

# Decision Making with Unknown Future Costs

**Yitian Chen**

A thesis submitted for the degree of

*Doctor of Philosophy*

The Australian National University

CIICADA Lab, Research School of Engineering



**Australian  
National  
University**

August 2025

© Copyright by Yitian Chen, 2025

All Rights Reserved

*To my parents, friends, and the wonderful universe we are intimately connected with*

## Declaration

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at the ANU or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at the ANU or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

*“My only enemy is time.”*

— Charlie Chaplin, 1972 Oscar Introduction Speech

*“Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning.”*

— Albert Einstein

*“Remain faithful to your first resolve, for only then can you reach the end; beginnings are simple, but enduring to the finish is rare.”*

— Buddhāvataṃsaka Sūtra

## Acknowledgements

Fifteen years ago, during a casual conversation with my dad, I asked him whether it was possible to build a time machine to travel to the past or the future. He answered, according to Einstein's theory of relativity, time travel would require moving faster than the speed of light. This is physically impossible. His answers, however, did not convince me. I did not see any connections between time travel and the speed of an object, nor why nothing could surpass a certain speed. He had no further explanation, but that question sparked my curiosity about whether a time machine could exist.

To prove the existence of a time machine design. As a first step, I decided to learn Einstein's special relativity for potential flaws in its assumptions or derivations. Although the chances of a middle school student disproving relativity were little (essentially zero), I considered failure acceptable, but not trying, on the other hand, was simply intolerable.

We all know how the story ended: I never discovered a way to prove the existence of a time machine design. However, the attempt to challenge relativity ignited my passion for research. To attempt the impossible task of disproving relativity, I had no choice but first master the mathematical foundation that underpinned it (calculus was used in the derivations of the book I had chosen). Before my school even introduced the concept of function, I struggled to digest the concepts from calculus, such as the  $\varepsilon$ - $\delta$  definition of limits and all its associated exercises, as well as with techniques for solving indefinite integrals (as a result, I gave up doing my homework and stopped listening in the class, sorry to all the teachers who got mad at me). Every step was difficult, yet those steps gave me a glimpse of the intellectual masterpieces that decode the world to uncover the fundamental principles of the universe, with nothing more than a pen and a piece of paper.

Time flies. What began as a simple curiosity, like my early fascination with the theory of special relativity in middle school has carried me all the way to completing my PhD. Along this journey, I have been fortunate to receive support, guidance, and companionship from many people, to whom I now wish to express my deepest gratitude.

I am profoundly thankful to my principal supervisor Iman Shames, for his invaluable guidance during my PhD studies. I have always been amazed by his research ideas, and I truly appreciate the inspiring research topic he introduced me to. He has consistently and patiently guided me whenever I was stuck on technical problems, showing me the right way to accurately express technical results and examples. I am

especially thankful for his advice and feedback on my ideas in the neural network problem formulation. Without his extraordinary conscientious mentoring and continuous inspiration, it would not have been possible for me to complete any of my research work. I am indebted to the insights and intuitions he shared in control and optimisation, which enabled me to view problems from different perspectives. I am also obliged for his help in arranging my research visit to Melbourne, which broadened my horizons and allowed me to meet wonderful people at the university.

I am indebted to my other supervisor Timothy Liam Molloy, who has been like my old brother that always willing to spend far more time than required discussing matters ranging from writing and research ideas to career development in academia. His advice and experiences have been a constant source of inspiration. From him, I have not only learned how to tackle technical problems and express research ideas, but also how to face and endure challenges with resilience. Many moments from our discussions inspired discoveries of my own. His support and motivation are truly difficult to overstate.

To Iman and Tim, I am also grateful for all the essays, novels, books, movies and comedies that you recommended. Although I have only managed to finish some of them so far, I will continue reading and watching to improve my sense of writing and to deepen my cultural understanding.

I am sincerely thankful to Philipp Braun for sharing his expertise in nonlinear controls, which broadened the scope of my work beyond linear systems. I also deeply appreciated his help in proofreading my recent journal paper and providing insightful feedback that significantly strengthened the quality of my mathematical proofs.

I would also like to thank Erik Verriest, my control systems lecturer during my exchange at the Georgia Institute of Technology. His courses combined rigorous mathematical derivations with real-life examples, showing how problems can be modelled and solved through mathematical control theory. This was the first time I gained such a clear picture of how control theory is constructed on rigorous mathematical foundations and built upon first principles. I will also never forget our discussions on the fact that the inverse Fourier transform of  $e^{-j\omega}$  is undefined in its usual sense, and under what conditions it could be made well-defined. My interactions with Erik laid me the foundation for my pursuit of an academic career.

Throughout my graduate studies, the faculty members of the CIICADA lab, especially Jochen Trumpf, Igor Vladimirov, Daoyi Dong, Robert Mahony, Ian Petersen, Brian Anderson, as well as the faculty members at the University of Melbourne, particularly Michael Cantoni, Peter Dower and Farhad Farokhi, provided me with

valuable advice and encouragement. Their generous help and feedback guided me through many challenges, especially when extending my results during my second year.

Among the members of CIICADAs, I was blessed to have the support of Yuen Man Pun throughout my PhD study. I am truly thankful for her encouragement and for sharing her perspective on optimisation problems. Every discussion we had gave me further insights into optimisation algorithms. I felt sorry for the amount of time she dedicated to patiently listening to my boundless ideas in neural network research and my random thoughts on other optimisation problems.

The generous technical and mental support of students from the CIICADA lab formed a very important component of my PhD journey. I thank Kanghong for dedicating his time to introducing me to control conferences and journals, patiently explaining their distinctions to me when I was a freshman in PhD. His impressive badminton skills also reminded me how much I was lacking in regular exercise whenever we played together. I thank Alex for sharing her reflections on her research experiences, which helped me understand different research styles in terms of problems and presentations. I am grateful to Matthew for his generous help with my presentations and proofs (and importantly, for sharing many jokes along the way). I appreciate Tian for his technical support in our F1TENTH project and our online control collaboration. His rich knowledge of cars has also helped me to diagnose my car when it failed to start when I was desperately seeking help, which even helped me diagnose my own car troubles when I was desperate for assistance. I am thankful to Weichao for generously sharing his experiences in academia, which gave me a clearer understanding of the interactions, joys, and challenges I may face in my potential future academic career. He also introduced me to bouldering, which became a valuable way for me to relieve pressure while doing my favourite kind of problem-solving through exercise. I feel fortunate to have made friends with Huawei, Weiting, Angela, Frank, Henry, June, Lachlan, Xingyu and Hanwen from the Audio and Acoustic Signal Processing group. The Friday dinner gathering with you all made me happy on Fridays during my graduate study. Our Friday dinner gatherings made my graduate study years in Canberra so much happier, to the point that I no longer thought of Canberra as a boring place (though Canberra might still think I am boring too). I would also like to extend my gratitude to other CIICADA members, including Amir Ali, Olivia, Yixiao, Angus, Shouliang, and Yun Yan, for their support and friendship. Special thanks also go to Haorui (my former flatmate) and Yijun (my former colleague at ANU) for their unwavering support and companionship throughout my journey.

Outside ANU, I was fortunate to have Ken Cai as a close friend during my first year in Australia. During times of immense pressure and frustration, he always did his utmost to pull me away from negative feelings and give me confidence in facing challenges. I would not have made it through the past ten years and their turning points without his help. Our countless discussions, ranging from self-exploration and awareness to logic, mathematics, computer science, and game design have given me the mental strength to keep going. During a game design project, I was also grateful to meet David, an amateur music composer. We spent countless hours discussing the nature of the world and the universe. His insights into religion, history, politics, and humanity enriched my perspective. Together, David and Ken made Sydney feel like a true home for me. But David, please drive carefully.

I thank Tom for constantly being my pool practice buddy, and Celia, Shaun, and Laura from my Saturday Avalon board game group for bringing me joy, especially when I was Mogana but pretending to be Percival. I am also grateful to Tsai Yuan for helping me improve my bouldering skills and for introducing me to many enjoyable karaoke and drinking experiences. I truly appreciate the meals made by Betty that were far more delicious than anything I could find in restaurants in Canberra.

I would also like to thank my girlfriend, who I prefer not to be named here, partly to avoid the so-called “breakup curse” as often joked about on social media. Our shared enjoyment of food and the countless conversations around it have brought me comfort and joy beyond my academic life. I am especially grateful for the way she has always embraced me at just the right moments when I was feeling down, and for the thoughtful care she gave me whenever I was unwell. Our eclectic talks, ranging from the mysteries of the universe to how the stock market works, never failed to “shed light” on my world, even if that light usually came in the middle of the night. Her love and support have been an irreplaceable part of my journey.

I would like to thank myself, as it was my own curiosity, persistence and endurance that carried me this far. My experiences in analysing, decomposing and solving countless complicated problems have shaped me into a capable problem solver, and my perseverance has kept me through the difficult moments along the way. Of course, I could not have come this far without all the support of the people I mentioned earlier, and above all, my family.

I am especially grateful to my grandmother, Xiufeng, who patiently taught me to read and write Chinese characters when I was a child. Her companionship filled my early years with warmth and happiness. Another source of joy came from my grandparents, Jinxing and Yuying, on my father’s side, who patiently helped me improve my Chinese chess skills, while sometimes accepting the rules I created,

which turned each game into moments of shared laughter. My uncle, Shuhua, was like a counsellor during my middle and high school years; his advice always healed and encouraged me, guiding me towards the right attitude in life. My father, Hao, is the true protagonist of the story I mentioned at the beginning. I deeply appreciate the inspiration he gave me. My earliest curiosity about the world was awakened by him, and I will never forget our weekends spent catching insects together. His passion for showing me the diversity of insect species deeply influenced my motivation to explore and understand the world. My mother, Hongzhan, has been my role model from the moment I was born. Every encouraging word from her has stood as a spiritual pillar behind me, reminding me that, in her eyes, I am and will always be the greatest person in the world. She has always been the one to support me whenever I felt exhausted by challenges, and I cannot imagine what my life would be like without her as my mother.

# Abstract

This thesis develops a unified framework for decision-making problems with unknown future costs, providing both theoretical guarantees and empirical evaluations of its performance. We begin by studying the online Linear Quadratic (LQ) optimal control problem for the cases where (i) future costs are unknown beyond a certain preview horizon and sequentially revealed over time; and (ii) costs are unknown and must be inferred from observed optimal trajectory data. We then extend the framework to dynamic LQ games with sequentially revealed (and potentially previewed) costs. In all settings, the proposed framework is based on predicting and tracking a candidate optimal trajectory using the available costs.

We begin by applying the proposed framework to the online LQ optimal control problem with sequentially revealed cost. We adopt the notion of regret as the decision quality measurement. We show that the regret of the proposed method is upper bounded by terms that decay exponentially fast as the preview horizon of future costs increases. Simulations verify this exponential decay and demonstrate that our controller outperforms state-of-the-art methods that do not leverage cost feedback.

We then consider the case where the costs must be inferred from observed optimal trajectory data. This is a new framework for solving the learning from demonstration problem. We establish a theoretical connection between the regret and the estimation error of the estimated optimal control gain. A regret bound is derived under an Extended Kalman Filter(EKF)-based parameter estimation scheme, and its performance is validated through numerical experiments.

We then apply this framework to a new dynamic LQ game problem, where the costs are sequentially revealed to the players (and may be previewed). We introduce the notion of *price of uncertainty* (PoU) that generalises the notion of regret to multi-agent settings. We establish bounds on the PoU incurred when all players are adopting the designed controller using our framework. Simulation results validate the theoretical bounds on PoU.

# List of Publications

## Conference Papers:

- Y. Chen, T. L. Molloy, T. Summers, and I. Shames, “Regret Analysis of Online LQR Control via Trajectory Prediction and Tracking,” in Proceedings of The 5th Annual Learning for Dynamics and Control Conference, ser. Proceedings of Machine Learning Research, vol. 211. PMLR, Jun. 2023, pp. 248–258.
- Y. Chen, T. L. Molloy, and I. Shames, “Two-Player Dynamic Potential LQ Games with Sequentially Revealed Costs,” accepted by Conference of Decision and Control, 2025.

## Journal Paper Drafts:

- Y. Chen, T. L. Molloy, T. Summers, and I. Shames, “Regret Analysis of Online LQR Control Using Trajectory Prediction and Tracking,” submitted to IEEE Transaction on Automatic Control, under the second round of review.
- Y. Chen, T. L. Molloy and I. Shames, “ $N$ -Player Dynamic Potential LQ Games with Sequentially Revealed Costs,” under preparation for Automatica submission.
- Y. Chen, T. Zhao, P. Braun, T. L. Molloy, and I. Shames, “Do as I Do: Online Linear-Quadratic Optimal Control with Sequentially Inferred Costs,” under preparation for IEEE Control Systems Letters submission.

---

# Contents

---

List of Figures . . . . .	xi
List of Tables . . . . .	xii
<b>1 Introduction</b>	<b>1</b>
1.1 Online LQ Optimal Control . . . . .	2
1.2 Online LQ Optimal Control with Sequentially Inferred Cost . . . . .	5
1.3 Dynamic Potential Game with Sequentially Revealed Costs . . . . .	5
1.4 Contributions . . . . .	7
1.5 Thesis Outline . . . . .	8
<b>2 Online LQ Optimal Control</b>	<b>10</b>
2.1 Related Works . . . . .	10
2.2 Problem Formulation . . . . .	13
2.3 Approach and Regret Analysis . . . . .	14
2.4 Regret Analysis of Deadbeat Tracking Controller . . . . .	19
2.5 Numerical Simulations . . . . .	23
2.6 Summary . . . . .	28
<b>3 Online LQ Optimal Control with Sequentially Inferred Cost</b>	<b>29</b>
3.1 Related Works . . . . .	29
3.2 Problem Formulation . . . . .	30
3.3 Implementation of Algorithm 3.1 . . . . .	33
3.4 Regret Analysis Under Cost Matrices Estimation . . . . .	36
3.5 Numerical Simulations . . . . .	40
3.6 Summary . . . . .	41
<b>4 Dynamic Potential Games with Sequentially Revealed Costs</b>	<b>44</b>
4.1 Related Works . . . . .	44
4.2 Problem Formulation . . . . .	45
4.3 Proposed Approach and PoU Analysis . . . . .	54
4.4 Relationship Between PoU and Feedback Nash Equilibrium . . . . .	56
4.5 Numerical Simulations . . . . .	57
4.6 Summary . . . . .	59

<b>5</b>	<b>Conclusions</b>	<b>61</b>
5.1	Summary of Contributions . . . . .	61
5.2	Future Research Directions . . . . .	63
	<b>Bibliography</b>	<b>67</b>
<b>A</b>	<b>Appendix. Proof for Online LQ Optimal Control</b>	<b>75</b>
A.1	Proof of Theorem 2.3.1 . . . . .	75
A.2	Proof of Proposition 2.3.2 . . . . .	80
A.3	Proof of Theorem 2.3.2 . . . . .	81
A.4	Proof of Theorem 2.4.1 . . . . .	87
A.5	Proof of Theorem 2.4.2 . . . . .	90
<b>B</b>	<b>Appendix. Proof for Online LQ Optimal Control with Sequentially Inferred Costs</b>	<b>94</b>
B.1	Preparatory Results for the Proof of Lemma 3.4.1 . . . . .	94
B.2	Proof of Lemma 3.4.1 . . . . .	99
B.3	Preparatory Results for the Proof of Theorem 3.4.1 . . . . .	100
B.4	Proof of Theorem 3.4.1 . . . . .	111
<b>C</b>	<b>Appendix. Proof for Dynamic Potential LQ Games</b>	<b>112</b>
C.1	Auxiliary Lemmas . . . . .	112
C.2	Proof of Theorem 4.3.1 . . . . .	126

---

# List of Figures

---

2.1	Average regret versus preview window length $W$ for disturbance-free linearised inverted pendulum with $T = 30$ . . . . .	25
2.2	Average regret versus preview window length for disturbance-free with $T = 30$ for random scalar systems on the left and random systems with $A \in \mathbb{R}^{9 \times 9}$ and $B \in \mathbb{R}^{9 \times 3}$ on the right. . . . .	25
2.3	Average regret versus preview window length with $T = 30$ for random scalar systems with disturbances $w_t \sim \mathcal{N}(0, 1)$ with $T = 30$ on the left and random systems with system matrices $A \in \mathbb{R}^{9 \times 9}$ , $B \in \mathbb{R}^{9 \times 3}$ , and disturbances $w_t \sim \mathcal{N}(0, I_{9 \times 9})$ on the right. . . . .	25
3.1	Copycat's estimated parameters $\hat{\theta}_t$ over time . . . . .	41
3.2	Comparison of expert and copycat trajectories. (Left) State trajectories $x_1$ and $x_2$ for both expert and copycat over time steps. (Right) Control input trajectories $u_1$ for the expert and the copycat. . . . .	42
3.3	$\frac{\text{Regret}_T(\{\nu_t\}_{t=0}^T)}{T}$ of the copycat with respect to different time horizons $T$ . . . . .	43
4.1	Performance measure $\text{PoU}_T$ in case $N = 3$ . (a) $\text{LogRelPoU}$ vs. Time Horizon with $W = 5$ ; (b) $\text{LogRelPoU}$ vs. Preview Window Length in $T = 26$ . . . . .	59

---

# List of Tables

---

2.1 Summary of related online optimal control works. Functions  $f_\tau, g_\tau$  denote state and control costs at time  $\tau$ , respectively. . . . . 10



---

# Introduction

---

In many real-world problems that range from portfolio optimisation [1], robotic planning [2] to demand-side management [3] and control systems [4–7], decisions must be made sequentially under uncertainty. In such settings, the decision maker often has access only to partial or delayed information about the environment. For example, only historical data and limited foresight of future conditions may be available. The fundamental challenge of sequential decision making under uncertainty has spurred extensive research across various disciplines, including engineering, economics and artificial intelligence [7–17]. In particular, a dominant research area for engineering related to sequential decision making is optimal control theory and dynamic game theory. The former one focuses on decision making by a single decision maker, while the latter one extends to the setting involving multiple decision makers.

Optimal control provides a formal approach to designing policies that minimise cumulative cost over time, subject to constraints imposed by system dynamics [18]. The archetypal optimal control problems assume that the decision maker has full knowledge of the system dynamics and cost functions across the entire decision horizon. However, in many practical settings, the decision maker does not have complete access to the full cost information, but still aims to achieve performance close to that of the optimal policy [19–21]. To capture these, we reformulate the optimal control problem into the following settings,

- i. The decision maker has access only to past and partial future cost information, rather than complete foresight
- ii. The decision maker aims to emulate an expert who optimises a certain objective, by online inferring the objective from the expert’s sequentially revealed trajectories.

In the absence of full cost information, the best achievable outcome for the decision maker is inherently suboptimal. Classical results in online convex optimisation provide rich theoretical frameworks for finding suboptimal solutions under cost uncertainty [22,23]. However, these works focus on decoupled decision variables and do

not account for system dynamics constraints arising from state evolution in control problems. To evaluate performance, we adopt the concept of regret, which measures the suboptimality of decisions made with cost uncertainty compared to an omniscient decision maker with full information. Several definitions of regret have been proposed in the literature [7–17], and we will review these variants to identify the most suitable one for our context.

While optimal control theory focuses on a single decision maker, many applications such as macroeconomic policy making [1] and demand-side management [3], involve multiple agents that each seek to optimise their own objective. The mathematical framework that generalises optimal control problems to multiple rational decision makers is noncooperative dynamic game theory [24]. In this framework, agents seek equilibria, where no agent can unilaterally improve their outcome. As in the single decision maker case, we impose informational constraints whereby each decision maker only observes past and partial future cost information. Moreover, we extend the notion of regret to the multi-agent setting, quantifying the suboptimality between the decisions made by agents with unknown future cost and those of an omniscient group of agents acting at their equilibrium.

In the following sections, we begin by introducing optimal control problems, followed by the two scenarios, (i) and (ii), of the online LQ optimal control problems. We then introduce the dynamic LQ potential games. Next, we summarise the main contributions of each chapter. Finally, we provide an outline for the thesis.

## 1.1 Online LQ Optimal Control

In this thesis, we begin by investigating decision making under unknown future costs in the single decision maker setting. We specifically consider a decision maker governed by linear time-invariant dynamics, aiming to minimise a cumulative quadratic costs. This leads to an online Linear Quadratic problem, which extends the archetypal Linear Quadratic (LQ) problem. We introduce the LQ optimal control problem in the following.

The LQ control problem is the archetypal optimal control problem. Consider a controllable linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad x_0 = \bar{x}_0 \quad (1.1)$$

where  $t$  is a non-negative integer,  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$  are *system matrices*,  $x_t \in \mathbb{R}^n$  are states,  $u_t \in \mathbb{R}^m$  are controls,  $w_t \in \mathbb{R}^n$  are disturbances, and  $\bar{x}_0 \in \mathbb{R}^n$  is

the initial state.

In the disturbance-free case,  $w_t = 0$  for all  $0 \leq t \leq T - 1$  and the LQ problem involves finding a feedback control policy  $\Pi^* := \{\pi_t^*\}_{t=0}^{T-1}$ , consisting of feedback control laws satisfying  $u_t = \pi_t^*(x_t)$ , to minimise the quadratic cost function

$$\sum_{t=0}^{T-1} x_{t+1}^\top Q_{t+1} x_{t+1} + u_t^\top R_t u_t, \quad (1.2)$$

subject to (1.1) for a given finite horizon  $1 \leq T < \infty$  and initial state  $\bar{x}_0$ . Here, the time-varying *cost matrices*  $Q_t \in \mathbb{S}_+^n$  and  $R_t \in \mathbb{S}_{++}^m$  are from the sets of positive semi-definite symmetric and positive definite symmetric matrices  $\mathbb{S}_+^n$  and  $\mathbb{S}_{++}^n$ , respectively. Under assumptions on the controllability of the system (1.1) and the positive (semi-)definiteness of the cost matrices, it is well-known that an optimal policy  $\Pi^*$  exists and is unique [18, Chapter 2.4]. We denote  $J_{LQ}^*$  as the (minimum) value of (1.2) under the optimal policy  $\Pi^*$ , and let  $x_t^*$  and  $u_t^*$  denote the associated optimal states and controls. In the (typical) case of stochastic disturbances,  $w_t$  for  $0 \leq t \leq T - 1$  are independent and identically distributed (i.i.d.) random variables with  $\mathbf{E}(w_t) = 0$  and  $\mathbf{E}(w_t w_t^\top) = \text{Cov}_w$  where  $\mathbf{E}(\cdot)$  is the expectation operator.

The stochastic LQ problem is then to design a feedback control policy  $\Pi^* := \{\pi_t^*\}_{t=0}^{T-1}$ , consisting of feedback control laws satisfying  $u_t = \pi_t^*(x_t, \{w_\tau\}_{\tau=0}^{t-1})$ <sup>1</sup>, that minimises the (expected) cost

$$\mathbf{E} \left( \sum_{t=0}^{T-1} x_{t+1}^\top Q_{t+1} x_{t+1} + u_t^\top R_t u_t \right), \quad (1.3)$$

subject to (1.1) where we note that at time  $t$  the state  $x_t$  and past disturbances  $\{w_\tau\}_{\tau=0}^{t-1}$  are available. Again, under assumptions on the controllability of the system (1.1) and the positive (semi-)definiteness of the cost matrices, the existence and uniqueness of the optimal policy  $\Pi^*$  are guaranteed [25, Chapter 5]. We denote  $J_{LQG}^*$  as the value of (1.3) under the optimal policy  $\Pi^*$ .

In classical LQ optimal control problems, the cost matrices  $\{Q_t\}_{t=0}^T$  and  $\{R_t\}_{t=0}^{T-1}$  are known *a priori*, and the optimal control policies can be found in closed form (cf. [18, Chapter 2] and [25, Chapter 5]). However, in many real-world applications, such as real-time energy pricing [26], building energy management [27], power control for wireless transmission [28], data-centre load balancing [29] and large-scale electric vehicle charging [30], decisions must be made with limited information about future costs.

---

<sup>1</sup>For time  $t = 0, 1, \dots, T - 1$ , the state  $x_t$  is a sufficient statistic for the disturbances  $\{w_\tau\}_{\tau=0}^{t-1}$ , but we retain them as an argument of the policy to highlight the control's dependence on them.

For example, suppose we wish to control a solar-powered uncrewed vehicle to efficiently track a planned trajectory  $\{x_t^{\text{track}}\}_{t=0}^T$  by using most control effort at times of the day and in weather conditions that lead to high solar irradiation levels in the vehicle's vicinity and less control effort when the irradiance is low. This objective can be modelled via cost matrices  $R_t$  that induce control costs  $u_t^\top R_t u_t$  that are inversely proportional to the solar irradiance level. Since weather conditions will influence the irradiance levels, the cost matrices  $R_t$  are time-varying, sequentially revealed, and may only be previewed over a short window of time. Such a problem was considered in [21] within the framework of online optimisation rather than optimal control by ignoring the constraints imposed by the vehicle's dynamics.

Motivated by problems with sequentially-revealed costs, we consider an *online* LQ control problem in which at any time  $t$ , only the cost matrices  $\{Q_{\tau+1}, R_\tau\}_{\tau=0}^{t+W}$  from the past and over a short preview window of (potentially zero) length  $0 \leq W \leq T-1$  are known to the decision maker (in addition to the initial state  $\bar{x}_0$ , horizon  $T$ , system matrices  $A$  and  $B$  and past disturbances  $\{w_\tau\}_{\tau=0}^{t-1}$ ). This corresponds to scenario (i) for online LQ control problems mentioned previously. The cost and disturbance information available to the decision maker at time  $t$  is thus

$$\mathcal{H}_{t,W} := \{\{Q_{\tau+1}, R_\tau\}_{\tau=0}^{t+W}, \mathcal{D}_t, \bar{x}_0\}, \quad (1.4)$$

where  $\mathcal{D}_t = \emptyset$  without the presence of disturbances and  $\mathcal{D}_t = \{w_\tau\}_{\tau=0}^{t-1}$  with the presence of disturbances in (1.1). Note that  $\mathcal{H}_{t,W}$  with  $t \geq T-1-W$  contains all the cost matrices usually assumed available in the classical LQR problem. The main focus of our work is to propose a novel framework for selecting controls  $u_t$  using the information available at time  $t$  (i.e.,  $\mathcal{H}_{t,W}$ ) such that the cost they incur compared to optimal controls selected in hindsight is bounded. Our framework adopts the notion of *dynamic regret* to capture the difference between sequentially incurred and optimal-in-hindsight (cumulative) costs. In contrast to previous treatments of online LQ control (e.g., [31]), our approach exploits the feedback of cost matrix information to predict and track the optimal trajectory. Furthermore, we establish corresponding regret bounds without consideration of “worst-case” costs.

The above introduces the archetypal LQ problems, the online LQ optimal control problem when the costs are *a priori* unknown to the decision maker. This corresponds to scenario (i) in the previous introduction. In the following, we introduce scenario (ii), the online LQ problem when costs are inferred online by observations of the optimal states and control trajectories of an expert.

## 1.2 Online LQ Optimal Control with Sequentially Inferred Cost

The previous section introduced a scenario of an online LQ problem in which cost matrices are revealed over time. In this section, in contrast, we introduce the online LQ optimal control problem when the cost matrices are no longer revealed perfectly. The decision maker has to learn the cost based on data from the expert's optimal control demonstration.

A motivating example that arises in the context of autonomous driving systems is to learn from human demonstrations. These systems aim to personalise their driving behaviour by adapting to individual driving styles. Human driving behaviour can be modelled as an LQ control problem [32], with the cost matrices capturing the individual driving style (e.g., tracking capabilities and vehicle comfort). However, these cost matrices are latent and not observable, even to the human driver. Therefore, direct cost matrix information will not be available, which requires the autonomous driving systems to infer cost matrices from the driving data released during the human demonstration. Most existing works in learning from demonstration focus on inferring parameters when full trajectories are given [citations]. However, this prevents the possibility of real-time learning on a moving vehicle, where the autonomous system needs to apply control without the cost function fully learned.

In this thesis, we will investigate this real-time learning and control problem by proposing a framework for a *copycat* that aims to mimic an *expert's* behaviour. We again adopt the notion of dynamic regret to quantify how similar the behaviour of the copycat is to the expert.

The above two sections introduced different scenarios of online LQ optimal control problems with uncertain cost matrices. The technical results will be discussed in Chapter 2 and 3, respectively. In the next section, we introduce the problem of decision-making with sequentially revealed costs, which extends scenario (i) from the single to the multiple decision-makers case. In particular, the costs and dynamics of the decision-makers form a noncooperative dynamic LQ potential game.

## 1.3 Dynamic Potential Game with Sequentially Revealed Costs

In this section, we introduce noncooperative dynamic LQ games and the problem of dynamic potential LQ games with sequentially revealed costs.

Noncooperative dynamic game theory is a mathematical framework for decision-making among rational players in dynamic environments [24, 33]. It has been widely adopted for modelling interactions between agents in applications including networked controls and communications [34, 35], economics [1, 36, 37], and power systems [38–41]. In these applications, the players try to minimise a cost functions that depend on decisions of their own and other players. The aim of the game is often set to seek a Nash equilibrium, where no player can gain by unilaterally changing its strategy. Finding the Nash equilibrium for general dynamic games is difficult due to the dependencies of coupled dynamics among different players. Discussions on properties related to Nash equilibrium under certain dynamic game structures have been discussed in [42–45].

*Dynamic Potential LQ Games* are a class of dynamic games in which feedback Nash equilibria can be determined by solving multivariate optimal control problems [46]. The recent development of dynamic potential LQ games has improved the tractability of noncooperative dynamic game models in many applications by enabling the use of well-established optimal-control solutions techniques to find Nash equilibria [3, 46]. It has been proven invaluable since it has been adopted for formulating many practical studies, such as energy demand-side management [3, 47], community battery management [48, 49], decentralised formation control of multi-vehicle systems [2] and macroeconomic policy making [1]. These problems are modelled as finite-horizon multiplayer linear quadratic dynamic games, where the full cost information is known to the players in advance.

From a theoretical standpoint, our consideration of dynamic potential games with sequentially revealed costs extends recent *online* LQ optimal control problems [8, 9, 12, 50] from a single-player setting to a multi-player noncooperative setting. The extension is, however, nontrivial. In online LQ control problems, the concept of *regret* is used to measure suboptimality of an algorithm against the optimal solution in hindsight. In a dynamic game setting, similar concepts of best-performance-in-hindsight are more difficult to define because the concept of optimality is itself ambiguous in noncooperative competitive settings, leading to equilibria solution concepts. Similarly, any proposed algorithms for solving online dynamic games must be tailored to the specific solution concept. To the best of our knowledge, these challenges have not been overcome in generalising results and algorithms from online LQ control problems to online LQ dynamic game problems. Importantly, while other decision quality measures have been explored in Markov games [51–53], these notions offer limited insight into LQ settings with continuous state and action spaces.

We consider feedback Nash equilibria as the solution concept of interest, and adopt and modify the notion of *price of uncertainty* (PoU) from static game [54] to dynamic game as an indication of decision quality.

In general, the existence and uniqueness of a Nash equilibrium (and hence the PoU performance measure) in a noncooperative dynamic game is not guaranteed, even if the cost functions of all players are strictly convex [24, Chapter 6, Section 6.2.2]. However, we shall impose specific conditions on the parameters of LQ feedback potential games so that we may define both our algorithm and performance guarantees with respect to a unique feedback Nash equilibrium.

## 1.4 Contributions

We propose a prediction and tracking framework for problems when the costs are *a priori* unknown. The framework comprises the decision maker predicting a candidate optimal trajectory, then tracking it. Specifically, our trajectory prediction step uses the feedback of the cost matrix information. As more cost matrix information is revealed, a trajectory that is closer to the optimal trajectory is tracked. This ensures that the control policy fully utilises all cost information received at each time step.

We first examine this framework in the online LQ control problem with sequentially revealed costs. We characterise the performance of this framework theoretically and numerically using the notion of dynamic regret. Our theoretical analysis shows that the incurred regret upper bound decays exponentially fast as the preview window length increases. We provide a sufficient condition under which our regret bound is less than that of the state-of-the-art methodology [12]. In simulations, we demonstrate the decays of regret when the preview horizon increases, and our proposed framework leads to controllers with improved performance compared to the state-of-the-art. In addition, our framework leads to methods that outperform those that do not use the feedback from cost matrix information.

Next, we apply this framework and use inverse optimal control to infer the cost matrices online. We establish the first dynamic regret upper bound under this proposed framework for the online LQ control problems, where costs are *a priori* unknown and must be inferred from an expert's sequentially revealed optimal trajectory.

Finally, we apply this framework in the dynamic potential LQ games with sequentially revealed costs. Since this is the first work that explores such a problem, we first extend the notion of dynamic regret from online LQ control problems to the dynamic potential game case. Inspired by the notion of *price of uncertainty* (PoU)

from static games, we develop a version of PoU for dynamic games as an indication of decision quality, and provide a connection between PoU and the notion of *price of anarchy* (PoA). We establish lower and upper bounds for the PoU incurred by the proposed framework. Our analysis shows that the PoU lower and upper bounds grow and decay, respectively, as the preview window length increases. In the single player case, the PoU upper bound specialises the regret bounds established in the online LQ control problem with sequentially revealed costs. Furthermore, we quantify how close a (general) control policy is to the feedback Nash equilibrium if it has a bounded PoU. Lastly, the numerical simulation shows the PoU decays exponentially, which matches our theoretical regret analysis.

## 1.5 Thesis Outline

The thesis consists of four chapters and is organised as follows:

In Chapter 2, we present our proposed new framework for solving online linear quadratic (LQ) control problems with time-varying cost matrices that are known only up to the current time or over a short preview window. Our framework involves using revealed cost matrices to predict the unknown optimal trajectory, and then using a tracking controller to drive the system towards this prediction. We adopted dynamic regret to measure and bound the resulting quality of control decisions with and without system disturbances, and present a constructive tracking controller design approach based on deadbeat control in a lifted space. Our analysis reveals that the regret of controllers designed using this proposed framework is upper bounded by terms that decay exponentially with the number of time steps over which future cost matrices can be previewed. Our simulations show that our proposed framework leads to controllers with improved performance compared to previously proposed online LQ methods that do not use revealed cost matrices. The exponential decay of dynamic regret with respect to the preview window length present in our regret bounds is also observed in simulations.

In Chapter 3, we investigate an online LQ control problem where the cost matrices depend on an unknown parameter. An *expert* agent that has access to the true parameter, sequentially reveals optimal trajectories to a *copycat* agent that is not aware of the true cost parameter. To estimate the unknown parameter, the copycat employs an extended Kalman filter (EKF) based method, using the optimal trajectories revealed by the expert. We then incorporate the estimated parameter into the proposed framework as the control policy for the copycat. We adopt dynamic regret to measure and bound the resulting quality of control decisions made by the copycat

---

against those of the expert. We established regret bounds associated with an EKF-based method. In our simulations, we show that the copycat's trajectory approaches the expert's optimal trajectory over time, while the regret remains constant as the time horizon increases.

In Chapter 4, we investigate a novel finite-horizon linear-quadratic (LQ) feedback potential dynamic game with a priori unknown cost matrices played between  $N$ -players. The cost matrices are revealed to the players sequentially, with the potential for future values to be previewed over a short time window. We propose an algorithm that enables the players to predict and track a feedback Nash equilibrium trajectory, and we measure the quality of their resulting decisions by introducing the concept of *price of uncertainty*. We show that under the proposed algorithm, the price of uncertainty is bounded by horizon-invariant constants. The constants are the sum of two terms; the first term decays exponentially as the preview window grows, and another that depends on the magnitude of the differences between the cost matrices for each player. Through simulations, we illustrate that the resulting price of uncertainty initially decays at an exponential rate as the preview window lengthens, then remains constant for large time horizons.

# Chapter 2. Online LQ Optimal Control

In this chapter, we introduce the online LQ optimal control problem with sequentially revealed costs. We begin by reviewing the related works for this problem, followed by the presentation of our proposed new framework, and the constructive deadbeat controller based on this framework. We then provide the regret analysis for both the framework and the associated deadbeat controller. Finally, we demonstrate the empirical performance of our proposed methods through numerical simulations, comparing the incurred regret with that of the state-of-the-art online LQ optimal control methods.

## 2.1 Related Works

Various formulations of online optimal control with unknown costs have been considered in systems and control theory [4–6, 10–12, 15, 55–57], machine learning and artificial intelligence [8, 13, 14, 58–60]. Table 2.1 summarises key features of representative works and outlines the contributions of this work. In the following, we compare and contrast our work with these existing works in more detail.

**Table 2.1:** Summary of related online optimal control works. Functions  $f_\tau, g_\tau$  denote state and control costs at time  $\tau$ , respectively.

Works	Systems	Costs	Optimal Time-varying Regret	Available Information	Causal Preview of Disturbances	Cost Matrix Feedback	Stochastic Disturbances	Regret w.r.t. Preview
[8, 10, 11]	Linear	Quadratic	$\times$ – only with respect to class of stabilising linear controllers in [8, Definition 3.4]	$\{\{Q_{\tau+1}, R_\tau\}_{\tau=0}^t, \{w_t\}_{\tau=0}^{t-1}, \bar{x}_0\}$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$
[12]	LTI	Quadratic	$\checkmark$	$\{\{Q_{\tau+1}, R_\tau\}_{\tau=0}^{t+W}, \{w_t\}_{\tau=0}^{t+W}, \bar{x}_0\}$	$\times$	$\times$	$\times$	$\checkmark$
[13]	Nonlinear-TI	Strongly Convex	$\checkmark$	$\{\{f_{\tau+1}, g_\tau\}_{\tau=0}^{t+W}, \{w_t\}_{\tau=0}^{t+W}, \bar{x}_0\}$	$\times$	$\times$	$\times$	$\checkmark$
[14]	LTI	Strongly Convex	$\checkmark$	$\{\{f_{\tau+1}, g_\tau\}_{\tau=0}^{t+W}, \bar{x}_0\}$	$\times$	$\times$	$\times$	$\checkmark$
[15]	LTI	Strongly Convex	$\checkmark$	$\{\{f_{\tau+1}, g_\tau\}_{\tau=0}^t, \bar{x}_0\}$	$\checkmark$	$\checkmark$	$\times$	$\times$
[17]	Nonlinear	Strongly Convex	$\checkmark$	$\{\{f_{\tau+1}, g_\tau\}_{\tau=0}^t, \{w_t\}_{\tau=0}^{t-1}, \bar{x}_0\}$	$\checkmark$	$\checkmark$	$\times$	$\times$
Ours	LTI	Quadratic	$\checkmark$	$\{\{Q_{\tau+1}, R_\tau\}_{\tau=0}^{t+W}, \{w_t\}_{\tau=0}^{t-1}, \bar{x}_0\}$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$

Recent work in [8, 10, 11] has focused on quadratic cost functions with linear systems

without preview of future costs, with *stabilising regret* used as the quality measure of decisions.<sup>1</sup> The notion of stabilising regret compares costs incurred by a proposed method and the least costs incurred by a fixed linear controller (i.e.,  $u_t = -Kx_t$  for  $\forall t$ ) chosen from a set of stabilising controllers when all cost information is given. However, optimal finite-horizon LQ control laws for time-varying costs are generically time-varying and may include unstable feedback gains at some time instances in the horizon. Thus, the aforementioned set of stabilising time-invariant controllers does not include the optimal control policy obtained from knowing all the cost matrices. Hence, a more natural measure of decision quality is the difference in cost incurred by decisions based on sequential cost information and decisions based on optimal control laws constructed with full cost information (only available in hindsight). This leads to the notion of *dynamic regret* from online optimisation [27, 28] which is a performance measure that compares the cumulative loss of an online algorithm against a sequence of optimal decision. However, the main point of difference between formulations found in the online optimisation literature and this chapter is the fact that optimal decisions and states in LQ optimal control problems are not just dependent on the cost at the decision time step, but are also coupled with the previous decisions and future and past costs.

Online LQ optimal control problems similar to the one considered in this chapter are studied in [12] via an approach inspired by model predictive control (MPC). Specifically, future cost matrices and disturbances are assumed to be available over a preview window of length  $W \geq 0$ . It is further assumed that the cost matrices  $Q_t$  and  $R_t$  satisfy uniform (in time) upper and lower bounds. The approach of [12] then involves formulating an MPC problem using the known future cost matrices and disturbances, and a pessimistic tail cost estimate based on the upper bounds of  $Q_t$  and  $R_t$ . Since the tail cost estimate is not updated using the cost-matrix information that becomes progressively available, the approach's performance and regret bounds are overly dependent on the ex ante assumptions made about the cost matrices. In contrast, we propose an alternative approach that uses revealed cost information to reduce the dependence of its performance on the ex ante assumptions made about the cost matrices.

The problem considered in [13] extends that of [12] by allowing non-quadratic cost functions and limited preview of future system matrices, disturbances, and costs. However, the dynamic regret upper bound presented in [13] is only valid when the preview window is longer than some minimum length, which is restrictive when

---

<sup>1</sup>As a side note, [8, Algorithm 1] may not even yield a feasible solution for a controllable linear systems, see [50, Footnote 2].

future information is unavailable. Moreover, similar to [12], the regret bound of [13] is overly reliant on assumptions related to bounded costs.

Considerations of online optimal control in the systems and control literature has addressed controller design with sequentially revealed costs under a variety of (potentially restrictive) technical conditions and *without* considering preview. For example, [15, 17] consider states and controls constrained within a (potentially time-varying) compact space, employ online convex optimisation algorithms, and relate dynamic regret upper bounds to path length to quantify changes in optimal controls or gains over consecutive time steps, without considering preview. Similarly, [56] considers the specialised control problem of tracking sequentially revealed targets, which is a special case of our online LQ optimal control problem without preview. The interplay between stability and dynamic regret has also been examined in [55].

Most recently, both [6] and [58] consider apparent generalisations of our online LQ optimal control problem, but under additional restrictive conditions. For example, [58] considers general nonlinear systems but with controller design simplified to selecting from a finite set of stabilising controllers, and regret defined with respect to the same set. We impose no such constraints on controller design and define regret with respect to all possible (uncountably infinite) controllers. Similarly, [6] considers a problem in which the controller only has partial or distributional knowledge of the costs. In contrast, we examine what is achievable when entire cost matrices are *a priori* unknown and sequentially revealed, without imposing any distributional assumptions.

**Contributions** This chapter proposes a framework for online LQ optimal control that uses cost-matrix feedback and causal disturbance information to predict and track the *a priori* unknown optimal trajectory. This framework enables us to develop: (i) methods that do not exploit knowledge of (potentially pessimistic) bounds on the cost matrices (in contrast to [12]); (ii) regret upper bounds that are well-defined when there are infinitely many available policies (in contrast to [58]); and, (iii) regret upper bounds that hold without the requirement of a minimum preview-horizon length (in contrast to [13]). Moreover, (i) we provide a sufficient condition under which our regret bound is less than that of the state-of-the-art methodology [12]; and (ii) demonstrate in simulation that our methods with cost-matrix feedback outperforms methods that do not exploit the feedback of cost information (e.g., those of [12] and [17]). We also provide (expected) regret bounds in the presence of stochastic disturbances.

**Outline** The rest of the chapter is organised as follows. In Section 2.2, we formulate the online LQ optimal control problem that we consider. In Section 2.3, we introduce our proposed framework and develop general dynamic regret bounds. In Section 2.4, we exploit the proposed framework to develop a deadbeat controller design methodology for online LQ optimal control with associated regret bounds. In Section 2.5, we provide numerical comparison results between algorithms designed with our framework and existing state-of-the-art approaches. Conclusions are provided in Section 2.6.

## 2.2 Problem Formulation

Consider the controllable LTI system (1.1), the quadratic cost function (1.2), and suppose the cost matrices  $Q_t$  and  $R_t$  satisfy the following assumption given  $T \geq 1$  and  $0 \leq W < T - 1$ .

**Assumption 2.2.1.** *There exist symmetric positive definite matrices  $Q_{min}, Q_{max}, R_{min}, R_{max}$  such that*

$$Q_{min} \preceq Q_{t+1} \preceq Q_{max}, \quad R_{min} \preceq R_t \preceq R_{max}, \quad (2.1)$$

for all  $0 \leq t \leq T - 1$  where  $F \preceq G$  denotes  $G - F$  being positive semi-definite for symmetric matrices  $F$  and  $G$ .

For the disturbance-free case, we aim to design a feedback policy  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  that selects controls  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$  for the system (1.1) using the state  $x_t$  and the currently available information  $\mathcal{H}_{t,W}$  in (1.4) with  $\mathcal{D}_t = \emptyset$ . Given  $\Pi$ , we consider the following notion of *dynamic regret* to measure decision quality

$$\text{Regret}_T(\Pi) := \sum_{t=0}^{T-1} x_{t+1}^\top Q_{t+1} x_{t+1} + u_t^\top R_t u_t - J_{LQ}^*, \quad (2.2)$$

where  $\{x_{t+1}, u_t\}_{t=0}^{T-1}$  with  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$  satisfies (1.1) and we recall that  $J_{LQ}^*$  is the cost in (1.2) under the optimal policy  $\Pi^*$  (with perfect knowledge of the cost matrices). This regret captures the difference between the cost incurred by the decision maker with partial cost information  $\mathcal{H}_{t,W}$  and that of the best decisions in hindsight with all information  $\mathcal{H}_{T-1,W}$ .

We specifically seek to design a *cost-feedback* control policy  $\Pi$ , that is independent of the bounds given in Assumption 2.2.1 (in contrast to [12]). We aim to show that the regret (as defined in (2.2)) associated with our proposed feedback policy is sublinear with respect to the time horizon  $T$  for the case where  $w_t = 0$  for  $0 \leq t \leq T - 1$ , i.e.,

$\text{Regret}_T(\Pi) \leq o(T)$ , where  $o(T)$  satisfies  $\lim_{T \rightarrow \infty} \frac{o(T)}{T} = 0$ .

For the case of stochastic disturbances, we consider designing a feedback policy  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  such that  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$  with  $\mathcal{D}_t = \{w_\tau\}_{\tau=0}^{t-1}$ . We use the notion of *expected regret* as the measure of decision quality

$$\text{ExpRegret}_T(\Pi) := \mathbf{E} \left( \sum_{t=0}^{T-1} x_{t+1}^\top Q_{t+1} x_{t+1} + u_t^\top R_t u_t \right) - J_{LQG}^*, \quad (2.3)$$

where  $\{x_{t+1}, u_t\}_{t=0}^{T-1}$  with  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$  satisfies (1.1) and we recall that  $J_{LQG}^*$  is the cost in (1.3) under the optimal policy  $\Pi^*$  (with perfect knowledge of the cost-matrices but not the disturbances).

We aim to show that our proposed control policy yields controls such that the *expected regret* satisfies  $\text{ExpRegret}_T(\Pi) \leq (C_{ER}\gamma^{2W} + C'_{ER})T$  for positive scalars  $C_{ER}, C'_{ER}$  and  $\gamma \in (0, 1)^2$ .

We begin by establishing regret upper bounds for the general case of the proposed control design framework. Later, we will provide a constructive approach to controller design leading to deadbeat controllers with regret bounds in a lifted space.

## 2.3 Approach and Regret Analysis

In this section, we introduce our proposed framework (or approach) for online LQ optimal control using revealed cost matrix information to predict and track optimal state and control trajectories. We also characterise its performance.

### 2.3.1 Cost-Feedback Controller via Prediction and Tracking

Our proposed online LQ approach involves first using the information available to the decision maker at each time  $t$  to *predict* the optimal state  $x_{t+1}^*$  and control  $u_t^*$  by solving a LQ optimal control problem of the form in (1.2) with (unknown) future cost matrices replaced by their most recently revealed values. We then use a tracking controller to *track* towards this prediction.

**Prediction of Optimal Trajectory without Disturbances** In the disturbance-free case, at each time  $t$ , we predict the optimal trajectory starting from the (known) initial state  $\bar{x}_0$  using the revealed cost matrices up to time  $t + W$

<sup>2</sup>The exact definition of  $\gamma$  will be presented in Theorem 2.3.1

and setting all (unknown) future matrices to equal their most recent (known) values at time  $t+W$ . Specifically, at time  $t$  where  $0 \leq t \leq T-W-1$ , define  $J_t(\cdot, \cdot)$  as

$$\begin{aligned} & J_t(\{\xi_{\tau+1}\}_{\tau=0}^{T-1}, \{v_\tau\}_{\tau=0}^{T-1}) \\ & := \sum_{k=0}^{t+W} [\xi_{k+1}^\top Q_{k+1} \xi_{k+1} + v_k^\top R_k v_k] + \sum_{k=t+1+W}^{T-1} [\xi_{k+1}^\top Q_{t+W+1} \xi_{k+1} + v_k^\top R_{t+W} v_k], \end{aligned}$$

and  $J_t(\{\xi_{\tau+1}\}_{\tau=0}^{T-1}, \{v_\tau\}_{\tau=0}^{T-1}) := J_T(\{\xi_{\tau+1}\}_{\tau=0}^{T-1}, \{v_\tau\}_{\tau=0}^{T-1})$  for  $T-W \leq t \leq T-1$ . We predict the optimal states and controls for  $0 \leq t \leq T-1$ , denoted  $(\{x_{\tau+1|t}\}_{\tau=0}^{T-1}, \{u_{\tau|t}\}_{\tau=0}^{T-1})$ , by solving

$$\begin{aligned} & \min_{(\{\xi_{\tau+1}\}_{\tau=0}^{T-1}, \{v_\tau\}_{\tau=0}^{T-1})} J_t(\{\xi_{\tau+1}\}_{\tau=0}^{T-1}, \{v_\tau\}_{\tau=0}^{T-1}) \\ & \text{subject to} \quad \xi_{\tau+1} = A\xi_\tau + Bv_\tau, \quad \xi_0 = \bar{x}_0. \end{aligned} \tag{2.4}$$

Recall that at time  $0 \leq t \leq T-1$ , the available information is  $\mathcal{H}_{t,W}$  and  $x_t$ . Importantly, solving (2.4) at each time  $t$  is equivalent to solving an LQ (regulator) problem, as detailed in the following proposition.

**Proposition 2.3.1.** *Let*

$$\begin{aligned} Q_{\tau+1|t} & := \begin{cases} Q_{\tau+1} & \text{if } \tau \leq \min(t+W, T-1) \\ Q_{t+W+1} & \text{if } \min(t+W, T-1) < \tau \leq T-1, \end{cases} \\ \text{and } R_{\tau|t} & := \begin{cases} R_\tau & \text{if } \tau \leq \min(t+W, T-1) \\ R_{t+W} & \text{if } \min(t+W, T-1) < \tau \leq T-1. \end{cases} \end{aligned}$$

For  $0 \leq t, \tau \leq T-1$ , the solution to the optimal control problem (2.4) is given by

$$\begin{aligned} P_{\tau|t} & = A^\top P_{\tau+1|t} A + Q_{\tau|t} + A^\top P_{\tau+1|t} B K_{\tau|t}, \\ K_{\tau|t} & := -(R_{\tau|t} + B^\top P_{\tau+1|t} B)^{-1} B^\top P_{\tau+1|t} A, \\ u_{\tau|t} & = K_{\tau|t} x_{\tau|t}, \quad x_{\tau+1|t} = A x_{\tau|t} + B u_{\tau|t} \end{aligned}$$

with  $P_{T|t} = Q_{T|t}$ . The matrices  $P_{\tau|t}$  for  $0 \leq t, \tau \leq T-1$  are positive definite under Assumption 2.2.1. Moreover, the control sequence that minimises the cost in (2.2) can be found by solving (2.4) given  $Q_{\tau+1|T-1}$  and  $R_{\tau|T-1}$  for  $0 \leq \tau \leq T-1$ .

*Proof.* See Appendix A.1. □

**Prediction of Optimal Trajectory with Disturbances** In the case of disturbances, recall that the decision maker knows the past disturbances  $\mathcal{D}_t = \{w_\tau\}_{\tau=0}^{t-1}$

at each time  $t$ , where  $0 < t \leq T - 1$ . We therefore compute the predicted states and controls,  $x_{t+1|t}$  and  $u_{t|t}$ , as elements of the sequences  $(\{x_{\tau+1|t}\}_{\tau=0}^{T-1}, \{u_{\tau|t}\}_{\tau=0}^{T-1})$ , by

$$\begin{aligned} u_{\tau|t} &= K_{\tau|t}x_{\tau|t} \text{ for } 0 \leq \tau \leq T - 1, \\ x_{\tau+1|t} &= Ax_{\tau|t} + Bu_{\tau|t} + w_{\tau} \text{ for } 0 \leq \tau \leq t - 1, \\ x_{\tau+1|t} &= Ax_{\tau|t} + Bu_{\tau|t} \text{ for } t \leq \tau \leq T - 1. \end{aligned} \quad (2.5)$$

Here the control gains  $K_{\tau|t}$  are those given by Proposition 2.3.1 that solve the LQ optimal control problem (2.4), and the predicted states  $x_{\tau+1|t}$  incorporate the (known) past disturbances  $\mathcal{D}_t$ .

**Tracking of Predicted Optimal Trajectory** The second part of our approach involves selecting controls,  $u_t$ , for the system (1.1) to track towards the predicted optimal states,  $x_{t+1|t}$ , and controls,  $u_{t|t}$ , obtained by solving either (2.4) in the disturbance-free case or (2.5) in the case of disturbances. To this aim, we propose a control policy  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  with

$$u_t = \pi_t(x_t, \mathcal{H}_{t,W}) = K_t(x_t - x_{t|t}) + u_{t|t}, \quad (2.6)$$

where  $K_t \in \mathbb{R}^{m \times n}$  are control gain matrices for which there exist constants  $C_f > 0$  and  $0 < q < 1$  such that

$$\left\| \prod_{\tau=t_0}^{t_1} (A + BK_{\tau}) \right\| \leq C_f q^{t_1 - t_0 + 1} \quad (2.7)$$

for any  $0 \leq t_0 \leq t_1 \leq T - 1$ , and  $\|\cdot\|$  denotes either the 2-norm of a vector or the spectral norm of a matrix, depending on its argument. Intuitively, such control gains  $K_t$  lead to contraction of the distance between  $x_{t+1}$  given by (1.1) and  $x_{t+1|t}$  given by either (2.4) or (2.5).

We now show that this approach leads to bounded regret in both disturbance-free and (stochastic) disturbance cases.

### 2.3.2 Regret Analysis without Disturbances

We first present a regret upper bound for the disturbance-free case with controls,  $u_t$ , generated by (2.6). For any matrix  $\Gamma$ , we further define  $\lambda_{\min}(\Gamma)$  as the minimum eigenvalue of  $\Gamma$  and  $\lambda_{\max}(\Gamma)$  as the maximum eigenvalue of  $\Gamma$ . We also require the following lemma establishing a property of the optimal control problem (2.4).

Specifically, we show that the matrices  $P_{\tau|t}$  from Proposition 2.3.1 are upper and lower bounded if the time-varying control gain is bounded and the cost matrices  $Q_\tau$  and  $R_\tau$  are upper and lower bounded.

**Lemma 2.3.1.** *Consider the optimal control problem (2.4), the controllable linear system (1.1), and suppose that Assumption 2.2.1 holds. Then there exists a positive definite matrix  $P_{max}$  such that  $Q_{min} \preceq P_{\tau|t} \preceq P_{max}$  for  $0 \leq t, \tau \leq T - 1$  where*

$$P_{max} = Q_{max} + A^\top P_{max} A - A^\top P_{max} B (R_{max} + B^\top P_{max} B)^{-1} B^\top P_{max} A. \quad (2.8)$$

*Proof.* See Appendix A.1. □

Finally, let us define the following constants:

$$C_K := \left\| (R_{min} + B^\top Q_{min} B)^{-1} \right\|^2 \left\| R_{max} B^\top \right\| \frac{\lambda_{max}^2(P_{max})}{\lambda_{min}(Q_{min})}, \quad (2.9)$$

$$\eta := \sqrt{1 - \frac{\lambda_{min}(Q_{min})}{\lambda_{max}(P_{max})}}, \quad D := \left\| R_{max} + B^\top P_{max} B \right\|, \quad (2.10)$$

$$\alpha := \lambda_{max}(A^\top P_{max} A), \quad \beta := \lambda_{min}(Q_{min}),$$

$$\gamma := \frac{\alpha}{\alpha + \beta}, \quad (2.11)$$

$$C := \frac{\lambda_{max}(P_{max})}{\lambda_{min}(Q_{min})}, \quad (2.12)$$

$$\alpha_K := 2 \left[ \left( \frac{\sigma_{max}(A)}{\sigma_{min}(B)} \right)^2 + \left( \frac{C_f + \|A\|}{\sigma_{min}(B)} \right)^2 \right], \quad (2.13)$$

along with the composite constants:  $g_1 := \eta\gamma(q(q - \eta\gamma))^{-1} - \eta q^{-1}(q - \eta)^{-1}$ ,  $g_2 := (\eta\gamma q^{-1}(q - \eta\gamma)^{-1})^2$ ,  $g_3 := (\eta q^{-1}(q - \eta)^{-1})^2$  and  $g_4 := (\gamma(1 - \gamma)^{-1})^2$ .

Our main result for the disturbance-free case follows.

**Theorem 2.3.1.** *For a given horizon  $T \geq 1$  and preview window length  $0 \leq W \leq T - 1$ , consider the linear system in (1.1) and the control policy  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  with  $\pi_t$  defined in (2.6) and (2.4). Under Assumption 2.2.1, the regret defined by (2.2) satisfies*

$$\text{Regret}_T(\Pi) \leq \gamma^{2W} \Psi, \quad (2.14)$$

where  $\gamma \in (0, 1)$  is defined in (2.11), and  $\Psi$  is a positive scalar that is monotonically increasing with respect to  $\|\bar{x}_0\|, D, \alpha_K, C, C_f, C_K, g_1, g_2, g_3$  and  $g_4$ .

*Proof.* See Appendix A.1. □

**Remark 2.3.1.** For any  $T$ , there exists a  $\Lambda \in \mathbb{R}$  that is independent of  $T$  such that  $\Psi \leq \Lambda$ . Thus,  $\overline{\lim}_{T \rightarrow \infty} \frac{\text{Regret}_T(\Pi)}{T} = 0$ , which  $\text{Regret}_T(\Pi) \leq o(T)$ . This implies that the control sequence described by (2.6) yields sublinear regret.

The regret upper bound presented in [12, Theorem 1, Equation (15)] also establishes exponential decay with respect to the preview window  $W$  and sublinear growth with respect to the time horizon  $T$  under their proposed control policy. However, their policy relies on prior knowledge of the extrema of the cost matrices  $Q_{t+1}$  and  $R_t$  for  $0 \leq t \leq T - 1$ , rather than operating solely on the realised costs. As a consequence, the resulting control policy can be overly conservative when the realised cost sequence deviates significantly from these extrema.

In the following proposition, we state a condition in terms of the bounds given in Assumption 2.2.1 and the cost matrices sequence extrema under which the bound in Theorem 2.3.1 is smaller than that of [12, Theorem 1, Equation (15)].

**Proposition 2.3.2.** Adopt the hypothesis of Theorem 2.3.1. If

$$\lambda_{\max}^{10}(Q_{\max}) \geq \frac{5 \left[ \left(1 + \frac{\alpha_K}{(1-\gamma)^2}\right) \left(\frac{1}{1-\eta^2}\right) + \frac{10C_f^2}{3q^2(q-\eta)^2} \left( \frac{1}{(q-\eta)^2(1-q)^2} + \frac{2}{(1-\eta^2)} \right) \right]}{6(C_K^2 \lambda_{\min}^2(R_{\min}) \lambda_{\min}^4(Q_{\min}))^{-1} \|A\|^2 \|B\|^2 \|BR_{\min}^{-1}B^\top\|^2}, \quad (2.15)$$

where  $Q_{\max}$  is given in Assumption 2.2.1, then the RHS of inequality in [12, Theorem 1, Equation (15)] is greater than the RHS of inequality (2.2) in Theorem 2.3.1.

*Proof.* See Appendix A.2. □

We next present a regret bound with disturbances.

### 2.3.3 Regret Analysis with Disturbances

To bound the regret of our approach (2.6) in the presence of disturbances, recall that we still consider the controls  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$  given by (2.6) but with  $x_{\tau|t}$  and  $u_{\tau|t}$  obtained by solving (2.5).

**Theorem 2.3.2.** For a given time horizon  $T \geq 1$  and preview window length  $0 \leq W \leq T - 1$ , consider the system defined by (1.1) and policy  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  with  $\pi_t$  defined in (2.6) and (2.5). Under Assumption 2.2.1, the expected regret defined by (2.3) satisfies

$$\text{ExpRegret}_T(\Pi) \leq (C_{ER}\gamma^{2W} + C'_{ER})T, \quad (2.16)$$

where  $\gamma \in (0, 1)$  is defined in (2.11), and  $C_{ER}$  and  $C'_{ER}$  are positive scalars that are

*monotonically increasing with respect to  $\text{Tr}(\text{Cov}_w), \|\bar{x}_0\|, D, \alpha_K, C, C_f, C_K, g_1, g_2, g_3$  and  $g_4$ .*

*Proof.* See Appendix A.3. □

Theorem 2.3.2 implies that in the disturbance case, the expected regret can, at worst, grow linearly with the horizon  $T$ .

The above provides a framework for online LQ optimal control. We next develop a possible approach to construct  $K_t$  in (2.6).

## 2.4 Regret Analysis of Deadbeat Tracking Controller

A natural question that arises in the design of the tracking controller (2.6) is the choice of the gain  $K_t$ . In principle, it could be chosen to minimise the regret bounds, (2.14) or (2.16), with cost matrices replaced by their bounds. However, this approach may not guarantee small empirical regret. Motivated by move blocking in MPC, in this section we instead propose using a deadbeat controller to track the predicted trajectory at every  $d$ -steps when the system (1.1) is  $d$ -step controllable.

### 2.4.1 Lifted Space

In this section, we suppose that the linear system (1.1) is  $d$ -step controllable [61, Equation (6.15)], i.e., there exists an integer  $1 \leq d \leq n$ , such that  $\text{rank}([B, AB, \dots, A^{d-1}B]) = n$ . Specifically, let  $d$  be the minimal integer that system (1.1) is  $d$ -step controllable. For time horizon  $T$  and preview horizon  $W$ , we assume that  $T \bmod d = 0$  and  $W \bmod d = 0$ . Let  $T_d := \frac{T}{d}$ ,  $W_d := \frac{W}{d}$ , and assume that  $T_d \geq 1, 0 \leq W_d \leq T_d$ . Let  $\oplus$  denote the direct sum operator, such that given a sequence of matrices  $U_i \in \mathbb{R}^{k_1 \times k_2}$ , for any positive integer  $p$ ,

$$\bigoplus_{i=1}^p U_i := \begin{bmatrix} U_1 & 0_{k_1 \times k_2} & \dots & 0_{k_1 \times k_2} \\ 0_{k_1 \times k_2} & U_2 & \dots & 0_{k_1 \times k_2} \\ \vdots & \ddots & \dots & \vdots \\ 0_{k_1 \times k_2} & \dots & \dots & U_p \end{bmatrix}. \text{ For } 0 \leq k_d, t_d \leq T_d, \text{ define the following}$$

notation for a ‘lifted space’ rewriting of (1.1), (1.2) and (1.3),

$$\begin{aligned}\hat{A} &:= I_{n(d+1)} - \begin{bmatrix} 0_{n,nd} & 0_{n,n} \\ \oplus_{j=0}^{d-1} A & 0_{nd,n} \end{bmatrix}, \hat{B} := \begin{bmatrix} 0_{n,md} \\ \oplus_{j=0}^{d-1} B \end{bmatrix}, \\ \Phi &:= \begin{bmatrix} 0_{n,nd} & I_n \end{bmatrix} \hat{A}^{-1} \begin{bmatrix} I_n \\ 0_{nd,n} \end{bmatrix}, \Gamma := \begin{bmatrix} 0_{n,nd} & I_n \end{bmatrix} \hat{A}^{-1} \hat{B}, \\ C_{k_d|t_d} &:= Q_{k_d|t_d}^{\frac{1}{2}}, \hat{C}_{k_d|t_d} = \begin{bmatrix} \oplus_{j=0}^{d-1} C_{dk_d+j|t_d} & 0_{nd,n} \end{bmatrix}, \\ \hat{R}_{k_d|t_d} &:= \oplus_{j=0}^{d-1} R_{dk_d+j|t_d},\end{aligned}\tag{2.17}$$

$$\Xi_{k_d|t_d} := \hat{C}_{k_d|t_d} \hat{A}^{-1} \begin{bmatrix} I_n \\ 0_{nd,n} \end{bmatrix},\tag{2.18}$$

$$\Delta_{k_d|t_d} := \hat{C}_{k_d|t_d} \hat{A}^{-1} \hat{B},\tag{2.19}$$

$$\tilde{Q}_{k_d|t_d} := \Xi_{k_d|t_d}^\top \Xi_{k_d|t_d} - \Xi_{k_d|t_d}^\top \Delta_{k_d|t_d} \tilde{R}_{k_d|t_d}^{-1} \Delta_{k_d|t_d}^\top \Xi_{k_d|t_d},$$

$$\tilde{R}_{k_d|t_d} := \hat{R}_{k_d|t_d} + \Delta_{k_d|t_d}^\top \Delta_{k_d|t_d},$$

$$\tilde{A}_{k_d|t_d} := \Phi - \Gamma \tilde{R}_{k_d|t_d}^{-1} \Delta_{k_d|t_d}^\top \Xi_{k_d|t_d}, \tilde{B} := \Gamma,$$

$$\hat{A}_w := \begin{bmatrix} A^{d-1} & A^{d-2} & \cdots & 1 \end{bmatrix}, \tilde{w}_t := \begin{bmatrix} w_t & w_{t+1} & \cdots & w_{t+d-1} \end{bmatrix}^\top.$$

The ‘unlifted’ space equations established in Proposition 2.3.1 correspond to the following ‘lifted’ space equations,

$$\begin{aligned}\tilde{P}_{T_d|t_d} &= \tilde{Q}_{T_d|t_d} = P_{dT_d|dt_d}, \\ (\tilde{R}_{\tau_d|t_d} + \tilde{B}^\top \tilde{P}_{\tau_d+1|t_d} \tilde{B}) \tilde{K}_{\tau_d|t_d} + \tilde{B}^\top \tilde{P}_{\tau_d+1|t_d} \tilde{A}_{\tau_d} &= 0,\end{aligned}\tag{2.20}$$

$$\tilde{x}_{\tau_d+1|t_d} = \tilde{A}_{\tau_d} \tilde{x}_{\tau_d|t_d} + \tilde{B} \tilde{u}_{\tau_d|t_d} + \hat{A}_w \tilde{w}_{\tau_d},\tag{2.21}$$

$$\tilde{u}_{\tau_d|t_d} = \tilde{K}_{\tau_d|t_d} \tilde{x}_{\tau_d|t_d}, \quad \tilde{x}_0|t_d = \tilde{x}_0.\tag{2.22}$$

By [5, Proposition 3], we have the following relationships

$$P_{d\tau_d|dt_d} = \tilde{P}_{\tau_d|t_d}, \quad x_{d\tau_d|dt_d} = \tilde{x}_{\tau_d|t_d},\tag{2.23}$$

and  $(\tilde{A}_{\tau_d}, \tilde{B})$  is one-step controllable for  $0 \leq \tau_d, t_d \leq T_d$ . We use these lifted space results to describe a constructive design of the tracking controller (2.6).

## 2.4.2 Proposed Method In Lifted Space

To introduce a lifted-space method for constructively designing the tracking controller (2.6), note that at time  $t_d$  where  $0 \leq t_d \leq T_d$ , we can predict the optimal trajectory using the available information given in  $\tilde{\mathcal{H}}_{t_d, W_d} (= \mathcal{H}_{dt_d, dW_d})$ . Let  $\tilde{x}_{t_d+1|t_d}$  denote the estimate of the optimal state at time  $t_d + 1$  based on  $\tilde{\mathcal{H}}_{t_d, W_d}$ . Given the

$d$ -step controllability of the system (1.1), we can then track (exactly) to the state  $\tilde{x}_{t_d+1|t_d}$  by time  $t_d$ . Our approach in lifted space follows.

### Prediction in Lifted Space

In the disturbance-free case, similar to (2.4), at time  $t_d$  let

$$\begin{aligned} & \tilde{J}_{t_d}(\{\tilde{x}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d}, \{\tilde{u}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d-1}) \\ & := \sum_{\tau_d=0}^{T_d-1} \tilde{x}_{\tau_d|t_d}^\top \tilde{Q}_{\tau_d|t_d} \tilde{x}_{\tau_d|t_d} + \tilde{u}_{\tau_d|t_d}^\top \tilde{R}_{\tau_d|t_d} \tilde{u}_{\tau_d|t_d} + \tilde{x}_{T_d|t_d}^\top \tilde{Q}_{T_d|t_d} \tilde{x}_{T_d|t_d}, \end{aligned}$$

where  $(\{\tilde{x}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d}, \{\tilde{u}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d-1})$  satisfy (2.21). We predict the optimal (lifted) trajectories  $\{\tilde{x}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d}, \{\tilde{u}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d-1}$  by solving

$$\min_{\{\tilde{u}_{k_d|t_d}\}_{k_d=0}^{T_d-1}} \tilde{J}_{t_d}(\{\tilde{x}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d}, \{\tilde{u}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d-1}) \quad (2.24)$$

subject to (2.21). Similarly, in the case of (stochastic) disturbances, similar to (2.5), at time  $t_d$ ,  $\tilde{\mathcal{H}}_{t_d, W_d}$  is available to the decision maker and we predict the optimal (lifted) trajectories  $\{\tilde{x}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d}, \{\tilde{u}_{\tau_d|t_d}\}_{\tau_d=0}^{T_d-1}$  via

$$\begin{aligned} \tilde{u}_{\tau_d|t_d} &= \tilde{K}_{\tau_d|t_d} \tilde{x}_{\tau_d|t_d} \text{ for } 0 \leq \tau_d \leq T_d - 1, \\ \tilde{x}_{\tau_d+1|t_d} &= \tilde{A}_{\tau_d} \tilde{x}_{\tau_d|t_d} + \tilde{B} \tilde{u}_{\tau_d|t_d} + \hat{A}_w \tilde{w}_{\tau_d} \text{ for } 0 \leq \tau_d \leq t_d - 1, \\ \tilde{x}_{\tau_d+1|t_d} &= \tilde{A}_{\tau_d} \tilde{x}_{\tau_d|t_d} + \tilde{B} \tilde{u}_{\tau_d|t_d} \text{ for } t_d \leq \tau_d \leq T_d - 1, \end{aligned} \quad (2.25)$$

with  $\tilde{w}_0 := [0, 0, \dots, 0]$ .

### Tracking in Lifted Space

Given the prediction of the optimal state trajectories (in lifted space), we drive the system state (1.1) to them in  $d$ -steps using a deadbeat controller in lifted space. For any vector  $x$ , let  $\dim(x)$  be the dimension of  $x$ , we use  $[x]_{i:j}$  where  $1 \leq i \leq j \leq \dim(x)$  to denote the (sub)vector consisting of the  $i$ -th to  $j$ -th components (inclusive) of  $x$ . For time  $0 \leq t \leq T - 1$ , let  $t_d := \lfloor \frac{t}{d} \rfloor$ ,  $\underline{t}_{m,d} := 1 + m(t \bmod d)$  and  $\bar{t}_{m,d} := m(1 + t \bmod d)$ , we proposed the policy  $\tilde{\Pi} = \{\tilde{\pi}_t\}_{t=0}^{T-1}$  with deadbeat controllers

$$u_t = \tilde{\pi}_t(\tilde{x}_{t_d}, \tilde{\mathcal{H}}_{t_d, W_d}) := [(\tilde{B}^\top \tilde{B})^{-1} \tilde{B}^\top (\tilde{x}_{t_d+1|t_d} - \tilde{A}_{t_d} \tilde{x}_{t_d})]_{\underline{t}_{m,d} : \bar{t}_{m,d}} \quad (2.26)$$

where the predicted states and controls are given by (2.24) in the disturbance-free case or (2.25) in the case of disturbances.

We next show that using deadbeat controllers (in the lifted space) yields regret bounds similar to those we established in Theorems 2.3.1 and 2.3.2.

### 2.4.3 Deadbeat Regret Analysis without Disturbances

In this section, we present the result for the disturbance-free case where the control inputs are generated by (2.26). As demonstrated in the following theorem, the prediction-tracking method in lifted space also provides a sublinear regret.

**Theorem 2.4.1.** *Given  $T \geq 1$  and  $W$  such that  $0 \leq W \leq T - 1$ ,  $T \bmod d = 0$  and  $W \bmod d = 0$ , consider the LTI system (1.1) and the (lifted-space) policy  $\tilde{\Pi} = \{\tilde{\pi}_t\}_{t=0}^{T-1}$  with  $\tilde{\pi}_t$  defined in (2.26). Under Assumption 2.2.1, the regret defined by (2.2) satisfies*

$$\text{Regret}_T(\tilde{\Pi}) < \tilde{\Psi} \gamma^{2\max(0, W-d)},$$

where  $\gamma \in (0, 1)$  is defined in (2.11) and  $\tilde{\Psi}$  is a positive scalar that is monotonically increasing with respect to  $\|\bar{x}_0\|$ ,  $d$ ,  $\|P_{\max}\|$ ,  $C$  and  $C_K$ , and monotonically decreasing with respect to  $\gamma$  and  $\eta$ .

*Proof.* See Appendix A.4. □

### 2.4.4 Deadbeat Regret Analysis with Disturbances

The result presented in the following theorem addresses the disturbance case.

**Theorem 2.4.2.** *For a given horizon  $T \geq 1$  and preview window length  $W$  where  $0 \leq W \leq T - 1$  satisfy  $T \bmod d = 0$  and  $W \bmod d = 0$ , consider the linear system defined by (1.1) and control policy  $\tilde{\Pi} = \{\tilde{\pi}_t\}_{t=0}^{T-1}$  with  $\tilde{\pi}_t$  defined in (2.26). Under Assumption 2.2.1, the expected regret defined by (A.13) satisfies*

$$\text{ExpRegret}_T(\tilde{\Pi}) < (\tilde{C}_{ER} \gamma^{2\max(0, W-d)} + \tilde{C}'_{ER})T,$$

where  $\gamma \in (0, 1)$  is defined in (2.11), and  $\tilde{C}_{ER}$  and  $\tilde{C}'_{ER}$  are positive scalars that are monotonically increasing with respect to  $\|\bar{x}_0\|$ ,  $d$ ,  $\|P_{\max}\|$ ,  $C$  and  $C_K$ , and monotonically decreasing with respect to  $\text{Tr}(A^d \text{Cov}_w A^{d\top})$ ,  $\gamma$ , and  $\eta$ .

*Proof.* See Appendix A.5. □

We next illustrate the (empirical) regret of our methods.

## 2.5 Numerical Simulations

We numerically<sup>3</sup> compare: (i) the state-of-the-art approach (**OnlineMPC**) of [12, Algorithm 1]; (ii) prediction-tracking controllers (**Tracking**) given by (2.6) with (time-invariant) stabilising gains  $K_t = K$ ; (iii) deadbeat controllers (**Deadbeat**) given by (2.26); (iv) the **Safe-OGD** approach of [17, Algorithm 1]; and, (v) simple constant stabilising feedback controllers (**Constant**) such that  $u_t = Kx_t$ . We use  $\text{Regret}_{T,W}$  to denote the average regret over trials of experiments given a time horizon  $T$  and preview window length  $W$ .

### 2.5.1 Linearised Inverted Pendulum

We first consider the (slightly modified) controllable linearised inverted pendulum system from [62, Chapter 2.13]:

$$x_{t+1} = \begin{pmatrix} 0.1 & 1 & 0 & 0 \\ 0 & -0.1818 & 2.6727 & 0 \\ 0.2 & 0 & 0 & 1 \\ 0 & -18.1818 & 31.1818 & 0 \end{pmatrix} x_t + \begin{pmatrix} 0 \\ 1.8182 \\ 0 \\ 4.5455 \end{pmatrix} u_t.$$

We vary the preview horizon  $W$  from 0 to 7, and the horizon  $T$  from 8 to 50. The cost matrices are chosen uniformly satisfying by Assumption 2.2.1 with  $Q_{min} = 10^3 I_{4 \times 4}$ ,  $Q_{max} = 10^4 I_{4 \times 4}$ ,  $R_{min} = 10$ , and  $R_{max} = 10^2$ . The gains  $K$  in the **Tracking** controller (2.6) and the **Constant** controller (i.e.,  $u_t = Kx_t$ ) are both chosen by placing the pole at  $p = [p_1, \dots, p_4]$  with  $p_i \sim \mathcal{U}(0, 0.5)$ ,  $1 \leq i \leq 4$ . We note that the system is 4-step controllable.

Figure 2.1 reports the average of  $\text{Regret}_{T,W}$  over 500 trials of the approaches with  $T = 30$ . Note that the **Safe-OGD** algorithm requires finding a control gain  $\bar{K}$  that satisfies  $\|A + B\bar{K}\| \leq 1 - \zeta$  where  $0 < \zeta < 1$ . However, in this scenario, there is no such  $\bar{K}$ . Therefore, the **Safe-OGD** algorithm is not applicable. In the disturbance-free case, the horizon  $T$  does not have a significant impact on the regret. Thus, limit the presentation to the case where  $T = 30$ . From Figure 2.1 we see that the (average) regret incurred by the simple **Constant** controller is much larger than regret incurred by all other methods and unsurprisingly independent of the preview window length. For short preview window lengths  $W \leq 3$ , the regret incurred by all methods is similar. However, for preview window lengths  $W > 3$ , the **Tracking** controller incurs a lower regret than all other methods, with it being  $10^3 - 10^5$  times smaller

<sup>3</sup>Code at <https://github.com/u7361886/OnlineLQR>

than `OnlineMPC`. The `Deadbeat` controller performs similarly to `OnlineMPC` across all preview window lengths, but performs worse than the (bespoke) `Tracking` controller. This result highlights that whilst the `Deadbeat` controller design of Section 2.4 is constructive, its relative inflexibility in tracking gain compared to the general `Tracking` controller (2.6) can come with an associated increase in regret. Moreover, the exponential decay of the average regret for the `Tracking` and `Deadbeat` controllers illustrates the relation between the preview window length  $W$  and the regret bounds established in Theorem 2.3.1 and 2.4.1, respectively.

To quantify the dispersion of this result, we adopt the maximum coefficient of variation. Let  $\text{StdRegret}_{T,W}$  denotes the standard deviation regret under given time horizon  $T$  and preview window length  $W$  over the 500 trials. We define the maximum coefficient of variation as

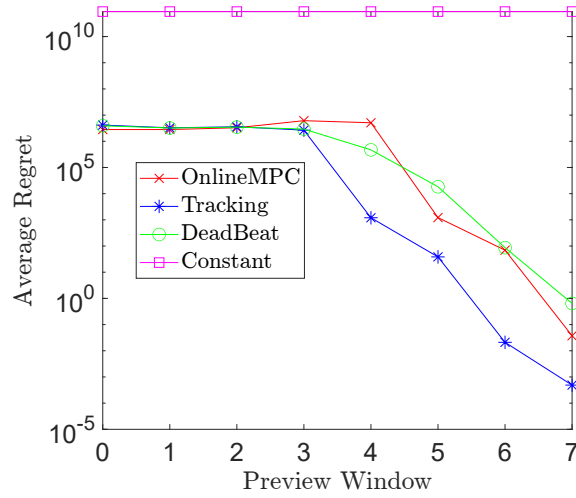
$$\text{MaxCV} := \max_W \frac{\text{StdRegret}_{T,W}}{\text{Regret}_{T,W}}. \quad (2.27)$$

Here, the preview window length  $W$  is restricted to the range  $[0, 7]$ , and the time horizon  $T = 30$ , specified by the experimental setting. In this experiment, the `MaxCV` for `Tracking`, `OnlineMPC` and `DeadBeat` are 2.035, 5.8486 and 2.3986, respectively.

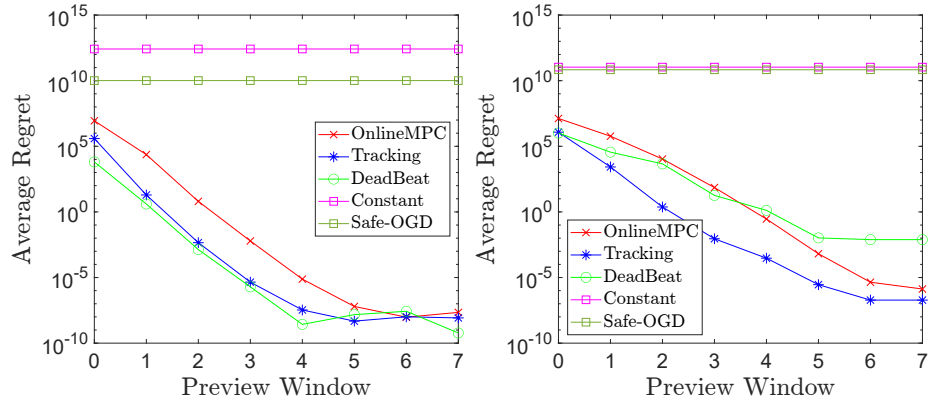
**Remark 2.5.1.** *For  $W = 0$ , the average regret incurred by `OnlineMPC` is different from the regret incurred by the `Tracking` controller. This discrepancy arises because `OnlineMPC` estimates the tail cost using  $P_{max}$ , whereas in the `Tracking` controller, at any given time, a candidate trajectory generated by solving an optimal control problem using the costs observed up to that time instance is tracked. The impact of the tail-cost construction for `OnlineMPC` on the observed performance is significant if the “future” costs are not close to their assumed extremal values of  $Q_{max}$  and  $R_{max}$ .*

## 2.5.2 Random Linear Systems without Disturbances

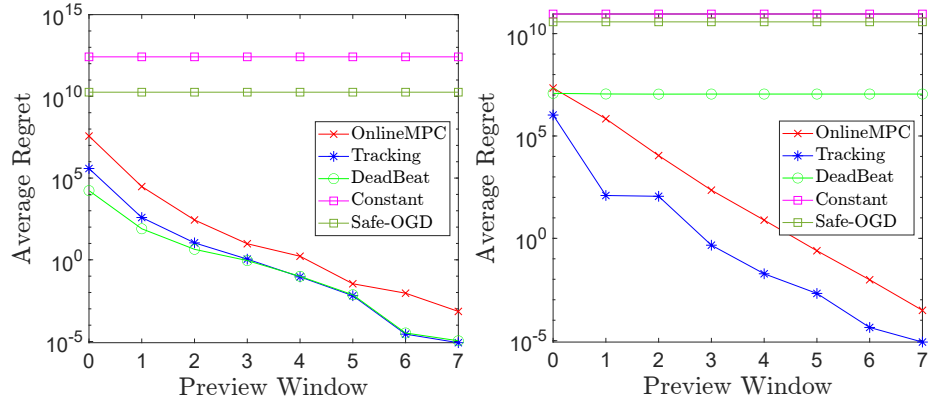
We next consider random linear systems with the elements of  $A$  and  $B$  drawn uniformly from the interval  $(0, 10)$ . We examine two cases: (i) scalar  $A$  and  $B$ ; (ii) matrix-valued  $A \in \mathbb{R}^{9 \times 9}$  and  $B \in \mathbb{R}^{9 \times 3}$ . For case (i), we randomly draw the system parameters according to  $A, B \sim \mathcal{U}(0.1, 0.9)$ . For case (ii), the elements are drawn uniformly from the interval  $(0, 10)$ , followed by normalisation of their singular values. The horizon  $T$  is varied between 10 and 50, and the preview horizon  $W$  varies from 0 to 9. The cost matrices are sampled uniformly in the ranges  $Q_{min} = 10^3 I_{n \times n}$ ,  $Q_{max} = 10^4 I_{n \times n}$ ,  $R_{min} = 10 I_{m \times m}$  and  $R_{max} = 10^2 I_{m \times m}$ , where the  $n$  and  $m$  are



**Figure 2.1:** Average regret versus preview window length  $W$  for disturbance-free linearised inverted pendulum with  $T = 30$ .



**Figure 2.2:** Average regret versus preview window length for disturbance-free with  $T = 30$  for random scalar systems on the left and random systems with  $A \in \mathbb{R}^{9 \times 9}$  and  $B \in \mathbb{R}^{9 \times 3}$  on the right.



**Figure 2.3:** Average regret versus preview window length with  $T = 30$  for random scalar systems with disturbances  $w_t \sim \mathcal{N}(0, 1)$  with  $T = 30$  on the left and random systems with system matrices  $A \in \mathbb{R}^{9 \times 9}$ ,  $B \in \mathbb{R}^{9 \times 3}$ , and disturbances  $w_t \sim \mathcal{N}(0, I_{9 \times 9})$  on the right.

chosen depending on the system matrices dimensions. Note that only 1-step controllable and 3-step controllable systems are used in simulations for cases (i) and (ii), respectively.

The gains  $K$  in the **Tracking** controller (2.6), the **Constant** controller (i.e.,  $u_t = Kx_t$ ) and the initial gain  $K_0$  in **Safe-OGD** are chosen by placing the pole(s) randomly uniformly in the circle with radius 0.5. In addition, we choose step size  $10^{-6}$ , impose a bound of 10 for  $\|K_t\|$  (i.e.,  $u_t = K_t x_t$ ), and constraint  $\|A + BK_t\|$  to be at most 0.9 during gradient descent and projection for **Safe-OGD**.

Figure 2.2 reports the average  $\text{Regret}_T$  from 500 trials of the four approaches in the case where  $A$  and  $B$  are scalars. Similar to Section 2.5.1, we again only show results for  $T = 30$  due to insignificant changes of regret in terms of  $T$ . The figure suggests that both our proposed **Tracking** and **Deadbeat** controllers outperform **OnlineMPC** and **Safe-OGD** (as well as the **Constant** controller) for all preview window lengths  $W \geq 0$ . Note that at any time step  $t$ , **Safe-OGD** only uses the cost function  $x_{t+1}^\top Q_{t+1} x_{t+1} + u_t^\top R_{t+1} u_t$ , therefore the regret remains the same regardless of the preview length. Moreover, in this experiment, the MaxCV for **Tracking**, **OnlineMPC** and **DeadBeat** are 4.3007, 4.3391 and 4.3039, respectively.

Figure 2.2 also reports the average  $\text{Regret}_T$  from 500 trials of the four approaches in the case where  $A \in \mathbb{R}^{9 \times 9}$  and  $B \in \mathbb{R}^{9 \times 3}$ , with gains  $K$  in the **Tracking** controller (2.6) and the **Constant** controller (i.e.,  $u_t = Kx_t$ ) are both chosen by placing the pole(s) at  $p = [p_1, p_2, \dots, p_9]$  with  $p_i \sim \mathcal{U}(0, 0.5)$ ,  $1 \leq i \leq 9$ . Similar to Section 2.5.1, we again only show results for  $T = 30$  due to insignificant changes of regret in terms of  $T$ . The figure suggests that both our proposed **Tracking** controllers outperform **OnlineMPC** (as well as the **Constant** controller and the **Safe-OGD** controller) for all preview window lengths  $W \geq 0$ . Moreover, the regret incurred by **Deadbeat** controller is slightly higher than the regret than **OnlineMPC** while still outperforming all other methods except the **Tracking** controller. To report the dispersion across the 500 trials of experiments, the MaxCV for **Tracking**, **OnlineMPC** and **DeadBeat** are 2.9032, 4.1016 and 4.3589, respectively.

Furthermore, the results from both figures suggest that the regret of both our proposed **Tracking** and **Deadbeat** controllers decay exponentially with the preview window length  $W$ , which align with the regret bounds in Theorems 2.3.1 and 2.4.1, respectively.

### 2.5.3 Random Linear Systems with Disturbances

We finally consider the two systems with system matrices selected according to

Section 2.5.2 with disturbances  $w_t \sim \mathcal{N}(0, 1)$  for  $0 \leq t \leq T - 1$ . Cost matrices are selected uniformly within bounds that are identical to Section 2.5.2. The selection of the preview window length, time horizon, the pole of gain  $K$  for the **Tracking** controller (2.6) and the **Constant** controller (i.e.,  $u_t = Kx_t$ ) are identical to the experiment in Section 2.5.2.

Figure 2.3 reports the average of regret from 500 trials of the five approaches for the scalar systems case (for  $T = 30$  due to the regret for other horizons  $T$  differing only by multiplicative factors). From Figure 2.3 we see that our proposed **Deadbeat** controller outperforms **OnlineMPC** (as well as the **Constant** and **Safe-OGD** controller) for all  $W \geq 0$ . Moreover, the regret incurred by our proposed **Tracking** controller is slightly higher than **OnlineMPC** while still outperforming all other methods except the **Deadbeat** controller. We note, however, that due to the presence of disturbances in these simulations the performance difference between the methods is smaller than that observed in the disturbance-free results of Figure 2.2.

The average of regret from 500 trials of the five approaches for matrix-valued systems  $A \in \mathbb{R}^{9 \times 9}$  and  $B \in \mathbb{R}^{9 \times 3}$  (for  $T = 30$  due to the regret for other horizons  $T$  differing only by multiplicative factors) is depicted in Figure 2.3 where it can be observed that our proposed **Tracking** controller outperforms all other methods for all  $W \geq 0$ . To quantify the dispersion across these experiments, for the random scalar systems case, the MaxCV for **Tracking**, **OnlineMPC** and **DeadBeat** are 7.1834, 8.4131 and 1.4144, respectively. In addition, for the random systems with  $A \in \mathbb{R}^{9 \times 9}$  and  $B \in \mathbb{R}^{9 \times 3}$ , the corresponding MaxCV values for **Tracking**, **OnlineMPC** and **DeadBeat** are 7.2512, 7.2453 and 6.2450, respectively.

The **Deadbeat** controller regret also exhibits exponential decay, albeit at a much slower rate and with much larger coefficient. This is due to the fact that when the candidate trajectory in (2.25) at time step  $t$  is computed, it only incorporates disturbance information up to time step  $t - d$ . As a result, the most recent  $d$ -steps of disturbances are ignored during tracking, causing the regret not to decay until the next step in the lifted space.

The exponential decay of the average regret for the **Tracking** and **Deadbeat** controllers with the preview window length  $W$  is consistent with the decay of the regret bounds in Theorems 2.3.2 and 2.4.2, respectively. As the number of controllable steps  $d$  increases, the terms  $\tilde{C}_{ER}$  and  $\tilde{C}'_{ER}$  in Theorem 2.4.2 grow exponentially, leading the expected regret to have a larger coefficient.

## 2.6 Summary

This chapter proposes a new online LQ optimal control design framework that uses cost matrix feedback and only causal disturbance information to predict and track an optimal trajectory. Contrary to the state-of-the-art, and are well defined when there are infinite control policies in the search space. Moreover, we do not require a minimum (nonzero) preview window length  $W$ . The effectiveness of controllers designed using our proposed framework compared to existing state-of-the-art and fixed stabilising controllers is demonstrated in numerical simulations.

# Chapter 3: Online LQ Optimal Control with Sequentially Inferred Cost

---

In this chapter, we introduce the online LQ optimal control problem via inferring costs from expert’s optimal demonstration. We begin by reviewing the related works for this problem. Then, we apply the framework designed in Chapter 2, with the cost inferred by inverse optimal control methods. We then provide the regret analysis of this framework. Finally, we demonstrate the empirical performance of our proposed methods through numerical simulations to validate the theoretical guarantee.

## 3.1 Related Works

To the best of our knowledge, the most closely related work is [6]. In their setting, the decision maker aims to replicate the performance of an  $N$ -step oracle, where the oracle solves an  $N$ -step receding horizon problem with known parameters. In contrast, their regret compares the performance between the decision maker and the  $N$ -step oracle.

Although the amount of studies that directly addressing our setting is limited, there are studies in inverse optimal control that are related to the parameter inference method of the copycat. For example, [63] proposed an unscented Kalman filter to estimate cost parameter online using the observed optimal trajectories, and [64] proposed an online computationally efficient extended Kalman filter (EKF) algorithm along with theoretical error bound guarantees. Motivated by these works, we adopt an EKF-based method for parameter estimation by the copycat and propose a solution for solving the online LQ optimal control problem where the cost matrices are inferred online.

**Notation** Let  $\mathbb{S}_{++}^n$  be the set of  $n \times n$  positive definite matrices. Define  $\mathbb{N}_{T-1} := \{0, 1, \dots, T-1\}$ . We use  $\|\cdot\|$  to denote the 2-norm of a vector or a matrix, depending on its argument. Similarly, the infinity norm is denoted by  $\|\cdot\|_\infty$ . For any symmetric matrix  $X \in \mathbb{R}^{n \times n}$ , we let  $\lambda_{\min}(X)$  and  $\lambda_{\max}(X)$  denote its smallest and largest eigenvalues, respectively. We use  $I_{n \times n}$  to denote a  $n$ -dimensional identity matrix. We use  $0_n$  to denote an  $n$ -dimensional vector where all elements are 0. Lastly, suppose  $M$  is a positive integer. For an arbitrary squared matrix sequence  $(E_i)_{i=1}^M$ , where each  $E_i$  for  $1 \leq i \leq M$  has the same dimension. For  $1 \leq p_1 \leq M$  and  $1 \leq p_2 \leq M$ , define the product operator as

$$\prod_{j=p_1}^{p_2} E_j := \begin{cases} E_{p_2} E_{p_2-1} \dots E_{p_1} & \text{if } p_1 < p_2 \\ E_{p_2} & \text{if } p_1 = p_2 \\ I & \text{if } p_1 > p_2, \end{cases}$$

For a symmetric matrix  $X \in \mathbb{R}^{n \times n}$  we note that  $\|X\| = |\lambda_{\max}(X)|$ .

## 3.2 Problem Formulation

Consider an *expert* controlling a discrete-time linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t, \quad (3.1)$$

with state  $x_t \in \mathbb{R}^n$ , input  $u \in \mathbb{R}^m$  and discrete time steps  $t \in \{0, 1, \dots, T-1\}$  for  $T \in \mathbb{N}$ . The system matrices  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$  are known to the expert, and the initial state satisfies  $x_0 = \bar{x}_0 \in \mathbb{R}^n$ .

Similarly, consider a *copycat* controlling the LTI system

$$\xi_{t+1} = A\xi_t + B\nu_t \quad (3.2)$$

with the same system matrices  $A$  and  $B$ , and initial state  $\xi_0 = \bar{x}_0$ , as (3.1), but with (different) states  $\xi_t \in \mathbb{R}^n$  and inputs  $\nu_t \in \mathbb{R}^m$ .

Throughout this chapter, we rely on the following controllability assumption.

**Assumption 3.2.1.** *The matrix pair  $(A, B)$  defined in (3.1) and (3.2), respectively, is stabilisable.*

To introduce the concept of optimality used in this chapter, we consider running

costs

$$Q : \mathbb{R}^p \rightarrow \mathbb{R}^{n \times n}, \quad R : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times m}, \quad (3.3)$$

characterized through a parameter  $\theta \in \mathbb{R}^p$ , i.e.,  $\theta \mapsto Q(\theta)$  and  $\theta \mapsto R(\theta)$ . For a sequence of inputs  $u_t \in \mathbb{R}^m$ ,  $t \in \{0, \dots, T-1\}$ , and an initial condition  $\bar{x}_0 \in \mathbb{R}^n$ , we define the costs incurred over the discrete-time window  $T$  as

$$J_T(\bar{x}_0, \{u_t\}_{t=0}^{T-1}, \theta) = \sum_{t=0}^{T-1} x_{t+1}^\top Q(\theta) x_{t+1} + u_t^\top R(\theta) u_t \quad (3.4)$$

s.t.  $x_0 = \bar{x}_0, \quad x_{t+1} = Ax_t + Bu_t.$

Based on the definition of the cost functions (3.4), minimal costs and a corresponding optimal input sequence are defined as

$$J_T^*(\bar{x}_0, \theta) = \min_{\{u_t\}_{t=0}^{T-1}} J_T(\bar{x}_0, \{u_t\}_{t=0}^{T-1}, \theta), \quad (3.5)$$

$$\{u_t^*(\bar{x}_0, \theta)\}_{t=0}^{T-1} = \operatorname{argmin}_{\{u_t\}_{t=0}^{T-1}} J_T(\bar{x}_0, \{u_t\}_{t=0}^{T-1}, \theta). \quad (3.6)$$

To ensure that (3.5) and (3.6) are well-defined we impose the following assumptions on functions  $Q(\cdot)$  and  $R(\cdot)$  in (3.3).

**Assumption 3.2.2.** *There exist Lipschitz continuous functions  $D^Q : \mathbb{R}^p \rightarrow \mathbb{R}^{n \times n}$  and  $D^R : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times m}$  with Lipschitz constant  $L_Q, L_R \in \mathbb{R}_{>0}$ , and positive parameters  $\varepsilon_Q, \varepsilon_R \in \mathbb{R}_{>0}$  such that the running costs in (3.3) can be written as*

$$Q(\theta) = D^Q(\theta)^\top D^Q(\theta) + \varepsilon_Q I_{n \times n}, \quad (3.7)$$

$$R(\theta) = D^R(\theta)^\top D^R(\theta) + \varepsilon_R I_{m \times m}, \quad (3.8)$$

where  $\varepsilon_Q$  and  $\varepsilon_R$  are positive scalars.

Lipschitz continuity of  $D^Q$  and  $D^R$  implies that the inequalities

$$\|D^Q(\theta_0) - D^Q(\theta_1)\| \leq L_Q \|\theta_0 - \theta_1\|, \quad (3.9)$$

$$\|D^R(\theta_0) - D^R(\theta_1)\| \leq L_R \|\theta_0 - \theta_1\|, \quad (3.10)$$

are satisfied for any  $\theta_0, \theta_1 \in \mathbb{R}^p$ . Since  $D^Q(\cdot)^\top D^Q(\cdot)$  and  $D^R(\cdot)^\top D^R(\cdot)$  are positive semi-definite, (3.7) and (3.8) ensures that  $Q(\cdot)$  and  $R(\cdot)$  are positive definite and bounded away from zero.

Instead of open-loop input sequences  $\{u_t\}_{t=0}^{T-1}$ , we are interested in time- and parameter-dependent feedback policies

$$\pi : \{0, \dots, T-1\} \times \mathbb{R}^{n+p} \rightarrow \mathbb{R}^m, \quad (t, x, \theta) \mapsto \pi_t(x, \theta). \quad (3.11)$$

and we note that based on the assumption on the system dynamics (3.1) and Assumption 3.2.2, there exists an optimal feedback policy such that

$$\pi_t^*(x_t, \theta) = u_t^*(\bar{x}_0, \theta), \quad t \in \{0, \dots, T-1\}. \quad (3.12)$$

In the following we will thus use the notation  $\pi_t^*(\cdot, \theta)$  and  $u_t^*(\cdot, \theta)$  interchangeably.

While the *expert* can solve (3.5) and (3.6) for arbitrary parameters  $\theta^* \in \mathbb{R}^p$  to obtain a corresponding feedback policy  $\pi_t^*(\cdot, \theta^*)$ , by assumption, the *copycat* is not aware of the *expert's* parameter selection, but intends to recover  $\theta^*$  and the performance of the feedback policy  $\pi_t^*(\cdot, \theta^*)$  based on output measurements

$$y_t = S_t \begin{bmatrix} x_t \\ u_t \end{bmatrix} + v_t, \quad (3.13)$$

where  $S_t \in \mathbb{R}^{q \times (n+m)}$  with  $q \in \{1, \dots, n+m\}$ , and  $v_t$  is unknown (potentially stochastic) noise.

In particular, at time  $t \in \{0, \dots, T-1\}$ , based on the observed measurements  $\{y_\tau\}_{\tau=0}^t$ , the *copycat* intends to infer the optimal parameter  $\theta^*$ , and thus the optimal policy used by the *expert* to define its feedback policy  $\pi_t^*(\cdot, \theta^*)$  used to generate the measurements (3.13). The *copycat* additionally relies on the assumption that it knows the *expert's* dynamics (3.1) through (3.2).

**Remark 3.2.1.** *An additional valid interpretation of the problem under consideration is, that the expert also does not know the optimal parameter  $\theta^*$ , but only knows a feedback policy  $\pi_t^*(\cdot, \theta^*)$  implicitly characterized through  $\theta^*$ . In this interpretation, the copycat intends to infer the parameter  $\theta^* \in \mathbb{R}^p$  to understand with respect to which running costs (3.3) the expert's feedback policy is optimal.*

Mirroring the definition (3.11), we denote the *copycat's* policy, which will be designed in the following sections, by

$$\kappa : \{0, \dots, T-1\} \times \mathbb{R}^{n+p} \rightarrow \mathbb{R}^m, \quad (t, \xi, \hat{\theta}) \mapsto \kappa_t(\xi, \hat{\theta}). \quad (3.14)$$

Here, we use  $\kappa$ ,  $\xi$  and  $\hat{\theta}$  to indicate a policy corresponding to the *copycat*, depending

on an estimated parameter  $\hat{\theta} \in \mathbb{R}^p$ . To be able to compare the performance of an input trajectory  $\{\nu_t\}_{t=0}^{T-1}$ , implemented by the *copycat*, with the optimal feedback policy  $\pi^*(\cdot, \theta^*)$ , we define the regret

$$\text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) := J_T(\bar{x}_0, \{\nu_t\}_{t=0}^{T-1}, \theta^*) - J_T^*(\bar{x}_0, \theta^*) \quad (3.15)$$

where  $J_T(\bar{x}_0, \{\nu_t\}_{t=0}^{T-1}, \theta^*)$  denotes the costs in (3.4) incurred through the input sequence  $\{\nu_t\}_{t=0}^{T-1}$ .

The overall algorithm to iteratively define the closed-loop input sequence  $\{\nu_t^{\text{cl}}\}_{t=0}^{T-1}$  implemented by the *copycat* is summarised and illustrated in Algorithm 3.1. In

---

**Algorithm 3.1** Expert-Copycat Online LQ Optimal Control

---

- 1: **Expert Initialization:**  $x_0 = \bar{x}_0$
  - 2: **Copycat Initialization:**  $\xi_0 = \bar{x}_0, \hat{\theta}_{-1} = \bar{\theta}$
  - 3: **for**  $t = 0$  **to**  $T - 1$  **do**
  - 4:   **Expert Selects Inputs:**
  - 5:      $u_t = \pi_t^*(x_t, \theta^*)$
  - 6:   **Copycat:**
  - 7:     Receives Measurements  $y_t = S_t \begin{bmatrix} x_t \\ u_t \end{bmatrix} + v_t$
  - 8:     Updates estimate:
- $$\hat{\theta}_t = f_t(\hat{\theta}_{t-1}, \{y_\tau\}_{\tau=0}^t)$$
- 9:     Selects input:  $\nu_t^{\text{cl}} = \kappa_t(\xi_t, \hat{\theta}_t)$
  - 10:   **Expert and Copycat Apply Inputs:**
  - 11:      $x_{t+1} = Ax_t + Bu_t$
  - 12:      $\xi_{t+1} = A\xi_t + B\nu_t^{\text{cl}}$
  - 13: **end for**
- 

Section 3.3, we focus on the definition of the function  $f_t(\cdot)$  in line 8 (which updates the parameter estimate  $\hat{\theta}_t$ ) and on the definition of the feedback policy  $\kappa_t(\cdot)$  in line 9. Based on the definition of  $\{\nu_t^{\text{cl}}\}_{t=0}^{T-1}$ , we analyse the regret

$$\text{Regret}_T(\{\nu_t^{\text{cl}}\}_{t=0}^{T-1}), \quad (3.16)$$

in Section 3.4, to derive upper bounds in terms of closed-loop performance estimates of the *copycat* controller.

### 3.3 Implementation of Algorithm 3.1

To be able to implement Algorithm 3.1 the function  $f_t(\cdot)$  in line 8 and the definition

$\kappa_t(\cdot)$  in line 9 needs to be made precise. Here, we rely on a generic observer design to define  $f_t(\cdot)$  and the definition of  $\kappa_t(\cdot)$  relies on the idea of a tracking controller.

### 3.3.1 Parameter Estimation $\hat{\theta}$

For the parameter update, we use a generic estimator of the form

$$\hat{\theta}_t = \hat{\theta}_{t-1} + L_t \left( y_t - S_t \begin{bmatrix} x_t(\hat{\theta}_t) \\ u_t(\hat{\theta}_t) \end{bmatrix} \right), \quad (3.17)$$

where  $\{u_\tau(\hat{\theta}_t)\}_{\tau=0}^{T-1} = \arg \min_{\{\tilde{u}_\tau\}_{\tau=0}^{T-1}} J(\bar{x}_0, \{\tilde{u}_\tau\}_{\tau=0}^{T-1}, \hat{\theta}_t)$ , state sequence  $\{x_\tau(\hat{\theta}_t)\}_{\tau=0}^{T-1}$  is generated by  $\{u_\tau(\hat{\theta}_t)\}_{\tau=0}^{T-1}$  from dynamic (3.1) and  $\{L_t\}_{t=0}^{T-1}$  defines observer gains specific to the estimator selection. While we do not insist on a specific observer gain selection, we rely on the following assumption.

**Assumption 3.3.1.** *For a given observer gain selection  $\{L_t\}_{t=0}^{T-1}$  and for a bounded disturbance sequence  $\{v_t\}_{t=0}^{T-1}$ ,  $\|v_t\| \leq v_\theta \in \mathbb{R}_{\geq 0}$ , there exist scalars  $C_\theta, r \in \mathbb{R}_{>0}$ ,  $\eta_\theta \in (0, 1)$  and  $\alpha \in \mathcal{K}$  such that the estimator of the form (3.17) satisfies <sup>1</sup>*

$$\|\hat{\theta}_t - \theta^*\| \leq C_\theta \|\hat{\theta}_0 - \theta^*\| \eta_\theta^t + \alpha(\|v_\theta\|), \quad \forall t \in \{0, \dots, T\} \quad (3.18)$$

for all  $\|\hat{\theta}_0 - \theta^*\| \leq r$ .

Under additional technical regularity assumptions on (3.5) and (3.6) as functions of  $\theta$ , a particular method with such estimation form satisfying Assumption 3.3.1 is the extended Kalman filter (EKF) [64, 65]. The estimate  $\hat{\theta}_t$  always lies in a compact set for  $\forall t \in \{0, 1, \dots, T-1\}$ . The EKF is used for the numerical experiments in Section 3.5.

### 3.3.2 Definition of the Policy $\kappa_t(\cdot)$

The feedback policy  $\kappa_t(\cdot)$  at time  $t \in \{0, \dots, T-1\}$  used by the copycat in line 9 of Algorithm 3.1 is comprised of two components: a *prediction* component and a *tracking* component.

For the prediction component at time  $t$  we consider  $\{\hat{\nu}_{\tau|t}^*(\bar{x}_0, \hat{\theta}_t)\}_{\tau=0}^{T-1}$  defined as the solution of the optimal control problem

$$\{\hat{\nu}_{\tau|t}^*(\bar{x}_0, \hat{\theta}_t)\}_{\tau=0}^{T-1} = \underset{\{\nu_\tau\}_{\tau=0}^{T-1}}{\operatorname{argmin}} J(\bar{x}_0, \{\nu_\tau\}_{\tau=0}^{T-1}, \hat{\theta}_t), \quad (3.19)$$

<sup>1</sup>A continuous function  $f : [0, a) \rightarrow [0, \infty)$  for some positive scalar  $a$  is said to belong to class  $\mathcal{K}$  if it is strictly increasing and  $f(0) = 0$ .

based on the current estimate  $\hat{\theta}_t$ . As highlighted before, existence and uniqueness of the optimal solution is guaranteed through the properties of  $Q(\cdot)$  and  $R(\cdot)$  and Assumption 3.2.2. Using the predicted input  $\{\hat{\nu}_{\tau|t}^*(\bar{x}_0, \hat{\theta}_t)\}_{\tau=0}^{T-1}$  the predicted state trajectory  $\{\hat{\xi}_{\tau|t}\}_{\tau=0}^{T-1}$  is defined through  $\hat{\xi}_{0|t} = \bar{x}_0$  and

$$\hat{\xi}_{\tau+1|t} = A\hat{\xi}_{\tau|t} + B\hat{\nu}_{\tau|t}^*, \quad \forall \tau \in \{0, \dots, T-1\} \quad (3.20)$$

based on the dynamics (3.2).

For the tracking component we define  $\Gamma \in \mathbb{R}^{m \times n}$  such that

$$A_{\text{cl}} = A + B\Gamma \quad (3.21)$$

is a Schur matrix and which implies the following corollary.

**Corollary 1.** *Let  $A_{\text{cl}} \in \mathbb{R}^{n \times n}$  be a Schur matrix. Then there exist  $C_q \in \mathbb{R}_{\geq 0}$  and  $\eta_q \in (0, 1)$  such that*

$$\|A_{\text{cl}}^t\| \leq C_q \eta_q^t \quad (3.22)$$

for all  $t \in \mathbb{N}$ .

*Proof.* According to [66, Thm. 5.8],  $A_{\text{cl}}$  Schur implies that there exists a positive definite matrix  $P$  such that

$$A_{\text{cl}}^\top P A_{\text{cl}} = P - I = (I - P^{-1})P \preceq \left(1 - \frac{1}{\lambda_{\max}(P)}\right) P \quad (3.23)$$

and where  $P_1 \leq P_2$  means that the matrix  $P_2 - P_1$  positive semi-definite. Using the estimate (3.23) iteratively, implies that

$$\lambda_{\min}(P) \|A_{\text{cl}}^t\|^2 \leq (A_{\text{cl}}^\top)^t P A_{\text{cl}}^t \leq \left(1 - \frac{1}{\lambda_{\max}(P)}\right)^t P \leq \lambda_{\max}(P) \left(1 - \frac{1}{\lambda_{\max}(P)}\right)^t$$

for all  $t \in \mathbb{N}$  and thus

$$\|A_{\text{cl}}^t\|^2 \leq \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \left(1 - \frac{1}{\lambda_{\max}(P)}\right)^t$$

completing the proof for  $C_q = \sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}$  and  $\eta_q = \sqrt{1 - \frac{1}{\lambda_{\max}(P)}}$ .  $\square$

Based on the definition of the prediction and the tracking controller components the overall copycat controller  $\kappa_t(\cdot)$  is defined as

$$\kappa_t(\xi_t, \hat{\theta}_t) = \Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{\nu}_{t|t}. \quad (3.24)$$

The definition of  $\kappa_t(\cdot)$  presented here, follows the ideas in [67] and in particular the definition [67, Eq. (14)].

Note that  $\Gamma \in \mathbb{R}^{n \times m}$  can be designed offline while  $\hat{\xi}_{t|t}$  and  $\hat{\nu}_{t|t}$  need to be updated online by solving (3.19) at every time step.

### 3.4 Regret Analysis Under Cost Matrices Estimation

In this section, we present an upper bound on the regret incurred under the control policy (3.24). We first define key operators and quantities that characterise the regret under this policy.

We begin by introducing the algebraic Riccati operator  $F_\tau : \mathbb{R}^p \rightarrow \mathbb{R}^{n \times n}$  and the associated control gain  $K_\tau : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times n}$ , both of which are standard in LQR formulations and will be used to express both the optimal and estimated control laws corresponding to (3.6).

For  $\tau \in \{0, \dots, T-1\}$ , we define

$$F_T(\theta) := Q(\theta) \tag{3.25}$$

$$F_\tau(\theta) := Q(\theta) + A^\top (F_{\tau+1}(\theta)^{-1} + BR(\theta)^{-1}B^\top)^{-1}A,$$

$$K_\tau(\theta) := (R(\theta) + B^\top F_{\tau+1}(\theta)B)^{-1}B^\top F_{\tau+1}(\theta)A, \tag{3.26}$$

where  $Q(\theta)$  and  $R(\theta)$  are defined in (3.7) and (3.8), respectively. Since  $Q(\theta)$  and  $R(\theta)$  are positive definite by definition,  $F_t$  are well-defined and positive definite for all  $\tau \in \{0, \dots, T-1\}$ . With the definition of  $K_\tau(\theta)$ , we can state the following result, providing a representation of the optimal input in terms of the feedback gains  $K_\tau(\theta)$ .

**Proposition 1.** *Under Assumptions 3.2.1 and 3.2.2, for  $\theta \in \mathbb{R}^p$  and  $\bar{x}_0 \in \mathbb{R}^n$  consider the optimal control problem (3.6). Then, with the definition (3.26), it holds that*

$$u_t^*(\bar{x}_0, \theta) = K_t(\theta)x_t, \quad x_{t+1} = \prod_{\tau=0}^t (A + BK_\tau(\theta))\bar{x}_0, \tag{3.27}$$

for all  $t \in \mathbb{N}_{T-1}$ .

*Proof.* The optimal control solution  $u_t^*(\bar{x}_0, \theta)$  can follow directly from [18, Chapter 2.4] by replacing  $Q_t$  and  $R_t$  with  $Q(\theta)$  and  $R(\theta)$ , respectively. This implies that

$x_{t+1} = (A + BK_t(\theta))x_t$ . Expanding this directly yields  $x_{t+1} = \prod_{\tau=0}^t (A + BK_\tau(\theta))\bar{x}_0$ .  $\square$

Using estimation  $\hat{\theta}_t$ , for  $\tau \in \{0, \dots, T-1\}$ , we define the plug-in estimated control gain  $\hat{K}_{\tau|t}$  and cost-to-go matrix  $\hat{P}_{\tau|t}$  as

$$\hat{K}_{\tau|t} := K_\tau(\hat{\theta}_t), \quad \hat{P}_{\tau|t} := F_\tau(\hat{\theta}_t). \quad (3.28)$$

In addition, for the optimal parameters  $\theta^*$ , we use the notation

$$K_t^* := K_t(\theta^*), \quad t \in \mathbb{N}_{T-1}. \quad (3.29)$$

To quantify the estimation error, for  $\tau \in \{0, \dots, T-1\}$  and  $t_1, t_2 \in \{0, \dots, T-1\}$  we introduce the following terms

$$\hat{\Phi}_{\tau|(t_1, t_2)} := \|\hat{K}_{\tau|t_1} - \hat{K}_{\tau|t_2}\|, \quad (3.30)$$

$$\bar{\Phi}_{\tau|t_1}^* := \|\hat{K}_{\tau|t_1} - K_\tau^*\|, \quad (3.31)$$

These quantities measure the variability in estimated control gains across time and the deviation from the optimal control gain, respectively.

We also need several constants that are related to eigenvalue and matrix norm bounds on the cost and cost-to-go matrices,

$$Q^* := Q(\theta^*), R^* := R(\theta^*), \hat{Q}_t := Q(\hat{\theta}_t), \quad (3.32a)$$

$$\hat{R}_t := R(\hat{\theta}_t), P_t^* := F_t(Q^*, R^*), \quad (3.32b)$$

$$\lambda_{\min}^Q := \min \lambda_{\min}(Q^*), \lambda_{\max}^P := \max_{\tau \in \mathbb{N}_{T-1}} \lambda_{\max}(P_\tau^*), \quad (3.32c)$$

$$\hat{\lambda}_{\min}^Q := \min_{t \in \mathbb{N}_{T-1}} \lambda_{\min}(\hat{Q}_t), \hat{\lambda}_{\max}^P := \max_{\tau, t \in \mathbb{N}_{T-1}} \lambda_{\max}(\hat{P}_{\tau|t}), \quad (3.32d)$$

$$D := \|R^*\| + \|B\|^2 \lambda_{\max}^P, \quad (3.32e)$$

$$\beta := \max_{\tau \in \mathbb{N}_{T-1}} \lambda_{\max}(A^\top P_\tau^* A), \hat{\beta} := \max_{\tau, t \in \mathbb{N}_{T-1}} \lambda_{\max}(A^\top \hat{P}_{\tau|t} A)$$

$$\gamma := \frac{\beta}{\beta + \lambda_{\min}^Q}, \hat{\gamma} := \frac{\hat{\beta}}{\hat{\beta} + \hat{\lambda}_{\min}^Q}, \quad (3.32f)$$

$$C := \sqrt{\frac{\lambda_{\max}^P}{\lambda_{\min}^Q}}, \hat{C} := \sqrt{\frac{\hat{\lambda}_{\max}^P}{\hat{\lambda}_{\min}^Q}} \quad (3.32g)$$

$$\eta := \sqrt{1 - \frac{1}{C}}, \hat{\eta} := \sqrt{1 - \frac{1}{\hat{C}}}, \quad (3.32h)$$

$$\Delta := \max_{t \in \mathbb{N}_{T-1}} \|K_t^* - \Gamma\|. \quad (3.32i)$$

Note that the matrices  $Q^*, P_\tau^*, \hat{Q}_{\tau|t}$  and  $\hat{P}_{\tau|t}$  are positive definite, implying  $\lambda_{\min}^Q, \lambda_{\max}^P, \hat{\lambda}_{\min}^Q$  and  $\hat{\lambda}_{\max}^P$  positive, therefore  $C, \hat{C} \in \mathbb{R}_{\geq 0}$ . Moreover,  $C \geq 1$  and  $\hat{C} \geq 1$ , this implies that  $1 - \frac{1}{C}$  and  $1 - \frac{1}{\hat{C}}$  are non-negative, which implies that  $\eta, \hat{\eta} \in \mathbb{R}_{\geq 0}$ .

We are now ready to present a key lemma that provides the relationship between the regret incurred by policy (3.24) to the estimation errors captured by  $\hat{\Phi}_{t|(p,q)}$  and  $\bar{\Phi}_{\tau|t}^*$ .

**Lemma 3.4.1.** *Let  $T \geq 1$  and consider the copycat system defined in (3.2) together with the control law (3.24) relying on the sensor measurements  $\{y_\tau\}_{\tau=0}^t$  in line 8, Algorithm 3.1. Moreover, assume that  $\Gamma \in \mathbb{R}^{n \times m}$  is selected such that  $A + B\Gamma$  is a Schur matrix and let Assumptions 3.2.2 and 3.3.1 be satisfied. Then the regret defined in (3.15) satisfies*

$$\text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) \leq 2D \|\bar{x}_0\|^2 \sum_{t=1}^{T-1} (\hat{C} \hat{\eta}^t \bar{\Phi}_{t|t}^*)^2 + (\Delta C_q \hat{C}^2 \|B\| \eta_q^{t-1} \sum_{j=1n=0}^t \sum_{j=1}^{j-1} \left(\frac{\hat{\eta}}{\eta_q}\right)^j \hat{\Phi}_{n|(j-1,j)})^2 \quad (3.33)$$

and where the individual components of the bound are defined in (3.32).

The bound in Lemma 3.4.1 characterizes the regret incurred by (3.24) in terms of

how accurately the optimal gains  $K_t(\theta^*)$ ,  $t \in \mathbb{N}_{T-1}$ , are estimated over time. The first term in the summation depends on the distance from the estimated control gain  $\hat{K}_{t|t}$ ,  $t \in \mathbb{N}_{T-1}$  to the optimal gains at each time step, while the second term in the summation depends on the difference of estimated gains defined in (3.26).

Building on Lemma 3.4.1, we now derive an explicit regret bound by using the convergence property from Corollary 1. In particular, we substitute upper bounds on  $\bar{\Phi}_{t|t}^*$  and  $\hat{\Phi}_{n|(j-1,j)}$  into the right-hand side of inequality (3.33), yielding a regret bound that depends on the previous stated constants and system parameters.

Before we present the main theorem, we need additional constants related to the estimation method, define in the following.

Under Assumption 3.3.1, it follows directly that  $\|\hat{\theta}_t\| \leq \|\theta^*\| + C_{\theta r} \eta_{\theta}^t + \alpha(\|v_{\theta}\|)$  for appropriately selected  $\hat{\theta}_0 \in \mathbb{R}^p$ . With the definition

$$\mathcal{D}_{\mathcal{O}} = \mathcal{B}_{C_{\theta r} + \alpha(\|v_{\theta}\|)}(\theta^*) \quad (3.34)$$

In addition, since  $\|D^Q(\cdot)\|$  and  $\|D^R(\cdot)\|$  defined in (3.7) and (3.8) are continuous, they attain their maximum in the compact set  $\mathcal{D}_{\mathcal{O}}$ , i.e.,

$$M_R := \max_{\theta \in \mathcal{D}_{\mathcal{O}}} \|D^R(\theta)\| \text{ and } M_Q := \max_{\theta \in \mathcal{D}_{\mathcal{O}}} \|D^Q(\theta)\| \quad (3.35)$$

are well-defined.

Our main result is stated as follows.

**Theorem 3.4.1.** *Adopt the hypothesis from Lemma 3.4.1, the regret satisfies*

$$\begin{aligned} & \text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) \\ & \leq 4D \left[ \hat{C}^2 C_{\phi^*}^2 ([C_{\theta} \|\hat{\theta}_0 - \theta^*\|]^2 E_1((\hat{\eta}\eta_{\theta})^2) + [\alpha(\|v_{\theta}\|)]^2 E_2(\hat{\eta}^2)) + (\Delta C_q \|B\| \hat{C}^2 C_{\hat{\phi}} \frac{\hat{\eta}}{\eta_q})^2 \right. \\ & \quad \left( \left( \frac{C_{\theta} \|\hat{\theta}_0 - \theta^*\|}{\eta_q - \hat{\eta}\eta_{\theta}} \right)^2 [E_1(\hat{\eta}\eta_{\theta}) + \eta_q (E_2(\hat{\eta}_q) - E_2(\hat{\eta}\eta_{\theta}))]^2 \right. \\ & \quad \left. \left. + \left( \frac{\alpha(\|v_{\theta}\|)}{\eta_q - \eta} \right)^2 [E_1(\hat{\eta}) + \eta_q (E_2(\hat{\eta}_q) - E_2(\hat{\eta}))]^2 \right) \right], \quad (3.36) \end{aligned}$$

where both  $\bar{C}$  and  $C_{\hat{\phi}}$  depend on  $M_Q, M_R$  and  $\|B\|$ , with  $\bar{C}$  and  $C_{\hat{\phi}}$  additionally depending on  $\lambda_{\min}(Q^*)$  and  $\min_{t \in \mathbb{N}_{T-1}} \lambda_{\min}(\hat{Q}_t)$ , respectively. The coefficients  $C_{\theta}$  and  $\eta_{\theta}$ , the vector  $v_{\theta}$ , and the function  $\alpha(\cdot)$  are defined in Assumption 3.3.1. The functions  $E_1(\cdot)$  and  $E_2(\cdot)$  are given by  $E_1(z) := \sum_{t=0}^{T-1} tz^t$  and  $E_2(z) := \sum_{t=0}^{T-1} z^t$  for  $z \in \mathbb{R}$ . All remaining coefficients are taken from the list in 3.32.

**Remark 3.4.1.** *The regret upper bound from the above theorem quantifies the relation between regret and the estimation error of cost matrices with EKF. From the right-hand side of (3.33) and the associated term defined under the inequality, the regret upper bound monotonically increases with respect to  $C_\theta$  and  $v_\theta$ . Thus, to potentially lower the regret via decreasing the regret upper bound, one way is through designing the EKF with small  $C_\theta$  and  $v_\theta$ . In particular, for the case of  $v_t = 0$  where  $v_t$  is the observation noise from the sensor, with the initial  $\hat{\theta}_0$  chosen sufficiently close to the true parameter  $\theta^*$ , we have  $v_\theta = 0$ . Further,  $C_\theta$  depends on the choice of filter parameters [65]. In addition, the regret bound can also be reduced by appropriately choosing the sequence of  $K_t$  to yield optimal  $C_q$  and  $\eta_q$  that reduces the right-hand side of (3.33).*

**Remark 3.4.2.** *Suppose  $0 < z < 1$ , then  $E_1(z) < \frac{1}{1-z^2}$  and  $E_2(z) < \frac{1}{1-z}$ . This implies that the regret upper bounded (3.36) is independent of the time horizon  $T$ . For fixed  $Q^*$  and  $R^*$ , it is expected that the regret grows up when  $T$  increases. Therefore, herustically, we will observe the regret grows monotonically then saturates at a level when  $T$  is sufficiently large.*

Next, we illustrate the (empirical) regret for our proposed methods.

### 3.5 Numerical Simulations

In this section, we numerically evaluate the regret incurred by the copycat under the proposed policy. We consider system matrices  $A = \begin{bmatrix} 0.7 & 0.2 \\ -0.1 & 0.8 \end{bmatrix}$  and  $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$  for both the expert and the copycat, with the initial condition being  $\bar{x}_0 = [100, 100]$ . The cost parameter for the expert is  $\theta^* = [5, 5]^\top$  and the cost matrices are computed by  $D^Q = \text{diag}([\theta^*]_1, [\theta^*]_2)$ ,  $D^R = 0$ ,  $\varepsilon_Q = 0.01$  and  $\varepsilon_R = 10$ , where  $[\cdot]_i$  denotes the  $i$ -th element of a vector, and  $\text{diag}(\cdot)$  returns a diagonal matrix from its arguments, with all the off-diagonal elements set to 0. For  $t \in \{0, 1, \dots, T-1\}$ , at each time step  $t$ , the sensor value for the copycat is given by  $y_t = \begin{bmatrix} I_{2 \times 2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}$ .

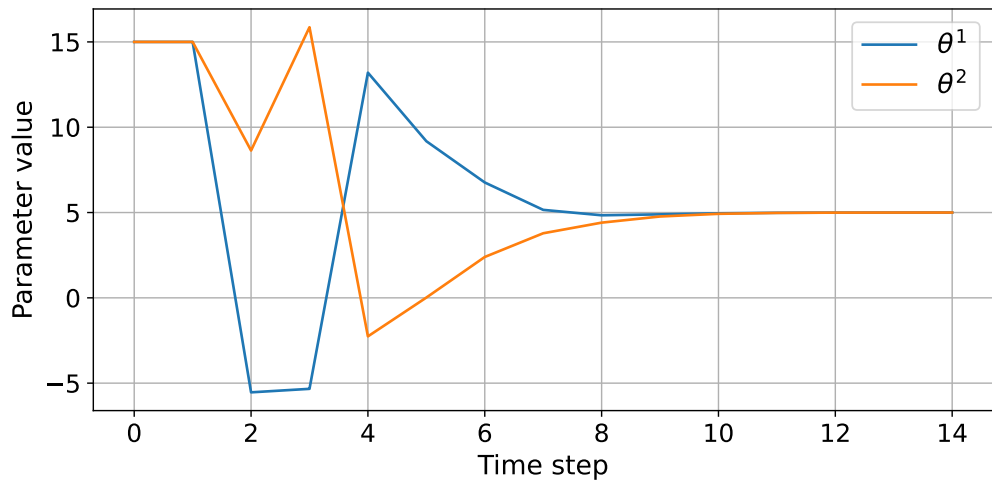
We demonstrate the performance of the copycat by illustrating the (i) convergence of  $\hat{\theta}_t$  to  $\theta^*$ ; (ii) comparison of trajectories generated by the expert and the copycat, and (iii) regret vs.  $T$ .

Figure 3.1 reports the estimated parameter value  $\hat{\theta}_t$  from the copycat for the case of  $T = 15$ . The plot indicates that under EKF, the value  $\hat{\theta}_t$  converges to the true value at time step  $t = 8$ .

Figure 3.2 reports the trajectories generated by the expert and copycat for  $T = 15$ .

The state and control trajectories are closely aligned after time step  $t = 4$ . This demonstrates that the copycat’s control actions closely follow those of the expert over time.

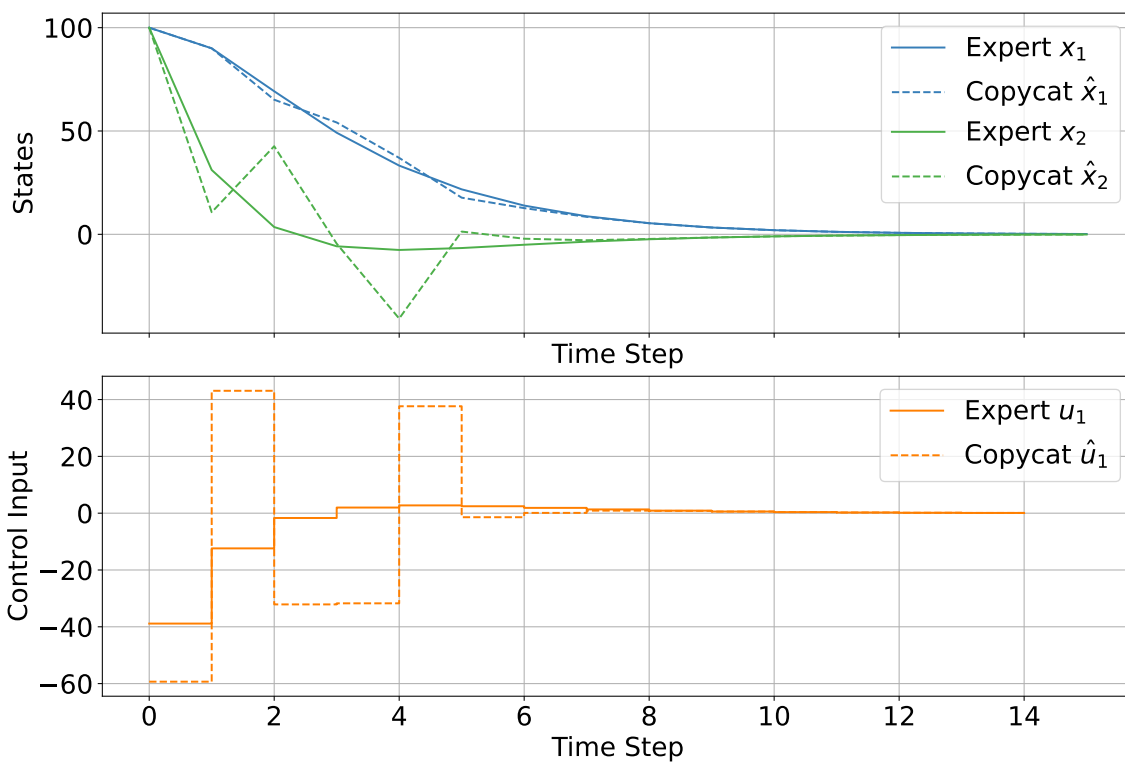
Figure 3.3 reports  $\frac{\text{Regret}_T(\{\nu_t\}_{t=0}^{T-1})}{T}$  incurred by the copycat under different time-horizon  $T$ . The regret saturates at a constant when  $T = 15$ . The result suggests that the regret is sublinear to  $T$ , which aligns with the regret bound in Theorem 3.4.1.



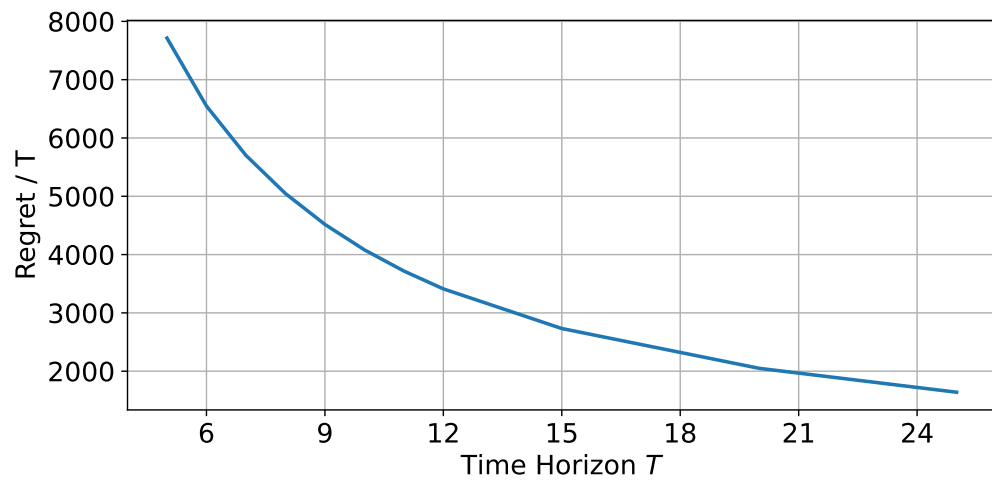
**Figure 3.1:** Copycat’s estimated parameters  $\hat{\theta}_t$  over time

## 3.6 Summary

This chapter introduces a new online LQ optimal control problem with sequentially inferred costs. We established the regret bounds associated with an EKF-based estimator. The numerical results demonstrate the effectiveness of our proposed estimation and control algorithm and show that the regret is sublinear with respect to the time horizon.



**Figure 3.2:** Comparison of expert and copycat trajectories. (Left) State trajectories  $x_1$  and  $x_2$  for both expert and copycat over time steps. (Right) Control input trajectories  $u_1$  for the expert and the copycat.



**Figure 3.3:**  $\frac{\text{Regret}_T(\{\nu_t\}_{t=0}^T)}{T}$  of the copycat with respect to different time horizons  $T$ .

---

# Chapter 4. $N$ -player Dynamic Potential LQ Games with Sequentially Revealed Costs

---

In this chapter, we introduce the dynamic potential LQ game problem with sequentially revealed costs. We begin by reviewing the related works for this problem. We then apply the framework proposed in Chapter 2 for this multi-agent setting. We extend the notion of dynamic regret from the single-player case to the multi-player case as *price of uncertainty* (PoU), then provide the PoU analysis of the framework. Finally, we demonstrate the empirical performance of our proposed methods through numerical simulations, validating the theoretical guarantee from the PoU analysis.

## 4.1 Related Works

Works concerning similar problems include repeated games with sequentially revealed costs [19], equilibrium tracking via online state measurement [68], and cost parameters estimation via sequentially revealed state measurement [20]. However, none of these works have studied the problem of a noncooperative LQ dynamic potential game with sequentially revealed costs. The only study that considers a similar setting as ours is [69, Chapter 2 and 3], where they investigate a linear-quadratic(LQ) pursuit evasion game in which each player aims to play at the feedback Nash equilibrium via minimising their own cost function despite lacking knowledge of their opponent's cost matrices. At each time  $t$ , the players aim to act at the feedback Nash equilibrium by estimating the cost-to-go matrices ( $Q_t^i$  and  $R_t^i$  from the Algebraic Riccati equation) for  $\tau \geq t$ , using state information potentially corrupted with measurement noise. However, this work does not consider using any notion that captures the performance quality in the absence of cost information from their

opponents.

## 4.2 Problem Formulation

This chapter is based on our preliminary work [70] for the two-player case. The key contributions of this chapter are:

1. The formulation of a novel  $N$ -player LQ dynamic feedback potential game problem with sequentially revealed costs and a price of uncertainty (PoU) performance measure;
2. The proposal of an algorithm that enables all players to predict and track a feedback Nash Equilibrium in an  $N$ -player dynamic potential game with sequentially revealed costs;
3. The lower and upper bounds on the PoU achieved by our proposed algorithm.
4. The relationship between PoU and price of anarchy (PoA), and how close the policy is to the feedback Nash equilibrium if it has a bounded PoU.
5. Numerical simulation on community battery scheduling problem based on [47]. Results show that the absolute value of PoU initially decays exponentially as the preview window increases, then remains constant for large time horizons, which matches our theoretical lower and upper bounds.

This chapter is structured as follows. In Section 4.2.1, we formulate the problem of  $N$ -player LQ dynamic feedback potential game with sequentially revealed costs. In Section 4.3, we present our proposed algorithm and lower and upper bounds on its price of uncertainty. In Section 4.5, we illustrate the performance of the proposed algorithm applied to a three-player community battery scheduling problem. We present concluding remarks and future directions in Section 4.6.

**Notation** For a matrix  $\mathbf{H}$  with  $N$  block-rows and  $N$  block-columns, i.e.,  $\mathbf{H} =$

$$\begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1N} \\ H_{21} & H_{22} & \cdots & H_{2N} \\ \vdots & \ddots & \vdots & \vdots \\ H_{N1} & H_{N2} & \cdots & H_{NN} \end{bmatrix},$$

we define the  $ij$ -th block of  $\mathbf{H}$  by  $[\mathbf{H}]_{ij} := H_{ij}$ . We use  $\|\cdot\|$  to denote the 2-norm of a vector or a matrix, depending on its argument. For any symmetric matrices  $F, G \in \mathbb{R}^{n \times n}$ ,  $F \preceq G$  denotes  $G - F$  being positive definite. Let  $\rho(\cdot)$  be the spectral radius operator. We use  $\lambda_{\min}(\cdot)$  and  $\lambda_{\max}(\cdot)$  to designate the minimal and the maximal eigenvalue of a symmetric matrix, respectively. Similarly,  $\sigma_{\min}(\cdot)$  and  $\sigma_{\max}(\cdot)$  denote the minimal and the maximal singular value of a matrix.

We use  $\lambda_{\min}^+(\cdot)$  and  $\sigma_{\min}^+(\cdot)$  to refer to the minimal positive eigenvalue and singular value of the matrix, respectively. In addition, for any symmetric matrices  $F \in \mathbb{R}^{n \times n}$  and  $G \in \mathbb{R}^{n \times m}$ , we consider convention  $\lambda_1(F) \geq \dots \geq \lambda_n(F)$  and  $\sigma_1(G) \geq \dots \geq \sigma_m(G)$ . For an integer  $k \geq 1$ , define  $\mathbb{N}_k$  to be the set  $\{1, 2, \dots, k\}$ . Let  $\mathbb{S}_{++}^n$  denote the set of  $n \times n$  positive definite matrix. Lastly, for a given  $T \geq 1$  and  $i \in \{1, 2, \dots, N\}$ , we consider  $\Pi_t$  being  $(\pi_{i,t})_{i=1}^N$ ,  $(\pi_t^i, \pi_t^{-i*})$  being  $(\pi_t^{1*}, \dots, \pi_t^{i-1*}, \pi_t^i, \pi_t^{i+1*}, \dots, \pi_t^{N*})$ , and we use  $\Pi_{n:m}$  to refer to the sequence of  $\Pi_n, \Pi_{n+1}, \dots, \Pi_m$  when  $1 \leq n \leq m \leq N$ , or  $\Pi_n$  when  $1 \leq m \leq n \leq N$ .

### 4.2.1 $N$ -Player Potential LQ Dynamic Game with Sequentially Revealed Costs

#### LQ dynamic potential game and its feedback Nash equilibrium

Let us first define a (non-cooperative)  $N$ -player LQ dynamic feedback game (LQ-DFG). Consider

$$x_{t+1} = Ax_t + \mathbf{B}\mathbf{u}_t, \quad x_1 = \bar{x}_1, \quad (4.1)$$

where  $t$  is a non-negative integer,  $A \in \mathbb{R}^{n \times n}$  and  $\mathbf{B} = [B^1, \dots, B^N]$ ,  $B^1, \dots, B^N \in \mathbb{R}^{n \times m}$  are *system matrices*,  $x_t \in \mathbb{R}^n$  is the state,  $\mathbf{u}_t = [u_t^{1\top}, \dots, u_t^{N\top}]^\top$  where  $u_t^1, \dots, u_t^N \in \mathbb{R}^m$  are player controls, and  $\bar{x}_1 \in \mathbb{R}^n$  is the initial state. Let  $\Lambda$  denote the set of all measurable maps from  $\mathbb{R}^n$  to  $\mathbb{R}^m$  and  $\Pi_t := (\pi_t^1, \dots, \pi_t^N)$ . Each player  $i$ ,  $i \in \{1, 2, \dots, N\}$ , at time  $t \in \mathbb{N}_{T-1}$ , selects a *feedback control policy*,  $\pi_t^i \in \Lambda$ , such that their controls satisfy  $u_t^i = \pi_t^i(x_t)$ , with the aim of minimising a quadratic cost function

$$J_T^i(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) := \sum_{t=1}^{T-1} x_{t+1}^\top Q_{t+1} x_{t+1} + \mathbf{u}_t^\top R_t^i \mathbf{u}_t \quad (4.2)$$

subject to (4.1) where  $\mathbf{u}_t = [\pi_t^1(x_t)^\top, \dots, \pi_t^N(x_t)^\top]^\top = \Pi_t(x_t)$ , and matrix  $R_t^i$  admits the following block structure:

$$R_t^i := \begin{bmatrix} [R_t^i]_{11} & [R_t^i]_{12} & \cdots & [R_t^i]_{1N} \\ [R_t^i]_{21} & [R_t^i]_{22} & \cdots & [R_t^i]_{2N} \\ \vdots & \ddots & \vdots & \vdots \\ [R_t^i]_{N1} & [R_t^i]_{N2} & \cdots & [R_t^i]_{NN} \end{bmatrix} \in \mathbb{R}^{Nm \times Nm}, \quad (4.3)$$

with  $[R_t^i]_{jk} \in \mathbb{R}^{m \times m}$  for  $j, k \in \{1, 2, \dots, N\}$  for all  $t \in \mathbb{N}_{T-1}$ .

We consider three examples.

**Example 1** (Community Battery). Consider a community battery example between a group of three users motivated by [3]. Let  $x_t := \tilde{x}_t - d_t$  be the state where  $\tilde{x}_t \in \mathbb{R}$  is the state of charge and  $d_t \in \mathbb{R}$  is the desire battery level at time  $t$ . The state dynamics is governed by

$$x_{t+1} = \begin{bmatrix} a & 0 \\ 0 & 0.9 \end{bmatrix} x_t + \begin{bmatrix} -b_1 & -b_2 & -b_3 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{u}_t, \quad (4.4)$$

where  $a \geq 1$  is the constant recharge rate of the battery, and  $u_t^i$  for  $i = \{1, 2, 3\}$  is user  $i$ 's electricity usage at time  $t$ .

The objective for each user is to minimise the total electricity and state of charge cost over a period. Similar to [49] or [3], such costs can be modelled in the quadratic form as (4.2), subject to the battery dynamics in (4.1), where matrix  $Q_t \in \mathbb{S}_{++}^2$  represents the degradation cost due to the battery deviates from the nominal level, and  $R_t^i \in \mathbb{R}^{3 \times 3}$  is the cost of electricity for the  $i$ -th user. This captures the objective of maintaining a desired battery level while minimising the individual energy usage costs. In realistic scenarios for a solar-powered community battery, weather-related external factors will influence the cost matrices  $Q_t$  and  $R_t^i$ , where the future quantities are known over a limited forecast horizon. Mathematically, at time  $t$ , agent  $i$  may have access to  $Q_\tau$  and  $R_\tau^i$  for  $1 \leq \tau \leq t + W$ , where  $W$  denotes the preview window of the future.

**Example 2** (Decentralised Formation Control). Consider a formation control problem considered in [2]. Let  $x_t := \tilde{x}_t - d_t$  be the state where  $\tilde{x}_t \in \mathbb{R}^N$  represent the positions of all  $N$  vehicles in a team at time  $t$ , with  $\tilde{x}_t^i \in \mathbb{R}$  denotes the positioning of the  $i$ -th vehicle, and  $d_t \in \mathbb{R}^N$  denotes the desire formation position at time  $t$  with  $d_t^i \in \mathbb{R}$  be the desire position of the  $i$ -th vehicle. The evolution of the states is governed by the linear relation

$$x_{t+1} = x_t + \mathbf{B} \mathbf{u}_t,$$

where  $u_t^i$  denotes the control of the  $i$ -th vehicle. The objective for each vehicle is to minimise the cumulative formation error and the energy consumption in the form of (4.2), where the matrix  $Q_t \in \mathbb{S}_{++}^N$  represents the formation error matrix that defined component-wise as  $[Q_t]_{ij} = \begin{cases} w_t^{ij} & \text{if } i = j \\ -w_t^{ij} & \text{if } i \neq j \end{cases}$  with  $w_t^{ij} \in \mathbb{R}$  be the error weights for  $i, j \in \{1, 2, \dots, N\}$ . Moreover, the energy consumption of the  $i$ -th vehicle is modelled by the quadratic cost  $\mathbf{u}_t^\top R_t^i \mathbf{u}_t$ . In practical implementations, such as decentralised formation control over solar-powered vehicles, the cost matrices  $Q_t$  and  $R_t^i$  depend

on environmental factors such as current solar irradiance level. Since weather conditions are only partially predictable, it is reasonable to assume that at time  $t$ , agent  $i$  has access to  $Q_\tau$  and  $R_\tau^i$  for  $1 \leq \tau \leq t + W$  with a preview horizon  $W$ .

**Example 3** (Macroeconomic Policy Making). Consider a generalised macroeconomic policy-making problem similar to that considered in [1]. Let  $x_t := [\tilde{x}_t, \tilde{x}_{t-1}]^\top \in \mathbb{R}^{2n}$  be the state where  $\tilde{x}_t$  is the value of monetary instrument and  $\mathbf{u}_t \in \mathbb{R}^m$  is the level of tax and spending policy at time  $t$ , respectively. The state and controls are governed by a linear dynamic (4.1), where the state transition matrix  $A$  maps how the lagged policy instruments affect the next period's lags, and the control matrix  $\mathbf{B}$  introduces the contribution of current policy decisions to the state.

The objective for each policy-maker is minimising the cost function in the form of (4.2), where  $Q_t \in \mathbb{S}_{++}^{2n}$  penalises rapid changes in the monetary instrument (reflecting the assumption that such changes are costly), and  $R_t^i \in \mathbb{R}^{2m \times 2m}$  is the cost associated with tax and spending policies for policy maker  $i$ . The cost matrices  $Q_t$  and  $R_t^i$  are influenced by the economic conditions that change over time. Given that economic environment is only partially predictable in the short term, it is reasonable to assume that at time  $t$ , each policy maker  $i$  has access to  $Q_\tau$  and  $R_\tau^i$  for  $1 \leq \tau \leq t + W$  with a preview horizon  $W$ .

The preceding examples illustrate the practical relevance of our dynamic LQ games with sequentially revealed costs. The systems and cost parameters under certain conditions outlined in [46] will fall within the class of dynamic LQ potential games. These conditions will be later formally stated in our assumptions.

The players do not (explicitly) cooperate, and thus we shall consider the solution that arises to be a *feedback Nash equilibrium*. To ease our presentation, let  $(\pi_t^{i*})_{i=1}^N$  denotes the feedback Nash equilibrium policy, and  $\Pi_t^* := (\pi_t^{1*}, \dots, \pi_t^{N*})$ . The feedback control policies  $(\pi_t^{i*})_{i=1, t=1}^{N, T-1}$  with  $\pi_t^{i*} \in \Lambda$ , where  $\Lambda$  is the set of measurable function, constituent a *feedback Nash equilibrium* if, for any given  $t \in \mathbb{N}_{T-1}$  and  $i \in \{1, 2, \dots, N\}$ , it satisfies the inequality below [24, Equation (3.27)-(3.28)]

$$J_T^i(\bar{x}_1, (\Pi_{1:t-1}, \Pi_t^*, \Pi_{t+1:T-1}^*)) \leq J_T^i(\bar{x}_1, (\Pi_{1:t-1}, (\pi_t^i, \pi_t^{-i*}), \Pi_{t+1:T-1}^*)), \quad (4.5)$$

for all feedback control policies  $(\pi_t^i)_{i=1, t=1}^{N, T-1}$  with  $\pi_t^i \in \Lambda$ . The state dynamics (4.1), player cost functions (4.2), and feedback Nash equilibrium solution concept (4.5), together define an  $N$ -player LQ-DFG. For notational convenience, define DFG as an operator that returns state and control sequences under a feedback Nash equilibrium

solution of this LQ-DFG, i.e.,

$$((x_t^*)_{t=1}^T, (\mathbf{u}_t^*)_{t=1}^{T-1}) = \text{DFG}(\mathcal{I}, T, A, \mathbf{B}), \quad (4.6)$$

where  $\mathbf{u}_t^* = \Pi_t^*(x_t^*)$ , and  $\mathcal{I} := (\bar{x}_1, (Q_{\tau+1}, (R_\tau^i)_{i=1}^N)_{\tau=1}^{T-1})$  is the tuple of the initial state and cost-function information necessary to compute a feedback Nash equilibrium.

Before presenting the parameter assumptions, we first introduce the concept of LQ dynamic potential games, starting with the definition of an LQ optimal control problem (LQ-OCP) as follows.

**Definition 4.2.1** (LQ-OCP). *An LQ-OCP is the problem of finding a policy  $(\bar{\Pi}_t)_{t=1}^{T-1}$  solving*

$$\begin{aligned} \min_{\bar{\Pi}} \quad & \sum_{t=1}^{T-1} x_{t+1}^\top \bar{Q}_{t+1} x_{t+1} + \mathbf{u}_t^\top \bar{R}_t \mathbf{u}_t \\ \text{s.t.} \quad & \mathbf{u}_t = \bar{\Pi}_t(x_t), \text{ and (4.1).} \end{aligned} \quad (4.7)$$

for a given positive integer  $T \geq 1$ , and positive definite cost matrices  $(\bar{Q}_t)_{t=1}^T$  and  $(\bar{R}_t)_{t=1}^{T-1}$ .

**Remark 4.2.1.** *A special case of the LQ game is  $[R_t^i]_{ii} = [R_t^j]_{jj}$ ,  $[R_t^i]_{ij} = [R_t^j]_{ji}$  and  $B^i = B^j$  for  $i, j \in \{1, 2, \dots, N\}$ . In this scenario, the LQ dynamic game reduces to an optimal control problem. In our work, we consider a more general setting beyond this optimal control case.*

We specifically leverage a recent result on the connection between LQ-DFGs and LQ-OCPs that gives rise to the following recently developed concept of linear-quadratic dynamic feedback potential games (LQ-DFPGs).

**Definition 4.2.2** (LQ-DFPG [46]). *An LQ-DFG is referred to as an LQ-DFPG, if there exists an LQ-OCP such that the solution of the LQ-OCP is a feedback Nash equilibrium of the LQ-DFG.*

We next introduce an assumption that an LQ-DFG with parameters  $A, \mathbf{B}, (Q_t)_{t=1}^T, (R_t^i)_{i=1, t=1}^{N, T-1}$  that satisfy the conditions given in [46, Theorem 6] is sufficient to constitute an LQ-DFPG.

**Assumption 4.2.1.** *Let  $Q_t \in \mathbb{S}_{++}^n$  and  $\Theta_t \in \mathbb{S}_{++}^{Nm}$ ,*

$$[R_t^i]_{ij} + B^{i\top} P_{t+1}^i B^j = ([R_t^j]_{ji} + B^{j\top} P_{t+1}^j B^i)^\top, \quad (4.8)$$

and

$$\mathbf{B}^\top P_t^i A = \mathbf{B}^\top P_t^j A, \quad (4.9)$$

where

$$P_T^i = Q_T, \quad (4.10)$$

$$[\Theta_t]_{ij} := [R_t]_{ij}^i + B^{i\top} P_{t+1}^i B^j, \quad (4.11)$$

$$K_t = -\Theta_t^{-1} \begin{bmatrix} B^1 P_{t+1}^1 \\ B^2 P_{t+1}^2 \\ \vdots \\ B^N P_{t+1}^N \end{bmatrix} A, \quad (4.12)$$

$$P_t^i = Q_t + K_t^\top R_t^i K_t + (A + BK_t)^\top P_{t+1}^i (A + BK_t). \quad (4.13)$$

for  $t \in \mathbb{N}_{T-1}$  and  $i, j \in \{1, 2, \dots, N\}$ .

**Remark 4.2.2.** Under Assumption 4.2.1, the LQ-DFG with parameters  $A, \mathbf{B}, (Q_t)_{t=1}^T, (R_t^i)_{i=1, t=1}^{N, T-1}$  is an LQ-DFPG and has a unique feedback Nash equilibria. The proof is identical to the proof of [46, Theorem 6].

The positive definiteness of  $\Theta_t$  guarantees that the diagonal block elements  $[R_t^i]_{ii} + B^{i\top} P_{t+1}^i B^i$  are positive definite. This ensures that the player  $i$ 's objective in (4.2) is strictly convex in  $u_t^i$ , and  $K_t^*$  is the unique feedback Nash equilibrium solution to the LQ-DFG given parameters  $(Q_t)_{t=1}^T, (R_t^i)_{i=1, t=1}^{N, T-1}$ . Equations (4.8)-(4.9) yield the relation of

$$\frac{\partial^2 J_{i,T}}{\partial u_{i,t} \partial u_{j,t}} = \left( \frac{\partial^2 J_{j,T}}{\partial u_{j,t} \partial u_{i,t}} \right)^\top, \quad (4.14)$$

which is a necessary and sufficient condition for the existence of a twice-differentiable potential function. This is analogue to the well-known result from [71, Theorem 4.5] or [46, Theorem 4], and we can ensure that the LQ-DFG parameters is an LQ-DFPG in the sense of Definition 4.2.2. Detail proof see Lemma C.1.1 from Appendix.

When the cost-matrices are known *a priori* to the players, the (Nash-equilibrium) solution of an LQ-DFPG can be found in closed form (cf. [46] and [24, Chapter 6]). However, in practice full information about the cost matrices over the whole time horizon  $T$  may be unavailable to the players in advance. Hence, in this chapter, we suppose that at any time  $t \in \mathbb{N}_{T-1-W}$  where  $W \in \mathbb{N}_{T-1} \cup \{0\}$ , only the initial condition of the system (4.1) and the (partial) sequences of cost matrices  $(Q_{\tau+1})_{\tau=1}^{t+W}$  and  $(R_\tau^i)_{i=1, \tau=1}^{N, t+W}$  are known to the players. Let the cost-function information available to the players at time  $t$  be

$$\mathcal{H}_t := (\bar{x}_1, (Q_{\tau+1}, (R_\tau^i)_{i=1}^N)_{\tau=1}^{\min(t+W, T-1)}). \quad (4.15)$$

Our focus will be to propose a novel control policy for each player in an LQ-DFPGs that uses the information available to them only at time  $t$ . For  $i \in \{1, 2, \dots, N\}$ , we specifically consider player-feedback control policies  $\pi_t^i(\cdot, \cdot)$  of the form

$$u_t^i = \pi_t^i(x_t, \mathcal{H}_t), \quad (4.16)$$

Note that the feedback Nash equilibrium solution  $\mathbf{u}_t^*$  can be written as  $\mathbf{u}_t^* = \pi_t^*(x_t^*, \mathcal{H}_{T-1})$ .

Motivated by the concept of the price of uncertainty (PoU) [54, Section 3.2] in static games, which captures discrepancies between expected payoffs in environments with complete and incomplete information, we introduce an analogous notion for dynamic feedback potential games to measure the difference between a decision made with causal cost information and a policy that achieves a feedback Nash equilibrium in hindsight.

**Definition 4.2.3** (Price of Uncertainty). *Given an LQ-DFPG, for a player feedback control policy  $(\Pi_t)_{t=1}^{T-1}$  with available information  $\mathcal{H}_t$  in the form of (4.16), we define the price of uncertainty (PoU) as*

$$PoU_T((\Pi_t)_{t=1}^{T-1}) := \frac{1}{N} \left[ \sum_{i=1}^N J_T^i(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) - \sum_{i=1}^N J_T^i(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) \right].$$

**Remark 4.2.3.** *The price of uncertainty coincides with dynamic regret in online optimal control (cf. [50, Eq. (9)]) when there is only a single player.*

Next, we establish the relationship between PoU and the notion of price of anarchy (PoA) of [72, Definition 2]. Define the PoA in the following.

**Definition 4.2.4** (Price of Anarchy). *For any  $N$ -player dynamic game that satisfies Assumption 4.2.1, define*

$$\tilde{J} := \min_{(\tilde{\Pi}_t)_{t=1}^{T-1}} \sum_{i=1}^N J_T^i(\bar{x}_1, (\tilde{\Pi}_t)_{t=1}^{T-1}),$$

*subject to dynamic (4.1). The Price of Anarchy (PoA) is*

$$PoA_T := \sum_{i=1}^N J_T^i(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) / \tilde{J}. \quad (4.17)$$

Note that in the original definition of PoA from [72, Definition 2] considers the Nash equilibrium that attained the smallest social cost. Under Assumption 4.2.1, there is a unique (feedback) Nash equilibrium, so we do not have to specify a Nash

equilibrium for computing the PoA. The following proposition relates PoA and PoU for any given policy.

**Proposition 2.** *For any policy  $(\Pi_t)_{t=1}^{T-1}$ , we have*

$$\tilde{J}(1 - PoA_T) \leq PoU_T((\Pi_t)_{t=1}^{T-1}).$$

**Remark 4.2.4.** *All the proofs can be found in the appendix.*

**Remark 4.2.5.** *A negative PoU indicates that the players have behaved cooperatively in retrospect under limited access to the cost parameters. In this case, the players may have sacrificed the chance of optimising their own pay-offs. Conversely, a policy that yields a large positive PoU suggests that players have acted noncooperatively. For example, in the context of demand-side management for a community battery, such a policy may reflect players selfishly drawing energy from the battery without caring about their own and their shared objective, potentially leading to energy depletion and undermining the shared objective of maintaining the battery's state of charge to a certain level. A policy that yields PoU near 0 indicates that the policy from players in retrospect aligned with their objectives while collective efficiency is met.*

We aim to show that the  $PoU$  associated with our proposed policy decays exponentially with respect to the preview window  $W$ , and is sublinear with respect to the time horizon  $T$ , i.e.,  $PoU_T((\Pi_t)_{t=1}^{T-1}) = o(T)$ , where  $o(T)$  satisfies  $\lim_{T \rightarrow \infty} \frac{o(T)}{T} = 0$ . This implies that, on average, the algorithm performs as well as the strategy that leads to feedback Nash equilibrium in hindsight, evaluated from the perspective of average costs.

We also require additional assumptions on the systems and cost parameters for our results. The following (mild) assumption is on the system dynamics matrices (4.1).

**Assumption 4.2.2.** *The state transition matrix  $A$  from (4.1) has full rank, and  $(A, \mathbf{B})$  is stabilisable.*

The next corollary states a property associated with cost matrices  $R_t^i$  for  $i \in \{1, 2, \dots, N\}$ .

**Corollary 2.** *Under Assumption 4.2.2, the matrix*

$$R_t^p := \begin{bmatrix} [R_t^1]_{11} & [R_t^1]_{12} & \cdots & [R_t^1]_{1N} \\ [R_t^2]_{21} & [R_t^2]_{22} & \cdots & [R_t^2]_{2N} \\ \vdots & \ddots & \vdots & \vdots \\ [R_t^N]_{N1} & [R_t^N]_{N2} & \cdots & [R_t^N]_{NN} \end{bmatrix} \quad (4.18)$$

is symmetric for  $t \in \mathbb{N}_{T-1}$ .

We further make assumptions on the conditions and bounds of time-varying cost matrices  $Q_t$  and  $R_t^i$ .

**Assumption 4.2.3.** *For any given  $t \in \mathbb{N}_{T-1}$ , there exist  $Q_{\min}, Q_{\max} \in \mathbb{S}_{++}^n$  and  $R_{\min}, R_{\max} \in \mathbb{S}_{++}^m$  that  $Q_t, R_t^p$  and  $R_t^i$  satisfy  $Q_{\min} \preceq Q_t \preceq Q_{\max}$ ,  $R_{\min} \preceq R_t^p \preceq R_{\max}$  and  $0 \preceq R_t^i \preceq R_{\max}$ .*

**Assumption 4.2.4.** *For any given  $t \in \mathbb{N}_{T-1}$ , let  $q_t := \frac{\sigma_{\max}(A)}{\sigma_{\min}^+(B)} \lambda_{\max}(R_t^p - R_t^1)$ . There exists a positive constant  $\varepsilon_Q$ , such that matrices  $Q_t$  and  $R_t^p$  satisfy  $\lambda_{\min}(Q_t) > |q_t| + \varepsilon_Q$ .*

Assumption 4.2.3 ensure that the eigenvalues for all cost matrices are positive and finite. This guarantees that all cost matrices  $(\bar{R}_t)_{t=1}^{T-1}$  in the LQ-OCP associated with the LQ-DFG are bounded and positive definite, which coincides with the standard assumption in LQR control literature, e.g., [50, Assumption 1], [4, Assumption 1]. See Corollary 4 in the Appendix. Assumption 4.2.4 ensures that the cost matrices  $(\bar{Q}_{t+1})_{t=1}^{T-1}$  in the LQ-OCP associated with the LQ-DFG are positive definite, see Lemma C.1.5 from the Appendix.

For a given preview horizon  $W \in \mathbb{N}_{T-1} \cup \{0\}$ , define

$$Q_{\tau+1|t} := \begin{cases} Q_{\tau+1} & \text{if } 1 \leq \tau \leq \min(t+W, T-1) \\ Q_{t+W+1} & \text{if } \min(t+W+1, T-1) \leq \tau \leq T-1, \end{cases} \quad (4.19)$$

and

$$R_{\tau|t}^i := \begin{cases} R_{\tau}^i & \text{if } 1 \leq \tau \leq \min(t+W, T-1) \\ R_{t+W}^i & \text{if } \min(t+W, T-1) \leq \tau \leq T-1, \end{cases} \quad (4.20)$$

for  $i \in \{1, 2, \dots, N\}$ . Our final assumption is then as follows.

**Assumption 4.2.5.** *The LQ-DFG with parameters  $(Q_{\tau|t})_{\tau=1}^T$  and  $(R_{\tau|t}^i)_{\tau=1, i=1}^{T-1, N}$  is an LQ-DFPG for all  $t \in \mathbb{N}_{T-1}$ .*

The above assumption ensures that when the last known cost matrices are repeated at time  $t$  for the next  $T-t$  stages, the LQ-DFG corresponds to an LQ-DFPG. In

many cases, we find that this assumption holds and can be verified using the structure of  $Q_{t+1}, (R_t^i)_{i=1}^N, A, \mathbf{B}$  for  $t \in \mathbb{N}_{T-1}$  (with knowledge of their specific numerical values). A particular example is presented in Section 4.5.

In the next section, we present our algorithm for finding controls and provide the *PoU* upper bound under such an algorithm.

### 4.3 Proposed Approach and PoU Analysis

In this section, we propose an algorithm for solving our proposed LQ-DFPG problem and analyse its PoU.

#### 4.3.1 Algorithm

Our algorithm involves two steps at each time step  $t$  (for each player) — a prediction step and a tracking step.

##### Prediction

We first predict a candidate feedback Nash equilibrium trajectory using the current cost matrices and setting the future unknown cost matrices equal to the known value at time  $t + W$ . That is, define  $\bar{\mathcal{H}}_t := (\bar{x}_1, (Q_{\tau+1|t})_{\tau=1}^{T-1}, (R_{\tau|t}^i)_{i=1}^N)_{\tau=1}^{T-1}$ , where  $Q_{\tau+1|t}, (R_{\tau|t}^i)_{i=1}^N$  are defined in (4.19) and (4.20). Note that the above cost matrices provide a valid structure for the LQ-DFG to be LQ-DFPG based on Assumption 4.2.5. Moreover, when  $T - W < t \leq T - 1$ ,  $\bar{\mathcal{H}}_t = \mathcal{I}$ . This suggests that when enough information is provided, we can find such candidate trajectory that is identical to the feedback Nash equilibrium trajectory obtained with full information  $\mathcal{I}$ . The predicted feedback Nash equilibrium trajectories given the information available at  $t$  are

$$((x_{\tau|t})_{\tau=1}^T, (\mathbf{u}_{\tau|t})_{\tau=1}^{T-1}) = \text{DFG}(\bar{\mathcal{H}}_t, T, A, \mathbf{B}), \quad (4.21)$$

where the operator DFG defined in (4.6) that uses  $\bar{\mathcal{H}}_t$  to generate the trajectories and controls corresponding to the feedback Nash equilibrium solution from  $N$  players.

### Tracking

Given the predicted feedback Nash equilibrium state and control trajectories (4.21), we propose tracking them using a feedback control policy of the form

$$\mathbf{u}_t = \begin{bmatrix} \pi_t^1(x_t, \bar{\mathcal{H}}_t) \\ \pi_t^2(x_t, \bar{\mathcal{H}}_t) \\ \vdots \\ \pi_t^N(x_t, \bar{\mathcal{H}}_{\perp}) \end{bmatrix} = \begin{bmatrix} [\underline{K}]_1 \\ [\underline{K}]_2 \\ \vdots \\ [\underline{K}]_N \end{bmatrix} (x_t - x_{t|t}) + \begin{bmatrix} [\mathbf{u}_{t|t}]_1 \\ [\mathbf{u}_{t|t}]_2 \\ \vdots \\ [\mathbf{u}_{t|t}]_N \end{bmatrix}, \quad (4.22)$$

where  $\underline{K} \in \mathbb{R}^{m \times n}$  is a gain matrix satisfies  $\rho(A + \mathbf{B}\underline{K}) < 1$ ,  $\pi_t^i(x_t, \bar{\mathcal{H}}_t)$  is the policy for the  $i$ -th player for  $i \in \{1, 2, \dots, N\}$ ,  $t \in \mathbb{N}_{T-1}$ ,  $[\underline{K}]_i \in \mathbb{R}^{m \times m}$  and  $[\mathbf{u}_{t|t}]_i \in \mathbb{R}^m$  have appropriate dimensions. In the next section, we show the *PoU* bounds for this algorithm. Intuitively, any LQ-DPG has a corresponding LQ-OCP. Thus, at each time  $t$ , there is a common value function  $\bar{V}_t(\cdot)$  for all players coordinating to minimise subject to the system dynamics (4.1). The state  $x_{t|t}$ , is the best possible minimiser for  $\bar{V}_t(\cdot)$  with information tuple  $\mathcal{H}_{t,W}$ .

#### 4.3.2 *PoU* Analysis

The following theorem establishes *PoU* lower and upper bounds for our proposed algorithm.

**Theorem 4.3.1** (*PoU* Bounds). *Consider the linear system (4.1), a given time horizon  $T \geq 1$ , a preview window length  $W \in \mathbb{N}_{T-1} \cup \{0\}$  and a number of players  $N \geq 1$ . Under Assumptions 4.2.1–4.2.5, with adopting policy in (4.22), the *PoU* defined by (4.17) satisfies*

$$-C'_1\gamma^{2W} - C'_2\gamma^W - C'_3\varepsilon_K < PoU_T((\Pi_t)_{t=1}^{T-1}) < C_1\gamma^{2W} + C_2\gamma^W + C_3\varepsilon_K, \quad (4.23)$$

where  $\gamma \in (0, 1)$  and  $\varepsilon_K$  is monotonically increasing w.r.t.  $\gamma$ ,  $\|\mathbf{B}\|$ ,  $\lambda_{\max}(R_{\max})$ ,  $\lambda_{\max}(Q_{\max})$  and  $\lambda_{\max}(R_{\max}) - \lambda_{\min}(R_{\min})$ . Moreover, constants  $C_1, C_2, C_3, C'_1, C'_2$  and  $C'_3$  are monotonically increasing w.r.t.  $\|\bar{x}_1\|$ ,  $\lambda_{\max}(R_{\max})$ ,  $\lambda_{\max}(Q_{\max})$  and  $\lambda_{\max}(R_{\max}) - \lambda_{\min}(R_{\min})$ , and the inverse of  $\rho(A + \mathbf{B}\bar{K})$  and  $\varepsilon_Q$ .

**Remark 4.3.1.** *Following Remark 4.2.1, in the special case of LQ optimal control,  $C'_2 = 0$ ,  $C'_3 = 0$ ,  $\varepsilon_K = 0$ , and  $C_2 = 0$ :*

$$0 \leq PoU_T((\Pi_t)_{t=1}^{T-1}) < C_1\gamma^{2W}.$$

**Remark 4.3.2.** *If  $C_1, C_2$  and  $C_3$  are equal to 0, this implies the corresponding*

coefficients  $C'_1, C'_2$  and  $C'_3$  are equal to 0. Consequently, if the upper bound is 0, then the PoU under policy (4.22) is 0.

**Remark 4.3.3.** Based on Proposition 2,  $\tilde{J}(1 - \text{PoA}_T)$  is the largest lower bound on PoU given an LQ-DFPG. Combining with Theorem 4.3.1, we immediately have  $\tilde{J}(1 - \text{PoA}_T) \leq -C'_1\gamma^{2W} - C'_2\gamma^W - C'_3\varepsilon_K$ .

Theorem 4.3.1 provides an insight into the relationship between PoU and the preview-window length  $W$ . Specifically, we see that for a fixed  $\mathcal{I}$ , the terms  $C'_1\gamma^{2W}, C'_2\gamma^W, C_1\gamma^{2W}$  and  $C_2\gamma^W$  in the PoU bound decays exponentially fast as the preview-window length  $W$  increases. Moreover, when  $W$  is sufficiently large, the PoU lower and upper bounds are dominated by  $-C'_3\varepsilon_K$  and  $C_3\varepsilon_K$ , respectively.

## 4.4 Relationship Between PoU and Feedback Nash Equilibrium

Suppose there exists  $(\hat{\Pi}_t)_{t=1}^{T-1}$  such that  $\text{PoU}((\hat{\Pi}_t)_{t=1}^{T-1}) \leq \delta_2$  for scalar  $\delta_2 \geq 0$ . Our goal is to quantify the relationship between this policy and the feedback Nash equilibrium policy  $(\Pi_t^*)_{t=1}^{T-1}$  in terms of  $\delta_2$ , coefficients that are system-dependent parameters, and the state trajectory generated by the policy  $(\hat{\Pi}_t)_{t=1}^{T-1}$ .

To proceed, we introduce the following notation. Let  $\{\hat{x}_t\}_{t=1}^{T-1}$  and  $\{\hat{\mathbf{u}}_t\}_{t=1}^{T-1}$  be the states and controls generated by applying  $(\hat{\Pi}_t)_{t=1}^{T-1}$  to the system (4.1). Define:

$$\begin{aligned}\tilde{\mathbf{u}}_t &:= K_t \hat{x}_t, \\ b_t(\hat{x}_t) &:= \frac{2}{N} [(R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) K_t + \mathbf{B}^\top P_{t+1}^i A] \hat{x}_t, \\ H_t &:= \frac{1}{N} (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}),\end{aligned}$$

$$\hat{\mathbf{U}} := \begin{bmatrix} \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \vdots \\ \hat{\mathbf{u}}_{T-1} \end{bmatrix}, \tilde{\mathbf{U}} := \begin{bmatrix} \tilde{\mathbf{u}}_1 \\ \tilde{\mathbf{u}}_2 \\ \vdots \\ \tilde{\mathbf{u}}_{T-1} \end{bmatrix}, \hat{b} := \begin{bmatrix} b_1(\hat{x}_1) \\ b_2(\hat{x}_2) \\ \vdots \\ b_{T-1}(\hat{x}_{T-1}) \end{bmatrix},$$

$\tilde{z} := \hat{\mathbf{U}} - \tilde{\mathbf{U}}$  and  $\tilde{H} := \text{diag}(H_1, H_2, \dots, H_{T-1})$ , where  $\text{diag}$  is the block-diagonal matrix concatenation operator.

The following proposition characterises an upper bound on  $\|\tilde{z}\|$  under the assumption that  $\text{PoU}((\hat{\Pi}_t)_{t=1}^{T-1}) \leq \delta_2$  and  $\|\hat{b}\| \leq \delta_x$  for some  $\delta_x \geq 0$ , potentially dependent on

state sequence  $(\hat{x}_t)_{t=1}^{T-1}$ . We bound  $\|\tilde{z}\|$  as a function of  $\delta_2, \delta_x$ , and system-dependent parameters.

**Proposition 3.** *Consider a policy  $(\hat{\Pi}_t)_{t=1}^{T-1}$  and associated state trajectory  $(\hat{x}_t)_{t=1}^{T-1}$  generated by applying it to (4.1). Suppose that there exists  $\delta_2 \geq 0$  and  $\delta_x \geq 0$ , such that*

$$PoU((\hat{\Pi}_t)_{t=1}^{T-1}) \leq \delta_2, \quad \text{and} \quad \|\hat{b}\|^2 \leq \delta_x.$$

Define  $\bar{\delta}_2 := \delta_2 + \frac{1}{2}\lambda_{max}(\tilde{H})\delta_x$ , then  $\|\tilde{z}\|^2 \leq \bar{\delta}_2$ .

Next, we discuss the sub-optimality of the policy  $(\hat{\Pi}_t)_{t=1}^{T-1}$  under the assumptions from the above proposition. Define the potential function:

$$\Phi(\bar{x}_1, (\bar{\Pi}_t)_{t=1}^{T-1}) := \sum_{t=1}^{T-1} x_{t+1}^\top \bar{Q}_t x_{t+1} + \mathbf{u}_t^\top \bar{R}_t \mathbf{u}_t, \quad (4.24)$$

where  $\mathbf{u}_t = \bar{\Pi}_t(x_t)$  for any given policy  $(\bar{\Pi}_t)_{t=1}^{T-1}$ , and  $(x_t, \mathbf{u}_t)$  satisfying (4.1). The matrices  $\bar{Q}_t$  and  $\bar{R}_t$  are the stage cost matrices defined in the corresponding LQ-OCP as in Definition 4.2.1.

**Proposition 4.** *Consider the potential function (4.24) with cost matrices  $(\bar{Q}_{t+1}, \bar{R}_t)_{t=1}^{T-1}$ . Define  $\bar{\varepsilon}_2 := \Delta_{max}\delta_2$ , and let  $\Delta_{max}$  be a positive real number monotonically increasing with  $\|R_{max}\|$  and  $\frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})}(\lambda_{max}(R_{max}) - \lambda_{min}(R_{min}))$ . Consider the policy  $(\hat{\Pi}_t)_{t=1}^{T-1}$  satisfying the conditions described in Proposition 3. Then,  $\Phi(\bar{x}_1, (\hat{\Pi}_t)_{t=1}^{T-1}) - \Phi(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) \leq \bar{\varepsilon}_2$ , where  $(\Pi_t^*)_{t=1}^{T-1}$  is the feedback Nash equilibrium policy satisfying (4.5).*

The above proposition shows that if the policy  $PoU((\hat{\Pi}_t)_{t=1}^{T-1})$  is upper bounded, then it yields  $\bar{\varepsilon}$  sub-optimality gap in the potential function. The sub-optimality gap  $\bar{\varepsilon}$  is inversely proportional to  $\delta_2$  and  $\delta_x$ . In other words, the policy  $(\hat{\Pi}_t)_{t=1}^{T-1}$  is close to the feedback Nash equilibrium policy if  $\delta_2$  and  $\delta_x$  are close to zero.

## 4.5 Numerical Simulations

In this section, we illustrate the performance of our proposed algorithm. We revisit the community battery application from Example 1 and provide conditions in which the system matrices and cost matrices satisfy Assumptions 4.2.1-4.2.5. Before specifying values of  $a, b_1, b_2$  and  $b_3$  for matrices  $A$  and  $\mathbf{B}$  defined in (4.4), consider the

cost matrices below

$$\begin{aligned} Q_t &= \begin{bmatrix} l_t & -d_t \\ -d_t & 0 \end{bmatrix}, \quad R_t^1 = \begin{bmatrix} r_{1,t} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \\ R_t^2 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & r_{2,t} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R_t^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & r_{3,t} \end{bmatrix}, \end{aligned} \quad (4.25)$$

where

$$\frac{b_1^2}{r_{1,t}} = \frac{b_2^2}{r_{2,t}} = \frac{b_3^2}{r_{3,t}} = \zeta_t. \quad (4.26)$$

It can be easily verified that the matrices  $A$  and  $\mathbf{B}$  satisfy Assumption 4.2.2. In the next proposition, we claim that such system matrices and the cost matrices will lead to an LQ potential dynamic game (Assumption 4.2.1) with satisfaction of Assumption 4.2.5.

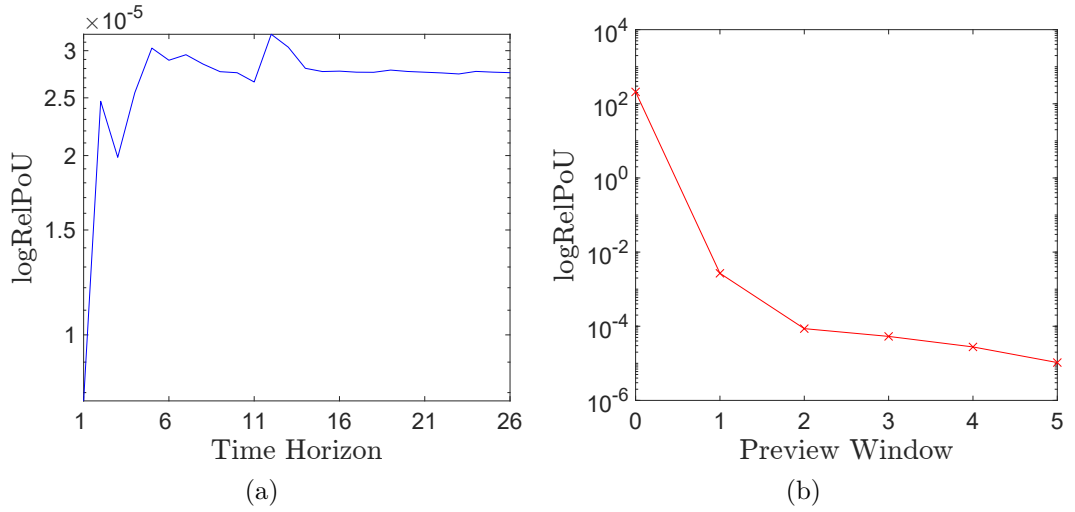
**Proposition 5.** *Consider matrices  $A$  and  $\mathbf{B}$  from Example 1, and cost matrices  $R_t^i, Q_t$  are defined in (4.25). For integers  $T \geq 1$ ,  $N \geq 1$ ,  $0 \leq t \leq T - 1$  and  $i, j \in \{1, 2, \dots, N\}$ , if  $\frac{b_i^2}{r_{i,t}} = \frac{b_j^2}{r_{j,t}}$ , then parameters  $P_t, \Theta_t$  defined in (4.13) and (4.11), respectively, satisfy conditions that described in Assumptions 4.2.1 and 4.2.5.*

For  $i \in \{1, 2, 3\}$ , by using the conditions of parameters  $b_i$  and  $r_{i,t}$  from the above proposition, we consider  $a = 1.6$ , and  $b_i \sim \mathcal{U}[1, 10]$  for  $i \in \{1, 2, 3\}$  in the simulations. The preview window  $W$  varies from 0 to 5, and the time horizon  $T$  ranges from 1 to 25. We consider cost matrices drawn randomly according to  $\zeta_t \sim \mathcal{U}[1, 16]$ ,  $l_t \sim \mathcal{U}[20, 110]$  and  $d_t \sim \mathcal{U}[15, 25]$  in (4.26), where  $\mathcal{U}[\cdot, \cdot]$  denotes the uniform distribution. Such choices of systems and cost parameters will lead the LQ potential dynamic game to satisfy Assumptions 4.2.3 and 4.2.4, along with the rest of the assumptions.

Before presenting our figures, we introduce a slightly modified notion of PoU, termed the *log-relative PoU*, that quantify the relative error between the decisions made from the algorithm and the (value of the) feedback Nash equilibrium solution. Specifically, for any sequence of policies  $(\Pi_t)_{t=1}^{T-1}$ , we define the *log relative PoU* as

$$\log \text{RelPoU}((\Pi_t)_{t=1}^{T-1}) := \log \left| \frac{\text{PoU}_T((\Pi_t)_{t=1}^{T-1})}{\frac{1}{N} \sum_{i=1}^N J_{i,T}(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1})} \right|,$$

where the policy  $(\Pi_t^*)_{t=1}^{T-1}$  generates the feedback Nash equilibrium trajectory as



**Figure 4.1:** Performance measure  $PoU_T$  in case  $N = 3$ . (a)  $LogRelPoU$  vs. Time Horizon with  $W = 5$ ; (b)  $LogRelPoU$  vs. Preview Window Length in  $T = 26$ .

defined in (4.5). In certain scenarios, the cost incurred by the actual feedback Nash equilibrium may be large due to the selection of the system and cost matrices. Thus, logarithmic relative PoU can be a more practical measure of the quality of decisions when we are interested in numerical results.

Figure 4.1(a) reports the *log relative average PoU* over 1000 trials with  $W = 5$ . From Figure 4.1(a) we see that the (average) log relative PoU appears to remain at  $3 \times 10^{-5}$  from  $T = 13$  when  $W = 5$ . This result illustrates that the average PoU will be bounded regardless of the choice of time-horizon, which validates the PoU bounds established in Theorem 4.3.1.

Figure 4.1(b) reports the *log relative average PoU* over 1000 trials with  $T = 26$ . From Figure 4.1(b) we see that the (average) log relative PoU shows an exponential decay as  $W$  grows from 0 to 5. The exponential decays of the (average) PoU illustrate the relation between the preview window  $W$  and the PoU bounds established in Theorem 4.3.1.

## 4.6 Summary

This chapter introduces the  $N$ -player dynamic potential LQ game with sequentially revealed costs problem, and proposes a novel algorithm for solving this problem, which extends the preliminary work in [70] from 2-player case to the general  $N$ -player setting. In addition, we establish the connection between the proposed notion of Price of Uncertainty (PoU) and the Price of Anarchy (PoA), and quantify the sub-optimality gap of the potential function for any policy whose PoU is upper

bounded. We provide numerical simulations to validate our PoU bounds established in Theorem 4.3.1.

---

# Conclusions

---

This thesis introduces a novel framework for decision making with unknown future costs, along with its theoretical guarantees and empirical evaluation across multiple settings. We first apply this framework to the online LQ control problem with sequentially revealed costs, this has discussed in Chapter 2. Then we apply this framework to the online LQ control problem, where the costs must be inferred from the expert's optimal trajectory data, this has discussed in Chapter 3. Lastly, we apply this framework to the dynamic potential LQ game problem, where it extends the setting in Chapter 2 to a multi-agent setting. In this chapter, we will summarise each chapter in detail and outline possible future research directions.

## 5.1 Summary of Contributions

In this section, we will summarise and discuss the research presented in this thesis.

### 5.1.1 Chapter 2: Online LQ Optimal Control Problem

In the first technical chapter of this thesis, we present a novel framework for decision making under unknown future costs, and apply it to the online LQ optimal control problem with sequentially revealed costs. Unlike the classical LQ control setting where cost matrices are known *a priori*, the online setting assumes that, at any time  $t$ , the controller only has access to past cost matrices and a short preview window of length  $W$ , i.e.,  $\{Q_{\tau+1}, R_{\tau}\}_{\tau=0}^{\min(t+W, T-1)}$ , with  $0 \leq W \leq T - 1$ .

To address this problem, we propose a framework that predicts a candidate optimal trajectory using the history  $\mathcal{H}_{t,W}$  defined in (1.4), and estimates the tail costs using the most recent available matrices  $Q_{t+W+1}$  and  $R_{t+W}$  for the horizon  $t + W \leq \tau \leq T - 1$ . The controller tracks the estimated next state  $x_{t+1|t}$  based on this prediction as given in (2.6). We adopt this framework to design a constructive deadbeat controller given in (2.26) and analyse its performance using the notion of dynamic regret (2.2),

as detailed in Theorems 2.3.1–2.4.2. These analyses show that the regrets are upper bounded by terms that decay exponentially with the preview horizon. Furthermore, we propose a sufficient condition in Proposition 2.3.2 that our framework achieves a tighter regret bound than the state-of-the-art method in [12].

We validate the regret analysis results of our methods through numerical simulations, comparing the proposed framework (2.6) and deadbeat controller (2.24) against algorithms from [12] and [17], as well as a trivial constant linear controller. The results confirm the exponential decay in regret predicted by theory and demonstrate that our methods consistently outperform those that do not exploit feedback from cost information.

In conclusion, Chapter 2 establishes both theoretical guarantees and empirical advantages of the proposed framework. It illustrates the benefit of exploiting feedback from cost information and paves the possibility for extending this framework to more general decision-making problems.

### 5.1.2 Chapter 3: Online LQ Optimal Control with Sequentially Inferred Costs

In Chapter 3, we apply the proposed framework to the online LQ optimal control problem, where the cost function is not directly observable but must be inferred from an expert’s optimal demonstrations in the form of optimal trajectories. This setting models practical scenarios such as learning from demonstration, where the decision maker has access only to observed behaviours and must infer the underlying objectives that generated them.

To address this problem, we incorporate an Extended Kalman Filter (EKF)-based estimator [64, 65] into our decision-making framework. At each time step, the controller uses the observed expert trajectories to estimate the unknown cost parameters, then adopts the control framework proposed in Chapter 2 to plan a candidate trajectory using this estimation and tracking towards it. We establish a regret bound that quantifies the impact of parameter estimation error on control performance, and derive a regret upper bound that holds under the EKF-based estimation scheme in Lemma 3.4.1 and Theorem 3.4.1. Our analysis shows that as the accuracy of the parameter estimates improves over time, the regret incurred by the controller decreases quadratically. The analysis also shows the regret yields a sublinear growth with respect to time. We validate our theoretical guarantee with numerical simulations. The simulation results in Figures 3.1 to 3.3 demonstrate that the proposed EKF-based controller effectively learns the underlying cost function and improves

its performance over time.

In conclusion, Chapter 3 illustrates how learning from expert demonstrations can be seamlessly integrated into online control using our proposed framework, and opens the door to future work on learning and control under uncertain or latent cost structures.

### 5.1.3 Chapter 4: Dynamic Potential LQ Games with Sequentially Revealed Costs

In Chapter 4, we apply the proposed framework to the dynamic potential LQ game with sequentially revealed costs problem, where it extends the setting from the single agent case in Chapter 2 to the multi-agent case. The formulation of this problem is introduced in detail in Section 4.2.

To address this problem, we introduce the notion of price of uncertainty (PoU) (4.17) as a performance measure defined in (4.17), that extends the notion of regret (2.2) from the online LQ optimal control problem to the dynamic game setting. We then extend our proposed framework for  $N$  players, as defined in (4.22), and apply it to the dynamic game problem formulated in 4.2. The PoU analysis establishes lower and upper bounds that are incurred under (4.22), where the bounds consist of terms that decay exponentially as the preview horizon of costs increases, and another term that depends on the magnitude of the differences between the cost matrices for each player (details are given in Theorem 4.3.1. Through simulations, we illustrate that the resulting price of uncertainty initially decays at an exponential rate as the preview window lengthens, then remains constant for large time horizons. This has illustrated in Figures 4.1(b) and 4.1(a), respectively.

## 5.2 Future Research Directions

In this thesis, the system dynamics for the controllers are linear time-invariant. A natural extension is to consider (potentially sequentially revealed) time-varying or nonlinear dynamics. Another immediate extension is to extend the cost functions from quadratic to non-quadratic settings. There are a few interesting long-term extensions to all our respective problems. We outline these in detail in the following.

### 5.2.1 Chapter 2: Online LQ Optimal Control

In Chapter 2, we provide a sufficient condition under which our regret bound is less than that of the state-of-the-art methodology. However, this sufficient condition does not necessarily imply superior empirical performance. The analysis of realised regret (rather than bounds) incurred by different control algorithms appears important, as it provides a more practical way to help people select an appropriate method for a corresponding scenario.

The proposed framework in (2.6) employs a time-varying control gain  $K_t$  to track towards the candidate trajectory. A particular choice of this gain is the deadbeat controller defined in (2.26). While deadbeat control ensures rapid convergence to the target trajectory, it is inherently a high-gain control (needs citation). Consequently, it can result in excessive control effort and be sensitive to disturbances, which leads to high costs. This has been shown empirically in Figures 2.3 where the regret decays in a significantly slower rate and with much larger regret coefficients compared to Figures 2.2. Therefore, a more effective approach would be to design the time-varying tracking controller adaptively, taking into account the time step  $t$  and the preview horizon  $W$ . This is motivated by the observation that, as the time step  $t$  increases or the preview horizon  $W$  lengthens, the candidate trajectory progressively gets closer to the optimal trajectory. The adaptive tracking strategies can potentially achieve better trade-offs between performance and robustness.

### 5.2.2 Chapter 3: Online LQ Optimal Control with Sequentially Inferred Costs

In Chapter 3, we used a constant controller for tracking the predicted trajectory in (3.24). As the copycat observes more expert trajectories, the candidate trajectory progressively converges toward the optimal one. This progression, similar to the observation in the previous section, motivates the development of adaptive, time-varying tracking controllers that are based on the closeness between the predicted and the expert's trajectories.

Throughout this work, we assumed that the copycat must act immediately after observing the optimal trajectory. However, in many learning from demonstration tasks, immediate imitation is not required. For example, in autonomous navigation tasks [73], the copycat may act after a delay, following the expert's demonstration. This delay allows the copycat to access additional data, enabling the use of smoothing techniques that can tighten the regret upper bound.

Moreover, we specifically consider the cost parameter that is fixed over time in the

problem formulation. A natural extension is to allow the cost parameter to evolve over time.

Another promising direction is to investigate the regret under the presence of disturbances for the linear system, which could provide robustness guarantees for the copycat's policy.

### **5.2.3 Chapter 4: Dynamic Potential LQ Games with Sequentially Revealed Costs**

In Chapter 4, we study the performance of our proposed framework within the context of dynamic potential LQ games. A general dynamic game can be decomposed into three components: a potential dynamic game component, a harmonic component and a non-strategic component. To extend our framework to more general dynamic game settings, a natural next step is to design controllers for dynamic LQ harmonic games. These controllers can then be integrated with those developed for dynamic LQ potential games, which enable the extension of our framework to a more general dynamic LQ game setting.

Another promising direction is to investigate the PoU in the presence of disturbances within the linear system, which would provide further insights into the robustness of the proposed framework.



---

# Bibliography

---

- [1] F. Kydland, “Noncooperative and Dominant Player Solutions in Discrete Dynamic Games,” *International Economic Review*, vol. 16, no. 2, pp. 321–335, 1975, publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University].
- [2] A. Aghajani and A. Doustmohammadi, “Formation control of multi-vehicle systems using cooperative game theory,” in *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, 2015, pp. 704–709.
- [3] S. Zazo, S. V. Macua, M. S. Fernández, and J. Zazo, “Dynamic Potential Games With Constraints: Fundamentals and Applications in Communications,” *IEEE Transactions on Signal Processing*, vol. 64, pp. 3806–3821, 2016.
- [4] J. Sun and M. Cantoni, “On receding-horizon approximation in time-varying optimal control,” May 2023, arXiv:2305.06010 [cs, eess, math].
- [5] —, “On Riccati contraction in time-varying linear-quadratic control,” May 2023, arXiv:2305.06003 [cs, eess, math].
- [6] I. Dogan, Z.-J. M. Shen, and A. Aswani, “Regret Analysis of Learning-Based MPC With Partially Unknown Cost Function,” *IEEE Transactions on Automatic Control*, vol. 69, no. 5, pp. 3246–3253, 2024.
- [7] M. Nonhoff, E. Dall’Anese, and M. A. Müller, “Online convex optimization for robust control of constrained dynamical systems,” 2024.
- [8] A. Cohen, A. Hassidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar, “Online Linear Quadratic Control,” *arXiv:1806.07104 [cs, stat]*, Jun. 2018, arXiv:1806.07104.
- [9] M. Akbari, B. Gharesifard, and T. Linder, “Logarithmic regret in online linear quadratic control using Riccati updates,” *Mathematics of Control, Signals, and Systems*, Apr. 2022.
- [10] S. Aggarwal, R. K. Velicheti, and T. Başar, “Learning to Control Under Communication Constraints,” *IEEE Control Systems Letters*, vol. 7, pp. 2137–2142, 2023.

- 
- [11] T.-J. Chang and S. Shahrampour, “Regret Analysis of Distributed Online LQR Control for Unknown LTI Systems,” *IEEE Transactions on Automatic Control*, vol. 69, no. 1, pp. 667–673, Jan. 2024.
- [12] R. Zhang, Y. Li, and N. Li, “On the Regret Analysis of Online LQR Control with Predictions,” in *2021 American Control Conference (ACC)*, May 2021, pp. 697–703, ISSN: 2378-5861.
- [13] Y. Lin, Y. Hu, G. Shi, H. Sun, G. Qu, and A. Wierman, “Perturbation-based Regret Analysis of Predictive Control in Linear Time Varying Systems,” in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 5174–5185.
- [14] Y. Li, X. Chen, and N. Li, “Online Optimal Control with Linear Dynamics and Predictions: Algorithms and Regret Analysis,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [15] M. Nonhoff, J. Köhler, and M. A. Müller, “Online convex optimization for constrained control of linear systems using a reference governor\*,” *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 2570–2575, 2023.
- [16] H. Zhou and V. Tzoumas, “Safe Non-Stochastic Control of Linear Dynamical Systems,” in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 5033–5038.
- [17] H. Zhou, Y. Song, and V. Tzoumas, “Safe Non-Stochastic Control of Control-Affine Systems: An Online Convex Optimization Approach,” *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 7873–7880, 2023.
- [18] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Courier Corporation, Feb. 2007.
- [19] K. Lu, “Online Distributed Algorithms for Online Noncooperative Games With Stochastic Cost Functions: High Probability Bound of Regrets,” *IEEE Transactions on Automatic Control*, vol. 69, no. 12, pp. 8860–8867, 2024.
- [20] L. Peters, V. Rubies-Royo, C. J. Tomlin, L. Ferranti, J. Alonso-Mora, C. Stachniss, and D. Fridovich-Keil, “Online and offline learning of player objectives from partial observations in dynamic games,” *The International Journal of Robotics Research*, vol. 42, no. 10, pp. 917–937, 2023.
- [21] S. Hosseini and M. Mesbahi, “Energy-Aware Aerial Surveillance for a Long-Endurance Solar-Powered Unmanned Aerial Vehicles,” *Journal of Guidance, Control, and Dynamics*, vol. 39, no. 9, pp. 1980–1993, 2016.

- 
- [22] E. Hazan, "Introduction to Online Convex Optimization," *Found. Trends Optim.*, vol. 2, no. 3–4, pp. 157–325, Aug. 2016, place: Hanover, MA, USA Publisher: Now Publishers Inc.
- [23] S. Shalev-Shwartz, *Online learning and online convex optimization*, ser. Foundations and trends in machine learning. Boston: Now, 2012, no. 4:2.
- [24] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.
- [25] D. Bertsekas, "Dynamic Programming and Optimal Control," Jan. 1995, vol. 1.
- [26] S.-J. Kim and G. B. Giannakis, "An Online Convex Optimization Approach to Real-Time Energy Pricing for Demand Response," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2784–2793, 2017.
- [27] A. Lesage-Landry, H. Wang, I. Shames, P. Mancarella, and J. A. Taylor, "Online Convex Optimization of Multi-Energy Building-to-Grid Ancillary Services," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2416–2431, 2020.
- [28] F. Amirnavaei and M. Dong, "Online Power Control Optimization for Wireless Transmission With Energy Harvesting and Storage," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4888–4901, 2016.
- [29] M. Lin, Z. Liu, A. Wierman, and L. L. H. Andrew, "Online algorithms for geographical load balancing," in *2012 International Green Computing Conference (IGCC)*, 2012, pp. 1–10.
- [30] S. Chen and L. Tong, "iEMS for large scale charging of electric vehicles: Architecture and optimal online scheduling," in *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, 2012, pp. 629–634.
- [31] Y. Li, S. Das, and N. Li, "Online Optimal Control with Affine Constraints," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8527–8537, May 2021.
- [32] H. Chu, L. Guo, Y. Yan, B. Gao, and H. Chen, "Self-Learning Optimal Cruise Control Based on Individual Car-Following Style," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6622–6633, 2021.
- [33] J. Engwerda, *LQ Dynamic Optimization and Differential Games*, Jan. 2005, publication Title: Journal of Banking & Finance - J BANK FINAN.
- [34] G. Perin and L. Badia, "Static and Dynamic Jamming Games Over Wireless Channels With Mobile Strategic Players," 2023.

- 
- [35] N. T. Thanh Van, N. C. Luong, S. Feng, H. T. Nguyen, K. Zhu, T. V. Luong, and D. Niyato, "Dynamic Network Service Selection in Intelligent Reflecting Surface-Enabled Wireless Systems: Game Theory Approaches," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 5947–5961, 2022.
- [36] L. Wu, Y. Xiong, K.-Z. Liu, and J. She, "A real-time pricing mechanism considering data freshness based on non-cooperative game in crowdsensing," *Information Sciences*, vol. 608, pp. 392–409, 2022.
- [37] Y. Yan and T. Hayakawa, "Stability and Stabilization of Nash Equilibrium for Uncertain Noncooperative Dynamical Systems With Zero-Sum Tax/Subsidy Approach," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 11 287–11 298, 2022.
- [38] F. Dong, D. Jin, X. Zhao, J. Han, and W. Lu, "A non-cooperative game approach to the robust control design for a class of fuzzy dynamical systems," *ISA Transactions*, vol. 125, pp. 119–133, 2022.
- [39] P. Scarabaggio, R. Carli, and M. Dotoli, "Noncooperative Equilibrium-Seeking in Distributed Energy Systems Under AC Power Flow Nonlinear Constraints," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 4, pp. 1731–1742, 2022.
- [40] P. Serna-Torre, V. Shenoy, D. Schoenwald, J. I. Poveda, and P. Hidalgo-Gonzalez, "Non-cooperative games to control learned inverter dynamics of distributed energy resources," *Electric Power Systems Research*, vol. 234, p. 110641, 2024.
- [41] H. Wu, H. Liu, Y. He, A. Y. Wu, and M. Ding, "Market bidding for multiple photovoltaic-storage systems: A two-stage bidding strategy based on a non-cooperative game," *Solar Energy*, vol. 271, p. 112438, 2024.
- [42] Y.-H. Ni, L. Liu, and X. Zhang, "Deterministic dynamic Stackelberg games: Time-consistent open-loop solution," *Automatica*, vol. 148, p. 110757, 2023.
- [43] S. Martin, I. Alvarez, and F. Lavallée, "Solving viability problems in dynamic games using individual strategies derived from guaranteed viability kernels: Application to an agricultural cooperative model," *Automatica*, vol. 167, p. 111752, 2024.
- [44] B. Nortmann, M. Sassano, and T. Mylvaganam, "Feedback Nash equilibria for scalar N-player linear quadratic dynamic games," *Automatica*, vol. 174, p. 112133, 2025.

- 
- [45] S. Aberkane and V. Dragan, “An addendum to the problem of zero-sum LQ stochastic mean-field dynamic games,” *Automatica*, vol. 153, p. 111007, 2023.
- [46] A. Prasad, P. S. Mohapatra, and P. V. Reddy, “On the Structure of Feedback Potential Difference Games,” *IEEE Transactions on Automatic Control*, pp. 1–8, 2023.
- [47] C. Wu, H. Mohsenian-Rad, J. Huang, and A. Y. Wang, “Demand side management for Wind Power Integration in microgrid using dynamic potential game theory,” in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*. IEEE, 2011, pp. 1199–1204.
- [48] C. P. Mediwaththe, E. R. Stephens, D. B. Smith, and A. Mahanti, “A Dynamic Game for Electricity Load Management in Neighborhood Area Networks,” *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1329–1336, 2016.
- [49] N. Kirsch, G. Salizzoni, T. Gorecki, and M. Kamgarpour, “A Distributed Game Theoretic Approach for Optimal Battery Use in an Energy Community,” *SIGENERGY Energy Inform. Rev.*, vol. 4, no. 4, pp. 207–213, Feb. 2025, place: New York, NY, USA Publisher: Association for Computing Machinery.
- [50] Y. Chen, T. L. Molloy, T. Summers, and I. Shames, “Regret Analysis of Online LQR Control via Trajectory Prediction and Tracking,” in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, N. Matni, M. Morari, and G. J. Pappas, Eds., vol. 211. PMLR, Jun. 2023, pp. 248–258.
- [51] D. Narasimha, K. Lee, D. Kalathil, and S. Shakkottai, “Multi-Agent Learning via Markov Potential Games in Marketplaces for Distributed Energy Resources,” in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6350–6357.
- [52] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras, “Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games,” in *International Conference on Learning Representations*, 2022.
- [53] Y. Tian, Y. Wang, T. Yu, and S. Sra, “Online learning in unknown markov games,” in *International conference on machine learning*. PMLR, 2021, pp. 10 279–10 288.
- [54] J. Grossklags, B. Johnson, and N. Christin, “The Price of Uncertainty in Security Games,” in *Economics of Information Security and Privacy*, T. Moore, D. Pym, and C. Ioannidis, Eds. Boston, MA: Springer US, 2010, pp. 9–32.
- [55] A. Karapetyan, A. Tsiamis, E. C. Balta, A. Iannelli, and J. Lygeros,

- “Implications of Regret on Stability of Linear Dynamical Systems,” *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 2583–2588, 2023.
- [56] A. Karapetyan, D. Bolliger, A. Tsiamis, E. C. Balta, and J. Lygeros, “Online Linear Quadratic Tracking With Regret Guarantees,” *IEEE Control Systems Letters*, vol. 7, pp. 3950–3955, 2023.
- [57] M. Nonhoff and M. A. Müller, “On the relation between dynamic regret and closed-loop stability,” *Systems & Control Letters*, vol. 177, p. 105532, 2023.
- [58] Y. Li, J. A. Preiss, N. Li, Y. Lin, A. Wierman, and J. S. Shamma, “Online switching control with stability and regret guarantees,” in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*. PMLR, Jun. 2023, pp. 1138–1151, iSSN: 2640-3498.
- [59] D. Baby and Y.-X. Wang, “Optimal dynamic regret in LQR control,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS ’22. Red Hook, NY, USA: Curran Associates Inc., 2024, event-place: New Orleans, LA, USA.
- [60] Y. Jedra and A. Proutiere, “Minimal Expected Regret in Linear Quadratic Control,” in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 10 234–10 321, iSSN: 2640-3498.
- [61] C. Chen, *Linear System Theory and Design*, ser. Linear System Theory and Design. Oxford University Press, 1999.
- [62] G. F. Franklin, A. Emami-Naeini, and J. D. Powell, *Feedback control of dynamic systems Gene F. Franklin, Stanford University, J. David Powell, Stanford University, Abbas Emami-Naeini, SC Solutions, Inc.*, eighth edition, global edition ed. New York, NY: Pearson, 2020.
- [63] S. Lecleach, M. Schwager, and Z. Manchester, “LUCIDGames: Online Unscented Inverse Dynamic Games for Adaptive Trajectory Prediction and Planning,” *IEEE Robotics and Automation Letters*, Apr. 2021.
- [64] T. Zhao and T. L. Molloy, “Extended Kalman Filtering for Recursive Online Discrete-Time Inverse Optimal Control,” in *2024 American Control Conference (ACC)*, 2024, pp. 1212–1218.
- [65] K. REIF, F. SONNEMANN, and R. UNBEHAUEN, “An EKF-Based Nonlinear Observer with a Prescribed Degree of Stability,” *Automatica*, vol. 34, no. 9, pp. 1119–1123, 1998.

- 
- [66] C. M. Kellett and P. Braun, *Introduction to nonlinear control: Stability, control design, and estimation*. Princeton University Press, 2023.
- [67] Y. Chen, T. L. Molloy, T. Summers, and I. Shames, “Regret Analysis of Online LQR Control via Trajectory Prediction and Tracking: Extended Version,” Nov. 2022.
- [68] G. Belgioioso, D. Liao-McPherson, M. H. de Badyn, S. Bolognani, R. S. Smith, J. Lygeros, and F. Dörfler, “Online Feedback Equilibrium Seeking,” *IEEE Transactions on Automatic Control*, vol. 70, no. 1, pp. 203–218, 2025.
- [69] T. D. Woodbury, “Estimation-Based Solutions to Incomplete Information Pursuit-Evasion Games,” PhD Thesis, 2019, iISBN: 9798834007807 Publication Title: ProQuest Dissertations and Theses.
- [70] Y. Chen, T. L. Molloy, and I. Shames, “Two-Player Dynamic Potential LQ Games with Sequentially Revealed Costs,” 2025.
- [71] D. Monderer and L. S. Shapley, “Potential Games,” *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, May 1996.
- [72] T. Başar and Q. Zhu, “Prices of Anarchy, Information, and Cooperation in Differential Games,” *Dynamic Games and Applications*, vol. 1, no. 1, pp. 50–73, Mar. 2011.
- [73] H. Liao, Y. Li, Z. Li, C. Wang, Z. Cui, S. E. Li, and C. Xu, “A Cognitive-Based Trajectory Prediction Approach for Autonomous Driving,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4632–4643, 2024.
- [74] A. C. Thompson, “On Certain Contraction Mappings in a Partially Ordered Vector Space,” *Proceedings of the American Mathematical Society*, vol. 14, no. 3, pp. 438–443, 1963.
- [75] K. Krauth, S. Tu, and B. Recht, “Finite-time Analysis of Approximate Policy Iteration for the Linear Quadratic Regulator,” *arXiv:1905.12842 [cs, math, stat]*, May 2019, arXiv: 1905.12842.
- [76] R. A. Horn, *Matrix analysis Roger A. Horn, Charles R. Johnson*, 2nd ed. Cambridge: Cambridge University Press, 2013.
- [77] H. Lee and Y. Lim, “Invariant metrics, contractions and nonlinear matrix equations,” *Nonlinearity*, vol. 21, no. 4, p. 857, Mar. 2008.



---

# Appendix. Proof for Online LQ Optimal Control

---

## A.1 Proof of Theorem 2.3.1

Before stating the proof of the theorem, we introduce several necessary propositions and lemmas describing properties of the matrices in Proposition 2.3.1. We first start with proofs of Proposition 2.3.1 and Lemma 2.3.1.

*Proof of Proposition 2.3.1:* Follows directly from [18, Chapter 2.4]. ■

*Proof of Lemma 2.3.1:* The proof is identical to that of [12, Appendix D, Proposition 11]. ■

The following lemmas reveal a matrix-norm upper bound for  $P_{\tau|t_0} - P_{\tau|t}$  and  $K_{\tau|t} - K_{\tau|t_0}$  for any  $0 \leq \tau \leq t \leq t_0 \leq T - 1$ . These upper bounds imply the exponential decays of  $\|K_{\tau|t} - K_{\tau|t_0}\|$  and  $\|P_{\tau|t_0} - P_{\tau|t}\|$  with respect to  $t - \tau$ .

**Lemma A.1.1.** *For any  $T \geq 1$  and  $0 \leq \tau \leq t \leq t_0 \leq T - 1$ . Consider  $\gamma$  defined by (2.11),  $Q_{\min}$  and  $P_{\max}$  defined by (2.1) and (2.8), respectively. The distance between  $P_{\tau|t}$  and  $P_{\tau|t_0}$  from Proposition 2.3.1 satisfies  $\|P_{\tau|t} - P_{\tau|t_0}\| \leq \frac{\lambda_{\max}^2(P_{\max})}{\lambda_{\min}(Q_{\min})} \gamma^{t+1-\tau}$ , and the distance between  $K_{\tau|t}$  and  $K_{\tau|t_0}$  satisfies  $\|K_{\tau|t} - K_{\tau|t_0}\| \leq C_K \gamma^{t-\tau}$ .*

*Proof.* Before proceeding with the proof, we first define the Thompson metric  $\delta_{\infty}(\cdot, \cdot)$  as  $\delta_{\infty}(X, Y) := \|\log(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})\|_{\infty}$  for positive semi-definite matrices  $X$  and  $Y$  [74, Section 2], where  $\|\cdot\|_{\infty}$  denotes the matrix infinity-norm. The following steps show that as  $t - \tau$  increases, under the metric  $\delta_{\infty}$ , the distance between  $P_{\tau|t}$  and  $P_{\tau|t_0}$  is contracting with a rate between 0 and 1. Since  $Q_{\tau}, R_{\tau}, P_{\tau|t_0}, P_{\tau|t}$  are positive definite, by applying [75, Lemma D.2], the recursive relation between  $\delta_{\infty}(P_{\tau|t}, P_{\tau|t_0})$  and  $\delta_{\infty}(P_{\tau+1|t_0}, P_{\tau+1|t})$  is such that  $\delta_{\infty}(P_{\tau|t}, P_{\tau|t_0}) \leq \gamma \delta_{\infty}(P_{\tau+1|t_0}, P_{\tau+1|t})$ , where  $\gamma$  is defined by (2.11). Furthermore, we have  $\delta_{\infty}(P_{\tau|t_0}, P_{\tau|t}) \leq \gamma^{t-\tau+1} \delta_{\infty}(P_{t+1|t_0}, P_{t+1|t})$ .

From [12, Lemma 6],

$$\delta_\infty(P_{\tau|t_0}, P_{\tau|t}) \leq \gamma^{t-\tau+1} \log\left(\frac{\lambda_{\max}(P_{\max})}{\lambda_{\min}(Q_{\min})}\right) =: c.$$

To upper bound  $\|P_{\tau|t_0} - P_{\tau|t}\|$ , we can replace  $A, B, U$  and  $c$  from [12, Lemma 7] with  $P_{\tau|t_0}, P_{\tau|t}, P$  and  $\gamma^{t-\tau+1} \log\left(\frac{\lambda_{\max}(P_{\max})}{\lambda_{\min}(Q_{\min})}\right)$ , respectively, that yields  $\|P_{\tau|t_0} - P_{\tau|t}\| \leq \gamma^{t-\tau+1} \frac{\lambda_{\max}^2(P_{\max})}{\lambda_{\min}(Q_{\min})}$ . To upper bound  $\|K_{\tau|t} - K_{\tau|t_0}\|$ , note that

$$\begin{aligned} \|K_{\tau|t} - K_{\tau|t_0}\| &= \|-(R_\tau + B^\top P_{\tau+1|t} B)^{-1} B^\top P_{\tau+1|t} A \\ &\quad + (R_\tau + B^\top P_{\tau+1|t_0} B)^{-1} B^\top P_{\tau+1|t_0} A\|. \end{aligned}$$

Let  $G_1 = (R_\tau + B^\top P_{\tau+1|t} B)$  and  $G_2 = (R_\tau + B^\top P_{\tau+1|t_0} B)$ , we have

$$\begin{aligned} \|K_{\tau|t} - K_{\tau|t_0}\| &= \|G_1^{-1} B^\top P_{\tau+1|t} - G_2^{-1} B^\top P_{\tau+1|t_0}\| \\ &= \|G_1^{-1} G_2^{-1} G_2 B^\top P_{\tau+1|t} - G_2^{-1} G_1^{-1} G_1 B^\top P_{\tau+1|t_0}\| \\ &\leq \|G_2^{-1} G_1^{-1}\| \| (G_1 G_2) (G_1^{-1} G_2^{-1}) G_2 B^\top P_{\tau+1|t} - G_1 B^\top P_{\tau+1|t_0} \| \\ &\leq \|G_2^{-1} G_1^{-1}\| (\| (G_1 G_2 - G_2 G_1) G_1^{-1} B^\top P_{\tau+1|t} \| + \| G_2 B^\top P_{\tau+1|t} - G_1 B^\top P_{\tau+1|t_0} \|) \\ &\leq \|G_2^{-1} G_1^{-1}\| \| G_2 B^\top P_{\tau+1|t} - G_1 B^\top P_{\tau+1|t_0} \|, \end{aligned} \tag{A.1}$$

where the last two inequalities are due to  $G_1$  and  $G_2$  being symmetric matrices,  $G_1 G_2 - G_2 G_1$  being skew symmetric, and the induced 2-norm of a skew symmetric matrix being 0. Moreover,

$$\begin{aligned} &\|G_2 B^\top P_{\tau+1|t} - G_1 B^\top P_{\tau+1|t_0}\| \\ &\leq \|R_\tau B^\top (P_{\tau+1|t} - P_{\tau+1|t_0})\| + \|B\| \|P_{\tau+1|t_0} (B B^\top) P_{\tau+1|t} - P_{\tau+1|t} (B B^\top) P_{\tau+1|t_0}\| \\ &\leq \|R_\tau B^\top\| \|P_{\tau+1|t} - P_{\tau+1|t_0}\|, \end{aligned} \tag{A.2}$$

where the last step is due to  $P_{\tau+1|t_0} (B B^\top) P_{\tau+1|t} - P_{\tau+1|t} (B B^\top) P_{\tau+1|t_0}$  being skew symmetric. Remember that  $C_K = \|G_2^{-1} G_1^{-1}\| \|R_{\max} B^\top\| \frac{\lambda_{\max}^2(P_{\max})}{\lambda_{\min}(Q_{\min})}$ . Substituting (A.2) into (A.1) gives that  $\|K_{\tau|t} - K_{\tau|t_0}\| \leq C_K \gamma^{t-\tau}$ .  $\square$

**Lemma A.1.2.** *For any  $T \geq 1$  and  $0 \leq \tau \leq t \leq T-1$ , consider matrices  $A$  and  $B$  from (1.1),  $K_{\tau|t}$  from Proposition 2.3.1,  $\eta$  defined by (2.10) and  $C$  defined by (2.12), respectively. For any  $0 \leq t_0 \leq t_1 \leq T-1$ , we have  $\|\prod_{\tau=t_0}^{t_1} (A + B K_{\tau|t})\| \leq C \eta^{t_1-t_0+1}$ .*

*Proof.* Similar to that of [12, Proposition 2].  $\square$

**Lemma A.1.3.** *Consider the optimal control problem (2.4) and suppose that Assumption 2.2.1 holds. Then the control gain matrix  $K_{\tau|t}$  for  $0 \leq t, \tau \leq T-1$  satisfies*

$$\|K_{\tau|t}\| \leq \frac{\sigma_{\max}(A)}{\sigma_{\min}(B)}.$$

*Proof.* By Assumption 2.2.1 and Proposition 2.3.1, respectively, matrices  $R_{\tau|t}$  and  $P_{\tau+1|t}$  are positive definite. Then, by [70, Lemma 6, Appendix], and the fact that  $B$  is full-rank, we have  $\sigma_{\min}(B) \neq 0$  and  $\|(R_{\tau|t} + B^\top P_{\tau+1|t} B)^{-1} P_{\tau+1|t} B\| \leq \frac{1}{\sigma_{\min}(B)}$ . Thus,

$$\|K_{\tau|t}\| \leq \|(R_{\tau|t} + B^\top P_{\tau+1|t} B)^{-1} P_{\tau+1|t} B\| \|A\| \leq \frac{\sigma_{\max}(A)}{\sigma_{\min}(B)}.$$

□

**Lemma A.1.4.** *Consider the tracking control gain  $K_t$  defined in (2.7) for  $0 \leq t \leq T - 1$ . Then the norm of  $K_t$  satisfies  $\|K_t\| \leq \frac{C_f + \|A\|}{\sigma_{\min}(B)}$ .*

*Proof.* By (2.7), we have  $\|A + BK_t\| \leq C_f$  for any  $0 \leq t \leq T - 1$ . This immediately implies  $\|BK_t\| \leq C_f + \|A\|$ . Then,

$$\begin{aligned} \|BK_t\| &= \sqrt{\lambda_{\max}(K_t^\top B^\top BK_t)} \stackrel{(i)}{=} \sqrt{\lambda_{\max}(BK_t K_t^\top B^\top)} \\ &\stackrel{(ii)}{\geq} \sqrt{\lambda_{\max}(K_t^\top K_t)} \sigma_{\min}(B) = \|K_t\| \sigma_{\min}(B). \end{aligned}$$

Step (i) uses the property that the maximum eigenvalue of  $H^\top H$  is the same as the maximum eigenvalue of  $HH^\top$  for any real matrix  $H$ . Step (ii) is a direct consequence of [76, Theorem 4.5.9] (taking  $\theta_k$  from the theorem as the minimal singular value of  $B$ ). Thus,  $\|K_t\| \leq \frac{C_f + \|A\|}{\sigma_{\min}(B)}$ . As  $B$  is full-rank,  $\sigma_{\min}(B) \neq 0$  and the bound is well-defined. □

The next lemma bounds the distance between the states generated by (2.6) and the optimal states generated by solving (1.2), together with the distance between the predicted states generated by (2.4) and the optimal states generated by (1.2).

**Lemma A.1.5.** *For any  $T \geq 1$ ,  $0 \leq W \leq T - 1$  and  $0 \leq t \leq T$ , consider state  $x_t$  that generated by (2.6) and the optimal state  $x_t^*$  as an element of the solution from solving (1.2). Then for  $C, C_K, \gamma$  and  $\eta$  from Theorem 2.3.1, the distance between state  $x_t$  and optimal state  $x_t^*$  satisfies*

$$\begin{aligned} \|x_t - x_t^*\| &\leq \tag{A.3} \\ &\frac{C^2 C_K \|\bar{x}_0\| \gamma^W}{\gamma - 1} (\eta^{t-1} \gamma (\gamma^t - 1)) + C_f \left( \frac{\eta \gamma}{q} \left( \frac{q^{t-1} - (\eta \gamma)^{t-1}}{q - \eta \gamma} \right) - \frac{\eta}{q} \left( \frac{q^{t-1} - \eta^{t-1}}{q - \eta} \right) \right). \end{aligned}$$

Moreover, the distance between the predicted state  $x_{t|t}$  as an element of solution from (2.4) and the optimal state  $x_t^*$  satisfies  $\|x_{t|t} - x_t^*\| \leq \frac{C^2 C_K \|\bar{x}_0\| \gamma^W}{\gamma - 1} \eta^{t-1} \gamma (\gamma^t - 1)$ .

*Proof.* The following proof provides closed-form expressions for the differences  $x_t - x_t^*$  and  $x_{t|t} - x_t^*$  using the recursive definition of  $x_t$  and  $x_{t|t}$  via the LTI system (1.1), then uses Lemma A.1.2 to establish upper bounds on their norms. Note that the computation of  $x_t$  depends on  $x_{t|t}$ , so we upper bound  $\|x_t - x_{t|t}\|$  and  $\|x_{t|t} - x_{t|T-1}\|$ , then use the relation  $x_t - x_t^* = x_t - x_{t|T-1} = x_t - x_{t|t} + x_{t|t} - x_{t|T-1}$  to upper bound  $\|x_t - x_t^*\|$  along with the triangular inequality. Define

$$y_t := x_t - x_{t|t} \text{ and } \theta_{i|p,q} := x_{i|p} - x_{i|q}, \quad (\text{A.4})$$

where  $0 \leq p \leq q \leq T-1$  and  $0 \leq i \leq T-1$ . Consequently,  $y_0 = 0$ , and via inspection we have  $y_{t+1} = \sum_{j=1}^{t+1} \prod_{\tau=0}^{t-j} (A + BK_\tau) \theta_{j|j-1,j}$ . A closed-form expression for  $\theta_{i|p,q}$ , noting  $\theta_{0|p,q} = 0$ , is then

$$\begin{aligned} \theta_{i+1|p,q} &= x_{i+1|p} - x_{i+1|q} \\ &= (A + BK_{i|p})x_{i|p} - (A + BK_{i|q})x_{i|q} \\ &= (A + BK_{i|p})(\theta_{i|p,q} + x_{i|q}) - (A + BK_{i|q})x_{i|q} \\ &= (A + BK_{i|p})\theta_{i|p,q} + B(K_{i|p} - K_{i|q})x_{i|q}. \end{aligned}$$

In the following, for any given matrices  $\{a_0, \dots, a_N\}$ , let  $I$  be the identity matrix, define the product operator

$$\prod_{n=m_1}^{m_2} a_n = \begin{cases} I & \text{if } m_1 > m_2, \\ a_{m_1} & \text{if } m_1 = m_2, \\ a_{m_1} a_{m_1+1} \cdots a_{m_2} & \text{if } m_1 < m_2. \end{cases}$$

The term  $\theta_{i+1|p,q}$  can thus be further expanded as

$$\theta_{i+1|p,q} = \sum_{n=0}^i \left( \prod_{m=n+1}^i (A + BK_{m|p}) \right) B(K_{n|p} - K_{n|q}) \left( \prod_{m=0}^{n-1} (A + BK_{m|q}) \right) \bar{x}_0.$$

By Lemma A.1.2, we have  $\|\prod_{m=n+1}^i (A + BK_{m|p})\| \leq C\eta^{i-n}$  and  $\|\prod_{m=0}^{n-1} (A + BK_{m|q})\| \leq C\eta^n$ . By Lemma A.1.1, we have  $\|B(K_{n|p} - K_{n|q})\| \leq C_K \gamma^{p-n}$ . Thus,  $\|\theta_{i+1|p,q}\| \leq \frac{C^2 C_K \gamma^p \eta^i}{1 - \frac{1}{\gamma}} (1 - \frac{1}{\gamma^{i+1}})$ . Choosing  $i = t$ ,  $p = t$ , and  $q = T-1$ , results in

$$\|\theta_{t|t,T-1}\| \leq \frac{C^2 C_K \|\bar{x}_0\| \eta^{t-1} \gamma^{W+1}}{\gamma - 1} (\gamma^t - 1). \quad (\text{A.5})$$

Moreover,  $\|\theta_{i|i-1,i}\| \leq \frac{C^2 C_K \|\bar{x}_0\| \eta^{i-1} \gamma^W}{\gamma - 1} (\gamma^i - 1)$ . Define  $\mu_{i,t} = \|\prod_{\tau=0}^{t-i-1} (A + BK_\tau)\|$ .

Combining the above, we have

$$\begin{aligned} \|x_t - x_{t|t}\| &\leq \sum_{i=1}^t \left\| \prod_{\tau=0}^{t-i-1} (A + BK_\tau) \theta_{|i-1+W, i+W} \right\| \\ &\leq \frac{C^2 C_K \|\bar{x}_0\| \gamma^W}{\gamma - 1} \sum_{i=1}^t \mu_{i,t} \eta^i (\gamma^i - 1). \end{aligned} \quad (\text{A.6})$$

$$\text{Thus, } \|x_t - x_t^*\| \leq \frac{C^2 C_K \|\bar{x}_0\| \gamma^W}{\gamma - 1} \left( \sum_{i=1}^t \mu_{i,t} \eta^i (\gamma^i - 1) + \eta^{t-1} \gamma (\gamma^t - 1) \right).$$

Furthermore, it is given that  $\mu_{s,r} = \|\prod_{\tau=s}^r (A + BK_\tau)\| \leq C_f q^{r-s}$  from the choice of gains  $K_\tau$  in (2.6).

Therefore, the upper bound of the distance between the state vector and the optimal state vector is given by

$$\|x_t - x_t^*\| \leq \frac{C^2 C_K \|\bar{x}_0\| \gamma^W}{\gamma - 1} \left( \eta^{t-1} \gamma (\gamma^t - 1) + C_f \left[ \frac{\eta \gamma}{q} \left( \frac{q^{t-1} - (\eta \gamma)^{t-1}}{q - \eta \gamma} \right) - \frac{\eta}{q} \left( \frac{q^{t-1} - \eta^{t-1}}{q - \eta} \right) \right] \right).$$

The proof is complete.  $\square$

**Lemma A.1.6** (Cost Difference Lemma [12]). *For any  $T \geq 1$  and  $0 \leq t \leq T$ , consider  $B$  from (1.1),  $R_t$  from (1.2),  $K_t^*$  be  $K_{t|T}$  and  $P_t^*$  be  $P_{t|T}$  from Proposition 2.3.1, respectively. Let  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  be the control policy defined in (2.26), states  $\{x_t\}_{t=0}^{T-1}$  be generated from control sequence  $\{u_t\}_{t=0}^{T-1}$  for the linear system (1.1) where each control being  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$ . Setting  $\bar{u}_t = K_t^* x_t$ , the regret defined by (2.2) satisfies*

$$\text{Regret}_T(\Pi) = \sum_{t=0}^{T-1} (u_t - \bar{u}_t)^\top (R_t + B^\top P_{t+1}^* B) (u_t - \bar{u}_t).$$

We also need the following elementary result.

**Lemma A.1.7.** *For any  $a_1, a_2, a_3 \in \mathbb{R}$ ,  $(a_1 + a_2 + a_3)^2 \leq \frac{10}{3}(a_1^2 + a_2^2 + a_3^2)$ .*

*Proof.* Note  $(a_1 + a_2 + a_3)^2 \leq 2(a_1 + a_2)^2 + 2a_3^2 \leq 4a_1^2 + 4a_2^2 + 2a_3^2$ . Similarly,  $(a_1 + a_2 + a_3)^2 \leq 4a_1^2 + 2a_2^2 + 4a_3^2$  and  $(a_1 + a_2 + a_3)^2 \leq 2a_1^2 + 4a_2^2 + 4a_3^2$ . Combining these gives  $(a_1 + a_2 + a_3)^2 \leq \frac{4+4+2}{3}(a_1^2 + a_2^2 + a_3^2)$ .  $\square$

*Proof of Theorem 2.3.1:* In light of Lemma A.1.7:

$$\begin{aligned} \|u_t - \bar{u}_t\|^2 &\leq \frac{10}{3} (\|K_{t|t} - K_t\|^2 \|x_{t|t} - x_t^*\|^2 \\ &\quad + \|K_t^* - K_t\|^2 \|x_t - x_t^*\|^2 + \|K_t^* - K_{t|t}\|^2 \|x_t^*\|^2). \end{aligned} \quad (\text{A.7})$$

By Lemmas A.1.1 and A.1.5, the summation of  $\|x_t - x_t^*\|^2$  from  $t = 1$  to  $t = T$  is upper bounded by

$$2\left(\frac{C^2 C_K \|\bar{x}_0\| \gamma^{W+1}}{(\gamma - 1)}\right)^2 \left( \gamma^2 S_T(\eta^2 \gamma^2) - 2\gamma S_T(\eta^2 \gamma) + S_T(\eta^2) \right) + \frac{10C_f^2}{3} \left( \left( \frac{\eta\gamma}{q(q - \eta\gamma)} - \frac{\eta}{q(q - \eta)} \right)^2 S_T(q^2) + \frac{(\eta\gamma)^2 S_T(\eta^2 \gamma^2)}{q^2(q - \eta\gamma)^2} + \frac{\eta^2 S_T(\eta^2)}{q^2(q - \eta)^2} \right). \quad (\text{A.8})$$

The summation of  $\|x_{t|t} - x_t^*\|^2$  from  $t = 1$  to  $t = T$  is upper bounded by

$$\gamma^2 S_T(\eta^2 \gamma^2) - 2\gamma S_T(\eta^2 \gamma) + S_T(\eta^2). \quad (\text{A.9})$$

By Lemma A.1.2, we have

$$\sum_{t=1}^T \|x_t^*\|^2 \leq (C^2 \|\bar{x}_0\|)^2 \eta^2 S_T(\eta^2). \quad (\text{A.10})$$

Substitute (A.8), (A.9) and (A.10) in (A.7). By Lemmas A.1.3 and A.1.4, we have  $\|K_{t|t} - K_t\|^2 \leq \alpha_K$  and  $\|K_t^* - K_t\|^2 \leq \alpha_K$ , where  $\alpha_K$  is defined in (2.13). From Lemma A.1.6 and taking summation of the RHS of (A.7) from  $t = 1$  to  $t = T$ , the  $\text{Regret}_T(\Pi)$  can be upper bounded by  $\gamma^{2W} \Psi$ , where

$$\begin{aligned} \Psi = & \frac{10D\|\bar{x}_0\|^2}{3} \left[ (\alpha_1 + \alpha_2) \frac{(C^2 C_K \gamma)^2}{(\gamma - 1)^2} [\gamma^2 S_T(\eta^2 \gamma^2) - 2\gamma S_T(\eta^2 \gamma) \right. \\ & + S_T(\eta^2)] + \frac{10C_f^2}{3} \left[ \left( \frac{\eta\gamma}{q(q - \eta\gamma)} - \frac{\eta}{q(q - \eta)} \right)^2 S_T(q^2) \right. \\ & \left. + \frac{(\eta\gamma)^2 S_T(\eta^2 \gamma^2)}{q^2(q - \eta\gamma)^2} + \frac{\eta^2 S_T(\eta^2)}{q^2(q - \eta)^2} \right] + (C_K C^2)^2 \eta^2 S_T(\eta^2) \Big]. \quad (\text{A.11}) \end{aligned}$$

The proof is complete. ■

## A.2 Proof of Proposition 2.3.2

Let  $F$  and  $F'$  represent the right-hand sides of (2.14) and [12, Theorem 1, Equation (15)], respectively. Note that

$$\begin{aligned} F' & \geq 4\|\bar{x}_0\|^2 D \|A\|^2 \|B\|^2 \lambda_{\max}^{10}(P_{\max}) C^4 \left( \frac{\|BR_{\min}^{-1}B^\top\|^2 (1 + \|BR_{\min}^{-1}B^\top\|^2) (\gamma^W + \eta^W)^2}{\lambda_{\min}^2(R_{\min}) \lambda_{\min}^4(Q_{\min}) (1 - \eta)^2} \right) \\ & \geq \frac{4\|\bar{x}_0\|^2 D \|A\|^2 \|B\|^2 \lambda_{\max}^{10}(Q_{\max}) C^4 \|BR_{\min}^{-1}B^\top\|^2 \gamma^{2W}}{\lambda_{\min}^2(R_{\min}) \lambda_{\min}^4(Q_{\min})}, \end{aligned}$$

and

$$F \leq \frac{10D\gamma^{2W}\|\bar{x}_0\|^2 C^4 C_K^2}{3} \left[ \left(1 + \frac{\alpha_K}{(1-\gamma)^2}\right) \left(\frac{1}{1-\eta^2}\right) + \frac{10C_f^2}{3q^2(q-\eta)^2} \left( \frac{1}{(q-\eta)^2(1-q)^2} + \frac{2}{(1-\eta^2)} \right) \right].$$

Thus, if

$$\lambda_{max}^{10}(Q_{max}) \geq \frac{5 \left[ \left(1 + \frac{\alpha_K}{(1-\gamma)^2}\right) \left(\frac{1}{1-\eta^2}\right) + \frac{10C_f^2}{3q^2(q-\eta)^2} \left( \frac{1}{(q-\eta)^2(1-q)^2} + \frac{2}{(1-\eta^2)} \right) \right]}{6(C_K^2 \lambda_{min}^2(R_{min}) \lambda_{min}^4(Q_{min}))^{-1} \|A\|^2 \|B\|^2 \|BR_{min}^{-1} B^\top\|^2},$$

then it follows that  $F \leq F'$ . ■

### A.3 Proof of Theorem 2.3.2

We first state a stochastic cost difference lemma.

**Lemma A.3.1** (Stochastic Cost Difference Lemma). *For any  $T \geq 1$  and  $0 \leq t \leq T$ , consider  $B$  from (1.1),  $R_t$  from (1.2), and let  $K_t^* = K_{t|T}$  and  $P_t^* = P_{t|T}$  with  $K_{t|T}$  and  $P_{t|T}$  given in Proposition 2.3.1. We further consider the random sequence  $\{w_t\}_{t=0}^{T-1}$  where  $w_t \in \mathbb{R}^n$ ,  $\mathbf{E}(w_t) = 0$  and  $\mathbf{E}(w_t w_t^\top) = Cov_w$ . Let  $\Pi = \{\pi_t\}_{t=0}^{T-1}$  be the control policy defined in (2.26), and the states  $\{x_t\}_{t=0}^{T-1}$  be generated by the system (1.1) with controls  $u_t = \pi_t(x_t, \mathcal{H}_{t,W})$ . The regret defined by (2.2) satisfies*

$$ExpRegret_T(\Pi) = \mathbf{E} \left[ \sum_{t=0}^{T-1} (u_t - \bar{u}_t)^\top (R_t + B^\top P_{t+1}^* B) (u_t - \bar{u}_t) \right],$$

where  $\bar{u}_t = K_t^* x_t$ .

*Proof.* With a slight abuse of notation, for any given  $x$ , define  $V_T(x) = x^\top Q_T x$ ,  $\Phi_t(x, u) := x^\top Q_t x + u^\top R_t u + \mathbf{E}(V_{t+1}(Ax + Bu + w_t))$ , and  $V_t(x) := \min_{\pi} \Phi_t(x, \pi(x))$ .

The expected regret in terms of  $\Phi_t(x_t, u_t)$  and  $V_t(x_t)$  is thus

$$\begin{aligned}
\text{ExpRegret}_T(\Pi) &= \mathbf{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_t x_t + u_t^\top R_t u_t + x_T^\top Q_T x_T\right] - V_0(\bar{x}_0) \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_t x_t + u_t^\top R_t u_t\right] + \sum_{t=0}^{T-1} \mathbf{E}(V_{t+1}(x_{t+1})) - \mathbf{E}(V_t(x_t)) \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_t x_t + u_t^\top R_t u_t + V_{t+1}(x_{t+1}) - V_t(x_t)\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} \Phi_t(x_t, u_t) - V_t(x_t)\right]. \tag{A.12}
\end{aligned}$$

Given state  $x_t$  and control  $u_t$ , following the proofs in [25, Chapter 5, Page 230], we have  $\Phi_t(x_t, u_t) = x_t^\top Q_t x_t + u_t^\top R_t u_t + \mathbf{E}(x_{t+1}^\top P_{t+1}^* x_{t+1} + \sum_{\tau=t}^{T-1} w_\tau^\top P_{\tau+1}^* w_\tau)$ ,

and  $V_t(x_t) = x_t^\top Q_t x_t + \bar{u}_t^\top R_t \bar{u}_t + \mathbf{E}((Ax_t + B\bar{u}_t)^\top P_{t+1}^* (Ax_t + B\bar{u}_t) + \sum_{\tau=t}^{T-1} w_\tau^\top P_{\tau+1}^* w_\tau)$ .

Substituting the above into (A.12), we have

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=0}^{T-1} \Phi_t(x_t, u_t) - V_t(x_t)\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_t x_t + u_t^\top R_t u_t - x_t^\top Q_t x_t - \bar{u}_t^\top R_t \bar{u}_t + \mathbf{E}(x_{t+1}^\top P_{t+1}^* x_{t+1} + \sum_{\tau=t}^{T-1} w_\tau^\top P_{\tau+1}^* w_\tau) \right. \\
&\quad \left. - \mathbf{E}((Ax_t + B\bar{u}_t)^\top P_{t+1}^* (Ax_t + B\bar{u}_t) - \sum_{\tau=t}^{T-1} w_\tau^\top P_{\tau+1}^* w_\tau)\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} (u_t - \bar{u}_t)^\top (R_t + B^\top P_{t+1}^* B)(u_t - \bar{u}_t) + 2(u_t - \bar{u}_t)^\top (R_t + B^\top P_{t+1}^* B)(\bar{u}_t + B^\top P_{t+1}^* Ax_t)\right] \\
&= \mathbf{E}\left[\sum_{t=0}^{T-1} (u_t - \bar{u}_t)^\top (R_t + B^\top P_{t+1}^* B)(u_t - \bar{u}_t)\right].
\end{aligned}$$

□

From the above lemma, we have

$$\text{ExpRegret}_T(\Pi) \leq D \sum_{t=0}^{T-1} \mathbf{E}(\|u_t - \bar{u}_t\|^2). \tag{A.13}$$

To establish the preceding lemmas for proving Theorem 2.3.2, we require the closed-form expression of  $x_{t|q}$ ,  $y_t$  and  $\theta_{t|p,q}$  for  $0 \leq t \leq T-1$  and  $0 \leq p \leq q \leq T-1$ . The state variable  $x_{t|q}$  can be expressed as

$$x_{t|q} = \prod_{j=0}^{t-1} (A + BK_{j|q}) \bar{x}_0 + \sum_{r=0}^{t-1} \left( \prod_{j=r+1}^{t-1} (A + BK_{j|q}) \right) w_r. \tag{A.14}$$

To calculate  $y_t$ , note that  $x_0 = x_{0|t} = \bar{x}_0$ , we again have  $y_0 = 0$  and  $\theta_{0|p,q} = 0$ . By repeating the calculation of  $y_{t+1}$  as in Lemma A.1.5, we have

$$y_{t+1} = \sum_{j=1}^{t+1} \prod_{\tau=0}^{t-j} (A + BK_{\tau}) \theta_{j|j-1,j}. \quad (\text{A.15})$$

To calculate  $\theta_{t|p,q}$ , for the case of  $1 \leq t \leq p$ ,

$$\begin{aligned} \theta_{t|p,q} &= (A + BK_{t-1|p})x_{t-1|p} + w_{t-1} \\ &\quad - (A + BK_{t-1|q})x_{t-1|q} - w_{t-1} \\ &= (A + BK_{t-1|p})(\theta_{t-1|p,q} + x_{t-1|q}) \\ &\quad - (A + BK_{t-1|q})x_{t-1|q} \\ &= (A + BK_{t-1|p})\theta_{t-1|p,q} + B(K_{t-1|p} - K_{t-1|q})x_{t-1|q} \\ &= \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + BK_{m|t-1}) \right) B(K_{n|t-1} - K_{n|t})x_{n|q}. \end{aligned} \quad (\text{A.16})$$

We also need the closed-form expression for the special case of  $p = t - 1$  and  $q = t$ ,

$$\begin{aligned} \theta_{t|t-1,t} &= (A + BK_{t-1|t})x_{t-1|t-1} - (A + BK_{t-1|t})x_{t-1|t} - w_{t-1} \\ &= (A + BK_{t-1|t-1})\theta_{t-1|t-1,t} + B(K_{t-1|t-1} - K_{t-1|t})x_{t-1|t} - w_{t-1} \\ &= \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + BK_{m|t-1}) \right) B(K_{n|t-1} - K_{n|t})x_{n|t} - w_{t-1}. \end{aligned} \quad (\text{A.17})$$

We use the above calculations to help us establish the following two lemmas, and they will prove helpful for bounding the expected regret.

**Lemma A.3.2.** *Consider i.i.d. random variables  $\{w_t\}_{t=0}^{T-1}$  where  $w_t \in \mathbb{R}^n$ ,  $\mathbf{E}(w_k) = 0$  and  $\mathbf{E}(w_k w_k^\top) = \text{Cov}_w$ . Consider the linear system (1.1), initial condition  $\bar{x}_0 = 0$  and policy  $\Pi$  given by (2.6), there exist positive scalars  $C_{R2}$  and  $C'_{R2}$  such that the expected regret defined by (A.13) satisfies*

$$\text{ExpRegret}_T(\Pi) \leq (C_{R2}\gamma^{2W} + C'_{R2})T. \quad (\text{A.18})$$

*Proof.* The technique of proving the upper bound of this expected regret is similar to how we established the upper bound for Regret for the disturbance-free case. By Lemma A.1.7, A.3.1 and (A.13), we have  $\text{ExpRegret}(\Pi) \leq D \sum_{t=0}^{T-1} \mathbf{E}(\|u_t - \bar{u}_t\|^2) = D \sum_{t=0}^{T-1} \mathbf{E}(\|(K_{t|t} - K)(x_{t|t} - x_t^*) - (K_t^* - K)(x_t - x_t^*) + (K_t^* - K_{t|t})x_t^*\|^2) \leq \frac{10D}{3} \sum_{t=0}^{T-1} \|(K_{t|t} - K)\|^2 \mathbf{E}(\|x_{t|t} - x_t^*\|^2) + \|K_t^* - K\|^2 \mathbf{E}(\|x_t - x_t^*\|^2) + \|K_t^* -$

$K_{t|t}\|^2 \mathbf{E}(\|x_t^*\|^2)$ . Term  $\|K_t^* - K_{t|t}\|^2$  can be upper bounded by using Lemma A.1.1, and the upper bound of  $\|K_{t|t} - K_t\|^2$ ,  $\|K_t^* - K_t\|^2$  are referring to  $\alpha_1$  and  $\alpha_2$  from Theorem 2.3.1. Therefore, we can prove the lemma by establishing the upper bound of  $\mathbf{E}(\|x_{t|t} - x_t^*\|^2)$ ,  $\mathbf{E}(\|x_t - x_t^*\|^2)$  and  $\mathbf{E}(\|x_t^*\|^2)$ .

We start with upper bounding  $\mathbf{E}(\|x_{t|t} - x_t^*\|^2)$ . Recall that  $x_{t|t} - x_t^* = \theta_{t|t,T}$  from definition (A.4). Substituting (A.14) to (A.16) and let  $\bar{x}_0 = 0$ , we have

$$\begin{aligned} & \mathbf{E}(\|\theta_{t|t,T-1}\|^2) \\ &= \mathbf{E}\left(\left\|\sum_{n=0}^{t-1} \sum_{r=0}^{n-1} \left(\prod_{m=n+1}^{t-1} (A + BK_{m|t})\right) B(K_{n|t} - K_n^*) \left(\prod_{j=r+1}^{n-1} (A + BK_j^*)\right) w_r\right\|^2\right) \\ &\leq \frac{(C^2 C_K)^2 \gamma^{2W} \eta^{2t} \gamma^{2t}}{\eta^2} \sum_{\substack{n_1=0 \\ n_2=0}}^{t-1} \sum_{r_1=0}^{n_1-1} \sum_{r_2=0}^{n_2-1} \frac{\mathbf{E}(w_{r_1} w_{r_2}^\top)}{\gamma^{n_1+n_2} \eta^{r_1+r_2}} =: \kappa_{w\theta}. \end{aligned} \quad (\text{A.19})$$

Furthermore, the above summation can be written as the form of  $\gamma^{2W}(C_{\kappa_{w\theta}} + L_{\kappa_{w\theta}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t}))$ , where  $C_{\kappa_{w\theta}}$  is a non-negative scalar that independent of  $t$ , and  $L_{\kappa_{w\theta}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t})$  is a linear combination of  $\eta^t, \gamma^t, \eta^{2t}$  and  $\gamma^{2t}$ .

To bound  $\mathbf{E}(\|x_t - x_t^*\|^2)$ , substituting (A.17) into (A.14) gives

$$\begin{aligned} & \mathbf{E}(\|x_t - x_t^*\|^2) \\ &\leq 2 \mathbf{E}(\|x_t - x_{t|t}\|^2) + 2 \mathbf{E}(\|x_{t|t} - x_t^*\|^2) \\ &\leq \mathbf{E}\left(2 \left\|\sum_{i=1}^{t-1} \prod_{\tau=0}^{t-i-1} (A + BK_\tau) \sum_{n=0}^{i-1} \sum_{r=0}^{n-1} \left(\prod_{m=n+1}^{i-1} (A + BK_{m|i-1+W})\right) \right. \right. \\ &\quad \left. \left. B(K_{n|i-1+W} - K_n^*) \left(\prod_{j=r+1}^{n-1} (A + BK_j^*)\right) w_r - w_{i-1}\right\|^2\right) + 2\kappa_{w\theta} \\ &\leq \kappa_{wx}, \end{aligned} \quad (\text{A.20})$$

where

$$\kappa_{wx} := 2\kappa_{w\theta} + 2(C^2 C_K)^2 \eta^{2t} \sum_{\substack{i_1, i_2=1 \\ n_1, n_2=0}}^{t-1} \sum_{r_1, r_2=0}^{n_1-1, n_2-1} \frac{\mathbf{E}(w_{r_1} w_{r_2}^\top) \gamma^{i_1+i_2-n_1-n_2}}{\eta^{r_1+r_2}}. \quad (\text{A.21})$$

Moreover,

$$\mathbf{E}(\|x_t^*\|^2) = \mathbf{E}\left(\left\|\sum_{r=0}^{t-1} \left[\prod_{j=r+1}^{t-1} (A + BK_j^*)\right] w_r\right\|^2\right) \leq \kappa_{wx^*}, \quad (\text{A.22})$$

where  $\kappa_{wx^*} := C^4 \sum_{\substack{r_1=0 \\ r_2=0}}^{t-1} \mathbf{E}(w_{r_1} w_{r_2}^\top) \eta^{2t-r_1-1} \eta^{2t-r_2-1}$ . Terms  $\kappa_{wx}$  and  $\kappa_{wx^*}$  are of the form  $C'_{\kappa_{wx}} + \gamma^{2W}(C_{\kappa_{wx}} + L_{\kappa_{wx}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t}))$  and  $C_{\kappa_{wx^*}} + L_{\kappa_{wx^*}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t})$ , respec-

tively, where  $C_{\kappa_{wx}}$ ,  $C_{\kappa_{wx}^*}$  and  $C'_{\kappa_{wx}}$  are non-negative scalars that are independent of  $t$ ,  $L_{\kappa_{wx}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t})$  and  $L_{\kappa_{wx}^*}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t})$  are linear combinations of  $\eta^t, \gamma^t, \eta^{2t}$  and  $\gamma^{2t}$ .

Recall  $\alpha_1, \alpha_2$  and  $D$  defined in Theorem 2.3.1. By applying the elementary inequality from Lemma A.1.7, we can upper bound the expected regret as follows,

$$\begin{aligned}
& \text{ExpRegret}_T(\Pi) \\
& \leq \frac{10D}{3} \sum_{t=0}^{T-1} (\alpha_1 \mathbf{E}(\|x_{t|t} - x_t^*\|^2) + \alpha_2 \mathbf{E}(\|x_t - x_t^*\|^2) + C_K^2 \gamma^{2W} \mathbf{E}(\|x_t^*\|^2)) \\
& \leq \frac{10D}{3} \sum_{t=0}^{T-1} \alpha_1 \kappa_{w\theta} + C_K^2 \gamma^{2W} \kappa_{wx^*} + \alpha_2 \kappa_{wx} \\
& = (C_{R2} \gamma^{2W} + C'_{R2})T. \tag{A.23}
\end{aligned}$$

where  $\gamma^{2W} C_{R2} = \frac{10D}{3} (C_K^2 \kappa_{wx^*} + \alpha_1 \kappa_{w\theta} + \alpha_2 (C_{\kappa_{wx}} + L_{\kappa_{wx}}(\eta^t, \gamma^t, \eta^{2t}, \gamma^{2t})))$  and  $C'_{R2} = \frac{10D}{3} \alpha_2 C'_{wx}$ . The last inequality holds by substituting (A.19) and (A.20) in (A.13).  $\square$

**Lemma A.3.3.** *Consider the sequence of i.i.d. random variables  $\{w_k\}_{k=0}^{T-1}$  where  $w_k \in \mathbb{R}^n$ ,  $\mathbf{E}(w_k) = 0$  and  $\mathbf{E}(w_k w_k^\top) = \text{Cov}_w$ . Consider the linear system (1.1) and policy  $\Pi$  from (2.6), then for any  $\bar{x}_0 \in \mathbb{R}^n$ , the expected regret defined in (A.13) satisfies  $\text{ExpRegret}_T(\Pi) \leq \text{RHS of (2.14)} + \text{RHS of (A.18)}$ .*

*Proof.* Let  $\kappa_{t,\theta}$ ,  $\kappa_{w\theta}$ ,  $\kappa_{t,x}$ ,  $\kappa_{wx}$ ,  $\kappa_{t,x^*}$ ,  $\kappa_{wx^*}$  and  $F_{exp}$  denote the RHS of (A.5), (A.19), (A.6), (A.21), (A.10), (A.22) and (A.18), respectively. Rewrite state  $x_{t|q}$  defined in (A.14) as  $x_{t|q} = \chi_{t|q}^0 + \chi_{t|q}^1$ , where  $\chi_{t|q}^0 = \prod_{j=0}^{t-1} (A + BK_{j|q}) \bar{x}_0$  and  $\chi_{t|q}^1 = \sum_{r=0}^{t-1} \left( \prod_{j=r+1}^{t-1} (A + BK_{j|q}) \right) w_r$ . Note that  $\chi_{t|q}^0$  does not contain any terms involving random variables. Then, we can calculate and upper bound  $\mathbf{E}(\|x_{t|t} - x_t^*\|^2)$  by

$$\begin{aligned}
& \mathbf{E}(\|x_{t|t} - x_t^*\|^2) \\
& = \mathbf{E}(\| \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + BK_{m|t}) \right) B(K_{n|t} - K_n^*) x_n^* \|^2) \\
& = \left\| \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + BK_{m|t}) \right) B(K_{n|t} - K_n^*) \chi_{n|T}^0 \right\|^2 \\
& \quad + \mathbf{E}(\| \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + BK_{m|t}) \right) B(K_{n|t} - K_n^*) \chi_{n|T}^1 \|^2) \\
& \leq \kappa_{t,\theta} + \kappa_{w\theta}, \tag{A.24}
\end{aligned}$$

where the last two lines hold since the expectation of  $\chi_{n|T}^1$ , and of cross terms between  $\chi_{n|T}^0$  and  $\chi_{n|T}^1$ , is 0.

To upper bound  $\mathbf{E}(\|x_t - x_t^*\|^2)$ , note that  $\mathbf{E}(\|x_t - x_t^*\|^2) \leq 2\mathbf{E}(\|x_t - x_{t|t}\|^2 + \|x_{t|t} - x_t^*\|^2)$ , and the upper bound of term  $\mathbf{E}(\|x_{t|t} - x_t^*\|^2)$  is given by (A.24). Therefore, it remains for us to upper bound  $\mathbf{E}(\|x_t - x_{t|t}\|^2)$ . By (A.15), we have

$$\mathbf{E}(\|x_t - x_{t|t}\|^2) = \mathbf{E}\left(\left\|\sum_{i=1}^{t-1} \prod_{\tau=0}^{t-i-1} (A + BK_\tau) \theta_{i|i-1,i}\right\|^2\right). \quad (\text{A.25})$$

By (A.17), for all  $1 \leq i \leq t-1$ ,  $\theta_{i|i-1,i}$  can be written as  $\sum_{n=0}^{i-1} \left( \prod_{m=n+1}^{i-1} (A + BK_{m|i-1}) \right) B(K_{n|i-1} - K_{n|i}) [\chi_{n|i}^0 + \chi_{n|i}^1] - w_{i-1}$ . For  $k \in \{0, 1\}$ , let

$$X_{i-1|i-1,i}^k := \sum_{n=0}^{i-1} \left( \prod_{m=n+1}^{i-1} (A + BK_{m|i-1}) \right) B(K_{n|i-1} - K_{n|i}) \chi_{n|i}^k.$$

Substitute the above into (A.25), we have

$$\begin{aligned} & \mathbf{E}(\|x_t - x_{t|t}\|^2) \\ &= \mathbf{E}\left(\left\|\sum_{i=1}^{t-1} \prod_{\tau=0}^{t-i-1} (A + BK_\tau) (X_{i-1|i-1,i}^0 + X_{i-1|i-1,i}^1 - w_{i-1})\right\|^2\right) \\ &= \left\|\sum_{i=1}^{t-1} \prod_{\tau=0}^{t-i-1} (A + BK_\tau) X_{i-1|i-1,i}^0\right\|^2 \end{aligned} \quad (\text{A.26})$$

$$+ \mathbf{E}\left(\left\|\sum_{i=1}^{t-1} \prod_{\tau=0}^{t-i-1} (A + BK_\tau) (X_{i-1|i-1,i}^1 - w_{i-1})\right\|^2\right). \quad (\text{A.27})$$

The above equality is due to (A.26) not containing any terms involving random variables. Note that (A.26) is upper bounded by  $\kappa_{t,x}$  given in (A.6), and (A.27) is upper bounded by  $\kappa_{wx}$  in (A.21). Therefore,

$$\mathbf{E}(\|x_t - x_{t|t}\|^2) \leq \kappa_{t,x} + \kappa_{wx}. \quad (\text{A.28})$$

Furthermore,

$$\begin{aligned} \mathbf{E}(\|x_t^*\|^2) &= \mathbf{E}(\|\chi_{t|T}^0 + \chi_{t|T}^1\|^2) \\ &= \|\chi_{t|T}^0\|^2 + \mathbf{E}(\|\chi_{t|T}^1\|^2) \leq \kappa_{t,x^*} + \kappa_{wx^*}. \end{aligned} \quad (\text{A.29})$$

Substituting (A.24), (A.28) and (A.29) into (A.13), we have

$$\begin{aligned} \text{ExpRegret}_T(\Pi) &\leq \frac{10D}{3} \sum_{t=0}^{T-1} (\alpha_1 \mathbf{E}(\|x_{t|t} - x_t^*\|^2) + \alpha_2 \mathbf{E}(\|x_t - x_t^*\|^2) + C_K^2 \gamma^{2W} \mathbf{E}(\|x_t^*\|^2)) \\ &\leq \frac{10D}{3} \sum_{t=0}^{T-1} \alpha_1 (\kappa_{\omega\theta} + \kappa_{t,\theta}) + \alpha_2 (\kappa_{t,x} + \kappa_{wx}) + C_K^2 \gamma^{2W} (\kappa_{wx^*} + \kappa_{t,x^*}). \end{aligned}$$

Recall that  $F$  and  $F_{exp}$  denote the RHS of (2.14) and the last term in the inequality (A.23), respectively. Rewriting  $F$  and  $F_{exp}$  with  $\kappa_{t,\theta}$ ,  $\kappa_{\omega\theta}$ ,  $\kappa_{t,x}$ ,  $\kappa_{\omega x}$ ,  $\kappa_{t,x^*}$ ,  $\kappa_{\omega x^*}$ , we have  $F = \frac{10D}{3} \sum_{t=0}^{T-1} \alpha_1 \kappa_{t,\theta} + \alpha_2 \kappa_{t,x} + C_K^2 \gamma^{2W} \kappa_{t,x^*}$  and  $F_{exp} = \frac{10D}{3} \sum_{t=0}^{T-1} \alpha_1 \kappa_{\omega\theta} + \alpha_2 \kappa_{\omega x} + C_K^2 \gamma^{2W} \kappa_{\omega x^*}$ . Therefore,  $\text{ExpRegret}_T(\Pi) \leq F + F_{exp}$ .  $\square$

Now, we are ready to prove Theorem 2.3.2.

*Proof of Theorem 2.3.2:* Based on the expression of  $F$  and  $F_{exp}$  from Theorem 2.3.1 and Lemma A.3.2, there exist positive scalars  $C_{R1}, C_{R2}$  and  $C'_{ER}$ , such that  $F \leq \gamma^{2W} C_{R1} T$  and  $F_{exp} \leq (\gamma^{2W} C_{R2} + C'_{ER}) T$  for  $T \geq 1$  and  $0 \leq W \leq T - 1$ . By Lemma A.3.3, the expected regret under control policy (2.6) yields  $\text{ExpRegret}_T(\Pi) \leq F + F_{exp} \leq \gamma^{2W} (C_{R1} + C_{R2} T) + C'_{ER} T$ . Let  $C_{ER} = C_{R1} + C_{R2}$ , we have  $\text{ExpRegret}_T(\Pi) \leq (\gamma^{2W} C_{ER} + C'_{ER}) T$ .  $\blacksquare$

## A.4 Proof of Theorem 2.4.1

To lay the groundwork for proving Theorem 2.4.1, we first study the special case of one-step controllable systems.

### One-Step Controllable Systems with No Disturbance

In this section, we consider the case where for any given  $x_0, x_1 \in \mathbb{R}^n$ , there exists a controller  $K$ , such that  $u = Kx_0$  and  $x_1 = Ax_0 + Bu$ . At each time  $t$ , we select  $K_t$  so that  $Ax_t + BK_t x_t = x_{t+1|t}$ , where  $x_{t+1|t}$  is calculated from (2.4) for the disturbance-free case. The control design is similar to the case with disturbances but with  $x_{t+1|t}$  predicted via (2.25).

**Proposition A.4.1.** *For any  $T \geq 1$ ,  $0 \leq W \leq T - 1$  and  $0 \leq t \leq T - 1$ . Consider optimal feedback gain  $K_t^*$  as  $K_{t|T}$  defined in Proposition 2.3.1. Let controls  $\{u_t\}_{t=0}^{T-1}$  and states  $\{x_t\}_{t=0}^T$  be generated by control policy (2.6) for the linear system (1.1). Then the square of the distance between control  $u_t$  and  $K_t^* x_t$  satisfies  $\|u_t - K_t^* x_t\|^2 < 2(C^2 C_K^2 + \frac{C_{KA}^2 C^4 C_K^2}{(1-\gamma)^2}) \eta^{2t} \gamma^{2W}$ , where constants  $C, C_K, \eta, \gamma$  are defined in Theorem 2.3.1 and  $C_{KA}$  is defined in Theorem 2.4.1.*

*Proof.* To calculate  $u_t - K_t^* x_t$ ,

$$\begin{aligned} u_t - K_t^* x_t &= (B^\top B)^{-1} B^\top (x_{t+1|t} - Ax_{t|t-1}) - K_t^* x_{t|t-1} \\ &= (B^\top B)^{-1} B^\top (Ax_{t|t} - Ax_{t|t-1} + BK_{t|t} x_{t|t}) - K_t^* x_{t|t-1} \\ &= (K_{t|t} - K_t^*) x_{t|t} + (K_t^* + (B^\top B)^{-1} B^\top A)(x_{t|t} - x_{t|t-1}), \end{aligned}$$

By  $\|x_{t|t}\| \leq C^2 \eta^{2t}$ ,  $\|x_{t|t} - x_{t|t-1}\|^2 < \frac{C^4 C_K^2 \gamma^{2W}}{(1-\gamma)^2} \eta^{2t}$ , the elementary inequality  $(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2)$  and  $\|K_{t|t} - K_t^*\|^2 \leq C_{KA}^2 \gamma^{2W}$ , we have  $\|u_t - K_t^* x_t\|^2 < 2(C^2 C_K^2 + \frac{C_{KA}^2 C^4 C_K^2}{(1-\gamma)^2}) \eta^{2t} \gamma^{2W}$ .  $\square$

Remember that  $\|R_t + B^\top P_{t+1} B\|^2 \leq D$  (cf. (2.10)). Invoking the Cost Difference Lemma (Lemma A.1.6) and Proposition A.4.1 that yield

$$\begin{aligned} \text{Regret}_T(\Pi) &\leq \sum_{t=0}^{T-1} \|R_t + B^\top P_{t+1} B\|^2 \|u_t - K_t^* x_t\|^2 \\ &< D \|\bar{x}_0\|^2 \sum_{t=0}^{T-1} 2(C^2 C_K^2 + \frac{C_{KA}^2 C^4 C_K^2}{(1-\gamma)^2}) \eta^{2t} \gamma^{2W} \\ &= 2D \|\bar{x}_0\|^2 \gamma^{2W} [C^2 C_K^2 + C_{KA}^2 \frac{C^4 C_K^2}{(1-\gamma)^2}] \frac{1 - \eta^{2T}}{1 - \eta^2}. \end{aligned} \quad (\text{A.30})$$

### Lifted Space Case

We start by stating the next two propositions to help us upper bound the regret for adopting the deadbeat tracking controller in disturbance free case.

**Proposition A.4.2.** *Suppose  $t$  is a positive integer. For any pair of real square matrix sequences of the same dimensions  $\{a_\tau\}_{\tau=0}^t$  and  $\{b_\tau\}_{\tau=0}^t$ , we have*

$$a_t a_{t-1} \cdots a_0 - b_t b_{t-1} \cdots b_0 = \sum_{m=0}^t \left[ \prod_{r=m+1}^t a_r \right] (a_m - b_m) \prod_{r=0}^{m-1} b_r.$$

*Proof.* Define  $\phi_{t+1} = a_t \phi_t$ ,  $\psi_{t+1} = b_t \psi_t$ , and  $\phi_0 = \psi_0 = I$ , where  $I$  is the identity matrix. We have

$$\begin{aligned} \left( \prod_{\tau=0}^t a_\tau - \prod_{\tau=0}^t b_\tau \right) &= \phi_{t+1} - \psi_{t+1} = a_t \phi_t - b_t \psi_t \\ &= a_t (\phi_t - \psi_t + \psi_t) - b_t \psi_t = a_t (\phi_t - \psi_t) + (a_t - b_t) \psi_t \\ &= \sum_{m=0}^t \left( \prod_{r=m+1}^t a_r \right) (a_m - b_m) \left( \prod_{r=0}^{m-1} b_r \right). \end{aligned}$$

The last equality follows from the recursive relation between  $\phi_{t+1} - \psi_{t+1}$  and  $\phi_t - \psi_t$

$\psi_t$ . □

**Proposition A.4.3.** *Suppose (1.1) is a  $d$ -step controllable system. Consider  $T \geq 1$  and  $0 \leq W \leq T - 1$  where  $T \bmod d = 0$  and  $W \bmod d = 0$ . For  $0 \leq \tau_d \leq T_d$ , the distance between matrices  $\tilde{K}_{\tau_d|\tau_d}$  and  $\tilde{K}_{\tau_d}^*$  as  $\tilde{K}_{\tau_d|T_d}$  defined in (2.20), respectively, satisfies  $\|\tilde{K}_{\tau_d|\tau_d} - \tilde{K}_{\tau_d}^*\|^2 < C_{\tilde{K}} \gamma^{2 \max(0, W-d)}$ , where  $C_{\tilde{K}} := \gamma^2 C^4 C_K^2 \frac{1-\gamma^2 + \|B\|^2}{1-\gamma^2} \frac{\gamma^{2d} - \eta^{2d}}{\gamma^2 - \eta^2}$ , with  $C, C_K, \gamma$  and  $\eta$  defined in (2.12), (2.9), (2.11) and (2.10), respectively.*

*Proof.* From [5, Equation (22)], we have

$$\tilde{K}_{\tau_d|\tau_d} = (\hat{R}_{\tau_d|\tau_d} + \Delta_{\tau_d|\tau_d}^\top \Delta_{\tau_d|\tau_d})^{-1} \Delta_{\tau_d|\tau_d}^\top \Xi_{\tau_d|\tau_d} + \begin{bmatrix} K_{d\tau_d|d\tau_d} \\ K_{d\tau_d+1|d\tau_d} (A_{d\tau_d|d\tau_d} + B_{d\tau_d|d\tau_d} K_{d\tau_d|d\tau_d}) \\ \vdots \\ K_{d(\tau_d+1)-1|\tau_d} (\prod_{n=d\tau_d}^{d(\tau_d+1)-1} A_{n|d\tau_d} + B_{n|d\tau_d} K_{n|d\tau_d}) \end{bmatrix}.$$

By definition of  $\hat{R}_{\tau_d|\tau_d}$ ,  $\Delta_{\tau_d|\tau_d}$  and  $\Xi_{\tau_d|\tau_d}$  per (2.17), (2.18) and (2.19), with the assumption on preview window length  $W$  in Theorem 2.4.1, we have  $(\hat{R}_{\tau_d|\tau_d} + \Delta_{\tau_d|\tau_d}^\top \Delta_{\tau_d|\tau_d})^{-1} \Delta_{\tau_d|\tau_d}^\top \Xi_{\tau_d|\tau_d} = (\hat{R}_{\tau_d|T_d} + \Delta_{\tau_d|T_d}^\top \Delta_{\tau_d|T_d})^{-1} \Delta_{\tau_d|T_d}^\top \Xi_{\tau_d|T_d}$ . Hence, we only need to bound the following term:

$$\begin{bmatrix} K_{d\tau_d|d\tau_d} - K_{d\tau_d}^* \\ K_{d\tau_d+1|d\tau_d} (A + BK_{d\tau_d|d\tau_d}) - K_{d\tau_d+1}^* (A + BK_{d\tau_d|d\tau_d}) \\ \vdots \\ K_{d(\tau_d+1)-1|\tau_d} [\prod_{n=d\tau_d}^{d(\tau_d+1)-1} (A + BK_{n|d\tau_d})] \\ - K_{d(\tau_d+1)-1|\tau_d} [\prod_{n=d\tau_d}^{d(\tau_d+1)-1} (A + BK_{n|d\tau_d})] \end{bmatrix}.$$

To this aim, note that the square of 2-norm of

$$K_{d\tau_d+r|d\tau_d} \prod_{n=d\tau_d}^{d\tau_d+r-1} (A + BK_{n|d\tau_d}) - \tilde{K}_{d\tau_d+r}^* \prod_{n=d\tau_d}^{d\tau_d+r-1} (A + BK_n^*) \quad (\text{A.31})$$

is less than or equal to  $C^4 C_K^2 \eta^{2r} (\gamma^{2(dW_d-r)} + \|B\|^2 \gamma^{2(dW_d-r)} \frac{1-\gamma^{2(r+1)}}{1-\gamma^2})$ . In light of Proposition A.4.2 and Lemma A.1.1:

$$\begin{aligned} & \|\tilde{K}_{\tau_d|\tau_d} - \tilde{K}_{\tau_d}^*\|^2 \\ & \leq \sum_{r=0}^{d-1} \|K_{d\tau_d+r|d\tau_d} \prod_{n=d\tau_d}^{d\tau_d+r-1} (A + BK_{n|d\tau_d}) - \tilde{K}_{d\tau_d+r}^* \prod_{n=d\tau_d}^{d\tau_d+r-1} (A + BK_n^*)\| \\ & < C^4 C_K^2 \gamma^2 \gamma^{2 \max(0, W-d)} \frac{1 - \gamma^2 + \|B\|^2}{1 - \gamma^2} \frac{\gamma^{2d} - \eta^{2d}}{\gamma^2 - \eta^2} =: C_{\tilde{K}}. \end{aligned}$$

□

The next lemma is the lifted space version of Lemma A.1.6.

**Lemma A.4.1** (Cost Difference Lemma in Lifted Space). *For any  $T \geq 1$  and  $0 \leq t \leq T$ , consider  $\tilde{B}$  from (2.21),  $\tilde{R}_t$  from (2.24),  $\tilde{K}_t^*$  be  $\tilde{K}_{\tau_d|T_d}$  in (2.22) and  $\tilde{P}_t^*$  be  $\tilde{P}_{t|T}$  from (2.23), respectively. Let  $\tilde{\Pi} = \{\tilde{\pi}_t\}_{t=0}^{T-1}$  be the control policy defined in (2.26), states  $\{\tilde{x}_{t_d+1}\}_{t_d=0}^{T_d-1}$  be generated from control sequence  $\{\tilde{u}_{t_d}\}_{t_d=0}^{T_d-1}$  for the linear system (2.21), where each control being  $\tilde{u}_{t_d} = [u_{dt_d}, u_{dt_d+1}, \dots, u_{dt_d+d-1}]$  and  $u_t$  is defined in (2.26). The regret defined by (2.2) satisfies*

$$\text{Regret}_T(\tilde{\Pi}) = \sum_{\tau_d=0}^{T_d-1} (\tilde{u}_{\tau_d} - \tilde{K}_{\tau_d}^* \tilde{x}_{\tau_d})^\top (\tilde{R}_{\tau_d} + \tilde{B}^\top \tilde{P}_{\tau_d+1} \tilde{B}) (\tilde{u}_{\tau_d} - \tilde{K}_{\tau_d}^* \tilde{x}_{\tau_d}).$$

To state the proof of Theorem 2.4.1, we define  $\hat{C}_{max} := [\oplus_{j=0}^{d-1} Q_{max}^{\frac{1}{2}}, 0_{nd,n}]$ ,  $\Delta_{max} := \hat{C}_{max} \hat{A}^{-1} \hat{B}$ , and  $\tilde{R}_{max} := \oplus_{j=0}^{d-1} R_{max} + \Delta_{max}^\top \Delta_{max}$ .

*Proof of Theorem 2.4.1:* By applying the Cost Difference Lemma in Lifted Space, the upper bound of  $\|\tilde{K}_{\tau_d|\tau_d} - \tilde{K}_{\tau_d}^*\|^2$  from Proposition A.4.3, the elementary inequality of  $(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2)$  and repeating the steps leading to (A.30) yields

$$\begin{aligned} \text{Regret}_T(\tilde{\Pi}) &\leq 2\tilde{D} \sum_{\tau_d=0}^{T_d-1} (\|\tilde{K}_{\tau_d|\tau_d} - \tilde{K}_{\tau_d}^*\|^2 \|\tilde{x}_{\tau_d|\tau_d}\|^2 \\ &\quad + \|\tilde{K}_{\tau_d}^* + (\tilde{B}^\top \tilde{B})^{-1} \tilde{B}^\top \tilde{A}_{\tau_d|\tau_d-1}\|^2 \|\tilde{x}_{\tau_d|\tau_d} - \tilde{x}_{\tau_d|\tau_d-1}\|^2) \\ &< \tilde{\Psi} \gamma^{2 \max(0, W-d)}. \end{aligned}$$

where  $\tilde{\Psi} := 2\tilde{D} \gamma^2 \|\bar{x}_0\|^2 [C^2 C_{\tilde{K}} + \tilde{C}_{KA} \frac{C^4 C_{\tilde{K}}^2}{(1-\gamma)^2}] \frac{1-\eta^{2T}}{1-\eta^{2d}}$  with  $C_K, \eta, \gamma, C$ , and  $C_{\tilde{K}}$  defined in (2.9), (2.10), (2.11), (2.12), and Proposition A.4.3, respectively, and  $\tilde{D} := \|\tilde{R}_{max}\| + d \|\tilde{B}\|^2 \|P_{max}\|$  and  $\tilde{C}_{KA} := \|\tilde{B}\| \|\hat{A}\| + d \frac{\sigma_{max}(A)}{\sigma_{min}(B)}$ . The above uses inequalities from Proposition A.4.3, and Lemmas A.1.2 and A.1.5, respectively.

## A.5 Proof of Theorem 2.4.2

Similar to the above regret analysis for the lifted space disturbance-free case, we first establish the regret for the one-step controllable case in the disturbance case.

### One-Step Controllable Systems with Disturbance

**Proposition A.5.1.** *For any  $T \geq 1$ , consider i.i.d. random variable  $\{w_t\}_{t=0}^{T-1}$  where  $w_t \in \mathbb{R}^n$ ,  $\mathbf{E}(w_t) = 0$  and  $\mathbf{E}(w_t w_t^\top) = \text{Cov}_w$ . We further consider  $0 \leq W \leq T-1$  and*

$1 \leq t \leq T$ . Suppose states  $x_{t|t}$  and  $x_{t|t-1}$  are determined by solving Problem (2.24) at time  $d \lfloor \frac{t}{d} \rfloor$  given  $\mathcal{H}_{t,W}$  and  $\mathcal{H}_{t-1,W}$ , respectively. Then,  $\mathbf{E}(\|x_{t|t}\|^2) < C^2 \|\bar{x}_0\|^2 (\eta^{2t} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2})$  and  $\mathbf{E}(\|x_{t|t} - x_{t|t-1}\|^2) < \frac{C^4 \|\bar{x}_0\|^2 \|B\|^2 C_K^2}{1-\gamma^2} (\eta^{2(t-1)} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2})$ , where constants  $C, C_K, \eta$  and  $\gamma$  are defined in Theorem 2.3.1.

*Proof.* By using

$$x_{t|t} = \prod_{\tau=0}^{t-1} (A + BK_{\tau|t}) \bar{x}_0 + \sum_{\tau=0}^{t-1} \prod_{n=t-1-\tau}^{t-1} (A + BK_{n|t}) w_{t-1-\tau},$$

we have  $\mathbf{E}(\|x_{t|t}\|^2) < C^2 \|\bar{x}_0\|^2 (\eta^{2t} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2})$ . To upper bound  $\mathbf{E}(\|x_{t|t} - x_{t|t-1}\|^2)$ , note

$$\begin{aligned} x_{t|t} - x_{t|t-1} &= w_{t-1} + B(K_{t-1|t} - K_{t-1|t-1})x_{t-1|t-1} + (A + BK_{t-1|t}) \\ &\quad \left[ \sum_{n=0}^{t-2} \prod_{m=n+1}^{t-2} (A + BK_{m|t}) B(K_{n|t} - K_{n|t-1}) x_{n|t-1} \right]. \end{aligned}$$

By adopting Lemma A.1.7,  $\mathbf{E}(\|x_{t|t} - x_{t|t-1}\|^2)$  can be upper bounded by

$$\begin{aligned} &\mathbf{E}(\|x_{t|t} - x_{t|t-1}\|^2) \\ &\leq \frac{10}{3} \mathbf{E}(\|(A + BK_{t-1|t})(x_{t-1|t} - x_{t-1|t-1})\|^2 + \|B(K_{t-1|t} - K_{t-1|t-1})x_{t-1|t-1}\|^2 + \|w_{t-1}\|^2) \\ &\leq \frac{10}{3} \left[ \text{Tr}(\text{Cov}_w) + \left( \frac{C_K^2 \gamma^{2W} \|\bar{x}_0\|^2 C^2}{1-\gamma^2} \right) (C^4 + \|B\|^2) (\eta^{2(t-1)} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2}) \right]. \end{aligned}$$

□

We also need a lifted-space expected cost difference lemma.

**Lemma A.5.1.** For any  $T \geq 1$  and  $0 \leq t \leq T$ , consider  $\tilde{B}$  from (2.21),  $\tilde{R}_t$  from (2.24),  $\tilde{K}_t^*$  be  $\tilde{K}_{\tau_d|T_d}$  in (2.22) and  $\tilde{P}_t^*$  be  $\tilde{P}_{t|T}$  from (2.23), respectively. We further consider the random sequence  $\{w_t\}_{t=0}^{T-1}$  where  $w_t \in \mathbb{R}^n$ ,  $\mathbf{E}(w_t) = 0$  and  $\mathbf{E}(w_t w_t^\top) = \text{Cov}_w$ . Let  $\tilde{\Pi} = \{\tilde{\pi}_t\}_{t=0}^{T-1}$  be the control policy defined in (2.26), states  $\{\tilde{x}_{t_d+1}\}_{t_d=0}^{T_d-1}$  be generated from control sequence  $\{\tilde{u}_{t_d}\}_{t_d=0}^{T_d-1}$  for the linear system (2.21), where each control being  $\tilde{u}_{t_d} = [u_{dt_d}, u_{dt_d+1}, \dots, u_{dt_d+d-1}]$  and  $u_t$  is defined in (2.26). The regret defined by (2.2) satisfies

$$\text{ExpRegret}_T(\tilde{\Pi}) = \mathbf{E} \left( \sum_{\tau_d=0}^{T_d-1} (\tilde{u}_{\tau_d} - \tilde{K}_{\tau_d}^* \tilde{x}_{\tau_d})^\top (\tilde{R}_{\tau_d} + \tilde{B}^\top \tilde{P}_{t+1} \tilde{B}) (\tilde{u}_{\tau_d} - \tilde{K}_{\tau_d}^* \tilde{x}_{\tau_d}) \right).$$

*Proof.* The proof mirrors that of Lemma A.3.1. □

The next theorem presents the expected regret upper bound in a one-step controllable system under policy (2.26).

**Theorem A.5.1.** *Adopt the hypothesis of Theorem 2.3.2. The expected regret satisfies  $\text{ExpRegret}_T(\tilde{\Pi}) < (C_{ERO}\gamma^{2W} + C'_{ERO})T$ , where policy  $\tilde{\Pi}$  is defined in (2.26) and  $C_{ERO} := 2\|\bar{x}_0\|^2 C_K^2 C^2 [1 + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2}] [1 + \frac{C_{KA}^2 C^2 \|B\|^2}{1-\gamma^2}]$ .*

*Proof.* From the previous lemma, we have that  $\text{ExpRegret}_T(\tilde{\Pi}) \leq D \sum_{t=0}^{T-1} \mathbf{E}(\|u_t - K_t^* x_t\|^2)$ . To bound  $\mathbf{E}(\|u_t - K_t^* x_t\|^2)$ , note

$$\begin{aligned} & \mathbf{E}(\|u_t - K_t^* x_t\|^2) \\ &= \mathbf{E}(\|(K_{t|t} - K_t^*)x_{t|t} + (K_t^* + (B^\top B)^{-1}A)(x_{t|t} - x_{t|t-1})\|^2) \\ &\leq 2(\|K_{t|t} - K_t^*\|^2 \mathbf{E}(\|x_{t|t}\|^2) + \|K_t^* + (B^\top B)^{-1}A\|^2 \mathbf{E}(\|x_{t|t} - x_{t|t-1}\|^2)). \end{aligned}$$

Based on Proposition A.5.1, the above inequality can be upper bounded by  $2\gamma^{2W} [C_K^2 C^2 \|\bar{x}_0\|^2 (\eta^{2t} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2}) + \frac{C_{KA}^2 C^4 \|\bar{x}_0\|^2 \|B\|^2 C_K^2}{1-\gamma^2} (\eta^{2(t-2)} + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2})]$ . Taking summation above that  $t$  varies from 0 to  $T-1$ , let  $C_{ERO} := \frac{20}{3} C_{KA}^2 \|B\|^2 C_K^2 \|\bar{x}_0\|^2 C^2 \left[ \|B\|^2 + \frac{C^4}{1-\gamma^2} \right] \left[ 1 + \frac{\text{Tr}(\text{Cov}_w)}{1-\eta^2} \right]$  and  $C'_{ERO} = \frac{20\text{Tr}(\text{Cov}_w)}{3}$ , the expected regret can be upper bound by  $\text{ExpRegret}_T(\tilde{\Pi}) < (C_{ERO}\gamma^{2W} + C'_{ERO})T$ .  $\square$

We need another proposition associate with the disturbances in the lifted space to proof Theorem 2.4.2.

**Proposition A.5.2.** *Suppose i.i.d. random variables  $w_n$  satisfy  $\mathbf{E}(w_n) = 0$  and  $\mathbf{E}(w_n w_n^\top) = \text{Cov}_w$  (also  $\mathbf{E}(w_n w_m^\top) = 0$  for  $n \neq m$ ), then  $\sum_{n=0}^{d-1} \mathbf{E}(\|A^n w_n\|^2) \leq \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{1 - \sum_{i=1}^d \sigma_i(A)}$ .*

*Proof.* Note that  $\sum_{n=0}^{d-1} \mathbf{E}(\|A^n w_n\|^2) = \sum_{n=0}^{d-1} \mathbf{E}(w_n^\top A^{n\top} A^n w_n) = \sum_{n=0}^{d-1} \mathbf{E}(\text{Tr}(A^{n\top} A^n w_n w_n^\top)) = \sum_{n=0}^{d-1} \text{Tr}(A^{n\top} A^n \text{Cov}_w) = \sum_{n=0}^{d-1} \text{Tr}(A^n \text{Cov}_w A^{n\top}) = \text{Tr}(\sum_{n=0}^{d-1} A^n \text{Cov}_w A^{n\top}) := G$ . Moreover,  $G - AGA^\top = I - A^d \text{Cov}_w A^{d\top}$ . Using the cyclic property of trace, we have  $\text{Tr}(G)(1 - \sum_{i=1}^d \sigma_i(A)) \leq \text{Tr}(G - AGA^\top) = \text{Tr}(I - A^d \text{Cov}_w A^{d\top})$ . Therefore,  $\text{Tr}(G) \leq \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{1 - \sum_{i=1}^d \sigma_i(A)}$ .  $\square$

*Proof of Theorem 2.4.2:* Similar to the proof of Theorem A.5.1 for the one-step controllable case, we have

$$\mathbf{E} \|\tilde{x}_{t_d}\|^2 \leq \mathbf{E} \left( \left\| \prod_{n=0}^{t_d-1} (\tilde{A} + \tilde{B} \tilde{K}_{n|t_d}) \tilde{x}_0 + \sum_{\tau_d=0}^{t_d-1} \prod_{n=t_d-1-\tau_d}^{t_d-1} (\tilde{A} + \tilde{B} \tilde{K}_{n|t_d}) \tilde{A}_w \tilde{w}_{t_d-1-\tau_d} \right\|^2 \right),$$

$$\begin{aligned}
& \mathbf{E}(\|\tilde{x}_{\tau_d|\tau_d} - \tilde{x}_{\tau_d|\tau_{d-1}}\|^2) \\
& \leq \frac{10}{3} [\|\tilde{B}(\tilde{K}_{\tau_{d-1}|\tau_d} - \tilde{K}_{\tau_{d-1}|\tau_{d-1}})\|^2 \mathbf{E}(\|\tilde{x}_{\tau_{d-1}|\tau_d}\|^2) + \mathbf{E}(\|\tilde{A}_w \tilde{w}_{\tau_{d-1}}\|^2) \\
& \quad + \|(\tilde{A} + \tilde{B}\tilde{K}_{\tau_{d-1}|\tau_{d-1}})\|^2 \mathbf{E}(\|\tilde{x}_{\tau_{d-1}|\tau_d} - \tilde{x}_{\tau_{d-1}|\tau_{d-1}}\|^2)].
\end{aligned}$$

By (A.24) and (A.28), we have

$$\begin{aligned}
& \|\tilde{B}(\tilde{K}_{\tau_{d-1}|\tau_d} - \tilde{K}_{\tau_{d-1}|\tau_{d-1}})\|^2 \mathbf{E}(\|\tilde{x}_{\tau_{d-1}|\tau_d}\|^2) \\
& \leq \|\tilde{B}\|^2 C^2 \|\bar{x}_0\|^2 \gamma^{2(1-d+dW_d)} (\eta^{2d(\tau_d-1)} + \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{(1-\eta^2)(1-\sum_{i=1}^d \sigma_i(A))}),
\end{aligned}$$

$$\begin{aligned}
& \mathbf{E}(\|(\tilde{A} + \tilde{B}\tilde{K}_{\tau_{d-1}|\tau_{d-1}})(\tilde{x}_{\tau_{d-1}|\tau_d} - \tilde{x}_{\tau_{d-1}|\tau_{d-1}})\|^2) \\
& \leq C^2 \eta^{2d} \frac{C^4 \|\bar{x}_0\|^2 \|B\|^2 C_K^2 \gamma^{2dW_d}}{1-\gamma^2} [\eta^{2d(\tau_d-2)} + \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{(1-\eta^2)(1-\sum_{i=1}^d \sigma_i(A))}],
\end{aligned}$$

and  $\mathbf{E}(\|\tilde{A}_w \tilde{w}_{\tau_{d-1}}\|^2) \leq \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{1-\sum_{i=1}^d \sigma_i(A)}$ . Therefore,

$$\begin{aligned}
& \text{ExpRegret}_T(\tilde{\Pi}) \\
& \leq \sum_{t_d=0}^{T_d-1} \mathbf{E}(\|(\tilde{K}_{t_d|t_d} - \tilde{K}_{t_d}^*) \tilde{x}_{t_d|t_d} + (\tilde{K}_{t_d}^* + (\tilde{B}^\top \tilde{B})^{-1} \tilde{A})(\tilde{x}_{t_d|t_d} - \tilde{x}_{t_d|t_{d-1}})\|^2) \\
& \leq (\tilde{C}_{ER} \gamma^{2 \min(0, dW_d-d)} + \tilde{C}'_{ER}) d T_d \\
& = (\tilde{C}_{ER} \gamma^{2 \max(0, W-d)} + \tilde{C}'_{ER}) T,
\end{aligned}$$

where  $\tilde{C}_{ER} := 2\gamma^2 C^2 \|\bar{x}_0\|^2 [1 + \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{1-\sum_{i=1}^d \sigma_i(A)}] [C_{\tilde{K}} + \frac{10\eta^{2d} C^4 \|B\|^2 C_K^2 \tilde{C}_{KA} \gamma^{2dW_d}}{3(1-\gamma^2)(1-\eta^2)}]$ , and  $\tilde{C}'_{ER} := \frac{20}{3} (1 + \frac{\text{Tr}(I - A^d \text{Cov}_w A^{d\top})}{1-\sum_{i=1}^d \sigma_i(A)}) (1 + \|B\|^2 C^2 \|\bar{x}_0\|^2)$ .  $\blacksquare$

# Appendix. Proof for Online LQ Optimal Control with Sequentially Inferred Costs

---

## B.1 Preparatory Results for the Proof of Lemma 3.4.1

In this section, we discuss auxiliary results and their proofs used in the proofs of the main results of this paper.

From Proposition 1, we observe that the states  $\hat{\xi}_{\tau|t}$  and  $x_{\tau}^*$  can be expressed as

$$\hat{\xi}_{\tau+1|t} := \prod_{m=0}^{\tau} (A + B\hat{K}_{m|t})\bar{x}_0, \quad (\text{B.1})$$

$$x_{\tau+1}^* := \prod_{m=0}^{\tau} (A + BK_m^*)\bar{x}_0, \quad (\text{B.2})$$

for  $\tau \in \mathbb{N}_{T-1}$  and where  $K_{\tau}^*$  and  $\hat{K}_{\tau|t}$  are defined in (3.29) and (3.26), respectively.

The following lemma provides upper bounds of  $\|\prod_{m=0}^{\tau} (A + B\hat{K}_{m|t})\|$  and  $\|\prod_{m=0}^{\tau} (A + BK_m^*)\|$ .

**Lemma B.1.1.** *Let Assumptions 3.2.1 and 3.2.2 be satisfied and let  $t, \tau \in \mathbb{N}_{T-1}$ ,  $\tau \leq t$ . Moreover, consider  $K_{\tau}^*$ ,  $\hat{K}_{\tau|t}$ ,  $\eta$ ,  $\hat{\eta}$   $C$  and  $\hat{C}$  defined in (3.28), (3.29), (3.32h)*

and (3.32g). Then, for all  $t_0, t_1 \in \mathbb{N}_{T-1}$ ,  $t_0 \leq t_1$ , it holds that

$$\left\| \prod_{\tau=t_0}^{t_1} (A + B\hat{K}_{\tau|t}) \right\| \leq \hat{C}\hat{\eta}^{t_1-t_0+1}, \quad (\text{B.3})$$

$$\left\| \prod_{\tau=t_0}^{t_1} (A + BK_{\tau}^*) \right\| \leq C\eta^{t_1-t_0+1}. \quad (\text{B.4})$$

*Proof.* The proof is identical to the proof of [12, Proposition 2] by replacing  $B_u, K_t$  from the proof of [12, Proposition 2] with  $B, \hat{K}_{\tau|t}$  defined in (3.1) and (B.1), respectively. The proof of the second inequality follows same procedures, but replacing  $K_t$  from the proof of [12, Proposition 2] with  $K_{\tau}^*$ .  $\square$

In the next lemma, we bound the distance between  $\hat{x}_{\tau|t}$  and  $x_{\tau}^*$  in terms of  $\hat{\Phi}_{t|(p,q)}$  and  $\bar{\Phi}_{\tau|t}^*$  defined in (3.30) and (3.31), respectively.

**Lemma B.1.2.** *Let the assumptions of Lemma B.1.1 be satisfied, let  $\Gamma \in \mathbb{R}^{n \times m}$  be defined such that  $A + B\Gamma$  is a Schur matrix, and recall the definitions of  $\hat{\xi}_{\tau|t}$  and  $x_{\tau}^*$  in (B.1), (B.2). Then it holds that<sup>1</sup>*

$$\|\hat{\xi}_{t|t} - x_t^*\| \leq C\hat{C}\hat{\eta}^{(t-1)}\|B\|\|\bar{x}_0\| \sum_{n=0}^{t-1} \frac{\eta^n}{\hat{\eta}^n} \bar{\Phi}_{n|t}^*, \quad (\text{B.5})$$

$$\|\xi_t - \hat{\xi}_{t|t}\| \leq C_q \hat{C}^2 \|B\| \|\bar{x}_0\| \frac{\eta_q}{\hat{\eta}} \eta_q^{t-1} \sum_{j=1}^t \sum_{n=0}^{j-1} \left(\frac{\hat{\eta}}{\eta_q}\right)^j \hat{\Phi}_{n|(j-1,j)}, \quad (\text{B.6})$$

where  $\hat{\Phi}_{n|(j-1,j)}$  and  $\bar{\Phi}_{n|t}^*$  are defined in (3.30) and (3.31), respectively.

*Proof.* For  $\tau, t \in \mathbb{N}_{T-1}$ , we define  $w_{\tau|t} := \hat{\xi}_{\tau|t} - x_{\tau}^*$ . Note that  $w_{0|t} = \xi_{0|t} - x_0^* = \bar{x}_0 - \bar{x}_0 = 0$  for all  $t \in \mathbb{N}_{T-1}$  according to (B.1), (B.2). When  $\tau = 1$ , we have

$$w_{1|t} = \hat{\xi}_{1|t} - x_1^* = (A + B\hat{K}_{0|t})\hat{\xi}_{0|t} - (A + BK_0^*)x_0^* = B(\hat{K}_{0|t} - K_0^*)\bar{x}_0. \quad (\text{B.7})$$

<sup>1</sup>The first inequality will not be used throughout the proof, but we still keep it and the proof.

For  $2 \leq \tau \leq T - 1$ , using (B.1) and (B.2) it holds that

$$\begin{aligned}
 w_{\tau|t} &= \hat{\xi}_{\tau|t} - x_t^* \\
 &= (A + B\hat{K}_{\tau-1|t})\hat{\xi}_{\tau-1|t} - (A + BK_{\tau-1}^*)x_{\tau-1}^* \\
 &= (A + B\hat{K}_{\tau-1|t})(w_{\tau-1|t} + x_{\tau-1}^*) - (A + BK_{\tau-1}^*)x_{\tau-1}^* \\
 &= (A + B\hat{K}_{\tau-1|t})w_{\tau-1|t} + B(\hat{K}_{\tau-1|t} - K_{\tau-1}^*)x_{\tau-1}^* \\
 &= (A + B\hat{K}_{\tau-1|t})w_{\tau-1|t} + B(\hat{K}_{\tau-1|t} - K_{\tau-1}^*) \prod_{n=0}^{\tau-2} (A + BK_n^*)\bar{x}_0
 \end{aligned}$$

and we have thus rewritten  $w_{\tau|t}$  in terms of  $w_{\tau-1|t}$  and  $\bar{x}_0$ . If  $\tau - 1 > 0$ , we can thus apply the same step to replace  $w_{\tau-1|t}$  with an expression depending on  $w_{\tau-2|t}$  and  $\bar{x}_0$ , i.e.,

$$\begin{aligned}
 w_{\tau|t} &= (A + B\hat{K}_{\tau-1|t}) \left( (A + B\hat{K}_{\tau-2|t})w_{\tau-2|t} + B(\hat{K}_{\tau-2|t} - K_{\tau-2}^*) \prod_{n=0}^{\tau-3} (A + BK_n^*)\bar{x}_0 \right) \\
 &\quad + B(\hat{K}_{\tau-1|t} - K_{\tau-1}^*) \prod_{n=0}^{\tau-2} (A + BK_n^*)\bar{x}_0 \\
 &= \prod_{n=\tau-2}^{\tau-1} (A + B\hat{K}_n|t) w_{\tau-2|t} \\
 &\quad + \sum_{n=\tau-1}^{\tau-1} \left( \prod_{m=n+1}^{\tau-1} (A + B\hat{K}_m|t) \right) B(\hat{K}_n|t - K_n^*) \left( \prod_{m=0}^{n-1} (A + BK_m^*) \right) \bar{x}_0.
 \end{aligned}$$

Using this argument iteratively and applied to  $\tau = t$ , it follows that

$$w_{t|t} = \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + B\hat{K}_m|t) \right) B(\hat{K}_n|t - K_n^*) \left( \prod_{m=0}^{n-1} (A + BK_m^*) \right) \bar{x}_0.$$

By using the triangle inequality and the fact that the 2-norm is sub-multiplicative, for  $2 \leq t \leq T - 1$ , we have

$$\begin{aligned}
 \|\hat{\xi}_{t|t} - x_t^*\| &\leq \\
 \sum_{n=0}^{t-1} \left\| \left( \prod_{m=n+1}^{t-1} (A + B\hat{K}_m|t) \right) \right\| &\|B(\hat{K}_n|t - K_n^*)\| \left\| \left( \prod_{m=0}^{n-1} (A + BK_m^*) \right) \right\| \|\bar{x}_0\|.
 \end{aligned}$$

By Lemma B.1.1,  $\|\prod_{m=n+1}^{t-1} (A + B\hat{K}_m|t)\| \leq \hat{C}\hat{\eta}^{t-1-n}$  and  $\|\prod_{m=0}^{n-1} (A + BK_m^*)\| \leq C\eta^n$ , and by recalling the definition (3.31), the last expression can be further upper

bounded by

$$\begin{aligned} \|\hat{\xi}_{t|t} - x_t^*\| &\leq \sum_{n=0}^{t-1} C\hat{C}\hat{\eta}^{(t-1)}\frac{\eta^n}{\hat{\eta}^n}\|B\|\|\hat{K}_{n|t} - K_n^*\|\|\bar{x}_0\| \\ &= C\hat{C}\hat{\eta}^{(t-1)}\|B\|\sum_{n=0}^{t-1}\frac{\eta^n}{\hat{\eta}^n}\|\hat{K}_{n|t} - K_n^*\|\|\bar{x}_0\| \\ &\leq C\hat{C}\hat{\eta}^{(t-1)}\|B\|\sum_{n=0}^{t-1}\frac{\eta^n}{\hat{\eta}^n}\bar{\Phi}_{n|t}\|\bar{x}_0\|. \end{aligned}$$

For  $t = 1$ , note that  $C \geq 1$  and  $\hat{C} \geq 1$  by (3.32g), using (B.7), yields

$$\|w_{1|1}\| = \|B(\hat{K}_{0|1} - K_0^*)\bar{x}_0\| \leq \|B\|\bar{\Phi}_{0|1}\|\bar{x}_0\|,$$

and the first inequality in Lemma B.1.2 follows.

For the second inequality in Lemma B.1.2 consider  $0 \leq t \leq T-1$ ,  $0 \leq p \leq q \leq T-1$ , and let

$$\phi_{t|p,q} = \hat{\xi}_{t|p} - \hat{\xi}_{t|q}. \quad (\text{B.8})$$

Similar to the calculation of  $w_{t|t}$ , it follows that

$$\begin{aligned} \phi_{t|(p,q)} &= \hat{\xi}_{t|p} - \hat{\xi}_{t|q} \\ &= (A + B\hat{K}_{t-1|p})\hat{\xi}_{t-1|p} - (A + B\hat{K}_{t-1|q})\hat{\xi}_{t-1|q} \\ &= (A + B\hat{K}_{t-1|p})(\phi_{t-1|(p,q)} + \hat{\xi}_{t-1|q}) - (A + B\hat{K}_{t-1|q})\hat{\xi}_{t-1|q} \\ &= (A + B\hat{K}_{t-1|p})\phi_{t-1|(p,q)} + B(\hat{K}_{t-1|p} - \hat{K}_{t-1|q})\hat{\xi}_{t-1|q} \\ &= (A + B\hat{K}_{t-1|p})\phi_{t-1|(p,q)} + B(\hat{K}_{t-1|p} - \hat{K}_{t-1|q})\prod_{n=0}^{t-2}(A + B\hat{K}_{n|q})\bar{x}_0. \end{aligned}$$

Expanding the above recursion as in the case of  $w_{\tau|t}$ , we have that

$$\phi_{t|(p,q)} = \sum_{n=0}^{t-1} \left( \prod_{m=n+1}^{t-1} (A + B\hat{K}_{m|p}) \right) B(\hat{K}_{n|p} - \hat{K}_{n|q}) \left( \prod_{m=0}^{n-1} (A + B\hat{K}_{m|q}) \right) \|\bar{x}_0\|.$$

Further, by applying the triangle inequality to the right-hand side of the above equality for  $j \in \mathbb{N}_{T-1}$ , we have

$$\|\phi_{j|(j-1,j)}\| \leq \hat{C}^2\hat{\eta}^{j-1}\|B\|\sum_{n=0}^{j-1}\hat{\Phi}_{n|(j-1,j)}\|\bar{x}_0\|. \quad (\text{B.9})$$

Next, we upper bound the distance between  $\xi_{t+1}$  and  $\hat{\xi}_{t+1|t+1}$  for  $t \in \mathbb{N}_{T-1}$ . Using

(3.24), we have

$$\xi_{t+1} = A\xi_t + B\nu_t = A\xi_t + B(\Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{\nu}_{t|t}).$$

By (3.19), we have  $\nu_{t|t} = \hat{K}_{t|t}\xi_{t|t}$ . Combining the calculations above, we can derive an estimate for  $\|\xi_{t+1} - \hat{\xi}_{t+1|t+1}\|$  as follows. Using the system dynamics (3.1), (3.2) and the definition of  $\phi_{\cdot|\cdot}$  in (B.8) it holds that

$$\begin{aligned} \phi_{t+1|(t+1,t+1)} &= \xi_{t+1} - \hat{\xi}_{t+1|t+1} = A\xi_t + B\nu_t - \hat{\xi}_{t+1|t+1} \\ &= A\xi_t + B(\Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{K}_{t|t}\hat{\xi}_{t|t}) - \hat{\xi}_{t+1|t+1}. \end{aligned}$$

With the definition of  $A_{\text{cl}}$  in (3.21) and the definition of  $\phi_{\cdot|\cdot}$  in (B.8) this expression can be further rewritten as

$$\begin{aligned} \phi_{t+1|(t+1,t+1)} &= (A + B\Gamma)\xi_t + B(\hat{K}_{t|t} - \Gamma)\hat{\xi}_{t|t} - (A + B\hat{K}_{t|t})\hat{\xi}_{t|t} + \hat{\xi}_{t+1|t} - \hat{\xi}_{t+1|t+1} \\ &= A_{\text{cl}}\xi_t + B(\hat{K}_{t|t} - \Gamma)\hat{\xi}_{t|t} - (A + B\Gamma + B(\hat{K}_{t|t} - \Gamma))\hat{\xi}_{t|t} + \hat{\xi}_{t+1|t} - \hat{\xi}_{t+1|t+1} \\ &= A_{\text{cl}}(\xi_t - \hat{\xi}_{t|t}) + \phi_{t+1|(t,t+1)} \\ &= A_{\text{cl}}\phi_{t|(t,t)} + \phi_{t+1|(t,t+1)} \end{aligned}$$

Thus, applying this argument iteratively, we can conclude that

$$\phi_{t+1|(t+1,t+1)} = \xi_{t+1} - \hat{\xi}_{t+1|t+1} = \sum_{j=1}^{t+1} A_{\text{cl}}^{t+1-j} \phi_{j|j-1,j}.$$

Therefore, from Corollary 1 the estimate  $\|\xi_{t+1} - \hat{\xi}_{t+1|t+1}\| \leq \sum_{j=1}^{t+1} C_q \eta_q^{t+1-j} \|\phi_{j|(j-1,j)}\|$  follows, which can be further rewritten as

$$\|\xi_{t+1} - \hat{\xi}_{t+1|t+1}\| \leq C_q \hat{C}^2 \|B\| \eta_q^{t+1} \hat{\eta}^{-1} \sum_{j=1}^{t+1} \sum_{n=0}^{j-1} \left(\frac{\hat{\eta}}{\eta_q}\right)^j \hat{\Phi}_{n|(j-1,j)}^* \|\bar{x}_0\|$$

using (B.9), and which completes the proof. □

Before we continue with a proof of Lemma 3.4.1 we state the following result from [50], providing an alternative representation of the regret in (3.15).

**Lemma B.1.3** (Regret representation [50, Lemma 11]). *Let  $T \geq 1$ , consider the linear system (3.2) and assume that Assumption 3.2.1 satisfied. Moreover, under Assumption 3.2.2 let  $R^*$ ,  $(K_t^*)_{t \in \mathbb{N}_T}$  and  $(P_t^*)_{t \in \mathbb{N}_T}$  in (3.32) be defined according to (3.25). For  $\xi_0 \in \mathbb{R}^n$  arbitrary, let  $(\bar{\nu}_t)_{t \in \mathbb{N}_{T-1}}$  be defined through  $\bar{\nu}_t = K_t^* \xi_t$  and the*

dynamics (3.2). Then, for  $\{\nu_t\}_{t=0}^{T-1}$  arbitrary, the regret in (3.15) satisfies

$$\text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) = \sum_{t=0}^{T-1} (\nu_t - \bar{\nu}_t)^\top (R^* + B^\top P_{t+1}^* B) (\nu_t - \bar{\nu}_t).$$

With the help of Lemma B.1.3, we proceed with a proof of Lemma 3.4.1, which establishes the regret bound (3.33).

## B.2 Proof of Lemma 3.4.1

*Proof of Lemma 3.4.1:* Let  $(\bar{\nu}_t)_{t \in \mathbb{N}_{T-1}}$  be defined according to Lemma B.1.3. Then, according to Lemma B.1.3, we have

$$\begin{aligned} \text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) &= \sum_{t=0}^{T-1} (\nu_t - \bar{\nu}_t)^\top (R^* + B^\top P_{t+1}^* B) (\nu_t - \bar{\nu}_t) \\ &\leq D \sum_{t=0}^{T-1} \|\nu_t - K_t^* \xi_t\|^2, \end{aligned} \quad (\text{B.10})$$

and where  $D$  is defined in (3.32e). By definition of  $\nu_t = \kappa_t(\xi_t, \hat{\theta}_t) = \Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{\nu}_{t|t}$  in (3.24), and by using the identity  $\hat{\nu}_{t|t} = \hat{K}_{t|t} \hat{\xi}_{t|t}$ , which follows from the representation in Proposition 1, we have

$$\begin{aligned} \|\nu_t - K_t^* \xi_t\| &= \|\Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{\nu}_{t|t} - K_t^* \xi_t\| \\ &= \|\Gamma(\xi_t - \hat{\xi}_{t|t}) + \hat{K}_{t|t} \hat{\xi}_{t|t} - K_t^* \xi_t\|. \end{aligned}$$

With the triangle inequality and the sub-multiplicativity of the 2-norm this expression can be upper bounded by

$$\begin{aligned} \|\nu_t - K_t^* \xi_t\| &= \|(\Gamma - K_t^*)(\xi_t - \hat{\xi}_{t|t}) + (\hat{K}_{t|t} - K_t^*) \hat{\xi}_{t|t}\| \\ &\leq \|(\Gamma - K_t^*)\| \|(\xi_t - \hat{\xi}_{t|t})\| + \|(\hat{K}_{t|t} - K_t^*)\| \|\hat{\xi}_{t|t}\|. \end{aligned}$$

With (B.1) (for  $\tau = t - 1$ ), inequality (B.3) in Lemma B.1.1, inequality (B.6) in Lemma B.1.2 as well as the definitions of  $\bar{\Phi}_{t|t}^*$  and  $\Delta$  in (3.31) and (3.32i), respectively, the following inequalities hold

$$\begin{aligned} \|\nu_t - K_t^* \xi_t\| &\leq \|(\hat{K}_{t|t} - K_t^*)\| \prod_{n=0}^{t-1} (A + B \hat{K}_{n|n}) \bar{x}_0 + \|(\Gamma - K_t^*)\| \|(\xi_t - \hat{\xi}_{t|t})\| \\ &\leq \|\bar{x}_0\| \left( \bar{\Phi}_{t|t}^* \hat{C} \hat{\eta}^t + \Delta C_q \hat{C}^2 \|B\| \eta_q^{t-1} \sum_{j=1}^t \sum_{n=0}^{j-1} \left(\frac{\hat{\eta}}{\eta_q}\right)^j \hat{\Phi}_{n|(j-1,j)} \right). \end{aligned}$$

Finally, combining this estimate with (B.10), we arrive at

$$\begin{aligned}
 & \text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) \\
 & \leq D \sum_{t=0}^{T-1} \|\nu_t - K_t^* \xi_t\|^2 \\
 & \leq 2D \|\bar{x}_0\|^2 \left( \sum_{t=1}^{T-1} (\hat{C} \hat{\eta}^t \bar{\Phi}_{t|t}^*)^2 + \left[ \Delta C_q \hat{C}^2 \|B\| \eta_q^{t-1} \sum_{j=1}^t \sum_{n=0}^{j-1} \left(\frac{\hat{\eta}}{\eta_q}\right)^j \hat{\Phi}_{n|(j-1,j)} \right]^2 \right)
 \end{aligned}$$

and where we have additionally used the fact that  $(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2)$  for  $a_1, a_2 \in \mathbb{R}$ . ■

### B.3 Preparatory Results for the Proof of Theorem 3.4.1

To go from the regret bound in Lemma 3.4.1 to the regret bound in Theorem 3.4.1 we need to derive estimates on  $\hat{\Phi}_{n|(j-1,j)}$  and  $\bar{\Phi}_{t|t}^*$ . To this end, we first recall the Thompson metric  $\delta_\infty(\cdot, \cdot)$  defined as

$$\delta_\infty(X, Y) := \|\log(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})\|_\infty \tag{B.11}$$

for positive semi-definite matrices  $X$  and  $Y$  [74, Sec. 2].

We first establish a relationship between  $\delta_\infty(\cdot, \cdot)$  and  $\|\cdot\|$ .

**Lemma B.3.1.** *For any  $X, Y \in \mathbb{S}_{++}^n$ , it holds that*

$$\|X - Y\| \leq \max(\lambda_{\max}(X), \lambda_{\max}(Y)) \delta_\infty(X, Y).$$

*Proof.* From [77, Rem. 2.2], we have

$$\delta_\infty(X, Y) = \max[\log(\lambda_{\max}(Y^{-\frac{1}{2}} X Y^{-\frac{1}{2}})), \log(\lambda_{\max}(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}))]$$

and thus

$$\begin{aligned}
 \delta_\infty(X, Y) &= \max \left( \log \left( \sup_{\|\xi\|=1} \frac{\xi^\top X \xi}{\xi^\top Y \xi} \right), \log \left( \sup_{\|\xi\|=1} \frac{\xi^\top Y \xi}{\xi^\top X \xi} \right) \right) \\
 &= \max_{\|\xi\|=1} |\log(\xi^\top X \xi) - \log(\xi^\top Y \xi)|.
 \end{aligned} \tag{B.12}$$

To proceed, we consider the mapping  $s \mapsto \log(\xi^\top [X + s(Y - X)] \xi)$ . By the mean

value theorem, there exists  $\bar{s} \in (0, 1)$ , such that

$$\max_{\|\xi\|=1} |\log(\xi^\top X \xi) - \log(\xi^\top Y \xi)| = \max_{\|\xi\|=1} \left| \frac{d}{ds} \log(\xi^\top [X + s(Y - X)] \xi) \right| \Big|_{s=\bar{s}}.$$

Thus, computing the derivative of the mapping evaluated at  $\bar{s}$ , we have

$$\max_{\|\xi\|=1} |\log(\xi^\top X \xi) - \log(\xi^\top Y \xi)| = \max_{\|\xi\|=1} \left| \frac{\xi^\top (Y - X) \xi}{\xi^\top [X + \bar{s}(Y - X)] \xi} \right|.$$

For any  $\|\xi\| = 1$  and any  $\bar{s} \in (0, 1)$ , we have

$$\begin{aligned} \xi^\top [X + \bar{s}(Y - X)] \xi &\leq \lambda_{\max}(X + \bar{s}(Y - X)) \\ &= \lambda_{\max}(\bar{s}Y + (1 - \bar{s})X). \end{aligned}$$

By Weyl's inequality (see [76, Theorem 4.3.1], for example) and the property  $\lambda_{\max}(\bar{s}H) = \bar{s}\lambda_{\max}(H)$  for  $\bar{s} \in [0, 1]$  and  $H \in \mathbb{S}_{++}^n$ , we have

$$\begin{aligned} \lambda_{\max}(\bar{s}Y + (1 - \bar{s})X) &\leq \bar{s}\lambda_{\max}(Y) + (1 - \bar{s})\lambda_{\max}(X) \\ &\leq \max(\lambda_{\max}(X), \lambda_{\max}(Y)). \end{aligned}$$

Therefore, combining all the above, yields

$$\begin{aligned} \delta_\infty(X, Y) &= \max_{\|\xi\|=1} |\log(\xi^\top X \xi) - \log(\xi^\top Y \xi)| \\ &\geq \max_{\|\xi\|=1} \frac{|\xi^\top (X - Y) \xi|}{\max(\lambda_{\max}(X), \lambda_{\max}(Y))} \\ &= \frac{|\lambda_{\max}(X - Y)|}{\max(\lambda_{\max}(X), \lambda_{\max}(Y))}, \end{aligned}$$

where the last equality follows from the fact that  $X - Y$  is symmetric and thus  $|\lambda_{\max}(X - Y)| = \|X - Y\|$  and which concludes the proof.  $\square$

According to [75, Lem. D.2], for  $V_1, X, Y \in \mathbb{S}_{++}^n$  it holds that  $0 < \frac{\delta_\infty(V_1+X, V_1+Y)}{\delta_\infty(X, Y)} < 1$ . Next, we extend this result to find an upper bound of  $\frac{\delta_\infty(V_1+X, V_2+Y)}{\delta_\infty(X, Y)}$  for any  $V_2 \in \mathbb{S}_{++}^n$ .

**Lemma B.3.2.** *Consider  $X, Y, V_1, V_2 \in \mathbb{S}_{++}^n$ . Then, with*

$$\begin{aligned} \alpha'_1 &:= \lambda_{\max}(V_1 + X), \quad \alpha'_2 := \lambda_{\max}(V_1 + Y), \\ \gamma &:= \frac{\max(\alpha'_1, \alpha'_2)}{\lambda_{\min}(V_1) + \max(\alpha'_1, \alpha'_2)} \end{aligned}$$

the Thomson metric in (B.11) satisfies

$$\delta_\infty(V_1 + X, V_2 + Y) \leq \gamma \delta_\infty(X, Y) + \frac{\|V_1 - V_2\|}{\min\{\lambda_{\min}(X + V_2), \lambda_{\min}(Y + V_1)\}}.$$

*Proof.* Due to the symmetry of the Thomson metric, i.e.  $\delta_\infty(X + V_1, Y + V_2) = \delta_\infty(Y + V_2, X + V_1)$ , we can assume without loss of generality that  $\sup_{\|\xi\|=1} \frac{\xi^\top(X+V_1)\xi}{\xi^\top(Y+V_2)\xi} \geq 1$ .

Using this assumption, from [77, Rem 2.2] and the expression in (B.12), we have

$$\delta_\infty(X + V_1, Y + V_2) = \log \left( \sup_{\|\xi\|=1} \frac{\xi^\top(X + V_1)\xi}{\xi^\top(Y + V_2)\xi} \right). \quad (\text{B.13})$$

Note that

$$\sup_{\|\xi\|=1} \frac{\xi^\top(X + V_1)\xi}{\xi^\top(Y + V_2)\xi} \leq \sup_{\|\xi\|=1} \frac{\xi^\top(X + V_1)\xi}{\xi^\top(X + V_2)\xi} \sup_{\|\xi\|=1} \frac{\xi^\top(X + V_2)\xi}{\xi^\top(Y + V_2)\xi},$$

and thus, proceeding with (B.13), we have

$$\begin{aligned} & \delta_\infty(X + V_1, Y + V_2) \\ & \leq \log \left( \sup_{\|\xi\|=1} \frac{\xi^\top(X + V_1)\xi}{\xi^\top(X + V_2)\xi} \right) + \log \left( \sup_{\|\xi\|=1} \frac{\xi^\top(X + V_2)\xi}{\xi^\top(Y + V_2)\xi} \right) \\ & \leq \log \left( 1 + \sup_{\|\xi\|=1} \frac{\xi^\top(V_1 - V_2)\xi}{\xi^\top(X + V_2)\xi} \right) + \delta_\infty(X + V_2, Y + V_2). \end{aligned}$$

Before we proceed with the last expression, we note that

$$\sup_{\|\xi\|=1} \frac{\xi^\top(V_1 - V_2)\xi}{\xi^\top(X + V_2)\xi} \leq \frac{|\lambda_{\max}(V_1 - V_2)|}{\lambda_{\min}(X + V_2)}.$$

and by applying the inequality  $\log(1 + r) \leq r$  for  $r \in \mathbb{R}_{\geq 0}$ , we have

$$\begin{aligned} \log \left( 1 + \sup_{\|\xi\|=1} \frac{\xi^\top(V_1 - V_2)\xi}{\xi^\top(X + V_2)\xi} \right) & \leq \frac{|\lambda_{\max}(V_1 - V_2)|}{\lambda_{\min}(X + V_2)} \\ & = \frac{\|V_1 - V_2\|}{\lambda_{\min}(X + V_2)}. \end{aligned}$$

Finally, by [75, Lem. D.2], we have  $\delta_\infty(X + V_2, Y + V_2) \leq \gamma \delta_\infty(X, Y)$  and combining the above inequalities we can conclude that

$$\delta_\infty(X + V_1, Y + V_2) \leq \frac{\|V_1 - V_2\|}{\lambda_{\min}(X + V_2)} + \gamma \delta_\infty(X, Y).$$

With the assumption  $\sup_{\|\xi\|=1} \frac{\xi^\top(X+V_1)\xi}{\xi^\top(Y+V_2)\xi} \geq 1$ , which we have used at the beginning of the proof, this completes the proof.  $\square$

With Lemma B.3.2, we are ready to establish upper bounds for  $\bar{\Phi}_{t|t}^*$  and  $\hat{\Phi}_{t|(t,t_0)}$ .

**Lemma B.3.3.** *For  $t, j \in \mathbb{N}_{T-1}$  and  $i \in \mathbb{N}_{j-1}$ , consider  $\hat{K}_{\cdot|j}$  and  $K^*$  defined in (3.28) and (B.2), respectively. There exist positive constants  $C_Q, C_R, \hat{C}_Q, \hat{C}_R \in \mathbb{R}_{>0}$ , such that*

$$\bar{\Phi}_{t|t}^* < C_Q \|\hat{Q}_t - Q^*\| + C_R \|\hat{R}_t - R^*\|, \quad (\text{B.14})$$

$$\hat{\Phi}_{i|(j-1,j)} < \hat{C}_Q \|\hat{Q}_{j-1} - \hat{Q}_j\| + \hat{C}_R \|\hat{R}_{j-1} - \hat{R}_j\|. \quad (\text{B.15})$$

Moreover, the constants  $C_Q$  and  $C_R$  depend on  $\lambda_{\min}(Q^*)$ ,  $\lambda_{\max}(Q^*)$ ,  $M_Q$ ,  $M_R$  and  $\|B\|$ , where  $M_Q$ ,  $M_R$  are defined in (3.35), and the constants  $\hat{C}_Q$  and  $\hat{C}_R$  depend on  $\min_{t \in \mathbb{N}_{T-1}} \lambda_{\min}(\hat{Q}_t)$ ,  $M_Q$ ,  $M_R$  and  $\|B\|$ , respectively.

*Proof.* We start with a derivation of (B.14). The proof relies on the following steps. We first apply Lemma B.3.1 to  $X = \hat{P}_{t|t}$  and  $Y = P_t^*$  to obtain a bound on  $\|\hat{P}_{t|t} - P_t^*\|$ . Then, as a second step, it will be sufficient to establish an upper bound on  $\|\hat{K}_{t|t} - K_t^*\|$  to show (B.14).

We now start upper bounding  $\delta_\infty(\hat{P}_{t|t}, P_t^*)$  as follows. For  $0 \leq t \leq T-1$ , define

$$H_1 := \min \left\{ \min_{t \in \mathbb{N}_{T-1}} [\lambda_{\min}(Q^* + A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A)]^{-1}, \right. \\ \left. \min_{t \in \mathbb{N}_{T-1}} [\lambda_{\min}(\hat{Q}_t + A^\top(P_{t+1}^{-1*} + BR^{-1*}B^\top)^{-1}A)]^{-1} \right\},$$

$$H_2 := \min \left\{ \min_{t \in \mathbb{N}_{T-1}} [\lambda_{\min}(\hat{P}_{t+1|t}^{-1} + BR^{*-1}B^\top)^{-1}]^{-1}, \right. \\ \left. \min_{t \in \mathbb{N}_{T-1}} [\lambda_{\min}(P_{t+1}^{-1*} + B\hat{R}_t^{-1}B^\top)^{-1}]^{-1} \right\},$$

$$\alpha_1 := \max_{t \in \mathbb{N}_{T-1}} (\lambda_{\max}(A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A), \\ \lambda_{\max}(A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A),$$

$$\hat{\alpha}_1 := \max_{t \in \mathbb{N}_{T-1}} (\lambda_{\max}(P_{t+1}^*), \lambda_{\max}(\hat{P}_{t+1|t})),$$

$$\hat{\beta}_1 := \min_{t \in \mathbb{N}_{T-1}} (\lambda_{\min}(BR^{*-1}B^\top), \lambda_{\min}(B\hat{R}_t^{-1}B^\top)),$$

$$\gamma_1 := \frac{\alpha_1}{\lambda_{\min}(Q^*) + \alpha_1}, \gamma_2 := \frac{\hat{\alpha}_1}{\hat{\alpha}_1 + \hat{\beta}_1}.$$

By definition of  $\hat{P}_{t|t}$  and  $P_t^*$ , we have

$$\delta_\infty(\hat{P}_{t|t}, P_t^*) = \delta_\infty(\hat{Q}_t + A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A, Q^* + A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A)$$

Then, applying Lemma B.3.2 to the above, we have

$$\begin{aligned} & \delta_\infty(\hat{Q}_t + A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A, Q^* + A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A) \\ & \leq \gamma_1 \delta_\infty(A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A, A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A) + H_1 \|\hat{Q}_t - Q^*\| \end{aligned} \quad (\text{B.16})$$

By [75, Lemma D.1 (i)], we have

$$\begin{aligned} & \delta_\infty(A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A, A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A) \\ & = \delta_\infty((\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}, (P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}). \end{aligned}$$

By substituting the above to (B.16), we have

$$\begin{aligned} & \gamma_1 \delta_\infty(A^\top(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top)^{-1}A, A^\top(P_{t+1}^{*-1} + BR^{*-1}B^\top)^{-1}A) + H_1 \|\hat{Q}_t - Q^*\| \\ & = \gamma_1 \delta_\infty(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top, P_{t+1}^{*-1} + BR^{*-1}B^\top) + H_1 \|\hat{Q}_t - Q^*\|. \end{aligned}$$

By applying Lemma B.3.2 to the above again, we further yield

$$\begin{aligned} & \gamma_1 \delta_\infty(\hat{P}_{t+1|t}^{-1} + B\hat{R}_t^{-1}B^\top, P_{t+1}^{*-1} + BR^{*-1}B^\top) + H_1 \|\hat{Q}_t - Q^*\| \\ & \leq \gamma_1(\gamma_2 \delta_\infty(\hat{P}_{t+1|t}^{-1}, P_{t+1}^{-1*}) + H_2 \|\hat{R}_t - R^*\|) + H_1 \|\hat{Q}_t - Q^*\|. \end{aligned}$$

Again by [75, Lemma D.1 (i)], we have  $\delta_\infty(\hat{P}_{t+1|t}^{-1}, P_{t+1}^{-1*}) = \delta_\infty(\hat{P}_{t+1|t}, P_{t+1}^*)$ .

Combining the above, yield

$$\delta_\infty(\hat{P}_{t|t}, P_t^*) \leq \gamma_1 \gamma_2 \delta_\infty(\hat{P}_{t+1|t}, P_{t+1}^*) + \gamma_1 H_2 \|\hat{R}_t - R^*\| + H_1 \|\hat{Q}_t - Q^*\|.$$

By expanding the recursion between  $\delta_\infty(\hat{P}_{t|t}, P_t^*)$  and  $\delta_\infty(\hat{P}_{t+1|t}, P_{t+1}^*)$  using the above inequality, we have

$$\begin{aligned} \delta_\infty(\hat{P}_{t|t}, P_t^*) & \leq [(\gamma_1 \gamma_2)^{T-t} \delta_\infty(\hat{Q}_t, Q^*) + H_1 \gamma_1 \|\hat{R}_t - R^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k] \\ & \quad + H_2 \|\hat{Q}_t - Q^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k. \end{aligned}$$

Note that  $\hat{P}_{\tau|t}$  for  $\tau \in \mathbb{N}_{T-1}$  are computing using  $\hat{Q}_t$  and  $\hat{R}_t$  according to (3.25)

(with the argument of  $F_\tau(\cdot)$  being  $\hat{\theta}_t$ ). Therefore, the above recursion expansion only involves  $\hat{Q}_t$  and  $\hat{R}_t$ , given estimation of  $\theta^*$  at time  $t$ .

In the above, we established an upper bound of  $\delta_\infty(\hat{P}_{t|t}, P_t^*)$ . By using Lemma B.3.1, we can establish an upper bound of  $\|\hat{P}_{t|t} - P_t^*\|$ . Let

$$H_P := \max_{t \in \mathbb{N}_T} (\lambda_{\max}(\hat{P}_{t|t}), \lambda_{\max}(P_t^*)),$$

then applying Lemma B.3.1 to  $\delta_\infty(\hat{P}_{t|t}, P_t^*)$ , we have

$$\begin{aligned} \|\hat{P}_{t|t} - P_t^*\| &\leq H_P \delta_\infty(\hat{P}_{t|t}, P_t^*) \\ &\leq H_P [(\gamma_1 \gamma_2)^{T-t} \delta_\infty(\hat{Q}_t, Q^*) + H_1 \gamma_1 \|\hat{R}_t - R^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k] + H_2 \|\hat{Q}_t - Q^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k. \end{aligned}$$

Furthermore, applying Lemma B.3.1 to  $\delta_\infty(\hat{Q}_t, Q^*)$ , let  $H_Q := \max_{t \in \mathbb{N}_{T-1}} (\lambda_{\max}(\hat{Q}_t), \lambda_{\max}(Q^*))$ , we have

$$\|\hat{Q}_t - Q^*\| \leq H_Q \delta_\infty(\hat{Q}_t, Q^*).$$

We continue establishing upper bound of  $\|\hat{P}_{t|t} - P_t^*\|$  as follows

$$\begin{aligned} &H_P [(\gamma_1 \gamma_2)^{T-t} \delta_\infty(\hat{Q}_t, Q^*) + H_1 \gamma_1 \|\hat{R}_t - R^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k] + H_2 \|\hat{Q}_t - Q^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k \\ &\leq [H_P H_Q (\gamma_1 \gamma_2)^{T-t} + H_2 \|\hat{Q}_t - Q^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k] + H_1 \gamma_1 \|\hat{R}_t - R^*\| \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k. \end{aligned}$$

By the property of the geometric sum, due to  $\gamma_1, \gamma_2 \in (0, 1)$ , the series  $\sum_{k=t}^{\infty} (\gamma_1 \gamma_2)^k$  converges to a finite value. Thus,

$$\sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k < \sum_{k=t}^{\infty} (\gamma_1 \gamma_2)^k = \frac{1}{1 - \gamma_1 \gamma_2}.$$

Additionally, we have  $(\gamma_1 \gamma_2)^{T-t} < 1$ . Thus, we have

$$\begin{aligned} &[H_P H_Q (\gamma_1 \gamma_2)^{T-t} + H_2 \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k] \|\hat{Q}_t - Q^*\| + H_1 \gamma_1 \sum_{k=t}^{T-1} (\gamma_1 \gamma_2)^k \|\hat{R}_t - R^*\| \\ &< [H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2}] \|\hat{Q}_t - Q^*\| + \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2} \|\hat{R}_t - R^*\|. \end{aligned}$$

Therefore,

$$\|\hat{P}_{t|t} - P_t^*\| < [H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2}] \|\hat{Q}_t - Q^*\| + \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2} \|\hat{R}_t - R^*\|.$$

Next, we establish an upper bound of  $\|\hat{K}_{t|t} - K_t^*\|$ . Note that

$$\|\hat{K}_{t|t} - K_t^*\| = \| -(\hat{R}_t + B^\top \hat{P}_{t+1|t} B)^{-1} B^\top \hat{P}_{t+1|t} A + (R^* + B^\top P_{t+1}^* B)^{-1} B^\top P_{t+1}^* A \|.$$

Let  $G_1 := (\hat{R}_t + B^\top \hat{P}_{t+1|t} B)$  and  $G_2 := (R^* + B^\top P_{t+1}^* B)$ , we have

$$\begin{aligned} \|\hat{K}_{t|t} - K_{t|t}\| &= \|G_1^{-1} B^\top \hat{P}_{t+1|t} - G_2^{-1} B^\top P_{t+1}^*\| \\ &= \|G_1^{-1} G_2^{-1} G_2 B^\top \hat{P}_{t+1|t} - G_2^{-1} G_1^{-1} G_1 B^\top P_{t+1}^*\|. \end{aligned} \quad (\text{B.17})$$

Note that

$$\begin{aligned} &G_1^{-1} G_2^{-1} G_2 B^\top \hat{P}_{t+1|t} - G_2^{-1} G_1^{-1} G_1 B^\top P_{t+1}^* \\ &= G_2^{-1} G_1^{-1} ((G_1 G_2)(G_1^{-1} G_2^{-1}) G_2 B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^*). \end{aligned}$$

Adding to and subtracting from  $G_2 B^\top \hat{P}_{t+1|t}$  the right-hand-side of the above, and note that  $G_2 B^\top \hat{P}_{t+1|t} = G_2 G_1 G_1^{-1} B^\top \hat{P}_{t+1|t}$ , we have

$$\begin{aligned} &((G_1 G_2)(G_1^{-1} G_2^{-1}) G_2 B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^*) \\ &= ((G_1 G_2) G_1^{-1} B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^* + G_2 B^\top \hat{P}_{t+1|t} - G_2 B^\top \hat{P}_{t+1|t}) \\ &= ((G_1 G_2) G_1^{-1} B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^* + G_2 B^\top \hat{P}_{t+1|t} - G_2 G_1 G_1^{-1} B^\top \hat{P}_{t+1|t}) \\ &= ((G_1 G_2 - G_2 G_1) G_1^{-1} B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^* + G_2 B^\top \hat{P}_{t+1|t}). \end{aligned}$$

Therefore, by the sub-multiplicative property of matrix norm, with the above calculations, (B.17) can be upper bounded by

$$\begin{aligned} &\|G_1^{-1} B^\top \hat{P}_{t+1|t} - G_2^{-1} B^\top P_{t+1}^*\| \leq \\ &\|G_2^{-1} G_1^{-1}\| \|((G_1 G_2 - G_2 G_1) G_1^{-1} B^\top \hat{P}_{t+1|t} - G_1 B^\top P_{t+1}^* + G_2 B^\top \hat{P}_{t+1|t})\| \end{aligned}$$

Applying the triangle inequality, we have

$$\begin{aligned} &\|G_2^{-1} G_1^{-1}\| (\|((G_1 G_2 - G_2 G_1) G_1^{-1} B^\top \hat{P}_{t+1|t} + G_1 B^\top P_{t+1}^* - G_2 B^\top \hat{P}_{t+1|t})\| \\ &\leq \|G_2^{-1} G_1^{-1}\| (\|((G_1 G_2 - G_2 G_1) G_1^{-1} B^\top \hat{P}_{t+1|t}\| + \|G_1 B^\top P_{t+1}^* - G_2 B^\top \hat{P}_{t+1|t}\|)) \end{aligned}$$

Since  $G_1$  and  $G_2$  are real symmetric matrices, we have  $G_1G_2 - G_2G_1$  being skew symmetric. Furthermore, by the property that the norm of a skew-symmetric matrix is zero, we have  $\|G_1G_2 - G_2G_1\| = 0$ . Therefore, by the sub-multiplicative property of matrix norm, we have

$$\begin{aligned} & \|G_2^{-1}G_1^{-1}\| \left( \|(G_1G_2 - G_2G_1)G_1^{-1}B^\top \hat{P}_{t+1|t}\| + \|G_1B^\top P_{t+1}^* - G_2B^\top \hat{P}_{t+1|t}\| \right) \\ & \leq \|G_2^{-1}G_1^{-1}\| \left( \|G_1G_2 - G_2G_1\| \|G_1^{-1}B^\top \hat{P}_{t+1|t}\| + \|G_2B^\top \hat{P}_{t+1|t} - G_1B^\top P_{t+1}^*\| \right) \\ & = \|G_2^{-1}G_1^{-1}\| \|G_2B^\top \hat{P}_{t+1|t} - G_1B^\top P_{t+1}^*\|. \end{aligned} \quad (\text{B.18})$$

Moreover, by substituting  $G_1$  and  $G_2$ , and applying the triangle inequality to the above, we have

$$\begin{aligned} & \|G_2B^\top \hat{P}_{t+1|t} - G_1B^\top P_{t+1}^*\| \\ & = \|(R^* + B^\top P_{t+1}^* B)B^\top \hat{P}_{t+1|t} - (\hat{R}_t + B^\top \hat{P}_{t+1|t} B)B^\top P_{t+1}^*\| \\ & = \|R^*B^\top \hat{P}_{t+1|t} - \hat{R}_t B^\top P_{t+1}^* + B^\top P_{t+1}^* BB^\top \hat{P}_{t+1|t} - B^\top \hat{P}_{t+1|t} BB^\top P_{t+1}^*\| \\ & \leq \|R^*B^\top (\hat{P}_{t+1|t} - P_{t+1}^*)\| + \|(R^* - \hat{R}_t)B^\top P_{t+1}^*\| + \\ & \quad + \|B\| \|P_{t+1}^* (BB^\top) \hat{P}_{t+1|t} - \hat{P}_{t+1|t} (BB^\top) P_{t+1}^*\| \\ & = \|R^*B^\top\| \|\hat{P}_{t+1|t} - P_{t+1}^*\| + \|(R^* - \hat{R}_t)\| \|B^\top P_{t+1}^*\|, \end{aligned} \quad (\text{B.19})$$

where again the last step is due to  $\hat{P}_{t+1|t}(BB^\top)P_{t+1}^* - P_{t+1}^*(BB^\top)\hat{P}_{t+1|t}$  being real skew symmetric and the matrix 2-norm is 0. Therefore,

$$\begin{aligned} & \|K_t^* - \hat{K}_{t|t}\| \\ & \leq \|G_2^{-1}G_1^{-1}\| \left( \|R^*B^\top\| \|\hat{P}_{t+1|t} - P_{t+1}^*\| + \|(R^* - \hat{R}_t)\| \|B^\top P_{t+1}^*\| \right) \\ & < \|G_2^{-1}G_1^{-1}\| \\ & \left( \|R^*B^\top\| \left( H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2} \right) \|\hat{Q}_t - Q^*\| + \|(\hat{R}_t - R^*)\| \left( \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2} + \|B^\top P_{t+1}^*\| \right) \right). \end{aligned}$$

Let

$$\begin{aligned} \lambda_{\max}^r & := \lambda_{\max}(R^*), \\ \lambda_{\max}^g & := \left( \max_{t \in \mathbb{N}_{T-1}} \lambda_{\max}[(\hat{R}_t + B^\top \hat{P}_{t+1|t} B)^{-1}], \lambda_{\max}[(R^* + B^\top P_{t+1}^* B)^{-1}] \right)^2, \text{ and} \\ \lambda_{\max}^p & := \max_{t \in \mathbb{N}_{T-1}} \lambda_{\max}(P_t^*). \end{aligned}$$

Note that,

$$\|G_2^{-1}G_1^{-1}\| \leq \lambda_{\max}^g \text{ and } \|R^*B\| \leq \lambda_{\max}^r \|B\|.$$

Furthermore, let

$$C_Q := \|B\| \lambda_{\max}^g \lambda_{\max}^r \left( H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2} \right),$$

$$C_R := \|B\| \lambda_{\max}^g \lambda_{\max}^p \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2},$$

we have

$$\begin{aligned} & \|G_2^{-1}G_1^{-1}\| \\ & \left[ \|R^*B^\top\| \left( H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2} \right) \|\hat{Q}_t - Q^*\| + \|(\hat{R}_t - R^*)\| \left( \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2} + \|B^\top P_{t+1}^*\| \right) \right] \\ & \leq C_Q \|\hat{Q}_t - Q^*\| + C_R \|\hat{R}_t - R^*\|. \end{aligned}$$

The upper bound of  $\bar{\Phi}_{i|(j-1,j)} = \|\hat{K}_{i|j-1} - \hat{K}_{i|j}\|$  will be followed by a very similar procedure to the above. We provide some key steps as follows. Let  $\hat{G}_1 := \hat{R}_{j-1} + B^\top \hat{P}_{i|j-1} B$  and  $\hat{G}_2 := \hat{R}_j + B^\top \hat{P}_{i|j} B$ , follow by the calculations in (B.17), (B.18) and (B.19), we have

$$\begin{aligned} \|\hat{K}_{i|j-1} - \hat{K}_{i|j}\| &= \|\hat{G}_1^{-1} \hat{G}_2^{-1} \hat{G}_2 B^\top \hat{P}_{i+1|j-1} - \hat{G}_2^{-1} \hat{G}_1^{-1} \hat{G}_1 B^\top \hat{P}_{i+1|j}\| \\ &\leq \|\hat{G}_2^{-1} \hat{G}_1^{-1}\| \|\hat{G}_2 B^\top \hat{P}_{i+1|j-1} - \hat{G}_1 B^\top \hat{P}_{i+1|j}\| \\ &\leq \|\hat{R}_j B^\top\| \|\hat{P}_{i|j-1} - \hat{P}_{i|j}\| + \|\hat{R}_j - \hat{R}_{j-1}\| \|B^\top \hat{P}_{i|j}\|. \end{aligned}$$

Let

$$\hat{\lambda}_{\max}^g := \left( \max_{i \leq k \leq T-1} \max_{j \in \mathbb{N}_{T-1}} (\hat{R}_j + B^\top \hat{P}_{i+1|j} B)^{-1} \right)^2,$$

$$\hat{\lambda}_{\max}^p := \max_{i,j \in \mathbb{N}_{T-1}} \lambda_{\max}(\hat{P}_{i|j}),$$

we have

$$\|\hat{G}_2^{-1} \hat{G}_1^{-1}\| \leq \hat{\lambda}_{\max}^g, \text{ and } \|\hat{R}_j B^\top\| \leq \hat{\lambda}_{\max}^r \|B\|.$$

Define

$$\begin{aligned}\hat{C}_Q &:= \|B\| \hat{\lambda}_{\max}^g \hat{\lambda}_{\max}^r \left( H_P H_Q + \frac{H_2}{1 - \gamma_1 \gamma_2} \right), \\ \hat{C}_R &:= \|B\| \hat{\lambda}_{\max}^g \hat{\lambda}_{\max}^p \frac{H_1 \gamma_1}{1 - \gamma_1 \gamma_2},\end{aligned}$$

combing all the above, yield

$$\|\hat{K}_{i|j-1} - \hat{K}_{i|j}\| \leq \hat{C}_Q \|\hat{Q}_{j-1} - \hat{Q}_j\| + \hat{C}_R \|\hat{R}_{j-1} - \hat{R}_j\|.$$

□

The above proposition suggests that  $\hat{\Phi}_{n|(j-1,j)}$  and  $\bar{\Phi}_{t|t}^*$  depend on the distances  $\|\hat{Q}_t - Q^*\|$ ,  $\|\hat{Q}_{j-1} - \hat{Q}_j\|$ ,  $\|\hat{R}_t - R^*\|$  and  $\|\hat{R}_{j-1} - \hat{R}_j\|$ , these are associated with errors  $\|\hat{\theta}_t - \theta^*\|$  and  $\|\hat{\theta}_{j-1} - \hat{\theta}_j\|$ .

To bound the terms  $\|\hat{Q}_t - Q^*\|$  and  $\|\hat{R}_t - R^*\|$  from the right-hand side of the inequality (3.36), we establishing related properties in the following. We first introduce an elementary inequality that is related to the matrix infinity norm.

**Proposition 6.** *For any real symmetric matrices  $H_1, H_2 \in \mathbb{R}^n$ , we have*

$$\|H_1^\top H_1 - H_2^\top H_2\| \leq (\|H_1\| + \|H_2\|) \|H_1 - H_2\|.$$

*Proof.* Note that

$$\begin{aligned}\|H_1^\top H_1 - H_2^\top H_2\| &\leq \|(H_1 + H_2)^\top (H_1 - H_2) + (H_1^\top H_2 - H_2^\top H_1)\| \\ &\leq \|(H_1 + H_2)^\top (H_1 - H_2)\| + \|(H_1^\top H_2 - H_2^\top H_1)\|\end{aligned}$$

Matrix  $(H_1^\top H_2 - H_2^\top H_1)$  is real skew-symmetric. For any skew-symmetric matrix  $H$ , we have  $\|H\| = 0$ . Thus,  $\|(H_1^\top H_2 - H_2^\top H_1)\| = 0$ , and

$$\begin{aligned}\|H_1^\top H_1 - H_2^\top H_2\| &= \|(H_1 + H_2)^\top (H_1 - H_2)\| \\ &\leq (\|H_1\| + \|H_2\|) \|H_1 - H_2\|.\end{aligned}$$

□

We next establish inequalities to characterise the upper bounds of  $\bar{\Phi}_{t|t}^*$  and  $\hat{\Phi}_{i|(j-1,j)}$  using the above Proposition 6 and Assumption 3.3.1.

**Lemma B.3.4.** *For any  $T \geq 1$ ,  $i, t \in \mathbb{N}_{T-1}$  and  $1 \leq j \leq T - 1$ , consider  $\bar{\Phi}_{t|t}^*$  and  $\hat{\Phi}_{i|(j-1,j)}$  defined in (3.31), we have the following inequalities*

$$\begin{aligned}\bar{\Phi}_{t|t}^* &\leq (C_Q \bar{M}_Q + C_R \bar{M}_R)(C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^t + \alpha(\|v_\theta\|)), \text{ and} \\ \hat{\Phi}_{i|(j-1,j)} &\leq 2(\hat{C}_Q \bar{M}_Q + \hat{C}_R \bar{M}_R)(C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^{j-1} + \alpha(\|v_\theta\|)),\end{aligned}$$

where  $L_Q$  and  $L_R$  are defined in (3.9),  $M_Q$  and  $M_R$  are defined in (3.35),  $C_\theta$ ,  $\eta_\theta$ ,  $\alpha(\cdot)$  and  $\|v_\theta\|$  are defined in Assumption 3.3.1,  $C_Q$ ,  $\hat{C}_Q$ ,  $\hat{C}_R$  and  $C_R$  are defined in Lemma B.3.3, with  $\bar{M}_Q := M_Q L_Q$  and  $\bar{M}_R := M_R L_R$ .

*Proof.* By definition of  $Q^*$  and  $\hat{Q}_t$  from (3.7), we have

$$\|\hat{Q}_t - Q^*\| = \|D^Q(\hat{\theta}_t)^\top D^Q(\hat{\theta}_t) - D^Q(\theta^*)^\top D^Q(\theta^*)\|.$$

By Proposition 6 and the Lipschitz property of  $D^Q$  in Assumption 3.2.2, the above can be upper bounded by

$$\begin{aligned}\|D^Q(\hat{\theta}_t)^\top D^Q(\hat{\theta}_t) - D^Q(\theta^*)^\top D^Q(\theta^*)\| &\leq (\|D^Q(\hat{\theta}_t)^\top\| + \|D^Q(\theta^*)^\top\|)\|D^Q(\hat{\theta}_t) - D^Q(\theta^*)\| \\ &\leq M_Q L_Q \|\hat{\theta}_t - \theta^*\|.\end{aligned}$$

Finally, by Assumption 3.3.1, we have  $\|\hat{\theta}_t - \theta^*\| \leq C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^t + \alpha(\|v_\theta\|)$ . This implies  $\|\hat{Q}_t - Q^*\| \leq \bar{M}_Q (C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^t + \alpha(\|v_\theta\|))$ .

Repeat the above steps to matrices  $\hat{R}_t$  and  $R^*$ , we have  $\|\hat{R}_t - R^*\| \leq \bar{M}_R (C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^t + \alpha(\|v_\theta\|))$ . Substitute the above to the right-hand-side of (B.14), we have

$$\bar{\Phi}_{t|t}^* \leq (C_Q \bar{M}_Q + C_R \bar{M}_R)(C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^t + \alpha(\|v_\theta\|)).$$

Repeat the above steps to upper bound  $\|\hat{Q}_j - \hat{Q}_{j-1}\|$  and  $\|\hat{R}_j - \hat{R}_{j-1}\|$ , we have

$$\begin{aligned}\|\hat{\theta}_j - \hat{\theta}_{j-1}\| &\leq C_\theta \|\hat{\theta}_0 - \theta^*\|(\eta_\theta^{j-1} + \eta_\theta^j) + 2\alpha(\|v_\theta\|) \\ &\leq 2(C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^{j-1} + \alpha(\|v_\theta\|)).\end{aligned}$$

Thus,

$$\hat{\Phi}_{t|(j-1,j)} \leq 2(\hat{C}_Q \bar{M}_Q + \hat{C}_R \bar{M}_R)(C_\theta \|\hat{\theta}_0 - \theta^*\|\eta_\theta^{j-1} + \alpha(\|v_\theta\|)).$$

□

## B.4 Proof of Theorem 3.4.1

By Lemma 3.4.1, the regret incurred by the copycat defined in (3.15) satisfies

$$\text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) \leq 2D\|\bar{x}_0\|^2 \sum_{t=1}^{T-1} (\hat{C}\hat{\eta}^t \bar{\Phi}_{t|t}^*)^2 + (\Delta C_q \hat{C}^2 \|B\| \eta_q^{t-1} \sum_{j=1}^t \sum_{i=0}^{j-1} (\frac{\hat{\eta}}{\eta_q})^j \hat{\Phi}_{n|(j-1,j)})^2.$$

Then, we substitute the upper bounds of  $\hat{\Phi}_{i|(j-1,j)}$  and  $\bar{\Phi}_{t|t}^*$  established in Lemma B.3.4 to the right-hand-side of the above, let  $E_1(z) := \sum_{t=0}^{T-1} tz^t$  and  $E_2(z) := \sum_{t=0}^{T-1} z^t$  for  $z \in \mathbb{R}$ , yields

$$\begin{aligned} & \text{Regret}_T(\{\nu_t\}_{t=0}^{T-1}) \\ & \leq 2D \sum_{t=0}^{T-1} (\hat{C}\hat{\eta}^t \bar{\Phi}_{t|t}^*)^2 + (\Delta C_q \|B\| \hat{C}^2 \eta_q^{t-1} \sum_{j=1}^t \sum_{i=0}^{j-1} (\frac{\hat{\eta}}{\eta_q})^j \hat{\Phi}_{i|(j-1,j)})^2 \\ & < 4D \left[ \hat{C}^2 C_{\phi^*}^2 (C_\theta^2 E_1((\hat{\eta}\eta_\theta)^2) + v_\theta^2 E_2(\hat{\eta}^2)) + (\Delta C_q \|B\| \hat{C}^2 C_{\hat{\phi}} \frac{\hat{\eta}}{\eta_q})^2 \right. \\ & \quad \left( \frac{C_\theta \|\hat{\theta}_0 - \theta^*\|}{\eta_q - \hat{\eta}\eta_\theta} \right)^2 [E_1(\hat{\eta}\eta_\theta) + \eta_q (E_2(\hat{\eta}_q) - E_2(\hat{\eta}\eta_\theta))]^2 \\ & \quad \left. + \left( \frac{\alpha(\|v_\theta\|)}{\eta_q - \eta} \right)^2 [E_1(\hat{\eta}) + \eta_q (E_2(\hat{\eta}_q) - E_2(\hat{\eta}))]^2 \right], \end{aligned}$$

where  $C_{\phi^*} = C_Q \bar{M}_Q + C_R \bar{M}_R$  and  $C_{\hat{\phi}} = 2(\hat{C}_Q \bar{M}_Q + \hat{C}_R \bar{M}_R)$ .

---

# Appendix. Proof for Dynamic Potential LQ Games

---

## C.1 Auxiliary Lemmas

Before stating the proof of Theorem 4.3.1, in the following, we introduce several necessary lemmas and propositions.

*Proof of Proposition 2:* For any policy  $(\Pi_t)_{t=1}^{T-1}$ , define  $J(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) := \sum_{i=1}^N J_T^i(\bar{x}_1, (\Pi_t)_{t=1}^{T-1})$ . The PoU can be rewritten as  $\text{PoU}_T(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) = J(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) - J(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1})$ . Note that

$$\begin{aligned} \tilde{J}(1 - \text{PoA}_T) &= \tilde{J} - J(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) \\ &\leq J(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) - J(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) \\ &= \text{PoU}_T(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}). \end{aligned}$$

This completes the proof of Proposition 2. ■

*Proof of Corollary 3:* Note that matrix  $A$  is full-rank due to Assumption 4.2.2. From (4.8) and (4.9), we have  $[R_t^i]_{ij} - [R_t^j]_{ji}^\top + B^{i\top}(P_{t+1}^i - P_{t+1}^j)B^j = [R_t^i]_{ij} - [R_t^j]_{ji}^\top = 0$ .

■

**Lemma C.1.1.** *for  $i \in \{1, 2, \dots, N\}$  and  $t \in \mathbb{N}_T$ , consider LQ-DFG parameters  $A, B^i, Q_t, R_t^i$  that satisfy Assumption 4.2.1. Then, the LQ-DFG parameters are an LQ-DFPG in the sense of Definition 4.2.2.*

*Proof.* At time instant  $T$ , set  $\bar{Q}_T = Q_T^1$ . By (4.10), we have  $\bar{P}_T = P_T^1$ , then by (4.9),

for  $i = \{1, \dots, N\}$ , we have  $\mathbf{B}^\top \bar{P}_T A = \mathbf{B}^\top \bar{P}_T^i A$ . Consider  $K_t$  in (4.12), define

$$\begin{aligned}\bar{R}_t &:= \Theta_t - \mathbf{B}^\top \bar{P}_{t+1} \mathbf{B}, \\ \bar{Q}_t &:= Q_t^1 + K_t^\top (R_t^1 - \bar{R}_t) K_t, \\ \bar{P}_t &:= \bar{Q}_t + K_t^\top \bar{R}_t K_t \\ \bar{K}_t &:= -(\bar{R}_t + \mathbf{B}^\top \bar{P}_{t+1} \mathbf{B})^{-1} \mathbf{B}^\top \bar{P}_t A,\end{aligned}\tag{C.1}$$

we claim that for  $i \in \{1, 2, \dots, N\}$ ,  $t \in \mathbb{N}_{T-1}$ , we have  $K_t = \bar{K}_t$  and  $\bar{P}_t = P_t^1$ , and as consequence  $\mathbf{B}^\top \bar{P}_t A = \mathbf{B}^\top P_t^1 A$ . We start with verifying the case of  $t = T - 1$ . Since  $\bar{R}_{T-1} = \Theta_{T-1} - \mathbf{B}^\top \bar{P}_T \mathbf{B}$ , by (4.12), we have

$$\begin{aligned}K_{T-1} &= -\Theta_{T-1}^{-1} \begin{bmatrix} B^{1\top} P_T^1 \\ B^{2\top} P_T^2 \\ \vdots \\ B^{N\top} P_T^N \end{bmatrix} A \\ &\stackrel{(i)}{=} -\Theta_{T-1}^{-1} \mathbf{B}^\top P_T^1 A \\ &\stackrel{(ii)}{=} -(\bar{R}_{T-1} + \mathbf{B}^\top \bar{P}_T \mathbf{B})^{-1} \mathbf{B}^\top \bar{P}_T A = \bar{K}_{T-1}.\end{aligned}$$

Steps (i) and (ii) follow from (4.9) and (4.8), respectively. Then, we have

$$\begin{aligned}\bar{P}_{T-1} - P_{T-1}^1 &= \bar{Q}_{T-1} - Q_{T-1}^1 + K_{T-1}^\top (\bar{R}_{T-1} - R_{T-1}^1) K_{T-1} \\ &\quad + (A + \mathbf{B} K_{T-1})^\top (\bar{P}_T - P_T^1) (A + \mathbf{B} K_{T-1}) \\ &= 0.\end{aligned}$$

Following a similar procedure as above, we can verify by induction that  $K_t = \bar{K}_t$  for  $t \in \mathbb{N}_{T-1}$ . The above proof is similar to the proof of [46, Theorem 6].  $\square$

The above lemma is the  $N$ -player case of [46, Theorem 6] when  $Q_t^i = Q_t^j$  for  $t \in \mathbb{N}_T, i, j \in \{1, \dots, N\}$ .

**Lemma C.1.2.** *Consider an LQ-OCP and an LQ-DFPG described by Definitions 4.2.1 and 4.2.2, respectively. Under Assumption 4.2.3, the feedback Nash Equilibrium for the LQ-DFPG defined in Definition 4.2.2 by parameters  $\{\bar{x}_1, \{Q_t\}_{t=1}^T, \{R_t^i\}_{i=1, t=1}^{N, T-1}\}$ , is identical to the solution of LQ-OCP defined in Definition 4.2.1, by parameters  $\{\bar{x}_1, \{\bar{Q}_t\}_{t=1}^T, \{\bar{R}_t\}_{t=1}^{T-1}\}$ , if for  $i, j \in \{1, 2, \dots, N\}$  and*

$t \in \mathbb{N}_{T-1}$ ,

$$\begin{aligned}
\bar{P}_T &= \bar{Q}_T, \\
\bar{\Theta}_t &= \bar{R}_t + \mathbf{B}^\top \bar{P}_{t+1} \mathbf{B}, \\
\bar{K}_t &= \bar{\Theta}_t^{-1} \mathbf{B}^\top \bar{P}_{t+1} A, \\
\bar{P}_t &= \bar{Q}_t + \bar{K}_t^\top \bar{R}_t \bar{K}_t + (A + \mathbf{B} \bar{K}_t)^\top \bar{P}_{t+1} (A + \mathbf{B} \bar{K}_t), \\
\bar{\Theta}_t &\in \mathbb{S}_{++}^{Nm}, \\
[R_t^i]_{ii} + B^{i\top} P_{t+1}^i B^i &= [\bar{R}_t]_{ii} + B^{i\top} \bar{P}_{t+1} B^i, \\
B^{i\top} P_{t+1}^i A &= B^{i\top} \bar{P}_{t+1} A, \\
[R_t^i]_{ij} + B^{i\top} P_{t+1}^i B^j &= [\bar{R}_t]_{ij} + B^{i\top} \bar{P}_{t+1} B^j, i \neq j, \\
[R_t^i]_{ij} + B^{i\top} P_{t+1}^i B^j &= ([R_t^j]_{ji} + B^{j\top} P_{t+1}^j B^i)^\top.
\end{aligned}$$

*Proof.* The proof of the above lemma follows the proof of [46, Theorem 5] with replacing the cost function that penalises decisions made by 2-players from [46, (21c)] to the objective function  $x_t^\top Q_t x_t + \sum_{n=1}^N \sum_{m=1}^N u_t^{n\top} [R_t^i]_{nm} u_t^m$  for an  $N$ -player setting. Then, applying [46, Theorem 3] establishes the result of the lemma.  $\square$

**Lemma C.1.3.** *For integer  $T \geq 1$ , define*

$$\Omega := \{(\bar{Q}_1, \bar{Q}_2, \dots, \bar{Q}_T, \bar{R}_1, \dots, \bar{R}_{T-1}) | \mathcal{K}\},$$

where  $\mathcal{K}$  is the set of conditions that, at time instant  $T$ ,  $\bar{Q}_T$  is such that  $\mathbf{B}^\top \bar{Q}_T A = \mathbf{B}^\top Q_T A$ . Consider  $\bar{P}_T$  that satisfies  $\mathbf{B}^\top \bar{P}_T A = \mathbf{B}^\top Q_T A$ , and  $t \in \mathbb{N}_{T-1}$ , define  $\bar{R}_t$  as

$$\bar{R}_t := \Theta_t - \mathbf{B}^\top \bar{P}_{t+1} \mathbf{B}, \quad (\text{C.2})$$

where  $\Theta_t \in \mathbb{S}_{++}^{Nm}$ , with  $\bar{Q}_t$  such that  $\bar{Q}_t = Q_t + K_t^\top (R_t^1 - \bar{R}_t) K_t$ , where  $K_t =$

$$\Theta_t^{-1} \begin{bmatrix} B^{1\top} P_{t+1}^1 \\ B^{2\top} P_{t+1}^2 \\ \vdots \\ B^{N\top} P_{t+1}^N \end{bmatrix} A. \text{ Then, every element in } \Omega \text{ leads to an LQ-OCP which is equivalent to an LQ-DFPG. Further, } \Omega \text{ is non-empty.}$$

*Proof.* The proof of the above Lemma is similar to the proof of [46, Theorem 7].  $\square$

To help us present preceding remarks and lemmas, with slight abuse of notations, we define the following operators associated with the computation of parameters in

Assumption 4.2.1,

$$K_t := \mathbf{K}((Q_{\tau+1}, (R_{\tau}^i)_{i=1}^N)_{\tau=t}^{T-1}, A, \mathbf{B}), \quad (\text{C.3})$$

$$\bar{R}_t := \bar{\mathbf{R}}((Q_{\tau+1}, (R_{\tau}^i)_{i=1}^N)_{\tau=t}^{T-1}, A, \mathbf{B}). \quad (\text{C.4})$$

The operators  $\mathbf{K}$  and  $\bar{\mathbf{R}}$  are defined through coupled discrete algebraic Riccati equations, as given in (4.13). These equations are used to compute the matrices  $P_t^i$  for each time step  $t \in \mathbb{N}_T$  and players  $i \in \{1, 2, \dots, N\}$ , based on the specified system dynamics and cost parameters. The returning matrices of the operators depend on the input arguments and the computed matrices  $P_t^i$ , following their respective formulas.

**Corollary 3.** *For  $1 \leq \tau \leq t \leq T - 1$ , under Assumptions 4.2.1, 4.2.2 and 4.2.5, consider  $\bar{R}_\tau$  defined in (C.2) and let  $\bar{R}_{\tau|t}$  be  $\bar{\mathbf{R}}((Q_{k+1|t}, (R_{k|t}^i)_{i=1}^N)_{k=\tau}^{T-1}, A, \mathbf{B})$ , we have  $\bar{R}_\tau = R_\tau^p$ ,  $\bar{R}_{\tau|t} = \bar{R}_\tau$  and  $\bar{R}_{\tau|t}$  symmetric.*

*Proof.* We first rewrite  $\bar{R}_t$  defined in (C.2) as

$$\begin{bmatrix} [R_t^1]_{11} & [R_t^1]_{12} & \cdots & [R_t^1]_{1N} \\ [R_t^2]_{21} & [R_t^2]_{22} & \cdots & [R_t^2]_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ [R_t^N]_{N1} & [R_t^N]_{N2} & \cdots & [R_t^N]_{NN} \end{bmatrix} + \begin{bmatrix} B^{1\top} P_{t+1}^1 \\ B^{2\top} P_{t+1}^2 \\ \vdots \\ B^{N\top} P_{t+1}^N \end{bmatrix} \mathbf{B} - \begin{bmatrix} B^{1\top} \bar{P}_{t+1} \\ B^{2\top} \bar{P}_{t+1} \\ \vdots \\ B^{N\top} \bar{P}_{t+1} \end{bmatrix} \mathbf{B}$$

By Assumption 4.2.2, matrix  $A$  defined in (4.1) is full-rank. By (4.13), (4.8) and (4.9) from Assumption 4.2.1, for  $i \in \{1, 2, \dots, N\}, t \in \mathbb{N}_T$ ,  $P_t^i$  are symmetric and  $B^{i\top} P_t^i = B^{i\top} P_t^j$ , we have  $[R_t^i]_{ij} = ([R_t^j]_{ji})^\top$ , and  $B^{i\top} P_{t+1}^i = B^{i\top} \bar{P}_{t+1}$ . Therefore,

$$\bar{R}_t = \begin{bmatrix} [R_t^1]_{11} & [R_t^1]_{12} & \cdots & [R_t^1]_{1N} \\ [R_t^2]_{21} & [R_t^2]_{22} & \cdots & [R_t^2]_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ [R_t^N]_{N1} & [R_t^N]_{N2} & \cdots & [R_t^N]_{NN} \end{bmatrix} = R_t^p. \text{ By Corollary 2, we have } \bar{R}_t \text{ is sym-}$$

metric. Moreover, by repeating similar procedures as the calculations above, we have

$$\bar{R}_{\tau|t} = \begin{bmatrix} [R_{\tau|t}^1]_{11} & [R_{\tau|t}^1]_{12} & \cdots & [R_{\tau|t}^1]_{1N} \\ [R_{\tau|t}^2]_{21} & [R_{\tau|t}^2]_{22} & \cdots & [R_{\tau|t}^2]_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ [R_{\tau|t}^N]_{N1} & [R_{\tau|t}^N]_{N2} & \cdots & [R_{\tau|t}^N]_{NN} \end{bmatrix}.$$

By  $R_{\tau|t}^i = R_\tau^i$  for  $1 \leq \tau \leq t \leq T - 1$ , we have  $\bar{R}_{\tau|t} = \bar{R}_\tau$ .  $\square$

**Lemma C.1.4.** *Suppose  $m \leq n$ . Consider matrices  $R \in \mathbb{S}_{++}^m$ ,  $P \in \mathbb{S}_{++}^n$  and  $F \in \mathbb{R}^{n \times m}$ , the 2-norm of matrix  $K = (R + F^\top PF)^{-1} F^\top P$  satisfies  $\|K\| \leq \frac{1}{\sigma_{\min}^+(F)}$ .*

*Proof.* Let  $\lambda_R := \lambda_{\min}(R)$ , we immediately have  $\lambda_R I \preceq R$ , and  $(R + F^\top PF)^{-1} \preceq (\lambda_R I + F^\top PF)^{-1}$ . By definition of matrix 2-norm  $\|\cdot\|$ , we have  $\|K\|^2 = \lambda_{\max}(K^\top K)$ . Therefore,

$$\lambda_{\max}(K^\top K) = \lambda_{\max}(PF(R + F^\top PF)^{-1}(R + F^\top PF)^{-1}F^\top P).$$

For any  $H_1, H_2 \in \mathbb{S}_{++}^n$  and  $G \in \mathbb{R}^{m \times n}$ , if  $H_1 \preceq H_2$ , we have

$$\lambda_{\max}(GH_1G^\top) \leq \lambda_{\max}(GH_2G^\top).$$

Applying this property, we have

$$\begin{aligned} & \lambda_{\max}(PF(R + F^\top PF)^{-1}(R + F^\top PF)^{-1}F^\top P) \\ &= \lambda_{\max}(PF(R + F^\top PF)^{-2}F^\top P) \\ &\leq \lambda_{\max}(PF(\lambda_R I + F^\top PF)^{-2}F^\top P) \end{aligned}$$

For any  $H \in \mathbb{R}^{n \times m}$ , we have  $\lambda_{\max}(H^\top H) = \lambda_{\max}(HH^\top)$ . Using this property for  $H = (\lambda_R I + F^\top PF)^{-1} F^\top P$ , we yield

$$\begin{aligned} & \lambda_{\max}(PF(\lambda_R I + F^\top PF)^{-2}F^\top P) \\ &= \lambda_{\max}((\lambda_R I + F^\top PF)^{-1}F^\top P^2F(\lambda_R I + F^\top PF)^{-1}). \end{aligned}$$

Let us consider the singular-value-decomposition (SVD) of  $F$  be  $F = U_F \Sigma_F V_F^\top$ , where  $U_F \in \mathbb{R}^{n \times n}$  and  $V_F \in \mathbb{R}^{m \times m}$  are orthogonal and  $\Sigma_F$  is a rectangular diagonal matrix. Consider  $r$  to be the rank of  $F$ , i.e.,  $\text{rank}(F) = r$ ,  $1 \leq r \leq \min(n, m)$ . The matrix  $\Sigma_F$  (from SVD of  $F$ ) can be written as

$$\Sigma_F = \begin{bmatrix} \Sigma_{Fr} & \mathbf{0}_{r, m-r} \\ \mathbf{0}_{n-r, r} & \mathbf{0}_{n-r, m-r} \end{bmatrix}$$

where  $\Sigma_{Fr} = \text{diag}(\sigma_1(F), \dots, \sigma_r(F))$ . By substituting the SVD of  $F$ , the previous step becomes

$$\begin{aligned} & \lambda_{\max}((\lambda_R I + F^\top PF)^{-1}F^\top P^2F(\lambda_R I + F^\top PF)^{-1}) \\ &= \lambda_{\max}((\lambda_R I + \Sigma_F^\top \bar{P} \Sigma_F)^{-1} \Sigma_F^\top \bar{P}^2 \Sigma_F (\lambda_R I + \Sigma_F^\top \bar{P} \Sigma_F)^{-1}), \end{aligned} \quad (\text{C.5})$$

where  $\bar{P} = U_F^\top P U_F$ . Rewrite  $\bar{P}$  as block matrix form

$$\bar{P} = \begin{bmatrix} [\bar{P}]_{11} & [\bar{P}]_{12} \\ [\bar{P}]_{21} & [\bar{P}]_{22} \end{bmatrix},$$

where  $[\bar{P}]_{11}$  and  $[\bar{P}]_{22}$  are square matrices comprised of the first  $r$  rows and columns and the last  $n - r$  rows and columns of  $\bar{P}$ , respectively. Since  $\Sigma_{Fr}$  is symmetric,  $\Sigma_F^\top \bar{P} \Sigma_F$  can be rewritten in block matrix form as

$$\begin{aligned} \Sigma_F^\top \bar{P} \Sigma_F &= \begin{bmatrix} \Sigma_{Fr} & \mathbf{0}_{r,n-r} \\ \mathbf{0}_{m-r,r} & \mathbf{0}_{n-r,n-r} \end{bmatrix} \begin{bmatrix} [\bar{P}]_{11} & [\bar{P}]_{12} \\ [\bar{P}]_{21} & [\bar{P}]_{22} \end{bmatrix} \begin{bmatrix} \Sigma_{Fr} & \mathbf{0}_{r,m-r} \\ \mathbf{0}_{n-r,r} & \mathbf{0}_{n-r,m-r} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{Fr} [\bar{P}]_{11} \Sigma_{Fr} & \mathbf{0}_{r,m-r} \\ \mathbf{0}_{m-r,r} & \mathbf{0}_{m-r,m-r} \end{bmatrix}. \end{aligned}$$

Similarly,

$$\Sigma_F^\top \bar{P}^2 \Sigma_F = \begin{bmatrix} \Sigma_{Fr} [\bar{P}]_{11}^2 \Sigma_{Fr} & \mathbf{0}_{r,m-r} \\ \mathbf{0}_{m-r,r} & \mathbf{0}_{m-r,m-r} \end{bmatrix}.$$

Substituting the block matrix form of  $\Sigma_F^\top \bar{P} \Sigma_F$  and  $\Sigma_F^\top \bar{P}^2 \Sigma_F$  to (C.5), we have

$$\begin{aligned} &\lambda_{\max}((\lambda_R I_{rr} + \Sigma_F^\top \bar{P} \Sigma_F)^{-1} \Sigma_F^\top \bar{P}^2 \Sigma_F (\lambda_R I_{rr} + \Sigma_F^\top \bar{P} \Sigma_F)^{-1}) \\ &= \lambda_{\max} \left( \begin{bmatrix} \lambda_R I_{rr} + \Sigma_{Fr} [\bar{P}]_{11} \Sigma_{Fr} & \mathbf{0}_{r \times m-r} \\ \mathbf{0}_{m-r \times r} & \sigma_R I_{m-r \times m-r} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{Fr} [\bar{P}^2]_{11} \Sigma_{Fr} & \mathbf{0}_{r \times m-r} \\ \mathbf{0}_{m-r \times r} & \mathbf{0}_{m-r \times m-r} \end{bmatrix} \right. \\ &\quad \left. \begin{bmatrix} \lambda_R I_{rr} + \Sigma_{Fr} [\bar{P}]_{11} \Sigma_{Fr} & \mathbf{0}_{r \times m-r} \\ \mathbf{0}_{m-r \times r} & \lambda_R I_{m-r \times m-r} \end{bmatrix}^{-1} \right) \\ &= \lambda_{\max}((\lambda_R I_{rr} + \Sigma_{Fr} [\bar{P}]_{11} \Sigma_{Fr})^{-1} \Sigma_{Fr} [\bar{P}^2]_{rr} \Sigma_{Fr} (\lambda_R I_{rr} + \Sigma_{Fr} [\bar{P}]_{11} \Sigma_{Fr})^{-1}) \\ &= \lambda_{\max}((\lambda_R \Sigma_{Fr}^{-1} [\bar{P}]_{11}^{-1} + \Sigma_{Fr})^{-1} (\lambda_R [\bar{P}]_{11}^{-1} \Sigma_{Fr}^{-1} + \Sigma_{Fr})^{-1}) \\ &= \lambda_{\max}(\Sigma_{Fr} (\lambda_R [\bar{P}]_{11}^{-1} + \Sigma_{Fr}^2)^{-2} \Sigma_{Fr}). \end{aligned}$$

Since  $[\bar{P}]_{11}$  is positive definite, this implies  $[\bar{P}]_{11}^{-1}$  is also positive definite. Therefore,

$$(\lambda_R [\bar{P}]_{11}^{-1} + \Sigma_{Fr}^2) \succeq \Sigma_{Fr}^2.$$

Taking the inverse of both sides of the above, this implies

$$\begin{aligned} &(\lambda_R [\bar{P}]_{11}^{-1} + \Sigma_{Fr}^2)^{-1} \preceq \Sigma_{Fr}^{-2}, \text{ and} \\ &(\lambda_R [\bar{P}]_{11}^{-1} + \Sigma_{Fr}^2)^{-2} \preceq \Sigma_{Fr}^{-4}. \end{aligned}$$

The last step above is due to the conclusion from [76, Chapter 8.1, Exercise 8.1.12].

Then, we left and right multiply  $\Sigma_{Fr}$  to both sides of the last inequality above, we have

$$\Sigma_{Fr}(\lambda_R[\bar{P}]_{11}^{-1} + \Sigma_{Fr}^2)^{-2}\Sigma_{Fr} \preceq \Sigma_{Fr}\Sigma_{Fr}^{-4}\Sigma_{Fr} = \Sigma_{Fr}^{-2}.$$

Lastly, by definition of singular value, for any real matrix  $H$ , the relationship between the maximum singular value and the maximum eigenvalue is  $\sigma_{max}^2(H) = \lambda_{max}(H^\top H)$ , we have

$$\lambda_{max}(\Sigma_{Fr}^{-2}) = (\sigma_{max}(\Sigma_{Fr}^{-1}))^2 = \frac{1}{(\sigma_{min}(\Sigma_{Fr}))^2} = \frac{1}{(\sigma_{min}^+(F))^2}.$$

Therefore,  $\|K\|^2 \leq \frac{1}{(\sigma_{min}^+(F))^2}$ .  $\square$

The next corollary establishes a matrix upper bound for the control gain  $K_{\tau|t}$  using the lemma above.

**Corollary 4.** *Suppose the cost matrices  $(Q_t)_{t=1}^T$  and  $(R_t^i)_{i=1,t=1}^{N,T-1}$  satisfy the conditions described in (4.8), (4.9), and Assumption 4.2.3. Consider operator  $K$  defined in (C.3). For a given preview horizon  $W \in \mathbb{N}_{T-1}$  and time step  $\tau \in \mathbb{N}_T$ , the control  $\mathbf{u}_{\tau|t}$  defined in (4.21) is given by  $\mathbf{u}_{\tau|t} = K_{\tau|t}x_{\tau|t}$ , where  $K_{\tau|t} = K((Q_{k+1|t}, (R_k^i)_{i=1}^N)_{k=\tau}^{T-1}, A, \mathbf{B})$ , and  $\|K_{\tau|t}\| \leq \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})}$ .*

*Proof.* By using [76, Theorem 4.5.9] (see also the discussion on [76, p. 284]), we have  $\|K_{\tau|t}\| \leq \sigma_{max}(A)\|(\bar{R}_{\tau|t} + \mathbf{B}^\top \bar{P}_{\tau+1|t} \mathbf{B})^{-1} \mathbf{B}^\top \bar{P}_{\tau+1|t}\|$ . By applying Lemma C.1.4, we can conclude that  $\|K_{\tau|t}\| \leq (\sigma_{min}^+(\mathbf{B}))^{-1} \sigma_{max}(A)$ .  $\square$

Intuitively, the maximal energy of the control gain in an LQ potential dynamic game is the energy of the deadbeat controller's gain.

**Lemma C.1.5.** *For any  $\tau, t \in \mathbb{N}_{T-1}$ ,  $\varepsilon_Q I \preceq \bar{Q}_{\tau|t} \preceq Q_{max} + [\frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})}(\lambda_{max}(R_{max}) - \lambda_{min}(R_{min}))]I$ , where  $\bar{Q}_{\tau|t}$  is defined in (C.1).*

*Proof.* Applying the Weyl's inequality to  $\bar{Q}_{\tau|t}$  yields

$$\lambda_{min}(\bar{Q}_{\tau|t}) = \lambda_n(\bar{Q}_{\tau|t}) \geq \lambda_n(Q_{\tau|t}) + \lambda_n(K_{\tau|t}^\top (R_{\tau|t}^1 - \bar{R}_{\tau|t}) K_{\tau|t}).$$

By [76, Theorem 4.5.9], there exists a positive scalar  $\theta_n$  such that  $\lambda_n(K_{\tau|t}^\top (R_{\tau|t}^1 - \bar{R}_{\tau|t}) K_{\tau|t}) = \theta_n \lambda_{min}(R_{\tau|t}^1 - \bar{R}_{\tau|t})$ , where  $\sigma_{min}(K_{\tau|t}) \leq \theta_n \leq \sigma_{max}(K_{\tau|t}) < \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})}$ .

The last inequality is due to Corollary 4. Moreover,

$$\max(0, \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})} \lambda_{max}(\bar{R}_t - R_t^1)) > -\lambda_{min}(K_{\tau|t}^\top (R_t^1 - \bar{R}_t) K_{\tau|t}).$$

Thus, by Assumption 4.2.4, we have

$$\begin{aligned} 0 < \varepsilon_Q &< \lambda_{min}(Q_{\tau|t}) - \max(0, \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})} \lambda_{max}(\bar{R}_t - R_t^1)) \\ &< \lambda_{min}(Q_{\tau|t} + K_t^\top (R_{\tau|t}^1 - \bar{R}_{\tau|t}) K_{\tau|t}) = \lambda_{min}(\bar{Q}_{\tau|t}). \end{aligned}$$

To prove the RHS, note that  $\varepsilon_Q I \preceq \bar{Q}_{\tau|t}$ ,  $Q_{\tau|t} \preceq Q_{max}$  and

$$K_{\tau|t}^\top (R_{\tau|t}^1 - R_{\tau|t}^p) K_{\tau|t} \preceq \left[ \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})} (\lambda_{max}(R_{max}) - \lambda_{min}(R_{min})) \right] I.$$

Combining the above inequalities, the proof is complete.  $\square$

**Definition C.1.1.** For any  $X, Y \in \mathbb{S}_{++}^n$ , define the operator  $\delta_\infty(\cdot, \cdot)$  as the Thompson metric defined in [74, Section 2], where  $\delta_\infty(X, Y) := \|\log(Y^{-\frac{1}{2}}XY^{-\frac{1}{2}})\|_\infty$ , and  $\|\cdot\|_\infty$  denotes the matrix infinity-norm.

The above corollary suggests that, if the costs and system matrices satisfy Assumption 4.2.4, then the matrices  $\bar{Q}_{\tau|t}$  are positive definite. The next lemma establishes the existence of constant matrices  $\bar{P}_{min}$   $\bar{P}_{max}$  independent of  $t$ , that serve as lower and upper bound for  $\bar{P}_t$  for  $t \in \mathbb{N}_T$ .

**Lemma C.1.6.** Consider  $A, \mathbf{B}$  from (4.1),  $\bar{R}_t$  and  $\bar{Q}_t$  from Lemma C.1.3 that satisfy Assumption 4.2.3. For all  $t \in \mathbb{N}_{T-1}$ , let  $\bar{P}_T = \bar{Q}_T$  and

$$\begin{aligned} \bar{K}_t &= -(R_t + B^\top \bar{P}_{t+1} B)^{-1} B^\top \bar{P}_{t+1} A, \\ \bar{P}_t &= \bar{Q}_t + \bar{K}_t^\top \bar{R}_t \bar{K}_t + (A + B \bar{K}_t)^\top \bar{P}_{t+1} (A + B \bar{K}_t), \end{aligned}$$

for  $t \in \mathbb{N}_{T-1}$ . There exist positive definite matrices  $\bar{P}_{min}$  and  $\bar{P}_{max}$  that is independent of  $t$ , such that  $\bar{P}_{min} \preceq \bar{P}_t \preceq \bar{P}_{max}$ .

*Proof.* By Lemma C.1.5, we have  $\bar{Q}_{min} \preceq \bar{Q}_t \preceq \bar{Q}_{max}$ . By Assumption 4.2.3 and Corollary 3, we have  $R_{min} \preceq \bar{R}_t \preceq R_{max}$ . Repeat the procedure that is identical to the proof of [12, Proposition 11] with matrices  $Q_{min}$ ,  $Q_{max}$ ,  $R_{max}$ ,  $Q_t^{max}$ ,  $R_t^{max}$  and  $P_{max}$  replaced by  $\bar{Q}_{min}$ ,  $\bar{Q}_{max}$ ,  $\lambda_{max}(R_{max})I$ ,  $\bar{Q}_{max}$ ,  $\lambda_{max}(R_{max})I$  and  $\bar{P}_{max}$ , respectively, completes the proof.  $\square$

**Lemma C.1.7.** For  $T, S, V_1, V_2 \in \mathbb{S}_{++}^n$ , if  $m \geq 1 + \frac{\lambda_{\max}(V_1 - V_2)}{\lambda_{\min}(T + V_2)}$ , then for any non-zero  $x$ , we have

$$\frac{x^\top(T + V_1)x}{x^\top(S + V_2)x} \leq m \frac{x^\top(T + V_2)x}{x^\top(S + V_2)x}. \quad (\text{C.6})$$

*Proof.* For any nonzero  $x$ , we have

$$m \geq 1 + \frac{\lambda_{\max}(V_1 - V_2)}{\lambda_{\min}(T + V_2)} \geq 1 + \frac{x^\top(V_1 - V_2)x}{x^\top(T + V_2)x} = \frac{x^\top(T + V_1)x}{x^\top(T + V_2)x}.$$

Due to  $x^\top(S + V_2)x > 0$  and the fact that  $T, V_2, V_1$  are positive definite, we have that  $m \frac{x^\top(T + V_2)x}{x^\top(S + V_2)x} \geq \frac{x^\top(T + V_1)x}{x^\top(S + V_2)x}$ .  $\square$

**Proposition 7.** For any  $X, Y \in \mathbb{S}_{++}^n$ ,

$$\delta_\infty(X, Y) = \max\left(\log\left(\sup_{\xi \neq 0} \frac{\xi^\top X \xi}{\xi^\top Y \xi}\right), \log\left(\sup_{\xi \neq 0} \frac{\xi^\top Y \xi}{\xi^\top X \xi}\right)\right).$$

*Proof.* Consider matrices  $X, Y \in \mathbb{S}_{++}^n$ . From [77, Remark 2.2.], we have  $\delta_\infty(X, Y) = \max(\lambda_{\max}(Y^{-1}X), \lambda_{\max}(XY^{-1})) = \max\left(\log\left(\sup_{\xi \neq 0} \frac{\xi^\top X \xi}{\xi^\top Y \xi}\right), \log\left(\sup_{\xi \neq 0} \frac{\xi^\top Y \xi}{\xi^\top X \xi}\right)\right)$ .  $\square$

**Remark C.1.1.** For  $T, S, V_1, V_2 \in \mathbb{S}_{++}^n$ , without the loss of generality, suppose  $\sup_{x \neq 0} \frac{x^\top(T + V_1)x}{x^\top(S + V_2)x} \geq 1$ . Based on Lemma C.1.6, C.1.7 and Proposition 7, consider a positive scalar  $m$  that satisfies  $m \geq 1 + \frac{\lambda_{\max}(V_1 - V_2)}{\lambda_{\min}(T + V_2)}$ . We now investigate  $\delta_\infty(T + V_1, S + V_2)$ . By using Proposition 7, we have

$$\begin{aligned} \delta_\infty(T + V_1, S + V_2) &= \log\left(\sup_{x \neq 0} \frac{x^\top(T + V_1)x}{x^\top(S + V_2)x}\right) \\ &\leq \log(m) + \log\left(\sup_{x \neq 0} \frac{x^\top(T + V_2)x}{x^\top(S + V_2)x}\right) \\ &\leq \log(m) + \delta_\infty(T + V_2, S + V_2) \\ &\leq \log(m) + r\delta_\infty(T, S), \end{aligned}$$

where  $r = \frac{\lambda_{\max}(T)}{\lambda_{\min}(V_2) + \lambda_{\max}(T)}$  and  $0 < r < 1$ .

Before presenting our next lemma that establishes bounds for  $\|\bar{P}_{\tau|t} - \bar{P}_{\tau|t_0}\|$  and

$\|\bar{K}_{\tau|t} - \bar{K}_{\tau|t_0}\|$  for  $1 \leq \tau \leq t \leq t_0 \leq T$ , we introduce the following constants:

$$\alpha := \lambda_{\max}(A^\top(\bar{P}_{\max}^{-1} + BR_{\max}^{-1}B^\top)^{-1}A), \quad (\text{C.7})$$

$$\gamma := \frac{\alpha}{\alpha + \varepsilon_Q}, \quad (\text{C.8})$$

$$h := \log\left(\frac{\lambda_{\max}(\bar{P}_{\max})}{\varepsilon_Q}\right), \quad (\text{C.9})$$

$$C_P = \frac{\lambda_{\max}^2(\bar{P}_{\max})}{\varepsilon_Q}, \quad (\text{C.10})$$

$$\begin{aligned} \bar{\omega} = \lambda_{\max}(A^\top(\bar{P}_{\max}^{-1} + BR_{\max}^{-1}B^\top)^{-1}) \\ - \lambda_{\min}(A^\top(\varepsilon_Q^{-1}I + BR_{\min}^{-1}B^\top)^{-1}), \end{aligned} \quad (\text{C.11})$$

$$\varepsilon_1 := \log\left(1 + \frac{\bar{\omega}}{\varepsilon_Q}\right), \quad (\text{C.12})$$

$$\varepsilon_P = \frac{\varepsilon_1 \lambda_{\max}(\bar{P}_{\max})(\exp(h) - 1)}{h(1 - \gamma)}, \quad (\text{C.13})$$

$$G_{\max} := \|(R_{\min} + \varepsilon_Q \mathbf{B}^\top \mathbf{B})^{-1}\|, \quad (\text{C.14})$$

$$C'_K = G_{\max}^2 \|R_{\max} \mathbf{B}^\top\| C_P, \quad (\text{C.15})$$

$$\varepsilon'_K = G_{\max}^2 \|R_{\max} \mathbf{B}^\top\| \varepsilon_P. \quad (\text{C.16})$$

The next lemma establishes the contraction of  $\|K_{\tau|t} - K_{\tau|t_0}\|$  with respect to  $\tau$ .

**Lemma C.1.8.** *For  $1 \leq \tau \leq t \leq t_0 \leq T$ , suppose  $W \in \mathbb{N}_{T-1}$  is the preview window length. Consider scalars  $\varepsilon_P, \varepsilon_K, C_P, C'_K$  and  $\gamma \in (0, 1)$  defined in (C.8), (C.10), (C.13), (C.15), and (C.16), respectively. The distance between  $\bar{P}_{\tau|t}$  and  $\bar{P}_{\tau|t_0}$  satisfies  $\|\bar{P}_{\tau|t} - \bar{P}_{\tau|t_0}\| \leq C_P \gamma^{t-\tau+W} + \varepsilon_P$ , and the distance between  $\bar{K}_{\tau|t}$  and  $\bar{K}_{\tau|t_0}$  satisfies  $\|\bar{K}_{\tau|t} - \bar{K}_{\tau|t_0}\| \leq C'_K \gamma^{t-\tau+1+W} + \varepsilon'_K$ .*

*Proof.* For  $1 \leq \tau \leq t \leq t_0 \leq T$ , by [75, Lemma D.2], Lemmas C.1.6 & C.1.7, and Remark C.1.1, we have

$$\begin{aligned} \delta_\infty(\bar{P}_{\tau|t}, \bar{P}_{\tau|t_0}) &= \delta_\infty(\bar{Q}_{\tau|t} + A^\top(\bar{P}_{\tau+1|t}^{-1} + B\bar{R}_{\tau|t}^{-1}B^\top)^{-1}A, \\ &\quad \bar{Q}_{\tau|t_0} + A^\top(\bar{P}_{\tau+1|t_0}^{-1} + B\bar{R}_{\tau|t_0}^{-1}B^\top)^{-1}A) \\ &\leq \gamma \delta_\infty(\bar{P}_{\tau+1|t}, \bar{P}_{\tau+1|t_0}) + \varepsilon_1 \\ &\leq \gamma^{t-\tau+W} \delta_\infty(\bar{P}_{t+W|t}, \bar{P}_{t+W|t_0}) + \varepsilon_1 \sum_{p=0}^{t-\tau+W} \gamma^p \\ &\leq C_{P_1} \gamma^{t-\tau+W} + \frac{\varepsilon_1}{1 - \gamma}, \end{aligned}$$

where  $\varepsilon_1$  is defined in (C.12). By monotonicity of function  $\frac{e^x - 1}{x}$  for  $x > 0$  and

$h > 0$ , we have  $\frac{\exp(\delta_\infty(\bar{P}_{\tau|t}, \bar{P}_{\tau|t_0}))^{-1}}{\delta_\infty(\bar{P}_{\tau|t}, \bar{P}_{\tau|t_0})} \leq \frac{\exp(h)-1}{h}$ . Thus,  $\|\bar{P}_{\tau|t} - \bar{P}_{\tau|t_0}\| < C_P \gamma^W + \varepsilon_P$ . By repeating procedures as the proof in [67, Lemma 8, (20) and (21)] by replacing  $B, R_\tau, P_{\tau+1|t}$  and  $P_{\tau+1|t_0}$  from [67, Lemma 8, (20) and (21)] to  $\mathbf{B}, \bar{R}_\tau, \bar{P}_{\tau+1|t}$  and  $\bar{P}_{\tau+1|t_0}$ , we have  $\|\bar{K}_{\tau|t} - \bar{K}_{\tau|t_0}\| < C'_K \gamma^{t-\tau-1+W} + \varepsilon'_K$ .  $\square$

**Remark C.1.2.** When  $\tau = t$  and  $t_0 = T$ , we have  $\|\bar{K}_{t|t} - \bar{K}_t^*\| \leq C'_K \gamma^{W-1} + \varepsilon'_K$ .

**Lemma C.1.9.** For integers  $1 \leq \tau \leq t_0 \leq t_1 \leq t \leq T-1$ , consider  $A, \mathbf{B}$  from (4.1) and  $\bar{K}_{\tau|t} = \mathbf{K}((Q_{k+1|t}, (R_{k|t}^i)_{i=1}^N)_{k=\tau}^{T-1}, A, \mathbf{B})$  where operator  $\mathbf{K}$  is defined in (C.3). Consider scalars  $C_{fb} = \frac{\lambda_{\max}(\bar{P}_{\max})}{\lambda_{\min}(\bar{Q}_{\min})}$  and  $\eta = \sqrt{1 - \frac{\lambda_{\min}(\bar{Q}_{\min})}{\lambda_{\max}(\bar{P}_{\max})}}$ , matrices  $\bar{K}_{\tau|t}$  satisfies  $\|\prod_{\tau=t_0}^{t_1} (A + \mathbf{B}\bar{K}_{\tau|t})\| \leq C_{fb} \eta^{t_1-t_0+1}$ .

The proof mirrors that of [12, Appendix E, Proposition 2].

**Lemma C.1.10.** For any  $T \geq 1$ ,  $W \in \mathbb{N}_{T-1} \cup \{0\}$  and  $t \in \mathbb{N}_T$ , consider state  $x_t$  that generated by control policy (4.22) and the state  $x_{t|t}$  as an element of the solution from (4.21). Let  $\varepsilon_K := \|\mathbf{B}\| \varepsilon'_K$  and  $C_K = \|\mathbf{B}\| C'_K$ , where  $C'_K$  and  $\varepsilon'_K$  are as in Lemma C.1.8. Then, the distance between state  $x_t$  and  $x_{t|t}$  satisfies

$$\|x_t - x_{t|t}\| \leq C_{fb}^2 C_q \|\bar{x}_1\| q^t \left[ \frac{C_K \gamma^W}{\gamma - 1} \left( \frac{1 - (\frac{\eta\gamma}{q})^t}{1 - \frac{\eta\gamma}{q}} - \frac{1 - (\frac{\eta}{q})^t}{1 - \frac{\eta}{q}} \right) + \varepsilon_K \left( \frac{(t-1)(\frac{\eta}{q})^{t+1} - t(\frac{\eta}{q})^t + \frac{\eta}{q}}{(1 - \frac{\eta}{q})^2} \right) \right].$$

where  $\gamma$  is defined in Lemma C.1.8,  $\eta$  is defined in Lemma C.1.9, together with  $q := \rho(A + \mathbf{B}\bar{K})$  and  $C_q := \sup_{n \geq 0} \frac{\|(A + \mathbf{B}\bar{K})\|^n}{q^n}$ .

*Proof.* Suppose  $M$  is a positive integer. For an arbitrary matrix sequence  $(a_i)_{i=1}^M$ . For  $1 \leq p_1 \leq M$  and  $1 \leq p_2 \leq M$ , define the product operator as

$$\prod_{j=p_1}^{p_2} a_j := \begin{cases} a_{p_2} a_{p_2-1} \cdots a_{p_1} & \text{if } p_1 < p_2 \\ a_{p_2} & \text{if } p_1 = p_2 \\ I & \text{if } p_1 > p_2, \end{cases}$$

Define  $\omega_t := x_t - x_{t|t}$ ,  $\theta_{\tau|p_1, p_2} := x_{\tau|p_1} - x_{\tau|p_2}$ , where  $\tau \leq p_1 \leq p_2 \leq T$ . Consequently,

$\theta_{1|p_1, p_2} = 0$ , and

$$\begin{aligned}
\theta_{\tau|p_1, p_2} &= x_{\tau|p_1} - x_{\tau|p_2} \\
&= (A + \sum_j^N B^j \bar{K}_{j, \tau|p_1}) x_{\tau-1|p_1} - (A + \sum_j^N B^j \bar{K}_{j, \tau|p_2}) x_{\tau-1|p_2} \\
&= (A + \sum_j^N B^j \bar{K}_{j, \tau|p_1}) \theta_{\tau-1|p_1, p_2} \\
&\quad + [\sum_{j=1}^N B^j (\bar{K}_{j, \tau-1|p_1} - \bar{K}_{j, \tau-1|p_2})] x_{\tau-1|p_2} \\
&\quad \vdots \\
&= \sum_{i=1}^{\tau-1} \left( \prod_{j=i+1}^{\tau-1} (A + \sum_{m=1}^N B^m \bar{K}_{m, j|p_1}) \right) \\
&\quad \left[ \sum_{m=1}^N B^m (\bar{K}_{m, i|p_1} - \bar{K}_{m, i|p_2}) \right] x_{i|p_2} \\
&= \sum_{i=1}^{\tau-1} \left( \prod_{j=i+1}^{\tau-1} (A + \sum_{m=1}^N B^m \bar{K}_{m, j|p_1}) \right) \left[ \sum_{m=1}^N B^m (\bar{K}_{m, i|p_1} - \bar{K}_{m, i|p_2}) \right] \\
&\quad \left( \prod_{n=1}^{i-1} (A + \sum_{m=1}^N B^m \bar{K}_{m|p_2}) \right) \bar{x}_1 \\
&= \sum_{i=1}^{\tau-1} \left( \prod_{j=i+1}^{\tau-1} (A + \mathbf{B} \bar{K}_{j|p_1}) \right) \left[ \mathbf{B}^\top (\bar{K}_{j|p_1}^\top - \bar{K}_{j|p_2}^\top) \right] \\
&\quad \left( \prod_{n=1}^{i-1} (A + \mathbf{B} \bar{K}_{n|p_2}) \right) \bar{x}_1.
\end{aligned}$$

Moreover,  $\omega_1 = 0$ , and for  $t \in \mathbb{N}_T$ , we have

$$\begin{aligned}
\omega_t &= (A + \sum_{j=1}^N B^j \underline{K}^j) (x_{t-1} - x_{t-1|t-1}) + x_{t|t-1} - x_{t|t} \\
&= \sum_{i=1}^t (A + \sum_{j=1}^N B^j \underline{K}^j)^{t-i} \theta_{i|i-1, i}.
\end{aligned}$$

We now investigate the dynamics of  $\theta_{\tau|p_1, p_2}$ . Note that  $\theta_{0|p_1, p_2} = 0$ , and

$$\begin{aligned}
\theta_{\tau+1|p_1, p_2} &= x_{\tau+1|p_1} - x_{\tau+1|p_2} \\
&= (A + \mathbf{B} \bar{K}_{\tau|p_1}) x_{\tau|p_1} - (A + \mathbf{B} \bar{K}_{\tau|p_2}) x_{\tau|p_2} \\
&= (A + \mathbf{B} \bar{K}_{\tau|p_1}) (\theta_{\tau|p_1, p_2} + x_{\tau|p_2}) \\
&\quad - (A + \mathbf{B} \bar{K}_{\tau|p_2}) x_{\tau|p_2} \\
&= (A + \mathbf{B} \bar{K}_{\tau|p_1}) \theta_{\tau|p_1, p_2} + \mathbf{B} (\bar{K}_{\tau|p_1} - \bar{K}_{\tau|p_2}) x_{\tau|p_2}.
\end{aligned}$$

This implies that

$$x_{\tau+1|p_1} - x_{\tau+1|p_2} = \sum_{n=1}^{\tau} \left( \prod_{m=n+1}^{\tau} (A + \mathbf{B}\bar{K}_{m|p_1}) \right) \mathbf{B}(\bar{K}_{n|p_1} - \bar{K}_{n|p_2}) \left( \prod_{m=1}^{n-1} (A + \mathbf{B}\bar{K}_{m|p_1}) \right) \bar{x}_1.$$

By Lemma C.1.9, we have  $\|\prod_{m=n+1}^{\tau} (A + \mathbf{B}\bar{K}_{m|p_1})\| \leq C_{fb}\eta^{\tau-n}$ . By Lemma C.1.8, we have  $\|\mathbf{B}(\bar{K}_{n|p_1} - \bar{K}_{n|p_2})\| \leq C_K\gamma^{p-n+W} + \varepsilon_K$ . Thus,

$$\begin{aligned} \|\theta_{\tau+1|p_1,p_2}\| &= \|x_{\tau+1|p_1} - x_{\tau+1|p_2}\| \\ &\leq C_{fb}^2 \|\bar{x}_1\| \left( \sum_{n=1}^{\tau} \eta^{\tau-n} (C_K\gamma^{p-n+W} + \varepsilon_K) \right) \\ &= C_{fb}^2 \|\bar{x}_1\| \left[ \frac{C_K\gamma^{p+W}\eta^{\tau}}{1 - \frac{1}{\gamma}} \left(1 - \left(\frac{1}{\gamma}\right)^{\tau+1}\right) + \varepsilon_K\tau\eta^{\tau} \right]. \end{aligned}$$

Choosing  $\tau = t, p_1 = t$  and  $p_2 = T$ , results in

$$\begin{aligned} \|\theta_{t|t,T}\| &\leq C_{fb}^2 \|\bar{x}_1\| \left[ \frac{C_K\gamma^{t+W}\eta^t}{1 - \frac{1}{\gamma}} \left(1 - \left(\frac{1}{\gamma}\right)^{t+1}\right) + \varepsilon_K t \eta^t \right] \\ &= C_{fb}^2 \|\bar{x}_1\| \left[ \frac{C_K\gamma^{1+W}\eta^t}{\gamma - 1} (\gamma^t - 1) + \varepsilon_K t \eta^t \right]. \end{aligned}$$

Moreover,  $\|\theta_{i|i-1,i}\| \leq C_{fb}^2 \|\bar{x}_1\| \left[ \frac{C_K\eta^{i-1}\gamma^W}{\gamma-1} (\gamma^i - 1) + \varepsilon_K (i-1)\eta^{i-1} \right]$ . Similar to the argument following [50, Lemma 10, (25)], we have that  $\|(A + \mathbf{B}\underline{K})^{t-i}\| \leq C_q q^{t-i}$ .

Conclude the above, we have

$$\begin{aligned} &\|x_t - x_{t|t}\| \\ &\leq \sum_{i=1}^t \|(A + \mathbf{B}\underline{K})^{t-i} \theta_{i|i-1,i}\| \\ &\leq C_{fb}^2 C_q \|\bar{x}_1\| \sum_{i=1}^t q^{t-i} \left[ \frac{C_K\eta^{i-1}\gamma^W}{\gamma-1} (\gamma^i - 1) + \varepsilon_K (i-1)\eta^{i-1} \right] \\ &= C_{fb}^2 C_q \|\bar{x}_1\| q^t \left[ \frac{C_K\gamma^W}{\gamma-1} \left( \frac{1 - \left(\frac{\eta\gamma}{q}\right)^t}{1 - \frac{\eta\gamma}{q}} - \frac{1 - \left(\frac{\eta}{q}\right)^t}{1 - \frac{\eta}{q}} \right) + \varepsilon_K \left( \frac{(t-1)\left(\frac{\eta}{q}\right)^{t+1} - t\left(\frac{\eta}{q}\right)^t + \frac{\eta}{q}}{\left(1 - \frac{\eta}{q}\right)^2} \right) \right]. \end{aligned}$$

□

Before presenting the Cost Difference Lemma that is essential for the proof of Theorem 4.3.1, we introduce the following definitions.

Consider any policies  $(\pi_t^i)_{i=1,t=1}^{N,T-1}$  and  $(\tilde{\pi}_t^i)_{i=1,t=1}^{N,T-1}$  where  $\pi_t^i, \tilde{\pi}_t^i \in \Lambda$ . We state the

convention  $\bar{\Pi}_t^{t_0} := (\Pi_\tau)_{\tau=t}^{t_0}$  for  $t, t_0 \in \mathbb{N}_{T-1}$  and  $t \leq t_0$ . We again define

$$\begin{aligned} g_t^i(x_t, \mathbf{u}_t) &:= x_t^\top Q_t x_t + \mathbf{u}_t^\top R_t^i \mathbf{u}_t, \\ V_{i,T,t}^{\bar{\Pi}_t^{T-1}}(x_t) &:= \begin{cases} \sum_{l=0}^{T-1-t} g_{t+l}^i(x_{t+1+l}^{\bar{\Pi}_{t+l}^{t+1}}, \Pi_{t+l}(x_{t+l})) & 1 \leq t < T-1 \\ 0 & t \geq T-1, \end{cases} \\ Q_{i,T,t}^{\bar{\Pi}_t^{T-1}}(x_t, \mathbf{u}_t) &:= g_t^i(x_t, \mathbf{u}_t) + V_{i,T,t+1}^{\bar{\Pi}_{t+1}^{T-1}}(x_{t+1}^{\bar{\Pi}_{t+1}^{t+1}}). \end{aligned}$$

where  $u_t^i = \pi_t^i(x_t)$  and  $x_{t+1+l}^{\bar{\Pi}_{t+l}^{t+1}}(x_{t+l}) = Ax_{t+l} + \mathbf{B}\Pi_{t+l}(x_{t+l})$  for  $i \in \{1, 2, \dots, N\}$ . Now we present our Cost Difference Lemma.

**Lemma C.1.11** (Cost Difference Lemma). *Given a positive integer  $T \geq 1$ , for  $t \in \mathbb{N}_{T-1}$  and  $i = \{1, 2, \dots, N\}$ , consider policies  $(\pi_t^i)_{t=1}^{T-1}, (\tilde{\pi}_t^i)_{t=1}^{T-1}$  such that  $\pi_t^i, \tilde{\pi}_t^i \in \Lambda$ . Let  $(\Pi_t)_{t=1}^{T-1} = (\pi_t^i)_{i=1, t=1}^{N, T-1}$  and  $(\tilde{\Pi}_t)_{t=1}^{T-1} = (\tilde{\pi}_t^i)_{i=1, t=1}^{N, T-1}$ . Then, we have*

$$J_{i,T}(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) - J_{i,T}(\bar{x}_1, (\tilde{\Pi}_t)_{t=1}^{T-1}) = \sum_{t=1}^{T-1} Q_{i,T,t}^{\tilde{\Pi}_t^{T-1}}(x_t, \mathbf{u}_t) - V_{i,T,t}^{\tilde{\Pi}_t^{T-1}}(x_t), \quad (\text{C.17})$$

where  $x_t$  and  $(u_t^i)_{i=1}^N$  satisfy (4.1).

*Proof.* Starting from the RHS of (C.17) we have:

$$\begin{aligned} & \sum_{t=1}^{T-1} Q_{i,T,t}^{\tilde{\Pi}_t^{T-1}}(x_t, \mathbf{u}_t) - V_{i,T,t}^{\tilde{\Pi}_t^{T-1}}(x_t) \\ &= \sum_{t=1}^{T-1} g_t(x_t, \mathbf{u}_t) + V_{i,T,t+1}^{\tilde{\Pi}_{t+1}^{T-1}}(x_{t+1}) - V_{i,T,t}^{\tilde{\Pi}_t^{T-1}}(x_t) \\ &= \underbrace{\sum_{t=1}^{T-1} g_t(x_t, \mathbf{u}_t)}_{J_{i,T}(\bar{x}_1, (\Pi_t)_{t=1}^{T-1})} - V_{i,T,1}^{\tilde{\Pi}_1^{T-1}}(x_1) \\ &= J_{i,T}(\bar{x}_1, (\Pi_t)_{t=1}^{T-1}) - J_{i,T}(\bar{x}_1, (\tilde{\Pi}_t)_{t=1}^{T-1}). \end{aligned}$$

□

**Proposition 8.** *For  $i \in \{1, 2, \dots, N\}$  and  $t \in \mathbb{N}_{T-1}$ , consider*

$$\begin{aligned} \Delta' &= \|\mathbf{B}\| \|\bar{P}_{max}\|, \Delta_1 = \|\bar{R}_{max}\| + \|\mathbf{B}\|^2 \|\bar{P}_{max}\|, \\ \Delta_2 &= \frac{\sigma_{max}(A)\Delta_1}{\sigma_{min}^+(\mathbf{B})} + \|A\| \|\mathbf{B}\| \|\bar{P}_{max}\|. \end{aligned}$$

We have the following inequalities

$$\|R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}\| \leq \Delta_1, \quad (\text{C.18})$$

$$\|(R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B})\bar{K}_t^* + \mathbf{B}^\top P_{t+1}^i A\| \leq \Delta_2. \quad (\text{C.19})$$

*Proof.* If costs  $(R_t^i)_{i=1,t=1}^{N,T-1}$  and  $\{Q_t\}_{t=1}^T$  from LQ-DFG is an LQ-DFPG. By Lemma C.1.2, we have  $\|\mathbf{B}^\top P_{t+1}^i\| = \|\mathbf{B}^\top \bar{P}_{t+1}\| \leq \Delta'$ , where  $\bar{P}_{t+1}$  is defined in Lemma C.1.2. Due to Assumption 4.2.3, matrix  $\|R_t^i\|$  is upper bounded uniformly w.r.t.  $t$  and  $i$ . Applying Corollary 4 and triangle inequality, we yield the inequalities (C.18) and (C.19).  $\square$

## C.2 Proof of Theorem 4.3.1

Consider  $(\pi_{i,t})_{i=1,t=1}^{N,T-1}$  as the control policy defined in (4.22), and  $(\tilde{\pi}_{i,t})_{i=1,t=1}^{N,T-1}$  as the control policy that compute the feedback Nash equilibrium defined in (4.5). By applying the Cost Difference Lemma (Lemma C.1.11), we have

$$\begin{aligned} & \text{PoU}_T((\mathbf{u}_t)_{t=1}^{T-1}) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} Q_{i,T,t}^{\tilde{\pi}_t^{T-1}}(x_t, \mathbf{u}_t) - V_{i,T,t}^{\tilde{\pi}_t^{T-1}}(x_t) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} x_t^\top Q_t x_t + \mathbf{u}_t^\top R_t^i \mathbf{u}_t + (Ax_t + \mathbf{B}\mathbf{u}_t)^\top P_{t+1}^i (Ax_t + \mathbf{B}\mathbf{u}_t) \\ &\quad - x_t^\top Q_t x_t - \tilde{\mathbf{u}}_t^\top R_t^i \tilde{\mathbf{u}}_t - (Ax_t + \mathbf{B}\tilde{\mathbf{u}}_t)^\top P_{t+1}^i (Ax_t + \mathbf{B}\tilde{\mathbf{u}}_t) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} \mathbf{u}_t^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \mathbf{u}_t - \tilde{\mathbf{u}}_t^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \tilde{\mathbf{u}}_t + 2x_t^\top A^\top P_{t+1}^i \mathbf{B} (\mathbf{u}_t - \tilde{\mathbf{u}}_t) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) (\mathbf{u}_t - \tilde{\mathbf{u}}_t) \\ &\quad + 2(\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \tilde{\mathbf{u}}_t + \mathbf{B}^\top P_{t+1}^i A x_t \right) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) (\mathbf{u}_t - \tilde{\mathbf{u}}_t) + 2(\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t^* + \mathbf{B}^\top P_{t+1}^i A \right) x_t. \end{aligned}$$

Since

$$\begin{aligned} \|\mathbf{u}_t - \tilde{\mathbf{u}}_t\| &= \|\underline{K}x_t + (\bar{K}_{t|t} - \underline{K})x_{t|t} - \bar{K}_t^* x_t\| \\ &= \|(\bar{K}_t^* - \underline{K})(x_{t|t} - x_t) + (\bar{K}_{t|t} - \bar{K}_t^*)x_{t|t}\| \\ &\leq \|\bar{K}_t^* - \underline{K}\| \|x_{t|t} - x_t\| + \|\bar{K}_{t|t} - \bar{K}_t^*\| \|x_{t|t}\|, \end{aligned}$$

and from the fact that for any  $a_1, a_2 \in \mathbb{R}$ ,  $(a_1 + a_2)^2 \leq 2(a_1^2 + a_2^2)$ , we have

$$\begin{aligned} \|\mathbf{u}_t - \tilde{\mathbf{u}}_t\|^2 &= \|\underline{K}x_t + (\bar{K}_{t|t} - \underline{K})x_{t|t} - \bar{K}_t^*x_t\|^2 \\ &\leq 2(\|\bar{K}_t^* - \underline{K}\|^2\|x_{t|t} - x_t\|^2 + \|\bar{K}_{t|t} - \bar{K}_t^*\|^2\|x_{t|t}\|^2), \end{aligned}$$

By Proposition 8, there exist  $\Delta_1 > 0$  and  $\Delta_2 > 0$ , such that

$$\begin{aligned} &\text{PoU}((\mathbf{u}_t)_{t=1}^{T-1}) \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) (\mathbf{u}_t - \tilde{\mathbf{u}}_t) \\ &\quad + 2(\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t^* + \mathbf{B}^\top P_{t+1}^i A \right) x_t \\ &\leq \sum_{t=1}^{T-1} (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left[ \sum_{i=1}^N \frac{1}{N} (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \right] (\mathbf{u}_t - \tilde{\mathbf{u}}_t) \\ &\quad + (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \sum_{i=1}^N \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t^* + \mathbf{B}^\top P_{t+1}^i A \right) x_t \\ &\leq \Delta_1 \sum_{t=1}^{T-1} \|\mathbf{u}_t - \tilde{\mathbf{u}}_t\|^2 + \Delta_2 \sum_{t=1}^{T-1} \|\mathbf{u}_t - \tilde{\mathbf{u}}_t\| (\|x_t - x_{t|t}\| + \|x_{t|t}\|) \\ &\leq 2\Delta_1 \sum_{t=1}^{T-1} \left( \|\bar{K}_t^* - \underline{K}\|^2 \|x_{t|t} - x_t\|^2 + \|\bar{K}_{t|t} - \bar{K}_t^*\|^2 \|x_{t|t}\|^2 \right) \\ &\quad + \Delta_2 \sum_{t=1}^{T-1} \left( \|\bar{K}_t^* - \underline{K}\| \|x_{t|t} - x_t\| + \|\bar{K}_{t|t} - \bar{K}_t^*\| \|x_{t|t}\| \right) \\ &\quad \left( \|x_t - x_{t|t}\| + \|x_{t|t}\| \right), \tag{C.20} \end{aligned}$$

By Lemma C.1.10, we have

$$\begin{aligned} \|x_t - x_{t|t}\| &\leq C_{fb}^2 \|\bar{x}_1\| q^t \left[ \frac{C_K \gamma^W}{\gamma - 1} \left( \frac{1 - (\frac{\eta}{q})^t}{1 - \frac{\eta}{q}} - \frac{1 - (\frac{\eta}{q})^t}{1 - \frac{\eta}{q}} \right) \right. \\ &\quad \left. + \varepsilon_K \left( \frac{(t-1)(\frac{\eta}{q})^{t+1} - t(\frac{\eta}{q})^t + \frac{\eta}{q}}{(1 - \frac{\eta}{q})^2} \right) \right]. \end{aligned}$$

To simplify the presentation of the main result, define

$$\begin{aligned}
C_* &:= \frac{\sigma_{max}(A)}{\sigma_{min}^+(\mathbf{B})} + \|\underline{K}\|, \\
D_K(\gamma^W, \varepsilon_K) &:= \frac{C_K \gamma^W q \eta (\gamma - 1)}{(q - \eta \gamma)(q - \eta)} + \frac{\varepsilon_K q \eta}{(q - \eta)^2}, \\
C_x &:= C_{fb}^2 D_K(\gamma^W, \varepsilon_K), \\
\Delta_a(\gamma^W, \varepsilon_K) &:= 2(\Delta_1 + \Delta_2) C_*^2 D_K(\gamma^W, \varepsilon_K) C_x^2, \\
\Delta_b(\gamma^W, \varepsilon_K) &:= 4\Delta_1 C_{fb}^2 (C_K'^2 \gamma^{2W} + \frac{\varepsilon_K^2}{\|\mathbf{B}\|^2}) + 2\Delta_2 (C_K' \gamma^W + \frac{\varepsilon_K}{\|\mathbf{B}\|}), \\
\Delta_c(\gamma^W, \varepsilon_K) &:= 2C_x C_{fb} \Delta_2 (C_K' \gamma^W + \frac{\varepsilon_K}{\|\mathbf{B}\|} + C_* D_K(\gamma^W, \varepsilon_K)),
\end{aligned}$$

where constants  $C_K'$ ,  $\varepsilon_K$ , and  $C_{fb}$  are from Lemmas C.1.8 to C.1.10. The distance between  $x_t$  and  $x_{t|t}$  can be simplified as  $\|x_t - x_{t|t}\| \leq C_x \|\bar{x}_1\| q^t$ . Continue calculations from (C.20), we have

$$\begin{aligned}
&\text{PoU}_T((\mathbf{u}_t)_{t=1}^{T-1}) \\
&< \|\bar{x}_1\|^2 \left[ \Delta_a(\gamma^W, \varepsilon_{K'}) \frac{1 - q^{2T}}{1 - q^2} + \Delta_b(\gamma^W, \varepsilon_{K'}) \frac{1 - \eta^{2T}}{1 - \eta^2} + \Delta_c(\gamma^W, \varepsilon_{K'}) \frac{1 - (q\eta)^T}{1 - q\eta} \right].
\end{aligned}$$

To simplify (C.20), let

$$\begin{aligned}
D_K &:= \frac{C_K q \eta (\gamma - 1)}{(q - \eta \gamma)(q - \eta)} + \frac{\varepsilon_K q \eta}{(q - \eta)^2}, \\
C_x &:= C_{fb}^2 D_K, \\
\Delta_a(z, y) &:= 2(\Delta_1 + \Delta_2) D_K(z, y) C_x^2, \\
\Delta_b(z, y) &:= 4\Delta_1 C_{fb}^2 (C_K^2 z^2 + y^2) + 2\Delta_2 (C_K z + y), \\
\Delta_c(z, y) &:= 2C_x C_{fb} \Delta_2 (C_K z + y + D_K),
\end{aligned}$$

and

$$\begin{aligned}
\Gamma_1(t) &:= C_*^2 D_K(\gamma^W, \varepsilon_K) C_x^2 q^{2t} + 2(C_K'^2 \gamma^{2W} + \varepsilon_K'^2) C_{fb}^2 \eta^{2t}, \\
\Gamma_2(t) &:= C_* D_K(\gamma^W, \varepsilon_K) C_x q^t + (C_K' \gamma^W + \varepsilon_K') C_{fb} \eta^t, \\
\Gamma_3(t) &:= C_x q^t + C_{fb} \eta^t.
\end{aligned}$$

By Lemmas C.1.8 and C.1.9, we have

$$\begin{aligned} \|\bar{K}_t^* - \bar{K}\|^2 \|x_{t|t} - x_t\|^2 + \|\bar{K}_{t|t} - \bar{K}_t^*\|^2 \|x_{t|t}\|^2 &\leq \|\bar{x}_1\|^2 \Gamma_1(t) \\ \|\bar{K}_t^* - \underline{K}\| \|x_{t|t} - x_t\| + \|\bar{K}_{t|t} - \bar{K}_t^*\| \|x_{t|t}\| &\leq \|\bar{x}_1\| \Gamma_2(t) \\ \|x_t - x_{t|t}\| + \|x_{t|t}\| &\leq \|\bar{x}_1\| \Gamma_3(t). \end{aligned}$$

Moreover, using the geometric series sum, observe

$$\sum_{t=1}^{T-1} \Gamma_1(t) \leq \bar{\Gamma}_1, \quad \sum_{t=1}^{T-1} \Gamma_2(t) \Gamma_3(t) \leq \bar{\Gamma}_2,$$

where

$$\begin{aligned} \bar{\Gamma}_1 &= C_*^2 D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + 2C_{fb}^2 (C_K'^2 \gamma^{2W} + \varepsilon_K'^2) \frac{1 - \eta^{2T}}{1 - \eta^2}, \\ \bar{\Gamma}_2 &= C_* D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + C_{fb}^2 (C_K' \gamma^W + \varepsilon_K') \frac{1 - \eta^{2T}}{1 - \eta^2} \\ &\quad + [D_K(\gamma^W, \varepsilon_K) C_x C_{fb} + C_x C_{fb} (C_K' \gamma^W + \varepsilon_K')] \frac{1 - (q\eta)^T}{1 - q\eta}. \end{aligned}$$

Note that the right hand side of (C.20) then can be bounded by  $\bar{\Gamma}$  where

$$\begin{aligned} \bar{\Gamma} &= \|\bar{x}_1\|^2 (2\Delta_1 \bar{\Gamma}_1 + \Delta_2 \bar{\Gamma}_2) \\ &= \|\bar{x}_1\|^2 \left[ \Delta_a(\gamma^W, \varepsilon_K') \frac{1 - q^{2T}}{1 - q^2} + \Delta_b(\gamma^W, \varepsilon_K') \frac{1 - \eta^{2T}}{1 - \eta^2} + \Delta_c(\gamma^W, \varepsilon_K') \frac{1 - (q\eta)^T}{1 - q\eta} \right] \end{aligned}$$

upper bounds the  $PoU$  by

$$\begin{aligned} &2\Delta_1 \|\bar{x}_1\|^2 \sum_{t=1}^{T-1} \left( C_*^2 D_K(\gamma^W, \varepsilon_K) C_x^2 q^{2t} + 2(C_K'^2 \gamma^{2W} + \varepsilon_K'^2) C_{fb}^2 \eta^{2t} \right) \\ &\leq 2\Delta_1 \|\bar{x}_1\|^2 \left( C_*^2 D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + 2C_{fb}^2 (C_K'^2 \gamma^{2W} + \varepsilon_K'^2) \frac{1 - \eta^{2T}}{1 - \eta^2} \right), \end{aligned}$$

and

$$\begin{aligned}
& \sum_{t=1}^{T-1} \left( \|\bar{K}_t^* - \underline{K}\| \|x_{t|t} - x_t\| + \|\bar{K}_{t|t} - \bar{K}_t^*\| \|x_{t|t}\| \right) \left( \|x_t - x_{t|t}\| + \|x_{t|t}\| \right) \\
& \leq \|\bar{x}_1\|^2 \sum_{t=1}^{T-1} \left( C_* D_K(\gamma^W, \varepsilon_K) C_x q^t + (C'_K \gamma^W + \varepsilon'_K) C_{fb} \eta^t \right) (C_x q^t + C_{fb} \eta^t) \\
& \leq \|\bar{x}_1\|^2 \left( C_* D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + C_{fb}^2 (C'_K \gamma^W + \varepsilon'_K) \frac{1 - \eta^{2T}}{1 - \eta^2} \right. \\
& \quad \left. + [D_K(\gamma^W, \varepsilon_K) C_x C_{fb} + C_x C_{fb} (C'_K \gamma^W + \varepsilon'_K)] \frac{1 - (q\eta)^T}{1 - q\eta} \right).
\end{aligned}$$

Conclude the above, the PoU can be upper bounded by,

$$\begin{aligned}
& \text{PoU}_T((\mathbf{u}_t)_{t=1}^{T-1}) \\
& < \|\bar{x}_1\|^2 \left[ 2\Delta_1 \left( C_* D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + 2C_{fb}^2 (C'_K \gamma^{2W} + \varepsilon'_K) \frac{1 - \eta^{2T}}{1 - \eta^2} \right) \right. \\
& \quad \left. + 2\Delta_2 \left( C_* D_K(\gamma^W, \varepsilon_K) C_x^2 \frac{1 - q^{2T}}{1 - q^2} + C_{fb}^2 (C'_K \gamma^W + \varepsilon'_K) \frac{1 - \eta^{2T}}{1 - \eta^2} \right) \right. \\
& \quad \left. + [C_* D_K(\gamma^W, \varepsilon_K) C_x C_{fb} + C_x C_{fb} (C'_K \gamma^W + \varepsilon'_K)] \frac{1 - (q\eta)^T}{1 - q\eta} \right] \\
& = \|\bar{x}_1\|^2 \left[ \Delta_a(\gamma^W, \varepsilon'_K) \frac{1 - q^{2T}}{1 - q^2} + \Delta_b(\gamma^W, \varepsilon'_K) \frac{1 - \eta^{2T}}{1 - \eta^2} + \Delta_c(\gamma^W, \varepsilon'_K) \frac{1 - (q\eta)^T}{1 - q\eta} \right].
\end{aligned}$$

By inspection, the PoU upper bound,  $\bar{\Gamma}$ , can be expressed as  $C_1 \gamma^{2W} + C_2 \gamma^W + C_3 \varepsilon_K$ , where  $C_1, C_2$  and  $C_3$  are monotonically increasing w.r.t.  $\|\bar{x}_1\|$ ,  $\lambda_{\max}(R_{\max})$ ,  $\lambda_{\max}(Q_{\max})$  and  $\lambda_{\max}(R_{\max}) - \lambda_{\min}(R_{\min})$ , and the inverse of  $\rho(A + \mathbf{B}\underline{K})$  and  $\varepsilon_Q$ .

We next find the lower bound of PoU. Let  $F_t := 2(\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t + \mathbf{B}^\top P_{t+1}^i A \right) x_t$ . Note that

$$\begin{aligned}
\text{PoU}((\Pi_t)_{t=1}^{T-1}) & \geq \sum_{t=1}^{T-1} F_t \\
& \geq \sum_{t=1}^{T-1} -2\|\mathbf{u}_t - \tilde{\mathbf{u}}_t\| \left\| (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t + \mathbf{B}^\top P_{t+1}^i A \right\| \|x_t\| \\
& \geq -\Delta_2 \bar{\Gamma}_2.
\end{aligned}$$

By inspection, the PoU lower bound  $-\Delta_2 \bar{\Gamma}_2$ , can be expressed as  $-C'_1 \gamma^{2W} - C'_2 \gamma^W - C'_3 \varepsilon_K$ , where  $C'_1, C'_2$  and  $C'_3$  are positive scalars that also monotonically increasing w.r.t.  $\|\bar{x}_1\|$ ,  $\lambda_{\max}(R_{\max})$ ,  $\lambda_{\max}(Q_{\max})$  and  $\lambda_{\max}(R_{\max}) - \lambda_{\min}(R_{\min})$ , and the inverse

of  $\rho(A + \mathbf{B}K)$  and  $\varepsilon_Q$ .

We begin with an elementary result to help us the derivation of an upper bound on  $\|\tilde{z}\|$ .

**Proposition 9.** *Consider  $Z \in \mathbb{S}_{++}^n$ ,  $c > 0$ , the maximal distance  $l$  between the origin to the set*

$$\begin{aligned} x^\top Zx + d^\top x &\leq c, \\ d^\top d &\leq \nu, \end{aligned}$$

satisfies  $\|x^*\|^2 \leq c + \frac{1}{2}\lambda_{\max}(Z^{-1})\nu$ .

*Proof.* The equation  $x^\top Zx + d^\top x = c_2$  can be rewritten as  $(x + \frac{1}{2}Z^{-1}d)^\top Z(x + \frac{1}{2}Z^{-1}d) = c + \frac{1}{4}d^\top Z^{-1}d$ . The maximal distance between the centre of the ellipse and the ellipse is  $c + \frac{1}{4}d^\top Z^{-1}d$ , and the distance between the origin to the centre of the ellipse is  $\frac{1}{4}d^\top Z^{-1}d$ . Therefore, by triangle inequality, the maximal distance between the origin and the ellipse is less than  $c + \frac{1}{2}d^\top Z^{-1}d$ , or  $c + \frac{1}{2}\lambda_{\max}(Z^{-1})\nu^2$  due to  $d^\top Z^{-1}d \leq \lambda_{\max}(Z^{-1})\nu$  when  $d^\top d \leq \nu$ .  $\square$

We now ready to prove Proposition 3.

*Proof.* As stated in Lemma C.1.11,  $\text{PoU}((\hat{\Pi}_t)_{t=1}^{T-1})$  can be written as

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T-1} (\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) (\mathbf{u}_t - \tilde{\mathbf{u}}_t) \\ &\quad + 2(\mathbf{u}_t - \tilde{\mathbf{u}}_t)^\top \left( (R_t^i + \mathbf{B}^\top P_{t+1}^i \mathbf{B}) \bar{K}_t^* + \mathbf{B}^\top P_{t+1}^i A \right) \hat{x}_t, \end{aligned}$$

where  $\mathbf{u}_t = \hat{\Pi}_t(\hat{x}_t)$ ,  $\tilde{\mathbf{u}}_t = K_t^* \hat{x}_t$  and  $\hat{x}_t$  is generated by  $(\hat{\Pi}_\tau)_{\tau=1}^{t-1}$ . Rewriting the above using  $\tilde{z}$ ,  $\tilde{H}$  and  $\tilde{b}$ , yields

$$\tilde{z}^\top \tilde{H} \tilde{z} + \tilde{b}^\top \tilde{z} \leq \delta_2.$$

Since  $\|\tilde{b}\|^2 \leq \delta_x$ , by Proposition 9,  $\|\tilde{z}\|^2 \leq \delta_2 + \frac{1}{2}\lambda_{\max}(\tilde{H})\delta_x$ .  $\square$

Before we proceed to prove Proposition 4, we state the cost different lemma for potential function.

**Lemma C.2.1** (Cost Difference Lemma). *For any  $T \geq 1$  and  $0 \leq t \leq T$ , consider  $B$  from (4.1),  $\bar{R}_t, \bar{P}_t$  from Lemma C.1.1, and  $K_t$  from (4.12), respectively. Let  $(\hat{\Pi}_t)_{t=1}^{T-1}$  be any feedback control policy among the players and  $(\Pi_t^*)_{t=1}^{T-1}$  be the feedback Nash equilibrium policy defined in (4.5). Further, states  $\{x_t\}_{t=0}^{T-1}$  be generated from control*

sequence  $\{u_t\}_{t=0}^{T-1}$  for the linear system (4.1) where each control being  $u_t = \hat{\Pi}_t(x_t)$ . Setting  $\bar{u}_t = K_t^* x_t$ , we have the following equality:

$$\Phi(\bar{x}_1, (\hat{\Pi}_t)_{t=1}^{T-1}) - \Phi(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) = \sum_{t=0}^{T-1} (u_t - \bar{u}_t)^\top (\bar{R}_t + B^\top \bar{P}_{t+1} B) (u_t - \bar{u}_t).$$

*Proof.* Proved in the manner of [12, Lemma 5] with  $Q_t$ ,  $R_t$ ,  $x_t^\pi$ ,  $u_t^\pi$ ,  $V_{t+1}$ ,  $B_u$ ,  $B_d$  replaced by  $\bar{Q}_t$ ,  $\bar{R}_t$ ,  $x_t$ ,  $u_t$ ,  $\bar{P}_{t+1}$ ,  $B$ ,  $0$ , respectively, where  $\bar{Q}_t$  is defined in the proof of Lemma C.1.1.  $\square$

Next, we prove Proposition 4 using Proposition 3.

*Proof.* From Lemma C.2.1 we have that

$$\Phi(\bar{x}_1, (\hat{\Pi}_t)_{t=1}^{T-1}) - \Phi(\bar{x}_1, (\Pi_t^*)_{t=1}^{T-1}) \leq \Delta_{max} \sum_{t=1}^{T-1} \|\hat{\Pi}_t(\hat{x}_t) - \Pi_t^*(\hat{x}_t)\|^2 = \Delta_{max} \|\tilde{z}\|^2 \leq \bar{\epsilon}$$

where Proposition 3 gives  $\|\tilde{z}\|^2 \leq \bar{\delta}_2$ .  $\square$

Lastly, we present the proof of Proposition 5.

*Proof.* Consider the case of  $N = 2$ . For  $t = T$ , since  $P_T^1 = P_T^2 = Q_T$ , and

$$\Theta_{T-1} = \begin{bmatrix} r_{1,T-1} & 0 \\ 0 & r_{2,T-1} \end{bmatrix} + \mathbf{B}^\top Q_T \mathbf{B},$$

we have  $\Theta_{T-1} \succ 0$ . Then, for  $t = T - 1$ ,

$$P_{T-1}^1 - P_{T-1}^2 = K_{T-1}^\top (R_{T-1}^1 - R_{T-1}^2) K_{T-1}.$$

We next claim that  $P_{T-1}^1 = P_{T-1}^2$ . Since

$$\begin{aligned} & \det(\Theta_{T-1})^2 \mathbf{B} \Theta_{T-1}^{-1} (R_{T-1}^1 - R_{T-1}^2) \Theta_{T-1}^{-1} \mathbf{B}^\top = \\ & \begin{bmatrix} -b_1 & -b_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_{2,T-1} + B^{2\top} P_{T-1}^2 B^{2\top} & B^{2\top} P_{T-1}^2 B^{1\top} \\ B^{1\top} P_{T-1}^1 B^{2\top} & r_{1,T-1} + B^{1\top} P_{T-1}^1 B^{1\top} \end{bmatrix} \\ & \begin{bmatrix} r_{1,T-1} & 0 \\ 0 & -r_{2,T-1} \end{bmatrix} \begin{bmatrix} r_{2,T-1} + B^{2\top} P_{T-1}^2 B^{2\top} & B^{2\top} P_{T-1}^2 B^{1\top} \\ B^{1\top} P_{T-1}^1 B^{2\top} & r_{1,T-1} + B^{1\top} P_{T-1}^1 B^{1\top} \end{bmatrix} \\ & \begin{bmatrix} -b_1 & 0 \\ -b_2 & 0 \end{bmatrix} \end{aligned}$$

Due to

$$\begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} h_1 & 0 \\ 0 & h_2 \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} = \begin{bmatrix} h_1\alpha^2 + h_2\beta^2 & h_1\alpha\beta + h_2\beta\gamma \\ h_1\alpha\beta + h_2\beta\gamma & h_1\beta^2 + h_2\gamma \end{bmatrix},$$

and

$$\begin{aligned} B^{2\top} P_{T-1} B^2 &= b_2^2 [Q_T]_{11}, \\ B^{1\top} P_{T-1} B^1 &= b_1^2 [Q_T]_{11}, \end{aligned}$$

substitute  $h_1 = r_{1,T-1}$ ,  $h_2 = -r_{2,T-1}$ , we have

$$\frac{h_1\alpha}{-h_2\gamma} = \frac{r_{1,T-1} r_{2,T-1} + b_2^2 [Q_T]_{11}}{r_{2,T-1} r_{1,T-1} + b_1^2 [Q_T]_{11}} = \frac{r_{1,T-1} r_{2,T-1}}{r_{2,T-1} r_{1,T-1}} = 1,$$

therefore,  $h_1\alpha\beta + h_2\beta\gamma = 0$ . Moreover,

$$\begin{bmatrix} -b_1 & -b_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \begin{bmatrix} -b_1 & 0 \\ -b_2 & 0 \end{bmatrix} = \begin{bmatrix} c_1 b_1^2 + c_2 b_2^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Here,  $c_1 = r_{1,T-1}(r_{2,T-1} + b_2^2 [Q_T]_{11})$ ,  $c_2 = -r_{2,T-1}(r_{1,T-1} + b_1^2 [Q_T]_{11})$ . Apply  $\frac{b_1^2}{b_2^2} = \frac{r_{1,T-1}}{r_{2,T-1}}$  again, we have  $c_1 b_1^2 + c_2 b_2^2 = 0$ . Therefore,  $K_{T-1}^\top (R_{T-1}^1 - R_{T-1}^2) K_{T-1} = 0$ .

Let's assume that  $P_{t+1}^1 - P_{t+1}^2 = 0$ . Due to

$$\Theta_t = \begin{bmatrix} r_{1,t} & 0 \\ 0 & r_{2,t} \end{bmatrix} + \mathbf{B}^\top P_{t+1}^1 \mathbf{B},$$

we have  $\Theta_t \succ 0$ , therefore  $\det(\Theta_t) > 0$ . Consider

$$K_t^\top (R_t^1 - R_t^2) K_t = \mathbf{B} \Theta_t^{-1} (R_t^1 - R_t^2) \Theta_t^{-1} \mathbf{B}^\top.$$

By repeating the steps above and using the relationship of  $B^{1\top} P_{t+1}^1 B^1 = b_1^2 [P_{t+1}^1]_{11}$ , we have

$$\frac{h_1\alpha}{-h_2\gamma} = \frac{r_{1,t} r_{2,t} + b_2^2 [P_{t+1}^2]_{11}}{r_{2,t} r_{1,t} + b_1^2 [P_{t+1}^1]_{11}} = 1,$$

and  $\frac{b_1^2}{b_2^2} = \frac{r_{1,t}}{r_{2,t}}$ . Therefore, by induction, we can conclude that for  $0 \leq t \leq T-1$ , parameters  $P_t$ ,  $\Theta_t$  defined in (4.13) and (4.11), respectively, satisfy Assumptions 4.2.1 and 4.2.5. For the case of  $N > 2$ , by setting  $\frac{b_i^2}{r_{i,t}} = \frac{b_j^2}{r_{j,t}}$  for  $i \in \{1, 2, \dots, N\}$ ,  $t \in \mathbb{N}_T$ ,

we have

$$R_t^1 - R_t^i = \begin{bmatrix} r_{1,t} & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & & & \\ 0 & \cdots & -r_{i,t} & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}.$$

We can assert the proposition by repeating the steps for the case of  $N = 2$ . □