Policy Gradient Methods for Linear Quadratic Problems



Huining Yang

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Acknowledgements

First and foremost, I am extremely grateful to those who have financially sponsored my PhD studies. This includes the Engineering and Physical Sciences Research Council (EPSRC) and the Centre for Doctoral Training (CDT) in Industrially Focused Mathematical Modelling (InFoMM), and my industrial partner bp plc.

I would like to express my deepest gratitude to my supervisor, Prof. Ben Hambly for all his support, encouragement, patience, and immense knowledge during my PhD studies. It has been my greatest honour and pleasure to be his student and the completion of this thesis would not have been possible without him. I would also like to thank my collaborator, Prof. Renyuan Xu from the University of Southern California for her valued support and advice which was very influential in my academic research and career plan. My sincere thanks also goes to Prof. Chris Breward, Prof. Álvaro Cartea, Prof. Sam Cohen, and Prof. Colin Please for providing guidance and feedback during my PhD. My gratitude extends to my industrial supervisors, Paul Barnes and Amr El Zanfally from bp plc, who always encouraged me in my research project. I would also like to express a special thanks to those who inspired me to pursue a PhD degree during my undergraduate studies at the University of Manchester: Prof. David Silvester, Prof. Goran Peskir, and Dr. Denis Denisov.

Last but not least, I would like to thank my parents, Weiqing Yang and Zhongjing Wei, who have given me constant support and love during my life. I also wish to express a special thanks to my friend, Yu Tian, for her insightful suggestions in my research and a cherished time spent together. My appreciation also goes out to my colleagues and friends from the Mathematical and Computational Finance research group: Dr. Jiawei Li, Jing Ye, Dr. Yufei Zhang, and Dr. Christina Zou for sharing their experience with me and helping me all through my studies.

Statement of Originality

I declare that this thesis contains no material which has been accepted or is currently being submitted for any other degree, diploma, certificate or other qualifications at the University of Oxford or elsewhere.

Chapter 1 of this thesis contains materials from a preprint [41]. Chapter 2 is based on a paper published in the SIAM Journal on Control and Optimization [40]. Chapter 3 is based on a preprint [39]. The three papers are co-authored with my supervisor Prof. Ben Hambly, and Prof. Renyuan Xu from the University of Southern California. Chapter 4 is also joint work and part of an ongoing research project with these co-authors.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Linear Quadratic Problems 2

1.3 Reinforcement Learning and Policy Gradient Methods 3

1.3.1 Set-up 4

1.3.2 Classification of RL Algorithms 7

1.4 Outline 9

1.5 Notation 10

2 Policy Gradient Methods for the Linear Quadratic Regulator **11**

2.1 Introduction 11

2.1.1 Related Work 13

2.2 Problem Set-up 15

2.2.1 Application: The Optimal Liquidation Problem 17

2.3 Exact Gradient Methods with Known Parameters 19

2.3.1 Regularity of the LQR Problem and Properties of the Gradient Descent Dynamics 22

2.3.2 Perturbation Analysis of Σ_K 26

2.3.3 Convergence and Complexity Analysis 29

2.4 Sample-based Policy Gradient Method with Unknown Parameters . . . 32

2.4.1 Perturbation analysis of $C(\boldsymbol{K})$ and $\nabla_t C(\boldsymbol{K})$ 36

2.4.2 Smoothing and the Gradient Descent Analysis 37

2.4.3 Proof of Theorem 2.4.4 41

2.4.4 Discussion 43

2.5 Numerical Experiments 44

2.5.1 Optimal Liquidation within the LQR Framework 44

2.5.2 Learning to Liquidate without Model Specification 49

2.5.3 Learning LQR in Higher Dimensions 52

3 Policy Gradient Methods for the Linear Quadratic Game 55

3.1 Introduction 55
3.1.1 Comparison to the Literature on General-sum LQGs. 57
3.2 Problem Set-up 58
3.3 The Natural Gradient Method with Known Parameters 61
3.3.1 Regularity of the LQG and Properties of the Gradient Descent Dynamics 67
3.3.2 Perturbation Analysis of the Second Moment Matrix 70
3.3.3 Convergence and Complexity Analysis 75
3.4 The Natural Gradient Method with Unknown Parameters 87
3.5 Numerical Experiments 89
3.5.1 Convergence of the Natural Policy Gradient Algorithm 89
3.5.2 Effect of the System Noise 92

4 Bargaining Problems 96

4.1 The LQG Model with Partial Information 96
4.1.1 A Two-player LQG Model with Kalman Filter 97
4.1.2 General Setting 100
4.1.2.1 Preliminary Results with Known Model Parameters . . 104
4.1.3 Nash Equilibrium Strategy and Value Function 106
4.1.3.1 Preliminary: End Game 106
4.1.3.2 Dynamic Programming Principle (DPP) 112
4.1.3.3 General Results by the DPP 116
4.1.4 Experimental Result of the Bargaining Model 122
4.2 The LQG Model with Reinforcement Learning 124
4.2.1 A Two-player LQG model 124
4.2.2 Experimental Results 126

5 Conclusion and Future Work 129

A Appendix for Chapter 2 131

A.1 Market Simulator for Linear Price Dynamics 131
A.2 Comparison between the Policy Gradient Method and Q-learning . . . 133
A.3 Proofs of Technical Results 135
A.3.1 Proofs in Section 2.3.1 135

A.3.2 Proofs in Section 2.3.2 140
A.3.3 Proofs in Section 2.3.3 141
A.3.4 Proofs in Section 2.4 143
A.3.5 Proofs in Section 2.5 155

B Appendix for Chapter 3 **156**
B.1 Proofs in Section 3.3.1 156
B.2 Proofs in Section 3.3.2 161

Bibliography **164**

Abstract

The linear quadratic framework is widely studied in the literature on stochastic control and game theory due to its simple structure, the existence of tractable solutions, and its applicability to various real-world problems. In this thesis we provide results on the theoretical convergence of policy gradient methods, which are the basis of many reinforcement learning algorithms, in both single-agent and multi-agent linear quadratic problems.

In the single-agent setting, we show the global convergence of such methods to the optimal solution in the case of known and unknown parameters. In the case of known parameters, the policy gradients can be calculated explicitly; while when the model parameters are unknown, we estimate the gradients from samples using a derivative-free method. We illustrate the performance of policy gradient methods in an optimal liquidation problem and show results for the case where we assume a model for the underlying dynamics and where we apply the method to the data directly.

In the multi-agent linear quadratic game, we show that the policy gradient method enjoys global convergence to the Nash equilibrium provided that there is a certain amount of noise in the system. We give a condition, essentially a lower bound on the covariance of the noise in terms of the model parameters, in order to guarantee convergence. The noise can either come from the underlying dynamics, or carefully designed explorations from the agents. We illustrate our results with numerical experiments to show that even in situations where the policy gradient method may not converge in the deterministic setting, the addition of noise leads to convergence.

The work in the thesis is motivated by bargaining and we conclude the thesis by constructing LQG models for bargaining. We consider the case where model parameters are known to the players while they do not have access to the true state, and where parameters are unknown and players learn the optimal strategy through repeated negotiations. We illustrate performance with numerical experiments.

List of Figures

2.1 Performance of the PPG algorithms (50 simulation scenarios). 47
2.2 Optimal inventory trajectory under different deadlines (200 simulation scenarios). 48
2.3 Inventory trajectories of AAPL under different ϕ (average across 200 simulation scenarios). 48
2.4 Original Almgren-Chriss framework versus LQR formulation under different ϵ (AAPL). 48
2.5 Empirical distribution of the relative performance on the test set. . . . 52
2.6 Performance of the policy gradient algorithms (50 simulation scenarios) 53
2.7 Performance of the policy gradient algorithm with unknown parameters under different step size η (50 simulation scenarios). 54

3.1 Normalized error under different σ^2 when $r = 0.25$ ($\eta_1 = \eta_2 = 0.1$ and $N = 5000$). 91
3.2 Normalized error under different σ^2 when $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$ and $M = 5000$). 91
3.3 Normalized error and trajectories of learned policies with $\sigma^2 = 0$ and $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$, $M = 10000$). The peak period between iterations 300-900 is indicated in grey. 92
3.4 Performance of the natural policy gradient algorithm with $r = 0.42$ and $\eta_1 = \eta_2 = 0.001$ ($M = 10000$). 92
3.5 Performance of the natural policy gradient algorithm with $r = 0.42$, $\eta_1 = 0.001$, and $\eta_2 = 0.01$ ($M = 10000$). 93
3.6 Performance of the natural policy gradient algorithm with $r = 0.16$ and $\eta_1 = \eta_2 = 0.001$ ($M = 10000$). 93
3.7 Performance of the natural policy gradient algorithm with unknown parameters ($\eta_1 = \eta_2 = 0.0005$). 94

3.8 Trajectory (indicated in red) of learned policy $K_0^1 = (K_{00}^1, K_{01}^1)$ ($\eta_1 = \eta_2 = 0.0005$, $M = 500$). 95

4.1 The true state (solid black lines) and state estimates (dashed blue lines) of the buyer and the seller. The top, middle, and bottom solid lines represent the true seller's offer, (the true) value of the project, and (the true) buyer's offers, respectively. The solid lines represent each player's estimates of the three true state variables. 123
4.2 An example of state trajectories under the Nash equilibrium with full information (solid lines) and partial information (dashed lines). The red dashed line represents the buyer's estimate of the project value. For comparison, the buyer's Nash equilibrium cost with full information is $J^B = 920$. 124
4.3 State trajectory under Nash equilibrium strategies. 127
4.4 Performance of the algorithm with known parameters for the bargaining model. 127
4.5 Performance of the algorithm with unknown parameters for the bargaining model. 128

A.1 Relationship between MFI and ΔS. (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019, $\Delta T = 1$min) 132
A.2 Average queue length (volume) of the first five levels on the limit buy side (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019 with 5000 samples uniformly sampled with natural time clock in each trading day.) 132
A.3 Q tables with 10 states and 10 actions: $q_0(s,a)$, $q_4(s,a)$ and $q_5(s,a)$ (from left to right). 134
A.4 Q tables with 100 states and 100 actions: $q_0(s,a)$, $q_4(s,a)$ and $q_5(s,a)$ (from left to right). 134
A.5 Comparison between Q-learning and the policy gradient method (log-log scale). (Average across 10 scenarios.) 135

Chapter 1

Introduction

This thesis focuses on the theoretical convergence of policy gradient methods, one of the most popular reinforcement learning methods, for linear quadratic problems which are fundamental problems in control theory. In this chapter, we will start with the motivation of the thesis and briefly introduce the general form of linear quadratic problems. We will also discuss reinforcement learning techniques and especially policy gradient methods. The outline of this thesis and the notation used throughout the thesis are provided at the end.

1.1 Motivation

This thesis is motivated by a bargaining problem of interest to our industrial partner, bp plc. An oil company such as bp needs to negotiate with a counterparty, usually a foreign government or another oil company, to gain the rights for oil exploration and extraction. We will consider a stylised setting where a single buyer and a single seller must agree on the value of a project. Here a project may consist of a series of activities, such as oil exploration and extraction. Each party has a reservation price or a target price they want to achieve. For the buyer the reservation price is the maximum they are willing to pay for the project and for the seller, it is the minimum they are willing to accept. The target price of a player is an agreed-upon price they want to achieve at the deadline, which depends on their view of the project's true value. Neither party knows the reservation price or the target price of the other party. A key question is to determine a party's optimal bidding strategy to secure the project along with the best price that can be achieved. This is particularly challenging when each party

has incomplete or partial information about their counterparty, under both external (e.g. market fluctuations) and internal (e.g. change of strategy inside the company) uncertainties.

This thesis will explore the classical linear quadratic framework and discuss how to find the optimal solution in linear quadratic problems using reinforcement learning techniques. In the final chapter we will show how to formulate the bargaining problem as linear quadratic problems and apply policy gradient methods to learn the optimal negotiation strategy.

1.2 Linear Quadratic Problems

Linear quadratic (LQ) problems, including the single-agent problem known as the *linear quadratic regulator* (LQR), and the multi-agent problem known as the *linear quadratic game* (LQG), are one of the most fundamental in optimal control theory. In an optimal control problem the aim is to find the optimal control or action in a dynamical system such that an objective function is optimised. In LQ problems, the state dynamics are described by a linear function of the current state and control, and the cost functions are quadratic in the current state and control. In this thesis, we focus on discrete-time LQ problems with stochastic dynamics over a finite horizon T ($T > 0$). More precisely, the state equation is given by

$$x_{t+1} = A_t x_t + \sum_{i=1}^{N} B_t^i u_t^i + w_t, \quad x_0 \sim \mathcal{D}, \quad t = 0, 1, \cdots, T-1, \tag{1.2.1}$$

where N ($N \geq 1$) is the number of players, $x_t \in \mathbb{R}^d$ is the state variable of the system with the initial state x_0 drawn from a distribution $\mathcal{D}$, $u_t^i \in \mathbb{R}^{k_i}$ is the control of player i ($i = 1, \ldots, N$) at time t and $\{w_t\}_{t=0}^{T-1}$ are zero-mean independent and identically distributed (i.i.d.) random variables which are independent of x_0. Throughout the thesis we consider the class of (adaptive) Markovian strategies as the admissible controls. The system parameters $A_t \in \mathbb{R}^{d \times d}$, $B_t^i \in \mathbb{R}^{d \times k_i}$, for $t = 0, 1, \cdots, T-1$ are referred to as system transition matrices. The objective of player i ($i = 1, \ldots, N$) is to minimise their finite time horizon value function:

$$\inf_{\{u_t^i\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right], \tag{1.2.2}$$

where the cost function c_t^i is given by

$$c_t^i(x_t, u_t^i) = x_t^\top Q_t^i x_t + (u_t^i)^\top R_t^i u_t^i, \quad t = 0, 1, \cdots, T-1, \tag{1.2.3}$$

with $c_T^i(x_T) = x_T^\top Q_T^i x_T$, where $Q_t^i \in \mathbb{R}^{d\times d}$ and $R_t^i \in \mathbb{R}^{k_i \times k_i}$ $(i = 1, \cdots, N)$ are the state and control cost (or weight) matrices, respectively.

When $N = 1$, the problem given by (1.2.1)-(1.2.3) is a single-agent LQR problem, and when $N \geq 2$, this formulation leads to an N-player general-sum (or non-zero-sum when $N = 2$) LQG. In a general-sum LQG, each player can have a different cost function which depends on their own parameters Q_t^i and R_t^i; a different but also well-studied setting is the zero-sum LQG, where $N = 2$ and instead of (1.2.3), players have the same objective function, that is, $c_t^i(x_t, u_t^i, u_t^j) = c_t^j(x_t, u_t^i, u_t^j) = x_t^\top Q_t^i x_t + (u_t^i)^\top R_t^i u_t^i - (u_t^j)^\top R_t^j u_t^j$ $(i \neq j)$. Player i wants to minimise the objective function while player j wants to maximise it.

LQ problems are important for a number of reasons:

1. there exists a closed-form analytical representation of the optimal feedback control (for each player) [15, 13] ;

2. they can be used to approximate other systems which are hard to analyze [82, 86];

3. they have been used in a wide variety of applications, including for example portfolio optimization [3] and optimal liquidation [7] in finance, resource allocation in energy markets [67, 89], and biological movement systems [52].

In this thesis, we will use the fact that the LQ framework is a relatively simple setting and analyze the behavior of reinforcement learning algorithms rigorously within this framework.

1.3 Reinforcement Learning and Policy Gradient Methods

The mathematical approach to many decision making problems in uncertain environments has traditionally been through modelling with stochastic processes and using techniques arising from stochastic control. The choice of models is often dictated by the need to balance tractability with applicability. Simple models lead to tractable

and implementable strategies in closed-form or that can be found through traditional numerical methods. However, these models sometimes oversimplify the mechanisms, which may result in strategies that are sub-optimal in practice. On the other hand, models that try to capture realistic features of the environments are much more complex and are often mathematically and computationally intractable using the classical tools of stochastic optimal control.

In recent years the availability of large amounts of data has revolutionized data processing and statistical modeling techniques in a wide range of industries and brought new theoretical and computational challenges. In contrast to the classical stochastic control approach, new ideas coming from reinforcement learning (RL) are being developed to make use of all this information. Reinforcement learning describes methods by which agents acting within some system might learn to make optimal decisions through repeated experience gained by interacting with the system. In practice there have been a number of recent successes in applying RL algorithms to improve decision-making in areas such as finance [61, 42, 78, 33], robotics and autonomous control [49, 37], and natural language processing [88, 54]. We refer to the survey [59] for a more detailed introduction to applications of RL in different fields.

1.3.1 Set-up

Markov decision processes (MDPs) provide a general framework for many reinforcement learning ideas. Since this thesis is concerned with finite horizon problems, we introduce MDPs with finite time horizon. Among many real-world decision-making problems with a finite horizon, two examples of particular interest in this thesis are negotiation rounds in bargaining, and optimal execution of the purchase or liquidation of a quantity of asset over a short time horizon. In these examples there may be a penalty at the terminal time, for instance if players do not close the deal in a negotiation, or the targeted amount is not completely executed by the deadline. We also mention the infinite time horizon setting, as there is a rich and vast RL literature dealing with MDPs with an infinite time horizon [71, 30].

Finite Horizon. We have a Markov process taking values in a state space $\mathcal{S}$, where we can influence its evolution by taking an action from a set $\mathcal{A}$. The aim is to optimise the expected total reward from $t = 0$ to time T by choosing a policy, that is a sequence of actions through time. We formulate this mathematically by defining the value

function $V_t^\star$ at time $t = 0, 1, \ldots, T$ to be

$$V_t^\star(s) = \sup_\Pi V_t^\Pi(s) := \sup_\Pi \mathbb{E}^\Pi \left[\sum_{u=t}^{T-1} r_u(s_u, a_u) + r_T(s_T) \middle| s_t = s \right], \quad \forall s \in \mathcal{S}, \qquad (1.3.1)$$

subject to

$$s_{u+1} \sim P_u(s_u, a_u), \quad a_u \sim \pi_u(s_u), \quad t \leq u \leq T-1.$$

Here $\mathbb{E}^\Pi$ denotes the expectation under the policy Π, where the probability measure $\mathbb{P}$ describes both the dynamics and the rewards in the MDP. We will write $\mathcal{P}(\mathcal{X})$ for the probability measures over a space $\mathcal{X}$. The state space $(\mathcal{S}, d_\mathcal{S})$ and the action space $(\mathcal{A}, d_\mathcal{A})$ are both complete separable metric spaces; $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ are the state and the action at time t; $P_u : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ denotes the transition function and $r_u(s, a)$ a real-valued random variable for each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ for $t \leq u \leq T-1$. Similarly $r_T(s)$, the terminal reward, is a real-valued random variable for all $s \in \mathcal{S}$. The notation $s_{t+1} \sim P_t(s_t, a_t)$ denotes that s_{t+1} is sampled from the distribution $P_t(s_t, a_t) \in \mathcal{P}(\mathcal{S})$. It can be shown that when rewards and transition probabilities are Markovian, the optimal policy $\Pi = \{\pi_t\}_{t=0}^\infty$ is also Markovian [70], in that it just depends on the current state, and can be either deterministic or randomized. Therefore hereafter we focus on the class of Markovian policies. For a deterministic policy, $\pi_t : \mathcal{S} \to \mathcal{A}$ maps the current state s_t to a deterministic action. On the other hand, a randomized policy $\pi_t : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps the current state s_t to a distribution over the action space. For a randomized policy π_t, we will use $\pi_t(s) \in \mathcal{P}(\mathcal{A})$ and $\pi_t(s, a)$ to represent, respectively, the distribution over the action space given state s and the probability of taking action a at state s. We also note that the set up is essentially the same if we consider the problem of minimizing costs rather than maximizing a reward.

The well-known Dynamic Programming Principle (DPP), that the optimal policy can be obtained by maximizing the reward from one step and then proceeding optimally from the new state, can be used to derive the following Bellman equation for the value function (1.3.1)

$$V_t^\star(s) = \sup_{a \in \mathcal{A}} \Big\{ \mathbb{E}[r_t(s, a)] + \mathbb{E}_{s' \sim P_t(s,a)}[V_{t+1}^\star(s')] \Big\},$$

with the terminal condition $V_T^\star(s) = \mathbb{E}[r_T(s)]$. We write the value function as

$$V_t^\star(s) = \sup_{a \in \mathcal{A}} Q_t^\star(s, a),$$

where the function $Q_t^\star$ (usually called the Q-function) is defined to be

$$Q_t^\star(s, a) = \mathbb{E}[r_t(s, a)] + \mathbb{E}_{s' \sim P_t(s,a)}[V_t^\star(s')],$$

the expected reward from taking action a at state s and then following the optimal policy thereafter. There is also a Bellman equation for the Q-function given by

$$Q_t^\star(s, a) = \mathbb{E}[r_t(s, a)] + \mathbb{E}_{s' \sim P_t(s,a)} \left[\sup_{a' \in \mathcal{A}} Q_{t+1}^\star(s', a') \right],$$

with the terminal condition $Q_T^\star(s, a) = \mathbb{E}[r_T(s)]$ for all $a \in \mathcal{A}$. If we can find the Q-function, we can then retrieve the optimal policy $\pi_t^*(s, a)$ (if it exists) from $Q_t^*(s, a)$, in that $\pi_t^*(s, a) = 1$ if $a = \arg\max_{a \in \mathcal{A}} Q_t^*(s, a)$ and 0 otherwise.

Infinite Horizon and Discounted Reward. MDPs with infinite horizon ($T = \infty$) are widely studied in the literature. In this setting, we aim to optimise the expected discounted reward and define the corresponding value function as

$$V^\star(s) = \sup_{\Pi} V^{\Pi}(s) := \sup_{\Pi} \mathbb{E}^{\Pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right],$$

subject to

$$s_{t+1} \sim P(s_t, a_t), \quad a_t \sim \pi_t(s_t),$$

where $\gamma \in (0, 1)$ is a discounting factor. Where this differs from the finite horizon case is that we no longer have a terminal reward, and the reward r and the transition dynamics P are often assumed to be time homogeneous in the literature [71].

The corresponding Bellman equation for the value function (1.3.2) is defined as

$$V^\star(s) = \sup_{a \in \mathcal{A}} \Big\{ \mathbb{E}[r(s, a)] + \gamma \, \mathbb{E}_{s' \sim P(s,a)}[V^\star(s')] \Big\}.$$

We write the value function as

$$V^\star(s) = \sup_{a \in \mathcal{A}} Q^\star(s,a),$$

where the Q-function is defined to be

$$Q^\star(s,a) = \mathbb{E}[r(s,a)] + \gamma\, \mathbb{E}_{s' \sim P(s,a)}[V^\star(s')].$$

There is also a Bellman equation for the Q-function given by

$$Q^\star(s,a) = \mathbb{E}[r(s,a)] + \gamma\, \mathbb{E}_{s' \sim P(s,a)} \sup_{a' \in \mathcal{A}} Q^\star(s',a').$$

Although in the literature some of the analysis for finite horizon MDPs was motivated by infinite horizon MDPs, our focus in this thesis is on the finite time horizon setting and we will not discuss the infinite horizon setting further.

1.3.2 Classification of RL Algorithms

We will explore RL methods to solve MDPs, which are implemented through RL algorithms. An RL algorithm includes one or more of the following components:

- a representation of a value function that provides a prediction of how good each state or each state/action pair is,
- a direct representation of the policy $\pi_t(s)$ or $\pi_t(s,a)$ $(t = 0, 1, \ldots, T-1)$,
- a model of the environment (the estimated transition function and the estimated reward function) in conjunction with a planning algorithm (any computational process that uses a model to create or improve a policy).

The first two components are related to what is called model-free RL. When the latter component is used, the algorithm is referred to as model-based RL.

In the MDP setting, model-based algorithms maintain an approximate MDP model by estimating the transition probabilities and the reward function, and derive a value function from the approximate MDP. The policy is then derived from the value function. Examples include [17, 48] and [81].

Unlike the model-based method, model-free algorithms *directly* learn a value (or state-value) function or the optimal policy without inferring the model. Model-free

algorithms can be further divided into two categories, value-based methods and policy-based methods. Policy-based methods explicitly build a representation of a policy and keep it in memory during learning. Examples of policy-based methods include REINFORCE [90], Trust Region Policy Optimization (TRPO) [74], Proximal Policy Optimization (PPO) [76] among others. As an alternative, value-based methods store only a value function without an explicit policy during the learning process. Examples include Q-learning [26] and SARSA [96]. In this case, the policy is implicit and can be derived directly from the value function (by picking the action with the best value).

Policy Gradient Methods. In this thesis our focus is on policy gradient methods. We first parametrize the policy π_t at each time t by a vector θ_t, thus we define $\pi_t(s, \cdot; \theta_t) \in \mathcal{P}(\mathcal{A})$, the probability distribution parameterized by θ_t over the action space given the state s at time t. Here we write $\Theta = (\theta_0, \theta_1, \ldots, \theta_{T-1})$ for the policy parameterisation over the whole time set, and use $\Pi(s, a; \Theta) = \{\pi_t(s, a; \theta_t)\}_{t=0}^{T-1}$ and Π_Θ interchangeably to represent the policy parameterized by Θ. We write μ for the initial distribution of the state. Then the policy objective function (to be maximised) is defined to be

$$J(\Theta) := \int_{\mathcal{S}} V_0^{\Pi_\Theta}(s)\mu(ds)$$

for RL with finite time horizon. In order to maximize this function, the policy parameter Θ is updated using the gradient ascent rule: for $t = 0, 1, \ldots, T-1$

$$\theta_t' = \theta_t + \beta\, \widehat{\nabla_{\theta_t} J(\Theta)}, \tag{1.3.2}$$

where β is the learning rate, and $\widehat{\nabla_{\theta_t} J(\Theta)}$ is an estimate of the gradient of $J(\Theta)$ with respect to θ_t. There are various ways to obtain the estimate $\widehat{\nabla_{\theta_t} J(\Theta)}$ such as Monte Carlo methods (e.g. REINFORCE [90]) and deep neural networks (e.g. DDPG [53]).

There have been some convergence results for policy gradient methods and their variants for MDPs (see for example [80, 66, 22]), LQ problems with an infinite time horizon (see for example [29, 93]), and (two-player) zero-sum LQGs with a finite horizon [94]. In Chapter 2 and 3 we will provide convergence results for policy gradient methods for the LQR and the N-player general-sum LQG with a finite horizon.

1.4 Outline

In this thesis, we provide the theoretical guarantees for global convergence of policy gradient methods for both the single-agent and the multi-agent LQ problems. We also model a classical bargaining problem as an LQG and solve the model using policy gradient methods.

In Chapter 2, we focus on single-agent LQR with stochastic dynamics over a finite horizon. We start with the policy gradient method with known model parameters and prove the global linear convergence of such an algorithm. In the setting of unknown model parameters, the policy gradient is estimated from samples using a derivative-free method, and the global linear convergence of this sample-based policy gradient method is also proved. In addition, the sublinear convergence of a projected policy gradient method is established in order to handle problems with constraints.

We illustrate the performance of the algorithm with two examples. The first example is the optimal liquidation of a holding in an asset. We show results for the case where we assume a model for the underlying dynamics and where we apply the method to the data directly. The empirical evidence suggests that the policy gradient method can learn the global optimal solution for a larger class of stochastic systems including the linear quadratic framework, and that it is more robust with respect to model misspecification when compared to a model-based approach. The second example is an LQR system in a higher dimensional setting with synthetic data.

In Chapter 3, we focus on a class of N-player general-sum LQGs and explore the convergence of the natural policy gradient method in this setting. This method can be viewed as a normalized version of the vanilla policy gradient descent method. We provide a global linear convergence guarantee for this approach in the setting of a finite time horizon and stochastic dynamics provided there is a certain level of noise in the system. In the presence of noise the agents can escape from traps such as those observed in [57] when the dynamics are deterministic. To the best of our knowledge, this is the first result of its kind in the multi-agent RL literature showing a provable convergence result for N-player general-sum LQGs.

To illustrate the performance of our algorithm, we first perform the natural policy gradient algorithm under the experimental set-up in [57] over a time finite horizon. The algorithm converges to the (unique) Nash equilibrium for appropriate initial policies and step sizes. The second example is a toy LQG example with synthetic data. We show that even in situations where the policy gradient algorithm may not converge in the deterministic setting, the addition of noise helps the algorithm to converge.

In Chapter 4, we model a two-player bargaining problem using LQGs, in which a buyer and a seller negotiate over the price of a project. At each round they exchange their offers (prices they propose) in order to agree on the value of the project. We propose simple cost functions capturing their risk appetites and reservation or target prices. We first consider an LQG model with partial information where players know the model parameters but they do not have access to the real state. They use the Kalman filter to generate state estimates. We derive the Dynamic Programming Principle and the Nash equilibrium in this case, where each player also estimates their opponent's state estimates based on their own information. We then study a different setting when model parameters are unknown to the players. We construct another LQG model using the framework discussed in Chapter 3 and apply the natural policy gradient method with known and unknown parameters to learn the optimal strategies.

1.5 Notation

For any matrix $Z = (Z_1, \cdots, Z_d) \in \mathbb{R}^{m\times d}$ with $Z_j \in \mathbb{R}^m$ $(j = 1, 2, \cdots, d)$, we denote by $Z^\top \in \mathbb{R}^{d\times m}$ the transpose of Z; $\sigma_{\min}(Z)$ and $\sigma_{\max}(Z)$ denote the minimal and maximal singular value of a matrix Z; $\|Z\|$ denotes the spectral norm of a matrix Z, which is defined as $\|Z\| := \sigma_{\max}(Z)$; $\mathrm{Tr}(Z)$ denotes the trace of a square matrix Z; $\|Z\|_F$ denotes the Frobenius norm of a matrix Z, defined as $\|Z\|_F := \sqrt{\mathrm{Tr}(Z^\top Z)}$; and $\mathrm{vec}(Z) = (Z_1^\top, \cdots, Z_d^\top)^\top$ denote the vectorized version of a matrix Z. For a sequence of matrices $\boldsymbol{D} = (D_0, \cdots, D_T)$, we define a new norm $|||\boldsymbol{D}|||$ as $|||\boldsymbol{D}||| = \sum_{t=0}^{T} \|D_t\|$, where $D_t \in \mathbb{R}^{m\times d}$ and $\|\cdot\|$ represents the spectral norm. Further we use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.

Chapter 2

Policy Gradient Methods for the Linear Quadratic Regulator

2.1 Introduction

In this chapter we consider the single-agent LQR problem. Until recently much of the work on the LQR problem has focused on solving for the optimal controls under the assumption that the model parameters are *fully known*. See the book of Anderson and Moore [9] for an introduction to the LQR problem with known parameters. However, assuming that the controller has access to all the model parameters is not realistic for many applications, and this has lead to the exploration of RL approaches to the problem.

Model-based and model-free RL approaches have both been applied to handle the LQR with unknown parameters. In the paradigm of the model-based approach, the controller estimates the unknown model parameters and then constructs a control policy based on the estimated parameters. The classical approach is the *certainty equivalence principle* [10]: the unknown parameters are estimated using observations (or samples), and a control policy is then designed by treating the estimated parameters as the truth. In the first step, the unknown model parameters can be estimated by standard statistical methods such as least-square minimization [25]. The second step is to show that when the estimated parameters are accurate enough, the policy using the "plug-in" estimates enjoys good theoretical guarantees of being close to optimal. See [25] and [31] for the optimal gap and sample complexities along this line and see [28] for the sample complexity with distributed robust learning. Another line of work in

the model-based regime focuses on *uncertainty* quantification. The controller updates their posterior belief or the confidence bounds on the unknown model parameters and then makes decisions in an online manner, see [1, 2, 27, 44, 64].

In model-free approaches, the controller learns the optimal policy directly via interacting with the system. As the optimal policy in the LQR problem is a linear function of the state, the aim is to determine this linear function. This is equivalent to learning a set of parameters in matrix form, called the policy matrix. Policy gradient method is a natural way to achieve this goal by applying the gradient descent method in the parameter space of the policy matrix. The paper [29] was the first to show that policy gradients converge to the global optimal solution with polynomial (in the relevant quantities) sample complexity. However, [29] focuses on the case where the only noise in the system is in the initial state, and the rest of the state transitions are deterministic. Other model-free methods for LQR include the Actor-Critic method [91] and least-squares temporal difference learning [83].

If the true system is indeed LQ, the model-based approaches (may) outperform the model-free approaches by fully utilizing the LQ structure. For example in the setting that the system transition matrices are unknown and the parameters in the cost function are known, [72] and [84] showed that model-based methods are (asymptotically) more sample-efficient than some popular model-free methods. However, we are often uncertain about whether the actual system is LQ in the learning setting; for instance there might be some small nonlinear terms in the system dynamics. Therefore, compared to the model-based approach, which strongly relies on the assumption that the stochastic system lies within the LQR framework and may, in practice, suffer from model mis-specification, the execution of the model-free algorithm does not rely on the assumptions of the model. It has been shown that the policy gradient method can learn the global optimal solution, not only for the LQR framework, but also for a more general class of deterministic systems in the setting of an infinite time horizon [16]. Thus the advantage of the model-free approach is that it is more robust against model mis-specification compared to the model-based approach.

Our Contributions. We now summarize our contributions. Motivated by many real-word decision-making problems with a fixed deadline and uncertainty in the underlying dynamics, such as the optimal liquidation problem that we discuss in Section 2, we extend the framework of [29] by incorporating a finite time horizon and sub-Gaussian noise (which includes Gaussian noise as a special case). In particular, we provide a global linear convergence guarantee and a polynomial sample complexity

guarantee for the policy gradient method in this setting with both known parameters (Theorem 2.3.3) and unknown parameters (Theorem 2.4.4). The analysis with known parameters paves the way for learning LQR with unknown parameters. In addition, numerically solving the Riccati equation with known parameters in high dimensions may suffer from computational inaccuracy. The policy gradient method provides a direct way of searching for the optimal solution with known parameters in this case, which may be of separate interest. Note that the optimal policy is time-invariant for the LQR with infinite time horizon, whereas the optimal policy is time-dependent with finite time horizon and hence harder to learn in general. With noise in the dynamics, we need more careful choices of the hyper-parameters to retrieve compatible sample complexities with noisy observations. In addition, when optimal polices need to satisfy certain constraints, we provide a global convergence result for the projected policy gradient method in Theorem 2.4.5. This is required in the context of our application to the optimal liquidation problem.

We will formulate the optimal liquidation problem over a fixed horizon as a noisy LQR problem which is essentially the classical Almgren-Chriss formulation [7]. The performance of the algorithm on NASDAQ ITCH data is assessed. As well as using the method within this modelling approach, we also consider the performance of the policy gradient method when applied directly to the data with an appropriate cost function. This improves the performance of the LQR/Almgren-Chriss solution and shows promising results for the use of the policy gradient method for problems that are 'close' to the LQR framework.

2.1.1 Related Work

Policy Gradient Methods for LQR Problems. Since the policy gradient method is our main focus, here we provide a review of the previous theoretical work on this method in various LQR settings and extensions. The first global convergence result for the policy gradient method to learn the optimal policy for LQR problems was developed in [29] in the setting of infinite horizon and deterministic dynamics. The work of [29] was extended in [16] to give global optimality guarantees of policy gradient methods for a larger class of control problem that includes the LQ case. In particular, this class of control problem satisfies a closure condition under policy improvement and convexity of policy improvement steps. The paper [18] considers policy gradient methods for LQR problems in terms of optimizing a real valued matrix function over the set of feedback gains. The extension of the policy gradient method to continuous-time can

be found in [19]. All of these methods are in the infinite horizon setting and without the addition of noise in the dynamics.

There has been some work on the case of noisy dynamics, but all in the setting of infinite horizon. In [35] the problem with a multiplicative noise was discussed, using a relatively straightforward extension of the deterministic dynamics considered in the original framework. In the case of additive noise [46] studies the global convergence of policy gradient and other learning algorithms for the LQR over an infinite time-horizon and with Gaussian noise. In particular, the policy considered in [46] is a randomized policy with Gaussian distribution. There is also [55] which studies derivative-free (zeroth-order) policy optimization methods for the LQR with bounded additive noise. Finally some other contributions can be found in [20, 93] for zero-sum LQR games and [21, 38] for mean-field LQR games.

Compared to [29], our technical difficulties are three-fold. First due to the time-dependent nature of the admissible policies over a finite-horizon and randomness from the system noise, we need additional conditions and analysis to guarantee the well-definedness of the state process, i.e., the non-degeneracy of the controlled $\mathbb{E}[x_t x_t^\top]$ matrices. This holds almost for free in the infinite horizon case with deterministic dynamics. Second, we need to take care of the additional randomness from the sub-Gaussian noise when developing the perturbation analysis and the gradient dominant condition. Third, we need more advanced concentration inequalities and tighter upper bounds to provide compatible sample complexity analysis in the unknown parameter case. See the more detailed discussion in Remark 2.4.12.

Optimal Liquidation. An early mathematical framework for the optimal liquidation problem is due to Almgren and Chriss [7]. In this problem a trader is required to liquidate a portfolio of shares over a fixed horizon. The selling of a large number of shares at once has both temporary and permanent impacts on the share price causing it to decrease. The trader therefore wishes to find a trading strategy which maximizes their return from, or alternatively, minimizes the cost of, the liquidation of the portfolio subject to a given level of risk.

This problem has been considered in many papers and extended in many directions. See for instance [5], [6] and [34]. We will cast this as an LQR problem and show how the policy gradient method is a powerful tool for solving this problem even without assumptions on the model.

More recently techniques from RL have been applied to the optimal liquidation problem. The first paper to do this was [61] where the authors showed promising results for this approach by designing a Q-learning based algorithm to optimally select price levels and passively place limit orders. This was further developed in [42] which designed a Q-learning based algorithm for liquidation within the standard Almgren-Chriss framework. For recent work incorporating deep learning see for example [12], [51], [62], and [95]. See [23] for a detailed review on RL with applications in finance and economics, and the references therein. However, all these works focus on the model-free setting without taking advantage of even weak modelling assumptions on the market dynamics. In addition, the performances of these proposed algorithms are validated only through empirical studies and no theoretical guarantee of convergence is provided.

The rest of this chapter is organized as follows. We introduce the mathematical framework and problem set-up in Section 2.2. The first step in our convergence analysis of the policy gradient method is to consider the case of known model parameters in Section 2.3. When parameters are unknown, the convergence results for the sample-based policy gradient method and projected policy gradient method are obtained in Section 2.4. Finally, the algorithm is applied to liquidation problem. See Sections 2.2.1 and 2.5 for the corresponding set-up and algorithm performance, respectively.

2.2 Problem Set-up

We consider the following LQR problem over a finite time horizon T,

$$\min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right], \tag{2.2.1}$$

such that for $t = 0, 1, \cdots, T-1$,

$$x_{t+1} = Ax_t + Bu_t + w_t,\ x_0 \sim \mathcal{D}. \tag{2.2.2}$$

Here $x_t \in \mathbb{R}^d$ is the state of the system with the initial state x_0 drawn from a distribution $\mathcal{D}$, $u_t \in \mathbb{R}^k$ is the control at time t which is assumed to be Markovian and $\{w_t\}_{t=0}^{T-1}$ are zero-mean i.i.d. noises which are independent from x_0. At this moment, we only assume x_0 and $\{w_t\}_{t=0}^{T-1}$ have finite second moments. That is, $\mathbb{E}[x_0 x_0^\top]$ and $W := \mathbb{E}[w_t w_t^\top]$ $(\forall\, t = 0, 1, \cdots, T-1)$ exist. The system parameters $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$

are referred to as system (transition) matrices; $Q_t \in \mathbb{R}^{d\times d}$ $(\forall\, t = 0, 1, \cdots, T)$ and $R_t \in \mathbb{R}^{k\times k}$ $(\forall t = 0, 1, \cdots, T-1)$ are matrices that parameterize the quadratic costs. Note that the expectation in (2.2.1) is taken with respect to both $x_0 \sim \mathcal{D}$ and w_t $(t = 0, 1, \cdots, T-1)$. We further denote by $\boldsymbol{u} := (u_0, \cdots, u_{T-1})$, $\boldsymbol{x} := (x_0, \cdots, x_T)$, $\boldsymbol{w} := (w_0, \cdots, w_{T-1})$, $\boldsymbol{Q} := (Q_0, \cdots, Q_T)$, and $\boldsymbol{R} := (R_0, \cdots, R_{T-1})$, the profile over the decision period T. We note that the more general LQR problem may also contain a drift term in the state equation and a cross term ($x_t^\top S_t u_t$ for some $S_t \in \mathbb{R}^{d\times k}$) in the cost function. In this thesis, we will focus on the LQR problem given in (2.2.1)-(2.2.2) for simplicity.

To solve the LQR problem (2.2.1)-(2.2.2), let us start with some conditions on the model parameters to assure the well-definedness of the problem.

Assumption 2.2.1 (Cost Parameter)**.** *Assume $Q_t \in \mathbb{R}^{d\times d}$, for $t = 0, 1, \cdots, T$, and $R_t \in \mathbb{R}^{k\times k}$, for $t = 0, 1, \cdots, T-1$, are positive definite matrices.*

We note that throughout the thesis, by saying that a matrix $M \in \mathbb{R}^{n\times n}$ is positive definite, we mean that M is symmetric and *strictly* positive definite, that is, for all non-zero $\boldsymbol{x}$ in $\mathbb{R}^n$, we have $\boldsymbol{x}^\top M \boldsymbol{x} > 0$. By saying that a matrix $M \in \mathbb{R}^{n\times n}$ is positive semi-definite, we mean that M is symmetric and for all non-zero $\boldsymbol{x}$ in $\mathbb{R}^n$, we have $\boldsymbol{x}^\top M \boldsymbol{x} \geq 0$.

Under Assumption 2.2.1, we can properly define a sequence of matrices $\{P_t^*\}_{t=0}^T$ as the solution to the following dynamic Riccati equation [15]:

$$P_t^* = Q_t + A^\top P_{t+1}^* A - A^\top P_{t+1}^* B \left(B^\top P_{t+1}^* B + R_t\right)^{-1} B^\top P_{t+1}^* A, \tag{2.2.3}$$

with terminal condition $P_T^* = Q_T$. The matrices $\{P_t^*\}_{t=0}^T$ can be found by solving the Riccati equations iteratively backwards in time. In particular with a slight modification of the initial state distribution in [15, Chapter 4.1], we have the following result.

Lemma 2.2.2 (Well-definedness and the Optimal Solution [15])**.** *Under Assumption 2.2.1,*

1. *The solution P_t^* to the Riccati equation* (2.2.3) *is positive definite, $\forall\, t = 0, 1, \cdots, T$;*
2. *Then the optimal control sequence $\{u_t\}_{t=0}^{T-1}$ is given by*

$$\begin{aligned} u_t &= -K_t^* x_t, \qquad \text{where} && (2.2.4)\\ K_t^* &= \left(B^\top P_{t+1}^* B + R_t\right)^{-1} B^\top P_{t+1}^* A. && (2.2.5) \end{aligned}$$

To find the optimal solution in the linear feedback form (2.2.4), we only need to focus on the following class of linear deterministic *admissible policies* in feedback form

$$u_t = -K_t x_t, \qquad t = 0, 1, \cdots, T-1, \tag{2.2.6}$$

which can be fully characterized by $\boldsymbol{K} := (K_0, \cdots, K_{T-1})$ where each $K_t \in \mathbb{R}^{k \times d}$ is a matrix.

2.2.1 Application: The Optimal Liquidation Problem

One application of the LQR framework (2.2.1)-(2.2.2) is the optimal liquidation problem. We give a slight variant of the setup of Almgren-Chriss [7]. Our aim is to liquidate an amount q_0 of an asset, with price S_0 at time 0, over the time period $[0, T]$ with trading decisions made at discrete time points $t = 0, 1, \ldots, T-1$. At each time t our decision is to liquidate an amount u_t of the asset. Any residual holding is then liquidated at time T. This will have two types of price impact. There will be a temporary price impact, caused when the order 'walks the book' and a permanent price impact as traders rearrange their positions in the light of the sell order. We will assume the impacts are linear in the number of traded shares.

We write S_t for the asset price at time t. This evolves according to a Bachelier model with a linear permanent price impact in that

$$S_{t+1} = S_t + \sigma Z_{t+1} - \gamma u_t,$$

where, for each $t = 1, \ldots, T$, Z_t is an independent standard normal random variable, σ is the volatility and γ is the permanent price impact parameter. The inventory process q_t records the current holding in the asset at time t. Thus we have

$$q_{t+1} = q_t - u_t.$$

Therefore, the two-dimensional state process is

$$\begin{pmatrix} S_{t+1} \\ q_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_t \\ q_t \end{pmatrix} + \begin{pmatrix} -\gamma \\ -1 \end{pmatrix} u_t + \begin{pmatrix} \sigma Z_{t+1} \\ 0 \end{pmatrix}. \tag{2.2.7}$$

When selling shares we incur a temporary price impact, parameter β, in that if,

at time t, we trade u_t of our asset then we obtain $\tilde{S}_t = S_t - \beta u_t$ per share. At the deadline, any residual holding q_T will be sold and we obtain $\tilde{S}_T = S_T - \beta q_T$ per share. Therefore the total revenue is $\sum_{t=0}^{T-1} u_t \tilde{S}_t + q_T \tilde{S}_T$, and C, the total cost of execution over $[0, T]$, is the book value at time 0 minus the revenue:

$$C = q_0 S_0 - \sum_{t=0}^{T-1} u_t \tilde{S}_t - q_T \tilde{S}_T.$$

In a similar way to [7], after summation by parts, we have

$$C = -\sigma \sum_{t=1}^{T} q_t Z_t - \frac{\gamma}{2} \sum_{t=0}^{T-1} u_t^2 + \frac{\gamma}{2} \left(q_0^2 - q_T^2 \right) + \beta \sum_{t=0}^{T-1} u_t^2 + \beta q_T^2.$$

The mean and variance of the total cost of execution are given by

$$\mathbb{E}(C) = \sum_{t=0}^{T-1} \delta u_t^2 + \delta q_T^2 + \frac{\gamma}{2} q_0^2, \quad \text{var}(C) = \sum_{t=1}^{T} \sigma^2 q_t^2,$$

where $\delta = \beta - \gamma/2$ summarizes the impact and is assumed positive.

Following Almgren-Chriss [7], we minimize the following cost function

$$C_{\text{AC}} = \min \left(\mathbb{E}(C) + \phi \, \text{var}(C) \right), \tag{2.2.8}$$

where ϕ is a parameter balancing risk versus return. For our LQR framework we take the cost function to be

$$\begin{aligned} C_{\text{LQR}}(\epsilon) &= \min \left(\mathbb{E}(C) + \phi \, \text{var}(C) + \epsilon \sum_{t=0}^{T} \mathbb{E}[S_t^2] \right) \\ &= \min \mathbb{E} \left[\sum_{t=0}^{T-1} \delta u_t^2 + \delta q_T^2 + \frac{\gamma}{2} q_0^2 + \phi \sum_{t=1}^{T} \sigma^2 q_t^2 + \epsilon \sum_{t=0}^{T} S_t^2 \right]. \end{aligned} \tag{2.2.9}$$

Note that the term $\epsilon \sum_{t=0}^{T} \mathbb{E}[S_t^2]$, with some small $\epsilon > 0$, serves as a regularization term (with no economic meaning) to guarantee Assumption 2.2.1 holds (more precisely, to guarantee that Q_t and Q_T are positive definite). In practice, we can show that the optimal solution with ϵ small is close to the Almgren-Chriss solution (when $\epsilon = 0$). We show empirically that $C_{\text{LQR}}(\epsilon)$ converges to $C_{\text{LQR}}(0)$ as ϵ tends to zero in Figure 2.4.

One can also check that having the ϵ in the matrices Q_t and Q_T introduces an error of $\mathcal{O}(\epsilon)$ in C_{LQR}. Since all the summations in 2.2.9 are finite and bounded, $C_{\mathrm{LQR}}(\epsilon)$ will converge to $C_{\mathrm{LQR}}(0)$ as ϵ tends to zero provided that u_t has finite second moment. In addition, the algorithm will still converge with $\epsilon = 0$. Thus, in the LQR formulation we have $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = (-\gamma, -1)^\top$, and $w_t = (\sigma Z_{t+1}, 0)^\top$ and the objective function has $Q_T = \begin{pmatrix} \epsilon & 0 \\ 0 & \delta + \phi\sigma^2 \end{pmatrix}$, $Q_t = \begin{pmatrix} \epsilon & 0 \\ 0 & \phi\sigma^2 \end{pmatrix}$ and $R_t = \delta$. It is easy to see that Q_t, for $t = 0, 1, \cdots, T$ and R_t for $t = 0, 1, \cdots, T-1$ are positive definite, hence Assumption 2.2.1 is satisfied.

We will show that the problem is well-defined and can be solved using the methods of this chapter with rigorous convergence guarantees.

2.3 Exact Gradient Methods with Known Parameters

In this section we assume all the parameters in the model, $\{Q_t\}_{t=0}^T$, $\{R_t\}_{t=0}^{T-1}$, A, B, are known. The analysis of exact gradient methods with known parameters paves the way for learning LQR with unknown parameters in Section 2.4. In addition, the policy gradient method provides an alternative way to solve the LQR problem when the parameters are fully known. In this setting the Riccati equation (2.2.3) is just solved backward in time. However this operation involves inverting large matrices when the problem is in high dimensions, which may lead to high computational cost and accumulation of computational errors.

Since an admissible policy can be fully characterized by $\boldsymbol{K}$, the cost of a policy $\boldsymbol{K}$ can be correspondingly defined as

$$C(\boldsymbol{K}) = \mathbb{E}\left[\sum_{t=0}^{T-1} \left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right], \tag{2.3.1}$$

where $\{x_t\}_{t=1}^T$ and $\{u_t\}_{t=0}^{T-1}$ are the dynamics and controls induced by following $\boldsymbol{K}$, starting with $x_0 \sim \mathcal{D}$. Recall that $\boldsymbol{K}^*$ is the optimal policy for the problem, in that

$$\boldsymbol{K}^* = \arg\min_{\boldsymbol{K}} C(\boldsymbol{K}), \tag{2.3.2}$$

subject to the dynamics (2.2.2).

Well-definedness of the State Process. To prove the global convergence of policy gradient methods, the essential idea is to show the *gradient dominance condition*, which states that $C(\boldsymbol{K}) - C(\boldsymbol{K}^*)$ can be bounded by $\|\nabla C(\boldsymbol{K})\|_F$ for any admissible policy $\boldsymbol{K}$. One of the key steps to guarantee this gradient dominance condition is the well-definedness of the second moment of the state. That is, $\mathbb{E}[x_t x_t^\top]$ is positive definite for $t = 0, 1, \cdots, T$. This condition holds almost for free for LQR problems with infinite time horizon and deterministic dynamics. The only condition needed there is the positive definiteness of $\mathbb{E}[x_0 x_0^\top]$ (See [29]). However, some effort needs to be made to ensure that the $\mathbb{E}[x_t x_t^\top]$ matrix is well-defined for LQR problems with finite horizon and stochastic dynamics. We show that this condition holds under moderate conditions.

Assumption 2.3.1 (Initial State and Noise Process)**.** *We assume that*

1. *Initial state:* $x_0 \sim \mathcal{D}$ *such that* $\mathbb{E}[x_0 x_0^\top]$ *is positive definite;*
2. *Noise:* $\{w_t\}_{t=0}^{T-1}$ *are i.i.d. and independent from* x_0 *such that* $\mathbb{E}[w_t] = 0$*, and* $W = \mathbb{E}[w_t w_t^\top]$ *is positive definite,* $\forall t = 0, 1, \cdots, T-1$.

We start by introducing some notation which will be used throughout the analysis. We define the matrix $\Sigma_t^{\boldsymbol{K}}$ as follows

$$\Sigma_t^{\boldsymbol{K}} = \mathbb{E}\left[x_t^{\boldsymbol{K}} (x_t^{\boldsymbol{K}})^\top\right], \ t = 0, 1, \cdots, T, \tag{2.3.3}$$

where $\{x_t^{\boldsymbol{K}}\}_{t=1}^T$ is a state trajectory generated by $\boldsymbol{K}$. Further define a matrix $\Sigma_{\boldsymbol{K}}$ as the sum of $\Sigma_t^{\boldsymbol{K}}$,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T} \Sigma_t^{\boldsymbol{K}} = \mathbb{E}\Big[\sum_{t=0}^{T} x_t^{\boldsymbol{K}} (x_t^{\boldsymbol{K}})^\top\Big]. \tag{2.3.4}$$

We will write $x_t = x_t^{\boldsymbol{K}}$ and $\Sigma_t^{\boldsymbol{K}} = \Sigma_t$ when no confusion may occur. Define $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}}$ to be the lower bound over all the minimum singular values of $\Sigma_t^{\boldsymbol{K}}$:

$$\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} := \min_t \sigma_{\min}(\Sigma_t^{\boldsymbol{K}}),$$

We also define $\underline{\sigma}_{\boldsymbol{X}}$ as

$$\underline{\sigma}_{\boldsymbol{X}} := \min\{\sigma_{\min}(\mathbb{E}[x_0 x_0^\top]), \sigma_{\min}(W)\}. \tag{2.3.5}$$

Then we have the following result and the proof can be found in Appendix A.3.1.

Lemma 2.3.2 (Well-definedness of the Second Moment Matrix)**.** *Under Assumption 2.3.1, we have* $\mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top]$ *is positive definite for* $t = 0, 1, \cdots, T$ *under any control policy* $\boldsymbol{K}$*. Therefore,* $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} \geq \underline{\sigma}_{\boldsymbol{X}} > 0$*.*

Lemma 2.3.2 implies that if the initial state and the noise driving the dynamics are non-degenerate, the second moment of the state dynamics are positive definite for any policy $\boldsymbol{K}$. However, the second moment matrix may be degenerate in many applications, especially when inventory processes are involved. We note that Part 1 of Assumption 2.3.1 can be partially relaxed, for example, $x_0 \in \mathbb{R}^d$ can be deterministic in some dimensions. In this case, some problem-dependent conditions are needed to guarantee that $\underline{\sigma}_{\boldsymbol{X}} > 0$ holds. See more discussion on the condition $\underline{\sigma}_{\boldsymbol{X}} > 0$ for the liquidation problem in Section 2.5.1. In the light of this we will assume $\underline{\sigma}_{\boldsymbol{X}} > 0$ in the analysis of the convergence of the algorithm in Sections 2.3 and 2.4.

Similarly, we define $\underline{\sigma}_{\boldsymbol{R}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ to be the smallest values of all the minimum singular values of $\boldsymbol{R}$ and $\boldsymbol{Q}$:

$$\underline{\sigma}_{\boldsymbol{R}} = \min_t \sigma_{\min}(R_t), \tag{2.3.6}$$

$$\underline{\sigma}_{\boldsymbol{Q}} = \min_t \sigma_{\min}(Q_t). \tag{2.3.7}$$

Under Assumption 2.2.1, we have $\underline{\sigma}_{\boldsymbol{R}} > 0$ and $\underline{\sigma}_{\boldsymbol{Q}} > 0$.

We denote the set of model parameters of the LQR as Θ:=$\{d,\ k,\ \frac{1}{\|A\|},\ \frac{1}{\|A\|+1},\ \|A\|,\ \frac{1}{\|B\|},\ \frac{1}{\|B\|+1},\ \|B\|,\ \frac{1}{\|\boldsymbol{R}\|},\ \frac{1}{\|\boldsymbol{R}\|+1},\ \|\boldsymbol{R}\|,\ \frac{1}{\|W\|},\ \frac{1}{\|W\|+1},\ \|W\|,\ \frac{1}{\underline{\sigma}_{\boldsymbol{Q}}},\ \frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1},\ \underline{\sigma}_{\boldsymbol{Q}},\ \frac{1}{\underline{\sigma}_{\boldsymbol{R}}},\ \frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1},\ \underline{\sigma}_{\boldsymbol{R}},\ \frac{1}{\underline{\sigma}_{\boldsymbol{X}}},\ \frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1},\ \underline{\sigma}_{\boldsymbol{X}},\ \|\boldsymbol{Q}\|,\ \mathbb{E}[x_0 x_0^\top]$, and $\frac{1}{\mathbb{E}[x_0 x_0^\top]}\}$. We write $\mathcal{H} = \{h \,|\, h$ are multivariate polynomials where the variables are in the set $\Theta\}$ and $\mathcal{H}(.)$ when there are other variables not only Θ.

Exact Gradient Descent. We consider the following *exact* gradient descent updating rule to find the optimal solution (2.3.2),

$$K_t^{n+1} = K_t^n - \eta \nabla_t C(\boldsymbol{K}^n),\ \forall\, 0 \leq t \leq T-1, \tag{2.3.8}$$

where n ($n \leq N$) is the number of iterations, N is the total number of iterations, $\nabla_t C(\boldsymbol{K}) = \frac{\partial C(\boldsymbol{K})}{\partial K_t}$ is the gradient of $C(\boldsymbol{K})$ with respect to K_t, and η is the step size. We further denote $\nabla C(\boldsymbol{K}) = (\nabla_0 C(\boldsymbol{K}), \cdots, \nabla_{T-1} C(\boldsymbol{K}))$.

Then, the main result for this setting is the following.

Theorem 2.3.3 (Global Convergence of Gradient Methods)**.** *Assume Assumption 2.2.1 holds. Further assume* $\underline{\sigma}_{\boldsymbol{X}} > 0$ *and* $C(\boldsymbol{K}^0)$ *is finite. Then, for an appropriate (constant) setting of the stepsize* $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$*, and for* $\epsilon > 0$*, if the total number of iterations* N *satisfies*

$$N \geq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{2\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{\epsilon},$$

the exact gradient descent method (2.3.8) *enjoys the following performance bound:*

$$C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) \leq \epsilon,$$

which is equivalent to $C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) = \mathcal{O}(e^{-N})$.

We note that the convergence of the cost function in Theorem 2.3.3 is equivalent to the convergence of the learned policy $\boldsymbol{K}$ to $\boldsymbol{K}^*$. To see this we can plug in $\boldsymbol{K}' = \boldsymbol{K}^N$ and $\boldsymbol{K} = \boldsymbol{K}^*$ in Lemma 2.3.7, which implies that the $C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*)$ can be rewritten as a function of $\boldsymbol{K}^N - \boldsymbol{K}^*$. Similarly one can check as $C(\boldsymbol{K}^N)$ gets close to $C(\boldsymbol{K})$ the policy gradient tends to zero, which is a consequence of Lemma 2.3.6.

We also note that $\Sigma_{\boldsymbol{K}^*}$ and $C(\boldsymbol{K}^*)$ are only used to analyze the theoretical guarantees for the convergence. In practice, we may conclude that the approximate solution is close enough to the optimal solution when the gradient term becomes very small. Therefore for practical evaluation, our stopping criterion does not necessarily need to rely on $\Sigma_{\boldsymbol{K}^*}$ and $C(\boldsymbol{K}^*)$.

The proof of Theorem 2.3.3 relies on the regularity of the LQR problem, some properties of the gradient descent dynamics, and the perturbation analysis of the second moment matrix of the controlled dynamics.

2.3.1 Regularity of the LQR Problem and Properties of the Gradient Descent Dynamics

Let us start with the analysis of some properties of the LQR problem (2.2.1)-(2.2.2). To start, Proposition 2.3.4 focuses on the well-definedness of the Ricatti system $\{P_t^{\boldsymbol{K}}\}_{t=0}^T$ induced by a control $\boldsymbol{K}$; Lemma 2.3.5 gives a representation of the gradient term; Lemma 2.3.6 and Lemma 2.3.7 provide the gradient dominance condition and a smoothness condition on the cost function $C(\boldsymbol{K})$ with respect to policy $\boldsymbol{K}$, respectively; and

finally, Lemma 2.3.8 gives two useful upper bounds on Ricatti system and $\mathbb{E}[x_t x_t^\top]$ matrices.

In the finite time horizon setting, define $P_t^{\boldsymbol{K}}$ as the solution to

$$P_t^{\boldsymbol{K}} = Q_t + K_t^\top R_t K_t + (A - BK_t)^\top P_{t+1}^{\boldsymbol{K}} (A - BK_t), \quad t = 0, 1, \cdots, T-1, \quad (2.3.9)$$

with terminal condition

$$P_T^{\boldsymbol{K}} = Q_T.$$

Note that (2.3.9) is equivalent to the Riccati equation (2.2.3) with optimal $K_t = K_t^*$ as given by (2.2.5). We have the following result on the well-definedness of $P_t^{\boldsymbol{K}}$ and the proof can be found in Appendix A.3.1.

Proposition 2.3.4. *Under Assumption 2.2.1, the matrices $P_t^{\boldsymbol{K}}$ for $t = 0, 1, \ldots, T$ derived from* (2.3.9) *are positive definite.*

To ease the exposition, we write $P_t^{\boldsymbol{K}}$ as P_t when there is no confusion. Then the cost of $\boldsymbol{K}$ can be rewritten as

$$C(\boldsymbol{K}) = \mathbb{E}_{x_0 \sim \mathcal{D}}\Big[x_0^\top P_0 x_0 + L_0\Big],$$

where, for $t = 0, 1, \cdots, T-1$,

$$L_t = L_{t+1} + \mathbb{E}[w_t^\top P_{t+1} w_t] = L_{t+1} + \mathrm{Tr}(W P_{t+1}), \quad (2.3.10)$$

with $L_T = 0$. To see this,

$$\begin{aligned}
&\mathbb{E}[x_0^\top P_0 x_0] + L_0 \\
&= \mathbb{E}\left[x_0^\top Q_0 x_0 + x_0^\top K_0^\top R_0 K_0 x_0 + x_0^\top (A - BK_0)^\top P_1 (A - BK_0) x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t\right] \\
&= \mathbb{E}\left[x_0^\top Q_0 x_0 + u_0^\top R_0 u_0 + x_1^\top P_1 x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1} w_t\right] \\
&= \mathbb{E}\Big[\sum_{t=0}^{T-1} \big(x_t^\top Q_t x_t + u_t^\top R_t u_t\big) + x_T^\top Q_T x_T\Big].
\end{aligned}$$

In addition, define

$$E_t^{\boldsymbol{K}} = (R_t + B^\top P_{t+1} B) K_t - B^\top P_{t+1} A,\ t = 0, 1, \cdots, T-1. \tag{2.3.11}$$

To ease the exposition, we write $E_t^{\boldsymbol{K}}$ as E_t when there is no confusion. Then we have the following representation of the gradient term.

Lemma 2.3.5. *The policy gradient has the following representation, for* $t = 0, 1, \cdots, T-1$,

$$\nabla_t C(\boldsymbol{K}) = 2\left(\left(R_t + B^\top P_{t+1} B\right) K_t - B^\top P_{t+1} A\right) \mathbb{E}\left[x_t x_t^\top\right] = 2 E_t^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}}.$$

Proof. Since

$$\begin{aligned} C(\boldsymbol{K}) &= \mathbb{E}\Big[x_0^\top P_0 x_0 + L_0\Big] \\ &= \mathbb{E}\Big[x_0^\top (Q_0 + K_0^\top R_0 K_0) x_0 + x_0^\top (A - B K_0)^\top P_1 (A - B K_0) x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t\Big], \end{aligned}$$

we have

$$\begin{aligned} \nabla_0 C(\boldsymbol{K}) = \frac{\partial C(\boldsymbol{K})}{\partial K_0} &= \mathbb{E}\Big[2 R_0 K_0 x_0 x_0^\top - 2 B^\top P_1 (A - B K_0) x_0 x_0^\top\Big] \\ &= 2 E_0 \mathbb{E}\Big[x_0 x_0^\top\Big] = 2 E_0 \Sigma_0. \end{aligned}$$

Similarly, $\forall\, t = 0, 1, \cdots, T-1$,

$$\nabla_t C(\boldsymbol{K}) = 2\left(\left(R_t + B^\top P_{t+1} B\right) K_t - B^\top P_{t+1} A\right) \mathbb{E}[x_t x_t^\top] = 2 E_t \mathbb{E}\Big[x_t x_t^\top\Big] = 2 E_t \Sigma_t,$$

where the expectation $\mathbb{E}$ is taken with respect to both initial distribution $x_0 \sim \mathcal{D}$ and noises $\boldsymbol{w}$. □

In classical optimization theory [29], gradient domination and smoothness of the objective function are two key conditions to guarantee the global convergence of the gradient descent methods. To prove that $C(\boldsymbol{K})$ is gradient dominated, we first prove Lemma 2.3.6, which indicates that for a policy $\boldsymbol{K}$, the distance between $C(\boldsymbol{K})$ and the optimal cost $C(\boldsymbol{K}^*)$ is bounded by the sum of the magnitude of the gradient $\nabla_t C(\boldsymbol{K})$ for $t = 0, 1, \cdots, T-1$.

Lemma 2.3.6. *Assume Assumption 2.2.1 and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Let $\boldsymbol{K}^*$ be an optimal policy and $C(\boldsymbol{K})$ be finite, then*

$$\underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{\mathrm{Tr}((E_t^{\boldsymbol{K}})^\top E_t^{\boldsymbol{K}})}{\|R_t + B^\top P_{t+1} B\|} \leq C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \mathrm{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K})),$$

where $\underline{\sigma}_{\boldsymbol{X}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ are defined in (2.3.5) *and* (2.3.6).

We defer the proof of Lemma 2.3.6 to Appendix A.3.1. Lemma 2.3.6 implies that when the gradient becomes small, the value of the objective function is close to $C(\boldsymbol{K}^*)$. Now we consider the smoothness condition of the objective function. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be smooth if f has continuous derivatives. An equivalent definition of smoothness [29] used in this thesis is the following

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \leq \frac{M}{2} \|x - y\|^2, \ \forall\, x, y \in \mathbb{R}^n,$$

for some finite constant M. In general, it is difficult to characterize the smoothness of $C(\boldsymbol{K})$, since it may blow up when $A - BK_t$ is large. Here we will prove that $C(\boldsymbol{K})$ is "almost" smooth, in the sense that when $\boldsymbol{K}'$ is sufficiently close to $\boldsymbol{K}$, $C(\boldsymbol{K}') - C(\boldsymbol{K})$ is bounded by the sum of the first and second order terms in $\boldsymbol{K} - \boldsymbol{K}'$.

Lemma 2.3.7 ("Almost Smoothness")**.** *Let $\{x_t'\}$ be the sequence of states for a single trajectory generated by $\boldsymbol{K}'$ starting from $x_0' = x_0$. Then, $C(\boldsymbol{K})$ satisfies*

$$\begin{aligned} C(\boldsymbol{K}') - C(\boldsymbol{K}) \ = \ & \sum_{t=0}^{T-1} \Big[2\,\mathrm{Tr}\left(\Sigma_t'(K_t' - K_t)^\top E_t\right) \\ & + \mathrm{Tr}\left(\Sigma_t'(K_t' - K_t)^\top (R_t + B^\top P_{t+1} B)(K_t' - K_t)\right) \Big], \end{aligned} \tag{2.3.12}$$

where $\Sigma_t' = \mathbb{E}\left[x_t'(x_t')^\top\right]$.

We defer the proof of Lemma 2.3.7 to Appendix A.3.1. To see why Lemma 2.3.7 is related to the smoothness, observe that when $\boldsymbol{K}'$ is sufficiently close to $\boldsymbol{K}$, in the sense that

$$\Sigma_t' \approx \Sigma_t + O(\|K_t - K_t'\|), \ \forall t = 0, 1, \cdots, T-1,$$

the first term in (2.3.12) will behave as $\mathrm{Tr}\left((K_t - K_t')\nabla_t C(\boldsymbol{K})\right)$ by Lemma 2.3.5, and the second term in (2.3.12) will be of second order in $K_t - K_t'$.

To utilize Lemmas 2.3.6 and 2.3.7 in the proof of Theorem 2.3.3, we need to further bound P_t and $\Sigma_{\boldsymbol{K}}$, which is provided below in Lemma 2.3.8. The proof can be found in Appendix A.3.1.

Lemma 2.3.8. *Assume Assumption 2.2.1 holds, and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Then we have*

$$\|P_t\| \le \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}},\ \|\Sigma_{\boldsymbol{K}}\| \le \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}},$$

where $\underline{\sigma}_{\boldsymbol{X}}$ and $\underline{\sigma}_{\boldsymbol{Q}}$ are defined in (2.3.5) *and* (2.3.7).

2.3.2 Perturbation Analysis of Σ_K

First, let us define two linear operators on symmetric matrices. For $X \in \mathbb{R}^{d\times d}$ we set

$$\mathcal{F}_{K_t}(X) = (A - BK_t)X(A - BK_t)^\top,$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X) := X + \sum_{t=0}^{T-1} \Pi_{i=0}^{t}(A - BK_i)\, X\, \Pi_{i=0}^{t}(A - BK_{t-i})^\top.$$

where Π denotes the repeated multiplication from the left, that is,

$$\Pi_{i=0}^{t}(A - BK_i) = (A - BK_t)(A - BK_{t-1})\cdots(A - BK_0).$$

If we write $\mathcal{G}_t = \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \cdots \circ \mathcal{F}_{K_0}$, then

$$\mathcal{G}_t(X) = \mathcal{F}_{K_t} \circ \mathcal{G}_{t-1}(X) = \Pi_{i=0}^{t}(A - BK_i)\, X\, \Pi_{i=0}^{t}(A - BK_{t-i})^\top, \quad (2.3.13)$$

$$\mathcal{T}_{\boldsymbol{K}}(X) = X + \sum_{t=0}^{T-1} \mathcal{G}_t(X). \quad (2.3.14)$$

We also define the induced norm for these operators as

$$\|T\| := \sup_X \frac{\|T(X)\|}{\|X\|},$$

where $T = \mathcal{F}_{K_t}, \mathcal{G}_t, \mathcal{T}_{\boldsymbol{K}}$ and the supremum is over all symmetric matrix X with non-zero spectral norm.

We first show the relationship between the operator $\mathcal{T}_{\boldsymbol{K}}$ and the quantity $\Sigma_{\boldsymbol{K}}$. The proof can be found in Appendix A.3.1.

Proposition 2.3.9. *For $T \geq 2$, we have that*

$$\Sigma_{\boldsymbol{K}} = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta(\boldsymbol{K}, W), \tag{2.3.15}$$

where $\Delta(\boldsymbol{K}, W) = \sum_{t=1}^{T-1} \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + T\,W$, with $D_{t,s} = \Pi_{u=s}^{t}(A - BK_u)$ (for $s = 1, 2, \cdots, t$), and $\Sigma_0 = \mathbb{E}\left[x_0 x_0^\top\right]$.

Let

$$\rho := \max\Big\{ \max_{0\leq t\leq T-1} \|A - BK_t\|, \max_{0\leq t\leq T-1} \|A - BK_t'\|, 1 + \xi \Big\}, \tag{2.3.16}$$

for some small constant $\xi > 0$. Then we have the following result on perturbations of $\Sigma_{\boldsymbol{K}}$.

Lemma 2.3.10 (Perturbation Analysis of Σ_K)**.** *Assume Assumption 2.2.1 holds. Then*

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}'}\right\| &\leq \left\|(\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}'})(\Sigma_0)\right\| + \|\Delta(\boldsymbol{K}, W) - \Delta(\boldsymbol{K}', W)\| \\ &\leq \frac{\rho^{2T} - 1}{\rho^2 - 1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\Big(2\rho\,\|B\|\,|||\boldsymbol{K} - \boldsymbol{K}'||| + \|B\|^2\,|||\boldsymbol{K} - \boldsymbol{K}'|||^2\Big).\end{aligned}$$

Remark 2.3.11. By the definition of ρ in (2.3.16), we have $\rho \geq 1 + \xi > 1$. This regularization term $1 + \xi$ is defined for ease of exposition. Alternatively, if we define $\rho := \max\Big\{ \max_{0\leq t\leq T-1} \|A - BK_t\|, \max_{0\leq t\leq T-1} \|A - BK_t'\| \Big\}$, a similar analysis can still be carried out by considering the different cases: $\rho < 1$, $\rho = 1$ and $\rho > 1$. Note that for the infinite horizon problem, the spectral radius of $A - BK$ needs to be smaller than 1 to guarantee the stability of the system (see [29]). In our setting with finite horizon, instability is not an issue and we do not need a condition on the boundedness of ρ. However, we will show later that ρ does appear in the sample complexity results. The smaller the ρ, the smaller the sample complexity.

The proof of Lemma 2.3.10 is based on the following Lemmas 2.3.12 and 2.3.13, which establish the Lipschitz property for the operators $\mathcal{F}_{K_t}$ and $\mathcal{G}_t$, respectively.

Lemma 2.3.12. *It holds that, $\forall\, t = 0, 1, \cdots, T-1$,*

$$\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \leq 2\|A - BK_t\|\|B\|\|K_t - K_t'\| + \|B\|^2\|K_t - K_t'\|^2. \tag{2.3.17}$$

We refer to [29, Lemma 19] for the proof of Lemma 2.3.12.

Recall the definition of $\mathcal{G}_t$ in (2.3.13) associated with $\boldsymbol{K}$, similarly let us define $\mathcal{G}'_t = \mathcal{F}_{K'_t} \circ \mathcal{F}_{K'_{t-1}} \circ \cdots \circ \mathcal{F}_{K'_0}$ for policy $\boldsymbol{K}'$. Then we have the following perturbation analysis for $\mathcal{G}_t$.

Lemma 2.3.13 (Perturbation Analysis for $\mathcal{G}_t$)**.** *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have that*

$$\sum_{t=0}^{T-1} \Big\| (\mathcal{G}_t - \mathcal{G}'_t)(\Sigma) \Big\| \le \frac{\rho^{2T}-1}{\rho^2-1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K'_t}\| \Big) \|\Sigma\|. \tag{2.3.18}$$

We defer the proof of Lemma 2.3.13 to Appendix A.3.2. The following perturbation analysis on $\mathcal{T}$ follows immediately from Lemma 2.3.13.

Corollary 2.3.14. *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have*

$$\Big\| (\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}'})(\Sigma) \Big\| \le \frac{\rho^{2T}-1}{\rho^2-1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K'_t}\| \Big) \|\Sigma\|, \tag{2.3.19}$$

where ρ is defined in (2.3.16).

Now we are ready for the proof of Lemma 2.3.10.

Proof of Lemma 2.3.10. Using Lemma 2.3.12,

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t} - \mathcal{F}_{K'_t}\| &= \sum_{t=0}^{T-1} \Big(2\|A - BK_t\|\|B\|\|K_t - K'_t\| + \|B\|^2 \|K_t - K'_t\|^2 \Big) \\
&\le 2\rho\|B\| \sum_{t=0}^{T-1} \|K_t - K'_t\| + \|B\|^2 \sum_{t=0}^{T-1} \|K_t - K'_t\|^2.
\end{aligned}$$

In the same way as for the proof of Lemma 2.3.13, we have, $\forall\, t = 1, \cdots, T-1$,

$$\sum_{s=1}^{t} \big\| D_{t,s} W D_{t,s}^\top - D'_{t,s} W (D'_{t,s})^\top \big\| \le \frac{\rho^{2T}-1}{\rho^2-1} \left(\sum_{s=0}^{t} \|\mathcal{F}_{K_s} - \mathcal{F}_{K'_s}\| \right) \|W\|. \tag{2.3.20}$$

By Proposition 2.3.9, Corollary 2.3.14, (2.3.14) and (2.3.20), we have

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}}-\Sigma_{\boldsymbol{K}'}\right\| &\leq \left\|(\mathcal{T}_{\boldsymbol{K}}-\mathcal{T}_{\boldsymbol{K}'})(\Sigma_0)\right\| + \sum_{t=1}^{T-1}\sum_{s=1}^{t}\left\|D_{t,s}WD_{t,s}^{\top}-D'_{t,s}W(D'_{t,s})^{\top}\right\| \\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t}-\mathcal{F}_{K'_t}\|\Big)\left(\|\Sigma_0\|+T\|W\|\right) \\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)\Big(2\rho\,\|B\|\,|||\boldsymbol{K}-\boldsymbol{K}'|||+\|B\|^2\,|||\boldsymbol{K}-\boldsymbol{K}'|||^2\Big).\end{aligned} \tag{2.3.21}$$

The last inequality holds since $\|\Sigma_0\| \leq \|\Sigma_{\boldsymbol{K}}\| \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$ by Lemma 2.3.8. □

2.3.3 Convergence and Complexity Analysis

We now provide the proof of Theorem 2.3.3 after two preliminary Lemmas.

Lemma 2.3.15. *Assume Assumption 2.2.1 holds, $\underline{\sigma}_{\boldsymbol{X}} > 0$, and that*

$$K'_t = K_t - \eta\nabla_t C(\boldsymbol{K}), \qquad \text{where} \tag{2.3.22}$$

$$\eta \leq \min\left\{\frac{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2T(\rho^{2T}-1)(2\rho+1)(C(\boldsymbol{K})+\underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|)\|B\|\max_t\{\|\nabla_t C(\boldsymbol{K})\|\}}, \frac{1}{2C_1}\right\}, \tag{2.3.23}$$

with

$$\begin{aligned}C_1 \;=\; &\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)\left(\frac{(2\rho+1)\|B\|(\rho^{2T}-1)}{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{X}}}\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\right) \\ &+\frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\sum_{t=0}^{T-1}\|R_t+B^{\top}P_{t+1}B\|.\end{aligned} \tag{2.3.24}$$

Then we have

$$C(\boldsymbol{K}')-C(\boldsymbol{K}^*) \leq \Big(1-2\eta\,\underline{\sigma}_{\boldsymbol{R}}\,\frac{{\underline{\sigma}_{\boldsymbol{X}}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K})-C(\boldsymbol{K}^*)\Big).$$

We defer the proof of Lemma 2.3.15 to Appendix A.3.3.

Lemma 2.3.16. *Assume Assumption 2.2.1 holds and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Then we have that*

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*)),$$

and that:

$$\sum_{t=0}^{T-1} \|K_t\| \leq \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}\Big(\sqrt{T \cdot \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*))} + \sum_{t=0}^{T-1} \|B^\top P_{t+1} A\|\Big).$$

Proof. Using Lemma 2.3.8 we have

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4 \sum_{t=0}^{T-1} \|\Sigma_t\|^2 \operatorname{Tr}(E_t^\top E_t) \leq 4\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \sum_{t=0}^{T-1} \operatorname{Tr}(E_t^\top E_t).$$

From Lemma 2.3.6 we have

$$\begin{aligned} C(\boldsymbol{K}) - C(\boldsymbol{K}^*) &\geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1} B\|} \operatorname{Tr}(E_t^\top E_t) \\ &\geq \frac{\underline{\sigma}_{\boldsymbol{X}}}{\max_t \|R_t + B^\top P_{t+1} B\|} \sum_{t=0}^{T-1} \operatorname{Tr}(E_t^\top E_t), \end{aligned} \tag{2.3.25}$$

and hence

$$\sum_{t=0}^{T-1} \|\nabla_t C(\boldsymbol{K})\|^2 \leq 4\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \frac{\max_t \|R_t + B^\top P_{t+1} B\|}{\underline{\sigma}_{\boldsymbol{X}}} (C(\boldsymbol{K}) - C(\boldsymbol{K}^*)).$$

For the second claim, using Lemma 2.3.6 again,

$$\begin{aligned} \sum_{t=0}^{T-1} \|K_t\| &= \sum_{t=0}^{T-1} \|(R_t + B^\top P_{t+1} B)^{-1} K_t (R_t + B^\top P_{t+1} B)\| \\ &\leq \sum_{t=0}^{T-1} \frac{1}{\sigma_{\min}(R_t)} \|K_t (R_t + B^\top P_{t+1} B)\| \leq \sum_{t=0}^{T-1} \frac{1}{\sigma_{\min}(R_t)} \Big(\|E_t\| + \|B^\top P_{t+1} A\|\Big) \\ &\leq \sum_{t=0}^{T-1} \left(\frac{\sqrt{\operatorname{Tr}(E_t^\top E_t)}}{\sigma_{\min}(R_t)} + \frac{\|B^\top P_{t+1} A\|}{\sigma_{\min}(R_t)} \right) \end{aligned}$$

$$\leq \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}\Big(\sqrt{T\cdot\sum_{t=0}^{T-1}\mathrm{Tr}(E_t^\top E_t)}+\sum_{t=0}^{T-1}\|B^\top P_{t+1}A\|\Big)$$
$$\leq \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}\Big(\sqrt{T\cdot\frac{\max_t\|R_t+B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}(C(\boldsymbol{K})-C(\boldsymbol{K}^*))}+\sum_{t=0}^{T-1}\|B^\top P_{t+1}A\|\Big).$$

The second inequality holds by the definition of E_t in (2.3.11), the second last step uses the Cauchy-Schwarz inequality, and the last inequality holds by (2.3.25). □

Proof of Theorem 2.3.3. In order to show the existence of a positive η such that (2.3.23) holds, it suffices to show there exists a positive lower bound on the RHS of (2.3.23). By Lemma 2.3.16 and the Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\| &\leq \sqrt{T\cdot\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|^2}\\ &\leq \sqrt{4T\cdot\Big(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2\frac{\max_t\|R_t+B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}(C(\boldsymbol{K})-C(\boldsymbol{K}^*))}.\end{aligned}\tag{2.3.26}$$

Note that if $d<ab+c$ for some $a>0$, $b>0$ $c>0$ and $d>0$, then $\frac{1}{d}>\frac{1}{(a+1)(b+1)(c+1)}$. Also $\frac{1}{a^n+1}>\frac{1}{(a+1)^n}$ for $a>0$ and $n\in\mathbb{N}^+$. Therefore, based on (2.3.24) and (2.3.26), $\frac{1}{C_1}$ is bounded below by polynomials in $\frac{1}{\rho}$, $\frac{1}{C(\boldsymbol{K})+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\boldsymbol{R}\|+1}$, $\frac{1}{\|W\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, and $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$.

Now we aim to show that $\frac{1}{\rho}$ is bounded below by some polynomials in the parameters. To see this, let us first show that ρ is bounded above by polynomials in $\|A\|$, $\|B\|$, $\|\boldsymbol{R}\|$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$. Since $\|B\|\|K_t'-K_t\|\leq\frac{\underline{\sigma}_{\boldsymbol{Q}}\underline{\sigma}_{\boldsymbol{X}}}{4C(\boldsymbol{K})}\leq\frac{1}{2}$ holds under the assumptions in Lemma 2.3.15, we have

$$\max_{0\leq t\leq T-1}\|A-BK_t'\|\leq\max_{0\leq t\leq T-1}\left(\|A-BK_t\|+\|B\|\,\|K_t'-K_t\|\right)\leq\max_{0\leq t\leq T-1}\|A-BK_t\|+\frac{1}{2},$$

thus

$$\begin{aligned}\rho&=\max\Big\{\max_{0\leq t\leq T-1}\|A-BK_t\|,\max_{0\leq t\leq T-1}\|A-BK_t'\|,\ 1+\xi\Big\}\\ &\leq\max\Big\{\max_{0\leq t\leq T-1}\|A-BK_t\|+\frac{1}{2},\ 1+\xi\Big\}\leq\max\Big\{\|A\|+\|B\|\sum_{t=0}^{T-1}\|K_t\|+\frac{1}{2},\ 1+\xi\Big\}.\end{aligned}\tag{2.3.27}$$

Given the bound on $\sum_{t=0}^{T-1}\|K_t\|$ by Lemma 2.3.16 and $\|P_t\| \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}}$ by Lemma 2.3.8, ρ is bounded above by polynomials in $\|A\|$, $\|B\|$, $\|\boldsymbol{R}\|$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$, or a constant $1+\xi$. Therefore $\frac{1}{\rho}$ is bounded below by polynomials in $\frac{1}{\|A\|+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\boldsymbol{R}\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{R}}$ and $\frac{1}{C(\boldsymbol{K})+1}$, or a constant $\frac{1}{1+\xi}$. Hence, by choosing $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ to be an appropriate polynomial in $\frac{1}{C(\boldsymbol{K}^0)}$, $\frac{1}{C(\boldsymbol{K}^0)+1}$, $\frac{1}{\|A\|+1}$, $\frac{1}{\|B\|+1}$, $\frac{1}{\|\boldsymbol{R}\|+1}$, $\frac{1}{\|W\|+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, and $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, (2.3.23) is satisfied, since by performing gradient descent, $C(\boldsymbol{K}^1) < C(\boldsymbol{K}^0)$. Therefore, by Lemma 2.3.15, we have

$$C(\boldsymbol{K}^1) - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\,\underline{\sigma}_{\boldsymbol{R}}\,\frac{{\underline{\sigma}_{\boldsymbol{X}}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)\Big),$$

which implies that the cost decreases at $n = 1$. Suppose that $C(\boldsymbol{K}^n) \leq C(\boldsymbol{K}^0)$, then the stepsize condition in (2.3.23) is still satisfied by Lemma 2.3.16. Thus, Lemma 2.3.15 can again be applied for the update at round $n+1$ to obtain:

$$C(\boldsymbol{K}^{n+1}) - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\,\underline{\sigma}_{\boldsymbol{R}}\,\frac{{\underline{\sigma}_{\boldsymbol{X}}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}^n) - C(\boldsymbol{K}^*)\Big).$$

For $\epsilon > 0$, provided $N \geq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{2\eta\,{\underline{\sigma}_{\boldsymbol{X}}}^2\,\underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0)-C(\boldsymbol{K}^*)}{\epsilon}$, we have

$$C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) \leq \epsilon.$$

□

2.4 Sample-based Policy Gradient Method with Unknown Parameters

In the setting with unknown parameters, the controller has only simulation access to the model; the model parameters, A, B, $\{Q_t\}_{t=0}^{T}$, $\{R_t\}_{t=0}^{T-1}$, are unknown. By using a zeroth-order optimization method to approximate the gradient, this section proves the policy gradient method with unknown parameters also leads to a global optimal policy, with both polynomial computational and sample complexities.

Note that in this section, when bounding the Frobenius norm of a matrix, we usually treat the matrix as a stacked vector. Therefore we denote by $D = k \times d$ the dimension of the corresponding vector formed from the $\boldsymbol{K}$ matrix for convenience in

the proofs. Therefore in each iteration $n = 1, 2, \cdots, N$, we can update the policy as, for $t = 0, 1, \cdots, T-1$,

$$K_t^{n+1} = K_t^n - \eta \widehat{\nabla_t C(\boldsymbol{K}^n)}, \tag{2.4.1}$$

where $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ is the estimate of $\nabla_t C(\boldsymbol{K}^n)$. We analyze the following Algorithm 1.

Algorithm 1 Policy Gradient Estimation with Unknown Parameters

1: **Input**: $\boldsymbol{K}$, number of trajectories m, smoothing parameter r, dimension D
2: **for** $i \in \{1, \ldots, m\}$ **do**
3: **for** $t \in \{0, \ldots, T-1\}$ **do**
4: Sample the (sub)-policy at time t: $\widehat{K}_t^i = K_t + U_t^i$ where U_t^i is drawn uniformly at random over matrices such that $\|U_t^i\|_F = r$.
5: Denote $\widehat{c}_t^{\,i}$ as the single trajectory cost with policy $(\boldsymbol{K}_{-t}, \widehat{K}_t^i) := (K_0, \cdots, K_{t-1}, \widehat{K}_t^i, K_t, \cdots, K_{T-1})$ starting from $x_0^i \sim \mathcal{D}$.
6: **end for**
7: **end for**
8: Return the estimates of $\nabla_t C(\boldsymbol{K})$ for each t:

$$\widehat{\nabla_t C(\boldsymbol{K})} = \frac{1}{m} \sum_{i=1}^{m} \frac{D}{r^2} \widehat{c}_t^{\,i} U_t^i. \tag{2.4.2}$$

Remark 2.4.1. [Zeroth-order Optimization Approach in the Sub-routine (2.4.2)] Note that (2.4.2) provides an estimate of the gradient $\nabla_t C(\boldsymbol{K})$ by Lemma 2.4.9. In (2.4.2), we adopt a zeroth-order optimization method, using only query access to a sample of the reward function $c(\cdot)$ at input points $\boldsymbol{K}$, without querying the gradients and higher order derivatives of $c(\cdot)$. The step (2.4.2) (in Algorithm 1) provides a procedure to find an (bounded bias) estimate $\widehat{\nabla C(\boldsymbol{K})}$ of $\nabla C(\boldsymbol{K})$.

The idea in (2.4.2) is to approximate the gradient of a function by only using the function values (see e.g. Lemma 2.1 in [32]). Observe that by a Taylor expansion to first order $\mathbb{E}[f(x+U)U] \approx \mathbb{E}[(\nabla f(x) \cdot U)U] = \nabla f(x) r^2 / D$ when $x \in \mathbb{R}^D$ and U is uniform over the surface of the ball of radius r in $\mathbb{R}^D$. Thus the gradient of the function f at x can be estimated by averaging over the samples $\frac{D}{r^2} f(x+U)U$.

Note that in Algorithm 1, we require mNT^2 samples to perform the policy gradient method N times.

To guarantee the global convergence of the sample-based algorithm (Algorithm 1), we propose some conditions on the distribution of x_0 and $\{w_t\}_{t=0}^{T-1}$, in addition to the finite second moment condition specified in Section 2.2.

Definition 2.4.2 ([85, Chapter 2]). *A zero-mean random variable X*

1. *is said to be sub-Gaussian with variance proxy σ^2 ($\sigma > 0$) and we write $X \in SG(\sigma^2)$ if its moment generating function satisfies $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.*

2. *is said to be sub-exponential with non-negative parameters (ν^2, α) and we write $X \in SE(\nu^2, \alpha)$, if $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right)$ for any λ such that $|\lambda| \leq \frac{1}{\alpha}$.*

We assume the initial distribution and the noise in the state process dynamics satisfy the following assumptions.

Assumption 2.4.3 (Initial State and Noise Process (II)). *1. Initial state: $x_0 = \widetilde{W}_0 z_0$ where $z_0 = (z_{0,1} \cdots, z_{0,d}) \in \mathbb{R}^d$ is a random vector with independent components $z_{0,i}$ which are sub-Gaussian, mean-zero, and have sub-Gaussian parameter σ_0^2; $\widetilde{W}_0 \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*

2. *Noise process: $w_t = \widetilde{W} v_t$ where $v_t := (v_{t,1} \cdots, v_{t,d}) \in \mathbb{R}^d$ are i.i.d. and independent from x_0. v_t has independent components $v_{t,i}$ which are sub-Gaussian, mean-zero, and have sub-Gaussian parameter σ_w^2, $\forall\, t = 0, 1, \cdots, T-1$. $\widetilde{W} \in \mathbb{R}^{d\times d}$ is an unknown and deterministic matrix.*

Note that Assumptions 2.3.1 and 2.4.3 serve different purposes in this chapter. Assumption 2.3.1 provides one sufficient condition to assure $\underline{\sigma}_{\boldsymbol{X}} > 0$. Assumption 2.4.3 is used to guarantee the convergence of the sample based algorithm (Algorithm 1).

In addition to the model parameters specified in Section 2.3, here we assume $\mathcal{H}(\cdot)$ includes polynomials that are also functions of σ_0, $\frac{1}{\sigma_0}$, $\frac{1}{\sigma_0+1}$, σ_w, $\frac{1}{\sigma_w}$, $\frac{1}{\sigma_w+1}$ $\|\widetilde{W}\|$,$\frac{1}{\|\widetilde{W}\|}$, $\frac{1}{\|\widetilde{W}\|+1}$, $\|\widetilde{W}_0\|$, $\frac{1}{\|\widetilde{W}_0\|}$, and $\frac{1}{\|\widetilde{W}_0\|+1}$.

Theorem 2.4.4. *Assume Assumptions 2.2.1 and 2.4.3 hold and further assume $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $C(\boldsymbol{K}^0)$ is finite. At every step the policy is updated as in (2.4.1), that is*

$$K_t^{n+1} = K_t^n - \eta \widehat{\nabla_t C(\boldsymbol{K}^n)},$$

with $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ and $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ is computed with hyper-parameters (r, m) such that $r < 1/\overline{h}_{radius}$ and $m > \overline{h}_{sample}$ with some fixed polynomials $\overline{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\overline{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. Then for $\epsilon > 0$, if we have

$$N \geq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}} \log \frac{C(\boldsymbol{K}^0) - C(\boldsymbol{K}^*)}{\epsilon},$$

it holds that $C(\boldsymbol{K}^N) - C(\boldsymbol{K}^*) \leq \epsilon$ *with high probability (at least* $1 - 2TN\left(\frac{\epsilon}{D}\right)^D$*).*

The proof of Theorem 2.4.4 is based on a perturbation analysis of $C(\boldsymbol{K})$ and $\nabla_t C(\boldsymbol{K})$, smoothing and the gradient descent analysis of the procedures in Algorithm 1. We provide the perturbation analysis and the smoothing analysis in Sections 2.4.1 and 2.4.2, respectively. We defer the proof of Theorem 2.4.4 to Section 2.4.3.

Projected Policy Gradient Method. In many situations constrained optimization problems arise and the *projected* gradient descent method is one popular approach to solve such problems. Recall the projection of a point $\boldsymbol{y} = (y_0, \cdots, y_{T-1})$ with $y_t \in \mathbb{R}^{k\times d}$ $(t = 0, 1, \cdots, T-1)$ onto a set $\mathcal{S} \subset \mathbb{R}^{k\times(T\times d)}$ is defined as

$$\Pi_{\mathcal{S}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}\in\mathcal{S}} \frac{1}{2}\sum_{t=0}^{T-1} \|x_t - y_t\|_F^2 . \tag{2.4.3}$$

Then the projected policy gradient (PPG) updating rule can be defined as

$$\boldsymbol{K}^{n+1} = \Pi_{\mathcal{S}}\left(\boldsymbol{K}^n - \eta\widehat{\nabla C(\boldsymbol{K}^n)}\right), \tag{2.4.4}$$

where $\widehat{\nabla C(\boldsymbol{K}^n)} = \left(\widehat{\nabla_0 C(\boldsymbol{K}^n)}, \cdots, \widehat{\nabla_{T-1} C(\boldsymbol{K}^n)}\right)$ denotes the estimate of $\nabla C(\boldsymbol{K}^n)$.

If the projection set $\mathcal{S}$ is convex and closed, the projection onto $\mathcal{S}$ is non-expansive, that is,
$\sum_{t=0}^{T-1} \|\widetilde{z}_t^1 - \widetilde{z}_t^2\|_F \leq \sum_{t=0}^{T-1} \|z_t^1 - z_t^2\|_F$ with $\widetilde{\boldsymbol{z}}^1 = \Pi_{\mathcal{S}}(\boldsymbol{z}^1)$ and $\widetilde{\boldsymbol{z}}^2 = \Pi_{\mathcal{S}}(\boldsymbol{z}^2)$. Given any policy matrix $\boldsymbol{K}$ and learning rate η, define the *gradient mapping for the projection operator*

$$G(\boldsymbol{K}) := \frac{\Pi_{\mathcal{S}}(\boldsymbol{K} - \eta\nabla C(\boldsymbol{K})) - \boldsymbol{K}}{2\eta}, \tag{2.4.5}$$

with $G(\boldsymbol{K}) = (G_0(\boldsymbol{K}), \cdots, G_{T-1}(\boldsymbol{K}))$. Note that the gradient mapping has been commonly adopted in the analysis of projected gradient descent methods in constrained optimization [60, 93]. A policy matrix $\widetilde{\boldsymbol{K}} \in \mathcal{S}$ is called a stationary point of $C(\cdot)$ if

$$\nabla C(\widetilde{\boldsymbol{K}})^\top(\boldsymbol{K} - \widetilde{\boldsymbol{K}}) \leq 0, \quad \forall \boldsymbol{K} \in \mathcal{S}. \tag{2.4.6}$$

It is well-known in the optimization literature that (2.4.6) holds if and only if $G(\widetilde{\boldsymbol{K}}) = 0$. We have the following sub-linear convergence result for the PPG version.

Theorem 2.4.5. *Assume Assumptions 2.2.1 and 2.4.3 hold, and the projection set of policies, denoted by $\mathcal{S}$, is convex and closed. Further assume $\boldsymbol{K}^* \in \mathcal{S}$, $\boldsymbol{K}^0 \in \mathcal{S}$, $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $C(\boldsymbol{K}^0)$ is finite. At every step the policy is updated as in (4.4), that is*

$$\boldsymbol{K}^{n+1} = \Pi_{\mathcal{S}} \left(\boldsymbol{K}^n - \eta \widehat{\nabla C(\boldsymbol{K}^n)} \right)$$

with $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ and $\widehat{\nabla_t C(\boldsymbol{K}^n)}$ $(t = 0, 1, \cdots, T-1)$ is computed with hyper-parameters (r, m) such that $r < 1/\widehat{h}_{radius}$ and $m > \widehat{h}_{sample}$ with some fixed polynomials $\widehat{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\widehat{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. Then the projected policy gradient method has a global sublinear convergence rate, that is, $\left\{ \frac{1}{N} \sum_{n=0}^{N-1} \left(\sum_{t=0}^{T-1} \|G_t(\boldsymbol{K}^n)\|_F^2 \right) \right\}_{N \geq 1}$ converges to 0 at rate $\mathcal{O}\left(\frac{1}{N}\right)$, where $G_t(\boldsymbol{K})$ is defined in (2.4.5).

The proof of Theorem 2.4.5 can be found in Appendix A.3.4.

Remark 2.4.6. We assume the projection step is perform accurately and the associated computational cost is of a separate interest and hence omitted here. The convergence result in Theorem 2.4.5 is described in terms of the sample complexity and to perform the projection step does not need extra samples.

2.4.1 Perturbation analysis of $C(\boldsymbol{K})$ and $\nabla_t C(\boldsymbol{K})$

This section shows that the objective function $C(\boldsymbol{K})$ and its gradient are stable with respect to small perturbations. The proofs of the following Lemmas can be found in Appendix A.3.4.

Lemma 2.4.7 ($C(\boldsymbol{K})$ Perturbation)**.** *Assume Assumptions 2.2.1 and 2.4.3 hold, $\underline{\sigma}_{\boldsymbol{X}} > 0$, and $\boldsymbol{K}'$ such that, $\forall\, t = 0, 1, \cdots, T-1$,*

$$\|K_t' - K_t\| \leq \min \left\{ \frac{(\rho^2 - 1)\, \underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{2T(\rho^{2T} - 1)(2\rho + 1)(C(\boldsymbol{K}) + \underline{\sigma}_{\boldsymbol{Q}}\, T\|W\|)\|B\|}, \|K_t\|, \frac{1}{T} \right\}, \quad (2.4.7)$$

where ρ is defined in (2.3.16)*. Then there exists a polynomial $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$|C(\boldsymbol{K}') - C(\boldsymbol{K})| \leq h_{cost} |||\boldsymbol{K}' - \boldsymbol{K}|||.$$

Lemma 2.4.8 ($\nabla_t C(\boldsymbol{K})$ Perturbation)**.** *Under the same assumptions as in Lemma 2.4.7, there exists a polynomial $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| \leq h_{grad}(\|\boldsymbol{K}' - \boldsymbol{K}\|), \text{ and } \|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\|_F \leq h_{grad}(\|\boldsymbol{K}' - \boldsymbol{K}\|_F).$$

2.4.2 Smoothing and the Gradient Descent Analysis

In this section, Lemma 2.4.9 provides the formula for the perturbed gradient term, Lemma 2.4.10 provides the concentration inequality for finite samples, and Lemma 2.4.11 provides the guarantees for the gradient approximation.

Recall that $D = k \times d$. Let $\mathbb{S}_r$ represent the uniform distribution over the points with norm r in dimension D, and $\mathbb{B}_r$ represent the uniform distribution over all points with norm at most r in dimension D. For each K_t ($t = 0, 1, \cdots, T-1$), the algorithm performs gradient descent on the following function:

$$C_t^r(\boldsymbol{K}) = \mathbb{E}_{V_t \sim \mathbb{B}_r}\left[C(\boldsymbol{K} + \boldsymbol{V}_t)\right], \tag{2.4.8}$$

where $\boldsymbol{V}_t := (0, \cdots, V_t, \cdots, 0)$ and $V_t \in \mathbb{R}^{k \times d}$.

Lemma 2.4.9. *Assume $C(\boldsymbol{K})$ is finite, then*

$$\nabla_t C_t^r(\boldsymbol{K}) = \frac{D}{r^2}\mathbb{E}_{U_t \sim \mathbb{S}_r}[C(\boldsymbol{K} + \boldsymbol{U}_t)U_t], \tag{2.4.9}$$

where $\boldsymbol{U}_t := (0, \cdots, U_t, \cdots, 0)$.

The proof of Lemma 2.4.9 is similar to the proof of [29, Lemma 29] and hence omitted.

We first state two facts on sub-Gaussian and sub-exponential random variables. Firstly, if X and Y are zero-mean independent random variables such that $X \in SG(\sigma_x^2)$ and $Y \in SG(\sigma_y^2)$, then $XY \in SE(\sigma_x\sigma_y, 4\sigma_x\sigma_y)$. Secondly, if $X_1, \cdots, X_n$ are zero-mean independent random variables such that $X_i \in SE(\nu_i^2, \alpha_i)$, then

$$\sum_{i=1}^n X_i \in SE\left(\sum_{i=1}^n \nu_i^2, \max_i \alpha_i\right).$$

Using the above two facts, we have the following.

Lemma 2.4.10. *Assume Assumptions 2.2.1 and 2.4.3 hold and $\underline{\sigma}_{\boldsymbol{X}} > 0$, then there exist polynomials $\nu \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ and $\alpha \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ such that*

$$\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right]$$

is sub-exponential with parameter (ν^2, α). Here $\{x_t\}_{t=0}^T$ is the dynamics under policy $\boldsymbol{K}$.

Proof. We first observe that, by direct calculation,

$$\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t\right) + x_T^\top Q_T x_T\right] = x_0^\top P_0 x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1} w_t. \tag{2.4.10}$$

Note that by (2.3.9) and Proposition 2.3.4, P_t is symmetric and positive definite. The Frobenius norm $\|\cdot\|_F$ and the spectral norm $\|\cdot\|$ of the matrix $P_t \in \mathbb{R}^{d\times d}$ have the following property:

$$\|P_t\| \le \|P_t\|_F \le \sqrt{d}\|P_t\|, \ \forall\, t = 0, 1, \cdots, T. \tag{2.4.11}$$

Let $\widehat{\sigma} = \max\{\sigma_0, \sigma_w\}$. Given the Hanson-Wright inequality (Theorem 2.5 in [4]),

$$\begin{aligned}
& \mathbb{P}\left(\left|w_t^\top P_{t+1} w_t - \mathbb{E}\left[w_t^\top P_{t+1} w_t\right]\right| \ge t\right) \\
= \ & \mathbb{P}\left(\left|v_t^\top (\widetilde{W}^\top P_{t+1} \widetilde{W}) v_t - \mathbb{E}\left[v_t^\top (\widetilde{W}^\top P_{t+1} \widetilde{W}) v_t\right]\right| \ge t\right) \\
\le \ & 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4 \|\widetilde{W}^\top P_{t+1}\widetilde{W}\|_F^2}, \frac{t}{\widehat{\sigma}^2\|\widetilde{W}^\top P_{t+1}\widetilde{W}\|}\right\}\right),
\end{aligned} \tag{2.4.12}$$

for some universal constant $c > 0$ which is independent of P_{t+1} and w_t.

Combining (2.4.11), (2.4.12) and Lemma 2.3.8,

$$\begin{aligned}
& \mathbb{P}\left(\left|w_t^\top P_{t+1} w_t - \mathbb{E}\left[w_t^\top P_{t+1} w_t\right]\right| \ge t\right) \\
\le \ & 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4\, d\, \|P_{t+1}\|^2 \|\widetilde{W}\|^4}, \frac{t}{\widehat{\sigma}^2 \|P_{t+1}\| \|\widetilde{W}\|^2}\right\}\right) \\
\le \ & 2\exp\left(-c\min\left\{\frac{t^2}{2\widehat{\sigma}^4 \|\widetilde{W}\|^4\, d\, C^2(\boldsymbol{K})/\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{t}{\widehat{\sigma}^2\|\widetilde{W}\|^2 C(\boldsymbol{K})/\,\underline{\sigma}_{\boldsymbol{X}}}\right\}\right).
\end{aligned}$$

Therefore the random variable $w_t^\top P_{t+1} w_t$ is sub-exponential with parameters $\left(\frac{\widehat{\sigma}^4 \|\widetilde{W}\|^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2 \|\widetilde{W}\|^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right)$. In the same way $x_0^\top P_0 x_0$ is sub-exponential with parameters $\left(\frac{\widehat{\sigma}^4 \|\widetilde{W}_0\|^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2 \|\widetilde{W}_0\|^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right)$. Let $\overline{\sigma} = \max\{\|\widetilde{W}_0\|, \|\widetilde{W}\|\}$. Since $\{w_t\}_{t=0}^{T-1}$ are i.i.d. and independent from x_0, we have (2.4.10) is sub-exponential with parameters

$$\left((T+1)\,\frac{\widehat{\sigma}^4 \overline{\sigma}^4 dC^2(\boldsymbol{K})}{c\,\underline{\sigma}_{\boldsymbol{X}}{}^2}, \frac{\widehat{\sigma}^2 \overline{\sigma}^2 C(\boldsymbol{K})}{2c\,\underline{\sigma}_{\boldsymbol{X}}}\right).$$

□

Define

$$\widetilde{\nabla}_t := \frac{1}{m}\sum_{i=1}^{m}\left(\frac{D}{r^2} C(\boldsymbol{K} + \boldsymbol{U}_t^i) U_t^i\right)$$

as the average of perturbed cost functions across m scenarios which is an empirical approximation of (2.4.9). Similarly, define

$$\widehat{\nabla}_t := \frac{1}{m}\sum_{i=1}^{m}\left(\frac{D}{r^2}\left[\sum_{t=0}^{T-1}\left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right] U_t^i\right) \tag{2.4.13}$$

as the average of perturbed and single-trajectory-based cost functions across m scenarios, which is the same as (2.4.2) in Algorithm 1. Note that in order to calculate $\widetilde{\nabla}_t$, we require access to $C(\boldsymbol{K} + \boldsymbol{U}_t^i)$, which involves the calculation of expectations with respect to unknown initial states and state noises. This may be restrictive in some settings. On the other hand, the calculation of $\widehat{\nabla}_t$ only involves single-trajectory-based cost functions.

Lemma 2.4.11. *Assume Assumptions 2.2.1 and 2.4.3 hold, and $\underline{\sigma}_{\boldsymbol{X}} > 0$. Given any ϵ, there are fixed polynomials $h_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ and $h_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ such that when $r \le 1/h_{radius}$, with $m \ge h_{sample}$ samples of $U_t^1, \cdots, U_t^m \sim \mathbb{S}_r$ for each $t = 0, \cdots, T-1$,*

$$\left\|\widetilde{\nabla}_t - \nabla_t C(\boldsymbol{K})\right\|_F \le \epsilon,$$

holds with high probability (at least $1 - \left(\frac{D}{\epsilon}\right)^{-D}$). In addition, there is a polynomial $h_{sample,2} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ such that when $r \le 1/h_{radius}$, with $m \ge h_{sample} + h_{sample,2}$ samples of $U_t^1, ..., U_t^m \sim \mathbb{S}_r$ for each $t = 0, \cdots, T-1$,

$$\left\|\widehat{\nabla}_t - \nabla_t C(\boldsymbol{K})\right\|_F \le \frac{3}{2}\epsilon,$$

holds with high probability (at least $1 - 2\left(\frac{D}{\epsilon}\right)^{-D}$*). Here, for each* $i = 1, 2, \cdots, m$*,* $\{x_t^i\}_{t=0}^T$ *and* $\{u_t^i\}_{t=0}^{T-1}$ *are the dynamics and controls for a single path sampled using policy* $\boldsymbol{K} + \boldsymbol{U}_t^i$*.*

Proof. Note that

$$\widetilde{\nabla}_t - \nabla_t C(\boldsymbol{K}) = (\nabla_t C_t^r(\boldsymbol{K}) - \nabla_t C(\boldsymbol{K})) + (\widetilde{\nabla}_t - \nabla_t C_t^r(\boldsymbol{K})),$$

where C_t^r is defined in (2.4.8).

For the first term, choose $h_{radius} = \max\{1/r_0, 4h_{grad}/\epsilon\}$ (r_0 is chosen later), where $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ is defined in Lemma 2.4.8. By Lemma 2.4.8 when $r \leq 1/h_{radius} \leq \epsilon/4h_{grad}$, for $\boldsymbol{V}_t := (0, \cdots, V_t, \cdots, 0)$ where $V_t \sim \mathbb{B}_r$, we have

$$\|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C(\boldsymbol{K})\|_F \leq h_{grad}\|\boldsymbol{V}_t\|_F \leq h_{grad}\frac{\epsilon}{4h_{grad}} = \frac{\epsilon}{4}. \tag{2.4.14}$$

Since $\nabla_t C_t^r(\boldsymbol{K}) = \mathbb{E}_{V_t \sim \mathbb{B}_r}[\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t)]$, we have

$$\|\nabla_t C(\boldsymbol{K} + \boldsymbol{V}_t) - \nabla_t C_t^r(\boldsymbol{K})\|_F \leq \frac{\epsilon}{4},$$

by (2.4.14) and the continuity of $\nabla_t C$. Therefore

$$\|\nabla_t C_t^r(\boldsymbol{K}) - \nabla_t C(\boldsymbol{K})\|_F \leq \|\nabla_t C(\boldsymbol{K}+\boldsymbol{V}_t) - \nabla_t C(\boldsymbol{K})\|_F + \|\nabla_t C(\boldsymbol{K}+\boldsymbol{V}_t) - \nabla_t C_t^r(\boldsymbol{K})\|_F \leq \frac{\epsilon}{2} \tag{2.4.15}$$

holds by triangle inequality. We choose r_0 such that for any $\boldsymbol{U}_t \sim \mathbb{S}_r$, we have that $C(\boldsymbol{K} + \boldsymbol{U}_t) \leq 2C(\boldsymbol{K})$. By Lemma 2.4.7, we can pick $1/r_0 = h_{cost}/C(\boldsymbol{K})$, then $|C(\boldsymbol{K} + \boldsymbol{U}_t) - C(\boldsymbol{K})| \leq r_0 \cdot h_{cost} \leq C(\boldsymbol{K})$.

For the second term, by Lemma 2.4.9, $\mathbb{E}[\widetilde{\nabla}_t] = \nabla_t C_t^r(\boldsymbol{K})$, and each individual sample is bounded by $2DC(\boldsymbol{K})/r$, so by the Operator-Bernstein inequality [36, Theorem 12] with

$$m \geq h_{sample} = \Theta\left(D\left(\frac{D \cdot C(\boldsymbol{K})}{r\epsilon}\right)^2 \log(D/\epsilon)\right),$$

we have

$$\mathbb{P}\left[\left\|\widetilde{\nabla}_t - \nabla_t C_t^r(\boldsymbol{K})\right\|_F \leq \frac{\epsilon}{2}\right] \geq 1 - \left(\frac{D}{\epsilon}\right)^{-D}. \tag{2.4.16}$$

Note that $h_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ since $1/r > h_{radius} \in \mathcal{H}(1/\epsilon, C(\boldsymbol{K}))$. Adding these two terms together and applying the triangle inequality gives the result.

For the second part, note that

$$\mathbb{E}_{x_0,\boldsymbol{w}}[\widehat{\nabla}_t] = \widetilde{\nabla}_t. \tag{2.4.17}$$

By Lemma 2.4.10,

$$\left[\sum_{t=0}^{T-1}\left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right]$$

is sub-exponential with parameters (ν^2, α). Therefore,

$$Z_i := \left(\frac{D}{r^2}\left[\sum_{t=0}^{T-1}\left((x_t^i)^\top Q_t x_t^i + (u_t^i)^\top R_t u_t^i\right) + (x_T^i)^\top Q_T x_T^i\right] U_t^i\right)$$

is sub-exponential with parameters $(\widetilde{\nu}^2, \widetilde{\alpha}) := \left(\frac{D}{r^2}\nu^2, \alpha\right)$. Then by Operator-Berinstein inequality [36, Theorem 12],

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z_1]\right\|_F \leq t\right] \geq 1 - 2D\exp\left(-m\frac{t^2}{2\widetilde{\nu}^2}\right),$$

when $t \leq \frac{\widetilde{\nu}^2}{\widetilde{\alpha}}$. That is, there exists a polynomial $h_{sample,2} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$ where

$$h_{sample,2} := h_{sample,2}\Big(D, \frac{1}{\epsilon}, \frac{1}{r}, \sigma_0, \sigma_w, \|\widetilde{W}_0\|, \|\widetilde{W}\|, C(\boldsymbol{K}), \frac{1}{\underline{\sigma}_{\boldsymbol{X}}}\Big) = \Theta\Big(D\left(\frac{\widetilde{\nu}}{\epsilon}\right)^2 \log(D/\epsilon)\Big),$$

such that when $m \geq h_{sample,2}$,

$$\mathbb{P}\left[\left\|\widehat{\nabla}_t - \widetilde{\nabla}_t\right\|_F \leq \frac{\epsilon}{2}\right] \geq 1 - \left(\frac{D}{\epsilon}\right)^{-D}. \tag{2.4.18}$$

Combining (2.4.18) with (2.4.15) and (2.4.16), we arrive at the desired result. □

2.4.3 Proof of Theorem 2.4.4

With the results in Section 2.4.1 and Section 2.4.2, now we are ready to prove the main theorem.

Proof of Theorem 2.4.4. By Lemma 2.3.15 and by choosing $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ such that the step size condition (2.3.23) is satisfied,

$$C(\boldsymbol{K}') - C(\boldsymbol{K}^*) \leq \Big(1 - 2\eta\, \underline{\sigma}_{\boldsymbol{R}}\, \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big).$$

Recall the definition of $\widehat{\nabla}_t$ in (2.4.13) and let $K_t'' = K_t - \eta\widehat{\nabla}_t$ be the iterate that uses the approximate gradient. We will show later that given enough samples, the gradient can be estimated with enough accuracy that makes sure

$$|C(\boldsymbol{K}'') - C(\boldsymbol{K}')| \leq \eta\, \underline{\sigma}_{\boldsymbol{R}}\, \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\epsilon. \tag{2.4.19}$$

That means as long as $C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \geq \epsilon$, we have

$$C(\boldsymbol{K}'') - C(\boldsymbol{K}^*) \leq \Big(1 - \eta\, \underline{\sigma}_{\boldsymbol{R}}\, \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big).$$

Then the same proof as that of Theorem 2.3.3 gives the convergence guarantee.

Now let us prove (2.4.19). First note that $C(\boldsymbol{K}'') - C(\boldsymbol{K}')$ is bounded. By Lemma 2.4.7, if $\|K_t'' - K_t'\| \leq \eta\, \underline{\sigma}_{\boldsymbol{R}}\, \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{\|\Sigma_{\boldsymbol{K}^*}\|} \cdot \epsilon/(T \cdot h_{cost})$, where $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ is the polynomial in Lemma 2.4.7, then (2.4.19) holds. To get this bound, recall $K_t' = K_t - \eta\nabla_t C(\boldsymbol{K})$ in (2.3.22) and writing $\nabla_t = \nabla_t C(\boldsymbol{K})$ for ease of exposition, observe that $K_t'' - K_t' = \eta(\nabla_t - \widehat{\nabla}_t)$, therefore it suffices to make sure

$$\|\nabla_t - \widehat{\nabla}_t\| \leq \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}}{T\|\Sigma_{\boldsymbol{K}^*}\| h_{cost}}\epsilon.$$

By Lemma 2.4.11, it is enough to pick $\overline{h}_{radius} = h_{radius}(3T\|\Sigma_{\boldsymbol{K}^*}\| h_{cost}(C(\boldsymbol{K}))/(2\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}\,\epsilon), C(\boldsymbol{K})) \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}))$, and

$$\overline{h}_{sample} = h_{sample}\Big(\frac{3h_{cost}(C(\boldsymbol{K}))\|\Sigma_{\boldsymbol{K}^*}\|}{2\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}\,\epsilon}, C(\boldsymbol{K})\Big) + h_{sample,2}\Big(\frac{3h_{cost}(C(\boldsymbol{K}))\|\Sigma_{\boldsymbol{K}^*}\|}{2\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}\,\epsilon}, C(\boldsymbol{K})\Big).$$

This gives the desired upper bound on $\|\nabla_t - \widehat{\nabla}_t\|$ with high probability (at least $1 - 2(\epsilon/D)^D$).

Since the number of steps is a polynomial, we have $TN = o(\epsilon^D)$. By the union

bound with probability at least

$$\left(1-2\left(\frac{\epsilon}{D}\right)^{D}\right)^{TN} \geq 1-2TN\left(\frac{\epsilon}{D}\right)^{D} \geq 1-\exp(-D),$$

we have $\|\nabla_t - \widehat{\nabla}_t\| \leq \frac{\underline{\sigma_{\boldsymbol{X}}}^2 \underline{\sigma_{\boldsymbol{R}}}}{T\|\Sigma_{\boldsymbol{K}^*}\|h_{cost}}\epsilon$, $\forall\, t = 0, 1, \cdots, T-1$. Therefore,

$$C(\boldsymbol{K}'') - C(\boldsymbol{K}^*) \leq \left(1-\eta\,\underline{\sigma_{\boldsymbol{R}}}\,\frac{\underline{\sigma_{\boldsymbol{X}}}^2}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\Big(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\Big). \tag{2.4.20}$$

This implies $C(\boldsymbol{K}'') < C(\boldsymbol{K})$. To guarantee that (2.4.20) holds at each iteration $n = 1, 2, \cdots, N$, it suffices to pick $\overline{h}_{radius} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$ and $\overline{h}_{sample} \in \mathcal{H}(1/\epsilon, \mathrm{C}(\boldsymbol{K}^0))$. The rest of the proof is the same as that of Theorem 2.3.3. Note again that in the smoothing, because the function value is monotonically decreasing, and by the choice of radius, all the function values encountered are bounded by $2C(\boldsymbol{K}^0)$, so the polynomials are indeed bounded throughout the algorithm. □

2.4.4 Discussion

Remark 2.4.12 (Comparison with [29])**.** The proofs of our main results, Theorems 2.3.3 and 2.4.4, are different to those from [29]. Firstly, to prove the gradient dominant condition, [29] only required conditions on the distribution of the initial position. However, we need conditions to guarantee the non-degeneracy of the $\mathbb{E}[x_t x_t^\top]$ matrix at any time. Secondly, the extra randomness from the sub-Gaussian noise needs to be taken care of in the perturbation analysis of $\Sigma_{\boldsymbol{K}}$. Finally, we need more advanced concentration inequalities to provide the number of samples and number of simulation trajectories that leads to the theoretical guarantee in the case with unknown parameters.

Remark 2.4.13 (Non-stationary Dynamics)**.** Note that our framework can be generalized to non-stationary dynamics, that is, for $t = 0, 1, \cdots, T-1$,

$$x_{t+1} = A_t x_t + B_t u_t + w_t,\ x_0 \sim \mathcal{D}. \tag{2.4.21}$$

with $\{A_t\}_{t=0}^{T-1}$ and $\{B_t\}_{t=0}^{T-1}$ time-dependent state parameters.

Remark 2.4.14 (Other Policy Gradient Methods)**.** Our convergence and sample complexity analysis could be applied to other policy gradient methods, including the Nat-

ural policy gradient method and the Gauss-Newton method, in the framework of the LQR with stochastic dynamics and finite horizons.

2.5 Numerical Experiments

The performance of the PPG algorithm (2.4.4) is demonstrated for the optimal liquidation problem with single asset and the empirical analysis of the policy gradient method (2.4.1) in higher dimensions is also provided with synthetic data. We will specifically focus on the following questions.

- In practice, how fast do the policy gradient algorithm and the PPG algorithm with known and unknown parameters converge to the true solution?
- How does the deadline (the finite horizon) influence the optimal policy?
- When the real-word system does not exactly follow the LQR framework, does the policy-gradient method outperform mis-specified LQR models?

This section is organized as follows. We demonstrate the performance of the PPG algorithms for the optimal liquidation problem with a single asset in the LQR framework in Section 2.5.1. We then show that without the LQR model specification, the learned policy from the policy gradient algorithm improves the Almgren-Chriss solution in Section 2.5.2. Finally, we test the performance of the algorithm with unknown parameters in high dimensions in Section 2.5.3.

Note the policy gradient method outperforms the Q-learning algorithm, a popular model-free method, in terms of both sample complexity and accuracy in our setting. An illustration in a one-dimensional example can be found in the Appendix A.2.

2.5.1 Optimal Liquidation within the LQR Framework

Recall the set up of the optimal liquidation problem in (2.2.1). By convention, we write the control in the feedback form as $u_t = -K_t x_t$. Writing $K_t = (k_t^1, k_t^2)$, we have $u_t = -k_t^1 S_t - k_t^2 q_t$, the state equation becomes

$$x_{t+1} = \begin{pmatrix} 1 + \gamma k_t^1 & \gamma k_t^2 \\ k_t^1 & 1 + k_t^2 \end{pmatrix} x_t + w_t.$$

In the liquidation problem, we assume $u_t \geq 0$ $(0 \leq t \leq T-1)$. That is, $k_t^1 \leq 0$ and $k_t^2 \leq 0$ $(0 \leq t \leq T-1)$.

Assumption 2.5.1 (Assumptions for the Optimal Liquidation Problems)**.** We assume

(1) $\gamma k_t^1 + k_t^2 > -1$ $(0 \leq t \leq T-1)$;

(2) $\beta > \frac{\gamma}{2}$.

Justification of the Assumption. Assumption 2.5.1-(1) is essential to ensure that the liquidation problem is well defined. First, $\gamma k_t^1 > -1$ makes sure that the stock price process $\{S_t\}_{t=0}^T$ is well-behaved:

$$\mathbb{E}[S_{t+1}] = \mathbb{E}[S_t] - \gamma \mathbb{E}[u_t] = (1 + \gamma k_t^1)\mathbb{E}[S_t] + \gamma k_t^2 q_t.$$

If $\gamma k_t^1 < -1$, then $\mathbb{E}[S_{t+1}] \leq 0$ since $k_t^2 \leq 0$. Second, $k_t^2 \geq -1$ guarantees that inventory will not be negative. Note that

$$q_{t+1} = q_t - (-k_t^1 S_t - k_t^2 q_t) = (1 + k_t^2) q_t + k_t^1 S_t.$$

If $k_t^2 \leq -1$ and $q_t > 0$, then $q_{t+1} < 0$. Assumption 2.5.1-(2) implies that the temporary market impact is "bigger" than one half of the permanent market impact, which is consistent with the empirical evidence [8] and assumptions in [7].

Learning to Liquidate. In practice, traders may not know the market impact parameter γ. But one can always take some $\bar{\gamma} > \gamma$ based on some basic understandings of the market and perform a PPG algorithm to the closed convex set $\mathcal{S}$:

$$\begin{aligned} \mathcal{S} := \{\boldsymbol{K} = (K_0, \cdots, K_{T-1}) \quad : \quad & K_t = (k_t^1, k_t^2),\ \gamma k_t^1 + k_t^2 \geq -1 + \zeta, \\ & k_t^1 \leq 0,\ k_t^2 \leq 0,\ \forall t = 0, \cdots, T-1\big\}, \end{aligned} \tag{2.5.1}$$

with some small parameter $\zeta > 0$.

In practice γ is usually on the order of $10^{-5} \sim 10^{-6}$ (See Table A.1 in Appendix A.1 based on the method in [24]) and hence a universal upper bound $\bar{\gamma}$ in (2.5.1) is not a strong assumption for a given portfolio of stocks to liquidate.

Proposition 2.5.2. *Assume $\boldsymbol{K} \in \mathcal{S}$ and Assumptions 2.2.1, 2.4.3 and 2.5.1 hold, we have $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $\{P_t^{\boldsymbol{K}}\}_{t=0}^T$ derived from* (2.3.9) *are positive definite for the optimal liquidation problem* (2.2.7) *and* (2.2.9)*.*

The proof of Proposition 2.5.2 is deferred to Appendix A.3.5. It is easy to check that the projection set $\mathcal{S}$ defined in (2.5.1) is convex and closed. Along with Proposition 2.5.2, the convergence result in Theorem 2.4.5 holds for the liquidation problem (2.2.7) and (2.2.9) as long as the conditions in Proposition 2.5.2 are satisfied.

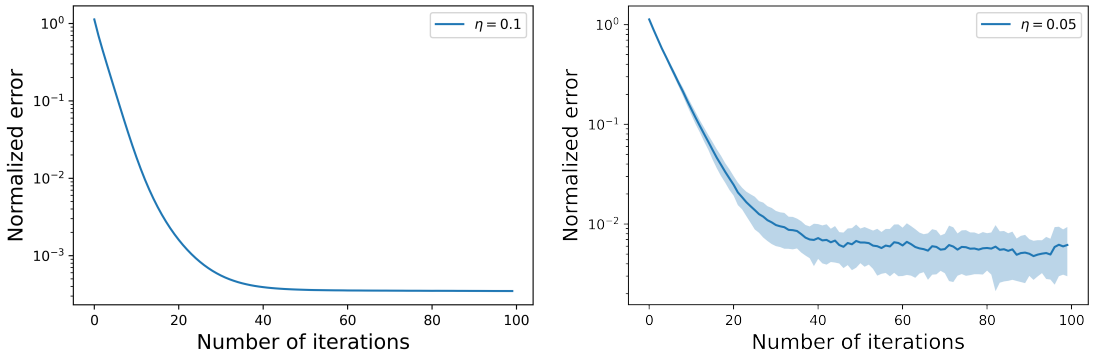
We test the performance of the PPG algorithm with projection set $\mathcal{S}$ on Apple (AAPL) and Facebook (FB) stocks. For the PPG algorithm with known parameters, the controller directly calculates the policy gradient using the expression given in Lemma (2.3.5) and update the policy using (2.3.8). For the PPG algorithm with unknown parameters, the controller estimates the policy gradient using Algorithm 1, where the market simulator gives the cost signal to the controller after each implementation of policy. The market simulator of the associated LQR framework is constructed with NASDAQ ITCH data and the details can be found in Appendix A.1. The controller then updates the policy using the rule (2.4.1).

Performance Measure. We use the following *normalized error* to quantify the performance of a given policy $\boldsymbol{K}$,

$$\text{Normalized error} = \frac{C(\boldsymbol{K}) - C(\boldsymbol{K}^*)}{C(\boldsymbol{K}^*)},$$

where $\boldsymbol{K}^*$ is the analytical solution calculated from (2.2.5) with P_t^* solved directly from the Riccati equation (2.2.3).

Set-up. (1) Parameters: $\phi = 5 \times 10^{-6}$ (for both AAPL and FB), $\epsilon = 10^{-8}$, $T = 10$; smoothing parameter $r = 0.6$, number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{1\times 2T}$ with $\{\boldsymbol{K}^0\}_{ij} = -0.2$ for all i, j, for both algorithms with known and unknown parameters; step sizes are indicated in the figures; $\gamma = 5 \times 10^{-5}$, $\zeta = 10^{-12}$ for the projection set. (2) Initialization: Assume the initial inventory q_0 follows $\mathcal{N}(500, 1)$. The small variance of the initial inventory distribution is used to guarantee the $\mathbb{E}[x_0 x_0^\top]$ is positive definite. In practice, the algorithm converges with deterministic initial inventories.



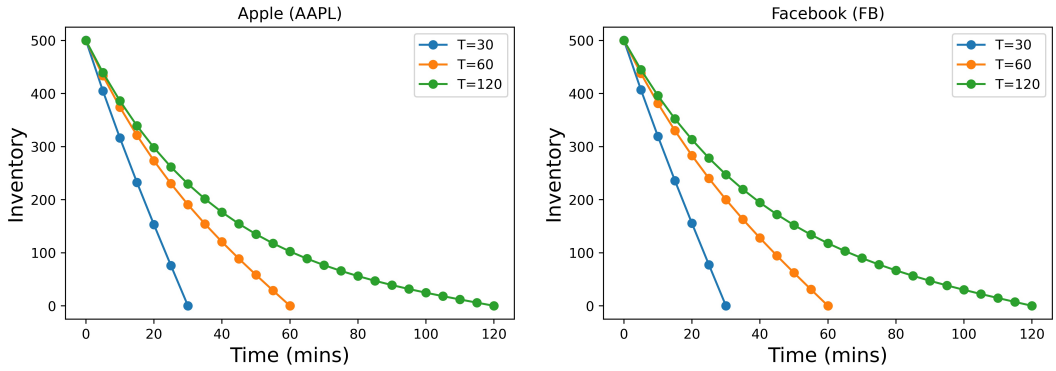
(a) PPG with known parameters. (b) PPG with unknown parameters.

Figure 2.1: Performance of the PPG algorithms (50 simulation scenarios).

Convergence. Both PPG algorithms with known parameters and unknown parameters show a reasonable level of accuracy within 50 iterations (that is the normalized error is less than 10^{-2}). The PPG algorithm with known parameters has almost no fluctuations across the 50 scenarios. By choosing $m = 200$, the performance of the PPG algorithm with unknown parameters is stable with relatively small fluctuations (see the blue area in Figure 2.1b) across the 50 scenarios.

Impact of the Deadline. The optimal policy is sensitive to the deadline in that the shapes of the optimal inventory trajectories are different with different deadlines. See Figure 2.2 for both AAPL and FB with $T = 30, 60$ and 120 minutes. The liquidation speed is almost linear when T is small; and it is faster in the initial trading phase and slower at the end when T is relatively large.

Impact of the Parameter ϕ. Recall that in (2.2.9) the parameter ϕ is used to balance the expected terminal wealth $\mathbb{E}[C]$ and the variance of the terminal wealth $\text{var}[C]$. To show the impact of ϕ, we set ϕ to be 10^{-4}, 10^{-5}, 10^{-6}, and 10^{-7} and show the corresponding inventory trajectories in Figure 2.3. The optimal liquidation speed is almost linear when ϕ is small, while it is faster in the initial trading phase and slower at the end when ϕ is relatively large.



(a) AAPL. (b) FB.

Figure 2.2: Optimal inventory trajectory under different deadlines (200 simulation scenarios).

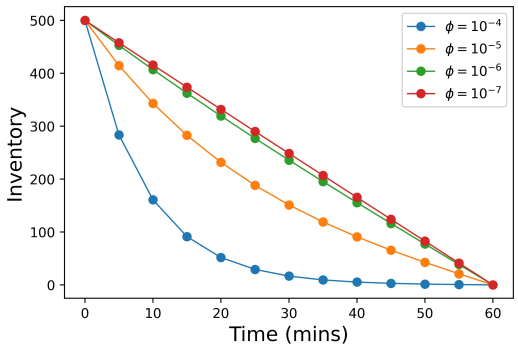
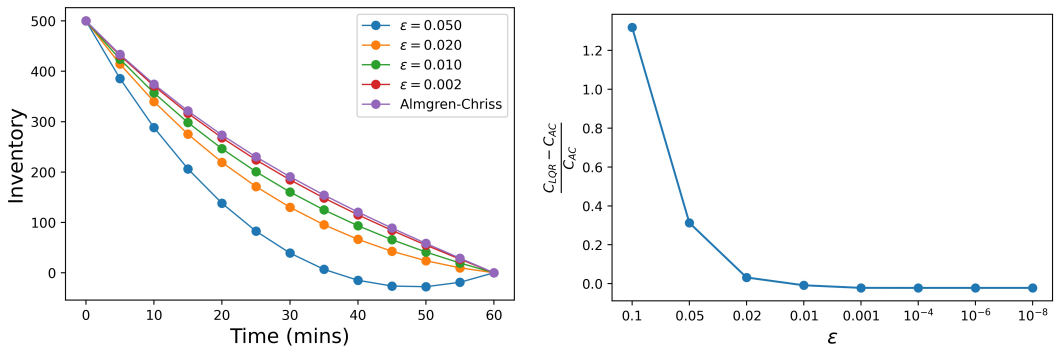


Figure 2.3: Inventory trajectories of AAPL under different ϕ (average across 200 simulation scenarios).



(a) Inventory trajectories. (b) Relative cost difference.

Figure 2.4: Original Almgren-Chriss framework versus LQR formulation under different ϵ (AAPL).

Impact of the Parameter ϵ. Recall that our liquidation formulation (2.2.9) differs from the Almgren-Chriss formulation (2.2.8) by an additional regularization term $\sum_{t=0}^{T} \epsilon S_t^2$. The role of this term is to enable the problem to be cast in the LQR framework and to guarantee the well-definedness of the Ricatti equation. From Figure 2.4a, the optimal policies and inventory trajectories are close to the Almgren-Chriss solution when $\epsilon \leq 0.01$. However, when $\epsilon = 0.05$, the optimal policy is far away from the Almgren-Chriss solution. We show the difference between C_{AC}, defined in (2.2.8), and $C_{\mathrm{LQR}}(\epsilon)$, defined in (2.2.9), in Figure 2.4b. We see that $C_{\mathrm{LQR}}(\epsilon)$ is close to C_{AC} when $\epsilon < 0.02$ and is markedly different from C_{AC} when $\epsilon \geq 0.02$. It is worth noticing that when $\epsilon = 0$, the algorithm does converge to the Almgren-Chriss solution in our setting although the convergence of the algorithm in this case is not guaranteed by our theoretical results.

2.5.2 Learning to Liquidate without Model Specification

In practice, the dynamics of the trading system may not be exactly those assumed in the LQR framework but we might expect that the policy gradient method could still perform well when the system is "nearly" LQ as the execution of the policy gradient method does not rely on the model specification. In this section, we consider liquidation problems in the Limit Order Book (LOB) setting. A LOB is a list of orders that a trading venue, for example the NASDAQ exchange, uses to record the interest of buyers and sellers in a particular financial instrument. There are two types of orders the buyers (sellers) can submit: a limit buy (sell) order with a preferred price for a given volume or a market buy (sell) order with a given volume which will be immediately executed with the best available limit sell (buy) orders. Here we perform the policy gradient method with unknown parameters to learn the optimal strategies to liquidate using market orders in the LOB.

We denote by S_t the mid-price of the asset at time t, that is the average of the best-bid price and best-ask price. At each time t, the decision is to liquidate an amount u_t of the asset. The action u_t will have an impact on the market, with possibly both temporary and permanent impacts. Unlike the LQR framework or the classical Almgren-Chriss model, where dynamics are assumed to follow some stochastic model, here we run the policy gradient method directly on the LOB without any assumption on how the mid-price S_t moves and what are the forms of the market impacts. Denote by $q_t = q_{t-1} - u_{t-1}$ the inventory at time t. We restrict the admissible controls to be of the linear feedback form $u_t = -K_t(S_t, q_t)^\top$ with some $K_t \in \mathbb{R}^{1\times 2}$.

The cost $c_t = \phi'(q_t - u_t)^2 - r_t(u_t)$ at time t consists of two parts. The first part $\phi'(q_t - u_t)^2$ is the holding cost of the inventory weighted by a parameter ϕ'. The quantity $r_t(u_t)$ is the amount we receive by liquidating u_t shares at time t. Note that $r_t(\cdot)$ may depend on S_t and other market observables. For example, if we liquidate $u_t = 1000$ shares of the asset with the market conditions given in Table 2.1, then the amount received would be

$$r_t(u_t) = 397 \times 200.1 + 412 \times 200.0 + (1000 - 397 - 412) \times 199.9 = 200020.6.$$

This transaction moves the best bid price two levels down. This is commonly referred to as the *temporary impact* of a market order.

Bid level	One	Two	Three	Four	Five
Bid price (USD)	200.1	200.0	199.9	199.8	199.7
Volume available	397	412	502	442	529

Table 2.1: One snapshot of the LOB.

Performance Metric: Implementation Shortfall [68].

$$\mathrm{IS}(\boldsymbol{u}) = \left(\sum_{t=0}^{T-1} c_t(u_t) + c_T \left(q_0 - \sum_{t=0}^{T-1} u_t \right) \right) - c_0(q_0). \tag{2.5.2}$$

The first term of (2.5.2) is the cost of implementing policy $\boldsymbol{u}$ over the horizon $[0, T]$. The second term is the cost when liquidating q_0 market orders at time 0. If we expect $\boldsymbol{u}$ is better than liquidating everything at time 0, then $\mathrm{IS}(\boldsymbol{u}) < 0$. A smaller implementation shortfall implies the strategy is more profitable.

We use the following *relative performance* (evaluated on a single trajectory) to compare the performance of two policies $\boldsymbol{u}^1$ and $\boldsymbol{u}^2$,

$$\text{Relative performance} = \frac{\mathrm{IS}(\boldsymbol{u}^2) - \mathrm{IS}(\boldsymbol{u}^1)}{|\mathrm{IS}(\boldsymbol{u}^2)|}.$$

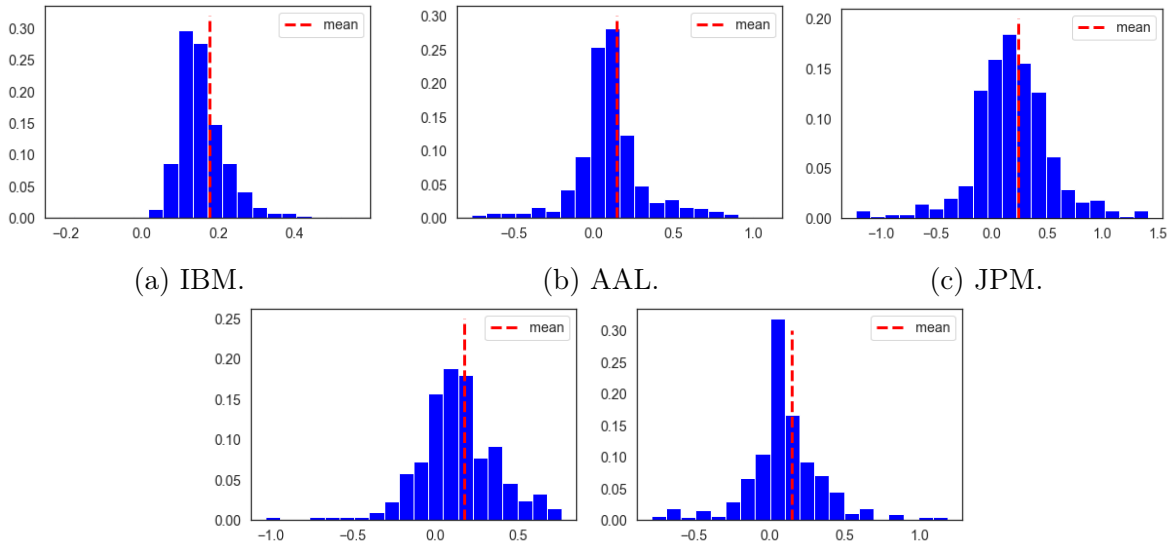
Experiment Set-up. We consider the LOB data consisting of the best 5 levels and we assume the trading frequency $\Delta = 1$ minute and the trading horizon $T = 10$ minutes. We perform a numerical analysis for five different stocks, Apple (AAPL), Facebook (FB), International Business Machines Corporation (IBM), American Airlines (AAL) and JP Morgan (JPM), during the period from 01/01/2019 to 12/31/2019.

The data is divided into two sets, a training set with data between 10:00AM-12:00AM 01/01/2019-08/31/2019 and a test set with data between 10:00AM-12:00AM 09/01/2019-12/31/2019.

We take $\phi' = 5 \times 10^{-6}$; $T = 10$; smoothing parameter $r = 0.4$; number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{1\times 20}$ with $(\boldsymbol{K}^0)_{ij} = -0.2$ for all i, j; and step size $\eta = 10^{-6}$. We assume the initial inventory follows $q_0 = 2000$. We compare the performance of the policy gradient method with the Almgren-Chriss solution with fitted parameters given in Table A.1 in the Appendix. In the Almgren-Chriss model, we set $\phi = \sigma^2\phi'$ to ensure a reasonable comparison.

Results. From Table 2.2 and Figure 2.5, the policy gradient method improves on the Almgren-Chriss solution by around 20% on five different stocks from different financial sectors. Note that the goal of the policy gradient method is to learn the global minimizer of the expected cost function, hence it is expected that the Almgren-Chriss solution could perform better than the policy gradient method for some sample trajectories, as shown in Figure 2.5. This result is compatible with the performance of the Q-learning algorithms [42]. The drawback of Q-learning algorithms is that the computational complexity is highly dependent on the size of the set of (discrete) states and actions, where as the policy gradient method can handle continuous states and actions.

We conjecture that the policy gradient method may be capable of learning the global "optimal" solution for a larger class of models that are "similar" to the LQR framework with stochastic dynamics and finite time horizon. In addition, as the policy gradient method is a model-free algorithm, it is more robust with respect to model mis-specification as compared to the Almgren-Chriss framework.



(a) IBM. (b) AAL. (c) JPM.

(d) FB. (e) AAPL.

Figure 2.5: Empirical distribution of the relative performance on the test set.

Asset	IBM	AAL	JPM	FB	AAPL
In sample	0.173	0.152	0.251	0.181	0.165
(std)	(0.09)	(0.27)	(0.31)	(0.32)	(0.31)
Out of sample	0.178	0.146	0.245	0.175	0.163
(std)	(0.08)	(0.29)	(0.36)	(0.24)	(0.37)

Table 2.2: Average relative performance of the policy gradient ($\boldsymbol{u}^1$) compared to Almgren-Chriss solution ($\boldsymbol{u}^2$).

2.5.3 Learning LQR in Higher Dimensions

In practice we can perform the policy gradient method for the optimal liquidation problem with multiple assets. However it is difficult to capture the cross impact and permanent impact with historical LOB data. Therefore we test the performance of the policy gradient method in higher dimensions on synthetic data consisting of a four-dimensional state variable and a two-dimensional control variable. The parameters are randomly picked such that the conditions for our LQR framework are satisfied.

Set-up. (1) Parameters:

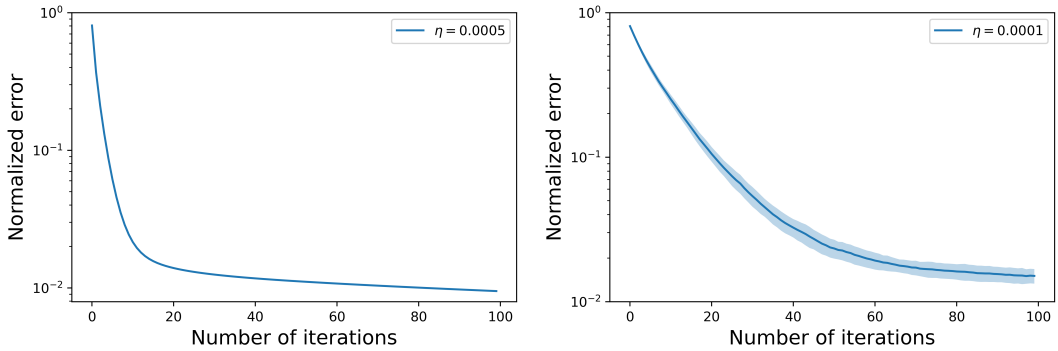
$$A = \begin{pmatrix} 0.5 & 0.05 & 0.1 & 0.2 \\ 0 & 0.2 & 0.3 & 0.1 \\ 0.06 & 0.1 & 0.2 & 0.4 \\ 0.05 & 0.2 & 0.15 & 0.1 \end{pmatrix}, \; B = \begin{pmatrix} -0.05 & -0.01 \\ -0.005 & -0.01 \\ -1 & -0.01 \\ -0.01 & -0.9 \end{pmatrix}, R_t = \begin{pmatrix} 0.4 & -0.25 \\ -0.25 & 0.7 \end{pmatrix},$$

$$Q_t = \begin{pmatrix} 1 & 0.2 & -0.005 & 0.015 \\ 0.2 & 1.1 & 0.15 & 0 \\ -0.05 & 0.15 & 0.9 & -0.08 \\ 0.015 & 0 & -0.08 & 0.88 \end{pmatrix}, \; W = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.3 \end{pmatrix},$$

$Q_T = Q_t$, $T = 10$; smoothing parameter $r = 1$, number of trajectories $m = 200$; initial policy $\boldsymbol{K}^0 \in \mathbb{R}^{2\times 40}$ with $\{\boldsymbol{K}^0\}_{ij} = 0.05$ for all i, j, for both known and unknown parameters;

(2) Initialization: We assume $x_0 = (x_0^1, x_0^2, x_0^3, x_0^4)^\top$ and x_0^i are independent. x_0^1, x_0^2, x_0^3, and x_0^4 are sampled from $\mathcal{N}(5, 0.1)$, $\mathcal{N}(2, 0.3)$, $\mathcal{N}(8, 1)$, $\mathcal{N}(5, 0.5)$.

Convergence. For the high-dimensional case, the normalized error falls below the threshold 10^{-2} within 80 iterations for the policy gradient algorithm with known parameters. It takes substantially more iterations for the policy gradient algorithm with unknown parameters to have an error near such a threshold, which is as expected.



(a) Known parameters ($\eta = 0.0005$).

(b) Unknown parameters ($\eta = 0.0001$).

Figure 2.6: Performance of the policy gradient algorithms (50 simulation scenarios)

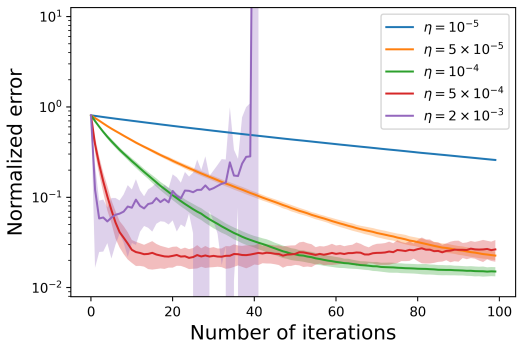


Figure 2.7: Performance of the policy gradient algorithm with unknown parameters under different step size η (50 simulation scenarios).

Outcomes from Varying the Parameter η. The performance of the policy gradient algorithm also depends on the values of the step size η. We show how the values of the step size $\eta \in [10^{-5}, 2 \times 10^{-3}]$ affect the convergence of the policy gradient algorithm with unknown parameters in Figure 2.7. A tiny step size leads to slow convergence (see the blue line when $\eta = 10^{-5}$) and a larger step size may cause divergence (see the purple line when $\eta = 2 \times 10^{-3}$).

Chapter 3

Policy Gradient Methods for the Linear Quadratic Game

3.1 Introduction

In Chapter 2 we have discussed policy gradient methods for single-agent LQR problem. In multi-agent ($N \geq 2$) decision-making problems, policy gradient methods and their variants have also achieved substantial empirical successes (see for example [77, 45]). However, there have been few results from a theoretical perspective showing why such a class of reinforcement learning algorithms performs well in multi-agent decision making problems. As a starting point, in this chapter we investigate LQGs which can be seen as a generalization of the LQR from a single agent to multiple agents. In an LQG, all agents jointly control a linear state process, which may be in high dimensions, where the control (or action) from each individual agent has a linear impact on the state process. Each agent optimizes a quadratic cost function which depends on the state process, the control from this agent and/or the controls from the opponents.

LQGs admit global Nash equilibria in the space of linear feedback policies and these equilibria can be found by solving a coupled set of Ricatti equations when system parameters are given. As such LQGs are a natural benchmark problem on which to test policy gradient algorithms in multi-agent settings when system parameters are unknown. Furthermore, policy gradient methods open up the possibility to develop new scalable approaches to finding or learning solutions to control problems even with constraints. Finally, the empirical results present in [57] imply that, even in this relatively straightforward LQ case (with linear dynamics, linear feedback policies, and

quadratic costs), policy gradient MARL would be unable to find the local Nash equilibrium in a non-negligible subset of problems. This further demonstrates the necessity of understanding under what circumstances policy gradient methods work for LQGs.

For LQGs with policy gradient algorithms, most of the existing literature has focused on zero-sum games with two players [20, 93, 94]. In the setting of deterministic dynamics and infinite time horizon, [93] proposed an alternating policy update scheme with a projection step and showed sublinear convergence for the algorithm. For a similar setting, [20] proposed a leader-follower type of policy gradient algorithm which is projection-free and enjoys a global sublinear convergence rate and asymptotically linear convergence rate. For the case of stochastic dynamics and finite time horizon, [94] provided the first sample complexity result for the global convergence of the policy gradient method with an alternating policy update scheme.

However, little theory has been developed for the more general class of LQGs with N players and general-sum cost functions. It has been documented that the policy gradient method may *fail to converge* in such a setting with deterministic dynamics due to the lack of theoretical guidance on how to properly choose the step size and the exploration scheme [57]. For a special class of N-player LQGs with homogeneous agents where agents interact with each other through a mean-field type of "deep state" which is the aggregated state position of all agents, [73] showed that the policy gradient method converges to the global Nash equilibrium. Up to now, as far as we are aware, providing theoretical guarantees for the convergence of the policy gradient method remains an open problem for general-sum LQGs with more than two players [58].

In this chapter, we explore the natural policy gradient method, which can be viewed as a normalized version of the vanilla policy gradient descent method, for a class of N-player general-sum LQGs. Our main result is Theorem 3.3.5 in which we provide a global linear convergence guarantee for this approach in the setting of a finite time horizon and stochastic dynamics provided there is a certain level of noise in the system. The noise can either come from the underlying dynamics or carefully designed explorations from the agents. Intuitively speaking, the noise can help the agents escape from traps such as those observed in [57] when the dynamics are deterministic. In addition, the noise can help agents find the correct descent direction for convergence to the desired Nash equilibrium point. To the best of our knowledge, this is the first result of its kind in the MARL literature showing a provable convergence result for N-player general-sum LQGs. From a technical perspective, the main difficulty is to quantify the perturbation of the individual's gradient term (see equation (3.3.49)) and

to control the descent of the individual's cost function (see equation (3.3.51)) in the presence of competition from the other $N-1$ agents.

We illustrate the performance of our algorithm with two examples. We first perform the natural policy gradient algorithm under the experimental set-up in [57] over a time finite horizon. The algorithm converges to the Nash equilibrium for appropriate initial policies and step sizes. The second example is a toy LQG example with synthetic data. The empirical results suggest that, in practice, the natural policy gradient algorithm can find the Nash equilibrium even if the level of system noise is lower than that required for our theoretical analysis.

3.1.1 Comparison to the Literature on General-sum LQGs.

With deterministic dynamics and infinite horizon, [57] provided some empirical examples where the policy gradient method fails to converge to the set of Nash equilibria (which may not be unique). These empirical examples motivate an examination of the possibility of applying policy gradient methods in the multi-agent environment. Here we explain the difference between our framework and the set-up in [57]. In addition, we offer some explanations for why the policy gradient method works in our framework whereas it fails to converge in [57].

- **Well-definedness of LQGs:** The existence and uniqueness of a Nash equilibrium is a prerequisite for the convergence of learning algorithms. In the setting of stochastic dynamics and finite horizon, the general-sum LQG has a unique Nash equilibrium solution under mild conditions [13]. For the setting with an infinite horizon, the existence of Nash equilibria can be proved under some stabilizability and detectability properties. However obtaining explicit and verifiable model conditions for stabilizability and detectability seems to be a quite challenging task let alone finding conditions for the uniqueness of the equilibrium [13]. It is not clear if there exists a unique Nash equilibrium for the setting considered in [57].

- **Self-exploration property of time-dependent policies:** For single-agent LQR problems, [14] highlighted that the time-dependent optimal feedback policy enjoys a self-exploration property in the finite time horizon setting. Namely, the time-dependent optimal feedback matrices ensure that the optimal state and control processes span the entire parameter space, which enables the design of efficient exploration-free learning algorithms. By contrast the optimal feedback

policy is time-invariant in the infinite time horizon setting and learning algorithms tend to have difficulty converging without efficient exploration schemes [56]. We believe a similar analogy holds for the game setting as well.

- **System noise:** In our setting we need Assumption 3.3.3 which implies that a certain level of system noise is essential for the convergence of the policy gradient method. We show in Section 3.5 that the circulating and divergence phenomenon described in [57] can be avoided with the addition of system noise. The noise can either come from the original system when the dynamics are stochastic (as suggested in Assumption 3.3.3) or agents can apply Gaussian exploration (as suggested in [56] for single-agent LQR problems with infinite time horizon).

In addition, [73] showed the global convergence of the policy gradient method for a mean-field type of LQG in the setting of infinite horizon and stochastic dynamics. In particular, agents are assumed to be homogeneous and are only able to interact through an aggregated state and action pair. With this special formulation, the uniqueness of the Nash equilibrium could be established (see [73, Theorem 1]) and the proof of convergence could be reduced to the single agent case (see [73, Theorem 2]). In this chapter, we focus on a more general LQG with no homogeneity assumption nor any restriction on the interactions.

The rest of the chapter is organized as follows. We introduce the mathematical framework and problem set-up in Section 3.2. The convergence analysis of the natural policy gradient method for the case of known model parameters is provided in Section 3.3. When parameters are unknown, the sample-based natural policy gradient method is discussed in Section 3.4. Finally, the algorithm is applied to two numerical examples in Section 3.5.

3.2 Problem Set-up

We consider the following N-player general-sum linear quadratic game (LQG) over a finite time horizon T. The state process evolves as

$$x_{t+1} = A_t x_t + \sum_{i=1}^{N} B_t^i u_t^i + w_t,\ t = 0, 1, \cdots, T-1, \tag{3.2.1}$$

where $x_t \in \mathbb{R}^d$ is the state of the system with the initial state x_0 drawn from a Gaussian distribution, $u_t^i \in \mathbb{R}^{k_i}$ is the control of player i at time t which is assumed to be Markovian and $\{w_t\}_{t=0}^{T-1}$ are zero-mean i.i.d. Gaussian random variables which are independent of x_0. The system parameters $A_t \in \mathbb{R}^{d\times d}$, $B_t^i \in \mathbb{R}^{d\times k_i}$, for $t = 0, 1, \cdots, T-1$ are referred to as system (transition) matrices. The objective of player i ($i = 1, \cdots, N$) is to minimise their finite time horizon value function:

$$\inf_{\{u_t^i\}_{t=0}^{T-1}} C^i(\boldsymbol{u}) := \inf_{\{u_t^i\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right], \tag{3.2.2}$$

where the cost function

$$c_t^i(x_t, u_t^i) = x_t^\top Q_t^i x_t + (u_t^i)^\top R_t^i u_t^i, \quad t = 0, 1, \cdots, T-1, \tag{3.2.3}$$

with $c_T^i(x_T) = x_T^\top Q_T^i x_T$, where $Q_t^i \in \mathbb{R}^{d\times d}$ and $R_t^i \in \mathbb{R}^{k_i\times k_i}$ ($i = 1, \cdots, N$) are matrices that parameterize the quadratic costs. Here we denote by $\boldsymbol{u} := (\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^N)$ a strategy profile with $\boldsymbol{u}^i := (u_0^i, \cdots, u_{T-1}^i)$ ($i = 1, \ldots, N$). Note that the expectation in (3.2.2) is taken with respect to both the initial state x_0 and the noise $\{w_t\}_{t=0}^{T-1}$. We also denote by $\boldsymbol{x} := (x_0, \cdots, x_T)$, $\boldsymbol{Q}^i := (Q_0^i, \cdots, Q_T^i)$, and $\boldsymbol{R}^i := (R_0^i, \cdots, R_{T-1}^i)$, for $i = 1, \cdots, N$.

Assumption 3.2.1 (Cost Parameter)**.** *Assume for $i = 1, \cdots, N$, $Q_t^i \in \mathbb{R}^{d\times d}$, for $t = 0, 1, \cdots, T$, and $R_t^i \in \mathbb{R}^{k_i\times k_i}$, for $t = 0, 1, \cdots, T-1$ are symmetric positive definite matrices.*

Assumption 3.2.2 (Initial State and Noise Process)**.** *Assume*

1. *Initial state: x_0 is Gaussian such that $\mathbb{E}[x_0 x_0^\top]$ is positive definite.*
2. *Noise: $\{w_t\}_{t=0}^{T-1}$ are i.i.d. Gaussian and independent from x_0 such that $\mathbb{E}[w_t] = 0$, and $W = \mathbb{E}[w_t w_t^\top]$ is positive definite, $\forall t = 0, 1, \cdots, T-1$.*

In N-player general-sum LQGs, we adopt the most popular notion, known as the Nash equilibrium, to define the optimal solution of the game. We give the formal definition of a Nash equilibrium as follows.

Definition 3.2.3 (Nash equilibrium)**.** *A strategy profile $\boldsymbol{u}^* = (\boldsymbol{u}^{1*}, \boldsymbol{u}^{2*}, \ldots, \boldsymbol{u}^{N*})$ with $\boldsymbol{u}^{i*} = \{u_t^{i*}\}_{t=0}^{T-1}$ ($i = 1, \ldots, N$) is a Nash equilibrium for the LQG given by* (3.2.1)-(3.2.3) *if for each player i and for any admissible strategy $\boldsymbol{u}^i$ of the player i, it holds*

that

$$C^i(\boldsymbol{u}^i, \boldsymbol{u}^{-i*}) \geq C^i(\boldsymbol{u}^{i*}, \boldsymbol{u}^{-i*}),$$

where $\boldsymbol{u}^{-i*}$ *denotes the Nash equilibrium strategies of all the* $N-1$ *players except* i*.*

Assumption 3.2.4 (Existence and Uniqueness of Solution)**.** *Assume there exists a unique solution* $\{K_t^{i*}\}_{t=0}^{T-1}$*, for* $i = 1, \cdots, N$ *to the following set of linear matrix equations:*

$$K_t^{i*} = \left(R_t^i + (B_t^i)^\top P_{t+1}^{i*} B_t^i\right)^{-1} (B_t^i)^\top P_{t+1}^{i*} \left(A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*}\right), \tag{3.2.4}$$

where $\{P_t^{i*}\}_{t=0}^{T}$ *are obtained recursively backwards from*

$$P_t^{i*} = Q_t^i + (K_t^{i*})^\top R_t^i K_t^{i*} + \left(A_t - \sum_{j=1}^{N} B_t^j K_t^{j*}\right)^\top P_{t+1}^{i*} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^{j*}\right), \tag{3.2.5}$$

with terminal condition $P_T^{i*} = Q_T^i$*.*

A similar assumption is adopted in [93] for a two-player zero-sum LQ game and in [73] for a homogeneous N-player game with mean-field interaction.

Remark 3.2.5. A sufficient condition for the unique solvability of (3.2.4) is the invertibility of the block matrix Φ_t, $t = 0, 1, \cdots, T-1$, with the ii-th block given by $R_t^i + (B_t^i)^\top P_{t+1}^{i*} B_t^i$ and the ij-th block given by $(B_t^i)^\top P_{t+1}^{i*} B_t^j$, where $i, j = 1, \cdots, N$ and $j \neq i$. See Remark 6.5 in [13].

Lemma 3.2.6 (Nash Equilibrium and Equilibrium Cost [13, Corollary 6.4])**.** *Assume Assumptions 3.2.1, 3.2.2, and 3.2.4 hold. Then for* $i = 1, 2, \ldots, N$*,*

1. *The Nash equilibrium strategy for player* i *is given by*

$$u_t^{i*} = -K_t^{i*} x_t, \quad t = 0, 1, \cdots T-1, \tag{3.2.6}$$

   where K_t^{i*} *is defined in* (3.2.4)*.*

2. *The Nash equilibrium cost for player* i *is*

$$\mathbb{E}[x_0^\top P_0^{i*} x_0 + N_0^{i*}] = \inf_{\{u_t^i\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right], \tag{3.2.7}$$

where $\{P_t^{i*}\}_{t=0}^T$ *are defined in* (3.2.5) *and*

$$N_t^{i*} = N_{t+1}^{i*} + \mathbb{E}[w_t^\top P_{t+1}^{i*} w_t] = N_{t+1}^{i*} + \mathrm{Tr}(W P_{t+1}^{i*}), \quad t = 0, 1, \cdots, T-1 \quad (3.2.8)$$

with terminal condition $N_T^{i*} = 0$.

To find the Nash equilibrium strategy in the linear feedback form (3.2.6), we only need to focus on the following class of linear admissible policies in feedback form

$$u_t^i = -K_t^i x_t, \qquad t = 0, 1, \cdots, T-1, \quad i = 1, \cdots, N$$

which can be fully characterized by $\boldsymbol{K}^i := (K_0^i, K_1^i, \cdots, K_{T-1}^i)$. We write $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$ for a collection of policies and

$$\boldsymbol{K}^* = (\boldsymbol{K}^{1*}, \cdots, \boldsymbol{K}^{N*}).$$

for the collection of optimal policies. We will use the notation

$$(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) = (\boldsymbol{K}^{1*}, \cdots, \boldsymbol{K}^{(i-1)*}, \boldsymbol{K}^i, \boldsymbol{K}^{(i+1)*}, \cdots, \boldsymbol{K}^{N*}),$$

for $i = 1, \cdots, N$ for player i's policy, when all other players use their optimal policies.

3.3 The Natural Gradient Method with Known Parameters

In this section, we provide a global linear convergence guarantee for the natural policy gradient method applied to the LQG (3.2.1) - (3.2.2). Throughout this section we assume all the parameters in the LQG, $\{A_t\}_{t=0}^{T-1}$, $\{B_t^i\}_{t=0}^{T-1}$, $\{Q_t^i\}_{t=0}^T$, and $\{R_t^i\}_{t=0}^{T-1}$ $(i = 1, \cdots, N)$, are known. The analysis with known parameters paves the way for learning LQGs with unknown parameters which is discussed in Section 3.4.

For LQGs, any (admissible) feedback policy can be fully characterized by a set of parameters $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$. Therefore, we can correspondingly define player i's

cost induced by the joint policy $\boldsymbol{K}$ as

$$C^i(\boldsymbol{K}) = \mathbb{E}\left[\sum_{t=0}^{T-1}\left(x_t^\top Q_t^i x_t + (K_t^i x_t)^\top R_t^i (K_t^i x_t)\right) + x_T^\top Q_T^i x_T\right],$$

where $\{x_t\}_{t=0}^T$ is the random path from the dynamics (3.2.1) induced by $\boldsymbol{K}$ starting with x_0.

Similarly to Chapter 2, we define the matrix $\Sigma_t^{\boldsymbol{K}}$, and let $\Sigma_{\boldsymbol{K}}$ be the sum of $\Sigma_t^{\boldsymbol{K}}$:

$$\Sigma_t^{\boldsymbol{K}} = \mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top], \quad \Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T} \Sigma_t^{\boldsymbol{K}}, \tag{3.3.1}$$

where $\{x_t^{\boldsymbol{K}}\}_{t=0}^T$ is a state trajectory generated by following a set of policies $\boldsymbol{K}$. We will write $x_t = x_t^{\boldsymbol{K}}$ when no confusion may occur. Define $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}}$ to be the lower bound over all the minimum singular values of $\Sigma_t^{\boldsymbol{K}}$:

$$\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} := \min_t \sigma_{\min}(\Sigma_t^{\boldsymbol{K}}),$$

We also define $\underline{\sigma}_{\boldsymbol{X}}$ as

$$\underline{\sigma}_{\boldsymbol{X}} := \min\{\sigma_{\min}(\mathbb{E}[x_0 x_0^\top]), \sigma_{\min}(W)\}. \tag{3.3.2}$$

Similarly, we define

$$\underline{\sigma}_{\mathrm{R},i} = \min_t \sigma_{\min}(R_t^i), \quad \underline{\sigma}_{\mathrm{Q},i} = \min_t \sigma_{\min}(Q_t^i),$$

and

$$\underline{\sigma}_{\boldsymbol{R}} = \min_i\{\underline{\sigma}_{\mathrm{R},i}\}, \quad \underline{\sigma}_{\boldsymbol{Q}} = \min_i\{\underline{\sigma}_{\mathrm{Q},i}\}. \tag{3.3.3}$$

We further define γ_A, γ_B, and γ_R as

$$\gamma_A = \max_{t=0,\cdots,T-1} \|A_t\|, \quad \gamma_B = \max_i\{\max_{t=0,\cdots,T-1} \|B_t^i\|\}, \quad \gamma_R = \max_i\{\max_{t=0,\cdots,T-1} \|R_t^i\|\}. \tag{3.3.4}$$

Under Assumption 3.2.1, we have $\underline{\sigma}_{\mathrm{R},i} \geq \underline{\sigma}_{\boldsymbol{R}} > 0$ and $\underline{\sigma}_{\mathrm{Q},i} \geq \underline{\sigma}_{\boldsymbol{Q}} > 0$, for $i = 1, \ldots, N$. For the well-definedness of the $\mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top]$ matrix, we have the following result.

Lemma 3.3.1. *Assume Assumption 3.2.2 holds. Then $\mathbb{E}[x_t x_t^\top]$ is positive definite for*

$t = 0, 1, \cdots, T$ *under any set of policies* $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$. *Therefore,* $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} \geq \underline{\sigma}_{\boldsymbol{X}} > 0$.

Proof. Let $\{x_t\}_{t=0}^T$ be the state trajectory induced by an arbitrary policy set $\boldsymbol{K}$. By Assumption 3.2.1 the matrix $\mathbb{E}[x_0 x_0^\top]$ is positive definite. For $t \geq 1$, we have from (3.2.1) and taking expectations

$$\mathbb{E}[x_t x_t^\top] = \left(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i\right) \mathbb{E}[x_{t-1} x_{t-1}^\top] \left(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i\right)^\top + \mathbb{E}[w_{t-1} w_{t-1}^\top].$$

Now as $(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i)\mathbb{E}[x_{t-1} x_{t-1}^\top](A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i)^\top$ is positive semi-definite and by assumption $\mathbb{E}[w_{t-1} w_{t-1}^\top]$ is positive definite, we have $\mathbb{E}[x_t x_t^\top]$ is positive definite and as a result $\underline{\sigma}_{\boldsymbol{X}} > 0$. In this case, we can simply take $\underline{\sigma}_{\boldsymbol{X}} = \min\{\sigma_{\min}(\mathbb{E}[x_0 x_0^\top]), \sigma_{\min}(W)\}$. □

We denote the set of model parameters of the LQG as $\widetilde{\Theta}$:={ $\frac{1}{(\sum_i k_i)+1}$, $\frac{1}{d+1}$, d, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\|W\|+1}$, $\frac{1}{\|\Sigma_0\|+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\|\boldsymbol{K}^*\|+1}$}. We write $\mathcal{H} = \{h \,|\, h$ are multivariate polynomials where the variables are in the set $\widetilde{\Theta}\}$ and $\mathcal{H}(.)$ when there are other variables not only $\widetilde{\Theta}$.

Natural Gradient Descent. We consider the following natural gradient descent updating rule for each player i ($i = 1, \cdots, N$), in a sequence of games for $m = 1, \ldots, M$,

$$K_t^{i,(m)} = K_t^{i,(m-1)} - \eta \nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})(\Sigma_t^{\boldsymbol{K}^{(m-1)}})^{-1}, \quad \forall\, 0 \leq t \leq T-1, \tag{3.3.5}$$

where $\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)}) = \frac{\partial C^i(\boldsymbol{K}^{(m-1)})}{\partial K_t^i}$ is the gradient of $C^i(\boldsymbol{K}^{(m-1)})$ with respect to K_t^i, and η is the step size.

Natural policy gradient methods [47] – and related algorithms such as trust region policy optimization [75] and the natural actor critic [69] – are some of the most popular and effective variants of the vanilla policy gradient methods in single-agent reinforcement learning. The natural policy gradient method performs better than the vanilla policy gradient method in practice since it takes the information geometry ([11, 47]) into consideration by normalizing the gradient term by $(\Sigma_t^{\boldsymbol{K}^{(m)}})^{-1}$ in (3.3.5). By doing so, the natural gradient method represents the steepest descent direction based on the underlying structure of the parameter space [47]. It is worth mentioning that [29] provided the first theoretical guarantee that the natural gradient method has an improved constant in the convergence rate compared to the vanilla version.

With access to the model parameters, players can directly calculate their own policy gradient and perform the natural gradient descent steps iteratively. See Algorithm 2 for the details. Unlike the nested-loop updates in some zero-sum LQG settings [92, 93] in which the inner-loop updates are sample inefficient, each player updates their policies simultaneously at each iteration in Algorithm 2. This simultaneous updating framework is a more realistic set-up for practical examples such as online auction bidding.

Algorithm 2 Natural Policy Gradient Method with Known Parameters

1: **Input**: Number of iterations M, time horizon T, initial policies $\boldsymbol{K}^{(0)} = (\boldsymbol{K}^{1,(0)}, \cdots, \boldsymbol{K}^{N,(0)})$, step size η, model parameters $\{A_t\}_{t=0}^{T-1}$, $\{B_t^i\}_{t=0}^{T-1}$, $\{Q_t^i\}_{t=0}^{T}$, and $\{R_t^i\}_{t=0}^{T-1}$ $(i = 1, \cdots, N)$.
2: **for** $m \in \{1, \ldots, M\}$ **do**
3: **for** $t \in \{T-1, \ldots, 0\}$ **do**
4: **for** $i \in \{1, \ldots, N\}$ **do**
5: Calculate the matrix $P_{t,i}^{\boldsymbol{K}^{(m-1)}}$ for $t = T-1, \ldots, 0$ by

$$P_{t,i}^{\boldsymbol{K}^{(m-1)}} = Q_t^i + (K_t^{i,(m-1)})^\top R_t^i K_t^{i,(m-1)} + \left(A_t - \sum_{i=1}^{N} B_t^i K_t^{i,(m-1)}\right)^\top P_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \left(A_t - \sum_{i=1}^{N} B_t^i K_t^{i,(m-1)}\right) \tag{3.3.6}$$

with $P_{T,i}^{\boldsymbol{K}^{(m-1)}} = Q_T^i$.
6: Calculate the matrix E_t^i by

$$E_{t,i}^{\boldsymbol{K}^{(m-1)}} = R_t^i K_t^{i,(m-1)} - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \left(A_t - \sum_{i=1}^{N} B_t^i K_t^{i,(m-1)}\right). \tag{3.3.7}$$

7: Update the policies using the natural gradient descent rule:

$$K_t^{i,(m)} = K_t^{i,(m-1)} - 2\eta E_{t,i}^{\boldsymbol{K}^{(m-1)}}. \tag{3.3.8}$$

8: **end for**
9: **end for**
10: **end for**
11: Return the iterates $\boldsymbol{K}^{(M)} = (\boldsymbol{K}^{1,(M)}, \cdots, \boldsymbol{K}^{N,(M)})$.

Remark 3.3.2. In Algorithm 2, during iteration m, each player first calculates the solution to the backward Riccati equation in (3.3.6) based on the observations from the previous iteration $m-1$, and obtains the policy gradients in (3.3.7). The natural

gradient method is then applied in (3.3.8) to update the policy for that iteration. Note that (3.3.8) does not involve $(\Sigma_t^{\boldsymbol{K}})^{-1}$ since, as we show later in Lemma 3.3.7, we have $2E_{t,i}^{\boldsymbol{K}} = \nabla_{K_t^i} C^i(\boldsymbol{K})(\Sigma_t^{\boldsymbol{K}})^{-1}$.

We now introduce the main assumption and the main result for the natural policy gradient method applied to the class of general-sum LQG. To start we define ρ^* as

$$\rho^* := \max\left\{ \max_{0\le t\le T-1} \left\| A_t - \sum_{i=1}^{N} B_t^i K_t^{i*} \right\|, 1+\delta \right\}, \tag{3.3.9}$$

for some small constant $\delta > 0$, and define $\psi := \max_i\{C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\}$. Now set

$$\bar{\rho} := \rho^* + N\gamma_B \sqrt{\frac{T\psi}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}} + \frac{1}{20T^2}. \tag{3.3.10}$$

Assumption 3.3.3 (System Noise)**.** *The system parameters satisfy the following inequality*

$$\frac{(\underline{\sigma}_{\boldsymbol{X}})^5}{\|\Sigma_{\boldsymbol{K}^*}\|} > 20(N-1)^2 T^2 d \,\frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}{}^2} \left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2. \tag{3.3.11}$$

Note that $\underline{\sigma}_{\boldsymbol{X}}$ appears on both sides of (3.3.11) and when $\underline{\sigma}_{\boldsymbol{X}}$ increases, the LHS of (3.3.11) increases while the RHS of (3.3.11) decreases. Therefore, Assumption 3.3.3 requires $\underline{\sigma}_{\boldsymbol{X}}$ to be large enough such that inequality (3.3.11) holds. This can be interpreted as ensuring that the system needs a certain level of noise for the learning agents to find the correct direction towards the Nash equilibrium. Assumption 3.3.3 also imposes conditions on the initial policy $\boldsymbol{K}^{(0)}$ as ψ depends on the difference between $C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$ and $C^i(\boldsymbol{K}^*)$.

Below we provide a set of parameters such that Assumption 3.3.3 is satisfied. In Section 3.5 we will show that, at least in some circumstances, the natural gradient method leads to the Nash equilibrium even when Assumption 3.3.3 is violated. This further demonstrates the power of the natural policy gradient method in practice.

Remark 3.3.4. An example where Assumption 3.3.3 is satisfied is given below. Note that to check that Assumption 3.3.3 holds in this example, we need to calculate $C^i(\boldsymbol{K}^*)$ with $\boldsymbol{K}^*$ solved from the Riccati equations given in (3.2.4)-(3.2.5), and also calculate the parameters $\bar{\rho}$, $\Sigma_{\boldsymbol{K}^*}$, and ψ accordingly.

- Parameters:

$$A_t = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix}, \quad W = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix},$$

$$Q_T^1 = Q_T^2 = Q_t^1 = Q_t^2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad R_t^1(t) = R_t^2(t) = 0.3,$$

  and $T = 1$.

- Initialization: Take $x_0 = (x_0^1, x_0^2)$ where x_0^1, x_0^2 are independent and sampled from $\mathcal{N}(0.3, 0.2)$ and $\mathcal{N}(0.4, 0.3)$ respectively. The initial policies are $\boldsymbol{K}^{1,(0)} = \boldsymbol{K}^{2,(0)} = (0.2, 0.01)$.

Theorem 3.3.5 (Global Convergence of the Natural Policy Gradient Method)**.** *Assume Assumptions 3.2.1, 3.2.2, 3.2.4 hold. We also assume Assumption 3.3.3 so that*

$$\widehat{\alpha} := \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - 20\,(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4 (\max_i\{C^i(\boldsymbol{K}^*)\} + \psi)^4}{\underline{\sigma}_{\boldsymbol{X}}{}^4\, \underline{\sigma}_{\boldsymbol{Q}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}} \left(\frac{\bar{\rho}^{2T} - 1}{\bar{\rho}^2 - 1}\right)^2 > 0. \quad (3.3.12)$$

Then there exists an $0 < \eta_0 \in \mathcal{H}(\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) + 1})$ *such that for* $\epsilon > 0, 0 < \eta < \eta_0$*, the cost function of the natural gradient method* (3.3.5) *satisfies*

$$\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \Big) \leq \epsilon,$$

whenever $M \geq \frac{1}{\widehat{\alpha}\eta} \log(\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))}{\epsilon})$.

The convergence rate developed in Theorem 3.3.5 is usually referred to as a linear convergence rate [63] as we show that $\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(m+1)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))}{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))} \leq r$ for some $r \in (0, 1)$, which is the result of Lemma 3.3.17. This leads to the result, seen in point 2. of the Theorem, that $\sum_{i=1}^N \big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \big) = O(e^{-M})$ and therefore this is also called an exponential convergence rate in the literature [63].

Compared to directly evaluating $C^i(\boldsymbol{K}^{(m)})$ along the training process $m = 1, 2, \cdots, M$, we observe that it is more natural to analyze the cost $C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*})$, with player i taking policy $\boldsymbol{K}^{i,(m)}$ and (as if) the other players were using optimal policies, as conditional optimality is the essence of a Nash equilibrium. This view point is the key

to establishing the global convergence result and this criterion is also used in all the key lemmas as well as the proof of the main theorem.

Note that our result can be easily generalized to the setting with agent-dependent learning rates as long as $\eta_i < \eta_0$ holds for each agent's step size η_i.

The proof of Theorem 3.3.5 relies on the regularity of the LQG problem, some properties of the gradient descent dynamics, and the perturbation analysis of the second moment matrix of the controlled dynamics.

3.3.1 Regularity of the LQG and Properties of the Gradient Descent Dynamics

We begin with the analysis of some properties of the N-player general-sum LQG (3.2.1)-(3.2.3). Our aim is to establish two key results Lemma 3.3.8 and Lemma 3.3.9 which provide the gradient dominance condition and a smoothness condition on the cost function $C^i(\boldsymbol{K})$ of player i with respect to the joint policy $\boldsymbol{K}$ from all players, respectively.

In the finite horizon setting, define $P_{t,i}^{\boldsymbol{K}}$ as the solution to

$$P_{t,i}^{\boldsymbol{K}} = Q_t^i + (K_t^i)^\top R_t^i K_t^i + \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right), \tag{3.3.13}$$

with terminal condition

$$P_{T,i}^{\boldsymbol{K}} = Q_T^i.$$

Lemma 3.3.6. *Assume Assumptions 3.2.1 holds. Then for $i = 1, \cdots, N$, $t = 0, \ldots, T$ the matrices $P_{t,i}^{\boldsymbol{K}}$ defined in* (3.3.13) *are positive definite.*

We defer the proof of Lemma 3.3.6 to Appendix B.1.

Player i's cost under the set of policies $\boldsymbol{K}$ can be rewritten as

$$C^i(\boldsymbol{K}) = \mathbb{E}\left[x_0^\top P_{0,i}^{\boldsymbol{K}} x_0 + N_{0,i}^{\boldsymbol{K}}\right],$$

where

$$N_{t,i}^{\boldsymbol{K}} = N_{t+1,i}^{\boldsymbol{K}} + \mathrm{Tr}(W P_{t+1,i}^{\boldsymbol{K}}), \quad N_{T,i}^{\boldsymbol{K}} = 0. \tag{3.3.14}$$

To see this,

$$\begin{aligned}
\mathbb{E}\left[x_0^\top P_{0,i}^{\boldsymbol{K}} x_0 + N_{0,i}^{\boldsymbol{K}}\right] &= \mathbb{E}\left[x_0^\top Q_0^i x_0 + x_0^\top (K_0^i)^\top R_0^i K_0^i x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t \right. \\
&\qquad \left. + x_0^\top (A_0 - \sum_{j=1}^{N} B_0^j K_0^j)^\top P_{1,i}^{\boldsymbol{K}} (A_0 - \sum_{j=1}^{N} B_0^j K_0^j) x_0\right] \\
&= \mathbb{E}\left[x_0^\top Q_0^i x_0 + (u_0^i)^\top R_0^i u_0^i + x_1^\top P_{1,i}^{\boldsymbol{K}} x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[c_0^i(x_0, u_0^i) + x_1^\top P_{1,i}^{\boldsymbol{K}} x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[c_0^i(x_0, u_0^i) + c_1^i(x_1, u_1^i) + x_2^\top P_{2,i}^{\boldsymbol{K}} x_2 + \sum_{t=2}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right].
\end{aligned}$$

In addition, for $t = 0, 1, \cdots, T-1$ and $i = 1, \cdots, N$, define

$$E_{t,i}^{\boldsymbol{K}} = R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right). \tag{3.3.15}$$

Then we have the following representation of the gradient terms.

Lemma 3.3.7. *The policy gradients have the following representation: for* $t = 0, 1, \cdots, T-1$,

$$\nabla_{K_t^i} C(\boldsymbol{K}) = 2\left(R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)\right) \Sigma_t^{\boldsymbol{K}} = 2E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}}, \tag{3.3.16}$$

where $E_{t,i}^{\boldsymbol{K}}$ *is defined in* (3.3.15).

Proof. Expressing the cost function in terms of K_t^i and suppressing the arguments of the cost c_t^i,

$$C^i(\boldsymbol{K}) = \mathbb{E}\left[\sum_{s=0}^{t-1} c_s^i + c_t^i + x_{t+1}^\top P_{t+1,i}^{\boldsymbol{K}} x_{t+1} + \sum_{s=t+1}^{T-1} w_s^\top P_{s+1,i}^{\boldsymbol{K}} w_s\right]$$

$$
\begin{aligned}
= \quad & \mathbb{E}\left[\sum_{s=0}^{t-1} c_s^i + x_t^\top (Q_t^i + (K_t^i)^\top R_t^i K_t^i) x_t + \sum_{s=t+1}^{T-1} w_s^\top P_{s+1,i}^{\boldsymbol{K}} w_s \right. \\
& \left. + \left(x_t^\top \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)^\top + w_t^\top\right) P_{t+1,i}^{\boldsymbol{K}} \left(\left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right) x_t + w_t\right)\right].
\end{aligned}
$$

Therefore, for $i = 1, \ldots, N$, the gradients are given by

$$
\nabla_{K_t^i} C^i(\boldsymbol{K}) = 2\left(R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)\right) \Sigma_t^{\boldsymbol{K}}.
$$

□

For the policy gradient method [29, 40] in the single-agent setting, gradient domination and smoothness of the objective function are two key conditions to guarantee the global convergence of the gradient descent methods. This is also the case for the N-player game setting. The gradient dominance condition for each player i is proved in Lemma 3.3.8, which indicates that for a policy $\boldsymbol{K}$, the distance between $C^i(\boldsymbol{K})$ and the optimal cost $C^i(\boldsymbol{K}^*)$ is bounded by the sum of the magnitudes of the gradients $\nabla_t C^i(\boldsymbol{K})$ for $t = 0, 1, \cdots, T-1$. The smoothness condition for each player i is proved in Lemma 3.3.9 where the difference between $C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ and $C^i(\boldsymbol{K})$ can be rewritten as a function of $\boldsymbol{K}^{i\prime} - \boldsymbol{K}^i$. The proofs of Lemmas 3.3.8 and 3.3.9 can be found in Appendix B.1.

Lemma 3.3.8 (Gradient Dominance)**.** *Assume Assumptions 3.2.1, 3.2.2, and 3.2.4 hold. Then, for player i, we have*

$$
\begin{aligned}
C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \quad &\leq \quad \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}\right) \\
&\leq \quad \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{R}}\,(\underline{\sigma}_{\boldsymbol{X}})^2} \sum_{t=0}^{T-1} \mathrm{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})\right),
\end{aligned}
$$

and

$$
C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \quad \geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} B_t^i\|} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}\right)
$$

$$\geq \frac{\underline{\sigma}_{\boldsymbol{X}}}{4\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2}\sum_{t=0}^{T-1}\frac{\operatorname{Tr}\left(\nabla_{K_t^i}C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top\nabla_{K_t^i}C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\right)}{\|R_t^i+(B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}B_t^i\|}.$$

Lemma 3.3.9 (Almost Smoothness)**.** *For two sets of policies $\boldsymbol{K}$ and $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i})$, we have that*

$$C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i})-C^i(\boldsymbol{K}) = \sum_{t=0}^{T-1}\Big[\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top(R_t^i+(B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}}B_t^i)(K_t^{i\prime}-K_t^i)\right) +2\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\right)\Big].$$

To utilize Lemmas 3.3.8 and 3.3.9 in the proof of Theorem 3.3.5, we need to further bound $P_{t,i}^{\boldsymbol{K}}$ and $\Sigma_{\boldsymbol{K}}$, which is provided below in Lemma 3.3.10. We defer the proof of Lemma 3.3.10 to Appendix B.1.

Lemma 3.3.10. *Assume Assumptions 3.2.1 and 3.2.2 hold. Then for $t=0,1,\ldots,T$ and $i=1,\ldots,N$*

$$\left\|P_{t,i}^{\boldsymbol{K}}\right\|\leq\frac{C^i(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}},\ \|\Sigma_{\boldsymbol{K}}\|\leq\frac{C^i(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}.$$

3.3.2 Perturbation Analysis of the Second Moment Matrix

Our aim in this section is to provide an explicit control of the change in the $\mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top]$ matrix after a change in policy $\boldsymbol{K}^i$. We begin by defining two linear operators on symmetric matrices. For $X\in\mathbb{R}^{d\times d}$ we set

$$\mathcal{F}_{K_t}(X):=\Big(A_t-\sum_{j=1}^N B_t^jK_t^j\Big)X\Big(A_t-\sum_{j=1}^N B_t^jK_t^j\Big)^\top, \tag{3.3.17}$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X):=X+\sum_{t=0}^{T-1}\Pi_{i=0}^t\Big(A_t-\sum_{j=1}^N B_t^jK_t^j\Big)\,X\,\Pi_{i=0}^t\Big(A_{t-i}-\sum_{j=1}^N B_{t-i}^jK_{t-i}^j\Big)^\top.$$

Here similar to Chapter 2, Π denotes the repeated multiplication from the left. If we write $\mathcal{G}_t^{\boldsymbol{K}} = \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \cdots \circ \mathcal{F}_{K_0}$, then the following relationships hold:

$$\mathcal{G}_t^{\boldsymbol{K}}(X) = \mathcal{F}_{K_t} \circ \mathcal{G}_{t-1}^{\boldsymbol{K}}(X) = \Pi_{i=0}^{t}\Big(A_t - \sum_{j=1}^{N} B_t^j K_t^j\Big)\, X\, \Pi_{i=0}^{t}\Big(A_{t-i} - \sum_{j=1}^{N} B_{t-i}^j K_{t-i}^j\Big)^{\top}, \tag{3.3.18}$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X) = X \,+\, \sum_{t=0}^{T-1} \mathcal{G}_t^{\boldsymbol{K}}(X). \tag{3.3.19}$$

When the policy $\boldsymbol{K}$ is clear we will write $\mathcal{G}_t = \mathcal{G}_t^{\boldsymbol{K}}$ and $\mathcal{F}_t = \mathcal{F}_{K_t}$. We first show the relationship between the operator $\mathcal{T}_{\boldsymbol{K}}$ and the quantity $\Sigma_{\boldsymbol{K}}$.

Proposition 3.3.11. *For $T \geq 2$, we have that*

$$\Sigma_{\boldsymbol{K}} = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta_{\boldsymbol{K}}(W), \tag{3.3.20}$$

where

$$\Delta_{\boldsymbol{K}}(W) = \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s} W D_{t,s}^{\top} + T\, W,$$

with $D_{t,s} = \Pi_{u=s}^{t}(A_u - \sum_{j=1}^{N} B_u^j K_u^j)$ (for $s = 1, 2, \cdots, t$), and $\Sigma_0 = \mathbb{E}\left[x_0 x_0^{\top}\right]$.

We defer the proof of Proposition 3.3.11 to Appendix B.2.

Given two policies $\boldsymbol{K}$ and $\boldsymbol{K}' = (\boldsymbol{K}^{1\prime}, \cdots, \boldsymbol{K}^{N\prime})$, let us define

$$\begin{aligned}\rho_{\boldsymbol{K},\boldsymbol{K}'} := \max\Bigg\{ &\max_i \Bigg\{ \max_{0\leq t\leq T-1} \Bigg\| A_t - B_t^i K_t^i - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*} \Bigg\| \Bigg\}, \max_{0\leq t\leq T-1} \Bigg\| A_t - \sum_{i=1}^{N} B_t^i K_t^i \Bigg\|, \\ &\max_i \Bigg\{ \max_{0\leq t\leq T-1} \Bigg\| A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*} \Bigg\| \Bigg\}, \\ &\max_i \Bigg\{ \max_{0\leq t\leq T-1} \Bigg\| A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j} \Bigg\| \Bigg\}, 1+\xi \Bigg\},\end{aligned} \tag{3.3.21}$$

for some small constant $\xi > 0$.

For any given policy $\boldsymbol{K}$, $\max_{0\leq t\leq T-1} \Big\| A_t - \sum_{j=1}^{N} B_t^j K_t^j \Big\|$ measures the radius of the state dynamics under policy $\boldsymbol{K}$. Thus $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ defined in (3.3.21) is the maximum radius of the policies $\boldsymbol{K}$, $(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})$, $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*})$, and $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ $(i = 1, 2, \cdots, N)$. We will

show later that the value (or the upper bound) of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ plays an essential role in the convergence analysis.

Remark 3.3.12. By the definition of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ in (3.3.21), we have $\rho_{\boldsymbol{K},\boldsymbol{K}'} \geq 1+\xi > 1$. This regularization term $1+\xi$ is introduced to simplify the presentation. Alternatively, if we remove this term from the definition of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$, a similar analysis can still be carried out by considering the different cases: $\rho_{\boldsymbol{K},\boldsymbol{K}'} < 1$, $\rho_{\boldsymbol{K},\boldsymbol{K}'} = 1$ and $\rho_{\boldsymbol{K},\boldsymbol{K}'} > 1$.

We now provide an upper bound for $\rho_{\boldsymbol{K},\boldsymbol{K}'}$.

Lemma 3.3.13. *Assume Assumption 3.2.4 holds. Then,*

$$\rho_{\boldsymbol{K},\boldsymbol{K}'} \leq \rho^* + N\gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}} \max_i \left\{ C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \right\}} + \gamma_B \max_i \left\{ \|K_t^{i\prime} - K_t^i\| \right\} \tag{3.3.22}$$

where ρ^* was defined in (3.3.9).

Proof. By Lemma 3.3.9, we have

$$\begin{aligned} & C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \\ = & \sum_{t=0}^{T-1} \left[\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} (K_t^i - K_t^{i*})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^*} B_t^i)(K_t^i - K_t^{i*}) \right) \right] \\ \geq & \ \underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}} \sum_{t=0}^{T-1} \|K_t^i - K_t^{i*}\|^2 \geq \frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{T} \left|\!\left|\!\left| \boldsymbol{K}^i - \boldsymbol{K}^{i*} \right|\!\right|\!\right|^2. \end{aligned} \tag{3.3.23}$$

Applying this we have

$$\begin{aligned} \left\| A_t - \sum_{j=1}^N B_t^j K_t^j \right\| & \leq \left\| A_t - \sum_{j=1}^N B_t^j K_t^{j*} \right\| + \sum_{j=1}^N \|B_t^j\| \, \|K_t^j - K_t^{j*}\| \\ & \leq \left\| A_t - \sum_{j=1}^N B_t^j K_t^{j*} \right\| + \gamma_B \sum_{j=1}^N \left|\!\left|\!\left| \boldsymbol{K}^j - \boldsymbol{K}^{j*} \right|\!\right|\!\right| \\ & \leq \left\| A_t - \sum_{j=1}^N B_t^j K_t^{j*} \right\| + N\gamma_B \sqrt{\frac{T \, \max_j \{ C^j(\boldsymbol{K}^j, \boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*) \}}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}}. \end{aligned} \tag{3.3.24}$$

Also, by the triangle inequality we have

$$\left\| A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j - B_t^i K_t^{i\prime} \right\| \leq \left\| A_t - \sum_{j=1}^{N} B_t^j K_t^j \right\| + \gamma_B \|K_t^{i\prime} - K_t^i\|. \tag{3.3.25}$$

and

$$\left\| A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^{i\prime} \right\| \leq \left\| A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^i \right\| + \gamma_B \|K_t^{i\prime} - K_t^i\|. \tag{3.3.26}$$

Finally, applying (3.3.23),

$$\begin{aligned} \left\| A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^i \right\| &\leq \left\| A_t - \sum_{j=1}^{N} B_t^j K_t^{j*} \right\| + \gamma_B \|K_t^{i*} - K_t^i\| \\ &\leq \left\| A_t - \sum_{j=1}^{N} B_t^j K_t^{j*} \right\| \\ &\quad + \gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} \big(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \big)}. \end{aligned} \tag{3.3.27}$$

Therefore combining (3.3.24)-(3.3.27), we obtain the statement (3.3.22). □

Recall the definition of $\mathcal{F}_t = \mathcal{F}_{K_t}$ and $\mathcal{G}_t = \mathcal{G}_t^{\boldsymbol{K}}$ in (3.3.17) and (3.3.18) associated with $\boldsymbol{K}$, similarly let us define $\mathcal{G}_t^{i\prime} = \mathcal{F}_{K_t^{i\prime}, K_t^{-i}} \circ \mathcal{F}_{K_{t-1}^{i\prime}, K_{t-1}^{-i}} \circ \cdots \circ \mathcal{F}_{K_0^{i\prime}, K_0^{-i}}$ for the set of policies $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ and write $\mathcal{F}_t^{i\prime} = \mathcal{F}_{K_t^{i\prime}, K_t^{-i}}$. Also recall that the operator norm is defined in Section 2.3.2. We now establish a perturbation analysis for $\mathcal{F}_t$ and $\mathcal{G}_t$.

Lemma 3.3.14. *For all $t = 0, 1, \cdots, T-1$, we have*

$$\|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \leq 2\rho_{\boldsymbol{K}, \boldsymbol{K}'} \gamma_B \|K_t^i - K_t^{i\prime}\|. \tag{3.3.28}$$

We defer the proof of Lemma 3.3.14 to Appendix B.2.

Lemma 3.3.15 (Perturbation Analysis for $\mathcal{G}_t$)**.** *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$*

and $i = 1, \cdots, N$*, we have that*

$$\sum_{t=0}^{T-1} \left\| (\mathcal{G}_t - \mathcal{G}_t^{i\prime})(\Sigma) \right\| \leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \Big) \|\Sigma\|.$$

We defer the proof of Lemma 3.3.15 to Appendix B.2.

Recall that γ_A, γ_B, and γ_R are defined in (3.3.4). Then we have the following perturbation analysis of $\Sigma_{\boldsymbol{K}}$.

Lemma 3.3.16 (Perturbation Analysis of $\Sigma_{\boldsymbol{K}}$)**.** *Assume Assumption 3.2.1 holds. Then*

$$\left\| \Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} \right\| \leq 2\gamma_B \, \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}(\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1)}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \left(\frac{C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) \left\| \left| \boldsymbol{K}^i - \boldsymbol{K}^{i\prime} \right| \right\|.$$

Proof. Using Lemma 3.3.14,

$$\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \quad \leq \quad 2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B \sum_{t=0}^{T-1} \|K_t^i - K_t^{i\prime}\|.$$

Define $D_{t,s}^{i\prime} = \Pi_{u=s}^{t}(A_u - B_u^i K_u^{i\prime} - \sum_{j=1, j\neq i}^{N} B_u^j K_u^j)$ (for $s = 1, 2, \cdots, t$). Then, in a similar way to the proof of Lemma 3.3.15, we have, $\forall\, t = 1, \cdots, T-1$,

$$\sum_{s=1}^{t} \left\| D_{t,s} W D_{t,s}^{\top} - D_{t,s}^{i\prime} W (D_{t,s}^{i\prime})^{\top} \right\| \leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \left(\sum_{s=0}^{t} \|\mathcal{F}_s - \mathcal{F}_s^{i\prime}\| \right) \|W\|. \quad (3.3.29)$$

By Proposition 3.3.11, (3.3.19) and (3.3.29), we have

$$\begin{aligned}
\left\| \Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} \right\| &\leq \left\| (\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})(\Sigma_0) \right\| + \sum_{t=1}^{T-1} \sum_{s=1}^{t} \left\| D_{t,s} W D_{t,s}^{\top} - D_{t,s}^{i\prime} W (D_{t,s}^{i\prime})^{\top} \right\| \\
&\leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \Big) \left(\|\Sigma_0\| + T\|W\| \right) \\
&\leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \left(\frac{C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) \left(2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B \left\| \left| \boldsymbol{K}^i - \boldsymbol{K}^{i\prime} \right| \right\| \right).
\end{aligned} \quad (3.3.30)$$

The last inequality holds since $\|\Sigma_0\| \leq \|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\| \leq \frac{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}}$ by Lemma 3.3.10. $\square$

3.3.3 Convergence and Complexity Analysis

We are now in a position to provide the proof of our main Theorem 3.3.5. This will follow from two important Lemmas. First, define

$$\rho_{\boldsymbol{K}} := \rho^* + N\gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}} \max_i \left\{ C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \right\}} + \frac{1}{20T^2}. \tag{3.3.31}$$

Further define g_1 and g_2 as follows:

$$g_1 := \frac{\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}, \tag{3.3.32}$$

and

$$g_2 := 20(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4 \max_i\{C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})\}^4}{{\underline{\sigma}_{\boldsymbol{Q}}}^2\, \underline{\sigma}_{\boldsymbol{R}}} \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \right)^2. \tag{3.3.33}$$

We also write $C^{i,-i*} = C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})$, $C^{i*} = C^i(\boldsymbol{K}^*)$ and $C^{i\prime,-i*} = C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*})$ to simplify notation.

Lemma 3.3.17 (One-step contraction)**.** *Assume Assumptions 3.2.1, 3.2.2, and 3.2.4 hold, and that*

$${\underline{\sigma}_{\boldsymbol{X}}}^5 > \frac{g_2}{g_1}. \tag{3.3.34}$$

Also assume the policy update step for player i at time t is given by

$$K_t^{i\prime} = K_t^i - \eta \nabla_{K_t^i} C^i(\boldsymbol{K})(\Sigma_t^{\boldsymbol{K}})^{-1}, \tag{3.3.35}$$

where

$$\eta \leq \min\left\{ I_1\,, I_2\,, \frac{1}{\underline{\sigma}_{\boldsymbol{R}}} \right\} \tag{3.3.36}$$

with

$$I_1 = \Bigg\{ 20T\frac{\rho_{\boldsymbol{K}}(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\left(\sum_{i=1}^{N} C^{i,-i*} + \underline{\sigma}_{\boldsymbol{Q}}\, T\|W\|\right)\gamma_B \max_i\{\max_t\{\|\nabla_{K_t^i} C^i(\boldsymbol{K})\|\}\}$$
$$+ \underline{\sigma}_{\boldsymbol{Q}}(\underline{\sigma}_{\boldsymbol{X}})^2 + 4\big(\gamma_R\,\underline{\sigma}_{\boldsymbol{X}} + (\gamma_B)^2\sum_{i=1}^{N} C^{i,-i*}\big)\sum_{i=1}^{N}\{C^{i,-i*}\}\Bigg\}^{-1} \cdot \underline{\sigma}_{\boldsymbol{Q}}(\underline{\sigma}_{\boldsymbol{X}})^2,$$

$$I_2 = \Bigg\{ \big(\max_i\{k_i\}\big)\frac{10T\sum_{i=1}^{N} C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\left(\gamma_R + (\gamma_B)^2\frac{\sum_{i=1}^{N} C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) + \frac{2d}{\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{10T\sum_{i=1}^{N} C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2$$
$$\left(\gamma_R + (\gamma_B)^2\frac{\sum_{i=1}^{N} C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2\Bigg\}^{-1} \cdot \frac{d}{80\,\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{10T\min_i\{C^{i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2.$$

Let $\alpha := \underline{\sigma}_{\boldsymbol{X}}\, g_1 - g_2/\,\underline{\sigma}_{\boldsymbol{X}}{}^4 > 0$. *Then, we have*

1. $\eta \in (0, \frac{1}{\alpha})$; *and*
2. *the following inequality holds*

$$\sum_{i=1}^{N}\big(C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big) \le (1-\alpha\eta)\left(\sum_{i=1}^{N}\big(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big)\right). \tag{3.3.37}$$

Remark 3.3.18. (1) In the one-step contraction analysis (Lemma 3.3.17), (3.3.34) imposes a condition on $C(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})$ which is associated with the current policy $\boldsymbol{K}$. In the analysis of the global convergence result (Theorem 3.3.5), (3.3.11) imposes a similar condition on $C(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$ which is associated with the initial policy $\boldsymbol{K}^{(0)}$. Condition (3.3.11) in Assumption 3.3.3 and the step size condition in Theorem 3.3.5 guarantee that, condition (3.3.34) holds for any $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ throughout the training process ($m = 1, 2, \cdots, M$). This further ensures that the one-step contraction analysis in Lemma 3.3.17 can be applied iteratively which leads to the global convergence result as stated in Theorem 3.3.5.
(2) Note that the numbers such as 20 and 80 that appear in I_1, I_2 are not arbitrary and, although some minor improvements can be made by optimizing at various stages, they enable us to obtain reasonable bounds.

Proof. We break this proof up into a series of steps.

Step 1: We first consider the consequences of the condition $\eta \leq \min\{I_1, I_2\}$. Straightforward calculations show that when condition $\eta \leq I_1$ is satisfied, the following inequalities hold:

1. $\forall i = 1, \cdots, N$,

$$\|K_t^{i\prime} - K_t^i\| = \eta \|\nabla_{K_t^i} C^i(\boldsymbol{K}) \Sigma_t^{-1}\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20T\gamma_B C^{i,-i*}}. \tag{3.3.38}$$

2. $\forall i = 1, \cdots, N$,

$$\begin{aligned} & \eta \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}} \frac{\gamma_B}{\underline{\sigma}_{\boldsymbol{X}}} \sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \right) \\ \leq\ & I_1 \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}} \frac{\gamma_B}{\underline{\sigma}_{\boldsymbol{X}}} T \max_t \{\|\nabla_{K_t^i} C^i(\boldsymbol{K})\|\} \right) \\ \leq\ & \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}. \end{aligned} \tag{3.3.39}$$

3. $\forall i = 1, \cdots, N$,

$$\eta \leq I_1 \ \leq\ \frac{\underline{\sigma}_{\boldsymbol{X}}}{\underline{\sigma}_{\boldsymbol{X}} + 2\frac{2C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}(\gamma_R + \gamma_B^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}})}. \tag{3.3.40}$$

In the case where $\eta \leq I_2$, we have $\forall i = 1, \cdots, N$

$$\begin{aligned} & 4\eta\, k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right) \\ & + 8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right)^2 \\ \leq\ & 4I_2 \left(k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right) \right. \\ & \left. + 2\frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right)^2 \right) \\ \leq\ & \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \min_i\{C^{i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \leq \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2. \end{aligned} \tag{3.3.41}$$

By (3.3.38) and Lemma 3.3.10 we have

$$\gamma_B \|K_t^{i\prime} - K_t^i\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20TC^{i,-i*}} \leq \frac{1}{20T^2}.$$

Therefore, we have $\rho_{\boldsymbol{K},\boldsymbol{K}'} \le \rho_{\boldsymbol{K}}$ by Lemma 3.3.13.

Step 2: We bound the norm of the second moment matrix of the state. By Lemma 3.3.14,

$$\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t^i,K_t^{-i*}} - \mathcal{F}_{K_t^{i\prime},K_t^{-i*}}\| \le 2\rho_{\boldsymbol{K}}\gamma_B \left(\sum_{t=0}^{T-1} \|K_t^i - K_t^{i\prime}\|\right),$$

and hence by Lemma 3.3.16, we have

$$\begin{aligned}
\left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| &\le \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\left(\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t^i,K_t^{-i*}} - \mathcal{F}_{K_t^{i\prime},K_t^{-i*}}\|\right)(\|\Sigma_0\| + T\|W\|) \\
&\le \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right) 2\rho_{\boldsymbol{K}}\gamma_B \left|\!\left|\!\left|\boldsymbol{K}^i - \boldsymbol{K}^{i\prime}\right|\!\right|\!\right| \\
&\le \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right) 2\rho_{\boldsymbol{K}}\gamma_B \frac{\eta}{\underline{\sigma}_{\boldsymbol{X}}} \sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \\
&\le \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}, \qquad (3.3.42)
\end{aligned}$$

where the last inequality holds by (3.3.39) when $\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K}\| > 0$. Note that when $\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| = 0$, we have $\left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| = 0 < \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}$ and hence (3.3.42) still holds in this case. Therefore, by Lemma 3.3.10, and noting $\underline{\sigma}_{\boldsymbol{X}} \le \|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|/T$,

$$\begin{aligned}
\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| &\le \left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} - \Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\| + \left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\| \le \frac{\underline{\sigma}_{\boldsymbol{X}}}{10} + \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} \\
&\le \frac{\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\|}{10T} + \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}, \qquad (3.3.43)
\end{aligned}$$

which leads to

$$\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| \le \frac{10T\, C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}. \qquad (3.3.44)$$

Step 3: We now bound $\|P_{t,i}^{\boldsymbol{K}} - P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|$ by $\sum_{j=1,j\neq i}^{N} \left|\!\left|\!\left|\boldsymbol{K}^j - \boldsymbol{K}^{j*}\right|\!\right|\!\right|$ where $\boldsymbol{K} = (\boldsymbol{K}^i, \boldsymbol{K}^{-i})$.

$$\begin{aligned}
&\left\|P_{t,i}^{\boldsymbol{K}} - P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\| \\
= &\left\|\Big(A_t - \sum_{j=1}^{N} B_t^j K_t^j\Big)^\top P_{t+1,i}^{\boldsymbol{K}} \Big(A_t - \sum_{j=1}^{N} B_t^j K_t^j\Big)\right.
\end{aligned}$$

$$
\begin{aligned}
&\left\| -\Big(A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}\Big)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Big(A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}\Big) \right\| \\
\leq\ & \left\| \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)^\top \Big(P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Big) \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big) \right\| \\
&+ \left\| \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big) \right. \\
&\left. -\Big(A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}\Big)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Big(A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}\Big) \right\| \\
\leq\ & \rho_{\boldsymbol{K}}^2 \left\| P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| + 2\rho_{\boldsymbol{K}}\gamma_B \left\| P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \sum_{j=1,j\neq i}^N \left\| K_t^j - K_t^{j*} \right\|, \qquad (3.3.45)
\end{aligned}
$$

where the last inequality holds by letting $X = P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}$, $Y = A_t - \sum_{j=1}^N B_t^j K_t^j$, and $Z = A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}$ in (B.2.2). Since $\left\| P_{T,i}^{\boldsymbol{K}} - P_{T,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| = 0$ holds at terminal time T, we obtain $\forall t = 0, \cdots, T-1$,

$$
\left\| P_{t,i}^{\boldsymbol{K}} - P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \leq \frac{2\gamma_B C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{\rho_{\boldsymbol{K}}(\rho_{\boldsymbol{K}}^{2(T-t)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t}^{T-1} \left(\sum_{j=1,j\neq i}^N \left\| K_s^j - K_s^{j*} \right\| \right). \qquad (3.3.46)
$$

Therefore, $\forall t = 0, 1, \cdots, T-1$,

$$
\begin{aligned}
&\left\| E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \\
=\ & \left\| (B_t^i)^\top \left(P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big) - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \sum_{j=1,j\neq i}^N B_t^j (K_t^j - K_t^{j*}) \right\| \\
&\leq \rho_{\boldsymbol{K}}\gamma_B \|P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\| + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \sum_{j=1,j\neq i}^N \|K_t^j - K_t^{j*}\| \qquad (3.3.47) \\
&\leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \Big(\frac{2\rho_{\boldsymbol{K}}^2(\rho_{\boldsymbol{K}}^{2(T-t-1)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t+1}^{T-1} \sum_{j=1,j\neq i}^N \|K_s^j - K_s^{j*}\| + \sum_{j=1,j\neq i}^N \|K_t^j - K_t^{j*}\| \Big) \qquad (3.3.48) \\
&\leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \Big(\frac{2(\rho_{\boldsymbol{K}}^{2(T-t)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t}^{T-1} \Big(\sum_{j=1,j\neq i}^N \|K_s^j - K_s^{j*}\| \Big) \Big) \\
&\leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \Big(\frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{j=1,j\neq i}^N \|\|\boldsymbol{K}^j - \boldsymbol{K}^{j*}\|\| \Big), \qquad (3.3.49)
\end{aligned}
$$

where (3.3.47) holds by (3.3.15), and (3.3.48) holds by (3.3.46).

Step 4: We can now estimate the cost difference between using $\boldsymbol{K}^i$ and the update $\boldsymbol{K}^{i\prime}$. By Lemma 3.3.9 we have

$$\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
&= \sum_{t=0}^{T-1} \Big[\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(K_t^{i\prime} - K_t^i)\right) \\
&+2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\Big].
\end{aligned}$$

Using the definition of $\boldsymbol{K}^{i\prime}$ and the expression for the gradient from Lemma 3.3.7

$$\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
&= \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot \right. \\
&\quad (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\big) \\
&\quad \left. -4\eta \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\right] \\
&= \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\right) \\
&\quad +8\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
&\quad +4\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
&\quad -4\eta \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
&\quad \left. -4\eta \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\right].
\end{aligned}$$

Now, letting $\omega^2 = \frac{2}{\underline{\sigma}_{\boldsymbol{X}}}$ in

$$2 \operatorname{Tr}(A^\top B) = \operatorname{Tr}(A^\top B + B^\top A) \leq \omega^2 \operatorname{Tr}(A^\top A) + \frac{1}{\omega^2} \operatorname{Tr}(B^\top B),$$

(which holds for any matrices A and B of the same dimension) we have

$$\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
&\leq \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\right) \\
&\quad +8\eta^2 \frac{\underline{\sigma}_{\boldsymbol{X}}}{4} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) + 8\eta^2 \frac{1}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot \right.
\end{aligned}$$

$$
\begin{aligned}
&(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big) \\
&+4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big) \\
&+4\eta \frac{1}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big) \\
&+4\eta \frac{\underline{\sigma}_{\boldsymbol{X}}}{4} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big) - 4\eta\, \underline{\sigma}_{\boldsymbol{X}} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big] \\
\le \sum_{t=0}^{T-1} \Big[&\Big(4\eta^2 \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \operatorname{Tr}\big(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\big) + \frac{8\eta^2}{\underline{\sigma}_{\boldsymbol{X}}} \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|^2 \cdot \\
&\operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big) + \frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big)\Big) \|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 \\
&+\big(2\eta^2\, \underline{\sigma}_{\boldsymbol{X}} + 4\eta^2 \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\| + \eta\, \underline{\sigma}_{\boldsymbol{X}} - 4\eta\, \underline{\sigma}_{\boldsymbol{X}}\big) \cdot \\
&\operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big]. \qquad (3.3.50)
\end{aligned}
$$

Now, using (3.3.44), we can bound the step size condition in (3.3.40) by

$$
\eta \le \frac{\underline{\sigma}_{\boldsymbol{X}}}{\underline{\sigma}_{\boldsymbol{X}} + 2\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|}.
$$

This gives

$$
2\eta^2\, \underline{\sigma}_{\boldsymbol{X}} + 4\eta^2 \left\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| \left\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\right\| + \eta\, \underline{\sigma}_{\boldsymbol{X}} - 4\eta\, \underline{\sigma}_{\boldsymbol{X}} \le -\eta\, \underline{\sigma}_{\boldsymbol{X}}\,.
$$

Hence, using this in (3.3.50), we have

$$
C^{i\prime,-i*} - C^{i,-i*} \le \eta\, h_{\text{diff}}^i \sum_{t=0}^{T-1} \|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 - \eta\, \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big), \qquad (3.3.51)
$$

where

$$
\begin{aligned}
h_{\text{diff}}^i := 4\eta\, k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) + 8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \cdot \\
\left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 + 4\frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2.
\end{aligned}
$$

Therefore, by (3.3.44), (3.3.49), Lemma 3.3.8, and Lemma 3.3.10,

$$
C^{i\prime,-i*} - C^{i,-i*} \;\le\; \eta\, h_{\text{diff}}^i\, T \left[\frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{j=1, j\neq i}^{N} |||\boldsymbol{K}^j - \boldsymbol{K}^{j*}|||\right]^2
$$

$$
\begin{aligned}
&-\eta\,\underline{\sigma}_{\boldsymbol{X}}\sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\\
\leq\;& \eta\,h_{\mathrm{glob}}\Big(\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|\Big)^2-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big(C^{i,-i*}-C^{i*}\big),
\end{aligned}
\tag{3.3.52}
$$

where

$$
\begin{aligned}
h_{\mathrm{glob}}=4T\Bigg[&\eta\,k_i\frac{10T\max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\left(\gamma_R+(\gamma_B)^2\frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right)\\
&+2\eta\frac{d}{\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{10T\max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2\left(\gamma_R+(\gamma_B)^2\frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2\\
&+\frac{d}{\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{10T\max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2\Bigg]\cdot\left[\frac{(\gamma_B)^2\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\right]^2.
\end{aligned}
\tag{3.3.53}
$$

Step 5: Finally we can establish the one step contraction. Using (3.3.52), we have

$$
\begin{aligned}
C^{i\prime,-i*}-C^{i*}&=C^{i\prime,-i*}-C^{i,-i*}+C^{i,-i*}-C^{i*}\\
&\leq\big(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\big(C^{i,-i*}-C^{i*}\big)+\eta\,h_{\mathrm{glob}}\Big(\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|\Big)^2.
\end{aligned}
\tag{3.3.54}
$$

Hence by Lemma 3.3.9 and (3.3.23), we have

$$
\sum_{j=1,j\neq i}^{N}\big(C^{j,-j*}-C^{j*}\big)\geq\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{T}\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|^2\geq\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}\left(\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|\right)^2}{T(N-1)},
$$

and thus

$$
C^{i\prime,-i*}-C^{i*}\leq\big(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\big(C^{i,-i*}-C^{i*}\big)+\;\eta\;h_{\mathrm{glob}}\frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\Big(\sum_{j=1,j\neq i}^{N}\big(C^{j,-j*}-C^{j*}\big)\Big).
\tag{3.3.55}
$$

Summing up (3.3.55) for $i = 1, \cdots, N$, we have

$$\sum_{i=1}^{N} \left(C^{i\prime,-i*} - C^{i*}\right) \leq \left(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} + \eta(N-1) h_{\text{glob}} \frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}\right) \left(\sum_{i=1}^{N} (C^{i,-i*} - C^{i*})\right). \tag{3.3.56}$$

Since $\eta \leq I_2$, we have (3.3.41) and then

$$h^i_{\text{diff}} \leq (4 + \frac{1}{20}) \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \leq 5 \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{Q}}}\right)^2,$$

and

$$h_{\text{glob}} \leq \bar{h}_{\text{glob}},$$

where $\bar{h}_{\text{glob}}$ is given by

$$\bar{h}_{\text{glob}} = 5\, T\, d\, \frac{\max_i\{C^{i,-i*}\}^4 (\gamma_B)^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\, \underline{\sigma}_{\boldsymbol{X}}{}^3} \left[\frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1}\right]^2. \tag{3.3.57}$$

Under condition (3.3.34), we have

$$\underline{\sigma}_{\boldsymbol{X}}\, g_1 - \frac{g_2}{\underline{\sigma}_{\boldsymbol{X}}{}^4} = \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - (N-1)^2 \bar{h}_{\text{glob}} \frac{T}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} > 0,$$

which indicates that $\alpha\eta > 0$. Since $\eta \leq \frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$, we have

$$\eta < \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} < \left(\frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - (N-1)^2 \bar{h}_{\text{glob}} \frac{T}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}\right)^{-1} = \left(\underline{\sigma}_{\boldsymbol{X}}\, g_1 - \frac{g_2}{\underline{\sigma}_{\boldsymbol{X}}{}^4}\right)^{-1}.$$

Therefore we have $\alpha\eta < 1$. Along with (3.3.56), we obtain the one-step contraction (3.3.37). □

Lemma 3.3.19. *Assume Assumptions 3.2.1, 3.2.2, and 3.2.4 hold. Then we have that for player i,*

$$\begin{aligned}\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\|^2 \leq\ & 8\Big\{\frac{\rho_{\boldsymbol{K}}^{2(T+1)} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1}\, T\, \|W\|\Big\}^2 \cdot \\ & \Big\{d \frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} \Big[\frac{(\gamma_B)^2 \sum_{i=1}^N C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1}\Big]^2 \cdot\end{aligned}$$

$$\Big(\sum_{j=1, j\neq i}^{N} (C^j(\boldsymbol{K}^j, \boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*)) \Big) + \frac{\underline{\sigma}_{\boldsymbol{X}}\, \gamma_R + (\gamma_B)^2 \sum_{i=1}^{N} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \right) \Big\}. \tag{3.3.58}$$

Proof. Using Lemma 3.3.7, we have

$$\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\|^2 \leq 4 \sum_{t=0}^{T-1} \mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}} \right) \leq 4 \left(\|\Sigma_{\boldsymbol{K}}\| \right)^2 \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}} \right), \tag{3.3.59}$$

and

$$\begin{aligned}
& \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}} \right) \\
= & \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}) \right) \\
\leq & 2 \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}) \right) \\
& + 2 \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right) \\
\leq & 2d \sum_{t=0}^{T-1} \left\| E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right\|^2 + 2 \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right).
\end{aligned} \tag{3.3.60}$$

By (3.3.23), (3.3.49), we have

$$\begin{aligned}
& \sum_{t=0}^{T-1} \left\| E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right\|^2 \\
\leq & T \left[\frac{(\gamma_B)^2 C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \left(\sum_{j=1, j\neq i}^{N} \left\| \left| \boldsymbol{K}^j - \boldsymbol{K}^{j*} \right| \right\| \right) \right]^2 \\
\leq & \frac{T^2 (N-1)}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} \left[\frac{(\gamma_B)^2 C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \right]^2 \Big(\sum_{j=1, j\neq i}^{N} (C^j(\boldsymbol{K}^j, \boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*)) \Big).
\end{aligned} \tag{3.3.61}$$

By Lemma 3.3.8 and Lemma 3.3.10, we have

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \leq \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2 C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\left(C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right). \tag{3.3.62}$$

By Proposition 3.3.11,

$$\begin{aligned}
\|\Sigma_{\boldsymbol{K}}\| &\leq \|\Sigma_0\| + \sum_{t=0}^{T-1}\|\mathcal{G}_t(\Sigma_0)\| + \sum_{t=1}^{T-1}\sum_{s=1}^{t}\|D_{t,s}WD_{t,s}^\top\| + T\|W\| \\
&\leq \|\Sigma_0\| + \|\Sigma_0\|\sum_{t=0}^{T-1}\rho_{\boldsymbol{K}}^{2(t+1)} + T\|W\|\sum_{s=1}^{T-1}\rho_{\boldsymbol{K}}^{2(T-s)} + T\|W\| \\
&\leq \frac{\rho_{\boldsymbol{K}}^{2(T+1)}-1}{\rho_{\boldsymbol{K}}^2-1}\|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\,T\,\|W\|.
\end{aligned} \tag{3.3.63}$$

Therefore, combining (3.3.59)-(3.3.63), we obtain the statement (3.3.58). □

Finally, we are ready to provide the proof for the main result based on Lemmas 3.3.17 and 3.3.19.

Proof of Theorem 3.3.5. We first show that the total cost for the N players decreases at round $m = 1$. Take $\boldsymbol{K} = \boldsymbol{K}^{(0)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(1)}$ in Lemma 3.3.17, we begin by showing that there exists a positive lower bound on the RHS of (3.3.36). By the Cauchy-Schwarz inequality and Lemma 3.3.19,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K}^{(0)})\| \leq \sqrt{T\cdot\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K}^{(0)})\|^2} \\
\leq\ & 2\sqrt{2}\left(\frac{\rho_{\boldsymbol{K}^{(0)}}^{2(T+1)}-1}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}^{(0)}}^{2T}-1}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\,T\,\|W\|\right)\cdot\left\{d\frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\cdot\right.\right. \\
&\left.\left.\frac{2(\rho_{\boldsymbol{K}^{(0)}}^{2T}-1)}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\right]^2 + \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\right\}^{1/2}\sqrt{T\left(\sum_{i=1}^N(C^{i,-i*}-C^{i*})\right)} \\
\leq\ & 2\sqrt{2}\left(\frac{\bar{\rho}^{2(T+1)}-1}{\bar{\rho}^2-1}\|\Sigma_0\| + \frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\,T\,\|W\|\right)\cdot\left\{d\frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\cdot\right.\right. \\
&\left.\left.\frac{2(\bar{\rho}^{2T}-1)}{\bar{\rho}^2-1}\right]^2 + \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\right\}^{1/2}\sqrt{T\left(\sum_{i=1}^N(C^{i,-i*}-C^{i*})\right)},
\end{aligned}$$

where the last inequality holds since, after performing one-step natural gradient descent $K_t^{i,(1)} = K_t^{i,(0)} - \eta \nabla_{K_t^i} C^i(\boldsymbol{K}^{(0)})(\Sigma_t^{\boldsymbol{K}^{(0)}})^{-1}$, we have

$$\rho_{\boldsymbol{K}^{(0)}} \quad \leq \quad \rho^* + N\gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}\psi} + \frac{1}{20T^2} = \bar{\rho} \tag{3.3.64}$$

where ρ^* and $\bar{\rho}$ are defined in (3.3.9) and (3.3.10), and $\psi := \max_i\{C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^{i*}\}$. (3.3.64) holds by Lemma 3.3.13.

Now we aim to show that $\frac{1}{\bar{\rho}}$ is bounded below by polynomials in some model parameters. Given that $\left\| A_t - \sum_{i=1}^N B_t^i K_t^{i*} \right\| \leq \gamma_A + \gamma_B |||\boldsymbol{K}^*|||$ and that Lemma 3.3.13, $\bar{\rho}$ can be bounded above by polynomials in T, N, γ_A, γ_B, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$, and $|||\boldsymbol{K}^*|||$, or a constant $1+\xi$. Therefore, along with the fact that $\frac{1}{d} > \frac{1}{(a+1)(b+1)(c+1)}$ for $d < ab + c$ with some $a, b, c, d > 0$ and $\frac{1}{a^n+1} > \frac{1}{(a+1)^n}$ for $a > 0$ and $n \in \mathbb{N}^+$, $\frac{1}{\bar{\rho}}$ is bounded below by polynomials in $\frac{1}{T+1}$, $\frac{1}{N+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, and $\frac{1}{|||\boldsymbol{K}^*|||+1}$, or a constant $\frac{1}{1+\xi}$. Similarly, I_1 can be bounded below by polynomials in $\frac{1}{d+1}$, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\sum_{i=1}^N C^{i,-i*}+1}$, $\frac{1}{\|W\|+1}$, $\frac{1}{\|\Sigma_0\|+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, and $\frac{1}{|||\boldsymbol{K}^*|||+1}$; and I_2 can be bounded below by polynomials in $\frac{1}{\sum_i k_i+1}$, $\frac{1}{d+1}$, d, $\frac{1}{\sum_{i=1}^N C^{i,-i*}+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$.

Hence, there exists $\eta_0 \in \mathcal{H}\left(\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})+1}\right)$ as an appropriate polynomial in $\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})+1}$, $\frac{1}{\sum_i k_i+1}$, $\frac{1}{d+1}$, d, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\|W\|+1}$, $\frac{1}{\|\Sigma_0\|+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, and $\frac{1}{|||\boldsymbol{K}^*|||+1}$, such that when $\eta < \eta_0$, the step size condition (3.3.36) is satisfied. Therefore, by Lemma 3.3.17, we have

$$\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(1)}, \boldsymbol{K}^{-i*}) - C^{i*}) \Big) \leq (1 - \widehat{\alpha}\eta) \sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^{i*}) \Big),$$

with $\widehat{\alpha}$ defined in (3.3.12), which implies the total cost of N players decreases at $m = 1$. Proceeding inductively, assume the following facts hold at round m:

- $C^i(\boldsymbol{K}^{i,(m-1)}, \boldsymbol{K}^{-i*}) - C^{i*} \leq \psi$ for $i = 1, 2, \cdots, N$;
- $\rho_{\boldsymbol{K}^{(m-1)}} \leq \bar{\rho}$;
- $\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big) \leq (1 - \widehat{\alpha}\eta) \sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m-1)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big)$.

Now we prove the above facts also hold in round $m+1$. Taking $\boldsymbol{K} = \boldsymbol{K}^{(m-1)}$ and

$\boldsymbol{K}' = \boldsymbol{K}^{(m)}$ in (3.3.55), we have

$$\begin{aligned} C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^{i*} &\leq \big(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\big(C^i(\boldsymbol{K}^{i,(m-1)}, \boldsymbol{K}^{-i*}) - C^{i*}\big) \\ &\quad + \eta\ h_{\text{glob}} \frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} \left(\sum_{j=1, j\neq i}^{N} (C^j(\boldsymbol{K}^{j,(m-1)}, \boldsymbol{K}^{-j*}) - C^{j*}) \right) \\ &\leq \big(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\psi + \eta\ h_{\text{glob}} \frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} (N-1)\psi \leq \psi. \end{aligned}$$

The last inequality holds since $1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} + \eta\ h_{\text{glob}} \frac{T(N-1)^2}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} < 1$ under Assumption 3.3.3. Therefore by (3.3.31), we have

$$\rho_{\boldsymbol{K}^{(m)}} = \rho^* + N\gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}} \max_i \left\{ C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^{i*} \right\}} + \frac{1}{20T^2} \leq \bar{\rho}.$$

Thus the step size condition (3.3.36) is still satisfied for $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(m+1)}$ with $\eta < \eta_0$, since $\rho_{\boldsymbol{K}^{(m)}} \leq \bar{\rho}$ and $\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) \leq \sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$. Therefore, Lemma 3.3.17 can be applied again for the update at round $m+1$ with $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(m+1)}$ to obtain:

$$\sum_{i=1}^{N} \Big(C^i(\boldsymbol{K}^{i,(m+1)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big) \leq (1 - \widehat{\alpha}\eta) \sum_{i=1}^{N} \Big(C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big).$$

For $\epsilon > 0$, provided $M \geq \frac{1}{\widehat{\alpha}\eta} \log \left(\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^{i*})}{\epsilon} \right)$, we have

$$\sum_{i=1}^{N} \Big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big) \leq \epsilon.$$

□

3.4 The Natural Gradient Method with Unknown Parameters

Based on the update rule in (3.3.5), it is straightforward to develop a *model-free version* of the natural policy gradient algorithm using *sampled data*. See Algorithm 3 for

the natural policy gradient method with unknown parameters. In contrast to the model-based case, where the gradient $\nabla C^i(\boldsymbol{K})$ and the matrix Σ_t^i can be calculated directly, these two terms can not be calculated in the model-free setting since the model parameters are unknown. Therefore, we propose to use a zeroth-order optimization method to estimate the gradient and an empirical second moment matrix to estimate the $\mathbb{E}[x_t x_t^\top]$ matrix (see (3.4.1)). Building upon the theories in Section 3.3, high-probability convergence guarantees (linear convergence rate and polynomial sample complexity) for the model-free counterpart can be established in the same way as for the LQR setting in Chapter 2.

Algorithm 3 Natural Policy Gradient Method with Unknown Parameters

Input: Number of iterations M, time horizon T, initial policy $\boldsymbol{K}^{(0)} = (\boldsymbol{K}^{1,(0)}, \cdots, \boldsymbol{K}^{N,(0)})$, step size η, number of trajectories L, smoothing parameters r_i, dimensions $D_i = k_i \times d$.
for $m \in \{1, \ldots, M\}$ **do**
  for $i \in \{1, \ldots, N\}$ **do**
    for $l \in \{1, \ldots, L\}$ **do**
      for $t \in \{0, \ldots, T-1\}$ **do**
        Sample the (sub)-policy at time t: $\widehat{K}_t^{i,l} = K_t^{i,(m-1)} + U_t^{i,l}$ where $U_t^{i,l}$ is drawn uniformly at random over matrices such that $\|U_t^{i,l}\|_F = r_i$.
        Denote $\widehat{c}_t^{i,l}$ as the single trajectory cost of player i with policy $(\widehat{\boldsymbol{K}}_{l,t}^{i,(m-1)}, \boldsymbol{K}^{-i,(m-1)})$ where $\widehat{\boldsymbol{K}}_{l,t}^{i,(m-1)} := (K_0^{i,(m-1)}, \cdots, K_{t-1}^{i,(m-1)}, \widehat{K}_t^{i,l}, K_t^{i,(m-1)}, \cdots, K_{T-1}^{i,(m-1)})$ starting from x_0^l.
        Denote $\widehat{\Sigma}_t^{i,l}$ as the sample of second moment of the state with $\widehat{\Sigma}_t^{i,l} = x_t^{i,l}(x_t^{i,l})^\top$.
      end for
    end for
  end for
  Obtain the estimates of $\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})$ and $\Sigma_t^{\boldsymbol{K}^{(m-1)}}$ for each i and t:

$$\widehat{\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})} = \frac{1}{L}\sum_{l=1}^{L} \frac{D_i}{r_i^2}\, \widehat{c}_t^{i,l}\, U_t^{i,l}, \qquad \widehat{\Sigma}_t^i = \frac{1}{L}\sum_{l=1}^{L} \widehat{\Sigma}_t^{i,l}. \tag{3.4.1}$$

  Update the policies using natural gradient descent rule:

$$K_t^{i,(m)} = K_t^{i,(m-1)} - \eta \widehat{\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})} (\widehat{\Sigma}_t^i)^{-1}. \tag{3.4.2}$$

end for
Return the iterates $\boldsymbol{K}^{(M)} = (\boldsymbol{K}^{1,(M)}, \cdots, \boldsymbol{K}^{N,(M)})$.

3.5 Numerical Experiments

We demonstrate the performance of the natural policy gradient algorithms with two examples. We will specifically focus on the following questions:

- In practice, how fast do the natural policy gradient algorithms converge to the true solution? How sensitive is the natural policy gradient algorithm to the step size and the initial policy?
- Do natural gradient methods converge when Assumption 3.3.3 is violated? How restrictive is Assumption 3.3.3 in practice?
- Can Theorem 3.3.5 provide any guidance on hyper-parameter tuning? In particular, does adding system noise improve the convergence?

The first example is a modified example from [57], in which they show that in the setting of infinite time horizon and deterministic dynamics, the (vanilla) policy gradient algorithms have no guarantees of even local convergence to the Nash equilibria with known parameters. We will show in Section 3.5.1 that, under the same experimental set-up but over a finite time horizon with stochastic dynamics, the natural policy gradient algorithm with known parameters finds the Nash equilibrium with properly chosen initial policies and step sizes. The second example is a two-player LQG with synthetic data (see Section 3.5.2). We will show that the system noise helps the natural policy gradient algorithm with unknown parameters to converge to the Nash equilibrium.

Performance Measure. We use the following *normalized error* to quantify the performance of a given pair of policies $(\boldsymbol{K}^1, \boldsymbol{K}^2)$: for $i = 1, 2$,

$$\text{Normalized error (of player } i) = \frac{C^i(\boldsymbol{K}^1, \boldsymbol{K}^2) - C^i(\boldsymbol{K}^{1*}, \boldsymbol{K}^{2*})}{C^i(\boldsymbol{K}^{1*}, \boldsymbol{K}^{2*})}. \tag{3.5.1}$$

3.5.1 Convergence of the Natural Policy Gradient Algorithm

For policy gradient MARL under the setting of N-player general-sum LQGs, some difficulties in convergence have been identified in some empirical studies. For example, Mazumdar et al. in [57] show by a counterexample that in the setting of infinite time horizon and deterministic dynamics, the (vanilla) policy gradient method avoids the Nash equilibria for a non-negligible subset of problems (with known parameters). In this section, we illustrate that in our setting with finite time horizon and stochastic

dynamics, the natural policy gradient algorithm (with known parameters) finds the (unique) Nash equilibrium under the same experimental set-up as in a modified example given in [57].

Set-up. We set up the model parameters and initialize the policies in the same way as Section 5.1 of [57]. Under the following set of parameters, there exists a unique Nash equilibrium since the sufficient condition in Remark 3.2.5 is satisfied.

1. Parameters: for $t = 1, \cdots, T-1$

$$A_t = \begin{bmatrix} 0.588 & 0.028 \\ 0.570 & 0.056 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad W = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

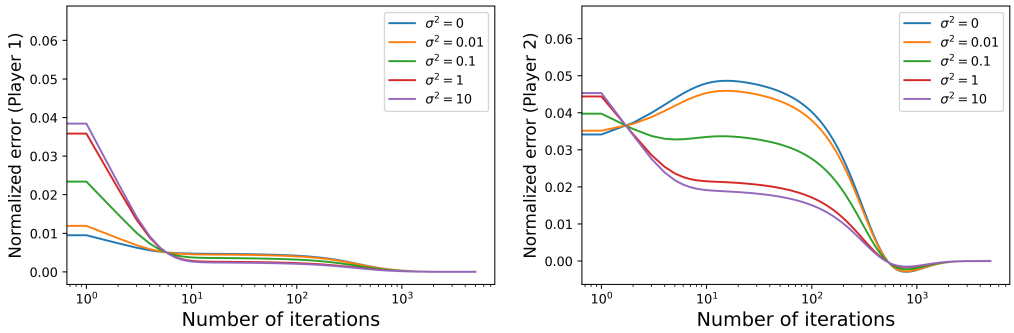
$$Q_T^1 = Q_t^1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_T^2 = Q_t^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.147 \end{bmatrix}, \quad R_t^1(t) = R_t^2(t) = 0.01,$$

   where $\sigma \in \mathbb{R}$ and $T = 10$.

2. Initialization: we assume the initial state distribution to be $[1, 1]^\top$ or $[1, 1.1]^\top$ with probability 0.5 each. We initialize both players' policies $K_t^{i,(0)} = (K_{t0}^{i,(0)}, K_{t1}^{i,(0)})$ such that $(K_{t0}^{i,(0)} - K_{t0}^{i*})^2 + (K_{t1}^{i,(0)} - K_{t1}^{i*})^2 \leq r^2$, where $K_t^{i*} = (K_{t0}^{i*}, K_{t1}^{i*})$ denotes the Nash equilibrium, and r is the radius of the ball centered at K_t^{i*} in which we initialize the policies.

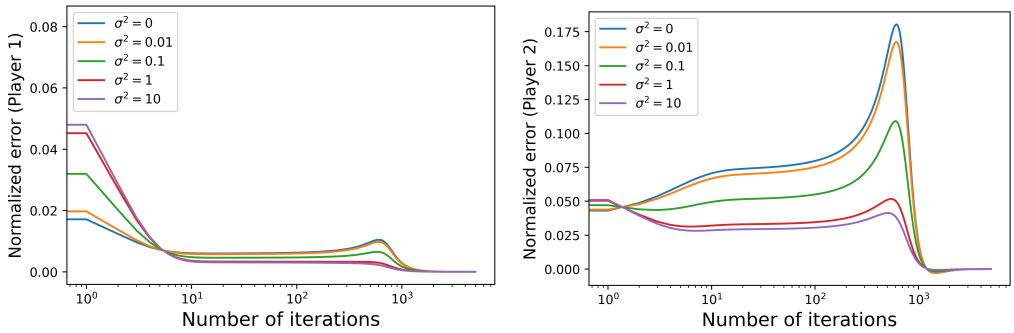
Convergence. The natural policy gradient algorithm shows a reasonable level of accuracy within 1000 iterations (i.e., the normalized error is less than 0.5%) for both players under different levels of system noise σ^2, which ranges from 0 (deterministic dynamics) to 10. See Figure 3.1 for the case where $r = 0.25$ and Figure 3.2 for the case where $r = 0.30$. We observe that when we initialize the policies in a larger neighborhood of the Nash equilibrium, it takes the natural policy gradient algorithm (with the same step size) more iterations to converge. There is a sharp peak in the normalized error for player 2 and this peak diminishes when the noise level increases.

In Figure 3.3, we show the normalized error and the corresponding trajectories of learned policies near the peak observed in Figure 3.2 under $\sigma = 0$. We denote by $K_0^i = (K_{00}^i, K_{01}^i)$ the learned policy at $t = 0$ of player i. The peak period (iterations 300-900) is indicated in grey in Figures 3.3a-3.3c, and the trajectories of learned policy K_0^i for the rest of the whole 10000 iterations are indicated in red in Figures 3.3b-3.3c.



(a) Player 1. (b) Player 2.

Figure 3.1: Normalized error under different σ^2 when $r = 0.25$ ($\eta_1 = \eta_2 = 0.1$ and $N = 5000$).

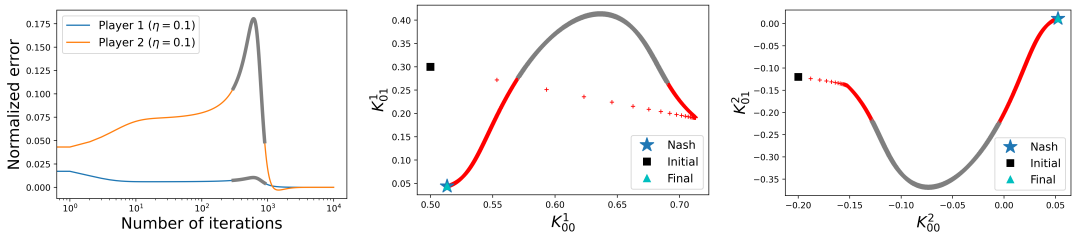


(a) Player 1. (b) Player 2.

Figure 3.2: Normalized error under different σ^2 when $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$ and $M = 5000$).

The natural policy gradient algorithm overshoots in the first few iterations but detects the right direction after about 500 iterations and eventually converges to the Nash equilibrium (see the blue star in Figures 3.3b-3.3c).

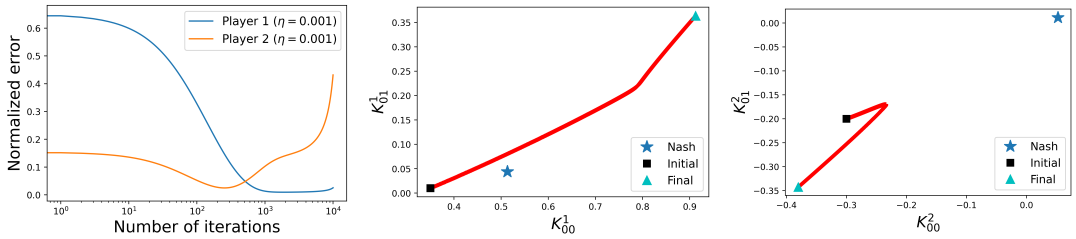
Performance under Deterministic Dynamics. We observe that under carefully chosen initial policies and step sizes, the natural policy gradient converges to the Nash equilibrium even with deterministic state dynamics ($\sigma = 0$). We first show the case when the natural policy gradient algorithm diverges with $r = 0.42$ and $\eta_1 = \eta_2 = 0.001$ in Figure 3.4 (a trajectory of 10000 iterations is indicated in red). However, either by adjusting the step size to $\eta_2 = 0.01$ (see Figure 3.5), or by initializing the policies from a smaller neighbourhood around the Nash equilibrium (see Figure 3.6), the natural policy gradient method converges to the Nash equilibrium. This further demonstrates that the theoretical result, along with its assumptions, in Theorem 3.3.5 could provide



(a) Normalized error. (b) Trajectory of K_0^1. (c) Trajectory of K_0^2.

Figure 3.3: Normalized error and trajectories of learned policies with $\sigma^2 = 0$ and $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$, $M = 10000$). The peak period between iterations 300-900 is indicated in grey.

insightful guidance on how to tune the hyper-parameters for practical examples.



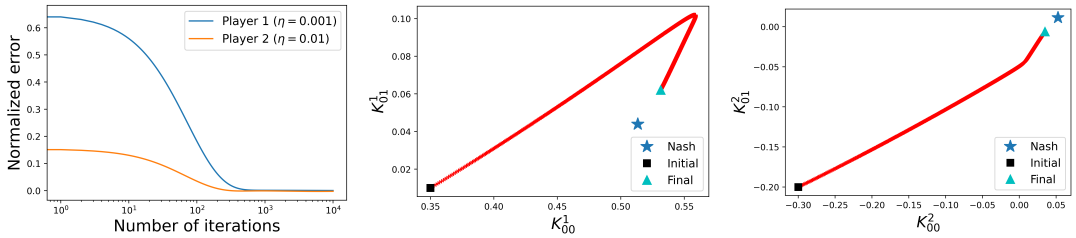
(a) Normalized error. (b) Trajectory of K_0^1. (c) Trajectory of K_0^2.

Figure 3.4: Performance of the natural policy gradient algorithm with $r = 0.42$ and $\eta_1 = \eta_2 = 0.001$ ($M = 10000$).

3.5.2 Effect of the System Noise

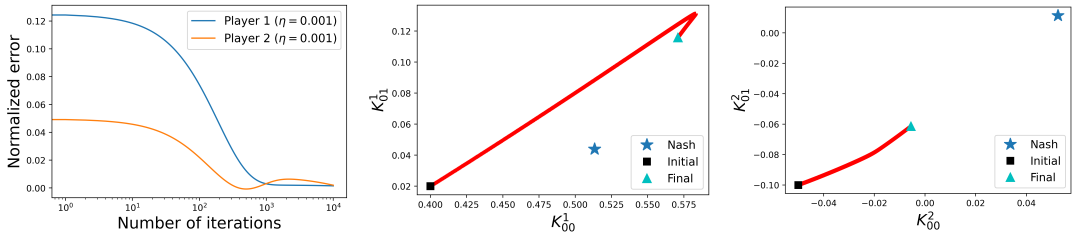
As illustrated in the theoretical analysis in Section 3.3, the system noise plays an important role in the convergence guarantee of the natural policy gradient algorithm. To test the sensitivity of the performance of this algorithm to the level of system noise, we apply the natural policy gradient algorithm with unknown parameters to a two-player LQG example with synthetic data consisting of a two-dimensional state variable and a one-dimensional control variable. The model parameters (except the level of noise σ^2 in W which we will discuss later) are randomly picked such that the conditions for our LQG framework are satisfied.

Set-up. We perform the natural policy gradient algorithm with synthetic data given as follows.



(a) Normalized error. (b) Trajectory of K_0^1. (c) Trajectory of K_0^2.

Figure 3.5: Performance of the natural policy gradient algorithm with $r = 0.42$, $\eta_1 = 0.001$, and $\eta_2 = 0.01$ ($M = 10000$).



(a) Normalized error. (b) Trajectory of K_0^1. (c) Trajectory of K_0^2.

Figure 3.6: Performance of the natural policy gradient algorithm with $r = 0.16$ and $\eta_1 = \eta_2 = 0.001$ ($M = 10000$).

1. Parameters: for $t = 1, \cdots, T-1$,

$$A_t = \begin{bmatrix} 0.588 & 0.28 \\ 0.57 & 0.56 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \quad W = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

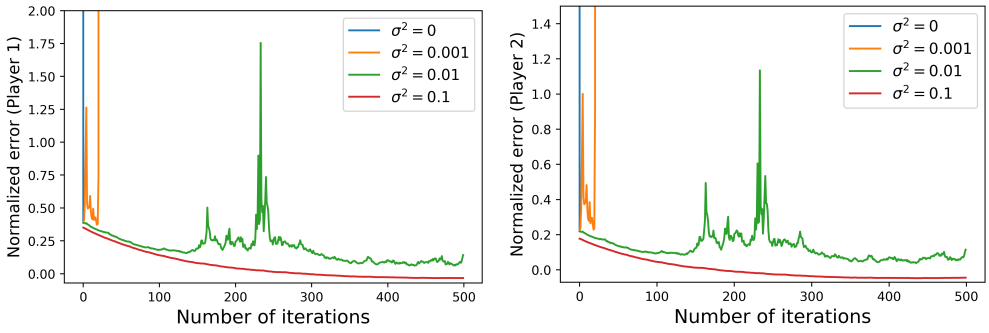
$$Q_T^1 = Q_t^1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_T^2 = Q_t^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.3 \end{bmatrix}, \quad R_t^1 = R_t^2 = 1,$$

where $\sigma \in \mathbb{R}_+$ and $T = 5$. The smoothing parameter is $r_1 = r_2 = 0.5$ and the number of trajectories is $L_1 = L_2 = 200$.

2. Initialization: we assume $x_0 = (x_0^1, x_0^2)$ with x_0^2 and x_0^2 independent and sampled from $\mathcal{N}(10, 2)$ and $\mathcal{N}(12, 3)$ respectively. The initial policy for player 1: $\boldsymbol{K}^1 \in \mathbb{R}^{1\times 10} = (K_0^{1,(0)}, \cdots, K_{T-1}^{1,(0)})$ with $K_t^{1,(0)} = [0.5, 0.5]$ for all t. The initial policy for player 2: $\boldsymbol{K}^2 \in \mathbb{R}^{1\times 10} = (K_0^{2,(0)}, \cdots, K_{T-1}^{2,(0)})$ with $K_t^{2,(0)} = [0.3, 0.15]$ for all t.

Convergence. To show that (even a low level of) the system noise can indeed help the natural policy gradient algorithm to find the Nash equilibrium, we vary the value

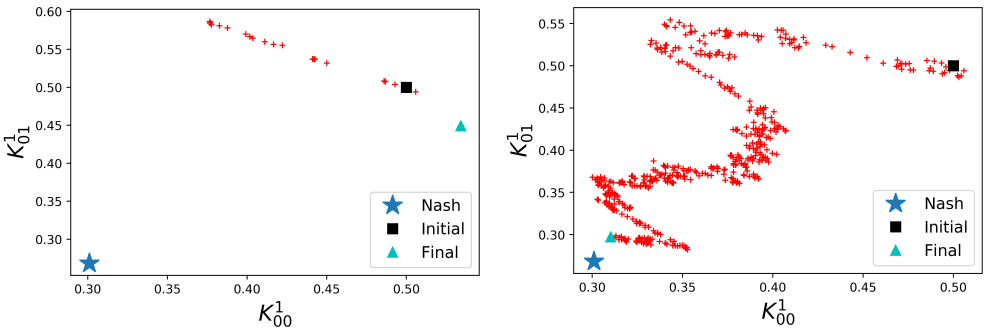
of σ^2 from 0 to 0.1 and show the normalized error for different values of σ^2 in Figure 3.7. The natural policy gradient algorithm diverges when $\sigma^2 \leq 0.0001$ for both players, and it starts to converge with large fluctuations when $\sigma^2 = 0.01$. When $\sigma^2 = 0.1$, the algorithm shows a reasonable accuracy within 300 iterations (that is the normalized error is less than 5%) for both players without fluctuations. It is worth pointing out that in fact, although Assumption 3.3.3 is violated when $\sigma^2 = 0.1$, the natural policy gradient algorithm converges to the Nash equilibrium. Finally, it is possible to make the algorithm converge when $\sigma^2 \leq 0.0001$ by adjusting the initial policy and the step size, as shown in Section 3.5.1. In practice we may not be able to change the variance of w_t directly, however the level of system noise can also be increased by adding (Gaussian) explorations to the agents' policies. See [43, 86] for more discussion of Gaussian exploration.



(a) Player 1. (b) Player 2.

Figure 3.7: Performance of the natural policy gradient algorithm with unknown parameters ($\eta_1 = \eta_2 = 0.0005$).

Trajectories of Learned Policies. We show the trajectories of the learned policy at $t = 0$ of player 1 by performing the natural policy gradient algorithm with unknown parameters under $\sigma^2 = 0.001$ and $\sigma^2 = 0.1$ in Figure 3.8 (with 500 iterations). In the case of a lower level of system noise ($\sigma^2 = 0.001$), the natural policy gradient does not detect the right direction towards the Nash equilibrium with $\eta_1 = \eta_2 = 0.0005$ in Figure 3.8a, whereas when the level of noise is increased to $\sigma^2 = 0.1$, the learned policy approaches the target within 500 iterations with the same step size.



(a) $\sigma^2 = 0.001$ (b) $\sigma^2 = 0.1$.

Figure 3.8: Trajectory (indicated in red) of learned policy $K_0^1 = (K_{00}^1, K_{01}^1)$ ($\eta_1 = \eta_2 = 0.0005$, $M = 500$).

Chapter 4

Bargaining Problems

Bargaining is an essential part of many human activities in the real world. As mentioned in Section 1.1, this thesis is motivated by bargaining and the case of interest to us is one in which two parties, the buyer and the seller, enter into negotiation rounds before a deadline, exchanging a sequence of offers in order to agree on the price (or value) of a project. Negotiations stop when both parties agree on the price of the project and they try to reach an agreement before the deadline. We aim to find their optimal strategies for minimising their negotiation costs. In this chapter, we construct LQ models for bargaining in two cases: (1) when model parameters are known to the players while they only have access to a noisy version of the true state, and (2) when two players do not know the model parameters and they can learn the optimal strategy through repeated negotiations with each other. We will start with an LQG model with partial information and derive the Nash equilibrium strategies in Section 4.1. We will then apply the framework discussed in Chapter 3 to construct another LQG model with unknown parameters and learn the Nash equilibrium through natural policy gradient methods in Section 4.2.

4.1 The LQG Model with Partial Information

When a seller and a buyer negotiate over a price of a project, the target price each player wants to achieve depends on the value of the project. However, the players have access to different information sets and thus will have different opinions on what the value of the project would be. To model this situation, in this section we consider an LQG bargaining model with partial information, where players update their belief of

the true state through Kalman filtering and determine their controls based on their state estimates. We will set up a two-player general-sum LQG model in Section 4.1.1, where players have a noisy observation of the true state. We will then consider the general framework of such games in Section 4.1.2 and derive the Dynamic Programming Principle (DPP) and the Nash equilibrium in Section 4.1.3. We finally present numerical experiments in 4.1.4 where we compare the Nash equilibrium strategies under partial information with the case of full information.

4.1.1 A Two-player LQG Model with Kalman Filter

Let us consider a game with two players, a buyer B and a seller S, who aim to reach an agreement on the value (or the price) of a project. The negotiation takes place over a finite period of time T. At each timestamp t, let $x_t^B \in \mathbb{R}$ be the price offered by the buyer and $x_t^S \in \mathbb{R}$ be the price offered by the seller. The dynamics of the offers follow

$$\begin{aligned} x_{t+1}^B &= x_t^B + u_t + \epsilon_t^B, \\ x_{t+1}^S &= x_t^S + v_t + \epsilon_t^S, \end{aligned}$$

with initial values x_0^B and x_0^S. Here $u_t \in \mathbb{R}$ is the change of the buyer's offers and $v_t \in \mathbb{R}$ is the change of the seller's offers at time t. In this chapter we also focus on the class of Markovian controls. Here ϵ_t^B and ϵ_t^S are i.i.d. random variables in both parties offers with $\epsilon_t^B \sim \mathcal{N}(0, W^B)$ and $\epsilon_t^S \sim \mathcal{N}(0, W^S)$, respectively. We note that ϵ_t^B and ϵ_t^S serve as regularization terms to guarantee the non-degeneracy of the state noise. Another way of thinking about it is to consider u_t and v_t as the *intended* change of their offers when players can not completely control the differences between their offers (for example, due to some external restrictions).

We further assume that the true value of the project $p_t \in \mathbb{R}$ (unknown to both buyer and seller) follows

$$p_{t+1} = p_t + w_t, \tag{4.1.1}$$

where $\{w_t\}_{t=0}^{T-1}$ is a sequence of i.i.d. Gaussian random variables with zero mean and variance W.

Both the buyer and the seller do not have access to the true value of the project. Instead, they could observe a noisy version of the true price with their private information.

Information Structure. At time $t = 0$, player B believes that the initial price

$$p_0 \sim \mathcal{N}(\widehat{p}_0^B, W_0^B), \tag{4.1.2}$$

and thereafter player B observes the following noisy price signal

$$z_{t+1}^B = p_{t+1} + w_{t+1}^B, \quad w_{t+1}^B \sim \mathcal{N}(0, G^B), \quad t = 0, 1, \cdots, T-2, \tag{4.1.3}$$

with $\{w_t^B\}_{t=1}^{T-1}$ a sequence of i.i.d. random variables. Similarly, player S believes that the initial price

$$p_0 \sim \mathcal{N}(\widehat{p}_0^S, W_0^S), \tag{4.1.4}$$

and thereafter player S observes the following noise price signal

$$z_{t+1}^S = p_{t+1} + w_{t+1}^S, \quad w_{t+1}^S \sim \mathcal{N}(0, G^S), \quad t = 0, 1, \cdots, T-2. \tag{4.1.5}$$

with $\{w_t^S\}_{t=1}^{T-1}$ a sequence of i.i.d. random variables that are independent from $\{w_t^B\}_{t=1}^{T-1}$.

Both players make their decisions based on the public and private information available to them. We write $\mathcal{Z}_t^B = \{z_s^B\}_{s=1}^t$ and $\mathcal{Z}_t^S = \{z_s^S\}_{s=1}^t$ as the private signals the buyer and the seller receives up to time t ($1 \le t \le T$), respectively. Further denote respectively $\mathcal{U}_t = \{u_s\}_{s=1}^t$ and $\mathcal{V}_t = \{v_s\}_{s=1}^t$ the control history from the buyer and seller up to time t.

We assume $\mathcal{H}_t^B$ is the information (or history) available to the buyer and $\mathcal{H}_t^S$ is the information available to the seller at time t for them to make decisions, where $\mathcal{H}_t^B$ and $\mathcal{H}_t^S$ follow:

$$\mathcal{H}_t^B = \{\widehat{p}_0^B, W_0^S, W_0^B\} \cup \mathcal{Z}_t^B \cup \mathcal{U}_{t-1} \cup \mathcal{V}_{t-1}, \tag{4.1.6}$$

$$\mathcal{H}_t^S = \{\widehat{p}_0^S, W_0^S, W_0^B\} \cup \mathcal{Z}_t^S \cup \mathcal{U}_{t-1} \cup \mathcal{V}_{t-1}. \tag{4.1.7}$$

Note that the covariance matrices $\{W_0^S, W_0^B\}$ are available to both players.

Posterior Update. We assume that each player updates their state estimate using the Kalman filter, and the updates do not directly involve the innovation of the other player (to avoid the other player's manipulation). With the prior distribution following (4.1.2), the buyer's posterior density $f_t^B(\cdot) = \mathbb{P}[p_t = \cdot | \mathcal{H}_t^B]$ is the density function of

a Gaussian distribution [50, Theorem 7.2.21] where the mean $\widehat{p}_t^B$ and the variance $\widehat{\Sigma}_t^B$ satisfy the following recursion relation:

$$\widehat{p}_{t+1}^B = \widehat{p}_t^B + L_{t+1}(z_{t+1}^B - \widehat{p}_t^B), \tag{4.1.8}$$

with initial condition $\widehat{p}_0^B$, and

$$\widehat{\Sigma}_{t+1}^B = (1 - L_{t+1}^B)(\widehat{\Sigma}_t^B + W), \tag{4.1.9}$$

with initial condition $\widehat{\Sigma}_0^B = W_0^B$. Here

$$L_t^B \;=\; (\widehat{\Sigma}_{t-1}^B + W)\left((\widehat{\Sigma}_{t-1}^B + W) + G^B\right)^{-1}.$$

Similarly the seller's posterior density is Gaussian with mean $\widehat{p}_t^B$ and variance $\widehat{\Sigma}_t^E$ where

$$\widehat{p}_{t+1}^S = \widehat{p}_t^S + L_{t+1}(z_{t+1}^S - \widehat{p}_t^S), \tag{4.1.10}$$

with initial condition $\widehat{p}_0^S$, and

$$\widehat{\Sigma}_{t+1}^S = (1 - L_{t+1}^S)(\widehat{\Sigma}_t^S + W), \tag{4.1.11}$$

with initial condition $\widehat{\Sigma}_0^S = W_0^S$ and

$$L_t^S = (\widehat{\Sigma}_{t-1}^S + W)\left((\widehat{\Sigma}_{t-1}^S + W) + G^S\right)^{-1}.$$

Cost Function. We formulate the buyer's and seller's objective in the game as

$$\begin{aligned}
&\min_{\{u_t\}_{t=0}^{T-1}} J^B(x_0^B, x_0^S; \widehat{p}_0^B, W_0^B) := \\
&\min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[\alpha_B\left(x_T^B - x_T^S\right)^2 + \beta_B\left(x_T^B - (1-c_B)p_T\right)^2 + \sum_{t=0}^{T-1} R_t^B(u_t)^2 \;\middle|\; \mathcal{H}_0^B\right], \\
&\min_{\{v_t\}_{t=0}^{T-1}} J^S(x_0^B, x_0^S; \widehat{p}_0^S, W_0^S) := \\
&\min_{\{v_t\}_{t=0}^{T-1}} \mathbb{E}\left[\alpha_S\left(x_T^B - x_T^S\right)^2 + \beta_S\left(x_T^S - (1+c_S)\,p_T\right)^2 + \sum_{t=0}^{T-1} R_t^S(v_t)^2 \;\middle|\; \mathcal{H}_0^S\right],
\end{aligned}$$

where $c_B \in (0,1)$ and $c_S > 0$ are the scalars that determine the buyer's and the seller's target price at terminal time T. $\alpha_B > 0$ and $\alpha_S > 0$ are the penalties of not reaching an agreement, and β_B and β_S are the penalties of not approaching their target prices. The quadratic terms $\alpha_S \left(x_T^B - x_T^S\right)^2$ and $\alpha_B \left(x_T^B - x_T^S\right)^2$ can be viewed as a relaxation of the hard constraint $x_T^B = x_T^S$. $R_t^B > 0$ and $R_t^S > 0$ are the cost parameters on adjusting the offer price at each time step. Since the control of each player represents the change in his offers, thus it measures the concession the player makes at each time. Therefore the terms $R_t^B(u_t)^2$ and $R_t^S(v_t)^2$ represent the penalty on the concession.

Both players have incentives to reach an agreement on the project at terminal time T. The desire to reach this agreement is characterized by the value of α_B and α_S, which may take different values for buyer and seller. The hard constraint $x_T^B = x_T^S$ can be recovered by letting α_B and α_S tend to infinity. The seller wants to sell the project higher than his *estimation* of the true price. Namely, the targeted price of the seller is set as $(1+c_S)\mathbb{E}[p_T|\mathcal{H}_T^S] = (1+c_S)\widehat{p}_T^S$. Similarly, the buyer has an incentive to purchase at a price lower than his estimated true price, which is denoted as $(1-c_B)\mathbb{E}[p_T|\mathcal{H}_T^B] = (1-c_B)\widehat{p}_T^B$. In practice, we may expect c_B and c_S to be some small positive numbers.

4.1.2 General Setting

Motivated by the example in Section 4.2, we consider a more general two-player game under partial information where controls from both players will have impacts on the dynamics which are not fully observable.

Let us consider a general setting for a game with two players, P and E. In the bargaining game player P corresponds to the buyer B and player E corresponds to the seller S. The state dynamics $x_t \in \mathbb{R}^n$ takes a linear form ($0 \le t \le T-1$):

$$x_{t+1} = A_t x_t + B_t u_t + C_t v_t + \Gamma_t w_t, \tag{4.1.12}$$

with initial value x_0 drawn from a Gaussian distribution, and the (Markovian) controls of P and E are $u_t \in \mathbb{R}^m$ and $v_t \in \mathbb{R}^k$, respectively. Here, for each t, the noise $w_t \in \mathbb{R}^d$ is i.i.d. sampled from $\mathcal{N}(0, W)$ with $W \in \mathbb{R}^{d\times d}$, $A_t \in \mathbb{R}^{n\times n}$, $B_t \in \mathbb{R}^{n\times m}$, $C_t \in \mathbb{R}^{n\times k}$, and $\Gamma_t \in \mathbb{R}^{n\times d}$.

Information Structure. At time $t = 0$, player P believes that the initial price is drawn from a Gaussian distribution

$$x_0 \sim \mathcal{N}(\widehat{x}_0^P, W_0^P), \tag{4.1.13}$$

and thereafter player P observes the following noisy price signal $z_t^P \in \mathbb{R}^p$:

$$z_{t+1}^P = H_{t+1}^P \, x_{t+1} + \, w_{t+1}^P, \quad w_{t+1}^P \sim \mathcal{N}(0, G^P), \quad t = 0, 1, \cdots, T-2, \tag{4.1.14}$$

with $\{w_t^P\}_{t=0}^{T-1}$ a sequence of i.i.d. random variables. Here $G^P \in \mathbb{R}^{p \times p}$ and $H_{t+1}^P \in \mathbb{R}^{p \times n}$. Similarly, player E believes that the initial price is drawn from a Gaussian distribution

$$x_0 \sim \mathcal{N}(\widehat{x}_0^E, W_0^E), \tag{4.1.15}$$

and thereafter player E observes the following noisy price signal $z_t^E \in \mathbb{R}^q$:

$$z_{t+1}^E = H_{t+1}^E \, x_{t+1} + \, w_{t+1}^E, \quad w_{t+1}^E \sim \mathcal{N}(0, G^E), \quad t = 0, 1, \cdots, T-2. \tag{4.1.16}$$

with $\{w_t^E\}_{t=0}^{T-1}$ a sequence of i.i.d. random variables. For simplicity we assume that $\{w_t^E\}_{t=0}^{T-1}$ are independent from $\{w_t^P\}_{t=0}^{T-1}$. In addition, we have $G^E \in \mathbb{R}^{q \times q}$ and $H_{t+1}^E \in \mathbb{R}^{q \times n}$.

We follow [65] to define the game with perfect, imperfect, and partial information:

(a) If the observation matrices H_t^P and H_t^E are the identity matrix and there is no measurement noise, that is, the covariance matrices $G^P = 0$ and $G^E = 0$, we have a game with perfect information. Therefore the players' observation is $z_t^E = z_t^P = x_t$.

(b) If the observation matrices H_t^P and H_t^E are the identity matrix and there is measurement noise, we have a game with imperfect information. Therefore, the players' observations are of the form

$$\begin{aligned} z_t^P &= x_t + w_t^P, \\ z_t^E &= x_t + w_t^E. \end{aligned}$$

(c) If the observation matrices are not the identity and there is measurement noise,

as in

$$\begin{aligned} z_t^P &= H_t^P x_t + w_t^P, \\ z_t^E &= H_t^E x_t + w_t^E, \end{aligned}$$

we have a game with partial information.

Similar to Section 4.1.1, we write $\mathcal{Z}_t^P = \{z_s^P\}_{s=1}^t$ and $\mathcal{Z}_t^E = \{z_s^E\}_{s=1}^t$ for the private signals players P and E receive up to time t ($1 \leq t \leq T$), respectively. Let $\mathcal{U}_t = \{u_s\}_{s=1}^t$ and $\mathcal{V}_t = \{v_s\}_{s=1}^t$ denote the control history from the buyer and seller up to time t, respectively.

We assume $\mathcal{H}_t^P$ is the information (or history) available to player P and $\mathcal{H}_t^E$ is the information available to player E for them to make decisions at time t, where $\mathcal{H}_t^P$ and $\mathcal{H}_t^E$ follow:

$$\mathcal{H}_t^P = \{\widehat{x}_0^P, W_0^P, W_0^E\} \cup \mathcal{Z}_t^P \cup \mathcal{U}_{t-1} \cup \mathcal{V}_{t-1}, \tag{4.1.17}$$

$$\mathcal{H}_t^E = \{\widehat{x}_0^E, W_0^P, W_0^E\} \cup \mathcal{Z}_t^E \cup \mathcal{U}_{t-1} \cup \mathcal{V}_{t-1}. \tag{4.1.18}$$

Note that the covariance matrices $\{W_0^P, W_0^E\}$ are known to both players.

Cost Function. We consider a non-zero sum game between player P and player E where they strive to minimise their own cost functions:

$$\min_{\{u_t\}_{t=0}^{T-1}} J^P(\widehat{x}_0^P, W_0^P) := \min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[x_T^\top Q_T^P x_T + \sum_{t=0}^{T-1} \left(x_t^\top Q_t^P x_t + u_t^\top R_t^P u_t \right) \Big| \mathcal{H}_0^P \right] \tag{4.1.19}$$

$$\min_{\{v_t\}_{t=0}^{T-1}} J^E(\widehat{x}_0^E, W_0^E) := \min_{\{v_t\}_{t=0}^{T-1}} \mathbb{E}\left[x_T^\top Q_T^E x_T + \sum_{t=0}^{T-1} \left(x_t^\top Q_t^E x_t + v_t^\top R_t^E v_t \right) \Big| \mathcal{H}_0^E \right] \tag{4.1.20}$$

with Q_t^P, Q_t^E ($t = 0, \ldots, T$) positive semi-definite, and R_t^P, R_t^E ($t = 0, \ldots, T-1$) positive definite.

We note that in the literature, the LQG with partial information given by (4.1.12)-(4.1.13)-(4.1.14)-(4.1.16)-(4.1.19)-(4.1.20) is also referred to as the *linear quadratic Gaussian game* (LQGG).

For the well-definedness of the game, we summarise the assumptions on the model parameters, initial state, and noise in Assumption 4.1.1.

Assumption 4.1.1 (Parameters, Initial State, and Noise)**.** *Assume for $i = P, E$,*

1. *The initial state x_0 is Gaussian such that $\mathbb{E}[x_0 x_0^\top]$ is positive definite;*

2. *$\{w_t\}_{t=0}^{T-1}$ and $\{w_t^i\}_{t=1}^{T-1}$ are zero-mean, i.i.d. Gaussian and independent from x_0 such that $\mathbb{E}[w_t w_t^\top] = W$ and $\mathbb{E}[w_t^i (w_t^i)^\top] = G^i$ are positive definite, and they are independent of each other;*

3. *The cost matrices Q_t^i $(t = 0, 1, \ldots, T)$ are positive semi-definite, and R_t^i $(t = 0, 1, \ldots, T-1)$ are positive definite.*

Sufficient Statistics (Decentralized Kalman Filtering). The sufficient statistics of player i $(i = E, P)$ at decision time $t = 0$ is $x_0 \sim N(\widehat{x}_0^i, W_0^i)$. For time $1 \le t \le T-1$, the distribution of the state x_t calculated by player i conditioning on the private information available to him at time t follows ([79, (5.3-39)-(5.3-42)] and [65]):

$$x_t \sim \mathcal{N}(\widehat{x}_t^i, \widehat{\Sigma}_t^i), \tag{4.1.21}$$

where

$$\begin{aligned}
\left(\widehat{x}_t^i\right)^- &= A_{t-1}\widehat{x}_{t-1}^i + B_{t-1}u_{t-1} + C_{t-1}v_{t-1}, && (4.1.22)\\
\left(\widehat{\Sigma}_t^i\right)^- &= A_{t-1}\widehat{\Sigma}_{t-1}^i A_{t-1}^\top + \Gamma_{t-1} W \Gamma_{t-1}^\top, && (4.1.23)\\
K_t^i &= \left(\widehat{\Sigma}_t^i\right)^- (H_t^i)^\top \left[H_t^i \left(\widehat{\Sigma}_t^i\right)^- (H_t^i)^\top + G^i\right]^{-1}, && (4.1.24)\\
\widehat{x}_t^i &= \left(\widehat{x}_t^i\right)^- + K_t^i \left[z_t^i - H_t^i \left(\widehat{x}_t^i\right)^-\right], && (4.1.25)\\
\widehat{\Sigma}_t^i &= \left(I - K_t^i H_t^i\right)\left(\widehat{\Sigma}_t^i\right)^-, && (4.1.26)
\end{aligned}$$

with initial condition $\widehat{x}_0^i = \widehat{x}_0^i$ and $\widehat{\Sigma}_0^i = W_0^i$. It is easy to see that $\left((\widehat{x}_t^i)^-, \left(\widehat{\Sigma}_t^i\right)^-\right)$ and $\left(\widehat{x}_t^i, \widehat{\Sigma}_t^i\right)$ are the estimates before and after the information update [79, (5.3-35)-(5.3-38)]:

$$\begin{aligned}
\mathbb{E}[x_t^i | \mathcal{H}_{t-1}] &= (\widehat{x}_t^i)^-,\\
\mathbb{E}[x_t^i | \mathcal{H}_t] &= (\widehat{x}_t^i),\\
\mathbb{E}\left[\left(x_t^i - (\widehat{x}_t^i)^-\right)\left(x_t^i - (\widehat{x}_t^i)^-\right)^\top \,\middle|\, \mathcal{H}_t\right] &= \left(\widehat{\Sigma}_t^i\right)^-,\\
\mathbb{E}\left[\left(x_t^i - (\widehat{x}_t^i)\right)\left(x_t^i - (\widehat{x}_t^i)\right)^\top \,\middle|\, \mathcal{H}_t\right] &= \left(\widehat{\Sigma}_t^i\right).
\end{aligned}$$

Therefore for the LQGG (4.1.12)-(4.1.13)-(4.1.14)-(4.1.16)-(4.1.19)-(4.1.20), the sufficient statistics for players E and P to make decisions at time t are $(\widehat{x}_t^P, \widehat{\Sigma}_t^P)$ and

$(\widehat{x}_t^E, \widehat{\Sigma}_t^E)$. The goal is to find the Nash equilibrium solution of the form $u_t^* = \gamma_t^{P*}(\widehat{x}_t^P, \widehat{\Sigma}_t^P)$ and $v_t^* = \gamma_t^{E*}(\widehat{x}_t^E, \widehat{\Sigma}_t^E)$ for $0 \leq t \leq T-1$, where γ_t^{P*} and γ_t^{E*} are strategy functions which map the estimated state and covariance to the Nash equilibrium solution. Since the expressions of $\widehat{\Sigma}_t^P$ and $\widehat{\Sigma}_t^E$ only involve model parameters which can be calculated directly, we write

$$u_t^* = \gamma_t^{P*}(\widehat{x}_t^P), \quad v_t^* = \gamma_t^{E*}(\widehat{x}_t^E) \tag{4.1.27}$$

with a slight abuse of notation.

4.1.2.1 Preliminary Results with Known Model Parameters

Player P's view on $\widehat{x}^E$. Both players update their belief of the true state process x_t but they do not have access to each other's estimate. Therefore, to solve the LQGG, it is important to first understand player P's view on $\widehat{x}^E$ as well as player E's view on $\widehat{x}^P$. When the model parameters (A_t, B_t, C_t, Γ_t, H_t^P, H_t^E, G^P, G^E, Q_t^P, Q_t^E, R_t^P and R_t^E) are available to both players, we have the following results.

Let us start at decision time $t = 0$. Player P models his initial belief $\widehat{x}_0^P$ of the initial state x_0 as

$$\widehat{x}_0^P = x_0 + e_0^P, \tag{4.1.28}$$

where x_0 is the true state and e_0^P is player P's estimation error with $e_0^P \sim \mathcal{N}(0, W_0^P)$ by (4.1.13). In addition, player E's belief $\widehat{x}_0^E$ of the initial state x_0 follows

$$\widehat{x}_0^E = x_0 + e_0^E, \tag{4.1.29}$$

where e_0^E is player E's estimation error with $e_0^E \sim \mathcal{N}(0, W_0^E)$ by (4.1.15). The Gaussian random variables e_0^E and e_0^P are independent by assumption.

From player P's perspective, $\widehat{x}^P$ is known but $\widehat{x}^E$ is a random variable. Subtracting (4.1.28) from (4.1.29), at time $t = 0$ player P believes that player E's state estimate is the random variable

$$\widehat{x}_0^E = \widehat{x}_0^P + e_0^E - e_0^P. \tag{4.1.30}$$

Therefore, from player P's perspective, E's estimate of the initial state x_0 is the Gaus-

sian random variable

$$\widehat{x}_0^E \sim \mathcal{N}(\widehat{x}_0^P, W_0^P + W_0^E). \tag{4.1.31}$$

Thus, at time $t = 0$ player P uses his private information $\widehat{x}_0^P$ and the public information (W_0^P, W_0^E) to calculate the distribution of the sufficient statistic $\widehat{x}_0^E$ of player E. Similarly, as far as player E is concerned, at time $t = 0$ the statistics of the initial state estimate $\widehat{x}_0^P$ of player P are

$$\widehat{x}_0^P \sim \mathcal{N}(\widehat{x}_0^E, W_0^P + W_0^E). \tag{4.1.32}$$

For general time step $0 \leq t \leq T-1$, we have the following result according to [65].

Theorem 4.1.2 ([65]). *For $0 \leq t \leq T-1$, player P's posterior distribution for the state estimate $\widehat{x}_t^E$ of player E is*

$$\widehat{x}_t^E \sim \mathcal{N}(\widehat{x}_t^P, \widehat{\Sigma}_t^{(E,P)}) \tag{4.1.33}$$

where the estimation error covariance matrix $\widehat{\Sigma}_t^{(E,P)}$ is defined as

$$\begin{aligned} \widehat{\Sigma}_t^{(E,P)} &= \mathbb{E}\left[\left(e_t^E - e_t^P\right)\left(e_t^E - e_t^P\right)^\top\right] \\ &= \widehat{\Sigma}_t^P + \widehat{\Sigma}_t^E - \widetilde{\Sigma}_t^{(E,P)} - \left(\widetilde{\Sigma}_t^{(E,P)}\right)^\top, \end{aligned} \tag{4.1.34}$$

in which $\widehat{\Sigma}_0^P = W_0^P$, $\widehat{\Sigma}_0^E = W_0^E$, and $\widetilde{\Sigma}_t^{(E,P)} = \mathbb{E}[e_t^P (e_t^E)^\top]$ satisfies a recursive linear matrix equation which is a Lyapunov equation:

$$\widetilde{\Sigma}_{t+1}^{(E,P)} = \left(I - K_{t+1}^P H_{t+1}^P\right)\left(A_t \widetilde{\Sigma}_t^{(E,P)} A_t^\top + \Gamma_t W \Gamma_t^\top\right)\left(I - K_{t+1}^E H_{t+1}^E\right)^\top \tag{4.1.35}$$

with initial condition $\widetilde{\Sigma}_0^{(E,P)} = 0$. Similarly, player E's posterior distribution for the state estimate $\widehat{x}_t^P$ of player P is

$$\widehat{x}_t^P \sim \mathcal{N}(\widehat{x}_t^E, \widehat{\Sigma}_t^{(E,P)}). \tag{4.1.36}$$

Before we derive the Nash equilibrium strategies for the LQGG, we first assume that the above $\widehat{\Sigma}_t^{(E,P)}$ is positive definite. This assumption is easy to check as $\widehat{\Sigma}_t^{(E,P)}$,

$\widehat{\Sigma}_t^P$, $\widehat{\Sigma}_t^E$, and $\widetilde{\Sigma}_t^{(E,P)}$ are independent of the observation signals and can be calculated offline.

Assumption 4.1.3. *Assume the estimation error covariance matrix* $\widehat{\Sigma}_t^{(E,P)}$ *defined in* (4.1.34) *is positive definite for all* $t = 0, 1, \ldots, T$.

4.1.3 Nash Equilibrium Strategy and Value Function

In this section, we show that the Nash equilibrium policy is unique and linear in the estimated states. Namely, there exist matrices F_t^E and F_t^P such that $\gamma_t^{E*}(\widehat{x}_t^E) = F_t^E \widehat{x}_t^E$ and $\gamma_t^{P*}(\widehat{x}_t^P) = F_t^P \widehat{x}_t^P$.

The derivation is motivated by [65] but the details are different since [65] focuses on a class of zero-sum games.

4.1.3.1 Preliminary: End Game

To start, we analyze the game at the last step $t = T-1$, which is also called the *End Game*.

At time $T-1$, define

$$\begin{aligned}\overline{J}_{T-1}^P\left(u_{T-1}; \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P\right) &:= u_{T-1}^\top R_{T-1}^P u_{T-1} \qquad (4.1.37)\\ &\quad + \mathbb{E}_{x_{T-1}, \widehat{x}_{T-1}^E, w_{T-1}, w_T^P}\left[x_{T-1}^\top Q_{T-1}^P x_{T-1} + x_T^\top Q_T^P x_T \middle| \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P\right],\\ \overline{J}_{T-1}^E\left(v_{T-1}; \widehat{x}_{T-1}^E, \widehat{\Sigma}_{T-1}^E\right) &:= v_{T-1}^\top R_{T-1}^E v_{T-1} \qquad (4.1.38)\\ &\quad + \mathbb{E}_{x_{T-1}, \widehat{x}_{T-1}^P, w_{T-1}, w_T^E}\left[x_{T-1}^\top Q_{T-1}^E x_{T-1} + x_T^\top Q_T^E x_T \middle| \widehat{x}_{T-1}^E, \widehat{\Sigma}_{T-1}^E\right].\end{aligned}$$

Note that player P's optimization problem $\min_{u_{T-1}} \overline{J}_{T-1}^P\left(u_{T-1}; \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P\right)$ depends on v_{T-1} through x_T. Here it is enough to consider $v_{T-1} = g(\widehat{x}_{T-1}^E)$ for some function $g : \mathbb{R}^n \to \mathbb{R}^k$ with $\widehat{x}_{T-1}^E$ a random variable following (4.1.33).

For the first term in the expectation in (4.1.37), since x_{T-1} does not depend on $\widehat{x}_{T-1}^E$, w_{T-1} and w_T^P, we have

$$\mathbb{E}_{x_{T-1}}\left[x_{T-1}^\top Q_{T-1}^P x_{T-1} \middle| \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P\right] = \mathrm{Tr}(Q_{T-1}^P \widehat{\Sigma}_{T-1}^P) + (\widehat{x}_{T-1}^P)^\top Q_{T-1}^P \widehat{x}_{T-1}^P. \quad (4.1.39)$$

Then, by (4.1.12) and $v_{T-1} = g(\widehat{x}_{T-1}^E)$, we have for the second term in (4.1.37)

$$\mathbb{E}_{x_{T-1}, \widehat{x}_{T-1}^E, w_{T-1}, w_T^P}\left[x_T^\top Q_T^P x_T \middle| \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P\right]$$

$$
\begin{aligned}
= \; & (\widehat{x}^P_{T-1})^\top A^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} + \mathrm{Tr}(A^\top_{T-1} Q^P_T A_{T-1} \widehat{\Sigma}^P_{T-1}) + u^\top_{T-1} B^\top_{T-1} Q^P_T B_{T-1} u_{T-1} \\
& + \mathbb{E}_{\widehat{x}^E_{T-1}} \left[g(\widehat{x}^E_{T-1})^\top (C^\top_{T-1} Q^P_T C_{T-1}) g(\widehat{x}^E_{T-1}) \,\middle|\, \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] \\
& + 2 u^\top_{T-1} B^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} + 2 \mathbb{E}_{\widehat{x}^E_{T-1}, x_{T-1}} \left[g(\widehat{x}^E_{T-1})^\top \, C^\top_{T-1} Q^P_T A_{T-1} x_{T-1} | \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] \\
& + 2 u^\top_{T-1} B^\top_{T-1} Q^P_T C_{T-1} \mathbb{E}_{\widehat{x}^E_{T-1}} \left[g(\widehat{x}^E_{T-1}) \, | \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] + \mathrm{Tr} \left(\Gamma^\top_{T-1} Q^P_T \Gamma_{T-1} W \right) . \quad (4.1.40)
\end{aligned}
$$

Player P calculates the expectations with respect to the random variable $\widehat{x}^E_{T-1}$. Similar to (4.1.30), we write $\widehat{x}^E_{T-1} = \widehat{x}^P_{T-1} + \widetilde{w}$, where the random variable $\widetilde{w} \equiv e^E_{T-1} - e^P_{T-1} \sim \mathcal{N}(0, \widehat{\Sigma}^{(E,P)}_{T-1})$. In other words, we have (4.1.33).

Now we calculate the expectations in (4.1.40)

$$
\mathbb{E}_{\widehat{x}^E_{T-1}} \left[g(\widehat{x}^E_{T-1}) \middle| \widehat{x}^P_{T-1} \right] = \mathbb{E}_{\widetilde{w}} \left[g(\widehat{x}^P_{T-1} + \widetilde{w}) \right] . \quad (4.1.41)
$$

Similarly,

$$
\begin{aligned}
& \mathbb{E}_{\widehat{x}^E_{T-1}} \left[g(\widehat{x}^E_{T-1})^\top (C^\top_{T-1} Q^P_T C_{T-1}) g(\widehat{x}^E_{T-1}) \,\middle|\, \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] \\
& \quad = \mathbb{E}_{\widetilde{w}} \left[g(\widehat{x}^P_{T-1} + \widetilde{w})^\top (C^\top_{T-1} Q^P_T C_{T-1}) g(\widehat{x}^P_{T-1} + \widetilde{w}) \right] . \quad (4.1.42)
\end{aligned}
$$

In addition, since $x_{T-1} = \widehat{x}^P_{T-1} - e^P_{T-1}$, the expectation:

$$
\begin{aligned}
& \mathbb{E}_{\widehat{x}^E_{T-1}, x_{T-1}} \left[g(\widehat{x}^E_{T-1})^\top C^\top_{T-1} Q^P_T A_{T-1} x_{T-1} \,\middle|\, \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] \\
= \; & \mathbb{E}_{\widetilde{w}, e^P_{T-1}} \left[g^\top (\widehat{x}^P_{T-1} + \widetilde{w}) C^\top_{T-1} Q^P_T A_{T-1} (\widehat{x}^P_{T-1} - e^P_{T-1}) \right] \\
= \; & \mathbb{E}_{\widetilde{w}} [g^\top (\widehat{x}^P_{T-1} + \widetilde{w})] C^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} \\
& - \mathbb{E}_{e^E_{T-1}, e^P_{T-1}} \left[g^\top (\widehat{x}^P_{T-1} + e^E_{T-1} - e^P_{T-1}) C^\top_{T-1} Q^P_T A_{T-1} e^P_{T-1} \right] \quad (4.1.43)
\end{aligned}
$$

Plugging in (4.1.41)-(4.1.43) into (4.1.40) gives

$$
\begin{aligned}
& \mathbb{E}_{x_{T-1}, \widehat{x}^E_{T-1}, w_{T-1}, w^P_T} \left[x^\top_T Q^P_T x_T \middle| \widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1} \right] \\
= \; & (\widehat{x}^P_{T-1})^\top A^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} + \mathrm{Tr}(A^\top_{T-1} Q^P_T A_{T-1} \widehat{\Sigma}^P_{T-1}) + u^\top_{T-1} B^\top_{T-1} Q^P_T B_{T-1} u_{T-1} \\
& + \mathbb{E}_{\widetilde{w}} \left[g(\widehat{x}^P_{T-1} + \widetilde{w})^\top (C^\top_{T-1} Q^P_T C_{T-1}) g(\widehat{x}^P_{T-1} + \widetilde{w}) \right] \\
& + 2 u^\top_{T-1} B^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} + 2 \mathbb{E}_{\widetilde{w}} [g^\top (\widehat{x}^P_{T-1} + \widetilde{w})] C^\top_{T-1} Q^P_T A_{T-1} \widehat{x}^P_{T-1} \\
& - 2 \mathbb{E}_{e^E_{T-1}, e^P_{T-1}} \left[g^\top (\widehat{x}^P_{T-1} + e^E_{T-1} - e^P_{T-1}) C^\top_{T-1} Q^P_T A_{T-1} e^P_{T-1} \right] \\
& + 2 u^\top_{T-1} B^\top_{T-1} Q^P_T C_{T-1} \mathbb{E}_{\widetilde{w}} \left[g(\widehat{x}^P_{T-1} + \widetilde{w}) \, \right] + \mathrm{Tr} \left(\Gamma^\top_{T-1} Q^P_T \Gamma_{T-1} W \right) . \quad (4.1.44)
\end{aligned}
$$

Player P minimises $\overline{J}^P_{T-1}\left(u_{T-1};\widehat{x}^P_{T-1},\widehat{\Sigma}^P_{T-1}\right)$ with respect to u_{T-1}. The first order condition leads to the following $u^*_{T-1}(\widehat{x}^P_{T-1}|g)$, which is player P's optimal response to player E's strategy g:

$$\begin{aligned} & u^*_{T-1}(\widehat{x}^P_{T-1}|g) = f^*(\widehat{x}^P_{T-1}|g) \\ = \; & -(R^P_{T-1}+B^\top_{T-1}Q^P_TB_{T-1})^{-1}B^\top_{T-1}Q^P_T\left(A_{T-1}\widehat{x}^P_{T-1}+C_{T-1}\mathbb{E}_{\widetilde{w}}\left[g(\widehat{x}^P_{T-1}+\widetilde{w})\right]\right). \end{aligned} \tag{4.1.45}$$

Similarly, player E calculates his optimal response to player's strategy $f(\widehat{x}^P_{T-1})$. Thus, given the information $\widehat{x}^E_{T-1}$, player E maximises his expected payoff $\overline{J}^E_{T-1}\left(v_{T-1};\widehat{x}^E_{T-1},\widehat{\Sigma}^E_{T-1}\right)$. The optimal response v^*_{T-1} follows:

$$\begin{aligned} & v^*_{T-1}(\widehat{x}^E_{T-1}|f) = g^*(\widehat{x}^E_{T-1}|f) \\ = \; & -(R^E_{T-1}+C^\top_{T-1}Q^E_TC_{T-1})^{-1}C^\top_{T-1}Q^E_T\left(A_{T-1}\widehat{x}^E_{T-1}+B_{T-1}\mathbb{E}_{\widetilde{w}}\left[f(\widehat{x}^E_{T-1}-\widetilde{w})\right]\right). \end{aligned} \tag{4.1.46}$$

Hence the Nash equilibrium strategies $f^*(\cdot)$ and $g^*(\cdot)$ of players P and E satisfy the set of two coupled equations (4.1.45)-(4.1.46):

$$f^*(\widehat{x}^P_{T-1}) = -(R^P_{T-1}+B^\top_{T-1}Q^P_TB_{T-1})^{-1}B^\top_{T-1}Q^P_T\left(A_{T-1}\widehat{x}^P_{T-1}+C_{T-1}\mathbb{E}_{\widetilde{w}}\left[g^*(\widehat{x}^P_{T-1}+\widetilde{w})\right]\right), \tag{4.1.47}$$

$$g^*(\widehat{x}^E_{T-1}) = -(R^E_{T-1}+C^\top_{T-1}Q^E_TC_{T-1})^{-1}C^\top_{T-1}Q^E_T\left(A_{T-1}\widehat{x}^E_{T-1}+B_{T-1}\mathbb{E}_{\widetilde{w}}\left[f^*(\widehat{x}^E_{T-1}-\widetilde{w})\right]\right). \tag{4.1.48}$$

For the multivariate Gaussian distribution with zero mean and positive definite covariance $\Sigma\in\mathbb{R}^{n\times n}$, we write the density function as $G(x;\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}}e^{-\frac{1}{2}x^\top\Sigma^{-1}x}$ for $x\in\mathbb{R}^n$. Using the convolution notation, we have

$$\mathbb{E}_{\widetilde{w}}[f(\widehat{x}^E_{T-1}-\widetilde{w})] = \left[f\star G(\widehat{\Sigma}^{E,P}_{T-1})\right](\widehat{x}^E_{T-1}), \tag{4.1.49}$$

$$\mathbb{E}_{\widetilde{w}}[g(\widehat{x}^P_{T-1}+\widetilde{w})] = \left[g\star G(\widehat{\Sigma}^{E,P}_{T-1})\right](\widehat{x}^P_{T-1}), \tag{4.1.50}$$

and hence (4.1.47)-(4.1.48) can be rewritten as

$$f^*(\widehat{x}^P_{T-1})$$

$$= -(R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T \left(A_{T-1} \widehat{x}^P_{T-1} + C_{T-1} \left[g^* \star G(\widehat{\Sigma}^{E,P}_{T-1}) \right] (\widehat{x}^P_{T-1}) \right), \tag{4.1.51}$$

$$g^*(\widehat{x}^E_{T-1})$$
$$= -(R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} C^\top_{T-1} Q^E_T \left(A_{T-1} \widehat{x}^E_{T-1} + B_{T-1} \left[f^* \star G(\widehat{\Sigma}^{E,P}_{T-1}) \right] (\widehat{x}^E_{T-1}) \right). \tag{4.1.52}$$

Plugging (4.1.52) into (4.1.51), for any $x \in \mathbb{R}^n$ we have

$$\begin{aligned} f^*(x) \;=\; & -(R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T A_{T-1} x \\ & +(R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T C_{T-1} (R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} \cdot \\ & \left\{ C^\top_{T-1} Q^E_T A_{T-1} x \star G + C^\top_{T-1} Q^E_T B_{T-1} f^* \star G \star G \right\} \end{aligned}$$

Since Gaussian kernel is self-similar [65], we have $G(\Sigma) \star G(\Sigma) = G(2\Sigma)$ and $x \star G(\Sigma) = x$ for any positive semi-definite Σ. Therefore,

$$\begin{aligned} f^*(\widehat{x}^P_{T-1}) \;=\; & (R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T \cdot \\ & \left[C_{T-1} (R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} C^\top_{T-1} Q^E_T - I \right] A_{T-1} \widehat{x}^P_{T-1} \\ & +(R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} \cdot \\ & B^\top_{T-1} Q^P_T C_{T-1} (R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} C^\top_{T-1} Q^E_T B_{T-1} f^* \star G(2\widehat{\Sigma}^{E,P}_{T-1}), \end{aligned} \tag{4.1.53}$$

for any $\widehat{x}^P_{T-1} \in \mathbb{R}^n$.

Similarly, we have

$$\begin{aligned} g^*(\widehat{x}^E_{T-1}) \;=\; & (R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} C^\top_{T-1} Q^E_T \cdot \\ & \left[B_{T-1} (R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T - I \right] A_{T-1} \widehat{x}^E_{T-1} \\ & +(R^E_{T-1} + C^\top_{T-1} Q^E_T C_{T-1})^{-1} \cdot \\ & C^\top_{T-1} Q^E_T B_{T-1} (R^P_{T-1} + B^\top_{T-1} Q^P_T B_{T-1})^{-1} B^\top_{T-1} Q^P_T C_{T-1} g^* \star G(2\widehat{\Sigma}^{E,P}_{T-1}), \end{aligned} \tag{4.1.54}$$

for any $\widehat{x}^E_{T-1} \in \mathbb{R}^n$.

Linear Policies. Since the Fredholm integral equations of the second kind, (4.1.53) and (4.1.54) are *linear* integral equations with linear forcing terms, the solution $f^*(\cdot)$ of (4.1.53) is linear in $\widehat{x}^P_{T-1}$ and the solution of $g^*(\cdot)$ of (4.1.54) is linear in $\widehat{x}^E_{T-1}$. Therefore, we look for $F^P_{T-1} \in \mathbb{R}^{k \times n}$ and $F^E_{T-1} \in \mathbb{R}^{m \times n}$ such that $f^*(x) = F^P_{T-1} x$ and

$g^*(x) = F^E_{T-1}x$. Plugging these two linear relationships in (4.1.53) and (4.1.54), we have

$$\begin{aligned} F^P_{T-1} &= \Big(I - (R^P_{T-1} + B^\top_{T-1}Q^P_T B_{T-1})^{-1}B^\top_{T-1}Q^P_T C_{T-1}(R^E_{T-1} + C^\top_{T-1}Q^E_T C_{T-1})^{-1}\cdot \\ &\quad C^\top_{T-1}Q^E_T B_{T-1}\Big)^{-1}(R^P_{T-1} + B^\top_{T-1}Q^P_T B_{T-1})^{-1}B^\top_{T-1}Q^P_T \cdot \\ &\quad \left[C_{T-1}(R^E_{T-1} + C^\top_{T-1}Q^E_T C_{T-1})^{-1}C^\top_{T-1}Q^E_T - I\right]A_{T-1}. \end{aligned} \tag{4.1.55}$$

Similarly, we have

$$\begin{aligned} F^E_{T-1} &= \Big(I - (R^E_{T-1} + C^\top_{T-1}Q^E_T C_{T-1})^{-1}C^\top_{T-1}Q^E_T B_{T-1}(R^P_{T-1} + B^\top_{T-1}Q^P_T B_{T-1})^{-1}\cdot \\ &\quad B^\top_{T-1}Q^P_T C_{T-1}\Big)^{-1}(R^E_{T-1} + C^\top_{T-1}Q^E_T C_{T-1})^{-1}C^\top_{T-1}Q^E_T \\ &\quad \left[B_{T-1}(R^P_{T-1} + B^\top_{T-1}Q^P_T B_{T-1})^{-1}B^\top_{T-1}Q^P_T - I\right]A_{T-1}. \end{aligned} \tag{4.1.56}$$

According to (4.1.47)-(4.1.48) and letting $f^*(x) = F^P_{T-1}x$ and $g^*(x) = F^E_{T-1}x$, we can simplify the above optimal policies to

$$F^P_{T-1} = -(R^P_{T-1} + B^\top_{T-1}Q^P_T B_{T-1})^{-1}B^\top_{T-1}Q^P_T\left(A_{T-1} + C_{T-1}F^E_{T-1}\right), \tag{4.1.57}$$

$$F^E_{T-1} = -(R^E_{T-1} + C^\top_{T-1}Q^E_T C_{T-1})^{-1}C^\top_{T-1}Q^E_T\left(A_{T-1} + B_{T-1}F^P_{T-1}\right). \tag{4.1.58}$$

Value Function of the End Game. With the linear policy, we can write (4.1.41), (4.1.42) and (4.1.43) as

$$\mathbb{E}_{\widehat{x}^E_{T-1}}\left[g(\widehat{x}^E_{T-1})\middle|\,\widehat{x}^P_{T-1}\right] = F^E_{T-1}\widehat{x}^P_{T-1}, \tag{4.1.59}$$

$$\begin{aligned} &\mathbb{E}_{x_{T-1}}\left[g(\widehat{x}^E_{T-1})^\top(C^\top_{T-1}Q^P_T C_{T-1})g(\widehat{x}^E_{T-1})\,\middle|\,\widehat{x}^P_{T-1}, \widehat{\Sigma}^P_{T-1}\right] \\ = \;& (\widehat{x}^P_{T-1})^\top(F^E_{T-1})^\top(C^\top_{T-1}Q^P_T C_{T-1})F^E_{T-1}\widehat{x}^P_{T-1} \\ &+ \mathrm{Tr}\left((F^E_{T-1})^\top(C^\top_{T-1}Q^P_T C_{T-1})F^E_{T-1}\widehat{\Sigma}^{E,P}_{T-1}\right), \end{aligned} \tag{4.1.60}$$

and

$$\begin{aligned} &\mathbb{E}_{e^E_{T-1}, e^P_{T-1}}\left[g^\top(\widehat{x}^P_{T-1} + e^E_{T-1} - e^P_{T-1})C^\top_{T-1}Q^P_T A_{T-1}e^P_{T-1}\right] = \\ &-\mathrm{Tr}\left(\widehat{\Sigma}^P_{T-1}A^\top_{T-1}Q^P_T C_{T-1}F^E_{T-1}\right) + \mathrm{Tr}\left((\widetilde{\Sigma}^{(E,P)}_{T-1})^\top A^\top_{T-1}Q^P_T C_{T-1}F^E_{T-1}\right). \end{aligned} \tag{4.1.61}$$

We further define the value functions as

$$\begin{aligned} V_{T-1}^P(\widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P; \widehat{\Sigma}_{T-1}^E, \widetilde{\Sigma}_{T-1}^{(E,P)}) &:= \min_{u_{T-1}} \overline{J}_{T-1}^P \left(u_{T-1}; \widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P \right), \\ V_{T-1}^E(\widehat{x}_{T-1}^E, \widehat{\Sigma}_{T-1}^E; \widehat{\Sigma}_{T-1}^P, \widetilde{\Sigma}_{T-1}^{(E,P)}) &:= \min_{v_{T-1}} \overline{J}_{T-1}^E \left(v_{T-1}; \widehat{x}_{T-1}^E, \widehat{\Sigma}_{T-1}^E \right). \end{aligned}$$

Plugging (4.1.59), (4.1.60) and (4.1.61) into (4.1.44) and taking $u_{T-1}^* = F_{T-1}^P \widehat{x}_{T-1}^P$, we have

$$\begin{aligned} & V_{T-1}^P(\widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P; \widehat{\Sigma}_{T-1}^E, \widetilde{\Sigma}_{T-1}^{(E,P)}) \\ = \; & (\widehat{x}_{T-1}^P)^\top A_{T-1} Q_T^P A_{T-1} \widehat{x}_{T-1}^P + \mathrm{Tr}(A_{T-1}^\top Q_T^P A_{T-1} \widehat{\Sigma}_{T-1}^P) \\ & + (\widehat{x}_{T-1}^P)^\top (F_{T-1}^P)^\top (R_{T-1}^P + B_{T-1}^\top Q_T^P B_{T-1}) F_{T-1}^P \widehat{x}_{T-1}^P \\ & + (\widehat{x}_{T-1}^P)^\top (F_{T-1}^E)^\top C_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \widehat{x}_{T-1}^P + \mathrm{Tr}\left((F_{T-1}^E)^\top C_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \widehat{\Sigma}_{T-1}^{(E,P)} \right) \\ & + 2(\widehat{x}_{T-1}^P)^\top (F_{T-1}^P)^\top B_{T-1}^\top Q_T^P A_{T-1} \widehat{x}_{T-1}^P + 2(\widehat{x}_{T-1}^P)^\top (F_{T-1}^E)^\top C_{T-1}^\top Q_T^P A_{T-1} \widehat{x}_{T-1}^P \\ & + 2\,\mathrm{Tr}\left(\widehat{\Sigma}_{T-1}^P A_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \right) - 2\,\mathrm{Tr}\left((\widetilde{\Sigma}_{T-1}^{(E,P)})^\top A_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \right) \\ & + 2(\widehat{x}_{T-1}^P)^\top (F_{T-1}^P)^\top B_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \widehat{x}_{T-1}^P + \mathrm{Tr}\left(\Gamma_{T-1}^\top Q_T^P \Gamma_{T-1}^P W \right) \\ & + (\widehat{x}_{T-1}^P)^\top Q_{T-1}^P \widehat{x}_{T-1}^P + \mathrm{Tr}(Q_{T-1}^P \widehat{\Sigma}_{T-1}^P), \end{aligned}$$

which is quadratic in $\widehat{x}_{T-1}^P$. Hence we can write the Nash equilibrium value (associated with the Nash equilibrium strategy (4.1.55) and (4.1.56)) as

$$V_{T-1}^P(\widehat{x}_{T-1}^P, \widehat{\Sigma}_{T-1}^P; \widehat{\Sigma}_{T-1}^E, \widetilde{\Sigma}_{T-1}^{(E,P)}) = \left(\widehat{x}_{T-1}^P\right)^\top \Pi_{T-1}^P \widehat{x}_{T-1}^P + c_{T-1}^P \tag{4.1.62}$$

with

$$\begin{aligned} \Pi_{T-1}^P &= A_{T-1} Q_T^P A_{T-1} + (F_{T-1}^P)^\top (R_{T-1}^P + B_{T-1}^\top Q_T^P B_{T-1}) F_{T-1}^P \\ &\quad + (F_{T-1}^E)^\top C_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E + 2(F_{T-1}^P)^\top B_{T-1}^\top Q_T^P A_{T-1} \\ &\quad + 2(F_{T-1}^E)^\top C_{T-1}^\top Q_T^P A_{T-1} + 2(F_{T-1}^P)^\top B_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E + Q_{T-1}^P \\ &= (F_{T-1}^P)^\top R_{T-1}^P F_{T-1}^P + \left(A_{T-1} + B_{T-1} F_{T-1}^P + C_{T-1} F_{T-1}^E \right)^\top Q_T^P \cdot \\ &\quad \left(A_{T-1} + B_{T-1} F_{T-1}^P + C_{T-1} F_{T-1}^E \right) + Q_{T-1}^P, \end{aligned} \tag{4.1.63}$$

and

$$c_{T-1}^P$$

$$
\begin{aligned}
&= \operatorname{Tr}(A_{T-1}^\top Q_T^P A_{T-1}\widehat{\Sigma}_{T-1}^P) + \operatorname{Tr}\left((F_{T-1}^E)^\top C_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E \widehat{\Sigma}_{T-1}^{(E,P)}\right) + \operatorname{Tr}(Q_{T-1}^P \widehat{\Sigma}_{T-1}^P) \\
&\quad +2\operatorname{Tr}\left((\widehat{\Sigma}_{T-1}^P - (\widetilde{\Sigma}_{T-1}^{(E,P)})^\top) A_{T-1}^\top Q_T^P C_{T-1} F_{T-1}^E\right) + \operatorname{Tr}\left(\Gamma_{T-1}^\top Q_T^P \Gamma_{T-1}^P W\right). \quad (4.1.64)
\end{aligned}
$$

Similarly, we have

$$
V_{T-1}^E(\widehat{x}_{T-1}^E, \widehat{\Sigma}_{T-1}^E; \widehat{\Sigma}_{T-1}^P, \widetilde{\Sigma}_{T-1}^{(E,P)}) = \left(\widehat{x}_{T-1}^E\right)^\top \Pi_{T-1}^E \widehat{x}_{T-1}^E + c_{T-1}^E \quad (4.1.65)
$$

with

$$
\begin{aligned}
\Pi_{T-1}^E &= A_{T-1} Q_T^E A_{T-1} + (F_{T-1}^E)^\top (R_{T-1}^E + C_{T-1}^\top Q_T^E C_{T-1}) F_{T-1}^E \\
&\quad +(F_{T-1}^P)^\top B_{T-1}^\top Q_T^E B_{T-1} F_{T-1}^P + 2(F_{T-1}^E)^\top C_{T-1}^\top Q_T^E A_{T-1} \\
&\quad +2(F_{T-1}^P)^\top B_{T-1}^\top Q_T^E A_{T-1} + 2(F_{T-1}^E)^\top C_{T-1}^\top Q_T^E B_{T-1} F_{T-1}^P + Q_{T-1}^E \\
&= (F_{T-1}^E)^\top R_{T-1}^E F_{T-1}^E + \left(A_{T-1} + B_{T-1} F_{T-1}^P + C_{T-1} F_{T-1}^E\right)^\top Q_T^E \cdot \\
&\quad \left(A_{T-1} + B_{T-1} F_{T-1}^P + C_{T-1} F_{T-1}^E\right) + Q_{T-1}^E, \quad (4.1.66)
\end{aligned}
$$

and

$$
\begin{aligned}
&c_{T-1}^E \\
&= \operatorname{Tr}(A_{T-1}^\top Q_T^E A_{T-1} \widehat{\Sigma}_{T-1}^E) + \operatorname{Tr}\left((F_{T-1}^P)^\top B_{T-1}^\top Q_T^E B_{T-1} F_{T-1}^P \widehat{\Sigma}_{T-1}^{(E,P)}\right) + \operatorname{Tr}(Q_{T-1}^P \widehat{\Sigma}_{T-1}^E) \\
&\quad + \operatorname{Tr}\left((\widehat{\Sigma}_{T-1}^E - \widetilde{\Sigma}_{T-1}^{(E,P)}) A_{T-1}^\top Q_T^E B_{T-1} F_{T-1}^P\right) + \operatorname{Tr}\left(\Gamma_{T-1}^\top Q_T^E \Gamma_{T-1}^P W\right). \quad (4.1.67)
\end{aligned}
$$

4.1.3.2 Dynamic Programming Principle (DPP)

Since players do not have access to the true state, their controls are related to their state estimates which causes extra randomness. Thus, unlike [65] which applies the standard DPP to the two-player zero-sum LQGG directly, we derive the individual DPP for the two-player general-sum LQGG under a fixed Markov strategy from the opponent.

At any given time $0 \le t \le T-1$, denote the value function of player P, under a fixed strategy $\gamma^E := \{\gamma_t^E(\widehat{x}_t^E)\}_{t=0}^\infty$ from player E, as

$$
V_t(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma^E) = \min_{\{u_s\}_{s=t}^T} \mathbb{E}\left[x_T^\top Q_T^P x_T + \sum_{s=t}^{T-1} u_t^\top R_t^P u_t + x_t^\top Q_t^P x_t \middle| \mathcal{H}_t\right] \quad (4.1.68)
$$

subject to (4.1.12) and γ^E, with the terminal value

$$\begin{aligned} V_T^P\left(\widehat{x}_T^P, \widehat{\Sigma}_T^P; \gamma^E\right) &= V_T^P\left(\widehat{x}_T^P, \widehat{\Sigma}_T^P\right) = \mathbb{E}_{x_T}\left[x_T^\top Q_T^P x_T \mid \widehat{x}_T^P, \widehat{\Sigma}_T^P\right] \\ &= \left(\widehat{x}_T^P\right)^\top Q_T^P \widehat{x}_T^P + \operatorname{Tr}\left(Q_T^P \widehat{\Sigma}_T^P\right). \end{aligned} \tag{4.1.69}$$

Now we show in the following theorem that a correction term is needed for the DPP in the two-player general-sum LQG with partial information.

Theorem 4.1.4 (Dynamic Programming Principle (DPP))**.** *For any given time* $0 \leq t \leq T-1$*, the value function* V_t^P *satisfies*

$$\begin{aligned} & V_t^P(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma^E) \\ = & \min_{u_t} \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)}\left[x_t^\top Q_t^P x_t + u_t^\top R_t^P u_t + V_{t+1}^P\left(\widehat{x}_{t+1}^P, \widehat{\Sigma}_{t+1}^P; \gamma^E\right) + d_{t+1}^P \middle| \widehat{x}_t^P\right], \end{aligned} \tag{4.1.70}$$

where the correction term d_{t+1}^P *is given by*

$$d_{t+1}^P = 2\mathbb{E}_{(e_t^P, e_t^E)}\left[\gamma_t^E(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P\right]. \tag{4.1.71}$$

In particular, when player E applies a linear strategy $\gamma_t^E(\widehat{x}_t^E) = F_t^E \widehat{x}_t^E$ *for some constant matrix* $F_t^E \in \mathbb{R}^{k\times n}$*, we have*

$$d_{t+1}^P = 2\operatorname{Tr}\left((F_t^E)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t \left(\widetilde{\Sigma}_t^{(E,P)} - \widehat{\Sigma}_t^P\right)\right). \tag{4.1.72}$$

The constant correction term d_{t+1}^P is a consequence of strategic interactions and partial observations. In (4.1.72) the terms K_t^P, H_t^P, and $\widehat{\Sigma}_t^P$ come from the Kalman filter, and the dependency on $\widetilde{\Sigma}_t^{(E,P)}$ is due to the fact that each player is uncertain about the state estimate of the other player, see Theorem 4.1.2. This term is equal to zero in the single-agent setting and in the multi-agent setting with full information.

Proof of Theorem 4.1.4. By definition of the value function in (4.1.68) we have,

$$\begin{aligned} & V_t(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma^E) \\ = & \min_{\{u_s\}_{s=t}^T} \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}}\left[x_T^\top Q_T^P x_T + \sum_{s=t}^{T-1} u_t^\top R_t^P u_t + x_t^\top Q_t^P x_t \middle| \widehat{x}_t^P\right] \end{aligned}$$

$$
\begin{aligned}
= \quad & \min_{u_t} \Bigg\{ \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \left[x_t^\top Q_t^P x_t + u_t^\top R_t^P u_t \middle| \widehat{x}_t^P \right] \\
& + \min_{\{u_s\}_{s=t+1}^T} \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \left[x_T^\top Q_T^P x_T + \sum_{s=t+1}^{T-1} \left(u_s^\top R_s^P u_s + x_s^\top Q_s^P x_s \right) \middle| \widehat{x}_t^P \right] \Bigg\}.
\end{aligned} \tag{4.1.73}
$$

Note that we have

$$
\begin{aligned}
& \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \left[x_T^\top Q_T^P x_T + \sum_{s=t+1}^{T-1} \left(u_s^\top R_s^P u_s + x_s^\top Q_s^P x_s \right) \middle| \widehat{x}_t^P \right] \qquad (4.1.74) \\
= \quad & \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \Bigg[x_T^\top Q_T^P x_T + \sum_{s=t+2}^{T-1} \left(u_s^\top R_s^P u_s + x_s^\top Q_s^P x_s \right) \\
& + (\widehat{x}_{t+1}^P)^\top Q_{t+1}^P \widehat{x}_{t+1}^P + u_{t+1}^\top R_{t+1}^P u_{t+1} + (e_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P - 2(\widehat{x}_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P \Big| \widehat{x}_t^P \Bigg].
\end{aligned}
$$

Given (4.1.22)-(4.1.26), we have

$$
\widehat{x}_{t+1}^P = A_t \widehat{x}_t^P + B_t u_t + C_t v_t + K_{t+1}^P H_{t+1}^P A_t (x_t - \widehat{x}_t^P) + K_{t+1}^P H_{t+1}^P \Gamma_t w_t + K_{t+1}^P w_{t+1}^P,
$$

and hence

$$
\begin{aligned}
e_{t+1}^P = \widehat{x}_{t+1}^P - x_{t+1} \quad &= \quad (I - K_{t+1}^P H_{t+1}^P) A_t (\widehat{x}_t^P - x_t) + (K_{t+1}^P H_{t+1}^P - I) \Gamma_t w_t + K_{t+1}^P w_{t+1}^P \\
&= \quad (I - K_{t+1}^P H_{t+1}^P) A_t \, e_t^P + (K_{t+1}^P H_{t+1}^P - I) \Gamma_t w_t + K_{t+1}^P w_{t+1}^P.
\end{aligned}
$$

Therefore, the last two terms in (4.1.74) satisfy

$$
\begin{aligned}
& \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \left[(e_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P - 2(\widehat{x}_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P \middle| \widehat{x}_t^P \right] \\
= \quad & \mathrm{Tr}(Q_{t+1}^P \widehat{\Sigma}_{t+1}^P) + 2\mathbb{E}_{(e_t^P, e_t^E)} \left[g(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top \, Q_{t+1}^P \, (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P \right] \\
& + \mathrm{Tr}(\Gamma_t^\top (I - K_{t+1}^P H_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) - \mathrm{Tr}((K_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P G^P) \\
& - \mathrm{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t \widehat{\Sigma}_t^P \right),
\end{aligned} \tag{4.1.75}
$$

which is independent from the control from player P at time t, namely, u_t. Therefore, the second minimisation in (4.1.73) is equivalent to the minimisation of the first four terms in (4.1.74).

On the other hand, minimising the first four terms in (4.1.74) gives

$$\begin{aligned}
&\min_{\{u_s\}_{t+1}^{T-1}} \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \Bigg[x_T^\top Q_T^P x_T + \sum_{s=t+2}^{T-1} \left(u_s^\top R_s^P u_s + x_s^\top Q_s^P x_s \right) \\
&\qquad\qquad + (\widehat{x}_{t+1}^P)^\top Q_{t+1}^P \widehat{x}_{t+1}^P + u_{t+1}^\top R_{t+1}^P u_{t+1} \Big| \widehat{x}_t^P \Bigg] \\
= \;& \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \Bigg[\min_{\{u_s\}_{t+1}^{T-1}} \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t+1}^{T-1}} \Bigg[x_T^\top Q_T^P x_T + \sum_{s=t+2}^{T-1} \left(u_s^\top R_s^P u_s + x_s^\top Q_s^P x_s \right) \\
&\qquad\qquad + (\widehat{x}_{t+1}^P)^\top Q_{t+1}^P \widehat{x}_{t+1}^P + u_{t+1}^\top R_{t+1}^P u_{t+1} \Big| \widehat{x}_{t+1}^P \Bigg] \Big| \widehat{x}_t^P \Bigg] \\
= \;& \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \Bigg[V_{t+1}(\widehat{x}_{t+1}^P) \\
&\qquad + \mathbb{E}_{e_{t+1}^P} \Big[-(e_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P + 2(\widehat{x}_{t+1}^P)^\top Q_{t+1}^P e_{t+1}^P \Big| \widehat{x}_{t+1}^P \Big] \Big| \widehat{x}_t^P \Bigg] \\
= \;& \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \Big[V_{t+1}(\widehat{x}_{t+1}^P) \Big| \widehat{x}_t^P \Big] - \mathrm{Tr}(Q_{t+1}^P \widehat{\Sigma}_{t+1}^P). \qquad (4.1.76)
\end{aligned}$$

The $\min_{\{u_s\}_{t+1}^{T-1}}$ commutes with $\mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)}$ since the control sequence after time $t+1$ $\{u_s\}_{t+1}^{T-1}$ is independent of the randomness from $(e_t^P, e_t^E, w_{t+1}^P, w_t)$. Plugging (4.1.75) and (4.1.76) into (4.1.73), we have

$$\begin{aligned}
& V_t(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma^E) \\
= \;& \min_{\{u_s\}_{s=t}^{T}} \mathbb{E}_{(e_s^P, e_s^E, w_{s+1}^P, w_s)_{s=t}^{T-1}} \left[x_T^\top Q_T^P x_T + \sum_{s=t}^{T-1} u_t^\top R_t^P u_t + x_t^\top Q_t^P x_t \Big| \widehat{x}_t^P \right] \\
= \;& \min_{u_t} \left\{ \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \left[x_t^\top Q_t u_t + u_s^\top R_s^P u_s + V_{t+1}^P \left(\widehat{x}_{t+1}^P \right) \Big| \widehat{x}_t^P \right] \right\} + d_{t+1}^P \qquad (4.1.77)
\end{aligned}$$

with the correction term d_{t+1}^P satisfies

$$\begin{aligned}
d_{t+1}^P \;=\; & 2\mathbb{E}_{(e_t^P, e_t^E)} \left[g(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P \right] \\
& + \mathrm{Tr}(\Gamma_t^\top (I - K_{t+1}^P H_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) - \mathrm{Tr}((K_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P G^P) \\
& - \mathrm{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t \widehat{\Sigma}_t^P \right) \\
=\; & 2\mathbb{E}_{(e_t^P, e_t^E)} \left[g(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P \right]
\end{aligned}$$

$$
\begin{aligned}
&+\operatorname{Tr}(\Gamma_t^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) - \operatorname{Tr}(\Gamma_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) \\
&- \operatorname{Tr}((K_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P G^P) - \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P\right) \\
&+ \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P A_t \widehat{\Sigma}_t^P\right) \\
= \;& 2\mathbb{E}_{(e_t^P, e_t^E)}\left[g(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P\right] \\
&+ \operatorname{Tr}(\Gamma_t^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) + \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P A_t \widehat{\Sigma}_t^P\right) \\
&+ \operatorname{Tr}\left(Q_{t+1}^P \left(\widehat{\Sigma}_{t+1}^P - \widehat{\Sigma}_{t+1}^{P,-}\right)\right) \qquad (4.1.78) \\
= \;& 2\mathbb{E}_{(e_t^P, e_t^E)}\left[g(\widehat{x}_t^P + e_t^E - e_t^P)^\top C_t^\top Q_{t+1}^P (K_{t+1}^P H_{t+1}^P - I) A_t e_t^P \mid \widehat{x}_t^P\right], \qquad (4.1.79)
\end{aligned}
$$

where (4.1.78) holds since for any matrix $\Pi \in \mathbb{R}^{n\times n}$,

$$
\begin{aligned}
&\operatorname{Tr}\left((K_{t+1}^P)^\top \Pi K_{t+1}^P G^P\right) + \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P\right) \\
&\quad + \operatorname{Tr}\left(\Gamma_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi K_{t+1}^P H_{t+1}^P \Gamma_t W\right) \\
&= \operatorname{Tr}\left((K_{t+1}^P)^\top \Pi K_{t+1}^P \left(H_{t+1}^P \Gamma_t W \Gamma_t^\top (H_{t+1}^P)^\top + G^P + H_{t+1}^P A_t \widehat{\Sigma}_t^P A_t^\top (H_{t+1}^P)^\top\right)\right) \\
&= \operatorname{Tr}\left((K_{t+1}^P)^\top \Pi K_{t+1}^P \left(H_{t+1}^P \widehat{\Sigma}_{t+1}^{P,-} (H_{t+1}^P)^\top + G^P\right)\right) = \operatorname{Tr}\left(\Pi \left(\widehat{\Sigma}_{t+1}^{P,-} - \widehat{\Sigma}_{t+1}^P\right)\right),
\end{aligned}
$$

and (4.1.79) holds since, by (4.1.22)-(4.1.26),

$$
\begin{aligned}
&\operatorname{Tr}\left(Q_{t+1}^P \left(\widehat{\Sigma}_{t+1}^P - \widehat{\Sigma}_{t+1}^{P,-}\right)\right) \\
= \;& -\operatorname{Tr}\left(Q_{t+1}^P K_{t+1}^P H_{t+1}^P \widehat{\Sigma}_t^{P,-}\right) \\
= \;& -\operatorname{Tr}\left(Q_{t+1}^P K_{t+1}^P H_{t+1}^P (A_t \widehat{\Sigma}_t^P A_t^\top + \Gamma_t W \Gamma_t^\top)\right) \\
= \;& -\operatorname{Tr}(\Gamma_t^\top Q_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W) - \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top Q_{t+1}^P A_t \widehat{\Sigma}_t^P\right).
\end{aligned}
$$

□

4.1.3.3 General Results by the DPP

Using the DPP formula in Theorem 4.1.4, we have the following result for the Nash equilibrium strategy and the corresponding value function for the game (4.1.12)-(4.1.13)-(4.1.14)-(4.1.16)-(4.1.19)-(4.1.20).

Theorem 4.1.5. *Assume Assumptions 3.2.4, 4.1.1, and 4.1.3 hold. The unique Nash*

equilibrium policy takes a linear format:

$$u_t^*(\widehat{x}_t^P) = F_t^P \widehat{x}_t^P, \tag{4.1.80}$$
$$v_t^*(\widehat{x}_t^E) = F_t^E \widehat{x}_t^E, \tag{4.1.81}$$

with

$$F_t^P = -(R_t^P + B_t^\top \Pi_{t+1}^P B_t)^{-1} B_t^\top \Pi_{t+1}^P (A_t + C_t F_t^E), \tag{4.1.82}$$

and

$$F_t^E = -(R_t^E + C_t^\top \Pi_{t+1}^E C_t)^{-1} C_t^\top \Pi_{t+1}^E (A_t + B_t F_t^P). \tag{4.1.83}$$

The corresponding value functions are quadratic $(0 \leq t \leq T)$*:*

$$V_t^P(\widehat{x}_t^P) = (\widehat{x}_t^P)^\top \Pi_t^P \widehat{x}_t^P + c_t^P, \tag{4.1.84}$$
$$V_t^E(\widehat{x}_t^E) = (\widehat{x}_t^E)^\top \Pi_t^E \widehat{x}_t^E + c_t^E, \tag{4.1.85}$$

with matrices $\Pi_t^P, \Pi_t^E \in \mathbb{R}^{n \times n}$ *given by*

$$\Pi_t^P = Q_t^P + (F_t^P)^\top R_t^P F_t^P + \big(A + B_t F_t^P + C_t F_t^E\big)^\top \Pi_{t+1}^P \big(A + B_t F_t^P + C_t F_t^E\big), \tag{4.1.86}$$

$$\Pi_t^E = Q_t^E + (F_t^E)^\top R_t^E F_t^E + \big(A_t + B_t F_t^P + C_t F_t^E\big)^\top \Pi_{t+1}^E \big(A_t + B_t F_t^P + C_t F_t^E\big), \tag{4.1.87}$$

and scalars $c_t^P, c_t^E, d_{t+1}^P, d_{t+1}^E \in \mathbb{R}$ *given by*

$$\begin{aligned} c_t^P = {} & c_{t+1}^P + d_{t+1}^P + \operatorname{Tr}\left(Q_t^P \widehat{\Sigma}_t^P\right) + \operatorname{Tr}\left((F_t^E)^\top C_t^\top \Pi_{t+1}^P C_t F_t^E \widehat{\Sigma}_t^{(E,P)}\right) \\ & + \operatorname{Tr}\left(\Gamma_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W\right) + \operatorname{Tr}\left((K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P G^P\right) \\ & + \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P\right) \\ & + 2 \operatorname{Tr}\left((F_t^E)^\top C_t^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t (\widehat{\Sigma}_t^P - \widetilde{\Sigma}_t^{(E,P)})\right), \end{aligned} \tag{4.1.88}$$

$$\begin{aligned} c_t^E = {} & c_{t+1}^E + d_{t+1}^E + \operatorname{Tr}\left(Q_t^E \widehat{\Sigma}_t^E\right) + \operatorname{Tr}\left((F_t^P)^\top B_t^\top \Pi_{t+1}^E B_t F_t^P \widehat{\Sigma}_t^{(E,P)}\right) \\ & + \operatorname{Tr}\left(\Gamma_t^\top (H_{t+1}^E)^\top (K_{t+1}^E)^\top \Pi_{t+1}^E K_{t+1}^E H_{t+1}^E \Gamma_t W\right) + \operatorname{Tr}\left((K_{t+1}^E)^\top \Pi_{t+1}^E K_{t+1}^E G^E\right) \\ & + \operatorname{Tr}\left(A_t^\top (H_{t+1}^E)^\top (K_{t+1}^E)^\top \Pi_{t+1}^E K_{t+1}^E H_{t+1}^E A_t \widehat{\Sigma}_t^E\right) \\ & + 2 \operatorname{Tr}\left((F_t^P)^\top B_t^\top \Pi_{t+1}^E K_{t+1}^E H_{t+1}^E A_t (\widehat{\Sigma}_t^E - (\widetilde{\Sigma}_t^{(E,P)})^\top)\right), \end{aligned} \tag{4.1.89}$$

and

$$\begin{aligned} d_{t+1}^P &= 2\,\mathrm{Tr}\left((F_t^E)^\top C_t^\top Q_{t+1}^P \left(K_{t+1}^P H_{t+1}^P - I\right) A_t \left(\widetilde{\Sigma}_t^{(E,P)} - \widehat{\Sigma}_t^P\right)\right), \\ d_{t+1}^E &= 2\,\mathrm{Tr}\left((F_t^P)^\top B_t^\top Q_{t+1}^E \left(K_{t+1}^E H_{t+1}^E - I\right) A_t \left((\widetilde{\Sigma}_t^{(E,P)})^\top - \widehat{\Sigma}_t^E\right)\right). \end{aligned}$$

The terminal conditions are

$$\Pi_T^P = Q_T^P, \quad \Pi_T^E = Q_T^E, \quad c_T^P = \mathrm{Tr}\left(Q_T^P \widehat{\Sigma}_T^P\right), \quad c_T^E = \mathrm{Tr}\left(Q_T^E \widehat{\Sigma}_T^E\right). \tag{4.1.90}$$

We note that the optimal policies F_t^P and F_t^E given in (4.1.82) and (4.1.83), and the Riccati equations given in (4.1.86) and (4.1.87), are the same as the optimal polices and Riccati equations in the case of full information ([13, Corollary 6.4]). However, the LQGG (defined in (4.1.12)-(4.1.13)-(4.1.14)-(4.1.16)-(4.1.19)-(4.1.20)) differs from the LQG with *full information* in [13, Corollary 6.4] in the sense that the Nash equilibrium strategy is linear in the *state estimate* rather than the *true state*, and the scalars c_t^P and c_t^E in the value function involve more terms due to the errors in state estimation.

Proof. At time $t = T$, by (4.1.69) we have the terminal conditions given in (4.1.90). In the end game analysis in Section 4.1.3.1, we have shown that at time $t = T-1$, the unique Nash equilibrium policy for each player is linear in his state estimate and the corresponding value function is quadratic in the state estimate, with F_{T-1}^P, F_{T-1}^E, Π_{T-1}^P, Π_{T-1}^E, c_{T-1}^P, and c_{T-1}^E given in (4.1.57), (4.1.58), (4.1.63), (4.1.66), (4.1.64), and (4.1.67).

At time $t = 0, 1, \cdots, T-2$, player P plays against the strategy $\gamma_t^E(\cdot)$ of player E. By Theorem 4.1.4, we have the DPP for player P

$$\begin{aligned} & V_t^P(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma_t^E) \\ = & \min_{u_t} \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^P, w_t)} \left[x_t^\top Q_t^P x_t + u_t^\top R_t^P u_t + V_{t+1}^P\left(\widehat{x}_{t+1}^P, \widehat{\Sigma}_{t+1}^P; \gamma_{t+1}^E\right) + d_{t+1}^P \middle| \widehat{x}_t^P\right], \end{aligned}$$

where

$$x_{t+1} = A_t x_t + B_t u_t + C_t(\gamma_t^E(\widehat{x}_t^E)) + \Gamma_t w_t, \tag{4.1.91}$$

and according to (4.1.22)-(4.1.26),

$$\widehat{x}_{t+1}^P = A_t\widehat{x}_t^P + B_t u_t + C_t(\gamma_t^E(\widehat{x}_t^P + \widetilde{w}_t)) + K_{t+1}^P(H_{t+1}^P\Gamma_t w_t + w_{t+1}^P - H_{t+1}^P A_t e_t^P). \quad (4.1.92)$$

Hence

$$\begin{aligned} V_t^P(\widehat{x}_t^P, \widehat{\Sigma}_t^P; \gamma_t^E) &= \min_{u_t}\Big\{u_t^\top R_t^P u_t + (\widehat{x}_t^P)^\top Q_t^P \widehat{x}_t^P + \mathrm{Tr}\left(Q_t^P\widehat{\Sigma}_t^P\right) \\ &\quad + \mathbb{E}_{\widetilde{w}_t, e_t^P, w_t, w_{t+1}^P}\left[V_{t+1}^P(A_t\widehat{x}_t^P + B_t u_t + C_t(\gamma_t^E(\widehat{x}_t^P + \widetilde{w}_t))\right. \\ &\quad \left. + K_{t+1}^P(H_{t+1}^P\,\Gamma_t w_t + w_{t+1}^P - H_{t+1}^P A_t e_t^P), \widehat{\Sigma}_{t+1}^P; \gamma_{t+1}^E) + d_{t+1}^P \Big| \widehat{x}_t^P\right]\Big\}, \end{aligned} \quad (4.1.93)$$

where

$$\widehat{x}_t^E = \widehat{x}_t^P + \widetilde{w}_t, \quad \widetilde{w}_t = e_t^E - e_t^P. \quad (4.1.94)$$

Similarly, the DPP for player E is

$$\begin{aligned} & V_t^E(\widehat{x}_t^E, \widehat{\Sigma}_t^E; \gamma_t^P) \\ = & \min_{v_t} \mathbb{E}_{(e_t^P, e_t^E, w_{t+1}^E, w_t)}\left[x_t^\top Q_t^E x_t + v_t^\top R_t^E v_t + V_{t+1}^E\left(\widehat{x}_{t+1}^E, \widehat{\Sigma}_{t+1}^E; \gamma_{t+1}^P\right) + d_{t+1}^E \Big| \widehat{x}_t^E\right], \end{aligned}$$

where

$$x_{t+1} = A_t x_t + B_t(\gamma_t^P(\widehat{x}_t^P)) + C_t v_t + \Gamma_t w_t, \quad (4.1.95)$$

and according to (4.1.22)-(4.1.26),

$$\widehat{x}_{t+1}^E = A_t\widehat{x}_t^E + B_t(\gamma_t^P(\widehat{x}_t^E - \widetilde{w}_t)) + C_t v_t + K_{t+1}^E(H_{t+1}^E\Gamma_t w_t + w_{t+1}^E - H_{t+1}^E A_t e_t^E). \quad (4.1.96)$$

Hence,

$$\begin{aligned} V_t^E(\widehat{x}_t^E, \Sigma_t^E; \gamma_t^P) &= \min_{v_t}\Big\{v_t^\top R_t^E v_t + (\widehat{x}_t^E)^\top Q_t^E \widehat{x}_t^E + \mathrm{Tr}\left(Q_t^E\widehat{\Sigma}_t^E\right) \\ &\quad + \mathbb{E}_{\widetilde{w}_t, e_t^E, w_t, w_{t+1}^E}\left[V_{t+1}^P(A_t\widehat{x}_t^E + C_t v_t + B_t(\gamma_t^P(\widehat{x}_t^E - \widetilde{w}_t)))\right. \\ &\quad \left. + K_{t+1}^E(H_{t+1}^E\,\Gamma_t w_t + w_{t+1}^E - H_{t+1}^E A_t e_t^E), \widehat{\Sigma}_{t+1}^P; \gamma_{t+1}^P) + d_{t+1}^E \Big| \widehat{x}_t^E\right]\Big\}, \end{aligned}$$

where

$$\widehat{x}_t^P = \widehat{x}_t^E - \widetilde{w}_t. \tag{4.1.97}$$

Similar to the end game analysis in Section 4.1.3.1, we can show that the optimal strategies at $t = T-2, T-3, \ldots, 0$ are linear in the state estimate,

$$\begin{aligned} \gamma_t^P(\widehat{x}_t^P) &= F_t^P \widehat{x}_t^P, & (4.1.98)\\ \gamma_t^E(\widehat{x}_t^E) &= F_t^E \widehat{x}_t^E, & (4.1.99) \end{aligned}$$

and the cost function is quadratic in the state estimate,

$$\begin{aligned} V_t^P(\widehat{x}_t^P) &= (\widehat{x}_t^P)^\top \Pi_t^P \widehat{x}_t^P + c_t^P, & (4.1.100)\\ V_t^E(\widehat{x}_t^E) &= (\widehat{x}_t^E)^\top \Pi_t^E \widehat{x}_t^E + c_t^E, & (4.1.101) \end{aligned}$$

for some F_t^P, F_t^E, Π_t^P, Π_t^E, c_t^P, and c_t^E to be determined. For player P, by substituting (4.1.98), (4.1.99), and (4.1.100) into the value function (4.1.93), they minimise their value function to find the optimal response to player E's strategy $v_t = F_t^E \widehat{x}_t^E$,

$$\begin{aligned} & V_t^P(\widehat{x}_t^P) \\ = \; & (\widehat{x}_t^P)^\top \Pi_t^P \widehat{x}_t^P + c_t^P \\ = \; & \min_{u_t} \Big\{ u_t^\top R_t^P u_t + (\widehat{x}_t^P)^\top Q_t^P \widehat{x}_t^P + \operatorname{Tr}\left(Q_t^P \widehat{\Sigma}_t^P\right) + \mathbb{E}_{\widetilde{w}_t, e_t^P, w_t, w_{t+1}^P} \big[(A_t \widehat{x}_t^P + B_t u_t \\ & + C_t F_t^E (\widehat{x}_t^P + \widetilde{w}_t) + K_{t+1}^P (H_{t+1}^P \Gamma_t w_t + w_{t+1}^P - H_{t+1}^P A_t e_t^P)) \cdot \Pi_{t+1}^P \cdot \\ & (A_t \widehat{x}_t^P + B_t u_t + C_t F_t^E (\widehat{x}_t^P + \widetilde{w}_t) + K_{t+1}^P (H_{t+1}^P \Gamma_t w_t + w_{t+1}^P - H_{t+1}^P A_t e_t^P)) \\ & + c_{t+1}^P + d_{t+1}^P \big| \widehat{x}_t^P \big] \Big\}. \end{aligned} \tag{4.1.102}$$

Now we calculate the expectations in (4.1.102),

$$\begin{aligned} & \mathbb{E}_{\widetilde{w}_t, e_t^P, w_t, w_{t+1}^P} [(C_t F_t^E (\widehat{x}_t^P + \widetilde{w}_t))^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t e_t^P | \widehat{x}_t^P] \\ = \; & \mathbb{E}_{e_t^E, e_t^P} [(C_t F_t^E (\widehat{x}_t^P + e_t^E - e_t^P))^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t e_t^P] \\ = \; & -\operatorname{Tr}((F_t^E)^\top C_t^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P) \\ & + \operatorname{Tr}((F_t^E)^\top C_t^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widetilde{\Sigma}_t^{(E,P)}). \end{aligned} \tag{4.1.103}$$

Substituting (4.1.103) into (4.1.102) gives

$$
\begin{aligned}
& V_t^P(\widehat{x}_t^P) \\
= \; & (\widehat{x}_t^P)^\top \Pi_t^P \widehat{x}_t^P + c_t^P \\
= \; & \min_{u_t} \Big\{ u_t^\top R_t^P u_t + (\widehat{x}_t^P)^\top Q_t^P \widehat{x}_t^P + \operatorname{Tr}\left(Q_t^P \widehat{\Sigma}_t^P\right) + (\widehat{x}_t^P)^\top A_t^\top \Pi_{t+1}^P A_t \widehat{x}_t^P \\
& + u_t^\top B_t^\top \Pi_{t+1}^P B_t u_t + (\widehat{x}_t^P)^\top (F_t^E)^\top C_t^\top \Pi_{t+1}^P C_t F_t^E \widehat{x}_t^P + \operatorname{Tr}\left((F_t^E)^\top C_t^\top \Pi_{t+1}^P C_t F_t^E \widehat{\Sigma}_t^{(E,P)}\right) \\
& + \operatorname{Tr}\left(\Gamma_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P \Gamma_t W\right) + \operatorname{Tr}\left((K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P G^P\right) \\
& + \operatorname{Tr}\left(A_t^\top (H_{t+1}^P)^\top (K_{t+1}^P)^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P\right) + 2(\widehat{x}_t^P)^\top A_t^\top \Pi_{t+1}^P B_t u_t \\
& + 2(\widehat{x}_t^P)^\top A_t^\top \Pi_{t+1}^P C_t F_t^E \widehat{x}_t^P + 2 u_t^\top B_t^\top \Pi_{t+1}^P C_t F_t^E \widehat{x}_t^P \\
& + 2 \operatorname{Tr}((F_t^E)^\top C_t^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widehat{\Sigma}_t^P) - 2 \operatorname{Tr}((F_t^E)^\top C_t^\top \Pi_{t+1}^P K_{t+1}^P H_{t+1}^P A_t \widetilde{\Sigma}_t^{(E,P)}) \\
& + c_{t+1}^P + d_{t+1}^P \Big\}
\end{aligned} \tag{4.1.104}
$$

Since the DPP correction term d_{t+1}^P is independent of the control u_t, the first order condition leads to the following optimal response:

$$
u_t^* = F_t^P \widehat{x}_t^P = -(R_t^P + B_t^\top \Pi_{t+1}^P B_t)^{-1} B_t^\top \Pi_{t+1}^P (A_t + C_t F_t^E) \widehat{x}_t^P. \tag{4.1.105}
$$

Similarly, player E minimises his value function to find his optimal response to player P's strategy $u_t = F_t^P \widehat{x}_t^P$. The optimal strategy v_t^* of player E is given by

$$
v_t^* = F_t^E \widehat{x}_t^E = -(R_t^E + C_t^\top \Pi_{t+1}^E C_t)^{-1} C_t^\top \Pi_{t+1}^E (A_t + B_t F_t^P) \widehat{x}_t^E. \tag{4.1.106}
$$

Plugging $u_t^* = F_t^P \widehat{x}_t^P$ into (4.1.104) and after direct calculation, we can rewrite the value function as

$$
V_t^P(\widehat{x}_t^P) = (\widehat{x}_t^P)^\top \Pi_t^P \widehat{x}_t^P + c_t^P, \quad t = 0, 1, \ldots, T-2
$$

with Π_t^P and c_t^P given in (4.1.86) and (4.1.88). Similarly, player E's value function can be rewritten as

$$
V_t^E(\widehat{x}_t^E) = (\widehat{x}_t^E)^\top \Pi_t^E \widehat{x}_t^E + c_t^E, \quad t = 0, 1, \ldots, T-2,
$$

with Π_t^E and c_t^E given in (4.1.87) and (4.1.89).

□

4.1.4 Experimental Result of the Bargaining Model

Recall that we have established a two-player bargaining model in Section 4.1.1. In this section, we will further discuss the model and present some numerical results. We set the penalty function in the model to be $R_t^i = \theta_i \exp(-\lambda_i t)$ where θ_i and λ_i are positive constants. Thus, the penalty for concession decays over time, which allows both players to be more flexible near the deadline in order to settle a deal. We also note that to ensure the well-definedness of the LQGG, we also add a small observation noise with mean zero and variance ϵ ($\epsilon > 0$) to the players' offers x_t^B and x_t^S. Thus if we write $x_t = (x_t^B, x_t^S, p_t)$ as the state, then each player observes

$$z_{t+1}^i = x_{t+1} + w_{t+1}^i$$

for $i = S, B$ where $w_{t+1}^i \sim \mathcal{N}(0, \overline{G}^i)$ with

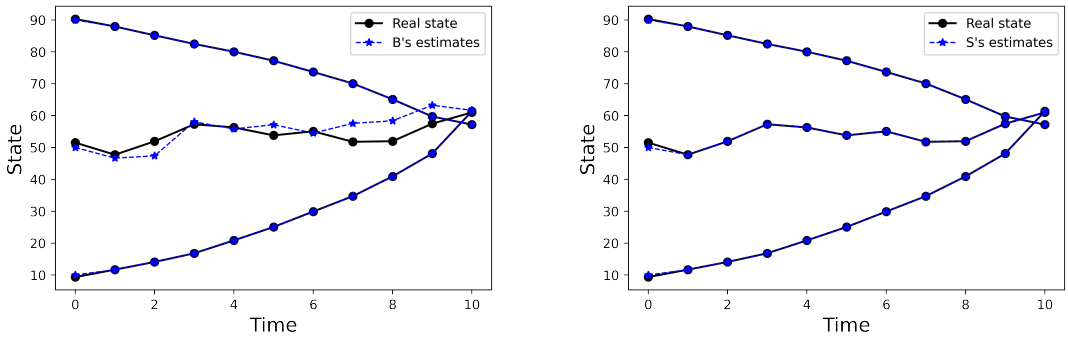
$$\overline{G}^i = \begin{bmatrix} \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & G^i \end{bmatrix}.$$

Set-up. We set up the model parameters and initial conditions as follows:

1. Parameters: $T = 10$, $W = 5$, $\alpha_B = 60$, $\alpha_S = 60$, $\beta_B = 10$, $\beta_S = 10$, $c_B = c_S = 0.05$, $\epsilon = 10^{-6}$, $G^B = 10$, $G^S = 0.01$. For the penalty function we set $\theta_i = 5$ and $\lambda_i = 0.1$.

2. Initial state: $x_0^B \sim \mathcal{N}(10, 3)$, $x_0^B \sim \mathcal{N}(90, 3)$, and $p_0 \sim \mathcal{N}(50, 5)$. We assume that players' belief of the initial state distribution are the same as the true distribution.

State estimates. We now plot the buyer's and the seller's state estimates obtained by the Kalman filter in Figure 4.1, where both players play the Nash equilibrium strategies. We can see that since the seller has a larger observation noise ($G^B > G^S$), he has a less accurate estimate of the project value.

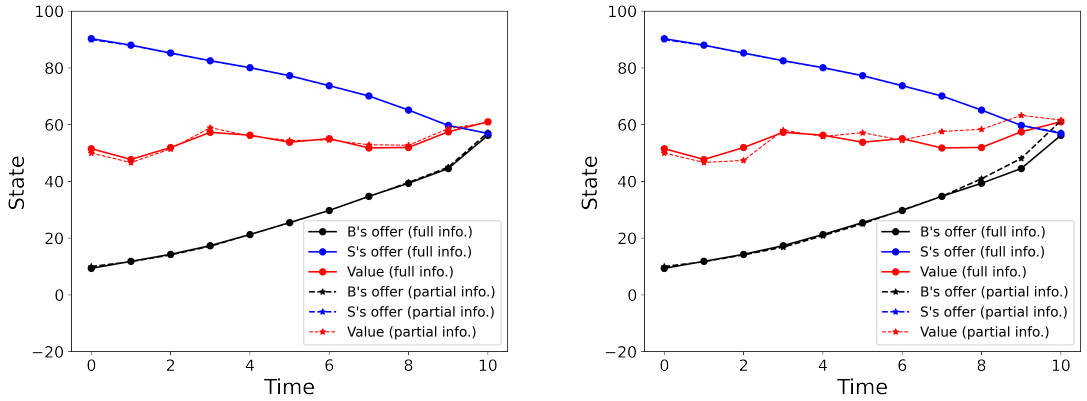


(a) Buyer's state estimates.

(b) Seller's state estimates.

Figure 4.1: The true state (solid black lines) and state estimates (dashed blue lines) of the buyer and the seller. The top, middle, and bottom solid lines represent the true seller's offer, (the true) value of the project, and (the true) buyer's offers, respectively. The solid lines represent each player's estimates of the three true state variables.

Effect of the Observation Noise. To study the effect of the observation noise on the negotiation strategies, we fix $G^S = 0.01$ and vary the variance of the buyer's observation noise G^B. In Figure 4.2, we provide an example to show the difference between the players' actual offers (blue and black dashed lines) under partial information with their offers under full information (blue and black solid lines). We can see that when $G^B = 1$, the buyer has a relatively accurate estimate of the project value (the red dashed line is close to the red solid line), therefore the final agreement price is similar to the case with full information; while when $G^B = 10$, in this example, the buyer's estimated project value is higher than the actual value, and thus he proposes a higher offer at the deadline. We also observe that in the case with full information the final offers do not cross each other while with partial information, there is a cross due to state estimation error and players may agree at the average of the two offers. In fact, the buyer's estimated value project could be either higher or lower than the actual value, thus higher observation noise leads to more risk of not reaching an agreement in the end and more variations in the final agreement price. In addition, we note that as G^B increases, the seller's cost J^S will also increase, but in smaller increments. Hence it is of everyone's interest to be less uncertain about the project value.



(a) $G^B = 1$. The buyer's cost is $J^B = 968$. (b) $G^B = 10$. The buyer's cost is $J^B = 2854$.

Figure 4.2: An example of state trajectories under the Nash equilibrium with full information (solid lines) and partial information (dashed lines). The red dashed line represents the buyer's estimate of the project value. For comparison, the buyer's Nash equilibrium cost with full information is $J^B = 920$.

4.2 The LQG Model with Reinforcement Learning

In Section 4.1 we have discussed an bargaining model under the assumption that the model parameters are known to both players. In reality, each player usually does not have access to his opponent's information and thus needs to learn the optimal negotiation strategies through repeated negotiations. In this section, we will cast the bargaining problem with unknown model parameters as an LQG problem using the framework discussed in Chapter 3. We will then show how, in a series of negotiations with the same counterparty, the natural policy gradient method (Algorithm 3) helps the players to improve their bargaining strategies without access to the model parameters.

4.2.1 A Two-player LQG model

Similar to Section 4.1, we set the control of each player i ($i = B, S$), denoted by u_t and v_t, to be change of the buyer's and the seller's offers, respectively, which measure the concession they make at time t. The state variables are their offers x_t^i and their valuations of the project y_t^i. The valuations can be considered as their reservation prices, that is the upper or lower limit of what they are prepared to offer: on the demand side, the reservation price is the highest price that a buyer is willing to pay; on the supply side, it is the lowest price at which a seller is willing to sell. There

is some uncertainty in their valuations due to both internal and external factors that are inherent in the problem, e.g. change in the economic climate or internal company policy. In contrast to the LQGG model studied in Section 4.1 where both players' target prices are linked to the true project value, here we allow each player to have a different reservation price with correlated noise. Therefore, we construct the following bargaining model with a four dimensional state variable $x_t = (x_t^B, x_t^S, y_t^B, y_t^S)^\top$ and a two dimensional control variable (u_t, v_t). The equations for the evolution of the state variable are

$$\begin{aligned} x_{t+1}^B &= x_t^B + u_t + \epsilon_t^B, \quad x_{t+1}^S = x_t^S + v_t + \epsilon_t^S, \\ y_{t+1}^B &= \gamma_B y_t^B + \sigma_B w_t^1, \quad y_{t+1}^S = \gamma_S y_t^S + \sigma_S(\sqrt{\rho} w_t^1 + \sqrt{1-\rho} w_t^2), \end{aligned}$$

where γ_i is the growth rate of y_t^i and σ_i is the volatility parameter, which measures the size of the uncertainty of y_t^i; T is the finite deadline. Similar to the model considered in Section 4.1.1, here ϵ_t^B and ϵ_t^S serve as regularization terms with $\epsilon_t^B \sim \mathcal{N}(0, W^B)$ and $\epsilon_t^S \sim \mathcal{N}(0, W^S)$ to guarantee the well-definedness of the LQG. We note that the Nash equilibrium strategies are still well behaved when $W^B = W^S = 0$ (see Figure 4.3). We assume that $\{w_t^1\}_{t=0}^{T-1}$ and $\{w_t^2\}_{t=0}^{T-1}$ are i.i.d. standard normal random variables and they are highly correlated through a constant $\rho \in (0.5, 1)$, since we believe that the players are likely to have a similar valuation of the product. We consider a simple quadratic cost function for the buyer, given by

$$\inf_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[\alpha_B(x_T^B - x_T^S)^2 + \beta_B(x_T^B - y_T^B)^2 + \sum_{t=0}^{T-1} R_t^B(u_t)^2\right],$$

and for the seller

$$\inf_{\{v_t\}_{t=0}^{T-1}} \mathbb{E}\left[\alpha_S(x_T^B - x_T^S)^2 + \beta_S(x_T^S - y_T^S)^2 + \sum_{t=0}^{T-1} R_t^S(v_t)^2\right],$$

where $\alpha_i > 0$ and $\beta_i > 0$ are weighting parameters. The function R_t^i is the penalty function representing the cost of control at time t. Therefore, when minimising the cost function, each player tries to settle a deal with their counterparty, gradually conceding toward their bottom line, while trying to make as little concession as possible.

Therefore, in the LQG formulation which we discussed in Section 3.2, we have

$$A_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r_B & 0 \\ 0 & 0 & 0 & r_S \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad w_t = \begin{bmatrix} \epsilon_t^B \\ \epsilon_t^S \\ \sigma_1 w_t^1 \\ \sigma_2(\sqrt{\rho} w_t^1 + \sqrt{1-\rho} w_t^2) \end{bmatrix},$$

$$Q_t^i = \mathbf{0},\ R_t^i = R_t^i(t),\ Q_T^B = \begin{bmatrix} \alpha_B + \beta_B & -\alpha_B & -\beta_B & 0 \\ -\alpha_B & \alpha_B & 0 & 0 \\ -\beta_B & 0 & \beta_B & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, Q_T^S = \begin{bmatrix} \alpha_S & -\alpha_S & 0 & 0 \\ -\alpha_S & \alpha_S + \beta_S & 0 & -\beta_S \\ 0 & 0 & 0 & 0 \\ 0 & -\beta_S & 0 & \beta_S \end{bmatrix}.$$

Note that in this model, Q_t^i and Q_T^i $(i = B, S)$ are positive semi-definite but not positive definite, Lemma 3.2.6 still holds under this extended assumption [13].

4.2.2 Experimental Results

We first show the Nash equilibrium of the LQG model in an example given in Figure 4.3. We then perform the natural policy gradient algorithm with known and unknown parameters to the bargaining model. For each case we plot the the state trajectory under the learned policies and the normalised error defined in (3.5.1). Note that although Assumptions 3.2.1 and 3.2.2 are not satisfied under this set of parameters, the algorithm will converge to the Nash equilibrium under proper conditions in both cases.

Set-up. We perform the natural policy gradient method under the following parameters:

1. Parameters: $\alpha_B = 6$, $\alpha_S = 7$, $\beta_B = 4$, $\beta_S = 4$, $\rho = 0.8$, $\sigma_B = 5$, $\sigma_S = 5$, $\gamma_B = 1$, $\gamma_S = 1$, $W^B = W^S = 0$, $T = 5$. The penalty function is set to be $R_t^i = \theta_i \exp(-\lambda_i t)$, with $\theta_B = \theta_S = 10$ and $\lambda_B = \lambda_S = 0.1$. For the algorithm with unknown parameters, the number of trajectories $L_B = L_S = 500$, and smoothing parameters $r_B = r_S = 0.5$.

2. Initialisation: $x_0^B \sim \mathcal{N}(6, 1)$, $x_0^S \sim \mathcal{N}(12, 1.5)$, $y_0^B \sim \mathcal{N}(8, 1)$, and $y_0^S \sim \mathcal{N}(7, 0.5)$. The initial policy for the buyer: $\boldsymbol{F}^B \in \mathbb{R}^{1\times 20} = (F_0^{B,(0)}, \cdots, F_{T-1}^{B,(0)})$ with $F_t^{B,(0)} = [0.1, -0.001, -0.1, -0.01]$ for all t. The initial policy for the seller: $\boldsymbol{F}^S \in \mathbb{R}^{1\times 20} = (F_0^{S,(0)}, \cdots, F_{T-1}^{S,(0)})$ with $F_t^{S,(0)} = [0.0001, -0.1, 0.05, 0.05]$ for all t.

Nash Equilibrium. The behavior of the buyer and the seller under their Nash equilibrium strategies is shown in Figure 4.3. The solution is calculated by solving the Ricatti system (3.2.5).

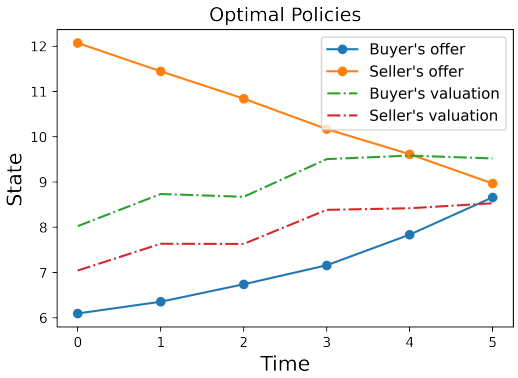
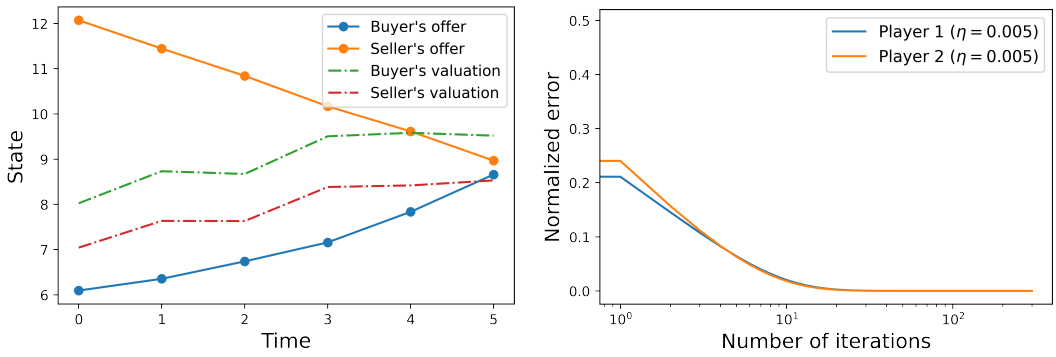


Figure 4.3: State trajectory under Nash equilibrium strategies.

Algorithm with Known Parameters. In the case of known parameters, the state trajectory (see Figure 4.4a) is very similar to the one under Nash equilibrium shown in Figure 4.3, and the algorithm is very efficient as the normalized error falls below 0.1 within 10 iterations (see Figure 4.4b).

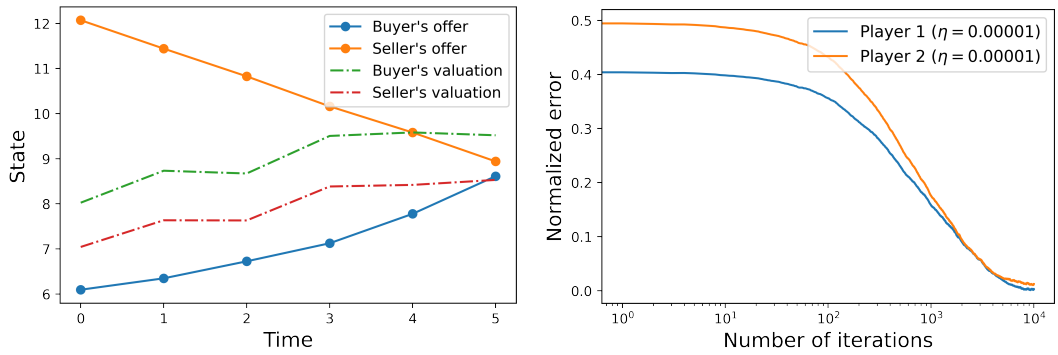


(a) State trajectory under learned policies. (b) Convergence.

Figure 4.4: Performance of the algorithm with known parameters for the bargaining model.

Algorithm with Unknown Parameters. The state trajectory under the learned policies (see Figure 4.5a) is very similar to the optimal one. However, it takes the algorithm much longer to converge compared with the case of known parameter – the normalized error falls below 0.1 within 2000 iterations (see Figure 4.5b). This indicates

that with fully unknown parameters, it takes a lot more efforts to learn an accurate strategy.



(a) State trajectory under learned policies.

(b) Convergence.

Figure 4.5: Performance of the algorithm with unknown parameters for the bargaining model.

Chapter 5

Conclusion and Future Work

In this thesis we studied the convergence of policy gradient methods for LQ problems in the setting of finite time horizon and stochastic state dynamics. We started with the single-agent LQR problem and proved that policy gradient methods enjoy linear convergence to the optimal solution in the case of known and unknown parameters. We also discussed the optimal liquidation problem and presented the experimental results with synthetic data and real market data. We then studied the N-player general-sum LQG and proved the global convergence of the natural policy gradient method to the Nash equilibrium. The theoretical convergence is guaranteed provided that there is a certain amount of noise in the system. We illustrated our results with numerical experiments to show that the addition of noise helps the algorithm to learn the Nash equilibrium strategy. Furthermore, we developed bargaining models using LQGs for the case where players have partial information but know the model parameters, and where players learn the optimal strategy with unknown model parameters.

It would be interesting to explore the following directions in the future:

- *Convergence of (natural) policy gradient methods for N-player general-sum LQGs in the case of unknown parameters.* In Chapter 3 we provide a provable convergence guarantee for the natural policy gradient method in the case of known parameters. It is likely to extend this result to the case when model parameters are unknown (Algorithm 3) following Chapter 2.

- *Convergence of policy gradient methods in mean-field games with a finite horizon.* We have studied LQ problems when $N < \infty$ as a starting point to analyse the performance of policy gradient methods. Extending the convergence results to

LQ or non-LQ mean-field games ($N = \infty$) is a natural direction for this research. Similar work has been done for the continuous-time LQ mean-field control and game [87] and discrete-time LQ mean-field control [21] problems.

- *Models for negotiation over multiple items among multiple players.* The bargaining model developed in Chapter 4 can be further generalised in various ways. For example, in reality there might be more than two players involved in a negotiation. Players may also negotiate over several items, in which case an important problem is how to design a cost function. We may also view this as an RL problem with multiple objectives. Other RL algorithms such as TRPO and PPO may also be applied to improve the learning efficiency.

- *RL methods for LQGs with partial information.* In Section 4.1 we consider an LQG model where each player knows the model parameters but has a noisy observation of the true state. In the future we may combine RL methods with the LQG and Kalman filter to learn the optimal strategy when model parameters are unknown to the players and players do not have access to the true state.

Appendix A

Appendix for Chapter 2

A.1 Market Simulator for Linear Price Dynamics

In the case of unknown parameters, the controller is assumed to have access to a simulator, which generates the cost signals using equation (2.2.9) with the estimated parameters from the NASDAQ ITCH data taken from Lobster[1]. We give the details of data we use in Figures A.1 and A.2.

Now we show how we estimate each of those parameters in (2.2.9) and summarize the estimated parameters in Table A.1.

Permanent Price Impact and Volatility. The model in (2.2.7) implies that prices changes are proportional to the *market-order flow imbalances* (MFI). We adopt the framework from [24], namely that the price change ΔS is given by

$$\Delta S = \gamma \, \mathrm{MFI} + \sigma \, \epsilon, \tag{A.1.1}$$

with $\mathrm{MFI} = M^b - M^s$ where M^s and M^b are the volumes of market sell orders and market buy orders respectively during a time interval $\Delta T = 5\text{mins}$ and $\epsilon \sim \mathcal{N}(0, 1)$. We then estimate γ and σ from the data.

[1] https://lobsterdata.com/

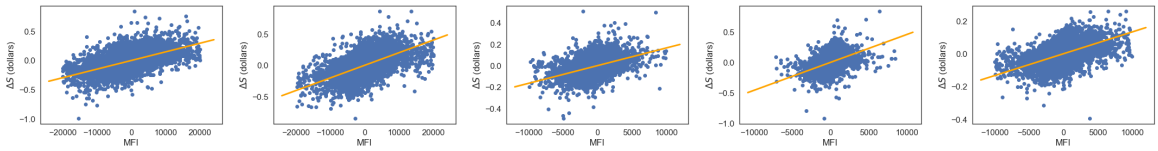


Figure A.1: Relationship between MFI and ΔS. (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019, $\Delta T = 1$min)

Temporary Price Impact. We assume the LOB has a flat shape with constant queue length l for the first few levels. Figure A.2 shows the average queue lengths for the first 5 levels so that our assumption is not too unreasonable. Therefore the following equation, on the amount received when we liquidate u shares with best bid price S, holds

$$u(S - \beta u) = \int_{S - \frac{u\Delta}{l}}^{S} lvdv.$$

Therefore we have $\beta = \frac{\Delta}{2l}$, where Δ is the tick size and l is the average queue length.

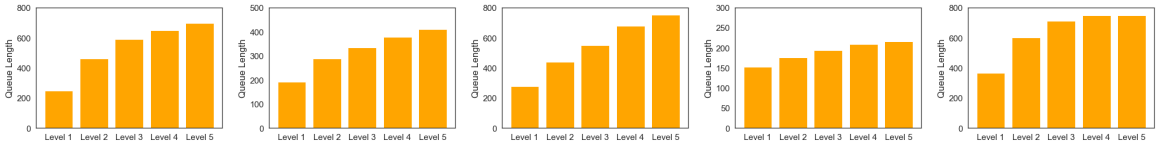


Figure A.2: Average queue length (volume) of the first five levels on the limit buy side (Example (from left to right): AAP, FB, JPM, IBM and AAL, 10:00AM-11:00AM 01/01/2019-08/31/2019 with 5000 samples uniformly sampled with natural time clock in each trading day.)

Parameter Estimation. See the estimates for AAPL, FB, IBM, JPM, and AAL in Table A.1.

Paramters/Stock	AAPL	FB	IBM	JPM	AAL
β	1.03×10^{-5}	1.30×10^{-5}	2.65×10^{-5}	9.28×10^{-6}	3.27×10^{-5}
γ	7.27×10^{-6}	1.40×10^{-5}	4.60×10^{-5}	1.65×10^{-5}	1.33×10^{-5}
σ	0.107	0.115	0.082	0.059	0.042

Table A.1: Parameter estimation from NASDAQ ITCH Data (10:00AM-11:00PM 01/01/2019-08/31/2019).

A.2 Comparison between the Policy Gradient Method and Q-learning

The computational complexity of Q-learning is highly dependent on the size of the set of the (discrete) states and actions. Therefore Q-learning is typically less suited to problems with continuous and unbounded states and actions. In order to apply Q-learning for such problems, we need to discretize the continuous state and action space. Intuitively speaking, Q-learning suffers from low accuracy when the discretization scheme is less refined (see Figures A.3 and A.5). On the other hand, the computational complexity grows quadratically when increasing the level of granularity of discretization (see Figures A.4 and A.5).

To demonstrate this view point, we compare the performance of the Q-learning algorithm with the policy gradient method on a one-dimensional LQR problem with finite horizon as suggested by the reviewer. (We would expect the deep Q-learning algorithm and the deep policy gradient method to have similar comparison results.)

Q learning update. We initialize the Q table $\{q_t^{(0)}(x,u)\}_{x,u,t}$ with all zeros. In the i-th iteration, we update the Q table for $t = 0, 1, \cdots, T-1$,

$$q_t^{(i)}(x,u) = (1-\tilde{\eta})\, q_t^{(i-1)}(x,u) + \tilde{\eta}\, \left[c_t(x,u) + \min_{u'}\ q_{t+1}^{(i)}(x',u') \right], \tag{A.2.1}$$

with terminal condition $q_T^{(i)}(x,u) = x^2 Q_T$. Here $c_t(x,u) = x^2 Q_t + u^2 R_t$ is the instantaneous cost at time t; x' is the next state simulated from the system when the agent takes an action u in state x at time t; and $\tilde{\eta} \in (0,1)$ is the learning rate.

Model set-up. We set $d = k = 1$, $T = 5$, $A = 1.0$, $B = 0.2$, $Q_t = 0.2$ for $t = 0, 1, 2, 3, 4$, $Q_T = 0.4$, $R_t = 0.1(t+1)$ for $t = 0, 1, \cdots, 4$, $w_t \sim \mathcal{N}(0, 0.1)$, and $x_0 \sim \mathcal{N}(0, 0.1)$.

Parameter set-up. To perform Q-learning, we uniformly partition the states and actions in $[-1, 1]$. We set the learning rate for Q-learning as $\tilde{\eta} = 0.1$. For the policy gradient method, we set the learning rate $\eta = 0.2$ and the number of trajectories in the zero-th optimization as $m = 50$.

Conclusion. From Figure A.5, we observe that

- For LQR with finite horizon, the policy gradient method outperforms Q-learning algorithms (with the size of actions and states varying from 10 to 100) in terms of both sample efficiency and accuracy.
- When increasing the size of the states and actions from 10 to 100, the accuracy of the Q-learning algorithm improves, however, it requires many more samples to converge.

To conclude, Q-learning is less suited to handling decision-making problems with continuous and unbounded states and actions. More advanced approximation techniques may be needed in this case [83].

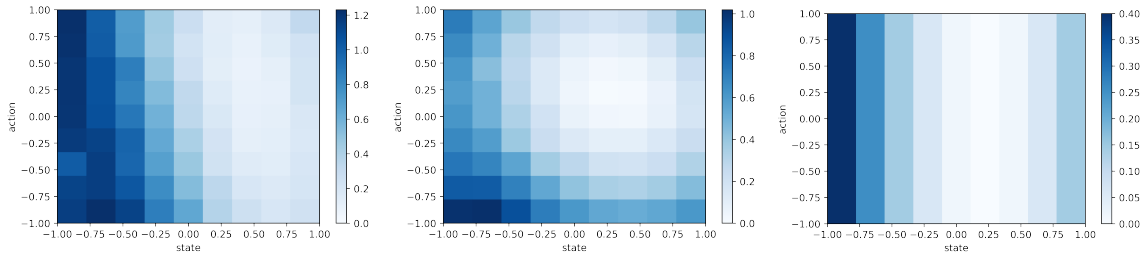


Figure A.3: Q tables with 10 states and 10 actions: $q_0(s, a)$, $q_4(s, a)$ and $q_5(s, a)$ (from left to right).

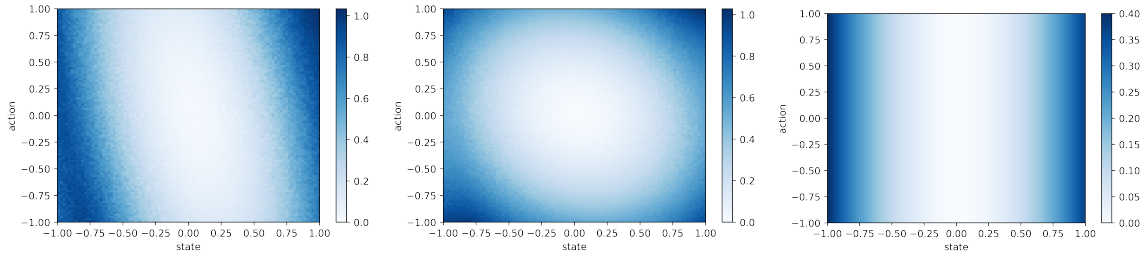


Figure A.4: Q tables with 100 states and 100 actions: $q_0(s, a)$, $q_4(s, a)$ and $q_5(s, a)$ (from left to right).

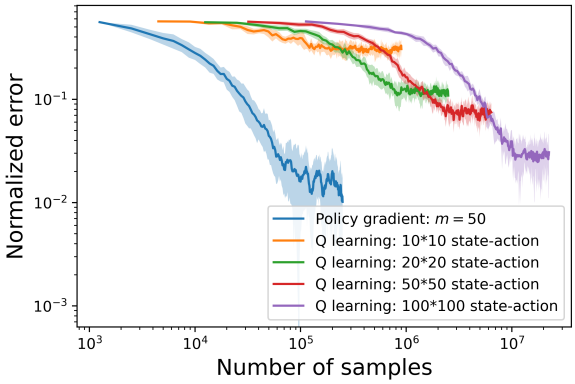


Figure A.5: Comparison between Q-learning and the policy gradient method (log-log scale). (Average across 10 scenarios.)

A.3 Proofs of Technical Results

We now give the proofs that were omitted in Chapter 2.

A.3.1 Proofs in Section 2.3.1

Proof of Lemma 2.3.2. Denote by $\{x_t\}_{t=0}^{T}$ the state trajectory induced by an arbitrary control $\boldsymbol{K}$. By Assumption 2.3.1 the matrix $\mathbb{E}[x_0 x_0^\top]$ is positive definite. For $t \geq 1$, we have

$$\mathbb{E}[x_t x_t^\top] = (A - BK_{t-1})\mathbb{E}[x_{t-1}x_{t-1}^\top](A - BK_{t-1})^\top + \mathbb{E}[w_{t-1}w_{t-1}^\top].$$

Now $(A - BK_{t-1})\mathbb{E}[x_{t-1}x_{t-1}^\top](A - BK_{t-1})^\top$ is positive semi-definite and $\mathbb{E}[w_{t-1}w_{t-1}^\top]$ is positive definite. Hence $\mathbb{E}[x_t x_t^\top]$ is positive definite and as a result $\underline{\sigma}_{\boldsymbol{X}} > 0$. The statement $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} \geq \underline{\sigma}_{\boldsymbol{X}}$ holds by the fact that, for real symmetric positive semi-definite A and B

$$\sigma_{\min}(A) + \sigma_{\min}(B) \leq \sigma_{\min}(A + B),$$

which is a consequence of the Weyl's inequality. □

Proof of Proposition 2.3.4. This can be proved by backward induction. For $t = T$, $P_T^{\boldsymbol{K}} = Q_T$ is positive definite since Q_T is positive definite. Assume $P_{t+1}^{\boldsymbol{K}}$ is positive definite for some $t + 1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P_t^{\boldsymbol{K}} z = z^\top Q_t\, z + z^\top K_t^\top R_t K_t z + z^\top \left(A - BK_t\right)^\top P_{t+1}^{\boldsymbol{K}} \left(A - BK_t\right) z > 0.$$

The last inequality holds since $z^\top Q_t\, z > 0$, $z^\top K_t^\top R_t K_t z \geq 0$ and $z^\top \left(A - BK_t\right)^\top P_{t+1}^{\boldsymbol{K}} \left(A - BK_t\right) z \geq 0$. By backward induction, we have $P_t^{\boldsymbol{K}}$ positive definite, $\forall\, t = 0, 1, \cdots, T$. □

To prove Lemma 2.3.6, let us start with a useful result for the value function. Define the value function $V_{\boldsymbol{K}}(x, \tau)$ for $\tau = 0, 1, \cdots, T-1$, as

$$V_{\boldsymbol{K}}(x,\tau) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=\tau}^{T-1}(x_t^\top Q_t x_t + u_t^\top R_t u_t) + x_T^\top Q_T x_T \middle| x_\tau = x\right] = x^\top P_\tau x + L_\tau,$$

with terminal condition

$$V_{\boldsymbol{K}}(x, T) = x^\top Q_T x,$$

where L_τ is defined in (2.3.10). We then define the Q function, $Q_{\boldsymbol{K}}(x, u, \tau)$ for $\tau = 0, 1, \cdots, T-1$ as

$$Q_{\boldsymbol{K}}(x,u,\tau) = x^\top Q_\tau x + u^\top R_\tau u + \mathbb{E}_{w_\tau}\left[V_{\boldsymbol{K}}(Ax + Bu + w_\tau, \tau + 1)\right],$$

and the advantage function

$$A_{\boldsymbol{K}}(x,u,\tau) = Q_{\boldsymbol{K}}(x,u,\tau) - V_{\boldsymbol{K}}(x,\tau).$$

Note that $C(\boldsymbol{K}) = \mathbb{E}_{x_0\sim\mathcal{D}}[V(x_0, 0)]$. Then we can write the difference of value functions between $\boldsymbol{K}$ and $\boldsymbol{K}'$ in terms of advantage functions.

Lemma A.3.1. *Assume $\boldsymbol{K}$ and $\boldsymbol{K}'$ have finite costs. Denote $\{x_t'\}_{t=0}^T$ and $\{u_t'\}_{t=0}^{T-1}$ as the state and control sequences of a single trajectory generated by $\boldsymbol{K}'$ starting from $x_0' = x_0 = x$, then*

$$V_{\boldsymbol{K}'}(x,0) - V_{\boldsymbol{K}}(x,0) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t', u_t', t)\right], \tag{A.3.1}$$

and $A_{\boldsymbol{K}}(x, -K_\tau' x, \tau) \quad = 2x^\top (K_\tau' - K_\tau)^\top E_\tau x + x^\top (K_\tau' - K_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)(K_\tau' - K_\tau)x$, where E_τ is defined in (2.3.11).

Proof. Denote by $c_t'(x)$ the cost generated by $\boldsymbol{K}'$ with a single trajectory starting from $x_0' = x_0 = x$. That is, $c_t'(x) = (x_t')^\top Q_t x_t' + (u_t')^\top R_t u_t'$, $t = 0, 1, \cdots, T-1$, and $c_T'(x) = (x_T')^\top Q_T x_T'$, with $u_t' = -K_t' x_t'$, $x_{t+1}' = Ax_t' + Bu_t' + w_t$, $x_0' = x$.

Therefore,

$$
\begin{aligned}
V_{\boldsymbol{K}'}(x,0) - V_{\boldsymbol{K}}(x,0) &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} c'_t(x)\right] - V_{\boldsymbol{K}}(x,0) \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} \left(c'_t(x) + V_{\boldsymbol{K}}(x'_t,t) - V_{\boldsymbol{K}}(x'_t,t)\right)\right] - V_{\boldsymbol{K}}(x,0) \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(c'_t(x) + V_{\boldsymbol{K}}(x'_{t+1},t+1) - V_{\boldsymbol{K}}(x'_t,t)\right)\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(Q_{\boldsymbol{K}}(x'_t,u'_t,t) - V_{\boldsymbol{K}}(x'_t,t)\right) \middle| x_0 = x\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t,u'_t,t) \middle| x_0 = x\right],
\end{aligned}
$$

where the third equality holds since $c'_T(x) = V_{\boldsymbol{K}}(x'_T, T)$ with the same single trajectory. For $u = -K'_\tau x$,

$$
\begin{aligned}
&A_{\boldsymbol{K}}(x, -K'_\tau x, \tau) \\
&= Q_{\boldsymbol{K}}(x, -K'_\tau x, \tau) - V_{\boldsymbol{K}}(x,\tau) \\
&= x^\top (Q_\tau + (K'_\tau)^\top R_\tau K'_\tau)x + \mathbb{E}_{w_\tau}\left[V_{\boldsymbol{K}}((A - BK'_\tau)x + w_\tau, \tau+1)\right] - V_{\boldsymbol{K}}(x,\tau) \\
&= x^\top (Q_\tau + (K'_\tau)^\top R_\tau K'_\tau)x + \left(x^\top (A - BK'_\tau)^\top P_{\tau+1}(A - BK'_\tau)x + \mathrm{Tr}(WP_{\tau+1}) + L_{\tau+1}\right) \\
&\quad - \left(x^\top P_\tau x + L_\tau\right) \\
&= x^\top (Q_\tau + (K'_\tau - K_\tau + K_\tau)^\top R_\tau (K'_\tau - K_\tau + K_\tau))x \\
&\quad + x^\top (A - BK_\tau - B(K'_\tau - K_\tau))^\top P_{\tau+1}(A - BK_\tau - B(K'_\tau - K_\tau))x \\
&\quad - x^\top (Q_\tau + K_\tau^\top R_\tau K_\tau + (A - BK_\tau)^\top P_{\tau+1}(A - BK_\tau))x \\
&= 2x^\top (K'_\tau - K_\tau)^\top ((R_\tau + B^\top P_{\tau+1}B)K_\tau - B^\top P_{\tau+1}A)x \\
&\quad + x^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1}B)(K'_\tau - K_\tau)x.
\end{aligned}
\tag{A.3.2}
$$

□

Proof of Lemma 2.3.6. First for any K'_τ, from (A.3.2),

$$\begin{aligned}
&A_{\boldsymbol{K}}(x, -K'_\tau x, \tau) \\
&= Q_{\boldsymbol{K}}(x, -K'_\tau x, \tau) - V_{\boldsymbol{K}}(x, \tau) \\
&= 2\,\mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top E_\tau) + \mathrm{Tr}(xx^\top (K'_\tau - K_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)(K'_\tau - K_\tau)) \\
&= \mathrm{Tr}\left(xx^\top (K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau)^\top (R_\tau + B^\top P_{\tau+1} B)\right. \\
&\quad \left.(K'_\tau - K_\tau + (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau)\right) - \mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau) \\
&\geq -\mathrm{Tr}(xx^\top E_\tau^\top (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau),
\end{aligned} \tag{A.3.3}$$

with equality holds when $K'_\tau = K_\tau - (R_\tau + B^\top P_{\tau+1} B)^{-1} E_\tau$. Then,

$$\begin{aligned}
C(\boldsymbol{K}) - C(\boldsymbol{K}^*) &= -\mathbb{E} \sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t^*, u_t^*, t) \leq \mathbb{E} \sum_{t=0}^{T-1} \mathrm{Tr}\left(x_t^* (x_t^*)^\top E_t^\top (R_t + B^\top P_{t+1} B)^{-1} E_t\right) \\
&\leq \|\Sigma_{\boldsymbol{K}^*}\| \sum_{t=0}^{T-1} \mathrm{Tr}(E_t^\top (R_t + B^\top P_{t+1} B)^{-1} E_t) \leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \mathrm{Tr}(E_t^\top E_t) \\
&\leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \mathrm{Tr}(\nabla_t C(\boldsymbol{K})^\top \nabla_t C(\boldsymbol{K})),
\end{aligned}$$

where $\underline{\sigma}_{\boldsymbol{X}}$ is defined in (2.3.5) and $\underline{\sigma}_{\boldsymbol{R}}$ is defined in (2.3.6). For the lower bound, consider $K'_t = K_t - (R_t + B^\top P_{t+1} B)^{-1} E_t$ where the equality holds in (A.3.3). Using $C(\boldsymbol{K}^*) \leq C(\boldsymbol{K}')$

$$\begin{aligned}
C(\boldsymbol{K}) - C(\boldsymbol{K}^*) \geq C(\boldsymbol{K}) - C(\boldsymbol{K'}) &= -\mathbb{E} \sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x'_t, u'_t, t) \\
&= \mathbb{E} \sum_{t=0}^{T-1} \mathrm{Tr}(x'_t (x'_t)^\top E_t^\top (R_t + B^\top P_{t+1} B)^{-1} E_t) \\
&\geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t + B^\top P_{t+1} B\|} \mathrm{Tr}(E_t^\top E_t).
\end{aligned} \tag{A.3.4}$$

□

Proof of Lemma 2.3.7. By lemma A.3.1 we have

$$\begin{aligned} C(\boldsymbol{K}') - C(\boldsymbol{K}) &= \mathbb{E}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}(x_t', -K_t' x_t', t)\right] \\ &= \sum_{t=0}^{T-1} \big(2\,\mathrm{Tr}(\Sigma_t'(K_t' - K_t)^\top E_t) \\ &\qquad + \mathrm{Tr}(\Sigma_t'(K_t' - K_t)^\top (R_t + B^\top P_{t+1} B)(K_t' - K_t))\big)\,. \end{aligned}$$

□

Proof of Lemma 2.3.8. For $t = 0, 1, \cdots, T$,

$$C(\boldsymbol{K}) \geq \mathbb{E}[x_t^\top P_t x_t] \geq \|P_t\| \sigma_{\min}(\mathbb{E}[x_t x_t^\top]) \geq \underline{\sigma}_{\boldsymbol{X}} \|P_t\|,$$

$$C(\boldsymbol{K}) = \sum_{t=0}^{T-1} \mathrm{Tr}(\mathbb{E}[x_t x_t^\top](Q_t + K_t^\top R_t K_t)) + \mathrm{Tr}(\mathbb{E}[x_T x_T^\top] Q_T) \geq \underline{\sigma}_{\boldsymbol{Q}} \mathrm{Tr}(\Sigma_{\boldsymbol{K}}) \geq \underline{\sigma}_{\boldsymbol{Q}} \|\Sigma_{\boldsymbol{K}}\|.$$

Therefore the statement in Lemma 2.3.8 follows provided that $\underline{\sigma}_{\boldsymbol{X}} > 0$ and Assumption 2.2.1 holds. □

Proof of Proposition 2.3.9. Recall that $\Sigma_t = \mathbb{E}\left[x_t x_t^\top\right]$. Note that

$$\begin{aligned} \Sigma_1 &= \mathbb{E}\left[x_1 x_1^\top\right] = \mathbb{E}\left[((A - B\,K_0)x_0 + w_0)\,((A - B\,K_0)x_0 + w_0)^\top\right] \\ &= (A - B\,K_0)\Sigma_0\,(A - B\,K_0)^\top + W = \mathcal{G}_0(\Sigma_0) + W. \end{aligned}$$

Now we first prove that

$$\Sigma_t = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top + W,\ \forall\, t = 2, 3, \cdots, T. \tag{A.3.5}$$

When $t = 2$,

$$\begin{aligned} \Sigma_2 &= \mathbb{E}\left[x_2 x_2^\top\right] = \mathbb{E}\left[((A - B\,K_1)x_1 + w_1)\,((A - B\,K_1)x_1 + w_1)^\top\right] \\ &= (A - B\,K_1)\Sigma_1\,(A - B\,K_1)^\top + W = \mathcal{G}_1(\Sigma_0) + (A - BK_1)W(A - BK_1)^\top + W, \end{aligned}$$

which satisfies (A.3.5). Assume (A.3.5) holds for $t \le k$. Then for $t = k+1$,

$$\begin{aligned}\mathbb{E}\left[x_{t+1}x_{t+1}^\top\right] &= \mathbb{E}\left[\left((A - B\,K_t)x_t + w_t\right)\left((A - B\,K_t)x_t + w_t\right)^\top\right] \\ &= (A - B\,K_t)\Sigma_t\,(A - B\,K_t)^\top + W = \mathcal{G}_t(\Sigma_0) + \sum_{s=1}^{t} D_{t,s}WD_{t,s}^\top + W.\end{aligned}$$

Therefore (A.3.5) holds, $\forall\, t = 1, 2, \cdots, T$. Finally,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T}\Sigma_t = \Sigma_0 + \sum_{t=0}^{T-1}\mathcal{G}_t(\Sigma_0) + \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s}WD_{t,s}^\top + TW = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta(\boldsymbol{K}, W).$$

□

A.3.2 Proofs in Section 2.3.2

Proof of Lemma 2.3.13. By direct calculation,

$$\|\mathcal{G}_t\| \le \rho^{2(t+1)}, \quad \text{and} \quad \|\mathcal{G}_t'\| \le \rho^{2(t+1)}. \tag{A.3.6}$$

Denote $\mathcal{F}_t = \mathcal{F}_{K_t}$ and $\mathcal{F}_t' = \mathcal{F}_{K_t'}$ to ease the exposition. Then for any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $t \ge 0$,

$$\begin{aligned}\|(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1})(\Sigma)\| &= \|\mathcal{F}_{t+1}'\circ\mathcal{G}_t'(\Sigma) - \mathcal{F}_{t+1}\circ\mathcal{G}_t(\Sigma)\| \\ &= \|\mathcal{F}_{t+1}'\circ\mathcal{G}_t'(\Sigma) - \mathcal{F}_{t+1}'\circ\mathcal{G}_t(\Sigma) + \mathcal{F}_{t+1}'\circ\mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1}\circ\mathcal{G}_t(\Sigma)\| \\ &\le \|\mathcal{F}_{t+1}'\circ\mathcal{G}_t'(\Sigma) - \mathcal{F}_{t+1}'\circ\mathcal{G}_t(\Sigma)\| + \|\mathcal{F}_{t+1}'\circ\mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1}\circ\mathcal{G}_t(\Sigma)\| \\ &= \|\mathcal{F}_{t+1}'\circ(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma)\| + \|(\mathcal{F}_{t+1}' - \mathcal{F}_{t+1})\circ\mathcal{G}_t(\Sigma)\| \\ &\le \|\mathcal{F}_{t+1}'\|\,\|(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma)\| + \|\mathcal{G}_t\|\,\|\mathcal{F}_{t+1}' - \mathcal{F}_{t+1}\|\,\|\Sigma\| \\ &\le \rho^2\|(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)}\|\mathcal{F}_{t+1}' - \mathcal{F}_{t+1}\|\|\Sigma\|.\end{aligned}$$

Therefore,

$$\|(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1})(\Sigma)\| \le \rho^2\|(\mathcal{G}_t' - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)}\|\mathcal{F}_{t+1}' - \mathcal{F}_{t+1}\|\|\Sigma\|. \tag{A.3.7}$$

Summing (A.3.7) up for $t \in \{1, 2, \cdots, T-2\}$ with $\|\mathcal{G}_0' - \mathcal{G}_0\| = \|\mathcal{F}_0' - \mathcal{F}_0\|$, we have

$$\sum_{t=0}^{T-1} \Big\|(\mathcal{G}_t - \mathcal{G}_t')(\Sigma)\Big\| \leq \frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_t - \mathcal{F}_t'\|\Big)\|\Sigma\|.$$

□

A.3.3 Proofs in Section 2.3.3

Proof of Lemma 2.3.15. Given (2.3.22) and condition (2.3.23), we have $\|K_t' - K_t\| = \eta\|\nabla_t C(\boldsymbol{K})\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})\|B\|}$. Therefore,

$$\|B\|\|K_t' - K_t\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})} \leq \frac{1}{2}. \tag{A.3.8}$$

The last inequality holds since $\underline{\sigma}_{\boldsymbol{X}} \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$ given by Lemma 2.3.8. Therefore, by Lemma 2.3.12,

$$\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \leq (2\rho+1)\|B\|\left(\sum_{t=0}^{T-1}\|K_t - K_t'\|\right). \tag{A.3.9}$$

By Lemmas 2.3.5 and 2.3.7,

$$\begin{aligned}
& C(\boldsymbol{K}') - C(\boldsymbol{K}) \\
= & \sum_{t=0}^{T-1}\Big[2\operatorname{Tr}\Big(\Sigma_t'(K_t'-K_t)^\top E_t\Big) + \operatorname{Tr}\Big(\Sigma_t'(K_t'-K_t)^\top(R_t + B^\top P_{t+1}B)(K_t'-K_t)\Big)\Big] \\
= & \sum_{t=0}^{T-1}\Big[-4\eta\operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top E_t\Big) + 4\eta^2\operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top(R_t + B^\top P_{t+1}B)E_t\Sigma_t\Big)\Big] \\
= & \sum_{t=0}^{T-1}\Big[-4\eta\operatorname{Tr}\Big((\Sigma_t' - \Sigma_t + \Sigma_t)\Sigma_t E_t^\top E_t\Big) + 4\eta^2\operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top(R_t + B^\top P_{t+1}B)E_t\Sigma_t\Big)\Big] \\
\leq & \sum_{t=0}^{T-1}\Big[-4\eta\operatorname{Tr}\Big(\Sigma_t E_t^\top E_t\Sigma_t\Big) + 4\eta\operatorname{Tr}((\Sigma_t' - \Sigma_t)\Sigma_t E_t^\top E_t\Sigma_t\Sigma_t^{-1}) \\
& + 4\eta^2\operatorname{Tr}\Big(\Sigma_t'\Sigma_t E_t^\top(R_t + B^\top P_{t+1}B)E_t\Sigma_t\Big)\Big] \\
\leq & \sum_{t=0}^{T-1}\Big[-4\eta\operatorname{Tr}\Big(\Sigma_t E_t^\top E_t\Sigma_t\Big) + 4\eta\frac{\|\Sigma_t' - \Sigma_t\|}{\sigma_{\min}(\Sigma_t)}\operatorname{Tr}\Big(\Sigma_t E_t^\top E_t\Sigma_t\Big)
\end{aligned}$$

$$
\begin{aligned}
&+4\eta^2\|\Sigma_t'(R_t+B^\top P_{t+1}B)\|\,\mathrm{Tr}\left(\Sigma_t E_t^\top E_t\Sigma_t\right)\Big] \\
\leq\ & -\eta\Big(1-\frac{\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}}-\eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|\Big)\cdot \\
& \sum_{t=0}^{T-1}\Big[\mathrm{Tr}(\nabla_t C(\boldsymbol{K})^\top\nabla_t C(\boldsymbol{K}))\Big]. \qquad \text{(A.3.10)}
\end{aligned}
$$

By Lemma 2.3.6, we have

$$
\begin{aligned}
C(\boldsymbol{K}')-C(\boldsymbol{K})\leq -\eta\Big(1-\frac{\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}}-\eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|\Big)\cdot \\
\Big(\frac{4\,\underline{\sigma}_{\boldsymbol{X}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\Big)\Big(C(\boldsymbol{K})-C(\boldsymbol{K}^*)\Big) \qquad \text{(A.3.11)}
\end{aligned}
$$

provided that

$$
1-\frac{\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}}-\eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|>0. \qquad \text{(A.3.12)}
$$

By (2.3.21), (2.3.22), and (A.3.8),

$$
\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|\leq\frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)\left(\eta(2\rho+1)\|B\|\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\right).
$$

Given the step size condition in (2.3.23), we have

$$
\begin{aligned}
\eta(2\rho+1)\|B\|\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\| &\leq \eta(2\rho+1)\|B\|\Big(T\cdot\max_t\{\|\nabla_t C(\boldsymbol{K})\|\}\Big) \\
&\leq \frac{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2(\rho^{2T}-1)(C(\boldsymbol{K})+\underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|)}. \qquad \text{(A.3.13)}
\end{aligned}
$$

Then, by Corollary 2.3.14 and (A.3.9),

$$
\frac{\|\Sigma_{\boldsymbol{K}'}-\Sigma_{\boldsymbol{K}}\|}{\underline{\sigma}_{\boldsymbol{X}}}\leq\frac{\rho^{2T}-1}{\rho^2-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t}-\mathcal{F}_{K_t'}\|\Big)\frac{\|\Sigma_0\|+T\|W\|}{\underline{\sigma}_{\boldsymbol{X}}}
$$

$$\leq \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\left(\sum_{t=0}^{T-1}\eta\|\nabla_t C(\boldsymbol{K}\|\right)\frac{C(\boldsymbol{K})+\underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|}{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}\leq\frac{1}{2},$$

where the last step holds by (A.3.13). Therefore, the bound of $\|\Sigma_{\boldsymbol{K}'}\|$ in (A.3.12) is given by

$$\|\Sigma_{\boldsymbol{K}'}\|\leq\|\Sigma_{\boldsymbol{K}'}-\Sigma_{\boldsymbol{K}}\|+\|\Sigma_{\boldsymbol{K}}\|\leq\frac{1}{2}\underline{\sigma}_{\boldsymbol{X}}+\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\leq\frac{1}{2}\|\Sigma_{\boldsymbol{K}'}\|+\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}},\tag{A.3.14}$$

which indicates that $\|\Sigma_{\boldsymbol{K}'}\|\leq\frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$. Therefore, (A.3.12) gives

$$\begin{aligned}
&1-\frac{\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}}-\eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|\\
&\geq 1-\frac{(\rho^{2T}-1)}{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)\left(\eta(2\rho+1)\|B\|\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\right)\\
&\quad-\eta\frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|\\
&=1-C_1\eta,
\end{aligned}$$

where C_1 is defined in (2.3.24). So if $\eta\leq\frac{1}{2C_1}$, then,

$$1-\frac{\sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|}{\underline{\sigma}_{\boldsymbol{X}}}-\eta\|\Sigma_{\boldsymbol{K}'}\|\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|\geq 1-C_1\eta\geq\frac{1}{2}>0.$$

Hence,$C(\boldsymbol{K}')-C(\boldsymbol{K})\leq-\frac{\eta}{2}\left(\frac{4\underline{\sigma}_{\boldsymbol{X}}{}^2\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\Big(C(\boldsymbol{K})-C(\boldsymbol{K}^*)\Big)$, and

$$\begin{aligned}
C(\boldsymbol{K}')-C(\boldsymbol{K}^*) &= (C(\boldsymbol{K}')-C(\boldsymbol{K}))+(C(\boldsymbol{K})-C(\boldsymbol{K}^*))\\
&\leq \left(1-2\eta\frac{\underline{\sigma}_{\boldsymbol{X}}{}^2\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\Big(C(\boldsymbol{K})-C(\boldsymbol{K}^*)\Big).
\end{aligned}$$

□

A.3.4 Proofs in Section 2.4

Before proceeding to the proof of Theorem 2.4.5, we show two important Lemmas which provide the intermediate steps. We first show the optimality condition for the

projection operator in Lemma A.3.2. We then show the one-step convergence result in Lemma A.3.3.

Lemma A.3.2 (Optimality Condition)**.** *Fix a policy matrix* $\boldsymbol{L}^1$ *and write* $\boldsymbol{L}^* = \Pi_{\mathcal{S}}(\boldsymbol{L}^1)$. *Then for any* $\boldsymbol{L}^0 \in \mathcal{S}$, *we have*

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^0 - L_t^*)(L_t^* - L_t^1)^\top\right) \geq 0. \tag{A.3.15}$$

Proof of Lemma A.3.2. We show condition (A.3.15) by contradiction. Assume condition (A.3.15) does not hold, then there exist some $\boldsymbol{L}^3 \in \mathcal{S}$ and some constant $b > 0$ such that

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right) = -b < 0. \tag{A.3.16}$$

Let

$$M = 1 + \frac{\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right)}{-2\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^* - L_t^1)^\top\right)} = 1 + \frac{\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right)}{2b} > 1,$$

and take

$$\overline{\boldsymbol{L}} = \frac{1}{M}\boldsymbol{L}^3 + \left(1 - \frac{1}{M}\right)\boldsymbol{L}^*.$$

By the convexity of $\mathcal{S}$, we have $\overline{\boldsymbol{L}} \in \mathcal{S}$. Hence from definition (2.4.3), we have

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^* - L_t^1)(L_t^* - L_t^1)^\top\right) \leq \sum_{t=0}^{T-1} \mathrm{Tr}\left((\overline{L}_t - L_t^1)(\overline{L}_t - L_t^1)^\top\right). \tag{A.3.17}$$

On the other hand,

$$\begin{aligned}
& \sum_{t=0}^{T-1} \mathrm{Tr}\left((\overline{L}_t - L_t^1)(\overline{L}_t - L_t^1)^\top\right) \\
= & \sum_{t=0}^{T-1} \mathrm{Tr}\left(\left(\frac{1}{M}(L_t^3 - L_t^*) + (L_t^* - L_t^1)\right)\left(\frac{1}{M}(L_t^3 - L_t^*) + (L_t^* - L_t^1)\right)^\top\right) \\
= & \sum_{t=0}^{T-1} \frac{1}{M^2} \mathrm{Tr}\left((L_t^3 - L_t^*)(L_t^3 - L_t^*)^\top\right) + \sum_{t=0}^{T-1} \mathrm{Tr}\left((L_t^* - L_t^1)(L_t^* - L_t^1)^\top\right)
\end{aligned}$$

$$+\sum_{t=0}^{T-1}\frac{2}{M}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^*-L_t^1)^\top\right). \tag{A.3.18}$$

By the definition of M we have,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^3-L_t^*)^\top\right)+2M\sum_{t=0}^{T-1}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^*-L_t^1)^\top\right)\\
=&\sum_{t=0}^{T-1}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^3-L_t^*)^\top\right)-2Mb\\
=&\sum_{t=0}^{T-1}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^3-L_t^*)^\top\right)-2b-\sum_{t=0}^{T-1}\operatorname{Tr}\left((L_t^3-L_t^*)(L_t^3-L_t^*)^\top\right)=-2b<0.
\end{aligned}$$

Thus substituting this in (A.3.18) contradicts (A.3.17) which completes the proof.

□

Lemma A.3.3. *Assume Assumption 2.1 holds,* $\underline{\sigma}_{\boldsymbol{X}}>0$, $\boldsymbol{K}\in\mathcal{S}$ *and that*

$$K_t'=K_t-\eta\nabla_t C(\boldsymbol{K}),\qquad \text{where} \tag{A.3.19}$$

$$\eta\le\min\left\{C_1,C_2\right\},\qquad \text{with} \tag{A.3.20}$$

$$C_1=\frac{(\rho^2-1)\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{4d\,T^2\sqrt{d+k}(\rho^{2T}-1)(2\rho+1)(C(\boldsymbol{K})+\underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|)\|B\|\max_t\{\|\nabla_t C(\boldsymbol{K})\|\}} \tag{A.3.21}$$

$$C_2=\frac{\underline{\sigma}_{\boldsymbol{Q}}}{8C(\boldsymbol{K})\sum_{t=0}^{T-1}\|R_t+B^\top P_{t+1}B\|}. \tag{A.3.22}$$

Take

$$\widetilde{\boldsymbol{K}}=\Pi_{\mathcal{S}}\left(\boldsymbol{K}'\right), \tag{A.3.23}$$

with $\boldsymbol{K}'=(K_0',\cdots,K_{T-1}')$ *and* K_t' *defined in* (A.3.19) $(t=0,1,\cdots,T-1)$. *Then we have*

$$2\eta\sum_{t=0}^{T}\|G_t(\boldsymbol{K})\|_F^2=2\eta\sum_{t=0}^{T}\operatorname{Tr}\left(G_t(\boldsymbol{K})^\top G_t(\boldsymbol{K})\right)\le C(\boldsymbol{K})-C(\widetilde{\boldsymbol{K}}).$$

Proof. By definition of $\widetilde{\boldsymbol{K}}$, and as $\boldsymbol{K} \in \mathcal{S}$, we have

$$\sum_{t=0}^{T-1} \operatorname{Tr}\left((\widetilde{K}_t - K'_t)(\widetilde{K}_t - K'_t)^\top\right) \leq \sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - K'_t)(K_t - K'_t)^\top\right). \tag{A.3.24}$$

Take $\boldsymbol{L}^1 = \boldsymbol{K}'$ and $\boldsymbol{L}^0 = \boldsymbol{K}$ in Lemma A.3.2, we have

$$\sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - \widetilde{K}_t)(\widetilde{K}_t - K'_t)^\top\right) \geq 0. \tag{A.3.25}$$

Combining (A.3.24) and (A.3.25) leads to

$$\sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - K'_t)(K_t - \widetilde{K}_t)^\top\right) \geq \sum_{t=0}^{T-1} \operatorname{Tr}\left((K_t - \widetilde{K}_t)(K_t - \widetilde{K}_t)^\top\right). \tag{A.3.26}$$

Given the definition (A.3.23), we have $G(\boldsymbol{K}) = \frac{\widetilde{\boldsymbol{K}} - \boldsymbol{K}}{2\eta}$ and $G_t(\boldsymbol{K}) = \frac{\widetilde{K}_t - K_t}{2\eta}$. By Lemmas 2.3.5 and 2.3.7,

$$\begin{aligned}
& C(\widetilde{\boldsymbol{K}}) - C(\boldsymbol{K}) \\
= & \sum_{t=0}^{T-1} \left[2 \operatorname{Tr}\left(\widetilde{\Sigma}_t(\widetilde{K}_t - K_t)^\top E_t\right) + \operatorname{Tr}\left(\widetilde{\Sigma}_t(\widetilde{K}_t - K_t)^\top (R_t + B^\top P_{t+1} B)(\widetilde{K}_t - K_t)\right)\right] \\
= & \sum_{t=0}^{T-1} \Big[4\eta \operatorname{Tr}\left((\widetilde{\Sigma}_t - \Sigma_t)(G_t(\boldsymbol{K}))^\top E_t\right) + 4\eta \operatorname{Tr}\left(\Sigma_t (G_t(\boldsymbol{K}))^\top E_t\right) \\
& + 4\eta^2 \operatorname{Tr}\left(\widetilde{\Sigma}_t (G_t(\boldsymbol{K}))^\top (R_t + B^\top P_{t+1} B) G_t(\boldsymbol{K})\right)\Big],
\end{aligned} \tag{A.3.27}$$

with $\widetilde{\Sigma}_t := \mathbb{E}[\widetilde{x}_t \widetilde{x}_t^\top]$ and $\{\widetilde{x}_t\}_{t=0}^{T-1}$ is the trajectory under policy $\widetilde{\boldsymbol{K}}$.

First, we have

$$\begin{aligned}
\sum_{t=0}^{T-1} \operatorname{Tr}\left(\Sigma_t (G_t(\boldsymbol{K}))^\top E_t\right) &= \sum_{t=0}^{T-1} \operatorname{Tr}\left((G_t(\boldsymbol{K}))^\top E_t \Sigma_t\right) \\
&= \frac{1}{4\eta^2} \sum_{t=0}^{T-1} \operatorname{Tr}((\widetilde{K}_t - K_t)^\top (K_t - K'_t)) \\
&\leq -\frac{1}{4\eta^2} \sum_{t=0}^{T-1} \operatorname{Tr}((\widetilde{K}_t - K_t)(\widetilde{K}_t - K_t)^\top) \\
&= -\operatorname{Tr}((G_t(\boldsymbol{K}))^\top (G_t(\boldsymbol{K}))),
\end{aligned}$$

in which the last inequality holds by (A.3.26).

Second given (A.3.19) and condition (A.3.21), we have

$$\|K_t' - K_t\| = \eta\|\nabla_t C(\boldsymbol{K})\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2\,T\sqrt{d+k}C(\boldsymbol{K})\|B\|}.$$

Therefore,

$$\begin{aligned}\sum_{t=0}^{T-1}\|B\|\|\tilde{K}_t - K_t\| &\leq \sum_{t=0}^{T-1}\|B\|\|\tilde{K}_t - K_t\|_F \leq \sum_{t=0}^{T-1}\|B\|\|K_t' - K_t\|_F \\ &\leq \sqrt{d+k}\|B\|\sum_{t=0}^{T-1}\|K_t' - K_t\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}}}{2C(\boldsymbol{K})} \leq \frac{1}{2}.\end{aligned}$$

The second inequality holds by (A.3.26) and the last inequality holds since $\underline{\sigma}_{\boldsymbol{X}} \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$ given by Lemma 2.3.8. By (2.3.21),

$$\begin{aligned}&\sum_{t=0}^{T-1}\|\widetilde{\Sigma}_t - \Sigma_t\| \\ \leq\ & \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\left(2\rho\,\|B\|\,|||\boldsymbol{K} - \widetilde{\boldsymbol{K}}||| + \|B\|^2\,|||\boldsymbol{K} - \widetilde{\boldsymbol{K}}|||^2\right) \\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\left(2(2\rho+1)\|B\|\sum_{t=0}^{T-1}\eta\|G_t(\boldsymbol{K})\|\right) \\ &\leq \frac{\underline{\sigma}_{\boldsymbol{X}}}{2d\,T^2\sqrt{d+k}\cdot\max_t\|\nabla_t C(\boldsymbol{K})\|}\sum_{t=0}^{T-1}\|G_t(\boldsymbol{K})\|, \qquad \text{(A.3.28)}\end{aligned}$$

where the last inequality holds by step size condition $\eta \leq C_1$. Hence

$$\begin{aligned}\text{(A.3.27)} \leq\ & \sum_{t=0}^{T-1}\Bigg[2\eta\frac{d\|\widetilde{\Sigma}_t - \Sigma_t\|}{\sigma_{\min}(\Sigma_t)}\|G_t(\boldsymbol{K})\|\,\|\nabla_t C(\boldsymbol{K})\| - 4\eta\,\mathrm{Tr}((G_t(\boldsymbol{K}))^\top(G_t(\boldsymbol{K}))) \\ &+4\eta^2\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\,\|R_t + B^\top P_{t+1}B\|\,\mathrm{Tr}\left((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\right)\Bigg] \\ \leq\ & 2\eta\frac{d}{\underline{\sigma}_{\boldsymbol{X}}}\Big(\sum_{t=0}^{T-1}\|\widetilde{\Sigma}_t - \Sigma_t\|\Big)\Big(\sum_{t=0}^{T-1}\|G_t(\boldsymbol{K})\|\Big)\Big(\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\Big) \\ &-4\eta\sum_{t=0}^{T-1}\mathrm{Tr}((G_t(\boldsymbol{K}))^\top(G_t(\boldsymbol{K})))\end{aligned}$$

$$
\begin{aligned}
&+4\eta^2\sum_{t=0}^{T-1}\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\,\|R_t+B^\top P_{t+1}B\|\,\mathrm{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big)\\
\leq\;& \frac{\eta}{T}\Big(\sum_{t=0}^{T-1}\|G_t(\boldsymbol{K})\|\Big)^2-4\eta\sum_{t=0}^{T-1}\mathrm{Tr}((G_t(\boldsymbol{K}))^\top(G_t(\boldsymbol{K})))\\
&+4\eta^2\sum_{t=0}^{T-1}\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\,\|R_t+B^\top P_{t+1}B\|\,\mathrm{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big)\\
\leq\;& \sum_{t=0}^{T-1}\big[\eta\|G_t(\boldsymbol{K})\|^2-4\eta\,\mathrm{Tr}((G_t(\boldsymbol{K}))^\top(G_t(\boldsymbol{K})))\\
&+4\eta^2\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\,\|R_t+B^\top P_{t+1}B\|\,\mathrm{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big)\big]\\
\leq\;& \sum_{t=0}^{T-1}\eta\left[-3+4\eta\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\,\|R_t+B^\top P_{t+1}B\|\right]\mathrm{Tr}\Big((G_t(\boldsymbol{K}))^\top G_t(\boldsymbol{K})\Big),
\end{aligned}
$$

where the third inequality holds by (A.3.28) and the fourth inequality holds by Cauchy-Schwarz inequality. By (A.3.26) we have $\sqrt{d+k}\sum_{t=0}^{T-1}\|\nabla_t C(\boldsymbol{K})\|\geq\sum_{t=0}^{T-1}\|G_t(\boldsymbol{K})\|$ and thus (A.3.28) $\leq\frac{\sigma_{\boldsymbol{X}}}{2}$ and

$$
\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\leq\big\|\Sigma_{\widetilde{\boldsymbol{K}}}-\Sigma_{\boldsymbol{K}}+\Sigma_{\boldsymbol{K}}\big\|\leq\frac{\sigma_{\boldsymbol{X}}}{2}+\|\Sigma_{\boldsymbol{K}}\|\leq\frac{\|\Sigma_{\widetilde{\boldsymbol{K}}}\|}{2}+\frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}.
$$

Thus $\|\Sigma_{\widetilde{\boldsymbol{K}}}\|\leq\frac{2C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}$. Therefore when $\eta\leq C_2$, we have

$$
\text{(A.3.27)}\leq-2\eta\sum_{t=0}^{T}\mathrm{Tr}((G_t(\boldsymbol{K}))^\top(G_t(\boldsymbol{K}))).
$$

□

Proof of Theorem 2.4.5. The key step in this proof is Lemma A.3.3 and it suffices to show that the projected policy gradient method enjoys sublinear convergence rate in the setting of known paramaters. This is because moving from the analysis for the case of known parameters to that for the case of unknown parameters follows the same procedure of policy gradient descent (without projection). In particular, the zeroth order estimation of the gradient term $\eta\nabla_t C(\boldsymbol{K})$ in (A.3.19) is the same for policy gradient method and projected policy gradient method.

We now show that the projected policy gradient method with known parameters enjoys a sublinear convergence rate. Since the step size conditions (A.3.20)-(A.3.22)

are independent of the term $G_t(\boldsymbol{K})$, the existence of η follows the analysis in Theorem 2.3.3. Hence when $\eta \in \mathcal{H}(\frac{1}{C(\boldsymbol{K}^0)+1})$ is an appropriate polynomial in $\frac{1}{C(\boldsymbol{K}^0)+1}$ and model parameters, by Lemma A.3.3, we have for any $N \in \mathbb{N}^+$,

$$\sum_{n=1}^{N}\Big(\sum_{t=0}^{T-1}\mathrm{Tr}((G_t^{\boldsymbol{K}^n})^\top(G_t^{\boldsymbol{K}^n}))\Big)\leq\frac{\sum_{n=1}^{N}C(\boldsymbol{K}^{n-1})-C(\boldsymbol{K}^n)}{2\eta}\leq\frac{C(\boldsymbol{K}^0)-C(\boldsymbol{K}^*)}{2\eta}. \tag{A.3.29}$$

Therefore $\left\{\frac{1}{N}\sum_{n=0}^{N-1}\left(\sum_{t=0}^{T-1}\|G_t(\boldsymbol{K}^n)\|_F^2\right)\right\}_{N\geq 1}$ converges at rate $\mathcal{O}\left(\frac{1}{N}\right)$, which thus completes the proof for the case of known parameters. □

Proof of Lemma 2.4.7. Under Assumption 2.4.3, we have $\mathbb{E}\left[x_0x_0^\top\right]=\widetilde{W}_0\mathbb{E}\left[z_0z_0^\top\right]\widetilde{W}_0^\top$, and $\left\|\mathbb{E}\left[x_0x_0^\top\right]\right\|\leq\sigma_0^2\|\widetilde{W}_0\|^2$. With the sub-Gaussian distributed noise, $W=\mathbb{E}\left[w_tw_t^\top\right]=\widetilde{W}\mathbb{E}\left[v_tv_t^\top\right]\widetilde{W}^\top$, then we have $\|W\|\leq\sigma_w^2\left\|\widetilde{W}^2\right\|$.
Denote $S_t=Q_t+K_t^TR_tK_t$, $\forall\, t=1,\cdots,T-1$. Thus, for $t=0,1,\cdots,T-2$,

$$\begin{aligned}\mathbb{E}[x_{t+1}^\top Q_{t+1}x_{t+1}+u_{t+1}^\top R_{t+1}u_{t+1}]&=\mathbb{E}[x_{t+1}^\top S_{t+1}x_{t+1}]=\mathrm{Tr}(\mathbb{E}[x_{t+1}x_{t+1}^\top]S_{t+1})\\&=\mathrm{Tr}\left(\mathcal{G}_t(\Sigma_0)S_{t+1}+\sum_{s=1}^{t}D_{t,s}WD_{t,s}^\top S_{t+1}+WS_{t+1}\right).\end{aligned}$$

The last equality holds by (A.3.5). Therefore,

$$\begin{aligned}&C(\boldsymbol{K}')-C(\boldsymbol{K})\\=\;&\underbrace{\mathbb{E}[x_0^\top(K_0')^\top R_0K_0'x_0-x_0^\top K_0^\top R_0K_0x_0]}_{(I)}+\underbrace{\sum_{t=0}^{T-2}\mathrm{Tr}\left(\mathcal{G}_t'(\Sigma_0)S_{t+1}'-\mathcal{G}_t(\Sigma_0)S_{t+1}\right)}_{(II)}\\&+\underbrace{\sum_{t=0}^{T-2}\mathrm{Tr}\Big(\sum_{s=1}^{t}\left(D_{t,s}'W(D_{t,s}')^\top S_{t+1}'-D_{t,s}WD_{t,s}^\top S_{t+1}\right)+W(S_{t+1}'-S_{t+1})\Big)}_{(III)}\\&+\underbrace{\mathrm{Tr}\Big(\mathcal{G}_{T-1}(\Sigma_0)Q_T-\mathcal{G}_{T-1}'(\Sigma_0)Q_T+\sum_{s=1}^{T-1}\Big(D_{T-1,s}'W(D_{T-1,s}')^\top Q_T-D_{T-1,s}WD_{T-1,s}^\top Q_T\Big)\Big)}_{(IV)}.\end{aligned}$$

For the first term, $(I)\leq\mathrm{Tr}(\mathbb{E}[x_0x_0^\top])\|(K_0')^\top R_0K_0'-K_0^\top R_0K_0\|$. For the second term

(II), since

$$\begin{aligned}\sum_{t=0}^{T-2}\left(\mathrm{Tr}\left(\mathcal{G}_t(\Sigma_0)S_{t+1}\right)\right) &= \mathbb{E}\left[\sum_{t=0}^{T-2}\left(\mathrm{Tr}\left(\Pi_{i=0}^{t}(A-BK_i)x_0x_0^\top\Pi_{i=0}^{t}(A-BK_{t-i})^\top S_{t+1}\right)\right)\right]\\ &\leq \mathrm{Tr}\left(\mathbb{E}\left[x_0x_0^\top\right]\right)\left\|\sum_{t=0}^{T-2}\mathcal{G}_t(S_{t+1})\right\|,\end{aligned}$$

we have, $(II)\leq \mathrm{Tr}\left(\mathbb{E}\left[x_0x_0^\top\right]\right)\left\|\sum_{t=0}^{T-2}\left(\mathcal{G}_t'\left(S_{t+1}'\right)-\mathcal{G}_t\left(S_{t+1}\right)\right)\right\|$.
We denote $\mathcal{G}_d:=\sum_{t=0}^{T-2}\left(\mathcal{G}_t'\left(S_{t+1}'\right)-\mathcal{G}_t\left(S_{t+1}\right)\right)$, then

$$\begin{aligned}\|\mathcal{G}_d\| &\leq \sum_{t=0}^{T-2}\Big\|\mathcal{G}_t'\left(Q_{t+1}+(K_{t+1}')^\top R_{t+1}K_{t+1}'\right)-\mathcal{G}_t\left(Q_{t+1}+(K_{t+1}')^\top R_{t+1}K_{t+1}'\right)-\\ &\quad \mathcal{G}_t\circ\left(K_{t+1}^\top R_{t+1}K_{t+1}-(K_{t+1}')^\top R_{t+1}K_{t+1}'\right)\Big\|\\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\left((2\rho+1)\|B\|\sum_{t=0}^{T-2}\|K_t-K_t'\|\right)\left(\sum_{t=1}^{T-1}\|Q_t+(K_t')^\top R_tK_t'\|\right)\\ &\quad +\sum_{t=0}^{T-2}\|\mathcal{G}_t\|\left\|(K_{t+1}')^\top R_{t+1}K_{t+1}'-K_{t+1}^\top R_{t+1}K_{t+1}\right\|\\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}\left((2\rho+1)\|B\|\sum_{t=0}^{T-2}\|K_t-K_t'\|\right)\cdot\\ &\quad \left(\sum_{t=1}^{T-1}\|Q_t+(K_t')^\top R_tK_t'-K_t^\top R_tK_t+K_t^\top R_tK_t\|\right)\\ &\quad +\frac{\rho^2(\rho^{2(T-1)}-1)}{\rho^2-1}\sum_{t=1}^{T-1}\left\|(K_t')^\top R_tK_t'-K_t^\top R_tK_t\right\|\\ &\leq \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}'-\boldsymbol{K}|||\left(|||\boldsymbol{Q}|||+|||\boldsymbol{K}|||^2\,|||\boldsymbol{R}|||\right)\\ &\quad +\left(\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}'-\boldsymbol{K}|||+\frac{\rho^2(\rho^{2(T-1)}-1)}{\rho^2-1}\right)\cdot\\ &\quad \sum_{t=1}^{T-1}\left\|(K_t')^\top R_tK_t'-K_t^\top R_tK_t\right\|. \qquad (A.3.30)\end{aligned}$$

where the second inequality holds by Lemma 2.3.13 and (A.3.9), and the third inequal-

ity holds by (A.3.6). For the first term in (III), we have

$$
\begin{aligned}
&\sum_{t=0}^{T-2} \operatorname{Tr}\left(\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top S'_{t+1} - D_{t,s} W D_{t,s}^\top S_{t+1}\right) \\
= &\sum_{t=0}^{T-2} \operatorname{Tr}\left(\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top (S'_{t+1} - S_{t+1}) + (D'_{t,s} W (D'_{t,s})^\top - D_{t,s} W D_{t,s}^\top) S_{t+1}\right) \\
\le &\Big(\sum_{t=0}^{T-2}\sum_{s=1}^{t} \operatorname{Tr}(W)\|D'_{t,s}\|^2\Big) \left\|\sum_{t=1}^{T-1} (K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\| \\
&+ \sum_{t=0}^{T-2} \left\|\sum_{s=1}^{t} D'_{t,s} W (D'_{t,s})^\top - D_{t,s} W D_{t,s}^\top\right\| \Big(\sum_{t=1}^{T-1} \operatorname{Tr}(Q_t) + \|K_t\|^2 \operatorname{Tr}(R_t)\Big) \\
\le &\operatorname{Tr}(W) \frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1} \left\|\sum_{t=1}^{T-1} (K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\| \\
&+ T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\Big(\operatorname{Tr}\big(\sum_{t=1}^{T-1} Q_t\big) + |||\boldsymbol{K}|||^2 \operatorname{Tr}\big(\sum_{t=1}^{T-1} R_t\big)\Big),
\end{aligned}
$$

where the last step holds by (2.3.20). The second term in (III) is bounded by

$$
\sum_{t=0}^{T-2} \operatorname{Tr}\Big(W(S'_{t+1} - S_{t+1})\Big) \le \operatorname{Tr}(W) \sum_{t=1}^{T-1} \left\|(K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\|.
$$

Similarly, by (2.3.20) and (A.3.9), (IV) is bounded by

$$
\begin{aligned}
(IV) \;\le\; &\operatorname{Tr}(\mathbb{E}[x_0 x_0^\top]) \sum_{t=0}^{T-1} \Big\|(\mathcal{G}'_t - \mathcal{G}_t)(Q_T)\Big\| \\
&+ \operatorname{Tr}\left(\sum_{s=1}^{T-1} D'_{T-1,s} W (D'_{T-1,s})^\top Q_T - D_{T-1,s} W D_{T-1,s}^\top Q_T\right) \\
\le\; &\operatorname{Tr}(\mathbb{E}[x_0 x_0^\top]) \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| \\
&+ \operatorname{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.
\end{aligned}
$$

Now we bound the term $\sum_{t=1}^{T-1} \left\|(K'_t)^\top R_t K'_t - K_t^\top R_t K_t\right\|$, which appears several times

in previous inequalities:

$$\begin{aligned}\sum_{t=1}^{T-1}\left\|(K_t')^\top R_t K_t' - K_t^\top R_t K_t\right\| &= \sum_{t=1}^{T-1}\left\|(K_t'-K_t+K_t)^\top R_t(K_t'-K_t+K_t) - K_t^\top R_t K_t\right\| \\ &\le \sum_{t=1}^{T-1}\|K_t'-K_t\|^2\|R_t\| + 2\|K_t\|\|R_t\|\|K_t'-K_t\| \\ &\le 3|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}'-\boldsymbol{K}|||.\end{aligned}$$

The last step holds since $\|K_t'-K_t\| \le \|K_t\|$ by assumption.

Therefore,

$$\begin{aligned}&|C(\boldsymbol{K}')-C(\boldsymbol{K})| \\ \le\ & \mathrm{Tr}(\mathbb{E}[x_0x_0^\top])\Big\{3|||\boldsymbol{K}|||\|R_0\|\,|||\boldsymbol{K}'-\boldsymbol{K}||| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,|||\boldsymbol{K}'-\boldsymbol{K}||| \\ &+\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}'-\boldsymbol{K}|||\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2\,|||\boldsymbol{R}|||\right) \\ &+\left(\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}'-\boldsymbol{K}||| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right)3|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}'-\boldsymbol{K}|||\Big\} \\ &+3\mathrm{Tr}(W)\Big(\frac{(T-1)(\rho^{2(T-1)}-1)}{\rho^2-1}+1\Big)|||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}'-\boldsymbol{K}||| \\ &+\left(T\frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}'-\boldsymbol{K}|||\right)\Big(\mathrm{Tr}\big(\sum_{t=1}^{T-1}Q_t\big)+|||\boldsymbol{K}|||^2\,\mathrm{Tr}\big(\sum_{t=1}^{T-1}R_t\big)\Big) \\ &+\mathrm{Tr}(Q_T)\frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}'-\boldsymbol{K}|||.\end{aligned}$$

By (2.3.27), Lemma 2.3.8, and Lemma 2.3.16, ρ is bounded above by polynomials in $\|A\|$, $\|B\|$, $|||\boldsymbol{R}|||$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$ and $C(\boldsymbol{K})$, or a constant $1+\xi$. Therefore, we rewrite the above inequality by

$$|C(\boldsymbol{K}')-C(\boldsymbol{K})| \le h_{CK}|||\boldsymbol{K}'-\boldsymbol{K}||| + h'_{CK}|||\boldsymbol{K}'-\boldsymbol{K}|||^2, \tag{A.3.31}$$

where $h_{CK} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ and $h'_{CK} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$ are polynomials in $C(\boldsymbol{K})$ and model parameters. Given assumption (2.4.7), we have $|||\boldsymbol{K}'-\boldsymbol{K}||| \le 1$ and hence

$$|||\boldsymbol{K}'-\boldsymbol{K}||| \ge |||\boldsymbol{K}'-\boldsymbol{K}|||^2.$$

Define $h_{cost} = h_{CK} + h'_{CK}$, then (A.3.31) gives

$$|C(\boldsymbol{K}') - C(\boldsymbol{K})| \leq h_{cost}|||\boldsymbol{K}' - \boldsymbol{K}|||,$$

with $h_{cost} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$. □

Proof of Lemma 2.4.8. Recall $\nabla_t C(\boldsymbol{K}) = 2E_t\Sigma_t$ and $W = \mathbb{E}\left[w_t w_t^\top\right] = \widetilde{W}\mathbb{E}\left[v_t v_t^\top\right]\widetilde{W}^\top$. We have,

$$\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| = \|2E'_t\Sigma'_t - 2E_t\Sigma_t\| \leq 2\|E'_t - E_t\|\|\Sigma'_t\| + 2\|E_t\|\|\Sigma'_t - \Sigma_t\|, \tag{A.3.32}$$

For the second term, by Lemma 2.3.6 and Cauchy-Schwarz inequality,

$$\|E_t\| \leq \sum_{t=0}^{T-1}\|E_t\| \leq \sum_{t=0}^{T-1}\sqrt{\mathrm{Tr}(E_t^\top E_t)} \leq \sqrt{T \cdot \frac{\max_t \|R_t + B^\top P_{t+1}B\|}{\underline{\sigma}_{\boldsymbol{X}}}\left(C(\boldsymbol{K}) - C(\boldsymbol{K}^*)\right)}. \tag{A.3.33}$$

By (A.3.7) and direct calculation, we have

$$\|(\mathcal{G}'_{t+1} - \mathcal{G}_{t+1})(\Sigma_0)\| \leq \rho^{2(t+1)}\left(\sum_{s=0}^{t+1}\|\mathcal{F}_{K'_s} - \mathcal{F}_{K_s}\|\,\|\Sigma_0\|\right).$$

By (A.3.9) and (2.3.20), for $t = 1, 2, \cdots, T-1$,

$$\begin{aligned}\|\Sigma'_t - \Sigma_t\| &\leq \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma_0)\| + \left\|\sum_{s=0}^{t-1} D_{t-1,s}WD_{t-1,s}^\top - D'_{t-1,s}W(D'_{t-1,s})^\top\right\| \\ &\leq \rho^{2t}(2\rho+1)\|B\|\,\|\Sigma_0\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| \; + \frac{(\rho^{2T}-1)}{\rho^2-1}(2\rho+1)\|B\|\,\|W\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.\end{aligned} \tag{A.3.34}$$

Therefore the second term in (A.3.32) is bounded by the product of (A.3.33) and (A.3.34).

Next we bound the first term in (A.3.32). Similar to (A.3.14), $\|\Sigma'_t\| \leq \|\sum_{t=0}^{T}\Sigma'_t\| = \|\Sigma_{\boldsymbol{K}'}\| \leq \|\Sigma'_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}}\| + \|\Sigma_{\boldsymbol{K}}\| \leq \frac{C(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}} + \|\Sigma_{\boldsymbol{K}}\|$. For $\|E'_t - E_t\|$, we first need a bound on

$\|P'_t - P_t\|$. Since $P_0 = S_0 + \sum_{t=0}^{T-2} \mathcal{G}_t(S_{t+1}) + \mathcal{G}_{T-1}(Q_T)$, by (A.3.30), we have

$$
\begin{aligned}
\|P'_t - P_t\| &\leq \|P'_0 - P_0\| \\
&\leq 3\|K_0\|\|R_0\|\|K'_0 - K_0\| + \|\mathcal{G}_d\| + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\Big(\sum_{t=0}^{T-1}\|K_t - K'_t\|\Big) \\
&\leq \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||\left(|||\boldsymbol{Q}||| + |||\boldsymbol{K}|||^2|||\boldsymbol{R}|||\right) \\
&\quad + 3\left(1 + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\,|||\boldsymbol{K}' - \boldsymbol{K}||| + \frac{\rho^2(1-\rho^{2(T-1)})}{\rho^2-1}\right)\cdot \\
&\quad |||\boldsymbol{K}|||\,|||\boldsymbol{R}|||\,|||\boldsymbol{K}' - \boldsymbol{K}||| \\
&\quad + \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|B\|\|Q_T\|\,|||\boldsymbol{K}' - \boldsymbol{K}|||.
\end{aligned}
\tag{A.3.35}
$$

Thus,

$$
\begin{aligned}
&\|E'_t - E_t\| \\
&= \left\|R_t(K'_t - K_t) - B^\top(P'_{t+1} - P_{t+1})A + B^\top(P'_{t+1} - P_{t+1})BK'_t + B^\top P_{t+1}B(K'_t - K_t)\right\| \\
&\leq \left(\|R_t\| + \|B\|^2\|P_0\|\right)|||\boldsymbol{K}' - \boldsymbol{K}||| + \|B\|\,\|P'_0 - P_0\|\,\|A\| + 2\|B\|^2\|P'_0 - P_0\|\,|||\boldsymbol{K}|||.
\end{aligned}
$$

Given the bound on $|||\boldsymbol{K}||| = \sum_{t=0}^{T-1}\|K_t\|$ in Lemma 2.3.16 and the bound on $\|P_t\|$ in Lemma 2.3.8, all the terms in (A.3.32) can be bounded by polynomials of related parameters multiplied by $|||\boldsymbol{K}' - \boldsymbol{K}|||$ and $|||\boldsymbol{K}' - \boldsymbol{K}|||^2$. Similarly to the proof of Lemma 2.4.7, we have $|||\boldsymbol{K}' - \boldsymbol{K}||| \leq 1$ and

$$
\|\nabla_t C(\boldsymbol{K}') - \nabla_t C(\boldsymbol{K})\| \leq h_{grad}|||\boldsymbol{K}' - \boldsymbol{K}|||,
$$

for some polynomial $h_{grad} \in \mathcal{H}(\mathrm{C}(\boldsymbol{K}))$.

□

A.3.5 Proofs in Section 2.5

Proof of Proposition 2.5.2. Denote $H_t := \begin{pmatrix} 1+\gamma k_t^1 & \gamma k_t^2 \\ k_t^1 & 1+k_t^2 \end{pmatrix}$. Since H_t has two eigenvalues 1 and $\gamma k_t^1 + k_t^2 + 1$, H_t is positive definite when $\gamma k_t^1 + k_t^2 > -1$ $(0 \le t \le T-1)$.

Then let us show the first claim by induction. Assume $\mathbb{E}[x_s x_s^\top]$ is positive definite for all $s \le t$, then

$$\begin{aligned} & \mathbb{E}[x_{t+1}x_{t+1}^\top] \\ = \; & \mathbb{E}[((A-BK_t)x_t + w_t)\left((A-BK_t)x_t + w_t\right)^\top] = \mathbb{E}[(H_t x_t + w_t)\left(H_t x_t + w_t\right)^\top] \\ = \; & \mathbb{E}[H_t x_t x_t^\top H_t^\top + w_t w_t^\top + w_t w_t^\top + 2H_t x_t w_t^\top] = H_t \mathbb{E}[x_t x_t^\top] H_t^\top + \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Hence $\mathbb{E}[x_{t+1}x_{t+1}^\top]$ is positive definite since $\mathbb{E}[x_t x_t^\top]$ is positive definite and H_t is positive definite. Therefore $\underline{\sigma}_{\boldsymbol{X}} > 0$.

The second claim can be proved by backward induction. For $t = T$, $P_T^{\boldsymbol{K}} = Q_T$ is positive definite since Q_T is positive definite. Assume $P_{t+1}^{\boldsymbol{K}}$ is positive definite for some $t+1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P_t^K z = z^\top Q_t\, z + z^\top K_t^\top R_t K_t z + z^\top H_t^\top P_{t+1}^{\boldsymbol{K}} H_t z > 0.$$

Note that H_t is positive definite when $\gamma k_t^1 + k_t^2 > -1$ and $1 + \gamma k_t^1 > 0$. The last inequality holds since Q_t and $H_t^\top P_{t+1}^{\boldsymbol{K}} H_t$ are positive definite, and $K_t^\top R_t K_t$ is positive semi-definite. Hence we have $P_t^{\boldsymbol{K}}$ positive definite for all $t = 0, 1, 2, \cdots, T$. □

Appendix B

Appendix for Chapter 3

We now give the proofs that were omitted in Chapter 3.

B.1 Proofs in Section 3.3.1

Proof of Lemma 3.3.6. We prove that the terms in the sequence $\{P_{t,i}^{\boldsymbol{K}}\}_{t=0}^{T}$ are positive definite for $i = 1, \cdots, N$ by backward induction. For $t = T$, $P_{T,i}^{\boldsymbol{K}} = Q_T^i$ is positive definite since Q_T^i is positive definite. Assume $P_{t+1,i}^{\boldsymbol{K}}$ is positive definite for some $t+1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P_{t,i}^{\boldsymbol{K}} z = z^\top Q_t^i z + z^\top (K_t^i)^\top R_t^i K_t^i z + z^\top \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right) z$$

$$> 0.$$

The last inequality holds since $z^\top Q_t^i z > 0$ and the other terms are non-negative under Assumption 3.2.1. By backward induction, we have $P_{t,i}^{\boldsymbol{K}}$ positive definite, $\forall\, t = 0, 1, \cdots, T$. □

To prove Lemma 3.3.8, let us start with a useful result for the value function. We define the value function for player i, $V_{\boldsymbol{K}}^i(x,t)$ for $t = 0, 1, \cdots, T-1$, as

$$V_{\boldsymbol{K}}^i(x,t) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{s=t}^{T-1} (x_s^\top Q_s^i x_s + (u_s^i)^\top R_s^i u_s^i) + x_T^\top Q_T^i x_T \,\middle|\, x_t = x\right] = x^\top P_{t,i}^{\boldsymbol{K}} x + N_{t,i}^{\boldsymbol{K}},$$

with terminal condition

$$V^i_{\boldsymbol{K}}(x,T) = x^\top Q^i_T x,$$

where $N^{\boldsymbol{K}}_{t,i}$ is defined in (3.3.14), and $\mathbb{E}_{\boldsymbol{w}}$ denotes the expectation over the noise $\boldsymbol{w}$. We then define the Q function, $Q^i_{\boldsymbol{K}}(x,u,t)$ for $u = (u^1, \cdots, u^N)$ as

$$Q^i_{\boldsymbol{K}}(x,u,t) = x^\top Q^i_t x + (u^i)^\top R^i_t u^i + \mathbb{E}_{w_t}\left[V^i_{\boldsymbol{K}}(A_t x + \sum_{j=1}^N B^j_t u^j + w_t, t+1)\right],$$

and the advantage function

$$A^i_{\boldsymbol{K}}(x,u,t) = Q^i_{\boldsymbol{K}}(x,u,t) - V^i_{\boldsymbol{K}}(x,t).$$

Here $\mathbb{E}_{w_t}$ denotes the expectation taken with respect to w_t. Note that $C^i(\boldsymbol{K}) = \mathbb{E}[V^i_{\boldsymbol{K}}(x_0,0)]$. We write

$$(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}) := (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^{i-1}, \boldsymbol{K}^{i\prime}, \boldsymbol{K}^{i+1}, \cdots, \boldsymbol{K}^N),$$

and the control sequences under the policy $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ as $\{u^{i\prime,-i}_t\}_{t=0}^{T-1}$ with

$$u^{i\prime}_t = -K^{i\prime}_t x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t, \quad \text{and} \quad u^j_t = -K^j_t x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t, \quad j \neq i. \tag{B.1.1}$$

Then we can write the difference between the cost functions of $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$ and $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ in terms of advantage functions.

Lemma B.1.1 (Cost Difference)**.** *Assume $\boldsymbol{K}$ and $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ have finite costs. Then*

$$V^i_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(x,0) - V^i_{\boldsymbol{K}}(x,0) = \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}}(x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t, u^{i\prime,-i}_t, t)\right], \tag{B.1.2}$$

where $\{u^{i\prime,-i}_t\}_{t=0}^{T-1}$ is defined in (B.1.1) *with $x_0^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} = x_0 = x$. For player i,*

$$\begin{aligned} A^i_{\boldsymbol{K}}(x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t, u^{i\prime,-i}_t, t) \;=\; & (x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t)^\top (K^{i\prime}_t - K^i_t)^\top (R^i_t + (B^i_t)^\top P^{\boldsymbol{K}}_{t+1,i} B^i_t)(K^{i\prime}_t - K^i_t)\, x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t \\ & + 2(x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t)^\top (K^{i\prime}_t - K^i_t)^\top E^{\boldsymbol{K}}_{t,i}\, x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t. \end{aligned}$$

Proof. Write $x'_t = x^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}_t$ for simplicity and denote by $c^{i\prime}_t(x)$ the cost of player i

generated by $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ with a single trajectory starting from $x'_0 = x_0 = x$. That is,

$$c_t^{i\prime}(x'_t) = (x'_t)^\top Q_t^i x'_t + (K_t^{i\prime} x_t^{i\prime})^\top R_t^i K_t^{i\prime} x_t^{i\prime}, \quad t = 0, 1, \cdots, T-1,$$

and

$$c_T^{i\prime}(x'_T) = (x'_T)^\top Q_T^i x'_T,$$

with

$$u_t^{i\prime} = -K_t^{i\prime} x'_t \quad , x'_{t+1} = A_t x'_t + B_t^i u_t^{i\prime} + \sum_{j=1, j\neq i}^{N} B_t^j u_t^j + w_t, \quad x'_0 = x.$$

Therefore

$$\begin{aligned}
V^i_{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}}(x,0) - V^i_{\boldsymbol{K}}(x,0) &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} c_t^{i\prime}(x'_t)\right] - V^i_{\boldsymbol{K}}(x,0) \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T} \left(c_t^{i\prime}(x'_t) - V^i_{\boldsymbol{K}}(x'_t, t)\right) + \sum_{t=1}^{T} V^i_{\boldsymbol{K}}(x'_t, t)\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(c_t^{i\prime}(x'_t) + V^i_{\boldsymbol{K}}(x'_{t+1}, t+1) - V^i_{\boldsymbol{K}}(x'_t, t)\right)\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(Q^i_{\boldsymbol{K}}(x'_t, u_t^{i\prime,-i}, t) - V^i_{\boldsymbol{K}}(x'_t, t)\right) \middle| x_0 = x\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}}(x'_t, u_t^{i\prime,-i}, t) \middle| x_0 = x\right],
\end{aligned}$$

where the third equality holds since $c_T^{i\prime}(x'_T) = V^i_{\boldsymbol{K}}(x'_T, T)$ with the same single trajectory.

For player i, the advantage function is given by

$$\begin{aligned}
& A^i_{\boldsymbol{K}}(x'_t, u_t^{i\prime,-i}, t) = Q^i_{\boldsymbol{K}}(x'_t, u_t^{i\prime,-i}, t) - V^i_{\boldsymbol{K}}(x'_t, t) \\
= \;& (x'_t)^\top (Q_t^i + (K_t^{i\prime})^\top R_t^i K_t^{i\prime}) x'_t + \mathbb{E}_{w_t}\Big[V^i_{\boldsymbol{K}}((A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j) x'_t + w_t, t+1)\Big] \\
& - V^i_{\boldsymbol{K}}(x'_t, t) \\
= \;& (x'_t)^\top (Q_t^i + (K_t^{i\prime})^\top R_t^i K_t^{i\prime}) x'_t + \Big((x'_t)^\top \Big(A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j\Big)^\top \\
& P^{\boldsymbol{K}}_{t+1,i}\Big(A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j\Big) x'_t + \mathrm{Tr}(W P^{\boldsymbol{K}}_{t+1,i}) + N^{\boldsymbol{K}}_{t+1,i}\Big) - \Big((x'_t)^\top P^{\boldsymbol{K}}_{t,i} x'_t + N^{\boldsymbol{K}}_{t,i}\Big)
\end{aligned}$$

$$
\begin{aligned}
&= (x_t')^\top (Q_t^i + (K_t^{i\prime} - K_t^i + K_t^i)^\top R_t^i (K_t^{i\prime} - K_t^i + K_t^i)) x_t' \\
&\quad + (x_t')^\top \Big(A_t - B_t^i (K_t^{i\prime} - K_t^i) - \sum_{j=1}^N B_t^j K_t^j \Big)^\top P_{t+1,i}^{\boldsymbol{K}} \Big(A_t - B_t^i (K_t^{i\prime} - K_t^i) - \sum_{j=1}^N B_t^j K_t^j \Big) x_t' \\
&\quad - (x_t')^\top \Big(Q_t^i + (K_t^i)^\top R_t^i K_t^i + \Big(A_t - \sum_{j=1}^N B_t^j K_t^j \Big)^\top P_{t+1,i}^{\boldsymbol{K}} \Big(A_t - \sum_{j=1}^N B_t^j K_t^j \Big) \Big) x_t' \\
&= (x_t')^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i) x_t' + 2 (x_t')^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}} x_t'.
\end{aligned}
$$

□

Now we are ready to prove Lemma 3.3.8.

Proof of Lemma 3.3.8. By Lemma B.1.1, we have

$$
\begin{aligned}
& A_{\boldsymbol{K}}^i (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}}, u_t^{i\prime, -i}, t) \\
= \; & (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i) x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} \\
& + 2 (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}} x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} \\
= \; & \operatorname{Tr}\big(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i) \big) \\
& + 2 \operatorname{Tr}\big(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}} \big) \\
= \; & \operatorname{Tr}\big(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i + (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}})^\top \cdot \\
& (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i + (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}) \big) \\
& - \operatorname{Tr}\big(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}} \big) \\
\geq \; & - \operatorname{Tr}\big(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}} \big),
\end{aligned}
$$

with equality holding when $K_t^{i\prime} = K_t^i - (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}$. Then, by Lemma B.1.1, letting $\{x_t^*\}_{t=0}^T$ and $u_t^* = (u_t^{1*}, \cdots, u_t^{N*})$ with $u_t^{i*} = -K_t^{i*} x_t^*$ denote the state and control sequences induced by the set of optimal policies $\boldsymbol{K}^*$, we have

$$
\begin{aligned}
& C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) = -\mathbb{E}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}^i (x_t^*, u_t^*, t) \right] \\
\leq \; & \mathbb{E}\left[\sum_{t=0}^{T-1} \operatorname{Tr}\left(x_t^* (x_t^*)^\top (E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right) \right] \\
\leq \; & \|\Sigma_{\boldsymbol{K}^*}\| \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right) \\
\leq \; & \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((\Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^{-1}\nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})(\Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^{-1}\right)\\
&\leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{R}}\,(\underline{\sigma}_{\boldsymbol{X}})^2} \sum_{t=0}^{T-1} \operatorname{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\right).
\end{aligned}
$$

For the lower bound, consider $K_t^{i\prime} = K_t^i - (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}$. Using $C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}) \geq C^i(\boldsymbol{K}^*)$ and letting $\{u_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\}_{t=0}^{T-1}$ denote the control sequence induced by $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*})$, we have

$$
\begin{aligned}
& C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\\
\geq\ & C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*})\\
=\ & -\mathbb{E}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}, u_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}, t)\right]\\
=\ & \mathbb{E}\left[\sum_{t=0}^{T-1} \operatorname{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\right]\\
\geq\ & \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\\
\geq\ & \frac{\underline{\sigma}_{\boldsymbol{X}}}{4\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2} \sum_{t=0}^{T-1} \frac{\operatorname{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\right)}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|}.
\end{aligned}
$$

□

Proof of Lemma 3.3.9. By Lemma B.1.1,

$$
\begin{aligned}
& C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}) - C^i(\boldsymbol{K})\\
=\ & \mathbb{E}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}}(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, u_t^{i\prime,-i}, t)\right]\\
=\ & \sum_{t=0}^{T-1}\Big[\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i)\right)\\
& +2\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\right)\Big].
\end{aligned}
$$

□

Proof of Lemma 3.3.10. It is straightforward to see that

$$C^i(\boldsymbol{K}) \geq \mathbb{E}[x_t^\top P_{t,i}^{\boldsymbol{K}} x_t] \geq \left\|P_{t,i}^{\boldsymbol{K}}\right\| \sigma_{\min}(\mathbb{E}[x_t x_t^\top]) \geq \underline{\sigma}_{\boldsymbol{X}} \left\|P_{t,i}^{\boldsymbol{K}}\right\|$$

and

$$\begin{aligned} C^i(\boldsymbol{K}) &= \sum_{t=0}^{T-1} \mathrm{Tr}\left(\mathbb{E}[x_t x_t^\top](Q_t^i + (K_t^i)^\top R_t^i K_t^i)\right) + \mathrm{Tr}\left(\mathbb{E}[x_T x_T^\top] Q_T^i\right) \\ &\geq \min_{t\in[0,T]} \sigma_{\min}(Q_t^i) \cdot \mathrm{Tr}(\Sigma_{\boldsymbol{K}}) \\ &\geq \underline{\sigma}_{\boldsymbol{Q}} \cdot \|\Sigma_{\boldsymbol{K}}\|. \end{aligned}$$

Then the statements in Lemma 3.3.10 follow since under Assumptions 3.2.1 and 3.2.2, we have $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $\underline{\sigma}_{\boldsymbol{Q}} > 0$. □

B.2 Proofs in Section 3.3.2

Proof of Proposition 3.3.11. Recall that $\Sigma_t^{\boldsymbol{K}} = \mathbb{E}\left[x_t x_t^\top\right]$ and note that

$$\begin{aligned} \Sigma_1^{\boldsymbol{K}} &= \mathbb{E}[x_1 x_1^\top] = \mathbb{E}\Big[\Big((A_0 - \sum_{j=1}^N B_0^j K_0^j)x_0 + w_0\Big)\Big((A_0 - \sum_{j=1}^N B_0^j K_0^j)x_0 + w_0\Big)^\top\Big] \\ &= \Big(A_0 - \sum_{j=1}^N B_0^j K_0^j\Big)\Sigma_0\Big(A_0 - \sum_{j=1}^N B_0^j K_0^j\Big)^\top + W = \mathcal{G}_0(\Sigma_0) + W. \end{aligned}$$

We first prove that for $t = 2, 3, \cdots, T$

$$\Sigma_t^{\boldsymbol{K}} = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top + W. \tag{B.2.1}$$

We have the result for $t = 1$, so assume (B.2.1) holds for $t \leq k$. Then for $t = k+1$,

$$\begin{aligned} \mathbb{E}[x_{t+1} x_{t+1}^\top] &= \mathbb{E}\Big[\Big((A_t - \sum_{j=1}^N B_t^j K_t^j)x_t + w_t\Big)\Big((A_t - \sum_{j=1}^N B_t^j K_t^j)x_t + w_t\Big)^\top\Big] \\ &= \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)\Sigma_t^{\boldsymbol{K}}\Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)^\top + W \end{aligned}$$

$$= \mathcal{G}_t(\Sigma_0) + \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + W.$$

Therefore (B.2.1) holds, $\forall\, t = 1, 2, \cdots, T$. Finally,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T} \Sigma_t^{\boldsymbol{K}} = \Sigma_0 + \sum_{t=0}^{T-1} \mathcal{G}_t(\Sigma_0) + \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + TW = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta_{\boldsymbol{K}}(W).$$

□

Proof of Lemma 3.3.14. For player i,

$$\begin{aligned} & (\mathcal{F}_t - \mathcal{F}_t^{i'})(X) \\ = \; & \Big(A_t - \sum_{j=1}^{N} B_t^j K_t^j\Big) X \Big(A_t - \sum_{j=1}^{N} B_t^j K_t^j\Big)^\top \\ & - \Big(A_t - B_t^i K_t^{i'} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j\Big) X \Big(A_t - B_t^i K_t^{i'} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j\Big)^\top, \end{aligned}$$

Then (3.3.28) follows by letting $Y = A_t - \sum_{j=1}^{N} B_t^j K_t^j$ and $Z = A_t - B_t^i K_t^{i'} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j$ in

$$YXY^\top - ZXZ^\top = \frac{(Y+Z)X(Y-Z)^\top + (Y-Z)X(Y+Z)^\top}{2} \tag{B.2.2}$$

and using the norm bound $\|AX\| \leq \|A\|\,\|X\|$. □

Proof of Lemma 3.3.15. By direct calculation,

$$\|\mathcal{G}_t^{i'}\| \leq \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2(t+1)}, \quad i = 1, \cdots, N. \tag{B.2.3}$$

Then for any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $t \geq 0$,

$$\begin{aligned} & \|(\mathcal{G}_{t+1}^{i'} - \mathcal{G}_{t+1})(\Sigma)\| \\ = \; & \|\mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t^{i'}(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\ = \; & \|\mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t^{i'}(\Sigma) - \mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t(\Sigma) + \mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\ \leq \; & \|\mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t^{i'}(\Sigma) - \mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t(\Sigma)\| + \|\mathcal{F}_{t+1}^{i'} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\ = \; & \|\mathcal{F}_{t+1}^{i'} \circ (\mathcal{G}_t^{i'} - \mathcal{G}_t)(\Sigma)\| + \|(\mathcal{F}_{t+1}^{i'} - \mathcal{F}_{t+1}) \circ \mathcal{G}_t(\Sigma)\| \\ \leq \; & \|\mathcal{F}_{t+1}^{i'}\|\,\|(\mathcal{G}_t^{i'} - \mathcal{G}_t)(\Sigma)\| + \|\mathcal{G}_t\|\,\|\mathcal{F}_{t+1}^{i'} - \mathcal{F}_{t+1}\|\,\|\Sigma\| \end{aligned}$$

$$\leq \quad \rho^2_{\boldsymbol{K},\boldsymbol{K}'}\|(\mathcal{G}^{i\prime}_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)}_{\boldsymbol{K},\boldsymbol{K}'}\|\mathcal{F}^{i\prime}_{t+1} - \mathcal{F}_{t+1}\|\|\Sigma\|.$$

Therefore,

$$\|(\mathcal{G}^{i\prime}_{t+1} - \mathcal{G}_{t+1})(\Sigma)\| \leq \rho^2_{\boldsymbol{K},\boldsymbol{K}'}\|(\mathcal{G}^{i\prime}_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)}_{\boldsymbol{K},\boldsymbol{K}'}\|\mathcal{F}^{i\prime}_{t+1} - \mathcal{F}_{t+1}\|\|\Sigma\|. \quad \text{(B.2.4)}$$

Summing (B.2.4) up for $t \in \{1, 2, \cdots, T-2\}$ with $\|\mathcal{G}^{i\prime}_0 - \mathcal{G}_0\| = \|\mathcal{F}^{i\prime}_0 - \mathcal{F}_0\|$, we have

$$\sum_{t=0}^{T-1}\Big\|(\mathcal{G}_t - \mathcal{G}^{i\prime}_t)(\Sigma)\Big\| \leq \frac{\rho^{2T}_{\boldsymbol{K},\boldsymbol{K}'} - 1}{\rho^2_{\boldsymbol{K},\boldsymbol{K}'} - 1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_t - \mathcal{F}^{i\prime}_t\|\Big)\|\Sigma\|.$$

□

Bibliography

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[2] Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[3] Marc Abeille, Alessandro Lazaric, Xavier Brokmann, et al. LQG for portfolio optimization. *Available at SSRN 2863925*, 2016.

[4] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20, 2015.

[5] Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, 2010.

[6] Robert Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003.

[7] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40, 2001.

[8] Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18(7):58–62, 2005.

[9] Brian D. O. Anderson and John B Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, 2007.

[10] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[11] J Andrew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1019–1024, August 2003.

[12] Wenhang Bao and Xiao-yang Liu. Multi-agent deep reinforcement learning for liquidation strategy analysis. *arXiv preprint arXiv:1906.11046*, 2019.

[13] Tamer Başar and Geert Jan Olsder. *Dynamic Non-Cooperative Game Theory*. Society for Industrial and Applied Mathematics, 1998.

[14] Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Available at SSRN 3848428*, 2021.

[15] Dimitri Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 3rd edition, 2005.

[16] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[17] Ronen I Brafman and Moshe Tennenholtz. R-MAX-A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[18] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[19] Jingjing Bu, Afshin Mesbahi, and Mehran Mesbahi. Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint arXiv:2006.09178*, 2020.

[20] Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

[21] René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

[22] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[23] Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38, 2021.

[24] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2014.

[25] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

[26] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(1), 2003.

[27] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.

[28] Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Efficient learning of distributed linear-quadratic control policies. *SIAM Journal on Control and Optimization*, 58(5):2927–2951, 2020.

[29] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[30] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.

[31] Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 72–80, 1997.

[32] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Society for Industrial and Applied Mathematics*, SODA '05, pages 385–394, USA, 2005.

[33] Sumitra Ganesh, Nelson Vadori, Mengda Xu, Hua Zheng, Prashant Reddy, and Manuela Veloso. Reinforcement learning for market making in a multi-agent dealer market. *arXiv preprint arXiv:1911.05892*, 2019.

[34] Jim Gatheral and Alexander Schied. Optimal trade execution under geometric Brownian motion in the Almgren and Chriss framework. *International Journal of Theoretical and Applied Finance*, 14(03):353–368, 2011.

[35] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning robust controllers for linear quadratic systems with multiplicative noise via policy gradient. *arXiv preprint arXiv:1905.13547*, 2019.

[36] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[37] Florent Guenter, Micha Hersch, Sylvain Calinon, and Aude Billard. Reinforcement learning for imitating constrained reaching movements. *Advanced Robotics*, 21(13):1521–1544, 2007.

[38] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. *arXiv preprint arXiv:2010.00145*, 2020.

[39] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods find the Nash equilibrium in N-player general-sum linear-quadratic games. *arXiv preprint arXiv:2107.13090*, 2021.

[40] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[41] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *arXiv preprint arXiv:2112.04553*, 2021.

[42] Dieter Hendricks and Diane Wilcox. A reinforcement learning extension to the Almgren-Chriss framework for optimal trade execution. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 457–464. IEEE, 2014.

[43] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[44] Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

[45] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2193–2201, 2018.

[46] Zeyu Jin, Johann Michael Schmitt, and Zaiwen Wen. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

[47] Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

[48] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002.

[49] Petar Kormushev, Sylvain Calinon, and Darwin G Caldwell. Reinforcement learning in robotics: Applications and real-world challenges. *Robotics*, 2(3):122–148, 2013.

[50] Panqanamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[51] Laura Leal, Mathieu Laurière, and Charles-Albert Lehalle. Learning a functional control for high-frequency finance. *arXiv preprint arXiv:2006.09611*, 2020.

[52] Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO*, pages 222–229, 2004.

[53] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[54] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*

2019, volume 57, pages 6309–6317. AAAI Press (Association for the Advancement of Artificial Intelligence), 2019.

[55] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR, 2019.

[56] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

[57] Eric Mazumdar, Lillian J Ratliff, Michael I Jordan, and S Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *AAMAS Conference proceedings*, 2020.

[58] Eric Mazumdar, Lillian J Ratliff, and S Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

[59] Muddasar Naeem, Syed Tahir Hussain Rizvi, and Antonio Coronato. A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*, 8:209320–209344, 2020.

[60] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[61] Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680, 2006.

[62] Brian Ning, Franco Ho Ting Ling, and Sebastian Jaimungal. Double deep Q-learning for optimal execution. *arXiv preprint arXiv:1812.06600*, 2018.

[63] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[64] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

[65] Meir Pachter. LQG dynamic games with a control-sharing information pattern. *Dynamic Games and Applications*, 7(2):289–322, 2017.

[66] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4026–4035. PMLR, 2018.

[67] Panagiotis Patrinos, Sergio Trimboli, and Alberto Bemporad. Stochastic MPC for real-time market-based optimal power dispatch. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 7111–7116. IEEE, 2011.

[68] Andre F Perold. The implementation shortfall: Paper versus reality. *Journal of Portfolio Management*, 14(3):4, 1988.

[69] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

[70] Martin L. Puterman. Chapter 8 markov decision processes. In *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, pages 331–434. Elsevier, 1990.

[71] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[72] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, 2019.

[73] Masoud Roudneshin, Jalal Arabneydi, and Amir G Aghdam. Reinforcement learning in nonzero-sum linear quadratic deep structured games: Global convergence of policy optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 512–517. IEEE, 2020.

[74] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

[75] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

[76] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[77] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[78] Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. Market making via reinforcement learning. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems*, AAMAS '18, pages 434–442, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[79] Robert F Stengel. *Optimal control and estimation*. Courier Corporation, 1994.

[80] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[81] Istvan Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, pages 1031–1038, 2010.

[82] Emanuel Todorov and Weiwei Li. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

[83] Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014, 2018.

[84] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083, 2019.

[85] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[86] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21:198–1, 2020.

[87] Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time. In *International Conference on Machine Learning*, pages 10772–10782. PMLR, 2021.

[88] William Yang Wang, Jiwei Li, and Xiaodong He. Deep reinforcement learning for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–21, 2018.

[89] Yasuaki Wasa, Kengo Sakata, Kenji Hirata, and Kenko Uchida. Differential game-based load frequency control for power networks and its integration with electricity market mechanisms. In *2017 IEEE Conference on Control Technology and Applications (CCTA)*, pages 1044–1049. IEEE, 2017.

[90] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

[91] Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: a case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019.

[92] Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10887–10895, 2021.

[93] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.

[94] Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34, 2021.

[95] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *The Journal of Financial Data Science*, 2(2):25–40, 2020.

[96] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. *Advances in Neural Information Processing Systems*, 32:8668–8678, 2019.