Spring 2019

Reinforcement Learning, Intelligent Control and their Applications in Connected and Autonomous Vehicles

Adedapo O. Odekunle

REINFORCEMENT LEARNING, INTELLIGENT CONTROL AND THEIR APPLICATIONS IN CONNECTED AND AUTONOMOUS VEHICLES

by

ADEDAPO ODEKUNLE

(Under the Direction of Weinan Gao)

ABSTRACT

Reinforcement learning (RL) has attracted large attention over the past few years. Recently, we developed a data-driven algorithm to solve predictive cruise control (PCC) and games output regulation problems. This work integrates our recent contributions to the application of RL in game theory, output regulation problems, robust control, small-gain theory and PCC. The algorithm was developed for H_∞ adaptive optimal output regulation of uncertain linear systems, and uncertain partially linear systems to reject disturbance and also force the output of the systems to asymptotically track a reference. In the PCC problem, we determined the reference velocity for each autonomous vehicle in the platoon using the traffic information broadcasted from the lights to reduce the vehicles' trip time. Then we employed the algorithm to design an approximate optimal controller for the vehicles. This controller is able to regulate the headway, velocity and acceleration of each vehicle to the desired values. Simulation results validate the effectiveness of the algorithms.

INDEX WORDS: Reinforcement learning, Adaptive control, Game theory, Small-gain theory, Robust control, Predictive cruise control

REINFORCEMENT LEARNING, INTELLIGENT CONTROL AND THEIR APPLICATIONS IN CONNECTED AND AUTONOMOUS VEHICLES

by

ADEDAPO ODEKUNLE

B.S., University of Lagos, Nigeria 2010

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

REINFORCEMENT LEARNING, INTELLIGENT CONTROL AND THEIR APPLICATIONS IN CONNECTED AND AUTONOMOUS VEHICLES

by

ADEDAPO ODEKUNLE

Major Professor:	Weinan Gao
Committee:	Masoud Davari
	Seungmo Kim
	Fernando Rios-Guitierrez

Electronic Version Approved:
May 2019

DEDICATION

To my family and friends, for their love and support.

ACKNOWLEDGMENTS

I owe a debt of gratitutde to my adviser, Professor Weinan Gao, for the great mentorship and the invaluable opportunity to study under his guidance at the Georgia Southern University. His experience-based advice, unique perspective, teaching style and effective advising approach is much appreciated.

I would like to thank my friends and family for their love, support and encouragement throughout this program.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF SYMBOLS

- $\mathbb{R}$ Set of real numbers.
- $\mathbb{R}_+$ Set of all non-negative real numbers.
- $\mathbb{Z}_+$ Set of non-negative integers
- $\mathbb{C}^-$ Open left-half complex plane
- $|\cdot|$ The Euclidean norm for vectors, or the induced matrix norm for matrices.
- $\|\cdot\|$ For any piecewise continous function $u : \mathbb{R}_+ \to \mathbb{R}^m$, $\|u\|$=sup$\{|u(t), t \geq 0|\}$
- $\otimes$ Kronecker product operator
- $\mathrm{vec}(\cdot)$ $\mathrm{vec}(A)$ is the mn-vector formed by stacking the columns of $A \in \mathbb{R}^{n\times m}$ on top of one another, or more precisely, starting with the first column and ending with the last column of A.
- $\ker(A)$ Kernel of A.
- TrA Trace of A.
- $\sigma(A)$ Complex spectrum of A
- $\mathrm{vecs}(\cdot)$ $\mathrm{vecs}(C) = [c_{11}, 2c_{12}, \cdots, 2c_{1m}, c_{22}, 2c_{23}, \cdots, 2c_{m-1,m}, c_{mm}]^T$.
- $\lambda_{\min}(C)$ Minimum eigenvalue of C.
- $\mathrm{vecv}(\cdot)$ $\mathrm{vecv}(v) = [v_1^2, v_1v_2, \cdots, v_1v_n, v_2^2, v_2v_3, \cdots, v_{n-1}v_n, v_n^2]^T$.

CHAPTER 1

INTRODUCTION

1.1 Reinforcement Learning

Reinforcement learning (RL) was first attributed to the learning behavior of human beings and other higher animals (Sutton & Barto, 1998). Looking at the nature of learning; we learn by interacting with our environment with no explicit teacher. However, there is a direct connection with the learning environment which produces a plethora of information about cause and effect, consequences of actions, and what to do to achieve some set goals. RL involves how intelligent agents modify their action for a better interaction with an uncertain environment to minimize some cost functional or maximize some accumulated reward. RL differs from supervised learning in a way that it uses trial and error method for its modification unlike supervised learning where training data are used. Trial and error search and delayed reward are two important features of RL (Sutton & Barto, 1998; Watkins, 1989). RL is defined as learning what to do and how to map situations to actions so as to maximize a numerical reward signal or minimize cost functional. RL refers to an actor or an agent that interacts with its environment and modifies its actions, or control policies based on stimuli received in response to its actions hence, it is an action based learning (Lewis & Vrabie, 2009). This implies a cause and effect relationship between actions and reward or cost. The actor will be able to differentiate rewards and lack of reward or cost during the learning process. The RL algorithms are designed based on the idea that successful control decisions are remembered by a reinforcement signal to increase the chances of their re-usability. RL is connected from a theoretical point of view with direct and indirect adaptive optimal control methods. A simple class of reinforcement learning methods is modeled using the Actor-Critic structure (Barto, Sutton, & Anderson, 1983) as depicted in Fig. 1.1. The actor component applies the control policy

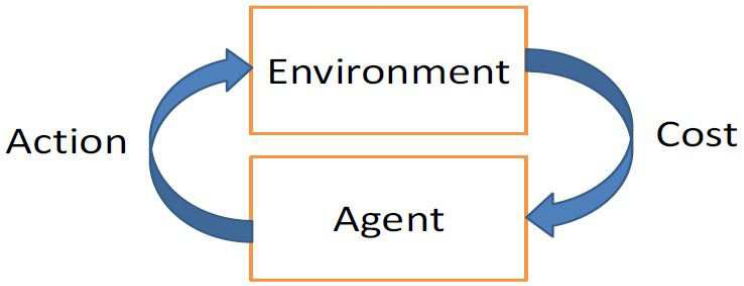

Figure 1.1: Simple Illustration of the RL

(action) to the environment, while the critic assesses the value of the control policy applied to the environment. An improvement on the previous value is sought to obtain a new policy by modifying or improving the action based on the assessment of the previous value. RL formed an important branch of machine learning theory and has been integrated into computational intelligence, computer science and control systems engineering literature as an effective way to study artificial intelligence (Mendel & McLaren, 1970; Minsky, 1961; Waltz & Fu, 1965; Sutton & Barto, 1998). This has brought about a lot of contributions to control engineering (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Odekunle, Chen, & Jiang, 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Odekunle, Gao, Anayor, Wang, & Chen, 2018; Gao & Jiang, 2016a; Vamvoudakis, 2014; Fan & Yang, 2016; Wang, Liu, Li, Luo, & Ma, 2016; Y. Jiang, Fan, Chai, Li, & Lewis, 2017).

1.1.1 The Development of Reinforcement Learning

The development of RL has two major phases. Firstly, RL was studied in computer science and operation research. In these fields, they employed both the policy iteration and value iteration (Sutton, 1990). Temporal difference methods were also integrated into RL . The well-known Q-learning method proposed by Watkins was also studied as a tool in

RL (Watkins, 1989). This Q-learning has a lot of similarities with the action-dependent Heuristic Dynamic Programming scheme proposed by Werbos (Werbos, 1989). Similar research work involving the framework of Markov decision processes are generally discrete in time and state-space (Sutton & Barto, 1998; Sutton, 1990; Barto et al., 1983; Waltz & Fu, 1965; Minsky, 1961). The second phase involves the design of real-time controllers, analysis techniques that yield guaranteed provable performance, stability and safety margins for dynamic systems (linear and nonlinear). The integration of stability theory and RL was introduced by Lewis (Lewis & Vrabie, 2009). This has a great advantage in obtaining an optimal control strategy iteratively by using online information without the need to solve algebraic Riccati equation (ARE) or the Hamilton-Jacobi-Bellman (HJB) equation for the linear and nonlinear systems respectively. This approach does not require having a full or partial knowledge of what the system dynamics are (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Odekunle, et al., 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Odekunle et al., 2018; Gao & Jiang, 2016a; Vamvoudakis, 2014; Fan & Yang, 2016; Wang et al., 2016). This data-driven approach is applicable to both the discrete time (DT) (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Lewis & Vrabie, 2009) and the continuous time (CT) systems (Gao, Odekunle, et al., 2018; Gao & Jiang, 2016a; Vamvoudakis, 2014; Fan & Yang, 2016; Wang et al., 2016).

The higher animals' intelligence is a very peculiar motivation to develop self-adaptive systems with a high level of intelligence. With the research trend in the modeling of human brain and neural system for complex engineering systems, there has been good success recorded (Y. Jiang et al., 2017; Mu, Ni, Sun, & He, 2017; Wang et al., 2016; H. Zhang, Cui, Zhang, & Luo, 2011; Vamvoudakis & Lewis, 2012; Al-Tamimi, Lewis, & Abu-Khalaf, 2007; Lewis & Vrabie, 2009), but still there are a lot of open issues in developing a perfect human intelligence-like systems. In other words, a clear understanding of the human intelligence and its extension to self-adaptive systems to perfectly mimic a level of intelligence

is still a great challenge to the academia. The major challenge is in designing intelligent systems with the capacity of "learning optimization" and "learning prediction" over time. In this work, data-driven algorithms via reinforcement learning is developed for the predictive cruise control of connected and autonomous vehicles, adaptive optimal control for the output regulation of uncertain linear systems and a robust adaptive optimal control for the output regulation of uncertain partially linear systems.

Generally, developing a control strategy for continuous time (CT) linear systems is difficult because of the need to solve the algebraic Riccati equation (ARE). Fortunately, RL approach gives us an opportunity to obtain the approximate solutions to the ARE without a priori knowledge of the system dynamics while the system stability is maintained (Werbos, 2009; Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Odekunle, et al., 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Odekunle et al., 2018; Gao & Jiang, 2016a). Recently, the RL approach in obtaining an optimal controller for intelligent systems has been gaining attention in the control engineering research circle and real-world applications (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Odekunle, et al., 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Odekunle et al., 2018; Gao & Jiang, 2016a). The application of RL for the predictive cruise control of connected and autonomous vehicles is still an open issue to the best of our knowledge. Also, much work has been done in its application in output regulation problems (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao, Liu, Odekunle, Jiang, et al., 2018; Gao & Jiang, 2016a), but none has extended to the zero-sum two game players. In this thesis, the detail solutions to this aforementioned problems is provided.

1.2 THESIS ORGANIZATION

The rest of this thesis includes two parts: Algorithms development and applications. We focus on output regulation problems and the application of RL in intelligent transportation systems.

Chapter 2 provides an RL based algorithms for solving the game output regulation problems and also analysis of its convergence and tracking ability are given. Simulation results are provided for methodology validation.

Chapter 3 gives a more general result for the application of RL in game output regulation problems by extending it to uncertain partially linear systems. A global robust optimal controller was designed. Also, nonlinear small-gain theory was applied to show the input-to-output stability for the closed-loop system. Simulation results were provided to validate the efficiency of the proposed methodology.

Chapter 4 studies the application of RL in the predictive cruise control of connected and autonomous vehicles. First, reference velocity is determined for each vehicle in the platoon. Second, the data-driven algorithm was developed to approximate the optimal control gains of a desired distributed controller. The obtained controller is able to regulate the headway, velocity and acceleration of each vehicle in an optimal sense.

Chapter 5 concludes the thesis.

1.3 PUBLICATIONS

The following papers have been published submitted during the course of my research work under Dr. Weinan Gao:

1. 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, 2018 16(5), 2273-2281.

2. W, Gao.; Y, Liu.; A, Odekunle.; Y, Yu.; Z.P, Jiang.; Y, Yu.; and P, Lu.; Cooperative and Adaptive Optimal Output Regulation of Discrete-time Multi-Agent Systems Using Reinforcement Learning, IEEE Conference on Real-time Computing and Robotics, 2018 348-353.

3. W, Gao.; A, Odekunle.; Y, Chen.; and Z.P, Jiang.; Predictive Cruise Control of Connected and Autonomous Vehicles via Reinforcement Learning, IET Control Theory and Applications, 2018 8pp.

4. A, Odekunle.; W, Gao.; C, Anayor.; X, Wang.; and Y, Chen.; Predictive Cruise Control of Connected and Autonomous Vehicles, IEEE Southeast Conference, 2018 1-3.

5. C, Anayor.; W, Gao.; and A, Odekunle.; Cooperative Adaptive Cruise Control of a Mixture of Human-driven and Autonomous Vehicles, IEEE Southeast Conference, 2018 1-6.

6. A, Odekunle.; W, Gao.; M, Davari.; Z.P, Jiang; Adaptive Optimal Output Regulation of Uncertain Linear Systems with Differential Games, Automatica (under revision).

7. A, Odekunle.; W, Gao; Data-Driven Global Robust Optimal Output Regulation of Uncertain Partially Linear Systems, IEEE/CAA Journal of Automatic Sinica (under review)

CHAPTER 2

DATA-DRIVEN H_∞ ADAPTIVE OPTIMAL OUTPUT REGULATION OF UNCERTAIN LINEAR SYSTEMS

2.1 ABSTRACT

The H_∞ output regulation problem, or game output regulation problem (GORP), is mainly concerned with the design of controllers to achieve asymptotic tracking while rejecting both modeled and unmodeled disturbances. With uncertain matrices in the state equation, this paper develops a novel data-driven adaptive optimal control approach solving the GORP of a class of continuous-time linear systems. A key strategy is to combine for the first time techniques from reinforcement learning (RL), H_∞ optimal control, and output regulation for data-driven control design. Different from the previous work in the present literature of the adaptive optimal output regulation, the feedforward matrix of the controlled plant is considered nontrivial. Theoretical analysis and simulation results demonstrate the efficacy of the developed data-driven control approach.

2.2 INTRODUCTION

Output regulation theory deals with the problems of designing a feedback controller to reject nonvanishing disturbances while forcing the output of a dynamical system to track a desired trajectory (Isidori, Marconi, & Serrani, 2003; Huang, 2004; Trentelman, Stoorvogel, & Hautus, 2002; Bonivento, Marconi, & Zanasi, 2001; Meng, Yang, Dimarogonas, & Johansson, 2015; Su & Huang, 2015; Serrani, Isidori, & Marconi, 2001) and many references therein—which have established the importance of the output regulation theory and its relevance to many real-world applications. Through minimizing a given cost function, the optimal control and the output regulation theories have been combined to design optimal output regulators (Saberi, Stoorvogel, Sannuti, & Shi, 2003; Krener, 1992). In the

framework of traditional output regulation, both the disturbance and the reference are generated by an autonomous system, named exosystem—wherein the unmodeled disturbance is usually not considered. Yaghmaie established a more practical case, i.e., H_∞ output regulation problem, in the presence of modeled and unmodeled disturbances (Yaghmaie, Movric, Lewis, Su, & Sebek, 2018). The H_∞ optimal control and output regulation techniques are efficiently combined for robust model-based control design.

The H_∞ optimal control is related to zero-sum differential games with two players; the controller (player 1) and the unmodeled disturbance (player 2) which are the minimizing and the maximizing players, respectively (Başar & Bernhard, 2008). Hence the H_∞ output regulation problem can also be called a game output regulation problem (GORP). Notice that most of the existing control solutions to output regulation problems and GORP are model-based, which means that an accurate knowledge of system model is absolutely needed. Since identifying a perfect system model is often time-consuming and costly, and involves modeling errors, it is imperative to develop data-driven controllers without relying on the system dynamics.

The terminology "data-driven" is originally from computer science, which has been in the vocabulary of the control community recently. Reinforcement learning (RL) is a practically sound data-driven adaptive optimal control technique. The RL is able to be used in learning the optimal control policy and value function based on the online state and input data, instead of system dynamics (Fan & Yang, 2016; Gao & Jiang, 2016b; Y. Jiang et al., 2017; Mu et al., 2017; Wang et al., 2016; H. Zhang et al., 2011; Rizvi & Lin, 2017). The extension to RL-based control design with differential games has been extensively studied as well (Vrabie, Vamvoudakis, & Lewis, 2013; Vamvoudakis & Lewis, 2012; Modares, Lewis, & Jiang, 2015; Wu & Luo, 2012; Al-Tamimi et al., 2007; Li, Liu, & Wang, 2014). However, most of these extensions focus on the adaptive optimal stabilization or tracking control. It is still an open problem to employ RL in developing data-driven

adaptive optimal control solutions to GORP. In our previous work, we have developed an adaptive optimal control approach to the output regulation problems based on RL (Gao & Jiang, 2016a). The solutions to both algebraic Riccati equation and regulator equation are iteratively approximated using online state and input data. The generalization to nonlinear systems and a class of multi-agent systems has been studied (Gao, Liu, Odekunle, Yu, & Lu, 2018; Gao & Jiang, 2018).

In this paper, we propose a novel data-driven adaptive optimal control approach to GORP. All of the matrices in the state equation are assumed unknown. The optimal feedforward and feedback control gains are approximated by collected real-time data. As the first contribution, this paper is the first time taking advantage of techniques from three separately studied areas: RL, H_∞ optimal control and output regulation. As the second contribution, this paper leverages RL to achieve asymptotic tracking while rejecting both modeled and unmodeled disturbances. The third contribution is that we consider the challenging and practical case that the feedforward matrix of the controlled plant is nontrivial, which is different from our previous work (Gao & Jiang, 2016b). This challenge is addressed by proposing a novel algorithm solving the regulator equations in the presence of the feedforward matrix. Interestingly, all the information needed by this methodology can be computed by online data, instead of the state and input matrices. Therefore, this methodology can be well applied as a part of proposed data-driven control solution to GORP.

The remainder of this note is organized as follows: In Section 2.3, we will formulate the GORP, and present basic results in output regulation, H_∞ optimal control, and game algebraic Riccati equation (GARE). In Section 2.4, we first propose a solution to regulator equations with known parameters. Then, an RL-based algorithm for solving the GORP will be developed and the convergence and tracking ability analyses will also be given therein. To ascertain the validity of the proposed algorithm in this paper, an illustrative example is examined in Section 2.5. Finally, conclusions and future work are contained in Section 2.6.

2.3 PROBLEM FORMULATION AND PRELIMINARIES

In this section, we formulate the zero-sum game output regulation problem (GORP) with two players. Then, some fundamentals of optimal control theory and a policy iteration technique to solve the corresponding game algebraic Riccati equation (GARE) and regulator equation will be reviewed.

2.3.1 PROBLEM FORMULATION

Consider a class of linear continuous-time systems described by

$$\dot{x} = Ax + B_1 u + B_2 \omega + Gv, \tag{2.1}$$

$$\dot{v} = Ev, \tag{2.2}$$

$$e = Cx + Du + Fv, \tag{2.3}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ the control input, and $v \in \mathbb{R}^q$ the state of the exosystem (3.2). The exosystem generates both the non-vanishing disturbance $g = Gv$ and the reference $y_0 = -Fv$ for the output of the plant $y = Cx + Du \in \mathbb{R}^r$. $e \in \mathbb{R}^r$ is the tracking error. $\omega \in \mathbb{R}^d$ is an unmodeled square integrable disturbance. $A \in \mathbb{R}^{n\times n}$, $B_1 \in \mathbb{R}^{n\times m}$, $B_2 \in \mathbb{R}^{n\times d}$, $C \in \mathbb{R}^{r\times n}$, $E \in \mathbb{R}^{q\times q}$ and $F \in \mathbb{R}^{r\times q}$, and $G \in \mathbb{R}^{n\times q}$ are constant matrices. Two standard assumptions are made throughout this paper.

Assumption 2.3.1. (A, B_1) *is stabilizable.*

Assumption 2.3.2. *The zero-transmission condition, i.e* $\text{rank}\begin{bmatrix} A-\lambda I & B_1 \\ C & D \end{bmatrix} = n + r,$ $\forall \lambda \in \sigma(E)$*, holds.*

The output regulation problem is mainly related to the design of a controller such that 1) the closed-loop system is exponentially stable at the origin when $v(t) \equiv 0$ and $\omega(t) \equiv 0$; and 2) the tracking error $e(t)$ asymptotically converges to zero for any initial conditions

$x(0)$ and $v(0)$ while $\omega(0) = 0$. In the absence of unmodeled disturbance ω, one can solve the output regulation problem by solving a regulator equation stated in Lemma 3.1.

Lemma 2.3.1. *((Francis, 1977)) Under Assumption 2.3.1, choose a control gain K such that $\sigma(A - B_1K) \subset \mathbb{C}^-$. The output regulation problem is solvable by the controller*

$$u = -Kx + Lv \tag{2.4}$$

if there exist matrices $X \in \mathbb{R}^{n\times q}, U \in \mathbb{R}^{m\times q}$ being solutions of the following regulator equations:

$$XE = AX + B_1U + G, \tag{2.5}$$

$$0 = CX + DU + F \tag{2.6}$$

with

$$L = U + KX. \tag{2.7}$$

Remark 2.3.1. *Assumption 3.3.1 ensures the solvability of regulator equations (2.5)–(2.6) for any matrices G, F; see (Huang, 2004).*

Recalling from (Krener, 1992), if the solution to the regulation equations (2.5)–(2.6) is not unique, one can find an optimal solution (X^*, U^*) by solving the following optimization problems.

Problem 2.3.1.

$$\min_{(X,U)} \mathrm{Tr}(X^T\bar{Q}X + U^T\bar{R}U) \tag{2.8}$$

$$\text{subject} \quad \text{to} \quad (2.5) - (2.6),$$

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

Letting $\bar{x} = x - X^*v$, $\bar{u} = u - U^*v$, it is direct to obtain the error system as follows

$$\dot{\bar{x}} = A\bar{x} + B_1\bar{u} + B_2\omega, \tag{2.9}$$

$$e = C\bar{x} + D\bar{u}. \tag{2.10}$$

Considering the effect of unmodeled disturbance ω, the GORP is defined as follows:

Definition 1. The GORP is solved if one designs the feedback controller

$$\bar{u} = -K^*\bar{x}, \tag{2.11}$$

and disturbance policy

$$\omega = N^*\bar{x}, \tag{2.12}$$

where the control gains K^* and N^* are obtained from the solution to the following constrained minimax problem:

Problem 2.3.2.

$$\min_{\bar{u}}\max_{\omega} \int_0^\infty (\bar{x}^T Q\bar{x} + \bar{u}^T R\bar{u} - \gamma^2\omega^T\omega)dt$$
$$\text{subject} \quad \text{to} \quad (2.9),$$

where $Q = Q^T > 0, R = R^T > 0$, $\gamma \geq \gamma^* \geq 0$. The γ^* is named H-infinity gain.

Remark 2.3.2. *Let the performance output be $z = \begin{bmatrix} \sqrt{Q}\bar{x} \\ \sqrt{R}\bar{u} \end{bmatrix}$. The system with the designed optimal control policy (2.11) achieves*

$$\int_0^\infty \|z\|^2 d\tau \leq \gamma^2 \int_0^\infty \|\omega\|^2 d\tau. \tag{2.13}$$

for any square integrable disturbance ω. The closed-loop system has $\mathcal{L}_2$ gain less than or equal to γ (Van Der Schaft, 1992).

Remark 2.3.3. *By Definition 1, it is checkable that the optimal feedback control input (2.11) can be written as*

$$u = -K^* x + L^* v, \tag{2.14}$$

while the feedforward control gain is $L^* = U^* + K^* X^*$. *The corresponding unmodeled disturbance policy (2.12) can be represented as*

$$\omega = N^* x - N^* X^* v. \tag{2.15}$$

2.3.2 REVIEW OF OPTIMAL CONTROL THEORY

Problem 2.3.2 is a standard linear quadratic regulator (LQR) design problem. By linear optimal control theory (Lewis, Vrabie, & Syrmos, 2012), the optimal feedback control gain K^* and disturbance gain N^* are

$$K^* = R^{-1} B_1^T P^*, \tag{2.16}$$

$$N^* = \gamma^{-2} B_2^T P^*, \tag{2.17}$$

respectively. These gains can obtained by solving for $P^* = P^{*T} > 0$ the following game algebraic Riccati equation (GARE)

$$\begin{aligned} A^T P^* + P^* A + Q \\ - P^* (B_1 R^{-1} B_1^T - \gamma^{-2} B_2 B_2^T) P^* = 0. \end{aligned} \tag{2.18}$$

Since (2.18) is nonlinear in P^*, it is usually difficult to directly solve P^* from (2.18). A model-based policy iteration algorithm—i.e., Algorithm 1—for solving GARE was established in (Modares et al., 2015; Wu & Luo, 2012), and is recalled below.

2.4 MAIN RESULTS

In this section, we first present a solution to regulator equations with known dynamics. Then, with unknown system matrices A, B_1, B_2, and G, we develop an RL algorithm to

Algorithm 1 Model-Based Algorithm for Solving GARE

1: Select a threshold $\epsilon_1 > 0$. Choose a stabilizing feedback control gain K_0, and a disturbance gain N_0.

2: $j \leftarrow 0$

3: **repeat**

4: Solve P_j from

$$\begin{aligned}(A-&B_1K_j + B_2N_j)^T P_j + P_j(A - B_1K_j + B_2N_j)\\ &+ Q + K_j^T R K_j - \gamma^2 N_j^T N_j = 0\end{aligned} \tag{2.19}$$

5: Solve K_{j+1} by

$$K_{j+1} = R^{-1}B_1^T P_j \tag{2.20}$$

6: Solve N_{i+1} by

$$N_{j+1} = \gamma^{-2}{B_2}^T P_j \tag{2.21}$$

7: $j \leftarrow j + 1$

8: **until** $|P_j - P_{j-1}| < \epsilon_1$

solve the regulator equation X^*, U^* and to approximate the optimal solution P^* to GARE, and optimal gains K^* and N^*. The convergence of the algorithm and the tracking ability of the closed-loop system are analyzed as well.

2.4.1 Solving regulator equations with known dynamics

Define two maps $\mathcal{S} : \mathbb{R}^{n\times q} \to \mathbb{R}^{n\times q}$ and $\bar{\mathcal{S}} : \mathbb{R}^{n\times q} \times \mathbb{R}^{m\times q} \to \mathbb{R}^{n\times q}$ by

$$\begin{aligned}\mathcal{S}(X) =& XE - AX,\\ \bar{\mathcal{S}}(X,U) =& XE - AX - BU, \quad X \in \mathbb{R}^{n\times q}, U \in \mathbb{R}^{m\times q}.\end{aligned}$$

Pick two constant matrices $X_1 \in \mathbb{R}^{n\times q}$ and $U_1 \in \mathbb{R}^{m\times q}$ such that

$$0 = CX_1 + DU_1 + F.$$

Then we select $X_i \in \mathbb{R}^{n\times q}$ and $U_i \in \mathbb{R}^{m\times q}$ for $i = 2, 3, \cdots, h+1$ such that all the vectors $\mathrm{vec}([X_i^T, U_i^T]^T)$ form a basis for $\ker(I_q \otimes [C, D])$, where h is the dimension of the null space of $I_q \otimes [C, D]$.

The following lemma shows a methodology to find the solution to regulator equations (2.5)–(2.6) with nontrivial feedforward matrix D.

Lemma 2.4.1. *The pair (X, U) solves the regulator equations (2.5)–(2.6) if and only if it satisfies the following matrix equation*

$$\mathcal{A}\chi = b, \tag{2.22}$$

where

$$\begin{aligned}
\mathcal{A} =& \begin{bmatrix} \mathrm{vec}([X_2^T, U_2^T]^T) & \cdots & \mathrm{vec}([X_{h+1}^T, U_{h+1}^T]^T) & -I_{nq} \\ \mathrm{vec}(\bar{\mathcal{S}}(X_2, U_2)) & \cdots & \mathrm{vec}(\bar{\mathcal{S}}(X_{h+1}, U_{h+1})) & 0_{nq\times nq} \end{bmatrix}, \\
\chi =& [\alpha_2, \cdots, \alpha_{h+1}, (\mathrm{vec}([X_1^T, U_1^T]^T))^T]^T, \\
b =& \begin{bmatrix} \mathrm{vec}([X_1^T, U_1^T]^T) \\ \mathrm{vec}(-\mathcal{S}(X_1, U_1) + G) \end{bmatrix}.
\end{aligned}$$

As it can be checked, each solution to (2.6) can be described by a sequence of $\alpha_2, \alpha_3, \cdots, \alpha_{h+1} \in \mathbb{R}$ as

$$(X, U) = (X_1, U_1) + \sum_{i=2}^{h+1} \alpha_i (X_i, U_i). \tag{2.23}$$

Combining (2.5) and (2.23), and based on the linearity of mapping $\bar{\mathcal{S}}$, we have

$$\bar{\mathcal{S}}(X, U) = \bar{\mathcal{S}}(X_1, U_1)) + \sum_{i=2}^{h+1} \alpha_i \bar{\mathcal{S}}(X_i, U_i) = G. \tag{2.24}$$

It immediately shows a pair (X, U) is a solution to (2.5)–(2.6) if and only if the pair satisfies (2.23)–(2.24). By vectorization, (2.23) can be written as

$$\sum_{i=2}^{h+1} \alpha_i \text{vec}\left(\begin{bmatrix} X_i \\ U_i \end{bmatrix}\right) - \text{vec}\left(\begin{bmatrix} X \\ U \end{bmatrix}\right) = -\text{vec}\left(\begin{bmatrix} X_1 \\ U_1 \end{bmatrix}\right), \tag{2.25}$$

while (2.24) can be written as

$$\sum_{i=2}^{h+1} \alpha_i \text{vec}(\bar{\mathcal{S}}(X_i, U_i)) = \text{vec}(-\mathcal{S}(X_1, U_1) + G). \tag{2.26}$$

Combining (2.25)–(2.26), we have (2.22). The proof is thus completed. □

2.4.2 Data-driven adaptive optimal controller design

Defining $\bar{x}_i = x - X_i v$, for $i = 0, 1, 2, \cdots, h+1$ with $X_0 = 0_{n\times q}$, from (2.1)–(2.2), we have

$$\begin{aligned}
\dot{\bar{x}}_i =& Ax + B_1 u + B_2 \omega + (G - X_i E)v \\
=& A(\bar{x}_i + X_i v) + B_1 u + B_2 w + (G - X_i E)v \\
=& A_j \bar{x}_i + B_1(K_j \bar{x}_i + u) + B_2(\omega - N_j \bar{x}_i) \\
&+ (G - \mathcal{S}(X_i))v,
\end{aligned} \tag{2.27}$$

where $A_j = A - B_1 K_j + B_2 N_j$.

Then, by equation (2.19), we have

$$
\begin{aligned}
&\bar{x}_i(t+\delta t)^T P_j \bar{x}_i(t+\delta t) - \bar{x}_i(t)^T P_j \bar{x}_i(t) \\
=&\int_t^{t+\delta t} \left[\bar{x}_i^T (A_j^T P_j + P_j A_j)\bar{x}_i + 2(u + K_j\bar{x}_i)^T B_1^T P_j \bar{x}_i \right. \\
&\left. +2v^T(G - \mathcal{S}(X_i))^T P_j \bar{x}_i + 2(\omega - N_j\bar{x}_i)^T B_2^T P_j \bar{x}_i\right] d\tau \\
=&-\int_t^{t+\delta t} \bar{x}_i^T (Q + K_j^T R K_j - \gamma^2 N_j^T N_j)\bar{x}_i d\tau \\
&+ 2\int_t^{t+\delta t} (u + K_j\bar{x}_i)^T R K_{j+1}\bar{x}_i d\tau \\
&+ 2\int_t^{t+\delta t} v^T (G - \mathcal{S}(X_i))^T P_j \bar{x}_i d\tau \\
&+ 2\gamma^2 \int_t^{t+\delta t} (\omega - N_j\bar{x}_i)^T N_{j+1}\bar{x}_i d\tau. \qquad (2.28)
\end{aligned}
$$

By Kronecker product representation, we obtain

$$
\begin{aligned}
&\bar{x}_i^T (Q + K_j^T R K_j - \gamma^2 N_j^T N_j)\bar{x}_i \\
&\quad = (\bar{x}_i^T \otimes \bar{x}_i^T)\mathrm{vec}(Q + K_j^T R K_j - \gamma^2 N_j^T N_j), \\
&v^T (G - \mathcal{S}(X_i))^T P_j \bar{x}_i \\
&\quad = (\bar{x}_i^T \otimes v^T)\mathrm{vec}((G - \mathcal{S}(X_i))^T P_j), \\
&(u + K_j\bar{x}_i)^T R K_{j+1}\bar{x}_i \\
&\quad = [(\bar{x}_i^T \otimes \bar{x}_i^T)(I_n \otimes K_j^T R) \\
&\qquad + (\bar{x}_i^T \otimes u^T)(I_n \otimes R)]\mathrm{vec}(K_{j+1}), \\
&(\omega - N_j\bar{x}_i)^T N_{j+1}\bar{x}_i \\
&\quad = [(\bar{x}_i^T \otimes \omega^T) - (\bar{x}_i^T \otimes \bar{x}_i^T)(I_n \otimes N_j^T)]\mathrm{vec}(N_{j+1}).
\end{aligned}
$$

Moreover, for positive integer s, define

$$\begin{aligned}
\delta_{\bar{x}_i\bar{x}_i} =&[\mathrm{vecv}(\bar{x}_i(t_1)) - \mathrm{vecv}(\bar{x}_i(t_0)), \mathrm{vecv}(\bar{x}_i(t_2))- \\
&\mathrm{vecv}(\bar{x}_i(t_1)), \cdots, \mathrm{vecv}(\bar{x}_i(t_s)) - \mathrm{vecv}(\bar{x}_i(t_{s-1}))]^T, \\
\Gamma_{\bar{x}_i\bar{x}_i} =&[\int_{t_0}^{t_1} \bar{x}_i \otimes \bar{x}_i d\tau, \int_{t_1}^{t_2} \bar{x}_i \otimes \bar{x}_i d\tau, \cdots, \int_{t_{s-1}}^{t_s} \bar{x}_i \otimes \bar{x}_i d\tau]^T, \\
\Gamma_{\bar{x}_i u} =&[\int_{t_0}^{t_1} \bar{x}_i \otimes u d\tau, \int_{t_1}^{t_2} \bar{x}_i \otimes u d\tau, \cdots, \int_{t_{s-1}}^{t_s} \bar{x}_i \otimes u d\tau]^T, \\
\Gamma_{\bar{x}_i v} =&[\int_{t_0}^{t_1} \bar{x}_i \otimes v d\tau, \int_{t_1}^{t_2} \bar{x}_i \otimes v d\tau, \cdots, \int_{t_{s-1}}^{t_s} \bar{x}_i \otimes v d\tau]^T, \\
\Gamma_{\bar{x}_i \omega} =&[\int_{t_0}^{t_1} \bar{x}_i \otimes \omega d\tau, \int_{t_1}^{t_2} \bar{x}_i \otimes \omega d\tau, \cdots, \int_{t_{s-1}}^{t_s} \bar{x}_i \otimes \omega d\tau]^T,
\end{aligned}$$

where $t_0 < t_1 < \cdots < t_s$ are positive integers. Hence, (4.12) implies the following linear equation

$$\Psi_{ij} \begin{bmatrix} \mathrm{vecs}(P_j) \\ \mathrm{vec}(K_{j+1}) \\ \mathrm{vec}((G - \mathcal{S}(X_i))^T P_j) \\ \mathrm{vec}(N_{j+1}) \end{bmatrix} = \Phi_{ij}, \tag{2.29}$$

where

$$\begin{aligned}
\Psi_{ij} =& [\delta_{\bar{x}_i\bar{x}_i}, -2\Gamma_{\bar{x}_i\bar{x}_i}(I_n \otimes (K_j^T R) - 2\Gamma_{\bar{x}_i u}(I_n \otimes R), \\
&- 2\Gamma_{\bar{x}_i v}, -2\gamma^2(\Gamma_{\bar{x}_i \omega} - \Gamma_{\bar{x}_i\bar{x}_i}(I_n \otimes N_i^T))], \\
\Phi_{ij} =& -\Gamma_{\bar{x}_i\bar{x}_i} \mathrm{vec}(Q + (K_j)^T R K_j - \gamma^2 N_i^T N_i).
\end{aligned}$$

Equation (2.29) can be uniquely solved when matrix Ψ_{ij} is of full column rank, i.e.,

$$\begin{bmatrix} \mathrm{vecs}(P_j) \\ \mathrm{vec}(K_{j+1}) \\ \mathrm{vec}((G - \mathcal{S}(X_i))^T P_j) \\ \mathrm{vec}(N_{j+1}) \end{bmatrix} = (\Psi_{ij}^T \Psi_{ij})^{-1} \Psi_{ij}^T \Phi_{ij}. \tag{2.30}$$

Remark 2.4.1. *Like in the previous work of others, an exploration noise ξ is needed in the learning phase to excite the system such that Ψ_{ij} has full column rank. Examples of such noise are random noise (Al-Tamimi et al., 2007) and sinusoidal signal (Y. Jiang & Jiang, 2012).*

From (2.29), one can compute G for $i = 0$ and $\mathcal{S}(X_i)$ for $i = 1, 2, \cdots, h+1$. From (2.20), for any $j > 0$, one can further obtain the map $\bar{\mathcal{S}}$ by

$$\bar{\mathcal{S}}(X_i, U_i) = \mathcal{S}(X_i) - P_j^{-1} K_{j+1}^T R U_i. \tag{2.31}$$

Thus, both $\mathcal{A}$ and b in (2.22) are computable. By Lemma 2.4.1, we can replace the constraint of Problem 2.3.1 by (2.22). Problem 2.3.1 is essentially a convex optimization problem that has been studied in the past literature (Boyd & Vandenberghe, 2004).

The data-driven RL algorithm for dealing with GORP is presented as follows.

The convergence of Algorithm 2 is discussed in the following Theorem.

Theorem 2.1. *The sequences $\{P_j\}_{j=0}^{\infty}$, $\{K_j\}_{j=1}^{\infty}$ and $\{N_j\}_{j=1}^{\infty}$ obtained from Algorithm 2 converge to P^* and K^*, and N^*, respectively.*

Given a stabilizing K_j, if $P_j = P_j^T$ is the solution to (2.19), K_i^{j+1} and N_i is determined by $K_{j+1} = R^{-1} B_1^T P_j$ and $N_{j+1} = \gamma^{-2} B_2^T P_j$, respectively. Let $T_{ij} = (G - \mathcal{S}(X_i))^T P_j$. By (2.28), we know that P_j, K_{j+1}, N_{j+1} and T_{ij} satisfy (2.29).

Therefore, the policy iteration (2.29) is equivalent to (2.19)–(2.21) in Algorithm 1. The convergence of Algorithm 1 has been proved in (Zhu, Modares, Peen, Lewis, & Yue, 2015; Wu & Luo, 2012). This ensures the convergence of Algorithm 2. The proof is completed.

Now, we are ready to show the tracking ability of the closed-loop system.

Theorem 2.2. *Considering the linear continuous-time system (2.1)–(2.3), let*

$$u = -K^{\dagger} x + L^{\dagger} v \tag{2.32}$$

Algorithm 2 Data-Driven Algorithm for Solving GORP

1: Select a threshold $\epsilon_2 > 0$. Compute matrices $X_0, X_1, \cdots, X_{h+1}$ and $U_0, U_1, \cdots, U_{h+1}$
2: Apply $u = -K_0 x + \xi$ on $[t_0, t_s]$ with (bounded) exploration noises ξ
3: $j \leftarrow 0, i \leftarrow 0$
4: **repeat**
5: Solve P_j, K_{j+1} and N_{j+1} from (3.44)
6: $j \leftarrow j + 1$
7: **until** $|P_j - P_{j-1}| < \epsilon_2$
8: Obtain the approximated optimal control gains $K^\dagger$ and $N^\dagger$, and approximated solution $P^\dagger$ to (3.14)
9: **repeat**
10: Solve $\mathcal{S}(X_i)$ from (3.44)
11: $i \leftarrow i + 1$
12: **until** $i = h + 2$
13: Obtain (X^*, U^*) by solving Problem 2.3.1
14: Obtain the approximate optimal feedforward control gain $L^\dagger = U^* + K^\dagger X^*$

and be the approximated optimal control with control gains $K^\dagger$ and $L^\dagger$ obtained from Algorithm 2. Then, the tracking error $e(t)$ asymptotically converges to zero.

The system (2.1)–(2.3) in closed-loop with the approximated optimal controller (2.32) implies the following error system

$$\begin{aligned}\dot{\bar{x}} =& A\bar{x} + B_1\bar{u} + B_2\omega \\ :=& (A - B_1K^\dagger)\bar{x} + B_2\omega, \\ e =& C\bar{x}.\end{aligned}$$

From the GARE, there exists a small enough ϵ_2 such that the following inequality holds

$$\begin{aligned} A^T P^\dagger + P^\dagger A + \frac{Q}{2} \\ - P^\dagger (B_1 R^{-1} B_1^T - \gamma^{-2} B_2 B_2^T) P^\dagger < 0. \end{aligned} \tag{2.33}$$

The inequality can be rewritten by

$$\begin{aligned} (A - BK^\dagger)^T P^\dagger + P^\dagger (A - BK^\dagger) + \frac{Q}{2} + (K^\dagger)^T R K^\dagger \\ + \gamma^{-2} P^\dagger B_2 B_2^T P^\dagger < 0. \end{aligned} \tag{2.34}$$

Choose a function $V = \bar{x}^T P^\dagger \bar{x}$, take the derivative around the solution of the error system

$$\begin{aligned} \dot{V} =& \bar{x}^T (A - B_1 K^\dagger)^T P^\dagger \bar{x} + \bar{x}^T P^\dagger (A - B_1 K^\dagger) \bar{x} \\ & + 2\bar{x}^T P^\dagger B_2 \omega \\ =& -\frac{1}{2} \bar{x}^T Q \bar{x} - \bar{u}^T R \bar{u} - \|\gamma^{-1} B_2^T P^\dagger \bar{x}\|^2 \\ & + 2\bar{x}^T P^\dagger B_2 \omega \\ \leq& -\frac{\lambda_{\min}(Q)}{2} \|\bar{x}\|^2 - \lambda_{\min}(R) \|\bar{u}\|^2 \\ & - \|\gamma^{-1} B_2^T P^\dagger \bar{x} - \gamma \omega\|^2 + \gamma^2 \|\omega\|^2 \end{aligned} \tag{2.35}$$

Integrating (2.35) from $t = 0$ to $t = \infty$, we have

$$\begin{aligned} & \int_0^\infty (\frac{\lambda_{\min}(Q)}{2} \|\bar{x}\|^2 + \lambda_{\min}(R) \|\bar{u}\|^2) dt \\ \leq& \gamma^2 \int_0^\infty \|\omega\|^2 dt + V(x(0)) - \lim_{t \to \infty} V(x(t)) \\ \leq& \gamma^2 \int_0^\infty \|\omega\|^2 dt + V(x(0)) \\ <& \infty \end{aligned} \tag{2.36}$$

Through Barbalat's lemma (Khalil, 2002), we have $\lim_{t \to \infty} \bar{u}(t) = 0$, $\lim_{t \to \infty} \bar{x}(t) = 0$. It is immediate to have $\lim_{t \to \infty} e(t) = \lim_{t \to \infty} C\bar{x}(t) = 0$. The proof is completed.

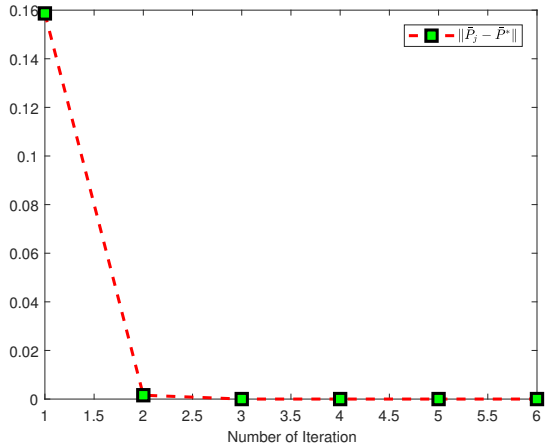

Figure 2.1: Convergence of P_j to its Optimal Value P^* during the Learning Process

Remark 2.4.2. *The exostate v is supposed measurable in this paper. when unmeasurable, we can use Gao's method to reconstruct the exostate with the knowledge of the minimal polynomial of E (Gao & Jiang, 2016a).*

2.5 Illustrative Example

To validate the effectiveness of the proposed data-driven Algorithm 2, we consider a continuous-time linear system with the following parameters:

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -3 \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 0.6 \end{bmatrix}, B_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, C = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T,$$
$$D = 1, E = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, F = \begin{bmatrix} -1 \\ 0 \end{bmatrix}^T, G = \begin{bmatrix} 0 & 0 \\ 0 & 0.5 \end{bmatrix}. \tag{2.37}$$

These system matrices A, B_1, B_2, and G are assumed to be unknown, the weight matrices (Q and R) are identity matrices, and γ is 11. For the purpose of simulation, the exploration noise is selected as a summation of sinusoidal waves with different frequencies.

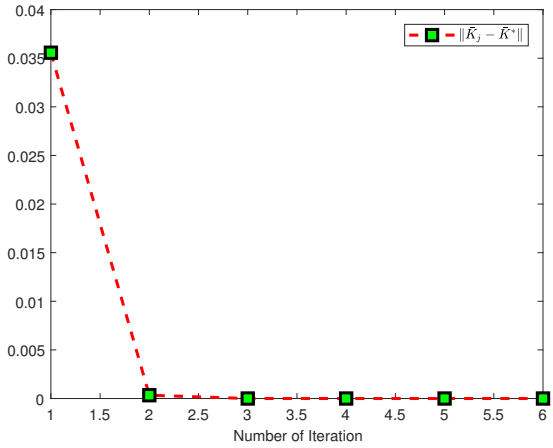

Figure 2.2: Convergence of K_j to its Optimal Value K^* during the Learning Process

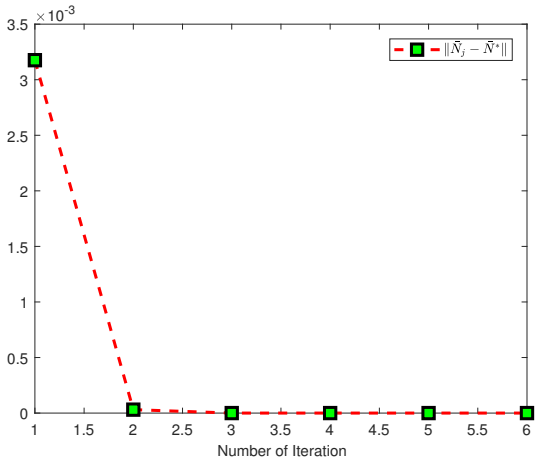

Figure 2.3: Convergence of N_j to its Optimal Value N^* during the Learning Process

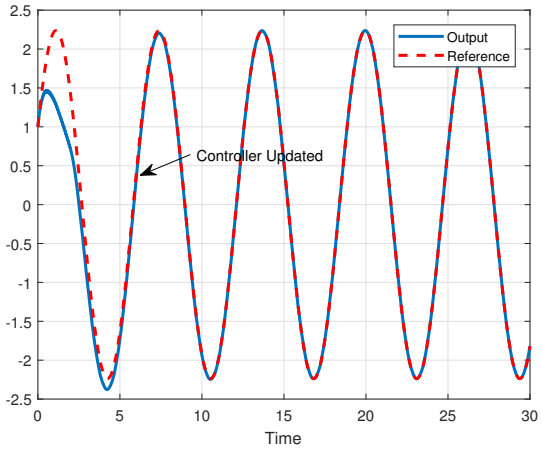

Figure 2.4: Trajectories of the Output and Reference

Convergence is achieved after 6 iterations when implementing the proposed Algorithm 2. The approximated control gains $K^{\dagger}$ and $N^{\dagger}$ captured and their optimal values K^{*} and N^{*} are:

$$K_6 = \begin{bmatrix} 0.309955 & 0.202105 \end{bmatrix},$$
$$K^{*} = \begin{bmatrix} 0.309954 & 0.202102 \end{bmatrix},$$

and

$$N_6 = \begin{bmatrix} 0.0326835 & 0.0154046 \end{bmatrix},$$
$$N^{*} = \begin{bmatrix} 0.0326834 & 0.0154047 \end{bmatrix}.$$

The comparison of feedforward gains is

$$L_6 = \begin{bmatrix} 0.309993 & 4.368750 \end{bmatrix},$$
$$L^{*} = \begin{bmatrix} 0.309954 & 4.368770 \end{bmatrix}.$$

Fig. 2.4 shows that the obtained data-driven approximated optimal controller makes the output of the plant to asymptotically track the given reference signal.

2.6 CONCLUSIONS

This paper proposes a novel reinforcement learning based approach to the adaptive optimal output regulation of linear systems with zero-sum differential games. A systematic data-driven control scheme is proposed for designing adaptive optimal trackers with guaranteed rejection of nonvanishing disturbance. Future work will be directed at generalizing the proposed method for the adaptive optimal tracking problem of uncertain nonlinear systems using differential games.

CHAPTER 3

DATA-DRIVEN GLOBAL ROBUST OPTIMAL OUTPUT REGULATION OF UNCERTAIN PARTIALLY LINEAR SYSTEMS

3.1 ABSTRACT

In this paper, a data-driven control approach is developed by reinforcement learning (RL) to solve the global robust optimal output regulation problem (GROORP) of partially linear systems with both static uncertainties and nonlinear dynamic uncertainties. By developing a proper feedforward controller, the GROORP is converted into a global robust optimal stabilization problem. A robust optimal feedback controller was designed which is able to stabilize the system in the presence of dynamic uncertainties. The closed-loop system is assured to be input-to-output stable regarding the static uncertainty as the external input. This robust optimal controller is numerically approximated via RL. Nonlinear small-gain theory is applied to show the input-to-output stability for the closed-loop system and thus solves the original GROORP. Simulation results validate the efficacy of the proposed methodology.

3.2 INTRODUCTION

The output regulation problem aims at designing control strategies to achieve the rejection of a nonvanishing disturbance and forces the output of dynamic systems to asymptotically track a desired reference. This problem has been tackled for linear systems since 1970s (Francis, 1977). Due to its relevance to many real-world applications, the output regulation problem for nonlinear systems has also attracted considerable attention with focus on either local, semi-global or global stabilization (Isidori & Byrnes, 1990; Huang & Rugh, 1990; Byrnes, Priscoli, Isidori, & Kang, 1997; Serrani et al., 2001; Huang, 2004).

In most existing output regulation problems, both the nonvanishing disturbance and

the reference are generated by an autonomous system, named exosystem—wherein the unmodeled disturbance is neglected. Yaghmaie considered a more generalized case; H_∞ output regulation problem, where both the nonvanishing and unmodeled disturbances were considered (Yaghmaie et al., 2018). This H_∞ optimal control and output regulation techniques are efficiently combined for robust model-based control design. The H_∞ optimal control can be formulated a zero-sum differential game involving two players; the controller (player 1) and the unmodeled disturbance (player 2) which are the minimizing and maximizing players, respectively (Başar & Bernhard, 2008). In this setting, one can solve the output regulation problem for the system with unmodeled disturbances which are static uncertainties. Considering a nonlinear system with dynamic uncertainties, the notion of input-to-state stability and small-gain theory (Gao & Jiang, 2015c, 2015a) have been employed to solve the global robust output regulation problems (Huang & Chen, 2004). However, we are not aware of any existing work on output regulation problems that take both static and dynamic uncertainties into consideration. Also, it is noteworthy that most of the existing control strategies to output regulation problems are model-based, which means that an accurate knowledge of system's model is absolutely needed.

Reinforcement Learning (RL) is a non-model-based and data-driven approach which solves optimal control problems via online state and input information (Lewis & Vrabie, 2009). RL has been used to design optimal feedback controllers for both continuous-time and discrete-time systems wherein optimal cost and feedback controllers are computed using online data (Y. Jiang & Jiang, 2012; Gao, Jiang, Jiang, & Chai, 2014; Lewis & Vrabie, 2009; Gao, Jiang, & Ozbay, 2015; Gao & Jiang, 2015b; B. Sun et al., 2018). Jiang proposed, a robust data-driven approach solve control problems in linear and nonlinear systems with dynamic uncertainties (Y. Jiang & Jiang, 2014). Gao extends the solution to global optimal output regulation problems by incorporating dynamic uncertainties in the system (Gao & Jiang, 2015b). The exact knowledge of system dynamics and dynamic

uncertainties are not required to design the robust optimal controllers.

This paper aims at proposing a novel data-driven solution to the global robust optimal output regulation problem (GROORP) for a class of partially linear composite systems. It is challenging since the system studied in this paper is with unknown dynamics, and both static and dynamic uncertainties. First, we convert the GROORP into a global robust optimal stabilization problem. Then a data-driven approach is developed to compute the robust adaptive optimal controller and disturbance policy via online input and state information. In the presence of dynamic uncertainty, the rejection of nonvanishing disturbance and the output trajectories asymptotically tracking the desired reference is achieved. With both the static and dynamic uncertainty, it is guaranteed that the closed-loop system is input-to-output stable with the static uncertainty acting as an the external input. Optimality and global output regulation are both achieved for the class of partially linear systems.

The remainder of this paper is organized as follows. In Section II, we briefly review the linear optimal output regulation problem and linear optimal control theory. Considering static and nonlinear dynamic uncertainties, we formulate the GROORP for a class of partially linear systems in Section III. An offline solution on the basis of nonlinear small-gain theory is proposed therein. In Section IV, the RL technique is employed to design a robust optimal controller via online data. Simulation results on a partially linear system are provided in Section V. Finally, concluding remarks are given in Section VI.

3.3 Robust optimal output regulation of linear systems

Considering a class of linear systems with nonvanishing disturbance and reference signals generated by linear exosystems, the robust optimal output regulation problem (ROORP) is formulated by minimizing both static and dynamic optimization problems. Then, we recall the basics of robust control and policy iteration (PI) technique. An approach explicitly solving the regulator equation is presented as well.

3.3.1 PROBLEM FORMULATION

To begin with, consider the linear system

$$\dot{x} = Ax + B_1 u + B_2 \omega + Dv, \tag{3.1}$$

$$\dot{v} = Ev, \tag{3.2}$$

$$y = Cx, \tag{3.3}$$

$$y_d = -Fv, \tag{3.4}$$

$$e = y - y_d \tag{3.5}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ the control input, and $v \in \mathbb{R}^q$ the state of the exosystem (3.2). The exosystem generates both the nonvanishing disturbance $\eta = Dv$ and the reference $y_0 = -Fv$ for the output of the plant $y = Cx \in \mathbb{R}^r$. $e \in \mathbb{R}^r$ is the tracking error. $\omega \in \mathbb{R}^d$ is an unmodeled square integrable disturbance. $A \in \mathbb{R}^{n\times n}$, $B_1 \in \mathbb{R}^{n\times m}$, $B_2 \in \mathbb{R}^{n\times d}$, $C \in \mathbb{R}^{r\times n}$, $E \in \mathbb{R}^{q\times q}$, $F \in \mathbb{R}^{r\times q}$, and $D \in \mathbb{R}^{n\times q}$ are system matrices with (A, B_1) stabilizable. Throughout this paper, the following assumptions are made.

Assumption 3.3.1. *The transmission zeros condition holds, i.e.,*

$$\operatorname{rank} \begin{bmatrix} A - \lambda I & B_1 \\ C & 0 \end{bmatrix} = n + r, \forall \lambda \in \sigma(E). \tag{3.6}$$

Assumption 3.3.2. *All the eigenvalues of E are simple with zero real part.*

Based on Assumptions 3.3.1-3.3.2, one can get the following technical result.

Theorem 3.1. *Let the feedback gain $K \in \mathbb{R}^{m\times n}$ be such that $\sigma(A - B_1 K) \in \mathbb{C}^-$. Then, if a controller is designed as $u = -K(x - Xv) + Uv$, where $X \in \mathbb{R}^{n\times q}$ and $U \in \mathbb{R}^{m\times q}$ solve the following equations:*

$$XE = AX + B_1 U + D,$$

$$0 = CX + F, \tag{3.7}$$

then the closed-loop linear system achieves disturbance rejection and asymptotic tracking.

Proof. Letting $\bar{x} = x - Xv, \bar{u} = u - Uv$ and using (3.7), we have

$$\begin{aligned}\dot{\bar{x}} &= Ax + B_1u + B_2\omega + Dv - XEv \\ &= Ax - B_1K(x - Xv) + (B_1U + D)v + B_2\omega - XEv \\ &= (A - B_1K)\bar{x} + B_2\omega. \end{aligned} \tag{3.8}$$

Since $\sigma(A - B_1K) \in \mathbb{C}^-$ and $w(t)$ is square integrable, we observe $\lim_{t\to\infty} \bar{x}(t) = 0$ and $\lim_{t\to\infty} \bar{u}(t) = 0$, which implies $\lim_{t\to\infty} e(t) = \lim_{t\to\infty} C\bar{x}(t) = 0$. The proof is completed.

Remark 3.3.1. *(3.7) is called the linear regulator equation. Assumption 3.3.1 is made such that (3.7) is solvable for any matrices D, F (Huang, 2004).*

Inspired by Gao, 2015d and Krener, 1992, we tackle the robust optimal output regulation problem (ROORP) by solving a static optimization Problem 3.3.1 to find the optimal solution (X^*, U^*) to (3.7) and a dynamic optimization Problem 3.3.2 to find the optimal gains K^* and N^*.

Problem 3.3.1.

$$\begin{aligned}&\min_{(X,U)} \mathrm{Tr}(X^T\bar{Q}X + U^T\bar{R}U), \\ &\text{subject} \quad \text{to} \quad (3.7)\end{aligned} \tag{3.9}$$

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

One can write the error system of (3.1)-(3.5) as:

$$\dot{\bar{x}}^* = A\bar{x}^* + B_1\bar{u}^* + B_2\omega, \tag{3.10}$$

$$e = C\bar{x}^* \tag{3.11}$$

where $\bar{x}^* = x - X^*v, \bar{u}^* = u - U^*v$.

Problem 3.3.2.

$$\min_{\bar{u}} \max_{\omega} \int_0^\infty [(\bar{x}^*)^T Q \bar{x}^* + (\bar{u}^*)^T \bar{u}^* - \gamma^{-2} \omega^T \omega] dt$$
$$\text{subject} \quad \text{to} \quad (3.10) - (3.11),$$

where $Q = Q^T > 0$, and $\gamma \geq \gamma^* \geq 0$. The γ^* is named by H_∞ gain.

Remark 3.3.2. *In order to solve the ROORP, we ought to design a control policy $u = -K^*(x - X^*v) + U^*v$ and a disturbance policy $\omega = N^*(x - X^*v)$ where optimal control gains K^* and N^*, and optimal regulator parameters X^* and U^* are achieved by solving optimization Problems 3.3.1 and 3.3.2. Theorem 3.1 ensures that the resultant closed-loop system achieves disturbance rejection and asymptotic tracking.*

Remark 3.3.3. *It is shown in (Gao & Jiang, 2016a, Remark 5) that the Problem 3.3.1 can be converted as a convex optimization problem with a quartic cost and linear constraints. The solution to the Problem 3.3.1 is unique given positive definite matrices $\bar{Q}$ and $\bar{R}$. The motivation of introducing the Problem 3.3.1 is to optimize the steady-state behavior the system.*

3.3.2 H_∞ CONTROL AND POLICY ITERATION (PI)

By linear optimal control theory, we design an optimal feedback controller $\bar{u}^* = -K^*\bar{x}^*$ and a disturbance policy $\omega^* = N^*\bar{x}^*$ to minimize the cost of Problem 3.3.2. The optimal feedback control gain K^* and disturbance gain N^* are

$$K^* = {B_1}^T P^*, \tag{3.12}$$

$$N^* = \gamma^{-2} {B_2}^T P^*, \tag{3.13}$$

respectively, with $P^* = P^{*T} > 0$ the unique solution to the following game algebraic Riccati equation (GARE)

$$\begin{aligned} &A^T P^* + P^* A + Q \\ &\quad - P^*(B_1 B_1^T - \gamma^{-2} B_2 B_2^T) P^* = 0. \end{aligned} \tag{3.14}$$

Remark 3.3.4. *From (3.12) to (3.14), computing K^* and N^* does not depend on X^*, U^*. Problems 3.3.1 and 3.3.2 can be solved separately.*

Lemma 3.3.1 ((Modares et al., 2015))**.** *Let $K_0 \in \mathbb{R}^{m\times n}$ be any stabilizing control gain, and let $N_0 \in \mathbb{R}^{d\times n}$ be a zero matrix. $P_j = P_j^T > 0$ is the solution to the following Lyapunov equation*

$$\begin{aligned} &(A - B_1 K_j + B_2 N_j)^T P_j + P_j (A - B_1 K_j + B_2 N_j) \\ &\quad + Q + K_j^T K_j - \gamma^2 N_j^T N_j = 0 \end{aligned} \tag{3.15}$$

where K_j and N_j, with $j = 1, 2, \cdots$, are defined by

$$K_j = B_1^T P_{j-1} \tag{3.16}$$

$$N_j = \gamma^{-2} B_2^T P_{j-1} \tag{3.17}$$

Then, the following properties hold:

1. $\sigma(A - B_1 K_j) \in \mathbb{C}^-$,
2. $P^* \le P_{j+1} \le P_j$,
3. $\lim_{j\to\infty} K_j = K^*$, $\lim_{j\to\infty} N_j = N^*$ *and* $\lim_{j\to\infty} P_j = P^*$.

3.3.3 SOLVING REGULATOR EQUATIONS

Define a Sylvester map $\mathcal{S} : \mathbb{R}^{n\times q} \to \mathbb{R}^{n\times q}$ by

$$\mathcal{S}(X) = XE - AX, X \in \mathbb{R}^{n\times q}. \tag{3.18}$$

If we choose a $X_1 \in \mathbb{R}^{n\times q}$ such that $CX_1+F=0$, and $X_i \in \mathbb{R}^{n\times q}$, for $i=2\cdots h+1$, such that all the $\mathrm{vec}(X_i)$ form a basis of $\ker(I_q \otimes C)$, where h is the nullity of $I_q \otimes C$, then a pair $(X_\dagger^0, U_\dagger^0)$ is a solution to the regulator equation (3.7) if and only if there exist $\alpha_2^0, \alpha_3^0, \cdots, \alpha_{h+1}^0 \in \mathbb{R}$ such that

$$\mathcal{S}(X_\dagger^0) = B_1 U_\dagger^0 + D, \tag{3.19}$$

$$X_\dagger^0 = X_1 + \sum_{i=2}^{h+1} \alpha_i^0 X_i. \tag{3.20}$$

If the solution is not unique, we find all linearly independent vectors $\mathrm{vec}(\begin{bmatrix} X_\dagger^k \\ U_\dagger^k \end{bmatrix})$ by seeking sequences $\alpha_i^k \in \mathbb{R}$ such that, for $k=1,2,\cdots,H$ with $H=q(m-r)$,

$$X_\dagger^k = \sum_{i=2}^{h+1} \alpha_i^k X_i, BU_\dagger^k = \sum_{i=2}^{h+1} \alpha_i^k \mathcal{S}(X_i). \tag{3.21}$$

Then, the solution set of (3.7) is equivalent to

$$\begin{aligned}\mathbb{S} =&\{(X,U)|X = X_\dagger^0 + \sum_{k=1}^{H} \beta_k X_\dagger^k, U = U_\dagger^0 + \sum_{k=1}^{H} \beta_k U_\dagger^k,\\ &\forall \beta_1, \beta_2, \cdots, \beta_H \in \mathbb{R}\}.\end{aligned} \tag{3.22}$$

If we compute $\mathcal{S}(X_i)$ for $i=0,1,\cdots,h+1$ by online data, the solution set of the regulator equation (3.7) is obtained with unknown system matrices. The proposed method for solving the regulator equation paves the way for online robust optimal controller design in Section 3.5.

3.4 Global Robust Optimal Output Regulation of Partially Linear Systems

In this section, we formulate the GROORP of a class of partially composite linear systems. An offline solution to the GROORP is given by developing a global robust optimal controller.

3.4.1 GROORP FORMULATION

Motivated by the class of partially linear systems in Saberi and Summers, 1990, we study a general class of perturbed partially linear systems:

$$\dot{\zeta} = g(\zeta, y, v), \tag{3.23}$$

$$\dot{x} = Ax + B_1[u + \Delta(\zeta, y, v)] + B_2\omega + Dv, \tag{3.24}$$

$$\dot{v} = Ev, \tag{3.25}$$

$$y = Cx, \tag{3.26}$$

$$y_d = -Fv, \tag{3.27}$$

$$e = y - y_d, \tag{3.28}$$

where $\zeta \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ represents the states of the dynamic uncertainty (3.23) and the exosystem (3.25), respectively. The functions $g(\zeta, y, v) : \mathbb{R}^p \times \mathbb{R}^r \times \mathbb{R}^q \to \mathbb{R}^p$, and $\Delta(\zeta, y, v) : \mathbb{R}^p \times \mathbb{R}^r \times \mathbb{R}^q \to \mathbb{R}^m$ are sufficiently smooth functions satisfying $g(0, 0, 0) = 0$ and $\Delta(0, 0, 0) = 0$. Suppose $A, B_1, B_2, D, g, \Delta$ are unknown with ζ unmeasurable (Saberi, Kokotovic, & Summers, 1990).

Remark 3.4.1. *The GROORP is solvable for the partially linear system (3.23)-(3.28) if a robust optimal controller is found by solving optimization Problems 3.3.1 and 3.3.2 such that for $v : [0, \infty) \to V$ with V being a prescribed compact set of $\mathbb{R}^q$, any initial conditions $\zeta(0), x(0)$, the trajectory of closed-loop system (3.23)-(3.28) exists and is bounded for any $t \geq 0$, and satisfies $\lim_{t\to\infty} e(t) = 0$ when $\omega \equiv 0$. And the system is input-to-output stable for nontrivial ω.*

3.4.2 OFFLINE SOLUTIONS TO GROORP

Let Σ_v be the class of piecewise functions from $[0, \infty)$ to V. Then two assumptions are made on the system (3.23)-(3.28):

Assumption 3.4.1. *A sufficiently smooth function $\boldsymbol{\zeta}(v)$ with $\boldsymbol{\zeta}(0) = 0$ exists satisfying the following equation for any $v \in \mathbb{R}^q$:*

$$\frac{\partial \boldsymbol{\zeta}(v)}{\partial v} Ev = g(\boldsymbol{\zeta}(v), y_d, v),$$
$$0 = \Delta(\boldsymbol{\zeta}(v), y_d, v). \tag{3.29}$$

Under equations (3.7) and (3.29), we write the error system of (3.23)-(3.28) by letting $\bar{\zeta} = \zeta - \boldsymbol{\zeta}(v)$,

$$\dot{\bar{\zeta}} = \bar{g}(\bar{\zeta}, e, v), \tag{3.30}$$
$$\dot{\bar{x}} = A\bar{x} + B_1[\bar{u} + \bar{\Delta}(\bar{\zeta}, e, v)] + B_2\omega, \tag{3.31}$$
$$e = C\bar{x}, \tag{3.32}$$

where

$$\bar{g}(\bar{\zeta}, e, v) = g(\zeta, y, v) - g(\boldsymbol{\zeta}(v), y_d, v),$$
$$\bar{\Delta}(\bar{\zeta}, e, v) = \Delta(\zeta, y, v) - \Delta(\boldsymbol{\zeta}(v), y_d, v).$$

Two assumptions are made on the dynamic uncertainty, i.e., $\bar{\zeta}$-system, with e as the input and $\bar{\Delta}$ as the output.

Assumption 3.4.2. *There exist a function σ_s of class $\mathcal{KL}$ and a function γ_s of class $\mathcal{K}$, both of which are independent of any $v \in \Sigma_v$ such that for any measurable locally essentially bounded e on $[0, T)$ with $0 < T \leq +\infty$ and any $v \in \Sigma_v$, $\bar{\zeta}(t)$ right maximally defined on $[0, T')(0 < T' \leq T)$ satisfies*

$$|\bar{\zeta}(t)| \leq \sigma_s(|\bar{\zeta}(0)|, t) + \gamma_s(\|[e^T_{[0,t]}, \bar{\Delta}^T_{[0,t]}]^T\|), \forall t \in [0, T'),$$

where $e_{[0,t]}$ and $\bar{\Delta}_{[0,t]}$ are the truncated functions of e and $\bar{\Delta}$ over $[0, t]$, respectively.

Assumption 3.4.3. *There exist a function σ_Δ of class $\mathcal{KL}$ and a function γ_Δ of class $\mathcal{K}$, both of which are independent of any $v \in \Sigma_v$ such that, for any initial state $\bar{\zeta}(0)$, any*

measurable locally essentially bounded e *on* $[0, T)$ *with* $0 < T \leq +\infty$ *and any* $v \in \Sigma_v$, $\bar{\Delta}(t)$ *right maximally defined on* $[0, T')(0 < T' \leq T)$ *satisfies*

$$|\bar{\Delta}(t)| \leq \sigma_{\Delta}(|\bar{\zeta}(0)|, t) + \gamma_{\Delta}(\|e_{[0,t]}\|), \forall t \in [0, T'). \tag{3.33}$$

Remark 3.4.2. *Assumptions 3.4.2 and 3.4.3 are made so that system (3.30) has strong unboundedness observability (SUO) (Z. P. Jiang, Teel, & Praly, 1994) with zero-offset and input-to-output stability (IOS) (Sontag, 2007) properties. Then by nonlinear small-gain theory, a controller exists to globally asymptotically stabilize the error system (Huang & Chen, 2004).*

Theorem 3.2. *Under Assumptions 3.4.2 and 3.4.3, let symmetric matrices* $Q \geq \gamma_x I_n, R = I_m$ *with* $\gamma_x > 0$. *If the gain function* $\gamma_{\Delta}(s)$ *satisfies the following inequality*

$$\gamma_{\Delta}(s) \leq (Id + \rho_1)^{-1} \circ \gamma_e^{-1} \circ (Id + \rho_2)^{-1}(s), \forall s \geq 0 \tag{3.34}$$

for $\gamma_e(s) = |C|\sqrt{1/\gamma_x}s$ *and* ρ_1, ρ_2 *of class* $\mathcal{K}_{\infty}$, *then, for any exostate* v, *the error system (3.30)-(3.32) in closed-loop with the optimal control policy* $\bar{u} = -K^*\bar{x}$ *is globally asymptotically stable when* $\omega \equiv 0$. *Moreover, when* ω *is nontrivial, the closed-loop system is input-to-output stable regarding* ω *as an input (Z. P. Jiang et al., 1994).*

Proof. The GARE can be rewritten as

$$\begin{aligned}&(A - B_1K^*)^T P^* + P^*(A - B_1K^*) + Q \\ &+ P^*B_1B_1^T P^* + \gamma^{-2}P^*B_2B_2^T P^* = 0\end{aligned} \tag{3.35}$$

Differentiating the Lyapunov function $V = \bar{x}^T P^* \bar{x}$ gives

$$\begin{aligned}
\dot{V} =& \bar{x}^T[(A - B_1K^*)^T P^* + P^*(A - B_1K^*)]\bar{x} + 2\bar{x}^T P^* B_1 \bar{\Delta} \\
&+ 2\bar{x}^T P^* B_2 \omega \\
=& - \bar{x}^T(Q + P^* B_1 B_1^T P^* + \gamma^{-2} P^* B_2 B_2^T P^*)\bar{x} \\
&+ 2\bar{x}^T P^* B_1 \bar{\Delta} + 2\bar{x}^T P^* B_2 \omega \\
\leq& - \bar{x}^T Q \bar{x} - |\bar{\Delta} - B_1^T P^* \bar{x}|^2 - |\gamma w - \gamma^{-1} B_2^T P^* \bar{x}|^2 \\
&+ |\bar{\Delta}|^2 + \gamma^2 |\omega|^2 \\
\leq& - \bar{x}^T Q \bar{x} + |\bar{\Delta}|^2 + \gamma^2 |\omega|^2 \\
\leq& - \gamma_x |\bar{x}|^2 + |\bar{\Delta}|^2 + \gamma^2 |\omega|^2
\end{aligned} \tag{3.36}$$

for any $t \geq 0$, we have

$$\begin{aligned}
V(t) \leq \exp\left(-\frac{\gamma_x}{\lambda_m(P^*)} t\right) V(0) + \frac{\lambda_m(P^*)}{\gamma_x} \|\bar{\Delta}\|^2 \\
+ \frac{\gamma^2 \lambda_m(P^*)}{\gamma_x} \|\omega\|^2.
\end{aligned} \tag{3.37}$$

An immediate consequence of the previous inequality is

$$\begin{aligned}
|\bar{x}(t)| \leq \exp\left(-\frac{\gamma_x}{2\lambda_m(P^*)} t\right) \sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)}} |\bar{x}(0)| \\
+ \sqrt{\frac{1}{\gamma_x}} \|\bar{\Delta}\| + \gamma \sqrt{\frac{1}{\gamma_x}} \|\omega\|, \quad \forall t \geq 0,
\end{aligned} \tag{3.38}$$

which implies that the $\bar{x}$-system with the pair $(\bar{\Delta}, \omega)$ as the input is input-to-state stable (Sontag, 1989). One can write

$$|e(t)| \leq \sigma_e(|\bar{x}_0|, t) + \gamma_e \|\bar{\Delta}\| + \gamma \gamma_e \|\omega\|, \tag{3.39}$$

where

$$\sigma_e(|\bar{x}_0|, t) = |C| \exp\left(-\frac{\gamma_x}{2\lambda_m(P^*)} t\right) \sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)}} |\bar{x}_0|$$

is a function of $\mathcal{KL}$ and $\gamma_e = |C|\sqrt{1/\gamma_x}$, which guarantees that the $\bar{x}$-system with e as output has SUO property with zero-offset and IOS properties (Z. P. Jiang et al., 1994). On the other hand, Assumptions 3.4.2 and 3.4.3 indicate that the $\bar{\zeta}$-system has SUO property with zero-offset and IOS properties with input-to-output gain function $\gamma_\Delta(s)$. By the nonlinear small-gain theory (Z. P. Jiang et al., 1994), under the following small-gain condition

$$(Id + \rho_2) \circ \gamma_e \circ (Id + \rho_1) \circ \gamma_\Delta(s) \leq s, \forall s \geq 0, \tag{3.40}$$

the error system (3.30)-(3.32) with $\bar{u} = -K^*\bar{x}$ is globally asymptotically stable at the origin if $\omega \equiv 0$. For a nontrivial square-integrable disturbance ω, one can achieve that the closed-loop system is input-to-output stable regarding ω as an external input. □.

3.4.3 SOLVABILITY OF GROORP

Now, we are ready to design a robust optimal controller to solve the GROORP of the partially linear system (3.23)-(3.28).

Theorem 3.3. *Under the conditions of Assumptions 3.3.1, 3.3.2-3.4.3, if weight matrices are chosen $Q = Q^T \geq \gamma_x I_n, R = I_m$ such that small-gain condition (3.40) holds, then the GROORP of the partially linear system (3.23)-(3.28) is solvable by the robust optimal controller $u = -K^*(x - X^*v) + U^*v$.*

Proof. By Theorem 3.2, the robust optimal feedback controller $\bar{u}^* = -K^*\bar{x}^*$ globally asymptotically stabilizes the error system (3.30)-(3.32) for any $v(t)$. Then, the trajectory of error system satisfies $\lim_{t\to\infty} \bar{\zeta}(t) = 0$ and $\lim_{t\to\infty} \bar{x}^*(t) = 0$ for $\omega \equiv 0$. We observe

$$\lim_{t\to\infty} e(t) = C\bar{x}^*(t) + (CX^* + F)v(t) = 0, \tag{3.41}$$

for any $x(0), \zeta(0)$. Also, it is checkable that the input-to-output stability of the closed-loop system still holds. The proof is completed.

3.5 RL ONLINE LEARNING

A novel online learning strategy is presented to solve X^*, U^* and online approximation of optimal values P^* and K^*. In Jiang's related work, he assumed Δ is available during the learning phase (Y. Jiang & Jiang, 2014). Defining $\bar{x}_i = x - X_i v$ for $i = 0, 1, 2, \cdots, h+1$ with $X_0 = 0_{n\times q}$, we have

$$\begin{aligned}\dot{\bar{x}}_i =& Ax + B_1(u+\Delta) + B_2\omega + (D - X_iE)v \\ =& A_j\bar{x}_i + B_1(K_j\bar{x}_i + z) + B_2(\omega - N_j\bar{x}_i) \\ &+ (D - \mathcal{S}(X_i))v, \end{aligned} \tag{3.42}$$

where $A_j = A - B_1K_j + B_2N_j$, $z = u + \Delta$.

Then

$$\begin{aligned}&\bar{x}_i(t+\delta t)^T P_j \bar{x}_i(t+\delta t) - \bar{x}_i(t)^T P_j \bar{x}_i(t) \\ =& \int_t^{t+\delta t} \left[\bar{x}_i^T(A_j^T P_j + P_j A_j)\bar{x}_i + 2(z + K_j\bar{x}_i)^T B_1^T P_j \bar{x}_i \right. \\ &\left. + 2v^T(D - \mathcal{S}(X_i))^T P_j \bar{x}_i + 2(\omega - N_j\bar{x}_i)^T B_2^T P_j \bar{x}_i\right] d\tau \\ =& -\int_t^{t+\delta t} \bar{x}_i^T(Q + K_j^T R K_j - \gamma^2 N_j^T N_j)\bar{x}_i d\tau \\ &+ 2\int_t^{t+\delta t} (z + K_j\bar{x}_i)^T R K_{j+1}\bar{x}_i d\tau \\ &+ 2\int_t^{t+\delta t} v^T(D - \mathcal{S}(X_i))^T P_j \bar{x}_i d\tau \\ &+ 2\gamma^2 \int_t^{t+\delta t} (\omega - N_j\bar{x}_i)^T N_{j+1}\bar{x}_i d\tau. \end{aligned} \tag{3.43}$$

For a large enough positive integer l and two vectors $a \in \mathbb{R}^{n_a}$, $b \in \mathbb{R}^{n_b}$, we define

$$\Gamma_{ab} = [\int_{t_0}^{t_1} a \otimes b \mathrm{d}\tau, \int_{t_1}^{t_2} a \otimes b \mathrm{d}\tau, \cdots, \int_{t_{l-1}}^{t_l} a \otimes b \mathrm{d}\tau]^T,$$

$$\delta_{\bar{x}_i\bar{x}_i} = [\mathrm{vecv}(\bar{x}_i(t_1)) - \mathrm{vecv}(\bar{x}_i(t_0)), \mathrm{vecv}(\bar{x}_i(t_2)) -$$

$$\mathrm{vecv}(\bar{x}_i(t_1)), \cdots, \mathrm{vecv}(\bar{x}_i(t_l)) - \mathrm{vecv}(\bar{x}_i(t_{l-1}))]^T,$$

where $t_0 < t_1 < \cdots < t_l$ are positive integers. (4.12) indicates the following equation.

$$\Psi_{ij} \begin{bmatrix} \mathrm{vecs}(P_j) \\ \mathrm{vec}(K_{j+1}) \\ \mathrm{vec}((D - \mathcal{S}(X_i))^T P_j) \\ \mathrm{vec}(N_{j+1}) \end{bmatrix} = \Phi_{ij}, \tag{3.44}$$

where

$$\begin{aligned} \Psi_{ij} &= [\delta_{\bar{x}_i \bar{x}_i}, -2\Gamma_{\bar{x}_i \bar{x}_i}(I_n \otimes (K_j^T R) - 2\Gamma_{\bar{x}_i z}(I_n \otimes R), \\ &\quad - 2\Gamma_{\bar{x}_i v}, -2\gamma^2(\Gamma_{\bar{x}_i \omega} - \Gamma_{\bar{x}_i \bar{x}_i}(I_n \otimes N_i^T))], \\ \Phi_{ij} &= -\Gamma_{\bar{x}_i \bar{x}_i} \mathrm{vec}(Q + (K_i^j)^T R K_i^j - \gamma^2 N_i^T N_i). \end{aligned}$$

Equation (3.44) is uniquely solved by the least squares method if the matrix Ψ_{ij} is of full column rank, i.e.,

$$\begin{bmatrix} \mathrm{vecs}(P_j) \\ \mathrm{vec}(K_i^{j+1}) \\ \mathrm{vec}((D - \mathcal{S}(X_i))^T P_j) \\ \mathrm{vec}(N_{j+1}) \end{bmatrix} = (\Psi_{ij}^T \Psi_{ij})^{-1} \Psi_{ij}^T \Phi_{ij}. \tag{3.45}$$

Note that D is computable by (3.45) given $\mathcal{S}(X_0) = 0$. If we seek a sequence $\alpha_2^0, \alpha_3^0, \cdots, \alpha_{h+1}^0 \in \mathbb{R}$ and a matrix $U_\dagger^0 \in \mathbb{R}^{m \times q}$ such that

$$\mathcal{S}(X_1) + \sum_{i=2}^{h+1} \alpha_i^0 \mathcal{S}(X_i) = P_j^{-1} K_{j+1} R U_\dagger^0 + D, \tag{3.46}$$

then $(X_\dagger^0, U_\dagger^0)$ is a solution to the regulator equation (3.7), where $X_\dagger^0 = X_1 + \sum_{i=2}^{h+1} \alpha_i^0 X_i$.

If the solution to (3.7) is not unique, we find all linearly independent vectors $\mathrm{vec}(\begin{bmatrix} X_\dagger^k \\ U_\dagger^k \end{bmatrix})$ by seeking sequences $\alpha_2^k, \alpha_3^k, \cdots, \alpha_{h+1}^k \in \mathbb{R}$ such that for $k = 1, 2, \cdots, H$ with $H =$

$q(m-r)$

$$X_\dagger^k = \sum_{i=2}^{h+1} \alpha_i^k X_i, \sum_{i=2}^{h+1} \alpha_i^k \mathcal{S}(X_i) = P_j^{-1} K_{j+1} R U_\dagger^k. \tag{3.47}$$

Then, we define a set:

$$\begin{aligned}\mathbb{S} =&\{(X,U)|X = X_\dagger^0 + \sum_{k=1}^{H} \beta_k X_\dagger^k, U = U_\dagger^0 + \sum_{k=1}^{H} \beta_k U_\dagger^k,\\ &\forall \beta_1, \beta_2, \cdots, \beta_H \in \mathbb{R}\}.\end{aligned} \tag{3.48}$$

Algorithm 3 RL Algorithm Algorithm for Solving GROORP

1: Select a K_0 such that $\sigma(A - B_1 K_0) \in \mathbb{C}^-$ and a threshold $\epsilon > 0$. Choose $Q = Q^T \geq \gamma_x I_n$ such that the small-gain condition holds. Compute trails $X_0, X_1, \cdots, X_{h+1}$

2: Employ $u = -K_0 x + \xi$ as the control input on $[t_0, t_l]$ with ξ an exploration noise.

3: $j \leftarrow 0, i \leftarrow 0$

4: **repeat**

5: Solve P_j, K_{j+1}, N_{j+1} from (3.45).

6: $j \leftarrow j + 1$

7: **until** $|P_j - P_{j-1}| < \epsilon$

8: Obtain the approximated optimal control gains K^* and N^*, and approximated solution P^* to (3.14)

9: **repeat**

10: Solve $\mathcal{S}(X_i)$ from (3.46)

11: $i \leftarrow i + 1$

12: **until** $i = h + 2$

13: Obtain (X^*, U^*) by solving Problem 3.3.1

14: The robust optimal controller $u = -K_j(x - X^* v) + U^* v$ and the optimal disturbance policy $\omega = N_j(x - X^* v)$ are computed.

Theorem 3.4. *Given a stabilizing $K_0 \in \mathbb{R}^{m\times n}$, if Ψ_{ij} is in full column rank for $i = 0, 1, \cdots, h+1$, $j \in \mathbb{Z}_+$, the sequences $\{P_j\}_{j=0}^{\infty}$, $\{K_j\}_{j=1}^{\infty}$ obtained from solving (3.45) converge to P^* and K^*, respectively.*

Proof. Given a stabilizing K_j, if $P_j = P_j^T$ is the solution of (4.9), K_{j+1} and N_{j+1} is determined by $K_{j+1} = R^{-1}B_1^T P_j$ and $N_{j+1} = \gamma^{-2}B_2^T P_j$, respectively. Let $T_j = (\mathcal{S}(X_i))^T P_j$. By (4.12), we know that P_j, K_{j+1} and T_j satisfy (3.45). On the other hand, let $P = P^T \in \mathbb{R}^{n\times n}$, $K \in \mathbb{R}^{m\times n}$, $N \in \mathbb{R}^{d\times n}$ and $T \in \mathbb{R}^{q\times n}$, such that

$$\Psi_{ij}\begin{bmatrix} \operatorname{vecs}(P) \\ \operatorname{vec}(K) \\ \operatorname{vec}(T) \\ \operatorname{vec}(N) \end{bmatrix} = \Phi_j.$$

Then, we have $P_j = P$, $K_{j+1} = K$, $N_{j+1} = N$, $T_j = T$. Moreover, P, K, N, T are unique when Ψ_{ij} is in full column rank. By Lemma 3.3.1, the convergence of P_j, K_j and N_j is proved.

3.6 EXAMPLE

Consider a partially linear system:

$$\begin{aligned}
\dot{\zeta} &= -\zeta^3 + \zeta e, \\
\dot{x} &= \begin{bmatrix} -1 & -2 \\ 0.5 & -2 \end{bmatrix} x + \begin{bmatrix} 2 \\ 2 \end{bmatrix} (u + v_1\zeta^2) + \begin{bmatrix} 1 \\ 4 \end{bmatrix} \omega, \\
\dot{v} &= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} v, \\
e &= x_1 + v_2.
\end{aligned}$$

In this example, for any $v \in \mathbb{R}^2$, $\boldsymbol{\zeta}(v) = 0$ satisfies the Assumption 3.4.1. Taking $V_\zeta = \zeta^2/2$, the derivative of V along the trajectories of the dynamic uncertainty is given by

$$\begin{aligned}\dot{V}_\zeta =& -\zeta^4 + \zeta^2 e \\ =& -0.5\zeta^4 - 0.5\zeta^4 + \zeta^2 e \\ \leq& -0.5\zeta^4, \quad \forall |\zeta| \geq \sqrt{\frac{|e|}{0.5}}\end{aligned} \tag{3.49}$$

Given the fact that $\Delta = \zeta^2$, it is checkable that Assumptions 3.4.2 and 3.4.3 are satisfied with gain function

$$\gamma_\Delta(s) = \frac{s}{0.5}.$$

If $\gamma_e(s) < 0.5s$, the error system (3.30)-(3.31) is guaranteed globally asymptotically stable at the origin. In this paper, we choose $Q = 5I_2, \gamma = 11$, the initial stabilizing feedback control gain matrix as $K_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$, the initial disturbance control gain as $N_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$, and the convergent criterion as $\epsilon = 10^{-8}$, and for $i = 1, 2, 3$, matrices X_i as

$$X_1 = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}, X_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, X_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The online data is collected from $t = 0s$ to $t = 15s$. After that, we iteratively compute the optimal values and convergence is attained after 6 iterations. Figs. 3.1-3.3 depicts the errors between P_j and P^*, between K_j and K^*, and N_j and N^*.

For $i = 0, 1, 2, 3$, we solve the linear map $\mathcal{S}(X)_i$ from online information. From (3.46) and (3.47), we get the set of unique solution of regulator equation, which is also the optimal solution (X^*, U^*):

$$X^* = \begin{bmatrix} 0.0000 & -1.0000 \\ 1.4999 & -1.0003 \end{bmatrix}, U^* = \begin{bmatrix} 0.9996 & -1.5002 \end{bmatrix}.$$

Then we get the robust optimal controller and optimal disturbance policy

$$u = -\begin{bmatrix} 0.8016 & 1.5239 \end{bmatrix} x + \begin{bmatrix} 3.2853 & -3.8262 \end{bmatrix} v,$$
$$\omega = \begin{bmatrix} -0.0268 & 0.0553 \end{bmatrix} x + \begin{bmatrix} 0.0829 & -0.0285 \end{bmatrix} v, \tag{3.50}$$

respectively. The learned controller is implemented after $t = 15s$. Fig. 3.4 depicts that the output of the plant asymptotically tracks the reference. Figs. 3.5-3.8 depict the trajectories of the states, the control input, the disturbance and the dynamic uncertainty respectively.

In order to validate the effect of disturbances on the cost, we change the disturbance input by

$$\omega = \frac{1}{2}\left(\begin{bmatrix} -0.0268 & 0.0553 \end{bmatrix} x + \begin{bmatrix} 0.0829 & -0.0285 \end{bmatrix} v\right). \tag{3.51}$$

We record the cost

$$\int_0^{500} [(\bar{x}^*)^T Q \bar{x}^* + (\bar{u}^*)^T \bar{u}^* - \gamma^{-2} \omega^T \omega] dt$$

for different disturbances until $t = 500s$. This is reasonable since the cost does not change significantly after $t > 500s$ for a stabilized system. It is obtained that the cost under disturbance (3.51) has reduced by 21.7296 compared with the cost under (3.50).

3.7 Conclusion

This paper proposes a novel control approach for global optimal output regulation of a class of partially linear systems with an exosystem and nonlinear dynamic uncertainties. By using reinforcement learning, a data-driven control strategy is proposed for designing robust adaptive optimal controllers and an optimal disturbance policy to achieve the rejection of nonvanishing disturbance, forcing the output to asymptotically track a desired output. The obtained simulation results ascertain the effectiveness of the proposed approach.

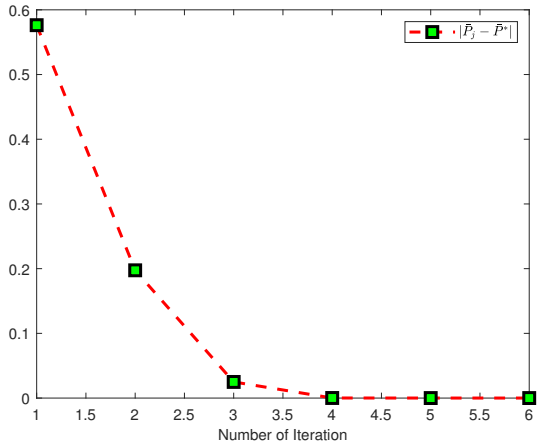

Figure 3.1: Convergence of P_j to its Optimal Value P^* during the Learning Process

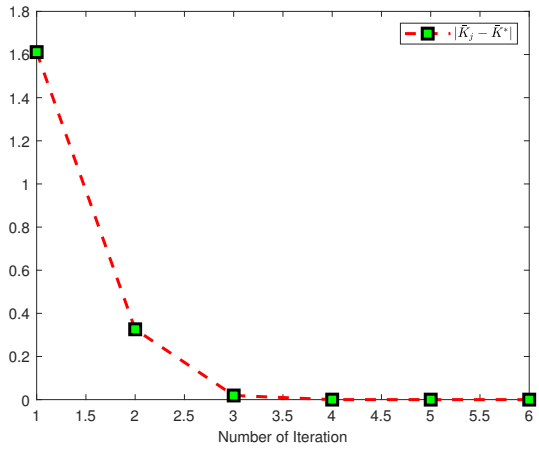

Figure 3.2: Convergence of K_j to its Optimal Value K^* during the Learning Process

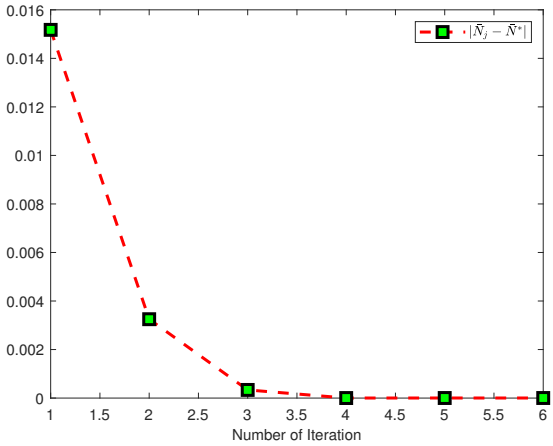

Figure 3.3: Convergence of N_j to its Optimal Value N^* during the Learning Process

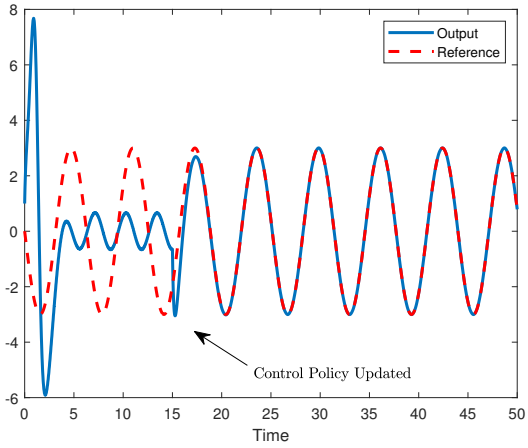

Figure 3.4: Trajectories of the Output and Reference

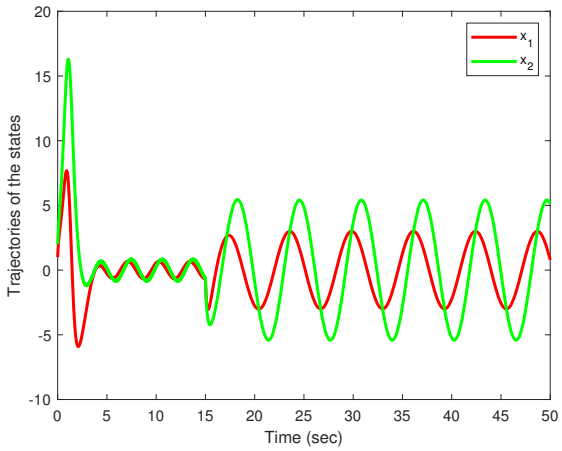

Figure 3.5: Trajectories of the States

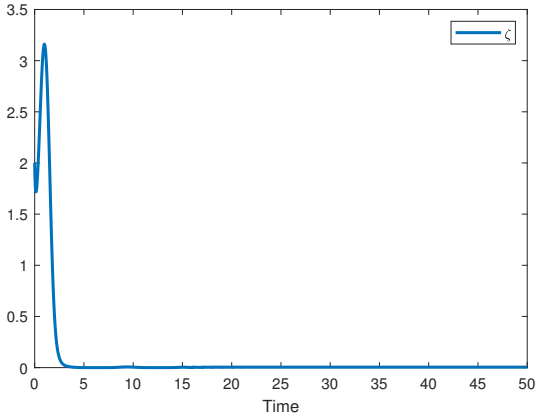

Figure 3.6: Trajectory of the Dynamic Uncertainty $\zeta(t)$

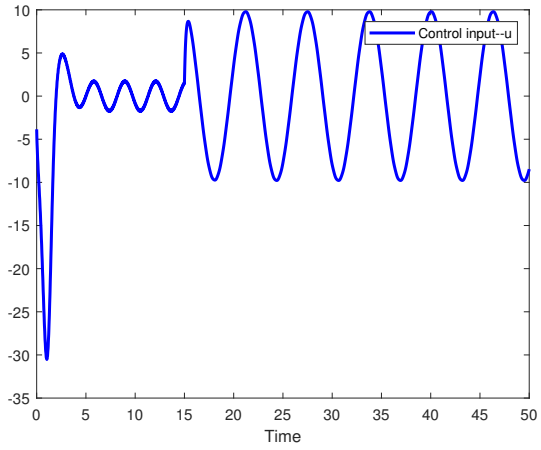

Figure 3.7: Trajectory of Control Input $u(t)$

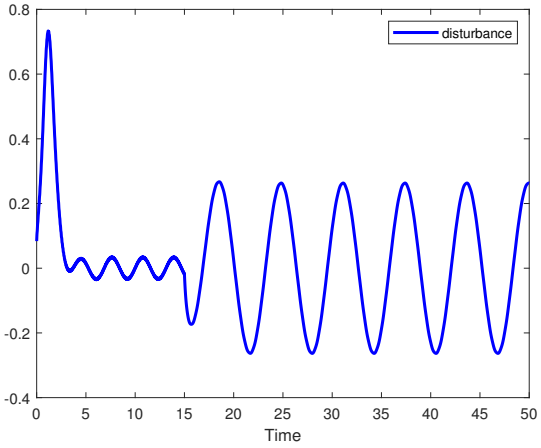

Figure 3.8: Trajectory of the Disturbance $\omega(t)$

CHAPTER 4

PREDICTIVE CRUISE CONTROL OF CONNECTED AND AUTONOMOUS VEHICLES

4.1 ABSTRACT

Predictive cruise control concerns designing controllers for autonomous vehicles using the broadcasted information from the traffic lights such that the idle time around the intersection can be reduced. This paper proposes a novel adaptive optimal control approach based on reinforcement learning to solve the predictive cruise control problem of a platoon of connected and autonomous vehicles. First, the reference velocity is determined for each autonomous vehicle in the platoon. Second, a data-driven adaptive optimal control algorithm was developed to approximate the optimal control gains of a desired distributed controller without the exact knowledge of system dynamics. The obtained controller is able to regulate the headway, velocity and acceleration of each vehicle in an optimal sense. The goal of trip time reduction is achieved without compromising vehicle safety and passenger comfort. Numerical simulations are presented to validate the efficacy of the proposed methodology.

4.2 INTRODUCTION

It is reported by the US Department of Transportation (DOT) that around ten percent of traffic delays are due to poor traffic signal scheduling (*The intelligent transportation systems for traffic signal control deployment benefits and lessons learned*, 2007), which results unnecessary fuel burning and heavy pollution to the public environment. Recently, several traffic intersections have been equipped with advanced traffic signal (ATS) control (X. Sun et al., 2016) to help drivers save fuel and increase the mobility. However, it is noteworthy that ATS has high a cost of implementation and maintenance, and usually fails to

provide the traffic light schedule. To solve the problems regarding traffic delay and pollution, the traveling vehicles need to have an insight on the timing of traffic lights switching which the traffic lights broadcast periodically to oncoming vehicles, i.e., let traffic light speak. This communication method between traffic lights and vehicles is called vehicle-to-infrastructure (V2I) communication. These vehicles are also allowed to exchange their information in a wireless manner, which refers to vehicle-to-vehicle (V2V) communication. Both V2I and V2V communications are crucial technologies in the field of connected vehicles.

Autonomous vehicle technologies aim at reducing fuel consumption and increasing traffic safety. By integration of the recent wireless vehicular networking technology in connected vehicles, the connected and autonomous vehicles (CAV) technology (Y. J. Zhang, Malikopoulos, & Cassandras, 2016; Talebpour & Mahmassani, 2016) is under extensive investigation, which is expected to prevent secondary crashes, reduce property damage and injury, congestions and emissions. There are several existing works that contribute to the development of CAV. For instance, cooperative adaptive cruise controllers (CACC) have been designed for a longitudinal platoon of CAV (Gao, Jiang, & Ozbay, 2017; Gao, Rios, Tong, & Chen, 2017; Oncu, Ploeg, van de Wouw, & Nijmeijer, 2014; Desjardins & Chaib-draa, 2011; Guo & Yue, 2014). The effectiveness of CACC on the safety, traffic flow, and environment has also been tested in different traffic scenarios with human-driven and autonomous vehicles (van Arem, van Driel, & Visser, 2006; Shladover, Su, & Lu, 2012). In Gao's previous work, he implemented an optimal CACC algorithm for buses on the exclusive bus lane of the Lincoln tunnel corridor (Gao, Jiang, Ozbay, & Gao, 2018). Micro-traffic simulation results have shown that, using the proposed algorithm, the travel times of buses on the exclusive bus lane are close to the present day travel times even when the traffic demand is increased by 30%. Cooperative vehicle intersection control (CVIC) is another approach of CAV (Lee & Park, 2012). The objective of CVIC is to let vehicles

automatically run across the intersection without requiring the traffic signals. Simulation results in (Lee & Park, 2012) demonstrate that CVIC is able to potentially decrease the traffic pollution and delay. However, to completely remove the traffic lights may not be easy to realize in the near future.

Recently, the predictive cruise control (PCC) has been gaining a lot of attention. The main idea is to reduce the idle time of vehicles by using the upcoming traffic signal information (Asadi & Vahidi, 2011; Kavurucu & Ensar, 2017; Alrifaee, Jodar, & Abel, 2015). The PCC can help promote a smooth traffic by decreasing the use of breaks, and increase the safety by excluding the red-light violation. However, there are two open issues of PCC in practice:

1) The existing PCC is designed via model-based control method, such as model predictive control (Alrifaee et al., 2015; Asadi & Vahidi, 2011). It is generally known that to obtain a system model accurately is hard work. The model-based control approach may even destabilize the system given an inaccurate model. Due to the significant development of information, communication and sensing technologies, the vehicles' position, velocity, and acceleration data and the upcoming traffic signal are available for feedback. In order to address this issue regarding the model uncertainty, data-driven control approaches can be employed to design PCCs.

2) Existing PCCs are usually studied for an individual vehicle. It will be interesting that this strategy can be generalized to a platoon of CAVs. There are two major challenges for this generalization. The first challenge is how to determine the reference velocity of each vehicle in the platoon such that safety is ensured while the vehicles in the platoon maintain desirable inter-vehicle distance. The second challenge is related to the control structure for CAVs. A platoon of CAVs can be considered as a class of autonomous multi-agent systems. We will design a distributed control strategy (Fang, Wu, & Yang, 2016), instead of centralized control. This is because traditional

centralized control design usually relies on the accessibility of all the agents (vehicles). Besides higher use of communication resources, centralized control is fragile as malfunction of only one agent may threaten the safety and reliability of the whole platoon.

In this paper, we propose a novel data-driven distributed PCC approach for a platoon of CAV. Our contributions are twofold: First, we propose a novel reference velocity determination approach based on the number and desired headways of vehicles, and traffic light schedule to reduce idle time for the whole platoon at stop lights. Second, without relying on the accurate knowledge of system dynamics, reinforcement learning (RL) is employed to obtain an optimal controller that regulates the headway, velocity and acceleration of CAV. RL is a practically-sound, data-driven computational approach which is inspired and adapted from human decision-making processes (Sutton & Barto, 1998; Vamvoudakis, 2014; Fan & Yang, 2016; Wang et al., 2016; Y. Jiang et al., 2017). It is essentially a direct adaptive optimal control approach that can be employed to learn or approximate the optimal control policy of the system without a priori knowledge of the system model while maintaining the system stability. Ozbay developed data-driven optimal controllers for CAVs on the highway via RL (Gao, Jiang, & Ozbay, 2017; Gao & Jiang, 2016a; Meng et al., 2015). In this paper, we will move forward to the PCC problem for urban traffic through taking the upcoming traffic signals into consideration.

The remainder of this paper is organized as follows. In Section 4.3, we formulate the procedure to obtain reference velocities for CAVs. A data-driven RL algorithm to learn optimal control gains of the distributed controller will be present in Section 4.4 to ensure all CAVs track their references. Section 4.5 includes simulation results using a platoon of four vehicles for illustration. Conclusions are drawn in Section 4.6.

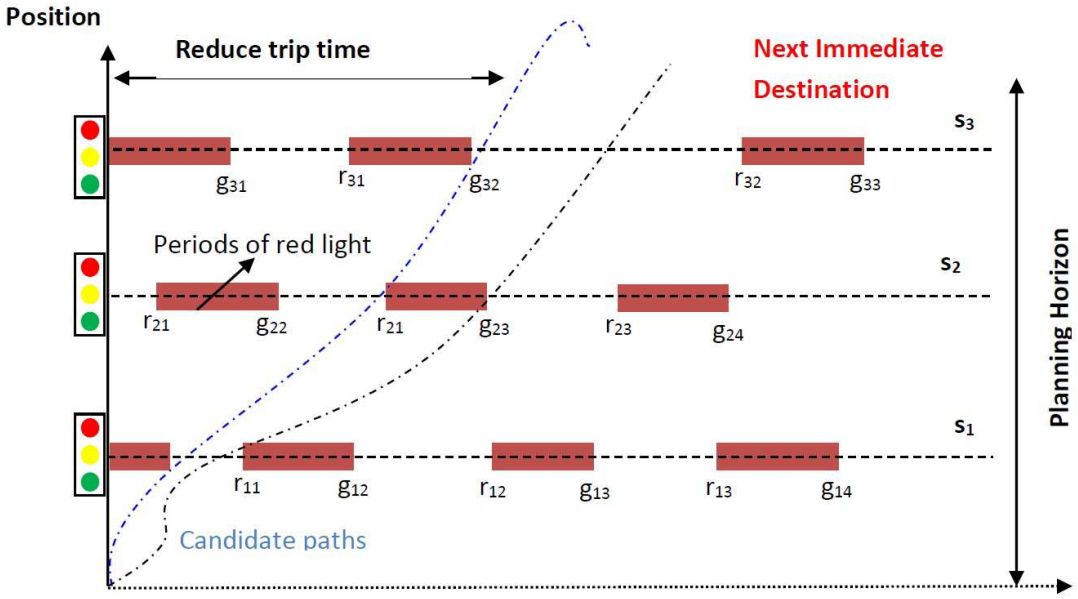

Figure 4.1: Space-Time graph

4.3 Reference Velocity Determination

In this section, we determine the reference velocity for the vehicular platoon based on the present cruising velocity, desired headway, vehicle length and information received from the impending traffic lights. Our interest is to find a maximum allowable velocity such that the vehicle arrives on green while avoiding waiting on red.

Fig. 4.1, a space-time graph, is depicted to aid the explanation. The instantaneous distance of the vehicle to the impending traffic light is assumed known. It is denoted as s_i where i is the traffic light index depicted in the Fig. 4.1. Periodically, the lights broadcast the light sequence schedule to upcoming vehicles. r_{ij} and g_{ij} are the jth red and green of the ith traffic light. For instance, the information $[g_{i1}, r_{i1}, g_{i2}, r_{i2}, g_{i3}, \cdots] = [30, 70, 100, 130, 160, \cdots]$ broadcasted to approaching vehicles is interpreted as follows. The ith light is presently on red which will elapse for 30s, then will be on green for 40s, then 30s on red and so forth. This time scheduling will be used to determine the reference

velocity to avoid stopping at red given the knowledge of the positions of vehicles. The reference velocity should be in $[V_{\min}, V_{\max}]$ where $V_{\min}$ is the local minimum velocity limit and the $V_{\max}$ is the allowed local maximum velocity.

The reference velocity for an individual vehicle can be determined via the following phases:

Phase 1 The velocity of the vehicle should stay in $[s_1/r_{1j}, s_1/g_{1j}]$ so that it is able to pass at the first green of the first light.

To avoid the red light depends on if there are intersection between these two intervals, i.e.,

$$v^* \in [\frac{s_1}{r_{1j}}, \frac{s_1}{g_{1j}}] \cap [V_{\min}, V_{\max}]$$

where v^* is the desired velocity. This is a mathematical description of the process stated above. This helps determine the possibility of passing during the next green. The process continues by letting $j \leftarrow j+1$ until the expression gives a nontrivial set. Set $j^* \leftarrow j$.

To get a better understanding of the scenario, assume the local velocity limits are $V_{\min} = 10m/s$, and $V_{\max} = 25m/s$. The distance to the traffic light $i = 1$ is $s_1 = 2000m$ and the following information is broadcasted

$$\begin{aligned} g_{11} =& 10s, r_{11} = 30s, g_{12} = 50s, \\ r_{12} =& 80s, g_{13} = 110s, r_{13} = 140s. \end{aligned}$$

Then, we have $[s_1/r_{11}, s_1/g_{11}] = [67, 200]m/s$. Obviously, this does not intersect with the interval $[V_{\min}, V_{\max}] = [10, 25]m/s$, which means this path is not a good choice if stopping at red is to be avoided, then we proceed to the next possible path, i.e., $[s_1/r_{12}, s_1/g_{12}] = [25, 40]m/s$ which does not intersect with the local velocity limits. Moving to the third interval $[s_1/r_{13}, s_1/g_{13}] = [14.3, 18]m/s$, it is found that

the intersection set between the third interval and the velocity limit is $[10, 18]m/s$. Therefore, the velocity of the vehicle is chosen $10 \leq v^* \leq 18$, which allows the vehicle to pass the first light without having to stop.

Phase 2 The above process is repeated for subsequent traffic lights $j^*+1, j^*+2, \cdots$ to obtain the best possible paths.

Phase 3 The intersection obtained from Phases 1 and 2 ranges from the minimum to the maximum possible velocities to reduce the number of stops at red since our objectives is to reduce trip time. We set the reference velocity to the highest possible velocities obtained from the intersection of ranges. The main objective of this system is to reduce the distance between the driving vehicles to increase traffic flow with minimum or no disturbances throughout the platoon of vehicles.

Considering that our interest is on a platoon of connected vehicles, the members of the platoon are connected through the vehicle-following objective. Every member follows its immediate preceding vehicle while maintaining a desired distance. Time headway spacing policy will be considered in this work. It can be mathematically represented as follows:

$$h_k^* = \tau_k v^* + d_k \tag{4.1}$$

where h_k^* is the desired headway between the front bumper of a vehicle k and the rear bumper of the immediate preceding vehicle $k-1$. d_k is the gap between the vehicles k and $k-1$ at standstill, τ_k is the headway time constant of vehicle k, and v^* is the vehicles' desired velocity.

The above three-phase approach can be generalized to determine the reference velocity of this platoon with some modifications. Suppose we have n autonomous vehicles with length of vehicle k being l_k, and all the vehicles are operating at their desired velocities. If the head of the leader passing through a specific position at $t = T$, then the tail of the last

vehicle will pass through the same position at $t = T + \Delta t(v^*)$, where

$$\Delta t(v^*) = \tau + \frac{d+l}{v^*}, \tau = \sum_{k=2}^{n} \tau_k, d = \sum_{k=2}^{n} d_k, l = \sum_{k=1}^{n} l_k.$$

In this setting, the desired velocity should satisfy the following inequality to allow the whole platoon go across the ith intersection and jth green light

$$v^* \geq \frac{s_i}{r_{ij} - \Delta t}.$$

When $\Delta t < r_{ij}$, this inequality is equivalent to

$$v^* \geq \frac{s_i + d + l}{r_{ij} - \tau}.$$

The velocity interval of the vehicles will be modified by

$$v^* \in [\frac{s_i + d + l}{r_{ij} - \tau}, \frac{s_i}{g_{ij}}] \cap [V_{\min}, V_{\max}],$$

which will be compared with the local velocity. The remaining part will be similar to the three-phase approach for an individual vehicle.

Remark 4.3.1. *Interestingly, for a large platoon, d and l may be large enough such that $\tau > r_{ij}$. Consequently, there is the possibility of insufficient time to allow all the vehicles to go through during one green light. An alternative method is to split the large platoon in multiple sub-platoons such that each platoon can be manipulated flexibly.*

4.4 Data-Driven Control Algorithm Design via RL

In this section, we propose a data-driven adaptive optimal control algorithm for CAV to track the determined reference velocity. To begin with, we define h_k and v_k as the headway and the velocity of the kth vehicle, respectively. Moreover, let $\Delta h_k = h_k - h_k^*$ and $\Delta v_k = v_{k-1} - v_k$. The vehicle dynamics can be represented by

$$\dot{x}_k = A_k x_k + B_k u_k + D x_{k-1} \tag{4.2}$$

where u_k is the control input of vehicle k, the system variables $x_k = [\Delta h_k, \Delta v_k, a_k]$ are the headway error, velocity error and acceleration of the vehicle.

System matrices $A_k \in \mathbb{R}^{3\times3}$, $B_k \in \mathbb{R}^{3\times1}$, $D \in \mathbb{R}^{3\times3}$ can be respectively defined as

$$A_k = \begin{bmatrix} 0 & 1 & -\tau_k \\ 0 & 0 & -1 \\ 0 & 0 & -\frac{1}{T_k} \end{bmatrix}, B_k = \begin{bmatrix} 0 \\ 0 \\ \frac{G_k}{T_k} \end{bmatrix}, D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.3}$$

with G_k and T_k uncertain system parameters.

Remark 4.4.1. *It is checkable that the pair (A_k, B_k) is stabilizable and all the eigenvalues of A_k stay in the closed left-half complex plane.*

Given the system (4.2), define a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. $\mathcal{V} = \{0, 1, 2, \cdots, n\}$ is the node set with the node 0 as the fictitious leader running in a constant velocity v^*, which implies that $x_0 = 0$. $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ refers to the edge set. $\mathcal{A} = [a_{kj}] \in \mathbb{R}^{(n+1)\times(n+1)}$ called the adjacency matrix. Denote $\mathcal{N}_k$ the set of all the nodes j such that $(j, k) \in \mathcal{E}$. The adjacency matrix has the property such that $a_{kj} > 0$ if $(j, k) \in \mathcal{E}$ and $a_{kj} = 0$ otherwise.

Our objective is to design a distributed state-feedback controller

$$u_k = -K_k^* \lambda_k \tag{4.4}$$

where, for $k = 1, 2, \cdots, n$, if $\mathcal{N}_k \neq \phi$, we let

$$\lambda_k = \sum_{j\in\mathcal{N}_k} a_{kj}(x_k - x_j). \tag{4.5}$$

Otherwise, we simply set $\lambda_k = x_k$. The control gain K_k^* is defined as

$$K_k^* = R_k^{-1} B_k^T P_k^* \tag{4.6}$$

where P_k^* is solved from the following algebraic Riccati equation (ARE)

$$A_k^T P_k + P_k A_k + Q_k - P_k B_k R_k^{-1} B_k^T P_K = 0. \tag{4.7}$$

From optimal control theory, K_k^* is an optimal control gain with respect to the following cost function

$$J = \int_0^\infty (x_k^T Q_k x_k + u_k^T R_k u_k) d\tau \tag{4.8}$$

and the system (4.2) in the absence of x_{k-1}, where $Q_k = Q_k^T \geq 0, R_k = R_k^T > 0$ with $(A_k, \sqrt{Q_k})$ observable.

If both A_k and B_k are accurately known, one can employ the classical linear optimal theory to seek the optimal control gain by solving the ARE (4.7).

Notice that the ARE is a nonlinear equation, solving it is computationally expensive, especially for a large scale system. A numerical algorithm is proposed by Kleinman (Kleinman, 1968) to approximate the solution P^*. The algorithm starts from a stabilizing K_{k0}. For each $l \in \mathbb{Z}_+$, we solve for the positive-definite matrix $P_{kl} = P_{kl}^T$ that satisfies the Lyapunov equation.

$$\begin{aligned}(A_k - B_k K_{kl})^T P_{kl} + P_{kl}(A_k - B_k K_{kl}) \\ + Q_k + K_{kl}^T R_k K_{kl} = 0.\end{aligned} \tag{4.9}$$

Then, we update $K_{k,l+1}$ by

$$K_{k,l+1} = R_k^{-1} B_k^T P_{kl}. \tag{4.10}$$

Kleiman proved that $\lim_{l\to\infty} P_{kl} = P_k^*$ and $\lim_{l\to\infty} K_{kl} = K_k^*$

Rewrite the state equation (4.2) as

$$\dot{x}_k = (A_k - B_k K_{kl}) x_k + B_k(K_{kl} x_k + u_k) + D x_{k-1}. \tag{4.11}$$

Then, by the Lyapunov equation (4.9), we have

$$
\begin{aligned}
& x_k^T(t+\delta t)P_{kl}x_k(t+\delta t) - x_k^T(t)P_{kl}x_k(t) \\
=& \int_t^{t+\delta t} x_k^T[(A_k - B_kK_{kl})^TP_{kl} + P_{kl}(A_k - B_kK_{kl})]x_k d\tau \\
& + 2\int_t^{t+\delta t} (u_k + K_{kl}x_k)^T B_k^T P_{kl}x_k d\tau \\
& + 2\int_t^{t+\delta t} x_{k-1}^T D^T P_{kl}x_k d\tau \\
=& -\int_t^{t+\delta t} x_k^T(Q_k + K_{kl}^T R_k K_{kl})x_k d\tau \\
& + 2\int_t^{t+\delta t} (u_k + K_{kl}x_k)^T R_k K_{k,l+1}x_k d\tau \\
& + 2\int_t^{t+\delta t} x_{k-1}^T D^T P_{kl}x_k d\tau. \qquad (4.12)
\end{aligned}
$$

Employing the idea of Kronecker product, the components in equation (4.12) can be simplified as

$$
\begin{aligned}
x_k^T(Q_k + K_{kl}^T R_k K_{kl})x_k =& (x_k^T \otimes x_k^T)\mathrm{vec}(Q_k + K_{kl}^T R_k K_{kl}), \\
(u_k + K_{kl}x_k)^T R_k K_{k,l+1}x_k =& [(x_k^T \otimes x_k^T)(I \otimes K_{kl}^T R_k) \\
& + (x_k^T \otimes u_k^T)(I \otimes R_k)]\mathrm{vec}(K_{k,l+1}), \\
x_{k-1}^T D^T P_{kl}x_k =& (x_k^T \otimes x_{k-1}^T)(I \otimes D^T)\mathrm{vec}(P_{kl}).
\end{aligned}
$$

For a large integer $s > 0$, define

$$\begin{aligned}
\alpha_k =&[\text{vecv}(x_k(t_1)) - \text{vecv}(x_k(t_0)), \text{vecv}(x_k(t_2)) \\
&- \text{vecv}(x_k(t_1)), \cdots, \text{vecv}(x_k(t_s)) - \text{vecv}(x_k(t_{s-1}))]^T, \\
f_k =&[\int_{t_0}^{t_1} x_k \otimes x_k d\tau, \int_{t_1}^{t_2} x_k \otimes x_k d\tau, \cdots, \int_{t_{s-1}}^{t_s} x_k \otimes x_k d\tau]^T, \\
g_k =&[\int_{t_0}^{t_1} x_k \otimes u_k d\tau, \int_{t_1}^{t_2} x_k \otimes u_k d\tau, \cdots, \int_{t_{s-1}}^{t_s} x_k \otimes u_k d\tau]^T, \\
h_k =&[\int_{t_0}^{t_1} x_k \otimes x_{k-1} d\tau, \int_{t_1}^{t_2} x_k \otimes x_{k-1} d\tau, \cdots, \\
&\int_{t_{s-1}}^{t_s} x_k \otimes x_{k-1} d\tau]^T
\end{aligned}$$

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

Equation (4.12) implies the following linear equation

$$\beta_{kl} \begin{bmatrix} \text{vec}(P_k) \\ \text{vec}(K_{k+1}) \end{bmatrix} = \gamma_{kl}, \tag{4.13}$$

where

$$\begin{aligned}
\beta_{kl} &= [\alpha_k - 2h_k(I \otimes D^T), -2f_k(I \otimes K_{kl}^T R_k) - 2g_k(I \otimes R_k)], \\
\gamma_{kl} &= -f_k \text{vec}(Q_k + K_{kl}^T R_k K_{kl}).
\end{aligned}$$

We can directly solve equation (4.13) if β_{kl} has full column rank

$$\begin{bmatrix} \text{vecs}(P_{kl}) \\ \text{vec}(K_{k,l+1}) \end{bmatrix} = (\beta_{kl}^T \beta_{kl})^{-1} \beta_{kl}^T \gamma_{kl}. \tag{4.14}$$

Remark 4.4.2. *It is shown in reference (Gao, Jiang, Lewis, & Wang, 2017) that, under some mild conditions of persistent excitation, given a stabilizing control gain K_{kl}, one can uniquely find P_{kl} and the improved control gain $K_{k,l+1}$ through collected online data $x_k(t), x_{k-1}(t)$ and $u_k(t)$. This process does not rely on the knowledge of system matrices A_k and B_k.*

Algorithm 4 RL Algorithm

$k \leftarrow 1$

repeat

  Apply an initial control policy $u_k = -K_{k0}x + \theta_k$ with exploration noise θ_k and $A_k - B_k K_{k0}$ a Hurwitz matrix

  $l \leftarrow 0$

  repeat

    Solve P_{kl} and $K_{k,l+1}$ from (4.14) via online input-state data.

    $l \leftarrow l + 1$

  until $\|P_{kl} - P_{k,l-1}\| < \epsilon_k$ with ϵ_k a small positive constant.

  $l^* \leftarrow l$

  Obtain the following suboptimal controller

$$u_l = -K_{k,l^*}\lambda_k \tag{4.15}$$

  $k \leftarrow k + 1$

until $k = n + 1$

Remark 4.4.3. *Like other ADP methods (Gao & Jiang, 2016a; Lewis & Vamvoudakis, 2011; Gao & Jiang, 2017), the exploration noise θ_k is introduced to excite the system such that the matrix β_{kl} has full column rank. Example of the exploration noise includes sinusoidal signals, and random noise.*

Theorem 4.1. *The sequences $\{P_{kl}\}_{l=0}^{\infty}$ and $\{K_{kl}\}_{l=1}^{\infty}$ computed by Algorithm 4 converge to P_k^* and K_k^*.*

Proof. Let $P_{kl} = (P_{kl})^T > 0$ be the unique solution to (4.9). $K_{k,l+1}$ is uniquely determined by $K_{k,l+1} = R_k^{-1}B_k^T P_{kl}$. On the other hand, letting $\hat{P}$ and $\hat{K}$ solve (4.13), then $P_{kl} = \hat{P}$, $K_{k,l+1} = \hat{K}$ are uniquely determined based on elegant choice of exploration noise. By

Kleiman's method, we have $\lim_{l\to\infty} K_{kl} = K_k^*$, $\lim_{k\to\infty} P_l = P_k^*$. The convergence of sequences $\{P_{kl}\}_{l=0}^{\infty}$ and $\{K_{kl}\}_{l=1}^{\infty}$ obtained by RL Algorithm 4 is thus technically guaranteed. □

The data-driving learning Algorithm 4 requires no prior knowledge of state matrices. Also, the convergence to the ideal optimal control gain obtained from the linear quadratic regulator (LQR) gives the implementation of this algorithm a high confidence level. Besides the convergence of the algorithm, the stability of the closed-loop system is another important component that need to be analyzed given the fact that the safety of vehicles is usually related with the system stability.

Theorem 4.2. *The CAV system (4.2) in closed-loop with the distributed controller (4.15) learned by data-driven control Algorithm 4 is asymptotically stable.*

Proof. The closed-loop system (4.2) with (4.15) can be written in a compact form

$$\dot{x} = A_L x \tag{4.16}$$

where the state $x = [x_1^T, x_2^T, \cdots, x_n^T]^T$. It is checkable that A_L is a block lower-triangular matrix with Hurwitz sub-matrices on the diagonal. This implies that A_L is a Hurwitz matrix. One can immediately see that the closed-loop system is exponentially stable. □

Recall that the cost function was defined in (4.8), Let the matrices Q and R be

$$Q = \text{blockdiag}(Q_1, Q_2, \cdots, Q_n),$$
$$R = \text{blockdiag}(R_1, R_2, \cdots, R_n).$$

By linear optimal control theory, the optimal controller is $u = -K^* x$ such that the cost function is minimized $J^* = x^T P^* x$, where

$$K^* = \text{blockdiag}(K_1^*, K_2^*, \cdots, K_n^*),$$
$$P^* = \text{blockdiag}(P_1^*, P_2^*, \cdots, P_n^*).$$

On the other side, there exists a bounded, positive definite matrix P_b such that the cost of the closed-loop system is $J_c = x^T P_b x$. By selecting μ as the largest eigenvalue of matrix $P^* P_b^{-1}$, we have $J_c \leq \mu^{-1} J^*$. As a result, the learned controller (4.15) is a suboptimal controller.

4.5 SIMULATION RESULTS

This section presents the result of simulation carried out to validate the effectiveness of the proposed algorithm. In this example, we consider a platoon of 4 autonomous vehicles with different dynamics. The communication topology of the vehicles is depicted in Fig. 4.2. The system parameters of vehicles are illustrated in Tab. 4.1. The initial position, velocity, and acceleration of vehicles are shown in Tab. 4.2. The allowed maximum and minimum velocity are $V_{\max} = 30m/s$ and $V_{\min} = 0m/s$. There are four traffic intersection located at $11000m, 12000m, 13000m$ and $14000m$, respectively. The specific timing information can be referred to Fig. 4.8.

We first determine the reference velocity for vehicles given the upcoming traffic light information based on the three-phase approach proposed in section 4.3. The reference velocity at each intersection is calculated as $30m/s$, $19.35m/s$, $22.22m/s$ and $11.76m/s$, respectively.

Then, the online input and state data are collected from $t = 0s$ to $t = 6s$. In the simulation, for $k = 1, 2, 3, 4$, $R_k = 1$ and $Q_k = \mathrm{diag}([0.01, 0.01, 0.01])$ are chosen as the weights in the performance index (4.8). The Algorithm 4 is implemented to approximate the optimal distributed control gain K_k^*. Figs. 4.3-4.6 depict the convergence of optimal values during the learning process. The convergence criterion is selected by $\epsilon_k = 10^{-10}$ for $k = 1, 2, 3, 4$. It is checkable that the convergence criterion is satisfied for all of vehicles with less than 10 iterations. We update by the learned near-optimal distributed control gains after $t = 6s$. The trajectories of vehicles using the designed PCC strategy are depicted in

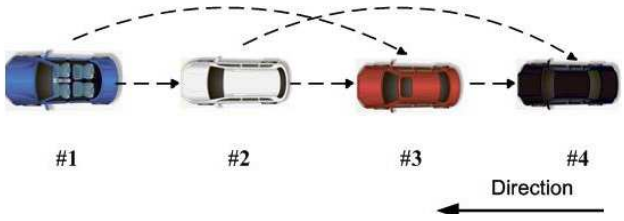

Figure 4.2: Communication Topology of Vehicles

Fig. 4.8 and its zoom-out Fig. 4.9. The velocities of vehicles are depicted in Fig. 4.7. It can be observed that, without accurate knowledge of system parameters, the designed controller is able to track the desired trajectory timely and reliably, which attests to the safety of the designed control policy. Moreover, by proper determination on reference velocity and design of data-driven controllers, one can see that the vehicles are able to go across the traffic intersection without unnecessary stopping which increases the traffic mobility.

4.6 Conclusions

This paper has studied the predictive cruise control problem for a platoon of connected and autonomous vehicles. Theoretical steps are proposed for planning of the reference velocity of the vehicles in the platoon. The reinforcement learning strategy is employed to develop a distributed optimal state-feedback controller. Simulation results show that the obtained controller is able to regulate the headway, velocity and acceleration of each vehicle to follow the desired paths while reducing the trip time of each vehicles.

Table 4.1: System Parameters

Parameter	Value	Parameter	Value	Parameter	Value
τ_1	1.5	G_{L1}	2.0	T_{L1}	0.20
τ_2	1.5	G_{L2}	2.2	T_{L2}	0.24
τ_3	1.5	G_{L3}	2.4	T_{L3}	0.29
τ_4	1.5	G_{L4}	2.6	T_{L4}	0.34
d_1	3	l_1	2.5		
d_2	3	l_2	3		
d_3	3	l_3	3		
d_4	3	l_4	3.5		

Table 4.2: Initial Values of Vehicles

Vehicle #	Position $[m]$	Velocity $[m/s]$	Acceleration $[m/s^2]$
#1	10000	29	0
#2	9900	27	0
#3	9700	29	0
#4	9400	32	0

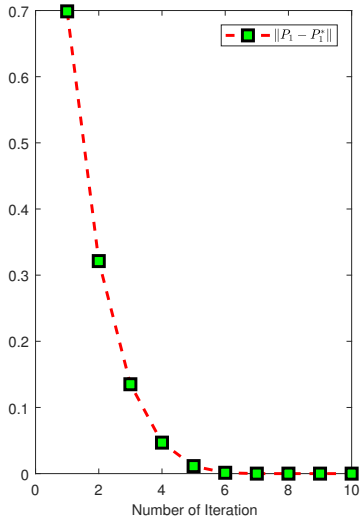

Figure 4.3: Convergence of the Optimal Value of Vehicle #1 during the Learning Process

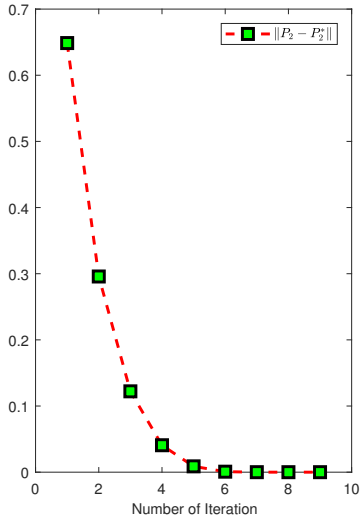

Figure 4.4: Convergence of the Optimal Value of Vehicle #2 during the Learning Process

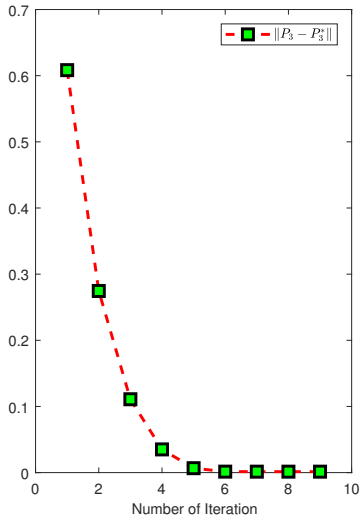

Figure 4.5: Convergence of the Optimal Value of Vehicle #3 during the Learning Process

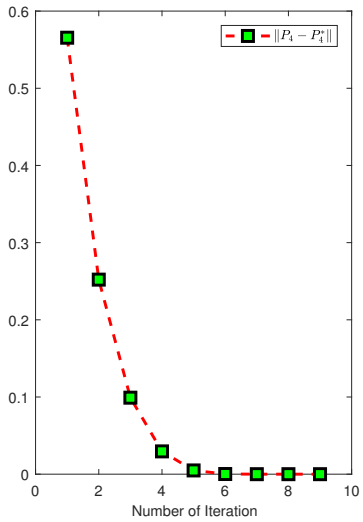

Figure 4.6: Convergence of the Optimal Value of Vehicle #4 during the Learning Process

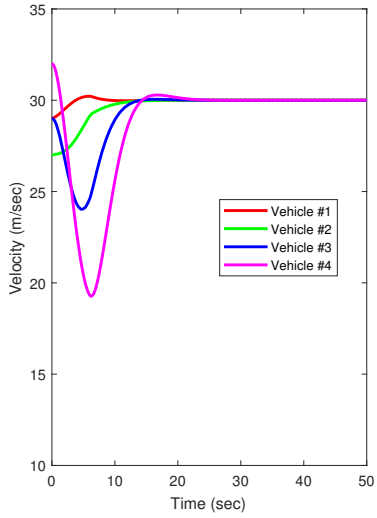

Figure 4.7: Velocity of Vehicles

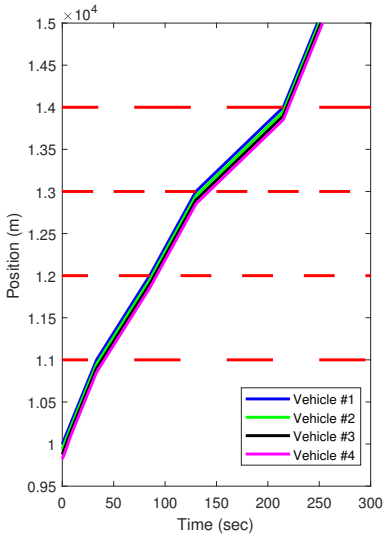

Figure 4.8: Trajectories of Vehicles

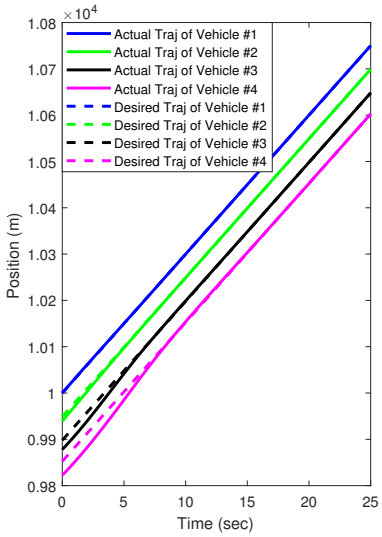

Figure 4.9: Trajectories of Vehicles (Zoom Out)

CHAPTER 5

CONCLUSIONS

We have developed RL methods for H_∞ output regulation problems for uncertain linear systems and uncertain partially linear systems. A novel data-driven adaptive optimal control approach was developed for solving the game output regulation problem of a continuous-time system. The obtained controller is able to asymptotically track a reference while both the modeled and unmodeled disturbances were rejected. The important feature of this computational mechanism is that it does not rely on any priori knowledge of the systems dynamics.

The RL methods were also developed for the predictive cruise control of connected and autonomous vehicles to design a controller which was able to regulate the headway, speed and the acceleration of the vehicles to their desired values.

REFERENCES

Alrifaee, B., Jodar, J. G., & Abel, D. (2015). Predictive Cruise Control for Energy saving in Reev using V2I Information. In *Mediterranean conference on control and automation* (p. 82-87). doi: 10.1109/MED.2015.7158733

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2007). Model-free Q-learning Designs for Linear Discrete-Time Zero-Sum Games with Application to H-Infinity Control. *Automatica*, *43*(3), 473-481.

Asadi, B., & Vahidi, A. (2011). Predictive Cruise Control: Utilizing Upcoming Traffic Signal Information for Improving Fuel Economy and Reducing Trip Time. *IEEE Transactions on Control Systems Technology*, *19*(3), 707-714.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems. *IEEE transactions on systems, man, and cybernetics*(5), 834–846.

Başar, T., & Bernhard, P. (2008). *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer Science & Business Media.

Bonivento, C., Marconi, L., & Zanasi, R. (2001). Output Regulation of Nonlinear Systems by Sliding Mode. *Automatica*, *37*(4), 535 - 542.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. New York, NY: Cambridge University Press.

Byrnes, C. I., Priscoli, F. D., Isidori, A., & Kang, W. (1997). Structurally Stable Output Regulation of Nonlinear Systems. *Automatica*, *33*(3), 369 - 385.

Desjardins, C., & Chaib-draa, B. (2011). Cooperative Adaptive Cruise Control: A Reinforcement Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1248-1260.

Fan, Q. Y., & Yang, G. H. (2016). Adaptive Actor-Critic Design-Based Integral Sliding-Mode Control for Partially Unknown Nonlinear Systems With Input Disturbances. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(1), 165-177. doi: 10.1109/TNNLS.2015.2472974

Fang, H., Wu, D., & Yang, T. (2016, Dec). Cooperative Management of A Lithium-ion Battery Energy Storage Network: A Distributed MPC Approach. In *2016 ieee 55th conference on decision and control (cdc)* (p. 4226-4232). doi: 10.1109/CDC.2016.7798911

Francis, B. (1977). The Linear Multivariable Regulator Problem. *SIAM Journal on Control and Optimization*, *15*(3), 486-505.

Gao, W., Jiang, Y., Jiang, Z. P., & Chai, T. (2014). Adaptive and Optimal Output Feedback Control of Linear Systems: An Adaptive Dynamic Programming Approach. In *Proceedings of the 11th world congress on intelligent control and automation* (p. 2085-2090). Shenyang, China.

Gao, W., & Jiang, Z.-P. (2015a). Global Optimal Output Regulation of Partially Linear Systems via Robust Adaptive Dynamic Programming. *IFAC-PapersOnLine*, *48*(11), 742–747.

Gao, W., & Jiang, Z. P. (2015b). Global Optimal Output Regulation of Partially Linear Systems via Robust Adaptive Dynamic Programming. In *Proc. 1st conference on modelling. identification and control of nonlinear systems* (Vol. 48, p. 742-747). Saint-Petersburg, Russia.

Gao, W., & Jiang, Z.-P. (2015c). Linear Optimal Tracking Control: An Adaptive Dynamic Programming Approach. In *American control conference (acc), 2015* (pp. 4929–4934).

Gao, W., & Jiang, Z. P. (2016a). Adaptive Dynamic Programming and Adaptive Optimal Output Regulation of Linear Systems. *IEEE Transactions on Automatic Control*,

61(12), 4164-4169. doi: http://dx.doi.org/10.1109/TAC.2016.2548662

Gao, W., & Jiang, Z.-P. (2016b). Adaptive Dynamic Programming and Adaptive Optimal Output Regulation of Linear Systems. *IEEE Transactions on Automatic Control*, *61*(12), 4164–4169.

Gao, W., & Jiang, Z. P. (2017). Learning-based Adaptive Optimal Tracking Control of Strict-feedback Nonlinear Systems. *IEEE Transactions on Neural Networks and Learning Systems,* in press. doi: 10.1109/TNNLS.2017.2761718

Gao, W., & Jiang, Z.-P. (2018). Learning-based Adaptive Optimal Tracking Control of Strict-feedback Nonlinear Systems. *IEEE transactions on neural networks and learning systems*, *29*(6), 2614–2624.

Gao, W., Jiang, Z. P., Lewis, F. L., & Wang, Y. (2017). Cooperative Optimal Output Regulation of Multi-agent Systems using Adaptive Dynamic Programming. In *Proceedings of the american control conference* (p. 2674-2679). Seattle, WA.

Gao, W., Jiang, Z. P., & Ozbay, K. (2015). Adaptive Optimal Control of Connected Vehicles. In *Proceedings of the 10th international workshop on robot motion and control* (p. 288-293). Poznan, Poland.

Gao, W., Jiang, Z. P., & Ozbay, K. (2017). Data-driven Adaptive Optimal Control of Connected vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *18*(5), 1122-1133. doi: 10.1109/TITS.2016.2597279

Gao, W., Jiang, Z. P., Ozbay, K., & Gao, J. (2018). Data-driven Cooperative Adaptive Cruise Control of Buses on the Exclusive Bus Lane of the Lincoln Tunnel Corridor. In *Trb annual meeting*. Washington, DC.

Gao, W., Liu, Y., Odekunle, A., Jiang, Z.-P., Yu, Y., & Lu, P. (2018). Cooperative and Adaptive Optimal Output Regulation of Discrete-Time Multi-Agent Systems Using Reinforcement Learning. In *2018 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 348–353).

Gao, W., Liu, Y., Odekunle, A., Yu, Y., & 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.

Gao, W., Odekunle, A., Chen, Y., & Jiang, Z.-P. (2018). Predictive Cruise Control of Connected and Autonomous Vehicles via Reinforcement Learning. *IET Control Theory & Applications*.

Gao, W., Rios, F., Tong, W., & Chen, L. (2017). Cooperative Adaptive Cruise Control of Connected and Autonomous Vehicles Subject to Input Saturation. In *Ieee annual ubiquitous computing, electronics and mobile communication conference (uemcon).* New York, NY.

Guo, G., & Yue, W. (2014). Sampled-data Cooperative Adaptive Cruise Control of Vehicles With Sensor Failures. *IEEE Transactions on Intelligent Transportation Systems*, *15*(6), 2404-2418.

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

Huang, J., & Chen, Z. (2004). A General Framework for Tackling the Output Regulation Problem. *IEEE Transactions on Automatic Control*, *49*(12), 2203-2218.

Huang, J., & Rugh, W. J. (1990). On a Nonlinear Multivariable Servomechanism Problem. *Automatica*, *26*(6), 963 - 972.

The intelligent transportation systems for traffic signal control deployment benefits and lessons learned (Tech. Rep.). (2007). Washington, DC: US Department of Transportation.

Isidori, A., & Byrnes, C. I. (1990). Output Regulation of Nonlinear Systems. *IEEE Transactions on Automatic Control*, *35*(2), 131-140.

Isidori, A., Marconi, L., & Serrani, A. (2003). *Robust Autonomous Guidance: An Internal Model Approach*. London: UK: Springer-Verlag.

Jiang, Y., Fan, J., Chai, T., Li, J., & Lewis, F. L. (2017). Data-driven Flotation Industrial Process Operational Optimal Control Based on Reinforcement Learning. *IEEE Transactions on Industrial Informatics,* in press. doi: 10.1109/TII.2017.2761852

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

Jiang, Y., & Jiang, Z. P. (2014). Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 882-893.

Jiang, Z. P., Teel, A. R., & Praly, L. (1994). Small-gain theorem for ISS systems and applications. *Mathematics of Control, Signals and Systems*, *7*(2), 95-120.

Kavurucu, Y., & Ensar, T. (2017). Predictive Cruise Control. In *2017 electric electronics, computer science, biomedical engineerings' meeting* (p. 1-4).

Khalil, H. K. (2002). *Nonlinear systems* (3rd ed.). NJ: Prentice Hall PTR.

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

Krener, A. J. (1992). The construction of optimal linear and nonlinear regulators. In A. Isidori & T. J. Tarn (Eds.), *Systems, models and feedback: Theory and applications* (Vol. 12, p. 301-322). Birkhauser Boston.

Lee, J., & Park, B. (2012). Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment. *IEEE Transactions on Intelligent Transportation Systems*, *13*(1), 81-90. doi: 10.1109/TITS.2011.2178836

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

41(1), 14-25.

Lewis, F. L., & Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, *9*(3), 32-50.

Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal control*. Hoboken, NJ: Wiley.

Li, H., Liu, D., & Wang, D. (2014). Integral Reinforcement Learning for Linear Continuous-Time Zero-Sum Games with Completely Unknown Dynamics. *IEEE Transactions on Automation science and engineering*, *11*(3), 706–714.

Mendel, J., & McLaren, R. (1970). Reinforcement-Learning Control and Pattern Recognition Systems. In *Mathematics in science and engineering* (Vol. 66, pp. 287–318). Elsevier.

Meng, Z., Yang, T., Dimarogonas, D. V., & Johansson, K. H. (2015). Coordinated output regulation of heterogeneous linear systems under switching topologies. *Automatica*, *53*, 362 - 368.

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30.

Modares, H., Lewis, F. L., & Jiang, Z.-P. (2015). H Tracking Control of Completely Unknown Continuous-Time systems via Off-Policy Reinforcement Learning. *IEEE Trans. Neural Netw. Learning Syst.*, *26*(10), 2550–2562.

Mu, C., Ni, Z., Sun, C., & He, H. (2017). Data-driven tracking control with adaptive dynamic programming for a class of continuous-time nonlinear systems. *IEEE Transactions on Cybernetics*, *47*(6), 1460-1470.

Odekunle, A., Gao, W., Anayor, C., Wang, X., & Chen, Y. (2018). Predictive Cruise Control of Connected and Autonomous Vehicles: An Adaptive Dynamic Programming Approach. In *Southeastcon 2018* (pp. 1–6).

Oncu, S., Ploeg, J., van de Wouw, N., & Nijmeijer, H. (2014). Cooperative adaptive cruise control: Network-aware analysis of string stability. *IEEE Transactions on Intelligent*

Transportation Systems, *15*(4), 1527-1537.

Rizvi, S. A. A., & Lin, Z. (2017). Output Feedback Reinforcement Q-Learning Control for the Discrete-time Linear Quadratic Regulator Problem. In *2017 ieee 56th annual conference on decision and control (cdc)* (pp. 1311–1316).

Saberi, A., Kokotovic, P., & Summers, S. (1990). Global stabilization of partially linear composite systems. *SIAM Journal of Control and Optimization*, *2*(6), 1491–1503.

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

Serrani, A., Isidori, A., & Marconi, L. (2001). Semiglobal nonlinear output regulation with adaptive internal model. *IEEE Transactions on Automatic Control*, *46*(8), 1178-1194.

Shladover, S., Su, D., & Lu, X.-Y. (2012). Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transportation Research Record*, *2324*, 63-70.

Sontag, E. D. (1989). Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, *34*(4), 435-443.

Sontag, E. D. (2007). Input to state stability: Basic concepts and results. In P. Nistri & G. Stefani (Eds.), *Nonlinear and optimal control theory* (p. 163-220). Berlin: Springer-Verlag.

Su, Y., & Huang, J. (2015). Cooperative global robust output regulation for nonlinear uncertain multi-agent systems in lower triangular form. *IEEE Transactions on Automatic Control*, *60*(9), 2378-2389.

Sun, B., He, M., Wang, Y., Gui, W., Yang, C., & Zhu, Q. (2018). A Data-Driven Optimal Control Approach for Solution Purification Process. *Journal of Process Control*, *68*, 171–185.

Sun, X., Chen, X., Qi, Y., Mao, B., Yu, L., & Tang, P. (2016). Analyzing the Effects of Different Advanced Traffic Signal Status Warning Systems on Vehicle Emission Reduc-

tions at Signalized Intersections. In *Transporation research board annual meeting.* Washington, DC.

Sutton, R. S. (1990). Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Machine learning proceedings 1990* (pp. 216–224). Elsevier.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA: MIT Press.

Talebpour, A., & Mahmassani, H. S. (2016). Influence of Connected and Autonomous Vehicles on Traffic Flow Stability and Throughput. *Transportation Research Part C: Emerging Technologies*, *71*, 143 - 163.

Trentelman, H., Stoorvogel, A., & Hautus, M. (2002). *Control theory for linear systems*. London, UK: Springer-Verlag.

Vamvoudakis, K. G. (2014). Event-triggered optimal adaptive control algorithm for continuous-time nonlinear systems. *IEEE/CAA Journal of Automatica Sinica*(3), 282-293.

Vamvoudakis, K. G., & Lewis, F. L. (2012). Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, *22*(13), 1460–1483.

van Arem, B., van Driel, C., & Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, *7*(4), 429-436.

Van Der Schaft, A. J. (1992). L2-gain analysis of nonlinear systems and nonlinear state-feedback h infinity control. *IEEE Transactions on Automatic Control*, *37*(6), 770-784.

Vrabie, D., Vamvoudakis, K. G., & Lewis, F. L. (2013). *Optimal adaptive control and differential games by reinforcement learning principles*. London, UK: Institution of

Engineering and Technology.

Waltz, M., & Fu, K. (1965). A Heuristic Approach to Reinforcement Learning Control Systems. *IEEE Transactions on Automatic Control*, *10*(4), 390–398.

Wang, D., Liu, D., Li, H., Luo, B., & Ma, H. (2016). An approximate optimal control approach for robust stabilization of a class of discrete-time nonlinear systems with uncertainties. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *46*(5), 713-717. doi: 10.1109/TSMC.2015.2466191

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards* (Unpublished doctoral dissertation). King's College, Cambridge.

Werbos, P. J. (1989). Neural Networks for Control and System Identification. In *Proceedings of the 28th ieee conference on decision and control,* (pp. 260–265).

Werbos, P. J. (2009). Intelligence in the Brain: A Theory of how it Works and how to Build It. *Neural Networks*, *22*(3), 200–212.

Wu, H., & Luo, B. (2012). Neural Network Based Online Simultaneous Policy Update Algorithm for Solving the HJI Equation in Nonlinear H_∞ Control. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(12), 1884-1895.

Yaghmaie, F. A., Movric, K. H., Lewis, F. L., Su, R., & Sebek, M. (2018). H Output Regulation of Linear Heterogeneous Multi-Agent Systems over Switching Graphs.

Zhang, H., Cui, L., Zhang, X., & Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, *22*(12), 2226-2236.

Zhang, Y. J., Malikopoulos, A. A., & Cassandras, C. G. (2016). Optimal control and coordination of connected and automated vehicles at urban traffic intersections. In *Proceedings of the american control conference* (p. 6227-6232). doi: 10.1109/ACC.2016.7526648

Zhu, L. M., Modares, H., Peen, G. O., Lewis, F. L., & Yue, B. (2015). Adaptive subopti-

mal output-feedback control for linear systems using integral reinforcement learning. *IEEE Transactions on Control Systems Technology*, *23*(1), 264-273.