Control Engineering Practice 121 (2022) 105042

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

Value iteration and adaptive optimal output regulation with assured convergence rate☆

Yi Jiang [a], Weinan Gao [b,*], Jing Na [c], Di Zhang [d], Timo T. Hämäläinen [d], Vladimir Stojanovic [e], Frank L. Lewis [f]

[a] Department of Biomedical Engineering, City University of Hong Kong, Hong Kong Special Administrative Region of China
[b] Department of Mechanical and Civil Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA
[c] Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China
[d] Faculty of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35, Jyväskylä, FIN-40014, Finland
[e] Faculty of Mechanical and Civil Engineering, University of Kragujevac, Kraljevo 36000, Serbia
[f] UTA Research Institute, the University of Texas at Arlington, Fort Worth, TX 76118, USA

ARTICLE INFO

Keywords:
Reinforcement learning
Optimal output regulation
Value iteration
Assured convergence rate
Adaptive dynamic programming

ABSTRACT

In this paper, we investigate the learning-based adaptive optimal output regulation problem with convergence rate requirement for disturbed linear continuous-time systems. An adaptive optimal control approach is proposed based on reinforcement learning and adaptive dynamic programming to learn the optimal regulator with assured convergence rate. The above-mentioned problem is successfully solved by tackling a static optimization problem to find the optimal solution to the regulator equations, and a dynamic and constrained optimization problem to obtain the optimal feedback control gain. Without requiring on the accurate system dynamics or a stabilizing feedback control gain, a novel online value iteration algorithm is proposed, which can learn both the optimal feedback control gain and the corresponding feedforward control gain using measurable data. Moreover, the output of the closed-loop system is guaranteed to converge faster or equal to a predefined convergence rate set by user. Finally, the numerical analysis on a LCL coupled inverter-based distributed generation system shows that the proposed approach can achieve desired disturbance rejection and tracking performance.

1. Introduction

To address an output regulation problem, designers need develop a regulator for the controlled plant to ensure the resultant closed-loop system rejects external disturbances and tracks some desired trajectories asymptotically. Model-based solutions to output regulation problems usually require accurate knowledge of the system model (Francis, 1977; Francis & Wonham, 1976). Practically, building perfect mathematical models for controlled plants may be hard or even impossible.

Instead of directly relying on the system model, some data-driven and adaptive approaches have been proposed using online data to adaptively tune the controller parameters and structure. For instance, He et al. (1993), Woo et al. (2000) have developed self-tuning PID controllers based on fuzzy logic. Wang and Huang (2005) have designed an adaptive dynamic surface control approach for uncertain strict-feedback nonlinear systems based on the neural networks. Park et al. (2009) have proposed an adaptive sliding mode control approach for nonholonomic wheeled mobile robots via neural network technology as well.

To guarantee the transient response of the closed-loop control systems, one usually requires the states asymptotically converge to the desired trajectory in an optimal sense, reinforcement learning (RL) and adaptive dynamic programming (ADP) techniques serve as powerful tools and have been introduced to learn an optimal regulator for unknown systems. See recent book, survey, tutorial and research papers (Gao et al., 2021; Jiang, Bian, & Gao, 2020; Kamalapurkar et al., 2018; Kiumarsi et al., 2018; Liu et al., 2021; Vamvoudakis & Kokolakis, 2020; Wang, Ha, & Qiao, 2020; Wang et al., 2017; Wang, Qiao, & Cheng, 2020). Among different strategies in the RL and ADP, policy iteration (PI) and value iteration (VI) are two successive approximation approaches to seek for the optimal control policy. In order to ensure

☆ This work was supported in part by U.S. National Science Foundation under grant CMMI-2138206, National Natural Science Foundation of China under grants 61922037 and 61873115, and Serbian Ministry of Education, Science and Technological Development (No. 451-03-9/2021-14/200108). The material in this paper was partially submitted to the 2022 Chinese Control and Decision Conference (CCDC), May 21–23, 2022, Hefei, China.

* Corresponding author.
E-mail addresses: yjian22@cityu.edu.hk (Y. Jiang), wgao@fit.edu (W. Gao), najing25@163.com, najing25@kust.edu.cn (J. Na), d.zhang@jyu.fi (D. Zhang), timo.t.hamalainen@jyu.fi (T.T. Hämäläinen), vladostojanovic@mts.rs (V. Stojanovic), lewis@uta.edu (F.L. Lewis).

https://doi.org/10.1016/j.conengprac.2021.105042
Received 26 August 2021; Received in revised form 23 November 2021; Accepted 16 December 2021
Available online 10 January 2022

its convergence, PI based RL algorithm usually initiates with an admissible control policy, which is stabilizing and ensures a finite cost. Unfortunately, in practice, it is usually hard to find such a satisfactory control policy, especially with unknown or inaccurate system dynamics. A major advantage of VI based RL algorithm is that one can begin with any control policy, which is more practical.

Generally, there are two alternative formulations to deal with the concerned adaptive optimal output regulation problem using RL techniques. The first formulation is to construct a linear quadratic tracker by introducing a discounted cost function. It was adopted in Modares and Lewis (2014a), Xue et al. (2021) for linear continuous-time (CT) systems, in Jiang, Fan, Chai, Lewis, and Li (2018), Kiumarsi et al. (2014), Wu et al. (2019), Xue et al. (2020) for linear discrete-time (DT) systems, in Modares and Lewis (2014b), Wang et al. (2021) for nonlinear CT systems, in Jiang et al. (2019), Jiang, Fan, Chai, Li, and Lewis (2018), Kiumarsi and Lewis (2015) for nonlinear DT systems. Another formulation to solve this problem is by splitting it into an adaptive optimal feedback control problem and an adaptive optimal feedforward control problem. This approach was adopted in Chen et al. (2019), Gao and Jiang (2016), Gao et al. (2018) for linear CT systems, in Fan et al. (2020), Gao and Jiang (2019), Jiang, Kiumarsi, et al. (2020) for linear DT systems, in Gao and Jiang (2018) for nonlinear CT systems, in Jiang, Fan, et al. (2020) for nonlinear DT systems. However, all the above approaches for linear CT systems are realized by adopting the PI based RL algorithm, which requires an initial admissible stabilizing feedback control gain to start learning (Kleinman, 1968). Besides asymptotic tracking, the convergence rate of the closed-loop system is usually required to be fast enough in practice. In Hong et al. (2002), the finite-time control of the robot system was studied through both state feedback and dynamic output feedback control. In Huang et al. (2017), finite-time controllers were proposed for underactuated spacecraft hovering in the absence of the radial or in-track thrust. In the above-mentioned references, the convergence rate is tuned by applying the finite-time controllers based on the system dynamics. However, it is not applicable to the case with unknown or inaccurate system dynamics.

To this end, we will propose a novel data-driven value iteration (VI) algorithm to handle the adaptive CT linear optimal output regulation problem (LO^2RP) with assured convergence rate in this paper. The proposed algorithm adaptively learns both optimal feedback and feedforward control gains using online measurable data. Moreover, the state of the system in closed-loop with the learned regulator is guaranteed to converge faster or equal to a predefined convergence rate. Notably, different from the PI based RL algorithms, VI based RL algorithms are free from an initial stabilizing control policy to initiate (see Al-Tamimi et al., 2008; Bian & Jiang, 2016).

This paper is outlined as follows. In Section 2, we formulate the LO^2RP with assured convergence rate. Section 3 provides the model-based solution to LO^2RP and a VI algorithm to approximate the optimal feedback control gain. In Section 4, a data-driven VI algorithm is presented to learn the optimal regulator with unknown system matrices. In Section 5, a numerical analysis based on a LCL coupled inverter-based distributed generation system is given to demonstrate the efficiency of the proposed learning-based output regulation approach. Section 6 contains the concluding remarks.

Notation. Throughout this paper, $\mathbb{R}$ and $\mathbb{N}$ denote respectively the set of real numbers and the set of positive natural numbers. For a $n \in \mathbb{N}$ and two matrices $X, Y \in \mathbb{R}^{n\times n}$, $X > 0$ ($X \geq 0$) means the matrix X is positive definite (positive semi-definite); $X > Y$ ($X \geq Y$) means $X - Y$ is positive definite (positive semi-definite); $\sigma(X)$ means the complex spectrum of matrix X. For brevity, denote $\mathbb{R}^n := \mathbb{R}^{n\times 1}$. Moreover, $\|\cdot\|$ denotes the Euclidean norm of vectors and the Frobenius norm of matrices; $\otimes$ denotes Kronecker product; $\mathbb{P}^n$ denotes the normed space of all n-by-n real symmetric matrices; $\mathbb{P}^n_+ := \{P \in \mathbb{P}^n : P \geq 0\}$. For $m, n \in \mathbb{N}$ and $X \in \mathbb{R}^{m\times n}$, X^{T} denotes the transpose of X, $\mathrm{vec}(X) = [x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}}]^{\mathrm{T}}$ with $x_i \in \mathbb{R}^m$ the columns of matrix X. For a symmetric matrix $X \in \mathbb{R}^{n\times n}$, $\mathrm{vecs}(X) = [x_{11}, 2x_{12}, \ldots, 2x_{1n}, x_{22}, 2x_{23}, \ldots, 2x_{(n-1)n}, x_{nn}]^{\mathrm{T}} \in \mathbb{R}^{(1/2)n(n+1)}$. For a vector $v \in \mathbb{R}^n$, $\mathrm{vecv}(v) = [v_1^2, v_1v_2, \ldots, v_1v_n, v_2^2, v_2v_3, \ldots, v_{n-1}v_n, v_n^2]^{\mathrm{T}} \in \mathbb{R}^{(1/2)n(n+1)}$.

2. Problem formulation

We start from the following disturbed linear time-invariant (LTI) CT system,

$$\dot{x}(t) = Ax(t) + Bu(t) + Dv(t), \tag{1}$$

$$y(t) = Cx(t), \tag{2}$$

where $x \in \mathbb{R}^{n_x}$, $u \in \mathbb{R}^{n_u}$, $v \in \mathbb{R}^{n_v}$ and $y \in \mathbb{R}^{n_y}$ are the state, the control input, the exostate, and the output, respectively. $A \in \mathbb{R}^{n_x\times n_x}$, $B \in \mathbb{R}^{n_x\times n_u}$, $D \in \mathbb{R}^{n_x\times n_v}$ and $C \in \mathbb{R}^{n_y\times n_x}$ are constant matrices. Following the output regulation framework, the exosystem is modeled by a linear LTI CT autonomous systems as follows,

$$\dot{v}(t) = Ev(t), \tag{3}$$

where $E \in \mathbb{R}^{n_v\times n_v}$ is a constant matrix. The reference signal can be computed as follows,

$$y_d(t) = -Fv(t), \tag{4}$$

where $y_d \in \mathbb{R}^{n_y}$ is the reference signal and $F \in \mathbb{R}^{n_y\times n_v}$.

The following standard assumptions are made throughout this paper.

Assumption 1. The pair (A, B) is controllable.

Assumption 2. The exosignal $v(t)$ is unmeasurable.

Assumption 3. The minimal polynomial of E is available.

Assumption 4. $\mathrm{rank}\left(\begin{bmatrix} A-\lambda I & B \\ C & 0 \end{bmatrix}\right) = n_x + n_y, \forall \lambda \in \sigma(E).$

Under Assumption 3, one can always find a new state $w \in \mathbb{R}^{n_w}$ and an autonomous system such that w and v are its state and output, which is shown as follows,

$$\dot{w}(t) = \hat{E}w(t), \tag{5}$$

$$v(t) = Gw(t), \tag{6}$$

where $\hat{E} \in \mathbb{R}^{n_w\times n_w}$ and $G \in \mathbb{R}^{n_v\times n_w}$ are constant matrices. Then, the controlled plant in (1) and its measurement output $e(t) := y(t) - y_d(t)$ can be formulated as follows,

$$\dot{x}(t) = Ax(t) + Bu(t) + \hat{D}w(t), \tag{7}$$

$$e(t) = Cx(t) + \hat{F}w(t), \tag{8}$$

where $\hat{D} = DG$ and $\hat{F} = FG$.

Our control goal is to develop an optimal regulator for the concerned disturbed LTI CT system in (7) such that the closed-loop system is globally asymptotically stable (GAS) in an optimal sense. Moreover, the convergence rate of the tracking error is fast then $\mathrm{e}^{-\gamma t}$, which means the following requirement needs to be satisfied,

$$\lim_{t\to\infty} \mathrm{e}^{\gamma t} e(t) = 0, \tag{9}$$

where the convergence rate criterion γ satisfies $0 \leq \gamma < \infty$. As a typical strategy to deal with output regulation problems, we present a feedback–feedforward controller as follows,

$$u(t) = -Kx(t) + Lw(t), \tag{10}$$

where $K \in \mathbb{R}^{n_u\times n_x}$ is the feedback control gain and $L \in \mathbb{R}^{n_u\times n_w}$ is the feedforward control gain, respectively. As discussed in Francis (1977),

Huang (2004), K is required to be designed to ensure that $A - BK$ is Hurwitz and L should satisfy the following equation,

$$L = U + KX, \tag{11}$$

where $X \in \mathbb{R}^{n_x \times n_w}$ and $U \in \mathbb{R}^{n_u \times n_w}$ solves of the following regulator equations,

$$X\hat{E} = AX + BU + \hat{D}, \tag{12}$$

$$0 = CX + \hat{F}. \tag{13}$$

Under the condition of Assumption 4, Eqs. (12)–(13) is solvable (Knobloch et al., 2012). The requirements of the LO^2RP focus on the following aspects: (1) the asymptotic tracking of the output; (2) the transient performance and the GAS of the resulting linear closed-loop control system. To satisfy these requirements, a static optimization Problem 1 needs to be solved to obtain the optimal solution of regulator equations (X^*, U^*), while a dynamic constrained optimization Problem 2 requires solved to obtain the optimal feedback control gain K^* and the corresponding feedforward control gain by $L^* = U^* + K^*X^*$. We first formulate the Problem 1 as follows:

Problem 1:

$$\begin{aligned} \min_{\bar{X}} \quad & \bar{X}^{\mathrm{T}} M \bar{X} \\ \text{s.t.} \quad & X\hat{E} = AX + BU + \hat{D} \\ & 0 = CX + \hat{F}, \end{aligned} \tag{14}$$

where $\bar{X} = \left[(\mathrm{vec}(X))^{\mathrm{T}}, (\mathrm{vec}(U))^{\mathrm{T}}\right]^{\mathrm{T}}$, $M = M^{\mathrm{T}} > 0$. ■

Under the solutions of (14), by denoting $\bar{x}(t) = \mathrm{e}^{\gamma t}(x(t) - Xw(t))$, $\bar{u}(t) = \mathrm{e}^{\gamma t}(u(t) - Uw(t))$ and $\bar{e}(t) = \mathrm{e}^{\gamma t}e(t)$ as the new state, input and error, respectively, a new CT system can be formulated as follows,

$$\begin{aligned} \dot{\bar{x}}(t) &= \gamma\bar{x}(t) + \mathrm{e}^{\gamma t}(\dot{x}(t) - X\dot{w}(t)) \\ &= \gamma\bar{x}(t) + \mathrm{e}^{\gamma t}(Ax(t) + Bu(t) + \hat{D}w(t) - X\hat{E}w(t)) \\ &= \gamma\bar{x}(t) + \mathrm{e}^{\gamma t}(Ax(t) + Bu(t) - (AX + BU)w(t)) \\ &= \bar{A}\bar{x}(t) + B\bar{u}(t), \end{aligned} \tag{15}$$

$$\begin{aligned} \bar{e}(t) &= \mathrm{e}^{\gamma t}Cx(t) + \mathrm{e}^{\gamma t}\hat{F}w(t) \\ &= \mathrm{e}^{\gamma t}Cx(t) - \mathrm{e}^{\gamma t}CXw(t) \\ &= C\bar{x}(t), \end{aligned} \tag{16}$$

with $\bar{A} = A + \gamma I$. Then, one can present the Problem 2 as below:

Problem 2:

$$\begin{aligned} \min_{\bar{u}} \quad & \int_0^\infty \left(\bar{x}^{\mathrm{T}}(\tau)Q\bar{x}(\tau) + \bar{u}^{\mathrm{T}}(\tau)R\bar{u}(\tau)\right) d\tau \\ \text{s.t.} \quad & \dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + B\bar{u}(t), \end{aligned} \tag{17}$$

where $Q = Q^{\mathrm{T}} > 0$ and $R = R^{\mathrm{T}} > 0$. ■

3. Solution to the LO^2RP with known system matrices

In this section, we present solutions to find K^* and L^* when the system matrices are available. That is, the model-driven solutions to Problems 1–2. To solve the Problem 1, a method of Lagrange multipliers is introduced to convert this problem to a static unconstrained optimization problem. A model-based VI algorithm is provided to obtain the optimal solution of the Problem 2. The results in this section will be helpful to develop data-driven VI algorithms to compute the optimal feedforward control gain and the optimal feedback control gain in Section 4.

To deal with the Problem 1, inspired by Gao and Jiang (2016), Jiang, Kiumarsi, et al. (2020), a Sylvester map $\Omega : \mathbb{R}^{n_x \times n_w} \to \mathbb{R}^{n_x \times n_w}$ is introduced, which is shown as follows,

$$\Omega(X) = X\hat{E} - AX. \tag{18}$$

Under the definition of this Sylvester map, a general solution with some unknown parameters of (12) can be easily established. Select a sequence $X_i \in \mathbb{R}^{n_x \times n_w}$ with $i = 0, 1, \ldots, m+1$, where $m = (n_x - n_y)n_w$, $X_0 = 0_{n_x \times n_w}$, $X_1 \in \mathbb{R}^{n_x \times n_w}$, so that $CX_1 = -\hat{F}$, and all $\mathrm{vec}(X_i)$ build a basis of $\ker(I_{n_w} \otimes C)$ with $i = 2, 3, \ldots, m+1$, that is $CX_i = 0$. Clearly, a general solution of (13) can be established as the following equation,

$$X = X_0 + X_1 + \sum_{i=2}^{m+1} \alpha_i X_i, \tag{19}$$

where $\alpha_i \in \mathbb{R}$. Besides, a general solution of (12) can be established as the following equation,

$$\Omega(X) = \Omega(X_1) + \sum_{i=2}^{m+1} \alpha_i \Omega(X_i) = BU + \hat{D}. \tag{20}$$

Thus, the regulator equations (12)–(13) has the following equivalent form by using the definition of the Sylvester map in (18) as follows,

$$\Lambda\chi = \xi, \tag{21}$$

where

$$\Lambda = \begin{bmatrix} \mathrm{vec}(\Omega(X_2)) & \cdots & \mathrm{vec}(\Omega(X_{m+1})) & 0 & -I_{n_w} \otimes B \\ \mathrm{vec}(X_2) & \cdots & \mathrm{vec}(X_{m+1}) & -I_{n_x n_w} & 0 \end{bmatrix},$$

$$\chi = \begin{bmatrix} \alpha_2 & \cdots & \alpha_{m+1} & \mathrm{vec}(X)^{\mathrm{T}} & \mathrm{vec}(U)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}},$$

$$\xi = \begin{bmatrix} \mathrm{vec}(-\Omega(X_1) + \hat{D}) \\ -\mathrm{vec}(X_1) \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}.$$

Then, by using row operation, Eq. (21) can be rewritten as the following equation,

$$\begin{bmatrix} \bar{\Lambda}_{11} & \bar{\Lambda}_{12} \\ \bar{\Lambda}_{21} & \bar{\Lambda}_{22} \end{bmatrix} \chi = \begin{bmatrix} \bar{\xi}_1 \\ \bar{\xi}_2 \end{bmatrix}, \tag{22}$$

with $\bar{\Lambda}_{21} \in \mathbb{R}^{m \times m}$ being a nonsingular matrix. Inspired by Gao and Jiang (2016), Jiang, Kiumarsi, et al. (2020), above equation can be rewritten as the following equation,

$$\Pi\bar{X} = \Psi, \tag{23}$$

where $\Pi = -\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\Lambda}_{22} + \bar{\Lambda}_{12}$ and $\Psi = -\bar{\Lambda}_{11}\bar{\Lambda}_{21}^{-1}\bar{\xi}_2 + \bar{\xi}_1$.

By using the method of Lagrange multipliers, the static constrained optimization problem in (14) can be converted as a static and unconstrained optimization problem, that is,

$$\min_{\bar{X}} \quad J = \bar{X}^{\mathrm{T}} M \bar{X} + \lambda^{\mathrm{T}}(\Pi\bar{X} - \Psi). \tag{24}$$

According to the optimization theory, we need to compute the partial derivative of J in (24) with respect to $\bar{X}$ and λ, respectively. They are

$$\frac{\partial J}{\partial \bar{X}} = 2M\bar{X} + \Pi^{\mathrm{T}}\lambda, \tag{25}$$

$$\frac{\partial J}{\partial \lambda^{\mathrm{T}}} = \Pi\bar{X} - \Psi. \tag{26}$$

By setting (25)–(26) equal to 0, one has the optimal solutions of Problem 1. They are

$$\begin{bmatrix} \bar{X} \\ \lambda \end{bmatrix} = \begin{bmatrix} 2M & \Pi^{\mathrm{T}} \\ \Pi & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \Psi \end{bmatrix}. \tag{27}$$

Problem 2 is a standard linear quadratic regulator (LQR) problem. By Zhou et al. (2008), the dynamic constrained optimization problem in (17) can be solved by designing the following controller,

$$\bar{u}(t) = -K^*\bar{x}(t), \tag{28}$$

where

$$K^* = R^{-1}B^{\mathrm{T}}P^* \tag{29}$$

and $P^* = (P^*)^{\mathrm{T}} > 0$ is the unique positive definite solution to the following CT algebraic Riccati equation (ARE),

$$P\bar{A} + \bar{A}^{\mathrm{T}}P + Q - PBR^{-1}B^{\mathrm{T}}P = 0. \tag{30}$$

One way to compute the K^* is to solve the CT ARE (30), directly. However, this CT ARE is a nonlinear function with respect to P.

Algorithm 1 Model-based VI Algorithm for LO^2RP with assured convergence rate

Initiation: Start with an arbitrary feedback control gain matrix K_0 and a positive semi-definite matrix P_0. Set $j \leftarrow 0$, $q \leftarrow 0$. Select a small threshold $\varepsilon > 0$, a sequence $X_i \in \mathbb{R}^{n_x \times n_w}$, a predefined convergence rate parameter $\gamma \geq 0$, bounded sets $\{B_q\}_{q=0}^{\infty}$ with nonempty interiors, and a sequence $\{\epsilon_j\}_{j=0}^{\infty}$ satisfying

$$B_q \subseteq B_{q+1}, \quad q \in \mathbb{N}, \quad \lim_{q\to\infty} B_q = \mathbb{P}_+^n, \tag{33}$$

$$\epsilon_j > 0, \quad \sum_{j=0}^{\infty} \epsilon_j = \infty, \quad \sum_{j=0}^{\infty} \epsilon_j^2 < \infty. \tag{34}$$

Optimal feedback control gain computation: Iterate the following three steps on j until the matrix sequence $\{P_j\}_{j=0}^{\infty}$ converges.

1. **Value Evaluation**: Solve $\tilde{P}_{j+1}$ from the following equation
$$\tilde{P}_{j+1} = P_j + \epsilon_j(\bar{A}^{\mathrm{T}}P_j + P_j\bar{A} + Q - K_j^{\mathrm{T}}RK_j); \tag{35}$$
2. **Policy Improvement**: Improve the feedback control gain K_{j+1} by
$$K_{j+1} = R^{-1}B^{\mathrm{T}}\tilde{P}_{j+1}; \tag{36}$$
3. If $\tilde{P}_{j+1} \notin B_q$, set $P_{j+1} \leftarrow P_0$, $q \leftarrow q+1$; else if $\|\tilde{P}_{j+1} - P_j\|/\epsilon_j < \varepsilon$, return $\tilde{P}_{j+1}$ and K_{j+1} as the approximations of P^* and K^*; else $P_{j+1} \leftarrow \tilde{P}_{j+1}$, $j \leftarrow j+1$ and go to the value evaluation.

Optimal feedforward control gain computation: Solve for the solutions (X, U) of the regulator equations (12)–(13) using Eq. (27). Then calculate the optimal feedforward gain matrix $L^* = U + K_jX$.

Another way to obtain K^* is using some iteration based approaches, including PI and VI based RL algorithm. In the PI approach (Kleinman, 1968), an initial stabilizing feedback control gain is required to achieve convergence. In the VI approach, this requirement is relaxed. The following algorithm shows how to use model-based VI approach to solve the LO^2RP and a lemma is given to show the convergence of this algorithm.

Lemma 1 (*Bian & Jiang, 2016*). *Consider* $\{P_j\}_{j=0}^{\infty}$ *and* $\{K_j\}_{j=0}^{\infty}$ *in Algorithm 1. Under Assumption 1, one has the following properties,*

$$\lim_{j\to\infty} P_j = P^*, \tag{31}$$

$$\lim_{j\to\infty} K_j = K^*. \tag{32}$$

Remark 1. The condition in Assumption 1 implies that $(\bar{A}, B)$ is controllable. If Assumption 1 is relaxed by the pair (A, B) is stabilizable, the convergence rate requirement is still achievable, but γ should be smaller than the opposite number of maximum uncontrollable stable eigenvalue of A so that $(\bar{A}, B)$ is still stabilizable.

Remark 2. Under the condition in Assumption 4, the regulator equations (12)–(13) are solvable. Based on the conclusion in Lemma 1, one has that Algorithm 1 can solve the LO^2RP. Moreover, this algorithm has two advantages. First, a stabilizing feedback control gain is no longer required to initiate the learning process. Second, compared with the model-based PI approach (Kleinman, 1968), there is no need to solve a Lyapunov equation at each iteration, which reduces the computational burden per iteration.

4. VI based adaptive optimal control for solving LO^2RP

We have proposed a model-based solution to the LO^2RP in Section 3. However, it is computed based on the exact values of the system matrices. In this section, we will develop a data-driven VI algorithm to learn the solution to the LO^2RP with known matrices C and F, which is realized by using the online states data of the concerned disturbed linear LTI CT system and the exo-system and input data.

By defining $\bar{x}_i(t) = \mathrm{e}^{\gamma t}(x(t) - X_iw(t))$ for $i = 0, 1, \ldots, m+1$, one has the following equation,

$$\begin{aligned}\dot{\bar{x}}_i(t) =& Ax(t) + Bu(t) + (\hat{D} - X_i\hat{E})w(t)\\ =& A\bar{x}_i(t) + Bu(t) + (\hat{D} - \Omega(X_i))w(t),\end{aligned} \tag{37}$$

where $\bar{w}(t) = \mathrm{e}^{\gamma t}w(t)$ and $\hat{u}(t) = \mathrm{e}^{\gamma t}u(t)$.

Along with the solutions of (37), one yields the following equation,

$$\begin{aligned}\frac{d}{dt}\left(\bar{x}_i^{\mathrm{T}}(t)P_j\bar{x}_i(t)\right) =& \left(\bar{A}\bar{x}_i(t) + B\hat{u}(t) + (\hat{D} - \Omega(X_i))\bar{w}(t)\right)^{\mathrm{T}}P_j\bar{x}_i(t)\\ &+ \bar{x}_i^{\mathrm{T}}(t)P_j\left(\bar{A}\bar{x}_i(t) + B\hat{u}(t) + (\hat{D} - \Omega(X_i))\bar{w}(t)\right)\\ =& \bar{x}_i^{\mathrm{T}}(t)H_j\bar{x}_i(t) + 2\hat{u}^{\mathrm{T}}(t)RK_j\bar{x}_i(t)\\ &+ 2\bar{x}_i^{\mathrm{T}}(t)P_j(\hat{D} - \Omega(X_i))\bar{w}(t),\end{aligned} \tag{38}$$

where $H_j = \bar{A}^{\mathrm{T}}P_j + P_j\bar{A}$ and $K_j = R^{-1}B^{\mathrm{T}}P_j$.

By taking the integration of Eq. (38), we observe that the following equation holds,

$$\begin{aligned}&\bar{x}_i^{\mathrm{T}}(t+\delta t)P_j\bar{x}_i(t+\delta t) - \bar{x}_i^{\mathrm{T}}(t)P_j\bar{x}_i(t)\\ =& \int_t^{t+\delta t} \bar{x}_i^{\mathrm{T}}(\tau)H_j\bar{x}_i(\tau)d\tau + \int_t^{t+\delta t} 2\hat{u}^{\mathrm{T}}(\tau)RK_j\bar{x}_i(\tau)d\tau\\ &+ \int_t^{t+\delta t} 2\bar{x}_i^{\mathrm{T}}(\tau)P_j(\hat{D} - \Omega(X_i))\bar{w}(\tau)d\tau,\end{aligned} \tag{39}$$

where $\delta t > 0$. Using Kronecker product representation, one obtains the following equations,

$$\begin{aligned}\bar{x}_i^{\mathrm{T}}(t)P_j\bar{x}_i(t) &= \left(\bar{x}_i^{\mathrm{T}}(t) \otimes \bar{x}_i^{\mathrm{T}}(t)\right)\mathrm{vec}(P_j),\\ \bar{x}_i^{\mathrm{T}}(\tau)H_j\bar{x}_i(\tau) &= \left(\bar{x}_i^{\mathrm{T}}(\tau) \otimes \bar{x}_i^{\mathrm{T}}(\tau)\right)\mathrm{vec}(H_j),\\ u^{\mathrm{T}}(\tau)RK_j\bar{x}_i(\tau) &= \left(\bar{x}_i^{\mathrm{T}}(\tau) \otimes u^{\mathrm{T}}(\tau)R\right)\mathrm{vec}(K_j),\\ \bar{x}_i^{\mathrm{T}}(\tau)P_j(\hat{D} - \Omega(X_i))w(\tau) &= \left(w^{\mathrm{T}}(\tau) \otimes \bar{x}_i^{\mathrm{T}}(\tau)\right)\mathrm{vec}(P_j(\hat{D} - \Omega(X_i))).\end{aligned}$$

Next, for positive integer s, define the following matrices,

$$\begin{aligned}\xi_{\bar{x}_i\bar{x}_i} =& \left[\mathrm{vecv}(\bar{x}_i(t_1)) - \mathrm{vecv}(\bar{x}_i(t_0)), \mathrm{vecv}(\bar{x}_i(t_2))\right.\\ &\left.-\mathrm{vecv}(\bar{x}_i(t_1)), \ldots, \mathrm{vecv}(\bar{x}_i(t_s)) - \mathrm{vecv}(\bar{x}_i(t_{s-1}))\right]^{\mathrm{T}},\\ \Gamma_{\bar{x}_i\bar{x}_i} =& \left[\int_{t_0}^{t_1} \mathrm{vecv}(\bar{x}_i(\tau))d\tau, \int_{t_1}^{t_2} \mathrm{vecv}(\bar{x}_i(\tau))d\tau, \ldots,\right.\\ &\left.\int_{t_{s-1}}^{t_s} \mathrm{vecv}(\bar{x}_i(\tau))d\tau\right]^{\mathrm{T}},\\ \Gamma_{\bar{x}_i\hat{u}} =& \left[\int_{t_0}^{t_1} \bar{x}_i(\tau) \otimes R\hat{u}(\tau)d\tau, \int_{t_1}^{t_2} \bar{x}_i(\tau) \otimes R\hat{u}(\tau)d\tau, \ldots,\right.\\ &\left.\int_{t_{s-1}}^{t_s} \bar{x}_i(\tau) \otimes R\hat{u}(\tau)d\tau\right]^{\mathrm{T}},\\ \Gamma_{\bar{w}\bar{x}_i} =& \left[\int_{t_0}^{t_1} \bar{w}(\tau) \otimes \bar{x}_i(\tau)d\tau, \int_{t_1}^{t_2} \bar{w}(\tau) \otimes \bar{x}_i(\tau)d\tau, \ldots,\right.\\ &\left.\int_{t_{s-1}}^{t_s} \bar{w}(\tau) \otimes \bar{x}_i(\tau)d\tau\right]^{\mathrm{T}},\end{aligned}$$

where $0 \leq t_0 < t_1 < \cdots < t_s$.

Using the above matrices, one can rewrite (39) as the following equation,

$$\left[\Gamma_{\bar{x}_i\bar{x}_i}, 2\Gamma_{\bar{x}_i\hat{u}}, 2\Gamma_{\bar{w}\bar{x}_i}\right]\begin{bmatrix}\mathrm{vecs}(H_j)\\ \mathrm{vec}(K_j)\\ \mathrm{vec}(P_j(\hat{D} - \Omega(X_i)))\end{bmatrix} = \xi_{\bar{x}_i\bar{x}_i}\mathrm{vecs}(P_j). \tag{40}$$

Note that the solution of (40) is unique if the rank condition in Lemma 2 is satisfied.

Lemma 2. *For $i = 0, 1, \dots, m+1$, (40) has an unique solution if the following rank condition holds,*

$$rank\left(\left[\Gamma_{\bar{x}_i\bar{x}_i}, 2\Gamma_{\bar{x}_i\hat{u}}, 2\Gamma_{\bar{w}\bar{x}_i}\right]\right) = \frac{n_x(n_x+1)}{2} + n_x(n_u+n_w).$$

By Lemma 2, one has that Eq. (40) has an unique solution by using the least squares (LS) method, that is,

$$\begin{bmatrix} \mathrm{vecs}(H_j) \\ \mathrm{vec}(K_j) \\ \mathrm{vec}(P_j(\hat{D} - \Omega(X_i))) \end{bmatrix} = \Theta_j \xi_{\bar{x}_i\bar{x}_i} \mathrm{vecs}(P_j), \tag{41}$$

where $\Theta_j = \left(\left[\Gamma_{\bar{x}_i\bar{x}_i}, 2\Gamma_{\bar{x}_i\hat{u}}, 2\Gamma_{\bar{w}\bar{x}_i}\right]^{\mathrm{T}}\left[\Gamma_{\bar{x}_i\bar{x}_i}, 2\Gamma_{\bar{x}_i\hat{u}}, 2\Gamma_{\bar{w}\bar{x}_i}\right]\right)^{-1}$ $\left[\Gamma_{\bar{x}_i\bar{x}_i}, 2\Gamma_{\bar{x}_i\hat{u}}, 2\Gamma_{\bar{w}\bar{x}_i}\right]^{\mathrm{T}}$.

Then, from Eq. (41), one can calculate P_j, K_j and $P_j(\hat{D} - \Omega(X_i))$ for $i = 0, 1, \dots, m+1$. For $i = 0$, we obtain $\hat{D}$ by $\hat{D} = P_j^{-1}P_j(\hat{D} - \Omega(X_i))$; for $i = 1, \dots, m+1$, we have $\Omega(X_i)$ by $\Omega(X_i) = -P_j^{-1}P_j(\hat{D} - \Omega(X_i)) - \hat{D}$. Moreover, one can calculate B by $B = P_j^{-1}K_j^{\mathrm{T}}R$. The matrix Λ in (21) can be rewritten as follows,

$$\Lambda = \begin{bmatrix} \mathrm{vec}(\Omega(X_2)) & \cdots & \mathrm{vec}(\Omega(X_{m+1})) & 0 \\ \mathrm{vec}(X_2) & \cdots & \mathrm{vec}(X_{m+1}) & -I_{n_x\times n_w} \\ -I_{n_w}\otimes P_j^{-1}K_j^{\mathrm{T}}R & & 0 & \end{bmatrix}. \tag{42}$$

Finally, an online VI based algorithm is given for the LO^2RP as Algorithm 2, and a theorem is provided to show the convergence of this algorithm.

Algorithm 2 Online VI Algorithm for the LO^2RP with assured convergence rate

Initiation: Start with an arbitrary feedback control gain matrix K_0 and a positive semi-definite matrix P_0. Set $j \leftarrow 0$, $q \leftarrow 0$, $i \leftarrow 0$. Select a small threshold $\varepsilon > 0$, a sequence $X_i \in \mathbb{R}^{n_x\times n_w}$, a predefined convergence rate parameter $\gamma \geq 0$, and a collection of bounded sets $\{B_q\}_{q=0}^{\infty}$ with nonempty interiors and a sequence $\{\epsilon_j\}_{j=0}^{\infty}$ satisfying (33) and (34). Employ $u(t) = K_0x(t) + \hat{e}(t)$ as the input on $[t_0, t_s]$.

Optimal feedback control gain computation: Iterate the following three steps on j until the matrix sequence $\{P_j\}_{j=0}^{\infty}$ converges.

1. Solve for matrices H_j and K_j using (40), then calculate (35) by using the following equation,
$$\tilde{P}_{j+1} = P_j + \epsilon_j(H_j + Q - K_j^{\mathrm{T}}RK_j); \tag{43}$$
2. If $\tilde{P}_{j+1} \notin B_q$, set $P_{j+1} \leftarrow P_0$, $q \leftarrow q+1$; else if $\|\tilde{P}_{j+1} - P_j\|/\epsilon_j < \varepsilon$, return P_j and K_j as the approximations of P^* and K^*; else $P_{j+1} \leftarrow \tilde{P}_{j+1}$, $j \leftarrow j+1$ and go to the value evaluation.

Optimal feedforward control gain computation: Let $j^* \leftarrow j$ and $i \leftarrow i+1$, repeat solving $\Omega(X_i)$ by using (41) until $i = m+1$. Then, solve for the solutions (X, U) of the regulator equations (12)–(13) using (27) and (42). At last, calculate the optimal feedforward gain $L^* = U + K_{j^*}X$.

Theorem 1. *Consider $\{P_j\}_{j=0}^{\infty}$ and $\{K_j\}_{j=0}^{\infty}$ obtained by Algorithm 2. Under Assumptions 1 and 4, if Lemma 2 is satisfied, one has the following properties,*

$$\lim_{j\to\infty} P_j = P^*, \tag{44}$$

$$\lim_{j\to\infty} K_j = K^*. \tag{45}$$

Proof. Under the rank condition in Lemma 2, Eq. (41) has a unique solution. Then, H_j, K_j and $P_j(\hat{D} - \Omega(X_i))$ obtained by Algorithm 2 must satisfy Eq. (39). This implies that $\tilde{P}_{j+1}$ and P_{j+1} are equivalent to the ones in the model-based VI Algorithm 1. By Lemma 1, the convergences of P_j and K_j is ensured. The proof of Theorem 1 is completed. □

Theorem 1 shows the convergence of the online VI algorithm, which uses online data to approximate the optimal feedback control gain and the corresponding feedforward control gain. Under the obtained control gains, one can design an approximate optimal controller to solve the LO^2RP. We will present the main results in the following theorem.

Theorem 2. *Consider the disturbed CT linear system in (1)–(2) with the exosystem in (3) and the reference signal in (4). Under Assumptions 1–4, the closed-loop system with the learned controller $u(t) = -K_{j^*}x(t) + L^*w(t)$ by Algorithm 2 is GAS. Moreover, the convergence rate of the state and tracking error is ensured to be no slower than $e^{-\gamma t}$.*

Proof. The system (1)–(2) with the controller $u(t) = -K_jx(t) + L^*w(t)$ is equivalent to the following system

$$\dot{\bar{x}}(t) = (\bar{A} - BK_j)\bar{x}(t), \tag{46}$$

$$\bar{e}(t) = C\bar{x}(t) + (CX + \hat{F})\bar{w}(t). \tag{47}$$

Given a stabilizing feedback control gain matrix K_j, one has that $\lim_{t\to\infty}\bar{x}(t) = 0$. This implies that $\lim_{t\to\infty}\bar{e}(t) = \lim_{t\to\infty}C\bar{x}(t) = 0$ and the convergence rate of $x(t)$ and $e(t)$ is no slower than $e^{-\gamma t}$. The proof of Theorem 2 is completed. ■

Remark 3. The probing noise $\hat{e}(t)$ introduced in the control input assures the rank condition in Lemma 2 is satisfied. Moreover, Algorithm 2 is subject to off-policy RL algorithms. The control policy that generates the online data is different from the control policy that is evaluated and improved, and interestingly the probing noise does not affect the accuracy of solutions at each iteration (Jiang, Kiumarsi, et al., 2020; Jiang et al., 2021; Kiumarsi et al., 2017; Li et al., 2018).

5. Application on a LCL coupled inverter-based distributed generation system

In the last bidecade, the distributed generation system has been under extensive investigations due to its short build time (see Ahmed et al., 2010, and references therein). In order to ensure the distributed generation system operates smoothly in the grid-connected mode, passive power LCL filters are usually introduced to connect the grid and the inverter, which formulates LCL-coupled inverter-based distributed generation systems. The control goal of these systems is designing a controller such that (1) the output current asymptotically follows the desired trajectory, (2) the effect from the grid voltage, the disturbance, is rejected. When the dynamics of distributed generation system model are accurately known, one can use the approaches developed in Ahmed et al. (2010) to realize the control goal. However, to identify the system dynamics and measure the grid voltage perfectly is almost impossible in practice due to the resistance, the inductance and the capacitance vary with circuit operation and temperature. Notably, the proposed VI algorithm is a good candidate to control distributed generation systems in an optimal sense with unknown system model and unmeasurable disturbance (see Figs. 5 and 6). Therefore, in this section, we apply the proposed approach to a LCL coupled inverter-based distributed generation system to illustrate the effectiveness. The system dynamics is shown as follows (Ahmed et al., 2010) and illustrated in Fig. 1,

$$V_I = I_LR_1 + L_1\frac{dI_L}{dt} + V_C,$$

$$I_L = I_O + C\frac{dV_C}{dt},$$

$$V_C = I_OR_2 + L_2\frac{dI_O}{dt} + V_G,$$

where the physical significance and the value of parameters V_I, I_L, R_1, L_1, V_C, I_O, C, R_2, L_2 and V_G can be found in Table 1. By considering $x := [I_L, V_C, I_O]^{\mathrm{T}}$, $y := I_O$ and $u := V_I$, $d := [0, 0, -V_G/L_2]^{\mathrm{T}}$ as the state

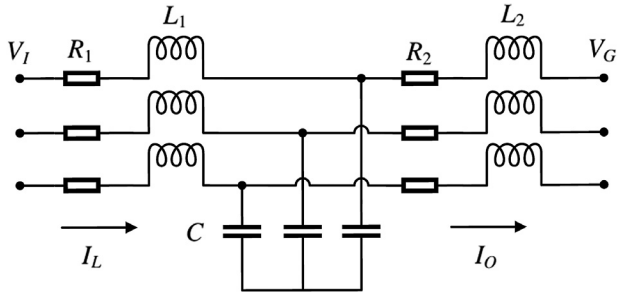

Fig. 1. Block diagram of LCL coupled inverter-based distributed generation system.

Table 1
Physical meaning and value of parameter.

Parameter	Physical meaning	Value	Unit
V_I	Input voltage	–	V
I_L	Inductor current	–	A
R_1	Filter resistor	0.2	Ω
L_1	Filter inductor	2.5	mH
V_C	Capacitor voltage	–	V
I_O	Output current	–	A
C	Capacitor	30	μF
R_2	Transformer resistor	2.5	Ω
L_2	Transformer inductor	5.5	mH
V_G	Grid voltage	–	V

vector, output, input and disturbance, one can obtain the following disturbed LTI CT state-space model,

$$\dot{x}(t)=\begin{bmatrix} -\frac{R_1}{L_1} & -\frac{1}{L_1} & 0 \\ \frac{1}{C} & 0 & -\frac{1}{C} \\ 0 & \frac{1}{L_2} & -\frac{R_2}{L_2} \end{bmatrix} x(t)+\begin{bmatrix} \frac{1}{L_1} \\ 0 \\ 0 \end{bmatrix} u(t)+d(t),$$

$$e(t)=x_3(t)-y_d(t),$$

with A, B and C in system (1)–(2) being

$$A=\begin{bmatrix} -0.08 & -0.4 & 0 \\ 33.33 & 0 & -33.33 \\ 0 & 0.1818 & -0.5 \end{bmatrix}, B=\begin{bmatrix} 0.4 \\ 0 \\ 0 \end{bmatrix},$$

$$C=\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

The reference signal $y_d(t)$ is chosen as $y_d(t)=-10\sin(100\pi t+\pi/3)$; V_G is the sum of sinusoidal waves with unknown phase and amplitude, but the frequencies are known and assumed to be 50 Hz and 60 Hz. Therefore, the exosystem (5) has the following form,

$$\dot{w}(t)=\begin{bmatrix} 0 & -100\pi & 0 & 0 \\ 100\pi & 0 & 0 & 0 \\ 0 & 0 & 0 & -120\pi \\ 0 & 0 & 120\pi & 0 \end{bmatrix} w(t).$$

The initial condition is $w(0)=[1,0,1,0]^{\mathrm{T}}$, and $\hat{F}$ is

$$\hat{F}=[5\sqrt{3},5,0,0]^{\mathrm{T}}.$$

Beside, in this section, we assume that $V_G=[1,0,2,0]w(t)$ and hence $\hat{D}$ is

$$\hat{D}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.5 & 0 & -1 & 0 \end{bmatrix}.$$

Note that $\hat{D}$ is only used to generate online data, and is unknown when implementing the Algorithm 2. We select the parameters in (14) and (17) as $Q=\bar{Q}=I$, $R=\bar{R}=1$ and $\gamma=0.5$. Therefore, the solutions

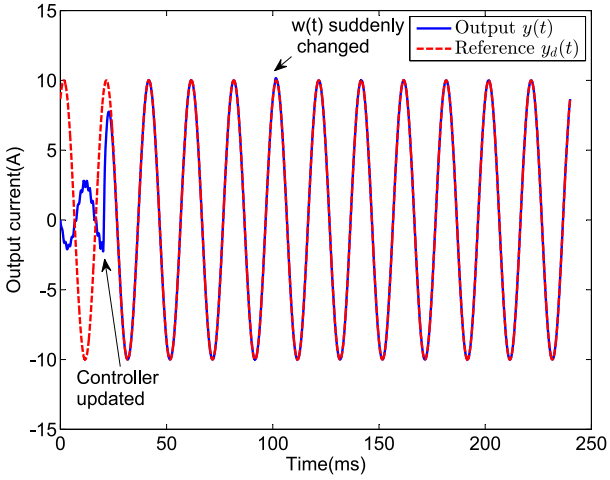

Fig. 2. Trajectories of output $y(t)$ and reference $y_d(t)$ via Algorithm 2.

of the regulator equations (12)–(13), and the idea values of P, K and L are

$$X=\begin{bmatrix} 8.6229 & 4.7131 & 0 & -0.0622 \\ 30.4419 & -3.9638 & 5.5 & 0 \\ 8.6603 & 5 & 0 & 0 \end{bmatrix},$$

$$U=\begin{bmatrix} 35.8687 & -9.7936 & 5.4414 & -0.0124 \end{bmatrix},$$

$$P^*=\begin{bmatrix} 26.5395 & 1.341 & -16.4614 \\ 1.341 & 0.5421 & -0.4992 \\ -16.4614 & -0.4992 & 45.3882 \end{bmatrix},$$

$$K^*=\begin{bmatrix} 10.6158 & 0.5364 & -6.5845 \end{bmatrix},$$

$$L^*=\begin{bmatrix} 86.7127 & 5.1906 & 8.3916 & -0.6727 \end{bmatrix}.$$

In the simulation experiment result, the parameters are selected as follows: $\varepsilon=0.01$, $\epsilon_j=0.1/\sqrt{j}$, $B_q=\{X>0\|\|X\|<30(q+1)\}$, $s=40$, $\hat{e}(t)=0.2\sin(15t)+\sin(5t)\sin(8t)$, $K_0=[-1,0,0]$, $x(0)=[0,0,0]^{\mathrm{T}}$, X_i are

$$X_0=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, X_1=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 5\sqrt{3} & 5 & 0 & 0 \end{bmatrix},$$

$$X_2=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, X_3=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$X_4=\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, X_5=\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$X_6=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, X_7=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$X_8=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, X_9=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then, one can compute the eigenvalues of the matrix $A-BK_0$ are $0.048+4.391i$, $0.048-4.391i$ and -0.176. Clearly, the initial feedback control gain matrix K_0 is not admissible stabilizing. In the simulation experiment, $t_0=0$s, $t_1=0.5$s, $t_2=1$s, …, $t_s=20$s. Therefore, the control input is $u(t)=K_0x(t)+\hat{e}(t)$ when $t\in[0,20]$ and updated as $u(t)=-K_{j^*}x(t)+L^*w(t)$ after 20s. Moreover, when $t>100$, the 60 Hz wave in the grid voltage disappears and hence $w_3(t)=0$ and $w_4(t)=0$. Based on these parameters and situations descriptions, the simulation experiment

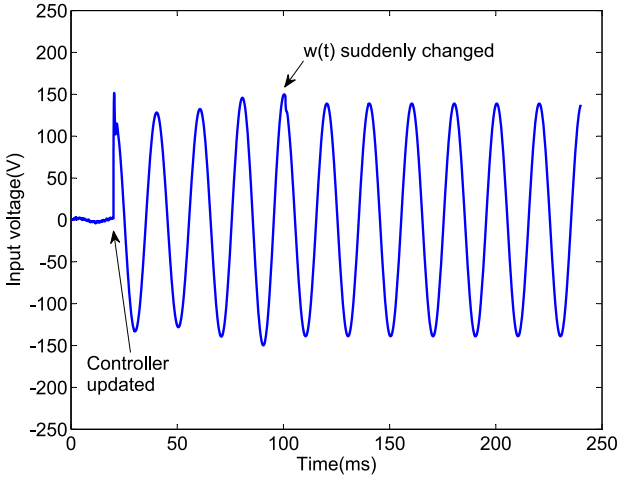

Fig. 3. Trajectory of input $u(t)$ via Algorithm 2.

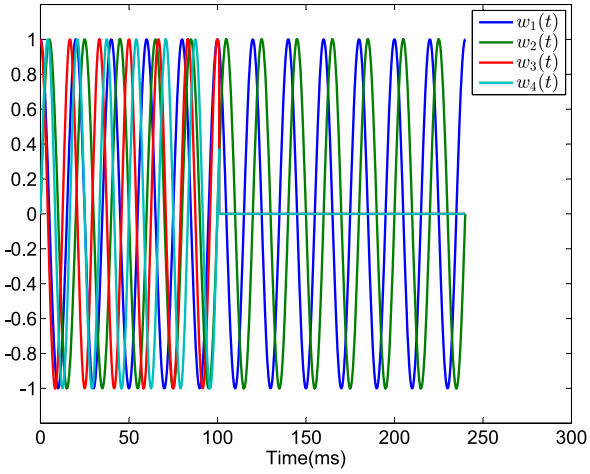

Fig. 4. Trajectory of $w(t)$.

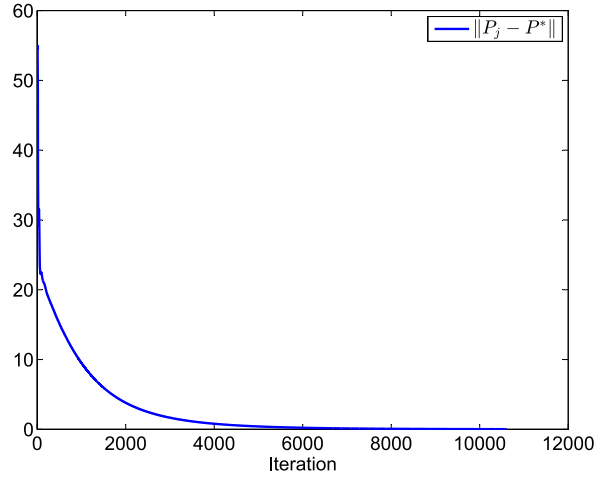

Fig. 5. Trajectory of $\|P_j - P^*\|$ via Algorithm 2.

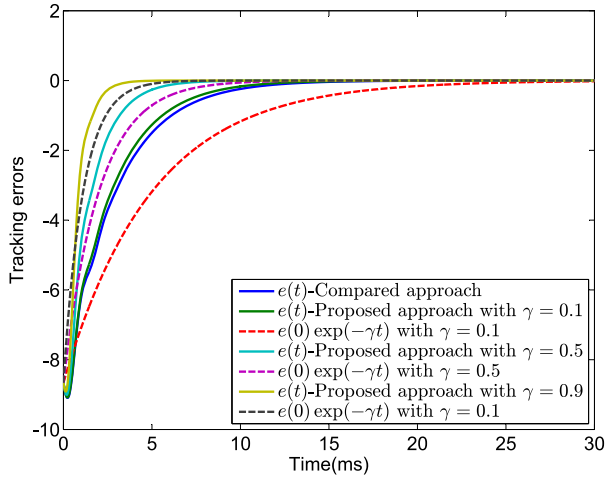

Fig. 6. Compared simulation results with standard PI based approach in Gao and Jiang (2016) and different convergence rate criterion γ.

result is shown. Figs. 2–3 show trajectories of output $y(t)$, reference $y_d(t)$ and input $u(t)$ via Algorithm 2. Fig. 4 shows the trajectory of $w(t)$ while Fig. 5 shows the trajectory of $\|P_j - P^*\|$ via Algorithm 2. It can be observed that the output of the controlled plant can be regulated via the learned feedback control gain and the learned feedforward control gain by Algorithm 2 even with $w(t)$ suddenly changing. The learned solutions of the regulator equations (12)–(13), and the learned values of X, U, P_j, K_j and L are

$$X = \begin{bmatrix} 8.6229 & 4.7131 & 0 & -0.0622 \\ 30.4419 & -3.9638 & 5.5 & 0 \\ 8.6603 & 5 & 0 & 0 \end{bmatrix},$$
$$U = \begin{bmatrix} 35.8687 & -9.7936 & 5.4414 & -0.0124 \end{bmatrix},$$
$$P_j = \begin{bmatrix} 26.5372 & 1.3408 & -16.4681 \\ 1.3408 & 0.5421 & -0.4999 \\ -16.4681 & -0.4999 & 45.3885 \end{bmatrix},$$
$$K_j = \begin{bmatrix} 10.6149 & 0.5363 & -6.5873 \end{bmatrix},$$
$$L = \begin{bmatrix} 86.6787 & 5.1733 & 8.3911 & -0.6727 \end{bmatrix}.$$

Then, we provide a compared simulation experiment with standard PI based approach in Gao and Jiang (2016) to show that the proposed approach can tune the convergence rate of $e(t)$ by regulating the convergence rate criterion γ, which has been proved in Theorem 2. If the initial control feedback gain is selected such that $A - BK_0$ is not Hurwitz, cannot learn the optimal feedback control gain and the solutions of the regulator equations (12)–(13). Therefore, we select an admissible initial control feedback gain for the compared approach. In the compared simulation, there are four tracking errors based on standard PI based approach in Gao and Jiang (2016) and the proposed approach with different convergence rate criterion γ, and we choose the same initial conditions as $x(0) = [0, 0, 0]^{\mathrm{T}}$ and $w(0) = [1, 0, 1, 0]^{\mathrm{T}}$. The compared simulation results with standard PI based approach in Gao and Jiang (2016) and different convergence rate criterion γ are depicted in Fig. 6. It can be observed from Fig. 6 that: (1) the convergence rate of the tracking errors based on the proposed approach is faster then that of the tracking error based on the compared approach; (2) the convergence rate of the tracking errors can be tuned by choosing different convergence rate criterion γ and is faster then $e^{-\gamma t}$.

At last, we provide a compared simulation experiment with linear quadratic tracking control approach in Modares and Lewis (2014a) to show that the proposed approach can result in performance advantages. In Modares and Lewis (2014a), the controller has the following form,

$$u(t) = -\bar{K}\begin{bmatrix} x(t) \\ w(t) \end{bmatrix} = R^{-1}\bar{B}^{\mathrm{T}}\bar{P}^*\begin{bmatrix} x(t) \\ w(t) \end{bmatrix},$$

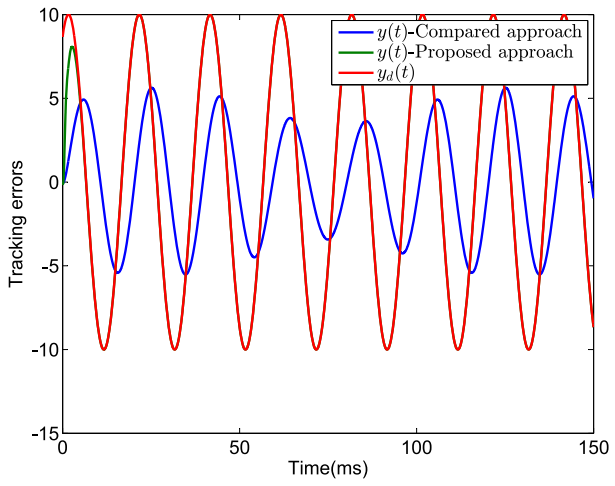

Fig. 7. Compared simulation results with linear quadratic tracking control approach in Modares and Lewis (2014a).

where $\bar{P}^*$ is the positive definite solution of the following linear quadratic tracking ARE,

$$\bar{P}T + T^{\mathrm{T}}\bar{P} + \bar{C}^{\mathrm{T}}\bar{Q}\bar{C} - \bar{P}\bar{B}R^{-1}\bar{B}^{\mathrm{T}}\bar{P} = 0,$$

with

$$T = \begin{bmatrix} A & \hat{D} \\ 0 & \hat{E} \end{bmatrix} - 0.5\alpha I, \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \bar{C} = \begin{bmatrix} C & \hat{F} \end{bmatrix}, \bar{Q} > 0.$$

In the compared simulation experiment, we select the parameters as $\alpha = 0.2$ and $\bar{Q} = 100000$ and we choose the same initial conditions as $x(0) = [0, 0, 0]^{\mathrm{T}}$ and $w(0) = [1, 0, 1, 0]^{\mathrm{T}}$. The compared simulation results with linear quadratic tracking control approach in Modares and Lewis (2014a) are depicted in Fig. 7. It can be observed from Fig. 7 that the output of the controlled plant can be regulated via Algorithm 2 while the compared approach in Modares and Lewis (2014a) can attenuate the disturbance but not make the tracking error asymptotically converge to zero.

6. Conclusion

In this paper, we studied the LO^2RP with assured convergence rate requirement under the challenges from the unknown system and exosystem dynamics. Without relying on the knowledge of system dynamics and initial stabilizing feedback control gain, a novel VI algorithm is proposed, which is capable of learning the optimal regulator using online data with a guaranteed convergence rate. An application to a LCL coupled inverter-based distributed generation system shows the efficiency of the learning-based output regulation approach.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmed, K. H., Massoud, A. M., Finney, S. J., & Williams, B. W. (2010). A modified stationary reference frame-based predictive current control with zero steady-state error for LCL coupled inverter-based distributed generation systems. *IEEE Transactions on Industrial Electronics*, *58*(4), 1359–1370.

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*(4), 943–949.

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

Chen, C., Modares, H., Xie, K., Lewis, F. L., Wan, Y., & Xie, S. L. (2019). Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics. *IEEE Transactions on Automatic Control*, *64*(11), 4423–4438.

Fan, J. L., Wu, Q., Jiang, Y., Chai, T. Y., & Lewis, F. L. (2020). Model-free optimal output regulation for linear discrete-time lossy networked control systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(11), 4033–4042.

Francis, B. A. (1977). The linear multivariable regulator problem. *SIAM Journal on Control and Optimization*, *15*(3), 486–505.

Francis, B. A., & Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, *12*(5), 457–465.

Gao, W., & Jiang, Z. P. (2016). Adaptive dynamic programming and adaptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control*, *61*(12), 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*(6), 2614–2624.

Gao, W., & Jiang, Z. P. (2019). Adaptive optimal output regulation of time-delay systems via measurement feedback. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(3), 938–945.

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*(10), 3581–3587.

Gao, W., Mynuddin, M., Wunsch, D. C., & Jiang, Z.-P. (2021). Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/TNNLS.2021.3069728, (in press).

He, S. Z., Tan, S. H., Xu, F. L., & Wang, P. Z. (1993). Fuzzy self-tuning of PID controllers. *Fuzzy Sets & Systems*, *56*(1), 37–46.

Hong, Y. G., Xu, Y. S., & Huang, J. (2002). Finite-time control for robot manipulators. *Systems & Control Letters*, *46*(4), 243–253.

Huang, J. (2004). *Nonlinear output regulation: Theory and applications*. SIAM.

Huang, X., Yan, Y., & Huang, Z. (2017). Finite-time control of underactuated spacecraft hovering. *Control Engineering Practice*, *68*, 46–62.

Jiang, Z.-P., Bian, T., & Gao, W. (2020). Learning-based control: A tutorial and some recent results. *Foundations and Trends in Systems and Control*, *8*(3), 176–284.

Jiang, Y., Fan, J. L., Chai, T. Y., & Lewis, F. L. (2019). Dual-rate operational optimal control for flotation industrial process with unknown operational model. *IEEE Transactions on Industrial Electronics*, *66*(6), 4587–4599. http://dx.doi.org/10.1109/TIE.2018.2856198.

Jiang, Y., Fan, J. L., Chai, T. Y., Lewis, F. L., & Li, J. N. (2018). Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4607–4620. http://dx.doi.org/10.1109/TNNLS.2017.2771459.

Jiang, Y., Fan, J. L., Chai, T. Y., Li, J. N., & Lewis, F. L. (2018). Data-driven flotation industrial process operational optimal control based on reinforcement learning. *IEEE Transactions on Industrial Informatics*, *14*(5), 1974–1989. http://dx.doi.org/10.1109/TII.2017.2761852.

Jiang, Y., Fan, J. L., Gao, W., Chai, T. Y., & Lewis, F. L. (2020). Cooperative adaptive optimal output regulation of discrete-time nonlinear multi-agent systems. *Automatica*, *121*, Article 109149.

Jiang, Y., Kiumarsi, B., Fan, J. L., Chai, T. Y., Li, J. N., & Lewis, F. L. (2020). Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. *IEEE Transactions on Cybernetics*, *50*(7), 3147–3156.

Jiang, Y., Zhang, K., Wu, J., Zhang, C. X., Xue, W. Q., Chai, T. Y., & Lewis, F. L. (2021). H_∞-based minimal energy adaptive control with preset convergence rate. *IEEE Transactions on Cybernetics*, http://dx.doi.org/10.1109/TCYB.2021.3061894, (in press).

Kamalapurkar, R., Walters, P., Rosenfeld, J., & Dixon, W. E. (2018). *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*. Springer.

Kiumarsi, B., & Lewis, F. L. (2015). Actor–critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(1), 140–151.

Kiumarsi, B., Lewis, F. L., & Jiang, Z. P. (2017). H_∞ Control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, *78*, 144–152.

Kiumarsi, B., Lewis, F. L., Modares, H., Karimpour, A., & Naghibi-Sistani, M.-B. (2014). Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, *50*(4), 1167–1175.

Kiumarsi, B., Vamvoudakis, K. G., Modares, H., & Lewis, F. L. (2018). Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2042–2062.

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

Knobloch, H. W., Isidori, A., & Flockerzi, D. (2012). *Topics in control theory, vol. 22*. Birkhäuser.

Li, J. N., Chai, T. Y., Lewis, F. L., Ding, Z. T., & Jiang, Y. (2018). Off-policy interleaved Q-learning: Optimal control for affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(5), 1308–1320.

Liu, D. R., Xue, S., Zhao, B., Luo, B., & Wei, Q. L. (2021). Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on*

Systems, Man, and Cybernetics: Systems, *51*(1), 142–160. http://dx.doi.org/10.1109/TSMC.2020.3042876.

Modares, H., & Lewis, F. L. (2014a). Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, *59*(11), 3051–3056.

Modares, H., & Lewis, F. L. (2014b). Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, *50*(7), 1780–1792.

Park, B. S., Yoo, S. J., Park, J. B., & Choi, Y. H. (2009). Adaptive neural sliding mode control of nonholonomic wheeled mobile robots with model uncertainty. *IEEE Transactions on Control Systems Technology*, *17*(1), 207–214.

Vamvoudakis, K. G., & Kokolakis, N.-M. T. (2020). Synchronous reinforcement learning-based control for cognitive autonomy. *Foundations and Trends in Systems and Control*, *8*(1–2), 1–175.

Wang, D., Ha, M. M., & Qiao, J. F. (2020). Data-driven iterative adaptive critic control toward an urban wastewater treatment plant. *IEEE Transactions on Industrial Electronics*, *68*(8), 7362–7369.

Wang, D., He, H. B., & Liu, D. R. (2017). Adaptive critic nonlinear robust control: A survey. *IEEE Transactions on Cybernetics*, *47*(10), 3429–3451.

Wang, D., & Huang, J. (2005). Neural network-based adaptive dynamic surface control for a class of uncertain nonlinear systems in strict-feedback form. *IEEE Transactions on Neural Networks*, *16*(1), 195–202.

Wang, D., Qiao, J. F., & Cheng, L. (2020). An approximate neuro-optimal solution of discounted guaranteed cost control design. *IEEE Transactions on Cybernetics*, http://dx.doi.org/10.1109/TCYB.2020.2977318, (in press).

Wang, D., Zhao, M. M., & Qiao, J. F. (2021). Intelligent optimal tracking with asymmetric constraints of a nonlinear wastewater treatment system. *International Journal of Robust and Nonlinear Control*, *31*(14), 6773–6787.

Woo, Z. W., Chung, H. Y., & Lin, J. J. (2000). A PID type fuzzy controller with self-tuning scaling factors. *Fuzzy Sets & Systems*, *115*(2), 321–326.

Wu, Q., Fan, J., Jiang, Y., & Chai, T. (2019). Data-driven dual-rate control for mixed separation thickening process in a wireless network environment. *Acta Automatica Sinica*, *45*(6), 1128–1141.

Xue, W. Q., Fan, J. L., Lopez, V. G., Li, J. N., Jiang, Y., Chai, T. Y., & Lewis, F. L. (2020). New methods for optimal operational control of industrial processes using reinforcement learning on two time-scales. *IEEE Transactions on Industrial Informatics*, *16*(5), 3085–3099.

Xue, W. Q., Fan, J. L., Mejia, V. G. L., Jiang, Y., Chai, T. Y., & Lewis, F. L. (2021). Off-policy reinforcement learning for tracking in continuous-time systems on two time-scales. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 4334–4346. http://dx.doi.org/10.1109/TNNLS.2020.3017461.

Zhou, B., Duan, G. R., & Lin, Z. L. (2008). A parametric Lyapunov equation approach to the design of low gain feedback. *IEEE Transactions on Automatic Control*, *53*(6), 1548–1554.