2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Adaptive Optimal Output-Feedback Control of Discrete-Time Systems based on Hybrid Iteration

Sitong Liu[1], Weinan Gao[1] and Zhong-Ping Jiang[2]

Abstract— **In this paper, a novel adaptive dynamic programming (ADP) algorithm, named output-feedback hybrid iteration (HI), is proposed to address the adaptive optimal control problem of discrete-time linear systems. The proposed output-feedback HI strategy learns the optimal control policy through two phases. First, a novel data-driven value-iteration (VI) scheme is employed to learn an admissible output-feedback control policy using the input/output data without relying on the knowledge of system matrices. Then, with the obtained admissible control policy, the optimal output feedback-control policy is approximated with an accelerated convergence rate through output-feedback policy iteration (PI). Online input/output data are utilized to reconstruct the full state of the system and integrated into the new output-feedback HI algorithm. Simulation results are presented and demonstrate the efficacy and practicality of the proposed output-feedback HI approach in comparison with traditional PI and VI techniques.**

I. INTRODUCTION

The aim of optimal control problems is to design optimal controllers with respect to a predefined performance index. The design process usually relies on the prior knowledge of system dynamics [1]. As an enhancement of optimal control, adaptive optimal control techniques have emerged, aiming to design optimal control policies when the system model is partially or completely unknown. Adaptive dynamic programming (ADP) represents a data-driven approach that does not require accurate knowledge of the system model. see, e.g., [2]–[8]. Rather than directly solving the algebraic Riccati equations (AREs) via the system model, ADP approaches the optimal control policy through iterations—thus reducing the computation burden associated with directly solving AREs.

There are two typical methods in the area of ADP, i.e., policy iteration (PI) and value iteration (VI). PI is usually quadratically convergent to the optimal control policy for linear systems [9]–[12]. For instance, a computational ADP method is proposed for adaptive optimal state-feedback control of continuous-time linear systems without requiring the knowledge of the state and input matrices [13]. The authors of [14] have combined ADP and the small-gain theory to propose robust adaptive dynamic programming to deal with dynamically perturbed nonlinear systems [15], [16]. However, an admissible control policy is required to initiate the learning process, which severely limits the practical implementation when the system models are not exactly known. On the contrary, the VI method eliminates the requirement of an initial admissible control policy but usually yields a slower convergence rate towards the optimal solution [17]–[19]. Recently, hybrid iteration has been proposed to solve the problems associated with the slow convergence from VI and the need for an admissible control policy from PI [20], [21].

It is noteworthy that the majority of the existing ADP results are based on state feedback. In other words, the complete state information has to be available when implementing these algorithms. It is imperative to develop ADP methods when dealing with unknown system dynamics and unmeasurable states. In order to overcome these challenges, an output-feedback ADP methodology has been proposed in [22] for discrete-time linear systems. The authors of [23] have proposed a novel computational output-feedback ADP approach for continuous-time linear systems with completely unknown dynamics via sampling-data strategy. Output-feedback Q-learning methods have been developed in [24], [25] to solve the discrete-time and continuous-time optimal control problems by combining Q-learning and PI. However, both PI-based and VI-based output-feedback control designs inherit their limitations from state-feedback designs. Specifically, output-feedback PI requires an admissible control policy, while output-feedback VI inherently converges more slowly. The main objective of this paper is to integrate VI and PI methods to overcome these limitations. The main challenge in accomplishing this objective is to find an admissible control policy with unknown system dynamics and unmeasurable state.

Compared with the state-of-the-art studies in ADP, the main contributions of this paper are summarized as follows.

1) We propose a novel output-feedback ADP method aimed at approximating the optimal control policy along with the corresponding value function with rigorous proof of stability and convergence.
2) We propose a sufficient condition to determine the admissibility of any control policy obtained from output-feedback VI.
3) We combine the HI and ADP for the output-feedback adaptive optimal control of linear dynamical systems.

The rest of this paper is organized as follows. Section

This paper was supported in part by the National Natural Science Foundation of China under Grant 62373090, while the work of the Zhong-Ping Jiang was supported in part by the NSF grant CNS-2227153

The corresponding author is Weinan Gao.

[1]Sitong. Liu and Weinan. Gao are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, Lianning, 110819, China, e-mails: 2370761@stu.neu.edu.cn, gaown@mail.neu.edu.cn.

[2]Zhong-Ping Jiang is with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA. e-mail: zjiang@nyu.edu.

II formulates the problem to be studied and presents some preliminaries related to optimal control and state reconstruction. Section III develops a novel adaptive optimal output-feedback control approach for discrete-time linear systems based on HI and ADP. Section IV contains simulation results to demonstrate the performance of the proposed controller. Finally, Section V closes the paper with some brief concluding remarks.

Notations. Throughout this paper, $\mathbb{Z}_+$ denotes the set of nonnegative integers. I_n denotes the identity matrix of dimension n. $|\cdot|$ denotes the induced norm operator for matrices and the Euclidean norm operator for vectors. $r(\cdot)$ denotes the rank for matrices. $\otimes$ denotes the Kronecker product operator. Given a matrix $A \in \mathbb{R}^{n\times m}$, with $a_i \in \mathbb{R}^n$ are the columns of A, $\text{vec}(A) = [a_1^T, a_2^T, \cdots, a_m^T]^T$. Given a vector $z \in \mathbb{R}^n$ and a matrix $P = P^T \in \mathbb{R}^{m\times m}$, $P \succ (\succeq) 0$ and $P \prec (\preceq) 0$ indicate that P is the positive definite (semidefinite) and negative definite (semidefinite), respectively. $\text{vecs}(P) = [p_{11}, 2p_{12}, \cdots, 2p_{1m}, p_{22}, 2p_{23}, \cdots, 2p_{m-1,m}, p_{mm}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}$, and $\text{vecv}(z) = \left[z_1^2,\ z_1z_2, \ldots, z_1z_n,\ z_2^2,\ z_2z_3, \ldots, z_{n-1}z_n,\ z_n^2\right]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}$.

II. Problem Statement and Preliminaries

Consider a class of discrete-time linear systems with single-input-single-output (SISO) as follows

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}$ is the control input vector, $y \in \mathbb{R}$ is the output vector, and $k \in \mathbb{Z}_+$ is the step. System matrices are $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^{1\times n}$ with the pair (A, B) controllable, the pair (A, C) observable, and the matrix A invertible.

For any $k \geq n$, the extended state equation using input/output sequences on time horizon $[k-n, k-1]$ can be written as

$$\begin{aligned} x_k &= A^n x_{k-n} + V(n)\bar{u}_{k-1,k-n}, \\ \bar{y}_{k-1,k-n} &= U(n)x_{k-n} + T(n)\bar{u}_{k-1,k-n} \end{aligned} \tag{2}$$

where

$$\begin{aligned} \bar{u}_{k_1,k-n} &= [u_{k-1}^{\mathrm{T}}, u_{k-2}^{\mathrm{T}}, \cdots, u_{k-n}^{\mathrm{T}}]^{\mathrm{T}}, \\ \bar{y}_{k-1,k-n} &= [y_{k-1}^{\mathrm{T}}, y_{k-2}^{\mathrm{T}}, \cdots, y_{k-n}^{\mathrm{T}}]^{\mathrm{T}}, \\ V(n) &= [B, AB, \cdots, A^{n-1}B], \\ U(n) &= [(CA^{n-1})^{\mathrm{T}}, \cdots, (CA)^{\mathrm{T}}, C^{\mathrm{T}}]^{\mathrm{T}}, \\ T(n) &= \begin{bmatrix} 0 & CB & CAB & \cdots & CA^{n-2}B \\ 0 & 0 & CB & \cdots & CA^{n-3}B \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & & 0 & CB \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The following lemma, Lemma 1, discusses and shows the methodology of reconstructing the state at step $k \geq n$ in terms of measured input/output data.

Lemma 1 ([26]). *Given a controllable and observable system (1), the system state is uniquely obtained in terms of measured input/output sequences by*

$$x_k = [M_u, M_y] \begin{bmatrix} \bar{u}_{k-1,k-n} \\ \bar{y}_{k-1,k-n} \end{bmatrix} = \Theta z_k \tag{3}$$

where $M_y = A^n U^{-1}(n)$, $M_u = V(n) - M_y T(n)$, $\Theta = [M_u, M_y]$ is full row rank, $z_k = [\bar{u}_{k-1,k-n}^T, \bar{y}_{k-1,k-n}^T]^T \in \mathbb{R}^{2n}$.

Remark 1. *It is checkable Θ must be full row rank under the condition of the pair (A, B) controllable and the pair (A, C) observable.*

Based on Lemma 1, one can generalize the idea of state-feedback ADP to tackle the output-feedback adaptive optimal control problem. For discrete-time system (1) and an arbitrary initial state $x_0 \in \mathbb{R}^n$, define the performance index as

$$J(x_0; u) = \sum_{k=0}^{\infty} \left(Qy_k^2 + Ru_k^2\right) \tag{4}$$

where $Q > 0$ and $R > 0$. A linear control law in the form of $u_k = -K^* x_k$, which minimizes the performance index, can be found by linear optimal control theory of discrete-time systems. In more detail, there must be a unique solution $P^* = (P^*)^{\mathrm{T}} \succ 0$ for the following ARE.

$$A^{\mathrm{T}}PA - P + C^{\mathrm{T}}QC - A^{\mathrm{T}}PB(R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA = 0, \tag{5}$$

and the optimal feedback control law K^* can be obtained as follows

$$K^* = (R + B^{\mathrm{T}}P^*B)^{-1}B^{\mathrm{T}}P^*A. \tag{6}$$

Solving P^* directly from (5) is difficult since (5) is nonlinear in P^*. In order to overcome this difficulty, one can leverage two methods for solving the ARE, i.e. PI and VI methods. The rest of this section recalls these approaches.

A. Policy Iteration

PI is an iterative approach that is proposed by Hewer in [27]; the details are as follows.

Lemma 2. *Given K_0 such that $A - BK_0$ is Schur stable, which means all of its eigenvalues are less than one in absolute value, and $P_j = P_j^{\mathrm{T}} \succ 0$ is the solution of the Lyapunov equation. The sequence $\{P_j\}_{j=0}^{\infty}$ is generated by the following equations for $j = 0, 1, 2, \cdots$,*

$$\begin{aligned} 0 =& (A - BK_j)^{\mathrm{T}}P_j(A - BK_j) - P_j \\ &+ C^{\mathrm{T}}QC + K_j^{\mathrm{T}}RK_j, \end{aligned} \tag{7}$$

$$K_{j+1} = (R + B^{\mathrm{T}}P_jB)^{-1}B^{\mathrm{T}}P_jA, \tag{8}$$

the following properties are satisfied:

1) *$A - BK_j$ is a Schur matrix for any $j \in \mathbb{Z}_+$.*
2) *$P^* \preceq P_{j+1} \preceq P_j$.*
3) *$\lim_{j\to\infty} K_j = K^*$, $\lim_{j\to\infty} P_j = P^*$.*

Algorithm 1 Model-based Hybrid Iteration

Select a small constant $\varepsilon > 0$ and a matrix $P_0 = (P_0)^{\mathrm{T}} \succeq 0$.
$j \leftarrow 0$.
repeat
  $K_j \leftarrow \left(B^{\mathrm{T}} P_j B + R\right)^{-1} B^{\mathrm{T}} P_j A$
  $P_{j+1} \leftarrow (A - BK_j)^{\mathrm{T}} P_j (A - BK_j) + C^{\mathrm{T}} Q C + K_j^{\mathrm{T}} R K_j$
  $j \leftarrow j + 1$
until $P_j - P_{j-1} \prec C^{\mathrm{T}} Q C$
$P_j \leftarrow P_{j-1}$
repeat
  $K_{j+1} \leftarrow \left(B^{\mathrm{T}} P_j B + R\right)^{-1} B^{\mathrm{T}} P_j A$
  Solve P_{j+1} from (7)
  $j \leftarrow j + 1$
until $|P_j - P_{j-1}| < \varepsilon$

B. Value Iteration

In contrast to PI, the VI does not require an initial stabilizing control policy. Instead, an arbitrary initial value matrix $P_0 = P_0^{\mathrm{T}} \succeq 0$ is employed to start the VI. The subsequent Lemma revisits the model-based VI strategy.

Lemma 3 ([28]). *For an arbitrary initial value matrix $P_0 \succeq 0$, the sequence $\{P_j\}_{j=0}^{\infty}$ converges to P^* as $j \to \infty$ by iterating the following step*

$$P_{j+1} = (A - BK_j)^{\mathrm{T}} P_j (A - BK_j) + K_j^{\mathrm{T}} R K_j^{\mathrm{T}} + C^{\mathrm{T}} Q C \tag{9}$$

where $K_j = (R + B^{\mathrm{T}} P_j B)^{-1} B^{\mathrm{T}} P_j A$.

In order to relax the limitation of slow convergence of VI and the reliance on stabilizing control policies for PI, the authors in [20] has proposed an HI approach; see [20] for more detail.

III. Output-Feedback Adaptive Optimal Controller Design Based on HI

In this section, we propose a novel data-driven HI approach for discrete-time systems via output-feedback control. Notably, the system matrices A, B and C are all unknown, the state x_k is unmeasurable, and an admissible control policy is unknown.

A. Phase 1: Learning Towards an Admissible Control Policy

Let two matrices of H_j and $\bar{H}_j$ take the form of

$$H_j = \begin{bmatrix} H_j^{11} & H_j^{12} \\ H_j^{21} & H_j^{22} \end{bmatrix} := \begin{bmatrix} B^{\mathrm{T}} P B & B^{\mathrm{T}} P A \\ A^{\mathrm{T}} P B & A^{\mathrm{T}} P A \end{bmatrix}, \tag{10}$$

$$\bar{H}_j = \begin{bmatrix} \bar{H}_j^{11} & \bar{H}_j^{12} \\ \bar{H}_j^{21} & \bar{H}_j^{22} \end{bmatrix} := \begin{bmatrix} B^{\mathrm{T}} P B & B^{\mathrm{T}} P A \Theta \\ \Theta^{\mathrm{T}} A^{\mathrm{T}} P B & \Theta^{\mathrm{T}} A^{\mathrm{T}} P A \Theta \end{bmatrix}. \tag{11}$$

From (9) and the control gain defined by $K_{j+1} = (R + B^{\mathrm{T}} P_j B)^{-1} B^{\mathrm{T}} P_j A$, one obtains

$$P_{j+1} = A^{\mathrm{T}} P_j A + C^{\mathrm{T}} Q C - A^{\mathrm{T}} P_j B (R + B^{\mathrm{T}} P_j B)^{-1} B^{\mathrm{T}} P_j A. \tag{12}$$

By collecting the input/output data (12) can be rewritten as follows.

$$x_{k+1}^{\mathrm{T}} C^{\mathrm{T}} Q C x_{k+1} = - x_{k+1}^{\mathrm{T}} \mathcal{F}(P_j) x_{k+1} + x_{k+1}^{\mathrm{T}} P_{j+1} x_{k+1} \tag{13}$$

where

$$\mathcal{F}(P_j) = A^{\mathrm{T}} P_j A - A^{\mathrm{T}} P_j B (R + B_j^P B)^{-1} B^{\mathrm{T}} P_j A$$

Further, $x_{k+1}^{\mathrm{T}} P_{j+1} x_{k+1}$ in (13) can be written as follows.

$$\begin{aligned} x_{k+1}^{\mathrm{T}} P_{j+1} x_{k+1} =& (A x_k + B u_k)^{\mathrm{T}} P_{j+1} (A x_k + B u_k) \\ =& \begin{bmatrix} u_k \\ x_k \end{bmatrix}^{\mathrm{T}} H_{j+1} \begin{bmatrix} u_k \\ x_k \end{bmatrix} = \begin{bmatrix} u_k \\ z_k \end{bmatrix}^{\mathrm{T}} \bar{H}_{j+1} \begin{bmatrix} u_k \\ z_k \end{bmatrix} \end{aligned} \tag{14}$$

Through substituting (10) and (14) into (13), the following equation can be obtained

$$\begin{aligned} y_{k+1}^{\mathrm{T}} Q y_{k+1} =& - x_{k+1}^{\mathrm{T}} [H_j^{22} - (H_j^{12})^{\mathrm{T}} (R + H_j^{11})^{-1} H_j^{12}] x_{k+1} \\ &+ \left(\begin{bmatrix} u_k \\ x_k \end{bmatrix} \otimes \begin{bmatrix} u_k \\ x_k \end{bmatrix} \right)^{\mathrm{T}} \mathrm{vec}(H_{j+1}) \\ =& - z_{k+1}^{\mathrm{T}} [\bar{H}_j^{22} - (\bar{H}_j^{12})^{\mathrm{T}} (R + \bar{H}_j^{11})^{-1} \bar{H}_j^{12}] z_{k+1} \\ &+ \left[\mathrm{vecv} \left(\begin{bmatrix} u_k \\ z_k \end{bmatrix} \right) \right]^{\mathrm{T}} \mathrm{vecs}(\bar{H}_{j+1}) \\ =& - \phi_{k+1}^j + (\psi_k)^{\mathrm{T}} \mathrm{vecs}(\bar{H}_{j+1}), \end{aligned} \tag{15}$$

where

$$\begin{aligned} \phi_{k+1}^j =& z_{k+1}^{\mathrm{T}} [\bar{H}_j^{22} - (\bar{H}_j^{12})^{\mathrm{T}} (R + \bar{H}_j^{11})^{-1} \bar{H}_j^{12}] z_{k+1}, \\ \psi_k =& \mathrm{vecv} \left(\begin{bmatrix} u_k^{\mathrm{T}} & z_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \right). \end{aligned}$$

Define

$$\begin{aligned} \Psi_j^V =& [\psi_{k_0}, \psi_{k_1}, \cdots, \psi_{k_s}]^{\mathrm{T}}, \\ \Phi_j^V =& [y_{k_0}^{\mathrm{T}} Q y_{k_0} + \phi_{k_0}^j, \cdots, y_{k_s}^{\mathrm{T}} Q y_{k_s} + \phi_{k_s}^j]^{\mathrm{T}}, \\ \Gamma_{\bar{z}} =& [\mathrm{vecv}(z_{k_0}), \mathrm{vecv}(z_{k_1}), \cdots, \mathrm{vecv}(z_{k_s})]^{\mathrm{T}}, \\ \Gamma_{zu} =& [z_{k_0} \otimes u_{k_0}, z_{k_1} \otimes u_{k_1}, \cdots, z_{k_s} \otimes u_{k_s}]^{\mathrm{T}}, \\ \Gamma_{\bar{u}} =& [\mathrm{vecv}(u_{k_0}), \mathrm{vecv}(u_{k_1}), \cdots, \mathrm{vecv}(u_{k_s})]^{\mathrm{T}}. \end{aligned}$$

where $n < k_0 < k_1 < \cdots < k_s$. Equation (13) indicates that

$$\Psi_j^V \mathrm{vecs}(\bar{H}_{j+1}) = \Phi_j^V. \tag{16}$$

It is worth to note that Ψ_j^V and Φ_j^V are matrices related to input and output data, and then we can use (16) to compute $\bar{H}_{j+1}$ for each iteration. The uniqueness of the solution to (16) can be guaranteed by a rank condition which will be discussed in the Lemma below.

Lemma 4 ([23]). *If there exists a positive integer s^* such that for all $s > s^*$*

$$r([\Gamma_{\bar{z}}, \Gamma_{zu}, \Gamma_{\bar{u}}]) = n(2n+1) + 2n + 1, \tag{17}$$

then (16) has a unique solution.

The primary goal of the first phase is to keep repeating (16) until a stabilizing control policy is found. This task is nontrivial, particularly in scenarios where the system

matrices are entirely unknown and the state information is unmeasurable. In order to tackle this grand challenge, this paper proposes a novel data-driven scheme to check the admissibility of control policy learned via VI. The following Lemma is helpful to determine the admissibility via online data.

Lemma 5. *Under the condition that A is invertible and Lemma 4, the following properties hold.*

1) *There always exists a matrix $Z_n \in \mathbb{R}^{2n\times n}$ such that $Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n$ is nonsingular, where all columns of Z_n are chosen from the sequence $\{z_{k_i}\}_{i=0}^{s}$.*
2) *For any $x \in \mathbb{R}^n$, one can always obtain a $\beta \in \mathbb{R}^n$ such that*

$$x = \Theta Z_n \beta, \tag{18}$$

where Z_n satisfies the property 1.

Proof. To prove the first property, under the condition of Lemma 4, there always exists an invertible matrix $\zeta \in \mathbb{R}^{2n\times 2n}$, where all columns of ζ are chosen from the sequence $\{z_{k_i}\}_{i=0}^{s}$. Since Θ is full row rank, the matrix $\Theta\zeta \in \mathbb{R}^{n\times 2n}$ is also full row rank. Then there are n linearly independent columns in the matrix $\Theta\zeta$. We call the matrix formed by these n linear independent columns as $\bar{Z}_n := \Theta Z_n \in \mathbb{R}^{n\times n}$. Obviously, all columns of Z_n can be found in ζ. The condition $r(\bar{Z}_n) = n$ implies that $r(Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n) = r(\bar{Z}_n^{\mathrm{T}}A^{\mathrm{T}}PA\bar{Z}_n) = r(A^{\mathrm{T}}PA)$. Under the condition that A is invertible, one has $r(A^{\mathrm{T}}PA) = n$. Therefore $r(Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n) = n$ which also implies that $Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n$ is nonsingular.

To prove the second property, since $Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n$ is nonsingular, one have $n = r(Z_n^{\mathrm{T}}\bar{H}_j^{22}Z_n) = r(Z_n^{\mathrm{T}}\Theta^{\mathrm{T}}A^{\mathrm{T}}PA\Theta Z_n) \leq r(\Theta Z_n) \leq n$, which proves that ΘZ_n is nonsingular. Obviously for any $x \in \mathbb{R}^n$, one can always find a sequence of parameters $\beta = [\beta_1, \beta_2, \cdots, \beta_n]^{\mathrm{T}}$ such that $x = \Theta Z_n \beta$. The proof is thus completed. □

Remark 2. *Under the condition of Lemma 4, one can always find a Z_n that satisfies Lemma 5 after making $C_n^s \in \mathbb{Z}_+$ selections from the sequence $\{z_{k_i}\}_{i=0}^{s}$, where*

$$C_n^s = \frac{s\times(s-1)\times\cdots\times(s-n+1)}{n!} \tag{19}$$

with ! representing the factorial. The resultant selections are called $Z_n^1, Z_n^2, \cdots, Z_n^{C_n^s}$ that will be used in Algorithm 2.

Define the matrix $\mathcal{H}_j$ as follows.

$$\mathcal{H}_j = \begin{bmatrix} \mathcal{H}_j^{1,1} & \mathcal{H}_j^{1,2} & \cdots & \mathcal{H}_j^{1,n} \\ \mathcal{H}_j^{2,1} & \mathcal{H}_j^{2,2} & \cdots & \mathcal{H}_j^{2,n} \\ \vdots & \ddots & \ddots & \vdots \\ \mathcal{H}_j^{n,1} & \cdots & \cdots & \mathcal{H}_j^{n,n} \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{20}$$

The element of the matrix $\mathcal{H}_j$ is computed by Z_n as follows, where $\bar{K}_j = (R + B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA\Theta := (R + \bar{H}_j^{11})^{-1}\bar{H}_j^{12}$.

$$\mathcal{H}_j^{l,w} = \begin{bmatrix} -\bar{K}_j z_l \\ z_l \end{bmatrix}^{\mathrm{T}} \bar{H}_j \begin{bmatrix} -\bar{K}_j z_w \\ z_w \end{bmatrix} - \begin{bmatrix} u_{l-1} \\ z_{l-1} \end{bmatrix}^{\mathrm{T}} \bar{H}_j \begin{bmatrix} u_{w-1} \\ z_{w-1} \end{bmatrix} \tag{21}$$

for any $l, w \in \{1, 2, \cdots, n\}$. Vectors z_l and z_w are the lth and wth columns in Z_n, respectively. $[u_{l-1}^{\mathrm{T}}, z_{l-1}^{\mathrm{T}}]^{\mathrm{T}}$ or $[u_{w-1}^{\mathrm{T}}, z_{w-1}^{\mathrm{T}}]^{\mathrm{T}}$ are online data at steps $l-1$ and $w-1$.

Based on Lemma 5, the following theorem is presented as a sufficient condition of admissibility of an output-feedback control policy.

Theorem 1. *If $\mathcal{H}_j$ solved by (20) and (21) is negative definite at the iteration $j \in \mathbb{Z}_+$, then $\bar{K}_j$ is guaranteed to be an admissible control gain.*

Proof. If $\mathcal{H}_j$ is negative definite, the following equation holds

$$\beta^{\mathrm{T}}\mathcal{H}_j\beta < 0, \forall\beta \neq 0 \tag{22}$$

which is equivalent to

$$\begin{bmatrix} -\bar{K}_j Z_n\beta \\ Z_n\beta \end{bmatrix}^{\mathrm{T}} \bar{H}_j \begin{bmatrix} -\bar{K}_j Z_n\beta \\ Z_n\beta \end{bmatrix} - (Z_n\beta)^{\mathrm{T}}\bar{P}_j Z_n\beta < 0 \tag{23}$$

where $\bar{P}_j = \Theta^{\mathrm{T}}P_j\Theta$. Under the second property of Lemma 5, one obtains

$$\begin{bmatrix} -K_j x \\ x \end{bmatrix}^{\mathrm{T}} H_j \begin{bmatrix} -K_j x \\ x \end{bmatrix} - x^{\mathrm{T}}P_j x < 0, \forall x \neq 0. \tag{24}$$

Upon substitution of (10) into (24), one reaches

$$x^{\mathrm{T}}[(A - BK_j)^{\mathrm{T}}P_j(A - BK_j) - P_j]x < 0, \forall x \neq 0. \tag{25}$$

One can immediately have $[A - BK_j]^{\mathrm{T}}P_j[A - BK_j] - P_j \prec 0$, which is enough to show that $A - BK_j$ is Schur stable based on the Lyapunov stability theory. It also implies that, in correspondence with $\mathcal{H}_j$, $\bar{K}_j$ is an admissible control gain. The proof is thus completed. □

Based on Theorem 1, one can obtain a sufficient condition to check the admissibility of any learned policy via VI. If an admissible control policy is obtained, it is feasible to switch to PI to accelerate the convergence rate.

B. Phase 2: Exploring the Optimal Control Policy

By collecting the input/output data, (1) and (7), one can obtain

$$\begin{aligned} & y_{k+1}^{\mathrm{T}}Qy_{k+1} + z_{k+1}^{\mathrm{T}}\bar{K}_j^{\mathrm{T}}R\bar{K}_j z_{k+1} \\ =& \begin{bmatrix} u_k \\ z_k \end{bmatrix}^{\mathrm{T}} \bar{H}_j \begin{bmatrix} u_k \\ z_k \end{bmatrix} - \begin{bmatrix} -\bar{K}_j z_{k+1} \\ z_{k+1} \end{bmatrix}^{\mathrm{T}} \bar{H}_j \begin{bmatrix} -\bar{K}_j z_{k+1} \\ z_{k+1} \end{bmatrix} \\ =& \left[\mathrm{vecv}\left(\begin{bmatrix} u_k \\ z_k \end{bmatrix}\right)\right]^{\mathrm{T}} \mathrm{vecs}(\bar{H}_j) \\ & - \left[\mathrm{vecv}\left(\begin{bmatrix} -\bar{K}_j z_{k+1} \\ z_{k+1} \end{bmatrix}\right)\right]^{\mathrm{T}} \mathrm{vecs}(\bar{H}_j) \\ =& [\lambda_{k+1}^j]^{\mathrm{T}}\mathrm{vecs}(\bar{H}_j) \end{aligned} \tag{26}$$

Algorithm 2 Output-feedback Hybrid Iteration

1: Select two constants $Q, R > 0$ and a stopping criterion $\epsilon > 0$.
2: $j \leftarrow 0$. $p \leftarrow 1$. $\bar{H}_j \leftarrow 0$. $\bar{K}_j \leftarrow 0$. $Z_n \leftarrow 0$.
3: Get $Z_n^1, Z_n^2, \cdots, Z_n^{C_n^s}$ from the sequence $\{z_{k_i}\}_{i=0}^s$.
4: **repeat**
5: Solve $\bar{H}_{j+1}$ from (16).
6: **if** $Z_n = 0$ **then**
7: **if** $|\bar{H}_{j+1}^{22}| \neq 0$ **then**
8: **repeat**
9: Calculate $|Z_n^{p\mathrm{T}} \bar{H}_{j+1}^{22} Z_n^p|$
10: $p \leftarrow p + 1$
11: **until** $|Z_n^{p\mathrm{T}} \bar{H}_{j+1}^{22} Z_n^p| \neq 0$
12: $Z_n \leftarrow Z_n^p$
13: $\bar{K}_{j+1} \leftarrow (R + \bar{H}_{j+1}^{11})^{-1} \bar{H}_{j+1}^{12}$.
14: Calculate the matrix $\mathcal{H}_{j+1}$ with Z_n, by (20) and (21).
15: $j \leftarrow j + 1$
16: **until** $\mathcal{H}_j \prec 0$.
17: $\bar{K}_j \leftarrow (R + \bar{H}_j^{11})^{-1} \bar{H}_j^{12}$.
18: **repeat**
19: Solve $\bar{H}_{j+1}$ from (27).
20: $\bar{K}_{j+1} \leftarrow (R + \bar{H}_{j+1}^{11})^{-1} \bar{H}_{j+1}^{12}$
21: $j \leftarrow j + 1$
22: **until** $|\bar{H}_j - \bar{H}_{j-1}| < \varepsilon$

where

$$\lambda_{k+1}^j = \mathrm{vecv}\left(\begin{bmatrix} u_k \\ z_k \end{bmatrix}\right) - \mathrm{vecv}\left(\begin{bmatrix} -\bar{K}_j z_{k+1} \\ z_{k+1} \end{bmatrix}\right).$$

Now, define

$$\begin{aligned} \delta_k^j &= y_k^{\mathrm{T}} Q y_k + z_k^{\mathrm{T}} \bar{K}_j^{\mathrm{T}} R \bar{K}_j z_k, \\ \Delta_j^P &= [\delta_{k_0}^j, \delta_{k_1}^j, \cdots, \delta_{k_s}^j], \\ \Lambda_j^P &= [\lambda_{k_0}^j, \lambda_{k_1}^j, \cdots, \lambda_{k_s}^j]^{\mathrm{T}}. \end{aligned}$$

Equation (26) implies

$$\Lambda_j^P \mathrm{vecs}\left(\bar{H}_{j+1}\right) = \Delta_j^P \tag{27}$$

where the uniqueness of the solution is also ensured by (17). Next, the output-feedback HI Algorithm is given in Algorithm 2 with the proof of convergence given in Theorem 2.

Theorem 2. *Under the condition of Lemma 4, the sequences $\{\bar{H}_j\}_{j=0}^{\infty}$ and $\{\bar{K}_j\}_{j=0}^{\infty}$ learned from Algorithm 2 converge to $\bar{H}^*$ and $\bar{K}^*$, respectively. where*

$$\begin{aligned} \bar{K}^* &= K^* \Theta \\ \bar{H}^* &= \begin{bmatrix} B^T P^* B & B^T P^* A \Theta \\ \Theta^T A^T P^* B & \Theta^T A^T P^* A \Theta \end{bmatrix}. \end{aligned}$$

Proof. If the rank condition (17) in Lemma 4 is satisfied, (16) is guaranteed to have a unique solution. Steps 5 and 13 are equivalent to Steps 4 and 5 of Algorithm 1. Besides, from Theorem 1, $\bar{K}_j$ calculated from the Step 18 is ensured as an admissible control policy. Under the rank condition of (17), $\bar{P}_j$ and $\bar{K}_j$ solved from Steps 19 and 20 are equivalent to Steps 10 and 11 in Algorithm 1. By [27], one has $\lim_{j\to\infty} K_j = K^*$ and $\lim_{j\to\infty} P_j = \bar{P}^*$, which indicates that $\lim_{j\to\infty} \bar{K}_j = \bar{K}^*$ and $\lim_{j\to\infty} \bar{H}_j = \bar{H}^*$. □

IV. Illustrative Example

This section evaluates the effectiveness of the proposed strategy using a second-order discrete-time linear system. Consider a second-order discrete-time linear system

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 1.0732 & 0.0288 \\ 0.0033 & 1.0047 \end{bmatrix} x_k + \begin{bmatrix} 0.0562 \\ 0.4128 \end{bmatrix} u_k \\ y_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} x_k, \end{aligned} \tag{28}$$

and Q and R are chosen to be 1. Letting $N = 2$. The initial conditions of the states $x_0 = \begin{bmatrix} 1 & 0.5 \end{bmatrix}^{\mathrm{T}}$, and the exploration noise is chosen as a sum of sinusoidal signals with different frequencies expressed in the form of $\eta = \sum_{i=1}^{100} \sin(\omega_i k)$ where ω_i are randomly selected from the range of $[-100, 100]$. The input/output data are collected from $k = 0$ to 40, and the output-feedback HI is started from $k = 40$. For Algorithm 2, set $\varepsilon = 0.1$. At 14th iterations, $\bar{K}_{14}$ corresponding to $\mathcal{H}_{14}$ satisfies Theorem 1. Thus, $\bar{K}_{14}$ is admissible and takes the following form.

$$\bar{K}_{14} = \begin{bmatrix} 0.2205 & -0.5528 & 14.24 & -13.37 \end{bmatrix}.$$

Applying this admissible $\bar{K}_{14}$ to output-feedback PI, the derived approximate optimal values are obtained at the 18th iterations, the approximated optimal control gain $\bar{K}_{18}$ and the optimal control gain $\bar{K}^*$ are shown as follows.

$$\begin{aligned} \bar{K}_{18} &= \begin{bmatrix} 0.2470 & -0.6327 & 16.1739 & -15.3036 \end{bmatrix}. \\ \bar{K}^* &= \begin{bmatrix} 0.2470 & -0.6327 & 16.1739 & -15.3036 \end{bmatrix}. \end{aligned}$$

Fig. 1 depicts the number of iterations required for the traditional VI and PI based output-feedback ADP, and illustrates the feasibility and efficiency of the HI. And it can be seen that the hybrid iteration converges to the optimal solution faster than the VI under the condition without an initial stabilizing control policy. Fig. 2 depicts the trajectories of the input and output data, where the control policy is updated at k = 40. Finally, one can notice the convergence of the input and output in an optimal sense.

V. Conclusion

This paper has proposed a novel computational output-feedback ADP approach—output-feedback HI—for linear systems. Notably, the proposed algorithm does not require the exact knowledge of system parameters or the full-state information. Additionally, it does not require any prior knowledge of a stabilizing control policy, while maintaining a fast convergence rate compared to traditional value iteration. Simulation results for a linear system clearly demonstrate the effectiveness of the proposed methodology.

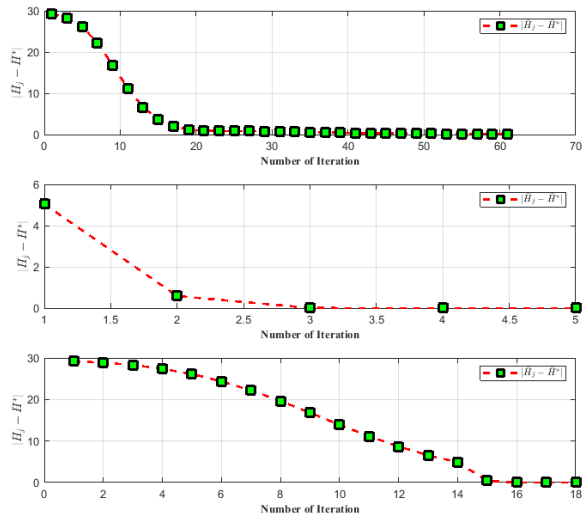

Fig. 1. Convergence of $\bar{H}_j$ under VI, PI and HI.

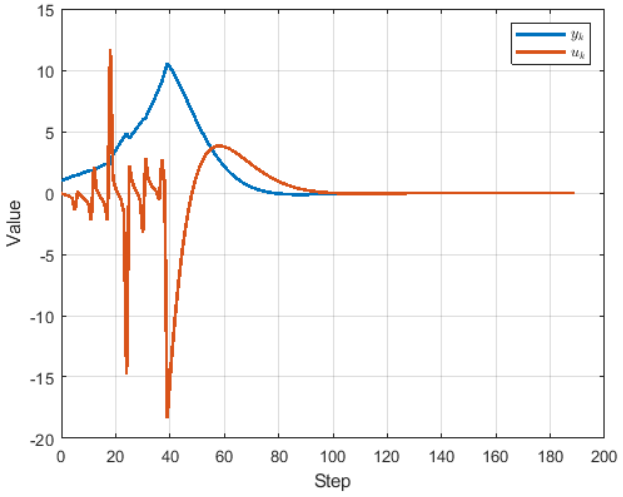

Fig. 2. Trajectory of output and input of the second-order linear system.

REFERENCES

[1] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ: Wiley, 2012.

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

[3] Y. Jiang and Z. P. Jiang, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2917–2929, 2015.

[4] Q. Wei, D. Liu, Y. Liu, and R. Song, "Optimal constrained self-learning battery sequential management in microgrid via adaptive dynamic programming," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 2, pp. 168–176, 2016.

[5] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*, ser. Communications and Control Engineering. Springer, 2013.

[6] 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 Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[7] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive dynamic programming with applications in optimal control*. Springer, 2017.

[8] Y. Yang, H. Modares, K. G. Vamvoudakis, W. He, C.-Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with approximation errors," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 762–13 773, 2022.

[9] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, 2014.

[10] B. Zhao, D. Wang, G. Shi, D. Liu, and Y. Li, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 10, pp. 1725–1735, 2017.

[11] F. Zhao and L. Zhao, "Adaptive optimal control for large-scale systems based on robust policy iteration," in *2022 34th Chinese Control and Decision Conference (CCDC)*, 2022, pp. 2704–2709.

[12] O. Qasem and W. Gao, "Robust policy iteration of uncertain interconnected systems with imperfect data," *IEEE Transactions on Automation Science and Engineering*, pp. 1–9, 2023.

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

[14] Z. P. Jiang, I. M. Mareels, and Y. Wang, "A Lyapunov formulation of the nonlinear small-gain theorem for interconnected ISS systems," *Automatica*, vol. 32, no. 8, pp. 1211 – 1215, 1996.

[15] Y. Jiang and Z. P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[16] ——, "A robust adaptive dynamic programming principle for sensorimotor control with signal-dependent noise," *Journal of Systems Science and Complexity*, vol. 28, pp. 261–288, 2015.

[17] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 840–853, 2016.

[18] T. Bian and Z.-P. Jiang, "Value iteration, adaptive dynamic programming, and optimal control of nonlinear systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016, pp. 3375–3380.

[19] W. Gao, M. Mynuddin, D. C. Wunsch, and Z. P. Jiang, "Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5229–5240, 2022.

[20] O. Qasem, W. Gao, and K. G. Vamvoudakis, "Adaptive optimal control of continuous-time nonlinear affine systems via hybrid iteration," *Automatica*, vol. 157, p. 111261, 11 2023.

[21] O. Qasem, H. Gutierrez, and W. Gao, "Experimental validation of data-driven adaptive optimal control for continuous-time systems via hybrid iteration: An application to rotary inverted pendulum," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 6, pp. 6210–6220, 2024.

[22] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 1, pp. 14–25, 2011.

[23] W. Gao, Y. Jiang, Z. P. Jiang, and T. Chai, "Adaptive and optimal output feedback control of linear systems: An adaptive dynamic programming approach," in *Proceedings of the 11th World Congress on Intelligent Control and Automation*, Shenyang, China, 2014, pp. 2085–2090.

[24] S. A. A. Rizvi and Z. Lin, "Output feedback optimal tracking control using reinforcement Q-learning," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 3423–3428.

[25] ——, "Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 3417–3422.

[26] W. Aangenent, D. Kostic, B. de Jager, R. van de Molengraft, and M. Steinbuch, "Data-based optimal control," in *Proceedings of the American Control Conference*, vol. 2, Portland, OR, 2005, pp. 1460–1465.

[27] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[28] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*. New York, NY: Oxford University Press Inc., 1995.