2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Adaptive Optimal Output Regulation of Discrete-Time Linear Systems: A Reinforcement Learning Approach

Sayan Chakraborty[1], Weinan Gao[2], Kyriakos G. Vamvoudakis[3], Zhong-Ping Jiang[1]

Abstract— **In this paper, we solve the optimal output regulation problem for discrete-time systems without precise knowledge of the system model. Drawing inspiration from reinforcement learning and adaptive dynamic programming, a data-driven solution is developed that enables asymptotic tracking and disturbance rejection. Notably, it is discovered that the proposed approach for discrete-time output regulation differs from the continuous-time approach in terms of the persistent excitation condition required for policy iteration to be unique and convergent. To address this issue, a new persistent excitation condition is introduced to ensure both uniqueness and convergence of the data-driven policy iteration. The efficacy of the proposed methodology is validated by an inverted pendulum on a cart example.**

I. Introduction

The problem of output regulation is a fundamental research topic in control theory and involves designing a feedback control law to achieve asymptotic tracking while rejecting disturbances. This problem is applicable to various fields, including engineering [1], [2], and [3]. In prior research, many authors have studied output regulation when the system dynamics are known, [4], [5], [6], [2], [7], and [8]. However, these studies require perfect knowledge of the system model, which is not always feasible. To address this limitation, researchers have developed model-free optimal control techniques using ideas from reinforcement learning [9] and adaptive dynamic programming (ADP) [10], [11], [12], [13], [14]. For instance, the authors of [15] introduced a novel policy iteration (PI) based optimal control technique that only requires partial knowledge of the system dynamics. In [16], the authors proposed an original model-free off-policy PI algorithm for optimal control of linear systems with completely unknown system dynamics. Other related works in the domain of model-free optimal control can be found in [17], [18], [19], [20], and recent developments in this area can be found in [21], [22], [23].

Recently, there has been an increasing interest in developing model-free techniques for output regulation. Authors in [24] proposed integrating ADP and output regulation theory to address asymptotic tracking and disturbance rejection. Later, a data-driven output regulation formulation was developed for non-linear systems in [25], which has since been applied to control various systems, such as a boiler-turbine system [26] and the longitudinal and lateral control of autonomous vehicles [27]. Recently, model-free techniques for discrete-time systems have gained significant attention, such as the work of [28] that addressed the problem of cooperative output regulation for a class of discrete-time multi-agent systems where the dynamics of all agents are considered unknown. Other recent works include [29], using Q-learning and output regulation to achieve tracking and disturbance rejection for multi-agent systems, and [30], developing an off-policy PI to solve discrete-time optimal output regulation problem. Another recent work by [31] addressing the problem of robust output regulation using reinforcement learning considering partial state measurements.

Most of the existing studies above utilize PI to compute the optimal controller. The convergence and uniqueness of the PI algorithm require the satisfaction of a persistence of excitation (PE) condition, which translates to requiring full column rank of the data matrix used in the PI. The PE condition is met by introducing probing noise to the system input during data collection [16]. For model-free discrete-time output regulation, however, it may be challenging to ensure that the data matrix used in the PI algorithm is full rank since the probing noise affects only the system states. Consequently, some columns of the data matrix used in the PI algorithm formed using only the states of the exosystem are unaffected by the probing noise. Therefore, careful selection of the rank condition is crucial for ensuring the convergence and uniqueness of the PI algorithm in discrete-time output regulation. This important issue has not been clearly addressed in the literature. This work aims to establish an appropriate rank condition that guarantees the convergence and uniqueness of the PI algorithm. Another difference from the existing literature is the consideration of feedthrough term in the plant output, which leads to a different data-driven approach to solve the regulator equations that is illustrated in this work. Additionally, to align model-based solutions of regulator equations with model-free techniques, existing methods [29], [30] may necessitate the state/plant matrix to be invertible. In this paper, we avoid such an assumption by a novel reformulation of the problem.

The remainder of the paper is structured as follows. Section II formulates the basic control objective and presents some model-based results on discrete-time linear optimal output regulator problem (LOORP). Section III presents a data-driven technique to solve the LOORP problem and provides the details on establishing a proper PE condition.

This work has been supported in part by the NSF under grant nos. EPCN-2210320, CPS-2227185, S&AS-1849198, CPS-1851588, and CPS-2227153.

[1]CAN Lab, New York University, Brooklyn, NY 11201 USA, sc8804@nyu.edu, zjiang@nyu.edu

[2]State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China, gaown@mail.neu.edu.cn

[3]The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, GA 30332-0150 USA, kyriakos@gatech.edu

Lastly, Section IV introduces 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 $v \in \mathbb{R}^n$, $\text{vecv}(v) = [v_1^2, v_1v_2, \cdots, v_1v_n, v_2^2, v_2v_3, \cdots, v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{(1/2)n(n+1)}$. $\mathbf{I}_n(\mathbf{0}_n)$ is the identity (zero) matrix of dimension n.

II. Problem Formulation and Background

A. Problem Formulation

Consider a class of discrete-time linear systems given as,

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{D}\mathbf{v}_k, \tag{1}$$

$$\mathbf{v}_{k+1} = \mathbf{E}\mathbf{v}_k, \tag{2}$$

$$\mathbf{e}_k = \mathbf{C}\mathbf{x}_k + \mathbf{J}\mathbf{u}_k + \mathbf{F}\mathbf{v}_k, \tag{3}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ is the state vector, $\mathbf{u}_k \in \mathbb{R}^m$ is the control input, $\mathbf{v}_k \in \mathbb{R}^q$ is the state of the exosystem (2), $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{r\times n}$, $\mathbf{D} \in \mathbb{R}^{n\times q}$, $\mathbf{E} \in \mathbb{R}^{q\times q}$, $\mathbf{F} \in \mathbb{R}^{r\times q}$, and $\mathbf{J} \in \mathbb{R}^{r\times m}$ are constant matrices, $\mathbf{d}_k = \mathbf{D}\mathbf{v}_k$ is the exogenous disturbance, $\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{J}\mathbf{u}_k$ is the output of the plant, $\mathbf{y}_{d_k} = -\mathbf{F}\mathbf{v}_k$ is the reference signal, and $\mathbf{e}_k \in \mathbb{R}^r$ is the tracking error. It is assumed that $\mathbf{v}_k$ is not measurable. Several other assumptions are as follows:

Assumption II.1. *The pair* $(\mathbf{A}, \mathbf{B})$ *is stabilizable.* □

Assumption II.2. $\text{rank}\begin{pmatrix}\begin{bmatrix}\mathbf{A}-\lambda\mathbf{I} & \mathbf{B}\\ \mathbf{C} & \mathbf{J}\end{bmatrix}\end{pmatrix} = n + r, \ \forall\lambda \in \sigma(\mathbf{E})$. □

Assumption II.3. *The minimal polynomial of* $\mathbf{E}$ *is known, which takes the form*

$$\alpha_m(s) = \prod_{i=1}^{N_1}(s-\lambda_i)^{a_i}\prod_{j=1}^{N_2}(s^2 - 2\mu_j s + \mu_j^2 + \omega_j^2)^{b_j}, \tag{4}$$

with degree $q_m \le q$*, where* a_i *and* b_j *are positive integers and* $\lambda_i, \mu_j, \omega_j \in \mathbb{R}$ *for* $i = 1, 2, ..., N_1, j = 1, 2, ..., N_2$. □

Remark 1. *Using Assumption II.3 one can always find a vector* $\mathbf{w}_k \in \mathbb{R}^{q_m}$ *and a matrix* $\hat{\mathbf{E}} \in \mathbb{R}^{q_m\times q_m}$ *such that:*

$$\mathbf{w}_{k+1} = \hat{\mathbf{E}}\mathbf{w}_k, \tag{5}$$

$$\mathbf{v}_k = \mathbf{G}\mathbf{w}_k, \forall k \ge 0, \tag{6}$$

with $\mathbf{G} \in \mathbb{R}^{q\times q_m}$ *an unknown constant matrix. Therefore, (1) and (3) are equivalent to:*

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \hat{\mathbf{D}}\mathbf{w}_k, \tag{7}$$

$$\mathbf{e}_k = \mathbf{C}\mathbf{x}_k + \mathbf{J}\mathbf{u}_k + \hat{\mathbf{F}}\mathbf{w}_k, \tag{8}$$

where $\hat{\mathbf{D}} = \mathbf{D}\mathbf{G}$ *and* $\hat{\mathbf{F}} = \mathbf{F}\mathbf{G}$. □

Remark 2. *Assumption II.2 guarantees the solvability of the regulator equations (10) and (11) for any* $\hat{\mathbf{F}}$, $\hat{\mathbf{D}}$. □

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

$$\mathbf{u}_k = -\mathbf{K}\mathbf{x}_k + \mathbf{L}\mathbf{w}_k, \tag{9}$$

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 (9) is globally exponentially stable at the origin, and
2) the tracking error $\mathbf{e}_k$ asymptotically converges to the origin.

Given that the designed controller is optimal with respect to a cost, the problem can be termed as a linear optimal output regulation problem (LOORP).

Theorem II.1. *([32]) Under Assumptions II.1 and II.3, choose* $\mathbf{K}$ *such that the closed-loop system is stable. The LORP is solvable by (9) if there exist* $\mathbf{X} \in \mathbb{R}^{n\times q_m}$ *and* $\mathbf{U} \in \mathbb{R}^{m\times q_m}$ *solutions to the following regulator equations:*

$$\mathbf{X}\hat{\mathbf{E}} = \mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{U} + \hat{\mathbf{D}}, \tag{10}$$

$$\mathbf{0} = \mathbf{C}\mathbf{X} + \mathbf{J}\mathbf{U} + \hat{\mathbf{F}}, \tag{11}$$

where the feedforward gain is given by

$$\mathbf{L} = \mathbf{U} + \mathbf{K}\mathbf{X}. \tag{12}$$

□

For any given initial conditions $\mathbf{x_0}$ and $\mathbf{w_0}$, if the controller given in (9) solves the LORP, one has $\lim_{k\to\infty}\mathbf{u}_k - \mathbf{U}\mathbf{w}_k = 0$ and $\lim_{k\to\infty}\mathbf{x}_k - \mathbf{X}\mathbf{w}_k = 0$. By solving the LOORP problem, we attempt to solve the problem of asymptotic tracking and disturbance rejection for discrete-time linear systems. Let $\mathbf{X}^\star$ and $\mathbf{U}^\star$ be the optimal solutions to the regulator equations (10) and (11) obtained by solving:

Problem II.1.

$$\min_{\mathbf{X},\mathbf{U}} \quad \text{Tr}\left(\mathbf{X}^\text{T}\bar{\mathbf{Q}}\mathbf{X} + \mathbf{U}^\text{T}\bar{\mathbf{R}}\mathbf{U}\right), \tag{13}$$

$$\text{subject to } (10) - (11),$$

where $\bar{\mathbf{Q}} = \bar{\mathbf{Q}}^\text{T} \succ 0$*, and* $\bar{\mathbf{R}} = \bar{\mathbf{R}}^\text{T} \succ 0$. □

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

$$\bar{\mathbf{x}}_{k+1} = \mathbf{A}\bar{\mathbf{x}}_k + \mathbf{B}\bar{\mathbf{u}}_k, \tag{14}$$

$$\mathbf{e}_k = \mathbf{C}\bar{\mathbf{x}}_k + \mathbf{J}\bar{\mathbf{u}}_k. \tag{15}$$

Problem II.2.

$$\min_{\bar{\mathbf{u}}} \quad J = \sum_{k=0}^{\infty}(\bar{\mathbf{x}}_k^\text{T}\mathbf{Q}\bar{\mathbf{x}}_k + \bar{\mathbf{u}}_k^\text{T}\mathbf{R}\bar{\mathbf{u}}_k), \tag{16}$$

$$\text{subject to } (14),$$

where $\mathbf{Q} = \mathbf{Q}^\text{T} \succeq 0$*,* $\mathbf{R} = \mathbf{R}^\text{T} \succ 0$*, and* $(\mathbf{A}, \sqrt{\mathbf{Q}})$ *is observable.* □

Remark 3. *After solving Problems II.1 and II.2, one can find the optimal controller* $\mathbf{u}_k^\star = -\mathbf{K}^\star\mathbf{x}_k + \mathbf{L}^\star\mathbf{w}_k$*. The design of the optimal feedback controller gain* $\mathbf{K}^\star$ *does not rely on the solution of the regulator equation* $\mathbf{X}^\star$ *and* $\mathbf{U}^\star$*. Thus, Problems II.1 and II.2 can be solved separately [2].* □

B. Model-based Approaches

1) Solution of Discrete-time LQR Problem: By solving the discrete-time LQR problem given in Problem II.2, one can obtain the optimal feedback gain $\mathbf{K}^\star$ as:

$$\mathbf{K}^\star = (\mathbf{R}+\mathbf{B}^{\mathrm{T}}\mathbf{P}^\star\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}^\star\mathbf{A}, \tag{17}$$

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

$$\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A}-\mathbf{P}+\mathbf{Q}-\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{B}(\mathbf{R}+\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{A}=0. \tag{18}$$

Note that, (18) is nonlinear in $\mathbf{P}$. It is usually difficult to solve especially for high-dimensional systems. A model-based PI technique to solve (18) [33] is shown in Algorithm 1 with $\mathbf{A}_j = \mathbf{A}-\mathbf{B}\mathbf{K}_j$.

Algorithm 1 Model-based PI

1: Select a stabilizing control policy $\mathbf{K}_0$ such that $\mathbf{A}-\mathbf{B}\mathbf{K}_0$ is Schur. Initialize $j \leftarrow 0$. Select a sufficiently small constant $\varepsilon > 0$.
2: **repeat**
3: Policy Evaluation:

$$\mathbf{A}_j^{\mathrm{T}}\mathbf{P}_j\mathbf{A}_j-\mathbf{P}_j+\mathbf{Q}+\mathbf{K}_j^{\mathrm{T}}\mathbf{R}\mathbf{K}_j=\mathbf{0}. \tag{19}$$

4: Policy Update:

$$\mathbf{K}_{j+1} = (\mathbf{R}+\mathbf{B}^{\mathrm{T}}\mathbf{P}_j\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{P}_j\mathbf{A}. \tag{20}$$

5: $j \leftarrow j+1$.
6: **until** $\|\mathbf{P}_j-\mathbf{P}_{j-1}\| < \varepsilon$.

2) Solution of Regulator Equations: We introduce a technique to solve the regulator equations (10) and (11) when the matrices $\mathbf{A},\mathbf{B}$, and $\hat{\mathbf{D}}$ are known. Define the Sylvester maps $S:\mathbb{R}^{n\times q_m}\to\mathbb{R}^{n\times q_m}$ and $\bar{S}:\mathbb{R}^{n\times q_m}\times\mathbb{R}^{m\times q_m}\to\mathbb{R}^{n\times q_m}$ as:

$$S(\mathbf{X}) = \mathbf{X}\hat{\mathbf{E}}-\mathbf{A}\mathbf{X}, \tag{21}$$

$$\bar{S}(\mathbf{X},\mathbf{U}) = \mathbf{X}\hat{\mathbf{E}}-\mathbf{A}\mathbf{X}-\mathbf{B}\mathbf{U}. \tag{22}$$

Select two constant matrices $\mathbf{X}_1$ and $\mathbf{U}_1$ such that $\mathbf{C}\mathbf{X}_1+\mathbf{J}\mathbf{U}_1+\hat{\mathbf{F}}=0$. Then select $\mathbf{X}_i$ and $\mathbf{U}_i$ for $i=2,3,\cdots,h+1$ such that all the vectors $\mathrm{vec}\left(\begin{bmatrix}\mathbf{X}_i\\ \mathbf{U}_i\end{bmatrix}\right)$ form a basis for $\ker(\mathbf{I}_{q_m}\otimes\begin{bmatrix}\mathbf{C} & \mathbf{J}\end{bmatrix})$, where $h=(n+m-r)q_m$ is the dimension of the null space of $(\mathbf{I}_{q_m}\otimes\begin{bmatrix}\mathbf{C} & \mathbf{J}\end{bmatrix})$. A general solution to (11) can be given as:

$$(\mathbf{X},\mathbf{U}) = (\mathbf{X}_1,\mathbf{U}_1)+\sum_{i=2}^{h+1}\alpha_i(\mathbf{X}_i,\mathbf{U}_i), \tag{23}$$

where $\alpha_i\in\mathbb{R}$. Then, (10) implies,

$$\bar{S}(\mathbf{X},\mathbf{U}) = \bar{S}(\mathbf{X}_1,\mathbf{U}_1)+\sum_{i=2}^{h+1}\alpha_i\bar{S}(\mathbf{X}_i,\mathbf{U}_i) = \hat{\mathbf{D}}. \tag{24}$$

Now, (23) and (24) can be written as:

$$\mathscr{A}\boldsymbol{\chi} = \mathbf{b}, \tag{25}$$

where

$$\mathscr{A} = \begin{bmatrix}\mathrm{vec}\left(\begin{bmatrix}\mathbf{X}_2\\ \mathbf{U}_2\end{bmatrix}\right) & \cdots & \mathrm{vec}\left(\begin{bmatrix}\mathbf{X}_{h+1}\\ \mathbf{U}_{h+1}\end{bmatrix}\right) & -\mathbf{I}\\ \mathrm{vec}(\bar{S}(\mathbf{X}_2,\mathbf{U}_2)) & \cdots & \mathrm{vec}(\bar{S}(\mathbf{X}_{h+1},\mathbf{U}_{h+1})) & \mathbf{0}\end{bmatrix},$$

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

$$\mathbf{b} = \begin{bmatrix}-\mathrm{vec}\left(\begin{bmatrix}\mathbf{X}_1\\ \mathbf{U}_1\end{bmatrix}\right)\\ \mathrm{vec}(-\bar{S}(\mathbf{X}_1,\mathbf{U}_1)+\hat{\mathbf{D}})\end{bmatrix}.$$

Following [24], (25) can be written as:

$$\begin{bmatrix}\bar{\mathscr{A}}_{11} & \bar{\mathscr{A}}_{12}\\ \bar{\mathscr{A}}_{21} & \bar{\mathscr{A}}_{22}\end{bmatrix}\boldsymbol{\chi} = \begin{bmatrix}\bar{b}_1\\ \bar{b}_2\end{bmatrix}, \tag{26}$$

where $\bar{\mathscr{A}}_{21}\in\mathbb{R}^{h\times h}$ is a nonsingular matrix. Then, the following result holds.

Lemma II.1. *([24]) A pair* $(\mathbf{X},\mathbf{U})$ *is a solution to the regulator equations if and only if it solves the following equation:*

$$\mathscr{M}\,\mathrm{vec}\left(\begin{bmatrix}\mathbf{X}\\ \mathbf{U}\end{bmatrix}\right) = \mathscr{N}, \tag{27}$$

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 II.1 can be reformulated as:

Problem II.3.

$$\min_{\mathbf{X},\mathbf{U}}\left(\begin{bmatrix}\mathrm{vec}(\mathbf{X})\\ \mathrm{vec}(\mathbf{U})\end{bmatrix}\right)^{\mathrm{T}}\begin{bmatrix}\mathbf{I}_{q_m}\otimes\bar{\mathbf{Q}} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{q_m}\otimes\bar{\mathbf{R}}\end{bmatrix}\left(\begin{bmatrix}\mathrm{vec}(\mathbf{X})\\ \mathrm{vec}(\mathbf{U})\end{bmatrix}\right), \tag{28}$$

subject to (27).

□

III. Data-driven Optimal Design

In this section, we develop an optimal data-driven technique to compute $\mathbf{P}^\star$ and $\mathbf{K}^\star$ to solve the discrete-time LQR problem in phase 1, and $\mathbf{X}^\star$ and $\mathbf{U}^\star$ that solves (10) and (11) in phase 2, when the matrices $\mathbf{A},\mathbf{B}$, and $\hat{\mathbf{D}}$ are unknown.

A. Phase 1: A Data-driven Solution

Define $\boldsymbol{\Pi}_i = \hat{\mathbf{D}}-S(\mathbf{X}_i)$ and consider,

$$\bar{\mathbf{x}}_{k,i} = \mathbf{x}_k-\mathbf{X}_i\mathbf{w}_k, i=0,1,\cdots,h+1, \tag{29}$$

where $\mathbf{X}_0=\mathbf{0}$. Then, from (29) and (7), we have:

$$\bar{\mathbf{x}}_{k+1,i} = \mathbf{A}\mathbf{x}_k+\mathbf{B}\mathbf{u}_k+\hat{\mathbf{D}}\mathbf{w}_k-\mathbf{X}_i\hat{\mathbf{E}}\mathbf{w}_k. \tag{30}$$

From (21), we have:

$$S(\mathbf{X}_i) = \mathbf{X}_i\hat{\mathbf{E}}-\mathbf{A}\mathbf{X}_i. \tag{31}$$

From (29), using (30) and (31), it holds that:

$$\bar{\mathbf{x}}_{k+1,i} = \mathbf{A}_j\bar{\mathbf{x}}_{k,i}+\mathbf{B}(\mathbf{u}_k+\mathbf{K}_j\bar{\mathbf{x}}_{k,i})+\boldsymbol{\Pi}_i\mathbf{w}_k. \tag{32}$$

Along the trajectories of (32), one can obtain that:

$$\bar{\mathbf{x}}_{k+1,i}^{\mathrm{T}}\mathbf{P}_j\bar{\mathbf{x}}_{k+1,i}-\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{P}_j\bar{\mathbf{x}}_{k,i} = \left[\mathbf{A}_j\bar{\mathbf{x}}_{k,i}+\mathbf{B}(\mathbf{u}_k+\mathbf{K}_j\bar{\mathbf{x}}_{k,i})+\boldsymbol{\Pi}_i\mathbf{w}_k\right]^{\mathrm{T}}\mathbf{P}_j\left[\mathbf{A}_j\bar{\mathbf{x}}_{k,i}+\mathbf{B}(\mathbf{u}_k+\mathbf{K}_j\bar{\mathbf{x}}_{k,i})+\boldsymbol{\Pi}_i\mathbf{w}_k\right]-\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{P}_j\bar{\mathbf{x}}_{k,i}. \tag{33}$$

Then, using (19) we have:

$$\bar{\mathbf{x}}_{k+1,i}^{\mathrm{T}}\mathbf{P}_j\bar{\mathbf{x}}_{k+1,i}-\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{P}_j\bar{\mathbf{x}}_{k,i}+\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{Q}_j\bar{\mathbf{x}}_{k,i}=2\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{\Gamma}_{1j}^{\mathrm{T}}\mathbf{u}_k+ 2\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{\Gamma}_{1j}^{\mathrm{T}}\mathbf{K}_j\bar{\mathbf{x}}_{k,i}-\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{K}_j^{\mathrm{T}}\mathbf{\Gamma}_{2j}\mathbf{K}_j\bar{\mathbf{x}}_{k,i}+\mathbf{u}_k^{\mathrm{T}}\mathbf{\Gamma}_{2j}\mathbf{u}_k+2\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\mathbf{\Theta}_{1ij}\mathbf{w}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{\Theta}_{2ij}\mathbf{w}_k+\mathbf{w}_k^{\mathrm{T}}\mathbf{\Theta}_{3ij}\mathbf{w}_k, \quad (34)$$

where $\mathbf{Q}_j=\mathbf{Q}+\mathbf{K}_j^{\mathrm{T}}\mathbf{R}\mathbf{K}_j$, $\mathbf{\Theta}_{1ij}=\mathbf{A}^{\mathrm{T}}\mathbf{P}_j\mathbf{\Pi}_i$, $\mathbf{\Theta}_{2ij}=\mathbf{B}^{\mathrm{T}}\mathbf{P}_j\mathbf{\Pi}_i$, $\mathbf{\Theta}_{3ij}=\mathbf{\Pi}_i^{\mathrm{T}}\mathbf{P}_j\mathbf{\Pi}_i$, $\mathbf{\Gamma}_{1j}=\mathbf{B}^{\mathrm{T}}\mathbf{P}_j\mathbf{A}$, $\mathbf{\Gamma}_{2j}=\mathbf{B}^{\mathrm{T}}\mathbf{P}_j\mathbf{B}$. Using the property of Kronecker product that $\mathrm{vec}(\mathbf{XYZ})=(\mathbf{Z}^T\otimes\mathbf{X})\mathrm{vec}(\mathbf{Y})$, we have:

$$\begin{aligned}&\big[(\bar{\mathbf{x}}_{k+1,i}^{\mathrm{T}}\otimes\bar{\mathbf{x}}_{k+1,i}^{\mathrm{T}})-(\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\otimes\bar{\mathbf{x}}_{k,i}^{\mathrm{T}})\big]\mathrm{vec}(\mathbf{P}_j)+(\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\otimes\bar{\mathbf{x}}_{k,i}^{\mathrm{T}})\mathrm{vec}(\mathbf{Q}_j)\\ &=\big[2(\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\otimes\mathbf{u}_k^{\mathrm{T}})+2(\bar{\mathbf{x}}_{k,i}^{\mathrm{T}}\otimes\bar{\mathbf{x}}_{k,i}^{\mathrm{T}})(\mathbf{I}_n\otimes\mathbf{K}_j^{\mathrm{T}})\big]\mathrm{vec}(\mathbf{\Gamma}_{1j})\\ &+\big[-(\mathbf{K}_j\bar{\mathbf{x}}_{k,i})^{\mathrm{T}}\otimes(\mathbf{K}_j\bar{\mathbf{x}}_{k,i})^{\mathrm{T}}+(\mathbf{u}_k^{\mathrm{T}}\otimes\mathbf{u}_k^{\mathrm{T}})\big]\mathrm{vec}(\mathbf{\Gamma}_{2j})\\ &+2(\mathbf{w}_k^{\mathrm{T}}\otimes\bar{\mathbf{x}}_{k,i}^{\mathrm{T}})\mathrm{vec}(\mathbf{\Theta}_{1ij})+2(\mathbf{w}_k^{\mathrm{T}}\otimes\mathbf{u}_k^{\mathrm{T}})\mathrm{vec}(\mathbf{\Theta}_{2ij})\\ &+(\mathbf{w}_k^{\mathrm{T}}\otimes\mathbf{w}_k^{\mathrm{T}})\mathrm{vec}(\mathbf{\Theta}_{3ij}).\end{aligned}$$

Now, by collecting data for the time sequence $k_0<k_1<\cdots<k_s$, we get

$$\mathbf{\Psi}_{1ij}\boldsymbol{\theta}_{1ij}=-\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}\mathrm{vec}(\mathbf{Q}_j), \quad (35)$$

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

$\boldsymbol{\theta}_{1ij}=\Big[\mathrm{vecs}(\mathbf{P}_j)^{\mathrm{T}},\mathrm{vec}(\mathbf{\Gamma}_{1j})^{\mathrm{T}},\mathrm{vecs}(\mathbf{\Gamma}_{2j})^{\mathrm{T}},\mathrm{vec}(\mathbf{\Theta}_{1ij})^{\mathrm{T}},$ $\mathrm{vec}(\mathbf{\Theta}_{2ij})^{\mathrm{T}},\mathrm{vecs}(\mathbf{\Theta}_{3ij})^{\mathrm{T}}\Big]^{\mathrm{T}}$,

$\mathbf{\Delta}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}=\Big[\mathrm{vecv}(\bar{\mathbf{x}}_{k_0+1,i})-\mathrm{vecv}(\bar{\mathbf{x}}_{k_0,i}),\cdots,\mathrm{vecv}(\bar{\mathbf{x}}_{k_s,i})-\mathrm{vecv}(\bar{\mathbf{x}}_{k_s-1,i})\Big]^{\mathrm{T}}$, $\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}=\Big[(\bar{\mathbf{x}}_{k_0,i}\otimes\bar{\mathbf{x}}_{k_0,i}),\cdots,(\bar{\mathbf{x}}_{k_s,i}\otimes\bar{\mathbf{x}}_{k_s,i})\Big]^{\mathrm{T}}$,

$\tilde{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}=\Big[\mathrm{vecv}(\mathbf{K}_j\bar{\mathbf{x}}_{k_0,i}),\cdots,\mathrm{vecv}(\mathbf{K}_j\bar{\mathbf{x}}_{k_s,i})\Big]^{\mathrm{T}}$,

$\mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}}=\Big[\bar{\mathbf{x}}_{k_0,i}\otimes\mathbf{u}_{k_0},\cdots,\bar{\mathbf{x}}_{k_s,i}\otimes\mathbf{u}_{k_s}\Big]^{\mathrm{T}}$,

$\mathbf{I}_{\mathbf{u},\mathbf{u}}=\Big[\mathrm{vecv}(\mathbf{u}_{k_0}),\cdots,\mathrm{vecv}(\mathbf{u}_{k_s})\Big]^{\mathrm{T}}$,

$\mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}=\Big[\mathbf{w}_{k_0}\otimes\bar{\mathbf{x}}_{k_0,i},\cdots,\mathbf{w}_{k_s}\otimes\bar{\mathbf{x}}_{k_s,i}\Big]^{\mathrm{T}}$,

$\mathbf{I}_{\mathbf{w},\mathbf{u}}=\Big[\mathbf{w}_{k_0}\otimes\mathbf{u}_{k_0},\cdots,\mathbf{w}_{k_s}\otimes\mathbf{u}_{k_s}\Big]^{\mathrm{T}}$,

$\mathbf{I}_{\mathbf{w},\mathbf{w}}=\Big[\mathrm{vecv}(\mathbf{w}_{k_0}),\cdots,\mathrm{vecv}(\mathbf{w}_{k_s})\Big]^{\mathrm{T}}$.

1) PE Condition: While collecting data for learning, it is a usual practice to incorporate a probing noise with the control input $\mathbf{u}_k=-\mathbf{K}_0\mathbf{x}_k+\mathbf{e}_k$ such that PE condition is satisfied [16]. Since the probing noise does not affect the exosystem, we 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:

$$\mathbf{w}_{k+1}=\hat{\mathbf{E}}\mathbf{w}_k=\begin{bmatrix}\cos(\theta)&-\sin(\theta)&0\\ \sin(\theta)&\cos(\theta)&0\\ 0&0&1\end{bmatrix}. \quad (36)$$

The state transition matrix can be obtained as:

$$\hat{\mathbf{E}}^k=\begin{bmatrix}\alpha&-\beta&0\\ \beta&\alpha&0\\ 0&0&1\end{bmatrix}, \quad (37)$$

where $\alpha=0.5[(c-ls)^k+(c+ls)^k]$, and $\beta=0.5[l(c-ls)^k-l(c+ls)^k]$, $s=\sin(\theta)$, $c=\cos(\theta)$, and $l=\sqrt{-1}$. Thus, the states of the exosystem have the following solutions:

$$w_{1,k}=\alpha w_{1,0}-\beta w_{2,0}, \quad (38)$$

$$w_{2,k}=\beta w_{1,0}+\alpha w_{2,0}, \quad (39)$$

$$w_{3,k}=w_{3,0}, \quad (40)$$

where $w_{i,0}$ for $i=1,2,3$ are the initial conditions. Now, the Kronecker product $\mathbf{w}_k^{\mathrm{T}}\otimes\mathbf{w}_k^{\mathrm{T}}$ has the unique components: $\mathrm{vecv}(\mathbf{w}_k)^{\mathrm{T}}=[w_{1,k}^2,\ w_{1,k}w_{2,k},\ w_{1,k}w_{3,k},\ w_{2,k}^2,\ w_{2,k}w_{3,k},\ w_{3,k}^2]$. Consider a constant $\gamma=\frac{w_{3,0}^2}{w_{1,0}^2+w_{2,0}^2}$. Now,

$$\gamma w_{1,k}^2+\gamma w_{2,k}^2=\gamma(\alpha^2+\beta^2)(w_{1,0}^2+w_{2,0}^2). \quad (41)$$

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 $\mathrm{vecv}(\mathbf{w}_k)$. This example shows that the matrix $\mathbf{I}_{\mathbf{w},\mathbf{w}}$ may not be full column rank, which makes the least squares problem in (35) non-unique. Thus to guarantee uniqueness we use $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ in (35), where $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ is constructed 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 $\mathrm{vecs}(\mathbf{\Theta}_{3ij})$ is also reduced.

Assumption III.1. *For $i=0,1,\cdots,h+1$ there exists a $s^\star\in\mathbb{Z}_+$ such that for all $s>s^\star$, and for any sequence $k_0<k_1<\cdots<k_s$:*

$$\mathrm{rank}([\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i},\mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}_i},\mathbf{I}_{\mathbf{u},\mathbf{u}},\mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i},\mathbf{I}_{\mathbf{w},\mathbf{u}},\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}])=\frac{n(n+1)}{2}+nm+\frac{m(m+1)}{2}+nq_m+mq_m+\frac{q_m(q_m+1)}{2}-N, \quad (42)$$

where N is the number of dependent columns of $\mathbf{I}_{\mathbf{w},\mathbf{w}}$. □

Remark 4. *A typical choice of $s^\star$ can be $s^\star\geq\frac{n(n+1)}{2}+nm+\frac{m(m+1)}{2}+nq_m+mq_m+\frac{q_m(q_m+1)}{2}$.* □

Proposition III.1. *Using $\bar{\mathbf{I}}_{\mathbf{w},\mathbf{w}}$ in (35) one can obtain:*

$$\bar{\mathbf{\Psi}}_{1ij}\bar{\boldsymbol{\theta}}_{1ij}=-\mathbf{I}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}\mathrm{vec}(\mathbf{Q}_j), \quad (43)$$

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

$\bar{\boldsymbol{\theta}}_{1ij}=\Big[\mathrm{vecs}(\mathbf{P}_j)^{\mathrm{T}},\mathrm{vec}(\mathbf{\Gamma}_{1j})^{\mathrm{T}},\mathrm{vecs}(\mathbf{\Gamma}_{2j})^{\mathrm{T}},\mathrm{vec}(\mathbf{\Theta}_{1ij})^{\mathrm{T}},$ $\mathrm{vec}(\mathbf{\Theta}_{2ij})^{\mathrm{T}},\mathrm{vec}(\bar{\mathbf{\Theta}}_{3ij})^{\mathrm{T}}\Big]^{\mathrm{T}}$. *Then, under Assumption III.1:*

(a) (43) has a unique solution.

(b) the sequences $\{\mathbf{P}_j\}_{j=0}^{\infty}$ and $\{\mathbf{K}_j\}_{j=0}^{\infty}$ obtained using Algorithm 2 converges to the optimal values $\mathbf{P}^\star$ and $\mathbf{K}^\star$, respectively.

Proof:

(a) Note that $\bar{\boldsymbol{\theta}}_{1ij}$ can be obtained from (43) using least squares and (42) guarantees that (43) can be uniquely solved (see [16], [24]).

(b) Given a stabilizing control gain $\mathbf{K}_j$, if $\mathbf{P}_j = \mathbf{P}_j^{\mathrm{T}}$ is the unique solution of (19), $\mathbf{K}_{j+1}$ is uniquely determined by $\mathbf{K}_{j+1} = (\mathbf{R}+\boldsymbol{\Gamma}_{2j})^{-1}\boldsymbol{\Gamma}_{1j}$. By (34), we know that $\mathbf{P}_j$, $\boldsymbol{\Gamma}_{1j}$, $\boldsymbol{\Gamma}_{2j}$, $\boldsymbol{\Theta}_{1ij}$, $\boldsymbol{\Theta}_{2ij}$, and $\bar{\boldsymbol{\Theta}}_{3ij}$ satisfy (43). Let, $\mathbf{P}$, $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, $\boldsymbol{\Theta}_{1i}$, $\boldsymbol{\Theta}_{2i}$, and $\bar{\boldsymbol{\Theta}}_{3i}$ of appropriate dimensions solve (43). Then, we have $\mathbf{P}_j = \mathbf{P}$, $\boldsymbol{\Gamma}_{1j} = \boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_{2j} = \boldsymbol{\Gamma}_2$, $\boldsymbol{\Theta}_{1ij} = \boldsymbol{\Theta}_{1i}$, $\boldsymbol{\Theta}_{2ij} = \boldsymbol{\Theta}_{2i}$, and $\bar{\boldsymbol{\Theta}}_{3ij} = \bar{\boldsymbol{\Theta}}_{3i}$. Then from part (a), it follows that $\mathbf{P}$, $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, $\boldsymbol{\Theta}_{1i}$, $\boldsymbol{\Theta}_{2i}$, and $\bar{\boldsymbol{\Theta}}_{3i}$ are unique. Thus, the PI in Algorithm 2 is same as (19) and (20). Thus, the theorem is proved by the equivalence of the two algorithms. ■

Algorithm 2 Phase-1 Learning

1: Compute matrices $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_{h+1}$.
2: Employ $\mathbf{u}_k = -\mathbf{K}_0\mathbf{x}_k + \eta_k$ as the input on the time horizon $[k_0, k_s]$, where $\mathbf{K}_0$ is initial stabilizing gain and η_k is the exploration noise.
3: For $i = 0, 1, \cdots, h+1$, compute $\bar{\boldsymbol{\Psi}}_{1ij}$ until the rank condition in (42) is satisfied. Let $i = 0, j = 0$.
4: Solve for $\bar{\boldsymbol{\theta}}_{1ij}$ from (43). Then, $\mathbf{K}_{j+1} = (\mathbf{R}+\boldsymbol{\Gamma}_{2j})^{-1}\boldsymbol{\Gamma}_{1j}$.
5: Let $j \leftarrow j+1$ and repeat Step 4. until $\|\mathbf{P}_j - \mathbf{P}_{j-1}\| \le \varepsilon_0$ for $j \ge 1$, where $\varepsilon_0 > 0$ is a predefined small threshold.

B. Phase 2: Data-driven Solution to the Regulator Equations

Using (32) and the property $\mathrm{vec}(\mathbf{XYZ}) = (\mathbf{Z}^T \otimes \mathbf{X})\mathrm{vec}(\mathbf{Y})$, one can obtain:

$$(\bar{\mathbf{x}}_{k,i}^{\mathrm{T}} \otimes \bar{\mathbf{x}}_{k+1,i}^{\mathrm{T}})\mathrm{vec}(\mathbf{P}_{j\star}) = (\bar{\mathbf{x}}_{k,i}^{\mathrm{T}} \otimes \bar{\mathbf{x}}_{k,i}^{\mathrm{T}})\mathrm{vec}(\mathbf{A}^{\mathrm{T}}\mathbf{P}_{j\star}) + (\bar{\mathbf{x}}_{k,i}^{\mathrm{T}} \otimes \mathbf{u}_k^{\mathrm{T}})\mathrm{vec}(\mathbf{B}^{\mathrm{T}}\mathbf{P}_{j\star}) + (\mathbf{w}_k^{\mathrm{T}} \otimes \bar{\mathbf{x}}_{k,i}^{\mathrm{T}})\mathrm{vec}(\mathbf{P}_{j\star}\boldsymbol{\Pi}_i), \quad (44)$$

where $\mathbf{P}_{j\star}$ is the approximated solution of the Riccati equation obtained from Algorithm 2. Using the data collected from Phase 1, one can obtain:

$$\boldsymbol{\Psi}_{2i}\boldsymbol{\theta}_{2i} = \boldsymbol{\Lambda}_i \mathrm{vec}(\mathbf{P}_{j\star}), \quad (45)$$

where, $\boldsymbol{\Psi}_{2i} = \left[\frac{1}{2}\bar{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i}, \mathbf{I}_{\bar{\mathbf{x}}_i,\mathbf{u}}, \mathbf{I}_{\mathbf{w},\bar{\mathbf{x}}_i}\right]$, $\boldsymbol{\theta}_{2i} = \Big[\mathrm{vecs}(\mathbf{A}^{\mathrm{T}}\mathbf{P}_{j\star} + \mathbf{P}_{j\star}\mathbf{A})^{\mathrm{T}}, \mathrm{vec}(\mathbf{B}^{\mathrm{T}}\mathbf{P}_{j\star})^{\mathrm{T}}, \mathrm{vec}(\mathbf{P}_{j\star}\boldsymbol{\Pi}_i)^{\mathrm{T}}\Big]^{\mathrm{T}}$, $\boldsymbol{\Lambda}_i = \Big[\bar{\mathbf{x}}_{k_0,i} \otimes \bar{\mathbf{x}}_{k_1,i}, \cdots, \bar{\mathbf{x}}_{k_{s-1},i} \otimes \bar{\mathbf{x}}_{k_s,i}\Big]^{\mathrm{T}}$, $\bar{\mathbf{I}}_{\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_i} = \Big[\mathrm{vecv}(\bar{\mathbf{x}}_{k_0,i}), \cdots, \mathrm{vecv}(\bar{\mathbf{x}}_{k_s,i})\Big]^{\mathrm{T}}$.

Remark 5. *Satisfying Assumption III.1 also implies that (45) has a unique solution for each $i = 0, \cdots, h+1$.* □

Algorithm 3 Phase-2 Learning

1: Set $\mathbf{X}_0 = \mathbf{0}$, $\mathbf{U}_0 = \mathbf{0}$. Then, $\boldsymbol{\Pi}_0 = \hat{\mathbf{D}}$.
2: For $i = 0$, obtain $\mathbf{B}$ and $\hat{\mathbf{D}}$ by solving (45).
3: Let $i \leftarrow i+1$, solve for $S(\mathbf{X}_i)$ from (45) until $i = h+1$.
4: Compute $\bar{S}(\mathbf{X}_i, \mathbf{U}_i) = S(\mathbf{X}_i) - \mathbf{B}\mathbf{U}_i$ for $i = 0, 1, \cdots, h+1$.
5: Obtain $\mathbf{X}^\star$ and $\mathbf{U}^\star$ by solving Problem II.3.
6: Obtain the feedforward gain as $\mathbf{L}_{j\star} = \mathbf{U}^\star + \mathbf{K}_{j\star}\mathbf{X}^\star$.

Remark 6. *From Phase 2, it is clear that the solution obtained for the regulator equations using the learning-based method is consistent with that of the model-based method.* □

Remark 7. *It is not difficult to show that the discrete-time linear system (1)-(3), under the approximate optimal control policy $\mathbf{u}_k^\star = -\mathbf{K}_{j\star}\mathbf{x}_k + \mathbf{L}_{j\star}\mathbf{w}_k$ has the following properties [24]:*

- *The closed loop system with $\mathbf{u}_k^\star$ is exponentially stable at the origin.*
- *The tracking error $\mathbf{e}_k \to 0$ as $k \to \infty$.* □

IV. Simulation Results and Discussion

In this section, we show the efficacy of the proposed algorithm by applying it to an inverted pendulum on a cart. Consider the following discrete-time system:

$$\mathbf{x}_{k+1} = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1-\frac{bT}{M} & -\frac{mgT}{M} & 0 \\ 0 & 0 & 1 & T \\ 0 & \frac{bT}{lM} & \frac{(M+m)gT}{lM} & 1 \end{bmatrix}\mathbf{x}_k + \begin{bmatrix} 0 \\ \frac{T}{M} \\ 0 \\ -\frac{T}{lM} \end{bmatrix}\mathbf{u}_k + \mathbf{d}_k, \quad (46)$$

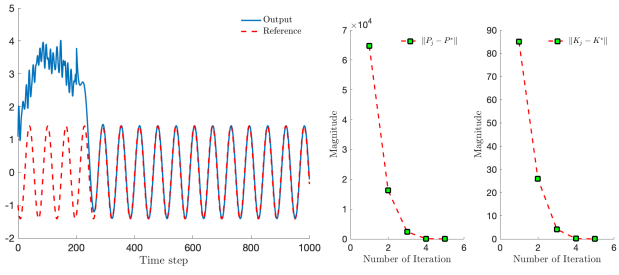
$$\mathbf{w}_{k+1} = \begin{bmatrix} \cos(0.1) & -\sin(0.1) & 0 \\ \sin(0.1) & \cos(0.1) & 0 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{w}_k, \quad (47)$$

$$\mathbf{e}_k = [1\ 0\ 0\ 0]\mathbf{x}_k + \mathbf{u}_k + [-1\ 0\ 0]\mathbf{w}_k. \quad (48)$$

For the meaning and value of the parameters refer to [34]. The upper 2×2 subsystem in (47) is used to generate the reference signal and the lower 1×1 subsystem in (47) is used to generate the disturbance. The initial conditions are given as $\mathbf{x}_0 = [0.5,\ 0,\ 0,\ 0]$, and $\mathbf{w}_0 = [-1,\ 0,\ 1]$. The system matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ are considered unknown. The weight matrices are chosen as $\mathbf{Q} = \mathbf{I}_4$, and $\mathbf{R} = 1$. The exploration noise in Algorithm 2 is chosen as the summation of sinusoidal waves with different frequencies. Using the learning data, Algorithm 2 converges with a tolerance of $\varepsilon_0 = 0.5$ to a neighborhood of the optimal values $\mathbf{P}^\star$ and $\mathbf{K}^\star$ in five iterations as shown in Fig. 1b. The optimal controller gain $\mathbf{K}^\star$ and the controller gain obtained from Algorithm 2 are given as:

$$\mathbf{K}^\star = [-0.9464, -25.7615, -73.3595, -13.2973], \quad (49)$$

$$\mathbf{K}_{5\star} = [-0.9464, -25.7615, -73.3597, -13.2974]. \quad (50)$$



(a) (b)

Fig. 1: (a) Output and reference trajectories, (b) Convergence of $\mathbf{P}_j$ to $\mathbf{P}^\star$ and $\mathbf{K}_j$ to $\mathbf{K}^\star$.

The optimal feedforward gain $\mathbf{L}^\star$ and the feedforward gain obtained from Algorithm 3 are given as:

$$\mathbf{L}^\star = [0.5244, 0.8036, 2.8678], \tag{51}$$

$$\mathbf{L}_{5\star} = [0.5244, 0.8036, 2.8679]. \tag{52}$$

V. Conclusion

The paper aims at solving the challenge of adaptive optimal output regulation in situations where the plant model of a discrete-time system is completely unknown. Our study reveals the fact that the design of the rank condition of the data matrix utilized in the PI algorithm is a critical factor for ensuring the convergence and uniqueness properties of the PI algorithm. This is because some of the columns of the data matrix are formed solely from the states of the exosystem, which are not affected by probing noise during data collection. To align model-based solutions of regulator equations with model-free techniques, existing methods may necessitate the state/plant matrix to be invertible. We address this issue by introducing a novel reformulation of the problem that removes the need for the invertibility assumption on the plant matrix. Also, we have illustrated the data-driven approach to solve the regulator equations in presence of feedthrough term in plant output. Finally, we provide numerical simulations to demonstrate the effectiveness of the proposed methodology. Future work will focus on extending the results to nonlinear systems.

Acknowledgement

We would like to thank Prof. Frank L. Lewis and Leilei Cui for their valuable suggestions.

References

[1] C. Bonivento, L. Marconi, and R. Zanasi, "Output regulation of nonlinear systems by sliding mode," *Automatica*, vol. 37, no. 4, pp. 535–542, 2001.

[2] J. Huang, *Nonlinear output regulation: theory and applications*. SIAM, 2004.

[3] H. L. Trentelman, A. A. Stoorvogel, M. Hautus, and L. Dewell, "Control theory for linear systems," *Appl. Mech. Rev.*, vol. 55, no. 5, pp. B87–B87, 2002.

[4] A. J. Krener, "The construction of optimal linear and nonlinear regulators," in *Systems, Models and Feedback: Theory and Applications*. Springer, 1992, pp. 301–322.

[5] A. Saberi, A. A. Stoorvogel, P. Sannuti, and G. Shi, "On optimal output regulation for linear systems," *International Journal of Control*, vol. 76, no. 4, pp. 319–333, 2003.

[6] W. Liu and J. Huang, "Output regulation of linear systems via sampled-data control," *Automatica*, vol. 113, p. 108684, 2020.

[7] Y. Yan and J. Huang, "Cooperative output regulation of discrete-time linear time-delay multi-agent systems," *IET Control Theory & Applications*, vol. 10, no. 16, pp. 2019–2026, 2016.

[8] R. Mantri, A. Saberi, Z. Lin, and A. A. Stoorvogel, "Output regulation for linear discrete-time systems subject to input saturation," *International Journal of Robust and Nonlinear Control*, vol. 7, no. 11, pp. 1003–1021, 1997.

[9] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[10] D. Bertsekas, *Dynamic programming and optimal control*. Athena scientific, 2012, vol. 1.

[11] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[12] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[13] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

[14] P. Werbos, "Beyond regression:" new tools for prediction and analysis in the behavioral sciences," *Ph. D. dissertation, Harvard University*, 1974.

[15] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

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

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

[18] Y. Jiang and Z.-P. Jiang, "Adaptive dynamic programming as a theory of sensorimotor control," *Biological cybernetics*, vol. 108, no. 4, pp. 459–473, 2014.

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

[20] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[21] 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.

[22] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 142–160, 2020.

[23] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[24] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 4164–4169, 2016.

[25] ——, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2614–2624, 2017.

[26] Q. Wei, X. Wang, Y. Liu, and G. Xiong, "Data-driven adaptive-critic optimal output regulation towards water level control of boiler-turbine systems," *Expert Systems with Applications*, vol. 207, p. 117883, 2022.

[27] L. Cui, K. Ozbay, and Z.-P. Jiang, "Combined longitudinal and lateral control of autonomous vehicles based on reinforcement learning," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 1929–1934.

[28] W. Gao, Y. Liu, A. Odekunle, 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.

[29] J. Li, Z. Xiao, P. Li, and J. Cao, "Robust optimal tracking control for multiplayer systems by off-policy q-learning approach," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 1, pp. 87–106, 2021.

[30] 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 Transactions on Cybernetics*, vol. 50, no. 7, pp. 3147–3156, 2019.

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

[32] B. A. Francis, "The linear multivariable regulator problem," *SIAM Journal on Control and Optimization*, vol. 15, no. 3, pp. 486–505, 1977.

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

[34] R. Gurumoorthy and S. Sanders, "Controlling non-minimum phase nonlinear systems-the inverted pendulum on a cart example," in *1993 American Control Conference*. IEEE, 1993, pp. 680–685.