

Adaptive Optimal Control of Linear Periodic Systems: An Off-Policy Value Iteration Approach

Bo Pang, *Student Member, IEEE*, and Zhong-Ping Jiang, *Fellow, IEEE*

Abstract—This paper studies the infinite-horizon adaptive optimal control of continuous-time linear periodic (CTLP) systems. A novel value iteration (VI) based off-policy ADP algorithm is proposed for a general class of CTLP systems, so that approximate optimal solutions can be obtained directly from the collected data, without the exact knowledge of system dynamics. Under mild conditions, the proofs on uniform convergence of the proposed algorithm to the optimal solutions are given for both the model-based and model-free cases. The VI-based ADP algorithm is able to find suboptimal controllers without assuming the knowledge of an initial stabilizing controller. Application to the optimal control of a triple inverted pendulum subjected to a periodically varying load demonstrates the feasibility and effectiveness of the proposed method.

Index Terms—Adaptive dynamic programming, linear periodic systems, optimal control, value iteration.

I. INTRODUCTION

Recently, reinforcement learning (RL) has invoked a lot of research interests from both researchers in academia and practitioners in industry, due to its successful applications to the design of intelligent computer GO player and many other intelligent agents learning tasks [1]. In RL, agents (or controllers) find optimal decisions (controls) from scratch through its interactions with an unknown environment. In spite of its popularity, most previous RL algorithms have their own limitations. Firstly, the underlying environments are described by Markov decision processes [1], where time is discrete and the state and input spaces are finite or countable. Secondly, stability and safety properties associated with the use of the obtained optimal policy are not considered and guaranteed. However, many physical systems are more naturally described by differential equations, where time is continuous and the state and input spaces are infinite and continuous. Stability and safety are also indispensable considerations in real-world applications, e.g., autonomous vehicles. To this end, over the past decade, another stream of RL algorithms has emerged, to solve the optimal control problems described by differential equations, without the exact knowledge of the system dynamics, and with stability guarantees; see, e.g., [2], [3] and numerous references therein. This class of RL algorithms are often coined adaptive dynamic programming (ADP), to be distinguished from those with Markov decision processes. The

interested reader can consult the books [2], [4] for several practical applications of ADP.

While significant progresses have been made in the development of ADP, most of the existing results are devoted exclusively to time-invariant systems. When problems arise from applications involving time-varying control systems, those ADP algorithms previously developed for time-invariant systems are not directly applicable. Recently, in [5] and [6], the finite-horizon optimal control problem was studied for time-varying systems by ADP. However, the corresponding infinite-horizon optimal control problem for time-varying systems described by differential equations has received scanty attention. There are several technical obstacles for this generalization. First, the stability analysis and control synthesis of time-varying systems are much more challenging than the case of time-invariant systems. Second, predicting the future evolution of the system trajectories becomes an intractable task for general time-varying systems, using only the historical data collected over a finite period of time, which is a key to the development of ADP algorithms. With these observations in mind, how to develop ADP algorithms to address the infinite-horizon optimal control problem of uncertain time-varying systems with guaranteed stability remains an open problem. In this paper, we take a step forward to study this longstanding unresolved issue. To this end, we will examine the infinite-horizon adaptive optimal control of continuous-time linear periodic (CTLP) systems. The analysis and control of linear period systems have played an important role in various applications. By exploiting the periodic time-varying nature, vibration is significantly actively suppressed in wind turbine system [7] and rotor-blade system [8]; through the design of periodic model predictive control strategies for periodic systems, better economic performance is achieved in building climate control [9], drinking water network [10] and nonisolated microgrid [11]; periodic feedback controller is reported to outperform a standard time-invariant feedback controller in online advertising [12]; to name a few. Orbital stabilization of time-invariant nonlinear systems can also be analyzed and designed using linear periodic systems, since linearization of nonlinear systems along a periodic orbit yields linear periodic systems [13, Section 5.1]. It should be emphasized that even for the class of CTLP systems, the design of ADP algorithm is a non-trivial task, as a result of the nonlinear dependence of system parameters on the time.

Inspired by the time-invariant results in [14], a novel value iteration (VI) based ADP algorithm is proposed for a class of CTLP systems in this paper, to find approximate optimal controllers without the exact knowledge of system dynamics

This work was partially supported by the National Science Foundation under Grants ECCS-1501044 and EPCN-1903781.

The authors are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 370 Jay Street, Brooklyn, NY 11201, USA (e-mail: bo.pang@nyu.edu; zjiang@nyu.edu).

and an initial stabilizing controller. The VI-based ADP is based on the asymptotic property of finite-horizon solution of the periodic Riccati equation (PRE). It is claimed in [15] that the solution of the PRE starting from a positive semidefinite initial matrix converges to the stabilizing solution of the same PRE, under certain conditions. However, it is pointed out by the authors of [16] and [17] that the proof of the claim in [15] is based on some wrong preliminary results (see Remark 1). In the present paper, we firstly give a new proof of the claim, and then present a VI-based ADP algorithm, using the Fourier basis approximation. It turns out that the VI-based ADP algorithm amounts to numerically solving the final value problem of a nonlinear differential equation, which only involves collected data and is independent of the exact system dynamics. In Section III, the uniform convergence of the VI-based ADP algorithm to the optimal solution of the corresponding optimal control problem is rigorously proved. In Section IV, the proposed VI-based ADP algorithm is applied to the adaptive optimal control of a triple inverted pendulum subjected to a periodically varying load, which demonstrates the effectiveness of the resulting algorithm. Section V closes the paper with some concluding remarks.

It is worth noting that there is a rich literature on optimal control (see, e.g., [7]–[11]) and on adaptive control (see, e.g., [18]–[20]) for linear periodic systems. However, they have been studied as two separate problems. That is, the optimal control solutions presented in [7]–[11] require the precise knowledge of the system dynamics, while the adaptive control results presented in [18]–[20] do not guarantee optimization of any prescribed cost function. Different from both groups of research, our proposed method finds the suboptimal solutions directly from the input/state data, without the exact knowledge of the system dynamics.

Notations: $\mathbb{R}$ ($\mathbb{R}_+$) is the set of (nonnegative) real numbers. $\mathbb{Z}_+$ is the set of nonnegative integers. $\mathbb{S}^n$ denotes the vector space of all n -by- n real symmetric matrices. $\otimes$ is the Kronecker product operator. $|\cdot|$ and $\|\cdot\|$ represent the Euclidean norm for vectors and the Frobenius norm for matrices, respectively. $[v]_j$ denotes the j th element of vector $v \in \mathbb{R}^n$. $[X]_{i,j}$ denotes the element in i th row and j th column of matrix $X \in \mathbb{R}^{m \times n}$. $[v]$ represents the largest integer no larger than $v \in \mathbb{R}$. $X^\dagger$ denotes the Moore-Penrose inverse of matrix X . $\sigma_{\min}(X)$ is the minimal singular value of matrix X .

II. PROBLEM FORMULATION AND PRELIMINARIES

Consider continuous-time linear periodic systems

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

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

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

Let $\Phi(t, t_0)$, $t > t_0$, $t_0 \in \mathbb{R}$ denote the state transition matrix of the unforced system of (1), with $u = 0$. In the setting of linear periodic system, the matrix $\Phi(t_0 + T, t_0)$

is known as the monodromy matrix. Its eigenvalues (also called characteristic multipliers) are independent of t_0 . $A(\cdot)$ is asymptotically stable if and only if its characteristic multipliers are inside the open unit disk. See [21] for the details.

The infinite-horizon periodic linear quadratic (PLQ) optimal control problem [17, Section 6.5.1.1] is to find a linear stabilizing control law $u(t) = -K(t)x(t)$, where $K(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{m \times n}$ is continuous and T -periodic, such that the following quadratic cost is minimized

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

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

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

Generally, the PRE (3) may admit many different kinds of solutions, among which two particular kinds are relevant to this paper.

Definition 1 ([17], [22]). Consider the real symmetric, periodic and positive semidefinite (SPPS) solutions satisfying PRE (3) over time interval $(-\infty, \infty)$.

- 1) $P_S(\cdot)$ is called strong solution, if the characteristic multipliers of $D_S(t) = A(t) - B(t)R^{-1}(t)B^T(t)P_S(t)$ belong to the closed unit disk.
- 2) $P_+(\cdot)$ is called stabilizing solution, if the characteristic multipliers of $D_+(t) = A(t) - B(t)R^{-1}(t)B^T(t)P_+(t)$ belong to the open unit disk.

Assumption 1. $(A(\cdot), B(\cdot))$ is stabilizable and $(A(\cdot), C(\cdot))$ is detectable [21, Theorem 4].

Under Assumption 1, the optimal solution to the infinite-horizon PLQ control problem exists and is unique [17, Theorems 6.5 and 6.12].

Lemma 1. There exists a unique SPPS solution $P^*(\cdot)$ of the PRE, and the corresponding closed-loop system is stable, if and only if Assumption 1 is satisfied. In addition, a) $P_S = P_+ = P^*$. b) the cost (2) is minimized by the optimal controller $u^*(t) = -K^*(t)x(t)$, with $K^*(t) = R^{-1}(t)B^T(t)P^*(t)$. c) the corresponding minimum cost is $J^*(t_0, \xi) = J(t_0, \xi, u^*(\cdot)) = \xi^T P^*(t_0)\xi$.

In general, it is difficult to obtain an analytic expression for $P^*(\cdot)$, which is a nonlinear matrix-valued function of time t . In this paper, Fourier basis functions are adopted to approximate different periodic functions. For a continuous and T -periodic function $f(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$, partial sums of its Fourier series representation are

$$f_N(x) = \frac{a_0}{2} + \sum_{i=1}^N (a_i \cos(\omega i x) + b_i \sin(\omega i x)),$$

where $\omega = 2\pi/T$, $N \in \mathbb{Z}_+$, $\{a_i\}_{i=0}^N$ and $\{b_i\}_{i=1}^N$ are Fourier coefficients. The following lemma gives the asymptotic property of using f_N to approximate f .

Lemma 2 ([23, Theorem 1.5.1]). *If f is T -periodic, continuous and piecewise continuously differentiable, then $f_N \rightarrow f$ uniformly, as $N \rightarrow \infty$.*

When matrices $A(\cdot)$ and $B(\cdot)$ are unknown, the optimal solution $P^*(\cdot)$ can hardly be obtained directly due to the nonlinearity of the PRE. In next section, VI is exploited to find approximate optimal controllers directly from the input/state data collected along the controlled system trajectories. As it can be directly checked, we have

Fact 1. For $X \in \mathbb{R}^{n \times m}$, $Y \in \mathbb{S}^n$, $v \in \mathbb{R}^n$,

$$|\text{vec}(X)| = \|X\|, \quad |\text{vecs}(Y)| = \|Y\|, \quad v^T Y v \equiv \tilde{v}^T \text{vecs}(Y),$$

where

$$\begin{aligned} \text{vec}(X) &= [X_1^T, X_2^T, \dots, X_m^T]^T, \\ \text{vecs}(Y) &= [y_{11}, \sqrt{2}y_{12}, \dots, \sqrt{2}y_{1m}, y_{22}, \sqrt{2}y_{23}, \\ &\quad \dots, \sqrt{2}y_{m-1,m}, y_{m,m}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}, \\ \tilde{v} &= [v_1^2, \sqrt{2}v_1v_2, \dots, \sqrt{2}v_1v_n, v_2^2, \sqrt{2}v_2v_3, \\ &\quad \dots, \sqrt{2}v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}, \end{aligned}$$

X_i is the i th column of X . In addition, there always exist operations $\text{vec}^{-1}(\cdot)$ and $\text{vecs}^{-1}(\cdot)$, such that $X = \text{vec}^{-1}(\text{vec}(X))$ and $Y = \text{vecs}^{-1}(\text{vecs}(Y))$, respectively.

III. VALUE ITERATION BASED ADAPTIVE DYNAMIC PROGRAMMING FOR CONTINUOUS-TIME LINEAR PERIODIC SYSTEMS

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

$$\begin{aligned} V(t, t_f, \xi, u(\cdot), G) &= x(t_f)^T G x(t_f) \\ &+ \int_t^{t_f} (|C(t)x(t)|^2 + u(t)^T R(t)u(t)) dt, \end{aligned}$$

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

$$\xi^T P(t; t_f, G) \xi = \min_u V(t, t_f, \xi, u(\cdot), G).$$

Generally, $P(\cdot; t_f, G)$ is not necessarily periodic, and different from the SPPS solutions in Definition 1, $P(\cdot; t_f, G)$ satisfies PRE (3) over time interval $(-\infty, t_f]$ [24, Section 6.1.4]. Next, it is shown that $P(t; t_f, G)$ with $G \geq 0$ will approach the SPPS solution $P^*(t)$ of the PRE (3), as time $t \rightarrow -\infty$.

Lemma 3. For any $0 \leq G_1 \leq G_2$ and $t < t_f$,

$$P(t; t_f, G_1) \leq P(t; t_f, G_2).$$

Proof. For any fixed $\xi \in \mathbb{R}^n$, and any measurable locally essentially bounded u ,

$$\xi^T P(t; t_f, G_1) \xi \leq V(t, t_f, \xi, u, G_1) \leq V(t, t_f, \xi, u, G_2).$$

Minimizing the above inequalities simultaneously over u , we obtain $\xi^T P(t; t_f, G_1) \xi \leq \xi^T P(t; t_f, G_2) \xi$. Since ξ is arbitrary, the proof is completed. $\square$

Theorem 1. Under Assumption 1, if $G = G^T \geq 0$, then

$$\lim_{t \rightarrow -\infty} (P(t; t_f, G) - P^*(t)) = 0. \quad (4)$$

Proof. See the Appendix. $\square$

Remark 1. In 1975, Hewer drew the same conclusion [15, Theorem 4.11] with Theorem 1. However, as pointed out in [17, Section 6.3.4] and [16], the proof of [15, Theorem 4.11] was based on [15, Theorem 3.7], which is shown wrong by a counterexample in [16, Section 2]. The proof of Theorem 1 in this paper is new and is included for the sake of completeness.

Note that the solutions of PRE (3) need not evolve according to the same time variable in system (1). To emphasize this point, in the rest of this paper we use $s \in \mathbb{R}$ for the algorithmic time, which is the time used in the PRE, while $t \in \mathbb{R}$ is reserved for the system evolution time, i.e. the time used in system (1). This separation of time variables is essential in the development of our proposed algorithm in the sequel (also see Remark 4).

Theorem 1 means that near-optimal solutions of $P^*(\cdot)$ can be found by solving the PRE (3) backward in time with boundary condition $G \geq 0$. Concretely, we can solve the following final value problem on interval $[0, s_f]$,

$$\begin{aligned} -\dot{P}(s) &= A^T(s)P(s) + P(s)A(s) + C^T(s)C(s) \\ &- P(s)B(s)R^{-1}(s)B^T(s)P(s), \quad P(s_f) = G, \end{aligned} \quad (5)$$

where $G = G^T \geq 0$, $P(s)$ is short for $P(s; s_f, G)$. By Theorem 1, if $s_f > 0$ is large, then $P(s)$ is close to $P^*(s)$ for s near 0. However, the *a priori* knowledge of $A(\cdot)$ and $B(\cdot)$ is still required to solve (5). Next, a VI-based off-policy ADP algorithm is proposed to solve (5) directly from input/state data, without the exact knowledge of $A(\cdot)$ and $B(\cdot)$.

Define matrix-valued functions

$$\begin{aligned} H(s, t) &= A^T(t)P(s) + P(s)A(t), \\ K(s, t) &= R^{-1}(t)B^T(t)P(s). \end{aligned} \quad (6)$$

Then (5) can be rewritten as,

$$\begin{aligned} -\dot{P}(s) &= H(s, s) + C^T(s)C(s) \\ &- K^T(s, s)R(s)K(s, s). \end{aligned} \quad (7)$$

By Theorem 1, as long as s_f is large enough, $H(s, s)$ and $K(s, s)$ with s near 0 will be good approximation to

$$\begin{aligned} H^*(s) &= A^T(s)P^*(s) + P^*(s)A(s), \\ K^*(s) &= R^{-1}(s)B^T(s)P^*(s). \end{aligned} \quad (8)$$

With Lemma 1, note that $K^*(\cdot)$ is the optimal control gain.

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

$$\begin{aligned} \text{vecs}(H(s, t)) &= W^H(s)F_N(t) + e_N^H(s, t), \\ \text{vec}(K(s, t)) &= W^K(s)F_N(t) + e_N^K(s, t), \end{aligned} \quad (9)$$

where $W^H(s) \in \mathbb{R}^{n_1 \times (2N+1)}$, $n_1 = n(n+1)/2$ and $W^K(s) \in \mathbb{R}^{n_2 \times (2N+1)}$, $n_2 = mn$ are Fourier coefficients at algorithmic time s ,

$$F_N(t) = [1, \cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t), \dots, \cos(N\omega t), \sin(N\omega t)]^T,$$

$e_N^H(s, t) \in \mathbb{R}^{n_1}$ and $e_N^K(s, t) \in \mathbb{R}^{n_2}$ are truncation errors.

Here is a preview of subsequent development of our algorithm. We aim at solving (5) (i.e. (7)) directly from input/state data. To this end, firstly, using (6) and (9), the PRE (5) (i.e. (7)) is transformed into an equivalent differential equation (15), in terms of Fourier coefficients $W^H(s)$ and $W^K(s)$. Secondly, by ignoring the truncation errors $e_N^H(s, t)$ and $e_N^K(s, t)$ in (15), a new differential equation (16) is derived in terms of $\hat{W}^H(s)$ and $\hat{W}^K(s)$, where only input/state data is involved. Thirdly, it is shown (in Lemma 6) that, $\hat{W}^H(s)$ and $\hat{W}^K(s)$ given by the new differential equation (16) are close to the original Fourier coefficients $W^H(s)$ and $W^K(s)$ on $[0, s_f]$ under suitable conditions, despite of the ignorance of the truncation errors. This implies that the problem of solving model-based PRE (5) (i.e. (7)) on $[0, s_f]$ can be approached by solving the data-based differential equation (16) on $[0, s_f]$. Thus then (16) is solved numerically on $[0, s_f]$, and the numerical solutions with time indexes near 0 are fitted by least squares to obtain two continuous functions $\bar{H}(s)$ and $\bar{K}(s)$. Finally, these two continuous functions are shown (in Theorem 2) to be close to the optimal solutions $H^*(s)$ and $K^*(s)$ in (8), under suitable conditions. This fulfills our goal.

With the above outline in mind, the rest of this section is devoted to presenting the details.

For $M \in \mathbb{Z}_+ \setminus \{0\}$, define $t_j = j\Delta t$, $j = 0, 1, 2, \dots, M$. Δt is the sampling interval. Assuming a measurable locally essentially bounded input $u_0(\cdot) : [0, t_M] \rightarrow \mathbb{R}^m$ is applied to system (1) to collect input/state data for learning, we have

$$\begin{aligned} \frac{dx^T(t)P(s)x(t)}{dt} &= \dot{x}^T(t)P(s)x(t) + x^T(t)P(s)\dot{x}(t) \\ &= x^T(t)H(s, t)x(t) + 2u_0^T(t)R(t)K(s, t)x(t). \end{aligned} \quad (10)$$

Integrating both sides of (10) from t_j to t_{j+1} , by Fact 1, we obtain

$$\begin{aligned} [\tilde{x}(t_{j+1}) - \tilde{x}(t_j)]^T \text{vecs}(P(s)) &= \\ \int_{t_j}^{t_{j+1}} \tilde{x}^T(t) \text{vecs}(H(s, t)) dt &+ \\ \int_{t_j}^{t_{j+1}} (x^T(t) \otimes (2u_0^T(t)R(t))) \text{vec}(K(s, t)) dt. \end{aligned} \quad (11)$$

Using (9), we can organize (11) for $j = 0, 1, 2, \dots, M-1$ into a single linear matrix equation

$$\Theta \begin{bmatrix} \text{vec}(W^H(s)) \\ \text{vec}(W^K(s)) \end{bmatrix} + E_N(s) = \Gamma_{\tilde{x}} \text{vecs}(P(s)), \quad (12)$$

where

$$\begin{aligned} \Theta &= [\Delta_0^T, \Delta_1^T, \dots, \Delta_{M-1}^T]^T, \quad \Delta_j = [I_{F\tilde{x},j}, I_{F\tilde{x}u,j}], \\ \Gamma_{\tilde{x}} &= [\delta_0^T, \delta_1^T, \dots, \delta_{M-1}^T]^T, \quad \delta_j = \tilde{x}^T(t_{j+1}) - \tilde{x}^T(t_j), \\ E_N(s) &= [e_{0,N}(s), e_{1,N}(s), \dots, e_{M-1,N}(s)]^T, \\ I_{F\tilde{x},j} &= \int_{t_j}^{t_{j+1}} F_N^T \otimes \tilde{x}^T dt, \\ I_{F\tilde{x}u,j} &= \int_{t_j}^{t_{j+1}} F_N^T \otimes x^T \otimes (2u_0^T R) dt, \\ e_{j,N}(s) &= \int_{t_j}^{t_{j+1}} \tilde{x}^T(t) e_N^H(s, t) dt \\ &+ \int_{t_j}^{t_{j+1}} (x^T(t) \otimes 2u_0^T(t)) e_N^K(s, t) dt. \end{aligned} \quad (13)$$

The following Assumption is imposed on the matrix Θ .

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

$$\frac{1}{M} \Theta^T \Theta \geq \alpha I_{(n_1+n_2)(2N+1)}. \quad (14)$$

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

Remark 2. Condition (14) appeared in the past literature of ADP [2], [4], [25], which is in the spirit of persistency of excitation (PE) in adaptive control. An exploration noise, such as sum of sinusoidal signals with diverse frequencies, can be added to u_0 , if needed, to satisfy (14).

Remark 3. Notice that by derivations from (10) to (12), input/state data on any time interval of length Δt satisfying (11) can be used to construct (12). To fulfill condition $|x(t)| \leq \beta$, one way is to restart the system at $|x(t_0)| \leq \beta$, whenever the boundedness condition is violated. The choice of t_0 depends on the problem at hand. In rotor-blade system [8], for example, a choice of t_0 can be the moment when the angular position of the rotor is zero.

The following lemma shows that equation (12) is differentiable in s . Its proof can be found in [26, Lemma 4].

Lemma 4. $W^H(\cdot)$, $W^K(\cdot)$, $e_N^H(\cdot, t)$, $e_N^K(\cdot, t)$ and $E_N(\cdot)$ in equations (9) and (12) are continuously differentiable in algorithmic time s .

By Lemma 4, equations (7) and (9), and Assumption 2, taking derivatives with respect to s on both sides of (12), we have

$$\begin{bmatrix} \text{vec}(\dot{W}^H(s)) \\ \text{vec}(\dot{W}^K(s)) \end{bmatrix} = \mathcal{H}(W(s), s) + \mathcal{G}(W(s), s), \quad (15)$$

where

$$\begin{aligned}\mathcal{H}(W(s), s) &= \Theta^\dagger \Gamma_{\bar{x}} [-W^H(s)F_N(s) - \text{vecs}(C^T(s)C(s)) \\ &\quad + \text{vecs}((\text{vec}^{-1}(W^K(s)F_N(s)))^T R(s) \\ &\quad \text{vec}^{-1}(W^K(s)F_N(s)))], \\ \mathcal{G}(W(s), s) &= \Theta^\dagger \left[-\dot{E}_N(s) \right. \\ &\quad + \Gamma_{\bar{x}} (\text{vecs}((\text{vec}^{-1}(W^K(s)F_N(s)))^T R(s) \text{vec}^{-1}(e_N^K(s, s)) \\ &\quad + (\text{vec}^{-1}(e_N^K(s, s)))^T R(s) \text{vec}^{-1}(W^K(s)F_N(s)) \\ &\quad \left. + (\text{vec}^{-1}(e_N^K(s, s)))^T R(s) \text{vec}^{-1}(e_N^K(s, s)) - e_N^H(s, s)) \right], \\ W(s) &= \left[(W^H(s))^T, (W^K(s))^T \right]^T.\end{aligned}$$

In (15), all the terms containing the truncation errors are grouped into $\mathcal{G}(W(s), s)$. If $\mathcal{G}(W(s), s)$ is ignored, we obtain the following differential equation

$$\begin{bmatrix} \text{vec}(\dot{W}^H(s)) \\ \text{vec}(\dot{W}^K(s)) \end{bmatrix} = \mathcal{H}(\hat{W}(s), s), \quad \hat{W}(s_f) = 0. \quad (16)$$

where

$$\hat{W}(s) = \left[(\hat{W}^H(s))^T, (\hat{W}^K(s))^T \right]^T.$$

Notice that no explicit system dynamics information is contained in (16). Thus if $\hat{W}(s)$ is close to $W(s)$, approximate optimal controllers are possible to be found in view of (9), directly from the collected data. Indeed, the following two lemmas show that $\hat{W}(s)$ can be made close to $W(s)$.

Remark 4. Through the derivations from (5) to (16), equation (10) is a critical step in getting rid of the knowledge of system dynamics. The validity of (10) is a result of distinguishing the algorithmic time s from the system evolution time t . This justifies the importance of the separation of time variables.

Assumption 3. The matrix-valued functions $A(\cdot)$ and $B(\cdot)$ are T -periodic and continuously differentiable on $\mathbb{R}$.

Lemma 5. Under Assumptions 1, 2 and 3, for any $-\infty < s' < s_f$:

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

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

Proof. See the Appendix. $\square$

Lemma 6. Let $G = 0$ in PRE (5). Under Assumptions 1, 2 and 3, for any $\epsilon > 0$ and $0 < s_f < \infty$, there exists $\bar{N} > 0$, such that $\forall N > \bar{N}$,

$$\begin{aligned}\sup_{s \in [0, s_f]} \|W^H(s) - \hat{W}^H(s)\| &< \epsilon, \\ \sup_{s \in [0, s_f]} \|W^K(s) - \hat{W}^K(s)\| &< \epsilon.\end{aligned}$$

Proof. See [26, Lemma 6]. $\square$

With Lemma 6, we can solve equation (16) by any convergent numerical method [27, Section 213] backward in time on $[0, s_f]$, to find approximate values of $W(s)$. Supposing in

the numerical method, the step size is h , and at step values $\{s_k\}_{k=0}^L$, $s_k = kh$, $hL = s_f$, the numerical solutions of (16) are computed, denoted by $\{\hat{W}_k^H\}_{k=0}^L$ and $\{\hat{W}_k^K\}_{k=0}^L$. Then we can define

$$\text{vecs}(\hat{H}_k) = \hat{W}_k^H F_N(s_k), \quad \text{vec}(\hat{K}_k) = \hat{W}_k^K F_N(s_k), \quad (17)$$

which are used as approximations to $H(s, s)$ and $K(s, s)$ at $\{s_k\}_{k=0}^L$, respectively. But (17) is discrete, which is not convenient. So we would like to fit the part of the data close to (8) in (17) to get two continuous functions $\bar{H}(\cdot)$ and $\bar{K}(\cdot)$. Supposing 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

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

Assumption 4. Given $N > 0$, there exist $\lfloor L/2 \rfloor > \bar{L}_0 > 2N + 1$ and $\alpha > 0$ (independent of N), such that for all $\lfloor L/2 \rfloor > \bar{L} > \bar{L}_0$, $s_{\bar{L}} > T$,

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

Under Assumption 4, over-determined least squares fittings are implemented on data sets $\{\mathcal{V}, \mathcal{U}\}$ and $\{\mathcal{W}, \mathcal{U}\}$, respectively, to get

$$\begin{aligned}\bar{H}(t) &= \text{vecs}^{-1}(\bar{W}^H F_N(t)), \\ \bar{K}(t) &= \text{vec}^{-1}(\bar{W}^K F_N(t)),\end{aligned} \quad (18)$$

where

$$(\bar{W}^H)^T = \mathcal{U}^\dagger \mathcal{V}, \quad (\bar{W}^K)^T = \mathcal{U}^\dagger \mathcal{W}. \quad (19)$$

We are in a position to state our main result on approximating the optimal solution of the infinite-horizon PLQ problem without the precise knowledge of system dynamics.

Theorem 2. Consider the infinite-horizon PLQ optimal control problem of system (1) with cost (2). Under Assumptions 1, 2, 3 and 4, 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 $\bar{L}$ is chosen to satisfy $s_{\bar{L}} > T$, $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$.

Proof. See the Appendix. $\square$

To sum up, our novel VI-based off-policy ADP algorithm is presented in Algorithm 1.

To give a clearer explanation of the derivations and convergence analysis of Algorithm 1, the relationships of different components in this section are summarized in Figure 1.

Remark 5. In (6) and (9), if $\lim_{s \rightarrow -\infty} P(s) - P^*(s) = 0$, $W(s)$ will converge to a unique periodic orbit. In view of Lemma 6, $\hat{W}(s)$ will be close to that periodic orbit. This suggests a heuristic procedure for the choice of parameters in Algorithm 1:

Algorithm 1: VI-based off-policy ADP

Choose $\Delta t > 0$, large enough $M > 0$, $N > 0$, $s_f > 0$, and small enough $h > 0$.
 Apply $u_0(\cdot) : [0, t_M] \rightarrow \mathbb{R}^m$ (with exploration noise) to system (1), collect the input/state data.
 Construct the data matrices Θ and $\Gamma_{\tilde{x}}$.
 Solve numerically (16) backward in time on $[0, s_f]$.
 Choose $\bar{L}$ satisfying $s_{\bar{L}} > T$, $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$.

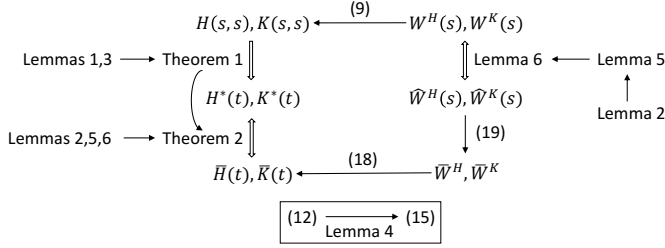


Fig. 1. Overview of derivations and convergence analysis of Algorithm 1.

- (1) Choose some $\Delta t > 0$, $N > 0$.
- (2) In view of Assumption 2, Θ should be at least full column rank. Thus $M \gg (n_1 + n_2)(2N + 1)$. Then $t_M = M\Delta t$.
- (3) Find an s_f , such that $\|\tilde{W}(s)\|$ given by (16) is almost periodic near $s = 0$. If such an s_f can not be found for large values of s_f , increase N and go back to step (2).
- (4) Set $h > 0$ such that $\lfloor \frac{s_f}{h} \rfloor \gg 2N + 1$.
- (5) Choose $\bar{L}$ satisfying $s_{\bar{L}} > T$, $\lfloor L/2 \rfloor > \bar{L} > 2N + 1$.

IV. SIMULATION RESULTS

Consider the triple inverted pendulum with a periodically varying load [28], modeled by system (1) with

$$A(t) = \begin{bmatrix} 0_3 & I_3 \\ A_{21}(t) & A_{22} \end{bmatrix}, B(t) = \begin{bmatrix} 0_3 \\ B_2 \end{bmatrix} \quad (20)$$

$$A_{21}(t) = \begin{bmatrix} (\gamma - 3) & (3 - \gamma) & -1 \\ (4 - \gamma) & 2(\gamma - 3) & (3 - \gamma) \\ -1 & (4 - \gamma) & (\gamma - 3) \end{bmatrix},$$

$$A_{22} = 0.5 \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, B_2 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix},$$

and states $x(t) = [\eta_1(t), \eta_2(t), \eta_3(t), \dot{\eta}_1(t), \dot{\eta}_2(t), \dot{\eta}_3(t)]^T$, inputs $u(t) = [u_1(t), u_2(t), u_3(t)]^T$, where $\gamma = 1 + 2\cos(t)$. For each $i = 1, 2, 3$, $\eta_i(\cdot)$ is the angle of the i th pendulum with respect to the vertical line; $\dot{\eta}_i(\cdot)$ is the corresponding angular velocity; $u_i(\cdot)$ is the control torque applied at the bottom of the i th pendulum.

If the system matrices in (20) are known, (5) can be solved to obtain a suboptimal controller, which is referred as model-based PLQ (MBPLQ) controller. However, the actual system

need not evolve exactly as (20). Suppose an extra periodically varying disturbance exists, which changes $A_{21}(t)$ in (20) to

$$\tilde{A}_{21}(t) = A_{21}(t) + \zeta(1 + \sin(3t))I_3,$$

where $\zeta > 0$ controls the magnitude of the disturbance. Algorithm 1 is applied to (20) with $\tilde{A}_{21}(t)$, to obtain ADP controllers. The simulation results with different choices of controllers, N , M , s_f and ζ are summarized in Table I. It is easy to check that Assumptions 1 and 3 are satisfied by system (20) and the choice of $C(\cdot)$ in Table I. In the simulation trials of ADP controllers, the following initial controller

$$[u_0(t)]_i = 0.2 * \sum_{j=1}^{500} \sin(\omega_{i,j}t), \quad i = 1, 2, 3$$

is applied to the system over time interval $[0, t_M]$ to collect data, where $\omega_{i,j}$ is drawn from a uniform distribution over $[-500, 500]$. To guarantee the collected data is bounded, we reset $x(t_j)$ to the initial state $x(t_0) = 0$, $t_0 = 0$ as long as $|x(t_j)| > 10$. In Table I, trial 1 yields the best result. Loss of stability and/or optimality occurs when one of the parameters N , M , s_f is not large enough, as shown in trials 2-6. Although MBPLQ controller is robust to extra disturbance with small magnitude in Trial 7, it is not robust to extra disturbance of large magnitude in trial 8, in contrast to the ADP controller. The different control gains generated by Algorithm 1 in trial 1 are compared with the optimal gains in Fig. 2. These simulation results demonstrate the viability of the theoretical results in the previous section.

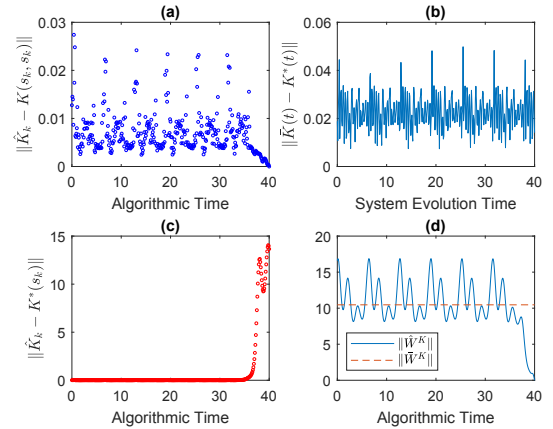


Fig. 2. Comparison of different control gains. $\bar{K}(\cdot)$ is the output of Algorithm 1; K_k is defined in (17); $K(\cdot)$ is generated by model-based VI (5); $K^*(\cdot)$ is the optimal control gain; $\tilde{W}^H$ is generated by (16); $\tilde{W}^K$ is given in (19).

V. CONCLUSION

An innovative VI-based ADP algorithm is proposed for CTLP systems in this paper, such that learning-based suboptimal controllers can be obtained from real-time data without the exact knowledge of system dynamics. The proposed algorithm does not assume an initial stabilizing controller, and is off-policy, which is easy-to-use and data-efficient. Convergence analysis is developed for the presented VI-based ADP algorithm. It is shown that, under mild conditions, the proposed

TABLE I
SIMULATION RESULTS OF THE TRIPLE INVERTED PENDULUM CONTROL UNDER DIFFERENT SETTINGS

Trials	Controller	N	M	s_f	ζ	Number of resetting	Stability	$\max_t \ \bar{K}(t) - K^*(t)\ $
1	ADP	6	800	40	1	24	Yes	0.0498
2	ADP	3	800	40	1	24	Yes	0.8784
3	ADP	1	800	40	1	25	No	64.9159
4	ADP	6	800	12	1	25	Yes	8.1861
5	ADP	6	800	8	1	24	No	146.4805
6	ADP	6	400	40	1	13	No	151.3061
7	MBPLQ	-	-	-	0.1	-	Yes	-
8	MBPLQ	-	-	-	1	-	No	-
Common parameters	$\Delta t = 0.2, h = 0.1, C = I_6, R = I_3, \bar{L} = \lfloor s_f / (3h) \rfloor$.							

algorithm generates a sequence of suboptimal controllers converging uniformly to the optimal solutions. In addition, the proposed adaptive optimal control method is successfully tested in a benchmark example of controlling a triple inverted pendulum with a periodically varying load. Our future work can be directed at extending the proposed methodology to the optimal output regulation problem as shown in [25].

ACKNOWLEDGMENT

It is a pleasure to thank the Associate Editor, anonymous reviewers and Tao Bian for their helpful and constructive comments.

APPENDIX

A. Proof of Theorem 1

The proof is divided into three cases: Case 1) $G > 0$; Case 2) $G = 0$; Case 3) $G \geq 0, G \neq 0$ with at least one zero eigenvalue.

Case 1): In this case, by Lemma 1 and [22, Corollary], we immediately obtain $\lim_{t \rightarrow -\infty} (P(t; t_f, G) - P^*(t)) = \lim_{t \rightarrow -\infty} (P(t; t_f, G) - P_S(t)) = 0$.

Case 2): In this case, for any $t \leq \tau_1 \leq \tau_2$ and any $\xi \in \mathbb{R}^n$, define $u^{(2)}(s) \triangleq -R(s)^{-1}B(s)^T P(s; \tau_2, 0)x^{(2)}(s)$, where $x^{(2)}(s), t \leq s \leq \tau_2$, is the solution of system (1) under control $u^{(2)}(s)$. By definition,

$$\begin{aligned} \xi^T P(t; \tau_1, 0)\xi &= \min_u V(t, \tau_1, \xi, u, 0) \leq V(t, \tau_1, \xi, u^{(2)}, 0) \\ &\leq V(t, \tau_2, \xi, u^{(2)}, 0) = \xi^T P(t; \tau_2, 0)\xi. \end{aligned}$$

Hence $P(t; \tau, 0)$ is nondecreasing as $\tau \rightarrow \infty$, when $t \leq \tau$. By Lemma 1, $P^*(t)$ is stabilizing solution of the PRE, which means $\xi^T P^*(t)\xi < \infty$, for any fixed $\xi \in \mathbb{R}^n$ and t . Again, by definition,

$$\begin{aligned} \xi^T P(t; \tau, 0)\xi &\leq V(t, \tau, \xi, u^*, 0) \\ &\leq \lim_{\bar{\tau} \rightarrow \infty} V(t, \bar{\tau}, \xi, u^*, 0) = \xi^T P^*(t)\xi < \infty. \end{aligned}$$

Therefore when $t \leq \tau$, $P(t; \tau, 0)$ is nondecreasing as $\tau \rightarrow \infty$, and bounded from the above. By the monotone convergence theorem, $\bar{P}(t; 0) := \lim_{\tau \rightarrow \infty} P(t; \tau, 0)$ exists. By periodicity,

$$P(t + T; \tau + T, 0) = P(t; \tau, 0).$$

Then we have

$$\bar{P}(t + T; 0) = \lim_{\tau \rightarrow \infty} P(t + T; \tau, 0) = \lim_{\tau \rightarrow \infty} P(t; \tau, 0) = \bar{P}(t; 0).$$

Furthermore, $P(t; \tau, 0)$ is symmetric and positive semidefinite for all $t \leq \tau$. This implies that $\bar{P}(t; 0)$ is a SPPS solution of the PRE. Due to the uniqueness of the SPPS solution, we have $\bar{P}(t; 0) = P^*(t)$. In $P(t; \tau, 0)$, $\tau \rightarrow \infty$ is equivalent to $t \rightarrow -\infty$, thus (4) holds when $G = 0$.

Case 3): In this case, we can always find a $\bar{G} > 0$, such that $0 \leq G \leq \bar{G}$, then (4) follows by Lemma 3 and the squeeze theorem [29, Theorem 3.3.6].

B. Proof of Lemma 5

For the proof of 1), please see [26, Lemma 5].

For 2), by definition of $\mathcal{G}(W(s), s)$, Lemma 2 and Assumption 2, for any $\epsilon_0 > 0$, there exists $\bar{N}_0 > 0$, such that $\forall N > \bar{N}_0$,

$$\begin{aligned} \sup_{s \in [s', s_f]} |\mathcal{G}(W(s), s)| &< M \|\Theta^\dagger\|_2 (1 + 2\beta^4) \epsilon_0 \\ &= \frac{M(1 + 2\beta^4)}{\sigma_{\min}(\Theta)} \epsilon_0 \leq \frac{1 + 2\beta^4}{\alpha} \epsilon_0. \end{aligned}$$

Note that α and β are independent of N , the proof is completed.

C. Proof of Theorem 2

For convenience, only the convergence of $\bar{H}$ to H^* is proved. The proof on the convergence of $\bar{K}$ to K^* is analogous. Set $G = 0$. For any $\epsilon_0 > 0$, by Lemma 5, there exists $\bar{N}_1 > 0$, such that $\forall N > \bar{N}_1$,

$$\sup_{s \in [0, s_f]} \|\text{vecs}^{-1}(W^H(s)F_N(s)) - H(s, s)\| < \frac{\epsilon_0}{4}.$$

By Lemma 6, there exists $\bar{N}_2 > 0$, such that $\forall N > \bar{N}_2$,

$$\sup_{s \in [0, s_f]} |\hat{W}^H(s)F_N(s) - W^H(s)F_N(s)| < \frac{\epsilon_0}{4}.$$

According to [27, Theorem 213B], we can choose a $\bar{h}_0$, such that for all step size $0 < h < \bar{h}_0$,

$$\sup_{k \in \{1, 2, \dots, L\}} |\text{vecs}(\hat{H}_k) - \hat{W}^H(s_k)F_N(s_k)| < \frac{\epsilon_0}{4}.$$

From Theorem 1, there exists a $\bar{s}_{f,0} > 0$, such that $\forall s_f > \bar{s}_{f,0}$,

$$\sup_{s \in [0, s_L]} \|H(s, s) - H^*(s)\| < \frac{\epsilon_0}{4}.$$

Set $\bar{N}_0 = \max\{\bar{N}_1, \bar{N}_2\}$. Applying the triangle inequality to the above inequalities and by Fact 1, we obtain that, $\forall s_f > \bar{s}_f, 0, N > \bar{N}_0$, any $0 < h < \bar{h}_0$, there are

$$\sup_{k \in \{1, 2, \dots, \bar{L}\}} \|\hat{H}_k - H^*(s_k)\| < \epsilon_0. \quad (21)$$

Now define

$$\mathcal{V}^* = [\text{vecs}(H^*(s_0)), \text{vecs}(H^*(s_1)), \dots, \text{vecs}(H^*(s_{\bar{L}}))]^T,$$

and express $H^*(s)$ by their Fourier series,

$$\text{vecs}(H^*(s)) = W_1^* F_N(s) + e_{1,N}^*(s).$$

In view of (19) and Assumption 4, we have

$$(W_1^*)^T = \mathcal{U}^\dagger(\mathcal{V}^* + E_{1,N}^*).$$

Thus we obtain

$$(W_1^* - \bar{W}^H)^T = \mathcal{U}^\dagger[(\mathcal{V}^* - \mathcal{V}) + E_{1,N}^*].$$

By the properties of matrix norms, we have

$$\begin{aligned} \frac{1}{\sqrt{n_1}} \|(W_1^* - \bar{W}^H)^T\| &\leq \|(W_1^* - \bar{W}^H)^T\|_2 \\ &\leq \|\mathcal{U}^\dagger\|_2 \|(\mathcal{V}^* - \mathcal{V}) + E_{1,N}^*\|. \end{aligned}$$

By (21), Lemma 2 and Assumption 4, for any $\epsilon_3 > 0$, there exist large enough $s_{f,3}$, N_3 , and small enough h_3 , such that

$$\|W_1^* - \bar{W}^H\| < \frac{\sqrt{n_1} \bar{L}}{\sigma_{\min}(\mathcal{U})} \epsilon_3 < \frac{\sqrt{n_1}}{\alpha} \epsilon_3.$$

As a result of the boundedness of $F_N(\cdot)$, 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 $0 < h < \bar{h}$,

$$\begin{aligned} \sup_{t \in \mathbb{R}} \|\bar{H}(t) - \text{vecs}^{-1}(W_1^* F_N(t))\| &< \frac{\epsilon}{2}, \\ \sup_{t \in \mathbb{R}} \|\text{vecs}^{-1}(W_1^* F_N(t)) - H^*(t)\| &< \frac{\epsilon}{2}. \end{aligned}$$

Again, using the triangle inequality completes the proof.

REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, Massachusetts: MIT Press, 2018.
- [2] Y. Jiang and Z. P. Jiang, *Robust Adaptive Dynamic Programming*. Hoboken, New Jersey: Wiley, 2017.
- [3] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec 2012.
- [4] F. L. Lewis and D. Liu, Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, New Jersey: Wiley-IEEE Press, 2013.
- [5] J. Fong, Y. Tan, V. Crocher, D. Oetomo, and I. Mareels, "Dual-loop iterative optimal control for the finite horizon LQR problem with unknown dynamics," *Syst. Control Lett.*, vol. 111, pp. 49–57, 2018.
- [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 Technol.*, vol. 17, no. 1, pp. 73–84, 2019.
- [7] I. Houtzager, J. van Wingerden, and M. Verhaegen, "Wind turbine load reduction by rejecting the periodic load disturbances," *Wind Energy*, vol. 16, no. 2, pp. 235–256, 2013.
- [8] J. Camino and I. Santos, "A periodic linear-quadratic controller for suppressing rotor-blade vibration," *J. Vib. Control*, vol. 25, no. 17, pp. 2351–2364, 2019.
- [9] R. Gondhalekar, F. Oldewurtel, and C. N. Jones, "Least-restrictive robust periodic model predictive control applied to room temperature regulation," *Automatica*, vol. 49, no. 9, pp. 2760–2766, 2013.
- [10] D. Limon, M. Pereira, D. M. De La Peña, T. Alamo, and J. M. Grosso, "Single-layer economic model predictive control for periodic operation," *J. Process Control*, vol. 24, no. 8, pp. 1207–1224, 2014.
- [11] M. Pereira, D. Limon, D. M. de la Peña, L. Valverde, and T. Alamo, "Periodic economic control of a nonisolated microgrid," *IEEE Trans. Ind. Electron.*, vol. 62, no. 8, pp. 5247–5255, 2015.
- [12] N. Karlsson, "Control of periodic systems in online advertising," in *Proc. IEEE 57th Conf. Decis. Control*. Miami, FL, USA: IEEE, 2018, pp. 5928–5933.
- [13] M. Farkas, *Periodic Motions*. New York: Springer Science & Business Media, 2013, vol. 104.
- [14] T. Bian and Z. P. Jiang, "Value iteration, adaptive dynamic programming, and optimal control of nonlinear systems," in *Proc. IEEE 55th Conf. Decis. Control*, Las Vegas, NV, USA, 2016, pp. 3375–3380.
- [15] G. Hewer, "Periodicity, detectability and the matrix Riccati equation," *SIAM J. Control Optim.*, vol. 13, no. 6, pp. 1235–1251, 1975.
- [16] S. Bittanti, G. Guardabassi, C. Maffezzoni, and L. Silverman, "Periodic systems: Controllability and the matrix Riccati equation," *SIAM J. Control Optim.*, vol. 16, no. 1, pp. 37–40, 1978.
- [17] 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, DE: Springer, 1991, ch. 6, pp. 127–162.
- [18] Z. Zhang and A. Serrani, "Adaptive robust output regulation of uncertain linear periodic systems," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 266–278, Feb 2009.
- [19] J.-X. Xu, "A new periodic adaptive control approach for time-varying parameters with known periodicity," *IEEE Trans. Autom. Control*, vol. 49, no. 4, pp. 579–583, April 2004.
- [20] K. S. Narendra and K. Esfandiari, "Adaptive identification and control of linear periodic systems using second-level adaptation," *Int. J. Adapt. Control Signal Process.*, vol. 33, no. 6, pp. 956–971, 2019.
- [21] S. Bittanti, "Deterministic and stochastic linear periodic systems," in *Time Series and Linear Systems*, S. Bittanti, Ed. Berlin, DE: Springer, 1986, ch. 5, pp. 141–182.
- [22] G. De Nicolao, "On the convergence to the strong solution of periodic Riccati equations," *Int. J. Control*, vol. 56, no. 1, pp. 87–97, 1992.
- [23] D. Anton, *A First Course in Harmonic Analysis*, 2nd ed. New York: Springer, 2005.
- [24] L. Daniel, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton, New Jersey: Princeton University Press, 2012.
- [25] W. Gao and Z. P. Jiang, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2614–2624, June 2018.
- [26] B. Pang and Z. P. Jiang, "Adaptive optimal control of linear periodic systems: An off-policy value iteration approach," *arXiv preprint arXiv:1901.08650*, 2019.
- [27] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*. Chichester, U.K.: Wiley, 2016.
- [28] S. C. Sinha and P. Joseph, "Control of general dynamic systems with periodically varying parameters via Lyapunov-Floquet transformation," *J. Dyn. Sys., Meas., Control*, vol. 116, no. 4, pp. 2439–2447, 1994.
- [29] H. H. Sohrab, *Basic Real Analysis*, 2nd ed. New York: Springer, 2014.