MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Output-Feedback Learning-based Adaptive Optimal Control of Nonlinear Systems

Gao, Weinan; Wang, Yebin; Vamvoudakis, Kyriakos

TR2026-028 March 03, 2026

Abstract

In this paper, we develop a novel solution to the output-feedback adaptive optimal control problem of general nonlinear nonaffine systems based on reinforcement learning (RL). This bridges a gap between RL and nonlinear output- feedback control theory by designing high-fidelity data-driven controllers to achieve disturbance rejection in an optimal sense. Based on the condition of uniform observability, the state is reconstructed by the retrospective input and output information, which can be seen as equivalent to a deadbeat observer. Both policy and value iteration algorithms are proposed to learn the optimal output-feedback control policy and value function. The convergence of the proposed algorithms and the practical stability of the closed-loop system with learned control policy are rigorously ensured even when the optimal value function is not positive definite.

Automatica 2026

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Output-Feedback Learning-based Adaptive Optimal Control of Nonlinear Systems

Weinan Gao, Yebin Wang, and Kyriakos G. Vamvoudakis

***Abstract*—In this paper, we develop a novel solution to the output-feedback adaptive optimal control problem of general nonlinear nonaffine systems based on reinforcement learning (RL). This bridges a gap between RL and nonlinear output-feedback control theory by designing high-fidelity data-driven controllers to achieve disturbance rejection in an optimal sense. Based on the condition of uniform observability, the state is reconstructed by the retrospective input and output information, which can be seen as equivalent to a deadbeat observer. Both policy and value iteration algorithms are proposed to learn the optimal output-feedback control policy and value function. The convergence of the proposed algorithms and the practical stability of the closed-loop system with learned control policy are rigorously ensured even when the optimal value function is not positive definite.**

***Index Terms*—Reinforcement learning; Output-Feedback control; Adaptive optimal control**

I. Introduction

Reinforcement learning (RL) is a biologically inspired approach that optimizes a cumulative reward based on agent-environment interactions. From a control theory viewpoint, RL is essentially a direct adaptive optimal control approach (Sutton et al., 1992). By using RL, one can numerically approximate the optimal controller of a dynamical system, without relying on accurate knowledge of the physics, see Vrabie et al. (2009); Wang et al. (2009); Jiang and Jiang (2012); Wang et al. (2015); Kamalapurkar et al. (2015); Gao and Jiang (2016); Bertsekas (2017); Zhu and Zhao (2017); Gao and Jiang (2018); Mukherjee et al. (2019); Yang et al. (2020); Bian and Jiang (2022); Gao and Jiang (2022); Qasem et al. (2023); Jiang et al. (2024); Yang et al. (2024); Xie et al. (2024); Jiang et al. (2025) and references therein. RL has found a variety of applications for autonomy (Vamvoudakis and Lewis, 2011; Odekunle et al., 2020; Wang et al., 2017; Kontoudis and Vamvoudakis, 2019; Gao et al., 2018), smart grids (Tang et al., 2015; Ni and Paul, 2019), connected vehicles (Gao et al., 2017), industrial processes (Jiang et al., 2018) and smart buildings (Wei et al., 2015).

Policy iteration (PI) (Kleinman, 1968; Sandell, 1974; Saridis and Lee, 1979; Jiang and Jiang, 2012; Vrabie et al., 2009) and value iteration (VI) (Bian and Jiang, 2016; Bertsekas, 2017; Al-Tamimi et al., 2008; Wei et al., 2018) are two successive approximation approaches that can learn the optimal value function and the corresponding optimal control policy through iterations. Contingent on the availability of an initial admissible control policy (finite cost and stabilizing), Saridis and Lee (1979) showed that the control policy learned by PI at any iteration preserves such admissibility. Different from PI, the assumption of admissible initial policy has been removed in VI. However, VI usually converges much slower than PI (Powell, 2007).

Most of the existing RL algorithms are based on full state-feedback information (Bertsekas, 2017; Bian and Jiang, 2022; Jiang et al., 2024; Yang et al., 2024). In other words, the realization of these algorithms requires complete accessibility to the full state information, which is hard or expensive to achieve for engineering and high-order systems. Output-feedback adaptive optimal controllers have been designed using RL for both discrete-time linear systems (Lewis and Vamvoudakis, 2011; Rizvi and Lin, 2019; Valadbeigi et al., 2020) and continuous-time linear systems (Modares et al., 2016; Xie et al., 2024). In terms of robust adaptive dynamic programming (Jiang and Jiang, 2014), Gao et al. (2016) developed a robust optimal controller with output-feedback for partially linear systems. Based on RL and neural network approximations, output-feedback controllers have been developed for nonlinear strict feedback systems (Xu et al., 2014; He and Jagannathan, 2005; Yang and Jagannathan, 2012; Jiang et al., 2025). However, to the best of our knowledge, there are not any output-feedback adaptive optimal algorithms for general nonlinear nonaffine systems.

This paper has established a theoretical framework for output-feedback adaptive optimal controller design of nonlinear nonaffine disturbed systems. Similar to the output regulation problems (Huang, 2004), the disturbance in this paper is generated by an external dynamic system, called exosystem. This formulation includes a class of nonlinear time-varying systems as a special case.

Contributions: The contributions of this paper are five-fold. First, the optimal output-feedback controller is developed for a class of nonlinear systems through nonlinear optimal control theory and state reconstruction. Specifically, based on the condition of uniform observability (Moraal and Grizzle, 1995), the state of nonlinear system is reconstructed through retrospective input and output signals, which is equivalent to a deadbeat observer. This renders the optimal control policy to be a composite function of optimal state-feedback control law and state reconstruction law (deadbeat observer). Second, in order to approximate the optimal control policy and value function, novel PI and VI approaches are proposed via output-

W. Gao is with State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. email: gaown@mail.neu.edu.cn

Y. Wang is with Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139, USA. email: yebinwang@ieee.org

K. Vamvoudakis is with Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA. email: kyriakos@gatech.edu

feedback and Q-learning (Watkins and Dayan, 1992). Up to now, the convergence of PI algorithms (Kleinman, 1968; Sandell, 1974; Saridis and Lee, 1979; Jiang and Jiang, 2012; Vrabie et al., 2009; Al-Tamimi et al., 2008; Jiang et al., 2024) has usually relied on one or more of the following assumptions: 1) the optimal value function, which is the solution to the Hamilton-Jacobi-Bellman (HJB) equation, is positive definite; 2) the stage cost incorporates a positive definite function of the states; 3) the system is free from disturbances and stabilizable. However, for output-feedback control of nonlinear systems subject to disturbances, the scenario is fundamentally different. This is because the value function may not be positive definite, and the stage cost often lacks a positive definite function of the states. Furthermore, the overall disturbed system may not be inherently stabilizable, as the exosystem generating external disturbances could be non-stabilizable. Consequently, ensuring convergence through a straightforward extension of existing successive approximation methods is not feasible. As our third contribution, we rigorously establish the convergence of output-feedback PI and VI algorithms for nonlinear nonaffine disturbed systems. Additionally, analyzing the stability of the nonlinear disturbed system under the learned approximate optimal control policy presents a grand challenge. The practical stability of some RL-based control systems has been demonstrated by using the optimal value function and a positive definite function of states to establish a specific inequality. However, due to the stage cost not being positive definite with respect to states, finding a suitable function to satisfy this inequality poses a significant hurdle. To overcome this obstacle, we employ LaSalle's invariance principle, converse Lyapunov theorem, and the concept of zero-state observability to guarantee the practical stability of the closed-loop system. Finally, considering the unknown system dynamics, we have provided online PI and VI algorithms based on approximation techniques. The input and output data are collected and used for learning. For each algorithm, we construct approximators to learn the value function and the control policy. Notably, the learned control policy is output-feedback that implicitly includes the deadbeat observer.

Structure: The remainder of the paper is structured as follows. In Section II, we formulate the output-feedback optimal control problem and provide a background on nonlinear optimal control. The state reconstruction of the nonlinear discrete-time system is also included in Section II. The stability of the equilibrium point of the closed-loop system given an optimal output-feedback control is analyzed in Section III. Model-based PI and VI algorithms are proposed in Section IV, while their data-driven version is presented in Section V. Both convergence of the proposed data-driven algorithms and the stability of the system in closed-loop with the learned controllers are rigorously analyzed in Section V as well. To demonstrate the effectiveness of the proposed data-driven algorithms, simulation results of two nonlinear systems are shown in Section VI. Finally, Section VII concludes and talks about future work.

II. Problem Formulation and Preliminaries

A. Problem Formulation

Consider a class of nonlinear nonaffine discrete-time systems $\forall k \in \mathbb{N}^+$,

$$w_{k+1} = a(w_k), \tag{1}$$

$$\begin{aligned} x_{k+1} &= f(w_k, x_k, u_k), \\ y_k &= h(w_k, x_k), \end{aligned} \tag{2}$$

where $\mathbb{N}^+$ represents the set of nonnegative integers, $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}$ is the control input, $y_k \in \mathbb{R}$ is the output that is available for measurement. The external input $w_k \in \mathbb{R}^p$ is generated by the exosystem (1). The function $a : \mathbb{R}^p \to \mathbb{R}^p$ is locally Lipschitz vanishing at the origin.

Our control objective is to design an adaptive optimal controller given (1)-(2) via output-feedback to minimize the following cost functional

$$J = \sum_{k=0}^{\infty} M(y_k, u_k), \tag{3}$$

where the function $M : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous, positive definite and proper. Moreover, for any fixed $y, \lambda \in \mathbb{R}$, the set $\{u \in \mathbb{R} | M(y, u) \leq \lambda\}$ is compact.

Throughout this paper, the following standard assumptions are needed.

Assumption 1: There exists a positively invariant and compact set (Khalil, 2002, Section 4.2), $\mathbb{W} \subset \mathbb{R}^p$, with respect to the exosystem (1). □

Assumption 2: The functions $f : \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $h : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$ are locally Lipschitz satisfying the condition that $f(w, 0, 0) = 0, h(w, 0) = 0$ for any $w \in \mathbb{W}$. □

Assumption 3: For any $(w_0, x_0) \in \mathbb{W} \times \mathbb{R}^n$, there exists uniquely a sequence of control inputs $\boldsymbol{u} = \{u_k\}_{k=0}^{\infty}$ minimizing the cost (3) with respect to the system (1)-(2), i.e., $\min_{\boldsymbol{u}} \sum_{k=0}^{\infty} M(h(w_k, x_k), u_k) := V^\star(w_0, x_0)$, where $V^\star$ is the optimal value function. □

Assumption 4: The system (2), which takes (u_k, w_k) as the input, y_k as the outputs, and x_k as the state, is uniformly observable. □

Remark 1: Assumption 1 implies that if a solution to the exosystem (1) belongs to the set $\mathbb{W}$ at step $k = 0$, then it belongs to $\mathbb{W}$ at all future steps. Assumption 2 enables the formulated system model (2) to include time-varying multiplicative disturbances. For instance, given $f = x_{1,k} \sin(k) + x_{2,k} \cos(k) + u_k,\ k \in \mathbb{N}^+$, one can denote the time-varying multiplicative disturbance as $w_k = \begin{bmatrix} \sin(k) & \cos(k) \end{bmatrix}^{\mathrm{T}}$ which can be generated by a linear exosystem (1) with

$$a(w) = \begin{bmatrix} \cos(1) & \sin(1) \\ -\sin(1) & \cos(1) \end{bmatrix} w$$

and an initial condition $w_0 = \begin{bmatrix} 0 & 1 \end{bmatrix}^{\mathrm{T}}$. □

Remark 2: The condition $f(w, 0, 0) = 0, h(w, 0) = 0$ for any $w \in \mathbb{W}$ in Assumption 2 can be relaxed by $f(0, 0, 0) = 0, h(0, 0) = 0$ if the exosystem (1) is exponentially stable. Under the relaxed condition, a vanishing additive disturbance can be included in the system model (2). In this

setting, one needs to combine w_k and x_k as a new state, and then solve an output-feedback stabilization problem of the system (1)-(2) in an optimal sense. Similar to Granzotto et al. (2021), the Assumption 3 is made such that there exists a sequence of control inputs to minimize the cost (3). The Assumption 4 is made due to the output feedback nature of the controller. □

Remark 3: The control objective in this paper is essentially to address an asymptotic regulation problem in an adaptive optimal sense. As a special case of nonlinear output regulation problems (Huang, 2004), the asymptotic regulation problem concerns with designing controllers to achieve disturbance rejection. □

Remark 4: The equation (1) is a nonlinear autonomous system, which can generate rich forms of disturbances w_k such as constants, exponential, and sinusoidal signals with arbitrary fixed or time-varying frequencies and initial phases, and their combinations. Equations (1)-(2) are a class of nonlinear systems which are common to model engineering systems in practice, such as the suspension systems of vehicles, multi-machine power systems affected by sinusoidal disturbances, and Van del Pol oscillators. □

Based on Assumptions 1-3, one can find the optimal value function $V^\star(w_k, x_k)$ and the optimal control policy $u_k^\star := u^\star(w_k, x_k)$ such that the cost (3), given (1)-(2) is minimized

$$V^\star(w_k, x_k) = \min_{u_k} \left(M(y_k, u_k) + V^\star(w_{k+1}, x_{k+1}) \right), \tag{4}$$

$$u^\star(w_k, x_k) = \arg\min_{u_k} \left(M(y_k, u_k) + V^\star(w_{k+1}, x_{k+1}) \right) \tag{5}$$

where (4) is the discrete-time Hamilton-Jacobi-Bellman (HJB) equation.

B. State Reconstruction of Nonlinear Systems

Let $U_{[k-n,k-1]}$, $W_{[k-n,k-1]}$ and $Y_{[k-n,k-1]}$ denote vectors of n consecutive measurements, respectively, i.e.,

$$\begin{aligned} U_{[k-n,k-1]} &= \left[u_{k-n}, u_{k-n+1}, \cdots, u_{k-1}\right]^{\mathrm{T}}, \\ W_{[k-n,k-1]} &= \left[w_{k-n}^{\mathrm{T}}, w_{k-n+1}^{\mathrm{T}}, \cdots, w_{k-1}^{\mathrm{T}}\right]^{\mathrm{T}}, \\ Y_{[k-n,k-1]} &= \left[y_{k-n}, y_{k-n+1}, \cdots, y_{k-1}\right]^{\mathrm{T}}. \end{aligned} \tag{6}$$

For the purpose of simplicity, denote

$$\begin{aligned} f^{w,u}(x) :=& f(w, x, u), \\ h^{w}(x) :=& h(w, x). \end{aligned}$$

In this way, the state can be reconstructed by sequences of measurements

$$x_k = \Phi(x_{k-n}, W_{[k-n,k-1]}, U_{[k-n,k-1]}), \tag{7}$$

$$Y_{[k-n,k-1]} = H(x_{k-n}, W_{[k-n,k-1]}, U_{[k-n,k-1]}), \tag{8}$$

where

$$\begin{aligned} &\Phi(x_{k-n}, W_{[k-n,k-1]}, U_{[k-n,k-1]}) := \\ &\quad f^{w_{k-1},u_{k-1}} \circ f^{w_{k-2},u_{k-2}} \circ \cdots \circ f^{w_{k-n},u_{k-n}}(x_{k-n}), \\ &H(x_{k-n}, W_{[k-n,k-1]}, U_{[k-n,k-1]}) := \\ &\qquad \begin{bmatrix} h^{w_{k-n}}(x_{k-n}) \\ \vdots \\ h^{w_{k-1}} \circ f^{w_{k-2},u_{k-2}} \circ \cdots \circ f^{w_{k-n},u_{k-n}}(x_{k-n}) \end{bmatrix} \end{aligned}$$

with " $\circ$ " denoting the composition.

Definition 1: The system (2) is said to be locally uniformly observable (Moraal and Grizzle, 1995) with respect to four sets $\bar{\mathbb{W}} \subset \mathbb{R}^{np}$, $\mathbb{X} \subset \mathbb{R}^n$, $\mathbb{Y} \subset \mathbb{R}^n$, and $\mathbb{U} \subset \mathbb{R}^n$, if the mapping $H^\star : \mathbb{X} \times \bar{\mathbb{W}} \times \mathbb{U} \to \mathbb{Y} \times \bar{\mathbb{W}} \times \mathbb{U}$ by $(x, W, U) \to (H(x, W, U), W, U)$ is injective. It is uniformly observable if $\bar{\mathbb{W}}, \mathbb{X}, \mathbb{U}$, and $\mathbb{Y}$ are sets of real numbers with proper dimensions. □

Remark 5: Note that all observable linear systems are uniformly observable. Moreover, we will show in Proposition 1 the uniform observability of a class of nonlinear strict feedback systems. □

Proposition 1: Based on Definition 1, a class of nonlinear strict feedback systems $\forall k \in \mathbb{N}^+$ in the form of

$$\begin{aligned} x_{1,k+1} =& f_1(x_{1,k}) + g_1 x_{2,k}, \\ &\vdots \\ x_{i,k+1} =& f_i(x_{1,k}, x_{2,k}, \cdots, x_{i,k}) + g_i x_{i+1,k}, \\ &\vdots \\ x_{n,k+1} =& f_n(x_{1,k}, x_{2,k}, \cdots, x_{n,k}) + g_n u_k, \\ y_k =& x_{1,k}, \end{aligned} \tag{9}$$

are uniformly observable, where for $i = 1, 2, \cdots, n$, $g_i \neq 0$.

Proof. To show this fact, we first write

$$\begin{aligned} x_{1,k-n} =& y_{k-n}, \\ x_{2,k-n} =& \frac{1}{g_1}(x_{1,k-n+1} - f_1(x_{1,k-n})) \\ :=& \Gamma_2(Y_{[k-n,k-n+1]}), \\ &\vdots \\ x_{i+1,k-n} =& \frac{1}{g_i}(x_{i,k-n+1} - f_i) := \Gamma_{i+1}(Y_{[k-n,k-n+i]}), \\ &\vdots \\ x_{n,k-n} =& \frac{1}{g_{n-1}}(x_{n-1,k-n+1} - f_{n-1}) := \Gamma_n(Y_{[k-n,k-1]}). \end{aligned} \tag{10}$$

For any two sequences of outputs $Y^a_{[k-n,k-1]}, Y^b_{[k-n,k-1]}$, we can obtain their corresponding states at the time $k-n$, i.e., x^a_{k-n} and x^b_{k-n} based on (10). Moreover, it can be checked from (10) that any $Y^a_{[k-n,k-1]} = Y^b_{[k-n,k-1]}$ implies that $x^a_{k-n} = x^b_{k-n}$. This proves the uniform observability of (9). □

Given the Definition 1 and Assumption 4, one can further obtain the following Lemma to reconstruct the system state using retrospective input and output data.

Lemma 2.1: Under Assumption 4, there exists a function $\Theta : \mathbb{R}^{n(p+2)} \to \mathbb{R}^n$ such that, for any applied $U_{[k-n,k-1]} \in \mathbb{R}^n$, $W_{[k-n,k-1]} \in \mathbb{R}^{np}$ and $Y_{[k-n,k-1]} \in \mathbb{R}^n$, the following equation holds

$$x_k = \Theta(z_k), \tag{11}$$

where $z_k = \begin{bmatrix} Y^{\mathrm{T}}_{[k-n,k-1]} & W^{\mathrm{T}}_{[k-n,k-1]} & U^{\mathrm{T}}_{[k-n,k-1]} \end{bmatrix}^{\mathrm{T}}$.

Proof. See the appendix. □

Remark 6: The state reconstruction (11) is essentially a deadbeat observer (Moraal and Grizzle, 1995). It is true that, the implementation of the data-driven Algorithms 3-4 does not rely on the knowledge of functions Θ which is related to the system dynamics f and h. □

The uniform observability also implies the system (2) with state x_k is zero-state observable, which will be discussed in Lemma 2.2.

Lemma 2.2: Under Assumptions 1, 2 and 4, for any $w_0 \in \mathbb{W}$, the only solution to $x_{k+1} = f(w_k, x_k, 0),\ k \in \mathbb{N}^+$ that can stay identically in the set $\{h(w,x) = 0\}$ is the trivial solution $x_k \equiv 0$.

Proof. See the appendix. □

In the following lemma, we will show that the value function $V^\star$ is not always positive definite.

Lemma 2.3: Under Assumptions 1-3, given that the set $\mathbb{W}$ contains a nonzero element, then the function $V^\star(w,x)$ is not positive definite with respect to its arguments (w,x).

Proof. The function $V^\star(w,x)$ is said to be positive definite if $V^\star(0,0) = 0$ and $V^\star(w,x) > 0$ for any $(w,x) \neq (0,0)$. Choose a nonzero $w_0 \in \mathbb{W}$ and a $x_0 = 0$. Let $u_k = 0$ for any $k \in \mathbb{N}^+$. Then we have $x_k = 0$ and $h(w_k, x_k) = 0$ for any $k \in \mathbb{N}^+$, which implies that $V^\star(w_0, 0) = 0$ and thus $V^\star(w,x)$ is not positive definite. The proof is completed. □

III. Output-Feedback Design and Stability

Based on the solution of the HJB equation (4) and the result of the state reconstruction (11), we can design the following output-feedback optimal control policy,

$$\bar{u}_k^\star := \bar{u}^\star(w_k, z_k) = u^\star(w_k, \Theta(z_k)) \tag{12}$$

which is equivalent to (5). Let $\bar{\mathcal{P}}$ denote the set of all continuous functions from $\mathbb{W} \times \mathbb{R}^n$ to $\mathbb{R}$. Each function $V(w,x) \in \bar{\mathcal{P}}$ has the property that, for any fixed $\bar{w} \in \mathbb{W}$, and $V(\bar{w}, x)$ is a positive definite function with respect to the argument x.

In the following theorem, we provide a sufficient condition to ensure that a control policy stabilizes the equilibrium point of the system (2).

Theorem 3.1: Suppose that Assumptions 1-3 hold. For any $V \in \bar{\mathcal{P}}$ and any $u(w,x)$ such that

$$\begin{aligned} V(w,x) \geq & M(h(w,x),u) + V(a(w), f(w,x,u)), \\ & \forall (w,x) \in \mathbb{W} \times \mathbb{R}^n, \end{aligned} \tag{13}$$

the equilibrium point $x = 0$ of the system (2) with input u is asymptotically stable for any sequence $\{w_k\}_{k=0}^\infty$ starting from $w_0 \in \mathbb{W}$.

Proof. See the appendix. □

Remark 7: Based on the definition of $V^\star$ and Lemma 2.2, one can observe that $V^\star \in \bar{\mathcal{P}}$. It is thus verifiable that the control (12) is a stabilizing control policy due to the fact that (12) is equivalent to (5) and the pair $(V^\star, u^\star)$ satisfies (13). By the definition of $\bar{\mathcal{P}}$, it is verifiable that $V^\star$ is not necessarily a positive definite function considering the fact that $V^\star(w,0) = 0$ for any $w \in \mathbb{W}$. For instance, a function $V(w,x) = (1 + w^2)x^2 \in \bar{\mathcal{P}}$ is not positive definite as $V(w,0) = 0$ for any w. □

IV. Policy and Value Iterations

The optimal output-feedback control policy (12) developed in Section III is based on the solution of the HJB equation (4), which is computationally expensive to solve directly since there is no closed-form solution. In this section, we are going to present two successive approximation techniques, namely a PI and a VI, to approximate the optimal control policy and the corresponding value function.

Define the following two functions, termed here as Q-functions, with respect to the system (1)-(2),

$$\begin{aligned} \mathcal{Q}_i(w_k, x_k, u_k) :=& M(y_k, u_k) + V_i(w_{k+1}, x_{k+1}), \\ \bar{\mathcal{Q}}_i(w_k, z_k, u_k) :=& M(y_k, u_k) + V_i(w_{k+1}, \Theta(z_{k+1})), \end{aligned} \tag{14}$$

where i stands for the iteration instant. From Lemma 2.1, we observe that $\mathcal{Q}_i(w_k, x_k, u_k) = \bar{\mathcal{Q}}_i(w_k, z_k, u_k)$.

Based on the aforementioned Q-functions, we present Algorithm 1 which is a Q-learning algorithm. We prove the convergence of Algorithm 1 in Theorem 4.1.

Algorithm 1 Output-Feedback PI Algorithm

1: Choose an initial admissible control policy $\bar{u}_1(w_k, z_k)$. $i \leftarrow 1$.
2: Repeat Steps $3 - 5$.
3: **Policy Evaluation**: Solve Q-function $\bar{\mathcal{Q}}_i$ with $\bar{\mathcal{Q}}_i(w, 0, 0) = 0, \forall w \in \mathbb{W}$ from

$$\begin{aligned} &\bar{\mathcal{Q}}_i(w_k, z_k, u_k) = \\ &\bar{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \bar{u}_i(w_{k+1}, z_{k+1})) + M(y_k, u_k). \end{aligned} \tag{15}$$

4: **Policy Improvement**: Improve the control policy by

$$\bar{u}_{i+1}(w_k, z_k) = \arg\min_u \bar{\mathcal{Q}}_i(w_k, z_k, u). \tag{16}$$

5: $i \leftarrow i + 1$

Theorem 4.1: Suppose that Assumptions 1-4 hold. Given an admissible control policy $\bar{u}_1$, consider $\bar{\mathcal{Q}}_i$ and $\bar{u}_{i+1}$ defined by (15)-(16). Then for any $i = 1, 2, \cdots$, the following statements hold:

1) $\bar{\mathcal{Q}}^\star(w,z,u) \leq \bar{\mathcal{Q}}_{i+1}(w,z,u) \leq \bar{\mathcal{Q}}_i(w,z,u)$.
2) $\bar{u}_{i+1}$ is admissible.
3) For each fixed (w,z,u), $\{\bar{\mathcal{Q}}_i(w,z,u)\}_{i=0}^\infty$ and $\{\bar{u}_i(w,z)\}_{i=0}^\infty$ converge to $\bar{\mathcal{Q}}^\star(w,z,u)$ and $\bar{u}^\star(w,z)$, respectively.

Proof. See the appendix. □

The PI algorithm requires an admissible control policy to initialize. If this control policy is not available, one can follow the VI algorithm given in Algorithm 2 to approximate the optimal control policy and value function.

The following Theorem 4.2 shows the convergence of the output-feedback VI Algorithm 2.

Theorem 4.2: Suppose that Assumptions 1-4 hold. Consider $\bar{\mathcal{Q}}_i$ and $\bar{u}_i$ defined by (17)-(18). For any $i = 1, 2, \cdots$, the following statements hold:

1) $\bar{\mathcal{Q}}_{i+1}(w,z,u) \geq \bar{\mathcal{Q}}_i(w,z,u)$.
2) For each fixed (w,z,u), $\{\bar{\mathcal{Q}}_i(w,z,u)\}_{i=0}^\infty$ and $\{\bar{u}_i(w,z)\}_{i=0}^\infty$ converge to $\bar{\mathcal{Q}}^\star(w,z,u)$ and $\bar{u}^\star(w,z)$, respectively.

Algorithm 2 Output-Feedback VI Algorithm

1: Choose an initial Q-function as $\bar{\mathcal{Q}}_0 = M(y_k, u_k)$, an initial control policy as $\bar{u}_0 \equiv 0$, and $i \leftarrow 0$.
2: Repeat Steps $3-5$.
3: Update the Q-function $\bar{\mathcal{Q}}_{i+1}$ from

$$\bar{\mathcal{Q}}_{i+1}(w_k, z_k, u_k) = \bar{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \bar{u}_i(z_{k+1})) + M(y_k, u_k). \quad (17)$$

4: Update the control policy by

$$\bar{u}_{i+1}(w_k, z_k) = \arg\min_u \bar{\mathcal{Q}}_{i+1}(w_k, z_k, u). \quad (18)$$

5: $i \leftarrow i+1$.

Proof. See the appendix. □

Remark 8: To implement Algorithms 1-2, we need denote y_k and z_{k+1} in terms of h, Θ, z_k and u_k. Note that, besides h, function Θ depends on the knowledge of system dynamics f. Therefore, Algorithms 1-2 are model-based algorithms. □

V. Data-Driven Realization

In this section, we will develop two data-driven algorithms to implement the PI and VI Algorithms 1-2 with approximation techniques. The proposed algorithms are model-free and they learn both the optimal control policy and Q-functions without the knowledge of the system dynamics $f(\cdot,\cdot,\cdot)$, $h(\cdot,\cdot)$, and $a(\cdot)$.

The uniform observability of the system (1)-(2) implies the unique reconstruction from the retrospective input and output to the state. Hence, the optimal state-feedback and output-feedback control policies, i.e., $u^\star(w_k, x_k)$ and $\bar{u}^\star(w_k, z_k)$, and their corresponding Q-functions are equivalent.

The Q-function $\bar{\mathcal{Q}}_i(w,z,u)$ can represented by $\bar{\mathcal{Q}}_i(w,z,u) = \sum_{j=0}^{\infty} \bar{s}_{i,j}\psi_j(w,z,u)$, where $\{\psi_j(w,z,u)\}_{j=0}^{\infty}$ is a sequence of linearly independent smooth basis functions, and for any $j = 0, 1, 2, \cdots, \bar{s}_{i,j} \in \mathbb{R}$. For the purpose of implementing Algorithms 1-2 using online input and output data, we approximate the Q-function $\bar{\mathcal{Q}}_i(w,z,u)$ by

$$\hat{\mathcal{Q}}_i(w,z,u) = \sum_{j=0}^{N_1} \hat{s}_{i,j}\psi_j(w,z,u), \quad (19)$$

where the approximation error is

$$\begin{aligned}&\bar{\mathcal{Q}}_i(w,z,u) - \hat{\mathcal{Q}}_i(w,z,u)\\ =&\sum_{j=0}^{N_1}(\bar{s}_{i,j} - \hat{s}_{i,j})\psi_j(w,z,u) + \sum_{j=N_1+1}^{\infty}\bar{s}_{i,j}\psi_j(w,z,u),\end{aligned}$$

and N_1 is a sufficiently large integer.

One can then replace the Q-function and control policy using their approximations, $\hat{\mathcal{Q}}_i$ and $\hat{\bar{u}}_i$, in the step of policy evaluation (15) in PI Algorithm, which induces the following approximation error

$$\begin{aligned}e_{i,k} =&\hat{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1}))\\ &+ M(y_k, u_k) - \hat{\mathcal{Q}}_i(w_k, z_k, u_k)\\ =&\sum_{j=0}^{N_1}\hat{s}_{i,j}\psi_j\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right)\\ &+ M(y_k, u_k) - \sum_{j=0}^{N_1}\hat{s}_{i,j}\psi_j(w_k, z_k, u_k). \quad (20)\end{aligned}$$

The weights $\hat{s}_{i,j}$ can be obtained in terms of least squares solutions (minimizing $\sum_{k=0}^{l} e_{i,k}^2$) under the following assumption.

Assumption 5: There exist $l_1 > 0$ and $\delta_1 > 0$, such that for any $l \geq l_1$, we have

$$\frac{1}{l+1}\sum_{k=0}^{l}\kappa_{i,k}^{\mathrm{T}}\kappa_{i,k} \geq \delta_1 I_{N_1+1}, \quad (21)$$

where

$$\kappa_{i,k}^{\mathrm{T}} = \begin{bmatrix} \psi_0(w_k, z_k, u_k) - \psi_0\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right)\\ \psi_1(w_k, z_k, u_k) - \psi_1\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right)\\ \vdots\\ \psi_{N_1}(w_k, z_k, u_k) - \psi_{N_1}\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right)\end{bmatrix}.$$

□

Now, we are ready to present the data-driven PI algorithm, Algorithm 3, to approximate the optimal control policy and Q-function.

Algorithm 3 Data-Driven Output-Feedback PI Algorithm

1: Choose a sufficiently small threshold $\epsilon > 0$. Employ an admissible control input $\bar{u}_1(w_k, z_k)$ to collect input and output data online. $i \leftarrow 1$.
2: **repeat**
3: Solve $\hat{s}_{i,j}$ from (20).
4: Update the control policy by

$$\hat{\bar{u}}_{i+1}(w_k, z_k) = \arg\min_u \hat{\mathcal{Q}}_i(w_k, z_k, u). \quad (22)$$

5: $i \leftarrow i+1$
6: **until** $\sum_{j=0}^{N_1}|\hat{s}_{i,j} - \hat{s}_{i-1,j}|^2 < \epsilon$.

In order to realize the VI algorithm 2, one can also replace the Q-function and control policy in (17) by their approximations.

$$\begin{aligned}e_{i,k} =&\hat{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1}))\\ &+ M(y_k, u_k) - \hat{\mathcal{Q}}_{i+1}(w_k, z_k, u_k)\\ =&\hat{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1}))\\ &+ M(y_k, u_k) - \sum_{j=0}^{N_1}\hat{s}_{i+1,j}\psi_j(w_k, z_k, u_k). \quad (23)\end{aligned}$$

The weights $\hat{s}_{i,j}$ can be obtained in terms of least squares solutions under the following assumption.

Assumption 6: There exist $l_2 > 0$ and $\delta_2 > 0$, such that, for any $l \geq l_2$, we have

$$\frac{1}{l+1}\sum_{k=0}^{l}\Psi_k^{\mathrm{T}}\Psi_k \geq \delta_2 I_{N_1+1}, \quad (24)$$

where

$$\Psi_k^{\mathrm{T}} = \begin{bmatrix} \psi_0(w_k, z_k, u_k) \\ \psi_1(w_k, z_k, u_k) \\ \vdots \\ \psi_{N_1}(w_k, z_k, u_k) \end{bmatrix}.$$

□

The data-driven VI Algorithm 4 is given as follows.

Algorithm 4 Data-Driven Output-Feedback VI Algorithm

1: Choose a sufficiently small threshold $\epsilon > 0$. Employ a control input to collect data online. $\bar{\mathcal{Q}}_0 = M(y_k, u_k)$, $\bar{u}_0 = \hat{\bar{u}}_0 \equiv 0$, $i \leftarrow 0$.
2: **repeat**
3: Solve $\hat{s}_{i+1,j}$ from (23).
4: Update the control policy by

$$\hat{\bar{u}}_{i+1}(z_k) = \arg\min_u \hat{\mathcal{Q}}_{i+1}(w_k, z_k, u). \tag{25}$$

5: $i \leftarrow i + 1$
6: **until** $\sum_{j=0}^{N_1} |\hat{s}_{i,j} - \hat{s}_{i-1,j}|^2 < \epsilon$.

The following Lemma is useful to show the convergence of data-driven PI Algorithm 3.

Lemma 5.1: Suppose that Assumptions 1-5 hold for each iteration i. There exists a compact set $\mathbb{D}_w \times \mathbb{D}_z \times \mathbb{D}_u \subset \mathbb{W} \times \mathbb{R}^{n(p+2)} \times \mathbb{R}$ such that $\lim_{N_1 \to \infty} \hat{\mathcal{Q}}_i(w, z, u) = \tilde{\mathcal{Q}}_i(w, z, u)$, $\lim_{N_1 \to \infty} \hat{\bar{u}}_{i+1}(w, z) = \tilde{u}_{i+1}(w, z), \forall (w, z, u) \in \mathbb{D}_w \times \mathbb{D}_z \times \mathbb{D}_u$, where $\tilde{\mathcal{Q}}_i$ and $\tilde{u}_{i+1}$ are respectively the accurate Q-function and the control policy updated by policy iteration given the control policy $\hat{\bar{u}}_i$, which are defined as follows

$$\begin{aligned} \tilde{u}_{i+1}(w_k, z_k) =& \arg\min_u \tilde{\mathcal{Q}}_i(w_k, z_k, u), \\ M(y_k, u_k) =& \tilde{\mathcal{Q}}_i(w_k, z_k, u_k) \\ & - \tilde{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})). \end{aligned} \tag{26}$$

Proof. See the appendix. □

Now we are ready to show the convergence of the Algorithm 3 in Theorem 5.1.

Theorem 5.1: Suppose that Assumptions 1-5 hold. Given any $\epsilon_1 > 0$, there exist positive integers $i^\star$ and $N_1^\star$, and a compact set $\mathbb{D}_w \times \mathbb{D}_z \times \mathbb{D}_u \subset \mathbb{W} \times \mathbb{R}^{n(p+2)} \times \mathbb{R}$ such that

$$\begin{aligned} \left| \sum_{j=0}^{N_1} \hat{s}_{i,j} \psi_j(w, z, u) - \bar{\mathcal{Q}}^\star(w, z, u) \right| \leq \epsilon_1, \\ \left| \hat{\bar{u}}_{i+1}(w, z) - \bar{u}^\star(w, z) \right| \leq \epsilon_1 \end{aligned} \tag{27}$$

for any $i \geq i^\star$ and any $(w, z, u) \in \mathbb{D}_w \times \mathbb{D}_z \times \mathbb{D}_u$, if $N_1 > N_1^\star$.

Proof. We will prove by induction that, for each $i \geq 1$ and given $\epsilon > 0$, there exists a large enough integer $N_1^\star$ such that

$$\begin{aligned} |\hat{\mathcal{Q}}_i(w, z, u) - \bar{\mathcal{Q}}_i(w, z, u)| \leq \epsilon, \\ |\hat{\bar{u}}_{i+1}(w, z) - \bar{u}_{i+1}(w, z)| \leq \epsilon, \end{aligned} \tag{28}$$

if $N_1 > N_1^\star$.

1) For $i = 1$, one has $\tilde{\mathcal{Q}}_i(w, z, u) = \bar{\mathcal{Q}}_i(w, z, u)$ and $\tilde{u}_{i+1}(w, z) = \bar{u}_{i+1}(w, z)$. The convergence is provable by Lemma 5.1.
2) Assume for some $i = j > 1$ that

$$\begin{aligned} \lim_{N_1 \to \infty} \hat{\mathcal{Q}}_j(w, z, u) =& \bar{\mathcal{Q}}_j(w, z, u), \\ \lim_{N_1 \to \infty} \hat{\bar{u}}_{j+1}(w, z) =& \bar{u}_{j+1}(w, z). \end{aligned}$$

Define

$$\begin{aligned} R_{j+1,k} =& \tilde{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \hat{\bar{u}}_{j+1}(w_{k+1}, z_{k+1})) \\ & - \tilde{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1})). \end{aligned}$$

By policy evaluation (15) and (26), we have

$$\begin{aligned} R_{j+1,k} =& \tilde{\mathcal{Q}}_{j+1}(w_k, z_k, u_k) - M(y_k, u_k) \\ & - \tilde{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1})) \\ =& \tilde{\mathcal{Q}}_{j+1}(w_k, z_k, u_k) - \bar{\mathcal{Q}}_{j+1}(w_k, z_k, u_k) \\ & + \bar{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1})) \\ & - \tilde{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1})) \\ =& [\tilde{\mathcal{Q}}_{j+1}(w_k, z_k, u_k) - \bar{\mathcal{Q}}_{j+1}(w_k, z_k, u_k)] \\ & - [\tilde{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1})) \\ & - \bar{\mathcal{Q}}_{j+1}(w_{k+1}, z_{k+1}, \bar{u}_{j+1}(w_{k+1}, z_{k+1}))]. \end{aligned}$$

By definition, we can check that the error $R_{j+1,k} \to 0$ as $N_1 \to \infty$. By Assumption 5, we observe that,

$$\begin{aligned} 0 =& \lim_{N_1 \to \infty} |\tilde{\mathcal{Q}}_{j+1}(w, z, u) - \bar{\mathcal{Q}}_{j+1}(w, z, u)|, \\ 0 =& \lim_{N_1 \to \infty} |\tilde{u}_{j+2}(w, z) - \bar{u}_{j+2}(w, z)|. \end{aligned}$$

Based on Lemma 5.1, we can show (28) for $i := j + 1$.

By Theorem 4.1, there always exists a $i^\star > 0$, such that

$$\begin{aligned} |\bar{\mathcal{Q}}_i(w, z, u) - \bar{\mathcal{Q}}^\star(w, z, u)| \leq \epsilon, \\ |\bar{u}_{i+1}(w, z) - \bar{u}^\star(w, z)| \leq \epsilon, \quad \forall i \geq i^\star. \end{aligned}$$

The proof is thus completed using the triangle inequality and setting $\epsilon_1 = 2\epsilon$. □

Let $\hat{\bar{u}}^\dagger(w_k, z_k)$ be the approximated optimal control policy learned by Algorithms 3 or 4. Based on Lemma 2.1, one can find a state-feedback control policy $\hat{u}^\dagger(w_k, x_k)$ such that $\hat{u}^\dagger(w_k, x_k) = \hat{u}^\dagger(w_k, \Theta(z_k)) := \hat{\bar{u}}^\dagger(w_k, z_k)$. For any w_0, the system (2) in closed-loop with the controller $\hat{\bar{u}}^\dagger(w_k, z_k)$ learned by Algorithms 3 or 4 can be represented by

$$\begin{aligned} x_{k+1} =& f(w_k, x_k, \hat{\bar{u}}^\dagger(w_k, z_k)) \\ =& f(w_k, x_k, \hat{u}^\dagger(w_k, x_k)) \\ =& f(a^k(w_0), x_k, \hat{u}^\dagger(a^k(w_0), x_k)) \\ :=& F_{w_0}^\dagger(k, x_k) \end{aligned} \tag{29}$$

where $a^k(w) = \underbrace{a \circ a \circ \cdots \circ a}_{k}(w)$.

We will show that the learned near-optimal control policy $\hat{\bar{u}}^\dagger(w_k, z_k)$ solves the problem of practical stabilization for system (2).

Theorem 5.2: Under the conditions in Theorem 5.1, for any $w_0 \in \mathbb{W}$ and any $\epsilon_2 > 0$ satisfying that $B(\epsilon_2) \subset \mathbb{D}_x$, there

exists a continuous function V_c such that, along the trajectory of the closed-loop system (29), we have

$$\begin{aligned}&\beta_1(|x|) \leq V_c(k,x) \leq \beta_2(|x|), \quad \forall x \in \mathbb{D}_x \\ &V_c(k+1, F^{\dagger}_{w_0}(k,x)) - V_c(k,x) < 0, \quad \forall x \in \mathbb{D}_x \backslash B(\epsilon_2)\end{aligned} \tag{30}$$

where β_1, β_2 are two functions of class $\mathcal{K}$, $B(\epsilon_2) = \{x \in \mathbb{R}^n : |x| < \epsilon_2\}$, $\mathbb{D}_x = \{x \in \mathbb{R}^n : x = \Theta(z), z \in \mathbb{D}_z\}$.

Proof. See the appendix. □

Remark 9: The convergence of data-driven VI Algorithm 4 can be ensured by following the similar logic when proving the convergence of data-driven PI Algorithm 3. One can firstly show the fact in the Lemma 5.1 by slightly changing $\tilde{u}_i$ by $\tilde{u}_{i+1}(w_k, z_k) = \arg\min_u \tilde{\mathcal{Q}}_{i+1}(w_k, z_k, u)$. After that, one can show that there exists large enough positive integer $i^\star$ and $N_1^\star$ such that the learned Q-function and control policy from Algorithm 4 is close enough to the optimal ones. □

Remark 10: Algorithms 3-4 are implemented based on the online measurement of w_k. If w_k is not measurable while the exosystem (1) is uniformly observable, one can use the output of the exosystem (1) to reconstruct w_k. Alternatively, if the exosystem is linear, one can use the minimal polynomial of the exosystem dynamics to reconstruct w_k; see Gao and Jiang (2016) and the simulation example 2 in this paper.

Remark 11: The Assumptions 5-6 are similar to conditions of persistency of excitation in adaptive control; see Jiang and Jiang (2014); Vamvoudakis and Lewis (2010). Practically, one can add some exploration noise into the employed control input in order to excite the system during the data collection such that these assumptions are satisfied. Note that Algorithms 3-4 are essentially off-policy RL algorithms. Their convergence is not affected by the existence of exploration noises. □

VI. Simulations

A. Example 1

Consider the following discrete-time system $\forall k \in \mathbb{N}^+$ with external disturbance w_k,

$$\begin{aligned}w_{k+1} =& ew_k, \\ x_{k+1} =& ax_k^{\mu_1} + dw_k^{\mu_2} + bu_k^3, \\ y_k =& x_k,\end{aligned} \tag{31}$$

where $a, b, d \neq 0, |e| < 1$ are unknown real numbers and μ_1, μ_2 are unknown positive odd integers. This satisfies the relaxed condition in Remark 2 where $f(0,0,0) = 0$. Note that, if one treats both $w_k \in \mathbb{R}$ and $x_k \in \mathbb{R}$ as states, the system (31) can be converted as follows

$$\begin{aligned}x_{k+1} =& \begin{pmatrix} ex_{1,k} \\ dx_{1,k}^{\mu_1} + ax_{2,k}^{\mu_2} + bu_k^3 \end{pmatrix}, \\ y_k =& x_{2,k}.\end{aligned} \tag{32}$$

By Definition 1, the system (32) is uniformly observable. The state of (32) can be represented in terms of retrospective inputs and outputs $\forall k \in \mathbb{N}^+$, i.e.,

$$x_{1,k} = \frac{e^2}{d}(y_{k-1} - ay_{k-2}^{\mu_2} - bu_{k-2}^3)^{1/\mu_1},$$

$$x_{2,k} = ay_{k-1}^{\mu_2} + bu_{k-1}^3 + e^{\mu_1}(y_{k-1} - ay_{k-2}^{\mu_2} - bu_{k-2}^3).$$

The cost to be minimized is defined as $J = \sum_{k=0}^{\infty}(y_k^2 + u_k^6)$. For the purpose of simulation, we set $a = 0.9, b = 1, d = 1, e = 0.9, \mu_1 = 3, \mu_2 = 1$. The initial value of state is chosen by $x_0 = \begin{bmatrix}-5 & 2\end{bmatrix}^{\mathrm{T}}$. The initial control policy is $\bar{u}_1 = 0$, which is also a stabilizing control policy can be used by PI algorithms. The exploration noise is selected as $n_k = \sqrt[3]{0.4}[\sin(k)+\sin(3k)+\sin(5k)+\sin(10k)+\sin(20k)+\sin(30k)]^{1/3}$.

The following basis functions are chosen for the Q-function

$$\begin{aligned}&[y_{k-1}^2, y_{k-1}y_{k-2}, y_{k-1}u_{k-1}^3, y_{k-1}u_{k-2}^3, y_{k-1}u_k^3, \\ &y_{k-2}^2, y_{k-2}u_{k-1}^3, y_{k-2}u_{k-2}^3, y_{k-2}u_k^3, u_{k-1}^6, \\ &u_{k-1}^3u_{k-2}^3, u_{k-1}^3u_k^3, u_{k-2}^6, u_{k-2}^3u_k^3, u_k^6, \\ &y_{k-1}^4, y_{k-1}^2y_{k-2}^2, y_{k-1}^2u_{k-1}^6, y_{k-1}^2u_{k-2}^6, y_{k-1}^3u_k^3, \\ &y_{k-2}^4, y_{k-2}^2u_{k-1}^6, y_{k-2}^2u_{k-2}^6, y_{k-2}^3u_{k-2}^3, u_{k-1}^1 2, \\ &u_{k-1}^6u_{k-2}^6, u_{k-1}^9u_k^3, u_{k-2}^1 2, u_{k-2}^9u_k^3].\end{aligned}$$

We use the initial control policy with exploration noise as the control input to collect online input and output data for the time step k from 0 to 60. Both data-driven PI and VI Algorithms 3-4 are applied to learn the optimal control policy and the Q-function, and convergence is attained once $\sum_{j=0}^{28} |\hat{s}_{i,j} - \hat{s}_{i-1,j}|^2 < 0.005$.

The approximated optimal output-feedback control policy learned by PI Algorithm 3 is

$$\begin{aligned}u_k^3 =& 1.748y_{k-1} - 0.9532y_{k-2} + 0.7654u_{k-1}^3 - 1.059u_{k-2}^3 \\ &- 7.158 \times 10^{-11}y_{k-1}^3 + 5.196 \times 10^{-11}y_{k-2}^3 \\ &- 3.441 \times 10^{-9}u_{k-1}^9 + 3.576 \times 10^{-9}u_{k-2}^9,\end{aligned} \tag{33}$$

while the control policy learned by VI Algorithm 4 is

$$\begin{aligned}u_k^3 =& 1.748y_{k-1} - 0.9532y_{k-2} + 0.7654u_{k-1}^3 - 1.059u_{k-2}^3 \\ &- 7.487 \times 10^{-14}y_{k-1}^3 + 7.487 \times 10^{-11}y_{k-2}^3 \\ &- 1.25 \times 10^{-13}u_{k-1}^9 + 9.433 \times 10^{-14}u_{k-2}^9.\end{aligned} \tag{34}$$

Under the PI Algorithm 3, Figure 1 depicts the norm of weights for the Q-function at each iteration, and Figure 2 shows respectively the trajectories of the input and the output of the closed-loop system. Figure 3 shows the comparison of Q-functions at the first and last iteration under the PI Algorithm 3. The Q-function in this example is a function of 5 arguments, $y_{k-1}, y_{k-2}, u_{k-1}, u_{k-2}, u_k$. In order to plot this function, we vary the value of y_{k-2} and u_k, and set the value of other arguments as 0. One can see from the Figure 3 that the Q-function has decreased after 5 iterations. For the sake of comparison, Figure 4 has been depicted using VI Algorithm 4. We see that the stopping criterion is achieved in only 5 iterations using PI, while it needs 11 iterations for VI. Figure 5 shows the comparison of Q-functions at the first and last iteration under the VI Algorithm 4. Different from the PI Algorithm 3, the Q-function has increased after 11 iterations. Moreover, we find that the function $\hat{\mathcal{Q}}_{11}$ in the Figure 5 is very close to $\hat{\mathcal{Q}}_5$ in the Figure 3, which implies that the Q-function learned by the PI Algorithm 3 is consistent with that learned by the VI Algorithm 4.

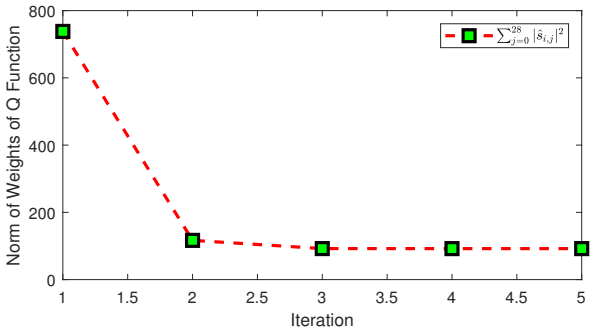

Fig. 1: Evolution of the weights of the Q-function under Algorithm 3.

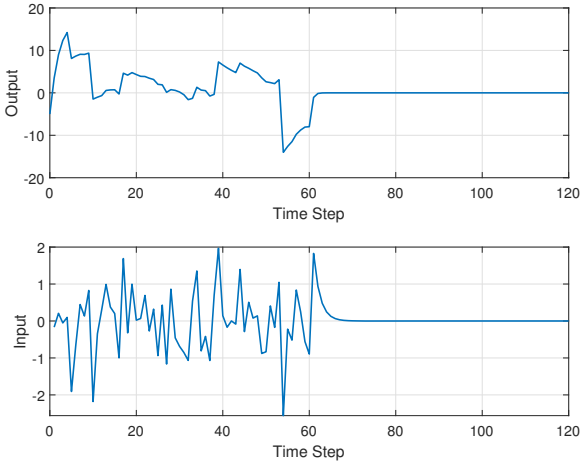

Fig. 2: Evolution of the system output and input trajectories under Algorithm 3.

We modify the cost as $J = \sum_{k=0}^{\infty}(y_k^2 + 0.005y_k^4 + u_k^6)$. By implementing the PI Algorithm 3, we obtain an approximated optimal control policy after 21 iterations

$$\begin{aligned} u_k^3 =& 1.689y_{k-1} - 0.9148y_{k-2} + 0.8162u_{k-1}^3 - 0.9538u_{k-2}^3 \\ &+ 1.924\times 10^{-4}y_{k-1}^3 + 3.662\times 10^{-4}y_{k-2}^3 \\ &+ 7.911\times 10^{-4}u_{k-1}^9 - 1.958\times 10^{-4}u_{k-2}^9, \end{aligned} \tag{35}$$

One can observe that the high-order basis functions in (35) affect more significantly than in (33)-(34). The corresponding evolution of weights of the Q function is shown in the Figure 6.

B. Example 2

In this example, we consider a Van der Pol oscillator (Byrnes et al., 1997; Gao and Jiang, 2018)

$$\begin{aligned} \dot{\xi}_1(t) =& \xi_2(t), \\ \dot{\xi}_2(t) =& -\xi_1(t) - 0.5\xi_2(t) - \xi_2^3(t) + \nu(t) \end{aligned}$$

with the exosystem being a harmonic oscillator

$$\begin{aligned} \dot{w}_1(t) =& w_2(t), \\ \dot{w}_2(t) =& -w_1(t). \end{aligned} \tag{36}$$

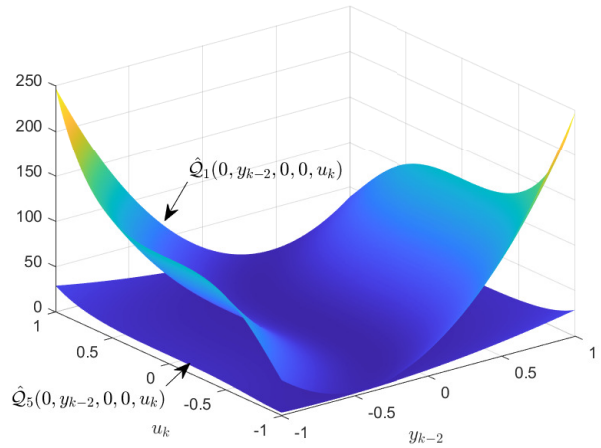

Fig. 3: Comparison of Q-functions at different iterations under Algorithm 3.

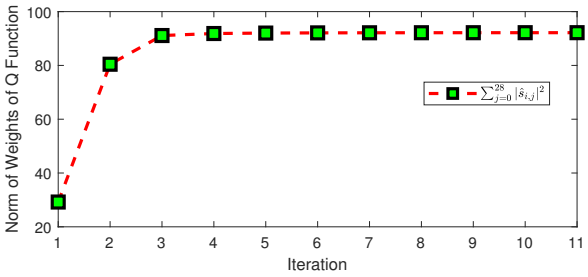

Fig. 4: Evolution of the weights of the Q-function under Algorithm 4.

Letting the states, input and output being $x_1 = \xi_1 - w_1$, $x_2 = \xi_2 - w_2$, $u = \nu - 0.5w_2 - w_2^3$ and $y = x_1$, we have

$$\begin{aligned} \dot{x}_1(t) =& x_2(t), \\ \dot{x}_2(t) =& -x_1(t) - 0.5x_2(t) - 3w_2^2(t)x_2(t) - 3w_2(t)x_2^3(t) \\ & - x_2^3(t) + u(t), \\ y(t) =& x_1(t). \end{aligned} \tag{37}$$

We discretize the exosystem (36) using zero-order holder method, and discretize the plant (37) via first-order Taylor method with the sampling period $T_s = 0.005s$. It is checkable that the discretized plant is uniformly observable. When deploying the learning Algorithm 3, it is important to recognize that the system dynamics are considered unknown. Both the state x and the external input w remain unmeasurable. The available data for online processing is limited to input and output measurements, as well as the frequency of the harmonic oscillator. In this setting, one can generate a vector $\hat{w}_k = [\cos(kT_s), \sin(kT_s)]^{\mathrm{T}}$ to ensure that there always exist a matrix G such that $w_k = G\hat{w}_k$ for any $k \in \mathbb{N}^+$. This will allow us to use the generated $\hat{w}_k$ (instead of w_k) along with online input and output measurement, u_k and y_k, for learning. For the purpose of simulation, the initial values of the state are $[x_{1,0}, x_{2,0}, w_{1,0}, w_{2,0}, \hat{w}_{1,0}, \hat{w}_{2,0}] = [2, -2, 0.5, -0.5, 1, 0]$. The cost function is $J = \sum_{k=0}^{\infty} 10^{-4}(y_k^2 + u_k^2)$. Similar to

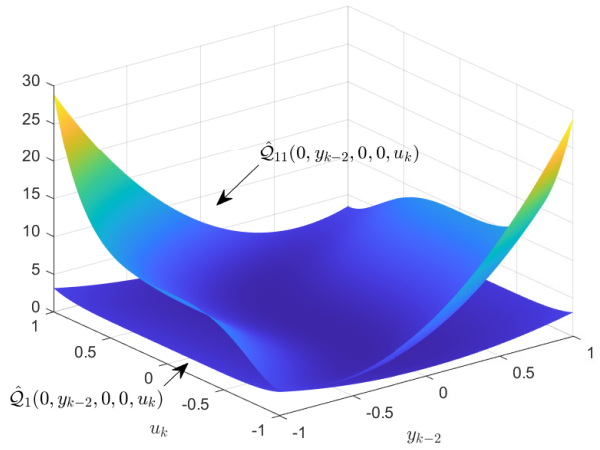

Fig. 5: Comparison of Q-functions at different iterations under Algorithm 4.

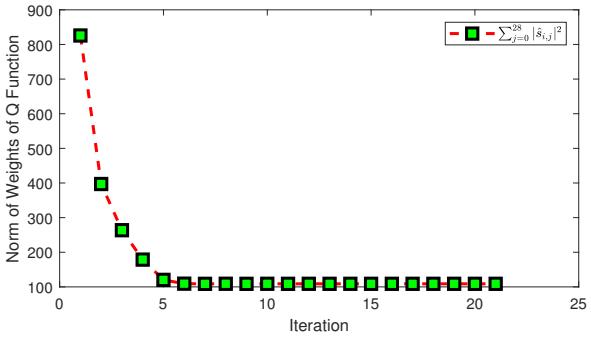

Fig. 6: Evolution of the weights of the Q-function under Algorithm 3 and the cost $J = \sum_{k=0}^{\infty}(y_k^2 + 0.005y_k^4 + u_k^6)$.

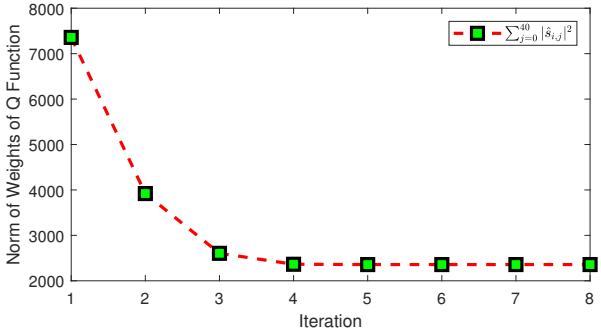

Fig. 7: Evolution of the weights of the Q-function of Example 2 under Algorithm 3.

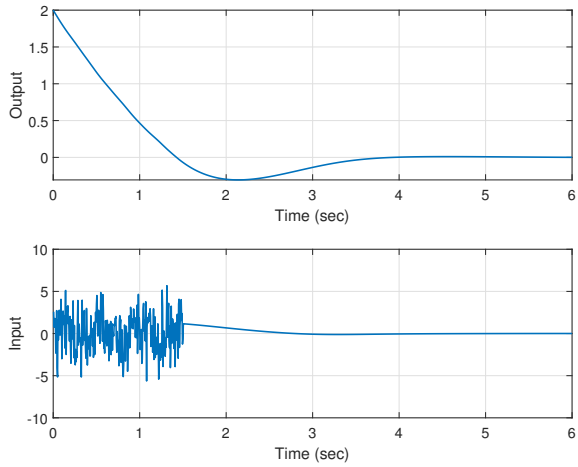

Fig. 8: System output and input trajectories of Example 2 under Algorithm 3.

Example 1, the exploration noise is chosen by a summation of sinusoidal signals with 10 different frequencies. The initial control policy is $\bar{u}_1 = 0$. We have chosen 41 polynomial functions as the basis for the Q-function. The input and output data are collected from $t = 0s$ to $t = 1.5s$ for learning the optimal output-feedback control policy. By using the PI Algorithm 3, the convergence is achieved after 8 iterations; see Figure 7. We have applied the learned control policy after $t = 1.5s$. The system input and output are depicted in Figure 8, where it can be seen that the learned control policy can regulate the output to the origin.

VII. Conclusion

This paper proposed fundamentally novel way to integrate reinforcement learning, output regulation, and nonlinear output-feedback theories to solve a longstanding open problem of output-feedback adaptive optimal control with disturbance rejection for nonlinear systems evolving in discrete-time. To overcome the challenge that the optimal value function is not positive definite, we have proposed novel online PI and VI algorithms with assured convergence. The whole process for implementation of the proposed algorithms does not rely on the knowledge of system dynamics and the state measurement. Simulation results show that both proposed online algorithms can approximate the optimal control policy and the value function based on only input and output data.

Appendix

Appendix. Proofs

Before proving the Lemma 2.1, we first show and prove the following Lemma.

Lemma A.1: Under Assumption 4, given retrospective measurements of the system (2), $(U_{[k-n,k-1]}, W_{[k-n,k-1]}, Y_{[k-n,k-1]})$, the state x_k is uniquely determined for any step $k \geq n$.

Proof Based on equations (7)-(8), if x_k and x_{k-n} are states of the system (2) at steps k and $k-n$, respectively, then the pair $(x, \underline{x}) := (x_k, x_{k-n})$ always solves the following equations

$$x = \Phi(\underline{x}, W_{[k-n,k-1]}, U_{[k-n,k-1]}), \tag{38}$$

$$Y_{[k-n,k-1]} = H(\underline{x}, W_{[k-n,k-1]}, U_{[k-n,k-1]}), \quad \forall k \geq n. \tag{39}$$

Based on Assumption 4 and Definition 1, $H^\star$ is an injective mapping, which implies that there exists at most one $\underline{x}$ satisfying the equation (39). One can further observe that there exists

at most one pair $(x, \underline{x})$ solving equations (38)-(39). Therefore, the solution of (38)-(39) is uniquely determined, which is equivalent to say that the state x_k is uniquely determined by $(U_{[k-n,k-1]}, W_{[k-n,k-1]}, Y_{[k-n,k-1]})$. The proof is thus completed. □

Lemma A.1 implies that the states can be determined uniquely based on retrospective inputs and outputs. We will find their relations explicitly. Definition 1 implies that, for any applied inputs $U_{[k-n,k-1]}, W_{[k-n,k-1]}$ and outputs $Y_{[k-n,k-1]}$, the state x_{k-n} can be uniquely solved by (8). In other words, there exist a function $\Theta_1 : \mathbb{R}^{n(p+2)} \to \mathbb{R}^n$ such that

$$x_{k-n} = \Theta_1(z_k), \quad \forall k \geq n. \tag{40}$$

From (7), one has,

$$\Theta(z_k) = \Phi(\Theta_1(z_k), W_{[k-n,k-1]}, U_{[k-n,k-1]}). \tag{41}$$

The proof is thus completed. □

It can be seen from (7)-(8) that

$$\begin{aligned} \Phi(0, W, 0) =& 0, \\ H(0, W, 0) =& 0, \end{aligned}$$

for any $W \in \underbrace{\mathbb{W} \times \cdots \times \mathbb{W}}_{n}$.

From Definition 1 and (41), one can see that

$$\Theta([0_n^{\mathrm{T}}, W, 0_n^{\mathrm{T}}]^{\mathrm{T}}) = 0$$

where 0_n is a zero column vector with n elements. It is true that any solution to $x_{k+1} = f(w_k, x_k, 0)$ that can stay identically in the set $\{h(w, x) = 0\}$ implies that $z_k = [0_n^{\mathrm{T}}, W_{[k-n,k-1]}, 0_n^{\mathrm{T}}]^{\mathrm{T}}$ for all k, which immediately shows that $x_k \equiv 0$. The proof is thus completed. □

By definition, for any fixed $\bar{w} \in \mathbb{W}$, there exist functions $\underline{\alpha}_{\bar{w}}, \bar{\alpha}_{\bar{w}}$ of class $\mathcal{K}$ such that $\underline{\alpha}_{\bar{w}}(|x|) \leq V(\bar{w}, x) \leq \bar{\alpha}_{\bar{w}}(|x|)$. Then, for any (w, x), there exist functions α_1, α_2 of class $\mathcal{K}$ such that

$$\begin{aligned} V(w, x) \leq& \max\{\bar{\alpha}_{\bar{w}}(|x|), \bar{w} \in \mathbb{W}\} := \alpha_2(|x|), \\ V(w, x) \geq& \min\{\underline{\alpha}_{\bar{w}}(|x|), \bar{w} \in \mathbb{W}\} := \alpha_1(|x|). \end{aligned} \tag{42}$$

The difference of the function V along the solutions of the system (1)-(2) with u yields,

$$\begin{aligned} \Delta V(w_k, x_k) =& V(w_{k+1}, x_{k+1}) - V(w_k, x_k) \\ \leq& -M(h(w_k, x_k), u_k) \leq 0. \end{aligned} \tag{43}$$

Choose constants $\gamma > 0$ and $c > 0$ such that the $\{|x| \leq \alpha_1^{-1}(c)\}$ is in the interior of a ball $B_\gamma(0) := \{x \in \mathbb{R}^n | \ |x| \leq \gamma\}$. Define a w-dependent set $\Omega_{w,c}$ by

$$\Omega_{w,c} = \{x \in B_\gamma(0) | V(w, x) \leq c\}.$$

Thus, for all $w \in \mathbb{W}$, we have

$$\{|x| \leq \alpha_2^{-1}(c)\} \subset \Omega_{w,c} \subset \{|x| \leq \alpha_1^{-1}(c)\} \subset B_\gamma(0). \tag{44}$$

Based on (43), it is clear that any solution starting in $(w, x) \in \{(w, x) \in \mathbb{W} \times \mathbb{R}^n | \ |x| \leq \alpha_2^{-1}(c)\} := \Omega$ stays in Ω. In other words, Ω is a positively invariant set with respect to the closed-loop system. Based on LaSalle's invariance principle (LaSalle, 1976), we have $\lim_{k\to\infty} \Delta V(w_k, x_k) = 0$.

Since $M \succ 0$, one has,

$$\begin{aligned} \Delta V^\star(w_k, x_k) = 0 \Rightarrow& h(w_k, x_k) = 0, u_k = 0 \\ \Rightarrow& x_{k+1} = f(w_k, x_k, 0), \ k \in \mathbb{N}^+. \end{aligned}$$

Finally, based on the zero-state observability of Lemma 2.2, one can conclude that $\lim_{k\to\infty} x_k = 0$ for any initial condition $(w_0, x_0) \in \Omega$.

The proof is thus completed. □

We prove 1) and 2) by induction.

- Let $i = 1$. Based on Lemma 2.1 and (14), we have $\bar{\mathcal{Q}}_1(w_k, z_k, u_k) = \mathcal{Q}_1(w_k, x_k, u_k)$ and $\bar{u}_1(w_{k+1}, z_{k+1}) = u_1(w_{k+1}, x_{k+1})$. Therefore, based on (14), equation (15) is equivalent to

$$\begin{aligned} 0 =& \mathcal{Q}_1(w_k, x_k, u_k) - \mathcal{Q}_1(w_{k+1}, x_{k+1}, u_1(w_{k+1}, x_{k+1})) \\ & - M(y_k, u_k) \\ =& V_1(w_{k+1}, x_{k+1}) - \mathcal{Q}_1(w_{k+1}, x_{k+1}, u_1(w_{k+1}, x_{k+1})) \\ =& V_1(w_{k+1}, x_{k+1}) - V_1(w_{k+2}, x_{k+2}) \\ & - M(y_{k+1}, u_1(w_{k+1}, x_{k+1})) \end{aligned} \tag{45}$$

By shifting the time index k by 1, we immediately have

$$V_1(w_k, x_k) = V_1(w_{k+1}, x_{k+1}) + M(y_k, u_1(w_k, x_k)) \tag{46}$$

where $x_{k+1} = f_k(u_1), \ k \in \mathbb{N}^+$. For simplicity, we denote $f_k(u) = f(w_k, x_k, u(w_k, x_k))$.

Along the solution of system (1)-(2) with u_1, we have $V_1(w_0, x_0) = \sum_{k=0}^{\infty} M(y_k, u_1(w_k, x_k))$. It can be checked that $V^\star \leq V_1$. Based on zero-state observability (see Lemma 2.2) and the fact that u_1 is an admissible control policy, we have $V_1(w, 0) = 0$, for any $w \in \mathbb{W}$. And $0 < V_1(w, x) < \infty$ for any nonzero and finite $x \in \mathbb{R}^n$, which implies that $V_1 \in \bar{\mathcal{P}}$.

Based on (16), we have

$$\begin{aligned} u_2(w_k, x_k) =& \bar{u}_2(w_k, z_k) \\ =& \arg\min_u \bar{\mathcal{Q}}_1(w_k, z_k, u) \\ =& \arg\min_u \mathcal{Q}_1(w_k, x_k, u). \end{aligned} \tag{47}$$

Through the definition of the value function V_1 and the Q-function $\mathcal{Q}_1$, for any $(w_k, x_k) \in \mathbb{W} \times \mathbb{R}^n$, one has the following inequality

$$\begin{aligned} 0 \leq& \mathcal{Q}_1(w_k, x_k, u_1) - \mathcal{Q}_1(w_k, x_k, u_2) \\ =& M(y_k, u_1) + V_1(a(w_k), f_k(u_1)) \\ & - M(y_k, u_2) - V_1(a(w_k), f_k(u_2)) \\ =& V_1(w_k, x_k) - M(y_k, u_2) - V_1(a(w_k), f_k(u_2)) \end{aligned} \tag{48}$$

which implies that the condition (13) holds.

From Theorem 3.1, the system (2) with u_2, i.e.,

$$\begin{aligned} x_{k+1} =& f_k(w_k, x_k, u_2(w_k, x_k)), \\ y_k =& h(w_k, x_k) \end{aligned} \tag{49}$$

is asymptotically stable at the origin for any sequences $\{w_k\}_{k=0}^{\infty}$ starting at $w_0 \in \mathbb{W}$. It implies that , for any $(w_0, x_0) \in \mathbb{W} \times \mathbb{R}^n$, the solution of (49) satisfies

$\lim_{k\to\infty} x_k = 0$, and thus $\lim_{k\to\infty} V_1(w_k, x_k) = 0$. Furthermore, based on the step of the policy evaluation (15) and its equivalency (46), at the second iteration, we have

$$V_2(w_k, x_k) - V_2(a(w_k), f_k(u_2)) = M(y_k, u_2). \quad (50)$$

Along the solution of systems (1) and (49), we have

$$\begin{aligned}
&V_1(w_0, x_0) - V_2(w_0, x_0)\\
=&\left[V_1(w_0, x_0) - \lim_{k\to\infty} V_1(w_k, x_k)\right]\\
&-\left[V_2(w_0, x_0) - \lim_{k\to\infty} V_2(w_k, x_k)\right]\\
=&\sum_{k=0}^{\infty}[V_1(w_k, x_k) - V_1(a(w_k), f_k(u_2))]\\
&-\sum_{k=0}^{\infty}[V_2(w_k, x_k) - V_2(a(w_k), f_k(u_2))]\\
=&\sum_{k=0}^{\infty}[V_1(w_k, x_k) - V_1(a(w_k), f_k(u_2))\\
&\qquad -M(y_k, u_2)]\\
\geq&0. \qquad (51)
\end{aligned}$$

We immediately obtain $V^\star(w, x) \leq V_2(w, x) \leq V_1(w, x)$, $V_2 \in \bar{\mathcal{P}}$ and $u_2(w, x) = \bar{u}_2(w, z)$ is admissible. This indicates that $\mathcal{Q}^\star(w, x, u) \leq \mathcal{Q}_2(w, x, u) \leq \mathcal{Q}_1(w, x, u)$ and $\bar{\mathcal{Q}}^\star(w, z, u) \leq \bar{\mathcal{Q}}_2(w, z, u) \leq \bar{\mathcal{Q}}_1(w, z, u)$.

- Suppose that 1) and 2) hold for $i = j \geq 1$, and $V_{j+1} \in \bar{\mathcal{P}}$. Similar to (47)-(48), we see that the condition (13) holds for u_{j+2}:

$$\begin{aligned}
&V_{j+1}(w_k, x_k) - M(y_k, u_{j+2}) - V_{j+1}(a(w_k), f_k(u_{j+2}))\\
&\geq 0.
\end{aligned}$$

From Theorem 3.1, the system (2) with u_{j+2} is asymptotically stable at the origin for any sequences $\{w_k\}_{k=0}^{\infty}$ starting at $w_0 \in \mathbb{W}$. Along the trajectory of this system, it is able to obtain

$$V_{j+1}(w_0, x_0) - V_{j+2}(w_0, x_0) \geq 0.$$

which implies that $V^\star(w, x) \leq V_{j+2}(w, x) \leq V_{j+1}(w, x)$, $V_{j+2} \in \bar{\mathcal{P}}$, $u_{j+2}(w, x) = \bar{u}_{j+2}(w, z)$ is admissible, $\mathcal{Q}^\star(w, x, u) \leq \mathcal{Q}_{j+2}(w, x, u) \leq \mathcal{Q}_{j+1}(w, x, u)$ and $\bar{\mathcal{Q}}^\star(w, z, u) \leq \bar{\mathcal{Q}}_{j+2}(w, z, u) \leq \bar{\mathcal{Q}}_{j+1}(w, z, u)$. Hence, 1) and 2) hold for $i = j + 1$.

In order to prove 3), by choosing sequentially the control input by (16), we see that the corresponding value functions form a uniformly converging monotonic sequence bounded by $V^\star$:

$$V_0 \geq V_1 \geq V_2 \cdots \geq V^\star.$$

Define the limit of the convergent sequence $\{V_i\}$ by V_e. We can see that V_e solves (4). Also, since $V^\star \leq V_e \leq V_1$, $V_e \in \bar{\mathcal{P}}$.

In the following Lemma, we will discuss the uniqueness of the solution to (4).

Lemma A.2: Under the Assumption 3, $V^\star$ is the unique solution to the HJB equation (4) on $\bar{\mathcal{P}}$.

Proof. The proof will be by contradiction. Suppose that there exists a function $V^a \neq V^\star$ on $\bar{\mathcal{P}}$ solving (4). Then we have

$$\begin{aligned}
V^a(w_k, x_k) =&\min_{u_k}\left(M(y_k, u_k) + V^a(w_{k+1}, x_{k+1})\right),\\
u^a(w_k, x_k) =&\arg\min_{u_k}\left(M(y_k, u_k) + V^a(w_{k+1}, x_{k+1})\right).
\end{aligned}$$

Along with (1)-(2) that correspond to an arbitrary control sequence $\{u_k^b\}_{k=0}^{\infty}$ in order for the cost (3) to be finite, we have

$$V_a(w_0, x_0) \leq \sum_{k=0}^{\infty} M(y_k, u_k^b),$$

which implies that $V_a(w_0, x_0) = \min_{\boldsymbol{u}} \sum_{k=0}^{\infty} M(y_k, u_k) := V^\star(w_0, x_0)$. This contradicts with $V^a \neq V^\star$. The proof is thus completed. □

By Lemma A.2, we have $V^\star = V_e$. By the definition of Q-function and Lemma 2.1, we can show the statement 3). The proof is thus completed. □

The system (1)-(2) can be rewritten as a combined nonlinear systems regarding (w_k, x_k) as the state and u_k as the input. Based on Lemma 2.1 and (14), we have $\bar{\mathcal{Q}}_i(w_k, z_k, u_k) = \mathcal{Q}_i(w_k, x_k, u_k)$ and $\bar{u}_i(w_{k+1}, z_{k+1}) = u_i(w_{k+1}, x_{k+1})$. Therefore, equation (17) is equivalent to

$$\begin{aligned}
0 =&\mathcal{Q}_{i+1}(w_k, x_k, u_k) - \mathcal{Q}_i(w_{k+1}, x_{k+1}, u_i(w_{k+1}, x_{k+1}))\\
&- M(y_k, u_k)\\
=&V_{i+1}(w_{k+1}, x_{k+1}) - \mathcal{Q}_i(w_{k+1}, x_{k+1}, u_i(w_{k+1}, x_{k+1}))\\
=&V_{i+1}(w_{k+1}, x_{k+1}) - V_i(w_{k+2}, x_{k+2})\\
&- M(y_{k+1}, u_i(w_{k+1}, x_{k+1})).
\end{aligned}$$

By shifting the time index k by 1, we immediately have

$$V_{i+1}(w_k, x_k) = V_i(w_{k+1}, x_{k+1}) + M(y_k, u_i) \quad (52)$$

where $x_{k+1} = f_k(u_i),\ k \in \mathbb{N}^+$.

Based on (18), we can write

$$\begin{aligned}
u_i(w_k, x_k) =&\bar{u}_i(w_k, z_k)\\
=&\arg\min_u \bar{\mathcal{Q}}_i(w_k, z_k, u)\\
=&\arg\min_u \mathcal{Q}_i(w_k, x_k, u)\\
=&\arg\min_u (M(y_k, u) + V_i(a(w_k), f_k(u))) \quad (53)
\end{aligned}$$

which is consistent with the VI algorithm presented in Bertsekas (2017).

Moreover, based on the definition of function M, it can be observed that, for any $(w, x) \in \mathbb{W} \times \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, the set $\{u \in \mathbb{R} | M(h(w, x), u) \leq \lambda\}$ is a compact subset of $\mathbb{R}$, which meets the compactness assumption in Bertsekas (2017). It has been shown in Bertsekas (2017) that the VI sequence $\{V_i\}_{i=0}^{\infty}$ converges pointwise to $V^\star$. Finally, based on Lemma 2.1 and (14), the convergence of the Q-function $\bar{\mathcal{Q}}_i$ and the output-feedback control policy $\bar{u}_i$ can also be ensured. The proof is thus completed. □

For $j = 0, 1, \cdots, N_1$, let $\tilde{s}_{i,j}$ be constant weights such that

$$\tilde{\mathcal{Q}}_i(w, z, u) = \sum_{j=0}^{\infty} \tilde{s}_{i,j}\psi_j(w, z, u). \quad (54)$$

Then, by combining (20) and (26), we have

$$\begin{aligned}
e_{i,k} =&\hat{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})) \\
&+ M(y_k, u_k) - \hat{\mathcal{Q}}_i(w_k, z_k, u_k) \\
=&\tilde{\mathcal{Q}}_i(w_k, z_k, u_k) - \hat{\mathcal{Q}}_i(w_k, z_k, u_k) \\
&+ \hat{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})) \\
&- \tilde{\mathcal{Q}}_i(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})) \\
=&\sum_{j=0}^{\infty} \tilde{s}_{i,j}\psi_j(w_k, z_k, u_k) - \sum_{j=0}^{N_1} \hat{s}_{i,j}\psi_j(w_k, z_k, u_k) \\
&+ \sum_{j=0}^{N_1} \hat{s}_{i,j}\psi_j\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right) \\
&- \sum_{j=0}^{\infty} \tilde{s}_{i,j}\psi_j\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right). \\
=&\kappa_{i,k}\Lambda_i + \zeta_{i,k}. \qquad (55)
\end{aligned}$$

where

$$\begin{aligned}
\Lambda_i =&[\tilde{s}_{i,0}, \cdots, \tilde{s}_{i,N_1}]^{\mathrm{T}} - [\hat{s}_{i,0}, \cdots, \hat{s}_{i,N_1}]^{\mathrm{T}}, \\
\zeta_{i,k} =& \sum_{j=N_1+1}^{\infty} \tilde{s}_{i,j}\psi_j(w_k, z_k, u_k) \\
&- \sum_{j=N_1+1}^{\infty} \tilde{s}_{i,j}\psi_j\left(w_{k+1}, z_{k+1}, \hat{\bar{u}}_i(w_{k+1}, z_{k+1})\right).
\end{aligned}$$

By collecting online data from $k = 0$ to $k = l$, one can construct the following equation based on (20)

$$\begin{bmatrix} e_{i,0} \\ e_{i,1} \\ \vdots \\ e_{i,l} \end{bmatrix} = -\underbrace{\begin{bmatrix} \kappa_{i,0} \\ \kappa_{i,1} \\ \vdots \\ \kappa_{i,l} \end{bmatrix}}_{\mathcal{A}} \underbrace{\begin{bmatrix} \hat{s}_{i,0} \\ \hat{s}_{i,1} \\ \vdots \\ \hat{s}_{i,N_1} \end{bmatrix}}_{\mathcal{X}} + \underbrace{\begin{bmatrix} M(y_0, u_0) \\ M(y_1, u_1) \\ \vdots \\ M(y_l, u_l) \end{bmatrix}}_{\mathcal{B}}. \qquad (56)$$

As noted after (20), the weights $\hat{s}_{i,0}, \cdots, \hat{s}_{i,N_1}$ can be found using least-squares, which is equivalent to determine the vector $\mathcal{X}$ such that $|\mathcal{A}\mathcal{X} - \mathcal{B}|^2 = \sum_{k=0}^{l} e_{i,k}^2$ is minimized. In this setting, one can guarantee based on (55) that the resultant sum of squares of errors satisfy

$$\begin{aligned}
\sum_{k=0}^{l} e_{i,k}^2 =& \min_{\hat{s}_{i,0}, \cdots, \hat{s}_{i,N_1}} \sum_{k=0}^{l} (\kappa_{i,k}\Lambda_i + \zeta_{i,k})^2 \\
&\leq \sum_{k=0}^{l} \zeta_{i,k}^2. \qquad (57)
\end{aligned}$$

Then, given the fact that $\sum_{k=0}^{l} \Lambda_i^{\mathrm{T}} \kappa_{i,k}^{\mathrm{T}} \kappa_{i,k} \Lambda_i = \sum_{k=0}^{l} (e_{i,k} - \zeta_{i,k})^2 \leq 4(l+1) \max_{0 \leq k \leq l} \zeta_{i,k}^2$, one can observe that

$$|\Lambda_i|^2 \leq \frac{4}{\delta_1} \max_{0 \leq k \leq l} \zeta_{i,k}^2.$$

It can be seen that $\lim_{N_1 \to \infty} \zeta_{i,k} = 0$. Therefore, we have $\lim_{N_1 \to \infty} \tilde{\mathcal{Q}}_i(w, z, u) - \hat{\mathcal{Q}}_i(w, z, u) = 0$, for any $(w, z, u) \in \mathbb{D}_w \times \mathbb{D}_z \times \mathbb{D}_u$. Based on the definition of $\hat{\bar{u}}_{i+1}$ and $\tilde{u}_{i+1}$, one can further have $\lim_{N_1 \to \infty} \hat{\bar{u}}_{i+1}(w, z) - \tilde{u}_{i+1}(w, z) = 0$. The proof is thus completed. □

First, we rewrite the system (2) in closed-loop with the optimal control policy $u^\star(w_k, x_k)$ as

$$\begin{aligned}
x_{k+1} =& f(w_k, x_k, u^\star(w_k, x_k)) \\
=& f(a^k(w_0), x_k, u^\star(a^k(w_0), x_k)) \\
:=& F_{w_0}^\star(k, x_k). \qquad (58)
\end{aligned}$$

Based on Theorem 3.1 and Khalil (2002), one can observe that the system (58) is uniformly asymptotically stable for any $w_0 \in \mathbb{W}$. By using the converse Lyapunov theorem for discrete-time systems (Jiang and Wang, 2002), there exists a smooth Lyapunov function V_c such that

$$\begin{aligned}
&\beta_1(|x|) \leq V_c(k, x) \leq \beta_2(|x|), \\
&V_c(k+1, F_{w_0}^\star(k, x)) - V_c(k, x) < -\beta_3(|x|),
\end{aligned}$$

for some $\beta_1, \beta_2, \beta_3$ of class $\mathcal{K}$.

As V_c is smooth and f is locally Lipschitz, there always exists a $L_1 > 0$ such that $|V_c(k+1, F_{w_0}^\star(k, x)) - V_c(k+1, F_{w_0}^\dagger(k, x))| \leq L_1|\hat{u}^\dagger(x) - u^\star(x)| = L_1|\hat{\bar{u}}^\dagger(z) - \bar{u}^\star(z)| \leq L_1\epsilon_1$ for any $x \in \mathbb{D}_x$ and $z \in \mathbb{D}_z$, which implies that

$$V_c(k+1, F_{w_0}^\dagger(k, x)) - V_c(k, x) < -\beta_3(|x|) + 2L_1\epsilon_1.$$

For any $\epsilon_2 > 0$, we can render a sufficiently small ϵ_1 so that $\beta_3(|x|) > 2L_1\epsilon_1$ for any $x \in \mathbb{D}_x \backslash B(\epsilon_2)$. The proof is thus completed.

REFERENCES

Al-Tamimi, A., Lewis, F.L., Abu-Khalaf, M., 2008. Discrete-time nonlinear hjb solution using approximate dynamic programming: Convergence proof. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 38, 943–949.

Bertsekas, D.P., 2017. Value and policy iterations in optimal control and adaptive dynamic programming. IEEE Transactions on Neural Networks and Learning Systems 28, 500–509.

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

Bian, T., Jiang, Z.P., 2022. Reinforcement learning and adaptive optimal control for continuous-time nonlinear systems: A value iteration approach. IEEE Transactions on Neural Networks and Learning Systems 33, 2781–2790.

Byrnes, C.I., Priscoli, F.D., Isidori, A., Kang, W., 1997. Structurally stable output regulation of nonlinear systems. Automatica 33, 369 – 385.

Gao, W., Jiang, Y., Jiang, Z.P., Chai, T., 2016. Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. Automatica 72, 37–45.

Gao, W., Jiang, Z.P., 2016. Adaptive dynamic programming and adaptive optimal output regulation of linear systems. IEEE Transactions on Automatic Control 61, 4164–4169.

Gao, W., Jiang, Z.P., 2018. Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems. IEEE

Transactions on Neural Networks and Learning Systems 29, 2614–2624.

Gao, W., Jiang, Z.P., 2022. Learning-based adaptive optimal output regulation of linear and nonlinear systems: An overview. Control Theory and Technology, in press .

Gao, W., Jiang, Z.P., Lewis, F.L., Wang, Y., 2018. Leader-to-formation stability of multi-agent systems: An adaptive optimal control approach. IEEE Transactions on Automatic Control 63, 3581 – 3587.

Gao, W., Jiang, Z.P., Ozbay, K., 2017. Data-driven adaptive optimal control of connected vehicles. IEEE Transactions on Intelligent Transportation Systems 18, 1122–1133.

Granzotto, M., Postoyan, R., Busoniu, L., Nesic, D., Daafouz, J., 2021. Finite-horizon discounted optimal control: Stability and performance. IEEE Transactions on Automatic Control 66, 550–565.

He, P., Jagannathan, S., 2005. Reinforcement learning-based output feedback control of nonlinear systems with input constraints. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 35, 150–154.

Huang, J., 2004. Nonlinear Output Regulation: Theory and Applications. SIAM, Philadelphia, PA.

Jiang, H., Zhou, B., Duan, G.R., 2024. Modified λ-policy iteration based adaptive dynamic programming for unknown discrete-time linear systems. IEEE Transactions on Neural Networks and Learning Systems 35, 3291–3301.

Jiang, Y., Chai, T., Chen, G., 2025. Output feedback-based adaptive optimal output regulation for continuous-time strict-feedback nonlinear systems. IEEE Transactions on Automatic Control 70, 767–782. doi:10.1109/TAC.2024.3441668.

Jiang, Y., Fan, J., Chai, T., Li, J., Lewis, F.L., 2018. Data-driven flotation industrial process operational optimal control based on reinforcement learning. IEEE Transactions on Industrial Informatics 14, 1974–1989.

Jiang, Y., Jiang, Z.P., 2012. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica 48, 2699–2704.

Jiang, Y., Jiang, Z.P., 2014. Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. IEEE Transactions on Neural Networks and Learning Systems 25, 882–893.

Jiang, Z.P., Wang, Y., 2002. A converse lyapunov theorem for discrete-time systems with disturbances. Systems & Control Letters 45, 49–58.

Kamalapurkar, R., Dinh, H., Bhasin, S., Dixon, W.E., 2015. Approximate optimal trajectory tracking for continuous-time nonlinear systems. Automatica 51, 40–48.

Khalil, H.K., 2002. Nonlinear Systems. 3rd ed., Prentice Hall PTR, NJ.

Kleinman, D., 1968. On an iterative technique for Riccati equation computations. IEEE Transactions on Automatic Control 13, 114–115.

Kontoudis, G.P., Vamvoudakis, K.G., 2019. Kinodynamic motion planning with continuous-time Q-learning: An online, model-free, and safe navigation framework. IEEE Transactions on Neural Networks and Learning Systems 30, 3803–3817.

LaSalle, J., 1976. The Stabilty of Dynamical Systems. SIAM, Philadelphia, PA.

Lewis, F.L., Vamvoudakis, K.G., 2011. Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41, 14–25.

Modares, H., Lewis, F.L., Jiang, Z., 2016. Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning. IEEE Transactions on Cybernetics 46, 2401–2410.

Moraal, P.E., Grizzle, J.W., 1995. Observer design for nonlinear systems with discrete-time measurements. IEEE Transactions on Automatic Control 40, 395–404.

Mukherjee, S., Chakrabortty, A., Bai, H., 2019. Block-decentralized model-free reinforcement learning control of two time-scale networks, in: 2019 American Control Conference (ACC), pp. 2233–2238.

Ni, Z., Paul, S., 2019. A multistage game in smart grid security: A reinforcement learning solution. IEEE Transactions on Neural Networks and Learning Systems 30, 2684–2695.

Odekunle, A., Gao, W., Davari, M., Jiang, Z.P., 2020. Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems. Automatica 112.

Powell, W.B., 2007. Approximate Dynamic Programming: Solving the curse of dimensionality. John Wiley & Sons, New York, NY.

Qasem, O., Gao, W., Vamvoudakis, K.G., 2023. Adaptive optimal control of continuous-time nonlinear affine systems via hybrid iteration. Automatica 157, 111261.

Rizvi, S.A.A., Lin, Z., 2019. Reinforcement learning-based linear quadratic regulation of continuous-time systems using dynamic output feedback. IEEE Transactions on Cybernetics, in press .

Sandell, N., 1974. On newton's method for riccati equation solution. IEEE Transactions on Automatic Control 19, 254–255.

Saridis, G.N., Lee, C.S.G., 1979. An approximation theory of optimal control for trainable manipulators. IEEE Transactions on Systems, Man and Cybernetics 9, 152–159.

Sutton, R., Barto, A.G., Williams, R.J., 1992. Reinforcement learning is direct adaptive optimal control. IEEE Control Systems Magazine 12, 19–22.

Tang, F., He, H., Wen, J., Liu, J., 2015. Power system stability control for a wind farm based on adaptive dynamic programming. IEEE Transactions on Smart Grid 6, 166–177.

Valadbeigi, A.P., Sedigh, A.K., Lewis, F.L., 2020. H_∞ static output-feedback control design for discrete-time systems using reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems 31, 396–406.

Vamvoudakis, K.G., Lewis, F.L., 2010. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 46, 878 – 888.

Vamvoudakis, K.G., Lewis, F.L., 2011. Multi-player non-zero-sum games: Online adaptive learning solution of coupled hamilton-jacobi equations. Automatica 47, 1556 – 1569.

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., Lewis, F., 2009. Adaptive optimal control for continuous-time linear systems based on policy iteration. Automatica 45, 477 – 484.

Wang, D., He, H., Zhong, X., Liu, D., 2017. Event-driven nonlinear discounted optimal regulation involving a power system application. IEEE Transactions on Industrial Electronics 64, 8177–8186.

Wang, D., Liu, D., Zhang, Q., Zhao, D., 2015. Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics. IEEE Transactions on Systems, Man, and Cybernetics: Systems 46, 1544–1555.

Wang, F.Y., Zhang, H., Liu, D., 2009. Adaptive dynamic programming: An introduction. IEEE Computational Intelligence Magazine 4, 39–47.

Watkins, C., Dayan, P., 1992. Q-learning. Machine Learning 8, 279–292.

Wei, Q., Lewis, F.L., Liu, D., Song, R., Lin, H., 2018. Discrete-time local value iteration adaptive dynamic programming: Convergence analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems 48, 875–891.

Wei, Q., Liu, D., Shi, G., 2015. A novel dual iterative Q-learning method for optimal battery management in smart residential environments. IEEE Transactions on Industrial Electronics 62, 2509–2518.

Xie, K., Zheng, Y., Jiang, Y., Lan, W., Yu, X., 2024. Optimal dynamic output feedback control of unknown linear continuous-time systems by adaptive dynamic programming. Automatica 163, 111601.

Xu, B., Yang, C., Shi, Z., 2014. Reinforcement learning output feedback NN control using deterministic learning technique. IEEE Transactions on Neural Networks and Learning Systems 25, 635–641.

Yang, Q., Jagannathan, S., 2012. Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators. IEEE Transactions on Systems Man and Cybernetics - PART B: Cybernetics 42, 377–390.

Yang, Y., Pan, Y., Xu, C.Z., Wunsch, D.C., 2024. Hamiltonian-driven adaptive dynamic programming with efficient experience replay. IEEE Transactions on Neural Networks and Learning Systems 35, 3278–3290.

Yang, Y., Vamvoudakis, K.G., Modares, H., Yin, Y., Wunsch, D.C., 2020. Safe intermittent reinforcement learning with static and dynamic event generators. IEEE Transactions on Neural Networks and Learning Systems, in press .

Zhu, Y., Zhao, D., 2017. Comprehensive comparison of online ADP algorithms for continuous-time optimal control. Artificial Intelligence Review , 1–17.