Integral Reinforcement Learning and Adaptive Inverse Optimal Control for Continuous-Time Dynamical Systems

Jae Young Lee

The Graduate School
Yonsei University
Department of Electrical and
Electronic Engineering

Integral Reinforcement Learning and Adaptive Inverse Optimal Control for Continuous-Time Dynamical Systems

A Dissertation
Submitted to the Department of
Electrical end Electronic Engineering
and the Graduate School of Yonsei University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Electrical and Electronic Engineering

Jae Young Lee

August 2015

This certifies that the dissertation of Jae Young Lee is approved.

Dissertation Supervisor: Jin Bae Park

Hyun Seok Yang

Euntai Kim

DaeEun Kim

Yoon Ho Choi

The Graduate School
Yonsei University
August 2015

To my lord God and my families, with special thanks to my wife and my sincere parents for his and her self-sacrifice and love.

For God so loved the world that he gave his one and only Son,
that whoever believes in him shall not perish but have eternal life
(John 3:16).

Abbreviations

ADHDP: Action-Dependent HDP

ARE: Algebraic Riccati Equation

CGFC: Cooperative Graphical Formation Control

CT: Continuous-Time

DOA: Domain Of Attraction

DT: Discrete-Time

DP: Dynamic Programming

DHP: Dual Heuristic Programming

GPI: Generalized PI

HDP: Heuristic Dynamic Programming

HJB: Hamilton-Jacobi-Bellman

I-GPI: Integral GPI

I-TD: Integral TD

I-PI: Integral PI

I-VI: Integral VI

ISS: Input-to-State Stability

IRL: Integral RL

LS: Least Squares

LTI: Linear Time-Invariant

LQR: Linear Quadratic Regulator

MPC: Model Predictive Control

MDP: Markov Decision Process

NN: Neural Network

PE: Persistency of Excitation

RL: Reinforcement Learning

PI: Policy Iteration

TD: Temporal Difference

VI: Value Iteration

Table of Contents

List of Figures

List of Tables

Abstract

Integral Reinforcement Learning and Adaptive Inverse Optimal Control for Continuous-Time Dynamical Systems

Jae Young Lee
Department of Electrical and
Eletronic Engineering
The Graduate School
Yonsei University

This dissertation studies integral reinforcement learning (IRL) and adaptive inverse optimal control for continuous-time (CT) dynamical systems. The ultimate goal of these series of researches is to develop the true adaptive optimal control scheme for the target dynamical systems, which have remained as a challenging problem for a long period in the fields of both control systems engineering and machine learning.

IRL is a family of RL methods to learn the optimal control law for unknown or partially unknown CT dynamical systems based on integral rewards. First, this dissertation introduces and analyzes various partially model-free IRL algorithms including integral policy iteration, integral value iteration, infinitesimal generalized policy iteration (GPI), and their generalization "integral GPI". By the mathematical analyses, a new classification of such IRL methods is established, and the conditions for closed-loop stability and monotone convergence are provided. Next, the I-PI algorithm is extended to propose a class of online IRL algorithms that efficiently use the probing signal to relax the model requirements and eliminate the negative effects of the probing signal on the learning algorithms. As a result, a model-free integral Q-learning and partially model-free explorized I-PI algorithms are proposed, both of which efficiently update the parameters while exploring the stable region. Several mathematical analyses and simulations are provided to verify the theoretical evidence and the performance of the IRL methods.

In the study of adaptive inverse optimal control, we focus on the cooperative graphical formation control problem of multiple mobile robots, each of which is modeled by a CT dynamical system. Both kinematics and dynamics of the mobile robots are transformed to consensus error plus velocity motion dynamics, which makes it possible to design the inverse optimal consensus and the adaptation law parts separately. This control-theoretic approach approximately provides the inverse optimality with respect to the given communication topology among the robots. By Lyapunov's and Hamiltonian analyses, the stability and inverse optimality are mathematically shown. Finally, a numerical simulation is given to support the theoretical statements and verify the performance under various scenarios.

Keywords: true adaptive optimal control, integral reinforcement learning, adaptive inverse optimal control, continuous-time, dynamical system, integral policy iteration, integral generalized policy iteration, integral Q-learning, mobile robot, graphical formation control, consensus, multi-agent system, optimality, inverse optimality, stability, convergence

Chapter 1

Introduction

This dissertation focuses on

1. a class of reinforcement learning (RL) for CT dynamical systems known as *integral RL* (IRL);

2. *adaptive inverse optimal design methodologies for control* of CT multi-agent dynamical systems with limited communications.

The academic spectrum of these interdisciplinary topics is extended from machine learning involved in *optimal decision-making problems* to control theories and applications related to *optimal control problems*. In fact, these two kinds of dynamic optimization problems coming from the different branches can be viewed in a unified manner, which help understand the dissertation.

1.1 Unified Viewpoint of Optimal Decision and Control

In machine learning fields, there have been tremendous researches on the design of a learning agent for its optimal behavior or decision in an unknown environment. This kind of design problem is often referred to as *an optimal decision-making problem* or *a RL problem*; Sutton and Barto suggested in their book [1] the terminologies in a RL problem as follows:

- **agent**: a learner or a decision maker;
- **environment**: everything outside the agent it interacts with;
- **state** (s_t): the variables describing the states of the environment;
- **action** (a_t): the input to the environment determined by the agent;

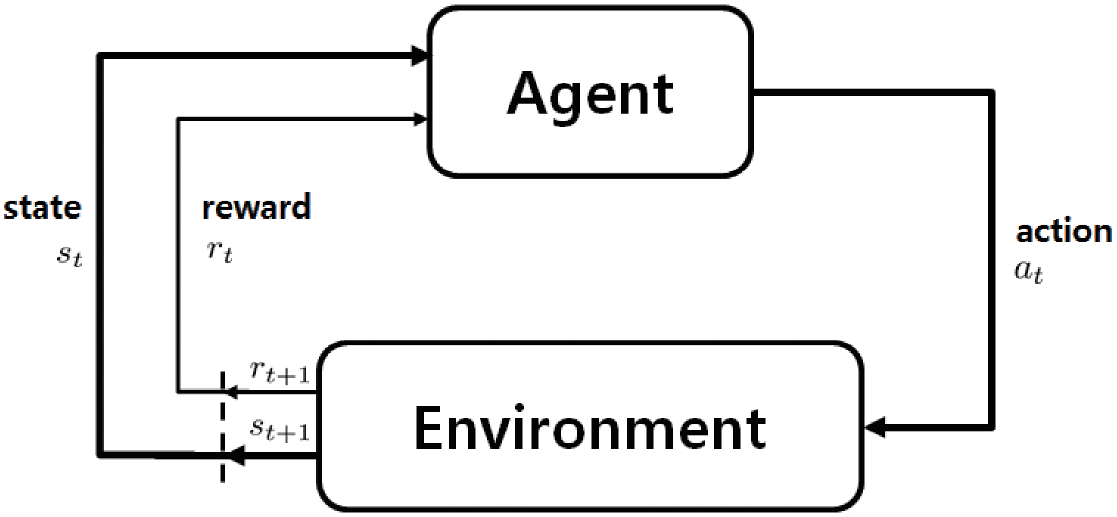


Figure 1.1: The agent-environment interaction in RL

- **policy**: a mapping from states to probabilities of selecting each possible action;
- **reward** (r_t): the numerical outcome the agent receives from the environment;
- **return** (R_t): the sum of the rewards $R_t = \sum_{k=t}^{\infty} r_{t+k}$.

In the RL problem, the agent and environment interact at discrete-time (DT) step t in such a way that the agent receives the state s_t and reward r_t from the environment and on that basis chooses an action a_t to be applied to the environment for the change of its state s_{t+1} in a proper way. This interaction between the agent and environment is described in Fig. 1.1. Now, the RL problem is described as a problem of finding the best policy that

$$Maximize \;\; R_t = \sum_{k=t}^{\infty} r_{t+k}.$$

On the other hand, optimal control theories and methods were developed in the fields of control system engineering for the design of a controller *in an optimal manner* for a given dynamical system and control objectives [2–4]. Similar to optimal decision-making problems, the terminologies in an optimal feedback control problem in CT domain can be defined as follows.

- **controller:** a control input generator;

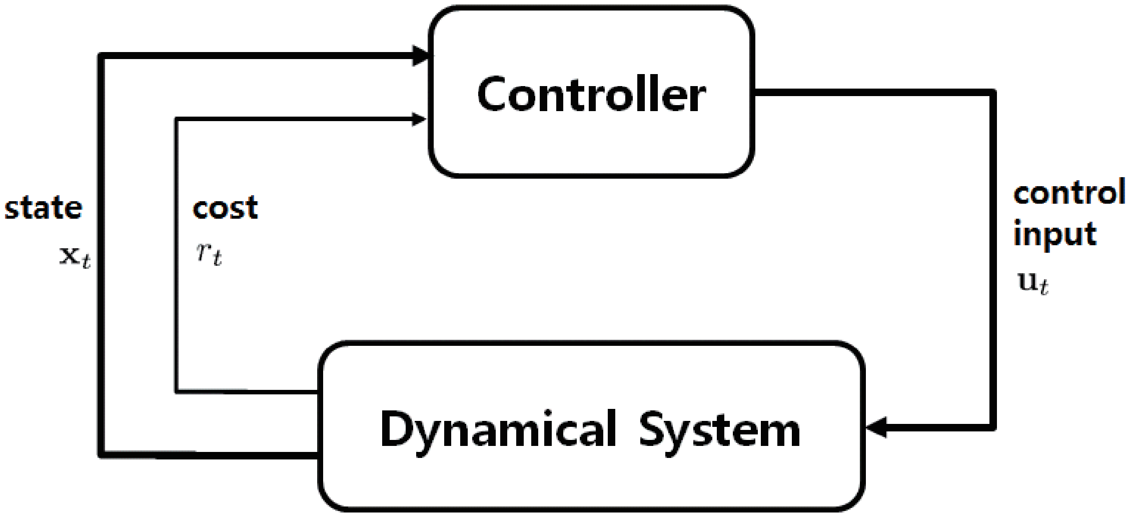


Figure 1.2: The interaction of controller and dynamical system in CT domain.

- **dynamical system:** an environment described by a differential equation;
- **state** ($\mathbf{x}_t$): the variables describing the states of the environment;
- **control input** ($\mathbf{u}_t$): the input to the dynamical system determined by a controller;
- **policy:** a function that generates the control input from the state;
- **cost** (r_t): the numerical outcome representing the instaneous control performance;
- **performance index** (V_t): the integral of the costs $V_t = \int_t^\infty r_\tau \, d\tau$.

In the CT optimal feedback control problem, the controller and the dynamical system interact continuously at each time instant t in such a way that the controller generates $\mathbf{u}_t$ to control the dynamical system in an optimal way based on the state and cost fed-back from the dynamical system. This interaction between the controller and dynamical system is also described in Fig. 1.2, and the solution to the problem is the optimal policy that

$$\textit{Minimize} \;\; V_t = \int_t^\infty r_\tau \, d\tau.$$

Here, the performance index under the optimal policy, called the optimal value function, satisfies the Hamilton-Jacobi-Bellman (HJB) equation known as Bellman optimality equation in machine learning fields [1, 5]. In optimal control problems, the performance index

Table 1.1: Terminologies in optimal control problems and their synonyms in RL problems

No.	Terminologies in control eng. fields	Synonyms in machine learning fields
1	controller	agent
2	dynamical system	environment
3	state, state variable	state
4	control input, control	action, decision
5	policy, control law, protocol	policy
6	cost	reward
7	performance index, cost functional	return
8	feedback control	interaction
9	adaptation, learning	learning
10	Hamiltonian equation, Bellman equation	Bellman equation
11	HJB equation	Bellman optimality equation
12	(optimal) value function	(optimal) value function
13	persistency of excitation (PE)	exploration

(the long-term cost-to-go function) specifies the desired performance with respect to the states and control inputs in the long run, implicitly balancing the amound of required control efforts and the desired transient response.

From the discussions with Figs. 1.1 and 1.2, one can see that both optimal feedback control and optimal decision-making problem can be viewed in an unified manner as an optimization problem of an agent/controller for a given environment and rewards/costs specifying the objectives of the agent/controller. The synonyms in both problems including those in Figs. 1.1 and 1.2 are summarized in Table 1.1.

1.2 Reinforcement Learning: A Historical Review

RL is a class of goal-directed learning algorithms that originate from and are inspired by biological animal learning mechanisms [1, 5–7]. A RL agent tries to learn the best policy by interacting with a given unknown environment to maximize its return, the sum of the rewards the agent receives, or to minimize the given cost functional [1,5,6]. Here, the RL agent and environment interact with each other in the exactly same way to that explained in the previous section (see Fig. 1.1 and 1.2).

1.2.1 Reinforcement Learning in Discrete-Time Domain

From the early 1980's up to date, on the basis of temporal difference (TD) prediction, there have been plenty of studies on RL methods in machine learning fields with special focus on the environment expressed by a finite Markov decision process (MDP) [1,6–11]. As a result, a variety of RL methods in finite MDPs have been proposed, including Sarsa [1], Q-learning [9], actor-critic methods [1,7], to name a few (see [8] for a survey and [1] for a comprehensive understanding). Here, a finite MDP is referred to as a DT system that has a finite number of discrete states and actions with state transition probabilities depending only on the current state and action [1,11].

There are also a number of generalizations of the RL methods in finite MDPs to those in general DT systems whose states or actions or both are "continuous" or "discrete but terribly many," e.g., see [1,6,10–17] for the RL methods developed from machine learning perspectives and [5,6,18–30] from control systems perspectives. In these methods, the value function or action-value function known as Q-function is approximated by a neural network (NN) or a general function approximator. In machine learning fields, the RL studies focus on how to discretize and represent the continuous state or action space to efficiently learn the optimal behavior and how to improve the performance by modifications of the learning rules. Combined with a deep learning NN, the recent study [17] in maching learning fields shows that the policy trained by the RL agent is superior to the professional human operator's policy, revealing the excellent performance of the RL methods. The other successful applications are shown in [12,16].

Reinforcement Learning for Discrete-Time Dynamical Systems

Focused on the environment described by a DT nonlinear dynamical system, Werbos [18] proposed several classes of adaptive dynamic programming (DP) methods such as heuristic dynamic programming (HDP), dual heuristic dynamic programming (DHP), action-dependent HDP (ADHDP), and their generalizations with their respective learning principles; Prokhorov and Wunsch [21] elaborated and simplified these ideas in a practical manner. Here, the term adaptive DP is often referred as a synonym of RL [5, 6, 11] since it has the actor-critic NN structure and the same purpose to RL: "solving the optimal control problems using TD learning"; ADHDP is actually a kind of Q-learning since it approximates the Q-function in forward time without using the knowledge of the system dynamics to learn the optimal solution. On the other hand, HDP and DHP approximate the value function and its derivatives, respectively, to learn the optimal solution, but both need a complete description of the dynamical system to run. There also exist many applications of these adaptive DP methods, e.g., power systems [26] and traffic flows [27] to name a few.

Stability and Convergence Issues

The key properties in a controlled dynamical system equipped with a RL agent are

1. closed-loop stability;
2. convergence near or to the optimal solution.

Here, closed-loop stability roughly means that under the given control, the state with small initial perturbations remains small for all time [31–33]. This is different from the concept of convergence that describes the behaviors only after a sufficiently large amount of time or iteration has passed [34]. There are a class of systems that converge to the desired operating point, but fail to be stabilized (see [31] for an example).

Policy Iteration, Generalized Policy Iteration, and Value Iteration

For a linear qudaratic regulator (LQR) problem, an optimal control problem with *a linear system* and *a quadratic cost*, Bradtke, Ydstie, and Barto [19] presented a Q-learning algorithm based on policy iteration (PI) and showed that the algorithm yields stable control sequences that converge to the optimal solution under the persistency of excitation, a similar concept of the sufficient number of visits in a finite MDP required for convergence of the RL policy to the optimal one [1, 9]. Here, PI is a class of algorithms consisting of the two processes called policy evaluation and policy improvement to sequentially find the optimal solution [5, 13]. In principle, the PI for a finite MDP is not implementable since it requires the infinite number of recursions named Bellman fixed point iterations in every policy evaluation step [1]. On the other hand, PI for the DT dynamical systems is realizable via least squares (LS), but at the expense of the need for an initial stabilizing policy to run [5, 19]. The Q-learning given in [19] for an LQR problem also assumed that the initial policy is stabilizing, which limits the use of the RL method in practical applications.

In a finite MDP, generalized PI (GPI) is the general idea of allowing the two consecutive steps of PI, policy evaluation and improvement, to be performed without completing the other step. [1, 35, 36]. Modified PI, formulated by Putermain and Shin [35] and van Nunen [36], is a typical example of this, where only the finite number of Bellman fixed-point iterations are performed to approximately implement the policy evaluation of PI. This idea of modified PI was also extended to the optimal control of DT dynamical systems with some convergence and stability analysis [5, 29].

When the modified PI performs only one Bellman fixed-point iteration at every policy evaluation of GPI, then it is called value iteration (VI) [1, 5]. Actually, it was originated in a finite MDP framework The convergence proof in a finite MDP framework was given in [1]. When applied to DT dynamical systems, VI has advantage over PI in that it does not need any initial stabilizing policy to run. Hence, the VI can be implemented regardless of whether the initial policy is stabilizing or not. The convergence proofs of VI for DT dynamical systems were given by Landelius [20] for an LQR problem and by Al-Tamimi [25] for a DT nonlinear optimal control problem with input-affine dynamics. With convergence proof, the ADHDP or Q-learning methods based on VI were also proposed in [20] for an LQR problem and [25] for a linear zero-sum game problem. Actually, all of the ADP methods mentioned above with the literature [18, 21, 26, 27] are also designed based on VI, rather than PI, but the proof of convergence in the general case is still an open problem, to the best author's knowledge, except the special cases, e.g., [24]. The model-free RL output feedback schemes were proposed in [28, 30].

1.2.2 Reinforcement Learning in Continuous-Time Domain

The RL ideas in DT domain are further extended to the general CT systems that have continuous states and actions [37–40]. Regardless to say, these CT extensions are necessary and meaningful in a practical manner since almost all real physical systems are modeled in CT, but to the best author's knowledge, the exact discretization is possible only for limited cases [33, 41, 42], e.g., linear systems [33]. However, the main barrier in extending the RL ideas from DT to CT is that the rare Q-function is hard to be expressed in a TD bootstrapping form. This is because the time difference in DT becomes *time differential* in CT domain, so there does not exists the next state in CT domain determined *only* by the current state and action pair; it is actually determined by the current state and the CT action during the time interval that continuously change the states for that period [33].

Several RL methods were proposed in CT domain without considering stability and convergence proofs. Baird [37] proposed Q-learning-like method called advantage updating, where the CT Bellman equation is discretized using the Euler's method and then to apply the idea of Q-learning. Based on the same discretization method, Doya [38] further extended the RL ideas in DT such as TD learning and eligibility trace [1] to develop his CT RL methods; the simulation studies for the pendulum/cart-pole swing-up tasks were also given in [38] to compare the CT RL algorithms. Hanselmann, Noakes, and Zaknich [39] extended the adaptive DP methods in DT [18, 21] to propose continous-time adaptive DP methods with fast update rules. Mehta and Meyn [40] revealed the relation between Q-function and Pontryagin's minimum principle and then proposed their Q-learning algorithm applicable to stochastic CT dynamical systems.

All of the above RL methods in CT domain did not consider their stability; the convergence to the optimal solution is still an open problem for those RL methods. To the best author's knowledge, the adaptive DP method given in [43] is the first one in CT whose stability and convergence were rigorously proven (see [44] for the proof). It is actually equivalent to the PI method [45–47] in the control communities that guarantees closed-loop stability and monotone decreasing convergence under an initial admissible policy (see also [48] for PI in CT for input-constrained case). Similar to Monte Carlo methods in DT [1], however, it is needed for the adaptive DP [43] to integrate or summing up the costs for all time and for each given initial condition by observing the states and control inputs. Moreover, for the relaxation of the requirements of the system drift dynamics, the derivatives of the state variables should be obtained, which is difficult and may contain undesirable noises that degrade the performance.

Integral Reinforcement Learning

By combining adaptive DP [43] with integral TD (I-TD), a class of RL methods in CT named integral RL (IRL) was presented with stability and convergence studies. The IRL methods are designed with the ideas of PI, GPI, and VI to find the online solution to a CT input-affine nonlinear optimal control problem with unknown system drift dynamics by minimizing or decreasing the associated I-TD error at each step. (see [5, 49] for a comprehensive survey and review).

For the IRL schemes developed from PI [43,45–47] in particular, which is called integral PI (I-PI) in this dissertation, the stability and convergence to the optimal solution were proven under the initial admissible policy [50]. The implementation methods based on LS and the Galerkin NN approximation were also presented in [50]. The IRL schemes derived from the ideas of GPI and VI are also shown in [51] and [49], respectively; they are named in this dissertation integral GPI (I-GPI) and integral VI (I-VI). While I-PI inherently needs an initial admissible (stabilizing) policy to run the algorithm, which is required for any PI method applied to dynamical systems, such an initial admissible policy is not required for I-GPI and VI as was done for GPI and VI for DT dynamical systems above. Unfortunately, the stability and convergence of I-GPI and VI are not fully investigated up to date, which restrict the use of the IRL algorithms.

1.3 Adaptive Optimal Control Theories

The RL methods in the previous section can be considered adaptive optimal control schemes in the control systems point of view; they adaptively find the optimal policy in an unknown environment and after the convergence of the policy, they are near-optimal with respect to the given cost functional or return. In this process, the balance between "exploration of the state or state-action space" and "exploitation of the learned policy" determines the degree between "the convergence to the optimal policy" and "the current performance." This is known as *exploration vs. exploitation dilemma* in both communities of RL [1, 6, 9] and adaptive control [52, 53].

1.3.1 Adaptive Control without Optimality

Adaptive control is a branch in control engineering fields regarding the design of a controller that equips with a adaptation/learning law to eliminate the parametric uncertainties so as to gradually improve the control performance against uncertainties [31, 52, 54–56]. There are a large number of adaptive control schemes that are either combined with well-designed state estimators [31,54] or directly designed via Lyapunov's stability analysis [31,52,54–56]. This dissertation focus on the latter approach, which is more involved in the stability-guaranteed design of control and adaptation laws. A number of intelligent adaptive control schemes, combined with NNs, were also reported in both cases [57–60].

On the other hand, considerable efforts have been made to design an adaptive controller which efficiently balances the exploitation and exploration for good transient performance by regulating the magnitudes of the exploratory random signals injected through the control input channels [52,53,61]. Here, the exploitation and exploration issues are directly connected to the notion of PE[1] as a dilemma between the satisfaction of PE (efficient exploration) and the satisfactory control performance (exploitation to improve the stability and state convergence).

The adaptive control schemes above contain the RL-like components such as "adaptation law" and "PE" that are similar to the terms "learning rule" and "exploration," respectively, as shown in Table 1.1. However, they failed to take a long-term performance index such as return and cost functional into considerations; they are just designed via cancellation of the unknown terms based on the short-term immediate costs.

1.3.2 Model-Based Optimal Control

In the optimal control approaches, Pontryagin's minimum principle and DP are two representative methods applicable to obtain the optimal policy minimizing a given performance index [2, 3, 5, 63]. Both optimal control approaches, however, are basically off-line and require complete knowledge of the system dynamics. In this dissertation, we focus on DP approaches.

The objective of DP is to solve the associated HBJ equation in backward time to obtain the optimal value function and policy. However, this is a formidable task even in the case of completely known dynamics due to the intractability of the HJB partial differential equation and a problem known as the curse of dimensionality.[2] The adaptive DP and RL

[1] The PE condition generally means that the corresponding signal is persistently changing. Hence, the satisfaction of PE introduces oscillatory states and control inputs, and even cause the escape of the stable region. However, without satisfaction of persistency of excitation, the learning parameters cannot converge to the true values [31, 52, 53, 56, 62].

[2] In DP, the computational complexity increases exponentially as the number of states grows linearly

methods shown in Section 1.2 are candidates to alleviate these problems.

Model Predictive Control

In control engineering fields, there are another approaches to alleviate the aforementioned difficulties in solving the HJB equation. One candidate is model-predictive control, also named receding-horizon control, a finite-horizon optimal control problem is formulated that approximates the given, possibly infinite-horizon, optimal control problem. Here, the finite-horizon problem is formulated in a way that it is solvable at every time step to yield the suboptimal control sequences. Though the relevant theories were well-developed and there are many successful applications, the suboptimal policy based on MPC undergoes model-dependency and should be properly designed to alleviate computational burden.

Inverse Optimal Control

Inverse optimal control is another approach to obtain an optimal policy [64–67]. In an inverse optimal approach, a stabilizing policy is given or designed *a priori*, and then the cost functional to be minimized by the given policy is determined *a posteriori*; The control Lyapunov function is involved in the inverse optimal design [64]. Since this inverse optimal approach does not need to solve the associated HJB equation, the aforementioned problems regarding the HJB equation do not arise both in the design and implementation steps. The design procedures, however, surely restrict the choice of the cost functional since it is not given *a priori*, but determined *a posteriori.* Whatever the cost functional is associated, an inverse optimal control poccess the same 'good' properties as the usual optimal controllers such as 90 degree phase margin and infinity gain margin in the case of LQR [63]. Though those advatanges above, the inverse optimal controller is designed only in an offline manner and should explicitly use the knowledge regarding the system dynamics.

[2], which usually limits the use of DP to the cases with the reasonably small number of states.

1.3.3 Adaptive Inverse Optimal Control

As briefly reviewed, adaptive control alone does not provide the optimality in the long run, and the optimal control is usually designed in an offline manner and do not provide the learning rules to improve the performance against the model uncertainties. To overcome these limitations, adaptive inverse optimal control schemes were studied as a combined concept of adaptive control and inverse optimal control [68–70]. In this design methodology, the so-called adaptive control Lyapunov function is constructed to yield both of the control and adaptation laws in an inverse optimal fashion. While the choice of the cost functional is restricted by the nature of inverse optimal control again, the adaptive inverse optimal control has advantages against RL methods that it can guarantees the optimal transient performance during online adaptation [68–70]. In RL approaches, the learning rules are designed only for the learning of the optimal solution *at the end*, so the transient responses are usually not optimal during the learning period.

1.4 Cooperative Graphical Formation Control

Formation control of multiple robots have received much attention for many years, and various approaches such as behavioral methods [71, 72], virtual structure [73, 74], leader-follower [75, 76], and graph-theoretic techniques [77, 78] have been developed. Among these approaches for formation control, graph-theoretic technique, called cooperative graphical formation control (CGFC) in this dissertation, has the highest degree of freedom of communication among the mobile agents. While the communication topology in the other approaches are determined and fixed in the design procedures, CGFC schemes allow to have arbitrary communication structure described by a graph satisfying some required properties (see Appendix C for a brief review of graph theory).

1.4.1 Consensus Theories for Multi-Agent Systems

The design of CGFC is based on the consensus theories, whose objective is to achieve consensus, meaning that all agents in the system reach to a common value [77]. However, the studies on the consensus theories were mainly done under the assumption that the dynamics of each agent is modeled by a linear system [77–84]. Moreover, in the case of optimal consensus algorithms, there is no research for *nonlinear* multi-agent systems to the best authors' knowledge. This is mainly due to the difficulties arising from the constraints on the communication topology of the group of the agents. Even for the linear optimal consensus protocols [79–81], the optimality of the proposed protocols has not been proven up to date due to those difficulties related to the communication constraints.

Inverse Optimal Consensus Protocols

The problem arising from the communication constraints in multi-agent optimal consensus can be solved by designing the protocols with inverse optimality. In this case, the minimizing performance index is determined *a posteriori* according to the given communication topology (and the designed protocol), so the aforementioned difficulties can be alleviated. Considering the LQR performance, inverse optimal consensus protocols have been studied for single integrator agents [82, 83] and identical linear time-invariant (LTI) agent dynamics [79–81, 84–86]. For single integrator agents, Cao and Ren [83] presented the optimal scaling factor and the optimal weighted adjacency matrix under undirected graph; Qu and Simaan [82] analyzed the inverse optimality under fixed and switching directed graph topologies. For identical LTI dynamics, Borrelli and Keviczky [79] proposed a number of sub-optimal consensus schemes under undirected graphs, and their related stability condition on algebraic connectivity. In [81] and [86], (sub-)optimal consensus protocols were proposed based on distributed estimation and game theories for general fixed digraph topologies. Strongly related to this note is a class of protocols designed with the solution of the simple ARE for both consensus [84] and synchronization [80, 84, 85].

Second-Order Consensus Protocols

The second-order consensus protocols were studied in [87,88]. In second-order consensus, each agent is assumed to have a double integrator dynamics. In [87], Ren and Beard introduced group velocity and proposed a second-order protocol that achieves "position consensus to a common value" and "velocity consensus to the given group velocity." This concept can be extended to our CGFC design. On the other hand, to the best author's knowledge, there is no inverse-optimal approach up to now for the second-order consensus.

1.4.2 Cooperative Graphical Formation Control for Mobile Robots

The CGFC of multiple mobile robots are mostly designed based on the linear consensus theory by virtue of dynamic feedback linearization [72,89] that converts the kinematics of a mobile robot into a simple double integrator. In [90] and [91], the authors proposed nonlinear CGFC of mobile robots by employing backstepping and consensus theory for nonholonomic systems [91]. However, most of the formation consensus methods for mobile robots did not consider the dynamics that drives the velocity inputs of the kinematics. In the recent preliminary work [92], an inverse optimal CGFC scheme was proposed, where both kinematics and dynamics of the mobile robots were considered. To the best authors' knowledge, this was the first to design the *nonlinear* inverse optimal GFC of mobile robots considering their kinematics and dynamics, but it fails to grant the desired group velocity to which all mobile agents' velocities converge. This is since the design was based on the first order consensus theory. Moreover, to the best authors' knowledge, there is no result on the adaptive inverse optimal design of CGFC for multiple mobile robots that guarantees the nonlinear inverse optimality of the whole closed-loop multi agent system under the exact parameter estimation.

1.5 Contributions of the Dissertation

As mentioned at the beginning of this chapter, this dissertation focus on the development and analysis of IRL methods for CT dynamical systems and the adaptive inverse optimal design of CGFC for multiple mobile robots. The main contributions can be summarized as the following three parts.

1. (**Analysis and Classifications of IRL Algorithms**) Noting that unlike I-PI, the stability and convergence of I-VI and I-GPI are still an open question, this dissertation analyze the stability and convergence of I-GPI, the most general IRL among I-PI, I-VI, and infinitesimal GPI (the differential limiting version of GPI). The analysis focus on the two convergence mode named

   - PI-mode convergence;
   - VI-mode convergence.

   As a result, a series of conditions are given for each convergence mode and stability. In addition, the analysis suggests

   - the new classification criteria of IRL algorithms

   in terms of the so-called update horizon. This new classification reveals the relation between the computational complexity and the learning speed of I-GPI (and IRL). These contributions are shown in **Chapter 4** and closely related to the journal paper [93] and conference papers [94–96] published during the Ph. D. period.

2. (**Explorized I-PI and Integral Q-Learning**) The IRL algorithms (I-PI, I-GPI, I-VI, and infinitesimal GPI) are partially model-free and require the complete knowledge of the input coupling terms in the system dynamics. In addition, there is no way to excite the state and input variables, which is truely necessary for online learning. In this part, the online IRL algorithms named as explorized I-PI and inte-

gral Q-learning are proposed based on I-PI by introducing a probing signal, called exploration, and advanced I-TD. Here,

- **integral Q-learning** is a model-free IRL that efficiently exploits exploration to relax the model requirements and to excite the state and input variables;
- **explorized I-PI** is a partially model-free IRL that efficiently learn the optimal solution at the expense of the requirement of knowledge of input coupling dynamics.

In addition, the conditions on the explorations are provided for input-to-state stability (ISS) and safe learning of the optimal solution. These contributions are partially shown in or closely related to the journal papers [97–99] and the conference papers [100, 101] published during the Ph. D. period. The corresponding main part of this dissertation is **Chapter 5**.

3. (**Adaptive Inverse Optimal CGFC for Multiple Mobile Robots**) Noting that there is no research on inverse optimal second-order consensus with group velocity, and adaptive inverse optimal approach in CGFC of multiple mobile robots,

- an adaptive inverse optimal design method of CGFC for mobile robot

is proposed with robots' kinematics and dynamics considerations. The proposed CGFC scheme is a union of the following inverse optimal and adaptive components, designed one-by-one and separately:

- inverse optimal second-order protocol for CGFC of mobile robots;
- inverse optimal torque inputs for CGFC of mobile robots with full dynamics;
- adaptation law co-design for compensating parametric uncertainties.

By Lyapunov's and Hamiltonian analyses, the stability and inverse optimality are mathematically shown for the proposed methods. These contributions are shown in

Chapter 6 and closely related to the journal papers [92, 102] and the conference paper [103] published during the Ph. D. period.

To verify the theoretical evidence and the performance of the methods, several numerical simulations are carried out for load-frequency control system (Chapter 4), the inverted pendulum (Chapter 5), and a group of mobile robots with various scenarios (Chapter 6).

1.6 Organization of Dissertation

This dissertation is organized as follows.

- **Chapter 2** summarizes the mathematical notations and backgrounds that are related in the rest of this dissertation.

- In **Chapter3**, the stability and (inverse) optimality theories for CT dynamical systems are reviewed and developed to employ them as the tools in the design and analyses of both IRL and adaptive inverse optimal controller in the dissertation.

- In the framework of CT LQR, **Chapter 4** introduces I-PI, I-VI, infinitesimal GPI, and their generalization I-GPI as the fundamental family of IRL algorithms that are applicable to the linear systems with unknown system matrix. Then, these fundamental IRL families are classified in terms of their update horizon and then analyzed to investigate the stability and convergence.

- In **Chapter 5**, explorized I-PI and integral Q-learning are proposed with a number of analysis in terms of its ISS and the effects of explorations in relation to the nonlinear I-PI and advanced I-TD.

- In **Chapter 6**, the adaptive inverse optimal design method of CGFC is proposed for mutiple mobile robots. By Lyapunov's and Hamiltonian analyses using the results in Chapter 3, the stability and optimality are mathematically shown for the proposed one.

- Finally, **Chapter 7** concludes the dissertation.

Chapter 2

Mathematical Notations and Backgrounds

In this chapter, the mathematical notations, concepts, and tools used or needed in the rest of this dissertation are summarized. All of the real vectors in $\mathbb{R}^n$ and real matrices in $\mathbb{R}^{n\times m}$ for any natural numbers $n, m \in \mathbb{N}$ are denoted with bold letters, and so are the vector- or matrix-valued functions. The set of nonnegative integers and real numbers are denoted by $\mathbb{Z}_+$ and $\mathbb{R}_+$, respectively.

2.1 Notations of Vectors and Matrices

In a Euclidean space $\mathbb{R}^n$,

$$\begin{cases} \|\mathbf{x}\| \text{ denotes any norm of a vector } \mathbf{x} \in \mathbb{R}^n; \\ \|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^T\mathbf{x}} \text{ is the Euclidean norm of a vector } \mathbf{x} \in \mathbb{R}^n; \\ \mathbf{0}_n \in \mathbb{R}^n \text{ is the zero vector in } \mathbb{R}^n; \\ \mathbf{1}_n := [\,1\;1\;\cdots 1\,]^T \text{ is the vector in } \mathbb{R}^n \text{ whose elements are all ones.} \end{cases}$$

In Euclidean matrix spaces $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{n\times n}$,

- $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ is the zero matrix in $\mathbb{R}^{m\times n}$;
- $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix in $\mathbb{R}^{n\times n}$.

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$,

- $\|\mathbf{A}\|$ is a matrix norm of $\mathbf{A}$ that is compatible with a vector norm $\|\mathbf{x}\|$ for $\mathbf{x} \in \mathbb{R}^n$ in a sense that $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$ holds;
- $\ker \mathbf{A} \subseteq \mathbb{R}^n$ indicates the null-space of $\mathbf{A}$, i.e., $\ker \mathbf{A} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}_n\}$.

For a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$,

- $\lambda_i(\mathbf{X})$ denotes the i-th eigenvalue of $\mathbf{X}$ with the absolute increasing order

$$\big|\mathrm{Re}\big(\lambda_1(\mathbf{X})\big)\big| \leq \big|\mathrm{Re}\big(\lambda_2(\mathbf{X})\big)\big| \leq \cdots \leq \big|\mathrm{Re}\big(\lambda_n(\mathbf{X})\big)\big|.$$

For a finite sequence of matrices $\big\{\mathbf{Y}_i \in \mathbb{R}^{p_i \times q_i}\big\}_{i=1}^{n}$ for some $p_i, q_i \in \mathbb{N}$, $\mathbf{diag}\{\mathbf{Y}_1, \cdots, \mathbf{Y}_n\}$ denotes a block-diagonal matrix of the form

$$\mathbf{diag}\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n\} := \begin{bmatrix} \mathbf{Y}_1 & \mathbf{0}_{p_1 \times q_2} & \cdots & \mathbf{0}_{p_1 \times q_n} \\ \mathbf{0}_{p_2 \times q_1} & \mathbf{Y}_2 & \cdots & \mathbf{0}_{p_2 \times q_n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{p_n \times q_1} & \mathbf{0}_{p_n \times q_2} & \cdots & \mathbf{Y}_n \end{bmatrix} \in \mathbb{R}^{(p_1+\cdots+p_n)\times(q_1+\cdots+q_n)}. \tag{2.1}$$

For any N-real vectors $\mathbf{x}_j \in \mathbb{R}^{n_j}$ $(j = 1, 2, \cdots, N)$, $\mathbf{col}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ is the column stacking operator defined as

$$\mathbf{col}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} := [\; \mathbf{x}_1^T \; \mathbf{x}_2^T \; \cdots \; \mathbf{x}_N^T \;]^T \in \mathbb{R}^{n_1+n_2+\cdots+n_N}.$$

With slight abuse of notation, the column stacking operator is also defined *for a real matrix* $\mathbf{X} = [\,\mathbf{x}_1 \,\vdots\, \mathbf{x}_2 \,\vdots\, \cdots \,\vdots\, \mathbf{x}_m\,] \in \mathbb{R}^{n \times m}$ with its i-th column $\mathbf{x}_i \in \mathbb{R}^n$ as

$$\mathbf{col}\{\mathbf{X}\} := \mathbf{col}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\} \in \mathbb{R}^{nm}.$$

For any two matrices $\mathbf{X}$ and $\mathbf{Y}$,

- $\mathbf{X} \otimes \mathbf{Y}$ denotes the Kronecker product of $\mathbf{X}$ and $\mathbf{Y}$.

Let the matrices $\mathbf{X} \in \mathbb{R}^{n \times m}$ and $\bar{\mathbf{Y}} \in \mathbb{R}^{(np) \times q}$ be partitioned as

$$\mathbf{X} = \Big[\mathbf{x}_{1r}^T \,\vdots\, \mathbf{x}_{2r}^T \,\vdots\, \cdots \,\vdots\, \mathbf{x}_{nr}^T\Big]^T \text{ and } \bar{\mathbf{Y}} = \Big[\mathbf{Y}_1^T \,\vdots\, \mathbf{Y}_2^T \,\vdots\, \cdots \,\vdots\, \mathbf{Y}_n^T\Big]^T,$$

where $\mathbf{x}_{ir} \in \mathbb{R}^{1 \times m}$ denotes the i-th row vector of $\mathbf{X}$, and $\mathbf{Y}_i \in \mathbb{R}^{p \times q}$ is the i-th submatrix of $\mathbf{Y}$ $(i = 1, 2, \cdots, n)$. Then, the Khatri-Rao product $\mathbf{X} * \bar{\mathbf{Y}}$ of the partitioned matrices $\mathbf{X}$

and $\bar{\mathbf{Y}}$ is defined as

$$\mathbf{X} * \bar{\mathbf{Y}} := \left[\begin{array}{c} \mathbf{x}_{1r} \otimes \mathbf{Y}_1 \\ \hdashline \mathbf{x}_{2r} \otimes \mathbf{Y}_2 \\ \hdashline \vdots \\ \hdashline \mathbf{x}_{nr} \otimes \mathbf{Y}_n \end{array}\right] = \begin{bmatrix} x_{11}\mathbf{Y}_1 & x_{12}\mathbf{Y}_1 & \cdots & x_{1m}\mathbf{Y}_1 \\ x_{21}\mathbf{Y}_2 & x_{22}\mathbf{Y}_2 & \cdots & x_{2m}\mathbf{Y}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}\mathbf{Y}_n & x_{n2}\mathbf{Y}_n & \cdots & x_{nm}\mathbf{Y}_n \end{bmatrix}.$$

To indicate that the Khatri-Rao product $\mathbf{X} * \bar{\mathbf{Y}}$ is a generalized Kronecker product for a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ and a given finite matrix sequence $\big\{\mathbf{Y}_i \in \mathbb{R}^{p\times q}\big\}_{i=1}^{n}$, we denote it by $\mathbf{X} \otimes \big\{\mathbf{Y}_i\big\}_{i=1}^{n}$, i.e.,

$$\mathbf{X} \otimes \big\{\mathbf{Y}_i\big\}_{i=1}^{n} := \mathbf{X} * \bar{\mathbf{Y}}.$$

Indeed, the Kronecker product $\mathbf{X} \otimes \mathbf{Y}$ is a special case of $\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^{n}$ with "$\mathbf{Y}_1 = \mathbf{Y}_2 = \cdots = \mathbf{Y}_n = \mathbf{Y}$". That is, $\mathbf{X} \otimes \mathbf{Y} = \mathbf{X} \otimes \big\{\mathbf{Y}\big\}_{i=1}^{n}$. The properties of both the Kronecker product $\mathbf{X} \otimes \mathbf{Y}$ and the Khatri-Rao product $\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^{n}$ are investigated and summarized in Appendix A.

Definition 2.1. *A symmetric matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is said to be positive definite (resp. positive semi-definite), denoted by $\mathbf{P} \succ \mathbf{0}_{n\times n}$ (resp. $\mathbf{P} \succeq \mathbf{0}_{n\times n}$) if*

$$\mathbf{x}^T\mathbf{P}\mathbf{x} > 0 \ (\textit{resp. } \mathbf{x}^T\mathbf{P}\mathbf{x} \geq 0) \textit{ for all nonzero vector } \mathbf{x} \in \mathbb{R}^n.$$

For any symmetric $\mathbf{P}$, $\mathbf{Q} \in \mathbb{R}^{n\times n}$, we denote

- $\mathbf{P} \succ \mathbf{Q}$ (resp. $\mathbf{P} \succeq \mathbf{Q}$) if $\mathbf{P} - \mathbf{Q}$ is positive definite (resp. positive semi-definite);
- $\mathbf{P} \prec \mathbf{Q}$ (resp. $\mathbf{P} \preceq \mathbf{Q}$) if $\mathbf{Q} - \mathbf{P}$ is positive definite (resp. positive semi-definite).

Related to the systems theory, a square matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ is said to be *Hurwitz* if every eigenvalue of $\mathbf{X}$ has strictly negative real part, i.e., $\mathrm{Re}\big[\lambda_i(\mathbf{X})\big] < 0$ for all $i \in \{1, 2, \cdots, n\}$; the matrix-time exponential of $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}$ is defined as an infinite series $e^{\mathbf{X}t} := \sum_{N=0}^{\infty}(\mathbf{X}t)^N/N!$ which converges to $\mathbf{0}_{n\times n}$ in the limit $t \to \infty$ if $\mathbf{X}$ is Hurwitz. Related to this is the following lemmas that will be used throughout the dissertation.

Lemma 2.1. *For any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ and any $t > 0$,*

$$e^{\mathbf{X}^T t}\, \mathbf{Y} e^{\mathbf{X}t} - \mathbf{Y} = \int_0^t e^{\mathbf{X}^T \tau}(\mathbf{X}^T\mathbf{Y} + \mathbf{Y}\mathbf{X})e^{\mathbf{X}\tau}\, d\tau. \tag{2.2}$$

Proof. From $\frac{d}{dt}e^{\mathbf{X}t} = \mathbf{X}e^{\mathbf{X}t} = e^{\mathbf{X}t}X$,

$$\int_0^t e^{\mathbf{X}^T \tau}(\mathbf{X}^T\mathbf{Y} + \mathbf{Y}\mathbf{X})e^{\mathbf{X}\tau}\, d\tau = \int_0^t \frac{d}{d\tau} e^{\mathbf{X}^T\tau}\mathbf{Y}e^{\mathbf{X}\tau}\, d\tau = e^{\mathbf{X}^T t}\, \mathbf{Y}e^{\mathbf{X}t} - \mathbf{Y},$$

which completes the proof. □

In addition, if $\mathbf{X}$ is Hurwitz, (2.2) can be simplified as

$$-\mathbf{Y} = \int_0^\infty e^{\mathbf{X}^T \tau}(\mathbf{X}^T\mathbf{Y} + \mathbf{Y}\mathbf{X})e^{\mathbf{X}\tau}\, d\tau \tag{2.3}$$

in the limit $t \to \infty$. This provides the explicit integral formula of the solution $\mathbf{Y} \in \mathbb{R}^{n\times n}$ of the Lyapunov equation shown below:

Lemma 2.2. *If $\mathbf{X} \in \mathbb{R}^{n\times n}$ is Hurwitz, then for any given $\mathbf{Y} \in \mathbb{R}^{n\times n}$, the Lyapunov equation $\mathbf{X}^T\mathbf{P} + \mathbf{P}\mathbf{X} = -\mathbf{Y}$ is uniquely solvable and its solution is given by*

$$\mathbf{P} = \int_0^\infty e^{\mathbf{X}^T\tau}\mathbf{Y}e^{\mathbf{X}\tau}\, d\tau. \tag{2.4}$$

Proof. Here, substituting (2.4) and using $\frac{d}{dt}e^{\mathbf{X}t} = \mathbf{X}e^{\mathbf{X}t} = e^{\mathbf{X}t}X$ and (2.3) proves that (2.4) is a solution to the Lyapunov equation $\mathbf{X}^T\mathbf{P} + \mathbf{P}\mathbf{X} = -\mathbf{Y}$. For the uniqueness of the solution $\mathbf{P} \in \mathbb{R}^{n\times n}$, see [104, Theorem 8.5.1]. □

2.2 Sets, Topology, and Functions in $\mathbb{R}^n$

For any two sets X and Y in $\mathbb{R}^n$,

- $X \subseteq Y$ (resp. $X \subset Y$) indicates that X is a subset (resp. a proper subset) of Y;
- ∂X denotes the boundary of X;
- $\bar{X}$ is the closure of X, i.e., the union of X and ∂X.

Note that the closure $\bar{X}$ of any set X is the smallest closed set containing X. For $r \in (0, \infty)$,

- $B_{\mathbf{z}}(r) := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{z}\| < r\}$ is an r-radius open ball in $\mathbb{R}^n$, centered at $\mathbf{z} \in \mathbb{R}^n$.

Definition 2.2. *A subset $\mathbb{S}$ of $\mathbb{R}^n$ is a linear subspace if it is closed under the vector addition and scalar multiplication, i.e., $\mathbf{x} + \mathbf{y} \in \mathbb{S}$ and $c \cdot \mathbf{x} \in \mathbb{S}$ for $\mathbf{x}$, $\mathbf{y} \in \mathbb{S}$ and $c \in \mathbb{R}$.*

Throughout the dissertation, $\mathbb{S}$ will be used to denote a linear subspace of $\mathbb{R}^n$ for some $n \in \mathbb{N}$. The followings are examples of a linear subspace that will be shown and play a central role in stability analysis in this dissertation.

Example 2.1. *The singleton $\mathbb{S} = \{\mathbf{0}_n\}$ a linear subspace; actually it is the smallest among the linear subspaces in $\mathbb{R}^n$. In this dissertation, this zero linear subspace $\mathbb{S} = \{\mathbf{0}_n\}$ is termed as the zero equilibrium or the zero equilibrium space.*

Example 2.2. *The null-space $\mathbb{S} = \ker \mathbf{A}$ of any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a linear subspace in $\mathbb{R}^n$. This type of linear subspace will be shown in Chapter 6.*

For a linear subspace $\mathbb{S} \subseteq \mathbb{R}^n$,

- $d(\mathbf{x}, \mathbb{S})$ denotes the distance function between $\mathbf{x} \in \mathbb{R}^n$ and the space $\mathbb{S}$, i.e.,

$$d(\mathbf{x}, \mathbb{S}) := \inf \big\{ \|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in \mathbb{S} \big\};$$

- $d_2(\mathbf{x}, \mathbb{S})$ is the distance function $d(\mathbf{x}, \mathbb{S})$ with the Euclidean norm $\| \cdot \| = \| \cdot \|_2$, i.e.,

$$d_2(\mathbf{x}, \mathbb{S}) := \inf \big\{ \|\mathbf{x} - \mathbf{y}\|_2 : \mathbf{y} \in \mathbb{S} \big\}.$$

With slight abuse of notations,

- $B_{\mathbb{S}}(\varepsilon) := \{\mathbf{x} \in \mathbb{R}^n : d(\mathbf{x}, \mathbb{S}) < \varepsilon\}$ is the ε-neighborhood of a subspace $\mathbb{S}$;

the closures of $B_{\mathbf{z}}(r)$ and $B_{\mathbb{S}}(\varepsilon)$ are denoted by $\bar{B}_{\mathbf{z}}(r)$ and $\bar{B}_{\mathbb{S}}(\varepsilon)$, respectively.

Definition 2.3. *A subset Ω of $\mathbb{R}^n$ is compact if it is closed and bounded.*

On the other hand, we impose the property of connectedness to the domains of functions. Here, the connected set is precisely defined as follows.

Definition 2.4. *An open set in $\mathbb{R}^n$ is said to be (path-)connected*[1] *if for every $\mathbf{x}$, $\mathbf{y} \in \mathcal{D}$, there is a continuous function $f_{\mathbf{xy}} : [0, 1] \to \mathcal{D}$ such that $f_{\mathbf{xy}}(0) = \mathbf{x}$ and $f_{\mathbf{xy}}(1) = \mathbf{y}$.*

[1] There is a general notion of connectedness in topology, but our analyses are sufficient to define it as a path-connectedness, which is a particular class of connectedness but more intuitive than its general definition.

In this dissertation, we denote

- $\mathfrak{D}(\mathbb{R}^n, \{\mathbf{0}_n\})$ the family of all open connected subsets that contain the origin $\mathbf{0}_n$.

This can be eailsy generalized for a linear subspace $\mathbb{S}$ as follows.

- $\mathfrak{D}(\mathbb{R}^n, \mathbb{S})$ is the family of all open connected subsets that contain the subspace $\mathbb{S}$.

Since $\{\mathbf{0}_n\} \subseteq \mathbb{S}$ for any linear subspace $\mathbb{S}$ of $\mathbb{R}^n$, we have $\mathfrak{D}(\mathbb{R}^n, \mathbb{S}) \subseteq \mathfrak{D}(\mathbb{R}^n, \{\mathbf{0}_n\})$. This definition can be further extended to the general case by declaring for $\mathcal{D}$, $\Omega \subseteq \mathbb{R}^n$

- $\mathfrak{D}(\mathcal{D}, \Omega)$ the family of all open connected subsets of $\mathcal{D}$ that contain $\Omega \cap \mathcal{D}$.

Here, if both $\mathcal{D}$ and Ω contain a linear subspace $\mathbb{S} \subseteq \mathbb{R}^n$ in common, then by $\mathcal{D} \subseteq \mathbb{R}^n$ and $\mathbb{S} \subseteq \Omega$, we have the following chain of inclusions:

$$\begin{array}{ccccc} \mathfrak{D}(\mathcal{D}, \Omega) & \subseteq & \mathfrak{D}(\mathbb{R}^n, \Omega) & & \\ \rotatebox[origin=c]{-90}{$\subseteq$} & & \rotatebox[origin=c]{-90}{$\subseteq$} & & \\ \mathfrak{D}(\mathcal{D}, \mathbb{S}) & \subseteq & \mathfrak{D}(\mathbb{R}^n, \mathbb{S}) & \subseteq & \mathfrak{D}(\mathbb{R}^n, \{\mathbf{0}_n\}). \end{array}$$

The other notations regarding the topology and connected subsets in in $\mathbb{R}^n$ are as follows.

- $\bar{\mathfrak{D}}(\mathcal{D}, \Omega)$ the family of closures of all open connected subsets in $\mathfrak{D}(\mathcal{D}, \Omega)$.
- $(\mathfrak{D} \cup \bar{\mathfrak{D}})(\mathcal{D}, \Omega)$ the family of all connected subsets X in $\mathbb{R}^n$ such that either $X \in \mathfrak{D}(\mathcal{D}, \Omega)$ or $X \in \bar{\mathfrak{D}}(\mathcal{D}, \Omega)$.

In this dissertation, the domains of most of the functions in $\mathbb{R}^n$ belong to $\mathfrak{D}(\mathbb{R}^n, \mathbb{S})$; otherwise specified, such a domain in $\mathfrak{D}(\mathbb{R}^n, \mathbb{S})$ is denoted by $\mathcal{D}$, i.e.,

- $\mathcal{D} \subseteq \mathbb{R}^n$ represents *a domain of a function in* $\mathbb{R}^n$ that belongs to $\mathfrak{D}(\mathbb{R}^n, \mathbb{S})$ for a linear subspace $\mathbb{S} \subseteq \mathbb{R}^n$.

Using this notation, we define

- $C^0(\mathcal{D})$ the set of all continuous functions on the domain $\mathcal{D}$;
- $C^1(\mathcal{D})$ the set of all continuously differentiable functions on the domain $\mathcal{D}$.

Definition 2.5. *A function $V : \mathcal{D} \to \mathbb{R}$ is said to be positive definite (resp. positive semi-definite) on a (connected) subset $\Omega \in (\mathfrak{D} \cup \bar{\mathfrak{D}})(\mathcal{D}, \{\mathbf{0}_n\})$, denoted by $V \succ 0$ (resp. $V \succeq 0$) on Ω, if*

1. *V is continuous on Ω, i.e., $V \in C^0(\Omega)$;*

2. *$V(\mathbf{0}_n) = 0$;*

3. *$V(\mathbf{x}) > 0$ (resp. $V(\mathbf{x}) \geq 0$) $\forall x \in \Omega \setminus \{0\}$.*

We simply say that V is positive definite (resp. positive semi-definite) if there is a subset $\Omega \in (\mathfrak{D} \cup \bar{\mathfrak{D}})(\mathcal{D}, \{\mathbf{0}_n\})$ such that V is positive definite (resp. positive semi-definite) on Ω.

The gradient of a real-valued function $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ in $C^1(\mathcal{D})$ is defined as

$$\nabla f(\mathbf{x}) := \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ \frac{\partial f(\mathbf{x})}{\partial x_2}, \ \cdots, \ \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^T \in \mathbb{R}^n,$$

where x_j $(1 \leq j \leq n)$ is the j-th element of $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^n$. For a vector-valued function $\mathbf{f}(\mathbf{x}) = \left[f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_m(\mathbf{x})\right]^T \in \mathbb{R}^m$, $\nabla \mathbf{f}$ or $\nabla \mathbf{f}(\mathbf{x})$ is meant to be a matrix-valued function of the first-order derivatives of the form

$$\nabla \mathbf{f}^T(\mathbf{x}) := \left[\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \cdots, \nabla f_m(\mathbf{x})\right] \in \mathbb{R}^{n \times m}.$$

Lemma 2.3. *If $V : \mathcal{D} \to \mathbb{R}$ is in $C^1(\mathcal{D})$, positive semi-definite, and*

$$V(\mathbf{x}) = 0 \iff \mathbf{x} \in \mathbb{S}, \tag{2.5}$$

where $\mathbb{S} \subseteq \mathbb{R}^n$ is a linear subspace such that $\mathcal{D} \in \mathfrak{D}(\mathbb{R}^n, \mathbb{S})$. Then,

$$\mathbf{x} \in \mathbb{S} \implies \nabla V(\mathbf{x}) = \mathbf{0}_n.$$

Proof. By "$V \succeq 0$" (see Definition 2.5) and the condition (2.5), all of $\mathbf{x}$ in the space $\mathbb{S}$ are the global minimums of $V(\mathbf{x})$. Hence, $\nabla V(\mathbf{x}) = 0$ for any $\mathbf{x} \in \mathbb{S}$ follows from $V \in C^1(\mathcal{D})$ and the fact that $\mathcal{D}$ contains $\mathbb{S}$ in its interior. □

Related to the systems theory and analysis, we define several classes of continuous functions called comparison functions [32] are defined in the following. These comparison functions can be provided as upper- and lower-bounding elements of the other continuos real-valued functions.

Definition 2.6. *A continuous function* $\alpha : [0, a) \to [0, \infty)$ *is of class* $\mathcal{K}$*, denoted by* $\alpha \in \mathcal{K}$*, if it is strictly increasing and* $\alpha(0) = 0$*; a class* $\mathcal{K}$ *function* α *is of class* $\mathcal{K}_\infty$*, denoted by* $\alpha \in \mathcal{K}_\infty$*, if* $a = \infty$ *and* $\lim_{r\to\infty} \alpha(r) = \infty$*.*

Definition 2.7. *A continuous function* $\beta : [0, a) \times [0, \infty) \to [0, \infty)$ *is of class* $\mathcal{KL}$*, denoted by* $\beta \in \mathcal{KL}$*, if* $\beta(\cdot, s) \in \mathcal{K}$ *for each fixed* s*, and for each fixed* r*,* $\beta(r, \cdot)$ *is decreasing and* $\beta(r, s) \to 0$ *as* $s \to \infty$*.*

Lemma 2.4. *Let* $V : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ *be a positive semi-definite function and* $\bar{B}_{\mathbb{S}}(r) \subseteq \mathcal{D}$ *for some* $r > 0$*. Suppose* $V(\mathbf{x}) = 0$ *whenever* $\mathbf{x} \in \mathbb{S}$*. Then, there exist real-valued continuous increasing functions* $\underline{\alpha}$ *and* $\bar{\alpha}$*, defined on* $[0, r]$*, such that* $\underline{\alpha}(0) = \bar{\alpha}(0) = 0$ *and*

$$\underline{\alpha}(d(\mathbf{x}, \mathbb{S})) \leq V(\mathbf{x}) \leq \bar{\alpha}(d(\mathbf{x}, \mathbb{S})). \tag{2.6}$$

Moreover, if the condition on V *is strengthened to*

$$\mathbf{x} \in \mathbb{S} \iff V(\mathbf{x}) = 0,$$

then $\underline{\alpha}$ *and* $\bar{\alpha}$ *can be chosen to belong to class* $\mathcal{K}$*. If* $\mathcal{D} = \mathbb{R}^n$*,* $\underline{\alpha}$ *and* $\bar{\alpha}$ *are defined on* $[0, \infty)$ *and satisfies* (2.6) *for all* $\mathbf{x} \in \mathbb{R}^n$*. If* $V(\mathbf{x}) \to \infty$ *as* $d(\mathbf{x}, \mathbb{S}) \to \infty$*, then* $\underline{\alpha}$ *and* $\bar{\alpha}$ *can be chosen to belong to class* $\mathcal{K}_\infty$*.*

Proof. See Appendix D.1. □

If V is given by $V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x}$ for some $\mathbf{P} \succ \mathbf{0}_{n\times n}$, then (2.6) follows from

$$\underline{\alpha} \cdot \|\mathbf{x}\|_2^2 \leq \mathbf{x}^T \mathbf{P} \mathbf{x} \leq \bar{\alpha} \cdot \|\mathbf{x}\|_2^2 \tag{2.7}$$

that holds for all $\mathbf{x} \in \mathbb{R}^n$, where $\underline{\alpha}$, $\bar{\alpha} > 0$ are chosen as $\underline{\alpha} = \lambda_1(\mathbf{P})$ and $\bar{\alpha} = \lambda_n(\mathbf{P})$. Here, (2.7) can be extended for $\mathbf{P} \succeq \mathbf{0}_{n\times n}$ as shown below.

Lemma 2.5. *For any* $\mathbf{P} \succeq \mathbf{0}_{n\times n}$*, there exist positive constants* $\underline{\alpha}$*,* $\bar{\alpha} > 0$ *such that*

$$\underline{\alpha} \cdot d_2^2(\mathbf{x}, \ker \mathbf{P}) \leq \mathbf{x}^T \mathbf{P} \mathbf{x} \leq \bar{\alpha} \cdot d_2^2(\mathbf{x}, \ker \mathbf{P}). \tag{2.8}$$

Proof. See Appendix D.2. □

Definition 2.8. *A function* $\mathbf{f} : \mathcal{D} \to \mathbb{R}^m$ *($m \in \mathbb{N}$) is locally Lipschitz continuous (on $\mathcal{D}$) if for any* $\mathbf{x} \in \mathcal{D}$*, there are* $r, L > 0$ *such that* $B_{\mathbf{x}}(r) \subseteq \mathcal{D}$ *and*

$$\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})\| \leq L \, \|\mathbf{y} - \mathbf{z}\| \;\; \forall \mathbf{y}, \mathbf{z} \in B_{\mathbf{x}}(r).$$

In what follows, we introduce the several classes of locally Lipschitz continuous functions on $\mathcal{D}$ as follows.

- $C_L^0(\mathcal{D})$: the set of all locally Lipschitz continuous functions;
- $C_L^1(\mathcal{D})$: the set of all functions $\mathbf{f} : \mathcal{D} \to \mathbb{R}^m$ whose first-order derivatives are locally Lipschitz continuous, i.e., $\nabla \mathbf{f} \in C_L^0(\mathcal{D})$;
- $C_{0+}^k(\mathcal{D})$ ($k = 1, 2$): the set of all positive definite functions in $C^k(\mathcal{D})$;
- $C_{L+}^k(\mathcal{D})$ ($k = 1, 2$): the set of all positive definite functions in $C_L^k(\mathcal{D})$;

These classes of functions have the following chain of inclusions (2.9).

$$\begin{array}{ccccccc} C_{L+}^1(\mathcal{D}) & \subset & C_{0+}^1(\mathcal{D}) & \subset & C_{L+}^0(\mathcal{D}) & \subset & C_{0+}^0(\mathcal{D}) \\ \cap & & \cap & & \cap & & \cap \\ C_L^1(\mathcal{D}) & \subset & C^1(\mathcal{D}) & \subset & C_L^0(\mathcal{D}) & \subset & C^0(\mathcal{D}) \end{array} \tag{2.9}$$

Lemma 2.6. *Suppose* $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$*,* $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^{n \times m}$*, and* $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^m$ *are defined and locally Lipschitz continuous on* $\mathcal{D}$*. Then,* $\mathbf{f} + \mathbf{G}\mathbf{u} \in C_L^0(\mathcal{D})$*.*

Proof. Suppose $\mathbf{x} \in \mathcal{D}$. Since $\mathbf{f}$, $\mathbf{G}$, and $\mathbf{u}$ are locally Lipschitz continuous on $\mathcal{D}$, there exist $r > 0$ and Lipschitz constants L_f, L_G, $L_u > 0$ such that $B_{\mathbf{x}}(r) \subseteq \mathcal{D}$ and for any $\mathbf{y}, \mathbf{z} \in B_{\mathbf{x}}(r)$, $\|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})\| \leq L_f \|\mathbf{y} - \mathbf{z}\|$, $\|\mathbf{G}(\mathbf{y}) - \mathbf{G}(\mathbf{z})\| \leq L_G \|\mathbf{y} - \mathbf{z}\|$, and $\|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{z})\| \leq L_u \|\mathbf{y} - \mathbf{z}\|$. Moreover, there are $g_M, \mu_M > 0$ such that $\|\mathbf{G}(\mathbf{y})\| \leq g_M$ and $\|\mathbf{u}(\mathbf{y})\| \leq \mu_M$ hold for all $\mathbf{y} \in B_{\mathbf{x}}(r)$. Therefore, letting $\mathbf{h} := \mathbf{f} + \mathbf{g}\mathbf{u}$ and using the Lipschitz inequalities and the properties of the norm, one can show that

$$\begin{aligned} \|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{z})\| &\leq \|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{z})\| + \|\mathbf{G}(\mathbf{y})\mathbf{u}(\mathbf{y}) - \mathbf{G}(\mathbf{z})\mathbf{u}(\mathbf{z})\| \\ &\leq L_f \|\mathbf{y} - \mathbf{z}\| + \|\mathbf{G}(\mathbf{y}) - \mathbf{G}(\mathbf{z})\| \|\mathbf{u}(\mathbf{y})\| + \|\mathbf{G}(\mathbf{z})\| \|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{z})\| \\ &\leq L_h \|\mathbf{y} - \mathbf{z}\|, \end{aligned}$$

where $L_h \equiv L_f + \mu_M L_g + g_M L_\mu$, and the proof is completed. □

Chapter 3

Stability and Optimal Control Theories for Continuous-Time Dynamical Systems

In this preliminary chapter, we discuss and investigate the theories on both stability and optimal control of CT dynamical systems. All of the dynamical systems in this dissertation are described by a class of of input-affine CT nonlinear dynamical systems of the form

$$\dot{\mathbf{x}}_\tau = \mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)\mathbf{u}(\mathbf{x}_\tau), \quad \mathbf{x}(t) = \mathbf{z} \in \mathcal{D} \subseteq \mathbb{R}^n \tag{3.1}$$

$$\text{where} \begin{cases} \mathbf{x} \in \mathbb{R}^n \text{ is the state variable;} \\ \mathbf{u} \in \mathbb{R}^m \text{ is a control policy to be determined;} \\ \mathbf{z} \text{ is the state value at given initial time instant } \tau = t; \\ \mathbf{f} : \mathcal{D} \to \mathbb{R}^n \text{ with } \mathbf{f}(\mathbf{0}_n) = \mathbf{0}_n \text{ is a vector-valued function in } C_L^0(\mathcal{D}); \\ \mathbf{G} : \mathcal{D} \to \mathbb{R}^{n \times m} \text{ is a matrix-valued function in } C_L^0(\mathcal{D}); \\ \mathcal{D} \in \mathfrak{D}(\mathbb{R}^n, \mathbb{S}) \text{ is the domain of the functions } \mathbf{f} \text{ and } \mathbf{G} \text{ that contains } \mathbb{S} \subset \mathbb{R}^n; \\ \mathbb{S} \subset \mathbb{R}^n \text{ is a linear subspace in } \mathbb{R}^n. \end{cases}$$

Notice that $\mathbf{x} = \mathbf{0}_n$ is an equilibrium point, and the state $\mathbf{x}$ may be zero in the linear subspace $\mathbb{S} \subset \mathbb{R}^n$, which is the generalized notion of the usual zero equilibrium point $\mathbb{S} = \{\mathbf{0}_n\}$. Throughout the dissertation, t indicates a specific time instant on $[0, \infty)$ and $\tau \in [t, \infty)$ will be used as the time variable after the specified time instant t. In addition, any function $\mathbf{x}(\tau)$ of time τ will be denoted as $\mathbf{x}(\tau)$, $\mathbf{x}_\tau$, or simply $\mathbf{x}$ for conciseness.

For a well-posed problem, we assume that there is a control policy $\mathbf{u}(\mathbf{x})$ that asymptotically stabilizes the system (3.1) to the equilibrium space $\mathbb{S}$. Here, the notion of a (control) policy and the stability under a (control) policy are precisely defined below.

Definition 3.1. *A control input function* $\mathbf{u} : \mathcal{D} \to \mathbb{R}^m$ *or its restriction on a subset* Ω *in* $\mathfrak{D}(\mathcal{D}, \mathbb{S})$ *is said to be a policy or a control policy (restricted on* Ω*) if*

1. $\mathbf{u}$ *is locally Lipschitz continuous on its domain;*

2. $\mathbf{u}(\mathbf{x}) = \mathbf{0}_m$ *whenever* $\mathbf{x} \in \mathbb{S}$.

Throughout the dissertation, $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ (and $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{0}_m)$) denotes the state trajectory $\mathbf{x}(\tau)$ at time $\tau \geq t$ generated by the system (3.1) with the initial condition $\mathbf{x}_t = \mathbf{z} \in \mathcal{D}$ and a policy $\mathbf{u}(\mathbf{x})$ (and the zero exploration $\mathbf{e}_\tau \equiv \mathbf{0}_m$[1]). For simplicity, we write $\mathbf{x}_\tau \equiv \mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ if $\mathbf{z}$ and $\mathbf{u}$ are well-understood in the context. Using these notations, we state the existence and uniqueness of the solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ and define a stabilizing policy as shown below.

Proposition 3.1. *For any policy* $\mathbf{u}$*, there is* $T_{max} \in (0, \infty]$ *such that the unique solution* $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ *exists* $\forall \tau \in [t, t + T_{max})$*. Moreover, for a compact subset* $\Omega \subset \mathcal{D}$*, if* $\mathbf{z} \in \Omega$ *and* $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}) \in \Omega$ *for all* $\tau \in [t, T_{max})$*, then* $T_{max} = \infty$*.*

Proof. The system (3.1) is autonomous, and $\mathbf{f} + \mathbf{G}\mathbf{u}$ is locally Lipschitz continuous on $\mathcal{D}$ by Lemma 2.6. Then, the proof can be done by the applications of Theorems 3.1 and 3.3 in [32]. □

Definition 3.2. *A policy* $\mathbf{u}(\mathbf{x})$ *is said to be*

- *stabilizing with respect to a linear subspace* $\mathbb{S}$ *if for any* $\varepsilon > 0$*, there is* $\delta(\varepsilon) \in (0, \varepsilon]$ *such that*
$$\mathbf{z} \in B_{\mathbb{S}}(\delta) \Longrightarrow \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}) \in B_{\mathbb{S}}(\varepsilon) \ \ \forall \tau \geq t;$$

- *asymptotically stabilizing with respect to* $\mathbb{S}$ *if it is stabilizing and there is* $r > 0$ *such that*
$$\lim_{\tau \to \infty} d(\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}), \mathbb{S}) = 0, \ \ \forall \mathbf{z} \in B_{\mathbb{S}}(r);$$

- *exponentially stabilizing with respect to* $\mathbb{S}$ *if there are* $r > 0$*,* $\beta > 0$ *and* $\kappa > 0$ *such that*
$$d(\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}), \mathbb{S}) \leq \beta e^{-\kappa(\tau - t)} d(\mathbf{z}, \mathbb{S}), \ \ \forall \tau \geq t, \ \ \forall \mathbf{z} \in B_{\mathbb{S}}(r).$$

[1] The exact meaning of the exploration $\mathbf{e}_\tau$ will be clear in Chapter 5. Until that, just ignore it.

If the respective conditions in Definition 3.2 are satisfied for the zero equilibrium space $\mathbb{S} = \{\mathbf{0}_n\}$, then we simply say that the policy $\mathbf{u}(\mathbf{x})$ is stabilizing, asymptotically stabilizing, and exponentially stabilizing, respectively.

Remark 3.1. *In the case $\mathbb{S} = \{\mathbf{0}_n\}$, by the first condition in Definition 3.2 and Proposition 3.1, the existence of the unique solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ for all $\tau \in [t, \infty)$ is guaranteed under a stabilizing policy $\mathbf{u}(\mathbf{x})$ since $B_{\mathbf{0}_n}(\delta) \subseteq B_{\mathbf{0}_n}(\varepsilon)$ and $\bar{B}_{\mathbf{0}_n}(\varepsilon)$ is compact. However, in the general case "$\mathbf{0}_n \subset \mathbb{S}$", the existence of the unique solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ is guaranteed only for a finite interval $[t, T_{max})$ by Proposition 3.1 since $\bar{B}_{\mathbb{S}}(\varepsilon)$ is not compact (it is unbounded). In this general case, it is just assumed throughout the dissertation that the unique solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ exists for any $\tau \in [0, \infty)$.*

In this dissertation, the region of attraction (ROA) $R_A(\mathbf{u})$ for $\mathbb{S} = \{\mathbf{0}_n\}$ is defined as

$$R_A(\mathbf{u}) := \big\{\mathbf{z} \in \mathcal{D} : \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}) \to \mathbf{0}_n \text{ as } \tau \to \infty\big\}$$

for an asymptotically stabilizing policy $\mathbf{u}(\mathbf{x})$. Here, the ROA is defined only for the zero equilibrium space $\mathbb{S} = \{\mathbf{0}_n\}$ for simplicity.

Lemma 3.1. *$R_A(\mathbf{u})$ is open, connected, and invariant. Moreover, the boundary $\partial R_A(\mathbf{u})$ is form by the trajectories $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u})$ for $\mathbf{z} \in \partial R_A(\mathbf{u})$.*

Proof. See Appendix C.16 in [32]. □

Since $\mathbf{0}_n$ is the equilibrium point of the dynamical system (3.1), the following lemma can be directly obtained from Lemma 3.1 and the definition of $R_A(\mathbf{u})$ for an asymptotically stabilizing policy $\mathbf{u}$.

Lemma 3.2. *$R_A(\mathbf{u}) \in \mathfrak{D}(\mathcal{D}, \{\mathbf{0}_n\})$. That is, $R_A(\mathbf{u})$ is an open connected subset of $\mathcal{D}$ that contains the origin $\mathbf{0}_n$.*

3.1 Lyapunov's Stability Theorems

Lyapunov's stability theorems [32] are representative tools to investigate the closed-loop stability of an equilibrium "$\mathbf{0}_n$" of dynamical systems. In this dissertation, the following generalized Lyapunov's theorem is established regarding a general linear subspace $\mathbb{S}$. This general Lyapunov's theorem includes the usual Lyapunov's theorem [32] for the zero equilibrium space $\mathbb{S} = \{\mathbf{0}_n\}$ as a special case.

Theorem 3.1. *Given a linear subspace $\mathbb{S}$ and a policy $\mathbf{u}(\mathbf{x})$ for the controlled system* (3.1)*, if there exists a C^1-positive semi-definite function $V : \Omega \to \mathbb{R}_+$, called a Lyapunov function, on a domain $\Omega \in \mathfrak{D}(\mathcal{D}, \mathbb{S})$ such that*

1) $V(\mathbf{x}) = 0 \Longleftrightarrow \mathbf{x} \in \mathbb{S}$;

2) $\dot{V}(\mathbf{x}) \preceq 0$ for all $\mathbf{x} \in \Omega$,

where $\dot{V}(\mathbf{x}) \equiv \nabla^T V(\mathbf{x})\big(\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}(\mathbf{x})\big)$, then the policy $\mathbf{u}(\mathbf{x})$ is stabilizing with respect to the equilibrium space $\mathbb{S}$. Moreover, if the second condition is strengthened to

2′) if there is a positive semi-definite function $W : \Omega \to \mathbb{R}_+$ such that

(a) $W(\mathbf{x}) = 0 \Longleftrightarrow \mathbf{x} \in \mathbb{S}$;

(b) $\dot{V}(\mathbf{x}) \preceq -W(\mathbf{x})$,

then $\mathbf{u}(\mathbf{x})$ is asymptotically stabilizing with respect to $\mathbb{S}$. If there are positive constants $\bar{k}_v$, $\underline{k}_v$, $\underline{k}_w$, $p > 0$ such that for all $\mathbf{x} \in \Omega$,

$$\begin{aligned} &\underline{k}_v d^p(\mathbf{x}, \mathbb{S}) \leq V(\mathbf{x}) \leq \bar{k}_v d^p(\mathbf{x}, \mathbb{S}), \\ &\underline{k}_w d^p(\mathbf{x}, \mathbb{S}) \leq W(\mathbf{x}) \end{aligned}$$

hold, then $\mathbf{u}(\mathbf{x})$ is exponentially stabilizing with respect to $\mathbb{S}$. If $\Omega = \mathcal{D} = \mathbb{R}^n$ and the assumptions hold globally, then $\mathbf{u}(\mathbf{x})$ is globally exponentially stabilizing with respect to $\mathbb{S}$.

Proof. Lemma 2.6 and the conditions on V and W imply the existence of class $\mathcal{K}$ functions $\underline{\alpha}_v$, $\bar{\alpha}_v$, and $\underline{\alpha}_w$ all defined in $[0, r]$ for some $r > 0$ guaranteeing $B_{\mathbb{S}}(r) \subseteq \Omega$ such that

$$\underline{\alpha}_v(d(\mathbf{x}, \mathbb{S})) \leq V(\mathbf{x}) \leq \bar{\alpha}_v(d(\mathbf{x}, \mathbb{S})), \tag{3.2}$$

$$\underline{\alpha}_w(d(\mathbf{x}, \mathbb{S})) \leq W(\mathbf{x}) \tag{3.3}$$

for all $\mathbf{x} \in B_{\mathbb{S}}(r)$. Then, the proof is completed by applying Theorem 4.1 in [105]. □

The above generalized rare theorem will be used in Chapter 6. In the other materials, it is sufficient to use the following Lyapunov's theorems that states the closed-loop stability *for the zero equilibrium* $\mathbb{S} = \{\mathbf{0}_n\}$. The proofs can be done by directly applying Theorem 3.1 (for a complete proof, see [32]).

Corollary 3.1. *Given a policy* $\mathbf{u}(\mathbf{x})$ *and the subspace* $\mathbb{S} = \{\mathbf{0}_n\}$ *for the system* (3.1)*, if there exists a* C^1*-positive definite function* $V : \Omega \to \mathbb{R}_+$ *on a domain* $\Omega \in \mathfrak{D}(\mathcal{D}, \{\mathbf{0}_n\})$ *such that*

$$\dot{V}(\mathbf{x}) \preceq 0 \ \ \textit{for all} \ \ \mathbf{x} \in \Omega,$$

then the policy $\mathbf{u}(\mathbf{x})$ *is stabilizing. Moreover, if the condition is strengthened to*

$$\dot{V}(\mathbf{x}) \prec 0 \ \ \textit{for all} \ \ \mathbf{x} \in \Omega,$$

then $\mathbf{u}(\mathbf{x})$ *is asymptotically stabilizing. In addition to this, if* $\Omega = \mathcal{D} = \mathbb{R}^n$ *and* V *is radially unbounded, then* $\mathbf{u}(\mathbf{x})$ *is globally asymptotically stabilizing.*

Corollary 3.2. *Given a policy* $\mathbf{u}(\mathbf{x})$ *and the subspace* $\mathbb{S} = \{\mathbf{0}_n\}$ *for the system* (3.1)*, if there exists a* C^1*-positive definite function* $V : \Omega \to \mathbb{R}_+$ *on a domain* $\Omega \in \mathfrak{D}(\mathcal{D}, \{\mathbf{0}_n\})$ *such that*

1) $\underline{k}_v \|\mathbf{x}\|^p \le V(\mathbf{x}) \le \bar{k}_v \|\mathbf{x}\|^p$

2) $\dot{V}(\mathbf{x}) \le -\underline{k}_w \|\mathbf{x}\|^p$

for all $\mathbf{x} \in \Omega$*, where* $\underline{k}_v$*,* $\bar{k}_v$*,* $\underline{k}_w$*, and* p *are some positive constants, then the policy* $\mathbf{u}(\mathbf{x})$ *is exponentially stabilizing. In addition to this, if* $\Omega = \mathcal{D} = \mathbb{R}^n$ *and the assumptions hold globally, then* $\mathbf{u}(\mathbf{x})$ *is globally exponentially stabilizing.*

3.2 Optimal Control of Dynamical Systems

The optimal control problems considered in this dissertation can be described in a general form consisting of the input-affine dynamics (3.1) and the performance index

$$V_{\mathbf{u}}(\mathbf{x}_t) = \int_t^\infty r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau))\, d\tau, \tag{3.4}$$

where $r(\mathbf{x}, \mathbf{u}) \in \mathbb{R}$ is the cost defined as $r(\mathbf{x}, \mathbf{u}) := S(\mathbf{x}) + \mathbf{u}^T \mathbf{R}(\mathbf{x}) \mathbf{u} \succeq 0$ for a positive semi-definite function $S : \mathcal{D} \to \mathbb{R}_+$ and a matrix-valued uniformly bounded smooth function $\mathbf{R} : \mathcal{D} \to \mathbb{R}^{m \times m}$ that is positive definite, uniformly for all $\mathbf{x} \in \mathcal{D}$. When the policy

$\mathbf{u}(\mathbf{x})$ is given and fixed for all $\tau \in [t, \infty)$, the performance index (3.4) is called the value function for the given policy. For the existence of $V_{\mathbf{u}}$, the policy $\mathbf{u}(\mathbf{x})$ needs to stabilize the system (3.1). However, since this is not sufficient for the existence of $V_{\mathbf{u}}$, the concept of admissibility is introduced as follows.

Definition 3.3. *A policy $\mathbf{u}(\mathbf{x})$ is admissible with respect to (3.4) and a subspace $\mathbb{S}$ if:*

1) it is asymptotically stabilizing with respect to $\mathbb{S}$;

2) there exists $r_a > 0$ such that $V_{\mathbf{u}}(\mathbf{z}) < \infty$ for all $\mathbf{z} \in B_{\mathbb{S}}(r_a)$.

If the subspace is given by $\mathbb{S} = \{\mathbf{0}_n\}$ and the conditions in Definition 3.3 hold for $\mathbb{S} = \{\mathbf{0}_n\}$, then we simply say that $\mathbf{u}(\mathbf{x})$ is admissible with respect to (3.4). In addition to this, if the performance index is clearly given in the context, then we just say in this dissertation that the policy $\mathbf{u}(\mathbf{x})$ is admissible.

The next theorem states the optimality conditions in addition to those for the closed-loop stability of the general subspace $\mathbb{S}$ under a policy $\mathbf{u}(\mathbf{x})$.

Theorem 3.2. *Consider the optimal control problem (3.1) and (3.4) and suppose that $S(\mathbf{x})$ in the performance index (3.4) has the following property:*

$$S(\mathbf{x}) = 0 \iff \mathbf{x} \in \mathbb{S}.$$

If there exists a C^1-positive semi-definite function $V^* : \Omega \to \mathbb{R}_+$ on a domain $\Omega \in \mathfrak{D}(\mathcal{D}, \mathbb{S})$ such that

1) $V^*(\mathbf{x}) = 0 \iff \mathbf{x} \in \mathbb{S}$;

2) the following HJB equation holds for all $\mathbf{x} \in \Omega$:

$$S(\mathbf{x}) + \nabla V^{*T}(\mathbf{x})\, \mathbf{f}(\mathbf{x}) - \frac{1}{4}\, \nabla V^{*T}(\mathbf{x})\, \mathbf{G}(\mathbf{x})\, \mathbf{R}^{-1}(\mathbf{x})\, \mathbf{G}^T(\mathbf{x})\, \nabla V^*(\mathbf{x}) = 0, \qquad (3.5)$$

then, the control input function $\mathbf{u}^*(\mathbf{x})$ given by

$$\mathbf{u}^*(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V^*(\mathbf{x})$$

is the optimal admissible policy that minimizes the performance index (3.4), and $V^*(\mathbf{x})$ is the corresponding optimal value function, i.e., $0 \preceq V^* \preceq V_{\mathbf{u}}$ for all admissible $\mathbf{u}$.

Proof. By Lemma 2.4 and the first and second conditions regarding V and S, respectively, there exist class $\mathcal{K}$ functions $\underline{\alpha}_v^*$, $\bar{\alpha}_v^*$, and $\underline{\alpha}_s$ such that

$$\underline{\alpha}_v^*(d(\mathbf{x},\mathbb{S})) \preceq V^*(\mathbf{x}) \preceq \bar{\alpha}_v^*(d(\mathbf{x},\mathbb{S})),$$
$$\underline{\alpha}_s(d(\mathbf{x},\mathbb{S})) \preceq S(\mathbf{x})$$

holds for all $\mathbf{x} \in \bar{B}_{\mathbb{S}}(r)$, where $\bar{B}_{\mathbb{S}}(r)$ is a closed ball in the interior of Ω; the application of [105, Theorem 4.1] completes the proof. □

Theorem 3.2 provides the optimality conditions with respect to the general subspace $\mathbb{S} \subset \mathbb{R}^n$ and will be used in Chapter 6. In the other parts, we will focus on the optimal control problems with respect to the equilibrium $\mathbb{S} = \{\mathbf{0}_n\}$. In this case, Theorem 3.2 can be simplified as in the following corollary.

Corollary 3.3. *Let the subspace* $\mathbb{S}$ *be given by* $\mathbb{S} = \{\mathbf{0}_n\}$ *and* $S(\mathbf{x})$ *be positive definite. If there exists a* C^1*-positive semi-definite function* $V^* : \Omega \to \mathbb{R}_+$ *on a domain* $\mathfrak{D}(\mathcal{D}, \{\mathbf{0}_n\})$ *such that the HJB equation* (3.5) *holds for all* $\mathbf{x} \in \Omega$*, then,*

$$\mathbf{u}^*(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V^*(\mathbf{x})$$

is the optimal admissible policy that minimizes the performance index (3.4)*, and* $V^*(\mathbf{x})$ *is the corresponding optimal value function.*

In the remaining subsection, we consider the inverse optimal input-dynamics extension technique under the assumptions that

Assumption 3.1. $V^*(\mathbf{x})$ *is* C^2 *on* $\bar{B}_{\mathbb{S}}(r)$*, where* $\bar{B}_{\mathbb{S}}(r)$ *is a closed ball in the interior of the domain of* V^**.*

Assumption 3.2. $\mathbf{R}(\mathbf{x})$ *is diagonal and* $\mathbf{G}(\mathbf{x})\mathbf{u}$ *is decomposed as*

$$\mathbf{G}(\mathbf{x})\mathbf{u} = \mathbf{G}_s(\mathbf{x})\mathbf{u}_s + \mathbf{G}_d(\mathbf{x})\mathbf{u}_d,$$

where $\mathbf{u}_s \in \mathbb{R}^{m_s}$ *and* $\mathbf{u}_d \in \mathbb{R}^{m_d}$ *with* $m_s + m_d = m$ *are called the static and dynamic feedback control inputs, respectively, and* $\mathbf{G}_s(\mathbf{x})$ *and* $\mathbf{G}_d(\mathbf{x})$ *are corresponding input-coupling matrix-valued functions.*

If $\mathbf{R}(\mathbf{x})$ is diagonal, there are diagonal positive definite matrices $\mathbf{R}_s(\mathbf{x})$ and $\mathbf{R}_d(\mathbf{x})$ such that the HJB equation (3.5) is expressed as

$$S + \nabla V^{*T}\,\mathbf{f}_s + \frac{1}{4}\,\nabla V^{*T}\,\mathbf{G}_s\,\mathbf{R}_s^{-1}\,\mathbf{G}_s^T\,\nabla V^* - \frac{1}{4}\,\nabla V^{*T}\,\mathbf{G}_s\,\mathbf{R}_s^{-1}\,\mathbf{G}_s^T\,\nabla V^* = 0, \tag{3.6}$$

where $\mathbf{f}_s(\mathbf{x}) := \mathbf{f}(\mathbf{x}) - \frac{1}{2}\mathbf{G}_s(\mathbf{x})\mathbf{R}_s^{-1}(\mathbf{x})\mathbf{G}_s^T(\mathbf{x})\nabla V^*(\mathbf{x})$. Hence, under Assumption 3.2, the optimal policy $\mathbf{u}^* = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{G}^T\nabla V^*$ can be decomposed as $\mathbf{u}^* = \mathbf{u}_s^* + \mathbf{u}_d^*$, where $\mathbf{u}_s^*$ and $\mathbf{u}_d^*$ are its static and dynamic parts given by

$$\begin{aligned}
\mathbf{u}_s^*(\mathbf{x}) &:= -\frac{1}{2}\mathbf{R}_s^{-1}(\mathbf{x})\mathbf{G}_s^T(\mathbf{x})\nabla V^*(\mathbf{x}),\\
\mathbf{u}_d^*(\mathbf{x}) &:= -\frac{1}{2}\mathbf{R}_d^{-1}(\mathbf{x})\mathbf{G}_d^T(\mathbf{x})\nabla V^*(\mathbf{x}),
\end{aligned}$$

respectively; the HJB equation (3.6) can be rewritten in terms of $\mathbf{u}_s^*$ as

$$r_s(\mathbf{x}, \mathbf{u}_s^*) + \nabla V^{*T}\,\mathbf{f}_s - \frac{1}{4}\,\nabla V^{*T}\,\mathbf{G}_s\,\mathbf{R}_s^{-1}\,\mathbf{G}_s^T\,\nabla V^* = 0, \tag{3.7}$$

where $r_s(\mathbf{x}, \mathbf{u}_s^*) := S(\mathbf{x}) + \mathbf{u}_s^{*T}\mathbf{R}_s(\mathbf{x})\mathbf{u}_s^*$. Keeping in mind these expressions and assuming that the static feedback control input $\mathbf{u}_s$ is given by $\mathbf{u}_s = \mathbf{u}_s^*(\mathbf{x})$, we consider the extended input-affine dynamics

$$\begin{cases}
\dot{\mathbf{x}} = \mathbf{f}_s(\mathbf{x}) + \mathbf{G}_d(\mathbf{x})\mathbf{u}_d\\
\dot{\mathbf{u}}_d = \mathbf{v}_d(\bar{\mathbf{x}}_d),
\end{cases} \tag{3.8}$$

where $\mathbf{v}_d \in \mathbb{R}^{m_d}$ is the control policy for the extended dynamics (3.8) and $\bar{\mathbf{x}}_d := \mathbf{col}\{\mathbf{x}, \mathbf{u}_d\} \in \mathbb{R}^{n+m_d}$ is its state vector. Defining $\bar{\mathbf{f}}_s(\mathbf{x})$ and $\bar{\mathbf{B}}_{0d}$ as

$$\bar{\mathbf{f}}_s(\mathbf{x}) := \begin{bmatrix}\mathbf{f}_s(\mathbf{x})\\ \mathbf{0}_{m_d}\end{bmatrix} \text{ and } \bar{\mathbf{B}}_{0d} := \begin{bmatrix}\mathbf{0}_{n\times m_d}\\ \mathbf{I}_{m_d}\end{bmatrix},$$

the system (3.8) can be rewritten as

$$\dot{\bar{\mathbf{x}}}_d = \bar{\mathbf{f}}_s(\mathbf{x}) + \bar{\mathbf{B}}_{0d}\mathbf{v}_d(\bar{\mathbf{x}}_d). \tag{3.9}$$

The objective of the inverse optimal input-dynamics extension is to design the extended dynamic control policy $\mathbf{v}_d(\bar{\mathbf{x}}_d)$ to asymptotically stabilize the extended system (3.8) in an optimal fashion with respect to the extended subspace $\mathbb{S}_e \subset \mathbb{R}^{n+m_d}$ defined as

$$\mathbb{S}_e := \Big\{ \bar{\mathbf{x}}_d = (\mathbf{x}, \mathbf{u}_d) \in \mathcal{D} \times \mathbb{R}^{m_d} : \mathbf{x} \in \mathbb{S} \text{ and } \mathbf{u}_d = \mathbf{u}_d^*(\mathbf{x}) \Big\}. \tag{3.10}$$

Since $\mathbf{u}_d^*(\mathbf{x})$ is a policy, we have $\mathbf{u}_d^*(\mathbf{x}) = \mathbf{0}_{m_d}$ whenever $\mathbf{x} \in \mathbb{S}$ by Definition 3.1. Hence, the stabilizing subspace $\mathbb{S}_e$ can be rewritten as

$$\mathbb{S}_e = \Big\{ (\mathbf{x}, \mathbf{u}_d) \in \mathcal{D} \times \mathbb{R}^{m_d} : \mathbf{x} \in \mathbb{S} \text{ and } \mathbf{u}_d = \mathbf{0}_n \Big\}.$$

The design of $\mathbf{v}_d(\bar{\mathbf{x}}_d)$ will be done based on the λ-scaled Q-function $Q_d^*(\bar{\mathbf{x}}_d; \lambda)$ for the dynamic feedback input part $\mathbf{u}_d$, which is defined as

$$Q_d^*(\bar{\mathbf{x}}_d; \lambda) := \lambda V^*(\mathbf{x}) + \nabla V^{*T}(\mathbf{x}) \mathbf{G}_d(\mathbf{x}) \mathbf{u}_d + \mathbf{u}_d^T \mathbf{R}_d(\mathbf{x}) \mathbf{u}_d,$$

where $\lambda > 0$ is a positive constant. Indeed, $Q_d^*(\bar{\mathbf{x}}_d; \lambda)$ satisfies the following properties

$$\begin{aligned} &1)\ \nabla_{\mathbf{u}_d} Q_d^*(\bar{\mathbf{x}}_d; \lambda) = \mathbf{0}_{m_d} \iff \mathbf{u}_d = \mathbf{u}_d^*(\mathbf{x}) \\ &2)\ \lim_{\lambda \to \infty} \frac{1}{\lambda} Q_d^*(\bar{\mathbf{x}}_d; \lambda) = V^*(\mathbf{x}) \end{aligned}$$

that are the similar properties to those of the existing Q-functions (see [19, 25, 98]). For the statement, under Assumption 3.2 let u_{dj}, $\mathbf{g}_{dj}(\mathbf{x})$, and $r_{dj}(\mathbf{x})$ be the j-th element of $\mathbf{u}_d$, $\mathbf{G}_d(\mathbf{x})$, and $\mathbf{R}_d(\mathbf{x})$, respectively. That is, $\mathbf{u}_d = [\, u_{d1} \;\; u_{d2} \;\; \cdots \;\; u_{dm_d} \,]^T$,

$$\begin{aligned} \mathbf{G}_d(\mathbf{x}) &= \Big[\, \mathbf{g}_{d1}(\mathbf{x}) \,\vdots\, \mathbf{g}_{d2}(\mathbf{x}) \,\vdots\, \cdots \,\vdots\, \mathbf{g}_{dm_d}(\mathbf{x}) \,\Big], \\ \mathbf{R}_d(\mathbf{x}) &= \mathbf{diag}\{ r_{d1}(\mathbf{x}),\ r_{d2}(\mathbf{x}),\ \cdots,\ r_{dm_d}(\mathbf{x}) \}. \end{aligned}$$

Moreover, define $\bar{S}_d(\bar{\mathbf{x}}_d) \in \mathbb{R}$ as

$$\bar{S}_d(\bar{\mathbf{x}}_d;\lambda) :=\lambda\, S(\mathbf{x}) + \mathbf{u}_d^T \boldsymbol{\Sigma}(\bar{\mathbf{x}}_d;\lambda)\mathbf{u}_d - (\nabla V^*)^T \boldsymbol{\Xi}(\bar{\mathbf{x}}_d)\Big(\mathbf{f}_c + \mathbf{G}_d\mathbf{u}_d\Big) - \mathbf{u}_d^T \mathbf{G}_d^T \nabla^2 V^* \mathbf{f}_c, \tag{3.11}$$

$$\text{where} \begin{cases} \boldsymbol{\Xi}(\bar{\mathbf{x}}_d) := \nabla \mathbf{G}_d(\mathbf{x})\mathbf{u}_d \equiv \sum_{j=1}^{m_d} u_{dj} \nabla \mathbf{g}_{dj} \\ \boldsymbol{\Sigma}(\bar{\mathbf{x}}_d;\lambda) := \lambda \mathbf{R}_d(\mathbf{x}) - \mathbf{G}_d^T(\mathbf{x})\nabla^2 V^*(\mathbf{x})\mathbf{G}_d(\mathbf{x}) - \boldsymbol{\Upsilon}(\bar{\mathbf{x}}_d) \\ \boldsymbol{\Upsilon}(\bar{\mathbf{x}}_d) := \mathbf{diag}\{(\nabla r_{d1})^T(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d), \;\cdots\;, (\nabla r_{dm_d})^T(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d)\}. \end{cases}$$

Assumption 3.3. *There exist positive constants $\underline{\lambda}$, r and class $\mathcal{K}$ functions $\underline{\alpha}_q$, $\underline{\alpha}_s$, defined on $[0, r]$, such that*

1. *$\bar{B}_{\mathbb{S}}(r) \in \bar{\mathfrak{D}}(\Omega, \mathbb{S})$, where $\Omega \in \mathfrak{D}(\mathcal{D}, \mathbb{S})$ is the domain of $V^*(\mathbf{x})$;*

2. *$\underline{\alpha}_q(d(\mathbf{x}, \mathbb{S})) \le Q_d^*(\bar{\mathbf{x}}_d;\underline{\lambda})$ and $\underline{\alpha}_s(d(\bar{\mathbf{x}}_d, \mathbb{S}_e)) \le \bar{S}_d(\bar{\mathbf{x}}_d;\underline{\lambda})$, $\forall \bar{\mathbf{x}}_d = (\mathbf{x}, \mathbf{u}_d) \in \bar{B}_{\mathbb{S}}(r) \times \mathbb{R}^{m_d}$.*

Now, the next theorem provides the inverse optimal policy $\mathbf{v}_d^*(\bar{\mathbf{x}}_d;\lambda)$ that asymptotically stabilizes the extended dynamics (3.8) in an optimal fashion.

Theorem 3.3. *Let $\mathbf{v}_d^*$ be the policy for the extended dynamics* (3.8) *given by*

$$\mathbf{v}_d^*(\bar{\mathbf{x}}_d;\lambda) = \lambda \mathbf{u}_d^*(\mathbf{x}) - \lambda \mathbf{u}_d \tag{3.12}$$

for a positive constant $\lambda > 0$. Then, under Assumptions 3.1, 3.2, and 3.3, the dynamic policy (6.32) *for any $\lambda \ge \underline{\lambda} > 0$ asymptotically stabilizes the extended system* (3.8) *with respect to the extended subspace $\mathbb{S}_e$ defined by* (3.10)*. Moreover, it is inverse optimal with respect to the performance index $J(\bar{\mathbf{x}}_d(0), \mathbf{v}(\cdot))$ given by*

$$\bar{V}_{\mathbf{v}_d}(\bar{\mathbf{x}}_d(0)) := \int_0^\infty \Big(\bar{S}_d(\bar{\mathbf{x}}_d;\lambda) + \lambda \cdot \mathbf{u}_s^{*T}(\mathbf{x})\mathbf{R}_s(\mathbf{x})\mathbf{u}_s^*(\mathbf{x}) \; + \lambda^{-1} \cdot \mathbf{v}_d^T \mathbf{R}_d(\mathbf{x})\mathbf{v}_d\Big)\, dt, \tag{3.13}$$

and $Q_d^*(\bar{\mathbf{x}}_d;\lambda)$ is the corresponding optimal value function.

Proof. See Appendix D.3. □

3.3 Properties of Value Functions and Hamiltonian

In this and the subsequent sections, we restrict our attention to a class of optimal control problems (3.1) and (3.4) such that

Assumption 3.4. $\mathbb{S} = \{\mathbf{0}_n\}$ *and* $S(\mathbf{x})$ *is positive definite.*

In all of the statements in the two subsequent sections, it is implicitly assumed that $\mathbb{S}$ and $S(\mathbf{x})$ satisfies Assumption 3.4. In this case, the value function $V_{\mathbf{u}}(\mathbf{x})$ and the Hamiltonian of an admissible policy $\mathbf{u}(\mathbf{x})$ possess the mathematical properties that are useful in the design of adaptive optimal control systems. In this section, we investigate those properties, and some of them are provided solely in this dissertation to the best author's knowledge. One of them is the following result regarding the extendability of the admissibility upon the ROA $R_A(\mathbf{u})$.

Proposition 3.2. *Suppose that* $\mathbf{u}(\mathbf{x})$ *is admissible. Then,* $V_{\mathbf{u}}(\mathbf{x}) < \infty$ *for all* $\mathbf{x} \in R_A(\mathbf{u})$*. Moreover,* $V_{\mathbf{u}}$ *is positive definite on* $R_A(\mathbf{u})$ *and* $V_{\mathbf{u}}(\mathbf{x}) \to \infty$ *as* $\mathbf{x} \to \partial R_A(\mathbf{u})$*,*

Proof. Since $\mathbf{u}(\mathbf{x})$ is admissible, it is asymptotically stabilizing. Let $\mathbf{z} \in R_A(\mathbf{u})$. Then, by the definition of the ROA and its invariance (see Lemma 3.1), $\mathbf{x}_\tau(\mathbf{z};\mathbf{u}) \in R_A(\mathbf{u})$ for all $\tau \geq t$, and there is $T_z > 0$ such that $\mathbf{x}_\tau(\mathbf{z};\mathbf{u}) \in B_{\mathbf{0}_n}(r_a)$ for all $\tau \geq t + T_z$, where $r_a > 0$ is given in Definition 3.3. Hence, one obtains

$$V_{\mathbf{u}}(\mathbf{z}) = \int_t^{t+T_z} r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau))\, d\tau \; + \int_{t+T_z}^{\infty} r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau))\, d\tau < \infty, \tag{3.14}$$

where $\mathbf{x}_\tau \equiv \mathbf{x}_\tau(\mathbf{z};\mathbf{u})$. Here, the first integral is finite since $\mathbf{x}_\tau \in R_A(\mathbf{u})$ and it is an integral over a finite time interval, and so is the second integral since it is equal to $V_{\mathbf{u}}(\mathbf{x}_{t+T_z})$ which is finite due to $\mathbf{x}_{t+T_z} \in B_{\mathbf{0}_n}(r_a)$ and the admissibility of $\mathbf{u}$. Since $\mathbf{z} \in R_A(\mathbf{u})$ is arbitrarily given, one has $V_{\mathbf{u}}(\mathbf{z}) < \infty$ for all $\mathbf{z} \in R_A(\mathbf{u})$. Next, since $r(\mathbf{x}, \mathbf{u}(\mathbf{x}))$ is positive definite on $\mathcal{D}$, so is $V_{\mathbf{u}}$ on $R_A(\mathbf{u}) \subseteq \mathcal{D}$ by (3.4), where $V_{\mathbf{u}}$ is finite on $R_A(\mathbf{u})$ by the first argument.

Finally, assume $\mathbf{z} \in \partial R_A(\mathbf{u})$. Then, by Lemma 3.1, $\mathbf{x}_\tau(\mathbf{z};\mathbf{u}) \in \partial R_A(\mathbf{u})$ $\forall \tau \geq t$, and it never converges to $\mathbf{0}_n$. Let $d \in (0, \infty]$ be given by $d := \inf_{\mathbf{x} \in \partial R_A(\mathbf{u})} r(\mathbf{x}, \mathbf{u}(\mathbf{x}))$. Then, $V_{\mathbf{u}}(\mathbf{z})$ for $\mathbf{z} \in \partial R_A(\mathbf{u})$ satisfies

$$V_{\mathbf{u}}(\mathbf{z}) = \int_t^{\infty} r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau))\, d\tau \geq \int_t^{\infty} \left(\inf_{\mathbf{x} \in \partial R_A(\mathbf{u})} r(\mathbf{x}, \mathbf{u}(\mathbf{x})) \right) d\tau = d \cdot \int_t^{\infty} 1\, d\tau = \infty,$$

implying that $V_{\mathbf{u}}(\mathbf{z}) = \infty$ for all $\mathbf{z} \in \partial R_A(\mathbf{u})$. Therefore, the continuity of $V_{\mathbf{u}}(\mathbf{x})$ on $R_A(\mathbf{u})$ implies that $V_{\mathbf{u}}(\mathbf{z}) \to \infty$ as $\mathbf{z} \to \partial R_A(\mathbf{u})$, and the proof is completed. □

For an admissible policy $\mathbf{u}$, let $C_0^k(\mathbf{u})$, $C_{0+}^k(\mathbf{u})$, and $C_{L+}^k(\mathbf{u})$ $(k = 0, 1)$ be the function spaces defined on the ROA $R_A(\mathbf{u})$ as

$$\begin{cases} C_0^k(\mathbf{u}) := \{V \in C^k(R_A(\mathbf{u})) : V(\mathbf{x}) \in \mathbb{R} \text{ and } V(\mathbf{0}_n) = 0\}, \\ C_{0+}^k(\mathbf{u}) := \{V \in C^k(R_A(\mathbf{u})) : V(\mathbf{x}) \in \mathbb{R} \text{ and } V \succ 0\}, \\ C_{L+}^k(\mathbf{u}) := \{V \in C_L^k(R_A(\mathbf{u})) : V(\mathbf{x}) \in \mathbb{R} \text{ and } V \succ 0\}, \end{cases}$$

For instance, $C_{L+}^1(\mathbf{u})$ is the set of all positive definite functions $V : R_A(\mathbf{u}) \to \mathbb{R}$ whose first derivatives are locally Lipschitz continuous. These function spaces have the following inclusions.

$$\begin{array}{ccccc} C_{L+}^1(\mathbf{u}) & \subset & C_{0+}^1(\mathbf{u}) & \subset & C_0^1(\mathbf{u}) \\ \cap & & \cap & & \cap \\ C_{L+}^0(\mathbf{u}) & \subset & C_{0+}^0(\mathbf{u}) & \subset & C_0^0(\mathbf{u}) \end{array} \tag{3.15}$$

by (2.9) and positive definiteness. By Proposition 3.2, one can see that $V_{\mathbf{u}}(\mathbf{x})$ is finite for all $\mathbf{x} \in R_A(\mathbf{u})$ and at least belongs to $C_{0+}^0(\mathbf{u})$. Moreover, Proposition 3.2 and Lemma 2.4 imply that for any $r_{\mathbf{u}} > 0$ and any admissible policy $\mathbf{u}$ satisfying $\bar{B}_{\mathbf{0}_n}(r_{\mathbf{u}}) \subset R_A(\mathbf{u})$, there exist $\underline{\alpha}_{\mathbf{u}}$, $\bar{\alpha}_{\mathbf{u}} \in \mathcal{K}$, defined on $[0, r_{\mathbf{u}}]$, such that

$$\underline{\alpha}_{\mathbf{u}}(\|\mathbf{x}\|) \leq V_{\mathbf{u}}(\mathbf{x}) \leq \bar{\alpha}_{\mathbf{u}}(\|\mathbf{x}\|) \tag{3.16}$$

for all $\mathbf{x} \in \bar{B}_{\mathbf{0}_n}(r_{\mathbf{u}})$. Similarly, since $S(\mathbf{x})$ is positive definite on $\mathcal{D}$, for any $r_d > 0$ satisfying $\bar{B}_{\mathbf{0}_n}(r_d) \subset \mathcal{D}$, there exist $\underline{\alpha}_s$, $\bar{\alpha}_s \in \mathcal{K}$, defined on $[0, r_d]$, such that

$$\underline{\alpha}_s(\|\mathbf{x}\|) \leq S(\mathbf{x}) \leq \bar{\alpha}_s(\|\mathbf{x}\|) \tag{3.17}$$

holds for all $\mathbf{x} \in \bar{B}_{\mathbf{0}_n}(r_d)$. Since $R_A(\mathbf{u}) \subseteq \mathcal{D}$ always holds, $r_d > 0$ can be chosen sufficiently large to satisfy $0 < r_{\mathbf{u}} \leq r_d$ for all admissible policies $\mathbf{u}$. These class $\mathcal{K}$ functions in (3.16) and (3.17) will be used in the analysis of the IRL algorithms in Chapter 5.

Next, define the Hamiltonian $\mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{p})$ as

$$\mathcal{H}(\mathbf{x}, \mathbf{u}, \mathbf{p}) := r(\mathbf{x}, \mathbf{u}) + \mathbf{p}^T(\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}), \tag{3.18}$$

and assume that $\mathbf{u}$ is admissible. Then, if $V_{\mathbf{u}} \in C^1_{0+}(\mathbf{u})$, it satisfies the following Hamiltonian equation for the nonlinear system (3.1):

$$\mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V_{\mathbf{u}}(\mathbf{x})) = 0, \ \ \forall \mathbf{x} \in R_A(\mathbf{u}), \tag{3.19}$$

which is actually the infinitesimal version of (3.4) and implies

$$\dot{V}_{\mathbf{u}}(\mathbf{x}_\tau) \equiv (\nabla V_{\mathbf{u}}(\mathbf{x}_\tau))^T \big(\mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)\mathbf{u}(\mathbf{x}_\tau)\big) = -r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau)) \prec 0. \tag{3.20}$$

for all $\mathbf{x} \in R_A(\mathbf{u})$. That is, the Hamiltonian equation (3.19) is actually a Lyapunov equation, where $V_{\mathbf{u}}(\mathbf{x})$ is the positive definite Lyapunov function for the system (3.1). Moreover, the solution $\nabla V_{\mathbf{u}}$ to the Hamiltonian equation (3.19) is unique over the function space $C^1_0(\mathbf{u})$.

Theorem 3.4. *For an admissible policy* $\mathbf{u}$*, if* $V_{\mathbf{u}} \in C^1_{0+}(\mathbf{u})$*, it is the unique solution to the Hamiltonian equation* (3.19) *over the function space* $C^1_0(\mathbf{u})$*.*

Proof. The proof will be done by contradiction. Assume for an admissible policy $\mathbf{u}$ that there exists another function $V \in C^1_0(\mathbf{u})$ satisfying the Hamiltonian equation

$$\mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x})) = 0 \ \ \ \forall \mathbf{x} \in R_A(\mathbf{u}), \ \ V(\mathbf{0}_n) = 0. \tag{3.21}$$

From (3.21), $r(\mathbf{x}, \mathbf{u}) \succ 0$ (by Assumption 3.4), and the definition of $\mathcal{H}$, we have

$$(\nabla V(\mathbf{x}))^T(\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}(\mathbf{x})) \prec 0, \ \ \forall \mathbf{x} \in R_A(\mathbf{u}) \setminus \{\mathbf{0}_n\},$$

which again implies $\nabla V(\mathbf{x}) \neq \mathbf{0}_n$ (and $\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}(\mathbf{x}) \neq \mathbf{0}_n$) $\forall \mathbf{x} \in R_A(\mathbf{u}) \setminus \{\mathbf{0}_n\}$.

Subtracting (3.21) from (3.19) yields

$$\begin{aligned}\mathcal{H}(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau), &\nabla V(\mathbf{x}_\tau)) - \mathcal{H}(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau), \nabla V_{\mathbf{u}}(\mathbf{x}_\tau)) \\ &= [\nabla V(\mathbf{x}_\tau) - \nabla V_{\mathbf{u}}(\mathbf{x}_\tau)]^T(\mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)\mathbf{u}(\mathbf{x}_\tau)) = 0, \qquad (3.22)\end{aligned}$$

which holds $\forall \mathbf{x}_\tau \in R_A(\mathbf{u})$. Since $R_A(\mathbf{u})$ is an invariant set by Lemma 3.1, $\mathbf{z} \in R_A(\mathbf{u})$ implies $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}) \in R_A(\mathbf{u})$ for all $\tau \geq t$. Therefore, the time integration of (3.22) over the entire interval $[0, \infty)$ yields $V(\mathbf{z}) = V_{\mathbf{u}}(\mathbf{z}) + c$ for a constant c, $\forall \mathbf{z} \in R_A(\mathbf{u})$. Here, $V(\mathbf{0}_n) = V_{\mathbf{u}}(\mathbf{0}_n) = 0$ results in $c = 0$ and thereby, $V(\mathbf{z}) = V_{\mathbf{u}}(\mathbf{z})$ is obtained for all $x \in R_A(\mathbf{u})$, a contradiction. Therefore, the value function $V_{\mathbf{u}}$ is the unique solution of (3.19) over the function space $C_0^1(\mathbf{u})$, the completion of the proof. □

The inclusion (3.15) and Theorem 3.4 imply the following corollary.

Corollary 3.4. *If the value function $V_{\mathbf{u}}$ for an admissible policy $\mathbf{u}$ is $C^1_{L^+}(\mathbf{u})$, then it is the unique solution to the Hamiltonian equation* (3.19) *over $C^1_{L^+}(\mathbf{u})$.*

The objective of the adaptive optimal control in this dissertation is to find the best admissible policy that minimizes (3.4), and the corresponding optimal value function $V^*(\mathbf{x})$. Minimizing the Hamiltonian $\mathcal{H}(\mathbf{x}, \mathbf{u}, \nabla V^*)$ with respect to $\mathbf{u}$, one can obtain the policy $\mathbf{u}^*(\mathbf{x})$ represented as

$$\mathbf{u}^*(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V^*(\mathbf{x}), \qquad (3.23)$$

Here, the admissibility and optimality of $\mathbf{u}^*$ given by (3.23) can be proven by the application of the following theorem.

Theorem 3.5. *For an admissible policy $\mathbf{u}$, let $V_{\mathbf{u}} : R_A(\mathbf{u}) \to \mathbb{R}$ be its corresponding value function. If $V_{\mathbf{u}} \in C^1_{L^+}(\mathbf{u})$, then any locally Lipschitz continuous function $\mathbf{u}^+ : \Omega \to \mathbb{R}^n$ on a domain $\Omega \in \mathfrak{D}(\mathcal{D}, R_A(\mathbf{u}))$ satisfying*

$$\mathbf{u}^+(\mathbf{x}) := -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V_{\mathbf{u}}(\mathbf{x}) \text{ on } R_A(\mathbf{u}) \qquad (3.24)$$

is an admissible policy (*restricted on* Ω)*. Moreover,*

- *$R_A(\mathbf{u})$ is an invariant subset of the ROA $R_A(\mathbf{u}^+)$ under $\mathbf{u}^+$, i.e.,*

$$\mathbf{z} \in R_A(\mathbf{u}) \implies \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+) \in R_A(\mathbf{u}) \;\; \forall \tau \geq t; \qquad (3.25)$$

- *for all* $\mathbf{x} \in R_A(\mathbf{u})$*, the value function* $V_{\mathbf{u}^+}(\mathbf{x})$ *is finite and satisfies*

$$0 \prec V_{\mathbf{u}^+}(\mathbf{x}) \leq V_{\mathbf{u}}(\mathbf{x}) < \infty. \tag{3.26}$$

Proof. First, "$\frac{1}{2}\mathbf{R}^{-1}\mathbf{G}^T\nabla V_{\mathbf{u}}$" in (3.24) is locally Lipschitz continuous by Lemma 2.6 since so are $\mathbf{R}^{-1}$, $\mathbf{G}$ and $\nabla V_{\mathbf{u}}$ ($\because$ $V_{\mathbf{u}} \in C^1_{L+}(\mathbf{u})$). Furthermore, $\mathbf{u}^+(\mathbf{0}_n) = \mathbf{0}_m$ results from Lemma 2.3 and (3.24). Hence, $\mathbf{u}^+$ is a policy. To show the admissibility of $\mathbf{u}^+$, consider the value function $V_{\mathbf{u}}(\mathbf{x})$ as a Lyapunov function candidate for the policy $\mathbf{u}^+$. Differentiating $V_{\mathbf{u}}(\mathbf{x})$ with respect to the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}^+(\mathbf{x})$ yields

$$\begin{aligned}\dot{V}_{\mathbf{u}}(\mathbf{x}; \mathbf{u}^+) &\equiv \nabla^T V_{\mathbf{u}}(\mathbf{x}) \cdot \big(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}^+(\mathbf{x})\big) \\ &= -\mathbf{S}(\mathbf{x}) - \mathbf{u}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{u}(\mathbf{x}) - 2\mathbf{u}^{+T}(\mathbf{x})\mathbf{R}(\mathbf{x})(\mathbf{u}^+(\mathbf{x}) - \mathbf{u}(\mathbf{x})),\ \forall \mathbf{x} \in R_A(\mathbf{u}),\end{aligned}$$

where (3.20) and (3.24) are substituted in the second equality, and $\mathbf{x} \equiv \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+)$. Then, the application of Young's inequality $2\mathbf{x}^T\mathbf{R}\mathbf{y} \leq \mathbf{x}^T\mathbf{R}\mathbf{x} + \mathbf{y}^T\mathbf{R}\mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ yields

$$\dot{V}_{\mathbf{u}}(\mathbf{x}) \leq -r(\mathbf{x}, \mathbf{u}^+(\mathbf{x})) \prec 0,\ \ \forall \mathbf{x} \in R_A(\mathbf{u}). \tag{3.27}$$

Therefore, by Corollary 3.1, the policy $\mathbf{u}^+$ given by (3.24) is asymptotically stabilizing.

To show the remaining part, fix $\mathbf{z} \in R_A(\mathbf{u})$. Then, by Proposition 3.2, there is $d > 0$ such that the compact subset $\Omega_d(\mathbf{u}) \subseteq R_A(\mathbf{u})$ defined as

$$\Omega_d(\mathbf{u}) := \{\mathbf{x} \in \mathcal{D} : V_{\mathbf{u}}(\mathbf{x}) \leq d\}$$

contains $\mathbf{z}$ in its interior. Moreover,

- $\Omega_d(\mathbf{u})$ is also invariant under the policy $\mathbf{u}^+$ by (3.27), where $\mathbf{x} \equiv \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+)$; this means that $\mathbf{z} \in \Omega_d(\mathbf{u}) \implies \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+) \in \Omega_d(\mathbf{u})\ \forall \tau \geq t$;
- since $V_{\mathbf{u}} \in C^1_{L+}(\mathbf{u})$ and $\Omega_d(\mathbf{u})$ is a compact subset of $R_A(\mathbf{u}) \subseteq \mathcal{D}$, $\nabla V_{\mathbf{u}}(\mathbf{x})$ exists and is finite for all $\mathbf{x} \in \Omega_d(\mathbf{u})$, meaning that so is $\dot{V}_{\mathbf{u}}(\mathbf{x})$.

Therefore, one can integrate (3.27) from '$t = 0$' to '∞' to obtain

$$\begin{aligned}V_{\mathbf{u}^+}(\boldsymbol{z}) = \int_0^\infty r(\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+), \mathbf{u}^+(\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}^+)))\, d\tau &\leq -\int_0^\infty \dot{V}_{\mathbf{u}}(\mathbf{x}_\tau(\mathbf{z}, \mathbf{u}^+))\, d\tau \\ &= V_{\mathbf{u}}(\mathbf{z}) - \lim_{\tau\to\infty} V_{\mathbf{u}}(\mathbf{x}_\tau(\mathbf{z}, \mathbf{u}^+)) < \infty.\end{aligned}$$

In the last inequality, we have used $V_{\mathbf{u}}(\mathbf{z}) < \infty$ and $\lim_{\tau\to\infty} V_{\mathbf{u}}(\mathbf{x}_\tau(\mathbf{z}, \mathbf{u}^+)) = 0$. Here, by asymptotic stability, $\lim_{\tau\to\infty} \mathbf{x}_\tau(\mathbf{z}, \mathbf{u}^+) = \mathbf{0}_n$ holds ($\because$ $\mathbf{z} \in \Omega_d(\mathbf{u})$). Hence, by the continuity of $V_{\mathbf{u}}(\mathbf{z})$ and $V_{\mathbf{u}}(\mathbf{0}_n) = 0$ (see Lemma 3.2 and Definition 2.5), $V_{\mathbf{u}}(\mathbf{x}_\tau(\mathbf{z}, \mathbf{u}^+)) \to 0$ as $\tau \to \infty$. Since all of these hold for any $\mathbf{z} \in R_A(\mathbf{u})$, we have (3.25) and, for all $\mathbf{x} \in R_A(\mathbf{u})$, (3.26) holds. This completes the proof. □

Corollary 3.5. *If $V^*(\mathbf{x})$ is the optimal value function whose derivatives are Lipschitz continuous on its domain, then $\mathbf{u}^*(\mathbf{x})$ given by (3.23) is the optimal admissible policy.*

Proof. Theorem 3.5 implies that $\mathbf{u}^*(\mathbf{x})$ given by (3.23) is an admissible policy and its value function, say $V^+(\mathbf{x})$, satisfies $0 \prec V^+(\mathbf{x}) \leq V^*(\mathbf{x})$. Since $V^*(\mathbf{x})$ is the optimal, $V^+(\mathbf{x})$ cannot be less than $V^*(\mathbf{x})$, so we have $V^+ = V^*$. □

Remark 3.2. *The contribution in this section is to investigate the properties of the value functions and Hamiltonian on the ROAs, where Proposition 3.2 plays a central role in extending the existing results in the literatures [45,46,106] to Theorem 3.5. This contribution on the investigation of the relations between the value function domain and the ROA will be used to propose the global version of policy iteration and IRL algorithms, which is another sole contribution of this dissertation.*

Substituting (3.23) into (3.19) and rearranging the equation yield the well-known HJB equation (3.5). Hence, the optimal policy given in (3.23) can be obtained by solving the HJB equation (3.5) numerically or analytically and then substituting $V^*(\mathbf{x})$. All of the IRL methods shown in Chapters 4, 4.5, and 5 are actually the online methods to solve the HJB equation (3.5), while the approach in Chapter 6 analytically finds the solution of the inverse optimal HJB equation (3.5) corresponding to the given well-designed asymptotocally stabilizing policy.

3.4 Policy Iteration on the Region of Attractions

The HJB equation (3.5) is a partial differential equation of dimension n, so its analytical solution is very hard and, in some cases, impossible to obtain. Due to this reason, a number of numerical methods have been proposed to solve the HJB equation, *e.g.*, [45,46,48,106–109]. policy iteration (PI) is one of these numerical methods that successively updates the value function and the policy by iterations to obtain the optimal value function $V^*(\mathbf{x})$. This idea of PI is the basic concepts of IRLs presented in this dissertation, and its direct extension gives birth to integral PI (I-PI) (see Sections 4.1 and 5.1 for more details), which is one of the fundamental IRL methods from which all of the IRL methods in Chapter 5 are derived.

In this section, using the properties of value functions shown in the previous section, I propose a PI algorithm, called *ideal PI*, that evaluates the value function on the DOA for the corresponding policy. The proposed one can be considered the extension of the domain of the existing local PI [106] upto the ROAs. Here, this extension to the ROA and its analytical results are also the sole contribution of this dissertation, by which the IRL methods shown in the previous work [97] is extended to the ROAs.

Algorithm 3.1 describes the proposed ideal PI, where it is assumed that priori to policy evaluation at each i-th iteration, the ROA $R_A(\mathbf{u}_i)$ of the i-th policy is exactly known, so that the information regarding it can be used in the algorithm. The main part of the ideal PI consists of the two consecutive steps named policy evaluation (line 3) and policy

Algorithm 3.1: Ideal Policy Iteration

```
   Input: an initial admissible policy $\mathbf{u}_0 : \mathcal{D} \to \mathbb{R}^m$.
   Output: the optimal solution $(\mathbf{u}^*, \mathbf{V}^*)$ satisfying (3.23) and (3.5).
1  $i \leftarrow 0$;
2  repeat
3      Policy Evaluation: find the value function $V_{\mathbf{u}_i} : R_A(\mathbf{u}_i) \to \mathbb{R}$ that belongs to
       $C^1_{L+}(\mathbf{u}_i)$ and satisfies
           $\mathcal{H}(\mathbf{x}, \mathbf{u}_i(\mathbf{x}), \nabla V_{\mathbf{u}_i}(\mathbf{x})) = 0 \quad \forall \mathbf{x} \in R_A(\mathbf{u}_i);$      (3.28)
4      Policy Improvement: update the next policy $\mathbf{u}_{i+1} : \Omega \to \mathbb{R}^n$ on a domain
       $\Omega \in \mathfrak{D}(\mathcal{D}, R_A(\mathbf{u}_i))$ whose restriction on $R_A(\mathbf{u}_i)$ satisfies
           $\mathbf{u}_{i+1}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\,\mathbf{G}^T(\mathbf{x})\nabla V_{\mathbf{u}_i}(\mathbf{x}) \;\; \forall \mathbf{x} \in R_A(\mathbf{u}_i);$      (3.29)
5      $i \leftarrow i + 1$;
6  until convergence is met.
```

improvement (line 4). The policy evaluation is the procedure to solve the Hamiltonian equation (3.19) on the domain $R_A(\mathbf{u})$ for the current policy $\mathbf{u} = \mathbf{u}_i$, which yields the corresponding value function $V_{\mathbf{u}_i} : R_A(\mathbf{u}_i) \to \mathbb{R}$ that belongs to the function space $C^1_{L+}(\mathbf{u}_i)$; the policy improvement is the update procedure to yield the next policy $\mathbf{u}_{i+1} : \Omega \to \mathbb{R}^m$ on a domain $\Omega \in \mathfrak{D}(\mathcal{D}, R_A(\mathbf{u}_i))$ that satisfies $\mathbf{u}_{i+1}(\mathbf{x}) = \mathbf{u}^+(\mathbf{x})$ for all $\mathbf{x} \in R_A(\mathbf{u}_i)$, where $\mathbf{u}^+$ is given by (3.24) in Theorem 3.24. To yield the optimal solution $(\mathbf{u}^*, V^*)$, these two procedures are repeated one after another until the convergence is met.

By the repetitive applications of Theorem 3.22 and Theorem 3.24 one after another to the sequences $\{V_{\mathbf{u}_i}\}_{i=0}^{\infty}$ and $\{\mathbf{u}_i\}_{i=0}^{\infty}$ generated by the ideal PI, the following corollary can be obtained.

Corollary 3.6. *The sequences of value functions $\{V_{\mathbf{u}_i}\}_{i=0}^{\infty}$ and policies $\{\mathbf{u}_i\}_{i=0}^{\infty}$, both generated by the ideal PI, have the followings:*

1. *the policy $\mathbf{u}_i$ is admissible for all $i \in \mathbb{Z}_+$;*
2. *for each $i \in \mathbb{Z}_+$, $R_A(\mathbf{u}_i)$ is the invariant subset of $R_A(\mathbf{u}_{i+1})$, i.e.,*
$$\mathbf{z} \in R_A(\mathbf{u}_i) \implies \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}_{i+1}) \in R_A(\mathbf{u}_i) \ \ \forall \tau \geq t; \tag{3.30}$$
3. *each value function $V_{\mathbf{u}_i}$ is the unique solution to the Hamiltonian equation* (3.28) *in $C^1_{L+}(\mathbf{u}_i)$, and for any $i \in \mathbb{Z}_+$ and any $N \in \mathbb{N}$,*
$$0 \prec V^*(\mathbf{x}) \leq V_{\mathbf{u}_{i+N}}(\mathbf{x}) \leq \cdots \leq V_{\mathbf{u}_{i+1}}(\mathbf{x}) \leq V_{\mathbf{u}_i}(\mathbf{x}) < \infty \ \ \forall \mathbf{x} \in R_A(\mathbf{u}_i). \tag{3.31}$$

Remark 3.3. *In Corollary 3.6, $V^*(\mathbf{x}) \leq V_{\mathbf{u}_{i+N}}(\mathbf{x})$ in* (3.31) *definitely holds since V^* is the optimal minimizing value function. Moreover, the second and third statements imply that the DOAs $R_A(\mathbf{u}_i)$ satisfy $R_A(\mathbf{u}_0) \subseteq \cdots \subseteq R_A(\mathbf{u}_i) \subseteq R_A(\mathbf{u}_{i+1}) \subseteq \cdots \subseteq R_A(\mathbf{u}^*)$.*

The ideal PI in Algorithm 3.1 requires the ROA $R_A(\mathbf{u}_i)$ at each i-th iteration, which is difficult to obtain and makes the algorithm complex, except the simple case $R_A(\mathbf{u}_i) = \mathbb{R}^n$; even for this simple case, the evaluations of $(V_{\mathbf{u}_i}, \mathbf{u}_i)$ over the whole state space $\mathbb{R}^n$ is a formidable task. Hence, the ideal PI is approximately done in practice by performing the policy evaluation and improvement only on a bounded subset of the initial ROA $R_A(\mathbf{u}_0)$ that is connected and contains $\mathbf{0}_n$ in its interior [48, 49, 106].

To investigate the local convergence behavior of the ideal PI, let Ω be any compact subset of $R_A(\mathbf{u}_0)$. Then, since the ROA is expanded as the iteration goes on (Remark 3.3), we have $\Omega \subset R_A(\mathbf{u}_i)$ for all $i \in \mathbb{Z}_+$. Moreover, by the third statement of Corollary 3.6, the sequence of value functions $\{V_{\mathbf{u}_i}\}_{i=0}^{\infty}$ generated by the ideal PI is monotonically decreasing on Ω. That is,

$$0 \prec V^*(\mathbf{x}) \leq \cdots \leq V_{\mathbf{u}_{i+1}}(\mathbf{x}) \leq V_{\mathbf{u}_i}(\mathbf{x}) \leq \cdots \leq V_{\mathbf{u}_0}(\mathbf{x}) < \infty \ \ \forall \mathbf{x} \in \Omega. \tag{3.32}$$

Since (3.32) states that $V_{\mathbf{u}_i}$ is monotonically decreasing and lower-bounded by zero, there is a function $\hat{V} : \Omega \to \mathbb{R}_+$ to which $V_{\mathbf{u}_i}$ converges pointwisely on Ω. For the convergence of $V_{\mathbf{u}_i} \to V^*$, however, some additional conditions are necessarily imposed as shown in the next theorem.

Theorem 3.6. *Let $\mathcal{F} : C^1_{L+}(\Omega) \to C^1_{L+}(\Omega)$ be the PI map defined as*

$$\mathcal{F}\{\hat{V}_{\mathbf{u}_{i+1}}\} := \hat{V}_{\mathbf{u}_i},$$

where $\hat{V}_{\mathbf{u}_i}$ is the restriction of the value function $V_{\mathbf{u}_i}$ on Ω. If $\mathcal{F}$ is continuous and $\hat{V}$ belongs to $C^1_{0+}(\Omega)$, then $\hat{V}_{\mathbf{u}_i} \to V^*$ uniformly on Ω.

Proof. Since $\hat{V} \in C^1_{0+}(\Omega)$, it is continuous on Ω, so the convergence $\hat{V}_{\mathbf{u}_i} \to \hat{V}$ is uniform on the compact set Ω by Dini's theorem. Similarly, we have the uniform convergence $\mathcal{F}(\hat{V}_{\mathbf{u}_i}) = \hat{V}_{\mathbf{u}_{i+1}} \to \hat{V}$ on Ω. Therefore, $\mathcal{F}(\hat{V}) = \hat{V}$ by continuity of $\mathcal{F}$, i.e., $\hat{V}$ is the fixed point of $\mathcal{F}$. Since the fixed point of $\mathcal{F}$ corresponds to the optimal solution V^* as shown in Corollary 3.5, we have $\hat{V} = V^*$ on Ω. □

Remark 3.4. *Notice that $V_{\mathbf{u}_i}(\mathbf{x})$ and $\mathbf{u}_i(\mathbf{x})$ in Algorithm 3.1 are always finite on a compact set Ω of $R_A(\mathbf{u}_0)$. Hence, some finite convergence criterion can be established in line 6 of Algorithm 5.1 to check the convergence, for example,*

$$\sup_{\mathbf{x} \in \Omega} \|\mathbf{u}_i(\mathbf{x}) - \mathbf{u}_{i-1}(\mathbf{x})\| < \varepsilon,$$

where $0 < \varepsilon \ll 1$ is an error tolerant constant.

Chapter 4

Classifications, Stability, and Convergence of Fundamental IRL

This chapter introduces and newly classifies the fundamental IRL methods applied to the CT LQR problem:

$$\text{minimize } V_{\mathbf{u}}(\mathbf{x}_t) = \int_t^{\infty} \left(\mathbf{x}_\tau^T \mathbf{S} \mathbf{x}_\tau + \mathbf{u}_\tau^T \mathbf{R} \mathbf{u}_\tau\right) d\tau \tag{4.1}$$

$$\text{subject to } \dot{\mathbf{x}}_\tau = \mathbf{A}\mathbf{x}_\tau + \mathbf{B}\mathbf{u}(\mathbf{x}_\tau), \;\; \mathbf{x}(t) = \mathbf{z} \in \mathbb{R}^n, \;\; \mathbb{S} = \{\mathbf{0}_n\} \tag{4.2}$$

and then analyze them regarding their stability and convergence. Here,

- $(\mathbf{A}, \mathbf{B})$ is a stabilizable pair of the matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$;
- $(\mathbf{S}, \mathbf{A})$ is a detectable pair of the matrices $\mathbf{S} \succeq \mathbf{0}_{n\times n}$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$;
- $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^m$ is a linear policy of the form $\mathbf{u}(\mathbf{x}) = -\mathbf{K}\mathbf{x}$ for some $\mathbf{K} \in \mathbb{R}^{m\times n}$.

Definition 4.1. *If a control input function $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^m$ is given in a form $\mathbf{u}(\mathbf{x}) = -\mathbf{K}\mathbf{x}$ for some constant gain matrix $\mathbf{K} \in \mathbb{R}^{m\times n}$, then it is called a linear policy. We also say that the gain matrix $\mathbf{K} \in \mathbb{R}^{m\times n}$ in $\mathbf{u} = -\mathbf{K}\mathbf{x}$ is a linear policy.*

For notational convenience, the matrices $\mathbf{A_K}$ and $\mathbf{S_K} \succeq \mathbf{0}_{n\times n}$ for a linear policy $\mathbf{K}$ are defined as $\mathbf{A_K} := \mathbf{A} - \mathbf{B}\mathbf{K}$ and $\mathbf{S_K} := \mathbf{S} + \mathbf{K}^T\mathbf{R}\mathbf{K}$, respectively. The classical linear control theory shows that if the closed-loop system matrix $\mathbf{A_K}$ is Hurwitz, then the linear policy $\mathbf{K}$ exponentially stabilizing the linear system (4.2). Here, we call this kind of policy Hurwitz rather than exponentially stabilizing.

Definition 4.2. *A linear policy $\mathbf{u} = -\mathbf{K}\mathbf{x}$ is said to be Hurwitz, or a Hurwitz poicy, if the closed-loop matrix $\mathbf{A_K}$ is Hurwitz.*

When the linear policy $\mathbf{u}$ is given and fixed, then the performance index $V_{\mathbf{u}}(\mathbf{x}_t)$ in (4.1) is called a value function for the given linear policy. Using the matrix notations $\mathbf{A_K}$ and $\mathbf{S_K}$, the value function $V_{\mathbf{u}}(\mathbf{x}_t)$ for a linear policy $\mathbf{u} = -\mathbf{K}\mathbf{x}$ can be expressed as

$$V_{\mathbf{u}}(\mathbf{x}_t) = \mathbf{x}_t^T \left(\int_t^\infty e^{\mathbf{A}_{\mathbf{K}}^T(\tau - t)} \mathbf{S_K}\, e^{\mathbf{A_K}(\tau - t)}\, d\tau \right) \mathbf{x}_t \equiv \mathbf{x}_t^T \mathbf{P_K} \mathbf{x}_t,$$

where $\mathbf{P_K} \succeq \mathbf{0}_{n \times n}$ is defined as

$$\mathbf{P_K} := \int_0^\infty e^{\mathbf{A}_{\mathbf{K}}^T \tau} \mathbf{S_K}\, e^{\mathbf{A_K}\tau}\, d\tau. \tag{4.3}$$

In this LQR case, the value function $V_{\mathbf{u}}(\mathbf{x}) = \mathbf{x}_t^T \mathbf{P_K} \mathbf{x}_t$ is positive semi-definite since so is $\mathbf{S}$. If $\mathbf{S}$ is positive definite, then it guarantees the positive definiteness of $\mathbf{P_K}$ for any Hurwitz policy $\mathbf{K}$. Moreover, (4.3) is always finite as lone as $\mathbf{K}$ is Hurwitz, so closed-loop stability always implies admissibility.

The LQR problem (4.1) and (4.2) is the special case of the nonlinear optimal control problem (3.1) and (3.4) with $\mathbf{f}(x) = \mathbf{A}\mathbf{x}$, $\mathbf{G}(\mathbf{x}) = \mathbf{B}$, $S(\mathbf{x}) = \mathbf{x}^T\mathbf{S}\mathbf{x}$, $\mathbf{R}(\mathbf{x}) = \mathbf{R}$, and $\mathbb{S} = \{\mathbf{0}_n\}$. On the other hand, it is slightly generalized from Assumption 3.4. in that $\mathbf{S}$ is positive *semi-definite*. By substituting $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, $\mathbf{G}(\mathbf{x}) = \mathbf{B}$, $\nabla V_{\mathbf{u}}(\mathbf{x}) = 2\mathbf{P}\mathbf{x}$, $\mathbf{u} = -\mathbf{K}\mathbf{x}$, and $r(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T\mathbf{S_K}\mathbf{x}$, the Hamiltonian equation (3.19) becomes the Lyapunov matrix equation $\mathcal{L}(\mathbf{K}, \mathbf{P_K}) = \mathbf{0}_{n\times n}$,[1] where the Lyapunov operator $\mathcal{L}(\mathbf{K}, \mathbf{P})$ is defined as

$$\mathcal{L}(\mathbf{K}, \mathbf{P}) := \mathbf{A}_{\mathbf{K}}^T\mathbf{P} + \mathbf{P}\mathbf{A_K} + \mathbf{S_K}. \tag{4.4}$$

Moreover, the policy update rule (3.24) can be expressed as $\mathbf{u}^+ = -\mathbf{K}^+\mathbf{x}$, where $\mathbf{K}^+ := \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P_K}$; the HJB equation (3.5) becomes the standard algebraic Riccati equation (ARE) $\mathcal{R}(\mathbf{P}^*) = 0$, where the Riccati operator $\mathcal{R}(\mathbf{P})$ is defined as

$$\mathcal{R}(\mathbf{P}) := \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{S}$$

[1] $\mathcal{L}(\mathbf{K}, \mathbf{P_K}) = \mathbf{0}_{n\times n}$ can be easily verified by substituting (4.3) into $\mathcal{L}(\mathbf{K}, \mathbf{P_K})$ and using the definition (4.4) and standard calculus.

and satisfies $\mathcal{R}(\mathbf{P}) = \mathcal{L}(\mathbf{K},\mathbf{P})|_{\mathbf{K}=\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}}$; $\mathbf{P}^*$ and $\mathbf{K}^*$ are optimal matrix solutions that satisfy $V^*(\mathbf{x}) = \mathbf{x}^T\mathbf{P}^*\mathbf{x}$ and $\mathbf{u}^* = -\mathbf{K}^*\mathbf{x}$, respectively. By the stabilizability and detectability of $(\mathbf{S},\mathbf{A},\mathbf{B})$, there exists the unique positive semi-definite solution $\mathbf{P}^*$ of the ARE $\mathcal{R}(\mathbf{P}^*) = 0$.

The fundamental IRL algorithms introduced in this chapter are integral PI (I-PI), integral generalized PI (I-GPI), integral value iteration (I-VI), and infinitesimal GPI; the sole contribution of this chapter is to suggest the new classification of these fundamental IRL algorithms shown in Section 4.4, and then analyze their stability and convergence along the new classifications.

Before the explorations of these IRL algorithms, it is necessary to review the PI for LQR problems (4.1) and (4.2) shown in Algorithm 4.1, which is the counterpart of the ideal PI described in Algorithm 3.1. This basic PI is closely related with all of the IRL methods. Roughly, Algorithm 4.2 can be considered as a revolving process to solve the Lyapunov equation $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_{\mathbf{K}_i}) = \mathbf{0}_{n\times n}$ in policy evaluation and update $\mathbf{K}_{i+1}$ by the rule $\mathbf{K}_{i+1} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{\mathbf{K}_i}$ in policy improvement. Since a Hurwitz policy is always admissible in the LQR case, the ideal PI applied to the LQR problems just needs a initial Hurwitz policy, rather than the admissible one. So, applying Corollary 3.6 and Theorem 3.6 to the ideal PI for LQR problems, we obtain the following monotone convergence result.

Corollary 4.1. *Assume that* $\mathbf{S}$ *is positive definite. Then, the matrix sequences* $\{\mathbf{P}_{\mathbf{K}_i}\}_{i=0}^{\infty}$ *and* $\{\mathbf{K}_i\}_{i=0}^{\infty}$ *generated by Algorithm 4.1 have the followings.*

1. $\mathbf{K}_i$ *is Hurwitz for all* $i \in \mathbb{Z}_+$*;*

2. $(\mathbf{P}_i,\mathbf{K}_i)$ *monotonically converges to the optimal solution* $(\mathbf{P}^*,\mathbf{K}^*)$ *in a sense that*

$$\begin{cases} \mathbf{0}_{n\times n} \prec \mathbf{P}^* \preceq \cdots \preceq \mathbf{P}_{\mathbf{K}_{i+1}} \preceq \mathbf{P}_{\mathbf{K}_i} \preceq \cdots \preceq \mathbf{P}_{\mathbf{K}_0}. \\ \lim_{i\to\infty} \mathbf{P}_{\mathbf{K}_i} = \mathbf{P}^* \text{ and } \lim_{i\to\infty} \mathbf{K}_i = \mathbf{K}^*. \end{cases} \tag{4.5}$$

Remark 4.1. *The statement in Corollary 4.1 is also proven in [93,110,111] for a positive semi-definite matrix* $\mathbf{S}$*. In addition, Kleinman has proven in his literature [110] that the convergence is quadratic, so there is* $\kappa > 0$ *such that* $\|\mathbf{P}_{\mathbf{K}_{i+1}} - \mathbf{P}^*\| \leq \kappa\|\mathbf{P}_{\mathbf{K}_i} - \mathbf{P}^*\|^2$ *for*

Algorithm 4.1: Ideal Policy Iteration for LQR Problems

```
Input: an initial Hurwitz policy $\mathbf{K} \in \mathbb{R}^{m \times n}$.
Output: $(\mathbf{K}^*, \mathbf{P}^*)$ satisfying $\mathbf{K}^* = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}^*$ and $\mathcal{R}(\mathbf{P}^*) = 0$.
$i \leftarrow 0$;
repeat
    Policy Evaluation: find the value function matrix $\mathbf{P}_{\mathbf{K}_i} \in \mathbb{R}^{n \times n}$ satisfying
        $\mathcal{L}(\mathbf{P}_{\mathbf{K}_i}) = \mathbf{0}_{n \times n}$;
    Policy Improvement: update the next policy $\mathbf{K}_{i+1} \in \mathbb{R}^{m \times n}$ by
        $\mathbf{K}_{i+1} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{\mathbf{K}_i}$;
    $i \leftarrow i + 1$;
until convergence is met.
```

all $i \in \mathbb{Z}_+$; in fact, the I-PI for LQR is equivalent to the Newton method of the form

$$\mathbf{P}_{\mathbf{K}_{i+1}} = \mathbf{P}_{\mathbf{K}_i} + \left(\mathcal{L}'_{\mathbf{K}_i, \mathbf{P}_{\mathbf{K}_i}}\right)^{-1} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{\mathbf{K}_i}), \tag{4.6}$$

where $\mathcal{L}'_{\mathbf{K}_i, \mathbf{P}_{\mathbf{K}_i}}$ is the Frechet derivative of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{\mathbf{K}_i})$ taken with respect to $\mathbf{P}_{\mathbf{K}_i}$. Since the nonlinear optimal control problem (3.1) *and* (3.4) *under Assumption 3.4 can be approximated as an LQR problem* (4.1) *and* (4.2) *near the origin $\mathbf{0}_n$, this quadratic convergence of PI for LQR problems also provides the same convergence property for the ideal PI (Algorithm 3.1) in a local region near the origin $\mathbf{0}_n$.*

Remark 4.2. *While Algorithm 4.1 is an offline method to solve the LQR problems and needs the exact information of the matrices $\mathbf{A}$ and $\mathbf{B}$, the fundamental IRL methods in this chapter can be implemented in online fashion and does not require the knowledge of the system matrix $\mathbf{A}$.*

Throughout this chapter, the closed-loop matrix $\mathbf{A}_{\mathbf{K}_i}$ generated by PI or any IRL methods is denoted by $\mathbf{A}_i \equiv \mathbf{A}_{\mathbf{K}_i}$ for notational convenience.

Algorithm 4.2: Integral Policy Iteration for LQR Problems

```
Input: an initial Hurwitz policy $\mathbf{u}_0 = -\mathbf{K}_0\mathbf{x}$ ($\mathbf{K} \in \mathbb{R}^{m\times n}$).
Output: $(\mathbf{K}^*, \mathbf{P}^*)$ satisfying $\mathbf{K}^* = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}^*$ and $\mathcal{R}(\mathbf{P}^*) = 0$.
$i \leftarrow 0$;
repeat
    Policy Evaluation: find the value function $V_{\mathbf{u}_i}(\mathbf{x}) = \mathbf{x}^T\mathbf{P}_{\mathbf{K}_i}\mathbf{x}$ satisfying
        $V_{\mathbf{u}_i}(\mathbf{x}_t) = \int_t^{t+T_s} \mathbf{x}_\tau^T\mathbf{S}_{\mathbf{K}_i}\mathbf{x}_\tau \, d\tau + V_{\mathbf{u}_i}(\mathbf{x}_{t+T_s}), \quad \forall \mathbf{x}_t \in \mathbb{R}^n, \qquad (4.7)$
    where $\mathbf{x}_\tau = e^{\mathbf{A}_i(t-\tau)}\mathbf{x}_t$;
    Policy Improvement: update the next policy $\mathbf{u}_{i+1} = -\mathbf{K}_{i+1}\mathbf{x}$ by
        $\mathbf{K}_{i+1} = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{\mathbf{K}_i}; \qquad (4.8)$
    $i \leftarrow i + 1$;
until convergence is met.
```

4.1 Integral Policy Iteration

I-PI is the most fundamental IRL algorithm that can be directly obtained from Algorithm 4.1. By integrating the Hamiltonian equation $\mathbf{x}^T\mathcal{L}(\mathbf{P}_{\mathbf{K}_i})\mathbf{x} = 0$ in time from t to $t + T_s$ along the trajectory generated by the linear system $\dot{\mathbf{x}} = \mathbf{A}_i\mathbf{x} \equiv (\mathbf{A} - \mathbf{B}\mathbf{K}_i)\mathbf{x}$ and then applying Lemma 2.2, one can show that

$$\begin{aligned}\mathbf{x}_t^T e^{\mathbf{A}_i^T T_s}\mathbf{P}_{\mathbf{K}_i}e^{\mathbf{A}_i T_s}\mathbf{x}_t - \mathbf{x}_t^T\mathbf{P}_{\mathbf{K}_i}\mathbf{x}_t &= \int_t^{t+T_s} \mathbf{x}_\tau^T(\mathbf{A}_i^T\mathbf{P}_{\mathbf{K}_i} + \mathbf{P}_{\mathbf{K}_i}\mathbf{A}_i)\mathbf{x}_\tau \, d\tau \\ &= -\int_t^{t+T_s} \mathbf{x}_\tau^T\mathbf{S}_{\mathbf{K}_i}\mathbf{x}_\tau \, d\tau.\end{aligned}$$

Substituting $\mathbf{x}_{t+T} = e^{\mathbf{A}_i T}\mathbf{x}_t$ and rearranging the equation yields the following temporal difference (TD) formula:

$$\mathbf{x}_t^T\mathbf{P}_{\mathbf{K}_i}\mathbf{x}_t = \int_t^{t+T} \mathbf{x}_\tau^T\mathbf{S}_{\mathbf{K}_i}\mathbf{x}_\tau \, d\tau + \mathbf{x}_{t+T}^T\mathbf{P}_{\mathbf{K}_i}\mathbf{x}_{t+T}. \tag{4.9}$$

Therefore, we obtain the I-PI for LQR problems (Algorithm 4.2) by replacing the Lyapunov matrix equation in policy evaluation of Algorithm 4.1 by (4.9) that will be solved in the algorithm to uniquely determine the matrix $\mathbf{P}_{\mathbf{K}_i}$. All of the other parts are same to the ideal PI applied to LQR problems (Algorithm 4.1).

Remark 4.3. *Since I-PI (Algorithm 4.2) is inherently equal to Algorithm 4.1, it possesses the same Hurwitz and 2^{nd}-order monotone convergence properties described in Corollary 4.1 and Remark 4.1. Moreover, I-PI can be online implementable without knowing the system matrix $\mathbf{A}$ (see [111]) while the ideal PI (Algorithm 4.1) cannot as mentioned in Remark 4.2. All of the IRL methods in this dissertation can be implemented in a similar manner to I-PI without knowing the drift dynamics, i.e., the system matrix $\mathbf{A}$, and hence can be considered a class of partially model-free adaptive optimal control methods.*

4.2 Integral Generalized Policy Iteration

I-PI presented in the previous section provides the 2^{nd}-order convergence speed as explained in Remark 4.3 and the policy is always improved as shown in (4.5) However, I-PI needs the initial Hurwitz policy to run, and its policy evaluation step can be computationally intractable when the difference $\boldsymbol{\phi}(\mathbf{x}_t) - \boldsymbol{\phi}(\mathbf{x}_{t+T_s})$ is not excited but remains constant for a long period. Here, $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^N$ is the activation function that approximately express the value function as $V_{\mathbf{u}_i}(\mathbf{x}) \approx \mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x})$ for some weight vector $\mathbf{w}_i \in \mathbb{R}^N$.

To overcome these limitations of I-PI, a class of IRL algorithms known as integral generalized PI (I-GPI) is proposed in [49]. Actually, I-GPI contains the other fundamental IRL methods—I-PI, I-VI, and infinitesimal GPI—as special or limiting cases, and the idea is to approximate the I-TD equation (4.7) by the finite k-number of Bellman fixed iterations, where $k \in \mathbb{N}$ is called *the iteration horizon*. Actually, this idea is originated from the similar concept of modified PI [35,36], where the TD equation for a finite Markov decision process (MDP) was approximated by the finite number of Bellman fixed iterations. Unlike I-PI, the I-GPI algorithm does not require any initial Hurwitz policy and instead of the difference "$\boldsymbol{\phi}(\mathbf{x}_t) - \boldsymbol{\phi}(\mathbf{x}_{t+T_s})$", the excitation of $\boldsymbol{\phi}(\mathbf{x}_t)$ itself is sufficient for the I-GPI

Algorithm 4.3: Generalized Integral Policy Iteration for LQR Problems

```
  Input:
    • an initial policy $\mathbf{u}_0 = -\mathbf{K}_0\mathbf{x}$ not necessarily Hurwitz;
    • an initial value function matrix $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$.
  Output: $(\mathbf{K}^*, \mathbf{P}^*)$ satisfying $\mathbf{K}^* = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}^*$ and $\mathcal{R}(\mathbf{P}^*) = 0$.
1 $i \leftarrow 0$;
2 repeat
      Approximate Policy Evaluation:
3         $V_{i|0}(\mathbf{x}) \leftarrow \mathbf{x}^T\mathbf{P}_i\mathbf{x}$;
4         for $j = 0, 1, 2, \cdots, k-1$ do
5             find the next approximation $V_{i|j+1}(\mathbf{x}) := \mathbf{x}^T\mathbf{P}_{i|j+1}\mathbf{x}$ by solving
                  $V_{i|j+1}(\mathbf{x}_t) = \int_t^{t+T_s} \mathbf{x}_\tau^T\mathbf{S}_{\mathbf{K}_i}\mathbf{x}_\tau \, d\tau + V_{i|j}(\mathbf{x}_{t+T_s}),$    (4.10)
              for all $\mathbf{x}_t \in \mathbb{R}^n$, where $\mathbf{x}_\tau = e^{\mathbf{A}_i(t-\tau)}\mathbf{x}_t$;
6         $\mathbf{P}_{i+1} \leftarrow \mathbf{P}_{i|k}$;
7     Policy Improvement: update the next policy $\mathbf{u}_{i+1} = -\mathbf{K}_{i+1}\mathbf{x}$ by
              $\mathbf{K}_{i+1} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{i+1};$    (4.11)
8     $i \leftarrow i + 1$;
9 until convergence is met.
```

algorithm to run.

The I-GPI algorithm for LQR problems is shown in Algorithm 4.3. In approximate policy evaluation (lines 3–6), the agent generates the finite matrix sequence $\{\mathbf{P}_{i|j}\}_{j=0}^{k}$ that sequentially satisfies $\mathbf{P}_{i|0} = \mathbf{P}_i$ and (4.10) for all $j \in \{0, 1, 2, \cdots, k-1\}$ and all $\mathbf{x}_t \in \mathbb{R}^n$. Then the next value function matrix is given by $\mathbf{P}_{i+1} = \mathbf{P}_{i|k}$. The policy improvement of I-GPI (line 7) is same to that of I-PI, but the agent updates the next policy $\mathbf{u}_{i+1}$ using $\mathbf{P}_{i+1}$ instead. These revolving policy evaluation and improvement steps are repeated again and again until convergence as the PI and I-PI methods in the previous two sections. Notice that $\mathbf{P}_{i+1} \neq \mathbf{P}_{\mathbf{K}_i}$ in general.

When it comes to the convergence of I-GPI, Vrabie [51] has shown in a nonlinear optimal control framework that if $\mathbf{u}_i$ is admissible, the approximated value function $V_{i|k}$ obtained in policy evaluation converges to the exact one $V_{\mathbf{u}_i}$ as $k \to \infty$. This implies that I-GPI with an initial admissible policy becomes I-PI in the limit $k \to \infty$. In LQR case, this means that $\mathbf{P}_{i|k}$ converges to $\mathbf{P}_{\mathbf{K}_i}$ as $k \to \infty$ under a Hurwitz policy $\mathbf{K}_i$, so Algorithm 4.3 with $k = \infty$ under an initial Hurwitz policy becomes equivalent to Algorithm 4.2. These convergence characteristics will be revisited in Section 4.4, where a new classification of the IRL methods is given with respect to the update horizon $\hbar \in \mathbb{R}_+$ defined as $\hbar := kT_s$.

4.3 Integral Value Iteration and Infinitesimal GPI

Integral value iteration (I-VI) is referred to as the I-GPI algorithm with $k = 1$ that executes the Bellman iteration (4.10) only one time at each iteration. Therefore, I-VI has the minimum computational costs while I-PI (I-GPI with $k = \infty$) requires theoretically the infinite number of Bellman iterations in policy evaluation. With this respect, the I-GPI has the two extreme tips—I-VI and I-PI corresponding to the minimum and maximum computational costs, respectively.

The policy evaluation of I-VI at i-th iteration can be viewed as a procedure to yield $\mathbf{P}_{i+1} \succeq \mathbf{0}_{n\times n}$ that satisfies

$$\mathbf{x}_t^T \mathbf{P}_{i+1} \mathbf{x}_t = \int_t^{t+T_s} \mathbf{x}_\tau^T \mathbf{S}_{\mathbf{K}_i} \mathbf{x}_\tau \, d\tau + \mathbf{x}_{t+T_s}^T \mathbf{P}_i \mathbf{x}_{t+T_s} \tag{4.12}$$

for all $\mathbf{x}_t \in \mathbb{R}^n$. Substituting $\mathbf{x}_\tau = e^{\mathbf{A}_i(\tau - t)}\mathbf{x}_t$ ($\tau \in [t, t+T_s]$) into (4.12) and rearranging it, we obtain the matrix formula "$\mathbf{P}_{i+1} - e^{\mathbf{A}_i^T T_s}\mathbf{P}_i e^{\mathbf{A}_i T_s} = \int_0^{T_s} e^{\mathbf{A}_i^T \tau}\mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \tau}\, d\tau$," which can be rewritten as

$$\mathbf{P}_{i+1} - \mathbf{P}_i = \int_0^{T_s} e^{\mathbf{A}_{\mathbf{K}_i}^T \tau} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) e^{\mathbf{A}_{\mathbf{K}_i}\tau} \tag{4.13}$$

by applying Lemma 2.1. Assuming $\mathbf{K}_0 = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_0$, then we have $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) = \mathcal{R}(\mathbf{P}_i)$ for all $i \in \mathbb{Z}_+$. So, dividing both sides of (4.13) by T_s and then limiting $T_s \to 0$ yields the following forward-in-time differential Riccati equation (DRE)

$$\dot{\mathbf{P}}_t = \mathcal{R}(\mathbf{P}_t), \;\; (0 \leq t < \infty), \tag{4.14}$$

which is the infinitesimal version of I-VI ($T_s \to 0$), and we call it in this dissertation "infinitesimal GPI" for LQR problems. In this limit case, under the zero initial condition $\mathbf{P}_0 = \mathbf{0}_{n\times n}$, it has been shown that $\mathbf{P}_t$ generated by the forward-in-time DRE monotonically converges to $\mathbf{P}^*$ with the monotonicity $\mathbf{0}_{n\times n} \preceq \mathbf{P}_{t_1} \preceq \mathbf{P}_{t_2} \preceq \mathbf{P}^*$ for $0 \leq t_1 \leq t_2$ [104, Theorem 16.4.3]. This type of monotone convergence is called VI-mode convergence and will be revisited and generalized in Section 4.5.

4.4 Equivalence Classes and Classifications of IRLs

By the mathematical analysis in terms of the update horizon $\hbar \in \mathbb{R}_+$ given by $\hbar = kT_s$, the product of the iteration horizon $k \in \mathbb{N}$ and the time horizon $T_s \in \mathbb{R}_+$, this section shows that any I-GPI algorithms that have the same update horizon $\hbar$ can be considered the same in the iteration domain, and that "infinitesimal GPI" and I-PI are the special cases of I-GPI in the limit $\hbar \to 0$ and $\hbar \to \infty$, respectively. From this result, the IRLs presented in this chapter are classified in terms of the update horizon $\hbar$.

The analysis will be done in LQR frameworks based on the dynamic programming (DP) operator $\mathcal{T}_{\mathbf{K}}^{T_s} : X \to X$, regarding the linear dynamics (4.2), defined on the space X of continuous functionals $V : \mathbb{R}^n \to \mathbb{R}$, at fixed time $t \geq 0$, as

$$\mathcal{T}_{\mathbf{K}}^{T_s} V(\mathbf{x}_t) := \int_t^{t+T_s} \mathbf{x}_\tau^T \mathbf{S}_{\mathbf{K}}\, \mathbf{x}_\tau \, d\tau + V(\mathbf{x}_{t+T_s}), \tag{4.15}$$

where $\mathbf{x}_\tau = e^{\mathbf{A}_{\mathbf{K}}(\tau - t)}\mathbf{x}_t$. This DP operator has the following property.

Lemma 4.1. *If $V(\mathbf{x}_t)$ is positive semi-definite, then so is its DP operation $\mathcal{T}_{\mathbf{K}}^{T_s} V(\mathbf{x}_t)$ for any policy $\mathbf{u} = -\mathbf{K}\mathbf{x}$ and any $T_s > 0$.*

Proof. The proof is done by the fact that the integral in (4.15) is positive semi-definite due to $\mathbf{S}_{\mathbf{K}} \succeq \mathbf{0}_{n\times n}$; the second term $V(\mathbf{x}_{t+T_s})$ is positive semi-definite by assumption. □

Next, the generalized DP operator $(\mathcal{T}_{\mathbf{K}}^{T_s})^k$ is defined for $k \in \mathbb{Z}_+$ and $T_s \in \mathbb{R}_+$ as

$$\begin{cases} (\mathcal{T}_{\mathbf{K}}^{T_s})^0 V(\mathbf{x}_t) := V(\mathbf{x}_t), \\ (\mathcal{T}_{\mathbf{K}}^{T_s})^{k+1} V(\mathbf{x}_t) := (\mathcal{T}_{\mathbf{K}}^{T_s})^k [\mathcal{T}_K^{T_s} V(\mathbf{x}_t)], \end{cases}$$

at fixed time $t \geq 0$. Here, k is referred to as the number of the DP operation $\mathcal{T}_{\mathbf{K}}^{T_s}[\cdot]$ and will be matched with the iteration horizon k in I-GPI. Indeed, the value function $V_{\mathbf{u}}(\mathbf{x}) = \mathbf{x}^T \mathbf{P}_{\mathbf{K}} \mathbf{x}$ for a Hurwitz policy $\mathbf{u} = -\mathbf{K}\mathbf{x}$ can be expressed as the I-TD form:

$$V_{\mathbf{u}}(\mathbf{x}_t) = \int_t^{t+T_s} \mathbf{x}_\tau^T \mathbf{S}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau + \underbrace{\int_{t+T_s}^{\infty} \mathbf{x}_\tau^T \mathbf{S}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau}_{=V_{\mathbf{u}}(\mathbf{x}_{t+T_s})} = \mathcal{T}_{\mathbf{K}}^{T_s} V_{\mathbf{u}}(\mathbf{x}_t). \tag{4.16}$$

In addition, the generalized DP operator has the following property:

Theorem 4.1. *For the update horizon $\hbar = kT_s$, the DP operator $\mathcal{T}_{\mathbf{K}}^{T_s}$ and its generalized operator $(\mathcal{T}_{\mathbf{K}}^{T_s})^k$ for a continuous functional $V(\mathbf{x})$ satisfy*

$$(\mathcal{T}_{\mathbf{K}}^{T_s})^k V(\mathbf{x}_t) = \mathcal{T}_{\mathbf{K}}^{\hbar} V(\mathbf{x}_t). \tag{4.17}$$

Moreover, if K is Hurwitz and $V(\mathbf{0}_n) = 0$, then $\lim_{\hbar \to \infty} \mathcal{T}_{\mathbf{K}}^{\hbar} V(\mathbf{x}_t) = V_{\mathbf{u}}(\mathbf{x}_t)$.

Proof. Consider the sequence $\{W_i(\mathbf{x}_t)\}_{i=0}^k$ of continuous functionals which is defined by $W_0(\mathbf{x}_t) := V(\mathbf{x}_t)$ and $W_i(\mathbf{x}_t) := \mathcal{T}_K^{T_s} W_{i-1}(\mathbf{x}_t)$ for $i = 1, 2, 3, \cdots, k$. Then, obviously, $(\mathcal{T}_{\mathbf{K}}^{T_s})^k V(\mathbf{x}_t) = W_k(\mathbf{x}_t)$ holds and thereby, one has

$$\begin{aligned} (\mathcal{T}_{\mathbf{K}}^{T_s})^k V(\mathbf{x}_t) &= \int_t^{t+T_s} \mathbf{x}_\tau^T \mathbf{Q}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau + \mathcal{T}_{\mathbf{K}}^{T_s} \big[\underbrace{(\mathcal{T}_{\mathbf{K}}^{T_s})^{k-2} V(\mathbf{x}_{t+T_s})}_{=W_{k-2}(\mathbf{x}_{t+T_s})} \big] \\ &= \int_t^{t+2T_s} \mathbf{x}_\tau^T \mathbf{Q}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau + \mathcal{T}_{\mathbf{K}}^{T_s} W_{k-3}(\mathbf{x}_{t+2T_s}) \\ &\ \vdots \\ &= \int_t^{t+(k-1)T_s} \mathbf{x}_\tau^T \mathbf{Q}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau + \mathcal{T}_{\mathbf{K}}^{T_s} W_0(\mathbf{x}_{t+(k-1)T_s}) \\ &= \int_t^{t+kT_s} \mathbf{x}_\tau^T \mathbf{Q}_{\mathbf{K}} \mathbf{x}_\tau \, d\tau + W_0(\mathbf{x}_{t+kT_s}) \\ &= \mathcal{T}_{\mathbf{K}}^{kT_s} V(\mathbf{x}_t), \end{aligned}$$

which completes the proof of (4.17). Moreover, assume that $\mathbf{K}$ is Hurwitz and $V(\mathbf{0}_n) = 0$.

Then, by continuity at $\mathbf{0}_n$, we have $\lim_{\hbar\to\infty} V(\mathbf{x}_{t+\hbar}) = 0$. Hence, by the definition (4.15) of DP operator,

$$\lim_{\hbar\to\infty} \mathcal{T}_{\mathbf{K}}^{\hbar} V(\mathbf{x}_t) = \int_t^{\infty} \mathbf{x}_\tau^T \mathbf{Q}_{\mathbf{K}}\, \mathbf{x}_\tau \, d\tau = V_{\mathbf{u}}(\mathbf{x}),$$

which completes the proof. □

Using the DP operator, the one-step Bellman iteration (4.10) at time $t \geq 0$ can be expressed as $V_{i|j+1}(\mathbf{x}_t) = \mathcal{T}_{\mathbf{K}_i}^{T_s} V_{i|j}(\mathbf{x}_t)$. Furthermore, using the generalized DP operator and applying Theorem 4.1, we can see that $V_{i+1}(\mathbf{x}) = V_{i|k}(\mathbf{x})$ obtained in approximate policy evaluation of I-GPI satisfies

$$V_{i+1}(\mathbf{x}_t) = \mathcal{T}_{\mathbf{K}_i}^{\hbar} V_i(\mathbf{x}_t) \tag{4.18}$$

for all $\mathbf{x} \in \mathbb{R}^n$. Moreover, if $\mathbf{u}_i = -\mathbf{K}_i\mathbf{x}$ is Hurwitz, then $V_{i+1} \to V_{\mathbf{u}_i}$ as $\hbar \to \infty$ by Theorem 4.1. From these observations, we have the following corollary regarding the equivalence classes of I-GPI including in the limit $\hbar \to \infty$.

Corollary 4.2. *The I-GPI algorithms that have the same update horizon $\hbar$ yield the same sequences $\{\mathbf{P}_i\}_{i=0}^{\infty}$ and $\{\mathbf{K}_i\}_{i=0}^{\infty}$, so they can be considered the equivalents in the iteration domain. Moreover, if $\mathbf{K}_0$ is Hurwitz, then, in the limit $\hbar \to \infty$, I-GPI (Algorithm 4.3) becomes the I-PI (Algorithm 4.2).*

Corollary 4.2 states that under Hurwitz $\mathbf{K}_i$, the difference $|V_{i+1}(\mathbf{x}_t) - V_{\mathbf{u}_i}(\mathbf{x}_t)|$ can be made arbitrarily small by increasing $\hbar$. Here, the update horizon $\hbar$ can be enlarged by either increasing k or T_s. However, the larger k is, the higher is the computational complexity; the larger T_s is, the slower the learning speed in the time domain is. Hence, there exists a trade-off between the computational complexity and learning speed in approximate policy evaluation, and one should carefully determine these parameters k, T_s, and of course, $\hbar$ $(= kT_s)$.

On the other hand, by the same procedure to (4.12)–(4.14) with T_s replaced by the update horizon $\hbar$, one can also show that in the limit $\hbar \to 0$, the GPI algorithms become the infinitesimal GPI (4.14). Therefore, considering in mind this and Corollary 4.2, all of the I-GPI algorithms can be classified in terms of the update horizon $\hbar$ as shown in Fig.

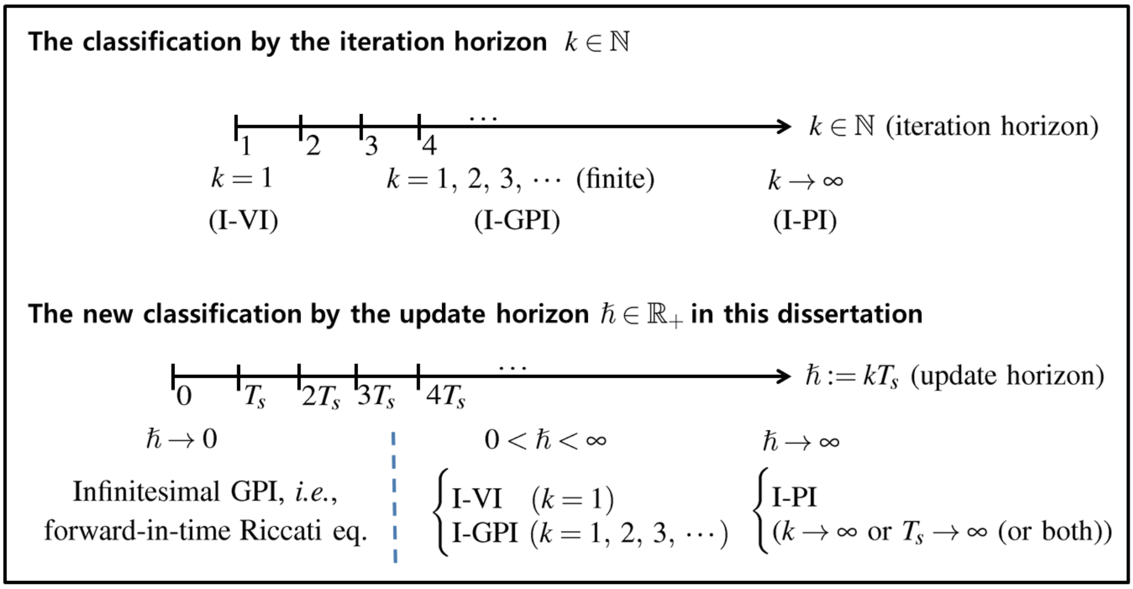

The classification by the iteration horizon $k \in \mathbb{N}$
1
2
3
4
$\cdots$
$k \in \mathbb{N}$ (iteration horizon)
$k = 1$
(I-VI)
$k = 1, 2, 3, \cdots$ (finite)
(I-GPI)
$k \to \infty$
(I-PI)
The new classification by the update horizon $\hbar \in \mathbb{R}_+$ in this dissertation
0
T_s
$2T_s$
$3T_s$
$4T_s$
$\cdots$
$\hbar := kT_s$ (update horizon)
$\hbar \to 0$
Infinitesimal GPI, *i.e.*,
forward-in-time Riccati eq.
$0 < \hbar < \infty$
I-VI $(k = 1)$
I-GPI $(k = 1, 2, 3, \cdots)$
$\hbar \to \infty$
I-PI
$(k \to \infty$ or $T_s \to \infty$ (or both))

Figure 4.1: The classifications of IRL algorithms in terms of k and $\hbar$.

4.1, where infinitesimal GPI is at one extreme tip ($\hbar \to 0$), and I-PI is at the other extreme tip of the spectrum ($\hbar \to \infty$); I-VI ($k = 1$) and I-GPI (fininte k) are posed on the middle of the spectrum. From the classification with respect to $\hbar$ (or Corollary 4.2), one can see the followings.

Remark 4.4. *the I-GPI algorithms with the different $k \in \mathbb{N}$ but the same update horizon $\hbar = kT_s$ are all equivalent so have the same convergence speed in the iteration domain $i \in \mathbb{Z}_+$ if it converges; hence, the computational complexity due to the large iteration horizon k can be lessened by increasing the time horizon T_s for the same convergence speed in the iteration domain.*

Remark 4.5. *I-GPI becomes I-PI as k or T_s (or both) go to ∞, which was not fully investigated when I-GPIs were classified with respect to $k \in \mathbb{N}$ as shown in Fig. 4.1; Only in terms of k has the equivalence of I-GPI and I-PI shown in the literature [51]. On the other hand, unlike I-GPIs, I-PI always generates the same sequences $\{\mathbf{P}_{\mathbf{K}_i}\}_{i=0}^{\infty}$ and $\{\mathbf{K}_i\}_{i=0}^{\infty}$ regardless of the sample period T_s, so all I-PI algorithms with different T_s show the same limiting behaviors when $\hbar \to \infty$.*

To verify that all I-GPI algorithms with the same update horizon $\hbar$ yield the same sequence $\{\mathbf{P}_i\}_{i=0}^{\infty}$ (see Corollary 4.2 and Remark 4.1), we simulated I-GPI (Algorithm 4.3)

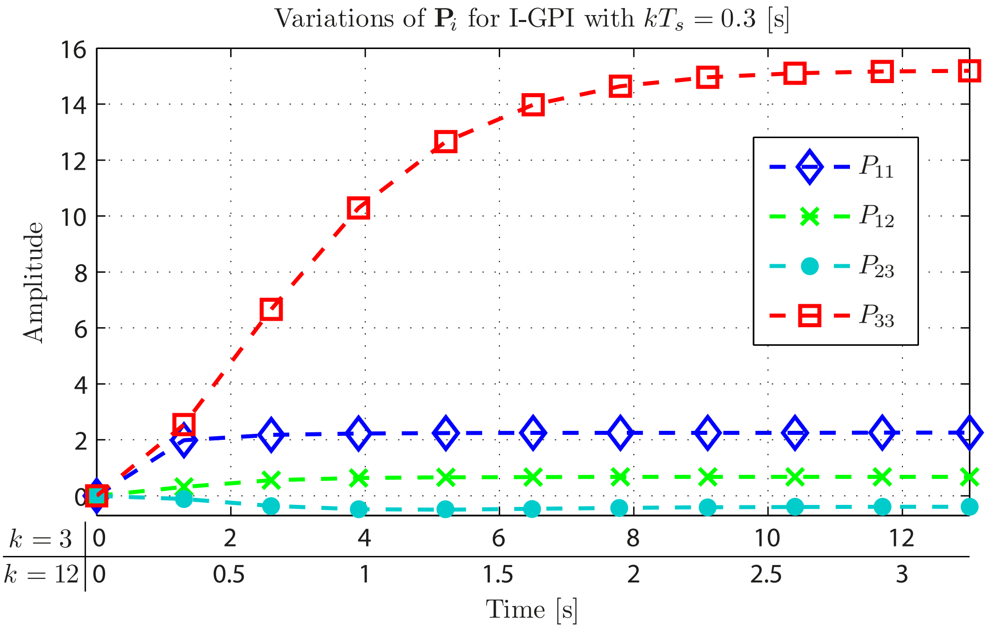


Figure 4.2: Variations of $\mathbf{P}_i$ for I-GPI with $\hbar = 0.3$ [s].

with the following LQR problem for the load frequency control system:

$$\mathbf{A} = \begin{bmatrix} -5 & 0 & -4 \\ 2 & -2 & 0 \\ 0 & 0.1 & -0.08 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ -0.1 \end{bmatrix}, \quad \mathbf{S} = \mathbf{diag}\{20, 10, 5\}, \quad \mathbf{R} = 0.15,$$

which is the same framework given by [112, Example 12.11], except that the governor speed regulation was set to 1.25 per unit. In the simulations, the policy evaluation of I-GPI is performed either by solving online least-squares problem at each iteration (see [93] for this), or by just conducting the equivalent matrix iteration (4.30). In either case, the policy evaluation yields the same value function matrix $\{\mathbf{P}_i\}_{i=0}^{\infty}$. In policy improvement, the next policy $\mathbf{K}_{i+1}$ is directely calculated by (4.11).

The simulation was performed with $(\mathbf{P}_0, \mathbf{K}_0) = (\mathbf{0}_{3\times3}, \mathbf{0}_{1\times3})$ for the same $\hbar = 0.3$ [s] and several different iteration horizons $k = 3, 6, 12, \cdots$, and the results are shown in Fig. 4.2, where the time axes were superposed and drawn only for the case of $k = 3, 12$. Note that all the sampling period T_s was set by the equation $kT_s = \hbar = 0.3$ [s], so the simulation results have different scales in the time domain. Also note that the policy evaluation of these simulations were performed by the LS method [93], where 12 data points were collected at each iteration to obtain the unique solution, and after every policy improvement step, the exploratory signal is injected for T_s seconds to excite the state variables. So, as can be seen from Fig. 4.2, the iteration was done every $13T_s$ [s]. The simulation results are consistent to the theory: one can see from Fig. 4.2 that all the I-GPI algorithms with the same $\hbar$ yield the same sequence $\{\mathbf{P}_i\}_{i=0}^{\infty}$, verifying in the iteration domain the equivalence of all the I-GPI methods that have the same $\hbar$.

4.5 Stability and Monotone Convergence Analysis

In the previous chapter, we see that the I-GPI (Algorithm 4.3) is the generalized one that includes the fundamental IRLs—I-PI, I-VI, and infinitesimal GPI—as special and limiting cases in the classification (see Fig. 4.1). Note that at the two extreme tips of the new classification, the sequence of matrices $\{\mathbf{P}_i\}_{i=0}^{\infty}$ monotonically converges to the optimal solution $\mathbf{P}^*$ in the following way:

- **(VI-Mode Convergence)** for infinitesimal GPI ($\hbar = 0$), $\mathbf{0}_{n\times n} \preceq \mathbf{P}_{t_1} \preceq \mathbf{P}_{t_2} \preceq \mathbf{P}^*$ for $0 \leq t_1 \leq t_2$ under $\mathbf{P}_0 = \mathbf{0}_{n\times n}$ (see also Section 4.3);
- **(PI-Mode Convergence)** for I-PI ($\hbar = \infty$), $\mathbf{0}_0 \preceq \mathbf{P}^* \preceq \mathbf{P}_{\mathbf{K}_{i+1}} \preceq \mathbf{P}_{\mathbf{K}_i}$ for all $i \in \mathbb{Z}_+$ under an initial Hurwitz policy.

Moreover, PI [113] (monotone decreasing) and value iteration (VI) [20, 25] (monotone increasing) for DT dynamical systems also have this monotone convergence property (see also [5]). In relation to the stability, this section shows these kinds of monotone convergence in I-GPI (Algorithm 4.3) and its policy evaluation (4.10) on the line of the new classification. As a first step, the two convergence modes above are precisely defined as follows.

Definition 4.3. *The sequence of matrices $\{\mathbf{P}_i \in \mathbb{R}^{n\times n}\}_{i=0}^{\infty}$ converges to $\mathbf{P}^* \in \mathbb{R}^{n\times n}$ in PI-mode (resp. in VI-mode) if it is monotonically decreasing (resp. increasing) in a sense that $\mathbf{P}^* \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i$ (resp. $\mathbf{P}_i \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}^*$) for all $i \in \mathbb{Z}_+$.*

For notational convenience in the analysis, we let $\mathbf{A}_i$ be the matrix of the i-th closed-loop system of I-GPI, i.e., $\mathbf{A}_i := \mathbf{A} - \mathbf{B}\mathbf{K}_i$ as in the previous chapter; we also define the $\mathbf{P}$-dependent control gain matrix $\mathbf{K}_\mathbf{P}$ for a given matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ as $\mathbf{K}_\mathbf{P} := \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}$. Here, the latter notation allows to write the matrix inequality $\mathcal{R}(\mathbf{P}) = \mathcal{L}(\mathbf{K}, \mathbf{P})|_{\mathbf{K}=\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}}$ in a simple form $\mathcal{R}(\mathbf{P}) = \mathcal{L}(\mathbf{K}_\mathbf{P}, \mathbf{P})$. The following lemma stated with $\mathbf{K}_\mathbf{P}$-notation will be extensively used in the analysis.

Lemma 4.2. *For any $\mathbf{P}$, $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$ and $\mathbf{K} \in \mathbb{M}^{m\times n}$, the Lyapunov and Riccati operators $\mathcal{L}(\cdot, \cdot)$ and $\mathcal{R}(\cdot)$ satisfy the followings:*

- $\mathcal{L}(\mathbf{K}, \mathbf{P}) - \mathcal{L}(\mathbf{K}, \mathbf{\Phi}) = \mathbf{A}_\mathbf{K}^T(\mathbf{P} - \mathbf{\Phi}) + (\mathbf{P} - \mathbf{\Phi})\mathbf{A}_\mathbf{K}, \qquad (4.19)$
- $\mathcal{L}(\mathbf{K}, \mathbf{P}) = \mathcal{R}(\mathbf{P}) + (\mathbf{K} - \mathbf{K}_\mathbf{P})^T\mathbf{R}(\mathbf{K} - \mathbf{K}_\mathbf{P}). \qquad (4.20)$

Proof. (4.19) can be easily verified by the definition of $\mathcal{L}(\mathbf{K}, \mathbf{P})$ (see (4.4)):

$$\mathcal{L}(\mathbf{K}, \mathbf{P}) = \mathbf{A}_\mathbf{K}^T\mathbf{P} + \mathbf{P}\mathbf{A}_\mathbf{K} + \mathbf{S}_\mathbf{K}.$$

For the proof of (4.20), note that $\mathcal{R}(\mathbf{P})$ can be represented in terms of $\mathbf{K}_\mathbf{P}$ as

$$\mathcal{R}(\mathbf{P}) = \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{K}_\mathbf{P}{}^T\mathbf{R}\mathbf{K}_\mathbf{P} + \mathbf{S},$$

and that $(\mathbf{K} - \mathbf{K}_\mathbf{P})^T\mathbf{R}(\mathbf{K} - \mathbf{K}_\mathbf{P}) = \mathbf{K}^T\mathbf{R}\mathbf{K} - \mathbf{K}_\mathbf{P}^T\mathbf{R}\mathbf{K} - \mathbf{K}^T\mathbf{R}\mathbf{K}_\mathbf{P} + \mathbf{K}_\mathbf{P}^T\mathbf{R}\mathbf{K}_\mathbf{P}$. Then, the proof is completed by substituting these into (4.20) and rearranging the equation. □

In the convergence analysis of I-GPI, the next lemma will be used to finalize every proof of monotone convergence of I-GPI. The lemma states that the convergent point corresponds to the optimal solution $(\mathbf{K}^*, \mathbf{P}^*)$.

Lemma 4.3. *Consider the sequences $\{\mathbf{P}_i\}_{i=0}^\infty$ and $\{\mathbf{K}_i\}_{i=0}^\infty$ generated by the I-GPI for LQR problems (Algorithm 4.3), and let $\{\mathbf{P}_i\}_{i=0}^\infty$ be a convergent sequence. Then, $\mathbf{P}_i$ and $\mathbf{K}_i$ converge to the optimal solutions $\mathbf{P}^*$ and $\mathbf{K}^*$, respectively.*

Proof. See Appendix. □

Together with Lemma 4.3, the next lemma is essentially needed for convergence analysis of I-GPI. The lemma relates $(\mathbf{K}_i, \mathbf{P}_i)$ with the optimal solution $(\mathbf{K}^*, \mathbf{P}^*)$.

Lemma 4.4. *The matrix sequence* $\{\mathbf{P}_i\}_{i=0}^{\infty}$ *generated by I-GPI (Algorithm 4.3) satisfies*

$$\mathbf{P}^* = \mathbf{P}_i + \int_0^{\infty} e^{\mathbf{A}_{\mathbf{K}^*}^T \tau} \left[\mathcal{R}(\mathbf{P}_i) + (\mathbf{K}^* - \mathbf{K}_i)^T \mathbf{R} \, (\mathbf{K}^* - \mathbf{K}_i)\right] e^{\mathbf{A}_{\mathbf{K}^*} \tau} \, d\tau. \tag{4.21}$$

Proof. The substitutions of $\mathbf{P} = \mathbf{P}_i$, $\mathbf{\Phi} = \mathbf{P}^*$, and $\mathbf{K} = \mathbf{K}^*$ into (4.19) and (4.20) in Lemma 4.2 yield

$$\mathcal{L}(\mathbf{K}^*, \mathbf{P}_i) = \mathbf{A}_{\mathbf{K}^*}^T (\mathbf{P}_i - \mathbf{P}^*) + (\mathbf{P}_i - \mathbf{P}^*) \mathbf{A}_{\mathbf{K}^*}, \tag{4.22}$$

$$\mathcal{L}(\mathbf{K}^*, \mathbf{P}_i) = \mathcal{R}(\mathbf{P}_i) + (\mathbf{K}^* - \mathbf{K}_{\mathbf{P}_i})^T \mathbf{R} (\mathbf{K}^* - \mathbf{K}_{\mathbf{P}_i}), \tag{4.23}$$

where $\mathcal{L}(\mathbf{K}^*, \mathbf{P}^*) = \mathcal{R}(\mathbf{P}^*) = 0$ is used, and the existence of the unique solution $(\mathbf{K}^*, \mathbf{P}^*)$ is guaranteed by the stabilizability and detectability of the triple $(\mathbf{S}, \mathbf{A}, \mathbf{B})$ [63]. Then, substituting (4.23) into (4.22) yields the following generalized Lyapunov equation:

$$\mathbf{A}_{\mathbf{K}^*}^T (\mathbf{P}_i - \mathbf{P}^*) + (\mathbf{P}_i - \mathbf{P}^*) \mathbf{A}_{\mathbf{K}^*} = \mathcal{R}(\mathbf{P}_i) + (\mathbf{K}^* - \mathbf{K}_{\mathbf{P}_i})^T \mathbf{R} (\mathbf{K}^* - \mathbf{K}_{\mathbf{P}_i}),$$

where the optimal policy $\mathbf{K}^*$ is Hurwitz. Therefore, the application of Lemma 2.2 to this generalized Lyapunov equation completes the proof. □

4.5.1 Monotone Convergence of Policy Evaluation Iterations

Corollary 4.2 states that $\mathbf{P}_{i|k}$ obtained by the policy evaluation iteration (4.10) at i-th step of I-GPI converges to $\mathbf{P}^*$ if $\mathbf{K}_i$ is Hurwitz. In this subsection, it will be shown that the convergence is actually monotone under certain conditions; its proof is based on the following lemma which shows the two equivalent matrix iterative formulas for the matrices $\mathbf{P}_{i|j}$ generated by the i-th approximate policy evaluation of I-GPI.

Lemma 4.5. *Any matrices* $\mathbf{P}_{i|l}$ *and* $\mathbf{P}_{i|l+\kappa}$ $(0 \le l \le l + \kappa < \infty)$ *generated by i-th approximate policy evaluation of I-GPI (Algorithm 4.3) satisfy*

- $$\mathbf{P}_{i|l+\kappa} - \mathbf{P}_{i|l} = \int_0^{\Delta h} e^{\mathbf{A}_i^T \tau} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l}) e^{\mathbf{A}_i \tau} \, d\tau, \tag{4.24}$$
- $$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l+\kappa}) = e^{\mathbf{A}_i^T \Delta h} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l}) e^{\mathbf{A}_i \Delta h}, \tag{4.25}$$

where $\Delta h := \kappa T_s$ *and* $\mathbf{K}_i$ *is a given policy at i-th iteration, not necessarily Hurwitz.*

Proof. First, note that $V_{i|l+\kappa}(\mathbf{x}_t) = \mathcal{T}_{\mathbf{K}_i}^{\Delta h} V_{i|l}(\mathbf{x}_t)$ holds by (4.10) and Theorem 4.1. Then, the substitution of the expansion

$$\begin{aligned}\mathcal{T}_{\mathbf{K}_i}^{\Delta h} V_{i|l}(\mathbf{x}_t) &= \int_0^{\Delta h} \mathbf{x}_{t+\tau}^T \mathbf{S}_{\mathbf{K}_i} \mathbf{x}_{t+\tau}\, d\tau + \mathbf{x}_{t+\Delta h}^T \mathbf{P}_{i|l}\, \mathbf{x}_{t+\Delta h} \\ &= \mathbf{x}_t^T \left[\int_0^{\Delta h} e^{\mathbf{A}_i^T \tau} \mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \tau}\, d\tau + e^{\mathbf{A}_i^T \Delta h} \mathbf{P}_{i|l}\, e^{\mathbf{A}_i \Delta h} \right] \mathbf{x}_t,\end{aligned}$$

and $V_{i|l+\kappa}(\mathbf{x}_t) = \mathbf{x}_t^T \mathbf{P}_{i|l+\kappa}\, \mathbf{x}_t$ into $V_{i|l+\kappa}(\mathbf{x}_t) = \mathcal{T}_{\mathbf{K}_i}^{\Delta h} V_{i|l}(\mathbf{x}_t)$ yields

$$\mathbf{P}_{i|l+\kappa} = e^{\mathbf{A}_i^T \Delta h} \mathbf{P}_{i|l} e^{\mathbf{A}_i \Delta h} + \int_0^{\Delta h} e^{\mathbf{A}_i^T \tau} \mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \tau}\, d\tau.$$

Then, adding and subtracting $\mathbf{P}_{i|l}$ on the right hand side and applying Lemma 2.1 proves (4.24). For the proof of (4.25), note that (4.19) in Lemma 4.2 implies

$$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l+\kappa}) = \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l}) + \mathbf{A}_i^T (\mathbf{P}_{i|l+\kappa} - \mathbf{P}_{i|l}) + (\mathbf{P}_{i|l+\kappa} - \mathbf{P}_{i|l}) \mathbf{A}_i,$$

where the second term of the right hand side satisfies

$$\mathbf{A}_i^T (\mathbf{P}_{i|l+\kappa} - \mathbf{P}_{i|l}) + (\mathbf{P}_{i|l+\kappa} - \mathbf{P}_{i|l}) \mathbf{A}_i = e^{\mathbf{A}_i^T \Delta h} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l}) e^{\mathbf{A}_i \Delta h} - \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|l}),$$

which is obtained by the substitution of (4.24) and the application of Lemma 2.1. Hence, one obtains (4.25), and the proof is completed. □

Based on Lemma 4.5, the monotone convergence theorem is provided as follows.

Theorem 4.2. *Consider the finite matrix sequence* $\{\mathbf{P}_{i|j}\}_{j=0}^k$ *generated by the i-th approximate policy evaluation of Algorithm 4.3. Suppose that* $\mathbf{K}_i$ *is Hurwitz so that* $\mathbf{P}_{i|k} \to \mathbf{P}_{\mathbf{K}_i}$ *as* $\hbar \to \infty$. *Then, for any* $\hbar \in \mathbb{R}_+$ *and* $j \in \{1, 2, \cdots, k\}$,

- $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|0}) \preceq \mathbf{0}_{n\times n}$ *implies* $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|j}) \preceq \mathbf{0}_{n\times n}$ *and*

$$\mathbf{P}_{\mathbf{K}_i} \preceq \mathbf{P}_{i|k} \preceq \cdots \preceq \mathbf{P}_{i|j} \preceq \cdots \preceq \mathbf{P}_{i|0}; \tag{4.26}$$

- $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|0}) \succeq \mathbf{0}_{n\times n}$ *implies* $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|j}) \succeq \mathbf{0}_{n\times n}$ *and*

$$\mathbf{P}_{i|0} \preceq \cdots \preceq \mathbf{P}_{i|j} \preceq \cdots \preceq \mathbf{P}_{i|k} \preceq \mathbf{P}_{\mathbf{K}_i}. \tag{4.27}$$

Proof. For the proof of the first part, suppose $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|0}) \preceq \mathbf{0}_{n\times n}$. Then, it satisfies

$$e^{\mathbf{A}_i^T t} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|0}) e^{\mathbf{A}_i t} \preceq \mathbf{0}_{n\times n} \;\; \forall t \geq 0.$$

Hence, $\mathbf{P}_{i|1} - \mathbf{P}_{i|0} \succeq \mathbf{0}_{n\times n}$ and $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|1}) \preceq \mathbf{0}_{n\times n}$ holds by (4.24) and (4.25) in Lemma 4.5 ($l = 0$, $\kappa = 1$), respectively. From $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|1}) \preceq \mathbf{0}_{n\times n}$, one also obtains $\mathbf{P}_{i|2} - \mathbf{P}_{i|1} \succeq \mathbf{0}_{n\times n}$ and $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|2}) \preceq \mathbf{0}_{n\times n}$ in the same manner, but with $l = 1$ and $\kappa = 1$ in Lemma 4.5. Continuing this procedure up to $l = k - 1$, all with $\kappa = 1$, yields (4.26), where the matrix inequality $\mathbf{P}_{\mathbf{K}_i} \preceq \mathbf{P}_{i|k}$ is obtained from $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|k}) \preceq \mathbf{0}_{n\times n}$ and (4.24) in Lemma 4.5 ($l = k$ and $\kappa \to \infty$); in this limit, $\mathbf{P}_{l+\kappa}$ converges to $\mathbf{P}_{\mathbf{K}_i}$ since $\mathbf{K}_i$ is assumed Hurwitz. This completes the proof of (4.26); the monotonicity (4.27) can be also proven by assuming $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i|0}) \succeq \mathbf{0}_{n\times n}$ and following the same procedure. □

4.5.2 Closed-Loop Stability and PI-Mode Convergence

Based on the results in Section 4.5.1, this section provides the closed-loop stability and PI-mode convergence results of I-GPI. For notational convenience, define the increments $\Delta\mathbf{P}_i$ and $\Delta\mathbf{K}_i$ as $\Delta\mathbf{P}_i := \mathbf{P}_{i+1} - \mathbf{P}_i$ and $\Delta\mathbf{K}_i := \mathbf{K}_{i+1} - \mathbf{K}_i$, respectively. Also, let $\mathbf{M}_{(i,\hbar)}$ be defined by

$$\mathbf{M}_{(i,\,\hbar)} := \int_0^{\hbar} e^{\mathbf{A}_i^T \tau} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) e^{\mathbf{A}_i \tau}\, d\tau. \tag{4.28}$$

Then, from (4.24) and (4.25) in Lemma 4.5 with $l = 0$ and $\kappa = k$, we obtain the following two equivalent matrix formulas

$$\Delta\mathbf{P}_i = \mathbf{M}_{(i,\hbar)}, \tag{4.29}$$

$$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) = e^{\mathbf{A}_i^T \hbar} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) e^{\mathbf{A}_i \hbar}, \tag{4.30}$$

where $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ for $i \geq 1$ in (4.30) satisfies $\mathcal{R}(\mathbf{P}_i) = \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ due to the policy improvement "$\mathbf{K}_i = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_i$" of I-GPI. In addition, (4.20) in Lemma 4.2 implies that the operators $\mathcal{R}(\mathbf{P}_{i+1})$ and $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ satisfy

$$\mathcal{R}(\mathbf{P}_{i+1}) = \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) - \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i. \tag{4.31}$$

This explains how the policy improvement step $\mathbf{K}_{i+1} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{i+1}$ influences the Riccati error $\mathcal{R}(\mathbf{P}_{i+1})$ through $\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i$, wherein $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ results from the approximate policy evaluation of I-GPI and satisfies (4.30).

In what follows, the general matrix inequality condition is given to guarantee the closed-loop stability of the next policy when the current policy is Hurwitz. This stability criterion will be a necessary condition for PI-mode convergence of I-GPI.

Theorem 4.3. *Assume $\mathbf{K}_i$ is Hurwitz and $\mathbf{P}_{i+1} \in \mathbb{R}^{n \times n}$ is positive semi-definite. If $\mathbf{K}_{i+1}$ updated by $\mathbf{K}_{i+1} = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_{i+1}$ satisfies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}}$, then $\mathbf{K}_{i+1}$ is also Hurwitz.*

Proof. See Appendix. □

From Theorem 4.3 and mathematical induction, we obtain the following corollary regarding stability propagation of I-GPI.

Corollary 4.3. *Suppose that $\mathbf{K}_0$ is Hurwitz and $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$. If the pair $(\mathbf{K}_{i+1}, \mathbf{P}_{i+1})$ and $\mathbf{K}_i$ satisfy*

$$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}} \;\; \forall i \in \mathbb{Z}_+, \tag{4.32}$$

then the policies $\{\mathbf{K}_i\}_{i=1}^{\infty}$ generated by I-GPI (Algorithm 4.3) are all Hurwitz.

Proof. Assume $\mathbf{P}_i \succeq \mathbf{0}_{n\times n}$, and $\mathbf{K}_i$ is Hurwitz and satisfies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}}$. Then, Lemma 4.1 implies $\mathbf{P}_{i+1} \succeq \mathbf{0}_{n\times n}$, and the application of Theorem 4.3 results in the Hurwitz policy $\mathbf{K}_{i+1}$. Hence, mathematical induction concludes that $\{\mathbf{K}_i\}_{i=1}^{\infty}$ are all Hurwitz. □

The condition $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \leq \mathbf{S}_{\mathbf{K}_{i+1}}$ in Corollary 4.3 provides stability during and after the learning phase, but the direct evaluation of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ requires the knowledge of the system matrix $\mathbf{A}$ at each iteration, while I-GPI does not. The next corollary provides another inequality condition for the closed-loop stability, which does not explicitly depend on $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$, so prevent I-GPI to require the knowledge of $\mathbf{A}$.

Corollary 4.4. *Suppose that the initial policy $\mathbf{K}_0$ is Hurwitz and $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$ satisfies $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{S}_{\mathbf{K}_0}$. Then, the policies $\{\mathbf{K}_i\}_{i=1}^{\infty}$ generated by I-GPI are all Hurwitz if $\mathbf{K}_i$ and $\mathbf{K}_{i+1}$ satisfy*

$$e^{\mathbf{A}_i^T \hbar}\mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \hbar} \preceq \mathbf{S}_{\mathbf{K}_{i+1}}, \quad \forall i \in \mathbb{Z}_+. \tag{4.33}$$

Proof. Assume that $\mathbf{K}_i$ is Hurwitz and $\mathbf{P}_i \succeq \mathbf{0}_{n\times n}$ satisfies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \preceq \mathbf{S}_{\mathbf{K}_i}$. Then, (4.30) and (4.33) implies

$$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) = e^{\mathbf{A}_i^T \hbar} \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) e^{\mathbf{A}_i \hbar} \preceq e^{\mathbf{A}_i^T \hbar} \mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \hbar} \preceq \mathbf{S}_{\mathbf{K}_{i+1}}. \tag{4.34}$$

So, since $\mathbf{P}_{i+1} \succeq \mathbf{0}_{n\times n}$ holds by Lemma 4.1, $\mathbf{K}_{i+1}$ is also Hurwitz by Theorem 4.3. Substituting the inequality (4.34) into (4.31) and rearranging it yields

$$\mathcal{L}(\mathbf{K}_{i+1}, \mathbf{P}_{i+1}) = \mathcal{R}(\mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}} - \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i \preceq \mathbf{S}_{\mathbf{K}_{i+1}}.$$

Therefore, mathematical induction with $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \leq \mathbf{S}_{\mathbf{K}_0}$ proves that the policies $\{\mathbf{K}_i\}_{i=1}^{\infty}$ are all Hurwitz, which completes the proof. □

Although condition (4.33) depends on the system matrix $\mathbf{A}$, it is contained only in the form of exponentials. By virtue of this fact, (4.33) can be easily checked without knowing the system matrix $\mathbf{A}$ (see [93]).

Now, the following theorem states that, under certain conditions satisfying (4.32), the policies $\{\mathbf{K}_i\}_{i=0}^{\infty}$ generated by I-GPI are all Hurwitz, and $\mathbf{P}_i \to \mathbf{P}^*$ in PI-mode.

Theorem 4.4. *Suppose that an initial Hurwitz policy $\mathbf{K}_0$ and $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$ satisfy*

$$\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_n.$$

Then, the following hold for all $i \in \mathbb{Z}_+$.

- ***(Stability)*** *The policy $\mathbf{K}_i$ is Hurwitz and satisfies the Lyapunov inequalities*

  $$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{0}_{n\times n} \text{ and } \mathcal{L}(\mathbf{K}_{i+1}, \mathbf{P}_{i+1}) \preceq -\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i.$$

  That is,

  $$\begin{cases} \mathbf{A}_i^T \mathbf{P}_{i+1} + \mathbf{P}_{i+1} \mathbf{A}_i \preceq -\mathbf{S}_{\mathbf{K}_i} \\ \mathbf{A}_{i+1}^T \mathbf{P}_{i+1} + \mathbf{P}_{i+1} \mathbf{A}_{i+1} \preceq -\mathbf{S}_{\mathbf{K}_{i+1}} - \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i. \end{cases} \tag{4.35}$$

- ***(PI-mode convergence)*** *The sequence $\{(\mathbf{P}_i, \mathbf{K}_i\}_{i=0}^{\infty}$ generated by I-GPI converge* *to the optimal solution* $(\mathbf{P}^*, \mathbf{K}^*)$ *with the following monotonicities*:

$$\begin{cases} \mathbf{0}_{n\times n} \preceq \mathbf{P}_{\mathbf{K}_i} \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i \\ \mathbf{0}_{n\times n} \preceq \mathbf{P}^* \preceq \cdots \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i \preceq \cdots \preceq \mathbf{P}_0. \end{cases} \tag{4.36}$$

- *(2^{nd}**-order monotone decreasing**) There exists* $c > 0$ *such that if* $(\mathbf{K}_{i+1}, \mathbf{P}_{i+1})$ *and* $\mathbf{K}_i$ *for some* $i \in \mathbb{Z}_+$ *satisfy*

$$-(\alpha_i - 1)\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i \preceq \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{0}_{n\times n}, \tag{4.37}$$

where $\alpha_i \geq 1$ *is a constant, then for all such* i, $\|\mathbf{P}_{i+1} - \mathbf{P}^*\| \leq c \cdot \alpha_i \|\mathbf{P}_i - \mathbf{P}^*\|^2$.

Proof. First, $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$ and Lemma 4.1 imply that $\mathbf{P}_i \succeq \mathbf{0}_{n\times n}$ for all $i \in \mathbb{Z}_+$ by mathematical induction. Next, assume $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \preceq 0$ for some $i \in \mathbb{Z}_+$. Then, we have $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq 0$ by (4.30) and substituting this into (4.31) yields

$$\mathcal{R}(\mathbf{P}_{i+1}) = \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) - \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i \preceq -\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i.$$

Hence, mathematical induction with $(\mathbf{K}_0, \mathbf{P}_0)$ satisfying $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$ implies that the Lyapunov inequalities $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{0}_{n\times n}$ and

$$\mathcal{L}(\mathbf{K}_{i+1}, \mathbf{P}_{i+1}) = \mathcal{R}(\mathbf{P}_{i+1}) \preceq -\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i \preceq \mathbf{0}_{n\times n}$$

hold for all $i \in \mathbb{Z}_+$.
(Proof of stability). $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{0}_{n\times n}$ implies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}}$, and by Corollary 4.3, one can conclude that for all $i \in \mathbb{Z}_+$, $\mathbf{K}_i$ is Hurwitz.
(Proof of monotone convergence). $\mathbf{P}_{i+1} = \mathbf{P}_{i|k}$ and (4.26) in Theorem 4.2 imply that $\mathbf{P}_{i+1}$ satisfies $\mathbf{0}_{n\times n} \preceq \mathbf{P}_{\mathbf{K}_i} \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i$, which holds for all $i \in \mathbb{Z}_+$ since $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \preceq \mathbf{0}_{n\times n}$ *for all* $i \in \mathbb{Z}_+$. Therefore, since it is monotonically decreasing and bounded by $\mathbf{0}_{n\times n}$, the sequence $\{\mathbf{P}_i\}_{i=0}^{\infty}$ monotonically converges. Let $\bar{\mathbf{P}}$ be the limit point of $\mathbf{P}_i$, i.e., $\bar{\mathbf{P}} := \lim_{i\to\infty} \mathbf{P}_i$. Then, $\bar{\mathbf{P}}$ satisfies $\mathbf{0}_{n\times n} \preceq \bar{\mathbf{P}} \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i$ for all $i \in \mathbb{Z}_+$, and Lemma 4.3 implies $\bar{\mathbf{P}} = \mathbf{P}^*$ and $\lim_{i\to\infty} \mathbf{K}_i = \mathbf{K}^*$, respectively. This proves the convergence $\mathbf{P}_i \to \mathbf{P}^*$ in PI-mode (with the monotonicity (4.36)).
(Proof of 2^{nd}-order convergence). First, note that one has $\mathbf{0}_{n\times n} \preceq \mathbf{P}_{i+1} - \mathbf{P}^*$ by (4.36), and (4.31) and (4.37) imply

$$-\alpha_i \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i \preceq \mathcal{R}(\mathbf{P}_{i+1}) \preceq -\Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i.$$

From this inequality and (4.21), one obtains

$$\begin{aligned}\mathbf{0}_{n\times n} \preceq \mathbf{P}_{i+1} - \mathbf{P}^* &\preceq -\int_0^\infty e^{\mathbf{A}_{\mathbf{K}^*}^T\tau}\mathcal{R}(\mathbf{P}_{i+1})e^{\mathbf{A}_{K^*}\tau}\, d\tau \\ &\preceq \alpha_i \int_0^\infty e^{\mathbf{A}_{\mathbf{K}^*}^T\tau}\Delta\mathbf{K}_i^T\mathbf{R}\Delta\mathbf{K}_i\, e^{\mathbf{A}_{\mathbf{K}^*}\tau}\, d\tau. \end{aligned} \tag{4.38}$$

By virtue of the fact that for $\mathbf{X}$, $\mathbf{Y} \in \mathbb{R}^{n\times n}$, "$\mathbf{0}_{n\times n} \preceq \mathbf{X} \preceq \mathbf{Y} \implies \|\mathbf{X}\| \le \|\mathbf{Y}\|$", one can take the matrix norm $\|\cdot\|$ on (4.38) and obtain the following inequality using the properties of the norm:

$$\begin{aligned}\|\mathbf{P}_{i+1} - \mathbf{P}^*\| &\le \alpha_i \int_0^\infty \left\| e^{\mathbf{A}_{\mathbf{K}^*}^T\tau}\Delta\mathbf{K}_i^T\mathbf{R}\Delta\mathbf{K}_i e^{\mathbf{A}_{\mathbf{K}^*}\tau} \right\| d\tau \\ &\le \alpha_i \underbrace{\left(\int_0^\infty \|e^{\mathbf{A}_{\mathbf{K}^*}\tau}\|^2\, d\tau\right)\|\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\|}_{=:c} \cdot \|\mathbf{P}_i - \mathbf{P}_{i+1}\|^2\end{aligned}$$

Now, the proof of the 2$^{\text{nd}}$-order monotone decreasing $\|\mathbf{P}_{i+1} - \mathbf{P}^*\| \le c \cdot \alpha_i\, \|\mathbf{P}_i - \mathbf{P}^*\|^2$ can be done by using the fact that by (4.36), $\mathbf{0}_{n\times n} \preceq \mathbf{P}_i - \mathbf{P}_{i+1} \preceq \mathbf{P}_i - \mathbf{P}^*$ holds for all $i \in \mathbb{Z}_+$, which again implies $\|\mathbf{P}_i - \mathbf{P}_{i+1}\| \le \|\mathbf{P}_i - \mathbf{P}_{K^*}\|$. □

The properties of I-GPI in PI-mode convergence shown in Theorem 4.4 are similar to those of I-PI which is equivalent to Kleinman's Newton method [110] (see (4.6)). Actually, the I-GPI algorithm can be considered an inexact Kleinman's Newton algorithm [114] with the residual "$\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$", which can be made arbitrarily small by increasing $\hbar$ in the approximate policy evaluation. Actually, the I-GPI can be represented by the quasi-Newton form:

$$\mathbf{P}_{i+1} = \mathbf{P}_i + \left(\mathcal{L}'_{\mathbf{K}_i,\mathbf{P}_i}\right)^{-1}\Big[\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) - \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})\Big],$$

which converges to the Newton method (4.6) as the residual $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ goes to zero ($\hbar \to \infty$). In this limit case, I-GPI becomes I-PI as mentioned in Section 4.5.1, and the Lyapunov inequalities in (4.35) of Theorem 4.4 become their respective Lyapunov equations of the form

$$\begin{cases} \mathbf{A}_i^T \mathbf{P}_{i+1} + \mathbf{P}_{i+1}\mathbf{A}_i = -\mathbf{S}_{\mathbf{K}_i} \\ \mathbf{A}_{i+1}^T \mathbf{P}_{i+1} + \mathbf{P}_{i+1}\mathbf{A}_{i+1} = -\mathbf{S}_{\mathbf{K}_{i+1}} - \Delta\mathbf{K}_i^T \mathbf{R} \Delta\mathbf{K}_i, \end{cases}$$

which provide the closed-loop stability at each i-th iteration and implies that

1) the residual $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ become zero;

2) $\mathcal{R}(\mathbf{P}_{i+1})$ satisfies $\mathcal{R}(\mathbf{P}_{i+1}) = -\Delta\mathbf{K}_i \mathbf{R} \Delta\mathbf{K}_i$ for all $i \in \mathbb{Z}_+$.

Here, the former guarantees the condition (4.37) with $\alpha_i = 1$, which implies the uniform 2^{nd}-order PI-mode convergence, and the latter implies $\mathcal{R}(\mathbf{P}_i) \preceq \mathbf{0}_{n\times n}$, which provides an alternative approach to the proof of monotone convergence of I-PI, as shown in this dissertation.

Regarding the closed-loop stability of I-GPI, under an initial Hurwitz policy $\mathbf{K}_0$, three matrix inequality conditions have been presented in this subsection and are summarized in Fig. 4.3 and in the following:

1) $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}}$ *in Corollary 4.3*: this condition is the most general stability condition among the three, and hence can be considered the sufficient condition of the other two;

2) $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{S}_{\mathbf{K}_0}$ *and* (4.33) *in Corollary 4.4*: this condition is rather restricted, but can be checked in online learning without using the knowledge of the matrix $\mathbf{A}$ (see [93]).

3) $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$ *for PI-mode convergence*: as shown in *Theorem 4.4*, this initial matrix condition ensures $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{0}_{n\times n}$ for all $i \in \mathbb{Z}_+$, which again implies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}}$ $\forall i \in \mathbb{Z}_+$. Therefore, the closed-loop stability in the entire iteration is automatically guaranteed so one does not need to check any matrix inequalities for stability except the initial one. For the first two conditions, the agent should check the inequality at every iteration step to maintain the stability.

4.5.3 Monotone Increasing and VI-Mode Convergence

This section discusses the monotone increasing and VI-mode convergence properties of I-GPI. These properties in VI-mode are the counterpart of PI-mode convergence and do

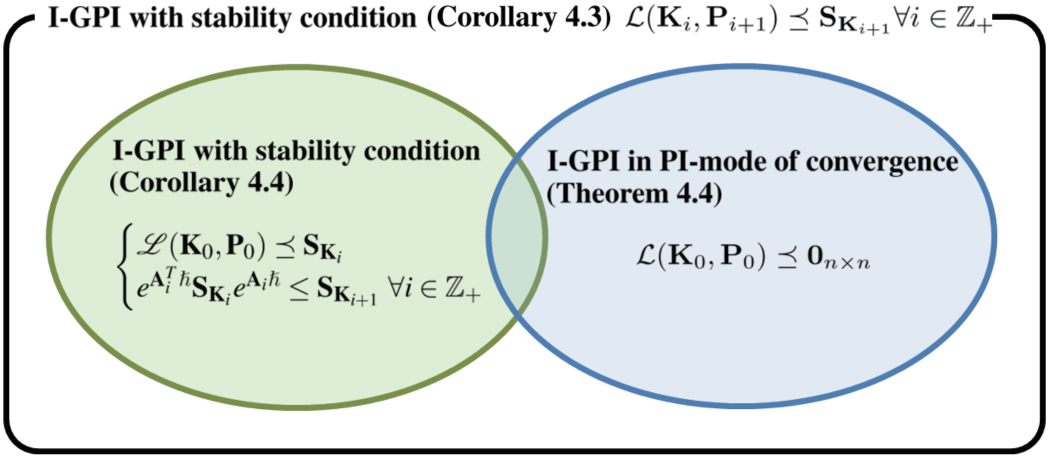

I-GPI with stability condition (Corollary 4.3) $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_{i+1}) \preceq \mathbf{S}_{\mathbf{K}_{i+1}} \forall i \in \mathbb{Z}_+$
I-GPI with stability condition (Corollary 4.4)
$\begin{cases} \mathscr{L}(\mathbf{K}_0,\mathbf{P}_0) \preceq \mathbf{S}_{\mathbf{K}_i} \\ e^{\mathbf{A}_i^T \hbar}\mathbf{S}_{\mathbf{K}_i} e^{\mathbf{A}_i \hbar} \leq \mathbf{S}_{\mathbf{K}_{i+1}} \ \forall i \in \mathbb{Z}_+ \end{cases}$
I-GPI in PI-mode of convergence (Theorem 4.4)
$\mathcal{L}(\mathbf{K}_0,\mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$

Figure 4.3: Summary of stability conditions and PI-mode convergence of I-GPI.

not require the initial policy $\mathbf{K}_0$ Hurwitz like the infinitesimal GPI and the VI in DT domain. In the discussions, the following analytical result plays a central role.

Theorem 4.5. *Consider the sequence of* $(\mathbf{P}_i,\mathbf{K}_i)$ *generated by I-GPI under*

$$\mathbf{0}_{n\times n} \preceq \mathbf{P_0} \preceq \mathbf{P}^*.$$

If $\mathbf{K}_i$ *and* $\mathbf{P}_i$ *satisfy* $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$, $\forall i \in \{0,1,2,\cdots,l-1\}$, *then the finite sequence* $\{\mathbf{P}_i\}_{i=0}^{l}$ *possesses the monotone increasing property:*

$$\mathbf{0}_{n\times n} \preceq \mathbf{P}_0 \preceq \cdots \preceq \mathbf{P}_i \preceq \mathbf{P}_{i+1} \preceq \cdots \preceq \mathbf{P}_l \preceq \mathbf{P}^*. \tag{4.39}$$

Moreover, if $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \preceq \mathbf{0}_{n\times n}$ $\forall i \in \mathbb{Z}_+$, *then* $(\mathbf{P}_i,\mathbf{K}_i)$ *converge to* $(\mathbf{P}^*,\mathbf{K}^*)$ *with the monotonicity* (4.39) *for all* $i \in \mathbb{Z}_+$.

Proof. For each $i \in \{0,1,2,\cdots,l-1\}$, $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \geq \mathbf{0}_{n\times n}$ and (4.27) in Theorem 4.2 implies $\mathbf{0}_{n\times n} \preceq \mathbf{P}_i \preceq \mathbf{P}_{i+1}$ (positive semi-definiteness comes from Lemma 4.1). Next, we obtain $\mathbf{P}^* \succeq \mathbf{P}_i$ $\forall i \in \{1,2,\cdots,l-1\}$ from (4.21) since $\mathcal{R}(\mathbf{P}_i) = \mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \succeq 0$ by assumption and $(\Delta\mathbf{K}_i^*)^T\mathbf{R}\,\Delta\mathbf{K}_i^* \succeq \mathbf{0}_{n\times n}$ ($\mathbf{P}^* \succeq \mathbf{P}_0$ is assumed for $i=0$). Rearranging all these inequalities yields $\mathbf{0}_{n\times n} \preceq \mathbf{P}_i \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}^*$, which holds for all $i \in \{0,1,2,\cdots,l-1\}$ by the assumption $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ for all such i. Therefore, we have (4.39), and the monotone convergence to the optimal solution can be directly proven by the assumption of $\mathcal{L}(\mathbf{K}_i,\mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ for all $i \in \mathbb{Z}_+$ and Lemma 4.3. □

This theorem with $\mathbf{0}_{n\times n} \preceq \mathbf{P}_0 \preceq \mathbf{P}^*$ and $\mathcal{L}(\mathbf{K}_0,\mathbf{P}_0) \succeq \mathbf{0}_{n\times n}$ obviously guarantees the monotone increasing (4.39) up to some finite $l \in \mathbb{N}$ (the trivial case is $l = 1$). For I-GPI methods with $\mathbf{P}_0 = \mathbf{0}_{n\times n}$, this monotone increasing is also valid for any given initial policy $\mathbf{K}_0$ not necessarily Hurwitz since $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) = \mathbf{K}_0^T\mathbf{R}\mathbf{K}_0 + \mathbf{S} \succeq \mathbf{0}_{n\times n}$ holds. On the other hand, for VI-mode convergence, I-GPI should generate the sequences $\{\mathbf{P}_i\}$ and $\{\mathbf{K}_i\}$ $(= \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_i)$, both of which satisfy $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ for all $i \in \mathbb{N}$. However, this is not attainable in general since, even in the case where $\mathbf{K}_i$ is Hurwitz that satisfies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succeq 0$ under $0 \preceq \mathbf{P}_i \preceq \mathbf{P}^*$, $\mathcal{R}(\mathbf{P}_{i+1})$ *can be indefinite or negative semi-definite for large* $\hbar$ *by* (4.31). This is because the residual $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ in (4.31) becomes zero as $\hbar \to \infty$ (see (4.30) and also Theorem 4.2). More obviously and intuitively, since $\mathbf{P}^*$ is the optimal solution, any $\mathbf{P}_{\mathbf{K}_i}$ for a Hurwitz $\mathbf{K}_i$ satisfies $0 \preceq \mathbf{P}^* \preceq \mathbf{P}_{\mathbf{K}_i}$, which in turn implies that $\mathbf{P}_i$ would not satisfy $\mathbf{0}_{n\times n} \preceq \mathbf{P}_i \preceq \mathbf{P}^*$ especially when $\hbar$ is large (an example of this case is $\mathbf{P}^* \preceq \mathbf{P}_i \preceq \mathbf{P}_{\mathbf{K}_i}$). Therefore, VI-mode convergence is not attainable in general.

In contrast, I-GPI methods with a sufficiently small $\hbar > 0$ can generate the sequence $\{\mathbf{P}_i\}$, which satisfies $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ for all $i \in \mathbb{N}$ and hence, converges in VI-mode according to Theorem 4.5. In this case, $(\mathbf{P}_0, \mathbf{K}_0)$ is required to satisfy $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \succ 0$, instead of $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \succeq 0$. To see this, assume $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succ 0$. Then, $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$ is also positive definite by (4.30) or Theorem 4.2, which again implies there is $\varepsilon_i > 0$ such that $\varepsilon_i I_n \preceq \mathcal{L}(\mathbf{K}_i, \mathbf{P}_{i+1})$. So, if $\Delta\mathbf{K}_i$ satisfies

$$\Delta\mathbf{K}_i^T\mathbf{R}\Delta\mathbf{K}_i \prec \varepsilon_i I_n, \tag{4.40}$$

then $\mathbf{P}_{i+1}$ satisfies $\mathcal{R}(\mathbf{P}_{i+1}) \succ 0$ by (4.31); the induction implies $\mathcal{R}(\mathbf{P}_i) \succ 0$ for all $i \in \mathbb{N}$, and the VI-mode convergence is guaranteed by Theorem 4.5. Here, since $\left\|\Delta\mathbf{K}_i^T\mathbf{R}\Delta\mathbf{K}_i\right\|$ can be made arbitrarily small by decreasing $\hbar > 0$ (notice $\Delta\mathbf{K}_i = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{M}_{(i,\hbar)}$ by (4.29) and $\mathbf{M}_{(i,\hbar)} \to 0$ as $\hbar \to 0$), the I-GPI with a sufficiently small $\hbar > 0$ yields $\Delta\mathbf{K}_i$ satisfying (4.40) and thereby, can generate the convergent sequence $\{\mathbf{P}_i\}$ in VI-mode by Theorem 4.5.

This VI-mode convergence can be also possible for infinitesimal GPI and I-VI with sufficiently small $T_s > 0$ since they belong to the special class of I-GPI with "$0 < \hbar \ll 1$" in

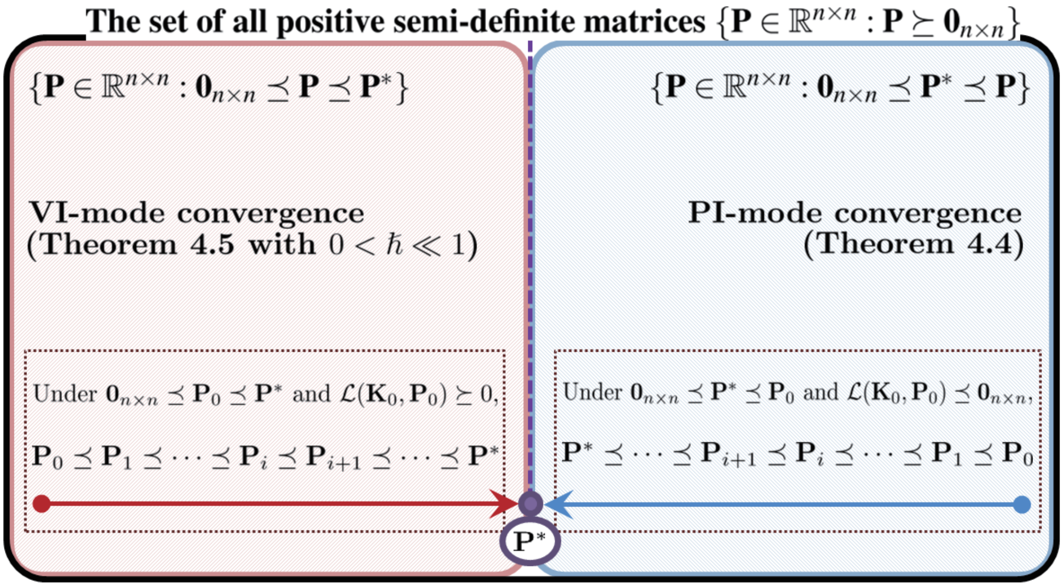

The set of all positive semi-definite matrices $\{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{P} \succeq \mathbf{0}_{n\times n}\}$
$\{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{0}_{n\times n} \preceq \mathbf{P} \preceq \mathbf{P}^*\}$
$\{\mathbf{P} \in \mathbb{R}^{n\times n} : \mathbf{0}_{n\times n} \preceq \mathbf{P}^* \preceq \mathbf{P}\}$
VI-mode convergence
(Theorem 4.5 with $0 < \hbar \ll 1$)
PI-mode convergence
(Theorem 4.4)
Under $\mathbf{0}_{n\times n} \preceq \mathbf{P}_0 \preceq \mathbf{P}^*$ and $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \succeq 0$,
$\mathbf{P}_0 \preceq \mathbf{P}_1 \preceq \cdots \preceq \mathbf{P}_i \preceq \mathbf{P}_{i+1} \preceq \cdots \preceq \mathbf{P}^*$
Under $\mathbf{0}_{n\times n} \preceq \mathbf{P}^* \preceq \mathbf{P}_0$ and $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$,
$\mathbf{P}^* \preceq \cdots \preceq \mathbf{P}_{i+1} \preceq \mathbf{P}_i \preceq \cdots \preceq \mathbf{P}_1 \preceq \mathbf{P}_0$
$\mathbf{P}^*$

Figure 4.4: PI- and VI-mode convergence of I-GPI.

the new classification (see Fig. 4.1). Actually, the VI-mode convergence of the infinitesimal I-GPI $\dot{\mathbf{P}}_t = \mathcal{R}(\mathbf{P}_t)$ $(0 \le t < \infty)$ under the zero initial condition $\mathbf{P}_0 = \mathbf{0}_{n\times n}$, discussed in Section 4.3, is the special case of that under $\mathbf{0}_{n\times n} \preceq \mathbf{P}_0 \preceq \mathbf{P}^*$ and $\mathcal{R}(\mathbf{P}_0) \succeq \mathbf{0}_{n\times n}$. Therefore, Theorem 4.5 also shows the monotone increasing and VI monotone convergence conditions of infinitesimal GPI in the general case "$\mathbf{0}_{n\times n} \preceq \mathbf{P}_0 \preceq \mathbf{P}^*$ and $\mathcal{R}(\mathbf{P}_0) \succeq \mathbf{0}_{n\times n}$" that contains the infinitesimal GPI with $\mathbf{P}_0 = \mathbf{0}_{n\times n}$ as a special case.

4.5.4 Convergence in PI-Mode versus VI-Mode

PI- and VI-mode convergence of I-GPI can be illustrated in Fig. 4.4 and, in reference to that, can be summarized as follows.

- In PI-mode, $\mathbf{P}_i$ remains in the region $\{\mathbf{0}_{n\times n} \preceq \mathbf{P}^* \preceq \mathbf{P}\}$ for all $i \in \mathbb{Z}_+$ and converges like PI methods, *e.g.*, I-PI for CT LQR (I-GPI in the limit $\hbar \to \infty$).

- In VI-mode, $\mathbf{P}_i$ is in the other region $\{\mathbf{0}_{n\times n} \preceq \mathbf{P} \preceq \mathbf{P}^*\}$ for all $i \in \mathbb{Z}_+$, and converges like "infinitesimal GPI" and "VI in DT domain".

Note that PI- and VI-mode convergence of I-GPI can be considered the generalizations of monotone convergence at $\hbar = 0$ (infinitesimal GPI) and $\hbar = \infty$ (I-PI) of the spectrum in the new classification. The convergence properties of I-VI ($k = 1$) and I-GPI (fininte k) posed on the middle of the spectrum are determined depending on the update horizon $\hbar$ and the matrix inequality conditions in this section. Since I-GPI with the same $\hbar$ are all equal in iteration domain as shown in the previous chapter, they have the same PI-mode or VI-mode convergence property if one of them has.

Remark 4.6. *While the choice of $\hbar > 0$ does not affect PI-mode convergence (Theorem 4.4), VI-mode convergence can be achieved only with sufficiently small $\hbar > 0$ (or in the limit $\hbar \to 0$) as discussed in Section 4.5.3; otherwise, $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ is not guaranteed after some finite step $i = l$ and in this case, Theorem 4.5 only implies I-GPI generates $\mathbf{P}_i$ that is monotonically increasing up to l. On the other hand, PI-mode convergence is possible even in the limit $\hbar \to 0$ and $\hbar \to \infty$ as long as the Hurwitz policy $\mathbf{K}_0$ and $\mathbf{P}_0 \succeq \mathbf{0}_{n\times n}$ satisfy $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$.*

4.5.5 Numerical Simulations

To verify and further discuss the stability and convergence properties of the I-GPI, we simulated I-GPI (Algorithm 4.3) with the LQR problem considered in Section 4.4. In the simulations, the policy evaluation of I-GPI is performed either by solving online least-squares problem at each iteration (see [93] for this), or by just conducting the equivalent matrix iteration (4.30). In either case, the policy evaluation yields the same value function matrix $\{\mathbf{P}_i\}_{i=0}^{\infty}$. In policy improvement, the next policy $\mathbf{K}_{i+1}$ is directly obtained by (4.11).

Simulation Example 1: PI-Mode Convergence

In this example, the initial conditions $\mathbf{P}_0$ and $\mathbf{K}_0$ were set to $\mathbf{P}_0 = \mathbf{diag}\{10, 10, 20\}$ and $\mathbf{K}_0 = \begin{bmatrix} 0, & 0, & -14 \end{bmatrix}$, respectively, so that the initial pair $(\mathbf{P}_0, \mathbf{K}_0)$ satisfies $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \preceq \mathbf{0}_{n\times n}$. Figs. 4.5 and 4.6 show the simulation results for $k = 5$ and $T_s = 20$ [ms]. In this case, as stated in Theorem 4.4 and can be seen from the figures, all the closed-loop systems are stable and $\mathbf{P}_i$ monotonically converges to $\mathbf{P}^*$ in PI-mode. Fig. 4.5 illustrates the state

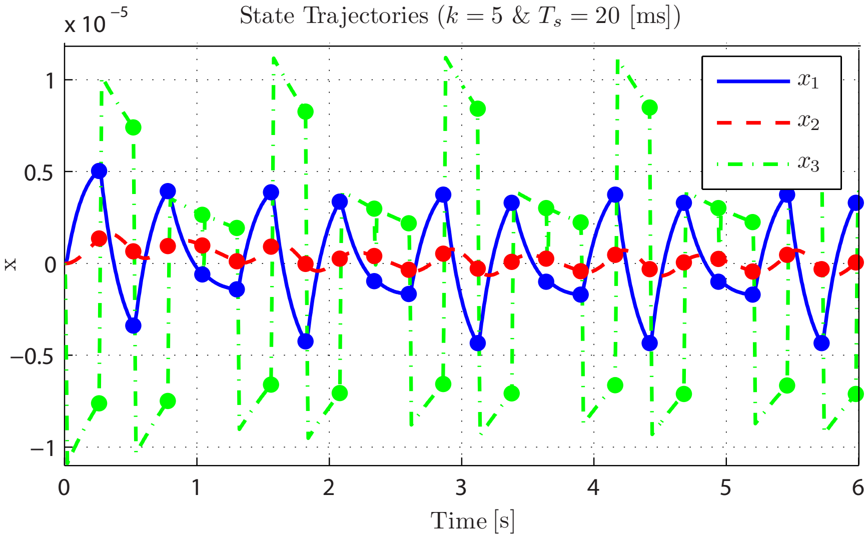
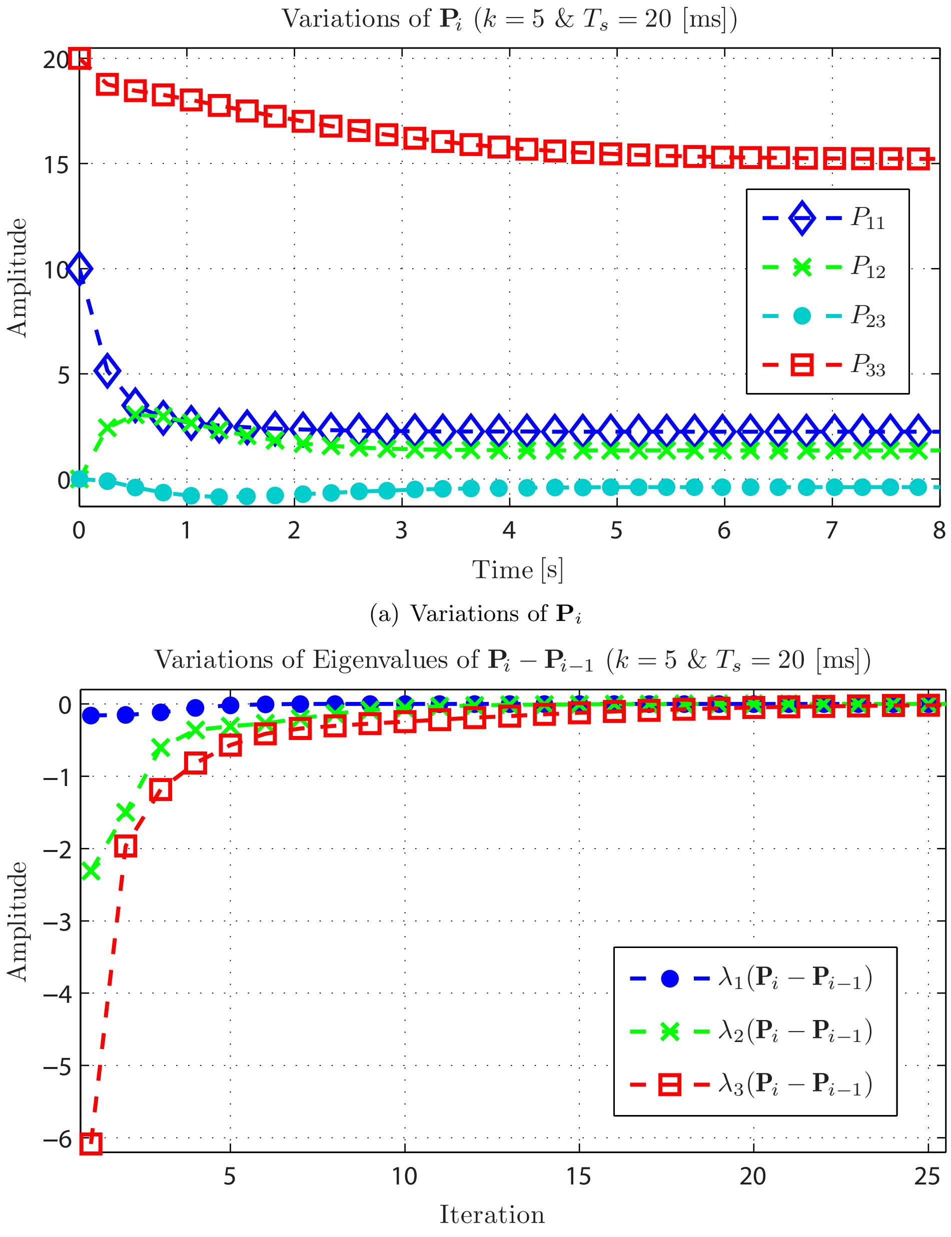


Figure 4.5: **(Example 1: PI-mode convergence)** Variations of state variable $\mathbf{x}_\tau$.

trajectories for the online LS implementation case [93], where the marked points indicate the time instant the policy was changed by the I-GPI agent. Here, the states rapidly vary right after the marked points due to the exploratory signal applied after every policy improvement. From this figure, one can see that the states remain in a small bounded region by the stability argument. In addition, Fig. 4.6(a) shows the convergence of $\mathbf{P}_i$ to $\mathbf{P}^*$, where the diagonals (P_{11} and P_{33}) are monotonically decreasing. This PI-mode convergence becomes obvious by Fig. 4.6(b), which shows the eigenvalues of the difference $\mathbf{P}_i - \mathbf{P}_{i-1}$ are always negative, implying $\mathbf{0}_{n\times n} \preceq \mathbf{P}_i \preceq \mathbf{P}_{i-1} \preceq \cdots \preceq \mathbf{P}_0$. Therefore, Fig. 4.6(a) and (b) exactly show PI-mode convergence stated in Theorem 4.4.



(a) Variations of $\mathbf{P}_i$

(b) Variations of $\lambda_k(\mathbf{P}_i - \mathbf{P}_{i-1})$

Figure 4.6: (**Example 1: PI-mode convergence**) Variations of (a) $\mathbf{P}_i$, and (b) eigenvalues of the difference $\mathbf{P}_i - \mathbf{P}_{i-1}$ for the I-GPI with $k = 5$ and $T_s = 20$ [ms]; the initial conditions are given by $\mathbf{P}_0 = \mathbf{diag}\{10, 10, 20\}$ and $\mathbf{u}_0 = 14x_3$.

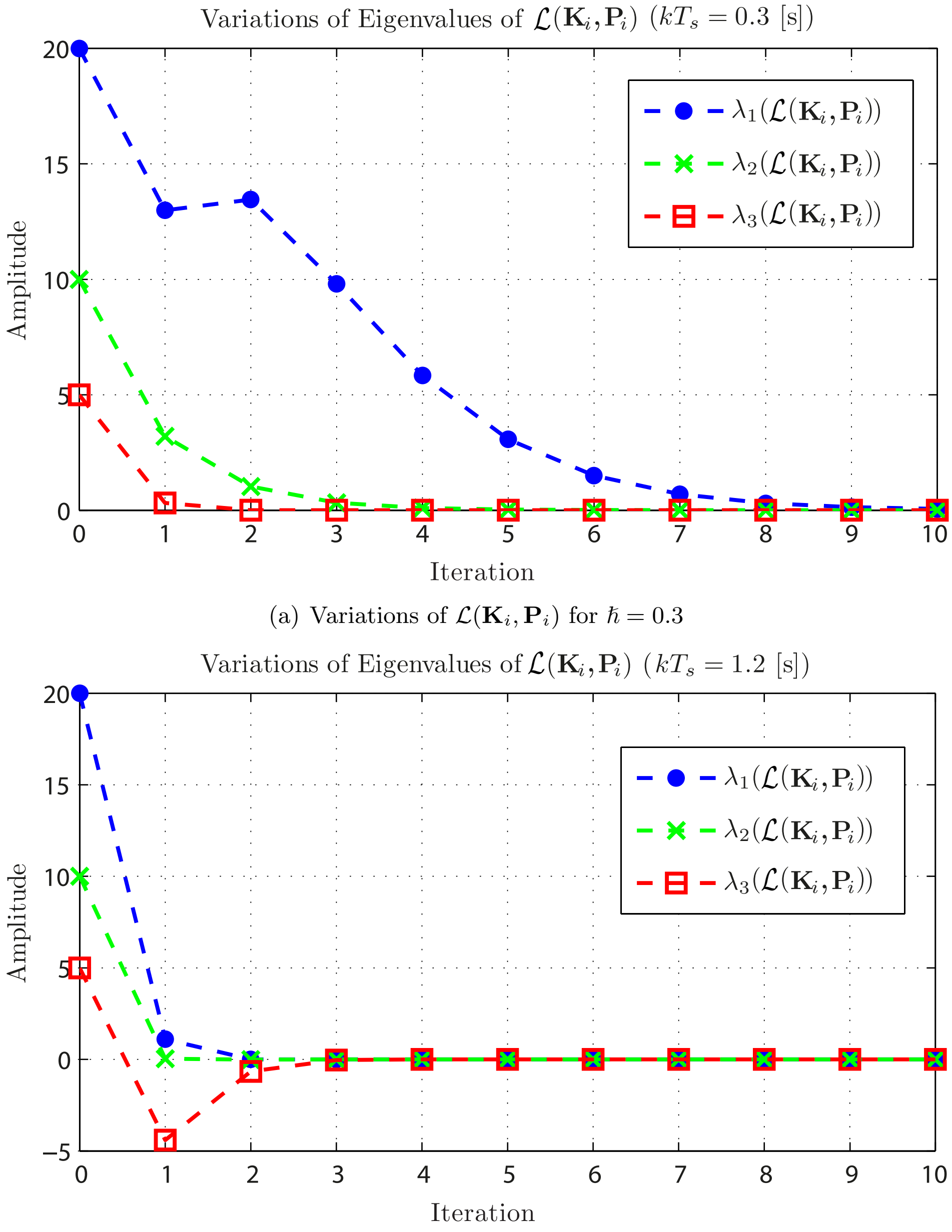
Simulation Example 2: VI-Mode Convergence

To further investigate the VI-mode convergence, an additional simulation was performed with $(\mathbf{P}_0, \mathbf{K}_0) = (\mathbf{0}_{3\times3}, \mathbf{0}_{1\times3})$ and $\hbar = 0.3$, 1.2 [s]. Then, both results for $\hbar = 0.3$ [s] and $\hbar = 1.2$ [s] were compared as shown in Fig. 4.7 and Table 4.1. In both simulations, T_s was set to $T_s = 0.1$ [s].

Fig. 4.7 shows the variations of eigenvalues of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$. In Fig. 4.7(a), it is shown that all the eigenvalues of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ remain positive for $\hbar = 0.3$ [s], implying VI-mode convergence by Theorem 4.5. Here, the convergence to $\mathbf{P}^*$ is verified by Fig. 4.2 and the monotonicity can be seen from Table 4.1, where the minimum eigenvalues of $\mathbf{P}_i - \mathbf{P}_{i-1}$ for $\hbar = 0.3$ [s] are all positive. This implies (4.39) with $l \to \infty$ in Theorem 4.5. On the other hand, in the case of $\hbar = 1.2$ [s], only the minimum eigenvalue of $\mathbf{P}_1 - \mathbf{P}_0$ ($i = 1$) is positive due to the initial condition $\mathcal{L}(\mathbf{K}_0, \mathbf{P}_0) \succeq \mathbf{0}_{n\times n}$, but the others are not due to the violations of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \succeq \mathbf{0}_{n\times n}$ for $i \geq 1$, as shown in Fig. 4.7(b) and Table 4.1 for $\hbar = 1.2$ [s]. Therefore, while $\mathbf{P}_i$ for $\hbar = 1.2$ [s] is actually shown to converge to $\mathbf{P}^*$ ($\because \mathcal{L}(\mathbf{K}_i, \mathbf{P}_i) \to 0$ by Fig. 4.7(b)), unlike the case with the small $\hbar = 0.3$ [s], the convergence is not monotone for this relatively large update horizon $\hbar = 1.2$ [s].

Table 4.1: **(Example 2: VI-mode convergence)** Variations of the minimum eigenvalue of $\mathbf{P}_i - \mathbf{P}_{i-1}$ for $\hbar = 0.3$ [s] and $\hbar = 1.2$ [s].

i	$\hbar = 0.3$ [s]	$\hbar = 1.2$ [s]	i	$\hbar = 0.3$ [s]	$\hbar = 1.2$ [s]
1	1.16e-00	1.59e-00	6	6.94e-08	-7.59e-06
2	3.53e-02	-2.30e-00	7	3.36e-09	-8.98e-08
3	1.07e-03	-2.76e-01	8	1.66e-10	-4.80e-10
4	3.72e-05	-1.38e-02	9	8.29e-12	-2.53e-11
5	1.52e-06	-3.97e-04	10	4.74e-13	-2.22e-12



(a) Variations of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ for $\hbar = 0.3$

(b) Variations of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ for $\hbar = 1.2$

Figure 4.7: (**Example 2: VI-mode convergence**) Variations of $\mathcal{L}(\mathbf{K}_i, \mathbf{P}_i)$ for the update horizons (a) $\hbar = 0.3$ [s] and (b) $\hbar = 1.2$ [s].

4.6 Summary

In this chapter, a family of partially model-free fundamental IRL algorithms including I-PI, I-VI, infinitesimal GPI, and their generalization "I-GPI" were presented in CT LQR framework and then classified in a new way in terms of the iteration horizon, the product of the iteration horizon involved in computational complexity and the time horizon determining the sampling period in time. In this new classification, the I-GPIs with the same update horizon are all equivalence classes in the iteration domain, implying the existence of the trade-off between the complexity and the sampling period. Then, the closed-loop stability and monotone convergence of I-GPI were investigated in relation to the update horizon. The main focus here were the two modes of convergence called VI- and PI-modes in convergence. These two convergence modes came from I-PI and infinitesimal GPI at the two extreme tips of the new classification and characterize the convergence behaviors of the fundamental IRLs. Here, it has been shown that PI-mode convergence guarantees the closed-loop stability and that VI-mode convergence is achieved only with the sufficiently small update horizon. Numerical simulations were conducted to support the theoretical foundations.

Chapter 5

Integral Reinforcement Learning with Invariant Explorations

This chapter introduces the IRL algorithms that efficiently and explicitly use the probing signal, injected to the target nonlinear dynamics through the control input channel, to solve the following optimal control problem:

$$\begin{aligned}&\text{minimize } J(\mathbf{x}_t, \mathbf{u}(\cdot)) = \int_t^\infty r(\mathbf{x}_\tau, \mathbf{u}_\tau)\, d\tau \\ &\text{subject to } \begin{cases} \dot{\mathbf{x}}_\tau = \mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)\mathbf{u}(\mathbf{x}_\tau), \ \ \mathbf{x}(t) = \mathbf{z} \in \mathcal{D} \subseteq \mathbb{R}^n \\ \mathbb{S} = \{\mathbf{0}_n\} \text{ and } S(\mathbf{x}) \succ 0 \text{ on } \mathcal{D} \end{cases}\end{aligned} \tag{5.1}$$

in online fashion, where $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$ with $\mathbf{f}(\mathbf{0}_n) = \mathbf{0}_n$ and $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^{n\times m}$ are nonlinear functions in $C^0_L(\mathcal{D})$; $\mathcal{D} \subseteq \mathbb{R}^n$ is the open connected domain of $\mathbf{f}$ and $\mathbf{G}$ that contains $\mathbf{0}_n$ in its interior; $r(\mathbf{x}, \mathbf{u}) := S(\mathbf{x}) + \mathbf{u}^T\mathbf{R}(\mathbf{x})\mathbf{u} \succ 0$ is the cost defined for a positive *definite* function $S : \mathcal{D} \to \mathbb{R}_+$ and a matrix-valued uniformly bounded smooth function $\mathbf{R} : \mathcal{D} \to \mathbb{R}^{m\times m}$ that is positive definite, uniformly for all $\mathbf{x} \in \mathcal{D}$. Here, the optimal control problem (5.1) is exatly same to the nonlinear optimal control problem (3.1) and (3.4) under Assumption 3.4 considered in Sections 3.3 and 3.4. The two IRL algorithms introduced in this chapter are the explorized variants of the ideal I-PI (Algorithm 5.1) introduced in the next section and derived from the ideal PI (Algorithm 3.1) in Section 3.1; the ultimate goal of the two IRL methods in this chapter is to find the solution $(V^*, \mathbf{u}^*)$ of the nonlinear optimal control problem (5.1) in online fashion when the nonlinear system (3.1) is explored by a known time-varying probing signal $\mathbf{e}_\tau$ so behaves according to the following $\mathbf{e}_\tau$-dependent nonlinear dynamics:

$$\dot{\mathbf{x}}_\tau = \mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)[\mathbf{u}(\mathbf{x}_\tau) + \mathbf{e}_\tau], \ \ \mathbf{x}(t) = \mathbf{z} \in \mathcal{D}, \tag{5.2}$$

where $\mathbf{e} : [t, \infty) \to \mathbb{R}^m$ is the probing signal, called an exploration. In this dissertation, the term 'exploration' is precisely defined as follows.

Definition 5.1. *A time function* $\mathbf{e} : [t, \infty) \to \mathbb{R}^m$ *is called an exploration if it is piecewise continuous and uniformly bounded for all* $t \geq 0$.

In this chapter, the state trajectory $\mathbf{x}_\tau$ at time $\tau \geq t$, will be denoted by $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ to indicate that it is generated by the explorized nonlinear system (5.2) under the given policy $\mathbf{u}(\mathbf{x})$ and exploration $\mathbf{e}_\tau$. Obviously, $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}) \equiv \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{0}_m)$ holds, which corresponds to the state trajectory $\mathbf{x}_\tau$ under the *zero* exploration $\mathbf{e}_\tau \equiv \mathbf{0}_m$.

In the IRL algorithms in this chapter, the use of the exploration $\mathbf{e}$ makes it possible to efficiently explore the state space in online fashion and more importantly, to relax the requirements of the knowledge of the input-coupling term $\mathbf{G}(\mathbf{x})$ that makes the IRL algorithms *model-free*. These improvements actually came from the ideal PI (Algorithm 5.1) combined with the concepts of explorations and temporal difference (TD) of RL in a finite MDP [1]. To develop such online IRL methods, however, these RL concepts have to be extended in the sense of admissiblility-guarantee and TD error compensations.

Together with the I-PI for the nonlinear system, we describe and organize these extended RL concepts named as "*invariant explorations*" and "*advanced I-TD*". Then, the target IRL algorithms will be derived using these extended RL concepts and nonlinear I-PI.

5.1 Nonlinear Integral Policy Iteration on ROAs

I-PI is a fundamental IRL algorithm to obtain the optimal solution $(\mathbf{u}^*, V^*)$ satisfying (3.23) and (3.5), without explicit use of the knowledge of the system drift dynamics $\mathbf{f}(\mathbf{x})$. Since I-PI can be applied to the nonlinear system (3.1) with completely unknown $\mathbf{f}(\mathbf{x})$, it is classified as a class of partially model-free IRL methods. The main idea of I-PI is to integrate the Hamiltonian equation (3.28) in PI from t to $t + T_s$ for some $T_s > 0$, which results in the following I-TD equation:

$$V_{\mathbf{u}}(\mathbf{x}_t) = \int_t^{t+T_s} r(\mathbf{x}_\tau, \mathbf{u}(\mathbf{x}_\tau))\, d\tau + V_{\mathbf{u}}(\mathbf{x}_{t+T_s}), \ \ \forall \mathbf{x}_t \in R_A(\mathbf{u}), \tag{5.3}$$

for an admissible policy $\mathbf{u}$, where $\mathbf{x}_\tau \equiv \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{0}_m)$ $(\tau \in [t, t + T_s])$ for $\mathbf{z} = \mathbf{x}(t)$.

The ideal I-PI is described in Algorithm 5.1, whose difference from the ideal PI (Algorithm 3.1) lies only in policy evaluation (line 3). While the value function in the ideal PI is obtained by solving the Hamiltonian equation (5.4), which requires the knowledge of $\mathbf{f}(\mathbf{x})$ and $\mathbf{G}(\mathbf{x})$, the ideal I-PI (Algorithm 5.1) finds the value function by solving the I-TD equation (5.4), where $\mathbf{f}(\mathbf{x})$ and $\mathbf{G}(\mathbf{x})$ are not explicitly shown, so they are not required to be known to execute policy evaluation of I-PI. On the other hand, $\mathbf{G}(\mathbf{x})$ should be

Algorithm 5.1: Ideal Integral Policy Iteration

```
Input: an initial admissible policy $\mathbf{u}_0 : \mathcal{D} \to \mathbb{R}^m$.
Output: the optimal solution $(\mathbf{u}^*, \mathbf{V}^*)$ satisfying (3.23) and (3.5).
1 $i \leftarrow 0$;
2 repeat
3     Policy Evaluation: find the value function $V_{\mathbf{u}_i} : R_A(\mathbf{u}_i) \to \mathbb{R}$ that belongs to
      $C^1_{L^+}(\mathbf{u}_i)$ and satisfies
          $V_{\mathbf{u}_i}(\mathbf{x}_t) = \int_t^{t+T_s} r(\mathbf{x}_\tau, \mathbf{u}_i(\mathbf{x}_\tau))\, d\tau + V_{\mathbf{u}_i}(\mathbf{x}_{t+T_s}), \ \ \forall \mathbf{z} \in R_A(\mathbf{u}_i),$   (5.4)
      where $\mathbf{x}_\tau \equiv \mathbf{x}(\mathbf{z}; \mathbf{u}_i)$ for $\mathbf{z} = \mathbf{x}(t)$ and an (admissible) policy $\mathbf{u}_i$;
4     Policy Improvement: update the next policy $\mathbf{u}_{i+1} : \mathcal{D} \to \mathbb{R}^n$ which is locally
      Lipschitz continuous and whose restriction on $R_A(\mathbf{u}_i)$ satisfies
          $\mathbf{u}_{i+1}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\, \mathbf{G}^T(\mathbf{x})\nabla V_{\mathbf{u}_i}(\mathbf{x}) \ \ \forall \mathbf{x} \in R_A(\mathbf{u}_i);$   (5.5)
5     $i \leftarrow i + 1$;
6 until convergence is met.
```

known to execute policy improvement of I-PI (see (3.29)); this restriction will be relaxed in Section 5.4.2. Since finding the solution $V_{\mathbf{u}}$ of (5.3) in the ideal I-PI is equivalent to solving (3.19) in the ideal PI (see Theorem 5.3 with $\mathbf{e} \equiv \mathbf{0}_m$ for this issue), the ideal I-PI have the same properties shown in Corollary 3.6 and Theorem 3.6 such as monotone improvement (3.31) and uniform convergence on any compact subset Ω of $R_A(\mathbf{u}_0)$.

5.2 Invariant Explorations and Input-To-State Stability

If the exploration $\mathbf{e}_\tau$ is given by a nonzero constant vector $\mathbf{c}$ such that $\mathbf{c} \notin \ker \mathbf{G}(\mathbf{x})$, then the equilibrium "$\mathbf{x} = \mathbf{0}_n$" of the nonlinear system (3.1) is no more the equilibrium of the explorized system (5.2). Moreover, a time-varying exploration $\mathbf{e}_\tau$ makes the system (5.2) non-autonomous. Hence, due to the non-zero (bounded) $\mathbf{e}_\tau$, the trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ for an admissible policy $\mathbf{u}$ and an initial condition $\mathbf{z} \in R_A(\mathbf{u})$ may escape the ROA $R_A(\mathbf{u})$, causing instability to the system. This is because unlike in a finite MDP [1] or linear dynamical systems [5, 19, 23, 93], $R_A(\mathbf{u})$ *is generally no more invariant under* $\mathbf{u}$ *and non-zero* $\mathbf{e}_\tau$. To prevent this pathological unstable situation due to $\mathbf{e}_\tau$, one should carefully design the exploration in a way that $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ is confined in $R_A(\mathbf{u})$ for all $\tau \geq t$. Here, the concept of invariant exploration plays an essential role in designing such $\mathbf{e}_\tau$.

Definition 5.2. *For a given admissible policy* $\mathbf{u}$*, let* $\Omega_{\mathcal{I}}(\mathbf{u})$ *be an invariant subset of* $R_A(\mathbf{u})$ *with respect to the autonomous non-explorized system* (3.1) *under* $\mathbf{u}(\mathbf{x})$*. Then, an exploration* $\mathbf{e}$ *is said to be invariant on* $\Omega_{\mathcal{I}}(\mathbf{u})$ *if*

$$\mathbf{z} \in \Omega_{\mathcal{I}}(\mathbf{u}) \implies \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e}) \in \Omega_{\mathcal{I}}(\mathbf{u}), \ \ \forall \tau \geq t. \tag{5.6}$$

Notice that the invariance (5.6) in Definition 5.2 is an extension of the invariance in autonomous systems to the explorized system (5.2). Moreover, the invariant exploration on a compact subset, say $\bar{\Omega}_{\mathcal{I}}(\mathbf{u})$, guarantees the existence of the unique solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ for all $\tau \geq t$ as shown below.

Proposition 5.1. *For an admissible policy* $\mathbf{u}$*, if the exploration* $\mathbf{e}$ *is invariant on a compact subset* $\bar{\Omega}_{\mathcal{I}}(\mathbf{u})$ *of* $R_A(\mathbf{u})$*, then for any* $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u})$*, the unique solution* $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ *exists* *for all $\tau \in [t, \infty)$.*

Proof. At fixed time τ, $\mathbf{e}_\tau$ is constant, so $\mathbf{f} + \mathbf{G}[\mathbf{u} + \mathbf{e}_\tau]$ is locally Lipschitz continuous on $\mathcal{D}$ by Lemma 2.6. In addition, the invariance of $\mathbf{e}$ guarantees that every solution $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ for $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u})$ entirely lies on $\bar{\Omega}_{\mathcal{I}}(\mathbf{u})$. Therefore, the proof can be done by applying Theorem 3.3 in [32]. □

The invariant exploration is actually related to the input-to-state stability (ISS) for the explorized system (5.2), which is the stability counterpart of Definition 3.2 and precisely defined as follows.

Definition 5.3. *For a given policy $\mathbf{u}$, the system (5.2) with an exploration $\mathbf{e}$ is said to be input-to-state stable on a subset $\Omega \in \mathfrak{D}(\mathcal{D}, \{\mathbf{0}_n\})$ if there exist $\alpha(\cdot)$, $\gamma(\cdot) \in \mathcal{K}$ and $\beta(\cdot, \cdot) \in \mathcal{KL}$ such that for any $\mathbf{z} \in \Omega$,*

$$\gamma\big(\|\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})\|\big) \leq \max\left\{\beta\big(\|\mathbf{z}\|, \, \tau - t\big), \;\; \alpha\bigg(\sup_{t \leq s \leq \tau} \|\mathbf{e}(s)\|\bigg)\right\}, \quad \forall \tau \geq t. \tag{5.7}$$

Now, consider the sequences $\{V_{\mathbf{u}_i}\}_{i=0}^{\infty}$ and $\{\mathbf{u}_i\}_{i=0}^{\infty}$ generated by the ideal I-PI (Algorithm 5.1). Then, the following theorem states the bounding condition of exploration $\mathbf{e}$ that guarantees the invariance of $\mathbf{e}$ and ISS on a compact set $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i})$ defined as

$$\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i}) := \big\{\mathbf{x} \in \mathcal{D} : V_{\mathbf{u}_i}(\mathbf{x}) \leq \bar{\alpha}_{\mathbf{u}_i}(r_{\mathbf{u}_i})\big\},$$

where $r_{\mathbf{u}_i} > 0$ is a constant satisfying $\bar{B}_{\mathbf{0}_n}(r_{\mathbf{u}_i}) \subset R_A(\mathbf{u}_i)$ and defining a closed interval $[0, r_{\mathbf{u}_i}]$ on which the class $\mathcal{K}$ functions $\underline{\alpha}(\cdot)$ and $\bar{\alpha}(\cdot)$ in (3.16) are defined.

Theorem 5.1. *Consider the sequences $\{\mathbf{u}_i\}$ and $\{V_{\mathbf{u}_i}\}$ generated by the ideal I-PI (Algorithm 5.1). If the exploration $\mathbf{e}$ satisfies*

$$\sup_{t \leq \tau < \infty} \|\mathbf{e}(\tau)\|_2 < \sqrt{\frac{\underline{\alpha}_s(r_{\mathbf{u}_i})}{\sup_{\mathbf{x} \in \mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))}}, \tag{5.8}$$

then under the policy $\mathbf{u}_i(\mathbf{x})$ or $\mathbf{u}_{i+1}(\mathbf{x})$,

1. *$\mathbf{e}(\tau)$ is invariant on $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i})$;*

2. *the explorized system (5.2) is input-to-state stable on $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i})$.*

Proof. See Appendix D.4. □

The next theorem is the counterpart of Theorem 5.1 of the global case $\mathcal{D} = R_A(\mathbf{u}_0) = \mathbb{R}^n$. In this case, the statement is dramatically simplified as shown below.

Theorem 5.2. *Suppose $S(\mathbf{x})$ is radially unbounded and $\mathcal{D} = R_A(\mathbf{u}_0) = \mathbb{R}^n$. Then, for under any $\mathbf{u}_i(\mathbf{x})$ generated by the ideal PI (Algorithm 5.1),*

- *any exploration $\mathbf{e}$ is invariant;*
- *ISS holds globally for any $\mathbf{z} \in \mathbb{R}^n$ and any (bounded) exploration $\mathbf{e}_\tau$.*

Proof. By the expansion property of the ROA shown in Remark 3.3, $R_A(\mathbf{u}_0) = \mathbb{R}^n$ implies that $R_A(\mathbf{u}_i) = \mathbb{R}^n$ for all $i \in \mathbb{Z}_+$. In this global case, $V_{\mathbf{u}_i}(\mathbf{x})$ is radially unbounded by Proposition 3.2, and so is $S(\mathbf{x})$ by assumption. Hence, $V_{\mathbf{u}_i}(\mathbf{x})$ is defined for all $\mathbf{x} \in \mathbb{R}^n$; the class $\mathcal{K}$ functions $\underline{\alpha}_s$, $\underline{\alpha}_{\mathbf{u}_i}$, $\bar{\alpha}_{\mathbf{u}_i}$, and their inverses all belong to $\mathcal{K}_\infty$ by Lemma 2.6 and [32, Lemma 4.2], so they are defined on $[0, \infty)$.

By the above argument, for any $\mathbf{z} \in \mathbb{R}^n$, there exists $r_{min1} > 0$ such that $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i})$ for all $r_{\mathbf{u}_i} > r_{min1}$. Moreover, for any given (bounded) exploration $\mathbf{e}$, if $r_{\mathbf{u}_i} > 0$ is chosen in the range

$$r_{\mathbf{u}_i} > \underline{\alpha}_s^{-1}\left(\left(\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))\right) \cdot \left(\sup_{t\le\tau<\infty} \|\mathbf{e}(\tau)\|_2^2\right)\right) := r_{min2},$$

then (5.8) holds. Therefore, for any $\mathbf{z} \in \mathbb{R}^n$ and any given exploration $\mathbf{e}_\tau$, if $r_{\mathbf{u}_i}$ is chosen sufficiently large so "$r_{\mathbf{u}_i} > \max\{r_{min1}, r_{min2}\}$" holds, then both $\mathbf{z} \in \Omega_{\mathcal{I}}(\mathbf{u}_i; r_{\mathbf{u}_i})$ and (5.8) hold. Since $\mathbf{z} \in \mathbb{R}^n$ is arbitrary, the application of Theorem 5.1 completes the proof. □

5.3 Advanced I-TD and Exploration Design Principles

If $\mathbf{x}_\tau$ is generated by (5.2) with non-zero exploration $\mathbf{e}$, then I-TDs (5.3) and (3.28) in policy evaluation of the I-PI do not function properly. Meanwhile, if $\mathbf{G}(\mathbf{x})$ is not known *a priori*, the next policy $\mathbf{u}_{i+1}$ cannot be updated by policy improvement of I-PI, either. To solve these two problems, the following $\mathbf{e}$-dependent advanced I-TD equation is devised from the I-TD equation (5.3):

$$V(\mathbf{x}_t) = \int_t^{t+T_s} \Big[r(\mathbf{x}, \mathbf{u}(\mathbf{x})) + 2\boldsymbol{\mu}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{e}(\tau) \Big] d\tau + V(\mathbf{x}_{t+T_s}), \tag{5.9}$$

where $(\mathbf{u}, \mathbf{e})$ is a given policy-exploration pair such that $\mathbf{u}$ is admissible and $\mathbf{e}$ is invariant on $R_A(\mathbf{u})$ under $\mathbf{u}$; $\mathbf{x}$ denotes the trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ for $\mathbf{z} = \mathbf{x}_t \in R_A(\mathbf{u})$. By solving the advanced I-TD equation (5.9), it is meant to find the locally Lipschitz continuous positive definite function $V \in C^1_{L+}(\mathbf{u})$ and an admissible policy $\boldsymbol{\mu} \in C^0_L(\mathcal{D})$ that satisfy (5.9) for all $\mathbf{z} \in R_A(\mathbf{u})$. The two IRL methods named *explorized I-PI* and *integral Q-learning* in this chapter will be designed based on the advanced I-TD. Compared to I-TD equation (5.3), the exploration cross-product term $\boldsymbol{\mu}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{e}(\tau)$ is added to cancel out the effects of $\mathbf{e}$ on I-TD and at the same time acquire the new policy $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{u}^+(\mathbf{x})$ without knowing $\mathbf{G}(\mathbf{x})$ *a priori*. Here, $\mathbf{u}^+(\mathbf{x})$ is the desired next policy defined in terms of $\mathbf{G}(\mathbf{x})$ and $\nabla V(\mathbf{x})$ as

$$\mathbf{u}^+(\mathbf{x}) := -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V(\mathbf{x}).$$

In the following discussions related to the advanced I-TD, it is assumed without loss of generality that the exploration $\mathbf{e}$ is T_s-periodic, i.e., $\mathbf{e}_\tau = \mathbf{e}_{\tau+T_s}$ for all $\tau \geq t$.

Theorem 5.3. *Finding $V \in C^1_{L+}(\mathbf{u})$ and an admissible policy $\boldsymbol{\mu} \in C^0_L(\mathcal{D})$ satisfying* (5.9) *for all $\mathbf{z} = \mathbf{x}_t \in R_A(\mathbf{u})$ is equivalent to solving*

$$\mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x})) = 2\boldsymbol{\varphi}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{e}_\tau \tag{5.10}$$

for all $\mathbf{x} \in R_A(\mathbf{u})$ and $\tau \in [t, t+T_s)$, where $\boldsymbol{\varphi} := \mathbf{u}^+ - \boldsymbol{\mu}$ is the policy approximation error function.

Proof. Note that since $\mathbf{u}$ is admissible and $\mathbf{e}$ is invariant on $R_A(\mathbf{u})$ under the policy $\mathbf{u}$, the trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}, \mathbf{e})$ lies entirely in $R_A(\mathbf{u})$, for all $\tau \geq t$. So, $V \in C^1_{L+}(\mathbf{u})$ satisfies

$$V(\mathbf{x}_{t+T}) - V(\mathbf{x}_t) = \int_t^{t+T} \dot{V}(\mathbf{x}_\tau)\, d\tau, \tag{5.11}$$

for any initial value $\mathbf{x}_t = \mathbf{z} \in R_A(\mathbf{u})$, where the time derivative $\dot{V}(\mathbf{x}_\tau)$ is given by

$$\dot{V}(\mathbf{x}_\tau) = \nabla^T V(\mathbf{x}_\tau) \cdot (\mathbf{f}(\mathbf{x}_\tau) + \mathbf{G}(\mathbf{x}_\tau)[\mathbf{u}(\mathbf{x}_\tau) + \mathbf{e}_\tau]). \tag{5.12}$$

Define $H(\mathbf{x}, \mathbf{e}) := \mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x})) - 2\boldsymbol{\varphi}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{e}$, where $\mathcal{H}(\cdot, \cdot, \cdot)$ is the Hamiltonian defined by (3.18). Then, substituting (5.11) and (5.12) into the I-TD (5.9), we obtain

$$\int_t^{t+T} H\big(\mathbf{x}_\tau(\mathbf{z};\mathbf{u},\mathbf{e}),\mathbf{e}_\tau\big)\ d\tau = 0. \tag{5.13}$$

Therefore, finding the solution of the advanced I-TD (5.9) for all $\mathbf{x}_t = \mathbf{z} \in \Omega$ is equivalent to solving (5.13) $\forall \mathbf{z} \in \Omega$. Since $\mathbf{x}_\tau(\mathbf{z};\mathbf{u},\mathbf{e}) \in R_A(\mathbf{u})$ for all $\tau \geq t + T$, following the same steps with starting time $t + MT$, instead of t, yields

$$\int_{t+MT}^{t+(M+1)T} H\big(\mathbf{x}_\tau(\mathbf{z};\mathbf{u},\mathbf{e}),\mathbf{e}_\tau\big)\ d\tau = 0,\ \ \forall M \in \mathbb{Z}_+.$$

Then, summing up the integrals for all $M \in \mathbb{Z}_+$, we obtain

$$h(t;\mathbf{z}) \equiv \int_t^{\infty} H\big(\mathbf{x}_\tau(\mathbf{z};\mathbf{u},\mathbf{e}),\mathbf{e}_\tau\big)\ d\tau = 0.$$

That is, $h(t;\mathbf{z}) = 0$ for all $t \geq 0$ and all $\mathbf{z} \in R_A(\mathbf{u})$. So, we have

$$0 \equiv \dot{h}(t;\mathbf{z}) = -H\big(\mathbf{x}_\tau(\mathbf{z};\mathbf{u},\mathbf{e}),\mathbf{e}_\tau\big)|_{\tau=t},$$

which implies that

$$H\big(\mathbf{z},\mathbf{e}_t\big) = 0,\ \ \forall t \geq 0 \text{ and } \forall \mathbf{z} \in R_A(\mathbf{u}). \tag{5.14}$$

Since e is T-periodic, *i.e.*, $\mathbf{e}_\tau = \mathbf{e}_{\tau+T}$ for all $\tau \geq t$, (5.14) is reduced to

$$H\big(\mathbf{z},\mathbf{e}_\tau\big) = 0,\ \ \forall \tau \in [t, t+T) \text{ and } \forall \mathbf{z} \in R_A(\mathbf{u}),$$

which is equivalent to (5.10) by the definition of $H(\mathbf{z},\mathbf{e})$. The proof of the opposite direction can be easily done by first integrating (5.10) and then substituting (5.11) and (5.12). □

Using Theorem 5.3, one can easily verify that if $V_\mathbf{u} \in C^1_{L+}(\mathbf{u})$, then

$$V(\mathbf{x}) = V_\mathbf{u}(\mathbf{x}),\ \ \boldsymbol{\mu}(\mathbf{x}) = \mathbf{u}^+(\mathbf{x})|_{V=V_\mathbf{u}} \tag{5.15}$$

are a solution to the advanced I-TD equation (5.9) and satisfy

$$\mathcal{H}(\mathbf{x},\mathbf{u}(\mathbf{x}),\nabla V(\mathbf{x})) = 0,\ \ \boldsymbol{\varphi}(\mathbf{x}) = \mathbf{0}_m,\ \ \forall x \in R_A(\mathbf{u}).$$

However, the solution may not be unique. For example, if $m = 1$ and $\mathbf{e}_\tau$ is constant and nonzero, i.e., $\mathbf{R}(\mathbf{x}) \equiv r$ for some $r > 0$ and $\mathbf{e}_\tau \equiv c$ for some $c \in \mathbb{R}\setminus\{0\}$ for all $\tau \in [t, t+T_s)$, then Theorem 5.3 implies that $\boldsymbol{\mu}$ can be obtained from $V(\mathbf{x})$, r, and c as

$$\boldsymbol{\mu}(\mathbf{x}) = \mathbf{u}^+(\mathbf{x}) + \mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x}))/rc.$$

This means that for a given $V(\mathbf{x})$, there are infinitely many solutions depending on the non-zero constant exploration $\mathbf{e}_\tau \equiv c$ unless $\mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x})) \equiv 0$. On the other hand, for the case when $\mathbf{G}(\mathbf{x})$ is known, $\boldsymbol{\mu} = \mathbf{u}^+$ can be substituted to (5.9) to obtain the following simplified advanced I-TD equation:

$$V(\mathbf{x}_t) - V(\mathbf{x}_{t+T_s}) = \int_t^{t+T_s} \big[r(\mathbf{x}, \mathbf{u}(\mathbf{x})) - \nabla^T V(\mathbf{x}) \cdot \mathbf{G}(\mathbf{x})\mathbf{e}_\tau\big]\, d\tau. \tag{5.16}$$

In this case, the solution $V = V_\mathbf{u}$ to (5.16) is unique as stated below.

Corollary 5.1. *Assume that $V_\mathbf{u} \in C^1_{L^+}(\mathbf{u})$. If $V \in C^1_{L^+}(\mathbf{u})$ is another solution to the advanced I-TD equation* (5.16)*, then $V = V_\mathbf{u}$ on their domain $R_A(\mathbf{u})$.*

Proof. The I-TD equation (5.16) is an advanced I-TD equation (5.9) with $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{0}_m$. So, Theorem 5.3 implies that $V \in C^1_{L^+}(\mathbf{u})$ satisfying (5.16) for all $\mathbf{x} \in R_A(\mathbf{u})$ is the solution of the Hamiltonian equation

$$\mathcal{H}(\mathbf{x}, \mathbf{u}(\mathbf{x}), \nabla V(\mathbf{x})) = 0, \ \ \forall \mathbf{x} \in R_A(\mathbf{u})$$

Then, the application of Theorem 3.4 concludes $V = V_\mathbf{u}$. □

If $\mathbf{G}(\mathbf{x})$ is not known *a priori*, then we cannot substitute $\boldsymbol{\mu} = \mathbf{u}^+$ to the advanced I-TD equation (5.9). In this general case, the uniqueness of (5.15) depends on the excitation condition. To see this, let $t_j \in [t, t+T_s]$ $(j = 0, 1, \cdots, L)$ be the time instants satisfying "$t_0 = t \le t_1 \le t_2 \le \cdots \le t_L = t + T_s$" and assume that $\mathbf{e}_\tau$ is piecewise constant and determined by

$$\mathbf{e}_\tau = \mathbf{c}_j, \ \ \forall \tau \in [t_j, t_{j+1}), \tag{5.17}$$

where $\{\mathbf{c}_j\}_{j=1}^L$ is a sequence of constant vectors in $\mathbb{R}^m$. We also define the $m \times (l-k)$ matrix $\mathbf{C}_{k:l}$ for $1 \le k \le l \le L$ as

$$\mathbf{C}_{k:l} = \left[\, \mathbf{c}_k \,\vdots\, \mathbf{c}_{k+1} \,\vdots\, \cdots \,\vdots\, \mathbf{c}_l \,\right].$$

Then, under the substitution of (5.17), the Hamiltonian equation (5.10) can be written as

$$\mathcal{H}(\mathbf{x},\mathbf{u}(\mathbf{x}),\nabla V(\mathbf{x})) = 2\boldsymbol{\varphi}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{c}_j,\ \ \forall \mathbf{x}\in R_A(\mathbf{u}),\ \ \forall j\in\{1,2,\cdots,L\}, \tag{5.18}$$

and we obtain the uniqueness condition of the solution (5.15) of the advanced I-TD (5.9).

Assumption 5.2. *There exist κ_1, $\kappa_2 > 0$ such that*

$$\kappa_1\mathbf{I}_m \preceq \sum_{j=1}^{L-1}(\mathbf{c}_j-\mathbf{c}_{j+1})(\mathbf{c}_j-\mathbf{c}_{j+1})^T \preceq \kappa_2\mathbf{I}_m.$$

Theorem 5.4. *Suppose $\mathbf{e}_\tau$ is given by* (5.17) *and $V_\mathbf{u}\in C^1_{L^+}(\mathbf{u})$. Then, the solution to the advanced I-TD* (5.9) *is uniquely determined by* (5.15) *under Assumption 5.2.*

Proof. By Theorem 5.3 and the above discussion, solving (5.9) for all $\mathbf{x}\in R_A(\mathbf{u})$ is equivalent to finding $V\in C^1_{L^+}(\mathbf{u})$ and a policy $\boldsymbol{\mu}\in C^0_L(\mathcal{D})$ satisfying (5.18) for all $\mathbf{x}\in R_A(\mathbf{u})$ and all $j\in\{1,2,\cdots,L\}$. From (5.18), we have $2(\mathbf{c}_j-\mathbf{c}_{j+1})^T\mathbf{R}(\mathbf{x})\boldsymbol{\varphi}(\mathbf{x}) = 0$ ($j=1,2,\cdots,L-1$). That is,

$$2(\mathbf{C}_{1:L-1}-\mathbf{C}_{2:L})^T\mathbf{R}(\mathbf{x})\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{0}_L. \tag{5.19}$$

From (5.19) and Assumption 5.2, $\boldsymbol{\varphi}(\mathbf{x})\equiv\mathbf{0}_m$ is obtained since Assumption 5.2 is equivalent to

$$\kappa_1 I \le (\mathbf{C}_{1:L-1}-\mathbf{C}_{2:L})(\mathbf{C}_{1:L-1}-\mathbf{C}_{2:L})^T \le \kappa_2 I,$$

which implies $\operatorname{rank}(\mathbf{C}_{1:L-1}-\mathbf{C}_{2:L}) = m$. Moreover, the substitution of $\boldsymbol{\varphi}(\mathbf{x})=\mathbf{0}_m$ into (5.18) yields $\mathcal{H}(\mathbf{x},\mathbf{u}(\mathbf{x}),\nabla V(\mathbf{x}))\equiv 0$. Therefore, the application of Theorem 3.4 proves $V=V_\mathbf{u}$, and we obtain $\boldsymbol{\mu}=\mathbf{u}^+|_{V=V_\mathbf{u}}$ from $\boldsymbol{\varphi}(\mathbf{x})\equiv\mathbf{0}_m$. □

Remark 5.1. *Note that for Assumption 5.2, there should exists a subsequence $\{\mathbf{c}_{j_k}\}_{k=1}^{m+1}$ whose difference $\{\mathbf{c}_{j_k}-\mathbf{c}_{j_{k+1}}\}_{k=1}^{m}$ is linearly independent; for this, $L\ge m+1$ is required. For instance, if $m=1$, two constants $c_1\neq c_2$ (*e.g.*, $c_1=1$ and $c_2=0$) are necessary to construct $\mathbf{e}_\tau$ without violating Assumption 5.2. Remember that "Assumption 5.2" is required to guarantee the uniqueness of the solution* (5.15) *as stated in Theorems 5.4.*

Remark 5.2. *If $t_{j+1}-t_j = T_s/L$ for all $j\in\{0,1,2,\cdots,L-1\}$, then Assumption 5.2 is equivalent to the existence of π_1, $\pi_2>0$ such that*

$$\pi_1\mathbf{I}_m \preceq \int_t^{t+(L-1)T_s/L}(\mathbf{e}_\tau-\mathbf{e}_{\tau+T_s/L})(\mathbf{e}_\tau-\mathbf{e}_{\tau+T_s/L})^T d\tau \preceq \pi_2\mathbf{I}_m. \tag{5.20}$$

This can be provided as a general condition on the exploration $\mathbf{e}$ *to guarantee the uniqueness of the solution* (5.15) *of the advanced I-TD* (5.9).

5.4 Explorized I-PI and Integral Q-learning

In this section, motivated by the advanced I-TD equations (5.9) and (5.16) in Section 5.3, the two online IRL algorithms named *explorized I-PI* and *integral Q-learning* are proposed, both of which exploit the exploration $\mathbf{e}_\tau$ to simultaneously excite the state variables and learn the next policy. Here, the former is partially model-free in a sense that the system drift dynamics $\mathbf{f}$ is not necessarily known to run it; the latter is model-free so that it can be applied to completely unknown dynamics $(\mathbf{f}, \mathbf{G})$. While I-PI in Section 5.1 is an off-line method, these partially/completely model-free IRL methods can run in online fashion even when the nonlinear system undergoes exploration $\mathbf{e}_\tau$. These online IRL methods are similar to the ideal I-PI (Algorithm 5.1) in principle, but different in practical manners as described below.

1. At each i-th policy evaluation and improvement steps of the ideal IRL, the IRL agent utilizes advanced I-TDs to find V_{i+1} and $\boldsymbol{\mu}_{i+1}$ satisfying $V_{i+1} \approx V^{\boldsymbol{\mu}_i}$ and $\boldsymbol{\mu}_{i+1} \approx \boldsymbol{\mu}_{i+1}^+$ on Ω_i, where $\boldsymbol{\mu}_{i+1}^+$ is given by

   $$\boldsymbol{\mu}_{i+1}^+(\mathbf{x}) := -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla V^{\boldsymbol{\mu}_i}(\mathbf{x}).$$

   Note that the IRL methods in this section can be considered the same to the ideal PI or I-PI in the iteration domain as long as the generated value functions and policies have no errors. Obviously, if $V_{i+1} = V^{\boldsymbol{\mu}_i}$, $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_{i+1}^+$ for all $i \in \mathbb{Z}_+$, and $\boldsymbol{\mu}_0 = \mathbf{u}_0$, then we have $V_{i+1} = V_{\mathbf{u}_i}$ and $\boldsymbol{\mu}_{i+1} = \mathbf{u}_{i+1}$ for all $i \in \mathbb{Z}_+$, where $(V_{\mathbf{u}_i}, \mathbf{u}_{i+1})$ is the sequence generated by the ideal PI or I-PI. Here, the approximation errors $|V_{\boldsymbol{\mu}_i}(\mathbf{x}) - V_{i+1}(\mathbf{x})|$ and $\|\boldsymbol{\mu}_{i+1}^+(\mathbf{x}) - \boldsymbol{\mu}_{i+1}(\mathbf{x})\|$ come from the advanced I-TDs, but can be made small by sufficiently exploring the state-space with well-designed explorations.

2. While the ideal I-PI cannot explore the state-space in online fashion, the explorized IRLs use invariant explorations to simultaneously excite the state variables in a stable manner. At each iteration, the IRL methods generate an invariant exploration $\mathbf{e}_\tau$ for stable learning (see Section 5.5 for more details).

5.4.1 Explorized Integral Policy Iteration

The first IRL method is partially model-free and named explorized I-PI in Algorithm 5.2. As can be seen from (5.21), explorized I-PI comes from the advanced I-TD (5.16) and is able to simultaneously excite the states during policy evaluation using the exploration $\mathbf{e}_\tau$. Unlike the IRL algorithms in Chapter 4, the I-TD equation (5.21) contains the explorized

Algorithm 5.2: Explorized Integral Policy Iteration

```
  Input: an initial admissible policy $\boldsymbol{\mu}_0 : \mathcal{D} \to \mathbb{R}^m$.
  Output: the optimal solution $(\mathbf{u}^*, \mathbf{V}^*)$.
1 $i \leftarrow 0$;
2 repeat
3     Policy Evaluation: Given an admissible policy $\boldsymbol{\mu}_i$ and given $\mathbf{z} \in R_A(\boldsymbol{\mu}_i)$,
          1. generate an exploration $\mathbf{e}_\tau$ that is invariant on $R_A(\boldsymbol{\mu}_i)$ under $\boldsymbol{\mu}_i$;
          2. find $V_{i+1} \in C^1_{L+}(\boldsymbol{\mu}_i)$ such that
             $$V_{i+1}(\mathbf{x}_t) - V_{i+1}(\mathbf{x}_{t+T_s}) = \int_t^{t+T_s} \big[r(\mathbf{x}, \boldsymbol{\mu}_i(\mathbf{x})) - \nabla^T V_{i+1}(\mathbf{x}) \cdot \mathbf{G}(\mathbf{x})\mathbf{e}_\tau\big]\, d\tau, \quad (5.21)$$
             where $\mathbf{x} \equiv \mathbf{x}_\tau(\mathbf{z}; \boldsymbol{\mu}_i, \mathbf{e})$;
4     Policy Improvement: update the next admissible policy $\boldsymbol{\mu}_{i+1} : \mathcal{D} \to \mathbb{R}^n$
      which is locally Lipschitz continuous and whose restriction on $R_A(\boldsymbol{\mu}_i)$ satisfies
             $$\boldsymbol{\mu}_{i+1}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\, \mathbf{G}^T(\mathbf{x}) \nabla V_{i+1}(\mathbf{x}) \;\; \forall \mathbf{x} \in R_A(\boldsymbol{\mu}_i);$$
5     $i \leftarrow i + 1$;
6 until convergence is met.
```

term "$\nabla^T V_{i+1}(\mathbf{x}) \cdot \mathbf{G}(\mathbf{x})\mathbf{e}_\tau$" to cancel out the effects of the exploration $\mathbf{e}_\tau$, but the knowledge of $\mathbf{G}(\mathbf{x})$ should be explicitly used for this exploration cancellation and policy improvement as shown in Algorithm 5.2. When $\mathbf{e}_\tau \equiv \mathbf{0}_m$, Algorithm 5.2 becomes the ideal I-PI described in Algorithm 5.1, provided that (5.21) holds for all $\mathbf{z} = \mathbf{x}_t \in R_A(\boldsymbol{\mu}_i)$. In explorized I-PI, $\mathbf{e}_\tau$ does not need to satisfy the excitation conditions such as that in Assumption 5.2 as neither does the advanced I-TD (5.16). By Corollary 5.1, if sufficiently explores the state space in a stable manner, explorized I-PI guarantees the uniqueness of the solution $V_i = V^{\boldsymbol{\mu}_i}$ for any given exploration $\mathbf{e}_\tau$, and one just need to efficiently explore the state-space using $\mathbf{e}_\tau$ without considering any excitation conditions *on* $\mathbf{e}_\tau$ in Section 5.3.

5.4.2 Model-Free Integral Q-learning

Integral Q-learning is the other online IRL algorithm proposed in this chapter, which is derived from the advanced I-TD equation (5.9) so can be implemented without knowing the system dynamics $(\mathbf{f}, \mathbf{G})$. Here, the exploration $\mathbf{e}_\tau$ plays a central role in relaxing the requirement of the knowledge of $\mathbf{G}(\mathbf{x})$. Algorithm 5.3 describes the processes of the proposed integral Q-learning, where the IRL agent generates an invariant exploration $\mathbf{e}$ under $\boldsymbol{\mu}_i$ and finds the solution $V_{i+1} \in C^1_{L^+}(\boldsymbol{\mu}_i)$ and $\boldsymbol{\mu}_{i+1} \in C^0_L(\mathcal{D})$, with $\boldsymbol{\mu}_{i+1}(\mathbf{0}_n) = \mathbf{0}_m$, of the advanced I-TD (5.22) all at the same time in the policy evaluation and improvement step.

For ease of explanation, it is assumed from now on that the exploration $\mathbf{e}$ applied to integral Q-learning is given by (5.17) for some constant vectors $\{\mathbf{c}_j\}_{j=1}^L$. In this case, for the uniqueness of the solution $(V_{i+1}, \boldsymbol{\mu}_{i+1}) = (V_{\boldsymbol{\mu}_i}, \boldsymbol{\mu}^+_{i+1})$, the vectors $\{\mathbf{c}_j\}_{j=1}^L$ should be carefully chosen so that they satisfy Assumption 5.2. In general cases, (5.20) can be an alternative to Assumption 5.2.

Algorithm 5.3: Model-Free Integral Q-Learning

```
Input: an initial admissible policy $\boldsymbol{\mu}_0 : \mathcal{D} \to \mathbb{R}^m$.
Output: the optimal solution $(\mathbf{u}^*, \mathbf{V}^*)$. satisfying (3.23) and (3.5).
$i \leftarrow 0$;
repeat
    Policy Evaluation & Improvement:
    Given an admissible policy $\boldsymbol{\mu}_i$ and given $\mathbf{z} \in R_A(\boldsymbol{\mu}_i)$,
        1. generate an exploration $\mathbf{e}_\tau$ that is invariant on $R_A(\boldsymbol{\mu}_i)$ under $\boldsymbol{\mu}_i$;
        2. find $V_{i+1} \in C^1_{L^+}(\boldsymbol{\mu}_i)$ and the next admissible policy $\boldsymbol{\mu}_{i+1} \in C^0_L(\mathcal{D})$ satisfying
           $$V_{i+1}(\mathbf{x}_t) - V_{i+1}(\mathbf{x}_{t+T_s}) = \int_t^{t+T_s} \Big[ r(\mathbf{x}, \boldsymbol{\mu}_i(\mathbf{x})) + 2\boldsymbol{\mu}_{i+1}^T(\mathbf{x})\mathbf{R}(\mathbf{x})\mathbf{e}_\tau \Big] d\tau \qquad (5.22)$$
           where $\mathbf{x} \equiv \mathbf{x}_\tau(\mathbf{z}; \boldsymbol{\mu}_i, \mathbf{e})$.
    $i \leftarrow i + 1$;
until convergence is met.
```

5.5 Exploration Design Considerations

In the previous section, we have emphasized the excitation condition (Assumption 5.3) for integral Q-learning to uniquely obtain $V_{i+1} \approx V_{\boldsymbol{\mu}_i}$ and $\boldsymbol{\mu}_{i+1} \approx \boldsymbol{\mu}_{i+1}^+$ at each iteration. On the other hand, explorized I-PI does not need such an excitation condition at the expense of model dependency on $\mathbf{G}(\mathbf{x})$. In this section, suppose that at i-th iteration, V_{i+1} and $\boldsymbol{\mu}_{i+1}$ has no error, i.e., $V_{i+1} = V_{\boldsymbol{\mu}_i}$ and $\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_{i+1}^+$ on $R_A(\boldsymbol{\mu}_i)$, and consider the next policy $\boldsymbol{\mu}_{i+1}$. Let $\underline{\alpha}_{i+1}$ and $\bar{\alpha}_{i+1}$ be of class $\mathcal{K}$ satisfying

$$\underline{\alpha}_{i+1}(\|\mathbf{x}\|) \leq V_{i+1}(\mathbf{x}) \leq \bar{\alpha}_{i+1}(\|\mathbf{x}\|).$$

In this ideal case, the explorized I-PI and integral Q-learning methods are equal to I-PI in the iteration domain, so $\boldsymbol{\mu}_{i+1}$ is admissible as long as so is $\boldsymbol{\mu}_i$. Moreover, Theorem 5.1

implies that if the exploration $\mathbf{e}$ at the $(i+1)$-th step is bounded by

$$\sup_{t \le \tau < \infty} \|\mathbf{e}(\tau)\| < \sqrt{\frac{\underline{\alpha}_s\big(r_{\boldsymbol{\mu}_i}\big)}{\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))}}, \tag{5.23}$$

where $r_{\boldsymbol{\mu}_i} > 0$ is chosen such that $\bar{\Omega}_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i}) \subset R_A(\boldsymbol{\mu}_i)$, then

1. $\mathbf{e}$ is invariant on $\bar{\Omega}_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$ under the next policy $\boldsymbol{\mu}_{i+1}$;
2. the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})(\mathbf{u}(\mathbf{x}) + \mathbf{e}_\tau)$ is input-to-state stable on $\bar{\Omega}_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$ under the next policy $\mathbf{u} = \boldsymbol{\mu}_{i+1}(\mathbf{x})$.

If $\mathcal{D} = R_A(\boldsymbol{\mu}_i) = \mathbb{R}^n$ and $S(\mathbf{x})$ is radially unbounded, then this ISS holds globally for any (bounded) exploration $\mathbf{e}_\tau$ by Theorem 5.2. Here, the condition $R_A(\boldsymbol{\mu}_i) = \mathbb{R}^n$ is approximately achived if

1. $R_A(\boldsymbol{\mu}_0) = \mathbb{R}^n$;
2. all of the policies generated by either of the IRL methods up to the i-th step are approximately equal to those generated by the ideal I-PI under the same initial admissible policy $\mathbf{u}_0 = \boldsymbol{\mu}_0$.

In case of that $\mathbf{e}$ is constructed from some constant vectors $\{\mathbf{c}_j\}_{j=1}^N$ and satisfies (5.17), the boundedness condition (5.23) is replaced by

$$\|\mathbf{c}_j\| < \sqrt{\frac{\underline{\alpha}_s\big(r_{\boldsymbol{\mu}_i}\big)}{\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))}}, \quad \forall j \in \{1, 2, \cdots, N\}. \tag{5.24}$$

Now, the remaining question is what to do when the state $\mathbf{x}_t$ is outside the invariant region $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$ but inside $R_A(\boldsymbol{\mu}_i)$ during online learning at i-th or $(i+1)$-th step. In this case, since $\mathbf{x}_t$ is outside $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$, the rare exploration $\mathbf{e}_\tau$ satisfying (5.7) is no more safe. In this particular case, the best way to preserve invariance and ISS is to apply the current policy with zero exploration $\mathbf{e} \equiv \mathbf{0}_m$ until some finite time $t' \ge t$ at which $\mathbf{x}_\tau$ enters into $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$, i.e., $\mathbf{x}_{t'} \in \Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$.[1] Then, as illustrated in Fig. 5.1,

[1] Since $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_{i+1}$ are admissible, there exists finite time $t' \in [t, \infty)$ such that $\mathbf{x}_\tau \equiv \mathbf{x}_\tau(\mathbf{z}; \boldsymbol{\mu}_i)$ for $\mathbf{z}$ on their region of attraction enters to the smaller set $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i}) \subset R_A(\boldsymbol{\mu}_i)$ at t' under the zero exploration $\mathbf{e}_\tau = \mathbf{0}_m$.

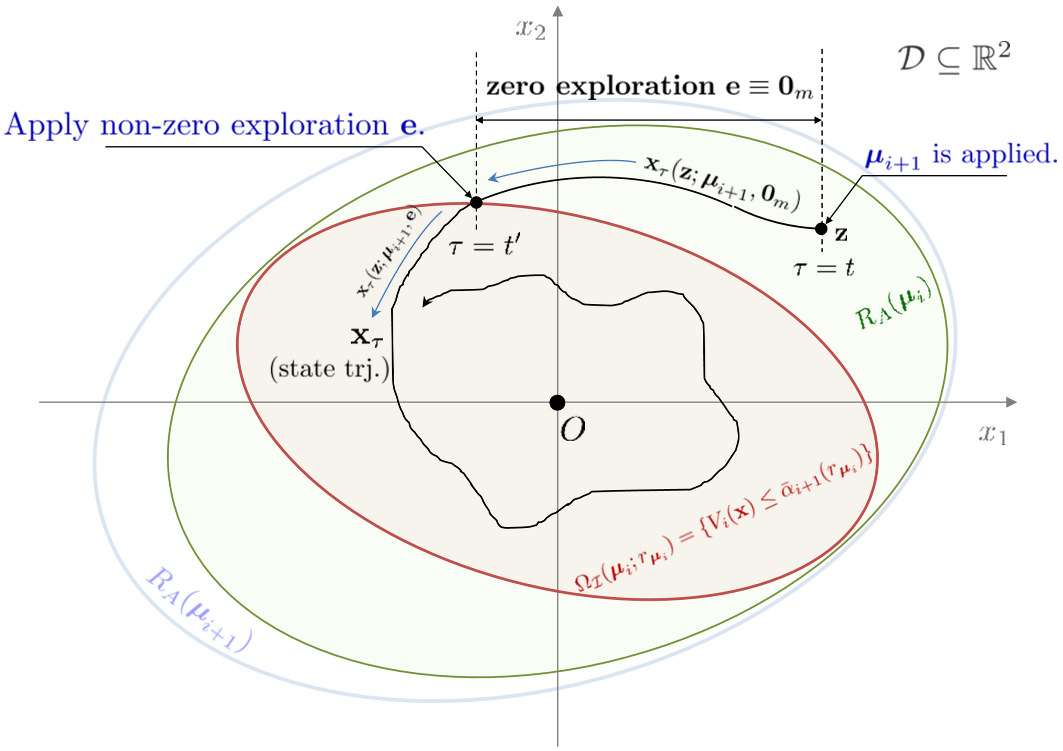

x_2
$\mathcal{D} \subseteq \mathbb{R}^2$
zero exploration $\mathbf{e} \equiv \mathbf{0}_m$
Apply non-zero exploration $\mathbf{e}$.
$\boldsymbol{\mu}_{i+1}$ is applied.
$\mathbf{x}_\tau(\mathbf{z}; \boldsymbol{\mu}_{i+1}, \mathbf{0}_m)$
$\mathbf{z}$
$\tau = t'$
$\tau = t$
$x_\tau(\mathbf{z}; \boldsymbol{\mu}_{i+1}, \mathbf{e})$
$\mathbf{x}_\tau$
(state trj.)
$R_A(\boldsymbol{\mu}_i)$
O
x_1
$\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i}) = \{V_i(\mathbf{x}) \leq \bar{\alpha}_{i+1}(r_{\boldsymbol{\mu}_i})\}$
$R_A(\boldsymbol{\mu}_{i+1})$

Figure 5.1: Switching zero-to-nonzero exploration strategy for invariance and ISS.

a non-zero exploration $\mathbf{e}$ satisfying (5.23) or (5.24) is applied thereafter which guarantees invariance and ISS on $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$. Note that in the global case ($\mathcal{D} = R_A(\boldsymbol{\mu}_i) = \mathbb{R}^n$) with radially unbounded $S(\mathbf{x})$, explorized I-PI or integral Q-learning can be performed without such consideration on the invariance and ISS. In the local case, such processes on the exploration $\mathbf{e}$ can be also removed when $\mathbf{e}$ is sufficiently small and $\mathbf{x}_\tau$ starts from a region near the origin that is small enough to be contained by $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\boldsymbol{\mu}_i})$.

5.6 Neural-Networks-Based Implementations

The partially model-free explorized I-PI (Algorithm 5.2) and the model-free integral Q-learning (Algorithms 5.3) can be implemented in the least-squares (LS) sense using neural networks (NNs) to approximate V_{i+1} and $\boldsymbol{\mu}_{i+1}$. To explain this, let $\{\phi_j^c \in C^1_{L^+}(\mathcal{D})\}_{j=1}^{\infty}$ and $\{\phi_j^a \in C^0_L(\mathcal{D}) : \phi_j^a(\mathbf{0}_n) = \mathbf{0}_m\}_{j=1}^{\infty}$ be the sequences of real-valued NN activation functions that are linearly independent and complete on their respective function spaces

$$\begin{cases} C^1_{L^+}(\mathcal{D}) = \{V \in C^1_L(\mathcal{D}) : V(\mathbf{x}) \in \mathbb{R}, V \text{ is positive definite }\}, \\ C^0_{L^0}(\mathcal{D}) := \{\mathbf{f} \in C^0_L(\mathcal{D}) : \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m, \mathbf{f}(\mathbf{0}_n) = \mathbf{0}_m\}. \end{cases}$$

Here, the superscripts 'a' and 'c' denote actor and critic, respectively. Using these activation functions, $V_{i+1} \in C^1_{L^+}(\mathcal{D})$ and a policy $\boldsymbol{\mu}_{i+1} \in C^0_{L^0}(\mathcal{D})$ are represented as

$$\begin{cases} V_{i+1}(\mathbf{x}) = \sum_{j=1}^{\infty} w_{ij}\phi^c_j(\mathbf{x}), \\ \boldsymbol{\mu}_{i+1}(\mathbf{x}) = \sum_{j=1}^{\infty} \mathbf{v}_{ij}\phi^a_j(\mathbf{x}), \end{cases} \tag{5.25}$$

respectively, where $w_{ij} \in \mathbb{R}$ and $\mathbf{v}_{ij} \in \mathbb{R}^m$ are weight vectors; we consider (N_c,N_a)-truncation of (5.25) as the NN expressions of V_{i+1} and $\boldsymbol{\mu}_{i+1}$:

$$\begin{cases} \hat{V}_{i+1}(\mathbf{x}) = \sum_{j=1}^{N_c} w_{ij}\phi^c_j(\mathbf{x}) \equiv \mathbf{w}_i^T\boldsymbol{\phi}_c(\mathbf{x}), \\ \hat{\boldsymbol{\mu}}_{i+1}(\mathbf{x}) = \sum_{j=1}^{N_a} \mathbf{v}_{ij}\phi^a_j(\mathbf{x}) \equiv \mathbf{V}_i^T\boldsymbol{\phi}_a(\mathbf{x}), \end{cases} \tag{5.26}$$

where $\begin{cases} \mathbf{w}_i := [\,w_{i1}, w_{i2}, \cdots, w_{iN_c}\,]^T \in \mathbb{R}^{N_c}, \quad \mathbf{V}_i := [\,\mathbf{v}_{i1}, \mathbf{v}_{i2}, \cdots, \mathbf{v}_{iN_a}\,]^T \in \mathbb{R}^{N_a \times m}, \\ \boldsymbol{\phi}_c(\mathbf{x}) := [\phi^c_1(\mathbf{x}), \cdots, \phi^c_{N_c}(\mathbf{x})]^T \in \mathbb{R}^{N_c}, \;\; \boldsymbol{\phi}_a(\mathbf{x}) := [\phi^a_1(\mathbf{x}), \cdots, \phi^a_{N_a}(\mathbf{x})]^T \in \mathbb{R}^{N_a}. \end{cases}$

Using these expressions, (5.25) can be rewritten as

$$\begin{cases} V_{i+1}(\mathbf{x}) = \mathbf{w}_i^T\boldsymbol{\phi}_c(\mathbf{x}) + \varepsilon^c_i(\mathbf{x}), \\ \boldsymbol{\mu}_{i+1}(\mathbf{x}) = \mathbf{V}_i^T\boldsymbol{\phi}_a(\mathbf{x}) + \varepsilon^a_i(\mathbf{x}), \end{cases} \tag{5.27}$$

where $\varepsilon^c_i(\mathbf{x})$ and $\varepsilon^a_i(\mathbf{x})$ are NN reconstruction errors. Note that if the domain is restricted to a compact subset of $R_A(\boldsymbol{\mu}_i) \subseteq \mathcal{D}$ such as $\Omega_{\mathcal{I}}(\boldsymbol{\mu}_i; r_{\mathbf{u}_i})$ that belongs to $R_A(\boldsymbol{\mu}_i)$, then there exist $N_c, N_a \in \mathbb{N}$ such that the NN errors ε^c_i and ε^a_i in (5.27) and $\nabla\varepsilon^c_i$ are all bounded on the compact set. Also note that each ϕ^c_j is positive definite for $V_{i+1} \succ 0$ and that $\boldsymbol{\phi}_a(\mathbf{0}_n) = \mathbf{0}_m$ for $\boldsymbol{\mu}_{i+1}(\mathbf{0}_n) = \mathbf{0}_m$.

Now, consider the LS implementation of integral Q-learning (Algorithm 5.3). In this case, substituting (5.27) into the advanced I-TD equation (5.22), we obtain

$$\delta_i(\mathbf{x}_t, \mathbf{e}) = \big[\boldsymbol{\phi}_c(\mathbf{x}_{t+T_s}) - \boldsymbol{\phi}_c(x_t)\big]^T \mathbf{w}_i + \int_t^{t+T_s} \Big[r(\mathbf{x}, \boldsymbol{\mu}_i(\mathbf{x})) + 2\boldsymbol{\phi}_a^T(\mathbf{x})\mathbf{V}_i R \mathbf{e}_\tau \Big] d\tau, \quad (5.28)$$

where $\delta_i(\mathbf{x}_t, \mathbf{e}) \in \mathbb{R}$ is the advanced I-TD error given by

$$\delta_i(\mathbf{x}_t, \mathbf{e}) = \varepsilon_i^c(\mathbf{x}_t) - \varepsilon_i^c(\mathbf{x}_{t+T_s}) - 2\int_t^{t+T_s} (\varepsilon_i^a(\mathbf{x}))^T \mathbf{R}(\mathbf{x})\mathbf{e}_\tau \, d\tau.$$

Define $\mathbf{v}_i \in \mathbb{R}^{N_a m}$ as $\mathbf{v}_i := \mathbf{col}\{\mathbf{V}_i\}$. Then, applying $\boldsymbol{\phi}_a^T(\mathbf{x})\mathbf{V}_i\mathbf{R}(\mathbf{x})\mathbf{e} = \big(\mathbf{R}(\mathbf{x})\mathbf{e} \otimes \boldsymbol{\phi}_a(\mathbf{x})\big)^T \mathbf{v}_i$ to (5.28) and then rearranging the equation, we obtain the following expression regarding (5.22):

$$\delta_i(\mathbf{x}_t; \mathbf{e}) = \boldsymbol{\psi}^T(\mathbf{x}_t; \mathbf{e}) \cdot \boldsymbol{\theta}_i + Z(\mathbf{x}_t; \boldsymbol{\mu}_i), \quad (5.29)$$

where $\boldsymbol{\theta}_i = \mathbf{col}\{\mathbf{w}_i, \mathbf{v}_i\}$ is the vector of unknown weights; $\boldsymbol{\psi}(\mathbf{x}_t; \mathbf{e})$ and $Z(\mathbf{x}_t; \boldsymbol{\mu}_i)$ are given in Table 5.1. The advanced I-TD equation (5.21) can be also formulated as (5.29) with $\boldsymbol{\theta}_i$, $\boldsymbol{\psi}(\mathbf{x}_t; \mathbf{e})$, and $Z(\mathbf{x}_t; \boldsymbol{\mu}_i)$ given in Table 5.1. Here, the advanced I-TD error $\delta_i(\mathbf{x}_t; \mathbf{e})$ for the explorized I-PI is omitted, but can be easily obtained by the similar procedure. In policy improvement of the explorized I-PI, the next neuro-policy $\hat{\boldsymbol{\mu}}_{i+1}$ can be updated by

$$\hat{\boldsymbol{\mu}}_{i+1}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}(\mathbf{x})\mathbf{G}^T(\mathbf{x})\nabla\boldsymbol{\phi}_c(\mathbf{x})\mathbf{w}_i \quad (5.30)$$

using $\hat{V}_{i+1}$ in (5.26), instead of V_{i+1}, as was done in [50].

Let N_θ be the number of elements of $\boldsymbol{\theta}_i$, *e.g.*, $N_\theta = N_c + N_a$ for (5.22). Then, we have N_θ unknowns in the 1-dimensional equation (5.29). In the implementations, $\boldsymbol{\theta}_i$ will be uniquely determined in LS sense. Define $\boldsymbol{\psi}[k]$, $\delta_i[k]$, and $Z[k]$ as

$$\begin{cases} \boldsymbol{\psi}[k] := \boldsymbol{\psi}(x_{t+kT}, \mathbf{e}), \\ \delta_i[k] := \delta_i(x_{t+kT}, \mathbf{e}), \\ Z[k] := Z(x_{t+kT}, \boldsymbol{\mu}_i). \end{cases}$$

Table 5.1: Functions and vectors of I-TD error equation (5.29) for online NN Implementations of explorized I-PI and integral Q-learning

Alg. No.	Process Type	Functions and Vectors in (5.29)
5.2	Policy Evaluation	$\boldsymbol{\theta}_i = \mathbf{w}_i$ $\boldsymbol{\psi}(\mathbf{x};\mathbf{e}) = \boldsymbol{\phi}_c(\mathbf{x}_{t+T_s}) - \boldsymbol{\phi}_c(\mathbf{x}_t) - \int_t^{t+T_s} \nabla^T \boldsymbol{\phi}_c(\mathbf{x}) \cdot \mathbf{G}(\mathbf{x})\mathbf{e}_\tau \, d\tau$ $Z(\mathbf{x}_t;\boldsymbol{\mu}_i) = \int_t^{t+T_s} r(\mathbf{x},\boldsymbol{\mu}_i(\mathbf{x}))\, d\tau$
5.3	Policy Evaluation & Improvement	$\boldsymbol{\theta}_i = \mathbf{col}\{\mathbf{w}_i, \mathbf{v}_i\}$ $\boldsymbol{\psi}(\mathbf{x};\mathbf{e}) = \mathbf{col}\Big\{ \boldsymbol{\phi}_c(\mathbf{x}_{t+T_s}) - \boldsymbol{\phi}_c(\mathbf{x}_t), \int_t^{t+T_s} 2\,\boldsymbol{\phi}_a(\mathbf{x}) \otimes (\mathbf{R}(\mathbf{x})\mathbf{e}_\tau)\, d\tau \Big\}$ $Z(\mathbf{x}_t;\boldsymbol{\mu}_i) = \int_t^{t+T_s} r(\mathbf{x},\boldsymbol{\mu}_i(\mathbf{x}))\, d\tau$

Then, referring $\mathbf{x}_{t+(k-1)T_s}$ as a starting point of the advanced I-TDs, the following generalized I-TD error equation can be derived from (5.29):

$$\delta_i[k] = \boldsymbol{\psi}^T[k] \cdot \boldsymbol{\theta}_i + Z[k], \tag{5.31}$$

which holds for any $k \in \mathbb{N}$ since $\mathbf{x}_\tau(\mathbf{z};\boldsymbol{\mu}_i,\mathbf{e})$ remains in an invariant region for all $\tau \geq t$ by the well-designed exloration $\mathbf{e}_\tau$. Suppose the data $(\boldsymbol{\psi}[k], Z[k])$ for $k = 1, 2, \cdots, N$ are all available, and define the LS error E to be minimized as $E^2 := \frac{1}{2}\sum_{k=1}^N \delta_i^2[k]$. Then, differentiating E^2 in terms of $\boldsymbol{\theta}_i$ with the substitution of (5.31) yields

$$\frac{\partial E^2}{\partial \boldsymbol{\theta}_i} = \sum_{k=1}^N \frac{\partial \delta_i[k]}{\partial \boldsymbol{\theta}_i}\delta_i[k] = \sum_{k=1}^N \Big[\boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k] \cdot \boldsymbol{\theta}_i + \boldsymbol{\psi}[k] Z[k] \Big].$$

Equating this to zero and rearranging the equation, we obtain the LS solution of the form

$$\boldsymbol{\theta}_{i,LS} = -\Big(\sum_{k=1}^N \boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k] \Big)^{-1} \Big(\sum_{k=1}^N \boldsymbol{\psi}[k] Z[k] \Big). \tag{5.32}$$

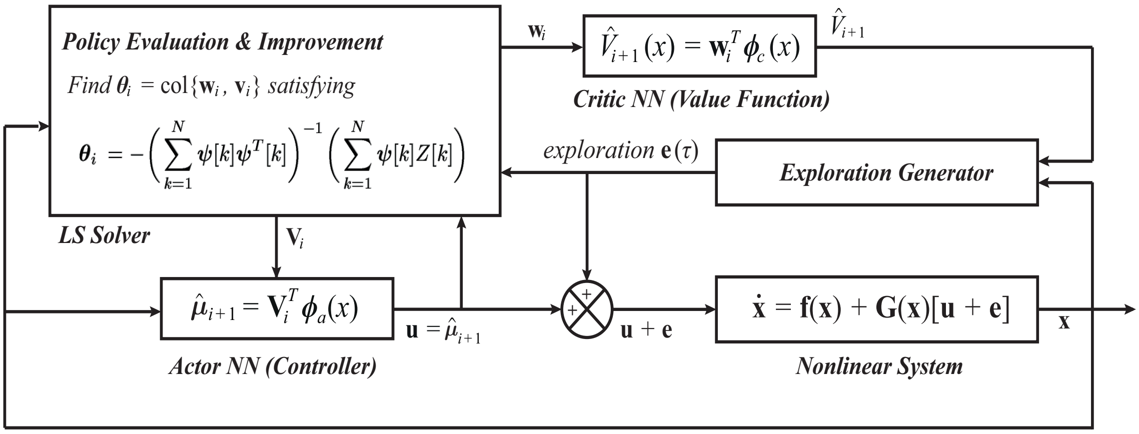
For the existence of the inverse in (5.32), we need the following excitation condition:

Assumption 5.3. *There exist κ_5, $\kappa_6 > 0$ such that*

$$\kappa_5 I \leq \sum_{k=1}^{N} \boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k] \leq \kappa_6 I.$$

Similar to Assumption 5.2, $N \geq L_\theta$ is necessary to satisfy Assumption 5.3, so at least L_θ-number of data should be collected to obtain the LS solution (5.32) at each iteration.

The whole control scheme for integral Q-learning and its LS implementation is demonstrated in Fig. 5.2. At each iteration, the LS solver collects the data needed to calculate $\boldsymbol{\psi}[k]$, $\delta_i[k]$, and $Z[k]$ for $k = 1, 2, \cdots, N$, and then finds the weight vectors $\mathbf{w}_i$ and $\mathbf{v}_i$ satisfying (5.32), both of which are transferred to the corresponding actor and critic NNs to update their weights. Here, the actor NN generates the control input; the output of the critic NN $\hat{V}_{i+1}(\mathbf{x})$ is used in the exploration generator module to calculate the bound (5.23) on the exploration $\mathbf{e}_\tau$. In exploration generator, the exploration $\mathbf{e}_\tau$ is constructed, and modified if necessary, that plays a key role in exciting the signal $\boldsymbol{\psi}(\mathbf{x}_t; \mathbf{e})$ in (5.32), and satisfies i) Assumption 5.2 (or (5.20)) and ii) the boundedness condition (5.23) for ISS and invariance of $\mathbf{e}$. The whole control scheme with explorized I-PI can be described in a similar manner by modifying LS solver and actor NN blocks.



Policy Evaluation & Improvement
Find $\theta_i = \text{col}\{\mathbf{w}_i, \mathbf{v}_i\}$ satisfying
$\theta_i = -\left(\sum_{k=1}^{N}\psi[k]\psi^T[k]\right)^{-1}\left(\sum_{k=1}^{N}\psi[k]Z[k]\right)$
LS Solver
$\mathbf{w}_i$
$\hat{V}_{i+1}(x) = \mathbf{w}_i^T\phi_c(x)$
$\hat{V}_{i+1}$
Critic NN (Value Function)
exploration $\mathbf{e}(\tau)$
Exploration Generator
$\mathbf{V}_i$
$\hat{\mu}_{i+1} = \mathbf{V}_i^T\phi_a(x)$
$\mathbf{u} = \hat{\mu}_{i+1}$
Actor NN (Controller)
+
+
$\mathbf{u} + \mathbf{e}$
$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})[\mathbf{u} + \mathbf{e}]$
$\mathbf{x}$
Nonlinear System

Figure 5.2: The whole control scheme with integral Q-learning (Algorithm 5.3) and its LS implementation.

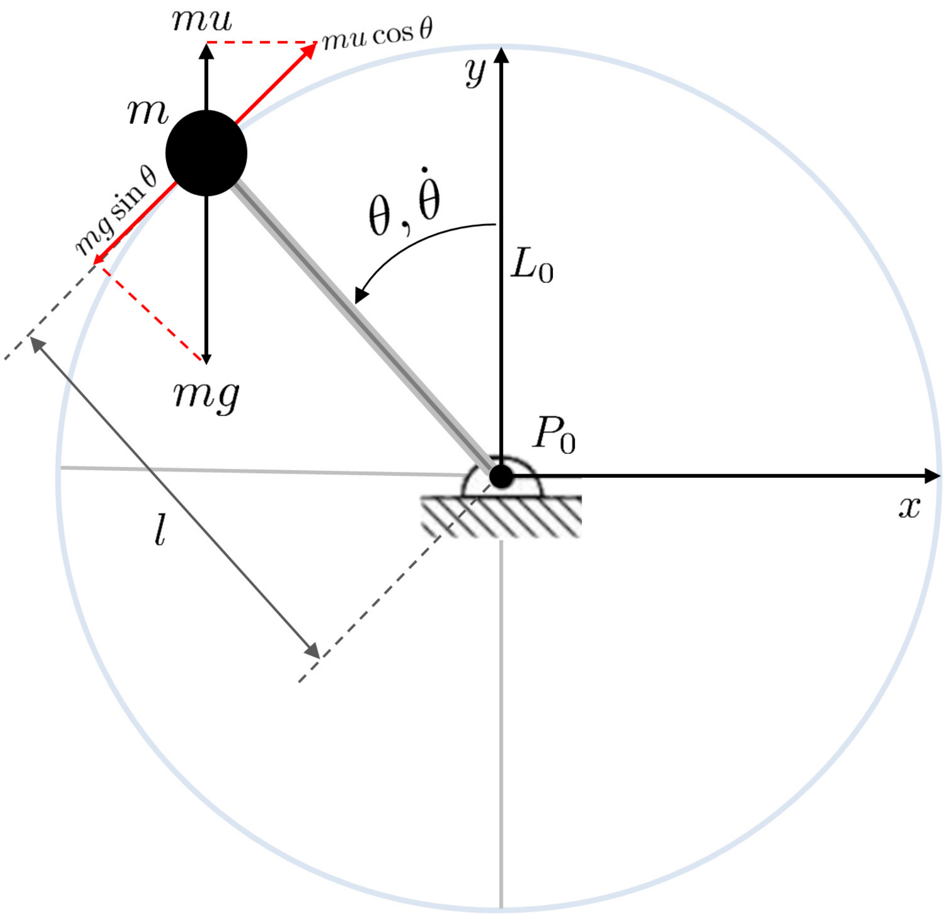


Figure 5.3: The free body diagram of the swing-up pendulum.

5.7 Numerical Simulations: Pendulum Examples

In this section, the integral Q-learning (Algorithm 5.3) is simulated to verify and discuss its performance. The simulation studies also show the practical remarks on the integral Q-learning method. In the simulations, we consider the adaptive optimal stabilization problem of the swing-up pendulum [38, 62] illustrated in Fig. 5.3 and modeled by

$$J\ddot{\theta} - mgl\sin\theta + mgu\cos\theta = 0, \tag{5.33}$$

where m is the mass of the pendulum, J is the moment of inertia with respect to the pivot point P_0, l is the distance from the pivot to the center of mass, θ is the angle from the vertical line L_0 to the pendulum in counter-clockwise direction, and g is the gravity constant. These parameters are given by $m = 1$ [kg], $l = 1$ [m], $J = 1$ [kg·m^2], and $g = 9.81$ [kg·m/s^2], but assumed completely unknown in the simulations. Defining the state variable $\mathbf{x} = [\,x_1,\ x_2\,]^T$ as $x_1 := \theta$ and $x_2 := \dot{\theta}$, the pendulum dynamics (5.33) can be rewritten in a form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})u$, where

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ J^{-1}mgl\sin x_1 \end{bmatrix} \text{ and } \mathbf{g}(\mathbf{x}) = \begin{bmatrix} 0 \\ -J^{-1}ml\cos x_1 \end{bmatrix}. \tag{5.34}$$

Note that the pendulum is locally controllable, i.e., its linearization near the equilibrium $\mathbf{x} = \mathbf{0}_2$:

$$\delta\dot{\mathbf{x}} = \mathbf{A}\delta\mathbf{x} + \mathbf{B}\delta u \tag{5.35}$$

is controllable, where $\delta\mathbf{x}$ and δu are small perturbations from $\mathbf{x} = \mathbf{0}_2$ and $u = 0$, and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and $\mathbf{B} \in \mathbb{R}^2$ are unknown matrices given by

$$\mathbf{A} := \nabla\mathbf{f}^T(\mathbf{x})|_{\mathbf{x}=\mathbf{0}_2} = \begin{bmatrix} 0 & 1 \\ mgl/J & 0 \end{bmatrix} \text{ and } \mathbf{B} := \nabla\mathbf{g}(\mathbf{x})|_{\mathbf{x}=\mathbf{0}_2} = \begin{bmatrix} 0 \\ ml/J \end{bmatrix},$$

which is obtained by approximating $\sin\theta \approx \theta$ and $\cos\theta \approx 1$ near $\theta = 0$. In the simulations, it is assumed that *the functions* $\mathbf{f}(\mathbf{x})$ *and* $\mathbf{g}(\mathbf{x})$ *are completely unknown*, but an initial admissible linear policy $\boldsymbol{\mu}_0(\mathbf{x}) = 30(x_1 + x_2)$ is given *a priori*; the sampling period $T_s > 0$ and the number of data N collected per iteration are set to $T_s = 50$ [ms] and $N = 50$, so that the LS solution $\theta_{i,LS} \in \mathbb{R}^5$ for $\theta = \mathbf{col}\{\mathbf{w}, \mathbf{v}\}$ is calculated by (5.32) every 2.25 [s]. The initial condition is given by $\mathbf{z} = [\,0.01,\ 0\,]^T$ to make the pendulum initially at a upright position, which is necessary since the initial policy guarantees asymptotic stability *only* in a local region of the equilibrium $\mathbf{x} = \mathbf{0}_2$.

5.7.1 Example 1: Adaptive Linear Quadratic Regulators

As a motivating example, we first apply the integral Q-learning to the linearized pendulum model (5.35) with the matrices $\mathbf{A}$ and $\mathbf{B}$ completely unknown and the cost

$$r(\mathbf{x}, u) = 10x_1^2 + 10x_2^2 + u^2.$$

In this LQR case, the policy u is linear in $\mathbf{x}$, and the value function is quadratic in $\mathbf{x}$ as shown in Chapter 4. So, we choose the activation functions in the critic and actor NNs $\hat{V}_{i+1}(\mathbf{x}) = \mathbf{w}_i^T \boldsymbol{\phi}_c(\mathbf{x})$ and $\hat{\mu}_{i+1}(\mathbf{x}) = \mathbf{v}_i^T \boldsymbol{\phi}_a(\mathbf{x})$ as

$$\boldsymbol{\phi}_c(\mathbf{x}) = \mathbf{col}\{x_1^2, x_1 x_2, x_2^2\} \text{ and } \boldsymbol{\phi}_a(\mathbf{x}) = \mathbf{col}\{x_1, x_2\}.$$

In this LQR case, if the system matrices $\mathbf{A}$ and $\mathbf{B}$ are known *a priori*, then the optimal parameters $\mathbf{w}^*$ and $\mathbf{v}^*$ can be obtained from the solution of the ARE as

$$\mathbf{w}^* = [\,70.0526,\ \ 40.2342,\ \ 7.0876\,]^T \text{ and } \mathbf{v}^* = [\,20.1171,\ \ 7.0876\,]^T. \tag{5.36}$$

Throughout the LQR learning, the exploration $e(\tau)$ described as

$$e(\tau) = \begin{cases} c \text{ for } \tau \in [t, t + NT_s/2) \\ -c \text{ for } \tau \in [t + NT_s/2, Nt + T_s) \end{cases} \tag{5.37}$$

for some $c > 0$ is applied until 22.5 [s] and then eliminated thereafter to see the convergence at the end.

LQR Simulations with Linearized and Nonlinear Pedulum Models ($c = 4$)

Figs. 5.4 and 5.5 shows the evolution of the NN weights and the state trajectories for $c = 4$, where we also plotted the simulation results of integral Q-learning applied to the nonlinear pendulum models under the same conditions and same LQ parameterizations. As can be seen from Fig. 5.4, the final weights

$$\mathbf{w}_f^{(1)} = [\,70.0526,\ \ 40.2342,\ \ 7.0876\,]^T \text{ and } \mathbf{v}_f^{(1)} = [\,20.1171,\ \ 7.0876\,]^T \tag{5.38}$$

for the LQR linear model case exactly matches with the optimal weights (5.36), which shows the convergence of interal Q-learning to the optimal solution. On the other hand, the nonlinear case exhibit weight convergence to a point slightly different from $\mathbf{w}^*$ and $\mathbf{v}^*$

as shown in Fig. 5.4; moreover, Fig. 5.5 describes that the nonlinear effects cause the state trajectories less perturbed from the origin $\mathbf{x} = \mathbf{0}_2$ than the case of linear pendulum model. In Fig. 5.4, the marked points indicate the time instant the next LS solution $\boldsymbol{\theta}_{i+1,LS}$ is calculated, and the oscillations were introduced due to the periodic exploration for online learning.

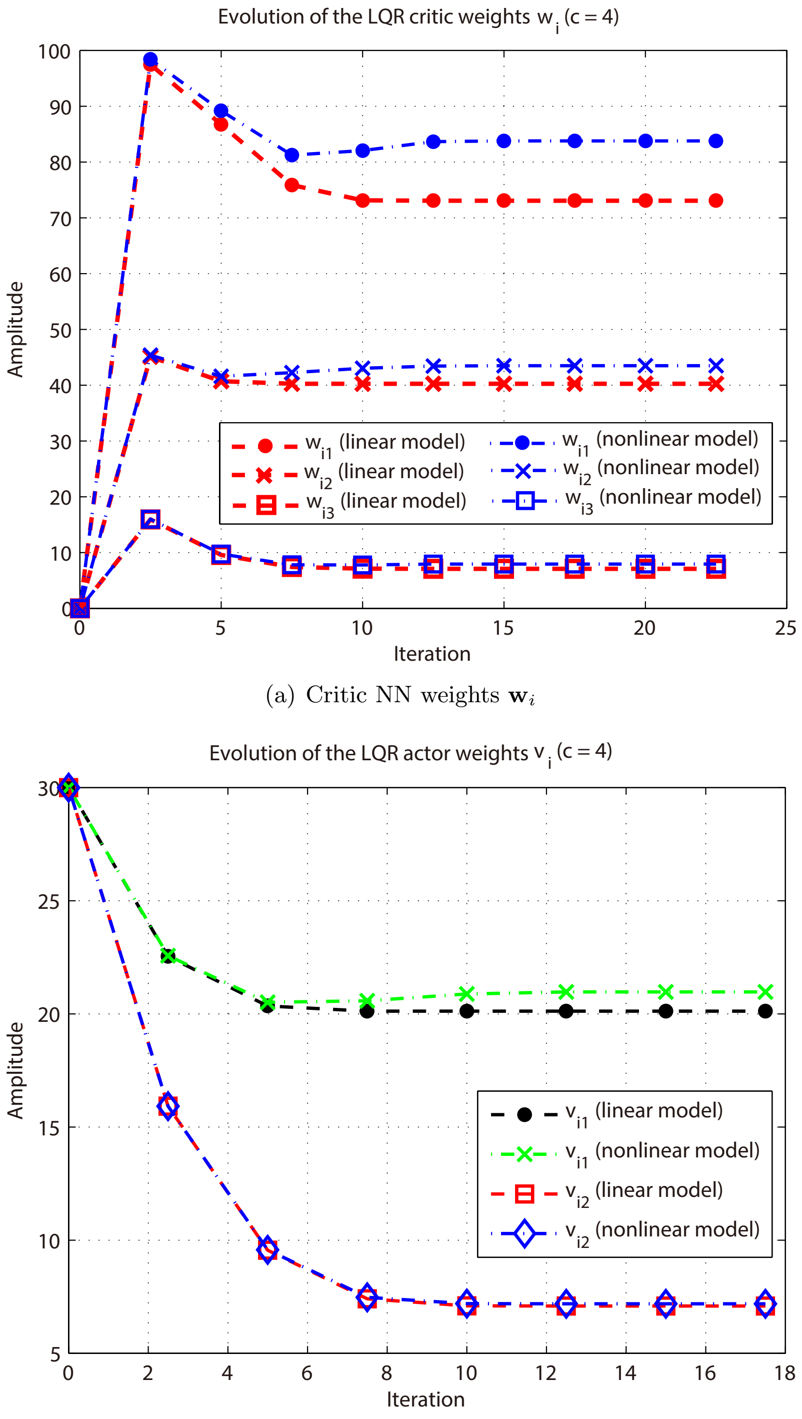
Discussions and Comparisons with the Nonlinear Pendulum Models ($c = 10$)

The parameter convergence to a point other than the LQR solution (5.36) in the nonlinear case ($c = 4$) in Fig. 5.4 may be the result of the compensation of the nonlinearities done by the RL agent to improve the control performance at the points other than the zero equilibrium $\mathbf{x} = \mathbf{0}_2$. This can be verified by comparing the performance of the final policies consisting of their stationary actor weights obtained by the integral Q-learning iteration at the ends. For this comparison, we have applied integral Q-learning with $c = 10$ to the nonlinear dynamics (5.33) and as a result obtained the final critic and actor weights as follows:

$$\mathbf{w}_f^{(2)} = [\,70.0526,\ \ 40.2342,\ \ 7.0876\,]^T \text{ and } \mathbf{v}_f^{(2)} = [\,25.0257,\ \ 7.5505\,]^T. \tag{5.39}$$

The final policies with the final actor weights $\mathbf{v}_f^{(1)}$ and $\mathbf{v}_f^{(2)}$ given in (5.38) and (5.39) are simulated for the initial condition $\theta_0 = \pi/2.5$ and $\dot{\theta}_0 = 0$ that is close to the horizontal (uncontrollable) position $\theta = \pi/2$ and far from the upright position $\theta = 0$. So, the nonlinear terms are highly excited at the beginning of the simulations.

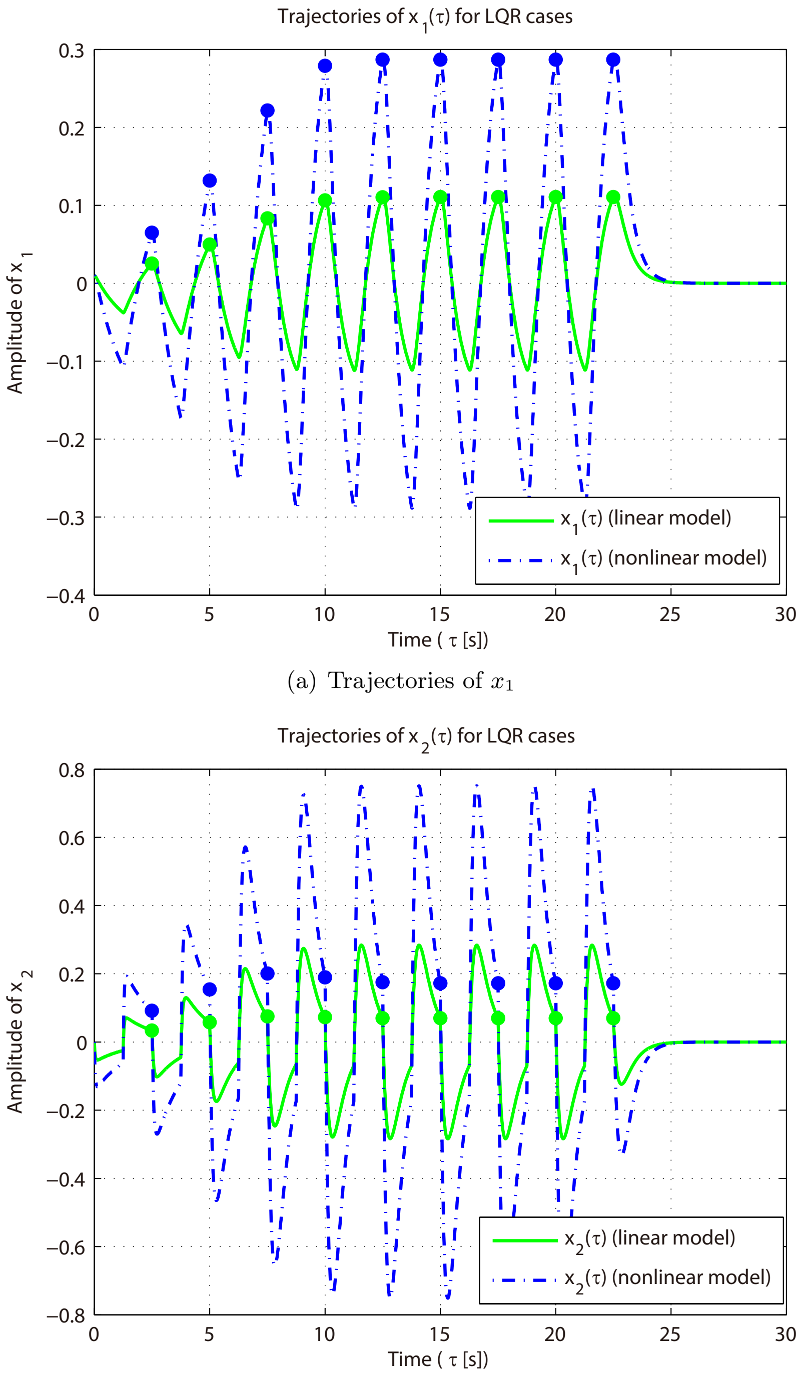
The state trajectories under the two final policies are shown in Fig. 5.6. While the policy trained with the linear model (5.35) under small exploration $c = 4$ fails to stabilize the pendulum to the upright position (Fig. 5.6(b)), the one trained with the nonlinear full dynamics (5.33) under the relatively large exploration $c = 10$ effectively control the pendulum to achieve the goal (Fig. 5.6(a)). From this observation, we can see that though the policy of the latter case is not optimal in a vicinity of the equilibrium, it is robust with



(a) Critic NN weights $\mathbf{w}_i$

(b) Actor NN weights $\mathbf{v}_i$

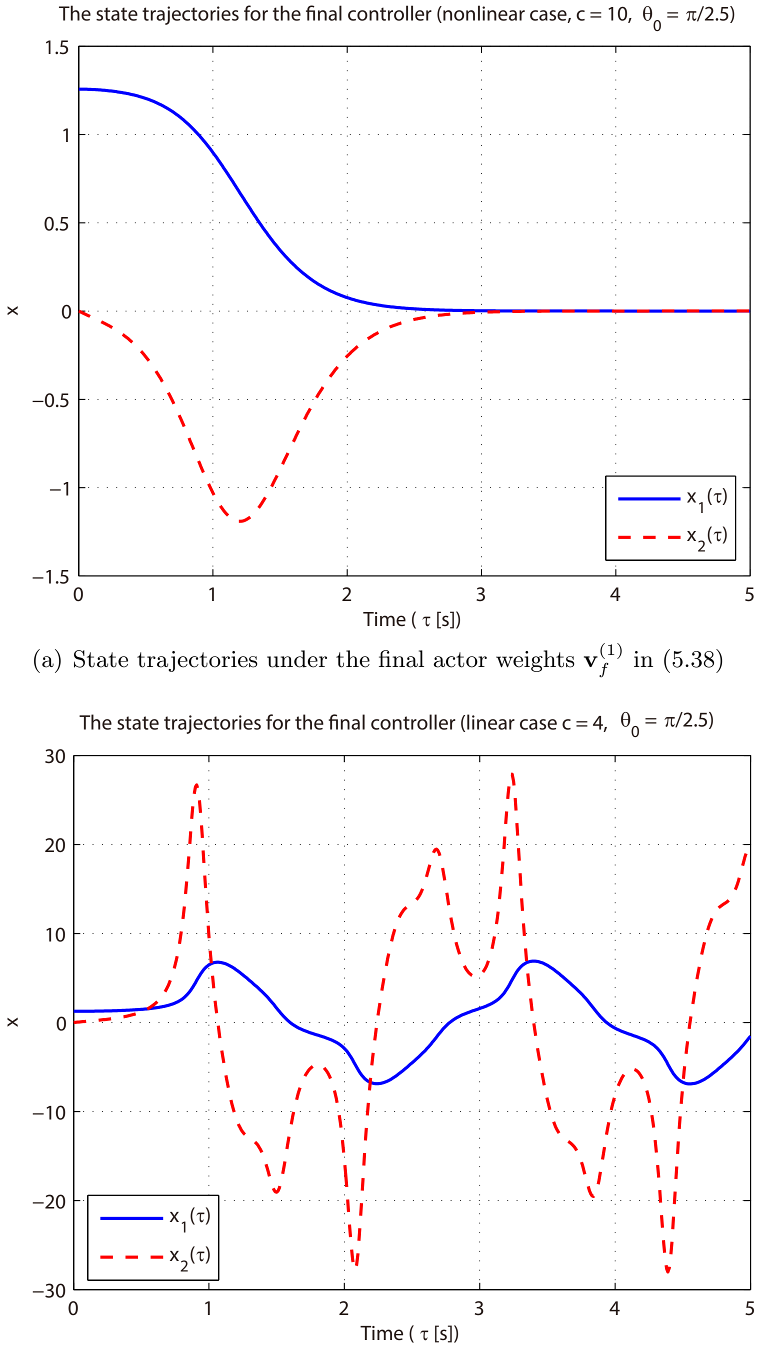
Figure 5.4: **(LQR examples: evolution of actor-critic NN weights)** Simulation results for the linear model (5.35) and the nonlinear dynamics (5.33) under $c = 4$ and the LQ structure of the policy and value function. The RL agent tunes the weights in nonlinear case slightly different from the linear one to compensate the nonlinear effects.



(a) Trajectories of x_1

(b) Trajectories of x_2

Figure 5.5: **(LQR examples: state trajectories)** Simulation results for the linear model (5.35) and the nonlinear dynamics (5.33) under $c = 4$ and the LQ structure of the policy and value function. For the same conditions, the states with the nonlinear model are less perturbed from the origin due to the nonlinearities.



(a) State trajectories under the final actor weights $\mathbf{v}_f^{(1)}$ in (5.38)

(b) State trajectories under the final actor weights $\mathbf{v}_f^{(2)}$ in 5.39

Figure 5.6: **(LQR examples: state trajectories in a nonlinear region)** State trajectories for the final linear policies with their weights $\mathbf{v}_f^{(1)}$ in (5.38) and $\mathbf{v}_f^{(2)}$ in (5.39), trained with the linear model (5.35) under $c = 4$ and the nonlinear dynamics (5.33) under $c = 10$, respectively. Though both weights are tuned with the same actor-critic NN structures, the latter explored the nonlinear regions during the learning period, which finally gives the robustness to nonlinearities in the pendulum.

respect to the nonlinearities it experienced during the learning period. On the other hand, since the nonlinear optimal control problem can be approximated by an LQR problem near the equilibrium, if the exploration is sufficiently small so the RL agent trains the actor only in a vicinity of $\mathbf{x} = \mathbf{0}_2$, then we can expect that the weights converge to a point close to the LQR solution (5.36) (see Fig. 5.4 as an example).

5.7.2 Example 2: Reinforcement Learning of Nonlinear Optimal Control

In this case, we consider the integral Q-learning applied to the nonlinear pendulum dynamics (5.33) with completely unknown functions $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ in (5.34) and the cost

$$r(\mathbf{x}, u) = 10x_1^4 + 10x_2^4 + 10x_1^2 + 10x_2^2 + u^2,$$

which is the 4th-order terms plus the same quadratic cost in the previous simulation. To approximate the high-order terms caused by nonlinearities in the pendulum dynamics and the terms x_1^4 and x_2^4 in the cost, the activation functions of the critic NN $\hat{V}_{i+1}(\mathbf{x}) = \mathbf{w}_i^T \boldsymbol{\phi}_c(\mathbf{x})$ were chosen as

$$\boldsymbol{\phi}_c(\mathbf{x}) = \mathbf{col}\{\underbrace{x_1^2,\, x_1x_2,\, x_2^2,}_{\text{2nd-order terms}}\, \underbrace{x_1^4,\, x_1^3x_2,\, x_1^2x_2^2,\, x_1x_2^3,\, x_2^4}_{\text{4th-order terms}}\} \in \mathbb{R}^8$$

To determine the actor NN structure, we assume it is known *a priori* that the first component of $\mathbf{g}(\mathbf{x})$ is zero, and that the second component is non-zero near the equilibrium $\mathbf{x} = \mathbf{0}_2$ by the local controllability of the pendulum. Then, we can see that under the approximation $\mathbf{g}(\mathbf{x}) \approx [\,0\ \ \beta\,]^T$ near the origin $\mathbf{x} = \mathbf{0}_2$ for some unknown $\beta \in \mathbb{R}$, we have

$$\begin{aligned}\mu_{i+1}(\mathbf{x}) = -\mathbf{g}^T(\mathbf{x})\nabla V_{i+1}(\mathbf{x}) &\approx -[\,0\ \ \beta\,] \cdot \nabla^T \boldsymbol{\phi}_c(\mathbf{x})\mathbf{w}_c = -\beta \cdot \nabla_{x_2}^T \boldsymbol{\phi}_c(\mathbf{x})\mathbf{w}_c. \\ &= -\beta \cdot \mathbf{w}_c^T\, \mathbf{col}\{0,\, x_1,\, 2x_2,\, 0,\, x_1^3,\, 2x_1^2x_2,\, 3x_1x_2^2,\, 4x_2^3\}.\end{aligned}$$

So, the activation function of the actor NN $\hat{\mu}_{i+1}(\mathbf{x}) = \mathbf{v}_i^T \boldsymbol{\phi}_a(\mathbf{x})$ can be chosen as

$$\boldsymbol{\phi}_a(\mathbf{x}) = \mathbf{col}\{x_1,\, x_2, x_1^3,\, x_1^2x_2,\, x_1x_2^2,\, x_2^3\},$$

which consists of all 1st- and 3rd-order terms of x_1 and x_2; the zero functions in $\nabla_{x_2}\boldsymbol{\phi}_c(\mathbf{x})$ are removed in the set of actor activations.

The the weight parameters and state/input trajectories under the biased exploration applied until $t = 22.5$ [s] and defined by

$$e(\tau) = \begin{cases} c \text{ for } \tau \in [t, t + NT_s/2) \\ 0 \text{ for } \tau \in [t + NT_s/2, Nt + T_s) \end{cases} \tag{5.40}$$

for $c = 4$ are shown in Figs. 5.7, 5.8, and 5.9. As shown in Figs. 5.7 and 5.8, all of the weights in the actor and critic NNs converge to a possibly near-optimal value after 12.5 [s]. Especially, the trajectories of the weights corresponding to the quadratic activation functions of the critic NN and the linear activations of the actor NN are almost same to

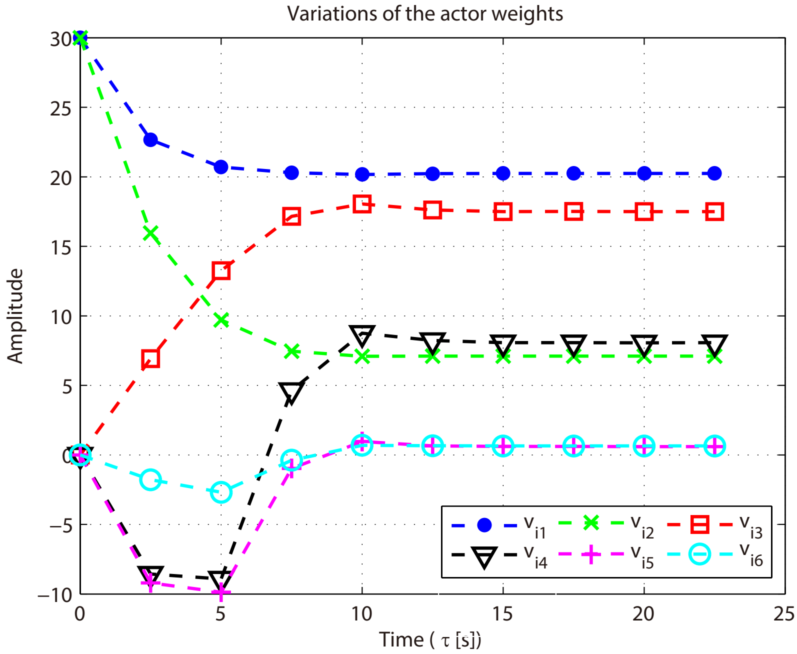
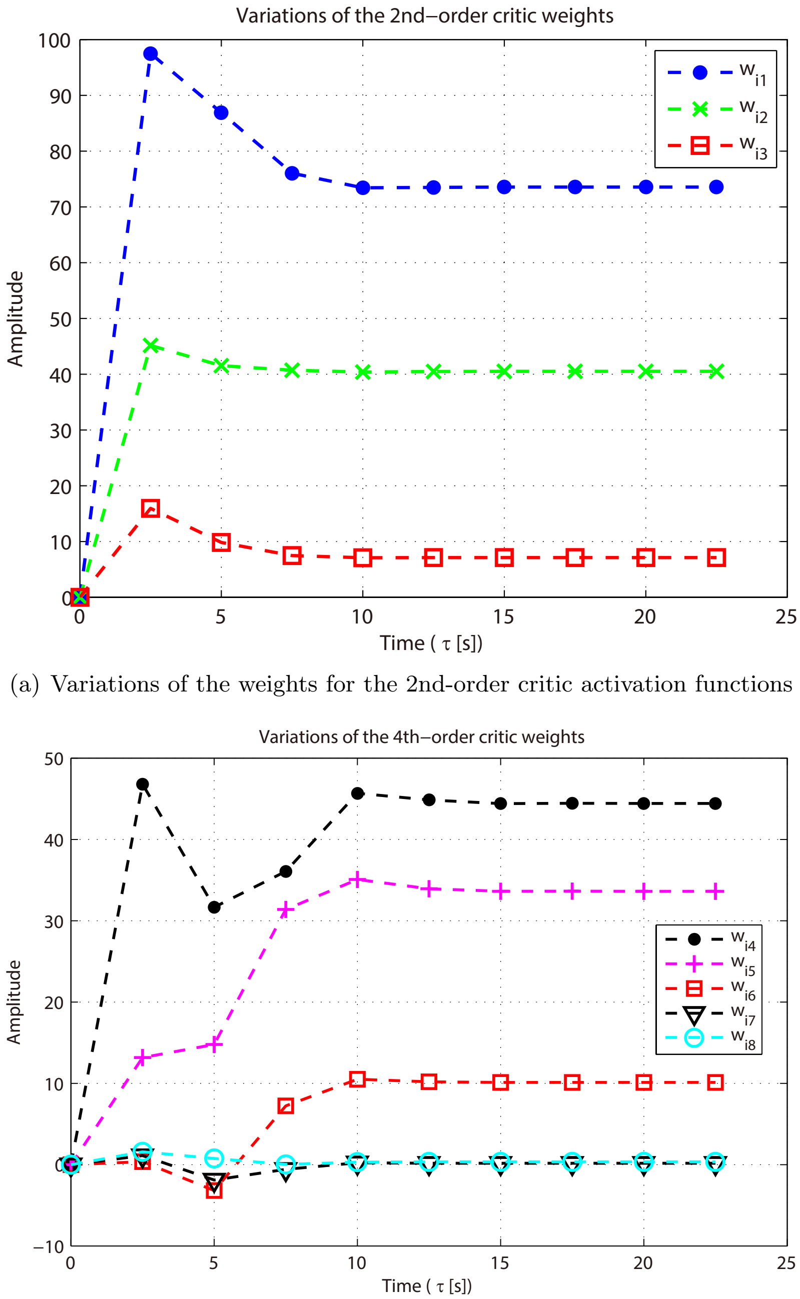


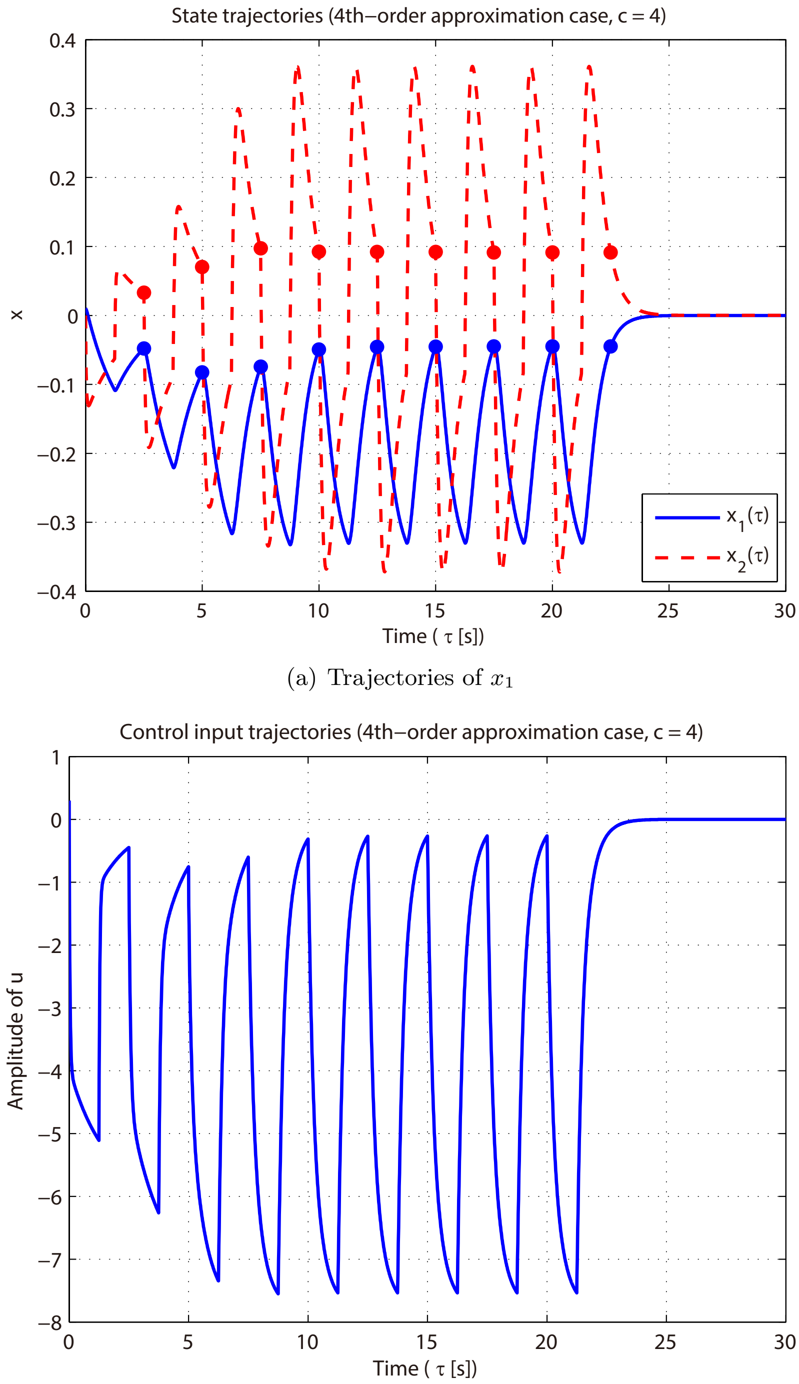
Figure 5.7: **(RL of nonlinear optimal control)** Variations of the actor NN weights.



(a) Variations of the weights for the 2nd-order critic activation functions

(b) Variations of the weights for the 4th-order critic activation functions

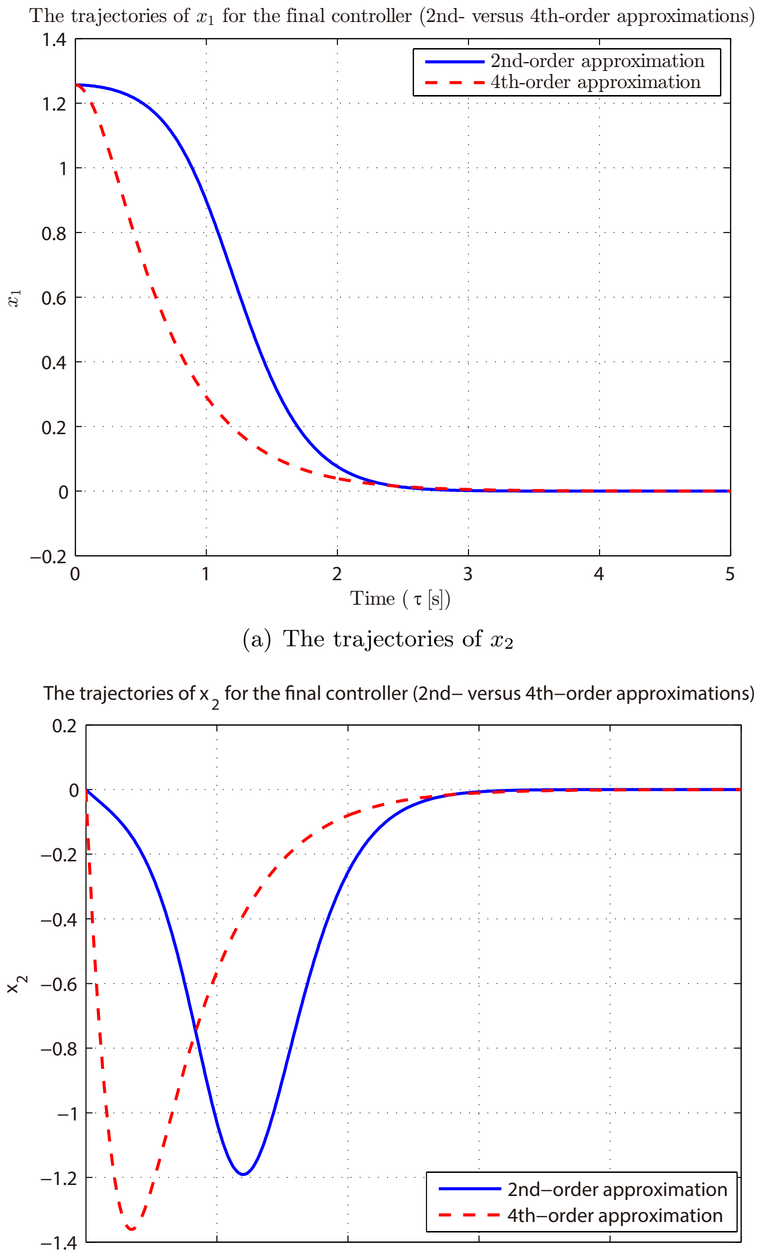
Figure 5.8: **(RL of nonlinear optimal control)** Variations of the critic NN weights: (a) the critic NN weights corresponding to the quadratic activation functions; (b) the critic NN weights corresponding to the 4th-order activation functions.



(a) Trajectories of x_1

(b) Trajectories of x_2

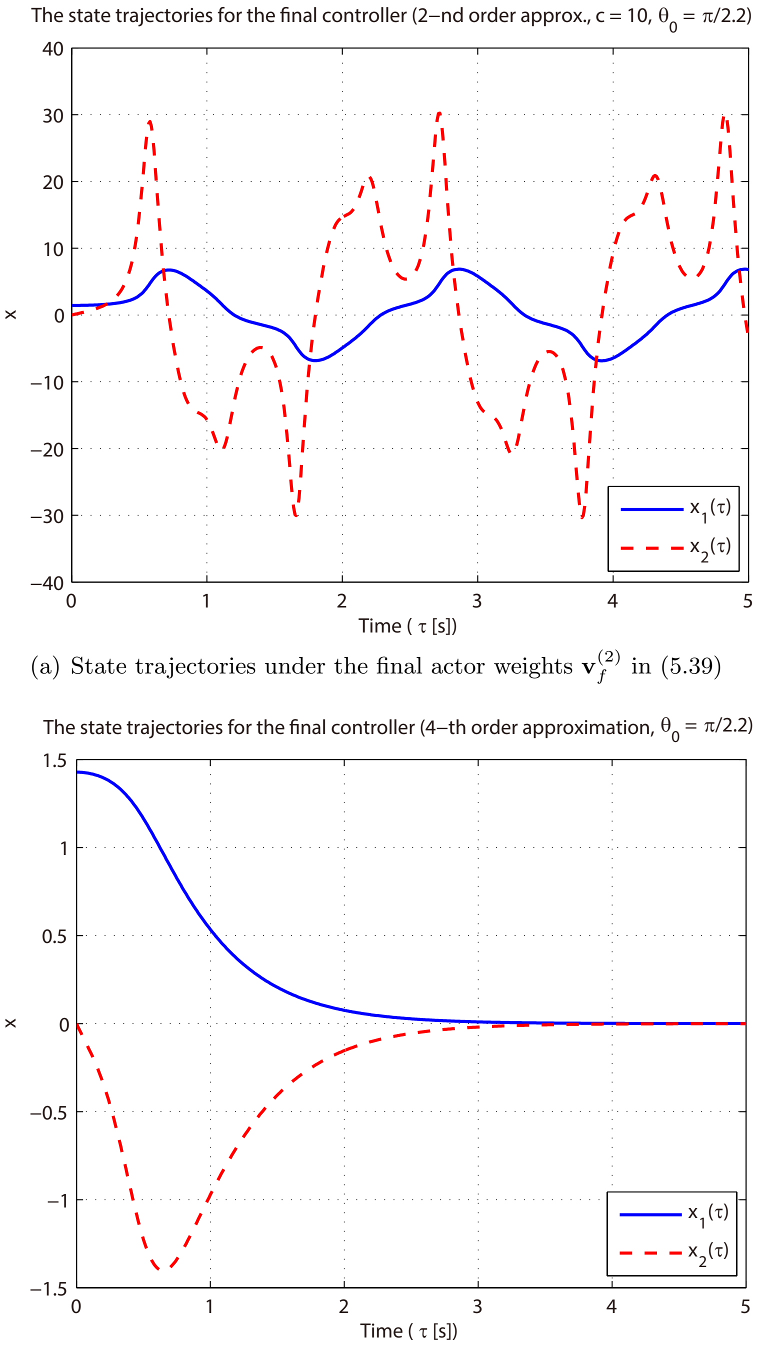
Figure 5.9: **(RL of nonlinear optimal control)** State and control input trajectories.



(a) The trajectories of x_2

(b) The trajectories of x_2

Figure 5.10: **(RL of nonlinear optimal control)** State trajectories for the initial condition $\theta = \pi/2.5$ and $\dot{\theta} = 0$ under the final policies with their weights $\mathbf{v}_f^{(2)}$ (2nd-order approximation) and $\mathbf{v}_f^{(3)}$ (4th-order approximation), both trained with the nonlinear dynamics (5.33) under $c = 10$ and $c = 4$, respectively.



(a) State trajectories under the final actor weights $\mathbf{v}_f^{(2)}$ in (5.39)

(b) State trajectories under the final actor weights $\mathbf{v}_f^{(3)}$

Figure 5.11: **(RL of nonlinear optimal control)** State trajectories for the initial condition $\theta = \pi/2.2$ and $\dot{\theta} = 0$ under the final policies with (a) their weights $\mathbf{v}_f^{(2)}$ in (5.39) and (b) $\mathbf{v}_f^{(3)}$, both tranied with the nonlinear dynamics (5.33) under $c = 10$ and $c = 4$, respectively.

those in Fig. 5.4. Actually, the final critic and actor weights in this case are given by

$$\begin{cases} \mathbf{w}_f^{(3)} = [\,73.5444,\;\; 40.5142,\;\; 7.1065,\;\; 44.4183,\;\; 33.6221,\;\; 10.1088,\;\; 0.1591,\;\; 0.3364\,]^T; \\ \mathbf{v}_f^{(3)} = [\,20.2389,\;\; 7.1012,\;\; 17.4999,\;\; 8.0659,\;\; 0.6003,\;\; 0.6496\,]^T. \end{cases}$$

Comparing the first three and two components of $\mathbf{w}_f^{(3)}$ and $\mathbf{v}_i^{(3)}$ with (5.36), one can see that the linear and quadratic weights definitely converge to a point near the LQR solution. On the other hand, due to the biased exploration (5.40), the state and input trajectories shown in Fig. (5.9) oscillates from a point which is of course different from the zero equilibrium $\mathbf{x} = \mathbf{0}_2$ and the zero input $u \equiv 0$.

The state trajectories for the initial condition $\theta = \pi/2.5$ and $\dot{\theta} = 0$ under the final policies with (a) the final actor weights $\mathbf{v}_f^{(2)}$ (2nd-order approximation) and (b) $\mathbf{v}_f^{(3)}$ (4th-order approximation) are illustrated in Fig. 5.10. Here, by the effects of the nonlinear terms in the actor, the states under the high-order actor weights $\mathbf{v}_f^{(3)}$ in (5.39) converge much faster than the second-order ones $\mathbf{v}_f^{(2)}$. Moreover, although the final actor weights $\mathbf{v}_f^{(3)}$ are learned with a smaller exploration, it is more robust than the former with respect to the system nonlinearities as shown in Figs. 5.11. Obviously, for the initial condition $\theta = \pi/2.2$ and $\dot{\theta} = 0$, the state trajectories under the final policies with $\mathbf{v}_f^{(3)}$ effectively converge to the up-right state (Fig. 5.11(a)), but those with $\mathbf{v}_f^{(2)}$ do not as shown in Fig. 5.11(b). Therefore, the introductions of the high-order terms in the critic and actor NNs finally enhance both control performance of robustness.

To properly learn the coefficients of the 3rd- and 4th-order terms (and higher-order terms) in the actor and critic NNs, it is necessary to properly apply a *biased* exploration as the one (5.40) in the previous simulations. Otherwise, the higher-order terms become unobservable and cannot be estimated near the equilibrium $\mathbf{x} = \mathbf{0}_2$. This is because of the fact that a nonlinear optimal control problem is approximated as an LQR problem where the activation functions of the critic and actor NNs are quadratic and linear, respectively; the other high-order terms in this case are either zeros or very small to be detected.

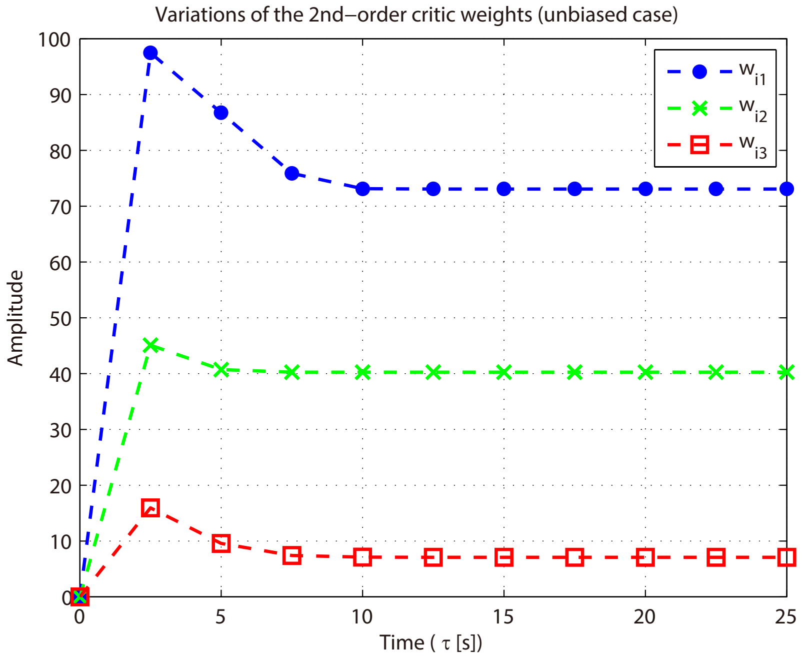


Figure 5.12: **(RL of nonlinear optimal control)** Variations of the critic NN weights corresponding the quadratic activation functions *under unbiased exploration.*

Figs. 5.12, 5.13, and 5.14 show the weight variations of the critic and actor NNs under the unbiased exploration (5.37) with $c = 4$. As can be seen from the figures, the 3rd-order and 4th-order coefficients in the actor and critic NNs do not converge but oscillates within bounded regions due to the unobservability of the high-order terms at $\mathbf{x} = \mathbf{0}_2$. On the other hand, the weights corresponding the quadratic and linear activation functions of the critic and actor NNs converge to a point close to the optimal solution $\mathbf{w}^*$ and $\mathbf{v}^*$ in (5.36) since they are observables near and every regions in the state space.

5.8 Summary

In this chapter, two online IRL methods that are able to explore the state space were proposed and analyzed based on the nonlinear I-PI and the concepts of both invariant explorations and advanced I-TD extended from the ideas of RL. These online IRL methods efficiently use the explorations to excite the necessary signals for online learning and, in integral Q-learning, to relax the model requirements; integral Q-learning provided the

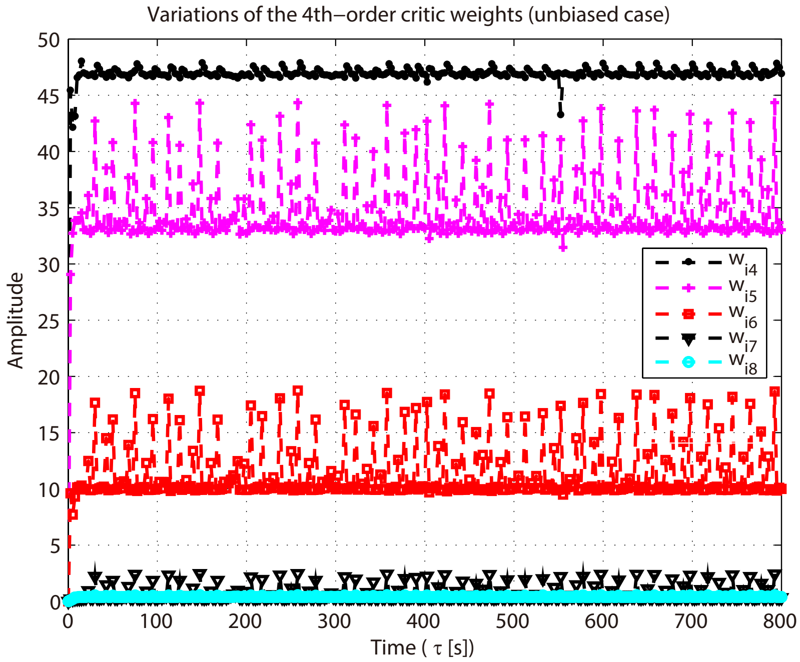


Figure 5.13: **(RL of nonlinear optimal control)** Variations of the critic NN weights corresponding the *4th-order* activation functions *under unbiased exploration.*

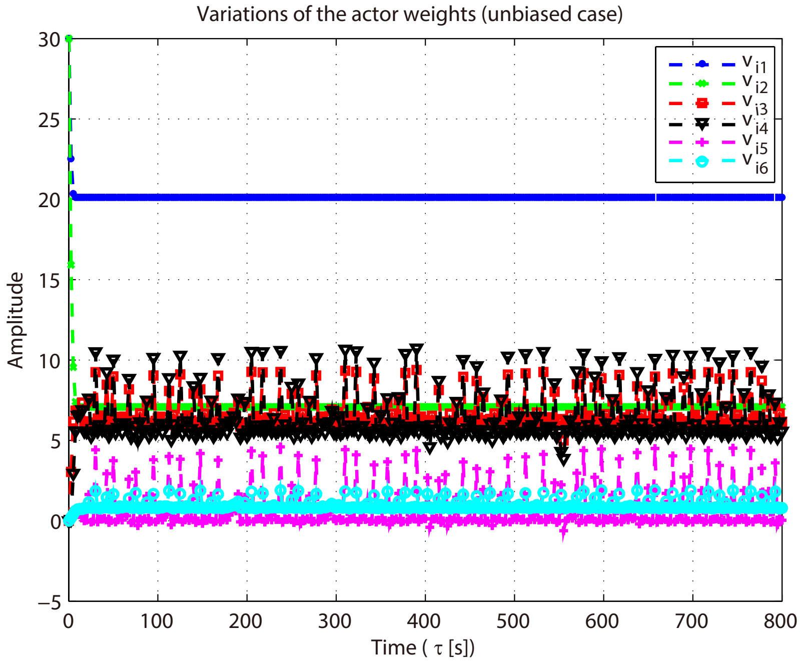


Figure 5.14: **(RL of nonlinear optimal control)** Variations of the actor NN weights *under unbiased exploration.*

model-free online learning solution for the CT nonlinear optimal control problems with unknown dynamics, while the other one named explorized I-PI was provided as an effective online solution when the input coupling terms in the dynamics are known. The properties such as ISS, uniqueness of advanced I-TD solution, and the convergence to the solution were studied in relation to the design of the exploration signal. Finally, numerical simulations for inverted pendulum were carried out to verify the performance of integral Q-learning, show a practical application example, and further study the algorithm.

Chapter 6

Adaptive Inverse Optimal Design of Cooperative Graphical Formation Control for Multiple Mobile Robots

In the previous two chapters, we studied adaptive optimal control methods from the perspectives of RL. In this chapter, from the control-theoretic perspectives, I present an adaptive inverse optimal design methodology for cooperative graphical formation control (CGFC) of multiple mobile robots whose communication topology is represented by an undirected graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$ (see Appendix C for a brief review of graph theory). In the design, the following kinematic and dynamic models are considered for each i-th mobile robot ($i \in \mathcal{N}$):

$$\textit{Kinematics:} \quad \begin{bmatrix} \dot{x}_i \\ \dot{y}_i \\ \dot{\theta}_i \end{bmatrix} = \begin{bmatrix} \nu_i \cos\theta_i \\ \nu_i \sin\theta_i \\ w_i \end{bmatrix}, \tag{6.1}$$

$$\textit{Dynamics:} \quad \begin{bmatrix} m_{i11} & m_{i12} \\ m_{i12} & m_{i11} \end{bmatrix} \begin{bmatrix} \dot{w}_{iR} \\ \dot{w}_{iL} \end{bmatrix} + \begin{bmatrix} b_i & \alpha_i w_i R_i^{-1} \\ -\alpha_i w_i R_i^{-1} & b_i \end{bmatrix} \begin{bmatrix} w_{iR} \\ w_{iL} \end{bmatrix} = \begin{bmatrix} \tau_{iR} \\ \tau_{iL} \end{bmatrix}, \tag{6.2}$$

where $[x_i, y_i]^T \in \mathbb{R}^2$ and $\theta_i \in \mathbb{R}$ are the position and the angle orientation of the i-th mobile robot; $\nu_i \in \mathbb{R}$ and $w_i \in \mathbb{R}$ are linear and angular velocities of the i-th robot; $w_{iL} \in \mathbb{R}$ and $w_{iR} \in \mathbb{R}$ denote the angular velocities of the left and right wheels of the robot; $\tau_{iR} \in \mathbb{R}$ and $\tau_{iL} \in \mathbb{R}$ represents the torques applied to the robot's left and right wheels, respectively.

In the dynamics (6.2), $b_i > 0$ denotes the damping coefficient, $R_i > 0$ is the half of the width of the i-th mobile robot, and $\alpha_i > 0$ is the constant defined as $\alpha_i := r_i^2 m_{ic} d_i / 2$, where r_i is the radius of the wheel, d_i is the distance from the center of mass P_i^c of the

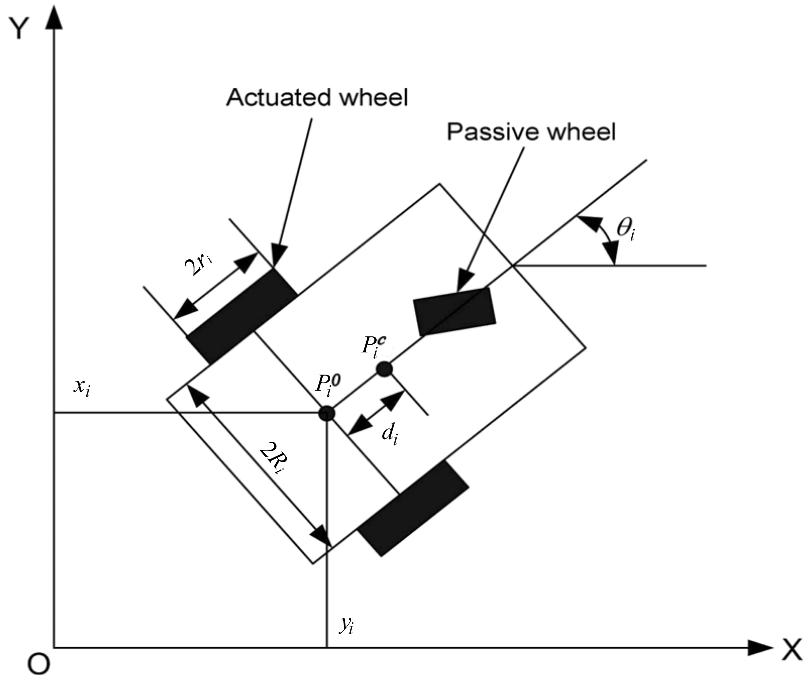


Figure 6.1: The spatial parameters of mobile robots.

i-th robot to the middle point P_i^0 between the right and left wheels, and m_{ic} is the mass of the body of the i-th mobile robot; the effective masses m_{i11} and m_{i12} of the i-th mobile robot are given by

$$\begin{cases} m_{i11} = I_i^w + r_i^2(m_i R_i^2 + I_i)/4R_i^2, \\ m_{i12} = r_i^2(m_i R_i^2 - I_i)/4R_i^2 \end{cases}$$

$$(I_i = m_{ic}\, d_i^2 + 2\, m_{iw}\, R_i^2 + I_i^c + 2\, I_i^m,\ m_i = m_{ic} + 2\, m_{iw}),$$

where m_{iw} is the mass of a wheel; I_i^c, I_i^w, and I_i^m are the moment of inertia of the body about the vertical axis through P_i^0, the wheel about the wheel axis, and the wheel about the wheel diameter, respectively. The velocities of the mobile robot (ν_i, w_i) and the angular velocities of the wheels $(w_{L,i}, w_{iR})$ have the relationship

$$\begin{cases} \nu_i = r_i(w_{iR} + w_{iL})/2, \\ w_i = r_i(w_{iR} - w_{iL})/2R_i. \end{cases} \tag{6.3}$$

6.1 Transformations of Mobile Robot Dynamic Models

For the design of the desired CGFC, the mobile robot's dynamics (6.1) and (6.2) needs to be transformed in terms of the consensus errors. For this, differentiate $\dot{x}$ and $\dot{y}$ in (6.1) to obtain

$$\ddot{x}_i = \dot{\nu}_i \cos\theta_i - \nu_i w_i \sin\theta_i \quad \text{and} \quad \ddot{y}_i = \dot{\nu}_i \sin\theta_i + \nu_i w_i \cos\theta_i,$$

which can be rewritten in the following nonlinear dynamic model of the mobile robot:

$$\begin{cases} \dot{\mathbf{q}}_i = \mathbf{v}_i \\ \dot{\mathbf{v}}_i = \mathbf{T}(\theta_i, \nu_i)\mathbf{u}_i \end{cases} \tag{6.4}$$

where $\mathbf{q}_i \in \mathbb{R}^2$ and $\mathbf{v}_i \in \mathbb{R}^2$ are the position and velocity vectors defined as $\mathbf{q}_i := [x_i, y_i]^T$ and $\mathbf{v}_i := \dot{\mathbf{q}}_i$, respectively; $\mathbf{u}_i := [\,\dot{\nu}_i \;\; w_i\,]^T \in \mathbb{R}^2$ denotes the effective control input for the system (6.4), and $\mathbf{T}(\theta_i, \nu_i) \in \mathbb{R}^{2\times 2}$ is the transformation matrix given by

$$\mathbf{T}(\theta_i, \nu_i) := \begin{bmatrix} \cos\theta_i & -\nu_i \sin\theta_i \\ \sin\theta_i & \nu_i \cos\theta_i \end{bmatrix} \tag{6.5}$$

whose inverse is given by

$$\mathbf{T}^{-1}(\theta_i, \nu_i) = \begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\nu_i^{-1} \sin\theta_i & \nu_i^{-1} \cos\theta_i \end{bmatrix}.$$

For the existence of the inverse $\mathbf{T}^{-1}(\theta_i, \nu_i)$, we assume that

Assumption 6.1. *There is a positive constant $\nu_{min} > 0$ such that*

$$\nu_i(t) \geq \nu_{min} > 0 \text{ for all } i \in \mathcal{N} \text{ and all } t \geq 0.$$

Lemma 6.1. *Under Assumption 6.1, $\nu_i(t)$, $\mathbf{T}(\theta_i, \nu_i)$, and $\mathbf{T}^{-1}(\theta_i, \nu_i)$ are represented in* *terms of* $\mathbf{v}_i$ *as*

$$\nu_i = \|\mathbf{v}_i\|,\ \mathbf{T}(\mathbf{v}_i) = \begin{bmatrix} \dot{x}_i/\|\mathbf{v}_i\| & -\dot{y}_i \\ \dot{y}_i/\|\mathbf{v}_i\| & \dot{x}_i \end{bmatrix},\ \textit{and}\ \mathbf{T}^{-1}(\mathbf{v}_i) = \frac{1}{\|\mathbf{v}_i\|}\begin{bmatrix} \dot{x}_i & \dot{y}_i \\ -\dot{y}_i/\|\mathbf{v}_i\| & \dot{x}_i/\|\mathbf{v}_i\| \end{bmatrix}$$

Proof. Note that $\dot{x} = \nu_i \cos\theta_i$, $\dot{y} = \nu_i \sin\theta_i$, and $\nu_i = \dot{x}_i \cos\theta_i + \dot{y}_i \sin\theta_i$ imply $\nu_i^2 = \|\mathbf{v}_i\|^2$, so we have $\nu_i = \|\mathbf{v}_i\|$ under Assumption 6.1. Then, the proof is completed by substituting $\nu_i = \|\mathbf{v}_i\|$, $\dot{x} = \nu_i \cos\theta_i$, and $\dot{y} = \nu_i \sin\theta_i$ into the definitions of $\mathbf{T}(\theta_i, \nu_i)$ and $\mathbf{T}^{-1}(\theta_i, \nu_i)$. □

By Lemma 6.1 and Assumption 6.1, the transformed system (6.4) can be rewritten as $\dot{\mathbf{q}}_i = \mathbf{v}_i$ and $\dot{\mathbf{v}}_i = \mathbf{T}(\mathbf{v}_i)\mathbf{u}_i$ without explicitly using the angle orientation θ_i. Next, to obtain the transformed model of the robot dynamics (6.2), differentiate (6.3) with respect to time, which yields

$$\begin{aligned} J_{i1}\dot{\nu}_i &= -b_i\nu_i + \alpha_i w_i^2 + r_i\tau_{i\nu} \\ J_{i2}\dot{w}_i &= -b_i w_i - \frac{\alpha_i}{R_i^2}\nu_i w_i + \frac{r_i}{R_i^2}\tau_{iw}, \end{aligned}$$

where $\tau_{i\nu}$ and τ_{iw} are defined by

$$\begin{cases} \tau_{i\nu} := (\tau_{iR} + \tau_{iL})/2 \\ \tau_{iw} := (\tau_{iR} - \tau_{iL})/2, \end{cases} \tag{6.6}$$

and J_{i1} and J_{i2} are defined by $J_{i1} := m_{i11} + m_{i12i}$ and $J_{i2} := m_{i11} - m_{i12}$, respectively. By the definitions, J_{i2} and J_{i2} satisfy $J_{i1} = I_i^w + r_i^2 I_i/2R_i^2$ and $J_{i2} = I_i^w + r_i^2 m_i/2$, respectively. Let $\rho_{i\nu}$ and ρ_{iw} be defined as

$$\begin{cases} \rho_{i\nu} := J_{i1}^{-1}(-b_i\nu_i + \alpha_i w_i^2 + r_i\tau_{i\nu}) \\ \rho_{iw} := J_{i2}^{-1}(-b_i w_i - R_i^{-2}\alpha_i\nu_i w_i + R_i^{-2}r_i\tau_{iw}). \end{cases} \tag{6.7}$$

Then, we finally obtain the following complete model for each i-th the mobile robot:

$$\text{Consensus model:} \quad \begin{cases} \dot{\mathbf{q}}_i = \mathbf{v}_i \\ \dot{\mathbf{v}}_i = \mathbf{T}(\mathbf{v}_i)\mathbf{u}_i, \end{cases} \tag{6.8}$$

$$\text{Dynamic model:} \quad \begin{cases} \dot{\nu}_i = \rho_{i\nu} \\ \dot{w}_i = \rho_{iw}. \end{cases} \tag{6.9}$$

Notice that all of the i-th robot's physical parameters are stuck together into $\rho_{i\nu}$ and ρ_{iw} in (6.9). This dramatically simplifies the whole design procedure, especially the procedure to derive the adaptation laws.

The whole design procedure in this chapter is divided into three steps that ultimately provide the actual feedback torque inputs τ_{iR} and τ_{iL} of each mobile robot in a stable, inverse optimal fashion. First, we design the inverse optimal effective control input $\mathbf{u}_i$ in (6.8). Second, the feedback control inputs $(\rho_{i\nu}, \rho_{iw})$ to (6.9) is derived using the inverse optimal derivative-free partial backstepping. Finally, the adaptation laws are designed to compensate the parametric uncertainties in $(\rho_{i\nu}, \rho_{iw})$, which results in the adaptive inverse optimal actual torque inputs τ_{iR} and τ_{iL} to each mobile robot for their CGFC.

6.2 Inverse Optimal Design of $(\dot{\nu}_i, w_i)$

As the first step, we focus on the consensus model (6.8) and design the effective control $\mathbf{u}_i = [\,\dot{\nu}_i \;\; w_i\,]^T$ of each i-th robot in an inverse optimal fashion. To define the desired formation and group velocity, let $\mathbf{d}_i = [\,d_{x,i} \;\; d_{y,i}\,]^T \in \mathbb{R}^2$ and $\mathbf{v}_g = [\,v_{x,g} \;\; v_{y,g}\,]^T \in \mathbb{R}^2$ be the consensus position vector of the i-th mobile robot and the group velocity vector, both of which has the following combined dynamics:

$$\dot{\mathbf{d}}_i = \mathbf{v}_g \quad \text{and} \quad \dot{\mathbf{v}}_g = \mathbf{0}_2. \tag{6.10}$$

Here, the relative position $\mathbf{d}_i - \mathbf{d}_j$ defines the desired formation between the i-th and the j-th robots, and the *nonzero* group velocity $\mathbf{v}_g$ states the desired nonzero speed and angular orientation of the group (here, we only consider the nonzero $\mathbf{v}_g$ to hold Assumption 6.1 in the limit $\mathbf{v}_i \to \mathbf{v}_g$). Then, differentiating the formation consensus errors $\boldsymbol{\delta}_i \in \mathbb{R}^2$ and $\boldsymbol{\eta}_i \in \mathbb{R}^2$ defined as

$$\boldsymbol{\delta}_i := \mathbf{q}_i - \mathbf{d}_i \quad \text{and} \quad \boldsymbol{\eta}_i := \mathbf{v}_i - \mathbf{v}_g, \tag{6.11}$$

substituting (6.8), (6.10), (6.11), and $\mathbf{T}(\mathbf{v}_i) = \mathbf{T}(\boldsymbol{\eta}_i + \mathbf{v}_g)$ yields the following formation consensus error dynamics:

$$\begin{cases} \dot{\boldsymbol{\delta}}_i = \boldsymbol{\eta}_i \\ \dot{\boldsymbol{\eta}}_i = \mathbf{T}(\boldsymbol{\eta}_i, \mathbf{v}_g)\mathbf{u}_i, \end{cases} \tag{6.12}$$

where $\mathbf{T}(\boldsymbol{\eta}_i, \mathbf{v}_g) := \mathbf{T}(\boldsymbol{\eta}_i + \mathbf{v}_g)$ with slight abuse of notations. Now, we propose

$$\mathbf{u}_i^* = \mathbf{T}^{-1}(\boldsymbol{\eta}, \mathbf{v}_g)\mathbf{u}_i^*, \tag{6.13}$$

as an inverse optimal effective control input to (6.12) for the formation and velocity consensuses, where $\mathbf{u}_i^* \in \mathbb{R}^2$ is the linear part of $\mathbf{u}_i$ given by

$$\mathbf{u}_i^* := -\mathbf{K} \cdot \sum_{j \in \mathcal{N}_i} a_{ij} \begin{bmatrix} \mathbf{q}_i - \mathbf{q}_j - \mathbf{d}_{ij} \\ \mathbf{v}_i - \mathbf{v}_j \end{bmatrix} - \gamma^{-1}\mathbf{Q}_v\,(\mathbf{v}_i - \mathbf{v}_g). \tag{6.14}$$

Here, $\gamma > 0$ is a positive constant, a_{ij} is the (i, j)-th element of the adjacency matrix $\mathcal{A}$ of the graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$, $\mathbf{Q}_v \in \mathbb{R}^{2\times 2}$ is a positive definite matrix, and $\mathbf{K} \in \mathbb{R}^{4\times 2}$ is the optimal gain matrix given by $\mathbf{K} := \gamma^{-1}\mathbf{B}_0^T\mathbf{P}$ for the solution to the ARE:

$$\mathbf{A}_c^T\mathbf{P} + \mathbf{P}\,\mathbf{A}_c + \mathbf{Q} - \frac{1}{\gamma}\mathbf{P}\,\mathbf{B}_0\mathbf{B}_0^T\mathbf{P} = \mathbf{0}_{4\times 4}. \tag{6.15}$$

where $\mathbf{Q} \in \mathbb{R}^{4\times 4}$ is a positive definite matrix, and the matrices $\mathbf{A}_c \in \mathbb{R}^{4\times 4}$ and $\mathbf{B}_0 \in \mathbb{R}^{4\times 2}$ are defined as follows:

$$\mathbf{B}_0 := \begin{bmatrix} \mathbf{0}_{2\times 2} \\ \mathbf{I}_2 \end{bmatrix} \text{ and } \mathbf{A}_c := \mathbf{A}_0 - \frac{1}{\gamma}\mathbf{B}_0\,\mathbf{Q}_v\mathbf{B}_0^T \text{ for } \mathbf{A}_0 := \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{I}_2 \\ \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} \end{bmatrix}.$$

To derive the global expression, define the (global) vectors of the formation consensus errors, and the effective control inputs as

$$\begin{aligned} &\boldsymbol{\delta} := \mathbf{col}\{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_N\}, \quad \boldsymbol{\eta} := \mathbf{col}\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_N\}, \quad \mathbf{u} := \mathbf{col}\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}, \\ &\mathbf{e}_i := \mathbf{col}\{\boldsymbol{\delta}_i, \boldsymbol{\eta}_i\}, \qquad\qquad \mathbf{e} := \mathbf{col}\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N\}, \end{aligned}$$

and notice that

1. the substitution of (6.13) into $\mathbf{u}_i$ and the use of $\mathbf{A}_0$ and $\mathbf{B}_0$ yields the $\mathbf{e}_i$-dynamics expression of (6.12) as $\dot{\mathbf{e}}_i = \mathbf{A}_0\mathbf{e}_i + \mathbf{B}_0\mathbf{u}_i^*$;
2. using the definition of $\mathbf{e}_i$ and the properties "$l_{ij} = -a_{ij}$ and $l_{ii} = \sum_{j\in\mathcal{N}_i} a_{ij}$" of the Laplacian matrix $\mathbf{L}$, the linear policy $\mathbf{u}_i^*$ given in (6.14) can be compactly rewritten as $\boldsymbol{\mu}_i^* = -\mathbf{K}\sum_{j=1}^N l_{ij}\mathbf{e}_j - \gamma^{-1}\mathbf{Q}_v\mathbf{B}_0^T\mathbf{e}_i$.

Then, using the Kronecker product and its properties, we obtain the global expression of the form

$$\begin{cases} \dot{\mathbf{e}} = \mathbf{A}_0^{\otimes}\mathbf{e} + \mathbf{B}_0^{\otimes}\mathbf{u}^*, \\ \boldsymbol{\mu}^* = -\big[(\mathbf{L}\otimes\mathbf{K}) + \gamma^{-1}(\mathbf{I}_N \otimes \mathbf{Q}_v\mathbf{B}_0^T)\big]\mathbf{e}, \end{cases} \tag{6.16}$$

where $\mathbf{A}_0^{\otimes} := \mathbf{I}_N \otimes \mathbf{A}_0$, $\mathbf{B}_0^{\otimes} := \mathbf{I}_N \otimes \mathbf{B}_0$, and $\mathbf{u}^* := \mathbf{col}\{\mathbf{u}_1^*, \mathbf{u}_2^*, \cdots, \mathbf{u}_N^*\}$.

Lemma 6.2. *The followings hold for the matrices* $\boldsymbol{\Pi}$*,* $\boldsymbol{\Theta}$*, and* $\boldsymbol{\Phi}$ *defined as*

$$\begin{cases} \boldsymbol{\Pi} := \mathbf{L}\otimes\mathbf{P} + \mathbf{I}_N \otimes \mathbf{B}_0\,\mathbf{Q}_v\mathbf{B}_0^T, \\ \boldsymbol{\Theta} := \boldsymbol{\Phi} + (\mathbf{L}^2 - \mathbf{L}) \otimes \mathbf{P}\mathbf{B}_0\mathbf{B}_0^T\mathbf{P}/\gamma, \\ \boldsymbol{\Phi} := (\mathbf{L}\otimes\mathbf{Q}) + \gamma^{-1}(\mathbf{I}_N \otimes \mathbf{B}_0\,\mathbf{Q}_v^2\mathbf{B}_0^T). \end{cases}$$

1) $(\mathbf{A}_0^{\otimes})^T\boldsymbol{\Pi} + \boldsymbol{\Pi}\mathbf{A}_0^{\otimes} + \boldsymbol{\Theta} - \boldsymbol{\Pi}\mathbf{B}_0^{\otimes}(\mathbf{B}_0^{\otimes})^T\boldsymbol{\Pi}/\gamma = \mathbf{0}_{4\times4};$ (6.17)

2) $\mathbf{u}^* = -\gamma^{-1}(\mathbf{B}_0^{\otimes})^T\boldsymbol{\Pi}\mathbf{e};$ (6.18)

3) $\ker\boldsymbol{\Pi} = \ker\boldsymbol{\Phi} = \ker(\mathbf{L}\otimes\mathbf{I}_4)\cap\ker(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{B}_0^T).$

Proof. The proof of (6.17) can be done using the Kronecker product properties and the definitions of the matrices as follows:

$$\begin{aligned}
&(\mathbf{A}_0^{\otimes})^T\boldsymbol{\Pi} + \boldsymbol{\Pi}\mathbf{A}_0^{\otimes} + \boldsymbol{\Theta} - \boldsymbol{\Pi}\mathbf{B}_0^{\otimes}(\mathbf{B}_0^{\otimes})^T\boldsymbol{\Pi}/\gamma \\
&= \mathbf{L}\otimes(\mathbf{A}_0^T\mathbf{P} + \mathbf{P}\mathbf{A}_0) + \mathbf{I}_N\otimes(\underbrace{\mathbf{A}_0^T\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T + \mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T\mathbf{A}_0}_{=\mathbf{0}_{4\times4}}) + \boldsymbol{\Theta} - \boldsymbol{\Pi}(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{B}_0^T)\boldsymbol{\Pi}/\gamma \\
&= \mathbf{L}\otimes(\mathbf{A}_c^T\mathbf{P} + \mathbf{P}\mathbf{A}_c) + \boldsymbol{\Theta} - \left(\mathbf{I}_N\otimes\frac{\mathbf{B}_0{\mathbf{Q}_v}^2\mathbf{B}_0^T}{\gamma}\right) - \left(\mathbf{L}^2\otimes\frac{\mathbf{P}\mathbf{B}_0\mathbf{B}_0^T\mathbf{P}}{\gamma}\right) \\
&= \mathbf{L}\otimes(\underbrace{\mathbf{A}_c^T\mathbf{P} + \mathbf{P}\mathbf{A}_c + \mathbf{Q} - \mathbf{P}\mathbf{B}_0\mathbf{B}_0^T\mathbf{P}/\gamma}_{=\mathbf{0}_{4\times4}\text{by the ARE (6.15)}}) \\
&= \mathbf{0}_{4\times4}.
\end{aligned}$$

Similarly, from $\mathbf{K} = \gamma^{-1}\mathbf{B}_0^T\mathbf{P}$, one can also show that

$$\begin{aligned}
\gamma^{-1}(\mathbf{B}_0^{\otimes})^T\boldsymbol{\Pi} &= \gamma^{-1}\big[(\mathbf{L}\otimes\mathbf{B}_0^T\mathbf{P}) + (\mathbf{I}_N\otimes\mathbf{B}_0^T\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T)\big] \\
&= (\mathbf{L}\otimes\mathbf{K}) + \gamma^{-1}(\mathbf{I}_N\otimes\mathbf{Q}_v\mathbf{B}_0^T),
\end{aligned}$$

which proves (6.18). On the other hand, the definition of $\boldsymbol{\Phi}$ and the Kroncker algebra imply

$$\boldsymbol{\Pi}\mathbf{x} = (\mathbf{I}_N\otimes\mathbf{P})(\mathbf{L}\otimes\mathbf{I}_4)\mathbf{x} + (\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T)\mathbf{x}, \quad \forall\mathbf{x}\in\mathbb{R}^{4N}. \tag{6.19}$$

Hence, $\boldsymbol{\Pi}\mathbf{x} = \mathbf{0}_{4N}$ always implies $\mathbf{x}\in\ker(\mathbf{L}\otimes\mathbf{I}_4)$ and $\mathbf{x}\in\ker(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T)$. Moreover, $\ker(\mathbf{B}_0\mathbf{Q}_v^k\mathbf{B}_0^T) = \ker(\mathbf{B}_0\mathbf{B}_0^T)$ holds for any $\mathbf{Q}_v \succ \mathbf{0}_{2\times2}$, so we have

$$\mathbf{x}\in\ker\boldsymbol{\Pi} \implies \mathbf{x}\in\ker(\mathbf{L}\otimes\mathbf{I}_4)\cap\ker(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{B}_0^T),$$

whose converse is also true by (6.19) and $\ker(\mathbf{B}_0\mathbf{Q}_v^k\mathbf{B}_0^T) = \ker(\mathbf{B}_0\mathbf{B}_0^T)$. Since

$$\boldsymbol{\Phi}\mathbf{x} = (\mathbf{I}_N\otimes\mathbf{Q})(\mathbf{L}\otimes\mathbf{I}_4)\mathbf{x} + \gamma^{-1}(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{Q}_v^2\mathbf{B}_0^T)\mathbf{x},$$

one can also show $\ker\boldsymbol{\Phi} = \ker(\mathbf{L}\otimes\mathbf{I}_4)\cap\ker(\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{B}_0^T)$, which completes the proof. □

Let $\mathbf{\Gamma}(\mathbf{v}_i) \in \mathbb{R}^{2\times 2}$ be defined as

$$\mathbf{\Gamma}(\mathbf{v}_i) := \gamma \cdot \mathbf{diag}\{1, \|\mathbf{v}_i\|^2\}. \tag{6.20}$$

Then, it implicitly depends on both $\boldsymbol{\eta}_i$ and $\mathbf{v}_g$ by $\mathbf{v}_i = \boldsymbol{\eta}_i + \mathbf{v}_g$. Define the global vector $\mathbf{v} \in \mathbb{R}^{2N}$ and the global matrices $\mathbf{B}_{\mathbf{v}}^{\otimes} \in \mathbb{R}^{4N\times 2N}$ and $\mathbf{\Gamma}_{\mathbf{v}}^{\otimes} \in \mathbb{R}^{2N\times 2N}$ as

$$\begin{cases} \mathbf{v} := \mathbf{col}\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\}, \\ \mathbf{B}_{\mathbf{v}}^{\otimes} := \mathbf{I}_N \otimes \big\{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\big\}_{i=1}^{N} \text{ and } \mathbf{\Gamma}_{\mathbf{v}}^{\otimes} := \mathbf{I}_N \otimes \big\{\mathbf{\Gamma}(\mathbf{v}_i)\big\}_{i=1}^{N}, \end{cases}$$

all of which depends on $\boldsymbol{\eta}$ and $\mathbf{v}_g$ by the rule $\mathbf{v}_i = \boldsymbol{\eta}_i + \mathbf{v}_g$.

Lemma 6.3. *The formation consensus error dynamics* (6.12) *can be represented in a global form*

$$\dot{\mathbf{e}} = \mathbf{A}_0^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}}^{\otimes}\mathbf{u}. \tag{6.21}$$

Moreover, the following equalities hold for $\mathbf{u}$*,* $\mathbf{B}_{\mathbf{v}}^{\otimes}$*, and* $\mathbf{\Gamma}_{\mathbf{v}}^{\otimes}$*:*

$$(\mathbf{A}_0^{\otimes})^T\mathbf{\Pi} + \mathbf{\Pi}\mathbf{A}_0^{\otimes} + \mathbf{\Theta} - \mathbf{\Pi}\,\mathbf{B}_{\mathbf{v}}^{\otimes}\big(\mathbf{\Gamma}_{\mathbf{v}}^{\otimes}\big)^{-1}\big(\mathbf{B}_{\mathbf{v}}^{\otimes}\big)^T\mathbf{\Pi} = \mathbf{0}_{4\times 4}, \tag{6.22}$$

$$\mathbf{u}^* = \big(\mathbf{\Gamma}_{\mathbf{v}}^{\otimes}\big)^{-1}\big(\mathbf{B}_{\mathbf{v}}^{\otimes}\big)^T\mathbf{\Pi}\mathbf{e}. \tag{6.23}$$

where $\mathbf{u}^* := \mathbf{col}\{\mathbf{u}_1^*, \mathbf{u}_2^*, \cdots, \mathbf{u}_N^*\}$.

Proof. First, note that (6.12) can be expressed as $\dot{\mathbf{e}}_i = \mathbf{A}_0\mathbf{e}_i + \mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\mathbf{u}_i$. Using the Khatri-Rao product and its property in Proposition A.6 in Appeidix A, this can be rewritten in terms of the global formation consensus error $\mathbf{e}$ as

$$\dot{\mathbf{e}} = (\mathbf{I}_N \otimes \mathbf{A}_0)\mathbf{e} + \big(\mathbf{I_N} \otimes \{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\}_{i=1}^{N}\big)\mathbf{u} = \mathbf{A}_0^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}}^{\otimes}\mathbf{u}$$

which proves (6.21). Moreover, (6.13) and the use of the Khatri-Rao product yield $\mathbf{u}^* = \big(\mathbf{I}_N \otimes \big\{\mathbf{T}(\mathbf{v}_i)\big\}_{i=1}^{N}\big)\mathbf{u}^*$. Thus, by the substitution of (6.18) in Lemma 6.2 and the operations of the Khatri-Rao product in Propositions A.6 and A.7 in Appendix A, one obtains

$$\begin{aligned}
\mathbf{u}^* &= \Big(\mathbf{I}_N \otimes \{\mathbf{T}^{-1}(\mathbf{v}_i)\}_{i=1}^N\Big)\Big(\mathbf{I}_N \otimes \gamma^{-1}\mathbf{B}_0^T\Big)\mathbf{\Pi}\mathbf{e} \\
&= \Big(\mathbf{I}_N \otimes \Big\{\gamma^{-1}\mathbf{T}^{-1}(\mathbf{v}_i)\mathbf{B}_0^T\Big\}_{i=1}^N\Big)\mathbf{\Pi}\mathbf{e} \\
&= \Big(\mathbf{I}_N \otimes \Big\{\mathbf{\Gamma}^{-1}(\mathbf{v}_i)\mathbf{T}^T(\mathbf{v}_i)\mathbf{B}_0^T\Big\}_{i=1}^N\Big)\mathbf{\Pi}\mathbf{e} \\
&= \Big(\mathbf{I}_N \otimes \{\mathbf{\Gamma}(\mathbf{v}_i)\}_{i=1}^N\Big)^{-1}\Big(\mathbf{I}_N \otimes \{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\}_{i=1}^N\Big)^T\mathbf{\Pi}\mathbf{e} \\
&= (\mathbf{\Gamma}_{\mathbf{v}}^{\otimes})^{-1}(\mathbf{B}_{\mathbf{v}}^{\otimes})^T\mathbf{\Pi}\mathbf{e}.
\end{aligned}$$

In the third equality, the matrix equality $\mathbf{T}(\mathbf{v}_i)\mathbf{\Gamma}^{-1}(\mathbf{v}_i)\mathbf{T}^T(\mathbf{v}_i) = \gamma^{-1}\mathbf{I}_2$ is substituted which can be easily verified using the definitions (6.5) and (6.20). This proves (6.23).

For the proof of (6.22), note that $\mathbf{T}\mathbf{\Gamma}^{-1}\mathbf{T}^T = \gamma^{-1}\mathbf{I}_2$ and the applications of Propositions A.6 and A.7 in Appendix A yield

$$\begin{aligned}
\gamma^{-1}\mathbf{B}_0^{\otimes}(\mathbf{B}_0^{\otimes})^T &= \mathbf{I}_N \otimes \Big\{\mathbf{B}_0\Big(\mathbf{T}(\mathbf{v}_i)\mathbf{\Gamma}^{-1}(\mathbf{v}_i)\mathbf{T}^T(\mathbf{v}_i)\Big)\mathbf{B}_0^T\Big\}_{i=1}^N \\
&= \Big(\mathbf{I}_N \otimes \{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\}_{i=1}^N\Big)\Big(\mathbf{\Gamma}_{\mathbf{v}}^{\otimes}\Big)^{-1}\Big(\mathbf{I}_N \otimes \{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\}_{i=1}^N\Big)^T \\
&= \mathbf{B}_{\mathbf{v}}^{\otimes}(\mathbf{\Gamma}_{\mathbf{v}}^{\otimes})^{-1}(\mathbf{B}_{\mathbf{v}}^{\otimes})^T.
\end{aligned}$$

Therefore, (6.22) can be obtained by the substitution of this into the ARE (6.17) in Lemma 6.2. This completes the proof. □

Lemma 6.4. *Suppose the algebraic connectivity $\lambda_2(\mathbf{L})$ of the undirected graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$ satisfies $\lambda_2(\mathbf{L}) \geq 1$. Then, $\mathbf{0}_{4N\times 4N} \preceq \mathbf{\Phi} \preceq \mathbf{\Theta}$.*

Proof. First, note that since the graph is undirected, its Laplacian $\mathbf{L}$ is positive semi-definite. So, one has $\mathbf{\Phi} \succeq \mathbf{0}_{4N\times 4N}$. Next, by the condition $\lambda_2(\mathbf{L}) \geq 1$, we have $\lambda_j(\mathbf{L}) \geq 1$ for all $j = 2, 3, \cdots, N$, and thus the matrix inequality $\mathbf{L}^2 - \mathbf{L} \succeq \mathbf{0}_{N\times N}$ holds. From this, one has $(\mathbf{L}^2 - \mathbf{L}) \otimes \mathbf{P}\mathbf{B}_0\mathbf{B}_0^T\mathbf{P}/\gamma \succeq \mathbf{0}_{4N\times 4N}$, and the proof is completed by the definition of $\mathbf{\Theta}$ in Lemma 6.2. □

Now, the following theorem is obtained that provides a condition on the graph for the inverse optimal formation and velocity consensuses.

Theorem 6.1. *Consider the group of mobile robots with the formation consensus error dynamics (6.12) whose communication topology is described by a simple undirected graph*

$\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$. *Suppose that each effective control input* $\mathbf{u}_i$ *in the dynamics* (6.12) *is given by* (6.13) *and* (6.14). *If the Laplacian* $\mathbf{L}$ *of the graph* $\mathcal{G}$ *satisfies*

$$\lambda_2(\mathbf{L}) \geq 1, \tag{6.24}$$

then the followings holds under Assumption 6.1.

1. ***(Formation and Velocity Consensus)*** *The mobile robot agents achieve the formation and velocity consensuses*

$$\lim_{\tau \to \infty} \|\mathbf{q}_i(\tau) - \mathbf{q}_j(\tau) - \mathbf{d}_{ij}\| = 0 \textit{ and } \lim_{\tau \to \infty} \|\mathbf{v}_i(\tau) - \mathbf{v}_g\| = 0 \tag{6.25}$$

exponentially. That is, there exist $\beta > 0$ *and* $\kappa > 0$ *such that for* $\tau \geq t$,

$$d(\mathbf{e}(\tau), \mathbb{S}) \leq \beta e^{-\kappa(\tau - t)} d(\mathbf{e}(t), \mathbb{S}),$$

where the consensuses are achieved on the subspace $\mathbb{S}$ *given by* $\mathbb{S} = \ker \mathbf{\Pi}$.

2. ***(Inverse Optimality)*** *The feedback control inputs* $\{\mathbf{u}_i^*\}_{i=1}^N$ *given in* (6.13) *and* (6.14) *are the inverse optimal solution that cooperatively minimize the performance index* $J(\mathbf{e}(0), \mathbf{u}(\cdot))$ *given by*

$$J(\mathbf{e}(t), \mathbf{u}(\cdot)) = \int_t^\infty \left(\mathbf{e}^T \mathbf{\Theta} \mathbf{e} + \sum_{i=1}^N \mathbf{u}_i^T \mathbf{\Gamma}(\mathbf{v}_i) \mathbf{u}_i \right) d\tau. \tag{6.26}$$

Proof. Substituting (6.18) in Lemma 6.2 into the linear dynamics (6.16), one obtains

$$\dot{\mathbf{e}} = \left(\mathbf{A}_0^\otimes - \gamma^{-1} \mathbf{B}_0^\otimes (\mathbf{B}_0^\otimes)^T \mathbf{\Pi} \right) \mathbf{e}, \tag{6.27}$$

where the matrix $\mathbf{\Pi}$ is positive semi-definite since so is the Laplacian $\mathbf{L}$ of the undirected graph $\mathcal{G}$ (see Lemma 6.2). Now, consider $V(\mathbf{e}) := \mathbf{e}^T \mathbf{\Pi} \mathbf{e}$ as the Lyapunov function candidate. Differentiating V with respect to the linear closed-loop system (6.27) and substituting (6.17) yields

$$\begin{aligned} \dot{V}(\mathbf{e}) &= \mathbf{e}^T \Big[(\mathbf{A}_0^\otimes)^T \mathbf{\Pi} + \mathbf{\Pi} \mathbf{A}_0^\otimes - 2\, \mathbf{\Pi} \mathbf{B}_0^\otimes (\mathbf{B}_0^\otimes)^T \mathbf{\Pi} / \gamma \Big] \mathbf{e} \\ &= -\mathbf{e}^T \Big[\mathbf{\Theta} + \mathbf{\Pi} \mathbf{B}_0^\otimes (\mathbf{B}_0^\otimes)^T \mathbf{\Pi} / \gamma \Big] \mathbf{e}. \end{aligned}$$

Then, by Lemma 6.4, $\dot{V}$ satisfies

$$\dot{V}(\mathbf{e}) \leq -\mathbf{e}^T \left(\boldsymbol{\Phi} + \boldsymbol{\Pi}\mathbf{B}_0^{\otimes}(\mathbf{B}_0^{\otimes})^T\boldsymbol{\Pi}/\gamma\right)\mathbf{e} \leq -\mathbf{e}^T\boldsymbol{\Phi}\mathbf{e}.$$

On the other hand, Lemma 2.5 implies there are positive constants $\underline{\alpha}$, $\bar{\alpha}$, $\underline{\beta}$, and $\bar{\beta} > 0$ such that

$$\underline{\alpha} \cdot d_2^2(\mathbf{e}, \ker\boldsymbol{\Phi}) \leq \mathbf{e}^T\boldsymbol{\Phi}\mathbf{e} \leq \bar{\alpha} \cdot d_2^2(\mathbf{e}, \ker\boldsymbol{\Phi}),$$
$$\underline{\beta} \cdot d_2^2(\mathbf{e}, \ker\boldsymbol{\Pi}) \leq V(\mathbf{e}) \leq \bar{\beta} \cdot d_2^2(\mathbf{e}, \ker\boldsymbol{\Pi}).$$

Hence, by Theorem 3.1 and the property "$\ker\boldsymbol{\Phi} = \ker\boldsymbol{\Pi}$" in Lemma 6.2, the policy given by (6.13) and (6.14) exponentially stabilizes the consensus error dynamics (6.12) to the subspace $\mathbb{S} = \ker\boldsymbol{\Pi}$. Next, consider the orthogonal decomposition

$$\mathbf{e}(\tau) = \mathbf{e}_r(\tau) + \mathbf{e}_n(\tau),$$

where $\mathbf{e}_r$ belongs to the row-space of $\boldsymbol{\Pi}$ and $\mathbf{e}_n \in \ker\boldsymbol{\Pi}$. Then, the exponential stability "$d(\mathbf{e}(\tau), \mathbb{S}) \leq \beta e^{-\kappa(\tau-t)}d(\mathbf{e}(t), \mathbb{S})$" implies the limit $\lim_{\tau\to\infty}\mathbf{e}_r(\tau) = \mathbf{0}_{4N}$, and by Lemma 6.2,

$$\mathbf{e}_n(\tau) \in \ker(\mathbf{L} \otimes \mathbf{I}_4) \cap \ker(\mathbf{I}_N \otimes \mathbf{B}_0\mathbf{B}_0^T) \quad \forall\tau \geq 0,$$

which implies that $\mathbf{e}_n(\tau)$ is represented as

$$\mathbf{e}_n(\tau) = \mathbf{z}_N \otimes \mathbf{col}\{\boldsymbol{\delta}_c(\tau), \mathbf{0}_2\}$$

for some $\mathbf{z}_N \in \ker\mathbf{L}$ and $\boldsymbol{\delta}_c(\tau) \in \mathbb{R}^2$. Since the condition (6.24) implies that $\mathbf{1}_N \in \mathbb{R}^N$ is the unique eigenvector for the zero eigenvalue "$\lambda_1(\mathbf{L}) = 0$", we have $\mathbf{z}_N = \mathbf{1}_N$. Thus, by the definitions of $\mathbf{e}$ and $\mathbf{e}_n \in \ker\boldsymbol{\Pi}$ and the convergence $\lim_{\tau\to\infty}\mathbf{e}_r(\tau) = \mathbf{0}_{4N}$, every formation consensus error $\boldsymbol{\delta}_i(\tau)$ converges to a common function $\boldsymbol{\delta}_c(\tau)$ as $\tau \to \infty$ and

$$\lim_{\tau\to\infty}\boldsymbol{\eta}_1(\tau) = \lim_{\tau\to\infty}\boldsymbol{\eta}_2(\tau) = \cdots = \lim_{\tau\to\infty}\boldsymbol{\eta}_N(\tau) = \mathbf{0}_2.$$

These obviously imply that for all $i, j \in \mathbb{N}$,

$$\lim_{\tau\to\infty}(\boldsymbol{\delta}_i(\tau) - \boldsymbol{\delta}_j(\tau)) = \lim_{\tau\to\infty}(\mathbf{q}_i(\tau) - \mathbf{q}_j(\tau) - \mathbf{d}_{ij}) = \mathbf{0}_2,$$
$$\lim_{\tau\to\infty}\mathbf{v}_i(\tau) = \mathbf{v}_g.$$

Therefore, the formation and velocity consensuses (6.25) are achieved under the polices $\{\mathbf{u}_i^*\}_{i=1}^N$ given by (6.13) and (6.14).

To show the inverse optimality, define the performance index $\mathcal{J}(\mathbf{e}(t), \mathbf{u}(\cdot))$ as

$$\mathcal{J}(\mathbf{e}(t), \mathbf{u}(\cdot)) = \int_t^\infty \left(\mathbf{e}^T \boldsymbol{\Theta} \mathbf{e} + \mathbf{u}^T \boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes} \mathbf{u}\right) \, d\tau.$$

Then, one can see that (6.22) in Lemma 6.3 is actually the matrix form of the nonlinear HJB equation with respect to $\mathcal{J}(\mathbf{e}(t), \mathbf{u}(\cdot))$ above and the system (6.21) in Lemma 6.3; $\mathbf{u}^* = -\big(\boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes}\big)^{-1}\big(\mathbf{B}_{\mathbf{v}}^{\otimes}\big)^T \boldsymbol{\Pi}\mathbf{e}$ is the corresponding optimal policy. Therefore, by Lemma 6.3 and the above discussions, $\{\mathbf{u}_i^*\}_{i=1}^N$ given by (6.23) is the inverse optimal formation consensus policies that cooperatively minimize $\mathcal{J}(\mathbf{e}(0), \mathbf{u}(\cdot))$. Moreover, the definition of $\boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes}$ and the operation of the Khatri-Rao product in Proposition A.6 in Appendix A yield $\mathcal{J}(\mathbf{e}(t), \mathbf{u}(\cdot)) = J(\mathbf{e}(t), \mathbf{u}(\cdot))$ since

$$\mathbf{u}^T \boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes} \mathbf{u} = \mathbf{u}^T \left(\mathbf{I}_N \otimes \{\boldsymbol{\Gamma}(\mathbf{v}_i)\}_{i=1}^N\right) \mathbf{u} = \sum_{i=1}^N \mathbf{u}_i \boldsymbol{\Gamma}(\mathbf{v}_i) \mathbf{u}_i,$$

which completes the proof. □

6.3 Inverse Optimal Cooperative Graphical Formation Control via Input-Dynamics Extension

Based on the the inverse optimal input-dynamics extension technique shown in Section 3.2 and the optimal policy $\mathbf{u}^*$ given by (6.13) and (6.14) in the previous section, this section presents an inverse optimal design method of the control inputs $(\rho_{i\nu}, \rho_{iw})$ of the dynamic model (6.9). The first step of this is to decompose $\mathbf{T}(\mathbf{v}_i)$ column-wisely as

$$\mathbf{T}(\mathbf{v}_i) = \left[\, \mathbf{t}_1(\mathbf{v}_i) \,\vdots\, \mathbf{t}_2(\mathbf{v}_i) \,\right],$$

where $\mathbf{t}_1(\mathbf{v}_i) = [\cos\theta_i \;\; \sin\theta_i\,]^T$ and $\mathbf{t}_2(\mathbf{v}_i) = \nu_i[-\sin\theta_i \;\; \cos\theta_i\,]^T$. Under Assumption 6.1, they are represented in terms of $\mathbf{v}_i$ as $\mathbf{t}_1(\mathbf{v}_i) = \|\mathbf{v}_i\|^{-1}[\,\dot{x}_i \;\; \dot{y}_i\,]^T$ and $\mathbf{t}_2(\mathbf{v}_i) = [-\dot{y}_i \;\; \dot{x}_i\,]^T$.

Next, note that $\mathbf{u}$ can be represented as the Khatri-Rao product "$\mathbf{u} = \mathbf{1}_N \otimes \{\mathbf{u}_i\}_{i=1}^N$" by its definition. Then, substituting $\mathbf{u}_i = [\,\dot{\nu}_i \;\; w_i\,]^T$ and using the definition of $\mathbf{B}_{\mathbf{v}}^{\otimes}$ and the properties of the Khatri-Rao product in Appendix A yields the following formula:

$$
\begin{aligned}
\mathbf{B}_{\mathbf{v}}^{\otimes}\mathbf{u} &= \Big(\mathbf{I}_N \otimes \big\{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\big\}_{i=1}^{N}\Big)\Big(\mathbf{1}_N \otimes \big\{\mathbf{u}_i\big\}_{i=1}^{N}\Big) \\
&= \Big(\mathbf{1}_N \otimes \big\{\mathbf{B}_0\mathbf{T}(\mathbf{v}_i)\mathbf{u}_i\big\}_{i=1}^{N}\Big) \\
&= \big(\mathbf{1}_N \otimes \big\{\mathbf{B}_0\mathbf{t}_1(\mathbf{v}_i)\dot{\nu}_i\big\}_{i=1}^{N}\big) + \big(\mathbf{1}_N \otimes \big\{\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_i)w_i\big\}_{i=1}^{N}\big) \\
&= \mathbf{B}_{\mathbf{v}1}^{\otimes}\dot{\boldsymbol{\nu}} + \mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w},
\end{aligned}
$$

where $\mathbf{B}_{\mathbf{v}k}^{\otimes} \in \mathbb{R}^{(4N)\times(2N)}$ $(k = 1, 2)$ and $\dot{\boldsymbol{\nu}}$, $\mathbf{w} \in \mathbb{R}^N$ are defined as

$$
\begin{aligned}
\mathbf{B}_{\mathbf{v}1}^{\otimes} &:= \mathbf{I}_N \otimes \big\{\mathbf{B}_0\mathbf{t}_1(\mathbf{v}_i)\big\}_{i=1}^{N}, \quad & \dot{\boldsymbol{\nu}} &:= \mathbf{diag}\{\nu_1,\, \nu_2.\, \cdots,\, \nu_N\}, \\
\mathbf{B}_{\mathbf{v}2}^{\otimes} &:= \mathbf{I}_N \otimes \big\{\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_i)\big\}_{i=1}^{N}, \quad & \mathbf{w} &:= \mathbf{diag}\{w_1,\, w_2,\, \cdots,\, w_N\}.
\end{aligned}
$$

Hence, the consensus error dynamics (6.21) can be rewritten in terms of $\dot{\boldsymbol{\nu}}$ and $\mathbf{w}$ as

$$
\dot{\mathbf{e}} = \mathbf{A}_0^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}1}^{\otimes}\dot{\boldsymbol{\nu}} + \mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w}. \tag{6.28}
$$

Similarly, since $\boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes}$ is diagonal, its definition (6.20) allows the following decomposition of the optimal policy $\mathbf{u}^* = \big(\boldsymbol{\Gamma}_{\mathbf{v}}^{\otimes}\big)^{-1}\big(\mathbf{B}_{\mathbf{v}2}^{\otimes}\big)^T\boldsymbol{\Pi}\mathbf{e}$ (see Lemma 6.3 for this formula):

$$
\mathbf{u}^* = \dot{\boldsymbol{\nu}}^* + \mathbf{w}^*
$$

for $\dot{\boldsymbol{\nu}}^* \in \mathbb{R}^N$ and $\mathbf{w}^* \in \mathbb{R}^N$ defined by

$$
\begin{cases}
\dot{\boldsymbol{\nu}}^* := \frac{1}{\gamma}\big(\mathbf{B}_{\mathbf{v}1}^{\otimes}\big)^T\boldsymbol{\Pi}\mathbf{e} \\
\mathbf{w}^* := \frac{1}{\gamma}\cdot(\mathbf{D}_{\mathbf{v}}^2)^{-1}\big(\mathbf{B}_{\mathbf{v}2}^{\otimes}\big)^T\boldsymbol{\Pi}\mathbf{e},
\end{cases} \tag{6.29}
$$

where $\mathbf{D}_{\mathbf{v}} := \mathbf{diag}\{\|\mathbf{v}_1\|_2,\, \|\mathbf{v}_2\|_2,\, \cdots,\, \|\mathbf{v}_N\|_2\}$. Moreover, the HJB matrix equation (6.22) can be also decomposed as

$$
(\mathbf{A}_s^{\otimes})^T\boldsymbol{\Pi} + \boldsymbol{\Pi}\mathbf{A}_s^{\otimes} + \boldsymbol{\Theta} + \frac{1}{\gamma}\boldsymbol{\Pi}\,(\mathbf{B}_{\mathbf{v}1}^{\otimes})\big(\mathbf{B}_{\mathbf{v}1}^{\otimes}\big)^T\boldsymbol{\Pi} - \frac{1}{\gamma}\boldsymbol{\Pi}\,(\mathbf{B}_{\mathbf{v}2}^{\otimes})\mathbf{D}_{\mathbf{v}}^{-2}\big(\mathbf{B}_{\mathbf{v}2}^{\otimes}\big)^T\boldsymbol{\Pi} = \mathbf{0}_{4\times 4},
$$

where $\mathbf{A}_s^{\otimes} := \mathbf{A}_0^{\otimes} - \gamma^{-1}\mathbf{B}_{\mathbf{v}1}^{\otimes}\big(\mathbf{B}_{\mathbf{v}1}^{\otimes}\big)^T\mathbf{\Pi}$.

Now, consider the following $(\dot{\boldsymbol{\nu}}, \mathbf{w})$-dynamics obtained by combining the dynamic model (6.9) for all $i \in \mathcal{N}$:

$$\begin{cases} \dot{\boldsymbol{\nu}} = \boldsymbol{\rho}_\nu \\ \dot{\mathbf{w}} = \boldsymbol{\rho}_w, \end{cases}$$

where $\boldsymbol{\rho}_\nu := \mathbf{col}\{\rho_{1\nu},\, \rho_{2\nu},\, \cdots,\, \rho_{N\nu}\}$ and $\boldsymbol{\rho}_w := \mathbf{col}\{\rho_{1w},\, \rho_{2w},\, \cdots,\, \rho_{Nw}\}$. From this and (6.28), one can see that the optimal policy $\dot{\boldsymbol{\nu}}^*$ given in (6.29) can be directly applied by letting $\boldsymbol{\rho}_\nu = \dot{\boldsymbol{\nu}}^*$, but the other optimal one $\mathbf{w}^*$ in (6.29) does not due to the presence of the integrator $\dot{\mathbf{w}} = \boldsymbol{\rho}_w$. For this reason, it is desirable to set $\dot{\boldsymbol{\nu}}$ as the static control inputs and $\mathbf{w}$ as the dynamic controls (see Section 3.2 for these terminologies and the reiview of the inverse optimal input-dynamics extension technique). Hence, substituting $\dot{\boldsymbol{\nu}}^*$ given in (6.29) into the $\boldsymbol{\nu}$-dynamics "$\dot{\boldsymbol{\nu}} = \boldsymbol{\rho}_\nu$" by letting $\boldsymbol{\rho}_\nu = \boldsymbol{\nu}^*$ and rearranging the equations, we finally obtain the following partially-closed-loop dynamics:

$$\begin{cases} \dot{\mathbf{e}} = \mathbf{A}_s^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w} \\ \dot{\mathbf{w}} = \boldsymbol{\rho}_w, \end{cases} \tag{6.30}$$

which is the counterpart of the extended dynamics (3.8). Let $\bar{\mathbf{e}}_s := \mathbf{col}\{\mathbf{e}, \mathbf{w}\}$ and define $\bar{\mathbf{A}}_s^{\otimes}$ and $\bar{\mathbf{B}}_0^{\otimes}$ as

$$\bar{\mathbf{A}}_s^{\otimes} := \left[\begin{array}{c:c} \mathbf{A}_s^{\otimes} & \mathbf{B}_{\mathbf{v}2}^{\otimes} \\ \hdashline \mathbf{0}_{N\times 4N} & \mathbf{0}_{N\times N} \end{array}\right] \text{ and } \bar{\mathbf{B}}_0^{\otimes} := \left[\begin{array}{c} \mathbf{0}_{4N\times N} \\ \hdashline \mathbf{I}_N \end{array}\right].$$

Then, the system (6.30) can be rewritten as

$$\dot{\bar{\mathbf{e}}}_s = \bar{\mathbf{A}}_s^{\otimes}\,\bar{\mathbf{e}}_s + \bar{\mathbf{B}}_0^{\otimes}\boldsymbol{\rho}_w \tag{6.31}$$

Now, the application of Theorem 3.3 to the system (6.31) (or (6.30)) yields the following theorem that states the semi-global asymptotic stability and inverse optimality.

Theorem 6.2. *Supose the graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$ is simple, undirected, and connected; let* *the functions $Q_d^*(\mathbf{v};\lambda)$ and $\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)$ be defined as*

$$Q_d^*(\mathbf{v};\lambda) = \bar{\mathbf{e}}_s^T \mathbf{Q}_d(\mathbf{v};\lambda)\bar{\mathbf{e}}_s \quad \textit{and} \quad \bar{S}_d(\bar{\mathbf{e}}_s;\lambda) = \bar{\mathbf{e}}_s^T \bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda)\bar{\mathbf{e}}_s$$

with the matrices $\mathbf{Q}_d(\mathbf{v};\lambda)$ and $\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda)$ given by

$$\begin{aligned}
\mathbf{Q}_d(\mathbf{v};\lambda) &:= \left[\begin{array}{c:c} \lambda\mathbf{\Pi} & \star \\ \hdashline (\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi} & \gamma\mathbf{D}_{\mathbf{v}}^2 \end{array}\right] \\
\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda) &:= \left[\begin{array}{c:c} \lambda\mathbf{\Theta} - (\mathbf{A}_s^{\otimes})^T\mathbf{\Xi}^T\mathbf{\Pi} - \mathbf{\Pi}\mathbf{\Xi}\mathbf{A}_s^{\otimes} & \star \\ \hdashline -\big(\mathbf{\Pi}\mathbf{\Xi}^T + (\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi}\big)\mathbf{A}_s^{\otimes} & \mathbf{\Sigma}(\bar{\mathbf{e}}_s;\lambda) \end{array}\right],
\end{aligned}$$

$$\textit{where}\begin{cases}
\mathbf{\Xi}(\mathbf{v},\mathbf{w}) := \nabla\mathbf{B}_{\mathbf{v}2}^{\otimes}(\mathbf{x})\mathbf{w} \equiv \sum_{i=1}^N w_i\mathbf{I}_N \otimes \big\{\delta_{ij}\cdot\mathbf{B}_0\nabla_{\mathbf{e}}\mathbf{t}_2(\mathbf{v}_j)\big\}_{j=1}^N \\
\qquad (\delta_{ij} : \textit{the Kronecker delta function, i.e., } \delta_{ij} = 1 \textit{ for } i = j \textit{ and } \delta_{ij} = 0 \textit{ otherwise}); \\
\mathbf{\Sigma}(\bar{\mathbf{e}}_s;\lambda) := \lambda\gamma\mathbf{D}_{\mathbf{v}}^2 - 2(\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi}\mathbf{B}_{\mathbf{v}2}^{\otimes} - \mathbf{\Upsilon}(\bar{\mathbf{e}}_s); \\
\mathbf{\Upsilon}(\bar{\mathbf{e}}_s) := 2\gamma\,\mathbf{diag}\{\mathbf{v}_1^T[\mathbf{A}_s^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w}]_1, \ \cdots \ , \mathbf{v}_N^T[\mathbf{A}_s^{\otimes}\mathbf{e} + \mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w}]_N\} \\
\qquad ([\mathbf{z}]_k : \mathbf{y}_k \in \mathbb{R}^2 \textit{ in the vector } \mathbf{z} = \mathbf{col}\{\mathbf{x}_1, \mathbf{y}_1, \cdots \mathbf{x}_N, \mathbf{y}_N\} \in \mathbb{R}^{4N} \textit{ for } \mathbf{x}_i,\ \mathbf{y}_i \in \mathbb{R}^2).
\end{cases}$$

Let the control $\boldsymbol{\rho}_w$ in the mobile robot's partially-closed-loop dynamics (6.30) *be given by $\boldsymbol{\rho}_w = \boldsymbol{\rho}_w^*$, where $\boldsymbol{\rho}_w^*$ is a policy of the form*

$$\boldsymbol{\rho}_w^*(\bar{\mathbf{e}}_s;\lambda) = \lambda\cdot(\mathbf{w}^*(\mathbf{x}) - \mathbf{w}) \quad (\lambda > 0). \tag{6.32}$$

Then, under Assumptions 6.1, for any initial condition $\bar{\mathbf{e}}_s \in \mathbb{R}^{n+m_d}$, there exists $\underline{\lambda} > 0$ such that for all $\lambda \geq \underline{\lambda}$,

1. *$Q_d^*(\mathbf{v};\lambda)$ and $\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)$ given in the theorem are positive semi-definite and satisfies Assumption 3.3;*

2. *the policy $\boldsymbol{\rho}_w^*(\bar{\mathbf{e}}_s;\lambda)$ asymptotically stabilizes the system* (6.30) *to the extended subspace $\mathbb{S}_e \subset \mathbb{R}^{5N}$ given by*

   $$\mathbb{S}_e := \big\{\bar{\mathbf{e}}_s = (\mathbf{e},\mathbf{w}) \in \mathbb{R}^{5N} : \mathbf{e} \in \ker\mathbf{\Pi} \textit{ and } \mathbf{w} \equiv \mathbf{w}^*\big\}$$

3. *$\boldsymbol{\rho}_w^*(\bar{\mathbf{e}}_s;\lambda)$ is inverse optimal with respect to $J(\bar{\mathbf{e}}_s(t),\mathbf{v}(\cdot))$ given by*

   $$J(\bar{\mathbf{e}}_s(t),\dot{\mathbf{w}}(\cdot)) := \int_0^\infty \Big(\bar{S}_d(\bar{\mathbf{e}}_s;\lambda) + \lambda\gamma\,\|\dot{\boldsymbol{\nu}}^*(\mathbf{e})\|_2^2 + \lambda^{-1}\gamma\cdot\dot{\mathbf{w}}^T\mathbf{D}_{\mathbf{v}}^2\dot{\mathbf{w}}\Big)\,d\tau,$$

   and $Q_d^*(\bar{\mathbf{e}}_s;\lambda) = \bar{\mathbf{e}}_s^T\mathbf{Q}_d(\mathbf{v};\lambda)\bar{\mathbf{e}}_s$ is its corresponding optimal value function.

Proof. The proof will be done by showing the existence of the lower bounds on λ that guarantees Assumption 3.3 and then applying Theorem 3.3 with radially unbounded value function $Q_d^*(\bar{\mathbf{e}}_s; \lambda)$ for sufficiently large $\lambda > 0$. For a sketch of the proof, see Appendix D.5. □

By the standard optimal control theory, $Q_d^*(\bar{\mathbf{e}}_s; \lambda)$ and $\bar{S}_d(\bar{\mathbf{e}}_s; \lambda)$ in Theorem 6.2 satisfy the HJB equation $(\nabla Q_d^*)^T \bar{\mathbf{A}}_s^{\otimes} \bar{\mathbf{e}}_s - \frac{\lambda}{4}(\nabla Q_d^*)^T \bar{\mathbf{B}}_0^{\otimes} \mathbf{D}_{\mathbf{v}}^{-2} (\bar{\mathbf{B}}_0^{\otimes})^T \nabla Q_d^* + \bar{S}_d + \lambda\gamma \, \|\dot{\boldsymbol{\nu}}^*\|_2^2 = 0$, which can be rearranged using $\boldsymbol{\rho}_w^* = -\frac{\lambda}{2} \mathbf{D}_{\mathbf{v}}^{-2} (\bar{\mathbf{B}}_0^{\otimes})^T \nabla Q_d^*$ and the notation $\boldsymbol{\rho}_\nu^* := \dot{\boldsymbol{\nu}}^*$ as

$$(\nabla Q_d^*)^T \bar{\mathbf{f}}_c^{\otimes} + \bar{S}_d + \lambda\gamma \, \|\boldsymbol{\rho}_\nu^*\|_2^2 + \lambda^{-1} \, (\boldsymbol{\rho}_w^*)^T \mathbf{D}_v^{-2} \boldsymbol{\rho}_w^* = 0, \tag{6.33}$$

where $\bar{\mathbf{f}}_c^{\otimes} := \bar{\mathbf{A}}_s^{\otimes} \bar{\mathbf{e}}_s + \bar{\mathbf{B}}_0^{\otimes} \boldsymbol{\rho}_w^*$ is the nonlinear function of the closed-loop dynamics $\dot{\bar{\mathbf{e}}}_s = \bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s)$. Here, the closed-loop function $\bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s)$ can be represented as

$$\bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s) = \bar{\mathbf{A}}_0^{\otimes} \bar{\mathbf{e}}_s + \bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes} \boldsymbol{\rho}_\nu^* + \bar{\mathbf{B}}_0^{\otimes} \boldsymbol{\rho}_w^*, \tag{6.34}$$

where $\bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes} := \begin{bmatrix} \mathbf{B}_{\mathbf{v}1}^{\otimes} \\ \hdashline \mathbf{0}_{N \times N} \end{bmatrix}$. Similarly, the $\bar{\mathbf{e}}_s$-dynamics (6.31) can be expressed in terms of $\boldsymbol{\rho}_\nu$ and $\boldsymbol{\rho}_w$

$$\dot{\bar{\mathbf{e}}}_s = \bar{\mathbf{A}}_0^{\otimes} \bar{\mathbf{e}}_s + \bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes} \boldsymbol{\rho}_\nu + \bar{\mathbf{B}}_0^{\otimes} \boldsymbol{\rho}_w. \tag{6.35}$$

These equations (6.33) through (6.35) play a key role in deriving the adaptation laws in the next section.

6.4 Adaptive Inverse Optimal Graphical Formation Control

In the previous two sections, the inverse optimal policy for CGFC is proposed under the assumption that the mobile robots' parameters are perfectly known. The resulting inverse optimal policy $(\boldsymbol{\rho}_\nu^*, \boldsymbol{\rho}_w^*)$ are given by

$$\boldsymbol{\rho}_\nu^* = \dot{\boldsymbol{\nu}}^*(\mathbf{x}), \quad \boldsymbol{\rho}_w^* = \lambda \mathbf{w}^*(\mathbf{x}) - \lambda \mathbf{w}, \tag{6.36}$$

where $\dot{\boldsymbol{\nu}}^*$ and $\mathbf{w}^*$ are given by (6.29) and (6.29), respectively. Define $\boldsymbol{\tau}_\nu$ and $\boldsymbol{\tau}_w$ in $\mathbb{R}^N$ as

$$\boldsymbol{\tau}_\nu := \mathbf{col}\{\tau_{1\nu}, \tau_{2\nu}, \cdots, \tau_{N\nu}\} \text{ and } \boldsymbol{\tau}_w := \mathbf{col}\{\tau_{1w}, \tau_{2w}, \cdots, \tau_{Nw}\}$$

and rewrite the input terms (6.7) in an aggregated form:

$$\begin{cases} \mathbf{H}_1 \boldsymbol{\rho}_\nu = -\mathbf{D}_\nu \mathbf{c}_1 - \mathbf{D}_w^2 \mathbf{z}_1 + \boldsymbol{\tau}_\nu \\ \mathbf{H}_2 \boldsymbol{\rho}_w = -\mathbf{D}_w \mathbf{c}_2 - \mathbf{D}_\nu \mathbf{D}_w \mathbf{z}_2 + \boldsymbol{\tau}_w \end{cases} \tag{6.37}$$

where $\mathbf{D}_\nu := \mathbf{diag}\{\nu_1, \cdots, \nu_N\}$, $\mathbf{D}_w := \mathbf{diag}\{w_1, \cdots, w_N\}$, and

$$\begin{cases} \mathbf{z}_k := [\, z_{1k} \;\; z_{2k} \;\; \cdots \;\; z_{Nk} \,]^T \text{ with } z_{ik} = (-1)^k \alpha_i / r_i \\ \mathbf{c}_k := [\, c_{1k} \;\; c_{2k} \;\; \cdots \;\; c_{Nk} \,]^T \text{ with } c_{ik} = b_i R_i^{2(k-1)} / r_i \\ \mathbf{H}_k := \mathbf{diag}\{h_{1k}, h_{2k}, \cdots, h_{Nk}\} \text{ with } h_{ik} = J_{ik} R_i^{2(k-1)} / r_i. \end{cases}$$

we also define $\mathbf{h}_k$ $(k = 1, 2)$ as $\mathbf{h}_k := [\, h_{1k} \;\; h_{2k} \;\; \cdots \;\; h_{Nk} \,]^T$ for notational convenience. If the parameters in $\mathbf{c}_k$, $\mathbf{z}_k$, and $\mathbf{h}_k$ are perfectly known, then the control law given by

$$\begin{cases} \boldsymbol{\tau}_\nu^* = \mathbf{D}_\nu \mathbf{c}_1 + \mathbf{D}_w^2 \mathbf{z}_1 + \mathbf{H}_1 \boldsymbol{\rho}_\nu^* \\ \boldsymbol{\tau}_w^* = \mathbf{D}_w \mathbf{c}_2 + \mathbf{D}_\nu \mathbf{D}_w \mathbf{z}_2 + \mathbf{H}_2 \boldsymbol{\rho}_w^* \end{cases}$$

achieves the inverse optimal asymptotic stabilization to the subspace $\mathbb{S}_e$ as shown in the previous section. However, such an inverse optimal control law cannot be generated exactly if there are parametric uncertainties in $\mathbf{c}_k$, $\mathbf{z}_k$, and $\mathbf{h}_k$. In this section, we derive the adaptation laws to cancel out such uncertainties to improve the performance of the controlled system.

6.4.1 Derivations of Stabilizing Adaptation Laws

In what follows, let $\hat{\mathbf{c}}_k$, $\hat{\mathbf{z}}_k$, and $\hat{\mathbf{h}}_k$ $(k = 1, 2)$ be the estimated matrices of $\mathbf{c}_k$, $\mathbf{z}_k$, and $\mathbf{h}_k$, respectively, and consider the adaptive control inputs $\hat{\boldsymbol{\tau}}_\nu$ and $\hat{\boldsymbol{\tau}}_w$ given by

$$\begin{cases} \hat{\boldsymbol{\tau}}_\nu = \mathbf{D}_\nu \hat{\mathbf{c}}_1 + \mathbf{D}_w^2 \hat{\mathbf{z}}_1 + \hat{\mathbf{H}}_1 \boldsymbol{\rho}_\nu^* \\ \hat{\boldsymbol{\tau}}_w = \mathbf{D}_w \hat{\mathbf{c}}_2 + \mathbf{D}_\nu \mathbf{D}_w \hat{\mathbf{z}}_2 + \hat{\mathbf{H}}_2 \boldsymbol{\rho}_w^*, \end{cases} \tag{6.38}$$

where $\hat{\mathbf{H}}_k := \mathbf{diag}\{\hat{h}_{k1}, \cdots, \hat{h}_{kN}\}$ $(k = 1, 2)$ for the elements $\hat{h}_{ki}$ of $\hat{\mathbf{h}}_k$ $(i \in \mathcal{N})$. Define the diagonal matrices $\mathbf{D}_{\rho_\nu^*}$ and $\mathbf{D}_{\rho_w^*}$ as

$$\mathbf{D}_{\rho_\nu^*} := \mathbf{diag}\{\rho_{\nu 1}^*, \cdots, \rho_{\nu N}^*\} \text{ and } \mathbf{D}_{\rho_w^*} := \mathbf{diag}\{\rho_{w1}^*, \cdots, \rho_{wN}^*\}.$$

Then, they satisfy

$$\begin{cases} \mathbf{H}_1 \boldsymbol{\rho}_\nu^* = \mathbf{D}_{\rho_\nu^*} \mathbf{h}_1, \;\; \hat{\mathbf{H}}_1 \boldsymbol{\rho}_\nu^* = \mathbf{D}_{\rho_\nu^*} \hat{\mathbf{h}}_1, \\ \mathbf{H}_2 \boldsymbol{\rho}_w^* = \mathbf{D}_{\rho_w^*} \mathbf{h}_2, \;\; \hat{\mathbf{H}}_2 \boldsymbol{\rho}_w^* = \mathbf{D}_{\rho_w^*} \hat{\mathbf{h}}_2. \end{cases}$$

Now, substituting $\boldsymbol{\tau}_\nu = \hat{\boldsymbol{\tau}}_\nu$ and $\boldsymbol{\tau}_w = \hat{\boldsymbol{\tau}}_w$ into (6.37) and rearranging the equation using the above properties of $\mathbf{D}_{\rho_\nu^*}$ and $\mathbf{D}_{\rho_w^*}$ yields

$$\begin{cases} \boldsymbol{\rho}_\nu = \boldsymbol{\rho}_\nu^* + \mathbf{H}_1^{-1} \tilde{\boldsymbol{\tau}}_\nu \\ \boldsymbol{\rho}_w = \boldsymbol{\rho}_w^* + \mathbf{H}_2^{-1} \tilde{\boldsymbol{\tau}}_w, \end{cases}$$

where $\tilde{\boldsymbol{\tau}}_\nu := \hat{\boldsymbol{\tau}}_\nu - \boldsymbol{\tau}_\nu^*$ and $\tilde{\boldsymbol{\tau}}_w := \hat{\boldsymbol{\tau}}_w - \boldsymbol{\tau}_w^*$ are the control input errors and can be represented in terms of the parametric errors $\tilde{\mathbf{c}}_k := \hat{\mathbf{c}}_k - \mathbf{c}_k$, $\tilde{\mathbf{z}}_k := \hat{\mathbf{z}}_k - \mathbf{z}_k$, and $\tilde{\mathbf{h}}_k := \hat{\mathbf{h}}_k - \mathbf{h}_k$ as

$$\begin{cases} \tilde{\boldsymbol{\tau}}_\nu = \mathbf{D}_\nu \tilde{\mathbf{c}}_1 + \mathbf{D}_w^2 \tilde{\mathbf{z}}_1 + \mathbf{D}_{\rho_\nu^*} \tilde{\mathbf{h}}_1 \\ \tilde{\boldsymbol{\tau}}_w = \mathbf{D}_w \tilde{\mathbf{c}}_2 + \mathbf{D}_\nu \mathbf{D}_w \tilde{\mathbf{z}}_2 + \mathbf{D}_{\rho_w^*} \tilde{\mathbf{h}}_2. \end{cases} \tag{6.39}$$

Now, substituting $\boldsymbol{\rho}_\nu = \boldsymbol{\rho}_\nu^* + \mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu$ and $\boldsymbol{\rho}_w = \boldsymbol{\rho}_w^* + \mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w$ into (6.35) and using (6.34), we finally obtain the closed-loop dynamics

$$\dot{\bar{\mathbf{e}}}_s = \bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s) + \bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes}\mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu + \bar{\mathbf{B}}_0^{\otimes}\mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w. \tag{6.40}$$

In the next theorem, the adaptation laws of the parameters in the torque input errors $(\tilde{\boldsymbol{\tau}}_\nu, \tilde{\boldsymbol{\tau}}_w)$ are given from the Lyapunov analysis. For the statement, define the parameter error vector $\mathbf{p} \in \mathbb{R}^{6N}$ as $\mathbf{p} := \mathbf{col}\{\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2\}$.

Theorem 6.3. *Suppose the graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$ is simple, undirected, and connected. Let the torque control $(\boldsymbol{\tau}_\nu, \boldsymbol{\tau}_w)$ be given by $\boldsymbol{\tau}_\nu = \hat{\boldsymbol{\tau}}_\nu$ and $\boldsymbol{\tau}_w = \hat{\boldsymbol{\tau}}_w$ for $(\hat{\boldsymbol{\tau}}_\nu, \hat{\boldsymbol{\tau}}_w)$ given in* (6.38)*, with the adaptation laws*

$$\begin{cases} \dot{\hat{\mathbf{c}}}_1 = -\mathbf{G}_{11}\mathbf{D}_\nu\boldsymbol{\phi}_1, & \dot{\hat{\mathbf{z}}}_1 = -\mathbf{G}_{12}\mathbf{D}_w^2\boldsymbol{\phi}_1, & \dot{\hat{\mathbf{h}}}_1 = -\mathbf{G}_{13}\mathbf{D}_{\rho_\nu^*}\boldsymbol{\phi}_1, \\ \dot{\hat{\mathbf{c}}}_2 = -\mathbf{G}_{21}\mathbf{D}_w\boldsymbol{\phi}_2, & \dot{\hat{\mathbf{z}}}_2 = -\mathbf{G}_{22}\mathbf{D}_{\nu w}\boldsymbol{\phi}_2, & \dot{\hat{\mathbf{h}}}_2 = -\mathbf{G}_{23}\mathbf{D}_{\rho_w^*}\boldsymbol{\phi}_2, \end{cases} \tag{6.41}$$

where $\mathbf{G}_{kj}$'s are diagonal positive gain matrices defined as $\mathbf{G}_{kj} := \mathbf{diag}\{g_{kj}^{(1)}, g_{kj}^{(2)}, \cdots, g_{kj}^{(N)}\}$ with each adaptation gain $g_{kj}^{(i)} > 0$ $(i \in \mathcal{N})$, and $\boldsymbol{\phi}_k(\bar{\mathbf{e}}_s)$ $(k = 1, 2)$ are vector-valued functions defined as

$$\boldsymbol{\phi}_1(\bar{\mathbf{e}}_s) := \frac{1}{2}(\mathbf{B}_{\mathbf{v}1}^{\otimes})^T\nabla_{\mathbf{e}}Q_d^*(\bar{\mathbf{e}}_s;\lambda) \quad \text{and} \quad \boldsymbol{\phi}_2(\bar{\mathbf{e}}_s) := \frac{1}{2}\nabla_{\mathbf{w}}Q_d^*(\bar{\mathbf{e}}_s;\lambda).$$

Then, under Assumptions 6.1, the adaptive policy $(\hat{\boldsymbol{\tau}}_\nu, \hat{\boldsymbol{\tau}}_w)$ stabilizes the adaptive system (6.40) *and* (6.41) *with respect to the extended adaptive subspace $\mathbb{S}_e^a \subset \mathbb{R}^{11N}$ defined as*

$$\mathbb{S}_e^a := \big\{(\mathbf{e}, \mathbf{w}, \mathbf{p}) \in \mathbb{R}^{(4+1+6)N} : \mathbf{e} \in \ker\boldsymbol{\Pi},\ \mathbf{w} \equiv \mathbf{0},\ \text{and } \mathbf{p} = \mathbf{0}_{6N}\big\}. \tag{6.42}$$

Moreover, for all $i \in \mathcal{N}$, the consensus $\|\mathbf{q}_i(\tau) - \mathbf{q}_j(\tau) - \mathbf{d}_{ij}\| \to 0$, $\|\mathbf{v}_i(\tau) - \mathbf{v}_g\| \to 0$, and $w_i(\tau) \to 0$, and the parameter convergence $\hat{c}_{i1}(\tau) \to c_{i1}$ are achieved in the limit $\tau \to \infty$.

Proof. For the proof, consider the Lyapunov function $V(\bar{\mathbf{e}}_s, \mathbf{p}; \lambda)$ given by

$$V(\bar{\mathbf{e}}_s, \mathbf{p}; \lambda) = \frac{1}{2}Q_d^*(\bar{\mathbf{e}}_s;\lambda) + \frac{1}{2}\sum_{k=1}^{2}\Big(\tilde{\mathbf{c}}_k^T\bar{\mathbf{H}}_{k1}^{-1}\tilde{\mathbf{c}}_k + \tilde{\mathbf{z}}_k^T\bar{\mathbf{H}}_{k2}^{-1}\tilde{\mathbf{z}}_k + \tilde{\mathbf{h}}_k^T\bar{\mathbf{H}}_{k3}^{-1}\tilde{\mathbf{h}}_k\Big),$$

where $\lambda > 0$ is sufficiently large to guarantee both Q_d^* and $\bar{S}_d$ given in Theorem 6.2 are positive semi-definite and satisfy Assumption 3.3; $\bar{\mathbf{H}}_{kj}$ $(k = 1, 2$ and $j = 1, 2, 3)$ are defined as $\bar{\mathbf{H}}_{kj} := \mathbf{G}_{kj}^{1/2}\mathbf{H}_k\mathbf{G}_{kj}^{1/2}$. Since both $\mathbf{G}_{kj}$ and $\mathbf{H}_k$ are diagonal positive definite, $\bar{\mathbf{H}}_{kj}$ can be expressed as

$$\bar{\mathbf{H}}_{kj} = \mathbf{H}_k\mathbf{G}_{kj} = \mathbf{G}_{kj}\mathbf{H}_k = \mathbf{G}_{kj}^{1/2}\mathbf{H}_k\mathbf{G}_{kj}^{1/2}. \tag{6.43}$$

First, note that the time-derivative of $Q_d^*(\bar{\mathbf{e}}_s;\lambda)$ along the trajectory generated by (6.40) satisfies

$$\begin{aligned}\frac{1}{2}\dot{Q}_d^*(\bar{\mathbf{e}}_s;\lambda) =& \frac{1}{2}\nabla Q_d^{*T}(\bar{\mathbf{e}}_s;\lambda)\cdot\big(\bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s) + \bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes}\mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu + \bar{\mathbf{B}}_0^{\otimes}\mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w\big)\\ \leq& -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda) + \frac{1}{2}\nabla Q_d^{*T}(\bar{\mathbf{e}}_s;\lambda)\cdot\bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes}\mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu + \frac{1}{2}\nabla Q_d^{*T}(\bar{\mathbf{e}}_s;\lambda)\cdot\bar{\mathbf{B}}_0^{\otimes}\mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w\\ =& -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda) + \frac{1}{2}\nabla_{\mathbf{e}} Q_d^{*T}(\bar{\mathbf{e}}_s;\lambda)\mathbf{B}_{\mathbf{v}1}^{\otimes}\mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu + \frac{1}{2}\nabla_{\mathbf{w}} Q_d^{*T}(\bar{\mathbf{e}}_s;\lambda)\mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w,\end{aligned}$$

where the HJB equation (6.33) is substituted. using the vector-valued functions $\boldsymbol{\phi}_k(\bar{\mathbf{e}}_s)$ $(k = 1, 2)$ defined in this theorem, the result can be compactly written as

$$\frac{1}{2}\dot{Q}_d^*(\bar{\mathbf{e}}_s;\lambda) \leq -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda) + \boldsymbol{\phi}_1^T\mathbf{H}_1^{-1}\tilde{\boldsymbol{\tau}}_\nu + \boldsymbol{\phi}_2^T\mathbf{H}_2^{-1}\tilde{\boldsymbol{\tau}}_w. \tag{6.44}$$

Next, we differentiate $V(\bar{\mathbf{e}}_s,\mathbf{p};\lambda)$ with respect to time and substitute (6.39), (6.43), and (6.44) as follows:

$$\begin{aligned}\dot{V}(\bar{\mathbf{e}}_s,\mathbf{p};\lambda) =& \dot{Q}_d^*(\bar{\mathbf{e}}_s;\lambda) + \sum_{k=1}^{2}\Big(\dot{\hat{\mathbf{c}}}_k^T\bar{\mathbf{H}}_{k1}^{-1}\tilde{\mathbf{c}}_k + \dot{\hat{\mathbf{z}}}_k^T\bar{\mathbf{H}}_{k2}^{-1}\tilde{\mathbf{z}}_k + \dot{\hat{\mathbf{h}}}_k^T\bar{\mathbf{H}}_{k3}^{-1}\tilde{\mathbf{h}}_k\Big)\\ \leq& -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda) + \sum_{k=1}^{2}\cdot\Big(\dot{\hat{\mathbf{c}}}_k^T\mathbf{G}_{k1}^{-1}\mathbf{H}_k^{-1}\tilde{\mathbf{c}}_k + \dot{\hat{\mathbf{z}}}_k^T\mathbf{G}_{k1}^{-1}\mathbf{H}_k^{-1}\tilde{\mathbf{z}}_k + \dot{\hat{\mathbf{h}}}_k^T\mathbf{G}_{k1}^{-1}\mathbf{H}_k^{-1}\tilde{\mathbf{h}}_k\Big)\\ &+ \boldsymbol{\phi}_1^T\mathbf{H}_1^{-1}\Big(\mathbf{D}_\nu\tilde{\mathbf{c}}_1 + \mathbf{D}_w^2\tilde{\mathbf{z}}_1 + \mathbf{D}_{\rho_\nu^*}\tilde{\mathbf{h}}_1\Big) + \boldsymbol{\phi}_2^T\mathbf{H}_2^{-1}\Big(\mathbf{D}_w\tilde{\mathbf{c}}_2 + \mathbf{D}_\nu\mathbf{D}_w\tilde{\mathbf{z}}_2 + \mathbf{D}_{\rho_w^*}\tilde{\mathbf{h}}_2\Big).\end{aligned}$$

Since $\mathbf{D}_{\rho_\nu^*}$, $\mathbf{D}_{\rho_w^*}$, $\mathbf{D}_\nu$, $\mathbf{D}_w$, $\mathbf{H}_k$ $(k = 1, 2)$ are all diagonal positive definite, they and their inverses are all commutable, which results in

$$\begin{aligned}&\dot{V}(\bar{\mathbf{e}}_s,\mathbf{p};\lambda)\\ &\leq -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)\\ &\quad + \big(\mathbf{D}_\nu\boldsymbol{\phi}_1 + \mathbf{G}_{11}^{-1}\dot{\hat{\mathbf{c}}}_1\big)^T\mathbf{H}_1^{-1}\tilde{\mathbf{c}}_1 + \big(\mathbf{D}_w^2\boldsymbol{\phi}_1 + \mathbf{G}_{12}^{-1}\dot{\hat{\mathbf{z}}}_1\big)^T\mathbf{H}_1^{-1}\tilde{\mathbf{z}}_1 + \big(\mathbf{D}_{\rho_\nu^*}\boldsymbol{\phi}_1 + \mathbf{G}_{13}^{-1}\dot{\hat{\mathbf{h}}}_1\big)^T\mathbf{H}_1^{-1}\tilde{\mathbf{h}}_1\\ &\quad + \big(\mathbf{D}_w\boldsymbol{\phi}_2 + \mathbf{G}_{21}^{-1}\dot{\hat{\mathbf{c}}}_2\big)^T\mathbf{H}_2^{-1}\tilde{\mathbf{c}}_2 + \big(\mathbf{D}_{\nu w}\boldsymbol{\phi}_2 + \mathbf{G}_{22}^{-1}\dot{\hat{\mathbf{z}}}_2\big)^T\mathbf{H}_2^{-1}\tilde{\mathbf{z}}_2 + \big(\mathbf{D}_{\rho_w^*}\boldsymbol{\phi}_2 + \mathbf{G}_{23}^{-1}\dot{\hat{\mathbf{h}}}_2\big)^T\mathbf{H}_2^{-1}\tilde{\mathbf{h}}_2.\end{aligned}$$

where $\mathbf{D}_{\nu w} := \mathbf{D}_\nu\mathbf{D}_w$. Therefore, we choose the adaptation laws as (6.41), which results in $\dot{V}(\bar{\mathbf{e}}_s,\mathbf{p};\lambda) \leq -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)$. Since $\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)$ satisfies Assumption 3.3 for sufficiently large

$\lambda > 0$ by Theorem 6.2, we have

$$\dot{V}(\bar{\mathbf{e}}_s, \mathbf{p}; \lambda) \leq -\frac{1}{2}\bar{S}_d(\bar{\mathbf{e}}_s; \lambda) \leq -\frac{1}{2}\underline{\alpha}_s(d(\bar{\mathbf{e}}_s; \mathbb{S}_e)) \preceq 0, \tag{6.45}$$

where $\mathbb{S}_e$ is the subspace defined in Theorem 6.2, and the application of Theorem 3.1 proves that $(\hat{\boldsymbol{\tau}}_\nu, \hat{\boldsymbol{\tau}}_w)$ with the adaptation laws (6.41) stabilizes the equilibrium subspace $\mathbb{S}_e^a \subset \mathbb{R}^{11N}$, defined in (6.42) , of the adaptive system for sufficiently large $\lambda > 0$. This implies that $\boldsymbol{\eta}$ $(= \mathbf{v} - \mathbf{v}_g)$, $\mathbf{w}$, and $\mathbf{p}$ are bounded, and $\mathbf{e}$ is bounded away from $\mathbb{S} = \ker \mathbf{\Pi}$. That is, $\bar{\mathbf{e}}_s$ is bounded away from $\mathbb{S}_e$ and $\mathbf{p}$ is bounded for all time. Since $V(\bar{\mathbf{e}}_s, \mathbf{p}; \lambda)$ is monotonically decreasing by (6.45) and lower-bounded by 0, it has a finite limit. Integrating $\dot{V} \leq -\bar{S}_d$ from t to $t + \tau$, we have

$$\int_t^{t+\tau} \bar{S}_d(\bar{\mathbf{e}}_s(\tau); \lambda) d\tau \leq V(\bar{\mathbf{e}}_s(t), \mathbf{p}(t); \lambda) - V(\bar{\mathbf{e}}_s(t+\tau), \mathbf{p}(t+\tau); \lambda),$$

so the limit $\lim_{\tau \to \infty} \int_t^{t+\tau} \bar{S}_d(\bar{\mathbf{e}}_s(\tau); \lambda)\, d\tau$ exists and is finite.

To apply Barbalat's lemma for the proof of the convergence, we establish the uniform continuity of $\bar{S}_d(\bar{\mathbf{e}}_s(\tau); \lambda)$ by showing that its time derivative is bounded. Since

$$\bar{\mathbf{e}}_s \in \mathbb{S}_e \implies \bar{\mathbf{A}}_0^{\otimes} \bar{\mathbf{e}}_s \in \mathbb{S}_e, \;\; \boldsymbol{\rho}_\nu^*(\mathbf{e}) = \mathbf{0}_{4N}, \text{ and } \boldsymbol{\rho}_w^*(\mathbf{e}) = \mathbf{0}_{4N},$$

and $\bar{\mathbf{B}}_{\mathbf{v}1}^{\otimes}$, we have $\bar{\mathbf{f}}_c^{\otimes}(\bar{\mathbf{e}}_s) \in \mathbb{S}_e$ whenever $\bar{\mathbf{e}}_s \in \mathbb{S}_e$ by (6.34). Moreover, since $\tilde{\boldsymbol{\tau}}_\nu$ and $\tilde{\boldsymbol{\tau}}_w$ are all bounded, (6.40) implies that $\dot{\bar{\mathbf{e}}}_s$ is bounded away from $\mathbb{S}_e$ whenever $\bar{\mathbf{e}}_s \in \mathbb{S}_e$. Also note that the matrices and their time derivatives in Theorem 6.2 are all bounded as shown below.

1. $\mathbf{\Xi}$ and its time derivative are bounded since $\mathbf{w}$ and $\dot{\mathbf{w}}$ are bounded, and

$$d\mathbf{t}_2(\mathbf{v}_j)/d\mathbf{v}_j = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

2. $\mathbf{\Upsilon}$ and its time derivative are all bounded since so are $\mathbf{v}_i$, $\mathbf{w}$, $\dot{\mathbf{w}}$, $\mathbf{A}_s^{\otimes}\mathbf{x}$, $\mathbf{B}_{\mathbf{v}2}^{\otimes}$ and all of their time derivatives.

3. Similarly, $\mathbf{\Sigma}$ and its time derivative are bounded since so are its matrix components.

(see the definitions and structures of the matrices in Theorem 6.2). By the above argument, it is established that the matrix $\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s; \lambda)$ and its time derivative are bounded. The time derivative of $\bar{S}_d(\bar{\mathbf{e}}_s;\lambda)$ is given by

$$\dot{\bar{S}}_d = 2\dot{\bar{\mathbf{e}}}_s^T\bar{\mathbf{S}}_d\bar{\mathbf{e}}_s + \bar{\mathbf{e}}_s^T\dot{\bar{\mathbf{S}}}_d\bar{\mathbf{e}}_s,$$

where $\dot{\bar{\mathbf{S}}}_d$ is given by

$$\dot{\bar{\mathbf{S}}}_d(\bar{\mathbf{e}}_s;\lambda) = \left[\begin{array}{c:c} -(\mathbf{A}_s^{\otimes})^T\dot{\mathbf{\Xi}}^T\mathbf{\Pi} - \mathbf{\Pi}\dot{\mathbf{\Xi}}\mathbf{A}_s^{\otimes} & \star \\ \hdashline -\big(\mathbf{\Pi}\dot{\mathbf{\Xi}}^T + (\dot{\mathbf{B}}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi}\big)\mathbf{A}_s^{\otimes} & \dot{\mathbf{\Sigma}}(\bar{\mathbf{e}}_s;\lambda) \end{array}\right].$$

Obvously, the structures of $\mathbf{S}_d$ and $\dot{\mathbf{S}}_d$ show that

$$\bar{\mathbf{e}}_s \in \mathbb{S}_e \implies \bar{\mathbf{e}}_s^T\dot{\bar{\mathbf{S}}}_d\bar{\mathbf{e}}_s = 0 \text{ and } \dot{\bar{\mathbf{e}}}_s^T\underbrace{\bar{\mathbf{S}}_d\bar{\mathbf{e}}_s}_{=\mathbf{0}_{5N}} = 0$$

Since $\bar{\mathbf{e}}_s$ and $\dot{\bar{\mathbf{e}}}_s$ are bounded away from $\mathbb{S}_e$, $\dot{\bar{S}}_d(\bar{\mathbf{e}}_s(\tau),\lambda)$ is bounded, which implies that $\bar{S}_d(\bar{\mathbf{e}}_s(\tau),\lambda)$ is uniformly continuous. Therefore, the application of Barbalat's lemma [32, Lemma 8.2] proves that $\bar{S}_d(\bar{\mathbf{e}}_s(\tau),\lambda) \to 0$ as $\tau \to \infty$, which implies that $d(\bar{\mathbf{e}}_s(\tau),\mathbb{S}_e) \to 0$ as $\tau \to \infty$, so the consensus $\|\mathbf{q}_i(\tau) - \mathbf{q}_j(\tau) - \mathbf{d}_{ij}\| \to 0$, $\|\mathbf{v}_i(\tau) - \mathbf{v}_g\| \to 0$, and $w_i(\tau) \to 0$ is achieved as $\tau \to \infty$. Furthermore, in the limit $\tau \to \infty$, we have $\mathbf{v}_i = \mathbf{v}_g$, $\dot{\boldsymbol{\eta}}_i = \mathbf{0}_2$, $w_i = 0$, $\rho_{i\nu}^* = 0$ for all $i \in \mathcal{N}$, so the velocity dynamics $\dot{\boldsymbol{\eta}}_{\boldsymbol{i}} = \mathbf{T}(\mathbf{v}_i)[\,\dot{\nu}_i \;\; w_i\,]^T$ becomes in the limit

$$\begin{aligned}\mathbf{0}_2 &= \mathbf{t}_1(\mathbf{v}_g)\dot{\nu}_i^{ss} \\ &= \mathbf{t}_1(\mathbf{v}_g)\cdot J_{i1}^{-1}(-b_i\nu_i^{ss} + \alpha_i(w_i^{ss})^2 + r_i\hat{\tau}_{i\nu}^{ss}) \\ &= \mathbf{t}_1(\mathbf{v}_g)\cdot h_{i1}\big(-\tilde{c}_{i1}^{ss}\nu_i^{ss} - \tilde{z}_{i1}^{ss}(w_i^{ss})^2 + \hat{h}_{i1}^{ss}\rho_{i\nu}^*\big) \\ &= -h_{i1}\cdot\mathbf{t}_1(\mathbf{v}_g)\|\mathbf{v}_g\|_2\cdot\tilde{c}_{i1}^{ss},\end{aligned}$$

where the superscript '*ss*' means '*steady-state*' and indicates the value or a time function of the variables *in the steady-state or in the limit* $\tau \to \infty$. Since $\mathbf{t}_1(\mathbf{v}_g) \neq \mathbf{0}_2$ and $\|\mathbf{v}_g\|_2 \neq 0$, the above equality in the limit necessarily implies $\tilde{c}_{i1}^{ss} = 0$, so $\hat{c}_{i1}(\tau) \to c_{i1}$ as $\tau \to \infty$. □

6.4.2 Decentralization and Simplification of Adaptation Laws

The stabilizing adaptation laws (6.41) in Theorem 6.3 is centralized and contains the complicated global vector functions $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$. In this subsection, we decentralize and simplify the adaptation laws (6.41) by analyzing $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ under the positive definite matrices $\mathbf{Q}$ and $\mathbf{Q}_v$ in the ARE (6.15) given by

$$\mathbf{Q} = \mathbf{diag}\{q_1\mathbf{I}_2, q_2\mathbf{I}_2\} \text{ and } \mathbf{Q}_v = q_v\mathbf{I}_2 \tag{6.46}$$

for some positive constants q_1, q_2, $q_v > 0$. As the first step, the next lemma states the matrix-components $\mathbf{P}_{11}$, $\mathbf{P}_{12}$, $\mathbf{P}_{22} \in \mathbb{R}^{2\times 2}$ of the decomposition

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \hline \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{array}\right] \tag{6.47}$$

can be expressed as as $\mathbf{P}_{11} = p_{11}\,\mathbf{I}_2$, $\mathbf{P}_{12} = p_{12}\,\mathbf{I}_2$, and $\mathbf{P}_{22} = p_{22}\,\mathbf{I}_2$ for some positive constants p_{11}, p_{12}, and p_{22}, which dramatically simplifies the proposed adaptation laws.

Lemma 6.5. *Under $\mathbf{Q}$ and $\mathbf{Q}_v$ given by* (6.46)*, $\mathbf{P}_{11}$, $\mathbf{P}_{12}$, and $\mathbf{P}_{22}$ in the decomposition* (6.47) *of the solution $\mathbf{P}$ of the ARE* (6.15) *are expressed in the diagonal forms*

$$\mathbf{P}_{11} = p_{11}\,\mathbf{I}_2, \mathbf{P}_{12} = p_{12}\,\mathbf{I}_2, \text{ and } \mathbf{P}_{22} = p_{22}\,\mathbf{I}_2$$

for p_{11}, p_{12}, $p_{22} > 0$ given by

$$\begin{cases} p_{12} = \sqrt{\gamma q_1}, \\ p_{22} = q_v\big(-1 + \sqrt{1 + \gamma(2p_{12} + q_2)/q_v^2}\big) = q_v\big(-1 + \sqrt{1 + \gamma(2\sqrt{\gamma q_1} + q_2)/q_v^2}\big), \\ p_{11} = \gamma^{-1}p_{12}(q_v + p_{22}) = \gamma^{-1}q_v\sqrt{\gamma q_1}\cdot\sqrt{1 + \gamma(2\sqrt{\gamma q_1} + q_2)/q_v^2}. \end{cases}$$

Proof. Substituting the block-wise expressions of the matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{Q}_v$, $\mathbf{A}_c$, and $\mathbf{B}_0$ into the ARE (6.15), we obtain

$$\begin{aligned} \mathbf{0}_{4\times 4} =& \mathbf{A}_c^T\mathbf{P} + \mathbf{P}\,\mathbf{A}_c + \mathbf{Q} - \frac{1}{\gamma}\,\mathbf{P}\,\mathbf{B}_0\mathbf{B}_0^T\mathbf{P} \\ =& \left[\begin{array}{c|c} \mathbf{0}_{2\times 2} & \star \\ \hline \mathbf{P}_{11} - \gamma^{-1}q_v\mathbf{P}_{12}^T & \mathbf{P}_{12}^T + \mathbf{P}_{12} - 2\gamma^{-1}q_v\mathbf{P}_{22} \end{array}\right] + \left[\begin{array}{c|c} q_1\mathbf{I}_2 & \star \\ \hline \mathbf{0}_{2\times 2} & q_2\mathbf{I}_2 \end{array}\right] - \frac{1}{\gamma}\cdot\left[\begin{array}{c|c} \mathbf{P}_{12}\mathbf{P}_{12}^T & \star \\ \hline \mathbf{P}_{22}\mathbf{P}_{12}^T & \mathbf{P}_{22}^2 \end{array}\right] \\ =& \left[\begin{array}{c|c} q_1\mathbf{I}_2 - \gamma^{-1}\mathbf{P}_{12}\mathbf{P}_{12}^T & \star \\ \hline \mathbf{P}_{11} - \gamma^{-1}q_v\mathbf{P}_{12}^T - \gamma^{-1}\mathbf{P}_{22}\mathbf{P}_{12}^T & \mathbf{P}_{12}^T + \mathbf{P}_{12} + q_2\mathbf{I}_2 - 2\gamma^{-1}q_v\mathbf{P}_{22} - \gamma^{-1}\mathbf{P}_{22}^2 \end{array}\right]. \end{aligned}$$

From the $(1,1)$-th block, we have $\mathbf{P}_{12} = \mathbf{P}_{12}^T = p_{12}\,\mathbf{I}_2$ with $p_{12} = \sqrt{\gamma q_1}$. Next, substituting $\mathbf{P}_{12} = \mathbf{P}_{12}^T = p_{12}\mathbf{I}_2$ and $\mathbf{P}_{22} = p_{22}\mathbf{I}_2$ into the $(2,2)$-th block yields the quadratic equation

$$p_{22}^2 + 2q_vp_{22} - \gamma(2p_{12} + q_2) = 0$$

whose positive solution exactly matches with the expression of p_{22}. Finally, substituting $\mathbf{P}_{12}^T = p_{12}\mathbf{I}_2$ and $\mathbf{P}_{22} = p_{22}\mathbf{I}_2$ into the $(2,1)$-th block completes the proof. □

Next, noticing that the Q-function $Q_d^*(\bar{\mathbf{e}}_s;\lambda) = \bar{\mathbf{e}}_s^T\mathbf{Q}_d(\mathbf{v};\lambda)\bar{\mathbf{e}}_s$ given in Theorem 6.2 is expressed as $Q_d^*(\bar{\mathbf{e}}_s;\lambda) = \lambda\mathbf{e}^T\mathbf{\Pi}\mathbf{e} + 2\mathbf{e}^T\mathbf{\Pi}\mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w} + \gamma\mathbf{w}^T\mathbf{D}_{\mathbf{v}}^2\mathbf{w}$, its partial derivatives $\nabla_{\mathbf{e}}Q_d^*$ and $\nabla_{\mathbf{w}}Q_d^*$ that are shown in $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, respectively, are expressed as

$$\frac{1}{2}\nabla_{\mathbf{e}}Q_d^* = \lambda\mathbf{\Pi}\mathbf{e} + \mathbf{\Pi}\mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w} + (\nabla_{\mathbf{e}}\mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w})^T\mathbf{\Pi}\mathbf{e} + \gamma\cdot(\nabla_{\mathbf{e}}\mathbf{w}\mathbf{D}_{\mathbf{v}}^2\mathbf{w})/2, \tag{6.48}$$

$$\frac{1}{2}\nabla_{\mathbf{w}}Q_d^* = (\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi}\mathbf{e} + \gamma\,\mathbf{D}_{\mathbf{v}}^2\mathbf{w}. \tag{6.49}$$

By the block-matrix operations, the terms in (6.48) can be deployed as follows.

$$\begin{aligned}
\lambda\mathbf{\Pi}\mathbf{e} =& \lambda\cdot\left[(\mathbf{L}\otimes\mathbf{P}) + (\mathbf{I}_N\otimes\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T)\right]\cdot\mathbf{col}\{\mathbf{e}_1,\cdots,\mathbf{e}_N\}\\
=&\lambda\begin{bmatrix} l_{11}\mathbf{P} & \cdots & l_{1N}\mathbf{P}\\ \vdots & \ddots & \vdots\\ l_{N1}\mathbf{P} & \cdots & l_{NN}\mathbf{P}\end{bmatrix}\begin{bmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_N\end{bmatrix} + \lambda\begin{bmatrix}\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T & & \\ & \ddots & \\ & & \mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T\end{bmatrix}\begin{bmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_N\end{bmatrix}\\
=&\begin{bmatrix}\sum_{j=1}^N l_{1j}\lambda\mathbf{P}\mathbf{e}_j + \lambda\mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_1\\ \vdots\\ \sum_{j=1}^N l_{Nj}\lambda\mathbf{P}\mathbf{e}_j + \lambda\mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_N\end{bmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathbf{\Pi}\mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w} =& \mathbf{\Pi}\cdot\mathbf{diag}\{\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_1),\cdots,\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_N)\}\cdot\mathbf{col}\{w_1,\cdots,w_N\}\\
=&\left(\begin{bmatrix} l_{11}\mathbf{P} & \cdots & l_{1N}\mathbf{P}\\ \vdots & \ddots & \vdots\\ l_{N1}\mathbf{P} & \cdots & l_{NN}\mathbf{P}\end{bmatrix} + \begin{bmatrix}\mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T & & \\ & \ddots & \\ & & \mathbf{B}_0\mathbf{Q}_v\mathbf{B}_0^T\end{bmatrix}\right)\begin{bmatrix}\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_1)w_1\\ \vdots\\ \mathbf{B}_0\mathbf{t}_2(\mathbf{v}_N)w_N\end{bmatrix}\\
=&\begin{bmatrix}\sum_{j=1}^N l_{1j}\mathbf{P}\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_j)w_j + \mathbf{B}_0\mathbf{Q}_v\mathbf{t}_2(\mathbf{v}_1)w_1\\ \vdots\\ \sum_{j=1}^N l_{Nj}\mathbf{P}\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_j)w_j + \mathbf{B}_0\mathbf{Q}_v\mathbf{t}_2(\mathbf{v}_N)w_N\end{bmatrix},
\end{aligned}$$

$$
\begin{aligned}
(\nabla_{\mathbf{e}}\mathbf{B}_{\mathbf{v}2}^{\otimes}\mathbf{w})^T\boldsymbol{\Pi}\mathbf{e} &= \begin{bmatrix} w_1\nabla_{\mathbf{e}}\mathbf{t}_2^T(\mathbf{v}_1)\mathbf{B}_0^T & \cdots & w_N\nabla_{\mathbf{e}}\mathbf{t}_2^T(\mathbf{v}_N)\mathbf{B}_0^T \end{bmatrix}\boldsymbol{\Pi}\mathbf{e} \\
&= \begin{bmatrix} w_1\mathbf{F} & & \\ & \ddots & \\ & & w_N\mathbf{F} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^N l_{1j}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_1 \\ \vdots \\ \sum_{j=1}^N l_{Nj}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_N \end{bmatrix} \\
&= \begin{bmatrix} w_1\left(\sum_{j=1}^N l_{1j}\mathbf{F}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{J}\mathbf{Q}_v\boldsymbol{\eta}_1\right) \\ \vdots \\ w_N\left(\sum_{j=1}^N l_{Nj}\mathbf{F}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{J}\mathbf{Q}_v\boldsymbol{\eta}_N\right) \end{bmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\gamma}{2}\cdot(\nabla_{\mathbf{e}}\mathbf{w}\mathbf{D}_{\mathbf{v}}^2\mathbf{w}) =& \frac{\gamma}{2}\cdot\nabla_{\mathbf{e}}\left(\sum_{i=1}^N(\mathbf{v}_i^T\mathbf{v}_i)w_i^2\right) \\
=& \gamma\cdot\mathbf{col}\{\mathbf{B}_0\mathbf{v}_1w_1^2,\cdots,\mathbf{B}_0\mathbf{v}_Nw_N^2\},
\end{aligned}
$$

where $\mathbf{J}\in\mathbb{R}^{2\times 2}$ and $\mathbf{F}\in\mathbb{R}^{4\times 4}$ are defined as

$$
\mathbf{J} := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ and } \mathbf{F} := \left[\begin{array}{c} \mathbf{0}_{2\times 4} \\ \hdashline \nabla_{\mathbf{v}_i}\mathbf{t}_2^T(\mathbf{v}_i)\mathbf{B}_0^T \end{array}\right] = \left[\begin{array}{c:c} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} \\ \hdashline \mathbf{0}_{2\times 2} & \mathbf{J} \end{array}\right].
$$

Hence, the regression function $\boldsymbol{\phi}_1 = \frac{1}{2}(\mathbf{B}_{\mathbf{v}1}^{\otimes})^T\nabla_{\mathbf{e}}Q_d^*$ can be written as

$$
\boldsymbol{\phi}_1 \equiv \left[\begin{array}{c} \phi_{11} \\ \hdashline \phi_{12} \\ \hdashline \vdots \\ \hdashline \phi_{1N} \\ \hdashline \end{array}\right] = \left[\begin{array}{c} \sum_{j=1}^N l_{1j}\psi^{[1]}(\mathbf{v}_1,\mathbf{e}_j,w_j) + \psi^{[2]}(\mathbf{e}_1,w_1) \\ \hdashline \sum_{j=1}^N l_{2j}\psi^{[1]}(\mathbf{v}_2,\mathbf{e}_j,w_j) + \psi^{[2]}(\mathbf{e}_2,w_2) \\ \hdashline \vdots \\ \hdashline \sum_{j=1}^N l_{Nj}\psi^{[1]}(\mathbf{v}_N,\mathbf{e}_j,w_j) + \psi^{[2]}(\mathbf{e}_N,w_N) \end{array}\right],
$$

$$
\text{where } \begin{cases} \psi^{[1]}(\mathbf{v}_i,\mathbf{e}_j,w_i,w_j) := [\,\mathbf{0}_2^T\;\;\mathbf{t}_1^T(\mathbf{v}_i)\,]\big((w_i\mathbf{F}+\lambda\mathbf{I}_4)\mathbf{P}\mathbf{e}_j + w_j\mathbf{P}\mathbf{B}_0\mathbf{t}_2(\mathbf{v}_j)\big) \\ \psi^{[2]}(\mathbf{e}_i,w_i) := [\,\mathbf{0}_2^T\;\;\mathbf{t}_1^T(\mathbf{v}_i)\,]\big(\mathbf{B}_0(w_i\mathbf{J}+\lambda\mathbf{I}_2)\mathbf{Q}_v\boldsymbol{\eta}_i + w_i\mathbf{B}_0\mathbf{Q}_v\mathbf{t}_2(\mathbf{v}_i) + \gamma\mathbf{B}_0\mathbf{v}_iw_i^2\big). \end{cases}
$$

Similarly, $(\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\boldsymbol{\Pi}\mathbf{e}$ and $\gamma\,\mathbf{D}_{\mathbf{v}}^2\mathbf{w}$ in (6.49) can be rearranged as follows.

$$(\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\boldsymbol{\Pi}\mathbf{e} = \begin{bmatrix} \mathbf{t}_2^T(\mathbf{v}_1)\mathbf{B}_0^T & & \\ & \ddots & \\ & & \mathbf{t}_2^T(\mathbf{v}_N)\mathbf{B}_0^T \end{bmatrix} \begin{bmatrix} \sum_{j=1}^N l_{1j}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_1 \\ \vdots \\ \sum_{j=1}^N l_{Nj}\mathbf{P}\mathbf{e}_j + \mathbf{B}_0\mathbf{Q}_v\boldsymbol{\eta}_N \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{t}_2^T(\mathbf{v}_1)\left(\sum_{j=1}^N l_{1j}\mathbf{B}_0^T\mathbf{P}\mathbf{e}_j + \mathbf{Q}_v\boldsymbol{\eta}_1\right) \\ \vdots \\ \mathbf{t}_2^T(\mathbf{v}_N)\left(\sum_{j=1}^N l_{Nj}\mathbf{B}_0^T\mathbf{P}\mathbf{e}_j + \mathbf{Q}_v\boldsymbol{\eta}_N\right) \end{bmatrix},$$

$$\begin{aligned} \gamma\,\mathbf{D}_{\mathbf{v}}^2\mathbf{w} &= \gamma\cdot\mathbf{diag}\{\|\mathbf{v}_1\|_2^2,\cdots,\|\mathbf{v}_N\|_2^2\}\cdot\mathbf{col}\{w_1,\cdots,w_N\} \\ &= \mathbf{col}\{\gamma\cdot\|\mathbf{v}_1\|_2^2\,w_1,\ \cdots,\ \gamma\cdot\|\mathbf{v}_N\|_2^2\,w_N\}, \end{aligned}$$

which results in the following expression of $\boldsymbol{\phi}_2$:

$$\boldsymbol{\phi}_2 \equiv \begin{bmatrix} \phi_{21} \\ \hdashline \phi_{22} \\ \hdashline \vdots \\ \hdashline \phi_{2N} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_2^T(\mathbf{v}_1)\big(\sum_{j=1}^N l_{1j}\mathbf{B}_0^T\mathbf{P}\mathbf{e}_j + \mathbf{Q}_v\boldsymbol{\eta}_1\big) + \gamma\|\mathbf{v}_1\|_2^2 w_1 \\ \hdashline \mathbf{t}_2^T(\mathbf{v}_2)\big(\sum_{j=1}^N l_{2j}\mathbf{B}_0^T\mathbf{P}\mathbf{e}_j + \mathbf{Q}_v\boldsymbol{\eta}_2\big) + \gamma\|\mathbf{v}_2\|_2^2 w_2 \\ \hdashline \vdots \\ \hdashline \mathbf{t}_2^T(\mathbf{v}_N)\big(\sum_{j=1}^N l_{Nj}\mathbf{B}_0^T\mathbf{P}\mathbf{e}_j + \mathbf{Q}_v\boldsymbol{\eta}_N\big) + \gamma\|\mathbf{v}_N\|_2^2 w_N \end{bmatrix}.$$

Now, let $\mathbf{t}_{1j} \equiv \mathbf{t}_1(\mathbf{v}_j)$ and $\mathbf{t}_{2j} \equiv \mathbf{t}_2(\mathbf{v}_j)$ for simplicity. Considering the expression (6.46) of $\mathbf{Q}$ and $\mathbf{Q}_v$ and revoking Lemma 6.5, one can see that $\mathbf{P}$ is expressed as

$$\mathbf{P} = \left[\begin{array}{c:c} p_{11}\mathbf{I}_2 & p_{12}\mathbf{I}_2 \\ \hdashline p_{12}\mathbf{I}_2 & p_{22}\mathbf{I}_2 \end{array}\right];$$

the block-matrix multiplications with the substitutions of this, $\mathbf{t}_{1i}^T\mathbf{t}_{2i} = 0$, $\mathbf{t}_{1i}^T\mathbf{v}_i = \nu_i$, and $\nu_i = \|\mathbf{v}_i\|_2$ by Assumption 6.1 yields

$$\begin{cases} \psi^{[1]}(\mathbf{v}_i, \mathbf{e}_j, w_j) = \mathbf{t}_{1i}^T\big(w_i\mathbf{J} + \lambda\mathbf{I}_2\big)\big(p_{12}\boldsymbol{\delta}_j + p_{22}\boldsymbol{\eta}_j\big) \\ \psi^{[2]}(\mathbf{e}_i, w_i) = q_v\mathbf{t}_{1i}^T(w_i\mathbf{J} + \lambda\mathbf{I}_2)\boldsymbol{\eta}_i + \gamma\|\mathbf{v}_i\|_2 w_i^2. \end{cases}$$

Hence, ϕ_{1i} can be written as

$$\phi_{1i} = \mathbf{t}_{1i}^T(w_i\mathbf{J} + \lambda\mathbf{I}_2)\bigg(\sum_{j=1}^N a_{ij}\Big(p_{12}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j) + p_{22}(\mathbf{v}_i - \mathbf{v}_j)\Big) + q_v\boldsymbol{\eta}_i\bigg) + \gamma\|\mathbf{v}_i\|w_i^2.$$

In a similar way, ϕ_{2i} can be expressed as

$$\phi_{2i} = \mathbf{t}_{2i}^T\bigg(\sum_{j=1}^N a_{ij}\Big(\mathbf{P}_{12}^T(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j) + \mathbf{P}_{22}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_j)\Big) + \mathbf{Q}_v\boldsymbol{\eta}_i\bigg) + \gamma\|\mathbf{v}_i\|_2^2 w_i.$$

Noting that $\gamma^{-1}\mathbf{T}^{-1}(\mathbf{v}_i) = \boldsymbol{\Gamma}^{-1}(\mathbf{v}_i)\mathbf{T}^T(\mathbf{v}_i)$ and the expression of $\mathbf{u}_i^* = (\dot{\nu}_i^*, w_i^*)$ given in (6.13) and (6.14) with $\mathbf{K} = \gamma^{-1}\mathbf{B}_0^T\mathbf{P}$ show that $\dot{\nu}_i^*$ and w_i^* can be represented in terms of $\mathbf{t}_{1i}$ and $\mathbf{t}_{2i}$ as

$$\begin{aligned} \dot{\nu}_i^* &= -\frac{\mathbf{t}_{1i}^T}{\gamma}\bigg(\sum_{j=1}^N a_{ij}\Big(p_{12}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j) + p_{22}(\mathbf{v}_i - \mathbf{v}_j)\Big) + q_v\boldsymbol{\eta}_i\bigg), \\ w_i^* &= -\frac{\mathbf{t}_{2i}^T}{\gamma\|\mathbf{v}_i\|_2^2}\bigg(\sum_{j=1}^N a_{ij}\Big(p_{12}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j) + p_{22}(\mathbf{v}_i - \mathbf{v}_j)\Big) + q_v\boldsymbol{\eta}_i\bigg). \end{aligned} \tag{6.50}$$

Hence, the w_i^*-expression in (6.50) dramatically simplifies ϕ_{2i} as

$$\phi_{2i} = \gamma\|\mathbf{v}_i\|_2^2\,(w_i - w_i^*).$$

Furthermore, noting that $\mathbf{t}_{1i}^T\mathbf{J} = \|\mathbf{v}_i\|_2^{-1}\mathbf{t}_{2i}^T$ and using (6.50), ϕ_{1i} can be simplified in a similar manner as

$$\begin{aligned} \phi_{1i} &= -\gamma(\|\mathbf{v}_i\|_2 w_i w_i^* + \lambda\dot{\nu}_i^*) + \gamma\|\mathbf{v}_i\|w_i^2 \\ &= -\lambda\gamma\dot{\nu}_i^* + \gamma w_i\|\mathbf{v}\|_2(w_i - w_i^*). \end{aligned}$$

Since the matrices in (6.38) and (6.41) are all diagonal, the control and adaptation laws can be written agent-wisely. Table 6.1 summarizes the final control law $(\tau_{i\nu}, \tau_{iw})$ equipped with the derived adaptation laws. While the derivations were complex, we finally obtain the simple expressions of both control and adaptation laws in a decentralized manner as shown in the table. Moreover, by (6.6), the actual torque inputs to the left and right wheels

Table 6.1: Adaptive inverse optimal control & adaptation laws with design equations

Control Laws

$$\begin{cases} \hat{\tau}_{iL} = \hat{\tau}_{i\nu} + \hat{\tau}_{iw} \\ \hat{\tau}_{iR} = \hat{\tau}_{i\nu} - \hat{\tau}_{iw} \end{cases} \text{with} \begin{cases} \hat{\tau}_{i\nu} = \hat{c}_{i1}\nu_i + \hat{z}_{i1} w_i^2 + \hat{h}_{i1}\dot{\nu}_i^* \\ \hat{\tau}_{iw} = \hat{c}_{i2} w_i + \hat{z}_{i2}\nu_i w_i + \lambda \hat{h}_{i2}(w_i^* - w_i) \end{cases}$$

Contol Parameters

* λ: a sufficiently large positive constant.

* $\mathbf{u}_i^* = (\dot{\nu}_i^*, w_i^*)$: the inverse optimal policy given by

$$\begin{bmatrix} \dot{\nu}_i^* \\ w_i^* \end{bmatrix} = -\frac{1}{\gamma}\mathbf{T}^{-1}(\mathbf{v}_i)\left(\sum_{j\in\mathcal{N}_i} a_{ij}\big(p_{12}(\mathbf{q}_i - \mathbf{q}_j - \mathbf{d}_{ij}) + p_{22}(\mathbf{v}_i - \mathbf{v}_j)\big) + q_v\left(\mathbf{v}_i - \mathbf{v}_g\right)\right).$$

Adaptation Laws

$$\begin{aligned} \dot{\hat{c}}_{1i} &= -g_{11}^{(i)}\phi_{1i}\nu_i, & \dot{\hat{z}}_{1i} &= -g_{12}^{(i)}\phi_{1i} w_i^2, & \dot{\hat{h}}_{1i} &= -g_{13}^{(i)}\phi_{1i}\dot{\nu}_i^*, \\ \dot{\hat{c}}_{2i} &= -g_{21}^{(i)}\phi_{2i} w_i, & \dot{\hat{z}}_{2i} &= -g_{22}^{(i)}\phi_{2i}\nu_i w_i, & \dot{\hat{h}}_{2i} &= -\lambda g_{23}^{(i)}\phi_{2i}(w_i^* - w_i), \end{aligned}$$

Adaptation Parameters

* $\mathbf{G}^{(i)} = \begin{bmatrix} g_{11}^{(i)} & g_{12}^{(i)} & g_{13}^{(i)} \\ g_{21}^{(i)} & g_{22}^{(i)} & g_{23}^{(i)} \end{bmatrix}$ with $g_{kj}^{(i)} > 0$: the adaptation gain matrix.

* (ϕ_{1i}, ϕ_{2i}): the regression function given by

$$\begin{aligned} \phi_{1i} &= -\lambda\gamma\dot{\nu}_i^* + \gamma w_i\|\mathbf{v}_i\|_2(w_i - w_i^*). \\ \phi_{2i} &= \gamma\|\mathbf{v}_i\|_2^2(w_i - w_i^*) \end{aligned}$$

where $\mathbf{T}(\mathbf{v}_i) = \left[\mathbf{t}_{1i} \,\vdots\, \mathbf{t}_{2i}\right]$.

can be obtained as $\tau_{iL} = \tau_{i\nu} - \tau_{iw}$ and $\tau_{iR} = \tau_{i\nu} + \tau_w$, respectively, as shown in Table 6.1.

6.5 Simulation Results

To verify the performance of the proposed adaptive inverse optimal control laws equipped with the adaptation laws both shown in Table 6.1, numerical simulations are carried out with the three mobile robots ($\mathcal{N} = \{1, 2, 3\}$) whose kinematic and dynamic models are given by (6.1) and (6.2), respectively. The parameters of the mobile robots and their true values used in the simulations are given in Table 6.2. In the simulations, their initial estimates are sampled by the following uniform distributions:

$$\begin{aligned}
&\hat{R}_i \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))R_i, && \hat{d}_i \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))d_i, && \hat{r}_i \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))r_i,\\
&\hat{m}_{ic} \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))m_{ic}, && \hat{m}_{iw} \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))m_{iw}, && \hat{b}_i \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))b_i,\\
&\hat{I}_i^c \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))I_i^c, && \hat{I}_i^w \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))I_i^w, && \hat{I}_i^m \sim (\sigma_d + \sigma_s \mathcal{U}(0,1))I_i^m,
\end{aligned}$$

where $\mathcal{U}(0,1)$ denotes the uniform distribution over the interval $[0,1]$, $\sigma_d = 0.5$, and $\sigma_s = 1$. From these estimates, the initial estimates of the control parameters $\hat{c}_{ik}$, $\hat{z}_{ik}$, and $\hat{h}_{ik}$ $(i \in \mathcal{N},\ k = 1,2)$ are calculated as

$$\hat{z}_{ik}(0) = (-1)^k \hat{\alpha}_i/\hat{r}_i, \quad \hat{c}_{ij}(0) = \hat{b}_i \hat{R}_i^{2(k-1)}/\hat{r}_i, \quad \hat{h}_{ik}(0) = \hat{J}_{ik}\hat{R}_i^{2(k-1)}/\hat{r}_i,$$

where $\hat{\alpha}_i :=:= \hat{r}_i^2 \hat{m}_{ic}\,\hat{d}_i/2$, $\hat{J}_{i1} := \hat{I}_i^w + \hat{I}_i/2\hat{R}_i^2$, and $\hat{J}_{i2} := \hat{I}_i^w + \hat{r}_i^2\hat{m}_i/2$. If the estimates of the mobile robots' parameter are exactly same to their true values, then so are the initial parameter estimates $\hat{c}_{ik}$, $\hat{z}_{ik}$, and $\hat{h}_{ik}$ in the adaptive optimal CGFC laws. In the simulations, the graph Laplacian $\mathbf{L}$ describing the communication topology among the mobile robots is given by

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix},$$

whose minimal positive eigenvalue $\lambda_2(\mathbf{L})$ called *the algebraic connectivity of the undirected graph* satisfies $\lambda_2(\mathbf{L}) = 1$, so (6.24) in Theorem 6.1 holds.

The matrices $\mathbf{Q}$ and $\mathbf{Q}_v$ and the positive constant γ in the ARE (6.15) are set to $\mathbf{Q} = \mathbf{diag}\{q_1\mathbf{I}_2, q_2\mathbf{I}_2\}$, $\mathbf{Q}_v = q_2\mathbf{I}_2$, and $\gamma = 1$ with $q_1 = q_2 = q_v = 1$, which yields the solution matrix $\mathbf{P} = \left[\begin{array}{c:c} p_{11}\mathbf{I}_2 & p_{12}\mathbf{I}_2 \\ \hdashline p_{12}\mathbf{I}_2 & p_{22}\mathbf{I}_2 \end{array}\right]$ with $p_{11} = 2$, $p_{12} = p_{22} = 1$ by Lemma 6.5. Hence, the optimal policy $(\dot{\nu}_i^*, w_i^*)$ in the simulations is given by

$$\begin{bmatrix} \dot{\nu}_i^* & w_i^* \end{bmatrix}^T = -\mathbf{T}^{-1}(\mathbf{v}_i)\Bigg(\sum_{j\in\mathcal{N}_i} a_{ij}\big((\mathbf{q}_i + \mathbf{v}_i) - (\mathbf{q}_j + \mathbf{v}_j) - \mathbf{d}_{ij}\big) + (\mathbf{v}_i - \mathbf{v}_g)\Bigg).$$

In the simulations, $\lambda > 0$ and the adaptation gains $g_{jk}^{(i)}$ are set to $\lambda = 10$ and $g_{jk}^{(i)} = 1$ for all $j = 1,2,3$, $k = 1,2$, and $i \in \mathcal{N}$; the initial poses $\boldsymbol{p}_i^0 = (x_i(0), y_i(0), \theta_i(0))$ and initial

Table 6.2: The parameters of the mobile robots in the simulation

Mobile Robot Parameters	True Values in Mobile Robot 1	True Values in Mobile Robot 2	True Values in Mobile Robot 3	Units
R_i	0.75	1.5	2	[m]
d_i	0.3	0.5	0.7	[m]
r_i	0.15	0.4	0.5	[m]
m_{ic}	30	50	55	[kg]
m_{iw}	1	2.5	3.5	[kg]
b_i	5	10	6.5	[kg·m^2/s]
I_i^c	15.625	40.625	35	[kg·m^2]
I_i^w	0.005	0.01	0.02	[kg·m^2]
I_i^m	0.0025	0.005	0.00625	[kg·m^2]

velocities $\boldsymbol{\xi}_i^0 = (\nu_i(0), w_i(0))$ of the mobile robots are assumed to be given by

$$\boldsymbol{p}_1^0 = (0, 0, \pi/6), \qquad \boldsymbol{p}_2^0 = (0, 1, 0), \qquad \boldsymbol{p}_3^0 = (1, 0, -\pi/6),$$
$$\boldsymbol{\xi}_1^0 = (2, 0), \qquad \boldsymbol{\xi}_2^0 = (1, 0), \qquad \boldsymbol{\xi}_3^0 = (3, 0).$$

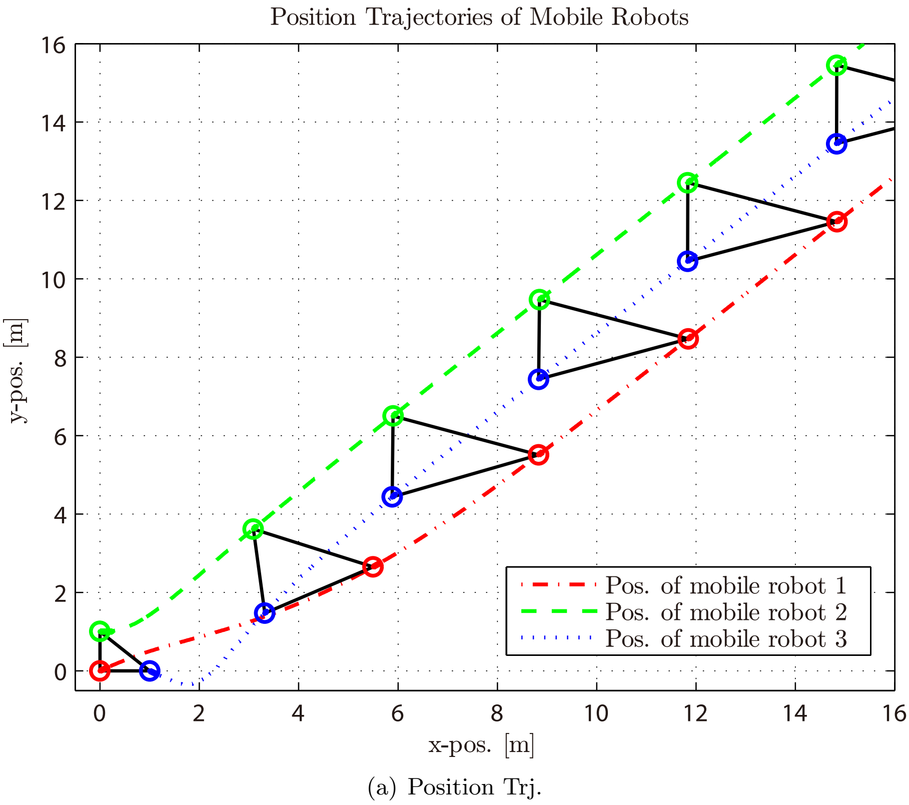
6.5.1 Simulation Example 1: Simple Case

At first, we consider the simple case, where the distance vectors $\mathbf{d}_i$ of the mobile robots and the group velocity $\mathbf{v}_g = [v_{g,x} \;\; v_{g,y}]^T$ are given by

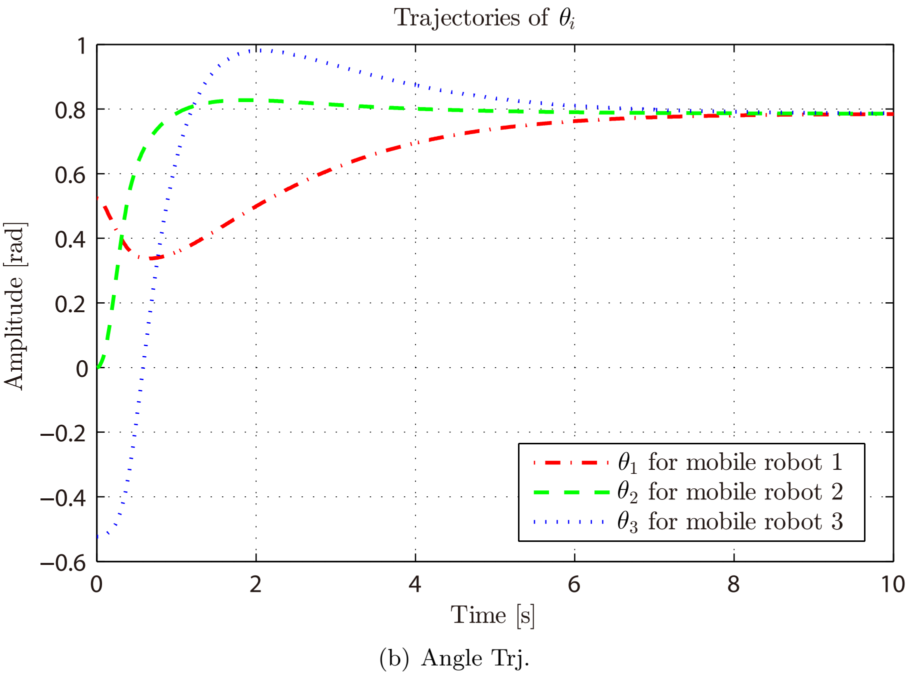
$$\mathbf{d}_1 = [3 \;\; 0]^T, \qquad \mathbf{d}_2 = [0 \;\; 1]^T, \qquad \mathbf{d}_3 = [0 \; -1]^T, \qquad \mathbf{v}_g = [1 \;\; 1]^T,$$

respectively. The simulation result for some samples of the initial parameter estimates is described in Figs. 6.2, 6.3, and 6.4, where Fig. 6.2 describes the trajectories of the poses of mobile robots, Fig. 6.3 llustrates the variations of mobile robots' linear and angular velocities (ν_i, w_i), and Fig. 6.4 is the parameter variations of $\hat{c}_{i1}$ in the mobile robots.

As shown in Figs. 6.2 and 6.3(a), the mobile robots driven by the proposed adaptive inverse optimal CGFC scheme shape and maintain the desired formation and ultimately move along the same group velocity $\mathbf{v}_g$. Notice that the group velocity $\mathbf{v}_g$ in this case can

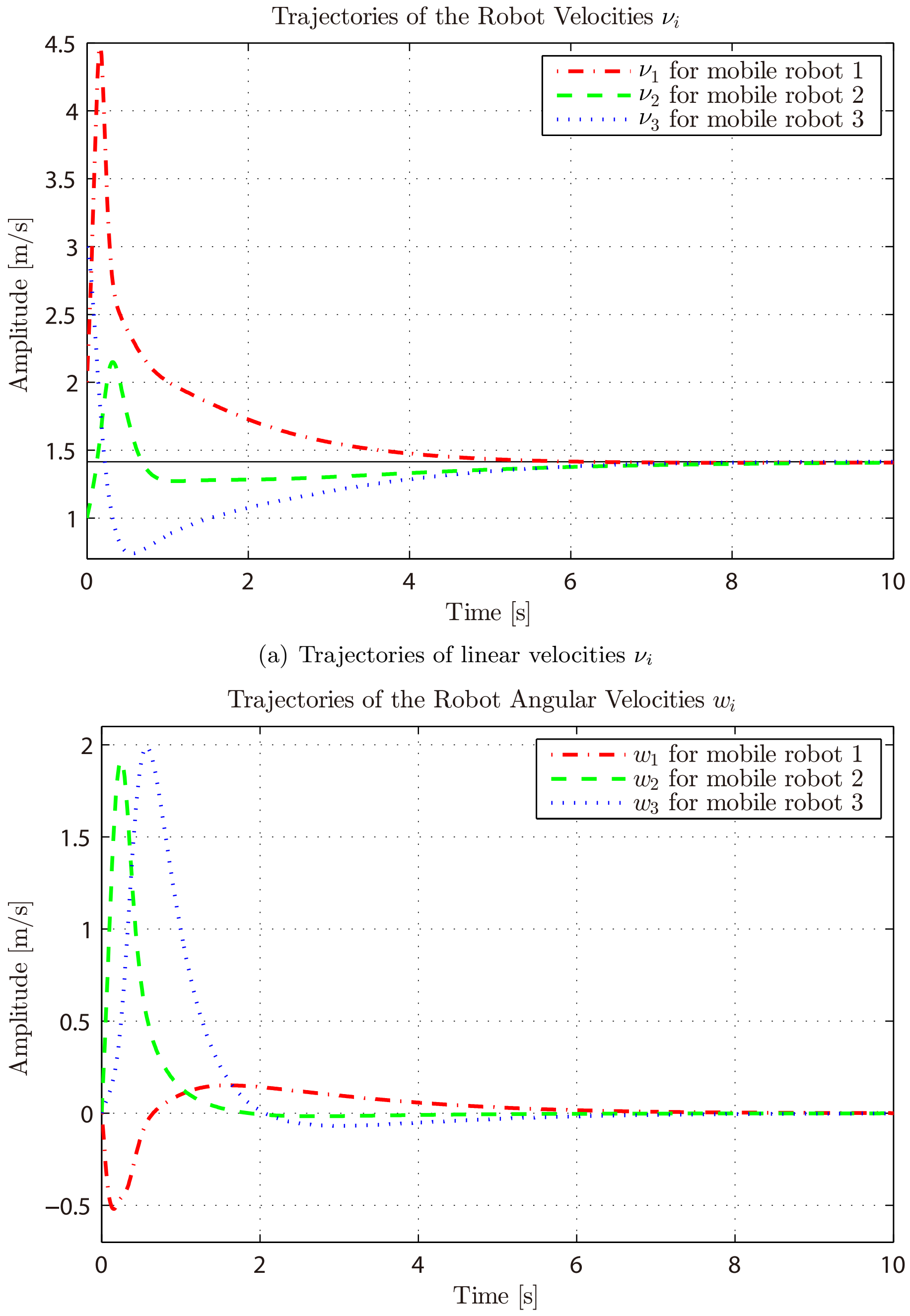


(a) Position Trj.



(b) Angle Trj.

Figure 6.2: **(Example 1)** Position and angle trajectories of mobile robots.



(a) Trajectories of linear velocities ν_i

(b) Trajectories of angular velocities w_i

Figure 6.3: **(Example 1)** The velocities (ν_i, w_i) of mobile robots: (a) ν_i and (b) w_i.

be rewritten as

$$\mathbf{v}_g = \sqrt{2} \cdot [\,\cos(\pi/4) \quad \sin(\pi/4)\,]^T \approx 1.414 \cdot [\,\cos(0.7854) \quad \sin(0.7854)\,]^T.$$

From Fig.6.2 and 6.3(a), one can see that the final angle orientations and the final group speeds of the robots are all approximately equal to $\pi/4$ and $\sqrt{2}$, respectively, implying the convergence $\mathbf{v}_i \to \mathbf{v}_g$ for all $i = 1, 2, 3$. As shown in Fig. 6.3(b) and Fig. 6.4, the angular velocity w_i and the paraeter estimate errors "$\hat{c}_{i1} - c_{i1}$" all converges to zeros, which coincides with Theorem 6.3. In Fig. 6.4, the three black solid lines indicate the respective true values c_{i1} of the estimates $\hat{c}_{i1}$. Here, note that

1. $w_i(\tau)$ should be necessarily zero in the limit $\tau \to \infty$ to drive the mobile robots finally with the constant same angle orientation $\theta \equiv \lim_{\tau\to\infty} \theta_i(\tau)$;

2. $\hat{c}_{i1}(\tau)$ becomes necessarily equal to c_{i1} in the velocity consensus $\mathbf{v}_i(\tau) \to \mathbf{v}_g$, and Theorem 6.3 states both $\hat{c}_{i1}(\tau) \to c_{i1}$ and $\mathbf{v}_i \to \mathbf{v}_g$ are achieved in the limit $\tau \to \infty$ without any additional conditions; unlike the arguments regarding the parameter convergence in the standard adaptive control theories [31, 55], the parameters $\hat{c}_{i1}$ converges to their true values *without any persistently exciting conditions* while the others $\hat{c}_{i2}$, $\hat{z}_{ik}$ and $\hat{h}_{ik}$ $(k = 1, 2)$, though not shown in the Figs., are just bounded by the stability argument in Theorem 6.3.

6.5.2 Simulation Example 2: Management of $\mathbf{v}_g$ and $\mathbf{d}_{ij}$'s

In this simluation, the profile of the group velocity $\mathbf{v}_g$ is given by

$$\mathbf{v}_g = \begin{cases} \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} & \text{for } 0 \leq \tau < 20 \text{ [s]}, \\ \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} + s_{20}(\tau) \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} & \text{for } 20 \text{ [s]} \leq \tau < 40 \text{ [s]}, \\ \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} + s_{40}(\tau) \begin{bmatrix} 0.0 \\ 0.5 \end{bmatrix} & \text{for } 40 \text{ [s]} \leq \tau < 60 \text{ [s]}, \\ \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} - s_{60}(\tau) \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} & \text{for } 60 \text{ [s]} \leq \tau < 80 \text{ [s]}, \\ \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} + s_{80}(\tau) \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} & \text{for } 80 \text{ [s]} \leq \tau < 100 \text{ [s]}, \\ \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - s_{100}(\tau) \begin{bmatrix} 0.0 \\ 0.5 \end{bmatrix} & \text{for } 100 \text{ [s]} \leq \tau \leq 120 \text{ [s]}, \end{cases} \tag{6.51}$$

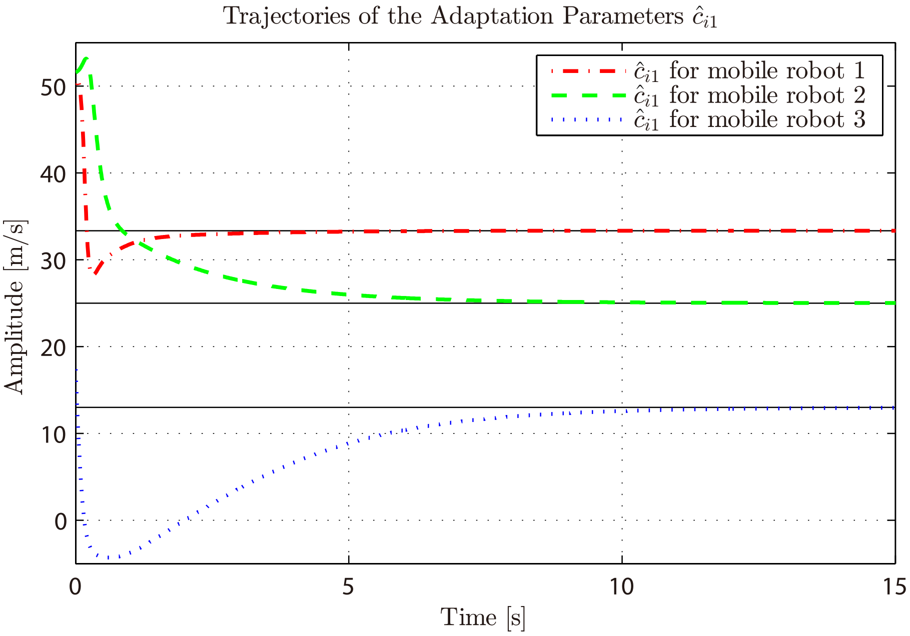
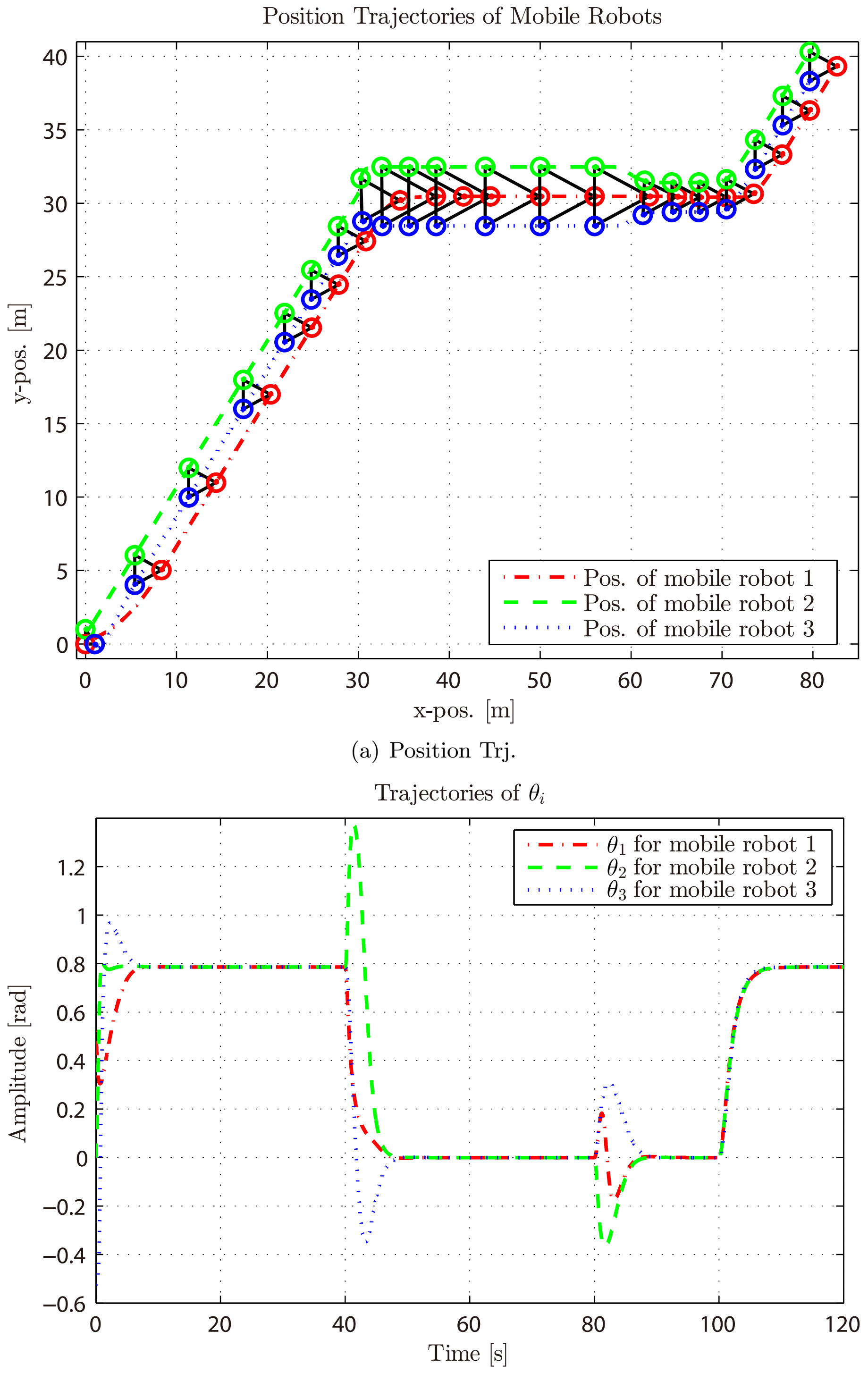


Figure 6.4: **(Example 1)** Trajectories of the parameter estimates $\hat{c}_{i1}$

Table 6.3: The profile of the desired formation $\mathbf{d}_{ij} = \mathbf{x}_{ij} - s_t(\tau)\mathbf{y}_{ij}$ in the simulation

$\mathbf{d}_{ij}$ $(= \mathbf{d}_{ji})$	Time interval	$\mathbf{x}_{ij}$	$\mathbf{y}_{ij}$
$\mathbf{d}_{12}$	0 [s] $\le \tau \le 40$ [s]	$[\;3\;\;-1\,]^T$	$[\;0\;\;\;0\,]^T$
	40 [s] $\le \tau \le 80$ [s]	$[\;6\;\;-2\,]^T$	$[\;3\;\;-1\,]^T$
	80 [s] $\le \tau \le 120$ [s]	$[\;0\;\;\;2\,]^T$	$[-3\;\;\;1\,]^T$
$\mathbf{d}_{23}$	0 [s] $\le \tau \le 40$ [s]	$[\;0\;\;\;2\,]^T$	$[\;0\;\;\;0\,]^T$
	40 [s] $\le \tau \le 80$ [s]	$[\;0\;\;\;4\,]^T$	$[\;0\;\;\;2\,]^T$
	80 [s] $\le \tau \le 120$ [s]	$[\;0\;\;\;2\,]^T$	$[\;0\;\;-2\,]^T$
$\mathbf{d}_{31}$	0 [s] $\le \tau \le 40$ [s]	$[-3\;\;-1\,]^T$	$[\;0\;\;\;0\,]^T$
	40 [s] $\le \tau \le 80$ [s]	$[-6\;\;-2\,]^T$	$[-3\;\;-1\,]^T$
	80 [s] $\le \tau \le 120$ [s]	$[-3\;\;-1\,]^T$	$[\;3\;\;\;1\,]^T$

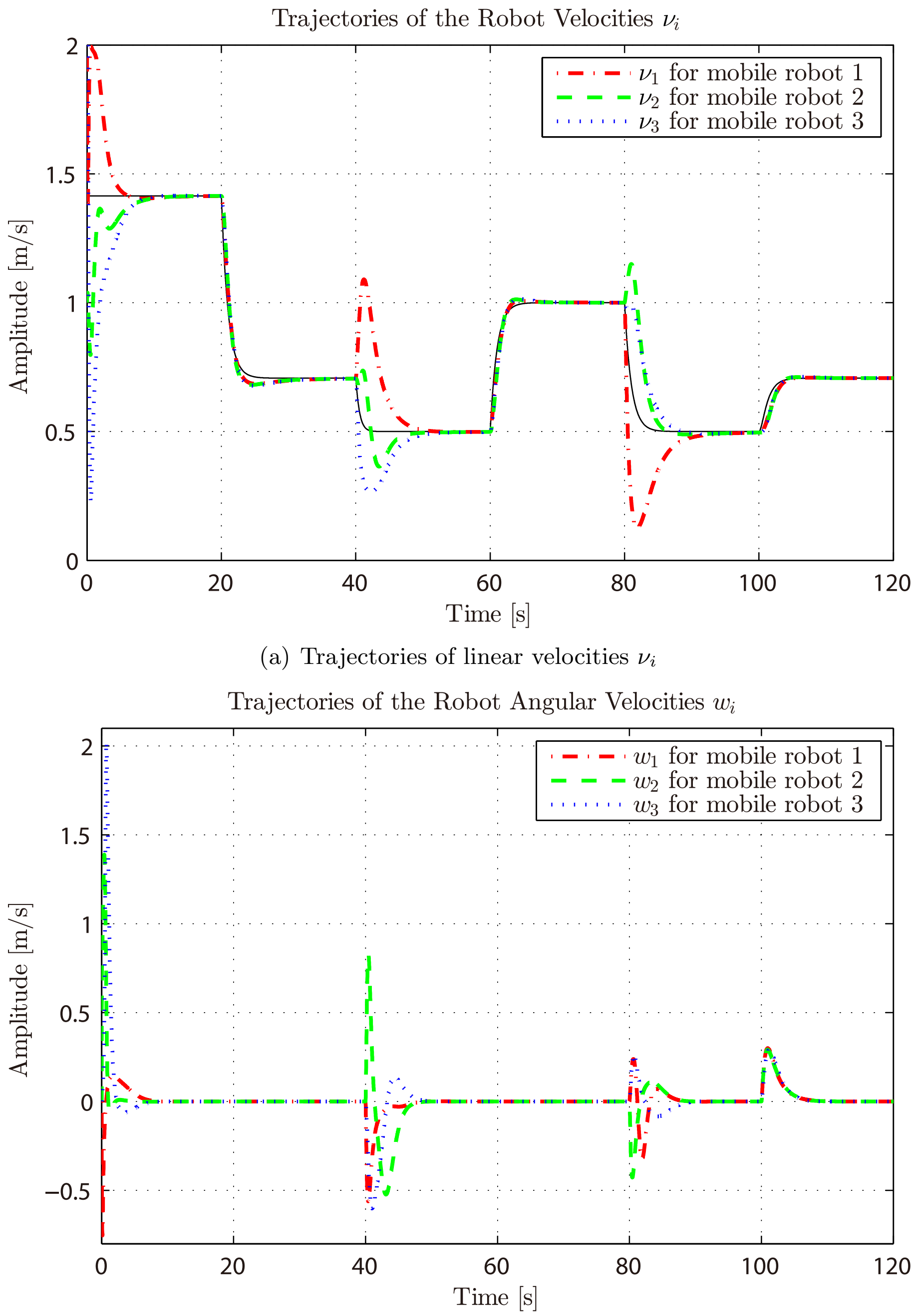
where $s_t(\tau) := \exp\{-\beta(\tau - t)\}$ with $\beta = 1$ is the smoothing function of the trajectory. In (6.51), the group velocity $\mathbf{v}_g$ for each time interval is of the form $\mathbf{v}_g = \mathbf{x} \pm s_t(\tau)\mathbf{y}$ for some vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ with the initial time $t > 0$ of the interval. This form is actually a smooth approximation of the hard profile $\mathbf{v}_g = \mathbf{x}$ for some vector $\mathbf{x}$ for each time interval. This hard profile of $\mathbf{v}_g$ can be obtained by replacing $s_t(\tau)$ by its limit "$\lim_{\tau\to\infty} s_t(\tau) = 0$". In a similar way, the profile of $\mathbf{d}_{ij}$ for each $i, j \in \{1, 2, 3\}$ is given in a form $\mathbf{d}_{ij} = \mathbf{x}_{ij} - s_t(\tau)\mathbf{y}_{ij}$, where $\mathbf{x}_{ij}$, $\mathbf{y}_{ij} \in \mathbb{R}^2$ are vectors given in Table 6.3.



(a) Position Trj.

(b) Angle Trj.

Figure 6.5: **(Example 2)** Position and angle trajectories of mobile robots.



(a) Trajectories of linear velocities ν_i

(b) Trajectories of angular velocities w_i

Figure 6.6: **(Example 2)** The velocities (ν_i, w_i) of mobile robots: (a) ν_i and (b) w_i.

The simulation result for some initial parameter estimates under the proposed adaptive inverse optimal CGFC scheme is shown in Figs. 6.5, 6.6, and 6.7, which represent the trajectories of the poses, the velocities (ν_i, w_i), and the parameters $\hat{c}_{i1}$ of the mobile

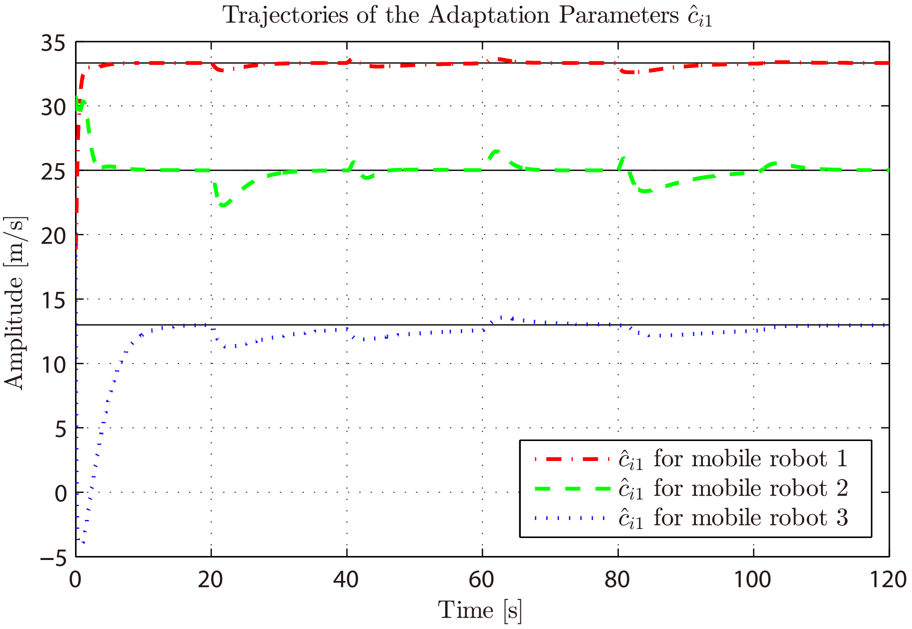


Figure 6.7: **(Example 2)** Trajectories of the parameter estimates $\hat{c}_{i1}$

robots, respectively. As shown in Figs. 6.5 and 6.6(a), the mobile robots effectively track the desired formation and group velocity profiles described by $\mathbf{d}_{ij}$ and $\mathbf{v}_g$; the black solid line in Fig. 6.6(a) indicates the desired speed given by the group velocity profile.

Moreover, as shown in Figs. 6.5(b) and 6.6(b), whenever the given group velocity profile changes its direction, the angular velocities of the robots are fluctuated from the zero steady-state to regulate their orientations and then converge to zero again; Fig. 6.6(a) and Fig. 6.7 also show that the linear velocities ν_i and the parameter estimates $\hat{c}_{i1}$ are perturbed only at the starting and changing points of $\mathbf{v}_g$ and $\mathbf{d}_{ij}$ and then converge to the desired point thereafter. From this, one can see that even in this complicated case, the parameter estimates $\hat{c}_{i1}$ are also regulated near the true values without imposing any persistently exciting conditions. If $\mathbf{v}_g$ and $\mathbf{d}_{ij}$ are constant as we have assumed in the design, then such perturbations shown in Fig. 6.6(a) and Fig. 6.7 are eliminated, and the estimates $\hat{c}_{i1}$ converges to their true values exactly. On the other hand, the other parameters in the control laws, though not plotted here, are just bounded by the stability argument in Theorem 6.2.

6.5.3 Simulation Example 3: Non-adaptive Inverse Optimal CGFC

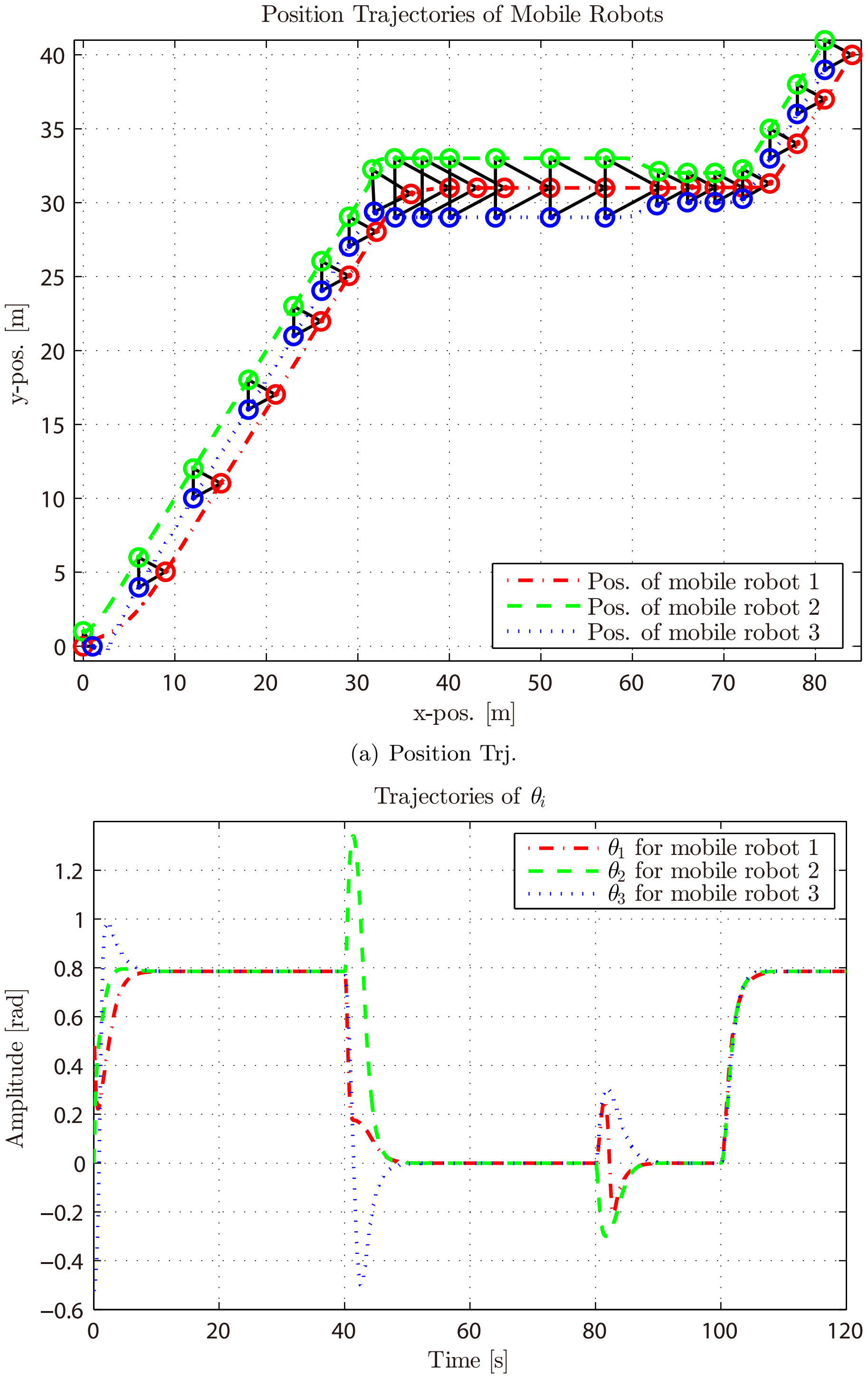
In this final example, we apply the inverse optimal CGFC scheme

$$\begin{cases} \tau_{i\nu}^* = c_{i1}\nu_i + z_{i1}w_i^2 + h_{i1}\dot{\nu}_i^* \\ \tau_{iw}^* = c_{i2}w_i + z_{i2}\nu_i w_i + \lambda h_{i2}(w_i^* - w_i) \end{cases}$$

to the mobile robots without adaptations of the parameters $\hat{c}_{ik}$, $\hat{z}_{ik}$, and $\hat{h}_{ik}$, where c_{ik}, z_{ik}, and h_{ik} are true values of the parameters. This simulation not only verifies the performance of the inverse optimal CGFC scheme, but makes it possible to compare the adaptive inverse optimal CGFC scheme in the previous example with the non-adaptive one. In the simulation, $\mathbf{v}_g$ and $\mathbf{d}_{ij}$ are given exactly same to those in the previous example.

The simulation results are plotted in Figs. 6.5 and 6.6. As shown in the figures, the proposed (non-adaptive) inverse optimal CGFC scheme $(\boldsymbol{\tau}_\nu^*, \boldsymbol{\tau}_w^*)$ control the mobile robots to efficiently follow the desired trajectories of their poses and velocities. On the other hand, there still exist the over- and under-shoots at the starting and changing points of $\mathbf{v}_g$ and/or $\mathbf{d}_{ij}$ as in the previous examples. This is because we have designed the proposed scheme under the assumption that $\mathbf{v}_g$ and $\mathbf{d}_{ij}$ are constant. This remains a future work of designing the same CGFC scheme with taking the time varying $\mathbf{v}_g$ and $\mathbf{d}_{ij}$ into considerations.

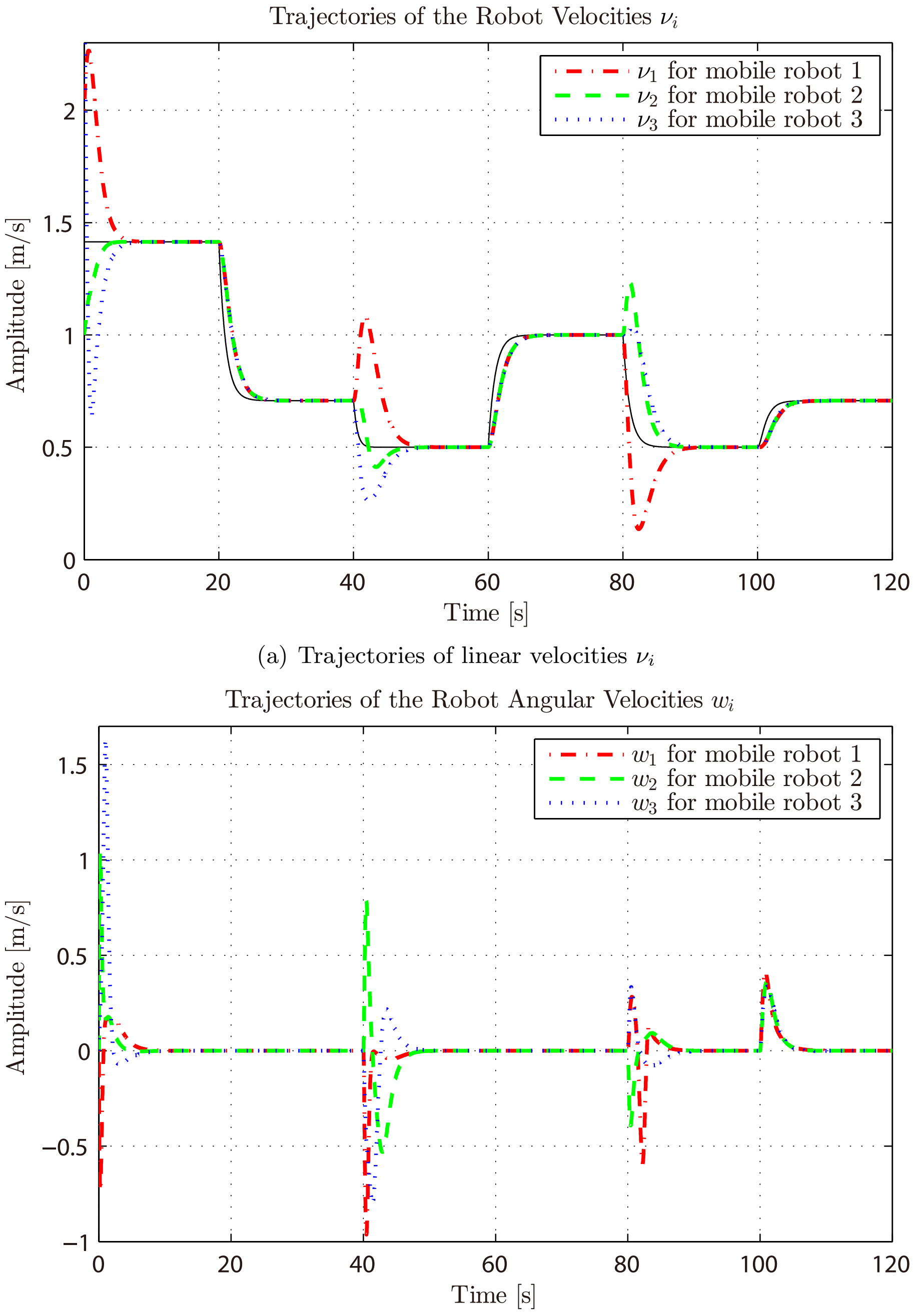
Comparing the simulation results of the adaptive scheme in the previous example with those of this non-adaptive schemes, one can see that the respective trajectories of the poses and velocities in both examples are almost same except values of the peak points. Therefore, though the parameters of the adaptive inverse optimal CGFC scheme in the previous example are not estimated exactly, the adaptive scheme approximately achieves the performance of the non-adaptive inverse optimal CGFC scheme.



(a) Position Trj.

(b) Angle Trj.

Figure 6.8: **(Example 3)** Position and angle trajectories of mobile robots.



(a) Trajectories of linear velocities ν_i

(b) Trajectories of angular velocities w_i

Figure 6.9: **(Example 3)** The velocities (ν_i, w_i) of mobile robots: (a) ν_i and (b) w_i.

6.6 Summary

From the control-theoretic perspectives, an adaptive inverse optimal CGFC was designed for multiple mobile robots described by CT dynamical systems with restricted information exchange. The kinematics and dynamics models of the mobile robots were transformed to the combined dynamics of consensus errors and velocity motions, which helps the design of the consensus-based inverse optimal control and the adaption parts separately. By Lyapunov and Hamiltonian analyses, it has been shown that

- the proposed scheme asymptotically achieves the desired formation and the desired group velocity under the undirected connected communication graph topology and adaptation laws;
- the proposed one is inverse optimal when the parametric uncertainties in the mobile robots are eliminated by the adaptation.

The simulation results were also provided to verify the performance of the proposed method under various scenarios.

Chapter 7

Conclusions

In this dissertation, IRL and adaptive inverse optimal control were studied as the candidates of the true adaptive optimal control for CT dynamical systems. As a preliminary offline algorithm, the ideal PI was proposed to introduce the fundamental IRLs in **Chapter 4** and to improve the explorized IRLs in **Chapter 5**, where the domains of the value functions in the existing offline PI were extended up to the ROAs. To develop the ideal PI above, the global properties of value functions were also studied on the ROAs. As a preliminary inverse optimal control scheme, an inverse optimal input-dynamics extension method is theoretically developed to employ it as a mathematical tool for the design of adaptive inverse optimal control in **Chapter 7**.

In **Chapter 4**, a family of partially model-free fundamental IRL algorithms including I-PI, I-VI, infinitesimal GPI, and their generalization "I-GPI" were presented in CT LQR framework and then classified in a new way in terms of the iteration horizon, the product of the iteration horizon involved in computational complexity and the time horizon determining the sampling period in time. In this new classification, the I-GPIs with the same update horizon are all equivalence classes in the iteration domain, implying the existence of the trade-off between the complexity and the sampling period. Then, in Chapter 4, the closed-loop stability and monotone convergence of I-GPI were investigated in relation to the update horizon. The main focus here were the two modes of convergence called VI- and PI-modes in convergence. These two convergence modes came from I-PI and infinitesimal GPI at the two extreme tips of the new classification and characterize the convergence behaviors of the fundamental IRLs. Here, it has been shown that PI-mode convergence guarantees the closed-loop stability and that VI-mode convergence is achieved only with the sufficiently small update horizon.

In **Chapter 5**, two online IRL methods that are able to explore the state space were proposed and analyzed based on the nonlinear I-PI and the concepts of both invariant explorations and advanced I-TD extended from the ideas of RL. These online IRL methods efficiently use the explorations to excite the necessary signals for online learning and, in integral Q-learning, to relax the model requirements; integral Q-learning provided the model-free online learning solution for the CT nonlinear optimal control problems with unknown dynamics, while the other one named explorized I-PI was provided as an effective online solution when the input coupling terms of the dynamics are known. The properties such as ISS, uniqueness of advanced I-TD solution, and the convergence to the solution were studied in relation to the design of the exploration signal.

In **Chapter 6**, from the control-theoretic perspectives, an adaptive inverse optimal CGFC was designed for multiple mobile robots described by CT dynamical systems with restricted information exchange. The kinematics and dynamics models of the mobile robots were transformed to the combined dynamics of consensus errors and velocity motions, which helps the design of the consensus-based inverse optimal control and the adaption parts separately. By Lyapunov and Hamiltonian analyses, it has been shown that

- the proposed scheme asymptotically achieves the desired formation and the desired group velocity under the undirected connected communication graph topology and adaptation laws designed by Lyapunov analysis;
- the proposed one is inverse optimal when the parametric uncertainties in the mobile robots are eliminated by adaptation.

In addition, the numerical simulations were performed to verify its performance and supports theoretical results.

Though there are still a number of future works we should focus on for true adaptive optimal control such as

- overcoming the exploration and exploitation dilemma,

- combining the RL methods with conventional adaptive (inverse optimal) controls,
- robustness with respect to the external disturbances,
- extending the results to the more general systems,
- stochastic considerations,

I believe that the works in this dissertation indeed make a progress toward developing the true adaptive optimal control of CT dynamical systems and bridge a gap among the interdisciplinary areas—reinforcement learning, adaptive control, and optimal control.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* Cambridge Univ Press, 1998.

[2] D. E. Kirk, *Optimal control theory: an introduction.* Dover Pubns, 2004.

[3] F. L. Lewis and V. L. Syrmos, *Optimal control.* John Wiley & Sons, 1995.

[4] K. Zhou, J. C. Doyle, K. Glover, *et al.*, *Robust and optimal control*, vol. 40. Prentice hall New Jersey, 1996.

[5] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[6] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, *Handbook of learning and approximate dynamic programming.* Wiley-IEEE Press, 2004.

[7] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 5, pp. 834–846, 1983.

[8] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[9] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.

[10] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000, ACM, 2009.

[11] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality.* Wiley-Interscience, 2007.

[12] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[13] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," *Advances in neural information processing systems*, pp. 1038–1044, 1996.

[14] W. D. Smart and L. P. Kaelbling, "Practical reinforcement learning in continuous spaces," in *ICML*, pp. 903–910, Citeseer, 2000.

[15] H. Van Hasselt and M. A. Wiering, "Reinforcement learning in continuous action spaces," in *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learnin (ADPRL)*, pp. 272–279, 2007.

[16] M. Riedmiller, T. Gabel, R. Hafner, and S. Lange, "Reinforcement learning for robot soccer," *Autonomous Robots*, vol. 27, no. 1, pp. 55–73, 2009.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[18] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches*, pp. 493–525, 1992.

[19] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *American Control Conference (ACC)*, vol. 3, pp. 3475–3479, 1994.

[20] T. Landelius, *Reinforcement learning and distributed local model synthesis.* PhD thesis, Linköping University Electronic Press, 1997.

[21] D. V. Prokhorov and D. C. Wunsch II, “Adaptive critic designs,” *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, 1997.

[22] F.-Y. Wang, H. Zhang, and D. Liu, “Adaptive dynamic programming: an introduction,” *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, 2009.

[23] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, “Model-free Q-learning designs for linear discrete-time zero-sum games with application to H_∞ control,” *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[24] A. Al-Tamimi, M. Abu-Khalaf, and F. L. Lewis, “Adaptive critic designs for discrete-time zero-sum games with application to control,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 1, pp. 240–247, 2007.

[25] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, “Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 4, pp. 943–949, 2008.

[26] W. Qiao, G. K. Venayagamoorthy, and R. G. Harley, “Optimal wide-area monitoring and nonlinear adaptive coordinating neurocontrol of a power system with wind power integration and multiple facts devices,” *Neural Networks*, vol. 21, no. 2, pp. 466–475, 2008.

[27] D. Zhao, X. Bai, F.-Y. Wang, J. Xu, and W. Yu, “DHP method for ramp metering of freeway traffic,” *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 4, pp. 990–999, 2011.

[28] F. L. Lewis and K. G. Vamvoudakis, “Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data,”

Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 41, no. 1, pp. 14–25, 2011.

[29] T. Y. Chun, J. B. Park, and Y. H. Choi, "Policy iteration-mode monotone convergence of generalized policy iteration for discrete-time linear systems," in *Control, Automation and Systems (ICCAS), 2013 13th International Conference on*, pp. 454–458, 2013.

[30] B. Kiumarsi, F. L. Lewis, M.-B. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," 2015.

[31] J.-J. E. Slotine and W. Li, *Applied nonlinear control.* Prentice-hall Englewood Cliffs, NJ, 1991.

[32] H. K. Khalil, *Nonlinear systems.* Prentice Hall, 2002.

[33] C.-T. Chen, *Linear system theory and design.* Oxford University Press, Inc., 1995.

[34] B. S. Thomson, J. B. Bruckner, and A. M. Bruckner, *Elementary real analysis.* ClassicalRealAnalysis. com, 2008.

[35] M. L. Puterman and M. C. Shin, "Modified policy iteration algorithms for discounted markov decision problems," *Management Science*, vol. 24, no. 11, pp. 1127–1137, 1978.

[36] J. A. E. E. van Nunen, "A set of successive approximation methods for discounted markovian decision problems," *Mathematical Methods of Operations Research*, vol. 20, no. 5, pp. 203–208, 1976.

[37] L. C. Baird III, "Reinforcement learning in continuous time: Advantage updating," in *Proc. Int. Conf. Neural Netw.*, vol. 4, pp. 2448–2453, 1994.

[38] K. Doya, "Reinforcement learning in continuous time and space," *Neural computation*, vol. 12, no. 1, pp. 219–245, 2000.

[39] T. Hanselmann, L. Noakes, and A. Zaknich, "Continuous-time adaptive critics," *IEEE Trans. Neural Networks*, vol. 18, no. 3, pp. 631–647, 2007.

[40] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Int. Conf. Decision and Control, held jointly with the Chinese Control Conference (CDC/CCC)*, pp. 3598–3605, 2009.

[41] T. Sakamoto, N. Hori, and Y. Ochi, "Exact linearization and discretization of nonlinear systems satisfying a Lagrange PDE condition," *Trans. Canadian Society for Mechanical Engineering*, vol. 35, no. 2, pp. 215–228, 2011.

[42] R. E. Mickens, "Nonstandard finite difference schemes for differential equations," *The Journal of Difference Equations and Applications*, vol. 8, no. 9, pp. 823–847, 2002.

[43] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, vol. 32, no. 2, pp. 140–153, 2002.

[44] J. J. Murray, C. J. Cox, and R. E. Saeks, "The adaptive dynamic programming theorem," in *Stability and Control of Dynamical Systems with Applications*, pp. 379–394, Springer, 2003.

[45] R. J. Leake and R.-W. Liu, "Construction of suboptimal control sequences," *SIAM Journal on Control*, vol. 5, no. 1, pp. 54–63, 1967.

[46] G. N. Saridis and C. S. G. Lee, "An approximation theory of optimal control for trainable manipulators," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 3, pp. 152–159, 1979.

[47] R. W. Beard, G. N. Saridis, and J. T. Wen, “Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation,” *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[48] M. Abu-Khalaf and F. L. Lewis, “Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach,” *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[49] D. Vrabie, *Online Adaptive Optimal Control For Continuous-time Systems.* Ph. D. Thesis, TX, USA: University of Texas at Arlington, 2010.

[50] D. Vrabie and F. L. Lewis, “Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems,” *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[51] D. Vrabie and L. Lewis, F., “Generalized policy iteration for continuous-time systems,” in *Proc. Int. Joint Conf. Neural Networks*, pp. 3224–3231, 2009.

[52] P. A. Ioannou and J. Sun, *Stable and robust adaptive control.* 1995.

[53] E. B. Kosmatopoulos, “Control of unknown nonlinear systems with efficient transient performance using concurrent exploitation and exploration,” *IEEE Trans. Neural Netw.*, vol. 21, no. 8, pp. 1245–1261, 2010.

[54] K. J. Åström and B. Wittenmark, *Adaptive control.* Courier Corporation, 2013.

[55] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and adaptive control design.* Wiley, 1995.

[56] K. S. Narendra and A. M. Annaswamy, *Stable adaptive systems.* Courier Corporation, 2012.

[57] G. Rovithakis and M. Christodoulou, “Adaptive control of unknown plants using dynamical neural networks,” *IEEE Trans. Systems, Man and Cybernetics*, vol. 24, no. 3, pp. 400–412, 1994.

[58] M. M. Polycarpou, “Stable adaptive neural control scheme for nonlinear systems,” *IEEE Trans. Automatic Control*, vol. 41, no. 3, pp. 447–451, 1996.

[59] K. S. Narendra and K. Parthasarathy, “Identification and control of dynamical systems using neural networks,” *IEEE Trans. Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.

[60] J. A. Farrell and M. M. Polycarpou, *Adaptive approximation based control: Unifying neural, fuzzy and traditional adaptive approximation approaches*, vol. 48. John Wiley & Sons, 2006.

[61] M. Fu and B. Barmish, “Adaptive stabilization of linear systems via switching control,” *IEEE Trans. Autom. Control*, vol. 31, no. 12, pp. 1097–1103, 1986.

[62] K. J. Åström and K. Furuta, “Swinging up a pendulum by energy control,” *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

[63] B. Anderson and J. B. Moore, *Optimal control: linear quadratic methods.* Prentice-Hall, Inc., 1989.

[64] R. Sepulchre, M. Janković, and P. Kokotović, *Constructive Nonlinear Control.*

[65] M. Krstić and Z.-H. Li, “Inverse optimal design of input-to-state stabilizing nonlinear controllers,” *IEEE Trans. Automatic Control*, vol. 43, no. 3, pp. 336–350, 1998.

[66] M. Krstić and P. Tsiotras, “Inverse optimal stabilization of a rigid spacecraft,” *IEEE Trans. Automatic Control*, vol. 44, no. 5, pp. 1042–1049, 1999.

[67] K. Ezal, Z. Pan, and P. Kokotovic, “Locally optimal and robust backstepping design,” *IEEE Trans. Automatic Control*, vol. 45, no. 2, pp. 260–271, 2000.

[68] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, “Nonlinear design of adaptive controllers for linear systems,” *Automatic Control, IEEE Transactions on*, vol. 39, no. 4, pp. 738–752, 1994.

[69] W. Luo, Y.-C. Chu, and K.-V. Ling, “Inverse optimal adaptive control for attitude tracking of spacecraft,” *IEEE Trans. Automatic Control*, vol. 50, no. 11, pp. 1639–1654, 2005.

[70] M. Krstic, “Optimal adaptive control—contradiction in terms or a matter of choosing the right cost functional?,” *IEEE Trans. Automatic Control*, vol. 53, no. 8, pp. 1942–1947, 2008.

[71] T. Balch and R. C. Arkin, “Behavior-based formation control for multirobot teams,” *IEEE Trans. Robotics and Automation*, vol. 14, no. 6, pp. 926–939, 1998.

[72] J. R. T. Lawton, R. W. Beard, and B. J. Young, “A decentralized approach to formation maneuvers,” *IEEE Trans. Robotics and Automation,*, vol. 19, no. 6, pp. 933–941, 2003.

[73] M. A. Lewis and K.-H. Tan, “High precision formation control of mobile robots using virtual structures,” *Autonomous Robots*, vol. 4, no. 4, pp. 387–403, 1997.

[74] K. D. Do, “Formation tracking control of unicycle-type mobile robots with limited sensing ranges,” *IEEE Trans. Control Systems Technology*, vol. 16, no. 3, pp. 527–538, 2008.

[75] J. P. Desai, J. P. Ostrowski, and V. Kumar, “Modeling and control of formations of nonholonomic mobile robots,” *IEEE Trans. Robotics and Automation*, vol. 17, no. 6, pp. 905–908, 2001.

[76] B. S. Park, J. B. Park, and Y. H. Choi, “Adaptive formation control of electrically driven nonholonomic mobile robots with limited information,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 4, pp. 1061–1075, 2011.

[77] W. Ren *et al.*, "Information consensus in multivehicle cooperative control," *IEEE Control systems magazine*, vol. 27, no. 2, pp. 71–82, 2007.

[78] J. A. Fax and R. M. Murray, "Information flow and cooperative control of vehicle formations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1465–1476, 2004.

[79] F. Borrelli and T. Keviczky, "Distributed lqr design for identical dynamically decoupled systems," *IEEE Trans. Autom. Control*, vol. 53, no. 8, pp. 1901–1912, 2008.

[80] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: state feedback, observer and output feedback," *IEEE Trans. Automat. Contr.*, vol. 56, no. 8, pp. 1948–1952, 2011.

[81] W. Dong, "Distributed optimal control of multiple systems," *Int. Journal of Control*, vol. 83, no. 10, pp. 2067–2079, 2010.

[82] Z. Qu and M. Simaan, "Inverse optimality of cooperative control for networked systems," in *Proc. Conf. Decision and Control, held jointly with the 28th Chinese Control Conference.*, pp. 1651–1658, 2009.

[83] Y. Cao and W. Ren, "Optimal linear-consensus algorithms: an LQR perspective," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 3, pp. 819–830, 2010.

[84] K. Hengster Movric and F. L. Lewis, "Cooperative optimal control for multi-agent systems on directed graph topologies," *IEEE Trans. Automat. Contr.*, vol. 59, no. 3, pp. 769–774, 2014.

[85] H. Zhang, F. L. Lewis, and Z. Qu, "Lyapunov, adaptive, and optimal design techniques for cooperative systems on directed communication graphs," *IEEE Trans. Industrial Electronics*, vol. 59, no. 7, pp. 3026–3041, 2012.

[86] E. Semsar-Kazerooni and K. Khorasani, “Multi-agent team cooperation: A game theory approach,” *Automatica*, vol. 45, no. 10, pp. 2205–2213, 2009.

[87] W. Ren and R. W. Beard, “Consensus algorithms for double-integrator dynamics,” *Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications*, pp. 77–104, 2008.

[88] W. Yu, G. Chen, M. Cao, and J. Kurths, “Second-order consensus for multiagent systems with directed topologies and nonlinear dynamics,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 3, pp. 881–891, 2010.

[89] G. Oriolo, A. De Luca, and M. Vendittelli, “WMR control via dynamic feedback linearization: design, implementation, and experimental validation,” *IEEE Trans. Control Systems Technology*, vol. 10, no. 6, pp. 835–852, 2002.

[90] W. Dong and J. A. Farrell, “Adaptive approximately optimal control of unknown nonlinear systems based on locally weighted learning,” in *Proc. Int. Conf. Decision and Control, held jointly with the Chinese Control Conference (CDC/CCC)*, pp. 345–350, 2009.

[91] W. Dong, “Flocking of multiple mobile robots based on backstepping,” *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 2, pp. 414–424, 2011.

[92] J. Y. Lee and Y. H. Choi, “LQ inverse optimal consensus protocol for continuous-time multi-agent systems and its applications to formation control,” *Journal of ICROS (in Korean)*, vol. 20, no. 5, pp. 526–532, 2014.

[93] J. Y. Lee, J. B. Park, and Y. H. Choi, “On integral generalized policy iteration for continuous-time linear quadratic regulations,” *Automatica*, vol. 50, no. 2, pp. 475–489, 2014.

[94] J. Y. Lee, J. B. Park, and Y. H. Choi, “Invariantly admissible policy iteration for a class of nonlinear optimal control problems,” *Submitted to Syst. Contol Lett. (available at http://arxiv.org/abs/1402.4187)*, 2014.

[95] J. Y. Lee, T. Y. Chun, J. B. Park, and Y. H. Choi, “On generalized policy iteration for continuous-time linear systems,” in *50th IEEE CDC and ECC*, pp. 1722–1728, 2011.

[96] J. Y. Lee, J. B. Park, and Y. H. Choi, “A novel generalized value iteration scheme for uncertain continuous-time linear systems,” in *49th IEEE Conf. Decision and Control*, pp. 4637–4642, 2010.

[97] J. Y. Lee, J. B. Park, and Y. H. Choi, “Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations,” *IEEE Trans. Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, 2015.

[98] J. Y. Lee, J. B. Park, and Y. H. Choi, “Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems,” *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.

[99] J. Y. Lee, J. B. Park, and Y. H. Choi, “Approximate dynamic programming for continuous-time linear quadratic regulator problems: relaxation of known input-coupling matrix assumption,” *Control Theory & Applications, IET*, vol. 6, no. 13, pp. 2063–2075, 2012.

[100] J. Y. Lee, J. B. Park, and Y. H. Choi, “Integral reinforcement learning with explorations for continuous-time nonlinear systems,” in *Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1042–1047, 2012.

[101] J. Y. Lee, J. B. Park, and Y. H. Choi, “Model-free approximate dynamic programming for continuous-time linear systems,” in *Proc. Int. Conf. Decision and Control, held jointly with the Chinese Control Conference (CDC/CCC)*, pp. 5009–5014, 2009.

[102] J. Y. Lee and Y. H. Choi, “Inverse optimal design of formation/velocity consensus protocolfor mobile robots based on LQ inverse optimal second-order consensus,” *Journal of ICROS (in Korean)*, vol. 21, no. 5, pp. 434–441, 2015.

[103] J. Y. Lee, Y. H. Choi, and J. B. Park, “Inverse optimal design of the distributed consensus protocol for formation control of multiple mobile robots,” in *IEEE 53rd Annual Conf. Decision and Control (CDC)*, pp. 2222–2227, 2014.

[104] P. Lancaster and L. Rodman, *Algebraic Riccati equations.* Oxford University Press, 1995.

[105] K. Hengster Movric, *Cooperative control of multi-agent systems stability, optimality and robustness.* PhD thesis, University of Texas at Arlington, 2013.

[106] R. W. Beard, *Improving the closed-loop performance of nonlinear systems.* PhD thesis, Rensselaer Polytechnic Institute, 1995.

[107] D. M. Adhyaru, I. N. Kar, and M. Gopal, “Bounded robust control of nonlinear systems using neural network-based HJB solution,” *Neural Computing and Applications*, vol. 20, no. 1, pp. 91–103, 2011.

[108] H. Alwardi, S. Wang, and L. S. Jennings, “An adaptive domain decomposition method for the Hamilton-Jacobi-Bellman equation,” *Journal of Global Optimization*, pp. 1361–1373, 2012.

[109] R. Munos, L. C. Baird, and A. W. Moore, “Gradient descent approaches to neural-net-based solutions of the Hamilton-Jacobi-Bellman equation,” in *International Joint Conf. Neural Networks (IJCNN)*, vol. 3, pp. 2152–2157, 1999.

[110] D. Kleinman, “On an iterative technique for Riccati equation computations,” *IEEE Trans. Autom. Cont.*, vol. 13, no. 1, pp. 114–115, 1968.

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

[112] H. Saadat, *Power system analysis.* McGraw-Hill Primis Custom, 2002.

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

[114] F. Feitzinger, T. Hylla, and E. W. Sachs, "Inexact kleinman-newton method for Riccati equations," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 2, pp. 272–288, 2009.

[115] G. Strang, *Linear algebra and its applications academic.* 2005.

Appendix A

Kronecker and Khatri-Rao Products

For any two matrices $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n \times m}$ and $\mathbf{Y} \in \mathbb{R}^{p \times q}$, the Kronecker product $\mathbf{X} \otimes \mathbf{Y}$ of $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathbf{X} \otimes \mathbf{Y} := \begin{bmatrix} x_{11}\mathbf{Y} & x_{12}\mathbf{Y} & \cdots & x_{1m}\mathbf{Y} \\ x_{21}\mathbf{Y} & x_{22}\mathbf{Y} & \cdots & x_{2m}\mathbf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}\mathbf{Y} & x_{n2}\mathbf{Y} & \cdots & x_{nm}\mathbf{Y} \end{bmatrix}. \tag{A.1}$$

This Kronecker product has the following well-known properties.

Proposition A.1. *For any real matrices* $\mathbf{X}$*,* $\mathbf{Y}$*,* $\mathbf{Z}$*, and any real number* $k \in \mathbb{R}$*,*

$$\cdot\ \textit{bilinearity:} \begin{cases} \mathbf{X} \otimes (\mathbf{Y} + \mathbf{Z}) = \mathbf{X} \otimes \mathbf{Y} + \mathbf{X} \otimes \mathbf{Z} \\ (\mathbf{X} + \mathbf{Y}) \otimes \mathbf{Z} = \mathbf{X} \otimes \mathbf{Z} + \mathbf{Y} \otimes \mathbf{Z}, \end{cases}$$

$$\cdot\ \textit{associativity:} \begin{cases} \mathbf{X} \otimes (\mathbf{Y} \otimes \mathbf{Z}) = (\mathbf{X} \otimes \mathbf{Y}) \otimes \mathbf{Z} \\ (k\mathbf{X}) \otimes \mathbf{Y} = \mathbf{X} \otimes (k\mathbf{Y}) = k(\mathbf{X} \otimes \mathbf{Y}). \end{cases}$$

Proposition A.2. *For any* $\mathbf{X} \in \mathbb{R}^{n \times m}$*,* $\mathbf{Y} \in \mathbb{R}^{p \times q}$*,* $\mathbf{Z} \in \mathbb{R}^{m \times l}$*, and* $\mathbf{W} \in \mathbb{R}^{q \times r}$*,*

- *transpose property:* $(\mathbf{X} \otimes \mathbf{Y})^T = \mathbf{X}^T \otimes \mathbf{Y}^T$
- *mixed-product property:* $(\mathbf{X} \otimes \mathbf{Y})(\mathbf{Z} \otimes \mathbf{W}) = \mathbf{X}\mathbf{Z} \otimes \mathbf{Y}\mathbf{W}$.

Proposition A.3. *If* $\mathbf{X} \in \mathbb{R}^{n \times n}$ *and* $\mathbf{Y} \in \mathbb{R}^{p \times p}$ *are invertible, then*

$$(\mathbf{X} \otimes \mathbf{Y})^{-1} = \mathbf{X}^{-1} \otimes \mathbf{Y}^{-1}.$$

Proposition A.4. *For any real vectors* $\mathbf{x} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^m$*, there is a permutation matrix* $\mathbf{U} \in \mathbb{R}^{(nm) \times (nm)}$ *such that*

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{U}(\mathbf{y} \otimes \mathbf{x}).$$

For the matrices $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $\bar{\mathbf{Y}} \in \mathbb{R}^{(np)\times q}$ partitioned as

$$\mathbf{X} = \left[\mathbf{x}_{1r}^T \,\vdots\, \mathbf{x}_{2r}^T \,\vdots\, \cdots \,\vdots\, \mathbf{x}_{nr}^T\right]^T \text{ and } \bar{\mathbf{Y}} = \left[\mathbf{Y}_1^T \,\vdots\, \mathbf{Y}_2^T \,\vdots\, \cdots \,\vdots\, \mathbf{Y}_n^T\right]^T,$$

where $\mathbf{x}_{ir} \in \mathbb{R}^{1\times m}$ denotes the i-th row vector of $\mathbf{X}$, and $\mathbf{Y}_i \in \mathbb{R}^{p\times q}$ is the i-th submatrix of $\mathbf{Y}$ $(i = 1, 2, \cdots, n)$, the Khatri-Rao product $\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^N$ (or $\mathbf{X} \otimes \bar{\mathbf{Y}}$), is defined as

$$\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^N := \begin{bmatrix} \mathbf{x}_{1r} \otimes \mathbf{Y}_1 \\ \hdashline \mathbf{x}_{2r} \otimes \mathbf{Y}_2 \\ \hdashline \vdots \\ \hdashline \mathbf{x}_{nr} \otimes \mathbf{Y}_n \end{bmatrix} = \begin{bmatrix} x_{11}\mathbf{Y}_1 & x_{12}\mathbf{Y}_1 & \cdots & x_{1m}\mathbf{Y}_1 \\ x_{21}\mathbf{Y}_2 & x_{22}\mathbf{Y}_2 & \cdots & x_{2m}\mathbf{Y}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}\mathbf{Y}_n & x_{n2}\mathbf{Y}_n & \cdots & x_{nm}\mathbf{Y}_n \end{bmatrix}. \tag{A.2}$$

As mentioned in Section 2.1, the Khatri-Rao product $\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n$ is a kine of a generalized Kronecker product, so some properties of the Kronecker product in Propositions A.1 through A.4 can be extended to the Khatri-Rao product $\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n$. For this we need the following lemma whose proof is trivial.

Lemma A.1. *For* $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n \in \mathbb{R}^{p\times q}$ *for some* $p, q \in \mathbb{N}$,

$$\mathbf{diag}\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n\} = \mathbf{I}_n \otimes \{\mathbf{Y}_i\}_{i=1}^n.$$

Lemma A.2. *For any invertible real matrices* $\mathbf{Y}_i$*'s* $(i = 1, 2, \cdots, n)$,

$$\big(\mathbf{diag}\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n\}\big)^{-1} = \mathbf{diag}\{\mathbf{Y}_1^{-1}, \mathbf{Y}_2^{-1}, \cdots, \mathbf{Y}_n^{-1}\}.$$

Proposition A.5. *For any* $\mathbf{X}, \mathbf{W} \in \mathbb{R}^{n\times m}$, $\mathbf{Y}_i, \mathbf{Z}_i \in \mathbb{R}^{p\times q}$ $(i = 1, 2, \cdots, n)$, *and* $k \in \mathbb{R}$,

$$\cdot \textit{ bilinearity: } \begin{cases} \mathbf{X} \otimes \{\mathbf{Y}_i + \mathbf{Z}_i\}_{i=1}^n = \mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n + \mathbf{X} \otimes \{\mathbf{Z}_i\}_{i=1}^n \\ (\mathbf{X} + \mathbf{W}) \otimes \{\mathbf{Y}_i\}_{i=1}^n = \mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n + \mathbf{W} \otimes \{\mathbf{Y}_i\}_{i=1}^n, \end{cases}$$

$$\cdot \textit{ associativity: } (k\mathbf{X}) \otimes \{\mathbf{Y}_i\}_{i=1}^n = \mathbf{X} \otimes \{k\mathbf{Y}_i\}_{i=1}^n = k\big(\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n\big).$$

Proof. The proof can be done by the definition (A.2) and the block-diagonal matrix multiplications. □

Proposition A.6. *For any* $\mathbf{X} \in \mathbb{R}^{n\times m}$, $\mathbf{Y}_i \in \mathbb{R}^{p\times q}$, *and* $\mathbf{Z}_i \in \mathbb{R}^{q\times r}$ $(i = 1, 2, \cdots, n)$,

$$\cdot \textit{ mixed-product property: } \big(\mathbf{I}_n \otimes \{\mathbf{Y}_i\}_{i=1}^n\big)\big(\mathbf{X} \otimes \{\mathbf{Z}_i\}_{i=1}^n\big) = \mathbf{X} \otimes \{\mathbf{Y}_i\mathbf{Z}_i\}_{i=1}^n \tag{A.3}$$

Proof. The proof can be done by using Lemma A.1, substituting the definitions of the Khatri-Rao product (A.2) and the block-diagonal operation $\mathbf{diag}\{\mathbf{Y}_1, \cdots, \mathbf{Y}_n\}$ in (2.1), and performing the block-matrix multiplications □

Proposition A.7. *If* $\mathbf{X} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n \in \mathbb{R}^{p\times p}$ *are all invertible, then*

$$\big(\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n\big)^{-1} = \big(\mathbf{X}^{-1} \otimes \mathbf{I}_p\big)\big(\mathbf{I}_n \otimes \{\mathbf{Y}_i^{-1}\}_{i=1}^n\big). \tag{A.4}$$

Proof. For the proof, suppose that $\mathbf{X}$ and $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ are all invertible. Then, by Proposition A.3 and Lemma A.2, $\mathbf{X} \otimes \mathbf{I}_p$ and $\mathbf{I}_n \otimes \{\mathbf{Y}_i\}_{i=1}^n$ are also invertible and their inverses are given by

$$\begin{cases} \big(\mathbf{X} \otimes \mathbf{I}_p\big)^{-1} = \mathbf{X}^{-1} \otimes \mathbf{I}_p \\ \big(\mathbf{I}_n \otimes \{\mathbf{Y}_i\}_{i=1}^n\big)^{-1} = \mathbf{I}_n \otimes \{\mathbf{Y}_i^{-1}\}_{i=1}^n. \end{cases}$$

Hence, the application of (A.3) and the property $(\mathbf{XB})^{-1} = \mathbf{B}^{-1}\mathbf{X}^{-1}$ yields

$$\begin{aligned} \big(\mathbf{X} \otimes \{\mathbf{Y}_i\}_{i=1}^n\big)^{-1} &= \big(\mathbf{X} \otimes \mathbf{I}_p\big)^{-1}\big(\mathbf{I}_n \otimes \{\mathbf{Y}_i\}_{i=1}^n\big)^{-1} \\ &= \big(\mathbf{X}^{-1} \otimes \mathbf{I}_p\big)\big(\mathbf{I}_n \otimes \{\mathbf{Y}_i^{-1}\}_{i=1}^n\big), \end{aligned}$$

which completes the proof. □

Appendix B

Schur Complement

For a symmetric matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ decomposed as $\mathbf{P} = \left[\begin{array}{c:c} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \hdashline \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{array}\right]$ with $\mathbf{P}_{11} \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{P}_{12} \in \mathbb{R}^{n_1 \times n_2}$, and invertible $\mathbf{P}_{22} \in \mathbb{R}^{n_2 \times n_2}$ with n_1 and n_2 satisfying $n_1 + n_2 = n$, its Schur complement $\mathcal{S}(\mathbf{P})$ is defined as

$$\mathcal{S}(\mathbf{P}) := \mathbf{P}_{11} - \mathbf{P}_{12}\mathbf{P}_{22}^{-1}\mathbf{P}_{12}^T.$$

Notice that any quadratic function $V(\mathbf{z}) = \mathbf{z}^T\mathbf{P}\mathbf{z}$ of $\mathbf{z} = \mathbf{col}\{\mathbf{x}, \mathbf{y}\} \in \mathbb{R}^n$ with the vectors $\mathbf{x}$ and $\mathbf{y}$ compatible with the block matrices of $\mathbf{P}$ can be represented in terms of the Schur complement of $\mathbf{P}$ as

$$\begin{aligned} V(\mathbf{z}) &= \mathbf{x}^T\mathbf{P}_{11}\mathbf{x} + 2\mathbf{x}^T\mathbf{P}_{12}\mathbf{y} + \mathbf{y}^T\mathbf{P}_{22}\mathbf{y} \\ &= \mathbf{x}^T\mathcal{S}(\mathbf{P})\mathbf{x} + \big(\mathbf{y} + \mathbf{P}_{22}^{-1}\mathbf{P}_{12}^T\mathbf{x}\big)^T\mathbf{P}_{22}\big(\mathbf{y} + \mathbf{P}_{22}^{-1}\mathbf{P}_{12}^T\mathbf{x}\big). \end{aligned} \tag{B.1}$$

From this Schur complement expression, the following well-known Schur complement lemma can be directly obtained.

Lemma B.1. $\mathbf{P}$ *is positive semi-definite iif so is* $\mathcal{S}(\mathbf{P})$*.*

Lemma B.2. $\operatorname{rank}\mathbf{P} = \operatorname{rank}\mathcal{S}(\mathbf{P}) + n_2$*.*

From Lemmas B.1 and B.2, one can prove the following Schur complement null-space lemma that is used in the proof of.

Lemma B.3. *If the nullities of both* $\mathbf{P}_{11}$ *and* $\mathcal{S}(\mathbf{P})$ *are same,* $\mathbf{P}$ *is positive semi-definite, and* $\ker\mathbf{P}_{11} \subseteq \ker\mathbf{P}_{12}$*, then*

$$\begin{aligned} &\ker\mathbf{P}_{11} = \ker\mathcal{S}(\mathbf{P}), \\ &\ker\mathbf{P} = \big\{\mathbf{z} = \mathbf{col}\{\mathbf{x}, \mathbf{y}\} \in \mathbb{R}^n : \mathbf{x} \in \ker\mathbf{P}_{11} \textit{ and } \mathbf{y} = \mathbf{0}_{n_2}\big\}. \end{aligned}$$

Proof. The condition "$\ker \mathbf{P}_{11} \subseteq \ker \mathbf{P}_{12}$" and the definition of $\mathcal{S}(\mathbf{P})$ show that

$$\mathbf{x} \in \ker \mathbf{P}_{11} \text{ implies } \mathbf{x} \in \ker \mathcal{S}(\mathbf{P}).$$

Since the dimensions of both null-spaces are same, the condition is necessary and sufficient, the proof of $\ker \mathbf{P}_{11} = \ker \mathcal{S}(\mathbf{P})$. Next, since $\mathbf{P}$ is positive semi-definite, so is $\mathcal{S}(\mathbf{P})$ by Lemma B.1. Hence, (B.1) implies that $\mathbf{z} \in \ker \mathbf{P}$ iif $\mathbf{x} \in \ker \mathcal{S}(\mathbf{P})$ and $\mathbf{y} = -\mathbf{P}_{22}^{-1}\mathbf{P}_{12}^{T}\mathbf{x}$. The former implies $\mathbf{z} \in \ker \mathbf{P}_{11}$; the latter and $\ker \mathbf{P}_{11} = \ker \mathcal{S}(\mathbf{P})$ show $\mathbf{y} = \mathbf{0}_{n_2}$, which completes the proof. □

Appendix C

Graph Theory

A graph $\mathcal{G}$ is a triple $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, \mathcal{A}\}$, where

- $\mathcal{N} := \{1, 2, \cdots, N\}$ is the node set;
- $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ represents the edge set;
- $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N \times N}$ denotes the weighted adjacency matrix whose elements a_{ij}'s are given by $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise.

The neighborhood $\mathcal{N}_i \subseteq \mathcal{N}$ of a node $i \in \mathcal{N}$ is defined as

$$\mathcal{N}_i := \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}.$$

A graph $\mathcal{G}$ is said to be simple if $a_{ii} = 0$ for all $i \in \mathcal{N}$, implying that $\mathcal{A}$ has zero diagonals; $\mathcal{G}$ is said to be undirected if $a_{ij} = a_{ji}$ for all $i, j \in \mathcal{N}$, which implies that $\mathcal{A}$ is symmetric. A path $\mathcal{P}_{i_1 i_2 \cdots i_l}$ of $\mathcal{G}$ is a subset of the edge $\mathcal{E}$ defined as $\mathcal{P}_{i_1 i_2 \cdots i_l} := \big\{(i_1, i_2), (i_2, i_3), \cdots, (i_{l-1}, i_l)\big\} \subseteq \mathcal{E}$, where l is called the length of the path $\mathcal{P}_{i_1 i_2 \cdots i_l}$.

Definition C.1. *A graph $\mathcal{G}$ is said to be connected if for each $i, j \in \mathcal{N}$, there is a path $\mathcal{P}_{i_1 i_2 \cdots i_l}$ from $i = i_1$ to $j = i_l$ for some length $l \in \mathbb{N}$.*

The Laplacian matrix $\mathbf{L} = [l_{ij}] \in \mathbb{R}^{N \times N}$ of a simple graph $\mathcal{G}$ is defined as $l_{ij} = -a_{ij}$ for $i \neq j$ and $l_{ii} = \sum_{j=1}^{N} a_{ij}$. If the graph $\mathcal{G}$ is undirected, then the associated Laplacian $\mathbf{L}$ is always positive semi-definite and have at least one zero eigenvalue associated with the eigenvector $\mathbf{1}_N \in \mathbb{R}^N$. That is,

$$0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L}).$$

Theorem C.1. *The graph $\mathcal{G}$ is simple, undirected, and connected if and only if $\lambda_2(\mathbf{L}) > 0$ and* $\operatorname{rank} \mathbf{L} = N - 1$.

Proof. See [77] with Gershigorin's circle theorem. □

Since the Laplacian matrix $\mathbf{L}$ is *symmetric* positive semi-definite and $\lambda_1(\mathbf{L}) = 0$, there is an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{N \times N}$ such that

$$\mathbf{U}^T \mathbf{L} \mathbf{U} = \left[\begin{array}{c:c} 0 & \mathbf{0}_{N-1}^T \\ \hdashline \mathbf{0}_{N-1} & \mathbf{\Lambda}_{N-1} \end{array}\right],$$

where $\mathbf{\Lambda} := \mathbf{diag}\{\lambda_2(\mathbf{L}), \cdots, \lambda_N(\mathbf{L})\}$ [115]. Hence, we have the following corollary.

Corollary C.1. *Let the graph $\mathcal{G}$ be simple, undirected, and connected. Then, there exists an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{N \times N}$ such that $\mathbf{L}$ is decomposed as*

$$\mathbf{U}^T \mathbf{L} \mathbf{U} = \left[\begin{array}{c:c} 0 & \mathbf{0}_{N-1}^T \\ \hdashline \mathbf{0}_{N-1} & \mathbf{\Lambda}_{N-1} \end{array}\right],$$

where $\mathbf{\Lambda}_{N-1} \in \mathbb{R}^{N \times N}$ is a diagonal positive definite matrix.

Appendix D

Proofs

D.1 Proof of Lemma 2.4

Suppose "$\mathbf{x} \in \mathbb{S}$ implies $V(\mathbf{x}) = 0$", and for $s \in [0, r]$, let $\psi(s)$ and $\phi(s)$ be defined as

$$\psi(s) := \inf\big\{V(\mathbf{x}) : s \leq d(\mathbf{x}, \mathbb{S}) \leq r\big\} \text{ and } \phi(s) := \sup\big\{V(\mathbf{x}) : d(\mathbf{x}, \mathbb{S}) \leq s\big\},$$

respectively. Then, $\psi(s)$ and $\phi(s)$ are continuous and increasing. Furthermore, they satisfy

$$\psi(d(\mathbf{x}, \mathbb{S})) \leq V(\mathbf{x}) = \phi(d(\mathbf{x}, \mathbb{S})) \tag{D.1}$$

for $\mathbf{x} \in B_{\mathbb{S}}(r)$; $\psi(0) = \phi(0) = 0$ holds by the assumption that $d(\mathbf{x}, \mathbb{S}) = 0$ implies $V(\mathbf{x}) = 0$. Hence, (2.6) holds for $\underline{\alpha} = \psi$ and $\bar{\alpha} = \phi$. Next, assume that the condition "$\mathbf{x} \in \mathbb{S}$ implies $V(\mathbf{x}) = 0$" is strengthened to

$$\mathbf{x} \in \mathbb{S} \iff V(\mathbf{x}) = 0.$$

Then, the definitions of ψ and ϕ show that $\psi(0) = \phi(0) = 0$ and $0 < \psi(s) \leq \phi(s)$ for all $s \in (0, r]$. Hence, they are positive definite on $[0, r]$. On the other hand, ψ and ϕ may not belong to class $\mathcal{K}$ since they are not necessarily strictly increasing. Take

$$\underline{\alpha}(s) = \frac{s}{s+1}\psi(s).$$

Then, it is strictly increasing since $s/(s+1)$ is strictly increasing and $\psi(s)$ is increasing and positive in $(0, r]$. In addition, $s/(s+1) \leq 1$ for $s \geq 0$ implies that $\underline{\alpha}(s) \leq \psi(s)$ for $s \in [0, r]$. Similarly, take the strictly increasing $\bar{\alpha}(s)$ as

$$\bar{\alpha}(s) = \phi(s) + \frac{s}{s+1}\phi(s)$$

that satisfies $\phi(s) \le \bar{\alpha}(s)$ for $s \in [0, r]$. Then, the above arguments show that $\underline{\alpha}$ and $\bar{\alpha}$ belong to class $\mathcal{K}$ and that (D.1) implies

$$\underline{\alpha}(d(\mathbf{x}, \mathbb{S})) \le \psi(d(\mathbf{x}, \mathbb{S})) \le V(\mathbf{x}) = \phi(d(\mathbf{x}, \mathbb{S})) \le \underline{\alpha}(d(\mathbf{x}, \mathbb{S}))$$

for $\mathbf{x} \in B_{\mathbb{S}}(r)$. For $\mathcal{D} = \mathbb{R}^n$, $r > 0$ can be arbitrarily large, so $\psi(s)$ and $\phi(s)$ defined for $s \in [0, \infty)$ as

$$\psi(s) := \inf \big\{ V(\mathbf{x}) : s \le d(\mathbf{x}, \mathbb{S}) \big\} \text{ and } \phi(s) := \sup \big\{ V(\mathbf{x}) : d(\mathbf{x}, \mathbb{S}) \le s \big\},$$

work for the proof. In addition, if $V(\mathbf{x}) \to \infty$ as $d(\mathbf{x}, \mathbb{S}) \to \infty$, then the function $\psi(s)$ and hence, $\underline{\alpha}(s)$ tends to infinity as $s \to \infty$. So, $\underline{\alpha}$ and $\bar{\alpha}$ can be chosen to belong class $\mathcal{K}^{\infty}$.

D.2 Proof of Lemma 2.5.

For the proof, consider the decomposition $\mathbf{x} = \mathbf{x}_r + \mathbf{x}_n$ of a vector $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{x}_r \in \mathbb{R}^n$ belongs to the row space of $\mathbf{P}$ and $\mathbf{x}_n \in \ker \mathbf{P}$. Since the row space and the null space of the same matrix are orthogonal complements of each other, for any $\mathbf{y} \in \ker \mathbf{P}$, $\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\|\mathbf{x}_r\|_2^2 + \|\mathbf{x}_n - \mathbf{y}\|_2^2}$ holds, and hence we obtain

$$d(\mathbf{x}, \ker \mathbf{P}) = \|\mathbf{x}_r\|_2 \tag{D.2}$$

by the definition of the distance function $d(\mathbf{x}, \mathbb{S})$ in Section 2.2. Next, without loss of generality, assume that $\mathbf{P} \ne \mathbf{0}_{n\times n}$ [1] and $\mathbf{P} \succeq \mathbf{0}_{n\times n}$. Then, every eigenvalues of $\mathbf{P}$ are real and nonnegative, and there is $N \in \{0, 1, 2, \cdots, n-1\}$ such that $\lambda_i(\mathbf{P}) > 0$ for all $i \in \{N+1, N+2, \cdots, n\}$. Hence, the singular value decomposition of $\mathbf{P}$ yields

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q}_n & \mathbf{Q}_r \end{bmatrix} \begin{bmatrix} \mathbf{0}_{N\times N} & \mathbf{0}_{N\times(n-N)} \\ \mathbf{0}_{(n-N)\times N} & \mathbf{\Lambda}_r \end{bmatrix} \begin{bmatrix} \mathbf{Q}_n^T \\ \mathbf{Q}_r^T \end{bmatrix} = \mathbf{Q}_r \mathbf{\Lambda}_r \mathbf{Q}_r^T, \tag{D.3}$$

[1] If $\mathbf{P} = \mathbf{0}_{n\times n}$, then the row space contains only the zero vector $\mathbf{0}_n$, so that $\mathbf{x}_r$ necessarily becomes to $\mathbf{0}_n$. Hence, $d(\mathbf{x}, \ker \mathbf{P}) = 0$ $\forall \mathbf{x} \in \mathbb{R}^n$ by (D.2), implying that any $\underline{\alpha}$, $\bar{\alpha} > 0$ satisfy (2.8).

where $\mathbf{Q}_n \in \mathbb{R}^{n\times N}$ and $\mathbf{Q}_r \in \mathbb{R}^{n\times(n-N)}$ are the orthogonal matrices whose columns are eigenvectors corresponding to the zero and non-zero eigenvalues, respectively; $\mathbf{\Lambda}_r := \mathbf{diag}\{\lambda_{N+1}(\mathbf{P}), \lambda_{N+2}(\mathbf{P}), \cdots, \lambda_n(\mathbf{P})\} \in \mathbb{R}^{(n-N)\times(n-N)}$. Here, the set of all nonzero eigenvectors in the columns of $\mathbf{Q}_r$ is an *orthonormal* basis of the row space of $\mathbf{P}$, implying that

$$\|\mathbf{Q}_r^T\mathbf{x}_r\|_2 = \|\mathbf{x}_r\|_2. \tag{D.4}$$

Let $\mathbf{z}_r \in \mathbb{R}^n$ be defined as $\mathbf{z}_r := \mathbf{Q}_r^T\mathbf{x}_r$. Then, (D.3) yields $\mathbf{x}_r^T\mathbf{P}\mathbf{x}_r = \mathbf{z}_r^T\mathbf{\Lambda}_r\mathbf{z}_r$ and thereby, we obtain

$$\lambda_{N+1}(\mathbf{P})\|\mathbf{z}_r\|_2^2 \leq \mathbf{x}_r^T\,\mathbf{P}\mathbf{x}_r \leq \lambda_n(\mathbf{P})\|\mathbf{z}_r\|_2^2. \tag{D.5}$$

On the other hand, $\mathbf{z}_r = \mathbf{Q}_r^T\mathbf{x}_r$ and (D.4) imply $\|\mathbf{z}_r\|_2 = \|\mathbf{x}_r\|_2$, and $\mathbf{x}_r^T\,\mathbf{P}\mathbf{x}_n = \mathbf{x}_n^T\,\mathbf{P}\mathbf{x}_n = 0$ implies $\mathbf{x}^T\,\mathbf{P}\mathbf{x} = \mathbf{x}_r^T\,\mathbf{P}\mathbf{x}_r$. Therefore, substituting these results and (D.2) into (D.5) and letting $\underline{\alpha} = \lambda_{N+1}(\mathbf{P})$ and $\bar{\alpha} = \lambda_n(\mathbf{P})$, one obtains (2.8), which completes the proof.

D.3 Proof of Theorem 3.3

The proof will be done by showing the corresponding conditions in Theorem 3.2. The partial derivatives of $Q_d^*(\bar{\mathbf{x}}_d;\lambda)$ with respect to $\mathbf{x}$ and $\mathbf{u}_d$ is evaluated as

$$\begin{aligned}\nabla_{\mathbf{x}}Q_d^*(\bar{\mathbf{x}}_d;\lambda) &= \lambda\nabla V^*(\mathbf{x}) + \nabla^2 V^{*T}(\mathbf{x})\mathbf{G}_d(\mathbf{x})\mathbf{u}_d + \mathbf{u}_d^T\nabla\mathbf{G}_d^T(\mathbf{x})\nabla V^{*T}(\mathbf{x}) + \mathbf{u}_d\nabla\mathbf{R}_d(\mathbf{x})\mathbf{u}_d \\ &= \lambda\nabla V^*(\mathbf{x}) + \nabla^2 V^{*T}(\mathbf{x})\mathbf{G}_d(\mathbf{x})\mathbf{u}_d + \sum_{j=1}^{m_d} u_{dj}\cdot\Big[\nabla\mathbf{g}_{dj}^T(\mathbf{x})\nabla V^*(\mathbf{x}) + u_{dj}\nabla r_{dj}(\mathbf{x})\Big]\end{aligned}$$

$$\nabla_{\mathbf{u}_d}Q_d^*(\bar{\mathbf{x}}_d;\lambda) = \mathbf{G}_d^T(\mathbf{x})\nabla V^*(\mathbf{x}) + 2\mathbf{R}_d(\mathbf{x})\mathbf{u}_d.$$

Hence, the term $(\nabla_{\mathbf{x}}Q_d^*(\bar{\mathbf{x}}_d))^T\big(\mathbf{f}_s(\mathbf{x}) + \mathbf{G}_d(\mathbf{x})\mathbf{u}_d\big)$ can be expanded as follows:

$$\begin{aligned}(\nabla_{\mathbf{x}}Q_d^*)^T\big(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d\big) &= \lambda\,(\nabla V^*)^T\big(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d\big) + \mathbf{u}_d^T\mathbf{G}_d^T(\nabla^2 V^*)\big(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d\big) \\ &\quad + \sum_{j=1}^{m_d} u_{dj}\cdot\Big[(\nabla V^*)^T\nabla\mathbf{g}_{dj} + u_{dj}\nabla^T r_{dj}\Big]\Big(\mathbf{f}_s + \mathbf{G}_d\mathbf{u}_d\Big).\end{aligned} \tag{D.6}$$

Similarly, the term $\lambda \cdot \nabla^T Q_d^*(\bar{\mathbf{x}}_d)\mathbf{B}_{0d}\mathbf{R}_d^{-1}(\mathbf{x})\mathbf{B}_{0d}^T \nabla Q_d^*(\bar{\mathbf{x}}_d)$ can be expanded using the expression of $\nabla_{\mathbf{u}_d} Q_d^*(\bar{\mathbf{x}}_d; \lambda)$ as

$$\begin{aligned}\lambda \cdot (\nabla Q_d^*)^T \mathbf{B}_{0d}\mathbf{R}_d^{-1}\mathbf{B}_{0d}^T(\nabla Q_d^*) &= \lambda \cdot (\nabla_{\mathbf{u}_d} Q_d^*)^T \mathbf{R}_d^{-1}(\nabla_{\mathbf{u}_d} Q_d^*) \\ &= \lambda \cdot \Big(\nabla^T V^* \mathbf{G}_d + 2\mathbf{u}_d^T \mathbf{R}_d\Big)\mathbf{R}_d^{-1}\Big(\mathbf{G}_d^T \nabla V^* + 2\mathbf{R}_d \mathbf{u}_d\Big).\end{aligned}$$

Then, rearranging the resultant equation multiplied by "1/4" yields

$$\begin{aligned}\frac{\lambda}{4}(\nabla Q_d^*)^T \mathbf{B}_{0d}\mathbf{R}_d^{-1}\mathbf{B}_{0d}^T(\nabla Q_d^*) &= \frac{\lambda}{4} \cdot \left(\nabla^T V^* \mathbf{G}_d \mathbf{R}_d^{-1}\mathbf{G}_d^T \nabla V^* + 4\mathbf{u}_d^T \mathbf{G}_d^T \nabla V^* + 4\mathbf{u}_d^T \mathbf{R}_d \mathbf{u}_d\right) \\ &= \lambda \cdot \left(\frac{1}{4}\nabla^T V^* \mathbf{G}_d \mathbf{R}_d^{-1}\mathbf{G}_d^T \nabla V^* + \mathbf{u}_d^T \mathbf{G}_d^T \nabla V^* + \mathbf{u}_d^T \mathbf{R}_d \mathbf{u}_d\right).\end{aligned} \tag{D.7}$$

Now, the substitutions of $(\nabla Q_d^*)^T \bar{\mathbf{f}}_s = (\nabla_{\mathbf{x}} Q_d^*)^T \big(\mathbf{f}_s + \mathbf{G}_d \mathbf{u}_d\big)$, (D.6), and (D.7), we have

$$\begin{aligned}&(\nabla Q_d^*)^T \bar{\mathbf{f}}_s - \frac{\lambda}{4}(\nabla Q_d^*)^T \bar{\mathbf{B}}_{0d}\mathbf{R}_d^{-1}\bar{\mathbf{B}}_{0d}^T(\nabla Q_d^*) \\ =&(\nabla_{\mathbf{x}} Q^*)^T \big(\mathbf{f}_s + \mathbf{G}_d \mathbf{u}_d\big) - \lambda \cdot \left(\frac{1}{4}\nabla^T V^* \mathbf{G}_d \mathbf{R}_d^{-1}\mathbf{G}_d^T \nabla V^* + \mathbf{u}_d^T \mathbf{G}_d^T \nabla V^* + \mathbf{u}_d^T \mathbf{R}_d \mathbf{u}_d\right) \\ =&\lambda\left(\nabla^T V^* \mathbf{f}_s - \frac{1}{4}\nabla^T V^* \mathbf{G}_d \mathbf{R}_d^{-1}\mathbf{G}_d^T \nabla V^*\right) + \mathbf{u}_d^T \mathbf{G}_d^T \nabla^2 V^* \big(\mathbf{f}_s + \mathbf{G}_d \mathbf{u}_d\big) + \lambda \mathbf{u}_d^T \mathbf{R}_d \mathbf{u}_d \\ &+ \sum_{j=1}^{m_d} u_{dj} \cdot \Big[\nabla^T V^* \nabla \mathbf{g}_{dj} + u_{dj} \nabla^T r_{dj}\Big]\Big(\mathbf{f}_c + \mathbf{G}_d \mathbf{u}_d\Big).\end{aligned}$$

Substituting the HJB equation (3.7) and rearranging it, we obtain

$$\begin{aligned}&(\nabla Q_d^*)^T \bar{\mathbf{f}}_s - \frac{\lambda}{4}(\nabla Q_d^*)^T \bar{\mathbf{B}}_{0d}\mathbf{R}_d^{-1}\bar{\mathbf{B}}_{0d}^T(\nabla Q_d^*) \\ =&-\lambda\left(S + \frac{1}{4}(\nabla V^*)^T \mathbf{G}_s \mathbf{R}_s^{-1}\mathbf{G}_s^T(\nabla V^*)\right) + \mathbf{u}_d^T \mathbf{G}_d^T \nabla^2 V^* \mathbf{f}_s \\ &- \mathbf{u}_d^T \Big[\lambda \mathbf{R}_d - \mathbf{G}_d^T (\nabla^2 V^*)\mathbf{G}_d\Big]\mathbf{u}_d + \sum_{j=1}^{m_d} u_{dj} \cdot \Big[\nabla^T V^* \nabla \mathbf{g}_{dj} + u_{dj} \nabla^T r_{dj}\Big]\Big(\mathbf{f}_c + \mathbf{G}_d \mathbf{u}_d\Big) \\ =&-\lambda\big(S + (\mathbf{u}_s^*)^T \mathbf{R}_s \mathbf{u}_s^*\big) + (\nabla V^*)^T \boldsymbol{\Xi}\Big(\mathbf{f}_c + \mathbf{G}_d \mathbf{u}_d\Big) + \mathbf{u}_d^T \mathbf{G}_d^T (\nabla^2 V^*)\mathbf{f}_c - \mathbf{u}_d^T \boldsymbol{\Sigma}(\bar{\mathbf{x}}_d; \lambda)\mathbf{u}_d \\ =&-\bar{S}_d(\bar{\mathbf{x}}_d; \lambda) - \lambda \cdot \mathbf{u}_s^{*T} \mathbf{R}_s \mathbf{u}_s^*\end{aligned}$$

which is obviously the HJB equation for the extended optimal control problem with the dynamics (3.9) and the performance index (3.13). Moreover, the dynamic policy $\mathbf{v}_d^*(\bar{\mathbf{x}}_d)$ can be rewritten as

$$\begin{aligned}\mathbf{v}_d^*(\bar{\mathbf{x}}_d;\lambda) &= -\frac{\lambda}{2}\mathbf{R}_d^{-1}\mathbf{G}_d^T(\nabla V^*) - \lambda\cdot(\mathbf{R}_d^{-1}\mathbf{R}_d)\mathbf{u}_d \\ &= -\frac{\lambda}{2}\cdot\mathbf{R}_d^{-1}\begin{bmatrix}\mathbf{0}_{m_d\times n} & \mathbf{I}_{m_d}\end{bmatrix}\begin{bmatrix}\nabla_{\mathbf{x}}Q_d^* \\ \mathbf{G}_d^T\nabla V^* + 2\mathbf{R}_d\mathbf{u}_d\end{bmatrix} \\ &= -\frac{1}{2}\cdot\big(\frac{1}{\lambda}\cdot\mathbf{R}_d\big)^{-1}\bar{\mathbf{B}}_{0d}^T\nabla Q_d^*.\end{aligned}$$

Let $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda)$ be defined as $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda) := \bar{S}_d(\bar{\mathbf{x}}_d;\lambda) + \lambda\cdot\mathbf{u}_s^{*T}(\mathbf{x})\mathbf{R}_s(\mathbf{x})\mathbf{u}_s^*(\mathbf{x})$. To proceed the proof, the following lemma is necessary.

Lemma D.1. *Suppose* $\lambda_1 < \lambda_2$. *Then,* $Q_d^*(\bar{\mathbf{x}}_d;\lambda_1) \preceq Q_d^*(\bar{\mathbf{x}}_d;\lambda_2)$ *and* $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda_1) \preceq \bar{S}_c(\bar{\mathbf{x}}_d;\lambda_2)$.

Proof of Lemma D.1. By the definitions of $Q_d^*(\bar{\mathbf{x}}_d;\lambda)$ and $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda)$, we have for $\lambda_1 < \lambda_2$

$$\begin{aligned}Q_d^*(\bar{\mathbf{x}}_d;\lambda_2) &= (\lambda_2-\lambda_1)V^*(\mathbf{x}) + Q_d^*(\bar{\mathbf{x}}_d;\lambda_1) \succeq Q_d^*(\bar{\mathbf{x}}_d;\lambda_1) \\ \bar{S}_c(\bar{\mathbf{x}}_d;\lambda_2) &= (\lambda_2-\lambda_1)\Big(S(\mathbf{x}) + \mathbf{u}_s^{*T}\mathbf{R}_s\mathbf{u}_s^* + \mathbf{u}_d^{*T}\mathbf{R}_d\mathbf{u}_d^*\Big) + \bar{S}_c(\bar{\mathbf{x}}_d;\lambda_1) \succeq \bar{S}_c(\bar{\mathbf{x}}_d;\lambda_1),\end{aligned}$$

which proves the lemma. □

Finally, Assumption 3.3 and Lemma D.1 imply $\forall\mathbf{x}\in\bar{B}_{\mathbb{S}}(r)\subset\Omega$ and $\forall\mathbf{u}_d\in\mathbb{R}^{m_d}$,

$$\begin{aligned}&\underline{\alpha}_q(d(\bar{\mathbf{x}}_d,\mathbb{S}_e)) \le Q_d^*(\bar{\mathbf{x}}_d;\underline{\lambda}) \le Q_d^*(\bar{\mathbf{x}}_d;\lambda) \\ &\underline{\alpha}_s(d(\bar{\mathbf{x}}_d,\mathbb{S}_e)) \le \bar{S}_d(\bar{\mathbf{x}}_d;\underline{\lambda}) \le \bar{S}_c(\bar{\mathbf{x}}_d;\underline{\lambda}) \le \bar{S}_c(\bar{\mathbf{x}}_d;\lambda)\end{aligned}$$

holds for any $\lambda \ge \underline{\lambda} > 0$, where $\Omega \subseteq \mathfrak{D}(\mathcal{D},\mathbb{S})$ is the domain of V^*. Hence, $Q_d^*(\bar{\mathbf{x}}_d;\lambda)$ and $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda)$ are positive semi-definite on $B_{\mathbb{S}}(r)\times\mathbb{R}^{m_d}$ and satisfy $Q_d^*(\bar{\mathbf{x}}_d;\lambda) = 0$ and $\bar{S}_c(\bar{\mathbf{x}}_d;\lambda) = 0$ iif $\bar{\mathbf{x}}_d \in \mathbb{S}_e$. Now that all the conditions in Theorem 3.2 are satisfied, it is proven by Theorem 3.2 that the dynamic policy $\mathbf{v}_d^*(\bar{\mathbf{x}}_d;\lambda)$ is the optimal admissible policy with respect to the performance index (3.13) and the extended subspace $\mathbb{S}_e$, and $Q_d^*(\bar{\mathbf{x}}_d;\lambda)$ is the corresponding optimal value function, the completion of the proof.

D.4 Proof of Theorem 5.1

Notice that by Theorem 5.3, the I-PI on the ROAs (Algorithm 5.1) and the ideal PI (Algorithm 3.1) generate the same sequence $(V_{\mathbf{u}_i}, \mathbf{u}_i)$ that converges to the optimal solution $(V^*, \mathbf{u}^*)$ under the conditions in Theorem 3.6. Hence, by Corollary 3.6, for each $i \in \mathbb{Z}_+$,

1. $\mathbf{u}_i$ is admissible on its ROA $R_A(\mathbf{u}_i)$;
2. $R_A(\mathbf{u}_i)$ is the invariant subset of $R_A(\mathbf{u}_{i+1})$, i.e.,

$$\mathbf{z} \in R_A(\mathbf{u}_i) \implies \begin{cases} \mathbf{x}_\tau(\mathbf{z}; \mathbf{u}_{i+1}, \mathbf{0}_m) \in R_A(\mathbf{u}_i) \ \ \forall \tau \geq t, \\ \lim_{\tau \to \infty} \mathbf{x}_\tau(\mathbf{z}, \mathbf{u}_{i+1}, \mathbf{0}_m) = \mathbf{0}_n; \end{cases} \tag{D.8}$$

3. each $V_{\mathbf{u}_i}$ is the unique solution to the Hamiltonian equation:

$$(\nabla V_{\mathbf{u}_i}(\mathbf{x}))^T (\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}_i(\mathbf{x})) = -r(\mathbf{x}, \mathbf{u}_i(\mathbf{x})). \tag{D.9}$$

   over the function space $C^1_{L+}(\mathbf{u}_i)$.

So, differentiating $V_{\mathbf{u}_i}(\mathbf{x})$ along the trajectory $\mathbf{x}(\mathbf{z}; \mathbf{u}_{i+1}, \mathbf{e})$ and substituting $2\mathbf{R}(\mathbf{x})\mathbf{u}_{i+1} = \mathbf{G}^T(\mathbf{x})\nabla V_{\mathbf{u}_i}(\mathbf{x})$ and (D.9) yields

$$\begin{aligned} \dot{V}_{\mathbf{u}_i}(\mathbf{x}) &= (\nabla V_{\mathbf{u}_i}(\mathbf{x}))^T \big(\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})[\mathbf{u}_{i+1}(\mathbf{x}) + \mathbf{e}]\big) \\ &= -r(\mathbf{x}, \mathbf{u}_i) - 2\mathbf{u}_{i+1}^T \mathbf{R}(\mathbf{x})\mathbf{e} - 2\mathbf{u}_{i+1}^T \mathbf{R}(\mathbf{x})(\mathbf{u}_{i+1} - \mathbf{u}_i). \end{aligned}$$

Applying Young's inequality $2\mathbf{x}^T\mathbf{R}(\mathbf{x})\mathbf{y} \leq \mathbf{x}^T\mathbf{R}(\mathbf{x})\mathbf{x} + \mathbf{y}^T\mathbf{R}(\mathbf{x})\mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, we obtain

$$\dot{V}_{\mathbf{u}_i}(\mathbf{x}) \leq -S(\mathbf{x}) + \mathbf{e}^T\mathbf{R}(\mathbf{x})\mathbf{e}. \tag{D.10}$$

Notice that $S(\mathbf{x})$ satisfies (3.17) for $r_s > 0$ such that $\bar{B}_{\mathbf{0}_n}(r_s) \subset \mathcal{D}$; $V_{\mathbf{u}_i}(\mathbf{x})$ satisfies (3.16) with $\mathbf{u} = \mathbf{u}_i$ for $r_{\mathbf{u}_i} > 0$ satisfying $\bar{B}_{\mathbf{0}_n}(r_{\mathbf{u}_i}) \subset R_A(\mathbf{u}_i)$. Since $R_A(\mathbf{u}_i) \subseteq \mathcal{D}$ holds by its definition, one can choose r_s such that $0 < r_{\mathbf{u}_i} \leq r_s$ holds. Hence, using $\underline{\alpha}_s(\|\mathbf{x}\|) \leq S(\mathbf{x})$

and $V_{\mathbf{u}_i}(\mathbf{x}) \le \bar{\alpha}_{\mathbf{u}_i}(\|\mathbf{x}\|)$, one obtains from the inequality (D.10)

$$\dot{V}_{\mathbf{u}_i} \le -(1-\theta)S(\mathbf{x}) - \theta \cdot \hat{\alpha}\big(V_{\mathbf{u}_i}(\mathbf{x})\big) + \bigg(\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))\bigg) \cdot \bigg(\sup_{t\le\tau<\infty} \|\mathbf{e}_\tau\|^2\bigg),$$

where $\hat{\alpha} := \underline{\alpha}_s \circ \bar{\alpha}_{\mathbf{u}_i}^{-1}$; $\theta \in (0,1)$ is a constant satisfying

$$\sup_{t\le\tau<\infty} \|e_\tau\|^2 < \theta \cdot \hat{\alpha}\big(d_i\big) \cdot \bigg(\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))\bigg)^{-1} \tag{D.11}$$

for $d_i := \bar{\alpha}_{\mathbf{u}_i}(r_{\mathbf{u}_i})$. Since we assume the exploration $\mathbf{e}$ satisfies (5.8), such θ always exists in $(0,1)$. Also notice that $\hat{\alpha} = \underline{\alpha}_s \circ \bar{\alpha}_{\mathbf{u}_i}^{-1}$ is a class $\mathcal{K}$ function defined on $[0, d_i]$. This is because we assume $0 < r_{\mathbf{u}_i} \le r_s$ without loss of generality, and the inverse $\bar{\alpha}_{\mathbf{u}_i}^{-1}(r)$ is a class $\mathcal{K}$ function defined for $r \in [0, d_i]$ by [32, Lemma 4.2]. Therefore, we have

$$\dot{V}_{\mathbf{u}_i}(\mathbf{x}) \le -(1-\theta)S(\mathbf{x}), \tag{D.12}$$

for all $\mathbf{x} \in R_A(\mathbf{u}_i)$ satisfying $V_{\mathbf{u}_i}(\mathbf{x}) \ge r_i$, where r_i is given by

$$r_i \equiv \hat{\alpha}^{-1}\Bigg(\theta^{-1} \cdot \bigg(\sup_{\mathbf{x}\in\mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x})\bigg) \cdot \bigg(\sup_{t\le\tau<\infty} \|\mathbf{e}_\tau\|^2\bigg)\Bigg) \tag{D.13}$$

and satisfies "$r_i < d_i$" by (D.11). Here, $r_i < d_i$ and (D.12) imply that $\dot{V}_{\mathbf{u}_i}(\mathbf{x})$ is negative definite on the compact subset $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i)$ given by

$$\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i) := \big\{\mathbf{x} \in \mathcal{D} : r_i \le V_{\mathbf{u}_i}(\mathbf{x}) \le d_i\big\}.$$

Hence, we have $\dot{V}_{\mathbf{u}_i} < 0$ on the boundary $\partial\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, d_i)$, implying that the state trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}_{i+1}, \mathbf{e})$ starting at any $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, d_i)$ at time $t > 0$ stays in $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, d_i)$ for all $\tau \ge t$. That is, $\mathbf{e}$ is invariant on $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, d_i) \equiv \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; d_i)$.

Next, applying (3.16) and (3.17) to (D.12) to prove ISS, we obtain

$$\dot{V}_{\mathbf{u}_i}(\mathbf{x}_\tau) \le -(1-\theta)\hat{\alpha}(V_{\mathbf{u}_i}(\mathbf{x}_\tau)) \tag{D.14}$$

$$\le -(1-\theta)\hat{\alpha}(r_i) \equiv -k < 0 \tag{D.15}$$

for all $\mathbf{x}_\tau \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i)$. Hence, (D.15) and the invariance of $\mathbf{e}$ on $\bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i)$ imply that for any $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i)$, there is $t' > t$ such that

$$\begin{cases} \mathbf{x}_\tau(\mathbf{z}, \mathbf{u}_{i+1}, \mathbf{e}) \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i) \text{ for all } \tau \in [t, t+t'), \\ \mathbf{x}_\tau(\mathbf{z}, \mathbf{u}_{i+1}, \mathbf{e}) \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, r_i) \text{ for all } \tau \geq t+t', \end{cases}$$

Assume $\hat{\alpha}(\cdot)$ is locally Lipshitz without loss of generality[2] and let v_τ be the solution to the scalar differential equation

$$\dot{v}_\tau = -(1-\theta)\hat{\alpha}(v_\tau)$$

under the initial condition $v(t) = V_{\mathbf{u}_i}(\mathbf{z})$. Then, [32, Lemma 3.4 and Lemma 4.4] and (D.14) show that there is $\beta_v \in \mathcal{KL}$, defined on $[0,\, d_i] \times [0, \infty)$, such that

$$V_{\mathbf{u}_i}(\mathbf{x}(\tau)) \leq v(\tau) = \beta_v\big(V_{\mathbf{u}_i}(\mathbf{z}),\, \tau - t\big),$$

for any initial condition $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; r_i, d_i)$ and all $\tau \in [t + t')$. Therefore, using (3.16) yields the following inequality:

$$\begin{aligned} \underline{\alpha}_{\mathbf{u}_i}\big(\|\mathbf{x}_\tau\|\big) \leq V_{\mathbf{u}_i}(\mathbf{x}_\tau) &\leq \beta_v\big(V_{\mathbf{u}_i}(\mathbf{z}),\, \tau - t\big) \\ &\leq \beta_v\big(\bar{\alpha}_{\mathbf{u}_i}(\|\mathbf{z}\|),\, \tau - t\big) \equiv \beta(\|\mathbf{z}\|,\, \tau - t\big), \end{aligned} \tag{D.16}$$

where $\beta(y, s) \equiv \beta_v\big(\bar{\alpha}_{\mathbf{u}_i}(y),\, s\big)$ is of class $\mathcal{KL}$ by [32, Lemma 4.2]. On the other hand, for all $\mathbf{x}_\tau \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; 0, r_i)$, we have $V_{\mathbf{u}_i}(\mathbf{x}_\tau) \leq r_i$, and from (3.16) and (D.13),

$$\underline{\alpha}_{\mathbf{u}_i}\big(\|\mathbf{x}_\tau\|\big) \leq \alpha\bigg(\sup_{t \leq s < \infty} \|\mathbf{e}(s)\|\bigg), \tag{D.17}$$

where $\alpha(y) \equiv \hat{\alpha}^{-1}\Big(y^2 \cdot \big(\theta^{-1} \cdot \sup_{\mathbf{x} \in \mathcal{D}} \lambda_1(\mathbf{R}(\mathbf{x}))\big)\Big)$ is of class $\mathcal{K}$ [32, Lemma 4.2].

[2] See the proof of [32, Theorem 4.9].

Finally, (D.16) and (D.17) imply that for all $\mathbf{z} \in \bar{\Omega}_{\mathcal{I}}(\mathbf{u}_i; d_i)$ and all $\tau \geq t$, the trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}_{i+1}, \mathbf{e})$ satisfies the inequality

$$\gamma\big(\|\mathbf{x}_\tau\|\big) \leq \max\left\{\beta\big(\|\mathbf{z}\|,\, \tau - t\big),\;\; \alpha\bigg(\sup_{t \leq s < \infty} \|\mathbf{e}(s)\|\bigg)\right\} \tag{D.18}$$

under (5.8), where $\gamma(\cdot) := \underline{\alpha}_{\mathbf{u}_i}(\cdot)$ is of class $\mathcal{K}$. Here, instead of $[t, \infty)$, the supremum on the right hand side can be chosen over $[t, \tau]$ since $\mathbf{x}_\tau$ depends only on $\mathbf{e}(s)$ for $t \leq s \leq \tau$. This completes the proof of the local ISS theorem under $\mathbf{u}_{i+1}$. In the case of $\mathbf{u}_i$, one can establish the inequality (D.10) by differentiating $V_{\mathbf{u}_i}(\mathbf{x})$ along the trajectory $\mathbf{x}_\tau(\mathbf{z}; \mathbf{u}_i, \mathbf{e})$ and then substituting (D.9). Then, following the procedure of the proof of the case $\mathbf{u}_{i+1}$ above, one obtains the same results.

D.5 Sketch of Proof of Theorem 6.2.

The proof starts by showing that $\bar{S}_d$ and Q_d^* in Section 3.2 are represented as $Q_d^*(\bar{\mathbf{e}}_s; \lambda) = \bar{\mathbf{e}}_s^T \mathbf{Q}_d(\mathbf{v}; \lambda) \bar{\mathbf{e}}_s$ and $\bar{S}_d(\bar{\mathbf{e}}_s; \lambda) = \bar{\mathbf{e}}_s^T \bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s; \lambda) \bar{\mathbf{e}}_s$. This can be easily verified as follows.

$$\begin{aligned}
Q_d^*(\bar{\mathbf{e}}_s; \lambda) =& \lambda \mathbf{e}^T \mathbf{\Pi} \mathbf{e} + 2\mathbf{e}^T \mathbf{\Pi} \mathbf{B}_{\mathbf{v}2}^{\otimes} \mathbf{w} + \gamma \mathbf{w}^T \mathbf{D}_{\mathbf{v}}^2 \mathbf{w} \\
=& \bar{\mathbf{e}}_s^T \mathbf{Q}_d(\mathbf{v}; \lambda) \bar{\mathbf{e}}_s \\
\bar{S}_d(\bar{\mathbf{e}}_s; \lambda) =& \lambda\, \mathbf{e}^T \mathbf{\Theta} \mathbf{e} + \mathbf{w}^T \mathbf{\Sigma}(\bar{\mathbf{e}}_s; \lambda) \mathbf{w} - 2\mathbf{w}^T (\mathbf{B}_{\mathbf{v}2}^{\otimes})^T \mathbf{\Pi} \mathbf{A}_s^{\otimes} \mathbf{e} - 2\mathbf{e}^T \mathbf{\Pi} \mathbf{\Xi}(\bar{\mathbf{e}}_s)(\mathbf{A}_s^{\otimes} \mathbf{e} + \mathbf{B}_{\mathbf{v}2}^{\otimes} \mathbf{w}) \\
=& \bar{\mathbf{e}}_s^T \bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s; \lambda) \bar{\mathbf{e}}_s.
\end{aligned}$$

The next step is to show that there is $\underline{\lambda}_Q > 0$ such that for all $\mathbf{v} \in \mathbb{R}^{2N}$ and all $\lambda \geq \underline{\lambda}_Q$, $\mathbf{Q}_d(\mathbf{v}; \lambda)$ is positive semi-definite and $\operatorname{rank} \mathbf{Q}_d(\mathbf{v}; \lambda) = 5N - 2$. For this, consider the matrix decomposition $\bar{\mathbf{\Pi}} := \bar{\mathbf{U}}^T \mathbf{\Pi} \bar{\mathbf{U}}$, where $\bar{\mathbf{U}} := \mathbf{U} \otimes \mathbf{I}_4$, and $\mathbf{U}$ is the orthogonal matrix given in Corollary C.1. Then, the structure of $\mathbf{\Pi}$, Corollary C.1, and the operations of the Khatri-Rao products in Appendix A show that $\bar{\mathbf{\Pi}}$ can be represented as $\bar{\mathbf{\Pi}} = \mathbf{diag}\{\mathbf{0}_{2\times 2}, \mathbf{P}\}$ for some positive *definite* matrix $\mathbf{P} \in \mathbb{R}^{(4N-2)\times(4N-2)}$. Now, revoking the Schur complement lemma (Lemmas B.1 and B.2), we know that $\mathbf{Q}_d$ is positive semi-definite and

$\operatorname{rank}\mathbf{Q}_d = 5N-2$ iif $\gamma\mathbf{D}_\mathbf{v}$ is positive definite, which is obvious under Assumption 6.1, and $\mathcal{S}(\mathbf{Q}_d)$ is positive semi-definite with $\operatorname{rank}\mathcal{S}(\mathbf{Q}_d) = 4N-2$. Here, $\mathcal{S}(\mathbf{Q}_d)$, the Schur complement of $\mathbf{Q}_d$ (see Appendix B), is given by $\mathcal{S}(\mathbf{Q}_d) = \lambda\mathbf{\Pi} - \gamma^{-1}\mathbf{\Pi}\mathbf{B}^{\otimes}_{\mathbf{v}2}\mathbf{D}_\mathbf{v}^{-2}(\mathbf{B}^{\otimes}_{\mathbf{v}2})^T\mathbf{\Pi}$, where $\mathbf{B}^{\otimes}_{\mathbf{v}2}\mathbf{D}_\mathbf{v}^{-2}(\mathbf{B}^{\otimes}_{\mathbf{v}2})^T$ is bounded under Assumption 6.1 since it consists of only the products of $\cos\theta_i$ and $\sin\theta_i$ $(i \in \mathcal{N})$. Now, the operations of Khatri-Rao products in Appendix A, the definitions of the matrices, and $\bar{\mathbf{\Pi}} = \mathbf{diag}\{\mathbf{0}_{2\times 2}, \mathbf{P}\}$ prove that its similarity transformation

$$\bar{\mathbf{U}}^T\mathcal{S}(\mathbf{Q}_d)\bar{\mathbf{U}} = \lambda\bar{\mathbf{\Pi}} - \frac{1}{\gamma}\bar{\mathbf{\Pi}}\bar{\mathbf{U}}^T\mathbf{B}^{\otimes}_{\mathbf{v}2}\mathbf{D}_\mathbf{v}^{2}(\mathbf{B}^{\otimes}_{\mathbf{v}2})^T\bar{\mathbf{U}}\bar{\mathbf{\Pi}}$$

can be compactly rewritten as $\bar{\mathbf{U}}^T\mathcal{S}(\mathbf{Q}_d)\bar{\mathbf{U}} = \mathbf{diag}\{\mathbf{0}_{2\times 2}, \lambda\bar{\mathbf{P}} - \bar{\mathbf{Y}}(\mathbf{v})\}$ for some positive defininte matrix $\bar{\mathbf{P}}$ and some bounded positive semi-definite matrix $\bar{\mathbf{Y}}(\mathbf{v})$. Then, since $\bar{\mathbf{Y}}(\mathbf{v})$ is bounded and $\bar{\mathbf{P}}$ is positive *definite*, there exists $\underline{\lambda}_Q > 0$ such that for all $\lambda \geq \underline{\lambda}_Q$, $\lambda\bar{\mathbf{P}} - \bar{\mathbf{Y}}(\mathbf{v})$ is positive definite for all $\mathbf{v} \in \mathbb{R}^{2N}$. For such λ, $\mathcal{S}(\mathbf{Q}_d)$ is positive semi-definite with its rank $4N-2$, which proves that so is $\mathbf{Q}_d(\mathbf{v};\lambda)$ for $\lambda \geq \underline{\lambda}_Q$ with its rank $5N-2$.

Notice that the nullities of $\mathcal{S}(\mathbf{Q}_d)$ and $\mathbf{\Pi}$ are '2' for $\lambda \geq \lambda_Q$ and that $\mathbf{e} \in \ker\mathbf{\Pi}$ implies $\mathbf{e} \in \ker(\mathbf{B}^{\otimes}_{\mathbf{v}2})^T\mathbf{\Pi}$ for all $\mathbf{v} \in \mathbb{R}^{2N}$. Hence, the application of Lemma B.3 proves $\ker\mathbf{Q}_d(\mathbf{v};\lambda) = \mathbb{S}_e$ $\forall\mathbf{v} \in \mathbb{R}^{2N}$ $\forall\lambda \geq \lambda_Q$. Here, the stabilizing subspace $\mathbb{S}_e$ satisfies

$$\mathbb{S}_e = \big\{\bar{\mathbf{e}}_s = (\mathbf{e},\mathbf{w}) \in \mathbb{R}^{5N} : \mathbf{e} \in \ker\mathbf{\Pi} \text{ and } \mathbf{w} \equiv \mathbf{0}_N\big\}.$$

This is because Definition 3.1 guarantees that the policy $\mathbf{w}^*(\mathbf{x})$ satisfies $\mathbf{w}^*(\mathbf{x}) = \mathbf{0}_N$ whenever $\mathbf{x} \in \mathbb{S}_e$. Finally, the application of Lemma 2.4 shows the existence of class $\mathcal{K}$ functions $\underline{\alpha}$ and $\bar{\alpha} > 0$ such that

$$\underline{\alpha}(d(\bar{\mathbf{e}}_s, \mathbb{S}_e)) \leq Q_d^*(\bar{\mathbf{e}}_s;\lambda) \leq \bar{\alpha}(d(\bar{\mathbf{e}}_s, \mathbb{S}_e)) \tag{D.19}$$

for all $\lambda \geq \underline{\lambda}$ and all $\bar{\mathbf{e}}_s \in \mathbb{R}^{5N}$.

Similarly to this, one can also show that for any $r > 0$, there is $\underline{\lambda}_s > 0$ such that for the function $S_d(\bar{\mathbf{e}}_s;\lambda)$, there exist class $\mathcal{K}$ functions $\underline{\beta}$ and $\bar{\beta}$ such that

$$\underline{\beta}(d(\bar{\mathbf{e}}_s,\mathbb{S}_e)) \le \bar{S}_d(\bar{\mathbf{e}}_s;\lambda) \le \bar{\beta}(d(\bar{\mathbf{e}}_s,\mathbb{S}_e)) \tag{D.20}$$

for all $\lambda \ge \underline{\lambda}$ and all $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$. In this case, the term $\lambda\mathbf{\Theta}$ in $\bar{\mathbf{S}}_d$ plays the same role to $\lambda\mathbf{\Pi}$ in $\mathbf{Q}_d$, and one can only show in this case that the terms in $\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda)$ are bounded not globally, but only in a ball $\bar{B}_{\mathbb{S}_e}(r)$, where $r > 0$ can be chosen arbitrarily. Though the structure of $\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda)$ is more complicated and $\bar{B}_{\mathbb{S}_e}(r)$ is non-compact, the boundedness of all the terms in $\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s;\lambda)$ can be shown using the structures and definitions of the matrices and the algebra of Katri-Rao products shown in Appendix A. Then, for any given $r > 0$, one can prove the existence of $\underline{\lambda}_s$ such that for all $\lambda \ge \underline{\lambda}_s$ and all $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$,

1. $\mathbf{\Sigma}(\bar{\mathbf{e}}_s;\lambda)$ is positive definite;

2. the Schur complement of $\bar{\mathbf{S}}_d$, $\mathcal{S}(\bar{\mathbf{S}}_d)$, is positive semi-definite with its rank $4N-2$ by the following expression of the transformed one:

$$\bar{\mathbf{U}}^T\mathcal{S}(\bar{\mathbf{S}}_d)\bar{\mathbf{U}} = \mathbf{diag}\{\mathbf{0}_{2\times 2}, \lambda\bar{\mathbf{P}} - \bar{\mathbf{Q}}(\bar{\mathbf{e}}_s;\lambda)\}, \tag{D.21}$$

   where $\bar{\mathbf{P}}$ is a positive definite matrix, and $\bar{\mathbf{Q}}(\bar{\mathbf{e}}_s;\lambda)$ is a symmetric matrix that is bounded on $\bar{B}_{\mathbb{S}_e}(r) \times [\underline{\lambda}_s,\infty) \subset \mathbb{R}^{5N}$.

Here, by the structures of $\mathbf{S}_d$, $\mathbf{\Pi}$, and $\mathbf{\Theta}$ with the property $\ker\mathbf{\Pi} = \ker\mathbf{\Theta}$ in Lemma 6.2, one can verify the expression (D.21). Moreover, $\mathbf{\Theta}$, $\mathbf{\Pi}$, $\mathcal{S}(\mathbf{S}_d)$ have the same nullity "2" for $\lambda \ge \underline{\lambda}_s$ and $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$.

Let $\bar{\mathbf{S}}_{d11}$ and $\bar{\mathbf{S}}_{d12}^T$ be the (1,1)- and (2,1)-th block elements of $\mathbf{S}_d$. That is,

$$\begin{aligned}\bar{\mathbf{S}}_{d11} &:= \mathbf{\Theta} - (\mathbf{A}_s^{\otimes})^T\mathbf{\Xi}^T\mathbf{\Pi} - \mathbf{\Pi}\mathbf{\Xi}\mathbf{A}_s^{\otimes}\\ \bar{\mathbf{S}}_{d12}^T &:= -\big(\mathbf{\Pi}\mathbf{\Xi}^T + (\mathbf{B}_{\mathbf{v}2}^{\otimes})^T\mathbf{\Pi}\big)\mathbf{A}_s^{\otimes}.\end{aligned}$$

Then, since $\bar{\mathbf{S}}_{d11} = \mathcal{S}(\bar{\mathbf{S}}_d) + \bar{\mathbf{S}}_{d12}\mathbf{\Sigma}^{-1}(\bar{\mathbf{e}}_s;\lambda)\bar{\mathbf{S}}_{d12}^T \succeq \mathcal{S}(\bar{\mathbf{S}}_d)$ by the definition of the Schur complement, $\bar{\mathbf{S}}_{d11}$ is also positive semi-definite, and its structure also show that $\bar{\mathbf{S}}_{d11}$ has the same nullity to $\mathcal{S}(\bar{\mathbf{S}}_d)$ for all $\lambda \geq \underline{\lambda}_s$ and $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$. On the other hand, the following can be established

$$\mathbf{x} \in \ker \mathbf{\Pi} \implies \mathbf{x} \in \ker \mathcal{S}(\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s; \lambda)) \text{ and } \mathbf{x} \in \ker \bar{\mathbf{S}}_{d11}(\bar{\mathbf{e}}_s; \lambda), \ \ \forall \lambda \geq \underline{\lambda}_s, \ \ \forall \bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$$

from the definition of $\mathbf{\Theta}$ and the kernel relation $\ker \mathbf{\Pi} \subset \ker \mathbf{A}_s^{\otimes}$. Since the dimensions of the null spaces of $\mathbf{\Pi}$, $\mathcal{S}(\mathbf{S}_d)$, and $\bar{\mathbf{S}}_{d11}$ for all $\lambda \geq \underline{\lambda}_s$ and all $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$ are all same, the converse of the statement is also true, so that we have

$$\ker \mathbf{\Pi} = \ker \mathcal{S}(\bar{\mathbf{S}}_d(\bar{\mathbf{e}}_s; \lambda)) = \ker \bar{\mathbf{S}}_{d11}(\bar{\mathbf{e}}_s; \lambda), \ \ \forall \lambda \geq \underline{\lambda}_s, \ \ \forall \bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r).$$

Moreover, from this, $\ker \mathbf{\Pi} \subset \ker \mathbf{A}_s^{\otimes}$, and the definition of $\bar{\mathbf{S}}_{d12}^T$, we easily obtain the kernel relation $\bar{\mathbf{S}}_{d11}^T \subseteq \bar{\mathbf{S}}_{d12}^T$. Hence, the application of Lemma B.3 proves $\ker \mathbf{S}_d(\bar{\mathbf{e}}_s; \lambda) = \mathbb{S}_e$ for $\lambda \geq \underline{\lambda}_s$ and $\bar{\mathbf{e}}_s \in \bar{B}_{\mathbb{S}_e}(r)$, and we obtain (D.20) by the application of Lemma 2.4.

Finally, (D.19) and (D.20) guarantees that Assumption 3.3 in Theorem 3.3. Moreover, since $r > 0$ can be arbitrarily chosen and $Q_d^*(\bar{\mathbf{e}}_s\lambda)$ is radially unbounded for $\lambda \geq \underline{\lambda}_Q$, one can choose $r > 0$ such that $\bar{B}_{\mathbb{S}_e}(r)$ contains a level set $\{Q_d^*(\bar{\mathbf{e}}_s\lambda) \leq c\}$ on which the initial condition $\mathbf{z}$ lies. This choice yields the lower bound $\underline{\lambda}_s$ on $\lambda > 0$ for (D.20), and the application of Theorem 3.3 with $\underline{\lambda} = \max\{\underline{\lambda}_Q, \underline{\lambda}_s\} > 0$ completes the proof.

국문요약

연속시간 동적 시스템을 위한 적분 강화학습과 적응 역최적 제어

본 학위논문에서는 연속시간 동적 시스템을 위한 적분 강화학습(integral reinforcement learning: IRL)과 적응 역최적 제어(adaptive inverse optimal control)에 관한 연구를 수행한다. 이러한 일련의 연구의 최종목표는 대상 동적 시스템에 대한 진정한 의미의 적응최적제어를 실현하는 것이다. 이러한 적응최적제어의 실현은, 제어공학과 기계학습 분야에서 긴 시간동안, 그리고 현재에도, 도전적인 과제로 남아있다.

적분 강화학습이란 미지의, 혹은 부분적으로 알려지지 않은, 연속시간 동적 시스템에 대한 최적제어 법칙을 적분형태의 보상을 이용하여 학습하는 강화학습의 일종이다. 먼저, 본 논문에서는 적분 정책반복법(integral policy iteration), 무한소 일반정책반복법(infinitesimal GPI), 적분 평가치 반복법(integral value iteration), 그리고 이들을 모두 포괄하는 적분 일반정책반복법(integral generalized policy iteration)등과 같은 다양한 준모델독립형(partially model-free) IRL에 대해 소개하고, 이를 분석한다. 수학적 분석을 통해 이들 방법에 대한 새로운 분류법을 제시하고, 폐루프 안정도와 단조수렴 조건을 제시한다. 다음으로, 적분 정책반복 알고리즘 기반으로, 탐색신호의 부정적 영향을 제거한 탐색화된(explorized) 적분 정책반복법과 모델독립형(model-free) IRL 방법인 적분 Q-학습법을 제안한다. 제안된 방법들은 안정한 상태공간 영역을 탐색하면서 파라미터들을 갱신할 수 있으며, 이를 통해 미지의 비선형 최적제어문제에 대한 솔루션을 실시간으로 도출할 수 있다. 마지막으로, 수학적 분석과 모의실험을 통해 제안된 기법의 성능과 이론적 성과를 최종 검증한다.

적응 역최적 제어에 관한 연구에서는, 연속시간 동적 시스템 모델을 갖는 다수 이동로봇

에 대한 협업 그래프 대형 제어문제 (cooperative graphical formation control problem)를 고려한다. 여기서 이동로봇의 기구학(kinematics)와 동역학(dynamics)은 역최적 일치제어와 적응법칙의 독립적 설계를 위한 일치제어 오차와 동역학 모델로 변환된다. 제안되는 기법은 최적성과 적응성을 함께 고려한 설계법으로, 역최적 2차 일치제어 방법, 역최적 동적 입력 확장기법, Lyapunov 함수 기반 적응법칙 설계법 등의 요소기술의 개발 및 결합을 통해 도출된다. 본 제어이론적 접근법에 의해 설계된 그래프 대형 제어 기법은, 주어진 통신 토폴로지에 대한 역최적성을 근사적으로 제공한다. 또한, Lyapunov와 Hamiltonian 분석을 통해 안정도와 파라미터 수렴성, 역최적성을 수학적으로 보이며, 모의실험을 통해 제안된 방법의 성능과 이론적 결과들을 다양한 시나리오에 대해 최종 검증한다.

핵심되는 말: 적응최적제어, 적분 강화학습, 적응 역최적 제어, 연속시간, 동적 시스템, 적분 정책반복법, 적분 일반정책반복법, 적분 Q-학습법, 이동로봇, 그래프 대형제어, 일치제어, 다개체 시스템, 최적성, 역최적성, 안정도, 수렴성