N^T \otimes \tilde{x}^T \mathrm{d}t, & \int_{t_0}^{t_1} F_N^T \otimes x^T \otimes 2u_0^T R\mathrm{d}t \\ \int_{t_1}^{t_2} F_N^T \otimes \tilde{x}^T \mathrm{d}t, & \int_{t_1}^{t_2} F_N^T \otimes x^T \otimes 2u_0^T R\mathrm{d}t \\ \vdots & \vdots \\ \int_{t_{M-1}}^{t_M} F_N^T \otimes \tilde{x}^T \mathrm{d}t, & \int_{t_{M-1}}^{t_M} F_N^T \otimes x^T \otimes 2u_0^T R\mathrm{d}t \end{bmatrix},$$

$$\Gamma_{\tilde{x}} = \begin{bmatrix} \tilde{x}^T(t_1) - \tilde{x}^T(t_0) \\ \tilde{x}^T(t_2) - \tilde{x}^T(t_1) \\ \vdots \\ \tilde{x}^T(t_M) - \tilde{x}^T(t_{M-1}) \end{bmatrix},$$

$$\begin{aligned} E_N(s) &= \left[e_{0,N}(s), e_{1,N}(s), \cdots, e_{M-1,N}(s)\right]^T, \\ e_{j,N}(s) &= \int_{t_j}^{t_{j+1}} \tilde{x}^T(t)e_N^{(1)}(s,t)dt \\ &\quad + \int_{t_j}^{t_{j+1}} \left(x^T(t) \otimes 2u_0^T(t)R(t)\right) e_N^{(2)}(s,t)\mathrm{d}t. \end{aligned} \tag{21}$$

Similar to Assumption 2, we make the following assumption on the data-based matrix Θ.

Assumption 5. *Given $N > 0$, there exist $\bar{M} > (n_1 + n_2)(2N + 1)$ and $\alpha > 0$ (independent of N), such that for all $M > \bar{M}$, $M \in \mathbb{Z}_+$,*

$$\frac{1}{M}\Theta^T\Theta \geq \alpha I_{(n_1+n_2)(2N+1)}. \tag{22}$$

Moreover, for all $t \in [0, t_M]$, $|x(t)| \leq \beta$, β independent of N.

Under Assumption 5, let $\Psi = (\Theta^T\Theta)^{-1}\Theta^T$, (20) can be rewritten as

$$\begin{bmatrix} \mathrm{vec}(X^{(1)}(s)) \\ \mathrm{vec}(X^{(2)}(s)) \end{bmatrix} = \Psi\left(\Gamma_{\tilde{x}}\mathrm{vecs}(P(s)) - E_N(s)\right). \tag{23}$$

Lemma 6. *$X^{(1)}(\cdot)$, $X^{(2)}(\cdot)$, $e_N^{(1)}(\cdot,t)$, $e_N^{(2)}(\cdot,t)$ and $E_N(\cdot)$ are continuously differentiable in algorithmic time s.*

Proof: From the definition (16), $H(s,t)$ and $\partial_s H(s,t)$ are continuous both in s and t. Then by Leibniz integral rule and the definition of Fourier coefficients, we have

$$[\dot{W}^{(1)}(s)]_{i,k} = \frac{2}{T}\int_{-T/2}^{T/2} [\mathrm{vecs}(\partial_s H(s,t))]_i p(k,t)dt, \tag{24}$$

where $i = 1, 2, \cdots, n_1$, $k = 1, 2, \cdots, 2N+1$ and

$$p(k,t) = \begin{cases} 1, & \text{if } k = 1 \\ \cos(\omega t k/2), & \text{if } k \text{ is even} \\ \sin(\omega t \lfloor k/2 \rfloor), & \text{if } k \text{ is odd and } k > 1 \end{cases}.$$

Thus by [23, Definition 10.1], $X^{(1)}(\cdot)$ is continuously differentiable in s. By (19), $e_N^{(1)}(\cdot, t)$ is continuously differentiable in s. With similar arguments, we know that $X^{(2)}(\cdot)$ and $e_N^{(2)}(\cdot, t)$ are continuously differentiable in s. Note that $e_N^{(1)}(s,t)$, $e_N^{(2)}(s,t)$, $\partial_s e_N^{(1)}(s,t)$ and $\partial_s e_N^{(2)}(s,t)$ are continuous both in s and t. Again, by Leibniz integral rule, (21) and [23, Definition 10.1], $E_N(\cdot)$ is continuously differentiable in s. This completes the proof. □

Lemma 6 allows us to take derivatives with respect to s on both sides of (23). Combined with the PRE (15) and definitions in (16), we obtain

$$\begin{bmatrix} \text{vec}(\dot{X}^{(1)}(s)) \\ \text{vec}(\dot{X}^{(2)}(s)) \end{bmatrix} = \mathcal{H}(X(s), s) + \mathcal{G}(X(s), s), \quad (25)$$

where

$$\mathcal{H}(X(s), s) = \Psi\Gamma_{\tilde{x}} \Big[-X^{(1)}(s)F_N(s) + \text{vecs}\Big((\text{vec}^{-1}(X^{(2)}(s)F_N(s)))^T R(s) \text{vec}^{-1}(X^{(2)}(s)F_N(s)) \Big) - \text{vecs}(C^T(s)C(s)) \Big],$$

$$X(s) = \left[\left(X^{(1)}(s) \right)^T, \left(X^{(2)}(s) \right)^T \right]^T,$$

and $\mathcal{G}(X(s), s)$ summarizes the effect of the truncation errors. If $\mathcal{G}(X(s), s)$ is ignored, we can define the following differential equation

$$\begin{bmatrix} \text{vec}(\dot{\hat{X}}^{(1)}(s)) \\ \text{vec}(\dot{\hat{X}}^{(2)}(s)) \end{bmatrix} = \mathcal{H}\left(\hat{X}(s), s \right), \qquad \hat{X}(s_f) = 0. \quad (26)$$

Lemma 7. *Under Assumptions 1, 4 and 5, for any $-\infty < s' < s_f$:*

1) *$e_N^{(1)}(s,t)$, $e_N^{(2)}(s,t)$, $\partial_s e_N^{(1)}(s,t)$, $\partial_s e_N^{(2)}(s,t)$ all converge uniformly to 0 on $[s', s_f] \times \mathbb{R}$, as $N \to \infty$.*
2) *for any $\epsilon > 0$, there exists $\bar{N} > 0$, such that $\forall N > \bar{N}$,*

$$\sup_{s \in [s', s_f]} |\mathcal{G}(X(s), s)| < \epsilon.$$

Proof: Please see the proof of Lemma 5 in [24].

□

Lemma 8. *Let $F = 0$ in PRE (15). Under Assumptions 1, 4 and 5, for any $\epsilon > 0$ and any $0 < s_f < \infty$, there exists $\bar{N} > 0$, such that $\forall N > \bar{N}$, $\forall s \in [0, s_f]$*

$$\|X^{(1)}(s) - \hat{X}^{(1)}(s)\| < \epsilon, \quad \|X^{(2)}(s) - \hat{X}^{(2)}(s)\| < \epsilon.$$

Proof: Please see the proof of Lemma 6 in [24]. □

Next, we can solve equation (26) by any numerical method backward in time on $[0, s_f]$. Suppose the numerical solutions of (26) are $\{\hat{X}_k^{(1)}\}_{k=0}^L$ and $\{\hat{X}_k^{(2)}\}_{k=0}^L$, with $s_0 = 0$, $s_L = s_f$, and the maximum step size $h > 0$. As long as s_f is large enough, we are able to choose a $\bar{L} \in \mathbb{Z}_+$, satisfying $s_{\bar{L}} > T$ and $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$. Define

$$\mathcal{U} = \begin{bmatrix} F_N(s_0) & F_N(s_1) & \cdots & F_N(s_{\bar{L}}) \end{bmatrix}^T,$$
$$\mathcal{V} = \begin{bmatrix} \text{vecs}(\hat{H}_0) & \text{vecs}(\hat{H}_1) & \cdots & \text{vecs}(\hat{H}_{\bar{L}}) \end{bmatrix}^T,$$
$$\mathcal{W} = \begin{bmatrix} \text{vec}(\hat{K}_0) & \text{vec}(\hat{K}_1) & \cdots & \text{vec}(\hat{K}_{\bar{L}}) \end{bmatrix}^T,$$
$$\text{vecs}(\hat{H}_k) = \hat{X}_k^{(1)} F_N(s_k), \text{vec}(\hat{K}_k) = \hat{X}_k^{(2)} F_N(s_k).$$

Let

$$\bar{H}(t) = \text{vecs}^{-1}(\bar{X}^{(1)} F_N(t)), \quad \bar{K}(t) = \text{vec}^{-1}(\bar{X}^{(2)} F_N(t)),$$

where

$$\begin{aligned} (\bar{X}^{(1)})^T &= (\mathcal{U}^T\mathcal{U})^{-1}\mathcal{U}^T\mathcal{V}, \\ (\bar{X}^{(2)})^T &= (\mathcal{U}^T\mathcal{U})^{-1}\mathcal{U}^T\mathcal{W}. \end{aligned} \quad (27)$$

The novel VI-based off-policy ADP algorithm is presented in Algorithm 1, whose convergence analysis is contained in following theorem.

Assumption 6. *Given $N > 0$, there exist $\lfloor L/2 \rfloor > \bar{L}_0 > 2N + 1$ and $\alpha > 0$ (independent of N), such that for all $\lfloor L/2 \rfloor > \bar{L} > \bar{L}_0$, $s_{\bar{L}} > T$,*

$$\frac{1}{\bar{L}} \mathcal{U}^T\mathcal{U} \geq \alpha I_{2N+1}.$$

Theorem 4. *Consider the infinite-horizon periodic linear quadratic optimal control problem of system (1) with cost function (2). Under Assumptions 1, 4, 5 and 6, for any $\epsilon > 0$, there exist $\bar{s}_f > 0$, $\bar{N} > 0$, $\bar{h} > 0$, such that $\forall s_f > \bar{s}_f$, $\forall N > \bar{N}$, any $0 < h < \bar{h}$, we have*

$$\sup_{t \in \mathbb{R}} \|\bar{H}(t) - H^*(t)\| < \epsilon, \quad \sup_{t \in \mathbb{R}} \|\bar{K}(t) - K^*(t)\| < \epsilon,$$

where $H^* = A^T P^* + P^* A, K^* = R^{-1} B^T P^*$, and $\bar{L}$ is chosen to satisfy $s_{\bar{L}} > T$, $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$.

Sketch of Proof: By Lemma 7, Lemma 8, Theorem 3, and the triangle inequalities, for any $\epsilon_0 > 0$, there exist large enough s_f, N, and small enough h, such that

$$\sup_{k \in \{1, 2, \cdots, \bar{L}\}} \|\hat{H}_k - H^*(s_k)\| < \epsilon_0.$$

This fact combined with equation (27) and Assumption 6 means $\bar{X}^{(1)}$ is close to the optimal value (same for $\bar{X}^{(2)}$). Then Theorem 4 is obtained by Lemma 2. □

Algorithm 1 VI-based off-policy ADP

1: Choose $\Delta t > 0$, large enough $M > 0$, $N > 0$, $s_f > 0$, and small enough $h > 0$.
2: Apply u_0 (with exploration noise) to system (1), collect input/state data to construct Θ and $\Gamma_{\tilde{x}}$.
3: Solve (26) on $[0, s_f]$ by any numerical method.
4: Choose $\bar{L}$ satisfying $s_{\bar{L}} > T$, $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$.
5: $(\bar{X}^{(2)})^T \leftarrow (\mathcal{U}^T\mathcal{U})^{-1}\mathcal{U}^T\mathcal{W}$.
6: $\bar{K}(t) \leftarrow \text{vec}^{-1}(\bar{X}^{(2)} F_N(t))$.
7: Approximate optimal control $\bar{u}(t) = -\bar{K}(t)x(t)$.

V. Numerical Example

In this section, the proposed methods are applied to the optimal control design of the well-known lossy Mathieu equation [15]. Consider the following second-order linear period system

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -(a - 2q\cos(\omega_p t)) & -2\zeta \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \quad (28)$$

where $\omega_p = 2$ is the pumping frequency; a represents the constant part of the dynamics; q is the pumping amplitude; and ζ is the damping ratio. Here, $T = \pi$.

Notice that the parameters a, q and ζ in (28) are not required to be known for the application of our learning methods. Here the parameters are assumed to satisfy the following condition:

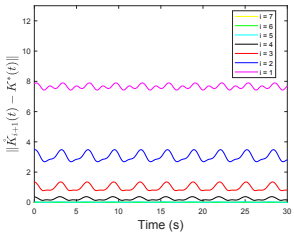
$$|a| < 3, \quad |q| < 3, \quad |\zeta| < 3. \quad (29)$$

In the simulation, the exploration noise is chosen as

$$u_e(t) = 0.2 \sum_{j=1}^{10} \sin(\omega_j t), \quad (30)$$

where ω_j is sampled from the uniform distribution over $[-10, 10]$. Other parameters are chosen as $C = I_2$, $R = 1$, $\epsilon = 0.01$, $N = 9$, $M = 100$, $\Delta t = 0.1$, $s_f = 20$, $h = 0.001$, and $\bar{L} = 8000$.

For the PI-based ADP, by [22, Theorem 4.9], a choice of initial controller gain $K_0 = [10, 7]$ stabilizes the system (28) with parameters satisfying (29). The convergence of PI-based ADP is shown in Figure 1. The convergence of Algorithm 1 is presented in Figure 1. In both cases, the underlying system parameters are chosen as $a = 2.5$, $q = 1$, and $\zeta = 0.2$. These results are consistent with our convergence analysis in previous sections.



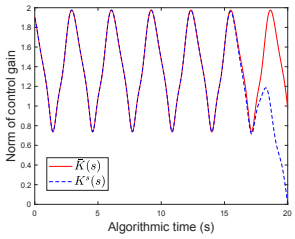


Fig. 1. PI-based ADP convergence Fig. 2. VI-based ADP convergence

VI. Conclusion

In this paper, using reinforcement learning techniques, PI-based and VI-based ADP algorithms are proposed to solve the adaptive LQR problem for CTLP systems. The two algorithms converge uniformly to the optimal solutions under mild conditions. Furthermore, both algorithms are off-policy, which is data-efficient. The case study in the lossy Mathieu equation has been used to demonstrate the efficacy of our new results in adaptive optimal control.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge: MIT Press, 2018.

[2] L. Buşoniu, T. de Bruin, D. Tolić, J. Kober, and I. Palunko, "Reinforcement learning for control: Performance, stability, and deep approximators," *Annual Reviews in Control*, vol. 46, pp. 8 – 28, 2018.

[3] F. L. Lewis and D. Liu, Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken: Wiley-IEEE Press, 2013.

[4] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. Hoboken: Wiley-IEEE Press, 2017.

[5] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348 – 360, 2016.

[6] B. Pang, T. Bian, and Z.-P. Jiang, "Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems," *Control Theory and Technology*, vol. 17, no. 1, pp. 18–29, 2019.

[7] K. S. Narendra and K. Esfandiari, "Adaptive control of linear periodic systems using multiple models," in *IEEE 57th Annual Conference on Decision and Control (CDC)*, Miami, FL, USA, 2018, pp. 589–594.

[8] J.-X. Xu, "A new periodic adaptive control approach for time-varying parameters with known periodicity," *IEEE Transactions on Automatic Control*, vol. 49, no. 4, pp. 579–583, 2004.

[9] J. A. Richards, *Analysis of Periodically Time-Varying Systems*. Heidelberg: Springer, 1983.

[10] S. Bittanti and P. Colaneri, *Periodic Systems: Filtering and Control*. London: Springer, 2009.

[11] M. L. Psiaki, "Magnetic torquer attitude control via asymptotic periodic linear quadratic regulation," *Journal of Guidance, Control, and Dynamics*, vol. 24, no. 2, pp. 386–394, Mar 2001.

[12] S. Bittanti and F. A. Cuzzola, "Periodic active control of vibrations in helicopters: a gain-scheduled multi-objective approach," *Control Engineering Practice*, vol. 10, no. 10, pp. 1043 – 1057, 2002.

[13] N. Karlsson, "Control of periodic systems in online advertising," in *IEEE 57th Annual Conference on Decision and Control (CDC)*, Miami, FL, USA, 2018, pp. 5928–5933.

[14] S. Bittanti, P. Colaneri, and G. De Nicolao, "The periodic Riccati equation," in *The Riccati Equation*, S. Bittanti, A. J. Laub, and J. C. Willems, Eds. Berlin: Springer, 1991, ch. 6, pp. 127–162.

[15] N. M. Wereley, "Analysis and control of linear periodically time varying systems," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1990.

[16] G. De Nicolao, "On the convergence to the strong solution of periodic riccati equations," *International Journal of Control*, vol. 56, no. 1, pp. 87–97, 1992.

[17] Z. Zhang and A. Serrani, "Adaptive robust output regulation of uncertain linear periodic systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 266–278, 2009.

[18] S. Bittanti, "Deterministic and stochastic linear periodic systems," in *Time Series and Linear Systems*, S. Bittanti, Ed. Berlin: Springer, 1986, ch. 5, pp. 141–182.

[19] D. Anton, *A First Course in Harmonic Analysis*, 2nd ed. New York: Springer, 2005.

[20] S. Bittanti, P. Bolzern, and P. Colaneri, "Stability analysis of linear periodic systems via the Lyapunov equation," in *9th IFAC World Congress, Budapest, Hungary*, 1984, pp. 213 – 216.

[21] I. Mareels and J. W. Polderman, *Adaptive Systems: An Introduction*. Boston: Birkhauser, 2012.

[22] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River: Prentice-Hall, 2002.

[23] W. Rudin, *Principles of Mathematical Analysis*. New York: McGraw-hill, 1976.

[24] B. Pang and Z.-P. Jiang, "Adaptive optimal control of linear periodic systems: An off-policy value iteration approach," *arXiv preprint arXiv:1901.08650v2*, 2019.