Control Theory Tech, Vol. 17, No. 1, pp. 73–84, February 2019

Control Theory and Technology

http://link.springer.com/journal/11768

Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems

Bo PANG [1†], Tao BIAN [2], Zhong-Ping JIANG [1]

1.*Control and Networks (CAN) Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, U.S.A.*

2.*Bank of America Merrill Lynch, One Bryant Park, New York, NY 10036, U.S.A.*

Received 15 August 2018; revised 24 October 2018; accepted 26 October 2018

Abstract

This paper studies data-driven learning-based methods for the finite-horizon optimal control of linear time-varying discrete-time systems. First, a novel finite-horizon Policy Iteration (PI) method for linear time-varying discrete-time systems is presented. Its connections with existing infinite-horizon PI methods are discussed. Then, both data-driven off-policy PI and Value Iteration (VI) algorithms are derived to find approximate optimal controllers when the system dynamics is completely unknown. Under mild conditions, the proposed data-driven off-policy algorithms converge to the optimal solution. Finally, the effectiveness and feasibility of the developed methods are validated by a practical example of spacecraft attitude control.

Keywords: Optimal control, time-varying system, adaptive dynamic programming, policy iteration (PI), value iteration (VI)

DOI https://doi.org/10.1007/s11768-019-8168-8

1 Introduction

Bellman's dynamic programming (DP) [1] has been successful in solving sequential decision making problems arising from areas ranging from engineering to economics. Despite a powerful theoretical tool in optimal control of dynamical systems [2, 3], besides the well-known "Curse of Dimensionality", the original DP is also haunted by the "Curse of Modeling" [4], i.e., the exact knowledge of the dynamical process under consideration is required. Reinforcement learning (RL) [5, 6] and adaptive dynamic programing (ADP) [7–12] are promising to deal with this problem. RL and ADP find approximate optimal laws by iteratively utilizing the data collected from the interactions between the controller and the plant, so that an explicit plant model is not needed.

†Corresponding author.

E-mail: bo.pang@nyu.edu. Tel.: +1 3472042834.

The work of B. Pang and Z.-P. Jiang has been supported in part by the National Science Foundation (No. ECCS-1501044).

Over the past decade, many papers have been devoted to this routine, for systems characterized by linear or nonlinear, differential or difference equations. For example, adaptive optimal controllers were proposed without the knowledge of system dynamics, by using policy iteration (PI) [7, 8, 13–15] or value iteration (VI) [16, 17], and many references therein.

However, most existing results focus only on the infinite-horizon optimal control problem for time-invariant systems. There are relatively few studies on the finite-horizon optimal control problem for time-varying systems. Although finite-horizon approximate optimal controllers were derived for linear systems in [18, 19], and nonlinear systems in [20–24], the authors of these papers assumed either the full knowledge of system dynamics, or time-invariant system parameters. Recently, several methods that do not require the exact knowledge of system dynamics have also been proposed for the finite-horizon control of time-varying systems. Extremum seeking techniques were applied to find approximate finite-horizon optimal open-loop control sequences for linear time-varying discrete-time (LTVDT) systems in [25, 26]. A dual-loop iteration algorithm was devised to obtain approximate optimal control for linear time-varying continuous-time (LTVCT) systems in [27]. Different from the time-invariant and infinite-horizon case where the optimal controller is stationary, in the time-varying and finite-horizon case the optimal controller is nonstationary. This brings new challenges to the design of data-driven, non-model based optimal controllers when the precise information of time-varying system dynamics is not available.

This paper considers the finite-horizon optimal control problem without the knowledge of system dynamics for LTVDT systems. Firstly, a novel finite-horizon PI method for LTVDT systems is presented. On one hand, the proposed finite-horizon PI method can be seen as the counterpart to the existing infinite-horizon PI methods, which may be found in [28] for LTVDT systems, in [29] for LTVCT systems, in [30] for linear time-invariant discrete-time (LTIDT) systems, and in [31] for linear time-invariant continuous-time (LTICT) systems, respectively. On the other hand, it is parallel to the finite-horizon PI for LTVCT systems in [32] (Theorem 8). Secondly, we prove that in the time-invariant case, as the time horizon goes to infinity, the finite-horizon PI method reduces to the infinite-horizon PI method for LTIDT systems. Thirdly, we propose data-driven off-policy finite-horizon PI and VI algorithms to find approximate optimal controllers when the system dynamics is unknown. The proposed data-driven algorithms are off-policy, by contrast, the methods in [25–27] were on-policy. In addition, the work [25–27] only considered special cases of linear systems or lacked convergence analysis. Finally, we simulate the proposed methods on a practical example of spacecraft attitude control with time-varying dynamics. The simulation results demonstrate the effectiveness and feasibility of our data-driven finite-horizon PI and VI algorithms.

The rest of this paper is organized as follows: Section 2 introduces the problem formulation and necessary preliminaries; Section 3 first presents the finite-horizon PI method, and then its connections with infinite-horizon PI method under certain conditions are revealed; Section 4 derives the data-driven finite-horizon PI and VI algorithms; in Section 5, the application of the proposed methods to spacecraft attitude control is provided; Section 6 concludes the whole paper.

Notations Throughout this paper, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{Z}_+$ denotes the set of nonnegative integers. $k \in \mathbb{Z}_+$ denotes the discrete time instance. $X(k)$ is denoted as X_k for short and the time index is always the first subscript if it has multiple subscripts, e.g., X_{k,i_1,i_2}. $\otimes$ is the Kronecker product operator. For matrix $A \in \mathbb{R}^{n\times m}$, $\mathrm{vec}(A) = [a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, \ldots, a_m^{\mathrm{T}}]^{\mathrm{T}}$, where a_i is the ith column of A. For symmetric matrix $B \in \mathbb{R}^{m\times m}$,

$$\mathrm{vecs}(B) = [b_{11}, 2b_{12}, \ldots, 2b_{1m}, b_{22}, 2b_{23}, \ldots, 2b_{m-1,m}, b_{m,m}]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}m(m+1)}.$$

For vector $v \in \mathbb{R}^n$,

$$\tilde{v} = [v_1^2, v_1v_2, \ldots, v_1v_n, v_2^2, v_2v_3, \ldots, v_{n-1}v_n, v_n^2]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}n(n+1)}.$$

$|\cdot|$ is the Euclidean norm for vectors and $\|\cdot\|$ represents the induced matrix norm for matrices.

2 Problem formulation and preliminaries

Consider the following linear time-varying discrete-time system:

$$x_{k+1} = A_k x_k + B_k u_k, \tag{1}$$

where $k \in [k_0, N)$ is the discrete time instant, $x_k \in \mathbb{R}^n$ is the system state, $u_k \in \mathbb{R}^m$ is the control input, $A_k \in \mathbb{R}^{n\times n}$ and $B_k \in \mathbb{R}^{n\times m}$ are time-dependent system matrices.

We are interested in finding a sequence of control inputs $\{u_k\}_{k=k_0}^{N-1}$, to minimize the following cost function with penalty on the final state

$$J(k_0, u, x_{k_0}) = x_N^{\mathrm{T}} F x_N + \sum_{j=k_0}^{N-1} (x_j^{\mathrm{T}} Q_j x_j + u_j^{\mathrm{T}} R_j u_j), \quad (2)$$

where $Q_j \in \mathbb{R}^{n\times n}$, $R_j \in \mathbb{R}^{m\times m}$, $F \in \mathbb{R}^{n\times n}$ are symmetric weighting matrices satisfying $Q_j \geqslant 0$, $R_j > 0$, and $F \geqslant 0$.

When the system matrices A_k and B_k are precisely known for all $k \in [k_0, N)$, this is the well-known finite-horizon linear quadratic regulator (LQR) problem (see [2, Pages 110–112]). The optimal control input is given by

$$u_k^* = -L_k^* x_k, \quad (3)$$

where the gain matrices L_k^* are given by the equation

$$L_k^* = (R_k + B_k^{\mathrm{T}} P_{k+1}^* B_k)^{-1} B_k^{\mathrm{T}} P_{k+1}^* A_k \quad (4)$$

and the optimal cost is given by

$$J^*(k, u^*, x_k) = x_k^{\mathrm{T}} P_k^* x_k$$

with the symmetric and positive semidefinite matrices P_k^* the solutions to the discrete-time Riccati equation

$$\begin{aligned} P_k^* = {} & Q_k + A_k^{\mathrm{T}} P_{k+1}^* A_k \\ & - A_k^{\mathrm{T}} P_{k+1}^* B_k (R_k + B_k^{\mathrm{T}} P_{k+1}^* B_k)^{-1} B_k^{\mathrm{T}} P_{k+1}^* A_k \end{aligned} \quad (5)$$

with the terminal condition $P_N^* = F$. Obviously, given a fixed matrix F, the matrix series $\{P_k^*\}_{k=k_0}^{N}$ are uniquely determined.

Remark 1 Note that (5) shares similar features with the VI for infinite-horizon discrete-time LQR problem [33, Theorem 17.5.3]. In Section 4.2, we will develop a VI-based finite-horizon ADP based on (5).

Remark 2 For the optimal control problem considered here, the optimal control u^* always exists, no matter whether system (1) is controllable, stabilizable or not. This fact can be verified by a straightforward application of DP to system (1) and cost function (2) (see [2, Pages 110–112]).

When A_k and B_k are unknown, due to the difference between (5) and the algebraic Riccati equation (ARE) in infinite-horizon control problems (see [2, Page 113]), existing infinite-horizon data-driven methods (see, e.g., [34, 35]) cannot be used here directly. Moreover, the methods in [25] and [26] are on-policy and only applied to linear systems with scalar input. By exploiting the property of the cost function, new off-policy methods that find the optimal control inputs without the knowledge of system matrices can be derived.

Now, applying a sequence $u^{(L)}$ of control inputs $u_k = -L_k x_k$, $k \in [k_0, N)$ with arbitrary gain matrices $\{L_k\}_{k=k_0}^{N-1}$ to system (1), we have

$$x_{k+1} = A_k^{(L)} x_k, \quad (6)$$

where $A_k^{(L)} = A_k - B_k L_k$. For any $k_1, k_2 \in [k_0, N]$, $k_1 > k_2$, by the definition of state-transition matrix Φ_L associated with the closed-loop system (6), we have

$$\begin{cases} x_{k_1} = \Phi_L(k_1, k_2) x_{k_2}, \\ \Phi_L(k_1, k_2) = A_{k_1-1}^{(L)} A_{k_1-2}^{(L)} \cdots A_{k_2}^{(L)}, \\ \Phi_L(k_1, k_1) = I. \end{cases} \quad (7)$$

Then the cost at time instance k, $k \in [k_0, N]$, can be rewritten as

$$J(k, u^{(L)}, x_k) = x_k^{\mathrm{T}} V_k^{(L)} x_k, \quad (8)$$

where $V_k^{(L)}$ is the value matrix associated with the gain matrices $\{L_k\}_{k=k_0}^{N-1}$:

$$\begin{aligned} V_k^{(L)} = {} & \Phi_L^{\mathrm{T}}(N, k) F \Phi_L(N, k) \\ & + \sum_{j=k}^{N-1} [\Phi_L^{\mathrm{T}}(j, k)(Q_j + L_j^{\mathrm{T}} R_j L_j) \Phi_L(j, k)]. \end{aligned} \quad (9)$$

Note that matrix $V_k^{(L)}$ is symmetric and positive semidefinite, since $F \geqslant 0$, $Q_j \geqslant 0$, $R_j > 0$. Furthermore, $V_N^{(L)} = F$.

By using (7) with $k_1 = k + 1$, $k_2 = k$ and (9), the following Lyapunov equation can be obtained:

$$V_k^{(L)} = (A_k^{(L)})^{\mathrm{T}} V_{k+1}^{(L)} A_k^{(L)} + Q_k + L_k^{\mathrm{T}} R_k L_k. \quad (10)$$

3 Finite-horizon policy iteration for linear time-varying discrete-time systems

For convenience, next, we use $A_k^{(i)}$ to denote the closed-loop system matrix at time instance k in ith iteration, i.e.,

$$A_k^{(i)} = A_k - B_k L_k^{(i)}. \quad (11)$$

Similarly, we use $V_k^{(i)}$ to denote the value matrix defined in (9) with respect to $\{L_k^{(i)}\}_{k=k_0}^{N-1}$, the state feedback gains in ith iteration.

3.1 Finite-horizon PI for LTVDT systems

Theorem 1 For system (1), consider the finite-horizon linear quadratic regulator problem with respect to the cost function (2).

1) Choose arbitrary initial gain matrices $\{L_k^{(0)}\}_{k=k_0}^{N-1}$, and let $i = 0$.

2) (Policy evaluation) Solve for $\{V_k^{(i)}\}_{k=k_0}^{N-1}$, by using the following Lyapunov equations with $V_N^{(i)} = F$:

$$V_k^{(i)} = (A_k^{(i)})^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i)} + Q_k + (L_k^{(i)})^{\mathrm{T}} R_k L_k^{(i)}. \tag{12}$$

3) (Policy improvement) Solve for $\{L_k^{(i+1)}\}_{k=k_0}^{N-1}$, by using the following equations:

$$L_k^{(i+1)} = (R_k + B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k)^{-1} B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k. \tag{13}$$

4) Let $i = i + 1$, and go to Step 2).

Then for all $k \in [k_0, N]$, it holds:

i) $P_k^* \leqslant V_k^{(i+1)} \leqslant V_k^{(i)}$, $\forall i \in \mathbb{Z}_+$.

ii) $\lim\limits_{i\to\infty} V_k^{(i)} = P_k^*$.

Proof i) Note that

$$\begin{aligned} A_k^{(i)} &= A_k^{(i+1)} - B_k(L_k^{(i)} - L_k^{(i+1)}),\\ L_k^{(i)} &= (L_k^{(i)} - L_k^{(i+1)}) + L_k^{(i+1)}. \end{aligned}$$

Substituting the above equations into (12) gives

$$\begin{aligned} V_k^{(i)} = {} & V_k^{(i+1)} + (A_k^{(i+1)})^{\mathrm{T}}(V_{k+1}^{(i)} - V_{k+1}^{(i+1)})A_k^{(i+1)}\\ & + (L_k^{(i)} - L_k^{(i+1)})^{\mathrm{T}}(R_k + B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k)(L_k^{(i)} - L_k^{(i+1)})\\ & + (L_k^{(i)} - L_k^{(i+1)})^{\mathrm{T}}(R_k L_k^{(i+1)} - B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i+1)})\\ & + (R_k L_k^{(i+1)} - B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i+1)})^{\mathrm{T}}(L_k^{(i)} - L_k^{(i+1)}). \end{aligned} \tag{14}$$

Note that, by (13), there is

$$\begin{aligned} 0 &= (R_k + B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k) L_k^{(i+1)} - B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k\\ &= (R_k L_k^{(i+1)} - B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i+1)}). \end{aligned}$$

Thus, the last two terms in expression (14) vanish. Then we have

$$\begin{aligned} & V_k^{(i)} - V_k^{(i+1)}\\ &= (A_k^{(i+1)})^{\mathrm{T}}(V_{k+1}^{(i)} - V_{k+1}^{(i+1)})A_k^{(i+1)}\\ &\quad + (L_k^{(i)} - L_k^{(i+1)})^{\mathrm{T}}(R_k + B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k)(L_k^{(i)} - L_k^{(i+1)}). \end{aligned} \tag{15}$$

Note that $(R_k + B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k)$ is always positive definite. Since $V_N^{(i)} = V_N^{(i+1)} = F$, by induction, $V_k^{(i)} \geqslant V_k^{(i+1)}$ for all i and k.

Since P_k^* is the value matrix associated with the optimal control gain matrices $\{L_k^*\}_{k=k_0}^{N-1}$, $x_k^{\mathrm{T}} P_k^* x_k \leqslant x_k^{\mathrm{T}} V_k^{(i+1)} x_k$, $\forall x_k \in \mathbb{R}^n$. Hence, $P_k^* \leqslant V_k^{(i+1)}$.

ii) By i), $V_k^{(i)} \geqslant V_k^{(i+1)} \geqslant P_k^*$. This implies sequence $\{V_k^{(i)}\}_{i=0}^{\infty}$ is nonincreasing and bounded from below. By a theorem on the convergence of a monotone sequence of self-adjoint operators (See [36, Pages 189–190]), $\lim\limits_{i\to\infty} V_k^{(i)}$ exists. Letting $i \to \infty$ in (12) and (13), we have by continuity

$$\begin{aligned} V_k^{(\infty)} = {} & A_k^{\mathrm{T}} V_{k+1}^{(\infty)} A_k + Q_k - (L_k^{(\infty)})^{\mathrm{T}} B_k^{\mathrm{T}} V_{k+1}^{(\infty)} A_k\\ & - A_k^{\mathrm{T}} V_{k+1}^{(\infty)} B_k L_k^{(\infty)}\\ & + (L_k^{(\infty)})^{\mathrm{T}}(B_k^{\mathrm{T}} V_{k+1}^{(\infty)} B_k + R_k) L_k^{(\infty)},\\ L_k^{(\infty)} = {} & (R_k + B_k^{\mathrm{T}} V_{k+1}^{(\infty)} B_k)^{-1} B_k^{\mathrm{T}} V_{k+1}^{(\infty)} A_k. \end{aligned}$$

Eliminating $L_k^{(\infty)}$ from the above equations yields

$$\begin{aligned} V_k^{(\infty)} = {} & Q_k + A_k^{\mathrm{T}} V_{k+1}^{(\infty)} A_k\\ & - A_k^{\mathrm{T}} V_{k+1}^{(\infty)} B_k (R_k + B_k^{\mathrm{T}} V_{k+1}^{(\infty)} B_k)^{-1} B_k^{\mathrm{T}} V_{k+1}^{(\infty)} A_k. \end{aligned}$$

This is exactly the Riccati equation. Due to $P_N^* = V_N^{(\infty)} = F$ and the uniqueness of solution to Riccati equation, $P_k^* = V_k^{(\infty)}$ follows. This completes the proof. □

Remark 3 Most existing numerical finite-horizon optimal control methods were developed for continuous-time models; see [18, 22, 23], to name a few. It is not clear how to adapt their model-based methods into data-driven algorithms. On the basis of Theorem 1, one can derive corresponding data-driven off-policy PI algorithm (see Section 4).

3.2 Connections with infinite-horizon PI

In this section, we investigate properties of the finite-horizon PI in Theorem 1 as the final time $N \to \infty$. It is shown that, as $N \to \infty$, the proposed finite-horizon PI reduces to the infinite-horizon PI, i.e., Hewer's algorithm proposed in [30].

Assumption 1 Throughout this section, we assume stationary system (A, B) and constant matrices $Q \geqslant 0$, $R > 0$, $F = 0$. In addition, (A, B) is controllable, and $(A, Q^{1/2})$ is observable.

For convenient reference, the infinite-horizon PI algorithm is summarized in the following lemma.

Lemma 1 If staring with an initial stabilizing control $u^{(0)} = -L^{(0)}x$, setting $i = 0$, the following three steps are iterated infinitely:

1) (Policy evaluation) Solve for $V^{(i)}$ from

$$V^{(i)} = (A^{(i)})^{\mathrm{T}} V^{(i)} A^{(i)} + Q + L^{(i)} R L^{(i)}, \tag{16}$$

where $A^{(i)} = A - BL^{(i)}$.

2) (Policy improvement) Obtain improved control gain

$$L^{(i+1)} = (R + B^{\mathrm{T}} V^{(i)} B)^{-1} B^{\mathrm{T}} V^{(i)} A. \tag{17}$$

3) Let $i = i + 1$, and go to 1).

Then,

i) $V^{(i)} \geqslant V^{(i+1)} \geqslant P^*$, $\forall i \in \mathbb{Z}_+$.

ii) $\lim\limits_{i\to\infty} V^{(i)} = P^*$, where symmetric matrix $P^* > 0$ is the unique solution to the ARE,

$$\begin{aligned} P^* = {} & A^{\mathrm{T}} P^* A + Q \\ & - A^{\mathrm{T}} P^* B (R + B^{\mathrm{T}} P^* B)^{-1} B^{\mathrm{T}} P^* A \end{aligned} \tag{18}$$

and the corresponding stationary optimal control $u^* = -L^* x$, where

$$L^* = (R + B^{\mathrm{T}} P^* B)^{-1} B^{\mathrm{T}} P^* A \tag{19}$$

minimizes the infinite-horizon cost function

$$J(k_0, u, x_{k_0}) = \sum_{j=k_0}^{\infty} (x_j^{\mathrm{T}} Q x_j + u_j^{\mathrm{T}} R u_j). \tag{20}$$

iii) $A^{(i)}$ is a Schur matrix for each $i \in \mathbb{Z}_+$. $V^{(i)}$ is the unique positive definite solution to Lyapunov equation (16), and the cost under control $u^{(i)} = -L^{(i)} x$ is given by

$$\begin{aligned} J(k_0, u^{(i)}, x_{k_0}) &= x_{k_0}^{\mathrm{T}} V^{(i)} x_{k_0} \\ &= \sum_{j=k_0}^{\infty} (x_j^{\mathrm{T}} Q x_j + (u_j^{(i)})^{\mathrm{T}} R u_j^{(i)}). \end{aligned} \tag{21}$$

In the rest of this section, let the pair $(V_k^{(i)}, L_k^{(i)})$ denote the solutions to equations (12) and (13), respectively; $(V^{(i)}, L^{(i)})$ denote solutions to equations (16) and (17), respectively; (P_k^*, L_k^*) denote the solutions to equations (5) and (4), respectively; (P^*, L^*) denote solutions to equations (18) and (19), respectively. Note that for fixed $V_N^{(i)} = F$ and free $V_{k_0}^{(i)}$, $N \to \infty$ implies $k \to -\infty$.

Theorem 2 In stationary case, if initial control gains are chosen as $L_k^{(0)} = L^{(0)}$, for all $k \in [k_0, N)$ in Theorem 1, where $L^{(0)}$ is the same with that in Lemma 1, and $F = 0$, then $\lim\limits_{k\to-\infty} V_k^{(i)} = V^{(i)}$, $\lim\limits_{k\to-\infty} L_k^{(i)} = L^{(i)}$, as $N \to \infty$. Thus, as N goes to the infinity, the finite-horizon PI in Theorem 1 reduces to the infinite-horizon PI in Lemma 1.

Proof Let $u_k^{(0)} = -L^{(0)} x_k$, for $\forall k \in [k_0, N)$. By the definition of $V_k^{(i)}$ in (9) and $F = 0$, for $\forall x_0 \in R^n$, $x_k = x_{k-1} = x_0$, $\forall k \in [k_0 + 1, N)$, we have

$$\begin{aligned} x_0^{\mathrm{T}} V_k^{(0)} x_0 &= \sum_{j=k}^{N-1} [x_j^{\mathrm{T}} Q x_j + u_j^{(0)} R u_j^{(0)}] \\ &\leqslant \sum_{j=k-1}^{N-1} [x_j^{\mathrm{T}} Q x_j + u_j^{(0)} R u_j^{(0)}] = x_0^{\mathrm{T}} V_{k-1}^{(0)} x_0. \end{aligned}$$

Note that the above inequality holds for all $x_0 \in R^n$, thus we know that $\{V_{N-1}^{(0)}, V_{N-2}^{(0)}, \ldots, V_{k_0}^{(0)}\}$ is a nondecreasing sequence. Furthermore, for every $k \leqslant N - 1$, $x_0^{\mathrm{T}} V_k^{(0)} x_0$ is bounded from above by the cost $x_0^{\mathrm{T}} V^{(0)} x_0$. Since for the same initial conditional x_0 and same stabilizing control $u^{(0)}$, $x_0^{\mathrm{T}} V_k^{(0)} x_0$ is equal to the sum of only first $N - k$ terms in equation (21). Thus as $k \to -\infty$, $\{V_{N-1}^{(0)}, V_{N-2}^{(0)}, \ldots, V_{-\infty}^{(0)}\}$ is a nondecreasing sequence and bounded from above by $V^{(0)}$. Again, by the theorem on the convergence of a monotone sequence of self-adjoint operators (See [36, Pages 189–190]), $\lim\limits_{k\to-\infty} V_k^{(0)}$ exists. Since A, Q and R are time invariant, letting $k \to -\infty$ in (12), we have

$$V_{-\infty}^{(0)} = (A^{(0)})^{\mathrm{T}} V_{-\infty}^{(0)} A^{(0)} + Q + L^{(0)} R L^{(0)}.$$

Obviously, this is the same form with Lyapunov equation (16). Due to the uniqueness of the solution to Lyapunov equation (16). We know $\lim\limits_{k\to-\infty} V_k^{(0)} = V^{(0)}$, and by (13), $\lim\limits_{k\to-\infty} L_k^{(0)} = L^{(0)}$. This means in the limiting case, policy evaluation step (12) and policy improvement step (13) in Theorem 1 reduce to (16) and (17) when $i = 0$, respectively. By induction, $\lim\limits_{k\to-\infty} V_k^{(i)} = V^{(i)}$ and $\lim\limits_{k\to-\infty} L_k^{(i)} = L^{(i)}$ hold for all $i > 0$. This proves that, in the limiting case $N \to \infty$, the proposed finite-horizon PI in Theorem 1 reduces to the Hewer's algorithm, i.e., Lemma 1. This completes the proof. □

Remark 4 There are two iteration variables in Theorem 1: time index k and algorithmic iteration index i. The relationships between different value-control pairs with respect to these two iteration variables are summarized in Fig. 1. The convergence of $(V_k^{(i)}, L_k^{(i)})$ to $(V^{(i)}, L^{(i)})$ as $k \to -\infty$ is proved in Theorem 2; the convergence of $(V_k^{(i)}, L_k^{(i)})$ to (P_k^*, L_k^*) is the main result in Theorem 1; the convergence of $(V^{(i)}, L^{(i)})$ to (P^*, L^*) is the Hewer's algorithm [30]; the proof of convergence of (P_k^*, L_k^*) to (P^*, L^*) can be found in [2, Proposition 3.1.1].

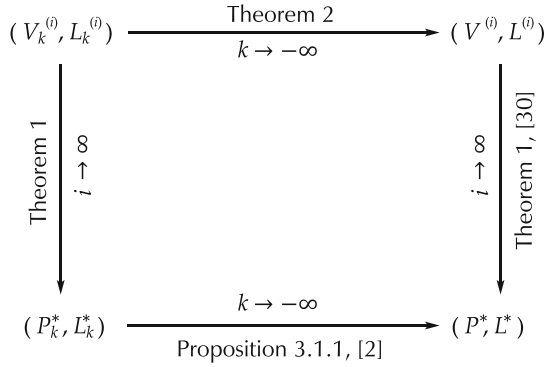


Fig. 1 Value-control pair relationships of policy iteration.

Remark 5 When system (1) is periodic, it is also true that the proposed finite-horizon PI reduces to the infinite-horizon PI for periodic systems as $N \to \infty$. The infinite-horizon PI for discrete-time periodic linear systems can be found in [37] (Theorem 3). In [38] (Theorem 3.1), it was shown that, the optimal control problem of discrete-time periodic linear system can be transformed into an equivalent optimal control problem of linear time-invariant system. Thus it is straightforward to extend Theorem 2 to periodic systems.

Remark 6 It is easy to see from [14] that in stationary case, (5) reduces to the infinite-horizon discrete-time VI, by defining $X_k = P^*_{N-k}$ in [33, Theorem 17.5.3].

4 Adaptive optimal controller design for linear time-varying discrete-time systems

This section applies adaptive dynamic programming to derive novel data-driven algorithms without precise knowledge of the system dynamics based on Theorem 1 and equation (5), respectively.

4.1 PI-based off-policy ADP

Suppose that a series of control inputs $\{u_k^{(0)}\}_{k=k_0}^{N-1}$ is applied to the system to generate data, and we are in the ith stage of the procedure in Theorem 1. Then (1) can be rewritten as

$$x_{k+1} = A_k^{(i)} x_k + B_k(L_k^{(i)} x_k + u_k^{(0)}). \tag{22}$$

By (22), we have

$$\begin{aligned} & x_{k+1}^{\mathrm{T}} V_{k+1}^{(i)} x_{k+1} - x_k^{\mathrm{T}} V_k^{(i)} x_k \\ &= x_k^{\mathrm{T}}((A_k^{(i)})^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i)} - V_k^{(i)}) x_k \\ &\quad + 2(L_k^{(i)} x_k + u_k^{(0)})^{\mathrm{T}} B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k^{(i)} x_k \\ &\quad + (L_k^{(i)} x_k + u_k^{(0)})^{\mathrm{T}} B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k (L_k^{(i)} x_k + u_k^{(0)}). \end{aligned} \tag{23}$$

Substituting (12) into equation (23) and rearranging the terms, we obtain

$$\begin{aligned} & x_k^{\mathrm{T}}(Q_k + (L_k^{(i)})^{\mathrm{T}} R_k L_k^{(i)}) x_k \\ &= x_k^{\mathrm{T}} V_k^{(i)} x_k - x_{k+1}^{\mathrm{T}} V_{k+1}^{(i)} x_{k+1} \\ &\quad + 2(L_k^{(i)} x_k + u_k^{(0)})^{\mathrm{T}} B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k x_k \\ &\quad + (u_k^{(0)} + L_k^{(i)} x_k)^{\mathrm{T}} B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k (u_k^{(0)} - L_k^{(i)} x_k). \end{aligned} \tag{24}$$

Using the properties of Kronecker product,

$$\begin{aligned} & a^{\mathrm{T}} W b = (b^{\mathrm{T}} \otimes a^{\mathrm{T}}) \mathrm{vec}(W), \\ & (AC) \otimes (BD) = (A \otimes B)(C \otimes D), \end{aligned}$$

where $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $W \in \mathbb{R}^{n \times m}$, A, B, C, D are matrices with compatible dimensions, the equation (24) can be rewritten as

$$\psi_k^{(i)} = \phi_k^{(i)} \begin{bmatrix} \mathrm{vecs}(V_k^{(i)}) \\ \mathrm{vec}(B_k^{\mathrm{T}} V_{k+1}^{(i)} A_k) \\ \mathrm{vecs}(B_k^{\mathrm{T}} V_{k+1}^{(i)} B_k) \\ \mathrm{vecs}(V_{k+1}^{(i)}) \end{bmatrix}, \tag{25}$$

where

$$\begin{aligned} \psi_k^{(i)} &= x_k^{\mathrm{T}}(Q_k + (L_k^{(i)})^{\mathrm{T}} R_k L_k^{(i)}) x_k, \\ \phi_k^{(i)} &= [\tilde{x}_k^{\mathrm{T}}, \delta_k^{(i)}, \widetilde{u_k^{(0)}}^{\mathrm{T}} - \widetilde{L_k^{(i)} x_k}^{\mathrm{T}}, -\tilde{x}_{k+1}^{\mathrm{T}}], \\ \delta_k^{(i)} &= 2[(x_k^{\mathrm{T}} \otimes x_k^{\mathrm{T}})(I_n \otimes (L_k^{(i)})^{\mathrm{T}} + (x_k^{\mathrm{T}} \otimes (u_k^{(0)})^{\mathrm{T}})]. \end{aligned}$$

Note that $\psi_k^{(i)}$ and $\phi_k^{(i)}$ are all known data matrices.

For fixed k, (25) is a degenerate linear equation, since there is only one data triad $(x_k, u_k^{(0)}, x_{k+1})$, yet at least $[\frac{1}{2}n(n+1) + mn + \frac{1}{2}m(m+1)]$ unknown variables to solve. Therefore, more data needs to be collected. To this end, suppose in total l groups of different control sequences $\{u_{k,j}^{(0)}\}_{k=k_0}^{N-1}$, $j = 1, 2, \ldots, l$ are applied to the system and corresponding data is recorded. Each group of control sequences can choose the following form:

$$u_{k,j}^{(0)} = -L_k^{(0)} x + w_{k,j}, \tag{26}$$

where $w_{k,j} \in \mathbb{R}^m$ is the exploration noise used to achieve sufficient excitation of the system. By defining the following data matrices:

$$\begin{aligned} \Gamma_{\tilde{x}_k} &= [\tilde{x}_{k,1}, \tilde{x}_{k,2}, \ldots, \tilde{x}_{k,l}]^{\mathrm{T}}, \\ \Gamma_{xx_k} &= [x_{k,1} \otimes x_{k,1}, x_{k,2} \otimes x_{k,2}, \ldots, x_{k,l} \otimes x_{k,l}]^{\mathrm{T}}, \\ \Gamma_{xu_{k,0}} &= [x_{k,1} \otimes u_{k,1}^{(0)}, x_{k,2} \otimes u_{k,2}^{(0)}, \ldots, x_{k,l} \otimes u_{k,l}^{(0)}]^{\mathrm{T}}, \end{aligned}$$

$$\Gamma_{\tilde{u}_{k,0}} = [\widetilde{u^{(0)}_{k,1}}, \widetilde{u^{(0)}_{k,2}}, \dots, \widetilde{u^{(0)}_{k,l}}]^{\mathrm{T}},$$

$$\Gamma_{\widetilde{L^{(i)}_k x_k}} = [\widetilde{L^{(i)}_k x_{k,1}}, \widetilde{L^{(i)}_k x_{k,2}}, \dots, \widetilde{L^{(i)}_k x_{k,l}}]^{\mathrm{T}},$$

$$\Delta^{(i)}_k = 2(\Gamma_{xx_k}(I_n \otimes (L^{(i)}_k)^{\mathrm{T}}) + \Gamma_{xu_{k,0}}),$$

we obtain the matrix equation

$$\Psi^{(i)} = \Phi^{(i)}\Theta^{(i)}, \tag{27}$$

where

$$\Theta^{(i)} = \begin{bmatrix} \theta^{(i)}_{k_0} \\ \theta^{(i)}_{k_0+1} \\ \vdots \\ \theta^{(i)}_{N-1} \end{bmatrix}, \quad \theta^{(i)}_k = \begin{bmatrix} \mathrm{vecs}(V^{(i)}_k) \\ \mathrm{vec}(B^{\mathrm{T}}_k V^{(i)}_{k+1} A_k) \\ \mathrm{vecs}(B^{\mathrm{T}}_k V^{(i)}_{k+1} B_k) \end{bmatrix}$$

and $\Psi^{(i)}, \Phi^{(i)}$ are given in (28) and (29) below.

$$\Phi^{(i)} = \begin{bmatrix} \Gamma_{\tilde{x}_{k_0}} & \Delta^{(i)}_{k_0} & (\Gamma_{\tilde{u}_{k_0,0}} - \Gamma_{\widetilde{L^{(i)}_{k_0} x_{k_0}}}) & -\Gamma_{\tilde{x}_{k_0+1}} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \Gamma_{\tilde{x}_{k_0+1}} & \Delta^{(i)}_{k_0+1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \Gamma_{\tilde{x}_{N-1}} & \Delta^{(i)}_{N-1} & (\Gamma_{\tilde{u}_{N-1,0}} - \Gamma_{\widetilde{L^{(i)}_{N-1} x_{N-1}}}) \end{bmatrix} \tag{28}$$

$$\begin{aligned} \Psi^{(i)} = [&\Gamma_{\tilde{x}_{k_0}} \mathrm{vecs}(Q_{k_0} + (L^{(i)}_{k_0})^{\mathrm{T}} R_{k_0} L^{(i)}_{k_0}) \quad \Gamma_{\tilde{x}_{k_0+1}} \mathrm{vecs}(Q_{k_0+1} + (L^{(i)}_{k_0+1})^{\mathrm{T}} R_{k_0+1} L^{(i)}_{k_0+1}) \quad \cdots \\ &\Gamma_{\tilde{x}_{N-1}} \mathrm{vecs}(Q_{N-1} + (L^{(i)}_{N-1})^{\mathrm{T}} R_{N-1} L^{(i)}_{N-1}) + \Gamma_{\tilde{x}_N} \mathrm{vecs}(F)]^{\mathrm{T}} \end{aligned} \tag{29}$$

If $\Phi^{(i)}$ is a full-column rank matrix, (27) can be uniquely solved, i.e.,

$$\Theta^{(i)} = ((\Phi^{(i)})^{\mathrm{T}}\Phi^{(i)})^{-1}(\Phi^{(i)})^{\mathrm{T}}\Psi^{(i)}. \tag{30}$$

As a result, the gain matrix $L^{(i+1)}_k$ can be obtained by using equation (13).

Now we are in the place to present the data-driven off-policy policy iteration algorithm.

Next, the convergence analysis of Algorithm 1 is presented.

Algorithm 1 (PI-based off-policy ADP)

Choose arbitrary $\{L^{(0)}_k\}^{N-1}_{k=k_0}$, threshold $\epsilon > 0$.
Run system (1) l times. In the jth run, use $\{u^{(0)}_{k,j}\}^{N-1}_{k=k_0}$ from (26) as control inputs, and collect the generated data.
Let $i \leftarrow 0$.
Repeat
  Compute $\Psi^{(i)}$, $\Phi^{(i)}$ by (28), (29) respectively;
  $\Theta^{(i)} \leftarrow ((\Phi^{(i)})^{\mathrm{T}}\Phi^{(i)})^{-1}(\Phi^{(i)})^{\mathrm{T}}\Psi^{(i)}$;
  $k \leftarrow k_0$;
  While $k < N$ **do**
    $L^{(i+1)}_k \leftarrow (R_k + B^{\mathrm{T}}_k V^{(i)}_{k+1} B_k)^{-1} B^{\mathrm{T}}_k V^{(i)}_{k+1} A_k$;
    $k \leftarrow k + 1$;
  $i \leftarrow i + 1$;
Until $\max_k \|V^{(i)}_k - V^{(i-1)}_k\| < \epsilon$.
Use $u_k = -L^{(i)}_k x_k$ as the approximate optimal control.

Theorem 3 If there exists an integer $l_0 > 0$, such that, for all $l > l_0$ and $k \in [k_0, N)$,

$$\begin{aligned} &\mathrm{rank}([\Gamma_{\tilde{x}_k}, \Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}]) \\ &= \frac{m(m+1)}{2} + mn + \frac{n(n+1)}{2} \end{aligned} \tag{31}$$

then,

1) $\Phi^{(i)}$ has full column rank for all $i \in \mathbb{Z}_+$.

2) the sequences $\{V^{(i)}_k\}^{\infty}_{i=0}$ and $\{L^{(i)}_k\}^{\infty}_{i=0}$ obtained by iteratively solving equation (27) and equation (13) converge to the optimal values P^*_k and L^*_k, respectively.

Proof For convenience, define variables

$$\Xi_v = [X^{\mathrm{T}}_v, Y^{\mathrm{T}}_v, Z^{\mathrm{T}}_v]^{\mathrm{T}},$$

$$\Phi^{(i)}_k = [\Gamma_{\tilde{x}_k}, \Delta^{(i)}_k, \Gamma_{\tilde{u}_{k,0}} - \Gamma_{\widetilde{L^{(i)}_k x_k}}],$$

$$\Psi^{(i)}_k = \Gamma_{\tilde{x}_k} \mathrm{vecs}(Q_k + (L^{(i)}_k)^{\mathrm{T}} R_k L^{(i)}_k),$$

where $X_v = \mathrm{vecs}(X_m)$, $Y_v = \mathrm{vec}(Y_m)$, $Z_v = \mathrm{vecs}(Z_m)$, $X_m \in \mathbb{R}^{n\times n}$ and $Z_m \in \mathbb{R}^{m\times m}$ are symmetric matrices, $Y_m \in \mathbb{R}^{m\times n}$, $k \in [k_0, N)$.

We first prove property 1). Obviously, $\Phi^{(i)}$ has full column rank if and only if $\Phi^{(i)}_k$ has full column rank for all $k \in [k_0, N)$. This is equivalent to show that linear equation

$$\Phi^{(i)}_k \Xi_v = 0$$

has unique solution $\Xi_v = 0$. We shall show that it is indeed the case.

According to the definition of $\Phi^{(i)}_k$ and the equation

(24), we have

$$\Phi_k^{(i)}\Xi_v = [\Gamma_{\tilde{x}_k}, 2\Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}]\begin{bmatrix}\mathrm{vecs}(\Lambda_1)\\ \mathrm{vec}(\Lambda_2)\\ \mathrm{vecs}(\Lambda_3)\end{bmatrix} = 0, \tag{32}$$

where

$$\begin{aligned}\Lambda_1 &= X_m + (L_k^{(i)})^{\mathrm{T}} Y_m + Y_m^{\mathrm{T}} L_k^{(i)} - (L_k^{(i)})^{\mathrm{T}} Z_m L_k^{(i)},\\ \Lambda_2 &= Y_m,\\ \Lambda_3 &= Z_m.\end{aligned}$$

From (31), we know that $[\Gamma_{\tilde{x}_k}, 2\Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}]$ has full column rank. This means the only solution to (32) is $\Lambda_1 = 0$, $\Lambda_2 = 0$, $\Lambda_3 = 0$. This is true if and only if $\Xi_v = 0$. Thus $\Phi_k^{(i)}$ always has full column rank. Therefore, $\Phi^{(i)}$ has full column ranks for all $i \in \mathbb{Z}_+$.

Now we prove property 2). By (24), it is easy to check that

$$\Phi_k^{(i)}\theta_k^{(i)} = \Psi_k^{(i)} + \Gamma_{\tilde{x}_{k+1}}\mathrm{vecs}(V_{k+1}^{(i)}). \tag{33}$$

Suppose now Ξ_v and symmetric matrix $W_m \in \mathbb{R}^{n\times n}$ satisfies

$$\Phi_k^{(i)}\Xi_v = \Psi_k^{(i)} + \Gamma_{\tilde{x}_{k+1}}\mathrm{vecs}(W_m).$$

By definitions of $\Phi_k^{(i)}$ and $\Psi_k^{(i)}$, this is equivalent to

$$[\Gamma_{\tilde{x}_k}, 2\Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}]\begin{bmatrix}\mathrm{vecs}(\Omega_1)\\ \mathrm{vec}(\Omega_2)\\ \mathrm{vecs}(\Omega_3)\end{bmatrix} = 0, \tag{34}$$

where

$$\begin{aligned}\Omega_1 &= X_m - Q_k - (L_k^{(i)})^{\mathrm{T}} R_k L_k^{(i)} + (L_k^{(i)})^{\mathrm{T}} Y_m + Y_m^{\mathrm{T}} L_k^{(i)}\\ &\quad - (L_k^{(i)})^{\mathrm{T}} Z_m L_k^{(i)} - A_k^{\mathrm{T}} W_m A_k,\\ \Omega_2 &= Y_m - B_k^{\mathrm{T}} W_m A_k,\\ \Omega_3 &= Z_m - B_k^{\mathrm{T}} W_m B_k.\end{aligned}$$

Again, $[\Gamma_{\tilde{x}_k}, 2\Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}]$ has full column rank. This means the only solution to (34) is $\Omega_1 = 0$, $\Omega_2 = 0$, $\Omega_3 = 0$. Substituting $Y_m = B_k^{\mathrm{T}} W_m A_k$, $Z_m = B_k^{\mathrm{T}} W_m B_k$ into $\Omega_1 = 0$, we obtain

$$X_m = Q_k + (L_k^{(i)})^{\mathrm{T}} R_k L_k^{(i)} + (A_k^{(i)})^{\mathrm{T}} W_m A_k^{(i)}, \tag{35}$$

which is exactly the Lyapunov equation (12). Since $V_N^{(i)} = F$ for all $i \in \mathbb{Z}_+$, Policy Iteration by (27) and (13) is equivalent to (12) and (13). By Theorem 1, the convergence is proved. □

Remark 7 Different from the infinite-horizon PI algorithms [30, 31, 35, 39], where the initial gain matrices must be stabilizing to guarantee the convergence of the algorithm, there is no restriction on the initial gain matrices in finite-horizon PI-based ADP. However, in practice, utilization of stabilizing initial gain matrices (if exist) will prevent the system states from becoming too large. This is beneficial for the sufficient excitation of the systems and numerical stability.

4.2 VI-based off-policy ADP

In this section, we develop a VI-based off-policy ADP scheme using (5).

For $k = k_0, \ldots, N-1$, define

$$H_k^* = \begin{bmatrix}\mathrm{vecs}(H_{k,1}^*)\\ \mathrm{vec}(H_{k,2}^*)\\ \mathrm{vecs}(H_{k,3}^*)\end{bmatrix} = \begin{bmatrix}\mathrm{vecs}(A_k^{\mathrm{T}} P_{k+1}^* A_k)\\ \mathrm{vec}(B_k^{\mathrm{T}} P_{k+1}^* A_k)\\ \mathrm{vecs}(B_k^{\mathrm{T}} P_{k+1}^* B_k)\end{bmatrix}.$$

Then we have

$$\begin{aligned}&x_{k+1}^{\mathrm{T}} P_{k+1}^* x_{k+1}\\ &= (A_k x_k + B_k u_k)^{\mathrm{T}} P_{k+1}^* (A_k x_k + B_k u_k)\\ &= x_k^{\mathrm{T}} H_{k,1}^* x_k + 2u_k^{\mathrm{T}} H_{k,2}^* x_k + u_k^{\mathrm{T}} H_{k,3}^* u_k.\end{aligned}$$

Similar to the last section, suppose l groups of different control (26) are applied to the systems to collect data. Above equation yields

$$\mathscr{F}_k H_k^* = \Gamma_{\tilde{x}_{k+1}}\mathrm{vecs}(P_{k+1}^*), \tag{36}$$

where

$$\mathscr{F}_k = [\Gamma_{\tilde{x}_k}, 2\Gamma_{xu_{k,0}}, \Gamma_{\tilde{u}_{k,0}}].$$

Starting with $P_N^* = F$, H_{N-1}^* can be obtained by the LS solution to (36). Next P_{N-1}^* can be computed by Riccati equation (5), i.e.,

$$P_k^* = Q_k + H_{k,1}^* - (H_{k,2}^*)^{\mathrm{T}}(R_k + H_{k,3}^*)^{-1} H_{k,2}^*. \tag{37}$$

Then we can further get H_{N-2}^* by knowing P_{N-1}^* in (36). In this way, all $\{P_k^*\}_{k=k_0}^N$ can be determined from the data. This procedure is summarized in Algorithm 2.

Theorem 4 If there exists an integer $l_0 > 0$, such that, for all $l > l_0$ and $k \in [k_0, N)$, (31) holds, then Algorithm 2 yields the optimal values $\{P_k^*\}_{k=k_0}^N$, $\{L_k^*\}_{k=k_0}^{N-1}$.

Remark 8 In this paper, two data-driven ADP methods are proposed. Compared with PI, VI is much easier to implement and only requires finite steps to find the optimal solution. However, all the $(N-k_0)$-step iterations must be executed cascaded to find the optimal solution, which is time consuming when N is large. On the contrary, in PI, the full sequence $\{V_k^{(i)}\}_{k=k_0}^{N}$ can be obtained in each learning iteration in parallel. Due to the fast convergence of PI, PI shows a better performance when N is large.

Algorithm 2 (VI-based off-policy ADP)

Run system (1) l times. In the jth run, use $\{u_{k,j}^{(0)}\}_{k=k_0}^{N-1}$ from (26) as control inputs, and collect the generated data.
$P_N^* \leftarrow F$;
$k \leftarrow N$;
While $k > k_0$ **do**
$H_{k-1}^* \leftarrow (\mathscr{F}_{k-1}^{\mathrm{T}}\mathscr{F}_{k-1})^{-1}\mathscr{F}_{k-1}^{\mathrm{T}}\Gamma_{\bar{x}_k}\mathrm{vecs}(P_k^*)$;
$L_{k-1}^* \leftarrow (R_{k-1}+H_{k-1,3}^*)^{-1}H_{k-1,2}^*$;
Solve P_{k-1}^* by (37);
$k \leftarrow k-1$;
Use $u_k = -L_k^* x_k$ as the approximate optimal control.

5 Application

In this section, we apply the algorithms presented in previous sections to the spacecraft attitude control problem. Due to a space structure extended in orbit [40] or on-orbit refueling [41], moment of inertia of the spacecraft will be time-varying. This can be modeled by the following continuous-time linear time varying system [42],

$$\dot{x} = A_{\mathrm{c}}(t)x + B_{\mathrm{c}}(t)u, \tag{38}$$

where $x = [q, \dot{q}]^{\mathrm{T}}$, $q = [\alpha, \beta, \gamma]^{\mathrm{T}}$, α, β, γ are the row angle, the pitch angle, the yaw angle of spacecrafts respectively,

$$A_{\mathrm{c}}(t) = \begin{bmatrix} 0 & I \\ -D(t) & -K(t) \end{bmatrix},\ B_{\mathrm{c}}(t) = \begin{bmatrix} 0 \\ L(t) \end{bmatrix},$$

$$D(t) = \begin{bmatrix} 0 & 0 & d(t) \\ 0 & 0 & 0 \\ -d(t) & 0 & 0 \end{bmatrix},\ K(t) = \begin{bmatrix} k_1(t) & 0 & 0 \\ 0 & k_2(t) & 0 \\ 0 & 0 & k_3(t) \end{bmatrix},$$

$$L(t) = \begin{bmatrix} \frac{1}{J_x(t)} & 0 & 0 \\ 0 & \frac{1}{J_y(t)} & 0 \\ 0 & 0 & \frac{1}{J_z(t)} \end{bmatrix},$$

$$\begin{cases} d(t) = \dfrac{\omega_0(J_y(t)-J_x(t)-J_z(t))}{J_x(t)}, \\ k_1(t) = \dfrac{4\omega_0^2(J_y(t)-J_z(t))}{J_x(t)}, \\ k_2(t) = \dfrac{3\omega_0^2(J_x(t)-J_z(t))}{J_y(t)}, \\ k_3(t) = \dfrac{\omega_0^2(J_y(t)-J_x(t))}{J_z(t)}, \end{cases}$$

$$\begin{cases} J_x(t) = 1070 + 15.5t, \\ J_y(t) = 2150 - 11t, \\ J_z(t) = 1300 - 7.5t. \end{cases}$$

$J_x(t), J_y(t), J_z(t)$ are the components of moment of inertia of spacecrafts with respect to a body coordinate system, and $\omega_0 = 0.0011$ rad/s is the orbital rate. We discretize system (38) by using Forward-Euler method with sampling time $h = 0.1$ s, and obtain the discrete-time linear time varying system,

$$x(k+1) = A_{\mathrm{d}}(k)x(k) + B_{\mathrm{d}}(k)u(k), \tag{39}$$

where

$$A_{\mathrm{d}}(k) = I + A_{\mathrm{c}}(kh)h,\ \ B_{\mathrm{d}}(k) = B_{\mathrm{c}}(kh)h.$$

Let final time $N = 300$, and choose weighting matrices $Q = 10I_6$, $R = I_3$. The proposed Algorithms 1 and 2 are applied to system (39). For Algorithm 1, the initial control gains are all zero matrices. To collect data that satisfies conditions (31), in the simulation, the elements of initial angles q_0 are assumed to be independently and uniformly distributed over $[-1, 1]$, while the elements of initial angle velocities $\dot{q}_0$ are assumed to be independently and uniformly distributed over $[-0.01, 0.01]$. The exploration noises are chosen as

$$[w_{k,j}]_r = 2(\sum_{s=1}^{500} \sin(\sigma_{s,j} \cdot k)),$$

where for rth component of control input, in jth trial, each $\sigma_{s,j}$ is independently drawn from the uniform dis-

tribution over $[-500, 500]$.

Algorithm 1 stops after 10 iterations on this task. Fig. 2 shows the convergence of $V_k^{(i)}$ to P_k^* in the PI case. One can find that, for fixed time k, $V_k^{(i)}$ converges monotonically to their optimal values, as predicted in Theorem 1. For Algorithm 2, the maximum difference between the LS solutions $\hat{P}_k^*$ given by Algorithm 2 and the optimal values P_k^*, measured by matrix 2-norm, is 1.6749×10^{-8}. The final approximate optimal control gains and the initial control gains (all zero matrix) are applied to system (39) with initial condition $q_0 = [0.0175, 0.0175, 0.0175]^{\mathrm{T}}$, $\dot{q}_0 = [0, 0, 0]^{\mathrm{T}}$. The state trajectories under these two control are shown in Figs. 3 and 4, respectively. The final control inputs are shown in Fig. 5. Therefore, the simulation results demonstrate the effectiveness of the proposed finite-horizon PI and VI algorithms.

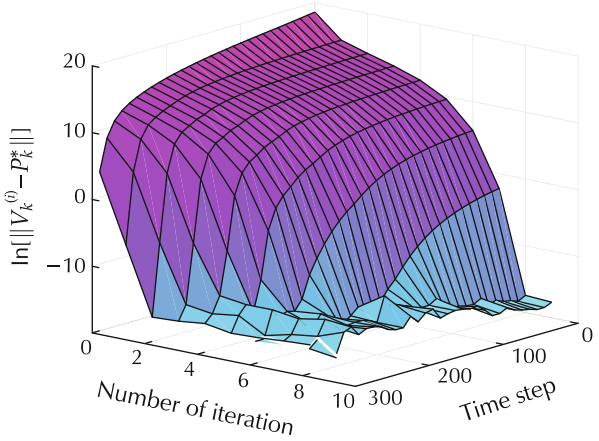


Fig. 2 Comparison of $V_k^{(i)}$ with its optimal value.

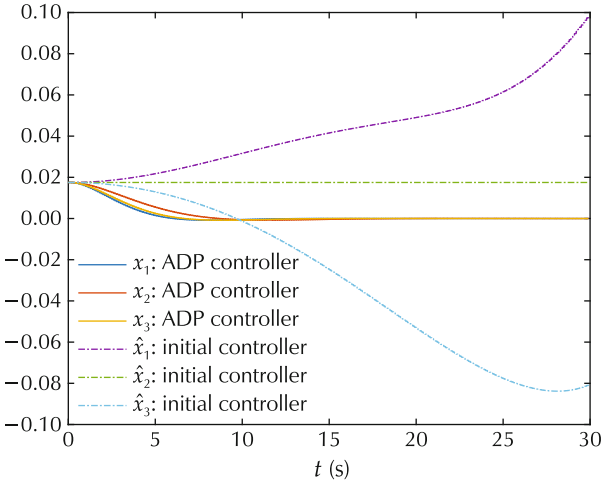


Fig. 3 Trajectories of angles of the spacecraft.

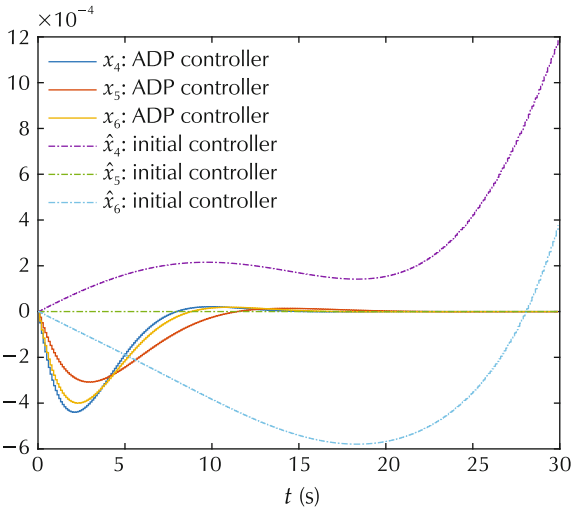


Fig. 4 Trajectories of angle velocities of the spacecraft.

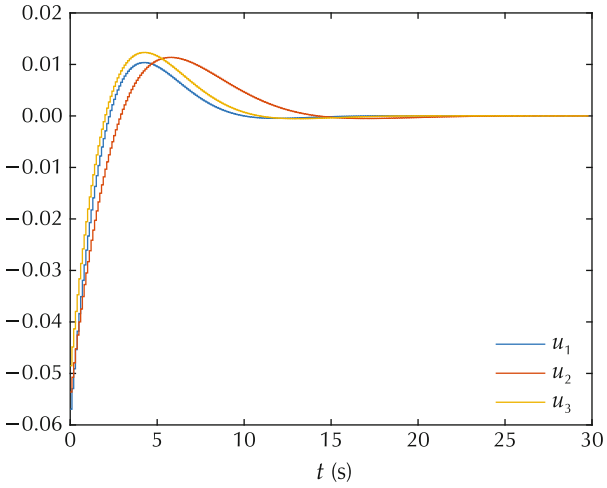


Fig. 5 Approximate optimal control inputs.

6 Conclusions

The popular policy iteration (PI) method has been extended for the finite-horizon optimal control problem of linear time-varying discrete-time systems in this paper. In the stationary case, its connections with the existing infinite-horizon PI methods are revealed. In addition, novel data-driven off-policy PI and VI algorithms are derived to find the approximate optimal controllers in the absence of the precise knowledge of system dynamics. It is shown via rigorous convergence proofs that under the proposed data-driven algorithms, the convergence of a sequence of suboptimal controllers to the optimal control is guaranteed under some mild conditions. The obtained results are validated by a case study in spacecraft attitude control.

References

[1] R. E. Bellman. *Dynamic Programming*. Princeton: Princeton University Press, 1957.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. 4th ed. Belmont: Athena Scientific, 2017.

[3] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton: Princeton University Press, 2011.

[4] D. P. Bertsekas, J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Belmont: Athena Scientific, 1996.

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

[6] C. Szepesvari. *Algorithms for Reinforcement Learning*. San Franscisco: Morgan and Claypool Publishers, 2010.

[7] Y. Jiang, Z. P. Jiang. *Robust Adaptive Dynamic Programming*. Hoboken: Wiley, 2017.

[8] F. L. Lewis, D. Liu (editors). *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken: Wiley, 2013.

[9] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, et al. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(6): 2042 – 2062.

[10] W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken: Wiley, 2011.

[11] D. Liu, Q. Wei, D. Wang, et al. *Adaptive Dynamic Programming with Applications in Optimal Control*. Berlin: Springer International Publishing, 2017.

[12] R. Kamalapurkar, P. Walters, J. Rosenfeld, et al. *Reinforcement Learning for Optimal Feedback Control: A Lyapunov Based Approach*. Berlin: Springer International Publishing, 2018.

[13] W. Gao, Z. P. Jiang. Adaptive dynamic programming and adaptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control*, 2016, 61(12): 4164 – 4169.

[14] D. Vrabie, K. G. Vamvoudakis, F. L. Lewis. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London: Institution of Engineering and Technology, 2013.

[15] M. Huang, W. Gao, Z. P. Jiang. Connected cruise control with delayed feedback and disturbance: An adaptive dynamic programming approach. *International Journal of Adaptive Control and Signal Processing*, 2017: DOI https://doi.org/10.1002/acs.2834.

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

[17] D. P. Bertsekas. Value and policy iterations in optimal control and adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28(3): 500 – 509.

[18] D. Kleinman, T. Fortmann, M. Athans. On the design of linear systems with piecewise-constant feedback gains. *IEEE Transactions on Automatic Control*, 1968, 13(4): 354 – 361.

[19] Q. M. Zhao, H. Xu, J. Sarangapani. Finite-horizon near optimal adaptive control of uncertain linear discrete-time systems. *Optimal Control Applications and Methods*, 2015, 36(6): 853 – 872.

[20] C. X. Mu, D. Wang, H. B. He. Data-driven finite-horizon approximate optimal control for discrete-time nonlinear systems using iterative HDP approach. *IEEE Transactions on Cybernetics*, 2018, 48(10): 2948 – 2961.

[21] A. Heydari, S. N. Balakrishnan. Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, 24(1): 145 – 157.

[22] R. Beard. *Improving the Closed-loop Performance of Nonlinear Systems*. Ph.D. dissertation. New York: Rensselaer Polytechnic Institute, 1995.

[23] T. Cheng, F. L. Lewis, M. Abu-Khalaf. A neural network solution for fixed-final time optimal control of nonlinear systems. *Automatica*, 2007, 43(3): 482 – 490.

[24] Q. M. Zhao, H. Xu, S. Jagannathan. Neural network-based finite-horizon optimal control of uncertain affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(3): 486 – 499.

[25] P. Frihauf, M. Krstic, T. Basar. Finite-horizon LQ control for unknown discrete-time linear systems via extremum seeking. *European Journal of Control*, 2013, 19(5): 399 – 407.

[26] S. J. Liu, M. Krstic, T. Basar. Batch-to-batch finite-horizon LQ control for unknown discrete-time linear systems via stochastic extremum seeking. *IEEE Transactions on Automatic Control*, 2017, 62(8): 4116 – 4123.

[27] J. Fong, Y. Tan, V. Crocher, et al. Dual-loop iterative optimal control for the finite horizon LQR problem with unknown dynamics. *Systems & Control Letters*, 2018, 111: 49 – 57.

[28] G. De Nicolao. On the time-varying Riccati difference equation of optimal filtering. *SIAM Journal on Control and Optimization*, 1992, 30(6): 1251 – 1269.

[29] E. Emre, G. Knowles. A Newton-like approximation algorithm for the steady-state solution of the riccati equation for time-varying systems. *Control Applications and Methods*, 1987, 8(2): 191 – 197.

[30] G. Hewer. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 1971, 16(4): 382 – 384.

[31] D. Kleinman. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 1968, 13(1): 114 – 115.

[32] D. Kleinman. *Suboptimal Design of Linear Regulator Systems Subject to Computer Storage Limitations*. Ph.D. dissertation. Cambridge: Massachusetts Institute of Technology, 1967.

[33] P. Lancaster, L. Rodman. *Algebraic Riccati Equations*. Oxford: Oxford University Press, 1995.

[34] S. J. Bradtke, B. E. Ydstie, A. G. Barto. Adaptive linear quadratic control using policy iteration. *Proceedings of the American Control Conference*, Baltimore: IEEE, 1994: 3475 – 3479.

[35] W. Gao, Y. Jiang, Z. P. Jiang, et al. Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. *Automatica*, 2016, 72: 37 – 45.

[36] L. V. Kantorovich, G. P. Akilov. *Functional Analysis in Normed Spaces*. New York: Macmillan, 1964.

[37] S. Bittanti, P. Colaneri, G. De Nicolao. The difference periodic Riccati equation for the periodic prediction problem. *IEEE Transactions on Automatic Control*, 1988, 33(8): 706 – 712.

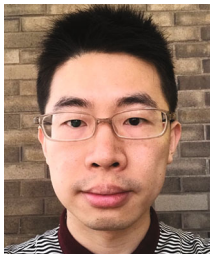
[38] Y. Yang. An efficient LQR design for discrete-time linear periodic system based on a novel lifting method. *Automatica*, 2018, 87: 383 – 388.

[39] Y. Jiang, Z. P. Jiang. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 2012, 48(10): 2699 – 2704.

[40] R. Okano, T. Kida. Stability and stabilization of extending space structures. *Transactions of the Society of Instrument and Control Engineers*, 2002, 38(3): 284 – 292.

[41] A. Long, M. Richards, D. E. Hastings. On-orbit servicing: a new value proposition for satellite design and operation. *Journal of Spacecraft and Rockets*, 2007, 44(4): 964 – 976.

[42] L. Zhang, G. R. Duan. Robust poles assignment for a kind of second-order linear time-varying systems. *Proceedings of the Chinese Control Conference*, Hefei: IEEE, 2012: 2602 – 2606.



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 with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, U.S.A. His research interests include optimal control, approximate/adaptive dynamic programming, and reinforcement learning. E-mail: bo.pang@nyu.edu.

Tao BIAN received the B.Eng. degree in Automation from Huazhong University of Science and Technology, Wuhan, China, in 2012, and the M.Sc. and the Ph.D. degree in Electrical Engineering from Tandon School of Engineering, New York University, Brooklyn, NY, in 2014 and 2017, respectively. He is currently a quantitative finance analyst, assistant vice president, at Bank of America Merrill Lynch, One Bryant Park, New York. His research interests include reinforcement learning, control and optimization of stochastic systems. E-mail: tbian@nyu.edu.

Zhong-Ping JIANG received the B.Sc. degree in Mathematics from the University of Wuhan, Wuhan, China, in 1988, the M.Sc. degree in Statistics from the University of Paris XI, Paris, France, in 1989, and the Ph.D. degree in Automatic Control and Mathematics from the École des Mines de Paris, Paris, in 1993. He is currently a Professor of electrical and computer engineering with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, U.S.A. He was a named a Highly Cited Researcher by Web of Science (2018) and has coauthored Stability and Stabilization of Nonlinear Systems (Springer, 2011), Nonlinear Control of Dynamic Networks (Taylor & Francis, 2014), Robust Adaptive Dynamic Programming (Wiley-IEEE Press, 2017) and Nonlinear Control Under Information Constraints (Science Press, 2018). His current research interests include stability theory, robust/adaptive/distributed nonlinear control, adaptive dynamic programming, and their applications to information, mechanical, and biological systems. Dr. Jiang is an IEEE Fellow and an IFAC Fellow. E-mail: zjiang@nyu.edu.