

Robust Output Regulation and Reinforcement Learning-based Output Tracking Design for Unknown Linear Discrete-Time Systems

Ci Chen, Lihua Xie, Yi Jiang, Kan Xie, and Shengli Xie

Abstract—In this paper, we investigate the optimal output tracking problem for linear discrete-time systems with unknown dynamics using reinforcement learning and robust output regulation theory. This output tracking problem only allows to utilize the outputs of the reference system and the controlled system, rather than their states, and differs from most existing tracking results that depend on the state of the system. The optimal tracking problem is formulated into a linear quadratic regulation problem by proposing a family of dynamic discrete-time controllers. Then, it is shown that solving the output tracking problem is equivalent to solving output regulation equations, whose solution, however, requires the knowledge of the complete and accurate system dynamics. To remove such a requirement, an off-policy reinforcement learning algorithm is proposed using only the measured output data along the trajectories of the system and the reference output. By introducing re-expression error and analyzing the rank condition of the parameterization matrix, we ensure the uniqueness of the proposed RL based optimal control via output feedback.

Index Terms—Reinforcement learning, robust output regulation, output tracking, adaptive optimal control.

I. INTRODUCTION

Output tracking, whose objective is to make the system output follow a desired reference trajectory, is a fundamental research topic of practical importance (see examples in [1]). One systematic way to approach the output tracking is to transform it into an output regulation problem, whose solution and corresponding control design date back to [2], where both the closed-loop stability and the asymptotic tracking of even an unbounded reference are achieved. Though elegant, such a solution is built on the knowledge of the complete and accurate system dynamics, making the output tracking design model dependent.

Reinforcement learning (RL) features making sequential decisions through interactions between the agent's actions and unknown environment [3], [4]. RL algorithms have been applied in the control field to solve optimal control problems for both discrete-time (DT) (see, e.g., [4], [5]) and continuous-time (CT) systems (see, e.g., [4]–[14]) without any knowledge of the system dynamics. RL-based methods in Chapter 11 of [4] were used in [15], [16] to handle the optimal tracking

control of linear and nonlinear DT systems. However, most RL algorithms are classified as on-policy learning, which assumes that the behaviour policy for generating the data for learning is the same as the target policy. Off-policy RL differs from the on-policy learning in its separating the target and behaviour policies. Off-policy RL for linear CT systems with unknown dynamics was given in Chapter 2 of [8] to solve the optimal regulation problem. A solution to the zero-sum game problem for the regulation of DT systems was given in [17] using off-policy RL. Together with RL and the output regulation theory, [18] and [19] gave tracking controllers for single and multiple DT systems and established the asymptotic stability of the tracking error, which is the DT version of [9]. Note that one essential assumption for [18], [19], as well as [9], is that not only the outputs of the reference and the controlled system but also their states are required in the optimal control design process. Therefore, a key challenge to be addressed is how to achieve the optimal output tracking of DT systems without using the reference state.

In contrast to full states feedback, output feedback utilizes the system output and adds extra flexibilities to control design. As for the output-feedback control, [20] utilized the state reconstruction to form an output-feedback RL for regulating the states of DT systems. Based on [20], off-policy H_∞ control of DT systems was given based on input and output data in the literature such as [21], [22]. The optimal tracking control using output-feedback RL algorithm was given in [23]. [24] gave a parameterization of the state of a DT system in terms of the input and output data for the output feedback LQR by Q -learning. Based on [24], [25] considered an output-feedback tracking control for DT systems with the assumption that the reference state is available during the learning and control processes. Utilizing the state reconstruction in [20], [26], [27] considered tracking-based RL algorithms for disturbance rejection with the augmented system being observable. It turns out that, in the state parametrization, the full row rank of the parametrization matrix is required, see [7], [12], [28]–[31] for example. In [29]–[31], sufficient conditions were established that guarantee such a rank condition in both the CT and DT settings. Within the framework of output-feedback RL, most of the existing output tracking results for DT systems are based on the linear output regulation theory, which, however, remains challenging for its extension to robust output regulation.

Motivated by the analysis above, we investigate the output tracking problem of DT systems using RL and robust output regulation theory based on the reference output. We use real-time data collected along the trajectories of the DT system and propose an RL algorithm that ensures not only the stability of the closed-loop system, but also the predefined system performance. The contributions of this paper are given as follows.

- 1) Compared to the linear output regulation based RL controllers for tracking DT systems [18], [19], [25]–[27], we amend the standard robust output regulation theory to propose a new family of robust DT controllers, based on which an optimal output tracking problem is formulated.
- 2) We invoke the CT work in [12] to propose a criterion for guaranteeing the parameterization matrix in the DT state

This work was supported in part by the Wallenberg-NTU Presidential Postdoctoral Fellowship and in part by National Natural Science Foundation of China under Grants 61703112, 61973087, and 61727810. (Corresponding author: Lihua Xie).

C. Chen and L. Xie are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. (e-mail: ci.chen@control.lth.se, ELHXIE@ntu.edu.sg).

Y. Jiang is with State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. (e-mail: yi-jiang.ezhou@gmail.com).

K. Xie and S. Xie are with School of Automation, Guangdong University of Technology, Guangdong Key Laboratory of IoT Information Technology, Guangzhou, 510006 China (e-mail: shlxie@gdut.edu.cn).

reconstruction to be full row rank, which leads to the uniqueness of the data-based optimal output-feedback DT controller. Inspired by [12], we establish the rank of the parameterization matrix for DT systems, and based on which we show that the controllability is a sufficient condition for ensuring the full row rank of the parameterization matrix for DT systems.

- 3) We do not need to compute analytic solutions to the output regulation equations. We take the DT state re-expression error into account during the optimal control learning. We add a phase of model-free pre-collection into the off-policy RL for reducing the adverse impact of the re-expression error.

Notation: Throughout this paper, given any square matrix \mathcal{M} , the notation $\lambda(\mathcal{M})$ indicates the spectrum of \mathcal{M} , $\rho(\mathcal{M})$ is its spectral radius, $\det(\mathcal{M})$ denotes its determinant, $\text{adj}(\mathcal{M})$ is its adjugate, $\mathcal{M} > 0$ ($\mathcal{M} \geq 0$) means that the matrix is positive definite (positive semi-definite), $\text{vec}(\mathcal{M}) = [\mathcal{M}_1^T, \mathcal{M}_2^T, \dots, \mathcal{M}_n^T]^T$ is a vector with \mathcal{M}_i being the i th column of \mathcal{M} , and $\text{vecs}(\mathcal{W}) = [\mathcal{W}_{11}, 2\mathcal{W}_{12}, \dots, 2\mathcal{W}_{1n}, \mathcal{W}_{22}, 2\mathcal{W}_{23}, \dots, 2\mathcal{W}_{n-1,n}, \mathcal{W}_{n,n}]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}$ with \mathcal{W}_{ij} being the entry at the i th row and j th column of \mathcal{W} . Given any vector $\mathcal{T} \in \mathbb{R}^n$, $\text{vecv}(\mathcal{T}) = [\mathcal{T}_1^2, \mathcal{T}_1\mathcal{T}_2, \dots, \mathcal{T}_1\mathcal{T}_n, \mathcal{T}_2^2, \mathcal{T}_2\mathcal{T}_3, \dots, \mathcal{T}_{n-1}\mathcal{T}_n, \mathcal{T}_n^2]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}$. Let the controllability matrix $\mathcal{C}(A_1, A_2) = [A_2, A_1A_2, A_1^2A_2, \dots, A_1^{n-1}A_2]$ with the dimension of A_1 being $n \times n$. The notation \otimes denotes the Kronecker product. The notations 0 and I , respectively, indicate a zero matrix and an identity matrix with appropriate dimensions.

II. PROBLEM FORMULATION AND PRELIMINARIES

A. Problem Formulation

We aim to study a class of DT dynamical systems modeled by

$$x(k+1) = Ax(k) + Bu(k), \quad (1)$$

$$y(k) = Cx(k), \quad (2)$$

where $x(k) \in \mathbb{R}^{r_n}$, $u(k) \in \mathbb{R}^{r_m}$, and $y(k) \in \mathbb{R}^{r_p}$ denote the system state, input, and output, respectively, and, $A \in \mathbb{R}^{r_n \times r_n}$, $B \in \mathbb{R}^{r_n \times r_m}$, and $C \in \mathbb{R}^{r_p \times r_n}$, involved in the system dynamics (1) and (2), are unknown constant matrices. The control objective is to drive $y(k)$ in (2) to follow a reference signal $y_d(k)$ given by

$$x_d(k+1) = Sx_d(k), \quad (3)$$

$$y_d(k) = Rx_d(k), \quad (4)$$

where $x_d(k) \in \mathbb{R}^{r_{qm}}$ and $y_d(k) \in \mathbb{R}^{r_p}$ respectively denote the reference state and output. Similar to the (A, B, C) in (1) and (2), $S \in \mathbb{R}^{r_{qm} \times r_{qm}}$ and $R \in \mathbb{R}^{r_p \times r_{qm}}$ are unknown constant matrices. The output tracking error between (2) and (4) is defined as

$$y_e(k) = y(k) - y_d(k). \quad (5)$$

For the purpose of control design, some assumptions on the dynamics of DT system (1)–(4) are made as follows.

Assumption 1: For system (1)–(2), (A, B) is controllable and (A, C) is observable.

Assumption 2: The eigenvalues of the matrix S are on or outside the unit circle.

Assumption 3: The minimal polynomial of S is known.

Assumption 4: $\text{rank}\left(\begin{bmatrix} A - \lambda_i I & B \\ C & 0 \end{bmatrix}\right) = r_n + r_p, \forall \lambda_i \in \lambda(S)$.

Assumption 1 is standard in the optimal control [4]. *Assumption 2* is made to rule out the trivial case wherein the matrix S in (3) is Schur (see Chapter 1 of [32]), namely, the reference state is asymptotically stable. *Assumptions 2–4* are standard in the output regulation literature.

Based on the system descriptions above, we focus on formulating an optimal tracking control problem that is solvable. To do this, we extend the work of CT systems [12] and construct a DT control protocol

$$z(k+1) = Fz(k) - Gy_e(k), \quad (6a)$$

$$u(k) = -Kx(k) - Hz(k) - Tz(k), \quad (6b)$$

where $z(k) \in \mathbb{R}^{r_p r_{qm}}$ is a dynamical signal driven by output error $y_e(k)$, (F, G) is an r_p -copy internal model of S^{-1} with $F \in \mathbb{R}^{r_p r_{qm} \times r_p r_{qm}}$ and $G \in \mathbb{R}^{r_p r_{qm} \times r_p}$, $T \in \mathbb{R}^{r_m \times r_p r_{qm}}$ is a newly proposed feedforward gain matrix compared to the standard DT controller by the output regulation theory [32], and $K \in \mathbb{R}^{r_m \times r_n}$ and $H \in \mathbb{R}^{r_m \times r_p r_{qm}}$ are gain matrices for solving the tracking problem to be specified later.

Substituting the dynamical controller (6) into the DT system (1) yields

$$\bar{x}(k+1) = \bar{A}(k)\bar{x}(k) + \bar{G}Rx_d(k), \quad (7a)$$

$$y_e(k) = \bar{C}\bar{x}(k) - Rx_d(k), \quad (7b)$$

where $\bar{x}(k) = [x(k)^T, z(k)^T]^T \in \mathbb{R}^{n_z}$ with $n_z = r_n + r_p r_{qm}$, and the system matrices $\bar{A} \in \mathbb{R}^{n_z \times n_z}$, $\bar{G} \in \mathbb{R}^{n_z \times r_p}$, and $\bar{C} \in \mathbb{R}^{r_p \times n_z}$ are, respectively, given as $\bar{A} = \begin{bmatrix} A - BK & -BH - BT \\ -GC & F \end{bmatrix}$, $\bar{G} = \begin{bmatrix} 0 \\ G \end{bmatrix}$, and $\bar{C} = [C, 0]$. Note that

$$\bar{A} = \underline{A} - \bar{B}[K, H], \quad (8)$$

where $\underline{A} = \begin{bmatrix} A & -BT \\ -GC & F \end{bmatrix} \in \mathbb{R}^{n_z \times n_z}$ and $\bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix} \in \mathbb{R}^{n_z \times r_m}$. By [12], the following properties of the system dynamics hold.

Lemma 1: Under *Assumption 4* and when (F, G) incorporates an r_p -copy internal model of S , we have that

- 1) given any matrix T , if (A, B) is stabilizable (controllable), then (\underline{A}, \bar{B}) is stabilizable (controllable).
- 2) given the matrix T such that (F, T) is observable with $r_p \geq r_m$, if (A, C) is detectable (observable), then the pair (\underline{A}, \bar{C}) is detectable (observable). In addition, if (A, B) is stabilizable (controllable), then (\underline{A}, \bar{B}) is stabilizable (controllable).

¹The design of (F, G) is available under *Assumption 3*; see Chapter 1 of [32].

Given a Schur matrix \bar{A} , under *Assumptions 1–2*, the equations $\bar{A}X + \bar{G}R = XS$ and $\bar{C}X = R$ have a common solution X [32]. Based on the matrix X , let

$$e(k) = \bar{x}(k) - Xx_d(k), \quad (9)$$

which leads to

$$e(k+1) = \bar{A}e(k), \quad (10)$$

where (7) and (9) are used. Substituting (8) into (10) yields

$$e(k+1) = \underline{A}e(k) + \bar{B}u_e(k), \quad (11a)$$

$$y_e(k) = \bar{C}e(k), \quad (11b)$$

$$u_e(k) = -\bar{K}e(k), \quad (11c)$$

where $\bar{K} = [K, H] \in \mathbb{R}^{r_m \times (r_m + r_p r_{q_m})}$ is the feedback gain matrix with K and H being from (6). Note that if $e(k)$ decays to zero, so does $y_e(k)$. Here, the matrix \bar{K} in (11c) is designed based on the following optimization problem.

Problem 1:

$$\min_{u_e} \sum_{i=k}^{\infty} (y_e^T(i)Qy_e(i) + u_e^T(i)\bar{R}u_e(i)) \quad (12)$$

subject to (11)

with $Q > 0$ and $\bar{R} > 0$.

B. Preliminaries on Optimal Control

The optimal tracking problem is solved if \bar{K} in (11c) is designed such that $e(k)$ in (11b) is stabilized to zero and, meanwhile, the performance index in (12) is minimized.

The optimal feedback gain matrix that solves the optimization problem (12) is labelled as \bar{K}^* and is given as follows (see Chapter 2 of [4]),

$$\bar{K}^* = (\bar{R} + \bar{B}^T P^* \bar{B})^{-1} \bar{B}^T P^* \underline{A}, \quad (13)$$

where P^* satisfies the DT algebraic Riccati equation (ARE)

$$\begin{aligned} \underline{A}^T P^* \bar{B} (\bar{R} + \bar{B}^T P^* \bar{B})^{-1} \bar{B}^T P^* \underline{A} \\ = \bar{C}^T \bar{Q} \bar{C} + \underline{A}^T P^* \underline{A} - P^* \end{aligned} \quad (14)$$

with $\bar{Q} = \text{diag}\{Q, 0\}$ and assuming that the stabilizability condition of (\underline{A}, \bar{B}) and the observability condition of (\underline{A}, \bar{C}) hold. Note that the detectability condition of (\underline{A}, \bar{C}) may not hold through the design of the output regulation-based standard DT controller (see Chapter 1 of [32]).

As for solving the DT ARE (14), a model-based algorithm with the knowledge of \underline{A} and \bar{B} was given in [33], and is recalled below.

Lemma 2: ([33]) Let \bar{K}^0 be a stabilizing gain matrix such that $\underline{A} - \bar{B}\bar{K}^0$ is Schur. Solve P^j from

$$P^j = \bar{Q} + (\bar{K}^j)^T \bar{R} \bar{K}^j + (\underline{A} - \bar{B}\bar{K}^j)^T P^j (\underline{A} - \bar{B}\bar{K}^j). \quad (15)$$

Update the policy as

$$\bar{K}^{j+1} = (\bar{R} + \bar{B}^T P^j \bar{B})^{-1} \bar{B}^T P^j \underline{A}. \quad (16)$$

Then,

- 1) $P^* \leq P^{j+1} \leq P^j$ for each $j = 1, 2, \dots$;
- 2) $\lim_{j \rightarrow \infty} \bar{K}^j = \bar{K}^*$, $\lim_{j \rightarrow \infty} P^j = P^*$.

III. OUTPUT-FEEDBACK RL FOR OPTIMAL OUTPUT TRACKING CONTROL OF DT SYSTEMS

This section is to solve *Problem 1* using the data collected along the trajectories of the controlled system and the reference output in the absence of any knowledge of the system dynamics. To achieve this, we use the following behavior policy to excite the DT system (1)

$$\bar{u}(k) = -\bar{K}^0 r(k) + \xi(k) - T\bar{z}(k), \quad (17)$$

where $r(k) = [x^T(k), \bar{z}^T(k)]^T \in \mathbb{R}^{n_z}$, \bar{K}^0 is an initial stabilizing gain, $\xi(k)$ is an exploration noise, and $\bar{z}(k) \in \mathbb{R}^{r_p r_{q_m}}$ is an alternative dynamical signal driven by the output error $y_e(k)$ as defined by $\bar{z}(k+1) = F\bar{z}(k) - Gy(k) + G\vartheta(k)$ with F and G being given from (6) and $\vartheta(k)$ being either the exploration noise or the reference's output $y_d(k)$. Here, $\bar{z}(k)$ is defined to differ from $z(k)$ in (6). We will specify $\vartheta(k)$ in our design later. Note that, if $\vartheta(k) = y_d(k)$ and $\xi(k) = 0$, then (17) is equivalent to (6). It follows from (1)–(4) and (17) that

$$r(k+1) = \underline{A}r(k) + \bar{B}\bar{u}(k) + \bar{G}\vartheta(k), \quad (18a)$$

$$y(k) = \bar{C}r(k). \quad (18b)$$

Based on the policy iteration of (15) and (16), (18) results in the off-policy Bellman equation for DT systems in the state-feedback form as

$$\begin{aligned} & r^T(k+1)P^{j+1}r(k+1) - r^T(k)P^{j+1}r(k) \\ &= -r^T(k)(\bar{Q} + (\bar{K}^j)^T \bar{R} \bar{K}^j)r(k) + \vartheta^T(k)\bar{G}^T P^{j+1} \bar{G}\vartheta(k) \\ & \quad + (-\bar{K}^j r(k) + \bar{u}(k))^T \bar{B}^T P^{j+1} \bar{B} (\bar{K}^j r(k) + \bar{u}(k)) \\ & \quad + 2\vartheta^T(k)\bar{G}^T P^{j+1} \bar{B}u(k) + 2r^T(k)\underline{A}^T P^{j+1} \bar{G}\vartheta(k) \\ & \quad + 2r^T(k)\underline{A}^T P^{j+1} \bar{B} (\bar{K}^j r(k) + \bar{u}(k)), \end{aligned} \quad (19)$$

where $\underline{A}^j = \underline{A} - \bar{B}\bar{K}^j$.

In (19), the system state $x(k)$ should be known. However, a variety of physical systems, including aircraft and power networks, only allow the measurement of the system output. Note that, within the output-feedback framework, the unknown term of $r(k)$ (or $x(k)$) prevents us from directly obtaining \bar{K}^{k+1} from (19). The next four subsections describe our output-feedback RL for solving *Problem 1*.

A. State Reconstruction

Since the system state $r(k)$ in (18) is not available for feedback control, this subsection provides a method to reconstruct $r(k)$ using input-output data from DT systems.

To reconstruct the DT state, we first generate the DT states $\zeta_{\bar{u}}(k) \in \mathbb{R}^{n_z r_m}$, $\zeta_y(k) \in \mathbb{R}^{n_z r_p}$, and $\zeta_{\vartheta}(k) \in \mathbb{R}^{n_z r_p}$ through three difference equations with zero initial conditions as

$$\zeta_{\bar{u}}(k+1) = (I_m \otimes A_{\zeta})\zeta_{\bar{u}}(k) + \bar{u}(k) \otimes b, \quad (20)$$

$$\zeta_y(k+1) = (I_p \otimes A_{\zeta})\zeta_y(k) + y(k) \otimes b, \quad (21)$$

$$\zeta_{\vartheta}(k+1) = (I_p \otimes A_{\zeta})\zeta_{\vartheta}(k) + \vartheta(k) \otimes b, \quad (22)$$

where A_{ζ} is a companion matrix with $-d_j$ for $j = 1, 2, \dots, n_z$ at the last row designed to make A_{ζ} Schur and $b = [0, 0, \dots, 0, 1]^T \in \mathbb{R}^{n_z}$.

The following result is attained by using the Luenberger observer for DT systems and extending [12], Chapters 3.6.1 and 4.5.4 of [34], [24].

Lemma 3: Consider the DT system (18). There exists a constant matrix $\bar{M} \in \mathbb{R}^{n_z \times r_{\bar{\zeta}}}$ satisfying

$$r(k) = \bar{M}\bar{\zeta}(k) + \omega(k), \quad (23)$$

where $\bar{\zeta}^T(k) = [\zeta_{\bar{u}}^T(k), \zeta_y^T(k), \zeta_{\bar{\theta}}^T(k)]^T \in \mathbb{R}^{r_{\bar{\zeta}}}$ with $r_{\bar{\zeta}} = n_z r_m + n_z r_p + n_z r_p$ and $\omega(k) = (\underline{A} - \bar{L}\bar{C})^k r(0)$.

The matrix \bar{M} in (23) is termed as a parameterization matrix as in [12]. The following result shows that the DT state reconstruction in (23) has a structural property that $\text{rank}(\bar{M})$ is linked to the three controllability matrices $\mathcal{C}(\underline{A}, \bar{B})$, $\mathcal{C}(\underline{A}, \bar{L})$, and $\mathcal{C}(\underline{A}, \bar{G})$. This is a necessary and sufficient condition for measuring the rank of the parameterization matrix for DT state reconstruction.

Lemma 4: $\text{rank}(\bar{M}) = \text{rank}([\mathcal{C}(\underline{A}, \bar{B}), \mathcal{C}(\underline{A}, \bar{L}), \mathcal{C}(\underline{A}, \bar{G})])$.

Proof: See Appendix A. \square

This lemma reveals that the parameterization matrix \bar{M} in the DT version of (23) has the same structural property as that in the CT version of [12]. Note that the state reconstruction in (23) is for a tracking problem, which is thus applicable to a regulation problem. Based on the property in Lemma 4, one has the following convergence result for the DT system (18).

Theorem 1: Consider the DT system (18) satisfying the controllability condition of (\underline{A}, \bar{B}) and the observability condition of (\underline{A}, \bar{C}) , $r(k) - \bar{M}\bar{\zeta}(k)$ asymptotically decays to zero.

Proof: Under the condition that (\underline{A}, \bar{B}) is controllable, it follows from Lemma 4 that $\text{rank}(\bar{M}) = n_z$. The matrix A_{ζ} in (20)–(22) is designed to be Schur by choosing appropriate coefficients d_j for $j = 1, 2, \dots, n_z$. Thus, the vector $\bar{\zeta}(k)$ in (23), formed by the difference equations (20)–(22), is known. In addition, since (\underline{A}, \bar{C}) is observable, then the eigenvalues of $\underline{A} - \bar{L}\bar{C}$ can be designed to be equal to those of A_{ζ} through choosing an appropriate matrix \bar{L} . By doing this, $\underline{A} - \bar{L}\bar{C}$ is Schur, based on which $r(k) - \bar{M}\bar{\zeta}(k)$ asymptotically decays to zero. This completes the proof. \square

The reason for seeking $\text{rank}(\bar{M}) = n_z$ is that the uniqueness of the approximate solution to the Bellman equation will depend on it (see Proof of Lemma 5 in the next subsection).

In the next two subsections, we will focus on how to use the input-output data to learn the optimal control gain matrix after taking the state reconstruction into account.

B. Off-Policy Bellman Equation in Output-Feedback Form

In this subsection, both \bar{M} and $\omega(k)$ are introduced to change the state-feedback Bellman equation (19) into an output-feedback form as

$$\begin{aligned} & \bar{\zeta}^T(k+1)\bar{P}^{j+1}\bar{\zeta}(k+1) - \bar{\zeta}^T(k)\bar{P}^{j+1}\bar{\zeta}(k) \\ &= -y^T Q y - \bar{\zeta}^T(k)\bar{M}^T(\bar{K}^j)^T \bar{R}\bar{K}^j \bar{M}\bar{\zeta}(k) \\ &+ (-\bar{K}^j \bar{M}\bar{\zeta}(k) + \bar{u}(k))^T \bar{B}^T P^{j+1} \bar{B} (\bar{K}^j \bar{M}\bar{\zeta}(k) + \bar{u}(k)) \\ &+ 2\vartheta^T(k) \bar{G}^T P^{j+1} \bar{B} u(k) + 2\bar{\zeta}^T(k) \bar{M}^T \underline{A}^T P^{j+1} \bar{G} \vartheta(k) \\ &+ 2\bar{\zeta}^T(k) \bar{M}^T \underline{A}^T P^{j+1} \bar{B} (\bar{K}^j \bar{M}\bar{\zeta}(k) + \bar{u}(k)) \\ &+ \vartheta^T(k) \bar{G}^T P^{j+1} \bar{G} \vartheta(k) + \bar{\chi}^{j+1}(t), \end{aligned} \quad (24)$$

where

$$\begin{aligned} & \bar{\chi}^{j+1}(k) \\ &= -2\omega^T(k+1)P^{j+1}\bar{M}\bar{\zeta}(k+1) - \omega^T(k+1)P^{j+1}\omega(k+1) \\ & - 2\omega^T(k)P^{j+1}\bar{M}\bar{\zeta}(k) - \omega^T(k)P^{j+1}\omega(k) \\ & - 2\omega^T(k)(\bar{K}^j)^T \bar{R}\bar{K}^j \bar{M}\bar{\zeta}(k) - \omega^T(k)(\bar{K}^j)^T \bar{R}\bar{K}^j \omega(k) \\ & - 2\omega^T(k)(\bar{K}^j)^T \bar{B}^T P^{j+1} \bar{B} \bar{K}^j \bar{M}\bar{\zeta}(k) \\ & - \omega^T(k)(\bar{K}^j)^T \bar{B}^T P^{j+1} \bar{B} \bar{K}^j \omega(k) \\ & + 2\omega^T(k)\underline{A}^T P^{j+1} \bar{G} \vartheta(k) + 2\omega^T(k)\underline{A}^T P^{j+1} \bar{B} \bar{K}^j \bar{M}\bar{\zeta}(k) \\ & + 2\omega^T(k)\underline{A}^T P^{j+1} \bar{B} \bar{K}^j \omega(k) + 2\omega^T(k)\underline{A}^T P^{j+1} \bar{B} u(k) \\ & + 2\bar{\zeta}^T(k) \bar{M}^T \underline{A}^T P^{j+1} \bar{B} \bar{K}^j \omega(k). \end{aligned} \quad (25)$$

Let $\mathcal{C}_{\bar{\zeta}} = [\text{vecv}(\bar{\zeta}(k_1)) - \text{vecv}(\bar{\zeta}(k_0)), \text{vecv}(\bar{\zeta}(k_2)) - \text{vecv}(\bar{\zeta}(k_1)), \dots, \text{vecv}(\bar{\zeta}(k_f)) - \text{vecv}(\bar{\zeta}(k_{f-1}))]^T$, $\mathcal{D}_{\bar{K}_o^j \bar{\zeta}} = [\text{vecv}(\bar{K}_o^j \bar{\zeta}(k_0)), \text{vecv}(\bar{K}_o^j \bar{\zeta}(k_1)), \dots, \text{vecv}(\bar{K}_o^j \bar{\zeta}(k_{f-1}))]^T$, $\mathcal{D}_{\vartheta \bar{\zeta}} = [\vartheta(k_0) \otimes \bar{\zeta}(k_0), \vartheta(k_1) \otimes \bar{\zeta}(k_1), \dots, \vartheta(k_{f-1}) \otimes \bar{\zeta}(k_{f-1})]^T$, $\mathcal{D}_{\bar{\zeta} \bar{\zeta}} = [\bar{\zeta}(k_0) \otimes \bar{\zeta}(k_0), \bar{\zeta}(k_1) \otimes \bar{\zeta}(k_1), \dots, \bar{\zeta}(k_{f-1}) \otimes \bar{\zeta}(k_{f-1})]^T$, $\mathcal{D}_{\bar{u} \vartheta} = [\bar{u}(k_0) \otimes \vartheta(k_0), \bar{u}(k_1) \otimes \vartheta(k_1), \dots, \bar{u}(k_{f-1}) \otimes \vartheta(k_{f-1})]^T$, and $\mathcal{D}_{\bar{\chi}^{j+1}} = [\bar{\chi}^{j+1}(t_0), \bar{\chi}^{j+1}(t_1), \dots, \bar{\chi}^{j+1}(t_s)]^T$. Besides, let $\bar{L}_P^{j+1} = \bar{M}^T P^{j+1} \bar{M}$, $\bar{L}_1^{j+1} = \bar{M}^T \underline{A}^T P^{j+1} \bar{B}$, $\bar{L}_2^{j+1} = \bar{B}^T P^{j+1} \bar{B}$, $\bar{L}_3^{j+1} = \bar{M}^T \underline{A}^T P^{j+1} \bar{G}$, $\bar{L}_4^{j+1} = \bar{G}^T P^{j+1} \bar{B}$, $\bar{L}_5^{j+1} = \bar{G}^T P^{j+1} \bar{G}$, and $\bar{K}_o^j = \bar{K}^j \bar{M}$. Now, (24) is rewritten as

$$\varrho_o^j \bar{L}_{vec} = \nu_o^j + \mathcal{D}_{\bar{\chi}^{j+1}}, \quad (26)$$

where

$$\begin{aligned} \bar{L}_{vec} &= [\text{vecs}^T(\bar{L}_P^{j+1}), \text{vec}^T(\bar{L}_1^{j+1}), \text{vecs}^T(\bar{L}_2^{j+1}), \\ & \text{vec}^T(\bar{L}_3^{j+1}), \text{vec}^T(\bar{L}_4^{j+1}), \text{vecs}^T(\bar{L}_5^{j+1})]^T, \end{aligned} \quad (27)$$

$$\begin{aligned} \varrho_o^j &= [\mathcal{C}_{\bar{\zeta}}, -2\mathcal{D}_{\bar{\zeta} \bar{\zeta}}(I \otimes (\bar{K}_o^j)^T) - 2\mathcal{D}_{\bar{u} \bar{\zeta}}, \\ & - \mathcal{D}_{\bar{u}} + \mathcal{D}_{\bar{K}_o^j \bar{\zeta}}, -2\mathcal{D}_{\vartheta \bar{\zeta}}, -2\mathcal{D}_{\bar{u} \vartheta}, -\mathcal{D}_{\vartheta}], \end{aligned} \quad (28)$$

$$\nu_o^j = -\mathcal{D}_{\bar{\zeta} \bar{\zeta}} \text{vec}((\bar{K}_o^j)^T \bar{R} \bar{K}_o^j) - \mathcal{D}_{yy} \text{vec}(Q). \quad (29)$$

From (23), if the initial state of $r(k)$ satisfies $r(0) = 0$, then $\omega(k)$ in (23) and $\mathcal{D}_{\bar{\chi}^{j+1}}$ in (26) are zeros. In the next subsection, we will take the non-zero initial state $r(0) \neq 0$ into account, and seek for approximating (26).

C. Solution to Output-Feedback Off-Policy Bellman Equation

In this subsection, we are to reduce the influence from non-zero initials and to give a sufficient condition for approximately solving the off-policy Bellman equation in the output-feedback form.

Here, $\mathcal{D}_{\bar{\chi}^{j+1}}$ is a nonlinear function of unknown terms $\omega(t)$, \bar{L}_1^{j+1} , and \bar{L}_2^{j+1} . It thus becomes difficult to obtain an accurate analytical solution from (26). Instead of directly solving (26), we turn to computing the following linear equation

$$\varrho_o^j \hat{\bar{L}}_{vec} = \nu_o^j, \quad (30)$$

where $\hat{\bar{L}}_{vec} = [\text{vecs}^T(\hat{\bar{L}}_P^{j+1}), \text{vec}^T(\hat{\bar{L}}_1^{j+1}), \text{vecs}^T(\hat{\bar{L}}_2^{j+1}), \text{vec}^T(\hat{\bar{L}}_3^{j+1}), \text{vec}^T(\hat{\bar{L}}_4^{j+1}), \text{vecs}^T(\hat{\bar{L}}_5^{j+1})]^T$, and the notation $(\hat{\cdot})$ is employed to differ the computed solution in (30) from the analytical one in (26). We have seen that (30) equals (26) if

$r(0) = 0$. In what follows, we focus on handling the non-zero case and give the following result on reducing the solution error between (26) and (30).

Theorem 2: Suppose that the non-zero matrix $\bar{\rho}^j$ in (28) is collected over the time interval $[k_0, k_f]$. If the starting time for the data collection k_0 is sufficiently large, then the computed solution from (30) can be considered as an approximate solution of (26) with the solution error being sufficiently small.

Proof: Consider two difference equations

$$v(s+1) = v(s) - \varepsilon(\varrho_o^j)^T (\varrho_o^j v(s) - \nu_o^j - \mathcal{D}_{\bar{\chi}^{j+1}}), \quad (31)$$

$$\hat{v}(s+1) = \hat{v}(s) - \varepsilon(\varrho_o^j)^T (\varrho_o^j \hat{v}(s) - \nu_o^j), \quad (32)$$

to solve (26) and (30) with $v(0) = \hat{v}(0) = 0$ and the constant ε satisfying

$$0 < \varepsilon < 2\rho^{-1}((\varrho_o^j)^T (\varrho_o^j)). \quad (33)$$

Bringing the algorithmic time s into (31) and (32) is to distinguish it from the system evolution time k used in (1)–(4).

It follows from the singular value decomposition that there exist matrices W and ϱ_w^j with $W^T W = W W^T = I$ and $(\varrho_w^j)^T \varrho_w^j > O$ such that [12], [35] $\varrho_o^j W = [\varrho_w^j \ O]$. Define

$$\bar{v}(s) = W^T v(s) = [\bar{v}_1^T(s), \bar{v}_2^T(s)]^T. \quad (34)$$

From (34), (31) is rewritten into

$$\begin{aligned} \bar{v}_1(s+1) &= \bar{v}_1(s) - \varepsilon(\varrho_w^j)^T \varrho_w^j \bar{v}_1(s) \\ &\quad + \varepsilon(\varrho_w^j)^T \nu_o^j + \varepsilon(\varrho_w^j)^T \mathcal{D}_{\bar{\chi}^{j+1}}, \end{aligned} \quad (35)$$

$$\bar{v}_2(s+1) = \bar{v}_2(s), \quad (36)$$

where $\bar{v}_2(s)$ is zero for any s . Hence, (35) is changed into

$$\begin{aligned} &\|\bar{v}_1(s+1)\| \\ &\leq \sqrt{|\lambda_{\max}(I - \varepsilon(\varrho_w^j)^T \varrho_w^j)|} \|\bar{v}_1(s)\| + \|\varepsilon(\varrho_w^j)^T \nu_o^j\| \\ &\quad + \|\varepsilon(\varrho_w^j)^T \mathcal{D}_{\bar{\chi}^{j+1}}\|. \end{aligned} \quad (37)$$

From (33), one has that $0 < \rho(\varepsilon(\varrho_w^j)^T \varrho_w^j) < 2$ such that $\rho(I - \varepsilon(\varrho_w^j)^T \varrho_w^j) < 1$. Therefore, given the constant ε and the matrix $(\varrho_w^j)^T \varrho_w^j$, there exists a constant ϵ so that $|\lambda_{\max}(I - \varepsilon(\varrho_w^j)^T \varrho_w^j)| < \epsilon < 1$. As a result, one changes (37) into

$$\begin{aligned} \|\bar{v}_1(s+1)\| &\leq \epsilon_1 \|\bar{v}_1(s)\| + \|\varepsilon(\varrho_w^j)^T \nu_o^j\| \\ &\quad + \|\varepsilon(\varrho_w^j)^T \mathcal{D}_{\bar{\chi}^{j+1}}\|, \end{aligned} \quad (38)$$

where $\epsilon_1 = \sqrt{|\lambda_{\max}(I - \varepsilon(\varrho_w^j)^T \varrho_w^j)|} < \sqrt{\epsilon} < 1$. It is noted that the term $\|\varepsilon(\varrho_w^j)^T \nu_o^j\|$ in (38) is bounded.

Now, we focus on the boundedness of $\mathcal{D}_{\bar{\chi}^{j+1}}$ in (38). Let α_i for $i = 0, 1, 2, 3, 4$ and ϵ_l for $l = 1, 2$ be certain positive constants. Since $\underline{A} - \bar{L}\bar{C}$ is Schur, $0 < |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})| < 1$ holds. From (25), one has

$$\|\mathcal{D}_{\bar{\chi}^{j+1}}\| \leq \|\bar{v}(s)\| \alpha_2 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}. \quad (39)$$

Substituting (39) into (38) yields

$$\|\bar{v}_1(s+1)\| \leq (\epsilon_1 + \alpha_3 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}) \|\bar{v}_1(s)\| + \bar{b}^j,$$

where \bar{b}^j is defined as an upper bound of $\|\varepsilon(\varrho_w^j)^T \mathcal{D}_{\bar{\chi}^{j+1}}\|$. Thus, with sufficiently large k_0 , $\alpha_3 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}$ is thus sufficiently small. There exists a positive constant ϵ_2 satisfying $0 < \epsilon_1 < \epsilon_2 < 1$ and $\epsilon_2 - \epsilon_1 > \alpha_3 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}$ such that

$$\|\bar{v}_1(s+1)\| \leq \epsilon_2 \|\bar{v}_1(s)\| + \bar{b}^j, \quad (40)$$

which leads to

$$\|\bar{v}_1(s)\| \leq \left(\|\bar{v}_1(0)\| - \frac{\bar{b}^j}{1 - \epsilon_2} \right) \epsilon_2^s + \frac{\bar{b}^j}{1 - \epsilon_2}. \quad (41)$$

Thus, $\bar{v}(s)$, $v(s)$, and $\mathcal{D}_{\bar{\chi}^{j+1}}$ are bounded for any s .

With the boundedness of $v(s)$, we now prove the convergence of the solution error between (26) and (30) in the remaining analysis. Similar to (34), we define

$$\hat{\hat{v}}(s) = W^T \hat{v}(s) = [\hat{\hat{v}}_1^T(s), \hat{\hat{v}}_2^T(s)]^T, \quad (42)$$

based on which one has

$$\hat{\hat{v}}_1(s+1) = \hat{\hat{v}}_1(s) - (\varrho_w^j)^T \varrho_w^j \hat{\hat{v}}_1(s) + (\varrho_w^j)^T \nu_o^j, \quad (43)$$

$$\hat{\hat{v}}_2(s+1) = \hat{\hat{v}}_2(s), \quad (44)$$

where $\hat{\hat{v}}(s)$ is zero. Thus, from (44), $\hat{\hat{v}}_2(s) = 0$ holds. Let the error be $\tilde{v}_1(s) = \bar{v}_1(s) - \hat{\hat{v}}_1(s)$. From (35) and (43), one has

$$\|\tilde{v}_1(s+1)\| \leq \epsilon_1 \|\tilde{v}_1(s)\| + \alpha_4 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}, \quad (45)$$

where the boundedness of $\mathcal{D}_{\bar{\chi}^{j+1}}$ is employed. By (45), one has

$$\begin{aligned} \|\tilde{v}_1(s)\| &\leq \left(\|\tilde{v}_{e_1}(0)\| - \frac{\alpha_4 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}}{1 - \epsilon_1} \right) |\epsilon_1|^s \\ &\quad + \frac{\alpha_4 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}}{1 - \epsilon_1}, \end{aligned} \quad (46)$$

where the constant ϵ_1 has been defined in (38) satisfying $0 < \epsilon_1 < 1$ and the term $\alpha_4 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}$ has been tuned sufficiently small under sufficiently large k_0 . From (34) and (42), one changes (46) into

$$\begin{aligned} \lim_{s \rightarrow \infty} \|\hat{v}(s) - v(s)\| &\leq \lim_{s \rightarrow \infty} \|W\| \|\hat{\hat{v}}(s) - \bar{v}(s)\| \\ &= \lim_{s \rightarrow \infty} \|\hat{\hat{v}}(s) - \bar{v}(s)\| \\ &= \frac{\alpha_4 |\lambda_{\max}(\underline{A} - \bar{L}\bar{C})|^{k_0}}{1 - \epsilon_1}, \end{aligned} \quad (47)$$

which implies that the solution of (26) converges to that of (30) with the error being sufficiently small by increasing k_0 . Therefore, the proof is completed. \square

Remark 1: *Theorem 2* shows that the solution error between (26) and (30) can be made smaller by choosing a larger starting time k_0 for the data collection. This prompts us to introduce an additional *Model-Free Pre-Collection Phase*. The necessity of the additional phase roots in the requirement of the convergence in the DT state reconstruction and the idea is inspired by the CT work of [12]. Details on how to coordinately execute the DT off-policy learning will be presented in *Algorithm 1*. \bullet

Remark 2: In contrast to [12] wherein differential equations are used to solve the linear equation, *Theorem 2* uses difference equations. In addition, we find that our result in *Theorem 2* and that in [36] are dual to each other in a certain sense.

To be specific, *Theorem 2* shows that a class of unknown nonlinear equations are approximately solved by a known linear difference equation, while [36] shows that an unknown linear difference equation is approximately solved by a class of nonlinear equations. •

The following result shows that how much data should we collect to achieve the optimal output tracking control within the output-feedback RL framework.

Lemma 5: The off-policy Bellman equation in the output-feedback form (30) over the time interval $[k_0, k_f]$, \hat{L}_i^{j+1} , for $i = 1, 2, \dots, 5$, can be uniquely solved, if

1) the collected input-output data at the time k_f satisfy

$$\begin{aligned} & \text{rank}([\mathcal{D}_{\bar{\zeta}\bar{\zeta}}, \mathcal{D}_{\bar{u}\bar{\zeta}}, \mathcal{D}_{\bar{u}}, \mathcal{D}_{\vartheta\bar{\zeta}}, \mathcal{D}_{\bar{u}\vartheta}, \mathcal{D}_{\vartheta}]) \\ &= \frac{1}{2}(n_{\bar{\zeta}}(n_{\bar{\zeta}} + 1) + r_m(r_m + 1) + r_p(r_p + 1)) \\ & \quad + n_{\bar{\zeta}}r_m + r_p n_{\bar{\zeta}} + r_p r_m; \end{aligned} \quad (48)$$

2) $\text{rank}(\bar{M}) = n_z$;

3) $r(0) = 0$.

Proof: See Appendix B. ◻

In *Lemma 5*, the condition $r(0) = 0$ is required. We now extend it to the condition $r(0) \neq 0$ by choosing the starting time k_0 to be sufficiently large. This is because the solution \hat{L}_i^{j+1} , for $i = 1, 2, \dots, 5$ in (30) under the condition $r(0) \neq 0$ converges to the unique solution in *Lemma 5* under the condition $r(0) = 0$ after recalling the result in *Theorem 2*.

D. Optimal Output Tracking Design via Output-Feedback RL

Our output-feedback off-policy learning algorithm is summarized in *Algorithm 1*, where the successive error of the control matrix $\|\hat{K}_o^{j+1} - \hat{K}_o^j\|$ is used for the stopping indicator since \hat{K}_o^{j+1} is solved uniquely under conditions in *Lemma 5*.

Based on the learned optimal control gain matrix \hat{K}_o^* from *Algorithm 1*, we achieve optimal output tracking control of DT systems via output feedback as below.

Theorem 3: Let the DT system (1)–(4) satisfy *Assumptions 1–4* and the output-feedback adaptive optimal output tracking DT controller be designed as

$$u(k) = -\hat{K}_o^* \bar{\zeta}(k) - T \bar{z}(k), \quad (50)$$

where $\bar{\zeta}$ is given in (23), $\bar{z}(k)$ is given in (17) with $\vartheta(k)$ being assigned as $y_d(k)$, and \hat{K}_o^* is the learnt optimal control gain matrix by *Algorithm 1*. Therefore, the state-oriented tracking optimization problem, *Problem 1*, is solved with the output tracking error $y_e(k)$ in (5) decaying to zero, asymptotically.

Proof: We first show that the learnt matrix \hat{K}_o^{j+1} in (49) converges to the ideal matrix $\bar{K}^* \bar{M}$ with \bar{K}^* and \bar{M} being defined in (13) and (23), respectively. With the condition in *Theorem 2* satisfied, \hat{L}_1^{j+1} and \hat{L}_2^{j+1} in (30) can be made to converge to \bar{L}_1^{j+1} and \bar{L}_2^{j+1} in (26), respectively. This, together with the uniqueness in *Lemma 5*, leads that the learned matrix $(\bar{R} + \hat{L}_2^{j+1})^{-1} (\hat{L}_1^{j+1})^T$ in (49) uniquely converges to $\bar{K}^{j+1} \bar{M}$, where \bar{K}^{j+1} is defined in (16). It follows from *Lemma 2* that \bar{K}^{j+1} converges to \bar{K}^* as the integer j gets larger. Note that \bar{K}^* is the unique solution satisfying the *Problem 1* under the controllability of (\bar{A}, \bar{B}) and the observability of (\bar{A}, \bar{C}) .

Algorithm 1 Output-Feedback Off-Policy RL for Optimal Output Tracking Control of DT Systems

- 1: **Initialize:** Let $j = 0$ and \bar{K}_o^j be a stabilizing gain. Let the pair (F, G) be an r_p -copy internal model of S .
 - 2: **Model-Free Pre-Collection Phase:** From (20)–(22), compute ζ_u , ζ_y , and ζ_ϑ over the time interval $[0, k_0)$, where k_0 is set sufficiently large.
 - 3: **Data-Collection Phase:** Apply the behavior policy $\bar{u}(k)$ in (17) with (F, T) being observable to the system (1) with $r_p \geq r_m$ over the time interval $[k_0, k_f]$ for collecting the input-output data $\mathcal{D}_{\bar{\zeta}\bar{\zeta}}$, $\mathcal{D}_{\bar{u}\bar{\zeta}}$, $\mathcal{D}_{\bar{u}}$, $\mathcal{D}_{\vartheta\bar{\zeta}}$, $\mathcal{D}_{\bar{u}\vartheta}$, and \mathcal{D}_{ϑ} .
 - 4: **if** (48) holds **then**
 - 5: **while** the stopping indicator $\|\hat{K}_o^{j+1} - \hat{K}_o^j\| \leq \varepsilon$ is not satisfied with $\varepsilon > 0$ being a small constant **do**
 - 6: Solve (30) and update the feedback gain as

$$\hat{K}_o^{j+1} = (\bar{R} + \hat{L}_2^{j+1})^{-1} (\hat{L}_1^{j+1})^T. \quad (49)$$
 - 7: **end while**
 - 8: **end if**
 - 9: **Optimal Control Phase:** The learned optimal control gain matrix \hat{K}_o^* is given as $\hat{K}_o^* = \hat{K}_o^{j+1}$.
-

Therefore, if \hat{K}_o^{j+1} converges, then the unique solution \hat{K}^{j+1} from (30) converges to $\bar{K}^* \bar{M}$. It reveals that the ideal matrix $\bar{K}^* \bar{M}$ is learned by the matrix \hat{K}_o^* from *Algorithm 1*. We then show the convergence of the output tracking error $y_e(k)$. With the learned optimal control gain matrix \hat{K}_o^* , the closed-loop system becomes $e(k+1) = \bar{A}_o^* e(k) + \bar{B}(\bar{A} - \bar{L}\bar{C})^k r(0)$, where $e(k)$ is given in (9) and both $\bar{A}_o^* = \bar{A} - \bar{B}\hat{K}_o^*$ and $\bar{A} - \bar{L}\bar{C}$ are Schur. Considering that $\lim_{k \rightarrow \infty} \bar{B}(\bar{A} - \bar{L}\bar{C})^k r(0) = 0$, one obtains that $\lim_{k \rightarrow \infty} e(k) = 0$ [36], based on which the output tracking error satisfies $\lim_{k \rightarrow \infty} y_e(k) = 0$ from (11b). This completes the proof. ◻

Remark 3: The RL-based controller in (50) is robust to system uncertainties, which corresponds to the linear robust output regulation in the literature such as [32]. That is, after the learning is completed, the proposed controllers are robust to some model uncertainties in the system dynamics matrices $A + \Delta A$, $B + \Delta B$, and $C + \Delta C$, where A , B , and C denote the nominal part of the plant; ΔA , ΔB , and ΔC represent the model uncertainties. •

IV. CONCLUSION

This paper investigated the output-feedback optimal output tracking problem for DT systems with unknown dynamics using the off-policy RL and robust output regulation theory. We have formulated the output tracking optimization problem based on the newly proposed dynamical DT controller in contrast to the standard DT controller by linear output regulation theory. We have shown that, by making use of the collected data along with the controlled system and the reference output, we are able to approximate the optimal output-feedback controller. We have studied the parameterization matrix and re-expression error so that the learned optimal controller has a satisfactory performance.

APPENDIX A
PROOF OF Lemma 4

Let us consider a standard Luenberger observer as

$$\begin{aligned}\hat{r}(k+1) &= \underline{A}\hat{r}(k) + \bar{B}\bar{u}(k) + \bar{G}\vartheta(k) \\ &\quad + \bar{L}(y(k) - \bar{C}\hat{r}(k))\end{aligned}\quad (51)$$

with \bar{L} being a $n_z \times r_p$ matrix so that $\hat{r}(k) - r(k)$ decays to zero with $r(k)$ given by (18). Similar to [24], (51) is rewritten as

$$\begin{aligned}\hat{r}(k) &= G_u(z)[\bar{u}(k)] + G_y(z)[y(k)] \\ &\quad + G_\vartheta(z)[\vartheta(k)] + (\mathcal{A}_o)^k \hat{r}(0),\end{aligned}\quad (52)$$

where $\mathcal{A}_o = \underline{A} - \bar{L}\bar{C}$, $G_{\bar{u}}(z)[\bar{u}(k)]$ is a time-domain DT signal represented by the frequency-domain representation $G_{\bar{u}}(z) = \begin{bmatrix} G_{1,1}^{\bar{u}}(z) & \cdots & G_{1,r_m}^{\bar{u}}(z) \\ \vdots & \ddots & \vdots \\ G_{n_z,1}^{\bar{u}}(z) & \cdots & G_{n_z,r_m}^{\bar{u}}(z) \end{bmatrix} \in \mathbb{R}^{n_z \times r_m}$. The entry $G_{i,j}^{\bar{u}}(z)$ at the i th row and j th column of $G_{\bar{u}}(z)$ is extended to

$$\begin{aligned}G_{i,j}^{\bar{u}}(z) &= \frac{1}{\det(zI - \mathcal{A}_o)} \left[g_{i,j,n_z-1}^{\bar{u}} z^{n_z-1} + g_{i,j,n_z-2}^{\bar{u}} z^{n_z-2} \right. \\ &\quad \left. + \cdots + g_{i,j,1}^{\bar{u}} z + g_{i,j,0}^{\bar{u}} \right].\end{aligned}\quad (53)$$

Here, $G_{\bar{u}}(z)[\bar{u}(k)]$ is interpreted as

$$\begin{aligned}G_{\bar{u}}(z)[\bar{u}(k)] &= \underbrace{\begin{bmatrix} g_{1,1,0}^{\bar{u}} & \cdots & g_{1,r_m,n_z-1}^{\bar{u}} & g_{1,r_m+1,0}^{\bar{u}} & \cdots & g_{1,n_z r_m,n_z-1}^{\bar{u}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{n_z,1,0}^{\bar{u}} & \cdots & g_{n_z,r_m,n_z-1}^{\bar{u}} & g_{n_z,r_m+1,0}^{\bar{u}} & \cdots & g_{n_z,n_z r_m,n_z-1}^{\bar{u}} \end{bmatrix}}_{M_{\bar{u}}} \\ &\quad \times \underbrace{[u(k)] \otimes [1, z, \dots, z^{n_z-1}]^T}_{\zeta_{\bar{u}}(k)} \frac{1}{\det(zI - \mathcal{A}_o)} \\ &\triangleq M_{\bar{u}} \zeta_{\bar{u}}(k).\end{aligned}\quad (54)$$

Similar definitions of $M_y \zeta_y(k)$ and $M_\vartheta \zeta_\vartheta(k)$ apply to $G_y(z)[y(k)]$ and $G_\vartheta(z)[\vartheta(k)]$, respectively. Thus, one obtains that

$$\hat{r}(k) = \bar{M} \bar{\zeta}(k) + (\mathcal{A}_o)^k \hat{r}(0).\quad (55)$$

From (51) and the z-transform operator, one has

$$\begin{aligned}\hat{r}(k) &= (\underline{A} - \bar{L}\bar{C})^k \hat{r}(0) + (zI - \mathcal{A}_o)^{-1} \bar{B}[\bar{u}(k)] \\ &\quad + (zI - \mathcal{A}_o)^{-1} \bar{L}[y(k)] + (zI - \mathcal{A}_o)^{-1} \bar{G}[\vartheta(k)],\end{aligned}$$

where

$$(zI - \mathcal{A}_o)^{-1} = \frac{\text{adj}(zI - \mathcal{A}_o)}{\det(zI - \mathcal{A}_o)},\quad (56)$$

$$\begin{aligned}\det(zI - \mathcal{A}_o) &= z^{n_z} + d_1 z^{n_z-1} + d_2 z^{n_z-2} + \cdots \\ &\quad + d_{n_z-1} z + d_{n_z},\end{aligned}\quad (57)$$

$$\begin{aligned}\text{adj}(zI - \mathcal{A}_o) &= B_0 z^{n_z-1} + B_1 z^{n_z-2} + \cdots \\ &\quad + B_{n_z-2} z + B_{n_z-1}.\end{aligned}\quad (58)$$

From (56) and (58), one has

$$\begin{aligned}\det(zI - \mathcal{A}_o) &= B_0 z^{n_z} + (B_1 - B_0 \mathcal{A}_o) z^{n_z-1} + (B_2 - B_1 \mathcal{A}_o) z^{n_z-2} \\ &\quad + \cdots + (B_{n_z-1} - B_{n_z-2} \mathcal{A}_o) z - B_{n_z-1} A.\end{aligned}\quad (59)$$

Using (57) and (59), the following equations $B_0 = I$ and $B_{i+1} = B_i \mathcal{A}_o + d_{i+1} I$ hold for $i = 0, 1, 2, \dots, n_z - 2$. For $i = 0$, one has $B_1 = B_0 \mathcal{A}_o + d_1 I = \mathcal{A}_o + d_1 I = \underline{A} - \bar{L}\bar{C} + d_1 I$. This leads to

$$\begin{aligned}\text{rank}([\![B_0, B_1]\bar{B}, [B_0, B_1]\bar{L}, [B_0, B_1]\bar{G}]\!]) \\ &= \text{rank}([\![B_0, \underline{A} - \bar{L}\bar{C} + d_1 I]\bar{B}, [B_0, \underline{A} - \bar{L}\bar{C} + d_1 I]\bar{L}, \\ &\quad [B_0, \underline{A} - \bar{L}\bar{C} + d_1 I]\bar{G}]\!]) \\ &= \text{rank}([\![\bar{B}, (\underline{A} - \bar{L}\bar{C})\bar{B} + d_1 \bar{B}], [\bar{L}, (\underline{A} - \bar{L}\bar{C})\bar{L} + d_1 \bar{L}], \\ &\quad [\bar{G}, (\underline{A} - \bar{L}\bar{C})\bar{G} + d_1 \bar{G}]]\!]) \\ &= \text{rank}([\![\bar{B}, \underline{A}\bar{B}, \bar{L}, \underline{A}\bar{L}, \bar{G}, \underline{A}\bar{G}]\!]),\end{aligned}\quad (60)$$

where the last equation is obtained using the fact that column operations do not change the rank of a matrix. Based on analysis in (60) with $i = 0$, proceeding the order i to be higher one by one, one has

$$\begin{aligned}\text{rank}([\![\mathcal{B}(\bar{B}), \mathcal{B}(\bar{L}), \mathcal{B}(\bar{G})]\!]) \\ &= \text{rank}([\![\mathcal{C}(\underline{A}, \bar{B}), \mathcal{C}(\underline{A}, \bar{L}), \mathcal{C}(\underline{A}, \bar{G})]\!])\end{aligned}\quad (61)$$

with $\mathcal{B}(\bar{B}) = [B_0, B_1, \dots, B_{n_z-1}]\bar{B}$, $\mathcal{B}(\bar{L}) = [B_0, B_1, \dots, B_{n_z-1}]\bar{L}$, and $\mathcal{B}(\bar{G}) = [B_0, B_1, \dots, B_{n_z-1}]\bar{G}$.

Based on (61), we next clarify that $\text{rank}(\bar{M}) = \text{rank}([\![\mathcal{B}(\bar{B}), \mathcal{B}(\bar{L}), \mathcal{B}(\bar{G})]\!])$ with \bar{M} being given in (55). Based on the definitions in (54) and (58), $\text{rank}(M_{\bar{u}}) = \text{rank}(\mathcal{B}(\bar{B}))$. It leads to $\text{rank}(M_y) = \text{rank}(\mathcal{B}(\bar{L}))$ and $\text{rank}(M_\vartheta) = \text{rank}(\mathcal{B}(\bar{G}))$. Hence, one has $\text{rank}(\bar{M}) = \text{rank}([\![\mathcal{B}(\bar{B}), \mathcal{B}(\bar{L}), \mathcal{B}(\bar{G})]\!])$. This, together with (61) leads to $\text{rank}(\bar{M}) = \text{rank}([\![\mathcal{C}(\underline{A}, \bar{B}), \mathcal{C}(\underline{A}, \bar{L}), \mathcal{C}(\underline{A}, \bar{G})]\!])$. Therefore, the proof is completed. \square

APPENDIX B
PROOF OF Lemma 5

In order to show the uniqueness of \hat{L}_i^{j+1} , for $i = 1, 2, \dots, 5$, in (30), it is equivalent to proving that

$$0 = \varrho_o^j \bar{\Xi}^v, \quad (62)$$

with $\bar{\Xi}^v = [\bar{W}^v, \bar{Y}_1^v, \bar{Y}_2^v, \bar{Y}_3^v, \bar{Y}_4^v, \bar{Y}_5^v]^T$ has a unique zero solution $[\bar{Y}_1^v, \bar{Y}_2^v, \bar{Y}_3^v, \bar{Y}_4^v, \bar{Y}_5^v]^T = 0$, where $\bar{W}^v = \text{vecs}(\bar{W}^m)$, $\bar{Y}_1^v = \text{vec}(\bar{Y}_1^m)$, $\bar{Y}_2^v = \text{vecs}(\bar{Y}_2^m)$, $\bar{Y}_3^v = \text{vec}(\bar{Y}_3^m)$, $\bar{Y}_4^v = \text{vec}(\bar{Y}_4^m)$, and $\bar{Y}_5^v = \text{vecs}(\bar{Y}_5^m)$ with $\bar{W}^m = (\bar{W}^m)^T$, $\bar{Y}_2^m = (\bar{Y}_2^m)^T$, and $\bar{Y}_5^m = (\bar{Y}_5^m)^T$. Define

$$\bar{Z}^m = (\bar{M} \bar{M}^T)^{-1} \bar{M} \bar{W}^m \bar{M}^T (\bar{M} \bar{M}^T)^{-1}, \quad (63)$$

where the property of $\text{rank}(\bar{M}) = n_z$ is employed. Under the condition (3) of Lemma 5, the approximation errors $\bar{\chi}^{j+1}(t)$ in (25) are zeros. Thus, it follows from (24) and (63) that (62) leads to

$$\begin{aligned}0 &= \mathcal{D}_{\bar{\zeta}} \text{vecs}(\bar{\kappa}_P) + 2\mathcal{D}_{\bar{u}\bar{\zeta}} \text{vec}(\bar{\kappa}_1) + \mathcal{D}_{\bar{u}} \text{vecs}(\bar{\kappa}_2) \\ &\quad + 2\mathcal{D}_{\vartheta\bar{\zeta}} \text{vec}(\bar{\kappa}_3) + 2\mathcal{D}_{\bar{u}\vartheta} \text{vec}(\bar{\kappa}_4) + \mathcal{D}_{\vartheta} \text{vecs}(\bar{\kappa}_5),\end{aligned}\quad (64)$$

where $\bar{\kappa}_P = \bar{M}^T[(\underline{A}^j)^T \bar{Z}^m \underline{A}^j - \bar{Z}^m] \bar{M} + (\bar{K}_o^j)^T (\bar{B}^T \bar{Z}^m \bar{B} - \bar{Y}_2^m) \bar{K}_o^j + (\underline{A}^T \bar{Z}^m \bar{B} - \bar{Y}_1^m) \bar{K}_o^j + (\bar{K}_o^j)^T (\underline{A}^T \bar{Z}^m \bar{B} - \bar{Y}_1^m)^T$, $\bar{\kappa}_1 = \underline{A}^T \bar{Z}^m \bar{B} - \bar{Y}_1^m$, $\bar{\kappa}_2 = \bar{B}^T \bar{Z}^m \bar{B} - \bar{Y}_2^m$, $\bar{\kappa}_3 = \underline{A}^T \bar{Z}^m \bar{G} - \bar{Y}_3^m$, $\bar{\kappa}_4 = \bar{G}^T \bar{Z}^m \bar{B} - \bar{Y}_4^m$, and $\bar{\kappa}_5 = \bar{G}^T \bar{Z}^m \bar{G} - \bar{Y}_5^m$.

The matrix $[\mathcal{D}_{\bar{\zeta}}, 2\mathcal{D}_{\bar{u}\bar{\zeta}}, \mathcal{D}_{\bar{u}}, 2\mathcal{D}_{\bar{\vartheta}\bar{\zeta}}, 2\mathcal{D}_{\bar{u}\bar{\vartheta}}, \mathcal{D}_{\bar{\vartheta}}]$ is full column rank if (48) holds. Thus, the solution to (64) is uniquely obtained as

$$\begin{aligned} &(\text{vecs}^T(\bar{\kappa}_P), \text{vecs}^T(\bar{\kappa}_1), \text{vecs}^T(\bar{\kappa}_2), \\ &\text{vecs}^T(\bar{\kappa}_3), \text{vecs}^T(\bar{\kappa}_4), \text{vecs}^T(\bar{\kappa}_5))^T = 0. \end{aligned} \quad (65)$$

Recalling $\text{rank}(\bar{M}) = n_z$, one further rewrites $\bar{\kappa}_P$ in (64) as $(\underline{A}^j)^T \bar{Z}^m \underline{A}^j - \bar{Z}^m = 0$, where \underline{A}^j is Schur. Therefore, \bar{Z}^m must be zeros, based on which \bar{Y}_i^v in (62), for $i = 1, 2, \dots, 5$, are also zeros. This implies that $\hat{\bar{L}}_i^{j+1}$ in (30), for $i = 1, 2, \dots, 5$, are unique. Note that since the non-square matrix \bar{M} in (63) is only full row rank, the zero solution of \bar{Z}^m does not ensure the zero solution of \bar{W}^m so that $\hat{\bar{L}}_P^{j+1}$ may not be unique. This completes the proof. \square

ACKNOWLEDGEMENT

The authors are thankful to Prof. Zongli Lin for his helpful comments and suggestions.

REFERENCES

- [1] B. L. Stevens, F. L. Lewis, and E. N. Johnson, *Aircraft control and simulation: dynamics, controls design, and autonomous systems*. John Wiley & Sons, 2015.
- [2] B. A. Francis, "The linear multivariable regulator problem," *SIAM Journal on Control and Optimization*, vol. 15, no. 3, pp. 486–505, 1977.
- [3] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT Press Cambridge, 1998.
- [4] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.
- [5] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive dynamic programming for control: algorithms and stability*. Springer Science & Business Media, 2012.
- [6] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. Institution of Engineering and Technology, 2013.
- [7] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2401–2410, 2016.
- [8] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. John Wiley & Sons, 2017.
- [9] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, 2016.
- [10] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. Dixon, *Reinforcement Learning for Optimal Feedback Control: A Lyapunov-Based Approach*. Springer, 2018.
- [11] C. Chen, H. Modares, K. Xie, F. L. Lewis, Y. Wan, and S. Xie, "Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4423–4438, 2019.
- [12] C. Chen, L. Xie, F. L. Lewis, and S. Xie, "Adaptive optimal output tracking of continuous-time systems via output-feedback-based reinforcement learning," *under review*, 2019.
- [13] C. Chen, F. L. Lewis, K. Xie, S. Xie, and Y. Liu, "Off-policy learning for adaptive optimal output synchronization of heterogeneous multi-agent systems," *Automatica*, vol. 119, p. 109081, 2020.
- [14] Z.-P. Jiang, T. Bian, W. Gao *et al.*, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, pp. 176–284, 2020.
- [15] B. Kiumarsi and F. L. Lewis, "Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 1, pp. 140–151, 2015.

- [16] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q -learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, 2014.
- [17] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, " H_∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.
- [18] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3147–3156, 2020.
- [19] W. Gao, Y. Liu, A. Odekinle, Y. Yu, and P. Lu, "Adaptive dynamic programming and cooperative output regulation of discrete-time multi-agent systems," *International Journal of Control, Automation and Systems*, vol. 16, no. 5, pp. 2273–2281, 2018.
- [20] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Trans. Syst. Man Cybern., Part B, Cybern.*, vol. 41, no. 1, pp. 14–25, Feb 2011.
- [21] J. Fan, Z. Li, Y. Jiang, T. Chai, and F. L. Lewis, "Model-free linear discrete-time system H_∞ control using input-output data," in *2018 International Conference on Advanced Mechatronic Systems*, Aug 2018, pp. 207–212.
- [22] S. A. A. Rizvi and Z. Lin, "Output feedback Q -learning for discrete-time linear zero-sum games with application to the H-infinity control," *Automatica*, vol. 95, pp. 213–221, 2018.
- [23] B. Kiumarsi, F. L. Lewis, M. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2770–2779, Dec 2015.
- [24] S. A. A. Rizvi and Z. Lin, "Output feedback Q -learning control for the discrete-time linear quadratic regulator problem," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 5, pp. 1523–1536, 2018.
- [25] —, "Output feedback optimal tracking control using reinforcement Q -learning," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 3423–3428.
- [26] W. Gao and Z.-P. Jiang, "Adaptive optimal output regulation via output-feedback: An adaptive dynamic programming approach," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 5845–5850.
- [27] —, "Adaptive optimal output regulation of time-delay systems via measurement feedback," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 3, pp. 938–945, 2018.
- [28] W. Gao, Y. Jiang, Z.-P. Jiang, and T. Chai, "Adaptive and optimal output feedback control of linear systems: An adaptive dynamic programming approach," in *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, 2014, pp. 2085–2090.
- [29] S. A. A. Rizvi and Z. Lin, "Output feedback adaptive dynamic programming for linear differential zero-sum games," *Automatica*, vol. 122, p. 109272, 2020.
- [30] —, "A note on state parameterizations in output feedback reinforcement learning control of linear systems," *under review*.
- [31] —, "Output feedback reinforcement learning control for linear systems," *Birkhauser, to appear*.
- [32] J. Huang, *Nonlinear output regulation: theory and applications*. SIAM, 2004.
- [33] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Trans. Autom. Control*, vol. 16, no. 4, pp. 382–384, 1971.
- [34] G. Tao, *Adaptive control design and analysis*. John Wiley & Sons, 2003.
- [35] T. Liu and J. Huang, "Adaptive cooperative output regulation of discrete-time linear multi-agent systems by a distributed feedback control law," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4383–4390, 2018.
- [36] J. Huang, "The cooperative output regulation problem of discrete-time linear multi-agent systems by the adaptive distributed observer," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1979–1984, 2016.