Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 10283–10288

Learning-Based Adaptive Optimal Output Regulation of Discrete-Time Linear Systems ⋆

Sayan Chakraborty * **Weinan Gao** ** **Leilei Cui** * **Frank L. Lewis** *** **Zhong-Ping Jiang** *

* *CAN Lab, New York University, Brooklyn, NY 11201 USA, (e-mail: sc8804, l.cui, zjiang @nyu.edu)*

** *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China (e-mail: gaown@mail.neu.edu.cn)*

*** *The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: lewis@uta.edu)*

Abstract- In this paper, we address the problem of model-free optimal output regulation of discrete-time systems that aims at achieving asymptotic tracking and disturbance rejection without the knowledge of the system parameters. Insights from reinforcement learning and adaptive dynamic programming are used to solve this problem. An interesting discovery is that the model-free discrete-time output regulation differs from the continuous-time counterpart in terms of the persistent excitation condition required to ensure the uniqueness and convergence of the policy iteration. In this work, we carefully establish the persistent excitation condition to ensure the uniqueness and convergence properties of the policy iteration.

Keywords: Adaptive control, approximate/adaptive dynamic programming, optimal control, discrete-time systems, discrete-time output regulation

1. INTRODUCTION

The output regulation problem is one of the most important topics in control theory. The output regulation problem aims at designing a feedback control law in order to achieve asymptotic tracking with disturbance rejection. The general mathematical formulation of this problem is applicable to many control problems arising from various disciplines like engineering, biology, etc; see Bonivento et al. (2001), Huang (2004), Trentelman et al. (2002) for instance. When the system dynamics is known, the problem of output regulation has been studied by many authors; see Krener (1992), Saberi et al. (2003), Liu and Huang (2020), Huang (2004), Yan and Huang (2016), Mantri et al. (1997). The above-mentioned studies however suffer from a common drawback of requiring the perfect knowledge of the system model. Model-free optimal control techniques are developed in the literature using the ideas from reinforcement learning (RL) (Sutton and Barto (2018)), and adaptive dynamic programming (ADP) (Bertsekas (2012)). Vrabie et al. (2009) proposed a novel policy iteration (PI) based optimal control technique that requires only the partial knowledge of system dynamics. Later, Jiang and Jiang (2012) proposed the first off-policy PI algorithm for the optimal control of linear systems with completely unknown system dynamics. More recently, this PI algorithm have been applied to linear parameter varying systems by Chakraborty et al. (2022), time-delay systems by Cui et al. (2022), and risk-sensitive optimal control by Cui and Jiang (2022).

The development of model-free techniques for output regulation has gained interest in the last decade. Gao and Jiang (2016) addressed the first model-free linear optimal output regulation problem (LOORP) for linear continuous-time (CT) systems. Recently, the model-free LOORP for discrete-time (DT) systems have gained interest. Gao et al. (2018) addressed the problem of cooperative output regulation for a class of DT multi-agent systems, where the dynamics of all the agents are considered unknown. Li et al. (2021) used Q-learning and output regulation to achieve tracking and disturbance rejection for multiplayer systems. Jiang et al. (2019) developed an off-policy PI to solve a special DT optimal output regulation problem. Chen et al. (2022) addressed the problem of robust output regulation using RL, where in addition to unknown system dynamics, partial state measurement is considered.

Policy iteration is a popular technique used by most of the studies mentioned above to compute the optimal controller. Persistence of excitation (PE) condition is an important criterion for guaranteeing the convergence and uniqueness of the PI algorithm. The PE condition is satisfied by incorporating an exploration/probing noise with the input while collecting data for learning (Jiang and Jiang (2012)). The PE condition is translated to requiring the full-rank condition of the data matrix used in the PI algorithm. As the probing noise affects the system states only, in case of model-free DT output regulation, it might be difficult to guarantee full-rank condition of the data matrix used in the PI algorithm. As in the case of DT output regulation, the data matrix used in the PI algorithm contains some columns that are formed using only the states of the exosystem which are not affected by the probing noise. Thus, in case of DT output regulation, the full-rank condition must be carefully stated such that the convergence and uniqueness of the PI algorithm is guaranteed. The existing literature, however, does not comment

⋆ This work has been supported partly by the NSF grant EPCN-1903781.

10.1016/j.ifacol.2023.10.912

on this important issue that arises in the DT output regulation formulation. In this work, we establish a proper rank condition such that the convergence and uniqueness of the PI algorithm is guaranteed. Also, an implicit assumption in the formulations given in a few works in the literature is that the state matrix $\mathbf{A}$ must be invertible in order to solve the regulator equations. In the formulation proposed in this paper, we avoid such an assumption by a novel reformulation of the problem.

The remainder of the paper is organized as follows: Section 2 formulates the basic control objective and presents some basic results on DT linear quadratic regulator (LQR) design. Section 3 presents a solution to the regulator equation with known parameters as well as a model-free technique to solve the LOORP problem. Section 4 provides the main results of the paper that includes redefining the rank condition of the data matrix in the PI algorithm to guarantee convergence and uniqueness properties. Lastly, Section 5 provides a numerical example to support the theoretical contributions of the paper.

Notations: Throughout this paper, $\mathbb{Z}_+$ denotes the set of non-negative integers, $\|.\|$ represents the spectral norm of matrices, $\sigma(\mathbf{W})$ is the complex spectrum of $\mathbf{W}$, $\otimes$ indicates the Kronecker product, $\text{vec}(\mathbf{T}) = \left[t_1^T, t_2^T, \cdots, t_m^T\right]^T$ with $t_i \in \mathbb{R}^r$ being the columns of $\mathbf{T} \in \mathbb{R}^{r\times m}$. For a symmetric matrix $\mathbf{P} \in \mathbb{R}^{m\times m}$, $\text{vecs}(\mathbf{P}) = [p_{11}, 2p_{12}, \cdots, 2p_{1m}, p_{22}, 2p_{23}, \cdots, 2p_{(m-1)m}, p_{mm}]^T \in \mathbb{R}^{(1/2)m(m+1)}$, for a column vector $v \in \mathbb{R}^n$, $\text{vecv}(v) = [v_1^2, v_1v_2, \cdots, v_1v_n, v_2^2, v_2v_3, \cdots, v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{(1/2)n(n+1)}$.

2. PROBLEM FORMULATION AND PRELIMINARIES

2.1 Problem Formulation

Consider the following discrete-time linear system given as:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{D}\mathbf{w}_k, \tag{1}$$
$$\mathbf{w}_{k+1} = \mathbf{E}\mathbf{w}_k, \tag{2}$$
$$\mathbf{e}_k = \mathbf{C}\mathbf{x}_k + \mathbf{F}\mathbf{w}_k, \tag{3}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state, $\mathbf{u}_k \in \mathbb{R}^m$ is the control input, and $\mathbf{w}_k \in \mathbb{R}^{q_m}$ is the state of the exosystem (2). $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{r\times n}$, $\mathbf{D} \in \mathbb{R}^{n\times q_m}$, $\mathbf{E} \in \mathbb{R}^{q_m\times q_m}$, and $\mathbf{F} \in \mathbb{R}^{r\times q_m}$ are constant matrices. $\mathbf{d}_k = \mathbf{D}\mathbf{w}_k$ is the exogenous disturbance, $\mathbf{y}_k = \mathbf{C}\mathbf{x}_k$ is the output of the plant, $\mathbf{y}_{dk} = -\mathbf{F}\mathbf{w}_k$ is the reference signal, and $\mathbf{e}_k \in \mathbb{R}^r$ is the tracking error.

Assumption 2.1. $(\mathbf{A}, \mathbf{B})$ *is stabilizable.*

Assumption 2.2. $rank\left(\begin{bmatrix}\mathbf{A}-\lambda\mathbf{I} & \mathbf{B}\\ \mathbf{C} & \mathbf{0}\end{bmatrix}\right) = n+r,\ \forall\lambda \in \sigma(\mathbf{E})$.

Remark 1. *Assumption 2.2 is a standard assumption to guarantee the existence of the solution to the regulator equations (5) and (6).*

In this paper, the discrete-time linear output regulation problem (LORP) is formulated by designing a controller of the form:

$$\mathbf{u}_k = -\mathbf{K}\mathbf{x}_k + \mathbf{L}\mathbf{w}_k, \tag{4}$$

where $\mathbf{K} \in \mathbb{R}^{m\times n}$ is the feedback gain and $\mathbf{L} \in \mathbb{R}^{m\times q_m}$ is the feedforward gain such that:

(1) the closed-loop system with the control law (4) is globally exponentially stable at the origin.
(2) the tracking error $\mathbf{e}_k$ asymptotically converges to zero.

If, in addition, the designed controller is optimal with respect to a cost function, the problem is termed as linear optimal output regulation problem (LOORP).

Theorem 2.1. *(Huang (2004)) Under Assumption 2.1, choose* $\mathbf{K}$ *such that the closed-loop system is stable. The LORP is solvable by the controller (4) if there exist* $\mathbf{X} \in \mathbb{R}^{n\times q_m}$, $\mathbf{U} \in \mathbb{R}^{m\times q_m}$ *solutions of the following regulator equations:*

$$\mathbf{X}\mathbf{E} = \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{U} + \mathbf{D}, \tag{5}$$
$$\mathbf{0} = \mathbf{C}\mathbf{X} + \mathbf{F}, \tag{6}$$

with the feedforward gain given as:

$$\mathbf{L} = \mathbf{U} + \mathbf{K}\mathbf{X}. \tag{7}$$

For any given initial condition $\mathbf{x_0}$ and $\mathbf{w_0}$, if the controller given in (4) solves the LORP, one can satisfy $\lim_{k\to\infty}\mathbf{u}_k - \mathbf{U}\mathbf{w}_k = 0$ and $\lim_{k\to\infty}\mathbf{x}_k - \mathbf{X}\mathbf{w}_k = 0$. By solving the LOORP problem in this paper, we attempt to solve the problem of asymptotic tracking and disturbance rejection for discrete-time linear systems. Let $\bar{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{X}^*\mathbf{w}_k$, and $\bar{\mathbf{u}}_k = \mathbf{u}_k - \mathbf{U}^*\mathbf{w}_k$, where $\mathbf{X}^*$ and $\mathbf{U}^*$ are the optimal solutions to the regulator equations (5) and (6) obtained by solving the following static optimization problem:

Problem 2.1.

$$\min_{\mathbf{X},\mathbf{U}} \quad Tr\left(\mathbf{X}^T\bar{\mathbf{Q}}\mathbf{X} + \mathbf{U}^T\bar{\mathbf{R}}\mathbf{U}\right), \tag{8}$$
$$\textit{subject to } (5)-(6),$$

where $\bar{\mathbf{Q}} = \bar{\mathbf{Q}}^T > 0$, and $\bar{\mathbf{R}} = \bar{\mathbf{R}}^T > 0$.

Using $\bar{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{X}^*\mathbf{w}_k$ and $\bar{\mathbf{u}}_k = \mathbf{u}_k - \mathbf{U}^*\mathbf{w}_k$, the following error system can be obtained:

$$\bar{\mathbf{x}}_{k+1} = \mathbf{A}\bar{\mathbf{x}}_k + \mathbf{B}\bar{\mathbf{u}}_k, \tag{9}$$
$$\mathbf{e}_k = \mathbf{C}\bar{\mathbf{x}}_k. \tag{10}$$

We solve the following dynamic optimization problem to find the optimal feedback gain $\mathbf{K}^*$:

Problem 2.2.

$$\min_{\bar{\mathbf{u}}} \quad J = \sum_{k=0}^{\infty}(\bar{\mathbf{x}}_k^{\mathrm{T}}\mathbf{Q}\bar{\mathbf{x}}_k + \bar{\mathbf{u}}_k^{\mathrm{T}}\mathbf{R}\bar{\mathbf{u}}_k), \tag{11}$$
$$\textit{subject to } (9),$$

where $\mathbf{Q} = \mathbf{Q}^T \geq 0$, $\mathbf{R} = \mathbf{R}^T > 0$, and $(\mathbf{A}, \sqrt{\mathbf{Q}})$ is observable.

Thus, solving Problems 2.1 and 2.2, one can find the optimal controller $\mathbf{u}_k^* = -\mathbf{K}^*\mathbf{x}_k + \mathbf{L}^*\mathbf{w}_k$.

Remark 2. *The design of optimal feedback controller gain* $\mathbf{K}^*$ *does not rely on the solutions* $\mathbf{X}^*$ *and* $\mathbf{U}^*$ *of the regulator equations. Thus, Problems 2.1 and 2.2 can be solved separately.*

2.2 Preliminaries

By solving the discrete-time LQR problem given in Problem 2.2, one can obtain the optimal feedback gain $\mathbf{K}^*$ as:

$$\mathbf{K}^* = (\mathbf{R} + \mathbf{B}^T\mathbf{P}^*\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}^*\mathbf{A}, \tag{12}$$

where $\mathbf{P}^* = \mathbf{P}^{*T} > 0$ is the unique solution of the following discrete-time algebraic Riccati equation:

$$\mathbf{A}^T\mathbf{P}\mathbf{A} - \mathbf{P} + \mathbf{Q} - \mathbf{A}^T\mathbf{P}\mathbf{B}(\mathbf{R} + \mathbf{B}^T\mathbf{P}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}\mathbf{A} = \mathbf{0}. \tag{13}$$

Note that, (13) is nonlinear in $\mathbf{P}$. Thus, it is usually difficult to directly solve (13) specially for high-dimensional systems. A model-based PI technique to solve (13) presented in Hewer (1971) is reproduced in Algorithm 1. Note that $\mathbf{A}_j = \mathbf{A} - \mathbf{B}\mathbf{K}_j$ in Algorithm 1.

Algorithm 1 Model-based PI

1: Select a stabilizing control policy $\mathbf{K}_0$ such that $\mathbf{A}-\mathbf{B}\mathbf{K}_0$ is a Schur matrix. Initialize $j \leftarrow 0$. Select a sufficiently small constant $\varepsilon > 0$.
2: **repeat**
3: Policy Evaluation:
$$\mathbf{A}_j^T\mathbf{P}_j\mathbf{A}_j - \mathbf{P}_j + \mathbf{Q} + \mathbf{K}_j^T\mathbf{R}\mathbf{K}_j = \mathbf{0}. \tag{14}$$
4: Policy Update:
$$\mathbf{K}_{j+1} = (\mathbf{R}+\mathbf{B}^T\mathbf{P}_j\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}_j\mathbf{A}. \tag{15}$$
5: $j \leftarrow j+1$.
6: **until** $\|\mathbf{P}_j - \mathbf{P}_{j-1}\| < \varepsilon$.

3. OPTIMAL OUTPUT REGULATOR DESIGN

In this section, we introduce a technique to solve the regulator equations (5) and (6). The matrices $\mathbf{A},\mathbf{B}$, and $\mathbf{D}$ are assumed to be unknown. At first, we present a model-based technique to solve the regulator equations (5) and (6). Then, we develop an optimal data-driven technique to compute $\mathbf{X}^*$ and $\mathbf{U}^*$ that solve (5) and (6), and $\mathbf{K}^*$ and $\mathbf{P}^*$ to solve the discrete-time LQR problem.

3.1 Model-Based Solution to the Regulator Equations

Define the Sylvester map $S : \mathbb{R}^{n\times q_m} \to \mathbb{R}^{n\times q_m}$ as:
$$S(\mathbf{X}) = \mathbf{X}\mathbf{E} - \mathbf{A}\mathbf{X}. \tag{16}$$
Pick a constant matrix $\mathbf{X}_1$ such that $\mathbf{C}\mathbf{X}_1 + \mathbf{F} = 0$. Then we select $\mathbf{X}_i$ for $i = 2,3,\cdots,h+1$ such that all the vectors $\text{vec}(\mathbf{X}_i)$ form a basis for $\ker(\mathbf{I}_{q_m}\otimes\mathbf{C})$, where $h = (n-r)q_m$ is the dimension of the null space of $\mathbf{I}_{q_m}\otimes\mathbf{C}$. A general solution to (6) can be given as:
$$\mathbf{X} = \mathbf{X}_1 + \sum_{i=2}^{h+1}\alpha_i\mathbf{X}_i, \tag{17}$$
where, $\alpha_i \in \mathbb{R}$. Then, (5) implies:
$$S(\mathbf{X}) = S(\mathbf{X}_1) + \sum_{i=2}^{h+1}\alpha_i S(\mathbf{X}_i) = \mathbf{B}\mathbf{U} + \mathbf{D}. \tag{18}$$
Now, (17) and (18) can be written as:
$$\mathscr{A}\boldsymbol{\chi} = \mathbf{b}, \tag{19}$$
where
$$\mathscr{A} = \begin{bmatrix} \text{vec}(S(\mathbf{X}_2)) & \cdots & \text{vec}(S(\mathbf{X}_{h+1})) & \mathbf{0} & -\mathbf{I}_{q_m}\otimes\mathbf{B} \\ \text{vec}(\mathbf{X}_2) & \cdots & \text{vec}(\mathbf{X}_{h+1}) & -\mathbf{I}_{nq_m} & \mathbf{0} \end{bmatrix}, \tag{20}$$
$$\boldsymbol{\chi} = \left[\alpha_2,\ \cdots,\ \alpha_{h+1},\ \text{vec}(\mathbf{X})^T,\ \text{vec}(\mathbf{U})^T\right]^T, \tag{21}$$
$$\mathbf{b} = \begin{bmatrix} \text{vec}(-S(\mathbf{X}_1)+\mathbf{D}) \\ -\text{vec}(\mathbf{X}_1) \end{bmatrix}. \tag{22}$$
Now, based on Gao and Jiang (2016), (19) can be written as:
$$\begin{bmatrix} \bar{\mathscr{A}}_{11} & \bar{\mathscr{A}}_{12} \\ \bar{\mathscr{A}}_{21} & \bar{\mathscr{A}}_{22} \end{bmatrix}\boldsymbol{\chi} = \begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \end{bmatrix}, \tag{23}$$
where $\bar{\mathscr{A}}_{21} \in \mathbb{R}^{h\times h}$ is a nonsingular matrix. Then, the following result holds.

Lemma 3.1. *A pair* $(\mathbf{X},\mathbf{U})$ *is a solution to the regulator equations if and only if it solves the following equation:*
$$\mathscr{M}\begin{bmatrix}\text{vec}(\mathbf{X}) \\ \text{vec}(\mathbf{U})\end{bmatrix} = \mathscr{N}, \tag{24}$$
where $\mathscr{M} = -\bar{\mathscr{A}}_{11}\bar{\mathscr{A}}_{21}^{-1}\bar{\mathscr{A}}_{22} + \bar{\mathscr{A}}_{12}$, $\mathscr{N} = -\bar{\mathscr{A}}_{11}\bar{\mathscr{A}}_{21}^{-1}\bar{b}_2 + \bar{b}_1$.

Thus, Problem 2.1 can be reformulated as:

Problem 3.1.
$$\min_{\mathbf{X},\mathbf{U}} \left(\begin{bmatrix}\text{vec}(\mathbf{X}) \\ \text{vec}(\mathbf{U})\end{bmatrix}\right)^T \begin{bmatrix}\mathbf{I}_{q_m}\otimes\bar{\mathbf{Q}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_m}\otimes\bar{\mathbf{R}}\end{bmatrix}\left(\begin{bmatrix}\text{vec}(\mathbf{X}) \\ \text{vec}(\mathbf{U})\end{bmatrix}\right), \tag{25}$$
subject to (24).

3.2 Model-Free Solution to the LQR Problem: Phase 1

Let us consider the following:
$$\bar{\mathbf{x}}_{k,i} = \mathbf{x}_k - \mathbf{X}_i\mathbf{w}_k, i = 0,1,\cdots,h+1, \tag{26}$$
where $\mathbf{X}_0 = \mathbf{0}$. Then, we have:
$$\bar{\mathbf{x}}_{k+1,i} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{D}\mathbf{w}_k - \mathbf{X}_i\mathbf{E}\mathbf{w}_k, \tag{27}$$
and
$$S(\mathbf{X}_i) = \mathbf{X}_i\mathbf{E} - \mathbf{A}\mathbf{X}_i. \tag{28}$$
From (26), using (27) and (28), it follows that:
$$\bar{\mathbf{x}}_{k+1,i} = (\mathbf{A}-\mathbf{B}\mathbf{K}_j)\bar{\mathbf{x}}_{k,i} + \mathbf{B}(\mathbf{u}_k + \mathbf{K}_j\bar{\mathbf{x}}_{k,i}) + (\mathbf{D} - S(\mathbf{X}_i))\mathbf{w}_k, \tag{29}$$
$$= \mathbf{A}_j\bar{\mathbf{x}}_{k,i} + \mathbf{B}(\mathbf{u}_k + \mathbf{K}_j\bar{\mathbf{x}}_{k,i}) + (\mathbf{D} - S(\mathbf{X}_i))\mathbf{w}_k. \tag{30}$$
Along the trajectories of (30), one can obtain that
$$\begin{aligned}&\bar{\mathbf{x}}_{k+1,i}^T\mathbf{P}_j\bar{\mathbf{x}}_{k+1,i} - \bar{\mathbf{x}}_{k,i}^T\mathbf{P}_j\bar{\mathbf{x}}_{k,i} \\ &= \left[\mathbf{A}_j\bar{\mathbf{x}}_{k,i} + \mathbf{B}(\mathbf{u}_k + \mathbf{K}_j\bar{\mathbf{x}}_{k,i}) + (\mathbf{D} - S(\mathbf{X}_i))\mathbf{w}_k\right]^T\mathbf{P}_j\left[\mathbf{A}_j\bar{\mathbf{x}}_{k,i} + \right. \\ &\qquad \left.\mathbf{B}(\mathbf{u}_k + \mathbf{K}_j\bar{\mathbf{x}}_{k,i}) + (\mathbf{D} - S(\mathbf{X}_i))\mathbf{w}_k\right] - \bar{\mathbf{x}}_{k,i}^T\mathbf{P}_j\bar{\mathbf{x}}_{k,i}.\end{aligned} \tag{31}$$
Then, using (14) we have:
$$\begin{aligned}&\bar{\mathbf{x}}_{k+1,i}^T\mathbf{P}_j\bar{\mathbf{x}}_{k+1,i} - \bar{\mathbf{x}}_{k,i}^T\mathbf{P}_j\bar{\mathbf{x}}_{k,i} + \bar{\mathbf{x}}_{k,i}^T\mathbf{Q}_j\bar{\mathbf{x}}_{k,i} \\ &= 2\bar{\mathbf{x}}_{k,i}^T\mathbf{A}^T\mathbf{P}_j\mathbf{B}\mathbf{u}_k + 2\bar{\mathbf{x}}_{k,i}^T\mathbf{A}^T\mathbf{P}_j\mathbf{B}\mathbf{K}_j\bar{\mathbf{x}}_{k,i} - \bar{\mathbf{x}}_{k,i}^T\mathbf{K}_j^T\mathbf{B}^T\mathbf{P}_j\mathbf{B}\mathbf{K}_j\bar{\mathbf{x}}_{k,i} \\ &\quad + \mathbf{u}_k^T\mathbf{B}^T\mathbf{P}_j\mathbf{B}\mathbf{u}_k + 2\bar{\mathbf{x}}_{k,i}^T\boldsymbol{\Theta}_{1ij}\mathbf{w}_k + 2\mathbf{u}_k^T\boldsymbol{\Theta}_{2ij}\mathbf{w}_k + \mathbf{w}_k^T\boldsymbol{\Theta}_{3ij}\mathbf{w}_k,\end{aligned} \tag{32}$$
where $\mathbf{Q}_j = \mathbf{Q} + \mathbf{K}_j^T\mathbf{R}\mathbf{K}_j$, $\boldsymbol{\Theta}_{1ij} = \mathbf{A}^T\mathbf{P}_j(\mathbf{D} - S(\mathbf{X}_i))$,
$\boldsymbol{\Theta}_{2ij} = \mathbf{B}^T\mathbf{P}_j(\mathbf{D} - S(\mathbf{X}_i))$, $\boldsymbol{\Theta}_{3ij} = (\mathbf{D} - S(\mathbf{X}_i))^T\mathbf{P}_j(\mathbf{D} - S(\mathbf{X}_i))$.
Now, by the property of Kronecker product that $\text{vec}(\mathbf{XYZ}) = (\mathbf{Z}^T\otimes\mathbf{X})\text{vec}(\mathbf{Y})$, we have:
$$\begin{aligned}&\left[(\bar{\mathbf{x}}_{k+1,i}^T\otimes\bar{\mathbf{x}}_{k+1,i}^T) - (\bar{\mathbf{x}}_{k,i}^T\otimes\bar{\mathbf{x}}_{k,i}^T)\right]\text{vec}(\mathbf{P}_j) + (\bar{\mathbf{x}}_{k,i}^T\otimes\bar{\mathbf{x}}_{k,i}^T)\text{vec}(\mathbf{Q}_j) \\ &= \left[2(\bar{\mathbf{x}}_{k,i}^T\otimes\mathbf{u}_k^T) + 2(\bar{\mathbf{x}}_{k,i}^T\otimes\bar{\mathbf{x}}_{k,i}^T)(\mathbf{I}_n\otimes\mathbf{K}_j^T)\right]\text{vec}(\mathbf{B}^T\mathbf{P}_j\mathbf{A}) + \\ &\quad \left[-(\mathbf{K}_j\bar{\mathbf{x}}_{k,i})^T\otimes(\mathbf{K}_j\bar{\mathbf{x}}_{k,i})^T + (\mathbf{u}_k^T\otimes\mathbf{u}_k^T)\right]\text{vec}(\mathbf{B}^T\mathbf{P}_j\mathbf{B}) + \\ &\quad 2(\mathbf{w}_k^T\otimes\bar{\mathbf{x}}_{k,i}^T)\text{vec}(\boldsymbol{\Theta}_{1ij}) + 2(\mathbf{w}_k^T\otimes\mathbf{u}_k^T)\text{vec}(\boldsymbol{\Theta}_{2ij}) + \\ &\quad (\mathbf{w}_k^T\otimes\mathbf{w}_k^T)\text{vec}(\boldsymbol{\Theta}_{3ij}).\end{aligned} \tag{33}$$
Collecting the data for the time sequence $k_0 < k_1 < \cdots < k_s$, we get
$$\boldsymbol{\Psi}_{1ij}\boldsymbol{\theta}_{1ij} = -\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}\text{vec}(\mathbf{Q}_j), \tag{34}$$
where $\boldsymbol{\Psi}_{1ij} = \Big[\boldsymbol{\Delta}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, -2\mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}} - 2\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}(\mathbf{I}_n\otimes\mathbf{K}_j^T), \tilde{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} - \mathbf{I}_{\mathbf{u},\mathbf{u}}$,
$-2\mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}, -2\mathbf{I}_{\mathbf{w},\mathbf{u}}, -\mathbf{I}_{\mathbf{w},\mathbf{w}}\Big]$,
$\boldsymbol{\theta}_{1ij} = \Big[\text{vecs}(\mathbf{P}_j)^T, \text{vec}(\mathbf{B}^T\mathbf{P}_j\mathbf{A})^T, \text{vecs}(\mathbf{B}^T\mathbf{P}_j\mathbf{B})^T, \text{vec}(\boldsymbol{\Theta}_{1ij})^T$,
$\text{vec}(\boldsymbol{\Theta}_{2ij})^T, \text{vecs}(\boldsymbol{\Theta}_{3ij})^T\Big]^T$,

$$\boldsymbol{\Delta}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} = \Big[\text{vecv}(\bar{\mathbf{x}}_{k_0+1,i}) - \text{vecv}(\bar{\mathbf{x}}_{k_0,i}), \cdots, \text{vecv}(\bar{\mathbf{x}}_{k_s,i}) - \text{vecv}(\bar{\mathbf{x}}_{k_s-1,i})\Big]^T \in \mathbb{R}^{s\times n(n+1)/2},$$

$$\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} = \Big[(\bar{\mathbf{x}}_{k_0,i}\otimes\bar{\mathbf{x}}_{k_0,i}), \cdots, (\bar{\mathbf{x}}_{k_s,i}\otimes\bar{\mathbf{x}}_{k_s,i})\Big]^T \in \mathbb{R}^{s\times n^2},$$

$$\tilde{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} = \Big[\text{vecv}(\mathbf{K}_j\bar{\mathbf{x}}_{k_0,i}), \cdots, \text{vecv}(\mathbf{K}_j\bar{\mathbf{x}}_{k_s,i})\Big]^T \in \mathbb{R}^{s\times m(m+1)/2},$$

$$\mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}} = \Big[\bar{\mathbf{x}}_{k_0,i}\otimes\mathbf{u}_{k_0}, \cdots, \bar{\mathbf{x}}_{k_s,i}\otimes\mathbf{u}_{k_s}\Big]^T \in \mathbb{R}^{s\times mn},$$

$$\mathbf{I}_{\mathbf{u},\mathbf{u}} = \Big[\text{vecv}(\mathbf{u}_{k_0}), \cdots, \text{vecv}(\mathbf{u}_{k_s})\Big]^T \in \mathbb{R}^{s\times m(m+1)/2},$$

$$\mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i} = \Big[\mathbf{w}_{k_0}\otimes\bar{\mathbf{x}}_{k_0,i}, \cdots, \mathbf{w}_{k_s}\otimes\bar{\mathbf{x}}_{k_s,i}\Big]^T \in \mathbb{R}^{s\times nq_m},$$

$$\mathbf{I}_{\mathbf{w},\mathbf{u}} = \Big[\mathbf{w}_{k_0}\otimes\mathbf{u}_{k_0}, \cdots, \mathbf{w}_{k_s}\otimes\mathbf{u}_{k_s}\Big]^T \in \mathbb{R}^{s\times mq_m},$$

$$\mathbf{I}_{\mathbf{w},\mathbf{w}} = \Big[\text{vecv}(\mathbf{w}_{k_0}), \cdots, \text{vecv}(\mathbf{w}_{k_s})\Big]^T \in \mathbb{R}^{s\times q_m(q_m+1)/2}.$$

Assumption 3.1. *For $i = 0,1,\cdots,h+1$ there exists a $s^* \in \mathbb{Z}_+$ such that for all $s > s^*$:*

$$\text{rank}([\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}_i}, \mathbf{I}_{\mathbf{u},\mathbf{u}}, \mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}, \mathbf{I}_{\mathbf{w},\mathbf{u}}, \bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}]) = \frac{n(n+1)}{2} + nm + \frac{m(m+1)}{2} + nq_m + mq_m + \frac{q_m(q_m+1)}{2} - N, \quad (35)$$

where N is the number of linearly dependent columns of $\mathbf{I}_{\mathbf{w},\mathbf{w}}$, and $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ is constructed by reducing those linearly dependent columns.

Remark 3. *A typical choice of s^* can be $s^* \geq \frac{n(n+1)}{2} + nm + \frac{m(m+1)}{2} + nq_m + mq_m + \frac{q_m(q_m+1)}{2}$. In Section 4, we discuss how the columns of $\mathbf{I}_{\mathbf{w},\mathbf{w}}$ can be linearly dependent by use of an example exosystem.*

Algorithm 2 Phase-1 Model-Free Policy Iteration

1: Compute matrices $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_{h+1}$.
2: Employ $\mathbf{u}_k = -\mathbf{K}_0\mathbf{x}_k + \boldsymbol{\eta}_k$ as the input on the time interval $[k_0, k_s]$, where $\mathbf{K}_0$ is an initial stabilizing gain and $\boldsymbol{\eta}_k$ is the exploration/probing noise.
3: For $i = 0,1,\cdots,h+1$, compute $\boldsymbol{\Delta}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}}, \mathbf{I}_{\mathbf{u},\mathbf{u}}, \mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}, \mathbf{I}_{\mathbf{w},\mathbf{u}}, \bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ until the rank condition in (35) is satisfied. Let $i = 0, j = 0$.
4: Solve for $\boldsymbol{\theta}_{1ij}$ from (34) using $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ in $\boldsymbol{\Psi}_{1ij}$. Then, $\mathbf{K}_{j+1} = (\mathbf{R} + \mathbf{B}^T\mathbf{P}_j\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}_j\mathbf{A}$.
5: Let $j \leftarrow j+1$ and repeat Step 4 until $\|\mathbf{P}_j - \mathbf{P}_{j-1}\| \leq \varepsilon_0$ for $j \geq 1$, where the constant $\varepsilon_0 > 0$ is a predefined small threshold.

3.3 Model-Free Solution to the Regulator Equations: Phase 2

Using (30), one can obtain:

$$\begin{aligned}&\bar{\mathbf{x}}_{k+1,i}^T\mathbf{P}_{j*}\bar{\mathbf{x}}_{k,i}\\ &= \bar{\mathbf{x}}_{k,i}^T\mathbf{A}^T\mathbf{P}_{j*}\bar{\mathbf{x}}_{k,i} + \mathbf{u}_k^T\mathbf{B}^T\mathbf{P}_{j*}\bar{\mathbf{x}}_{k,i} + \bar{\mathbf{x}}_{k,i}^T\mathbf{P}_{j*}(\mathbf{D} - S(\mathbf{X}_i))\mathbf{w}_k, \quad (36)\end{aligned}$$

where $\mathbf{P}_{j*}$ is the approximated solution of the Riccati equation obtained from Algorithm 2. Now, using Kronecker product:

$$\begin{aligned}&(\bar{\mathbf{x}}_{k,i}^T\otimes\bar{\mathbf{x}}_{k+1,i}^T)\text{vec}(\mathbf{P}_{j*}) = (\bar{\mathbf{x}}_{k,i}^T\otimes\bar{\mathbf{x}}_{k,i}^T)\text{vec}(\mathbf{A}^T\mathbf{P}_{j*})\\ &+ (\bar{\mathbf{x}}_{k,i}^T\otimes\mathbf{u}_k^T)\text{vec}(\mathbf{B}^T\mathbf{P}_{j*}) + (\mathbf{w}_k^T\otimes\bar{\mathbf{x}}_{k,i}^T)\text{vec}(\mathbf{P}_{j*}(\mathbf{D} - S(\mathbf{X}_i))). \quad (37)\end{aligned}$$

Using the data collected from Phase 1, one can obtain:

$$\boldsymbol{\Psi}_{2i}\boldsymbol{\theta}_{2i} = \boldsymbol{\Lambda}_i\text{vec}(\mathbf{P}_{j*}), \quad (38)$$

where, $\boldsymbol{\Psi}_{2i} = \Big[\bar{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}}, \mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}\Big]$, $\boldsymbol{\theta}_{2i} = \Big[(1/2)\text{vecs}(\mathbf{A}^T\mathbf{P}_{j*} + \mathbf{P}_{j*}\mathbf{A})^T, \text{vec}(\mathbf{B}^T\mathbf{P}_{j*})^T, \text{vec}(\mathbf{P}_{j*}(\mathbf{D} - S(\mathbf{X}_i)))^T\Big]^T$, $\boldsymbol{\Lambda}_i = \Big[\bar{\mathbf{x}}_{k_0,i}\otimes\bar{\mathbf{x}}_{k_1,i}, \cdots, \bar{\mathbf{x}}_{k_{s-1},i}\otimes\bar{\mathbf{x}}_{k_s,i}\Big]^T$, $\bar{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} = \Big[\text{vecv}(\bar{\mathbf{x}}_{k_0,i}), \cdots, \text{vecv}(\bar{\mathbf{x}}_{k_s,i})\Big]^T$.

Assumption 3.2. *There exists a $s^* \in \mathbb{Z}_+$ such that for all $s > s^*$:*

$$\text{rank}([\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}}, \mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}]) = \frac{n(n+1)}{2} + nm + nq_m. \quad (39)$$

Using the rank condition in (39), the least squares problem in (38) can be uniquely solved for $i = 0,1,\cdots,h+1$ to obtain the Sylvester maps $S(\mathbf{X}_i)$. When $i = 0$, $\mathbf{X}_0 = \mathbf{0}$ and one can obtain matrices $\mathbf{B}$ and $\mathbf{D}$. Once, $S(\mathbf{X}_i)$'s, $\mathbf{B}$ and $\mathbf{D}$ are obtained, $\mathscr{M}$ and $\mathscr{N}$ are completely known. Thus, one can solve Problem 3.1 to obtain the solution of the regulator equation. Once $\mathbf{X}^*$ and $\mathbf{U}^*$ are obtained by solving Problem 3.1, one can obtain the feedforward gain as $\mathbf{L}_{j*} = \mathbf{U}^* + \mathbf{K}_{j*}\mathbf{X}^*$, where $\mathbf{K}_{j*}$ is obtained from Algorithm 2. Note that we do not require the information of $\mathbf{A}$ to solve Problem 3.1.

Remark 4. *Note that Assumptions 3.1 and 3.2 are like persistency of excitation in the adaptive control (Jiang and Jiang (2017); Vamvoudakis and Lewis (2010)).*

Remark 5. *As mentioned before, some works in the literature have an implicit assumption that the state matrix $\mathbf{A}$ must be invertible in order to solve the regulator equation. Note that, in the formulation given in this section, we do not need this assumption. One just needs to solve for $S(\mathbf{X}_i)$'s, $\mathbf{B}$, and $\mathbf{D}$ using the Phase 2 learning, then solve Problem 3.1 to obtain the solution for the regulator equations. In other words, the solution obtained for the regulator equations using the learning based method is consistent with that of the model-based method.*

4. CONVERGENCE AND UNIQUENESS ANALYSIS

Since the probing noise in Algorithm 2 does not affect the exosystem, one cannot guarantee the full rank condition of the matrix $\mathbf{I}_{\mathbf{w},\mathbf{w}}$. Consider the following example of an exosystem that generates sinusoidal disturbance and a constant reference:

$$\mathbf{w}_{k+1} = \mathbf{E}\mathbf{w}_k = \begin{bmatrix}\cos(\theta) & -\sin(\theta) & 0\\ \sin(\theta) & \cos(\theta) & 0\\ 0 & 0 & 1\end{bmatrix}\mathbf{w}_k = \begin{bmatrix}c & -s & 0\\ s & c & 0\\ 0 & 0 & 1\end{bmatrix}\mathbf{w}_k. \quad (40)$$

The state transition matrix can be obtained as:

$$\mathbf{E}^k = \begin{bmatrix}\alpha & -\beta & 0\\ \beta & \alpha & 0\\ 0 & 0 & 1\end{bmatrix}, \quad (41)$$

where $\alpha = 0.5[(c-l.s)^k + (c+l.s)^k]$, and $\beta = 0.5[l.(c-l.s)^k - l.(c+l.s)^k]$, $l = \sqrt{-1}$.

Thus, the states of the exosystem have the following solutions:

$$w_{1,k} = \alpha w_{1,0} - \beta w_{2,0}, \tag{42}$$
$$w_{2,k} = \beta w_{1,0} + \alpha w_{2,0}, \tag{43}$$
$$w_{3,k} = w_{3,0}, \tag{44}$$

where $w_{1,0}, w_{2,0}$, and $w_{3,0}$ are the initial conditions. Now, the kronecker product $\mathbf{w}_k^T \otimes \mathbf{w}_k^T$ has the unique components: $\text{vecv}(\mathbf{w}_k) = [w_{1,k}^2,\ w_{1,k}w_{2,k},\ w_{1,k}w_{3,k},\ w_{2,k}^2,\ w_{2,k}w_{3,k},\ w_{3,k}^2]$.

Consider a constant $\gamma = \frac{w_{3,0}^2}{w_{1,0}^2 + w_{2,0}^2}$. Now,

$$\gamma w_{1,k}^2 + \gamma w_{2,k}^2 = \gamma(\alpha^2 + \beta^2)(w_{1,0}^2 + w_{2,0}^2). \tag{45}$$

Since $\cos(\theta) = \frac{e^{l\theta} + e^{-l\theta}}{2}$, $\sin(\theta) = \frac{e^{l\theta} - e^{-l\theta}}{2l}$, one can obtain $\alpha = 0.5[(e^{-l\theta})^k + (e^{l\theta})^k]$, $\beta = 0.5[l.(e^{-l\theta})^k - l.(e^{l\theta})^k]$. It is easy to see that $\alpha^2 + \beta^2 = 1$. Thus, $\gamma(w_{1,k}^2 + w_{2,k}^2) = w_{3,k}^2$. This shows the dependence of components of $\text{vecv}(\mathbf{w}_k)$. Hence, the matrix $\mathbf{I}_{\mathbf{w},\mathbf{w}}$ is not full rank. Thus, one cannot guarantee that the data matrix $\mathbf{I}_{\mathbf{w},\mathbf{w}}$ constructed with the states of the exosystem has full rank.

By the above discussion, we incorporate $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ in (34) by reducing the linearly dependent columns of $\mathbf{I}_{\mathbf{w},\mathbf{w}}$. Since $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ has less number of columns, the size of $\text{vecs}(\mathbf{\Theta}_{3ij})$ is also reduced. Note that $\text{vecs}(\mathbf{\Theta}_{3ij})$ is not an essential unknown to be learned. Thus, it neither effects the solution of Ricatti equation nor the solution of the regulator equation.

Theorem 4.1. *Using $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ in (34) one can obtain:*

$$\bar{\mathbf{\Psi}}_{1ij}\bar{\boldsymbol{\theta}}_{1ij} = -\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}\text{vec}(\mathbf{Q}_j), \tag{46}$$

where $\mathbf{\Psi}_{1ij} = \Big[\mathbf{\Delta}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, -2\mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}} - 2\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}(\mathbf{I}_n \otimes \mathbf{K}_j^T), \tilde{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} - \mathbf{I}_{\mathbf{u},\mathbf{u}}, -2\mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}, -2\mathbf{I}_{\mathbf{w},\mathbf{u}}, -\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}\Big]$, *and*

$\bar{\boldsymbol{\theta}}_{1ij} = \Big[\text{vecs}(\mathbf{P}_j)^T, \text{vec}(\mathbf{B}^T\mathbf{P}_j\mathbf{A})^T, \text{vecs}(\mathbf{B}^T\mathbf{P}_j\mathbf{B})^T, \text{vec}(\mathbf{\Theta}_{1ij})^T, \text{vec}(\mathbf{\Theta}_{2ij})^T, \text{vec}(\bar{\mathbf{\Theta}}_{3ij})^T\Big]^T$. *Then, under the Assumption 3.1:*

(a) (46) has a unique solution.
(b) the sequence $\{\mathbf{P}_j\}_{j=0}^{\infty}$ and $\{\mathbf{K}_j\}_{j=0}^{\infty}$ obtained using Algorithm 2 converges to the optimal values $\mathbf{P}^*$ and $\mathbf{K}^*$, respectively.

Proof. (a) Note that $\bar{\boldsymbol{\theta}}_{1ij}$ can be obtained from (46) using least squares. Under Assumption 3.1, $\bar{\mathbf{\Psi}}_{1ij}$ has full rank. Thus $\bar{\boldsymbol{\theta}}_{1ij}$ is unique.

(b) Given a stabilizing control gain $\mathbf{K}_j$, if $\mathbf{P}_j = \mathbf{P}_j^T$ is the unique solution of (14), $\mathbf{K}_{j+1}$ is uniquely determined by $\mathbf{K}_{j+1} = (\mathbf{R} + \mathbf{B}^T\mathbf{P}_j\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}_j\mathbf{A}$. Let $\mathbf{\Gamma}_{1j} = \mathbf{B}^T\mathbf{P}_j\mathbf{A}$, and $\mathbf{\Gamma}_{2j} = \mathbf{B}^T\mathbf{P}_j\mathbf{B}$. By (32) we know that $\mathbf{P}_j$, $\mathbf{\Gamma}_{1j}$, $\mathbf{\Gamma}_{2j}$, $\mathbf{\Theta}_{1ij}$, $\mathbf{\Theta}_{2ij}$, and $\bar{\mathbf{\Theta}}_{3ij}$ satisfy (46). Let, $\mathbf{P}$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Theta}_{1i}$, $\mathbf{\Theta}_{2i}$, and $\bar{\mathbf{\Theta}}_{3i}$ of appropriate dimensions solve (46). Then, we have $\mathbf{P}_j = \mathbf{P}$, $\mathbf{\Gamma}_{1j} = \mathbf{\Gamma}_1$, $\mathbf{\Gamma}_{2j} = \mathbf{\Gamma}_2$, $\mathbf{\Theta}_{1ij} = \mathbf{\Theta}_{1i}$, $\mathbf{\Theta}_{2ij} = \mathbf{\Theta}_{2i}$, and $\bar{\mathbf{\Theta}}_{3ij} = \bar{\mathbf{\Theta}}_{3i}$. Then from part (a), we know that $\mathbf{P}$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Theta}_{1i}$, $\mathbf{\Theta}_{2i}$, and $\bar{\mathbf{\Theta}}_{3i}$ are unique. Thus, the PI in Algorithm 2 is same as Algorithm 1. Thus, the theorem is proved by the equivalence of the two algorithms. □

5. RESULTS AND DISCUSSION

We show the efficacy of the proposed algorithm by a numerical example. Consider the following discrete-time system:

$$\mathbf{x}_{k+1} = \begin{bmatrix} 0.3417 & -0.6217 & 0.0364 \\ 0.0622 & 0.9630 & -0.0982 \\ 0.0004 & 0.0098 & 0.9896 \end{bmatrix}\mathbf{x}_k + \begin{bmatrix} 0.6218 \\ 0.0365 \\ 0.0001 \end{bmatrix}\mathbf{u}_k + \mathbf{d}_k, \tag{47}$$

$$\mathbf{w}_{k+1} = \begin{bmatrix} \cos(0.01) & -\sin(0.01) & 0 & 0 \\ \sin(0.01) & \cos(0.01) & 0 & 0 \\ 0 & 0 & \cos(0.1) & -\sin(0.1) \\ 0 & 0 & \sin(0.1) & \cos(0.1) \end{bmatrix}\mathbf{w}_k, \tag{48}$$

$$\mathbf{e}_k = [0\ 0\ 1]\mathbf{x}_k + [5\sqrt{3}\ 5\ 0\ 0]\mathbf{w}_k. \tag{49}$$

The upper 2×2 subsystem in (48) is used to generate the reference signal and the lower 2×2 subsystem in (48) is used to generate the disturbance. The initial conditions are given as $\mathbf{x}_0 = [1,\ 2,\ 3]$, and $\mathbf{w}_0 = [1,\ 0,\ 1,\ 0]$. The system matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ are considered unknown. The weight matrices $\mathbf{Q}$, and $\mathbf{R}$ are chosen as identity matrices. The initial stabilizing controller gain is $\mathbf{K}_0 = [-0.8727,\ -0.9849,\ -0.1354]$. The exploration noise in Algorithm 2 is chosen as the summation of sinusoidal waves with different frequencies.

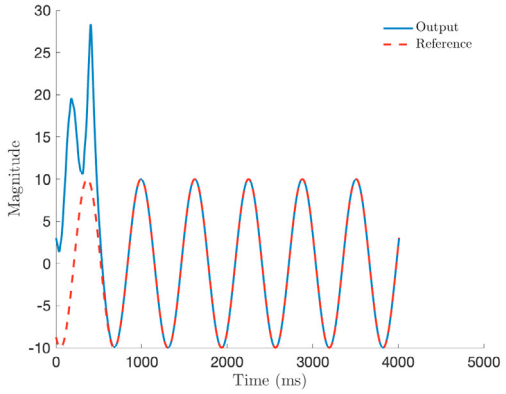

Figure 1. Output and reference trajectories.

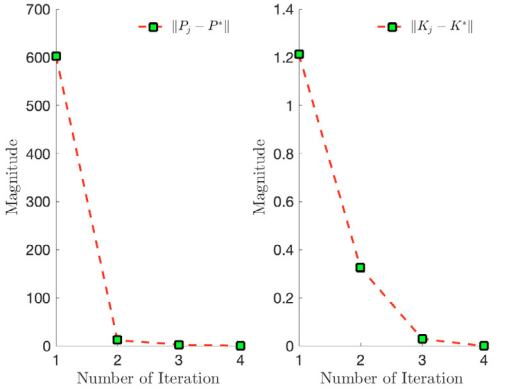

Figure 2. Convergence of $\mathbf{P}_j$ to $\mathbf{P}^*$, and $\mathbf{K}_j$ to $\mathbf{K}^*$.

Using the learning data, Algorithm 2 converges with a tolerance of $\varepsilon_0 = 0.05$ to a neighborhood of the optimal values $\mathbf{P}^*$ and $\mathbf{K}^*$ in 4 iterations as shown in Fig. 2. The optimal controller gain

$\mathbf{K}^\star$ and the controller gain obtained from Algorithm 2 are given as:

$$\mathbf{K}^\star = [0.1972 \ 0.1488 \ -0.1688], \quad (50)$$

$$\mathbf{K}_4 = [0.1973 \ 0.1489 \ -0.1684]. \quad (51)$$

The optimal feedforward gain $\mathbf{L}^\star$ and the feedforward gain obtained from Phase 2 (Section 3.3) are given as:

$$\mathbf{L}^\star = [-24.2085 \ 0.7679 \ 0 \ 0], \quad (52)$$

$$\mathbf{L}_4 = [-24.2131 \ 0.7659 \ 0 \ 0]. \quad (53)$$

6. CONCLUSION

This paper addresses the problem of discrete-time output regulation when the system parameters are unknown. It was shown that the rank condition of the data matrix used in the PI algorithm must be carefully chosen in order to guarantee the convergence and uniqueness properties of the PI algorithm. This is crucial as certain columns of the data matrix are constructed using only the states of the exosystem which are not affected by the probing noise during data collection. Also, the existing methodologies in the literature implicitly assumes the invertibility of the state matrix in order to solve the regulator equation. Thus, in case where the state matrix is not full rank, the model-based and model-free techniques for solving the regulator equation will yield different results. This issue is also addressed in this work by a novel reformulation of the problem that avoids the invertibility assumption on the state matrix. Finally, numerical simulation is provided to demonstrate the validity of the proposed methodology. Future work will focus on extending the results to nonlinear systems.

REFERENCES

Bertsekas, D. (2012). *Dynamic programming and optimal control*, volume 1. Athena scientific.

Bonivento, C., Marconi, L., and Zanasi, R. (2001). Output regulation of nonlinear systems by sliding mode. *Automatica*, 37(4), 535–542.

Chakraborty, S., Cui, L., Ozbay, K., and Jiang, Z.P. (2022). Automated lane changing control in mixed traffic: An adaptive dynamic programming approach. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 1823–1828. IEEE.

Chen, C., Xie, L., Jiang, Y., Xie, K., and Xie, S. (2022). Robust output regulation and reinforcement learning-based output tracking design for unknown linear discrete-time systems. *IEEE Transactions on Automatic Control*.

Cui, L. and Jiang, Z.P. (2022). A reinforcement learning look at risk-sensitive linear quadratic gaussian control. *arXiv preprint arXiv:2212.02072*.

Cui, L., Pang, B., and Jiang, Z.P. (2022). Learning-based adaptive optimal control of linear time-delay systems: A policy iteration approach. *arXiv preprint arXiv:2210.00204*.

Gao, W. and 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., Liu, Y., Odekunle, A., Yu, Y., and Lu, P. (2018). Adaptive dynamic programming and cooperative output regulation of discrete-time multi-agent systems. *International Journal of Control, Automation and Systems*, 16(5), 2273–2281.

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

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

Jiang, Y., Kiumarsi, B., Fan, J., Chai, T., Li, J., and Lewis, F.L. (2019). Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. *IEEE Transactions on Cybernetics*, 50(7), 3147–3156.

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

Jiang, Y. and Jiang, Z.P. (2017). *Robust adaptive dynamic programming*. John Wiley & Sons.

Krener, A.J. (1992). The construction of optimal linear and nonlinear regulators. In *Systems, Models and Feedback: Theory and Applications*, 301–322. Springer.

Li, J., Xiao, Z., Li, P., and Cao, J. (2021). Robust optimal tracking control for multiplayer systems by off-policy q-learning approach. *International Journal of Robust and Nonlinear Control*, 31(1), 87–106.

Liu, W. and Huang, J. (2020). Output regulation of linear systems via sampled-data control. *Automatica*, 113, 108684.

Mantri, R., Saberi, A., Lin, Z., and Stoorvogel, A.A. (1997). Output regulation for linear discrete-time systems subject to input saturation. *International Journal of Robust and Nonlinear Control*, 7(11), 1003–1021.

Saberi, A., Stoorvogel, A.A., Sannuti, P., and Shi, G. (2003). On optimal output regulation for linear systems. *International Journal of Control*, 76(4), 319–333.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement learning: An introduction*. MIT press.

Trentelman, H.L., Stoorvogel, A.A., Hautus, M., and Dewell, L. (2002). Control theory for linear systems. *Appl. Mech. Rev.*, 55(5), B87–B87.

Vamvoudakis, K.G. and Lewis, F.L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5), 878–888.

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., and Lewis, F.L. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, 45(2), 477–484.

Yan, Y. and Huang, J. (2016). Cooperative output regulation of discrete-time linear time-delay multi-agent systems. *IET Control Theory & Applications*, 10(16), 2019–2026.