REINFORCEMENT LEARNING FOR LINEAR QUADRATIC CONTROL UNDER MULTIPLICATIVE NOISE: POLICY METHODS, SCALABILITY, AND ENTROPY REGULARIZATION

HAORAN ZHANG

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Applied Mathematics

Reinforcement Learning for Linear Quadratic Control under Multiplicative Noise: Policy Methods, Scalability, and Entropy Regularization

Haoran Zhang

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

June 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signature)

Haoran Zhang (Name of student)

Dedicated to my beloved parents, my puppy Cartoon, and my uncle in heaven

Abstract

The discrete time Linear Quadratic (LQ) control problem has been a cornerstone of control theory, and it provides an optimization framework for regulating linear dynamical systems, with quadratic cost functions used for minimization. With a closed-form solution from the algebraic Riccati equation (ARE), the LQ problem offers unique tractability in optimal control problems. However, in many practical scenarios, the true system parameters and noise characteristics are unknown, rendering model-based solutions infeasible. Consequently, reinforcement learning (RL) has become a powerful data-driven approach for policy optimization in the LQ problem under parameter uncertainty. Recent advances have highlighted its versatility with applications in the field of dynamic portfolio management, financial derivative pricing, and biomedical engineering. This thesis is organized into three sections. It focuses on key aspects of RL for the discrete time LQ problem and hence highlights their relevance to financial decision-making problems. In addition, the thesis contributes to expanding the single-agent LQ problem to multi-agent settings to address challenges related to scalability and decentralized control structures. Throughout, the emphasis is on advancing theoretical insights and offering practical tools for data-driven control in financial and engineering applications. Examples include the mean-variance (MV) problem and the optimal liquidation problem. Each section offers theoretical insights and practical contributions to enhance policy optimization techniques under diverse contexts.

Part I studies finite horizon LQ control with mixed noise through a perturbance-wise viewpoint that unifies the classical model, constrained control case, and a distributionally robust extension calibrated from an unknown noise distribution framework. By augmenting the affine controller into a linear feedback form, I derive an implementable policy gradient method that accommodates a nonzero noise mean estimated from samples. I establish global decrease and convergence guarantees under constant step sizes chosen as explicit low-order functions of problem parameters. Numerical studies in MV portfolio allocation and dynamic benchmark tracking validate stable convergence and illustrate sensitivity trade-offs across horizon length, trading frictions, and estimation windows.

Part II focuses on an iterative policy optimization (IPO) framework for multi-agent LQ control problems, where the system is subject to multiplicative noise and governed by a Shannon entropy-regularized cost function. It demonstrates that IPO achieves a super-linear convergence rate in the neighborhood of the optimal policy. The applications of MV optimization and optimal liquidation problems underscore the utility of IPO in financial decision-making contexts. This chapter explores the understudied scenario in which multiplicative noise simultaneously influences both control and state variables. It fills a gap in the extensive policy iteration literature by analyzing how noise impacts model efficiency in a multi-agent case. The findings establish a theoretical basis for enhancing the effectiveness of RL algorithms in the multi-agent LQ problem.

Part III presents an innovative, data-driven policy iteration approach for LQ problem under Tsallis-entropy regularization. This study utilizes Tsallis entropy to balance exploration and sparsity in control laws, contrasting with the prevailing emphasis on Shannon entropy in the previous literature. The methodology utilizes Q-functions estimated by least-squares methods with instrumental variables, offering a robust solution for problems associated with state/control-multiplicative noise.

This study tackles significant deficiencies in the integration of Tsallis entropy with data-driven methods, thereby opening new directions for enhancing RL-based policy in stochastic control systems.

Keywords: Reinforcement learning, policy gradient, policy iteration, Tsallis entropy, linear quadratic control, Riccati equation, financial decision-making, multi-agent

Acknowledgements

The course of our lives has always been determined by luck and choice, but they are frequently disregarded. It must be an extraordinary journey if there is a period in your existence during which you possess both. I may have a family and a career in a few years, and when I reflect on my PhD studies in Hong Kong, I will undoubtedly consider it an excellent choice and a four-year period of great luck. I have been able to embark on this academic path that is full of unknowns because of the encouragement and support I have received from all of you over the past four years. This success is a direct result of your presence. I eagerly anticipate your continued presence in the future. I have never been particularly intelligent, at least not to the extent that would enable me to conduct scientific research in the field of applied mathematics. However, I have no regret for the choice to pursue my PhD rather than working in the industry after completing my master's degree. This experience has added a sense of fulfillment, completion, and radiance to my life.

I would like to express my sincere gratitude to Prof. Li Xun, my chief supervisor. He is the individual who provided me with the opportunity to pursue my PhD degree. He is patient with me when I am lacking in related math knowledge and provides guidance on how to address my deficiencies. In addition, Prof. Li also offers me technical guidance and targeted directions that are tailored to my strengths and weaknesses. Furthermore, Prof. Li presents me with an adequate amount of study and research latitude, which enables me to feel supported and at ease throughout

my PhD study. In contrast, I think that I would experience physical and mental exhaustion, as well as struggle and suffer. I am appreciative of the time, ideas, and contacts he has contributed, which are instrumental in the success of my PhD study.

I would also like to express thanks to my co-supervisor, Prof. Yu Xiang, who is a youthful and highly thoughtful professor. I am very appreciative of the scientific instruction and motivational concepts he has offered me. He is highly detail-oriented and frequently concentrates on aspects that I have not yet considered. This assistance is both opportune and beneficial to me. I not only broadened my research horizons but also saved a significant amount of time that could have been spent on fruitless endeavors during my discussions with Prof. Yu. Furthermore, I earnestly hope that Prof. Yu will make significant strides in his research and personal career. I am genuinely lucky to have met you.

In the end, I can never say thank you enough to my parents. I am truly lucky to have you in my life, as your affection and concern for me have persisted for the past twenty-seven years, much like water moistens all things. Your unwavering support, both mental and physical, is essential in my choice to pursue a PhD with its successful completion. I am like a kite that is soaring through the sky in your eyes, eliciting feelings of pride and contentment in you. Conversely, I am constantly reminded of your being behind me, which provides me with a sense of comfort and motivation. Certainly, I will never forget my adorable puppy, Cartoon. Whenever I arrive home feeling exhausted, you always rush to me with a joyful tail wag, and your appearance dispels my anxieties. My uncle passed away after a serious illness from us during my PhD years. Every time I returned home, you were the one to pick me up at the airport, and you were the one to remain with me the night before I departed for Hong Kong. I must be your nephew that you are proud of, and I will accompany my cousin (your daughter) on my run to a brighter future. Last but not least, I am so lucky to have my close friends in PolyU, Dr. Luo Sheng and Dr. Sun Bei, who have

been instrumental in helping me continue and complete my research. Your support has been invaluable, and I am deeply grateful. I am so thankful to each and every one of you for your contributions to my thesis. I am extremely lucky to have you all in my life.

Contents

List of Figures

List of Tables

List of Notations

$\mathbb{R}$	Real number field;
$\mathbb{R}_+$	The subset of $\mathbb{R}$ consisting of positive elements;
$Z^\top$	The transpose of matrix Z;
$\|Z\|$	The spectral norm of a matrix Z;
$\|Z\|_F$	The Frobenius norm of a matrix Z;
$\mathrm{Tr}(Z)$	The trace of a matrix Z;
$\sigma_{\min}(Z)$	The minimum singular value of a square matrix Z;
$\sigma_{\max}(Z)$	The maximum singular value of a square matrix Z;
$\mathcal{F}_t$	The filtration (information set) at time t;
$\|\|\|\boldsymbol{D}\|\|\|$	The sum of $\|D_i\|$ for a sequence of matrices $D = (D_0, \ldots, D_i, \ldots, D_T)$;
$\otimes$	The Kronecker product;
$\mathbf{0}_{n\times m}$	The $n \times m$ zero matrix;
$\mathbf{I}_n$	The $n \times n$ identity matrix;
$\mathrm{vec_s}(Z)$	The symmetrized vectorization of matrix Z;
$\mathrm{unvec_s}$	The inverse of $\mathrm{vec_s}$.

Chapter 1

Introduction

1.1 Introduction

The linear quadratic (LQ) problem constitutes a foundational pillar of control theory. The primary aim is to identify an optimal control strategy for regulating a linear dynamical system, whose dynamics are linear in the state and control input, to minimize a quadratic cost function (Abeille et al. (2016)). The optimal control input can be expressed as a constant feedback gain that operates on the system state, while the gain is obtained by solving the algebraic Riccati equation (ARE). The LQ problem is a canonical example among optimal control problems that admit a closed-form analytical description of the optimal feedback control (Almgren and Chriss (2001), Li and Todorov (2004)). When the dynamics exhibit nonlinearity and are difficult to analyze, an LQ approximation can be constructed by locally expanding. This approximation is locally accurate to the original problem (Benigno and Woodford (2012), Gu (2011), McKean (1966)).

In recent years, the LQ problem has attracted substantial interest due to its versatile applications among multiple domains, including dynamic portfolio management and optimization (Abeille et al. (2016), Cui et al. (2014b), Fu et al. (2010), Gao et al. (2015), Li et al. (2002), Li and Ng (2000)), financial derivative pricing (Baz and Chacko (2004), Costa et al. (2005), Primbs and Sung (2009), Zuhlsdorff (2001))

and optimal liquidation (Ackermann et al. (2021), Almgren and Chriss (2001)). This widespread applicability has heightened focus on the LQ control problem within these fields (Elliott et al. (2013), Ni et al. (2015)). Motivated by these applications, this thesis focuses on LQ control with multiplicative noise and realistic feasibility requirements. I further investigate learning-based approaches for computing implementable feedback gains when model parameters are unknown or only partially observed.

1.1.1 Related Study

Multiplicative Noise LQ Problem and Applications. Before focusing on multiplicative noise models, we briefly review constrained LQ methods that address implementability requirements in practical applications. LQ problems are valued for their ability to produce explicit control policies through ARE solutions, which offer a robust framework for control and optimization in both continuous and discrete time systems. However, integrating real-world constraints, such as economic regulations, risk factors, or physical limitations, into these models introduces complexities that often preclude closed-form solutions. This challenge has driven the development of advanced methodologies, including semi-analytical solutions (Gao et al. (2015)), parametric programming in Bemporad et al. (2002), and model predictive control (MPC) in Mayne (2014) and Mesbah (2016). These approaches enable approximation or numerical solution of constrained LQ problems and support realistic settings with inequalities and other implementability constraints on states and controls.

In many modern applications, however, uncertainty does not enter the dynamics as additive noise. Instead, the system matrices may fluctuate randomly, leading to multiplicative noise models in which the noise intensity scales with the current state and/or control. Such models naturally arise in engineering systems with random gains and in financial settings where random returns multiply the invested position or wealth (Basin et al. (2006), Gershon and Shaked (2006), Rami et al. (2001)).

Compared with additive noise LQ settings, multiplicative noise changes the nature of stability and solvability, typically requiring mean-square (second-moment) analysis and yielding generalized Riccati-type characterizations (Ni et al. (2015), Rami et al. (2001)). This structure appears prominently in finance-motivated LQ formulations such as dynamic mean-variance (MV) portfolio optimization and optimal liquidation in which stochastic returns, frictions, and implementation constraints play an essential role (Ackermann et al. (2021), Almgren and Chriss (2001), Cui et al. (2014b), Gao et al. (2015), Zhou and Li (2000)).

The presence of multiplicative noise fundamentally couples optimal control with second-moment behavior: stability and performance are typically studied in a mean-square sense, and solvability may still hold even when the quadratic penalties on state and/or control are indefinite (Mukaidani and Xu (2017), Ni et al. (2015), Rami et al. (2001)). A major application domain is dynamic MV portfolio optimization, which extends Markowitz's classical static portfolio selection to dynamic settings (Cui et al. (2014b), Gao et al. (2015), Li and Ng (2000), Rubinstein (2002), Zhou and Li (2000)). In such models, multiplicative noise naturally arises from random asset returns, while constraints and market frictions motivate constrained MV formulations; several works provide explicit characterizations and structural conditions under which these problems are posed (Czichowsky and Schweizer (2013), Fu et al. (2010), Sun and Wang (2006), Wu et al. (2018)). Beyond portfolio selection, LQ-style modeling also underpins optimal liquidation problems that seek to balance execution costs and risk in the presence of stochastic price fluctuations (Almgren and Chriss (2001), Ackermann et al. (2021)). The optimal liquidation problem concerns executing a large position over a fixed time horizon in which a single large order can represent a substantial fraction of daily trading volume and induce significant market impact (Alfonsi et al. (2008), Becherer et al. (2018)). To reduce execution cost and adverse price impact, the order is typically split into smaller trades and scheduled

over time; the goal is to choose an execution policy (e.g., trading rate and timing) that minimizes a functional cost capturing both temporary and permanent price impact as well as risk (Almgren and Chriss (2001)). Recent work has also explored learning-based execution methods, including policy gradient and Q-learning, to learn such execution policies from data; in many cases, these approaches are evaluated primarily through empirical studies, and rigorous convergence or optimality guarantees remain limited. Together, these model-based results form an essential baseline for this thesis: they identify what structure can be retained under multiplicative uncertainty and constraints, and they motivate later discussions of computation and implementation under additional information or modeling limitations. These results motivate my later focus on decentralized information structures and learning-based computation when model parameters are unavailable.

Multi-Agent LQ and Decentralized Control. Many real-world systems are inherently multi-agent because decision-making authority, sensing, and actuation are distributed across multiple subsystems or stakeholders. These systems consist of multiple decision-makers interacting through shared dynamics and common or partially aligned objectives. In this context, decentralization refers to an information constraint: each agent chooses its action based on local observations and possibly limited neighbor communication rather than the full global state. The multi-agent LQ framework extends classical LQ problems to model such scenarios. Applications include autonomous vehicles, power grids, financial networks, and sensor arrays, where coordination among agents is essential. The extension of LQ control to multi-agent systems has led to significant interest in decentralized frameworks in which each agent determines its control inputs based solely on local information. For instance, Wang et al. (2017) proposed a decentralized LQ regulator in which each agent optimizes its own strategy using neighbor information under a consensus objective. Subsequent work has expanded on these ideas to develop suboptimal or observer-based strategies

that reduce computational complexity and preserve privacy (Zhang et al. (2023)). In multi-agent settings, cooperative decision-making becomes a central challenge. Solutions can be broadly categorized, but the core paradigm is using Centralized Training with Decentralized Execution (CTDE) while leveraging centralized access to full global states and actions during training and maintaining fully decentralized execution policies at runtime (Wilk et al. (2024)). CTDE relaxes the information requirement during training by allowing centralized access to global trajectories, while the learned policies are constrained to use only local information at execution time. This setup enables scalability while preserving some coordination through intermittent synchronization or shared rewards (Abouheaf et al. (2014)). In Chapter 4, we adopt a CTDE-like approach in our decentralized multi-agent LQ setting: we jointly train policies using global information, then enforce a shared decentralized feedback gain matrix so that each agent can act using only its local state at execution time.

Reinforcement Learning. Reinforcement learning (RL) is a core paradigm in machine learning, focusing on how agents take actions in an environment to maximize cumulative rewards over time (Sutton and Barto (2018)). RL is distinguished by its ability to address sequential decision-making problems in which the system dynamics are partially known or completely unknown. The framework of RL typically revolves around the Markov Decision Process (MDP), comprising states, actions, rewards, and a transition function, which collectively define the agent-environment interaction (Puterman (2014)). A key advantage of RL is its flexibility: agents learn optimal policies through trial-and-error interactions rather than requiring explicit system models. This adaptability has enabled researchers to find RL applications in diverse fields, from robotics and game playing to finance and control systems (Levine et al. (2020), Silver et al. (2017)).

Within RL, two primary methodologies dominate: model-based methods, which

attempt to explicitly estimate the underlying transition dynamics; and model-free methods, which directly learn policies or value functions without requiring transition kernel estimation (Jin et al. (2020), Wulfmeier et al. (2015)). While model-based methods can often converge faster given accurate models, their dependence on accurate system knowledge limits their applicability in uncertain or complex environments. Conversely, model-free methods, such as policy gradient and Q-learning algorithms, have gained prominence for their robustness in environments where the system dynamics are opaque or highly variable (Haarnoja et al. (2018)).

RL has made significant strides in addressing LQ problems, a classical control challenge in engineering and economics that is often solved analytically via ARE (Anderson and Moore (2007), Kalman et al. (1960)). Traditionally, LQ solutions assume complete knowledge of model parameters, enabling the derivation of optimal feedback control laws in closed form. However, in practical applications, these assumptions often break down due to system uncertainties, measurement noise, or parameter variability (Bu et al. (2020)).

By integrating RL into the LQ framework, agents can learn optimal control policies directly from data. Model-free RL methods, in particular, have demonstrated considerable promise for solving LQ problems without requiring precise parameter knowledge. Methods such as policy gradient (Fazel et al. (2018)) and least-squares policy iteration (Hambly et al. (2021)) have been adapted to discrete time LQ settings, which achieve convergence guarantees under various assumptions. These methods leverage trial-and-error interaction to iteratively refine control policies, and they extend naturally to stochastic dynamics and entropy-regularized objectives. The synergy between RL and LQ problems opens new avenues for solving practical control problems in engineering, robotics, and financial systems, whose analytical solutions are either infeasible or suboptimal in the presence of real-world complexities.

Policy Gradient Method. Policy gradient methods optimize a parameterized policy by ascending the gradient of expected return; in LQ control, the parameters are often taken to be the feedback gain itself (Fazel et al. (2018)). The policy gradient approach aims to optimize the policy function by adjusting parameters based on the gradient of the expected reward, which allows for dynamic learning in complex or unknown environments without the need to solve static equations. This shift highlights a wider trend in control theory, moving away from algebraic and iterative methods toward learning-based algorithms that develop optimal control strategies using model-free techniques. These strategies combine dynamic programming principles with RL to create advanced adaptive control systems capable of handling continuous and discrete time systems without prior knowledge of the system dynamics (Jiang and Jiang (2012), Lee and Hu (2018)).

The study of policy gradient methods has gained significant momentum in RL, particularly for solving LQ problems across various settings. The foundational work in this area showcased the global convergence of the policy gradient method in infinite horizon and deterministic dynamics, thus establishing a precedent for subsequent research (Fazel et al. (2018)). Further work has expanded the policy gradient method to account for stochastic elements, such as noise systems, in the system dynamics, which required adaptations of the deterministic framework to maintain convergence guarantees and to tackle the intricacies over an infinite or finite horizon (Bu et al. (2019), Gravell et al. (2020), Jin et al. (2020), Malik et al. (2019), Pang and Jiang (2022)). This progression reflects the dynamic nature of policy gradient research and underscores both its potential and the need for continued study into its convergence properties and performance guarantees across deterministic and stochastic models, infinite and finite horizons, and theoretical and practical applications (Bhandari and Russo (2024), Bu et al. (2020), Wang et al. (2020)). Algorithms have been developed that converge in both parameter estimation and optimal control. The subject

of randomized adaptive control algorithms for LQ systems has been significantly improved by recent advances from Agarwal et al. (2019) and Basei et al. (2022). Computing the optimal policy in the LQ problem without solving the ARE is often challenging because of the dynamic nature of the control problem, notwithstanding the importance of the cost function. In recent times (Balakrishnan and Vandenberghe (2003), Guo et al. (2023), Jiang and Jiang (2012), Lee et al. (2012), Lee and Hu (2018)), there has been a notable increase in the enthusiasm for developing optimal control techniques directly. This rise has been influenced by the ideas of RL and dynamic programming, which provide data-driven solutions that do not rely on models. Moreover, the juxtaposition of model-based with model-free methods in RL highlights the nuanced trade-offs between robustness against model mis-specification and the exploitation of the LQ structure for enhanced sample efficiency. I build on these developments by studying policy gradient-type updates for LQ control under multiplicative noise and additional structural constraints arising from decentralized or single-agent implementation.

Entropy Regularization. Entropy regularization is prevalent in RL, as evidenced by Auer and Ortner (2006), Littlestone and Warmuth (1994), Neu et al. (2017), Sutton and Barto (2018) and Zhao et al. (2019). Entropy regularization is frequently employed to encourage exploration and improve convergence. Research has shown that incorporating entropy regularization leads to a smoother and more consistent landscape, allowing for the use of higher learning rates. As a result, this process can lead to faster convergence during the training phase. The notion of entropy regularization is crucial for achieving a balance between exploration and exploitation in RL. This concept has been highlighted in the works of Haarnoja et al. (2017) and Haarnoja et al. (2018). Sometimes the learning objective includes Tsallis entropy as a regularizer, introduced by Tsallis (1988) and Tsallis (2011), to represent the agent's action distribution. This approach generalizes Shannon entropy for non-extensive

systems in which standard additivity is not applicable, including systems characterized by long-range interactions, fractal structures, or non-equilibrium dynamics. This generalization alters the optimal action distribution to be q-Gaussian rather than Gaussian, as established by the Shannon entropy in Donnelly and Jaimungal (2024) and Wang et al. (2020). In RL, Choy et al. (2020) and Lee et al. (2018) propose a method for obtaining sparse control policies through a specific application of Tsallis entropy.

This thesis investigates RL and data-driven methods for the discrete time LQ control problem, with particular emphasis on settings that arise in practice, including multiplicative noise, implementability constraints, scalability, and entropy-regularized objectives. Throughout the thesis, several applications, e.g., dynamic MV portfolio optimization and optimal liquidation, are used as representative application case studies to illustrate both the modeling and algorithmic implications. This thesis is organized into three parts.

Part I addresses three cases of discrete time stochastic LQ optimal control problems with multiplicative noise over a finite horizon, including classical, scalar state constrained, and unknown noise distribution cases. Our primary objective is to investigate the convergence properties of policy gradient methods applied to their linear systems that are characterized by both state and control multiplicative noise. The theoretical foundation and problem framework are introduced. This study investigates the policy gradient method by assuming the noise distribution with sampled means and covariances, and presents findings with real data on convergence for the applications on the MV problem and benchmark tracking problem. The existing literature provides a robust foundation from which to extensively examine analytical solutions for LQ problems with constraints, unknown noise distribution, and convergence proofs for policy gradient methods within the unconstrained LQ problem affected by additive noise. However, a critical gap remains: no prior studies have

integrated these aspects to assess the performance of policy gradients solving the generalized LQ problem, incorporating multiplicative noise and different variations. This study fills this gap by demonstrating the robustness and versatility of the policy gradient method across diverse system dynamics. We refine existing proofs to align with the complex structure of our model, thereby contributing a novel theoretical dimension to the current body of research.

Part II focuses on the LQ problem with multiplicative noise and demonstrates how the iterative policy optimization (IPO) method, a reliable technique for resolving this problem, ensures convergence with a theoretical guarantee. Our contributions are that we rigorously analyze the IPO method and establish that it achieves a super-linear convergence rate if it enters a local region around the optimal policy. We demonstrate the practicality of the IPO method through its application to the MV optimization and the optimal liquidation problem, showing promising results for practical financial decision-making problems. There are many existing studies on the LQ problem and policy iteration learning method, but the case where the noise is influenced by the control and state value is rarely discussed. Therefore, an explicit result for the context of RL-based policy and improving the efficiency of algorithms that depend on policy optimization is needed for related groups with investment goals.

Part III explores the application of Tsallis entropy in the multiplicative LQ problems. This part introduces a novel, model-free, and fully data-driven policy iteration algorithm designed specifically for the multiplicative LQ problem under the Tsallis entropy framework. While the existing literature predominantly focuses on Shannon entropy, our work extends the scope by leveraging Q-functions combined with instrumental variables employed in least squares learning. These methodological advancements address gaps in the literature where the integration of Tsallis entropy with data-driven approaches remains underexplored. By doing so, we provide a foun-

dational result that enhances RL-based policy and opens new avenues for improving algorithmic efficiency in stochastic control problems.

These three parts form a progression from model-based policy optimization under structured uncertainty, to scalable decentralized learning under entropy regularization, and finally to fully data-driven entropy-regularized control. The common goal is to compute implementable policies for LQ systems under multiplicative noise when model knowledge and execution structure become progressively more restrictive.

1.2 Organization of the Thesis

The remainder of the thesis is organized as follows.

Chapter 2 summarizes the basic model for the LQ problem and the definition of Shannon entropy and Tsallis entropy. The finance-related applications, that is, the MV problem and optimal liquidation problem, are also discussed in this chapter.

Chapter 3 investigates the convergence of policy gradient methods in LQ control models with multiplicative noise. The study blends existing frameworks with the RL method, accommodating both state and control multiplicative noise to address an unknown noise distribution system with sampled means and covariance. By deriving rigorous proofs and validating convergence, the results enhance the theoretical understanding of stochastic control.

Chapter 4 presents the IPO method for discounted entropy-regularized LQ problems across an infinite time horizon with multiplicative noise. The study establishes its global linear convergence and the super-linear convergence of the IPO method near the optimal policy, demonstrating its effectiveness in scalar state and high-dimensional scenarios. Practical applications to MV optimization and optimal liquidation problems validate the method's efficiency in financial decision-making.

Chapter 5 develops RL in this setting by examining the behavior of the least

squares estimator based on the Q-function for the multiplicative noise LQ problem with Tsallis entropy. The algorithm of such a policy iteration method is rigorously established, offering insights into optimal control.

Chapter 6 provides concluding remarks and discusses future work.

Chapter 2

Preliminary

This chapter provides a concise introduction to the essential mathematical principles of the basic LQ problem with multiplicative noise and the entropy preliminaries. The typical real-life examples, including the MV problem and the optimal liquidation problem, along with a compilation of requisite results utilized throughout the thesis, are discussed.

2.1 LQ Problem

Define the admissible policy set as $\Pi = \{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{U})\}$, with $\mathcal{X}$ the state space, $\mathcal{U}$ the action space, and $\mathcal{P}(\mathcal{U})$ the space of probability measures on action space $\mathcal{U}$. Here, each admissible policy $\pi \in \Pi$ maps a state $x \in \mathcal{X}$ to a randomized action in $\mathcal{U}$.

The state dynamic equation remains the same for both infinite and finite horizons:

$$x_{t+1} = Ax_t + Bu_t + (Cx_t + Du_t)w_t,\ x_0 \sim \mathcal{D}_0, \tag{2.1.1}$$

where $\{w_t\}_{t\geqslant 0}$ is the noise sequence and $\mathcal{D}_0$ denotes the distribution of the initial state x_0. A relaxed policy $\pi \in \Pi$ specifies a conditional distribution $\pi(\cdot|x)$ over controls given the current state x, and we write $u_t \sim \pi(\cdot|x_t)$. The decision-maker aims to find an optimal policy by minimizing the following objective function

$$\min_{\pi\in\Pi} \mathbb{E}_{x_0\sim\mathcal{D}_0}\left[J_\pi(x)\right], \tag{2.1.2}$$

with value function J_π for the infinite horizon case given by

$$J_\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^\top Q x_t + u_t^\top R u_t \right) \middle| x_0 = x \right], \tag{2.1.3}$$

where $t = 0, 1, 2, \cdots$. The discount factor is denoted by $\gamma \in (0, 1)$. Now, for the finite horizon case, the value function changes to the following:

$$J_\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \left(x_t^\top Q_t x_t + u_t^\top R_t u_t \right) + x_T^\top Q_T x_T \middle| x_0 = x \right], \tag{2.1.4}$$

where T is the total number of time steps for $t = 0, 1, 2, \cdots, T-1$.

Here $x_t \in \mathbb{R}^m$ is the state of the system, and $u_t \in \mathbb{R}^n$ is the control at time t following a policy π. In addition, $\{w_t\}_{t=0}^{\infty}$ are zero-mean independent and identically distributed (i.i.d.) noises. We assume that $\{w_t\}_{t=0}^{\infty}$ have finite second moments. That is, $\mathrm{Tr}(W) < \infty$ with $W = \mathbb{E}\left[w_t w_t^\top\right]$ for any $t = 0, 1, 2, \cdots$. The matrices $A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times n}$ define the system's (transition) dynamics. The matrices $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ are positive definite, which parameterize the quadratic costs. The expectations in (2.1.3) and (2.1.4) are taken with respect to the control $u_t \sim \pi(\cdot|x_t)$ and system noise w_t for $t \geqslant 0$.

Remark 2.1. *The model (2.1.1) allows multiplicative noise through the term $(Cx_t + Du_t)w_t$. For completeness, we also allow an additive exogenous noise term. In Chapter 3, we consider an additional additive noise besides the multiplicative noise term that is treated as a mixed noise specialization and is used there explicitly.*

2.2 Entropy Regularization

Shannon Entropy. A given admissible policy $\pi \in \Pi$ assigns to each state x a probability density (or distribution) $\pi(\cdot|x)$ on $\mathcal{U}$. The corresponding Shannon entropy

at state x, originally introduced by Shannon (1948) (see Guo et al. (2023) for an application in entropy-regularized control), is defined as:

$$\mathcal{H}(\pi(\cdot|x)) = -\int_{\mathcal{U}} \pi(u|x) \log \pi(u|x) du \quad \forall x \in \mathcal{X}, \tag{2.2.5}$$

where $\mathbb{E}[-\log \pi(u|x)]$ can be used to generalize the above.

Tsallis Entropy. To enhance exploration throughout the control selection process, we modify the performance criterion by integrating a reward based on entropy. We employ Tsallis entropy as a measure to quantify the degree of uncertainty in the control distribution. This concept was initially presented by Tsallis (1988). Before presenting the Tsallis entropy $\mathcal{T}_q$, it is essential to construct the q-exponential and q-logarithm functions, as well as the multivariate q-Gaussian within the framework of Tsallis statistics (Hashizume et al. (2024); Tsallis (1994)). The variable q is designated as a deformation parameter (Gell-Mann and Tsallis (2004); Borges (2004); Lavagno and Swamy (2002)). Since $q \to 1$, the formula reduces to the Shannon entropy. In this study, we assume that the value of $q \in (0, 1)$ for simplicity.

Definition 1 (q-Exponential and q-Logarithm functions)**.**

$$\begin{aligned} \exp_q(x) &= [1 + (1-q)x]_+^{\frac{1}{1-q}}, \quad x \in \mathbb{R}, \\ \log_q(x) &= \frac{x^{1-q} - 1}{1-q}, \qquad\qquad x > 0, \end{aligned} \tag{2.2.6}$$

where $[x]_+ = \max(x, 0)$*. These functions have the inverse property that*

$$\begin{aligned} \exp_q\left(\log_q(x)\right) &= x, \quad x > 0, \\ \log_q\left(\exp_q(x)\right) &= x, \quad x > -\frac{1}{1-q}. \end{aligned} \tag{2.2.7}$$

Definition 2 (Multivariate q-Gaussian)**.** *A q-Gaussian $N_q(\mu, \Sigma)$ is an n-dimensional random variable characterized by its density function as follows:*

$$\varphi(x) = \frac{1}{Z_q} \exp_q \left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{(n+4)-(n+2)q} \right), \tag{2.2.8}$$

where

$$Z_q = \det(\Sigma)^{\frac{1}{2}} \left(\pi \frac{(n+4)-(n+2)q}{1-q} \right)^{\frac{n}{2}} \frac{\Gamma\left(\frac{2-q}{1-q}\right)}{\Gamma\left(\frac{2-q}{1-q} + \frac{n}{2}\right)},$$

with $\Gamma(\cdot)$ denoting the Gamma function.

The associated q-Gaussian has compact support for $q < 1$, so large magnitude controls have density zero outside an ellipsoid $(x-\mu)^\top \Sigma^{-1}(x-\mu) \leqslant R_q^2$, whereas the Gaussian policy under Shannon entropy has full support on $\mathbb{R}^n$. Subsequently, we define Tsallis entropy, which functions as an extension of Shannon entropy.

Definition 3 (Tsallis entropy and q-Entropy)**.** *Tsallis entropy $\mathcal{T}_q$ is defined as*

$$\mathcal{T}_q(f) = -\int f(x)^q \log_q f(x) dx. \tag{2.2.9}$$

The deformed Tsallis entropy (Bao and Sakaue (2022)) or q-entropy $\mathcal{H}_q$ is defined as

$$\mathcal{H}_q(f) = -\frac{1}{2-q} \left(\int f(x) \log_q f(x) dx - 1 \right). \tag{2.2.10}$$

In the following chapter, we will use the deformed q-entropy for the sake of notational simplicity.

This definition demonstrates the equivalence of Tsallis entropy and deformed q-entropy. For the sake of notation, the deformed q-entropy will serve as a regularization term in this study. For a given admissible policy $\pi \in \Pi$, the corresponding

Tsallis entropy is defined as

$$\mathcal{H}_q(\pi(\cdot|x)) = -\frac{1}{2-q}\left(\int_{\mathcal{U}} \pi(u|x)\log_q \pi(u|x)du - 1\right), \quad \forall x \in \mathcal{X}, \tag{2.2.11}$$

where, in a similar way, $-\frac{1}{2-q}\left(\mathbb{E}[\log_q \pi(u|x)] - 1\right)$ is equivalent for calculation.

2.3 Applications

2.3.1 MV Problem

Consider an investor who enters the capital market with initial wealth x_0 and invests simultaneously in one riskless asset and n risky assets with a finite time horizon. Let r_t be a given deterministic return of the riskless asset at time period t and $e_t = (e_t^1, \cdots, e_t^n)^\top$ the vector of random returns of the n risky assets at period t. We assume that vectors e_t, for $t = 0, 1, \cdots, T$, are statistically independent. The only information known about the random return vector e_t is its first two moments: its mean $\mathbb{E}(e_t) = (\mathbb{E}[e_t^1], \mathbb{E}[e_t^2], \cdots, \mathbb{E}[e_t^n])^\top$ and its covariance $\mathrm{Cov}(e_t) = \mathbb{E}\left[(e_t - \mathbb{E}e_t)(e_t - \mathbb{E}e_t)^\top\right]$. Clearly, $\mathrm{Cov}(e_t)$ is positive semidefinite, that is, $\mathrm{Cov}(e_t) \geqslant 0$. We interpret r_t and each component of e_t as a gross return factor over period t, hence $r_t > 0$. Equivalently, if one uses net returns $\bar{r}_t$, then $r_t = 1 + \bar{r}_t$. This thesis will establish the necessary and sufficient criteria for the solvability of multi-period MV portfolio optimization based on the general theory of LQ problems. To proceed, let x_t be the wealth of the investor at the beginning of the t-th period, and let u_t^i, $i = 1, 2, \cdots, n$, be the amount invested in the i-th risky asset at period t, and x_t be the wealth level in period t. As discussed in Markowitz (1952), Cui et al. (2014a), and Wu et al. (2018), supposing that the investor decides to end the portfolio at time T, then adopts the following MV formulation to guide his investment:

$$(\text{MV}): \quad \min_{\{u_t\}_{t=0}^{T-1}} \text{Var}\left[x_T\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} u_t^\top R u_t\right] \tag{2.3.12a}$$

$$\text{s.t. } \mathbb{E}\left[x_T\right] = d, \tag{2.3.12b}$$

$$x_{t+1} = r_t x_t + P_t u_t, \quad t = 0, \cdots, T-1, \tag{2.3.12c}$$

where $u_t \triangleq (u_t^1, u_t^2, \cdots, u_t^n)^\top$ and $P_t \triangleq (P_t^1, P_t^2, \cdots, P_t^n) = (e_t^1 - r_t, e_t^2 - r_t, \cdots, e_t^n - r_t)$ is the excess return vector. d is a predetermined price level for the expected value of terminal wealth $\mathbb{E}[x_T]$, with $d \geqslant x_0 \prod_{t=0}^{T-1} r_t$, which requires that the target wealth level must exceed the terminal wealth level obtained from fully investing in the risk-free asset.

Remark 2.2. *The quadratic term $u_t^\top R u_t$ is included to model trading frictions or position-regularization and it also ensures strong convexity in the control, which improves the well-posedness and stability of LQ model. The matrix $R \succeq 0$ is taken as an exogenously specified parameter chosen by the modeler; setting $R = \mathbf{0}_{n\times n}$ recovers the classical MV formulation.*

Consider a Lagrange function by introducing Lagrangian multiplier $2\lambda \in \mathbb{R}$ for (2.3.12b):

$$\begin{aligned}
&\text{Var}\left[x_T\right] + 2\lambda\left(\mathbb{E}\left[x_T\right] - d\right) + \mathbb{E}\left[\sum_{t=0}^{T-1} u_t^\top R u_t\right] \\
=&\mathbb{E}\left[(x_T - d)^2 + 2\lambda\left(x_T - d\right)\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} u_t^\top R u_t\right],
\end{aligned}$$

where $\text{Var}\left[x_T\right] = \mathbb{E}\left[x_T^2\right] - \mathbb{E}\left[x_T\right]^2 = \mathbb{E}\left[\left(x_T - d\right)^2\right]$ on the feasible set (2.3.12b). This

further leads to the following equivalent problem (MV-λ_1):

$$\begin{aligned}(\text{MV-}\lambda_1): \quad & \min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[(x_T-(d-\lambda))^2\right]+\mathbb{E}\left[\sum_{t=0}^{T-1} u_t^\top R u_t\right]-\lambda^2 \\ & \text{s.t. } x_{t+1}=r_t x_t+P_t u_t, \quad t=0,\cdots,T-1.\end{aligned} \tag{2.3.13}$$

The aforementioned auxiliary problem can be reformulated as a separable linear quadratic control problem (MV-λ_2), by replacing the state variable x_t with $y_t + \frac{d-\lambda}{\prod_{k=t}^{T-1} r_k}$:

$$\begin{aligned}(\text{MV-}\lambda_2): \quad & \min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}\left[y_T^2\right]+\mathbb{E}\left[\sum_{t=0}^{T-1} u_t^\top R u_t\right] \\ & \text{s.t. } y_{t+1}=r_t y_t+P_t u_t, \quad t=0,\cdots,T-1.\end{aligned} \tag{2.3.14}$$

In certain real-world scenarios, numerous investors, particularly institutions such as pension funds or university endowments, may find it practical to assess cumulative risk over all time periods when considering long-term (or infinite horizon) models, rather than focusing solely on a single future date. This type of problem is frequently encountered in long-term asset management models in which the objective is to incorporate continual risk management and the investor's risk aversion.

$$(\text{MV-}\infty): \quad \min_{u_t} \sum_{t=0}^{\infty} \text{Var}\left[x_t\right]+\mathbb{E}\left[\sum_{t=0}^{\infty} u_t^\top R u_t\right] \tag{2.3.15a}$$

$$\text{s.t. } \mathbb{E}\left[x_t\right]=d_t, \text{ and} \tag{2.3.15b}$$

$$x_{t+1}=r_t x_t+P_t u_t, \tag{2.3.15c}$$

for $t=0,\cdots,\infty$. The level d_t is a predetermined price level for the expected wealth $\mathbb{E}[x_t]$ at time t, with $d_t \geqslant x_0 \prod_{k=0}^{t-1} r_k$. Similarly, we introduce a Lagrangian multiplier

$2\lambda_t \in \mathbb{R}$ for (2.3.15b):

$$\begin{aligned}
&\sum_{t=0}^{\infty} \mathrm{Var}\left[x_t\right] + \sum_{t=0}^{\infty} 2\lambda_t \left(\mathbb{E}\left[x_t\right] - d_t\right) + \mathbb{E}\left[\sum_{t=0}^{\infty} u_t^\top R u_t\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \left(x_t - d_t\right)^2 + 2\lambda_t \left(x_t - d_t\right)\right] + \mathbb{E}\left[\sum_{t=0}^{\infty} u_t^\top R u_t\right],
\end{aligned}$$

where $\mathrm{Var}\left[x_t\right] = \mathbb{E}\left[x_t^2\right] - \mathbb{E}\left[x_t\right]^2 = \mathbb{E}\left[\left(x_t - d_t\right)^2\right]$. This further leads to the following equivalent problem (MV-∞_1):

$$\begin{aligned}
(\text{MV-}\infty_1): \quad &\min_{u_t} \mathbb{E}\left[\sum_{t=0}^{\infty} \left(x_t - \left(d_t - \lambda_t\right)\right)^2 + u_t^\top R u_t - \lambda_t^2\right] \\
&\text{s.t. } x_{t+1} = r_t x_t + P_t u_t, \quad t = 0, \cdots, \infty.
\end{aligned} \tag{2.3.16}$$

The aforementioned auxiliary problem can be reformulated as a separable linear quadratic control problem (MV-∞_2) by replacing the state variable x_t with $y_t + d_t - \lambda_t$:

$$\begin{aligned}
(\text{MV-}\infty_2): \quad &\min_{u_t} \mathbb{E}\left[\sum_{t=0}^{\infty} y_t^2 + u_t^\top R u_t\right] \\
&\text{s.t. } y_{t+1} = r_t y_t + P_t u_t, \quad t = 0, \cdots, \infty.
\end{aligned} \tag{2.3.17}$$

Remark 2.3. *The undiscounted formulation is an infinite horizon model background, while we always add a discount factor $\gamma \in (0,1)$ to ensure the objective remains finite, while the discounted formulation is the one actually used later. Economically, alternative criteria are also common in portfolio problems, such as long-run growth rate or long-run average cost objectives; these can be more appropriate when one is interested in asymptotic wealth growth rather than discounted utility. We do not pursue these alternatives here since our goal is to cast the problem into a discounted LQ framework that admits tractable analysis and serves as a consistent baseline for the subsequent chapters.*

In dynamic decision-making problems, especially those involving trade-offs over time, time inconsistency often arises when future preferences do not align with current optimal plans. This phenomenon has been widely studied and metaphorically described as the "random shoe" behavior of Wall Street, where optimal strategies keep shifting unpredictably (Guo et al. (2017)).

The MV problem is time-inconsistent because it involves a nonlinear function of the square of expected terminal wealth rather than the expected value of a nonlinear function, thus violating the principle of dynamic programming (Bjork and Murgoci (2010)). For the finite horizon case in Ni et al. (2017), an open-loop time-consistent equilibrium control can be constructed by leveraging the dynamic structure of the system. However, in the infinite horizon case, we use the precommitment solution; in other words, the numerical results reported in the research are based on simulation experiments rather than real financial data.

2.3.2 Optimal Liquidation Problem

We follow the setup of Almgren and Chriss (2001) and the example in Hambly et al. (2021) for an optimal liquidation problem with a slight variant. Our aim is to liquidate an amount q of an asset, with security price S_0 at time 0, so that the initial market value of our position is qS_0. The trading decisions are made at discrete time points $t = 0, 1, \cdots, \infty$. Our initial holding is $q_0 = q$ and liquidation at time T for $T \to \infty$ requires $q_\infty = 0$. At each time t our decision is to liquidate an amount u_t of the asset. This process will have two types of price impact. Temporary impact refers to temporary imbalances in supply and demand caused by our trading, leading to temporary price movements away from the equilibrium. Permanent impact means changes in the equilibrium price due to our trading, which remains at least for the life of our liquidation.

We write S_t for the asset price at time t and assume the impacts are linear in the

number of traded shares. Then, the security price of the asset is modeled by

$$S_{t+1} = S_t + \sigma_S Z_{t+1} - \beta_t^{\text{pi}}\, u_t,$$

where the $\sigma_S Z_{t+1}$ term is the exogenous noise on price, which is independent of control to guarantee risky asset price still changes when $u_t = 0$. We have the volatility σ_S and standard normal random variable Z_{t+1}. In addition, $\mathbb{E}(\beta_t^{\text{pi}}) = \beta^{\text{pi}}$ and $\text{Var}(\beta_t^{\text{pi}}) = \sigma^2_{\beta^{\text{pi}}}$ for each $t = 0, 1, \cdots, \infty$ and β_t^{pi} is the permanent price impact parameter. The inventory process q_t records the current holding in the asset at time t. Thus we have

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

Assumption 2.1. *For the infinite liquidation model, we typically assume*

(i) Full liquidation in the limit:

$$q_t \downarrow 0, \quad \sum_{t=0}^{\infty} u_t = q_0.$$

(ii) Square-summability (so variance is finite):

$$\sum_{t=0}^{\infty} u_t^2 < \infty, \quad \sum_{t=1}^{\infty} q_t^2 < \infty, \quad \sum_{t=0}^{\infty} (q_{t+1} u_t)^2 < \infty.$$

(iii) A mild boundary condition:

$$\lim_{T\to\infty} q_T S_T = 0.$$

Therefore, the two-dimensional state process is

$$\begin{bmatrix} S_{t+1} \\ q_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S_t \\ q_t \end{bmatrix} + \begin{bmatrix} -\beta^{\text{pi}} \\ -1 \end{bmatrix} u_t + \left[\begin{bmatrix} -1 \\ 0 \end{bmatrix} u_t \right] \varepsilon_t + \begin{bmatrix} \sigma_S Z_{t+1} \\ 0 \end{bmatrix}, \tag{2.3.18}$$

where $\beta_t^{\rm pi} = \beta^{\rm pi} + \varepsilon_t$ and ε_t is an independent normal random variable with zero mean and variance $\sigma^2_{\beta^{\rm pi}}$. When selling shares, we incur a temporary price impact, parameter $\beta^{\rm ti}$, in that if at time t we trade u_t of our asset, then we obtain the actual price $\tilde{S}_t = S_t - \beta^{\rm ti}\, u_t$ per share, while this temporary effect does not appear in the next security price S_t. Therefore the total trading revenue is $\sum_{t=0}^{\infty} u_t \tilde{S}_t$. The total cost ($TC$) of trading is the difference between the initial book value and the revenue:

$$\begin{aligned}
TC &= qS_0 - \sum_{t=0}^{\infty} u_t \tilde{S}_t \\
&= q_0 S_0 - \sum_{t=0}^{\infty} u_t S_t + \beta^{\rm ti} \sum_{t=0}^{\infty} u_t^2 \\
&= -\sum_{t=1}^{\infty} q_t \left(S_t - S_{t-1}\right) + \beta^{\rm ti} \sum_{t=0}^{\infty} u_t^2 \\
&= -\sigma_S \sum_{t=1}^{\infty} q_t Z_t + \beta^{\rm pi} \sum_{t=1}^{\infty} q_t u_{t-1} + \sum_{t=0}^{\infty} q_{t+1} u_t \varepsilon_t + \beta^{\rm ti} \sum_{t=0}^{\infty} u_t^2 \\
&= -\sigma_S \sum_{t=1}^{\infty} q_t Z_t + \sum_{t=0}^{\infty} q_{t+1} u_t \varepsilon_t + \delta \sum_{t=0}^{\infty} u_t^2 + \frac{\beta^{\rm pi}}{2} q_0^2,
\end{aligned}$$

where $\delta = \beta^{\rm ti} - \frac{\beta^{\rm pi}}{2}$ summarizes the impact and is assumed positive. The second equation substitutes $\tilde{S}_t = S_t - \beta^{\rm ti}\, u_t$; the third equation uses the boundary condition; the fourth equation substitutes $S_t - S_{t-1} = \sigma_S Z_{t+1} - \beta^{\rm pi}_{t-1} u_{t-1} - \varepsilon_{t-1} u_{t-1}$, and in the last equation, we use $\sum_{t=1}^{\infty} q_t u_{t-1} = \frac{1}{2}\left(q_0^2 - \sum_{t=0}^{\infty} u_t^2\right)$. The mean and variance of the total cost of trading are given by

$$\mathbb{E}(TC) = \delta \sum_{t=0}^{\infty} u_t^2 + \frac{\beta^{\rm pi}}{2} q_0^2 = \frac{1}{2} \beta^{\rm pi}\, q^2 + \delta \sum_{t=0}^{\infty} u_t^2,$$

$$\mathrm{Var}(TC) = \sigma_S^2 \sum_{t=0}^{\infty} q_t^2.$$

Next, we generate the following objective function:

$$TC_{LQ} = \min_{u}(\mathbb{E}(TC) + \phi \operatorname{Var}(TC)), \tag{2.3.19}$$

where ϕ is a parameter balancing risk versus return.

Remark 2.4. *Linear price impact is a tractable first-order approximation but can be inaccurate for large orders, does not capture transient impact and resilience, and may generate unrealistic prices. If β_t^{pi} is modeled as Gaussian, it may become negative or excessively large; in practice, one may assume $\beta_t^{pi} > 0$ a.s. (e.g., truncated normal, or lognormal specifications) or enforce a price floor.*

2.3.3 Dynamic Benchmark Tracking Problem

To manage a small basket of sector ETFs and track a benchmark (e.g., SPY) over a short horizon, the objective is to minimize (i) deviation from the benchmark and (ii) trading cost. In practice, the ideal sector exposures that best replicate the benchmark drift over time due to composition and sector changes. We trade to keep up with those drifting targets. This benchmark tracking problem is well known to admit stochastic LQ formulations; see, for example, Yao et al. (2006) about stochastic LQ tracking and later dynamic trading papers that cast portfolio rebalancing with quadratic costs in LQ form (Moallemi and Sağlam (2017) and Gârleanu and Pedersen (2013)). Define the exposure deviation (state), trades (control), and noise term as follows:

- State (dimension = number of sector ETFs): $x_t \in \mathbb{R}^m$ is the exposure deviation at time t, which is actual sector holdings, in shares or exposure units, minus the target exposures that best track the benchmark at t.
- Control: $u_t \in \mathbb{R}^m$ is the trade volume change (buy/sell) in sector indices. Positive means buy.

- The target exposure drift as exogenous noise: let $a_t^* \in \mathbb{R}^m$ be time-varying target exposures that best explain benchmark returns at time t, estimated from data. One-step change $w_t = a_{t+1}^* - a_t^*$ is an unpredictable, zero-mean additive noise to the exposure deviation dynamics.

The dynamics are

$$x_{t+1} = x_t - u_t + w_t,$$

where $A_t = \mathbf{I}_m$, $B_t = -\mathbf{I}_m$, $D_t = \mathbf{I}_m$. We use a quadratic tracking objective with

$$Q_t = \lambda_{\text{step}}\, \sigma_{\text{step}}^2\, \mathbf{I}_m, \quad R_t = \text{diag}\,(c_1, \ldots, c_m)\,, \quad Q_T = \bar{q}\mathbf{I}_m(\bar{q} > 0).$$

Here $\lambda_{\text{step}} > 0$ scales tracking risk via per-step variance σ_{step}^2 ; R_t collects per ETF trading cost weights $c_j > 0$; a large $\bar{q}$ enforces terminal closeness to the benchmark.

Performance Measure. We use the following normalized error to quantify the performance of a given policy C_{policy} ,

$$\text{Normalized error} = \frac{C_{\text{policy}} - C_{\text{optimal}}^*}{C_{\text{policy}}},$$

where C_{policy} represents the sample results obtained from each trial or iteration, and C_{optimal}^* is the optimal cost established for a specific case. We say that the method converges if the normalized error tends to 0 as the iteration index tends to ∞.

Chapter 3

Policy Learning for Perturbance-Wise Linear Quadratic Control Problem

This chapter introduces policy gradient methods tailored to finite horizon multiplicative noise environments. Mirroring the analytical rigor found in Fazel et al. (2018) and Hambly et al. (2021), we expand upon the framework of LQ optimal control, particularly with the RL method, to broader variations with unknown noise distribution. Building upon the three models proposed by classical LQ, the scalar state constrained LQ from Wu et al. (2018), and the unknown noise distribution cases with distributionally robust optimization (DRO) problem from Kim and Yang (2023), we adapt the conceptual framework to accommodate both constrained linear systems and unknown noise distribution systems with both state and control multiplicative noise. This modification extends the model's applicability, diverging from the basic LQ focus to embrace a broader spectrum of system complexities. Moreover, we apply the policy gradient method proof, as established by Hambly et al. (2021), to our enriched model. While Hambly et al. (2021) provided a foundational approach to policy gradient methods, our work takes a significant leap by validating their theoretical results under the new conditions posed by our advanced model. This approach not only tests the robustness of their proofs but also showcases the versatility of the

policy gradient method across varying system dynamics. We have further developed these proofs to align with the intricate landscape of our model, which introduces a novel theoretical dimension to the existing body of work. The synergy between the sophisticated model structure and the robust policy gradient method culminates in a rigorous proof that enhances the theoretical understanding of stochastic control in multiplicative noise.

3.1 Problem Setup

3.1.1 Classical LQ Problem

Based on the LQ problem in Section 2.1, we set a finite time horizon denoted as T and consider the discrete time linear stochastic dynamic system with (2.1.4) for $t = 0, 1, \cdots, T-1$, and

$$x_{t+1} = A_t x_t + B_t u_t + D_t w_t, \ x_0 \sim \mathcal{D}_0. \tag{3.1.1}$$

Note that $w_t \in \mathbb{R}^k$ and (3.1.1) is the mixed noise specialization of the general multiplicative noise model, see Remark 2.1.

We now apply the dynamic programming algorithm to define a sequence of positive definite matrices $\{P_t^*\}_{t=0}^T$ as the solution to the dynamic ARE:

$$P_t^* = A_t^\top P_{t+1}^* A_t - A_t^\top P_{t+1}^* B_t (B_t^\top P_{t+1}^* B_t + R_t)^{-1} B_t^\top P_{t+1}^* A_t + Q_t, \tag{3.1.2}$$

with terminal condition $P_T^* = Q_T$. The matrices $\{P_t^*\}_{t=0}^T$ can be found by solving the ARE iteratively backwards in time.

The optimal control sequence $\{u_t^*\}_{t=0}^{T-1}$ is given by

$$u_t^* = -K_t^* x_t, \tag{3.1.3}$$

where

$$K_t^* = (B_t^\top P_{t+1}^* B_t + R_t)^{-1} B_t^\top P_{t+1}^* A_t. \tag{3.1.4}$$

To identify the optimal solution in the linear feedback form (3.1.3), it is essential to concentrate on the specific class of linear admissible policies in feedback form

$$u_t = -K_t x_t, \quad t = 0, 1, \cdots, T-1. \tag{3.1.5}$$

3.1.2 LQ Problem with Constraints

When x is a scalar state, we can take into account the following set of general control constraints, then find the optimal solution for this case, which is a piecewise linear function, adapted from Wu et al. (2018):

$$\mathcal{C}_t(x_t) = \{u_t \in \mathbb{R}^n \,|H_t u_t \leqslant d_t |\, x_t|\}, \tag{3.1.6}$$

where $H_t \in \mathbb{R}^{s\times n}$ and $d_t \in \mathbb{R}^s$ are given matrices and vectors. In order to ensure the feasibility, we impose the following standing assumption: for each $t = 0, \cdots T-1$, the set $\{K \in \mathbb{R}^n \mid H_t K \leqslant d_t,\ H_t K \leqslant \mathbf{0}_{s\times 1}\}$ is nonempty. The parametrization in (3.1.6) is flexible and can capture a variety of practically relevant constraints. A few representative examples are:

(i) Nonnegativity/nonpositivity. If we choose $H_t = -\mathbf{I}_n$ and $d_t = \mathbf{0}_{n\times 1}$, then (3.1.6) enforces $u_t \geqslant \mathbf{0}_{n\times 1}$(component-wise). Similarly, taking $H_t = \mathbf{I}_n$ and $d_t = \mathbf{0}_{n\times 1}$ yields $u_t \leqslant \mathbf{0}_{n\times 1}$.

(ii) State-dependent box constraints. Suppose we are given vectors $\underline{d}_t, \overline{d}_t \in \mathbb{R}^n$. The requirement $\underline{d}_t |x_t| \leqslant u_t \leqslant \overline{d}_t |x_t|$ can be represented by setting $H_t = \begin{bmatrix} \mathbf{I}_n \\ -\mathbf{I}_n \end{bmatrix}$ and $d_t = \begin{bmatrix} \overline{d}_t \\ -\underline{d}_t \end{bmatrix}$.

(iii) Cone-type constraints. A general conic constraint of the form $H_t u_t \geqslant \mathbf{0}_{s\times 1}$, for a prescribed matrix H_t, is also encompassed by this framework.

(iv) Unconstrained controls. The unconstrained case $u_t \in \mathbb{R}^n$ is recovered by taking $H_t = \mathbf{0}_{s\times n}$ and $d_t = \mathbf{0}_{s\times 1}$ so that (3.1.6) imposes no restriction on u_t.

Before deriving the explicit solution to the problem, it is convenient to introduce the following state-independent set associated with $\mathcal{C}_t(x_t)$: $\mathcal{K}_t = \{K \in \mathbb{R}^n \mid H_t K \leqslant d_t\}$. Additionally, three auxiliary optimization problems are introduced for the interval, as follows:

$$\begin{aligned} &\min_{u\in\mathcal{C}_t(\xi)} p_t(u,\xi,y,z), \\ &\min_{K\in\mathcal{K}_t} \hat{p}_t(K,y,z), \\ &\min_{K\in\mathcal{K}_t} \bar{p}_t(K,y,z), \end{aligned} \tag{3.1.7}$$

where $\xi \in \mathbb{R}$ is a $\mathcal{F}_t$-measurable random variable, and $y, z \in \mathbb{R}_+$ are $\mathcal{F}_{t+1}$-measurable random variables. The corresponding objective functions are defined by

$$\begin{aligned} p_t(u,\xi,y,z) = \mathbb{E}_t\Big[&\xi^2 Q_t + u^\top R_t u + (A_t\xi + B_t u + D_t w_t)^2 \\ &\cdot \big(y\cdot\ \mathbf{1}_{\{A_t\xi+B_t u+D_t w_t\geqslant 0\}} + z\cdot\ \mathbf{1}_{\{A_t\xi+B_t u+D_t w_t<0\}}\big)\Big], \end{aligned} \tag{3.1.8}$$

$$\begin{aligned} \hat{p}_t(K,y,z) = \mathbb{E}_t\Big[&Q_t + K^\top R_t K + (A_t + B_t K + D_t w_t)^2 \\ &\cdot \big(y\cdot\ \mathbf{1}_{\{A_t+B_t K+D_t w_t\geqslant 0\}} + z\cdot\ \mathbf{1}_{\{A_t+B_t K+D_t w_t<0\}}\big)\Big], \end{aligned} \tag{3.1.9}$$

$$\begin{aligned} \bar{p}_t(K,y,z) = \mathbb{E}_t\Big[&Q_t + K^\top R_t K + (A_t - B_t K - D_t w_t)^2 \\ &\cdot \big(y\cdot\ \mathbf{1}_{\{A_t-B_t K-D_t w_t\leqslant 0\}} + z\cdot\ \mathbf{1}_{\{A_t-B_t K-D_t w_t>0\}}\big)\Big]. \end{aligned} \tag{3.1.10}$$

The function $p_t(u,\xi,y,z)$ is the one-step dynamic programming objective; it combines the stage cost at time t with the continuation value at time $t+1$. The two continuation coefficients (y,z) correspond to the piecewise quadratic form of the value function induced by the absolute value state constraint structure. The functions $\hat{p}_t$

and $\bar{p}_t$ are p_t after substituting the linear feedback form $u = -K\xi$, so minimizing them over K yields the backward recursions for $(\hat{P}_t, \bar{P}_t)$.

An explicit solution for the problem can be developed. We introduce two random sequences, $\hat{P}_0, \hat{P}_1, \cdots, \hat{P}_T$ and $\bar{P}_0, \bar{P}_1, \cdots, \bar{P}_T$, defined recursively in reverse as follows:

$$\hat{P}_t = \min_{K_t \in \mathcal{K}_t} \hat{p}_t \left(K_t, \hat{P}_{t+1}, \bar{P}_{t+1} \right), \tag{3.1.11}$$

$$\bar{P}_t = \min_{K_t \in \mathcal{K}_t} \bar{p}_t \left(K_t, \hat{P}_{t+1}, \bar{P}_{t+1} \right), \tag{3.1.12}$$

where function $\hat{p}_t$ and $\bar{p}_t$ are defined, respectively, in (3.1.9) and (3.1.10) for $t = T-1, \cdots, 0$, with the boundary conditions of $\hat{P}_T = \bar{P}_T = Q_T$. It is evident that $\hat{P}_t$ and $\bar{P}_t$ are $\mathcal{F}_t$-measurable random variables. The optimal control policy for the problem at time t is characterized by a linear feedback policy,

$$u_t^*(x_t) = \begin{cases} \hat{K}_t^* x_t, & x_t \geqslant 0, \\ -\bar{K}_t^* x_t, & x_t < 0, \end{cases} \tag{3.1.13}$$

where $\hat{K}_t^*$ and $\bar{K}_t^*$ are defined as

$$\hat{K}_t^* = \arg \min_{K_t \in \mathcal{K}_t} \hat{p}_t \left(K_t, \hat{P}_{t+1}, \bar{P}_{t+1} \right),$$

$$\bar{K}_t^* = \arg \min_{K_t \in \mathcal{K}_t} \bar{p}_t \left(K_t, \hat{P}_{t+1}, \bar{P}_{t+1} \right),$$

where $\{\hat{P}_t\}_{t=0}^T$ and $\{\bar{P}_t\}_{t=0}^T$ are given in (3.1.11) and (3.1.12), respectively, and $\hat{P}_t > 0$ and $\bar{P}_t > 0$ for $t = 0, \cdots, T-1$. Furthermore, the optimal objective value of the problem is

$$x_0^2 \left(\hat{P}_0 \cdot \mathbf{1}_{\{x_0 \geqslant 0\}} + \bar{P}_0 \cdot \mathbf{1}_{\{x_0 < 0\}} \right). \tag{3.1.14}$$

The above theory suggests that the systems represented by (3.1.11) and (3.1.12) serve a function analogous to that of the ARE in the classical LQ optimal problem. Moreover, in the absence of control constraints, these two equations will converge into the single equation presented in (3.1.2).

3.1.3 LQ Problem with Unknown Noise Distribution

Consider the stochastic program

$$J^* = \min_{\pi \in \Pi} \mathbb{E}_{\mathbb{P}} \left[J_\pi \left(x, u | w \right) \right], \tag{3.1.15}$$

where $\mathbb{P}$ is the noise distribution supported on Ω. In applications, the distribution $\mathbb{P}$ cannot be accessed explicitly. Instead, we assume that we observe a finite number M i.i.d. samples from $\mathbb{P}$ and collect them in the training set $\widehat{\Omega}_M = \{\widehat{w}_i\}_{i=1}^M \subseteq \Omega$. Before it is realized, this dataset can be regarded as a random element in Ω^M with distribution $\mathbb{P}^M$.

A data-based control decision $\widehat{u}_M$ is any feasible control rule constructed from the information contained in $\widehat{\Omega}_M$. Its true performance is measured by the expected cost $\mathbb{E}_{\mathbb{P}} \left[J_\pi \left(x, \widehat{u}_M | w \right) \right]$, that is, the expectation with respect to a fresh noise draw w, independent of the training set. Since $\mathbb{P}$ is unknown, we aim to derive high confidence upper bounds on this out-of-sample cost. Concretely, for a random bound $\widehat{J}_M$ and a prescribed significance level β, we would like to ensure that

$$\mathbb{P}^M \left\{ \widehat{\Omega}_M : \mathbb{E}_{\mathbb{P}} \left[J_\pi \left(x, \widehat{u}_M | w \right) \right] \leqslant \widehat{J}_M \right\} \geqslant 1 - \beta. \tag{3.1.16}$$

Here $\widehat{J}_M$ plays the role of a performance certificate, while the probability on the left-hand side quantifies its reliability.

Because it is impossible to evaluate the true out-of-sample performance, we instead seek a data-driven solution with a low certificate and high reliability $\widehat{J}_M$, which are small and reliable in the above sense. A natural first step is to approximate $\mathbb{P}$ by the empirical measure with sample data $\widehat{w}_{t,1}, \cdots, \widehat{w}_{t,M}$ from $\mathbb{P}$:

$$\widehat{\mathbb{P}}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{\widehat{w}_{t,i}}, \tag{3.1.17}$$

where $\delta_{\widehat{w}_i}$ is the Dirac mass at $\widehat{w}_i$. If one were to treat $\widehat{\mathbb{P}}_M$ as the true distribution in the controller design, the resulting policy could be highly sensitive to sampling noise and may perform poorly when $\mathbb{P}$ differs from $\widehat{\mathbb{P}}_M$. To hedge against such misspecification, we adopt the distributionally robust counterpart

$$\widehat{J}_M = \min_{\pi \in \Pi} \max_{\mathbb{Q} \in \widehat{\mathcal{P}}_M} \mathbb{E}_{\mathbb{Q}} \left[J_\pi \left(x, u | w \right) \right], \tag{3.1.18}$$

where $\widehat{\mathcal{P}}_M$ denotes an ambiguity set of probability measures that are deemed plausible given the observed samples. In this chapter, we construct $\widehat{\mathcal{P}}_M$ as a ball centered at the empirical distribution $\widehat{\mathbb{P}}_M$, measured in the Wasserstein metric. We will show that the optimal value $\widehat{J}_M$ and any optimal data-driven control $\widehat{u}_M$ solving this distributionally robust problem enjoy two desirable properties: a finite sample performance guarantee (Theorem 3.2) and asymptotic consistency as the sample size M tends to infinity (Theorem 3.3).

Wasserstein Metric and Ambiguity Set

Let $\mathcal{M}(\Omega)$ denote the set of all probability measures $\mathbb{Q}$ on Ω such that $\mathbb{E}_{\mathbb{Q}}[\|w\|] = \int_\Omega \|w\| \mathbb{Q}(dw) < \infty$.

Definition 4 (Wasserstein metric)**.** *For $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\Omega)$, the 2-Wasserstein metric $W_2 \left(\mathbb{Q}_1, \mathbb{Q}_2 \right)$ is defined as*

$$W_2 \left(\mathbb{Q}_1, \mathbb{Q}_2 \right) = \left(\inf \int_{\Omega \times \Omega} (\|w_1 - w_2\|)^2 \Pi_w \left(dw_1, dw_2 \right) \right)^{1/2},$$

where the infimum is taken over all joint distributions Π_w of w_1 and w_2 whose marginals are $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively.

More generally, one can define the p-Wasserstein metric ($p \geqslant 1$) by using the cost $\|w_1 - w_2\|^p$. In what follows, we restrict attention to $p = 2$. To quantify ambiguity

set around the empirical law $\widehat{\mathbb{P}}_M$, we work with the W_2-neighborhood

$$\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_M) = \left\{\mathbb{Q} \in \mathcal{M}(\Omega) : W_2(\widehat{\mathbb{P}}_M, \mathbb{Q}) \leqslant \varepsilon_r\right\}.$$

Given a confidence parameter $\beta \in (0,1)$, we calibrate with the radius ε_r so that

$$\mathbb{P}^M\left\{\mathbb{P} \in \mathbb{B}_{\varepsilon_r}(\widehat{\mathbb{P}}_M)\right\} \geqslant 1-\beta.$$

In other words, $\mathbb{B}_{\varepsilon_r}(\widehat{\mathbb{P}}_M)$ is a W_2-ball centered at the empirical measure that, with probability at least $1-\beta$ over the draw of the M samples, contains the unknown data distribution. In particular, for any confidence $1-\beta > 0$, the goal is to choose an ambiguity ball radius ε_r so that the true distribution in this ball has confidence $1-\beta$.

We now impose a mild regularity assumption on the noise distribution.

Assumption 3.1 (Light-tailed distribution)**.** *There exists an exponent $a > 1$ such that*

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left(\|w\|^a\right)\right] = \int_\Omega \exp\left(\|w\|^a\right)\mathbb{P}(dw) < \infty.$$

Assumption 3.1 essentially requires the tail of the distribution $\mathbb{P}$ to decay at an exponential rate. This process ensures that standard measure concentration inequalities apply, allowing us to control the deviation of empirical quantities from their expectations with high probability.

Theorem 3.1 (Measure concentration)**.** *If Assumption 3.1 holds, we have*

$$\mathbb{P}^M\left\{W_2(\mathbb{P}, \widehat{\mathbb{P}}_M) \geqslant \varepsilon_r\right\} \leqslant \begin{cases} c_1 \exp\left(-c_2 M \varepsilon_r^{\max\{m,2\}}\right) & \text{if } \varepsilon_r \leqslant 1, \\ c_1 \exp\left(-c_2 M \varepsilon_r^a\right) & \text{if } \varepsilon_r > 1, \end{cases}$$

for all $M \geqslant 1$, $m \neq 2$, and $\varepsilon > 0$, where c_1, c_2 are positive constants, while c_2 depends on the p-norm.

The proofs of Theorem 3.1 are taken from Fournier and Guillin (2015).

Using Theorem 3.1, we can estimate the radius of the smallest Wasserstein ball that contains $\mathbb{P}$ with confidence $1-\beta$ for some prescribed $\beta \in (0,1)$ as:

$$\varepsilon_M(\beta) = \begin{cases} \left(\frac{\log\left(c_1\beta^{-1}\right)}{c_2 M}\right)^{1/\max\{m,2\}} & \text{if } M \geqslant \frac{\log\left(c_1\beta^{-1}\right)}{c_2}, \\ \left(\frac{\log\left(c_1\beta^{-1}\right)}{c_2 M}\right)^{1/a} & \text{if } M < \frac{\log\left(c_1\beta^{-1}\right)}{c_2}. \end{cases}$$

Note that the Wasserstein ball with radius $\varepsilon_M(\beta)$ can thus be viewed as a confidence set for the unknown true distribution, as in statistical testing (Mohajerin Esfahani and Kuhn (2018) and Wainwright (2019)). We now connect the above construction to the distributionally robust control problem (3.1.18) introduced earlier.

Theorem 3.2 (Finite sample guarantee)**.** *Let $\beta \in (0,1)$ be fixed and define the ambiguity set $\widehat{\mathcal{P}}_M = \mathbb{B}_{\varepsilon_M(\beta)}(\widehat{\mathbb{P}}_M)$. Suppose Assumption 3.1 holds and let $\widehat{J}_M$ and $\widehat{u}_M$ denote, respectively, the optimal value and an optimal policy of the distributionally robust problem (3.1.18). Then, the finite sample guarantee (3.1.16) holds.*

We finally discuss the asymptotic behavior as the number of samples grows.

Theorem 3.3 (Asymptotic consistency)**.** *Let β_M be a sequence in $(0,1)$ such that $\sum_{M=1}^{\infty} \beta_M < \infty$ and $\lim_{M\to\infty} \varepsilon_M(\beta_M) = 0$ for $M \in \mathbb{N}$. For each M, define the ambiguity set $\widehat{\mathcal{P}}_M = \mathbb{B}_{\varepsilon_M(\beta_M)}(\widehat{\mathbb{P}}_M)$ and let $\widehat{J}_M$ and $\widehat{u}_M$ represent the optimal value and an optimal policy of the problem (3.1.18). Under Assumption 3.1, one can show that $\widehat{J}_M$ goes down to J^* as M goes to ∞, where J^* is the optimal value of (3.1.15).*

The proof of Theorem 3.2 and Theorem 3.3 is based on Mohajerin Esfahani and Kuhn (2018).

Solve Minimax LQ Problem with Wasserstein Penalty

When tackling the distributionally robust optimization problem (3.1.18), it is useful to note that a direct dynamic programming approach over distributions is typically computationally prohibitive in high dimensions in contrast to Riccati-type recursions for classical LQ problems. Following the regularization approach in the literature (Kim and Yang (2023)), we instead modify the cost functional by introducing a Wasserstein penalty term.

For a given policy π, define the penalized cost

$$J_\pi^W(x,u|w) = \mathbb{E}_\pi\left[x_T^\top Q_T x_T + \sum_{t=0}^{T-1}\left(x_t^\top Q_t x_t + u_t^\top R_t u_t - \lambda W_2(\widehat{\mathbb{P}}_M, \mathbb{Q})^2\right)\middle| x_0 = x\right],$$

where $\lambda > 0$ is the penalty parameter. By varying λ, we can ascertain how conservative the resulting control policy is: larger values of λ penalize distributions that are far from the empirical law more heavily. The corresponding penalized minimax problem is

$$\min_{\pi\in\Pi}\max_{\mathbb{Q}\in\widehat{\mathcal{P}}_M}\mathbb{E}_\mathbb{Q}\left[J_\pi^W(x,u|w)\right], \tag{3.1.19}$$

which we use as a tractable surrogate for the original DRO formulation (3.1.18). Let π^* denote an optimal policy of (3.1.19) for a given λ, and define

$$V(x;\lambda) = \inf_{\pi\in\Pi}\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_M} J_\pi^W(x,u|w)$$

as the optimal value of the penalized minimax problem. Then, for a suitable choice of λ^*, one can show a guaranteed cost property of the form

$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_M}\left[J_\pi^W(x,u|w)\,|\pi^*(\lambda^*)\right] \leqslant \lambda^*{\varepsilon_M}^2(\beta) + V(x;\lambda^*),$$

where $\varepsilon_M(\beta)$ is the ambiguity radius defined earlier. To analyze this penalized prob-

lem over a finite horizon, we introduce the value function

$$V_t(x) = \inf_{\pi\in\Pi} \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_M} \mathbb{E}_{\mathbb{Q}} \left[\sum_{s=t}^{T-1} \left(x_s^\top Q_s x_s + u_s^\top R_s u_s - \lambda W_2(\widehat{\mathbb{P}}_M, \mathbb{Q})^2 \right) + x_T^\top Q_T x_T \middle| x_t = x \right],$$

which represents the minimal worst case expected cost-to-go from stage t onward, given $x_t = x$. By construction, the scalar value $V(x;\lambda)$ can be identified with $V_0(x;\lambda)/T$. The dynamic programming principle then yields the recursion

$$V_t(x) = x^\top Q x + \inf_u \sup_{\mathbb{Q}\in\mathcal{M}(\Omega)} \left[u^\top R u - \lambda W_2(\widehat{\mathbb{P}}_M, \mathbb{Q})^2 + \int_\Omega V_{t+1}(A_t x + B_t u + D_t w)\mathbb{Q}(dw) \right]$$

for $t = 0, \cdots, T-1$, and $V_T(x) = x_t^\top Q_T x_t$. The inner maximization over probability measures $\mathbb{Q}$ is infinite-dimensional and thus not amenable to direct numerical treatment. To obtain a tractable reformulation, we invoke Kantorovich duality for the Wasserstein metric (Kim and Yang (2023) and Beiglböck and Pratelli (2012)), and arrive at

$$V_t(x) = x^\top Q x + \inf_u \left[u^\top R u + \frac{1}{M} \sum_{i=1}^{M} \sup_{w\in\Omega} \{ V_{t+1}(A_t x + B_t u + D_t w) - \lambda \|\widehat{w}_{t,i} - w\|^2 \} \right], \tag{3.1.20}$$

which replaces the maximization over measures by a finite sum of pointwise optimization problems over w. We denote the empirical mean and covariance of the noise at stage t by

$$\bar{w}_t = \mathbb{E}_{\widehat{\mathbb{P}}_M}[w_t] \text{ and } \Sigma_t^w = \mathbb{E}_{\widehat{\mathbb{P}}_M}\left[w_t w_t^\top\right]. \tag{3.1.21}$$

In the subsequent Lemma 3.1, we derive an explicit solution to the minimax problem associated with (3.1.19) under the following condition on the penalty parameter.

Assumption 3.2. *The penalty parameter satisfies $\lambda > \bar{\lambda}_t$ for all $t \geqslant 1$, where $\bar{\lambda}_t$ is the maximum eigenvalue of $D_t^\top P_t D_t$.*

This result ensures that the resulting quadratic forms are strictly concave in the disturbance variable, which will be crucial for solving the inner maximization in closed form.

Lemma 3.1. *Suppose that*

$$V_{t+1}(x) = x^\top P_{t+1} x + 2r_{t+1}^\top x + q_{t+1}$$

for some nonnegative $P_{t+1} \in \mathbb{R}^{m\times m}$, $r_{t+1} \in \mathbb{R}^m$, *and* $q_{t+1} \in \mathbb{R}$. *We further assume that the penalty parameter satisfies* $\lambda > \bar{\lambda}_{t+1}$, *where* $\bar{\lambda}_{t+1}$ *is the maximum eigenvalue of* $D_t^\top P_{t+1} D_t$. *Let* $\Phi(w) = V_{t+1}(A_t x + B_t u + D_t w) - \lambda \|\widehat{w}_{t,i} - w\|^2$ *and find* $\frac{\partial \Phi(w)}{\partial w} = 0$. *Then, the inner maximization problem* $\sup_{w\in\Omega} \Phi(w)$ *in (3.1.20) has a unique maximizer* $w_t^* = (w_{t,1}^*, \cdots, w_{t,M}^*)$, *defined as*

$$w_{t,i}^* = \Delta_t'(D_t^\top P_{t+1}(A_t x + B_t u) + D_t^\top r_{t+1} + \lambda \widehat{w}_{t,i}), \tag{3.1.22}$$

where $\Delta_t' = \left(\lambda \mathbf{I}_k - D_t^\top P_{t+1} D_t\right)^{-1}$. *Substituting (3.1.21) and (3.1.22) into (3.1.20), we can obtain the optimal value function with dynamic programming principle*

$$\begin{aligned} V_t(x) = {} & x^2 \left(Q_t + A_t^\top P_{t+1} \Delta_t A_t\right) + 2x \left(A_t \Delta_t r_{t+1} + \lambda A_t P_{t+1} D_t \Delta_t' \bar{w}_t\right) \\ & + \inf_u \left\{u^\top (R_t + B_t^\top P_{t+1} \Delta_t B_t) u + 2u^\top \left[(B_t^\top P_{t+1} \Delta_t A_t) x + B_t^\top \Delta_t r_{t+1} + \lambda B_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t\right]\right\} \\ & + r_{t+1}^\top D_t \Delta_t' (D_t^\top r_{t+1} + 2\lambda \bar{w}_t) + \mathrm{Tr}\left(\lambda^2 \Delta_t' \Sigma_t^w - \lambda \Sigma_t^w\right) + q_{t+1}, \end{aligned}$$

where $\Delta_t = \mathbf{I}_m + D_t \Delta_t' D_t^\top P_t$. *One can find that the outer minimization problem in (3.1.20) has a unique minimizer*

$$u^* = -K_t^* x - L_t^*,$$

where

$$\begin{aligned} K_t^* &= (R + B_t^\top P_{t+1} \Delta_t B_t)^{-1} B_t^\top P_{t+1} \Delta_t A_t, \\ L_t^* &= (R + B_t^\top P_{t+1} \Delta_t B_t)^{-1} (B_t^\top \Delta_t r_{t+1} + \lambda B_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t). \end{aligned}$$

Thus, the explicit derivation with this linear structure yields the following Riccati equation:

$$
\begin{aligned}
P_t &= Q_t + A_t^\top P_{t+1}\Delta_t A_t - A_t^\top \Delta_t P_{t+1} B_t (R_t + B_t^\top P_{t+1}\Delta_t B_t)^{-1} \\
&\quad B_t^\top P_{t+1}\Delta_t A_t, \\
r_t &= A_t \Delta_t r_{t+1} + \lambda A_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t - A_t^\top \Delta_t P_{t+1} B_t (R_t \\
&\quad + B_t^\top P_{t+1}\Delta_t B_t)^{-1} (B_t^\top \Delta_t r_{t+1} + \lambda B_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t), \\
q_t &= q_{t+1} + r_{t+1}^\top D_t \Delta_t' (D_t^\top r_{t+1} + 2\lambda \bar{w}_t) \\
&\quad + \operatorname{Tr}\left(\lambda^2 \Delta_t' \Sigma_t^w - \lambda \Sigma_t^w\right) - (B_t^\top \Delta_t r_{t+1} + \lambda B_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t)^\top \\
&\quad (R_t + B_t^\top P_{t+1}\Delta_t B_t)^{-1} (B_t^\top \Delta_t r_{t+1} + \lambda B_t^\top P_{t+1} D_t \Delta_t' \bar{w}_t)
\end{aligned}
$$

with the terminal conditions $P_T = Q_T$, $r_T = \mathbf{0}_{m\times 1}$, *and* $q_T = 0$.

Suppose that Assumption 3.2 holds. Then, the matrices P_t are well-defined and the value function can be expressed as

$$V_t(x) = x^\top P_t x + 2r_t^\top x + q_t, \quad t = 0, \cdots, T.$$

Furthermore, the regularized problem (3.1.19) in the finite horizon case has a unique optimal policy, defined as

$$u^* = -K_t^* x - L_t^*, \quad t = 0, \cdots, T-1.$$

3.2 Gradient Dynamics and Convergence Analysis

In the gradient analysis, we treat the noise contribution via its empirical moments $\bar{w}_t$ and Σ_t^w as fixed (data-driven) quantities and perform the minimization over u while deferring the maximization over w. By the envelope theorem, since these moments correspond to the unique maximizer of the inner problem, their dependence on the

control K does not affect the gradient. Consequently, $\bar{w}_t$ and Σ_t^w act as constants in the Bellman expansion and do not introduce implicit derivatives in the cost gradient.

We represent an affine controller $u_t = -K_t x_t - L_t$ as a single linear map by augmenting the state with a constant 1, fixing the policy convention explicitly to avoid sign errors:

$$u_t = -\widehat{K}_t \widehat{x}_t, \quad \widehat{K}_t = \begin{bmatrix} K_t & L_t \end{bmatrix}, \quad \widehat{x}_t = \begin{bmatrix} x_t \\ 1 \end{bmatrix}.$$

Embedding the original number into augmented matrices yields

$$\widehat{A}_t = \begin{bmatrix} A_t & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{1\times m} & 1 \end{bmatrix}, \quad \widehat{B}_t = \begin{bmatrix} B_t \\ \mathbf{0}_{1\times n} \end{bmatrix}, \quad \widehat{D}_t = \begin{bmatrix} D_t \\ \mathbf{0}_{1\times k} \end{bmatrix}, \quad \widehat{Q}_t = \begin{bmatrix} Q_t & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{1\times m} & 0 \end{bmatrix}.$$

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

$$C(\widehat{K}) = \mathbb{E}\left[\sum_{t=0}^{T-1} \left(\widehat{x}_t^\top \widehat{Q}_t \widehat{x}_t + u_t^\top R_t u_t\right) + \widehat{x}_T^\top \widehat{Q}_T \widehat{x}_T\right], \tag{3.2.23}$$

subject to the dynamics

$$\widehat{x}_{t+1} = \widehat{A}_t \widehat{x}_t + \widehat{B}_t u_t + \widehat{D}_t w_t.$$

Recall that $\widehat{K}^*$ is the optimal policy for the problem

$$\widehat{K}^* = \arg\min_{\widehat{K}} C(\widehat{K}).$$

Assumption 3.3 (Initial State and Noise Process)**.** *We make the following assumptions:*

(i) Initial state: The initial state satisfies $x_0 \sim \mathcal{D}_0$, where $\mathbb{E}\big[\widehat{x}_0 \widehat{x}_0^\top\big]$ is positive definite.

(ii) Noise process: The sequence $\{w_t\}_{t=0}^{T-1}$ *consists of i.i.d. random vectors that are independent of* x_0*. Their empirical mean and covariance are denoted by*

$$\bar{w}_t = \mathbb{E}_{\widehat{\mathbb{P}}_M}[w_t], \quad \Sigma_t^w = \mathbb{E}_{\widehat{\mathbb{P}}_M}[w_t w_t^\top],$$

where $\Sigma_t^w \succ 0$ *for all* $t = 0, 1, \ldots, T-1$.

We next introduce several quantities that will be used in our analysis. Specifically, we define $\underline{\sigma}_X$ as the uniform lower bound on the smallest singular values of the state covariance matrices $\mathbb{E}[\widehat{x}_t \widehat{x}_t^\top]$. Similarly, $\underline{\sigma}_R$ and $\underline{\sigma}_Q$ denote the smallest singular values of the control and state cost matrices, respectively:

$$\underline{\sigma}_X = \min_t \ \sigma_{\min}\big(\mathbb{E}[\widehat{x}_t \widehat{x}_t^\top]\big), \tag{3.2.24a}$$

$$\underline{\sigma}_R = \min_t \ \sigma_{\min}(R_t), \tag{3.2.24b}$$

$$\underline{\sigma}_Q = \min_t \ \sigma_{\min}(\widehat{Q}_t). \tag{3.2.24c}$$

By Assumption 3.3, it follows that $\underline{\sigma}_R > 0$ and $\underline{\sigma}_Q > 0$.

3.2.1 Explicit Characterization of the Policy Gradient

We study an exact gradient descent scheme for computing the optimal feedback policy. The update rule is given by

$$\widehat{K}_t^{n+1} = \widehat{K}_t^n - \eta \nabla_t C(\widehat{K}^n), \quad \forall\ 0 \leqslant t \leqslant T-1, \tag{3.2.25}$$

where n denotes the iteration index, $\eta > 0$ is the step size, and $\nabla_t C(\widehat{K}) = \frac{\partial C(\widehat{K})}{\partial \widehat{K}_t}$ is the gradient of the cost function with respect to $\widehat{K}_t$.

We write $\nabla C(\widehat{K}) = (\nabla_0 C(\widehat{K}), \cdots, \nabla_{T-1} C(\widehat{K}))$ for compactness. Let $\{\widehat{x}_t\}_{t=0}^T$ denote the trajectory under the current policy $\widehat{K}$. Define the state mean and covariance at time t as

$$\mu_t = \mathbb{E}\left[\widehat{x}_t^\top\right] \text{ and } \Sigma_t = \mathbb{E}\left[\widehat{x}_t \widehat{x}_t^\top\right], \quad t = 0, 1, \cdots, T. \tag{3.2.26}$$

We further define the aggregated quantities over the horizon: a matrix $\Sigma_{\widehat{K}}$ as the sum of Σ_t,

$$\Sigma_{\widehat{K}} = \sum_{t=0}^{T} \Sigma_t = \mathbb{E}\left[\sum_{t=0}^{T} \widehat{x}_t \widehat{x}_t^\top\right], \tag{3.2.27}$$

and a matrix $\mu_{\widehat{K}}$ as the sum of μ_t,

$$\mu_{\widehat{K}} = \sum_{t=0}^{T} \mu_t = \mathbb{E}\left[\sum_{t=0}^{T} \widehat{x}_t^\top\right]. \tag{3.2.28}$$

In the finite time horizon setting, define $\widehat{P}_t^{\widehat{K}}$, $\widehat{r}_t^{\widehat{K}}$ and $\widehat{q}_t^{\widehat{K}}$ as the solution to the backward recursions

$$\begin{aligned}
\widehat{P}_t^{\widehat{K}} &= \widehat{Q}_t + \widehat{K}_t^\top R_t \widehat{K}_t + (\widehat{A}_t - \widehat{B}_t \widehat{K}_t)^\top \widehat{P}_{t+1}^{\widehat{K}} (\widehat{A}_t - \widehat{B}_t \widehat{K}_t), \\
\widehat{r}_t^{\widehat{K}} &= (\widehat{A}_t - \widehat{B}_t \widehat{K}_t)^\top \widehat{r}_{t+1}^{\widehat{K}} + (\widehat{A}_t - \widehat{B}_t \widehat{K}_t)^\top \widehat{P}_{t+1}^{\widehat{K}} \widehat{D}_t \bar{w}_t, \\
\widehat{q}_t^{\widehat{K}} &= \widehat{q}_{t+1}^{\widehat{K}} + 2(\widehat{r}_{t+1}^{\widehat{K}})^\top \widehat{D}_t \bar{w}_t + \mathrm{Tr}(\widehat{D}_t^\top \widehat{P}_{t+1}^{\widehat{K}} \widehat{D}_t \Sigma_t^w),
\end{aligned} \tag{3.2.29}$$

for $t = 0, 1, \cdots, T-1$ with terminal conditions

$$\widehat{P}_T^{\widehat{K}} = \begin{bmatrix} P_T & \mathbf{0}_{m\times 1} \\ \mathbf{0}_{1\times m} & 0 \end{bmatrix}, \quad \widehat{r}_T^{\widehat{K}} = \begin{bmatrix} r_T \\ 0 \end{bmatrix}, \quad \widehat{q}_T^{\widehat{K}} = q_T,$$

where $P_T = Q_T$, $r_T = \mathbf{0}_{m\times 1}$, and $q_T = 0$. To simplify, we remove $\widehat{K}$-superscript when there is no confusion. Then the cost of $\widehat{K}$ can be rewritten as

$$C(\widehat{K}) = \mathbb{E}_{x_0 \sim \mathcal{D}_0}\left[\widehat{x}_0^\top \widehat{P}_0 \widehat{x} + 2\widehat{r}_0^\top \widehat{x}_0 + \widehat{q}_0\right].$$

To see this,

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{x}_0^\top \widehat{P}_0 \widehat{x} + 2\widehat{r}_0^\top \widehat{x}_0 + \widehat{q}_0\right] \\
&= \mathbb{E}\Big[\widehat{x}_0^\top \widehat{Q}_0 \widehat{x}_0 + \widehat{x}_0^\top \widehat{K}_0^\top R_0 \widehat{K}_0 \widehat{x} + \widehat{x}_0^\top (\widehat{A}_0 - \widehat{B}_0 \widehat{K}_0)^\top \widehat{P}_1 (\widehat{A}_0 - \widehat{B}_0 \widehat{K}_0)\widehat{x}_0 \\
&\quad + 2\widehat{r}_1^\top (\widehat{A}_t - \widehat{B}_t \widehat{K}_t)\widehat{x}_0 + 2\bar{w}_t^\top \widehat{D}_t^\top \widehat{P}_1 (\widehat{A}_t - \widehat{B}_t \widehat{K}_t)\widehat{x}_0 \\
&\quad + \widehat{q}_1 + 2\widehat{r}_1^\top \widehat{D}_0 \bar{w}_0 + \mathrm{Tr}(\widehat{D}_0^\top \widehat{P}_1 \widehat{D}_0 \Sigma_0^w)\Big] \\
&= \mathbb{E}\left[\widehat{x}_0^\top \widehat{Q}_0 \widehat{x}_0 + u_0^\top R_0 u_0 + \widehat{x}_1^\top \widehat{P}_1 \widehat{x}_1 + 2\widehat{r}_1^\top \widehat{x}_1 + \widehat{q}_1\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \left(\widehat{x}_t^\top \widehat{Q}_t \widehat{x}_t + u_t^\top R_t u_t\right) + \widehat{x}_T^\top Q_T \widehat{x}_T\right].
\end{aligned}
$$

Lemma 3.2 (Gradient Representation)**.** *Define for each* $t = 0, 1, \ldots, T-1$,

$$
\begin{aligned}
E_t &= (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)\widehat{K}_t - \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t, \\
F_t &= \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{D}_t \bar{w}_t + \widehat{B}_t^\top \widehat{r}_{t+1}.
\end{aligned}
\tag{3.2.30}
$$

Now we have the following representation of the gradient term with respect to the policy $\widehat{K}_t$,

$$
\nabla_t C(\widehat{K}) = 2E_t \Sigma_t - 2F_t \mu_t.
$$

Proof. We illustrate the derivation for $t = 0$, and the argument extends identically to all $t = 0, 1, \ldots, T-1$. By definition,

$$
\begin{aligned}
\nabla_0 C(\widehat{K}) &= \frac{\partial C(\widehat{K})}{\partial \widehat{K}_0} \\
&= \mathbb{E}\left[2R_0 \widehat{K}_0 \widehat{x}_0 \widehat{x}_0^\top - 2\widehat{B}_0^\top \widehat{P}_1 (\widehat{A}_0 - \widehat{B}_0 \widehat{K}_0)\widehat{x}_0 \widehat{x}_0^\top - 2\widehat{B}_0^\top \widehat{r}_1 \widehat{x}_0^\top - 2\widehat{B}_0^\top \widehat{P}_1^\top \widehat{D}_0 \bar{w}_0 \widehat{x}_0^\top\right] \\
&= 2E_0 \mathbb{E}\left[\widehat{x}_0 \widehat{x}_0^\top\right] - 2F_0 \mathbb{E}\left[\widehat{x}_0^\top\right] \\
&= 2E_0 \Sigma_0 - 2F_0 \mu_0.
\end{aligned}
$$

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

$$\begin{aligned}\nabla_t C(\widehat{K}) &= 2\left((R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t)\widehat{K}_t - \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{A}_t\right)\mathbb{E}\left[\widehat{x}_t\widehat{x}_t^\top\right] \\ &\quad - \left(\widehat{B}_t^\top \widehat{P}_{t+1}\widehat{D}_t\bar{w}_t + \widehat{B}_t^\top \widehat{r}_{t+1}\right)\mathbb{E}\left[\widehat{x}_t^\top\right] \\ &= 2E_t\mathbb{E}\left[\widehat{x}_t\widehat{x}_t^\top\right] - 2F_t\mathbb{E}\left[\widehat{x}_t^\top\right] \\ &= 2E_t\Sigma_t - 2F_t\mu_t,\end{aligned}$$

where the expectation $\mathbb{E}$ is taken with respect to both initial distribution $x_0 \sim \mathcal{D}_0$ and noises w. □

In the following lemma, we establish the analytical form of the advantage function, which serves as the foundation for subsequent results. Building upon this, Lemma 3.4 shows that for any given policy $\widehat{K}$, the optimality gap $C(\widehat{K}) - C(\widehat{K}^*)$ is bounded by the sum of the magnitude of the gradient $\nabla_t C(\widehat{K})$ for $t = 0, 1, \cdots, T-1$.

Let us start with a useful result for the value function. The value function $V_{\widehat{K}}(\widehat{x}, \tau)$ for $\tau = 0, 1, \cdots, T-1$ can be defined as

$$V_{\widehat{K}}(\widehat{x}, \tau) = \mathbb{E}\left[\sum_{t=\tau}^{T-1}\left(\widehat{x}_t^\top \widehat{Q}_t\widehat{x}_t + u_t^\top R_t u_t\right) + \widehat{x}_T^\top \widehat{Q}_T\widehat{x}_t \middle| \widehat{x}_\tau = x\right] = \widehat{x}^\top \widehat{P}_\tau\widehat{x} + 2\widehat{r}_\tau\widehat{x} + \widehat{q}_\tau,$$

with terminal condition

$$V_{\widehat{K}}(\widehat{x}, T) = \widehat{x}^\top \widehat{Q}_T\widehat{x}.$$

We then define the Q-function, $Q_{\widehat{K}}(\widehat{x}, u, \tau)$ for $\tau = 0, 1, \cdots, T-1$ as

$$Q_{\widehat{K}}(\widehat{x}, u, \tau) = \widehat{x}^\top \widehat{Q}_\tau\widehat{x} + u^\top R_\tau u + \mathbb{E}_{w_\tau}\left[V_{\widehat{K}}\left(\widehat{A}_\tau\widehat{x} + \widehat{B}_\tau u + \widehat{D}_\tau w_\tau, \tau + 1\right)\right],$$

and the advantage function

$$A_{\widehat{K}}(\widehat{x}, u, \tau) = Q_{\widehat{K}}(\widehat{x}, u, \tau) - V_{\widehat{K}}(\widehat{x}, \tau).$$

Note that $C(\widehat{K}) = \mathbb{E}_{x_0\sim\mathcal{D}_0}[V(\widehat{x}_0, 0)]$. We can then write the difference of value functions between $\widehat{K}$ and $\widehat{K}'$ in terms of advantage functions in Lemma 3.3.

Lemma 3.3. *Assume $\widehat{K}$ and $\widehat{K}'$ have finite costs. Denote $\{\widehat{x}'_t\}_{t=0}^{T}$ and $\{u'_t\}_{t=0}^{T-1}$ as the state and control sequences of a single trajectory generated by $\widehat{K}'$ starting from $\widehat{x}'_0 = \widehat{x}_0 = \widehat{x}$, then*

$$V_{\widehat{K}'}(\widehat{x}, 0) - V_{\widehat{K}}(\widehat{x}, 0) = \mathbb{E}_w \left[\sum_{t=0}^{T-1} A_{\widehat{K}} \left(\widehat{x}'_t, u'_t, t \right) \right], \tag{3.2.31}$$

and $A_{\widehat{K}}(\widehat{x}, -\widehat{K}'_\tau \widehat{x}, \tau) = 2\widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top E_\tau \widehat{x} + \widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top (R_\tau + \widehat{B}_\tau^\top \widehat{P}_{\tau+1} \widehat{B}_\tau)(\widehat{K}'_\tau - \widehat{K}_\tau)\widehat{x} - 2\widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top F_\tau$, where E_τ and F_τ are defined in (3.2.30).

Proof. Denote by $c'_t(\widehat{x})$ the cost generated by $\widehat{K}'$ with a single trajectory starting from $\widehat{x}'_0 = \widehat{x}_0 = \widehat{x}$. That is, $c'_t(\widehat{x}) = (\widehat{x}'_t)^\top \widehat{Q}_t \widehat{x}'_t + (u'_t)^\top R_t u'_t, t = 0, 1, \cdots, T-1$, and $c'_T(\widehat{x}) = (\widehat{x}'_T)^\top Q_T \widehat{x}'_T$, with $u'_t = -\widehat{K}'_t \widehat{x}'_t$, $\widehat{x}'_{t+1} = \widehat{A}_t \widehat{x}'_t + \widehat{B}_t u'_t + \widehat{D}_t w_t$, $\widehat{x}_0 = \widehat{x}$. Therefore,

$$\begin{aligned}
V_{\widehat{K}'}(\widehat{x}, 0) - V_{\widehat{K}}(\widehat{x}, 0) &= \mathbb{E}_w \left[\sum_{t=0}^{T} c'_t(\widehat{x}) \right] - V_{\widehat{K}}(\widehat{x}, 0) \\
&= \mathbb{E}_w \left[\sum_{t=0}^{T} \left(c'_t(\widehat{x}) + V_{\widehat{K}} \left(\widehat{x}'_t, t \right) - V_{\widehat{K}} \left(\widehat{x}'_t, t \right) \right) \right] - V_{\widehat{K}}(\widehat{x}, 0) \\
&= \mathbb{E}_w \left[\sum_{t=0}^{T-1} \left(c'_t(\widehat{x}) + V_{\widehat{K}} \left(\widehat{x}'_{t+1}, t+1 \right) - V_{\widehat{K}} \left(\widehat{x}'_t, t \right) \right) \right] \\
&= \mathbb{E}_w \left[\sum_{t=0}^{T-1} \left(Q_{\widehat{K}} \left(\widehat{x}'_t, u'_t, t \right) - V_{\widehat{K}} \left(\widehat{x}'_t, t \right) \right) \middle| \widehat{x}_0 = \widehat{x} \right] \\
&= \mathbb{E}_w \left[\sum_{t=0}^{T-1} A_{\widehat{K}} \left(\widehat{x}'_t, u'_t, t \right) \middle| \widehat{x}_0 = \widehat{x} \right],
\end{aligned}$$

where the third equality holds since $c'_T(\widehat{x}) = V_{\widehat{K}} \left(\widehat{x}'_t, T \right)$ with the same single trajec-

tory. For $u = -\widehat{K}'_\tau \widehat{x}$, we have

$$
\begin{aligned}
&A_{\widehat{K}}(\widehat{x}, -\widehat{K}'_\tau \widehat{x}, \tau) \\
&= Q_{\widehat{K}}(\widehat{x}, -\widehat{K}'_\tau \widehat{x}, \tau) - V_{\widehat{K}}(\widehat{x}, \tau) \\
&= \widehat{x}^\top (\widehat{Q}_\tau + (\widehat{K}'_\tau)^\top R_\tau \widehat{K}'_\tau)\widehat{x} + \mathbb{E}_{w_\tau}\left[V_{\widehat{K}}((\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} + \widehat{D}_\tau w_\tau, \tau+1)\right] - V_{\widehat{K}}(\widehat{x}, \tau) \\
&= \widehat{x}^\top (\widehat{Q}_\tau + (\widehat{K}'_\tau)^\top R_\tau \widehat{K}'_\tau)\widehat{x} + \mathbb{E}\Big[\widehat{x}^\top (\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} + w_\tau^\top \widehat{D}_\tau^\top \widehat{P}_{\tau+1}\widehat{D}_\tau w_\tau \\
&\quad + 2w_\tau^\top \widehat{D}_\tau^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} + 2\widehat{r}_{\tau+1}^\top((\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} + \widehat{D}_\tau w_\tau) + \widehat{q}_{\tau+1}\Big] \\
&\quad - \mathbb{E}\left[\widehat{x}^\top \widehat{P}_\tau \widehat{x} + 2\widehat{r}_\tau^\top \widehat{x} + \widehat{q}_\tau\right] \\
&= \widehat{x}^\top (\widehat{Q}_\tau + (\widehat{K}'_\tau)^\top R_\tau \widehat{K}'_\tau)\widehat{x} + \widehat{x}^\top (\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} \\
&\quad + 2w_\tau^\top \widehat{D}_\tau^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} + 2\widehat{r}_{\tau+1}^\top(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}'_\tau)\widehat{x} - (\widehat{x}^\top \widehat{P}_\tau \widehat{x} + 2\widehat{r}_\tau^\top \widehat{x}) \\
&= \widehat{x}^\top (\widehat{Q}_\tau + (\widehat{K}'_\tau - \widehat{K}_\tau + \widehat{K}_\tau)^\top R_\tau (\widehat{K}'_\tau - \widehat{K}_\tau + \widehat{K}_\tau))\widehat{x} \\
&\quad + \widehat{x}^\top (\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau - \widehat{B}_\tau(\widehat{K}'_\tau - \widehat{K}_\tau))^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau - \widehat{B}_\tau(\widehat{K}'_\tau - \widehat{K}_\tau))\widehat{x} \\
&\quad + 2\bar{w}_\tau^\top \widehat{D}_\tau^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau - \widehat{B}_\tau(\widehat{K}'_\tau - \widehat{K}_\tau))\widehat{x} + 2\widehat{r}_{\tau+1}^\top(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau - \widehat{B}_\tau(\widehat{K}'_\tau - \widehat{K}_\tau))\widehat{x} \\
&\quad - \widehat{x}^\top (\widehat{Q}_\tau + \widehat{K}_\tau^\top R_\tau \widehat{K}_\tau + (\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau)^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau))\widehat{x} \\
&\quad - 2\widehat{r}_{\tau+1}^\top(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau)\widehat{x} - 2\bar{w}_\tau^\top \widehat{D}_\tau^\top \widehat{P}_{\tau+1}(\widehat{A}_\tau - \widehat{B}_\tau \widehat{K}_\tau)\widehat{x} \\
&= 2\widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top ((R_\tau + \widehat{B}_\tau^\top \widehat{P}_{\tau+1}\widehat{B}_\tau)\widehat{K}_\tau - \widehat{B}_\tau^\top \widehat{P}_{\tau+1}\widehat{A}_\tau)\widehat{x} \\
&\quad + \widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top (R_\tau + \widehat{B}_\tau^\top \widehat{P}_{\tau+1}\widehat{B}_\tau)(\widehat{K}'_\tau - \widehat{K}_\tau)\widehat{x} \\
&\quad - 2\widehat{x}^\top (\widehat{K}'_\tau - \widehat{K}_\tau)^\top \widehat{B}_\tau^\top (\widehat{P}_{\tau+1}\widehat{D}_\tau \bar{w}_\tau + \widehat{r}_{\tau+1}).
\end{aligned}
\tag{3.2.32}
$$

□

Lemma 3.4. *Let $\widehat{K}^*$ be an optimal policy and $C(\widehat{K})$ be finite; then*

$$
(\underline{\sigma}_X + 1)\sum_{t=0}^{T-1} \frac{\mathrm{Tr}(E_t^\top E_t) + F_t^\top F_t}{\|R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t\|} \leqslant C(\widehat{K}) - C(\widehat{K}^*) \leqslant \frac{\|\Sigma_{\widehat{K}^*}\| + 1}{4{\underline{\sigma}_X}^2 \underline{\sigma}_R} \sum_{t=0}^{T-1} \left(\nabla_t C(\widehat{K})^\top \nabla_t C(\widehat{K})\right),
$$

where $\underline{\sigma}_X$ and $\underline{\sigma}_Q$ are defined in (3.2.24a) and (3.2.24b).

Proof. First, for any $\widehat{K}'_\tau$, from (3.2.32),

$$\begin{aligned}
&A_{\widehat{K}}(\widehat{x}, -\widehat{K}'_\tau\widehat{x}, \tau)\\
&= Q_{\widehat{K}}(\widehat{x}, -\widehat{K}'_\tau\widehat{x}, \tau) - V_{\widehat{K}}(\widehat{x}, \tau)\\
&= 2\widehat{x}^\top(\widehat{K}'_\tau - \widehat{K}_\tau)^\top(E_\tau\widehat{x} - F_\tau) + \widehat{x}^\top(\widehat{K}'_\tau - \widehat{K}_\tau)^\top(R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)(\widehat{K}'_\tau - \widehat{K}_\tau)\widehat{x}\\
&= ((\widehat{K}'_\tau - \widehat{K}_\tau)\widehat{x} + (R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}(E_\tau\widehat{x} - F_\tau))^\top(R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)\\
&\quad((\widehat{K}'_\tau - \widehat{K}_\tau)\widehat{x} + (R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}(E_\tau\widehat{x} - F_\tau))\\
&\quad - (E_\tau\widehat{x} - F_\tau)^\top(R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}(E_\tau\widehat{x} - F_\tau)\\
&\geqslant -\operatorname{Tr}(\widetilde{x}\widetilde{x}^\top\widetilde{E}_\tau^\top(R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}\widetilde{E}_\tau),
\end{aligned} \tag{3.2.33}$$

where $\widetilde{x}_\tau = \begin{bmatrix}\widehat{x}_\tau\\ 1\end{bmatrix}$ and $\widetilde{E}_\tau = \begin{bmatrix}E_\tau & -F_\tau\end{bmatrix}$. The equality holds when $\widehat{K}'_\tau = \widehat{K}_\tau - (R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}E_\tau$ and $(R_\tau + \widehat{B}_\tau^\top\widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}F_\tau = 0$. Then,

$$\begin{aligned}
C(\widehat{K}) - C(\widehat{K}^*) &= -\mathbb{E}\left[\sum_{t=0}^{T-1} A_K\left(\widehat{x}_t^*, u_t^*, t\right)\right]\\
&\leqslant \mathbb{E}\left[\sum_{t=0}^{T-1}\operatorname{Tr}\left(\widetilde{x}_t^*(\widetilde{x}_t^*)^\top\widetilde{E}_t^\top(R_t + \widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t)^{-1}\widetilde{E}_t\right)\right]\\
&\leqslant \left(\|\Sigma_{\widehat{K}*}\| + 1\right)\sum_{t=0}^{T-1}\operatorname{Tr}\left(\widetilde{E}_t^\top(R_t + \widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t)^{-1}\widetilde{E}_t\right)\\
&\leqslant \frac{\|\Sigma_{\widehat{K}*}\| + 1}{\underline{\sigma}_R}\sum_{t=0}^{T-1}\operatorname{Tr}\left(\widetilde{E}_t^\top\widetilde{E}_t\right)\\
&\leqslant \frac{\|\Sigma_{\widehat{K}*}\| + 1}{4\underline{\sigma}_X^2\underline{\sigma}_R}\sum_{t=0}^{T-1}\operatorname{Tr}\left(\nabla_t C(\widehat{K})^\top\nabla_t C(\widehat{K})\right),
\end{aligned}$$

where $\underline{\sigma}_X$ is defined in (3.2.24a) and $\underline{\sigma}_R$ is defined in (3.2.24b). For the lower bound,

consider $\widehat{K}'_\tau = \widehat{K}_\tau - (R_\tau + \widehat{B}_\tau^\top \widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}E_\tau$ and $(R_\tau + \widehat{B}_\tau^\top \widehat{P}_{\tau+1}\widehat{B}_\tau)^{-1}F_\tau = 0$, where the equality holds in (3.2.33). Using $C(K^*) \leqslant C(\widehat{K}')$,

$$\begin{aligned} C(\widehat{K}) - C(\widehat{K}^*) &\geqslant C(\widehat{K}) - C(\widehat{K}') \\ &= -\mathbb{E}\left[\sum_{t=0}^{T-1} A_{\widehat{K}}\left(\widehat{x}'_t, u'_t, t\right)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{T-1} \operatorname{Tr}\left(\widetilde{x}'_t(\widetilde{x}'_t)^\top \widetilde{E}_t^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t)^{-1}\widetilde{E}_t\right)\right] \\ &\geqslant (\underline{\sigma}_X + 1)\sum_{t=0}^{T-1} \frac{\operatorname{Tr}(E_t^\top E_t) + F_t^\top F}{\|R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t\|}. \end{aligned} \tag{3.2.34}$$

□

Having obtained the explicit policy gradient, we next investigate its structural properties. Lemma 3.5 demonstrates an "almost smoothness" condition of the cost landscape, which shows that when $\widehat{K}'$ is sufficiently close to $\widehat{K}$, $C(\widehat{K}') - C(\widehat{K})$ is bounded by the sum of the first- and second-order terms in $\widehat{K} - \widehat{K}'$.

Lemma 3.5. ("Almost Smoothness"). *Let $\{\widehat{x}'_t\}$ be the sequence of states for a single trajectory generated by $\widehat{K}'$ starting from $\widehat{x}'_0 = \widehat{x}_0$. Then, $C(\widehat{K})$ satisfies*

$$\begin{aligned} C(\widehat{K}') - C(\widehat{K}) = \sum_{t=0}^{T-1} \Big[&2\operatorname{Tr}(\Sigma'_t(\widehat{K}'_t - \widehat{K}_t)^\top E_t) + \operatorname{Tr}(\Sigma'_t(\widehat{K}'_t - \widehat{K}_t)^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t)(\widehat{K}'_t - \widehat{K}_t)) \\ &-2\mu'_t(\widehat{K}'_t - \widehat{K}_t)^\top F_t \Big], \end{aligned}$$

where $\Sigma'_t = \mathbb{E}\left[\widehat{x}'_t(\widehat{x}'_t)^\top\right]$ and $\mu'_t = \mathbb{E}\left[(\widehat{x}'_t)^\top\right]$.

Proof. By Lemma 3.3, we have

$$
\begin{aligned}
& C(\widehat{K}') - C(\widehat{K}) \\
= \; & \mathbb{E}\left[\sum_{t=0}^{T-1} A_{\widehat{K}}(\widehat{x}'_t, -\widehat{K}'_t\widehat{x}'_t, t)\right] \\
= \; & \sum_{t=0}^{T-1}\Big[2\operatorname{Tr}(\Sigma'_t(\widehat{K}'_t - \widehat{K}_t)^\top E_t) + \operatorname{Tr}(\Sigma'_t(\widehat{K}'_t - \widehat{K}_t)^\top(R_t + \widehat{B}_t^\top \widehat{P}_{t+1}\widehat{B}_t)(\widehat{K}'_t - \widehat{K}_t)) \\
& -2\mu'_t(\widehat{K}'_t - \widehat{K}_t)^\top F_t\Big].
\end{aligned}
$$

□

The first term and third term can be added as $\operatorname{Tr}((\widehat{K}_t - \widehat{K}'_t)\nabla_t C(\widehat{K}))$ by Lemma 3.2, while the second term is the second order of $\widehat{K}_t - \widehat{K}'_t$. We further bound P_t and $\Sigma_{\widehat{K}}$, which is provided below in Lemma 3.6.

Lemma 3.6. *For every $t = 0, 1, \cdots, T$ and for the aggregated quantities, we have*

$$
\|P_t\| \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_X}, \quad \|\Sigma_{\widehat{K}}\| \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_Q}, \quad \|\mu_{\widehat{K}}\| \leqslant \sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}},
$$

where $\underline{\sigma}_X$ and $\underline{\sigma}_Q$ are defined in (3.2.24a) and (3.2.24c).

Proof. For $t = 0, 1, \cdots, T$,

(i) Bound on $\|\widehat{P}_t\|$,

$$
C(\widehat{K}) \geqslant \mathbb{E}\left[\widehat{x}_t^\top \widehat{P}_t \widehat{x}_t\right] \geqslant \|\widehat{P}_t\|\left(\mathbb{E}\left[x_t^2\right]\right) \geqslant \underline{\sigma}_X \|\widehat{P}_t\|.
$$

(ii) Bound on $\|\Sigma_{\widehat{K}}\|$,

$$
C(\widehat{K}) = \sum_{t=0}^{T-1} \operatorname{Tr}\left(\mathbb{E}[\widehat{x}_t\widehat{x}_t^\top](\widehat{Q}_t + \widehat{K}_t^\top R_t \widehat{K}_t)\right) + \operatorname{Tr}\left(\mathbb{E}[\widehat{x}_T\widehat{x}_T^\top]\widehat{Q}_T\right) \geqslant \underline{\sigma}_Q \|\Sigma_{\widehat{K}}\|.
$$

Rearranging yields the second inequality.

(iii) Bound on $\|\mu_{\widehat{K}}\|$, first extract the mean contribution from the cost. Note

$$\Sigma_t = \mathrm{Cov}\,(\widehat{x}_t) + \mu_t^\top \mu_t \succeq \mu_t^\top \mu_t.$$

So for each t,

$$\mathrm{Tr}\left(\mathbb{E}[\widehat{x}_t \widehat{x}_t^\top]\widehat{Q}_t\right) \geqslant \mu_t \widehat{Q}_t \mu_t^\top \geqslant \underline{\sigma}_Q \|\mu_t\|^2 .$$

Summing over $t = 0, \cdots, T-1$ gives

$$C(\widehat{K}) \geqslant \underline{\sigma}_Q \sum_{t=0}^{T-1} \|\mu_t\|^2 \quad \Longrightarrow \quad \sum_{t=0}^{T-1} \|\mu_t\|^2 \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_Q}.$$

Now apply vector Cauchy-Schwarz, i.e., the inequality $\|\sum_t v_t\|^2 \leqslant (T+1)\sum_t \|v_t\|^2$:

$$\|\mu_{\widehat{K}}\|^2 \leqslant T \sum_{t=0}^{T} \|\mu_t\|^2 \leqslant \frac{TC(\widehat{K})}{\sigma_Q}.$$

Taking square roots yields the third inequality:

$$\|\mu_{\widehat{K}}\| \leqslant \sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}}.$$

□

3.2.2 Perturbation Analysis of State Statistics

This section examines how the state statistics, namely the covariance $\Sigma_{\widehat{K}}$ and mean $\mu_{\widehat{K}}$, respond to changes in the policy $\widehat{K}$. We extend the operator level bounds to derive a comprehensive perturbation analysis of $\Sigma_{\widehat{K}}$ and $\mu_{\widehat{K}}$, which quantifies the robustness of state evolution under policy updates.

First, let us define two linear operators on symmetric matrices. For $X \in \mathbb{R}^{(m+1)\times(m+1)}$ and $Y \in \mathbb{R}^{1\times(m+1)}$ we set

$$\mathcal{F}_{\widehat{K}_t}(X) = (\widehat{A}_t - \widehat{B}_t\widehat{K}_t)X(\widehat{A}_t - \widehat{B}_t\widehat{K}_t)^\top, \quad \mathcal{F}^{\mu}_{\widehat{K}_t}(Y) = Y(\widehat{A}_t - \widehat{B}_t\widehat{K}_t)^\top,$$

and

$$\mathcal{T}_{\widehat{K}}(X) = X + \sum_{t=0}^{T-1} \Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)X\Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)^\top,$$

$$\mathcal{T}^{\mu}_{\widehat{K}}(Y) = Y + \sum_{t=0}^{T-1} Y\Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)^\top.$$

If we write $\mathcal{G}_t = \mathcal{F}_{\widehat{K}_t} \circ \mathcal{F}_{\widehat{K}_{t-1}} \circ \cdots \circ \mathcal{F}_{\widehat{K}_0}$ and $\mathcal{G}^{\mu}_t = \mathcal{F}^{\mu}_{\widehat{K}_t} \circ \mathcal{F}^{\mu}_{\widehat{K}_{t-1}} \circ \cdots \circ \mathcal{F}^{\mu}_{\widehat{K}_0}$ then

$$\mathcal{G}_t(X) = \mathcal{F}_{\widehat{K}_t} \circ \mathcal{G}_{t-1}(X) = \Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)X\Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)^\top, \tag{3.2.35}$$

$$\mathcal{G}^{\mu}_t(Y) = \mathcal{F}^{\mu}_{\widehat{K}_t} \circ \mathcal{G}^{\mu}_{t-1}(Y) = Y\Pi_{i=0}^{t}(\widehat{A}_i - \widehat{B}_i\widehat{K}_i)^\top, \tag{3.2.36}$$

$$\mathcal{T}_{\widehat{K}}(X) = X + \sum_{t=0}^{T-1} \mathcal{G}_t(X), \tag{3.2.37}$$

$$\mathcal{T}^{\mu}_{\widehat{K}}(Y) = Y + \sum_{t=0}^{T-1} \mathcal{G}^{\mu}_t(Y). \tag{3.2.38}$$

We first show the relationship between the operator $\mathcal{T}_{\widehat{K}}$ and the quantity $\Sigma_{\widehat{K}}$, which pave the way for the relation between $\mathcal{T}^{\mu}_{\widehat{K}}$ and $\mu_{\widehat{K}}$.

Proposition 3.1. *For $T \geqslant 2$,*

$$\Sigma_{\widehat{K}} = \mathcal{T}_{\widehat{K}}(\Sigma_0) + \phi(\widehat{K}, \widehat{\phi}),$$

$$\mu_{\widehat{K}} = \mathcal{T}^{\mu}_{\widehat{K}}(\mu_0) + \phi^{\mu}(\widehat{K}, \widehat{\phi}^{\mu}).$$

Define $(i)\phi(\widehat{K}, \widehat{\phi}) = \sum_{t=1}^{T-1}\sum_{i=1}^{t} \Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)\widehat{\phi}_{i-1}\Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top + \sum_{t=0}^{T-1}\widehat{\phi}_t$, *with* $\widehat{\phi} = 2\widehat{D}\bar{w}\mu(\widehat{A} - \widehat{B}\widehat{K})^\top + \widehat{D}\Sigma^w\widehat{D}^\top$ *and* $\Sigma_0 = \mathbb{E}[\widehat{x}_0\widehat{x}_0^\top]$; $(ii)\phi(\widehat{K}, \widehat{\phi}^{\mu}) = \sum_{t=1}^{T-1}\sum_{i=1}^{t} \widehat{\phi}^{\mu}_{i-1}\Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top + \sum_{t=0}^{T-1}\widehat{\phi}^{\mu}_t$, *with* $\widehat{\phi}^{\mu} = \bar{w}^\top\widehat{D}^\top$ *and* $\mu_0 = \mathbb{E}[\widehat{x}_0^\top]$.

Proof. We only need to prove for $\Sigma_{\widehat{K}}$ since the proof of $\mu_{\widehat{K}}$ follows the same method.

Recall that $\Sigma_t = \mathbb{E}\left[\widehat{x}_t \widehat{x}_t^\top\right]$. Note that

$$\begin{aligned}
\Sigma_1 &= \mathbb{E}[\widehat{x}_1 \widehat{x}_1^\top] \\
&= \mathbb{E}\left[\left((\widehat{A}_0 - \widehat{B}_0\widehat{K}_0)\widehat{x}_0 + \widehat{D}_0 w_0\right)\left((\widehat{A}_0 - \widehat{B}_0\widehat{K}_0)\widehat{x}_0 + \widehat{D}_0 w_0\right)^\top\right] \\
&= (\widehat{A}_0 - \widehat{B}_0\widehat{K}_0)\Sigma_0(\widehat{A}_0 - \widehat{B}_0\widehat{K}_0)^\top + 2\widehat{D}_0\bar{w}_0\mu_0(\widehat{A}_0 - \widehat{B}_0\widehat{K}_0)^\top + \widehat{D}_0\Sigma_0^w\widehat{D}_0^\top \\
&= \mathcal{G}_0(\Sigma_0) + \widehat{\phi}_0.
\end{aligned}$$

Now we first prove that

$$\Sigma_t = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{i=1}^{t-1} \Pi_{s=i}^{t-1}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)\widehat{\phi}_{i-1}\Pi_{s=i}^{t-1}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top + \widehat{\phi}_{t-1}, \tag{3.2.39}$$

for $t = 2, 3, \cdots, T$. When $t = 2$,

$$\begin{aligned}
\Sigma_2 &= \mathbb{E}[\widehat{x}_2 \widehat{x}_2^\top] \\
&= \mathbb{E}\left[\left((\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)\widehat{x}_1 + \widehat{D}_1 w_1\right)\left((\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)\widehat{x}_1 + \widehat{D}_1 w_1\right)^\top\right] \\
&= (\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)\Sigma_1(\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)^\top + 2\widehat{D}_1\bar{w}_1\mu_1(\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)^\top + \widehat{D}_1\Sigma_1^w\widehat{D}_1^\top \\
&= \mathcal{G}_1(\Sigma_0) + (\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)\widehat{\phi}_0(\widehat{A}_1 - \widehat{B}_1\widehat{K}_1)^\top + \widehat{\phi}_1,
\end{aligned}$$

which satisfies (3.2.39). Assume (3.2.39) holds for $t \leqslant k$. Then for $t = k + 1$,

$$\begin{aligned}
\mathbb{E}[\widehat{x}_{t+1}\widehat{x}_{t+1}^\top] &= \mathbb{E}\left[\left((\widehat{A}_t - \widehat{B}_t\widehat{K}_t)\widehat{x}_t + \widehat{D}_t w_t\right)\left((\widehat{A}_t - \widehat{B}_t\widehat{K}_t)\widehat{x}_t + \widehat{D}_t w_t\right)^\top\right] \\
&= (\widehat{A}_t - \widehat{B}_t\widehat{K}_t)\Sigma_t(\widehat{A}_t - \widehat{B}_t\widehat{K}_t)^\top + 2\widehat{D}_t\bar{w}_t\mu_t(\widehat{A}_t - \widehat{B}_t\widehat{K}_t)^\top + \widehat{D}_t\Sigma_t^w\widehat{D}_t^\top \\
&= \mathcal{G}_t(\Sigma_0) + \sum_{i=1}^{t} \Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)\widehat{\phi}_{i-1}\Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top + \widehat{\phi}_t.
\end{aligned}$$

Therefore (3.2.39) holds, $\forall t = 1, 2, \cdots, T$. Finally,

$$\begin{aligned}\Sigma_{\widehat{K}} &= \sum_{t=0}^{T} \Sigma_t \\ &= \Sigma_0 + \sum_{t=0}^{T-1} \mathcal{G}_t(\Sigma_0) + \sum_{t=1}^{T-1}\sum_{i=1}^{t} \Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)\widehat{\phi}_{i-1}\Pi_{s=i}^{t}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top + \sum_{t=0}^{T-1}\widehat{\phi}_t \\ &= \mathcal{T}_{\widehat{K}}(\Sigma_0) + \phi(\widehat{K}, \widehat{\phi}).\end{aligned}$$

□

Let

$$\rho = \max\left\{\max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\|, \max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}'_t\|, 1+\xi\right\}, \tag{3.2.40}$$

for some small constant $\xi > 0$. Then we have the following result on perturbations of $\Sigma_{\widehat{K}}$ and $\mu_{\widehat{K}}$. Lemma 3.7 and 3.8 establish Lipschitz continuity of the associated operators $\mathcal{F}_{\widehat{K}_t}$ (and $\mathcal{F}^{\mu}_{\widehat{K}_t}$) as well as $\mathcal{G}_t$ (and $\mathcal{G}^{\mu}_t$).

Lemma 3.7. *It holds that* $\forall t = 0, 1, \cdots, T-1$,

$$\begin{aligned}\|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}'_t}\| &\leqslant 2\|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\|\|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\| + \|\widehat{B}_t\|^2\|\widehat{K}_t - \widehat{K}'_t\|^2, \\ \|\mathcal{F}^{\mu}_{\widehat{K}_t} - \mathcal{F}^{\mu}_{\widehat{K}'_t}\| &\leqslant \|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\|.\end{aligned} \tag{3.2.41}$$

Proof. Let $\Delta\widehat{K}_t = \widehat{K}_t - \widehat{K}'_t$ and $\widehat{A}^{\mathrm{cl}}_t = \widehat{A}_t - \widehat{B}_t\widehat{K}_t$. Then

$$\widehat{A}'^{\mathrm{cl}}_t = \widehat{A}_t - \widehat{B}_t\widehat{K}'_t = \widehat{A}^{\mathrm{cl}}_t + \widehat{B}_t\Delta\widehat{K}_t.$$

Thus,

$$\begin{aligned}\mathcal{F}_{\widehat{K}'}(X) &= (\widehat{A}^{\mathrm{cl}}_t + \widehat{B}_t\Delta\widehat{K}_t)X(\widehat{A}^{\mathrm{cl}}_t + \widehat{B}_t\Delta\widehat{K}_t)^\top \\ &= \widehat{A}^{\mathrm{cl}}_t X(\widehat{A}^{\mathrm{cl}})^\top_t + 2\widehat{A}^{\mathrm{cl}}_t X(\Delta\widehat{K}_t)^\top\widehat{B}^\top_t + \widehat{B}_t\Delta\widehat{K}_t X(\Delta\widehat{K}_t)^\top\widehat{B}^\top_t,\end{aligned}$$

and

$$\mathcal{F}^{\mu}_{\widehat{K}'}(Y) = Y(\widehat{A}^{\mathrm{cl}}_t + \widehat{B}_t\Delta\widehat{K}_t)^\top = Y(\widehat{A}^{\mathrm{cl}})^\top_t + Y(\Delta\widehat{K}_t)^\top\widehat{B}^\top_t.$$

The differences are

$$\begin{aligned}
\mathcal{F}_{\widehat{K}}(X) - \mathcal{F}_{\widehat{K}'}(X) &= -2\widehat{A}^{\mathrm{cl}}_t X(\Delta\widehat{K})^\top_t\widehat{B}^\top_t - \widehat{B}_t\Delta\widehat{K}_t X(\Delta\widehat{K}_t)^\top\widehat{B}^\top_t,\\
\mathcal{F}^{\mu}_{\widehat{K}}(Y) - \mathcal{F}^{\mu}_{\widehat{K}'}(Y) &= -Y(\Delta\widehat{K}_t)^\top\widehat{B}^\top_t.
\end{aligned}$$

Take operator norms and use submultiplicativity $\|MN\| \leqslant \|M\|\|N\|$:

$$\begin{aligned}
\|\mathcal{F}_{\widehat{K}}(X) - \mathcal{F}_{\widehat{K}'}(X)\| &\leqslant 2\|\widehat{A}^{\mathrm{cl}}_t\|\|\widehat{B}_t\|\|\Delta\widehat{K}_t\|\|X\| + \|\widehat{B}_t\|^2\|\Delta\widehat{K}\|^2\|X\|,\\
\|\mathcal{F}^{\mu}_{\widehat{K}}(Y) - \mathcal{F}^{\mu}_{\widehat{K}'}(Y)\| &\leqslant \|\widehat{B}_t\|\|\Delta\widehat{K}_t\|\|Y\|.
\end{aligned}$$

Setting $\|X\| = 1$ and $\|Y\| = 1$ gives

$$\begin{aligned}
\|\mathcal{F}_{\widehat{K}} - \mathcal{F}_{\widehat{K}'}\| &\leqslant 2\|\widehat{A}^{\mathrm{cl}}_t\|\|\widehat{B}_t\|\|\Delta\widehat{K}_t\| + \|\widehat{B}_t\|^2\|\Delta\widehat{K}_t\|^2,\\
\|\mathcal{F}^{\mu}_{\widehat{K}} - \mathcal{F}^{\mu}_{\widehat{K}'}\| &\leqslant \|\widehat{B}_t\|\|\Delta\widehat{K}_t\|,
\end{aligned}$$

that is,

$$\begin{aligned}
\|\mathcal{F}_{\widehat{K}} - \mathcal{F}_{\widehat{K}'}\| &\leqslant 2\|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\|\|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\| + \|\widehat{B}_t\|^2\|\widehat{K}_t - \widehat{K}'_t\|^2,\\
\|\mathcal{F}^{\mu}_{\widehat{K}} - \mathcal{F}^{\mu}_{\widehat{K}'}\| &\leqslant \|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\|.
\end{aligned}$$

□

Recall the definition of $\mathcal{G}_t$ in (3.2.35) associated with K; similarly, let us define $\mathcal{G}'_t = \mathcal{F}_{K'_t} \circ \mathcal{F}_{K'_{t-1}} \circ \cdots \circ \mathcal{F}_{K'_0}$ for policy K'. Then we have the following perturbation analysis for $\mathcal{G}_t$.

Lemma 3.8. (Perturbation Analysis for $\mathcal{G}_t$ and $\mathcal{G}^{\mu}_t$) *For any symmetric matrix* Σ

and vector μ*,*

$$\sum_{t=0}^{T-1} \|(\mathcal{G}_t - \mathcal{G}'_t)(\Sigma)\| \leqslant \frac{\rho^{2T}-1}{\rho^2-1} \left(\sum_{t=0}^{T-1} \|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}'_t}\| \right) \|\Sigma\|, \text{ and}$$

$$\sum_{t=0}^{T-1} \|(\mathcal{G}_t^\mu - (\mathcal{G}_t^\mu)')(\mu)\| \leqslant \frac{\rho^{T}-1}{\rho-1} \left(\sum_{t=0}^{T-1} \|\mathcal{F}_{\widehat{K}_t}^\mu - \mathcal{F}_{\widehat{K}'_t}^\mu\| \right) \|\mu\|.$$

Proof. By direct calculation,

$$\|\mathcal{G}_t\| \leqslant \rho^{2(t+1)}, \quad \|\mathcal{G}'_t\| \leqslant \rho^{2(t+1)} \quad \text{and} \quad \|\mathcal{G}_t^\mu\| \leqslant \rho^{t+1}, \quad \|(\mathcal{G}_t^\mu)'\| \leqslant \rho^{t+1}. \tag{3.2.42}$$

Denote $\mathcal{F}_t = \mathcal{F}_{\widehat{K}_t}$ and $\mathcal{F}'_t = \mathcal{F}_{\widehat{K}'_t}$ to ease the exposition. Then for any matrix Σ and $t \geqslant 0$,

$$\begin{aligned}
\left\|\left(\mathcal{G}'_{t+1} - \mathcal{G}_{t+1}\right)(\Sigma)\right\| &= \left\|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\right\| \\
&= \left\|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) + \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\right\| \\
&\leqslant \left\|\mathcal{F}'_{t+1} \circ \mathcal{G}'_t(\Sigma) - \mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma)\right\| + \left\|\mathcal{F}'_{t+1} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\right\| \\
&= \left\|\mathcal{F}'_{t+1} \circ (\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\right\| + \left\|\left(\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\right) \circ \mathcal{G}_t(\Sigma)\right\| \\
&\leqslant \left\|\mathcal{F}'_{t+1}\right\| \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \|\mathcal{G}_t\| \left\|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\right\| \|\Sigma\| \\
&\leqslant \rho^2 \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)} \left\|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\right\| \|\Sigma\|.
\end{aligned}$$

We define $\mathcal{F}_t^\mu = \mathcal{F}_{\widehat{K}_t}^\mu$ and $(\mathcal{F}_t^\mu)' = (\mathcal{F}_{\widehat{K}_t}^\mu)'$ and can get similar inequality result for μ. Therefore,

$$\left\|\left(\mathcal{G}'_{t+1} - \mathcal{G}_{t+1}\right)(\Sigma)\right\| \leqslant \rho^2 \|(\mathcal{G}'_t - \mathcal{G}_t)(\Sigma)\| + \rho^{2(t+1)} \left\|\mathcal{F}'_{t+1} - \mathcal{F}_{t+1}\right\| \|\Sigma\|, \tag{3.2.43}$$

$$\left\|\left((\mathcal{G}_{t+1}^\mu)' - \mathcal{G}_{t+1}^\mu\right)(\mu)\right\| \leqslant \rho \left\|\left((\mathcal{G}_{t+1}^\mu)' - \mathcal{G}_{t+1}^\mu\right)(\mu)\right\| + \rho^{t+1} \left\|(\mathcal{F}_{t+1}^\mu)' - \mathcal{F}_{t+1}^\mu\right\| \|\mu\|. \tag{3.2.44}$$

Summing (3.2.43) and (3.2.44) up for $t \in \{1, 2, \cdots, T-2\}$ with $\|\mathcal{G}'_0 - \mathcal{G}_0\| = \|\mathcal{F}'_0 - \mathcal{F}_0\|$

and $\|(\mathcal{G}_0^\mu)' - \mathcal{G}_0^\mu\| = \|(\mathcal{F}_0^\mu)' - \mathcal{F}_0^\mu\|$,

$$\sum_{t=0}^{T-1} \|(\mathcal{G}_t - \mathcal{G}_t')(\Sigma)\| \leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_t - \mathcal{F}_t'\|\right)\|\Sigma\|,$$

$$\sum_{t=0}^{T-1} \|(\mathcal{G}_t^\mu - (\mathcal{G}_t^\mu)')(\mu)\| \leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_{\widehat{K}_t}^\mu - \mathcal{F}_{\widehat{K}_t'}^\mu\|\right)\|\mu\|.$$

□

The following perturbation analysis on $\mathcal{T}$ follows immediately from Lemma 3.8.

Corollary 3.1. *For any symmetric matrix $\Sigma \in \mathbb{R}^{(m+1)\times(m+1)}$ and matrix $\mu \in \mathbb{R}^{1\times(m+1)}$, we have*

$$\|(\mathcal{T}_{\widehat{K}} - \mathcal{T}_{\widehat{K}'})(\Sigma)\| \leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}_t'}\|\right)\|\Sigma\|,$$

$$\|(\mathcal{T}_{\widehat{K}}^\mu - \mathcal{T}_{\widehat{K}'}^\mu)(\mu)\| \leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_{\widehat{K}_t}^\mu - \mathcal{F}_{\widehat{K}_t'}^\mu\|\right)\|\mu\|,$$

where ρ is defined in (3.2.40).

Now we are ready to present the Lemma 3.9.

Lemma 3.9. (Perturbation Analysis of $\Sigma_{\widehat{K}}$ and $\mu_{\widehat{K}}$). *We have*

$$\begin{aligned}
\|\Sigma_{\widehat{K}} - \Sigma_{\widehat{K}'}\| &\leqslant \left\|\left(\mathcal{T}_{\widehat{K}} - \mathcal{T}_{\widehat{K}'}\right)(\Sigma_0)\right\| + \|\phi(\widehat{K},\widehat{\phi}) - \phi(\widehat{K}',\widehat{\phi})\| \\
&\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right) \\
&\quad \left(2\rho\|\widehat{B}\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'||| + \|\widehat{B}\|^2\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||^2\right) + 2\|\widehat{D}\|\|\bar{w}\|\|\mu\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||, \text{ and} \\
\|\mu_{\widehat{K}} - \mu_{\widehat{K}'}\| &\leqslant \|(\mathcal{T}_{\widehat{K}}^\mu - \widehat{\mathcal{T}}_{K'}^\mu)(\mu_0)\| + \|\phi(\widehat{K},\widehat{\phi}^\mu) - \phi(\widehat{K}',\widehat{\phi}^\mu)\| \\
&\leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}} + T\|\widehat{D}\|\|\bar{w}\|\right)\|\widehat{B}\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||,
\end{aligned}$$

where $\|\widehat{B}\| = \max_t\{\|\widehat{B}_t\|\}$ *and the same definition for* $\|\widehat{D}\|, \|\bar{w}\|, \|\mu\|$ *and* $\|\Sigma^w\|$.

Proof. Using Lemma 3.7,

$$\sum_{t=0}^{T-1} \|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}'_t}\| = \sum_{t=0}^{T-1} \left(2\|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\|\|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\| + \|\widehat{B}_t\|^2\|\widehat{K}_t - \widehat{K}'_t\|^2\right)$$

$$\leqslant 2\rho\|\widehat{B}\| \sum_{t=0}^{T-1} \|\widehat{K}_t - \widehat{K}'_t\| + \|\widehat{B}\|^2 \sum_{t=0}^{T-1} \|\widehat{K}_t - \widehat{K}'_t\|^2, \text{ and}$$

$$\sum_{t=0}^{T-1} \|\mathcal{F}^{\mu}_{\widehat{K}_t} - \mathcal{F}^{\mu}_{\widehat{K}'_t}\| = \sum_{t=0}^{T-1} \left(\|\widehat{B}_t\|\|\widehat{K}_t - \widehat{K}'_t\|\right) \leqslant \|\widehat{B}\| \sum_{t=0}^{T-1} \|\widehat{K}_t - \widehat{K}'_t\|.$$

In the same way as for the proof of Lemma 3.8, for $t = 1, \cdots, T-1$, we have

$$\sum_{i=1}^{t} \left\|\Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)\widehat{\phi}_{i-1}\Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top - \Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}'_s)\widehat{\phi}_{i-1}\Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}'_s)^\top\right\|$$

$$\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{i=0}^{t} \|\mathcal{F}_{\widehat{K}_i} - \mathcal{F}_{\widehat{K}'_i}\|\|\widehat{\phi}_{i-1}\|\right)$$

$$\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{i=0}^{t} \|\mathcal{F}_{\widehat{K}_i} - \mathcal{F}_{\widehat{K}'_i}\|\right)(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|), \text{ and}$$

$$\sum_{i=1}^{t} \left\|\widehat{\phi}^{\mu}_{i-1}\Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}_s)^\top - \widehat{\phi}^{\mu}_{i-1}\Pi^t_{s=i}(\widehat{A}_s - \widehat{B}_s\widehat{K}'_s)^\top\right\|$$

$$\leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sum_{i=0}^{t} \|\mathcal{F}_{\widehat{K}_i} - \mathcal{F}_{\widehat{K}'_i}\|\|\widehat{\phi}^{\mu}_{i-1}\|\right)$$

$$\leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sum_{i=0}^{t} \|\mathcal{F}_{\widehat{K}_i} - \mathcal{F}_{\widehat{K}'_i}\|\right)\|\widehat{D}\|\|\bar{w}\|, \tag{3.2.45}$$

where we define $\widehat{\phi}_{i-1} = \mathbf{0}_{(m+1)\times(m+1)}$ and $\widehat{\phi}^{\mu}_{i-1} = \mathbf{0}_{1\times(m+1)}$ when $i = 0$. By Proposition

3.1, Corollary 3.1, (3.2.38) and (3.2.45), we have

$$
\begin{aligned}
\|\Sigma_{\widehat{K}} - \Sigma_{\widehat{K}'}\| &\leqslant \left\|\left(\mathcal{T}_{\widehat{K}} - \widehat{\mathcal{T}}_{K'}\right)(\Sigma_0)\right\| + \|\phi(\widehat{K}, \widehat{\phi}) - \phi(\widehat{K}', \widehat{\phi})\| \\
&\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1} \|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}'_t}\|\right)\left(\|\Sigma_0\| + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right) \\
&\quad + \sum_{t=0}^{T-1} 2\|\widehat{D}\|\|\bar{w}\|\|\mu\|\|\widehat{K}_t - \widehat{K}'_t\| \\
&\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(K)}{\underline{\sigma}_Q} + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right) \\
&\quad \left(2\rho\|\widehat{B}\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'||| + \|\widehat{B}\|^2\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||^2\right) + 2\|\widehat{D}\|\|\bar{w}\|\|\mu\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||, \\
\|\mu_{\widehat{K}} - \mu_{\widehat{K}'}\| &\leqslant \|(\mathcal{T}^{\mu}_{\widehat{K}} - \widehat{\mathcal{T}}^{\mu}_{K'})(\mu_0)\| + \|\phi(\widehat{K}, \widehat{\phi}^{\mu}) - \phi(\widehat{K}', \widehat{\phi}^{\mu})\| \\
&\leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sum_{t=0}^{T-1} \|\mathcal{F}^{\mu}_{\widehat{K}_t} - \mathcal{F}^{\mu}_{\widehat{K}'_t}\|\right)\left(\|\Sigma_0\| + T\|\widehat{D}\|\bar{w}\|\right) \\
&\leqslant \frac{\rho^{T}-1}{\rho-1}\left(\sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}} + T\|\widehat{D}\|\|\bar{w}\|\right)\|\widehat{B}\|\ |||\hat{\mathbf{K}} - \hat{\mathbf{K}}'|||.
\end{aligned}
\tag{3.2.46}
$$

The last inequalities for both holds since $\|\Sigma_0\| \leqslant \|\Sigma_{\widehat{K}}\| \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_Q}$ and $\|\mu_0\| \leqslant \|\mu_{\widehat{K}}\| \leqslant \sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}}$ by Lemma 3.6. □

3.2.3 Convergence and Complexity Guarantees

We can provide the global convergence Theorem 3.4 after two preliminary Lemmas about admissible step-size conditions and properties for the cumulation of cost sequence gradient and policy $\widehat{K}$.

Lemma 3.10. *We have*

$$
\widehat{K}'_t = \widehat{K}_t - \eta \nabla_t C(\widehat{K}), \tag{3.2.47}
$$

where

$$\eta \leqslant \min\left\{\frac{(\rho^2-1)(\rho-1)\underline{\sigma}_Q\sqrt{\underline{\sigma}_Q}\underline{\sigma}_X}{2T(Dn_1+Dn_2)\|\widehat{B}\|\max_t\{\|\nabla_t C(\widehat{K})\|\}},\ \frac{1}{2C_1}\right\}, \tag{3.2.48}$$

with

$$\begin{aligned}
Dn_1 &= \left(C(\widehat{K}) + \underline{\sigma}_Q T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right)(2\rho+1)(\rho^{2T}-1)(\rho-1)\sqrt{\underline{\sigma}_Q},\\
Dn_2 &= \left(\sqrt{TC(\widehat{K})} + \sqrt{\underline{\sigma}_Q}T\|\widehat{D}\|\|\bar{w}\|\right)(\rho^T-1)(\rho^2-1)\underline{\sigma}_Q,\\
C_1 &= \left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right)\left(\frac{(2\rho+1)\|\widehat{B}\|(\rho^{2T}-1)}{(\rho^2-1)\,\underline{\sigma}_X}\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right)\\
&\quad + \frac{2\|\widehat{D}\|\|\bar{w}\|\|\mu\|}{\underline{\sigma}_X}\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\| + \left(\sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}} + T\|\widehat{D}\|\|\bar{w}\|\right)\left(\frac{\|\widehat{B}\|(\rho^T-1)}{(\rho-1)\underline{\sigma}_X}\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right)\\
&\quad + \frac{2C(\widehat{K})}{\underline{\sigma}_Q}\sum_{t=0}^{T-1}\|R_t + \widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t\|.
\end{aligned} \tag{3.2.49}$$

Then we have

$$C(\widehat{K}') - C(\widehat{K}^*) \leqslant \left(1 - 2\eta\underline{\sigma}_R\frac{\underline{\sigma}_X^2}{\|\Sigma_{\widehat{K}^*}\|+1}\right)\left(C(\widehat{K}) - C(\widehat{K}^*)\right).$$

Proof. Given (3.2.47) and condition (3.2.48), we have $\|\widehat{K}_t' - \widehat{K}_t\| = \eta\|\nabla_t C(\widehat{K})\| \leqslant \frac{\underline{\sigma}_Q\underline{\sigma}_X}{2C(\widehat{K})\|\widehat{B}\|}$. Therefore,

$$\|\widehat{B}_t\|\|\widehat{K}_t' - \widehat{K}_t\| \leqslant \frac{\underline{\sigma}_Q\underline{\sigma}_X}{2C(\widehat{K})} \leqslant \frac{1}{2}. \tag{3.2.50}$$

The last inequality holds since $\underline{\sigma}_X \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_Q}$ given by Lemma 3.6. Therefore, by Lemma 3.7,

$$\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t} - \mathcal{F}_{K_t'}\| \leqslant (2\rho+1)\|\widehat{B}\|\left(\sum_{t=0}^{T-1}\|\widehat{K}_t - \widehat{K}_t'\|\right). \tag{3.2.51}$$

Define $\widetilde{\Sigma}_t = \begin{bmatrix}\Sigma_t \\ \mu_t\end{bmatrix}$, $\widetilde{\Sigma}'_t = \begin{bmatrix}\Sigma'_t \\ \mu'_t\end{bmatrix}$ and $\widetilde{E}_t = \begin{bmatrix}E_t & -F_t\end{bmatrix}$. From Lemmas 3.2 and 3.5, we have

$$
\begin{aligned}
& C(\widehat{K}') - C(\widehat{K}) \\
&= \sum_{t=0}^{T-1} \left[2 \operatorname{Tr}\left(\widetilde{\Sigma}'_t (\widehat{K}'_t - \widehat{K}_t)^\top \widetilde{E}_t \right) + \operatorname{Tr}\left(\Sigma'_t (\widehat{K}'_t - \widehat{K}_t)^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)(\widehat{K}'_t - \widehat{K}_t) \right) \right] \\
&= \sum_{t=0}^{T-1} \left[-4\eta \operatorname{Tr}\left(\widetilde{\Sigma}'_t \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \right) + 4\eta^2 \operatorname{Tr}\left(\Sigma'_t \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t) \widetilde{E}_t \widetilde{\Sigma}_t \right) \right] \\
&= \sum_{t=0}^{T-1} \left[-4\eta \operatorname{Tr}\left((\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t + \widetilde{\Sigma}_t) \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \right) + 4\eta^2 \operatorname{Tr}\left(\Sigma'_t \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t) \widetilde{E}_t \widetilde{\Sigma}_t \right) \right] \\
&\leqslant \sum_{t=0}^{T-1} \Big[-4\eta \operatorname{Tr}\left(\widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \right) + 4\eta \operatorname{Tr}\left((\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t) \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \widetilde{\Sigma}_t^\dagger \right) \\
&\quad + 4\eta^2 \operatorname{Tr}\left(\Sigma'_t \widetilde{\Sigma}_t^\top \widetilde{E}_t^\top (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t) \widetilde{E}_t \widetilde{\Sigma}_t \right) \Big] \\
&\leqslant \sum_{t=0}^{T-1} \Bigg[-4\eta \operatorname{Tr}\left(\widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \right) + 4\eta \frac{\|\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t\|}{\sigma_{\min}(\Sigma_t)} \operatorname{Tr}\left(\widetilde{\Sigma}_t \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \right) \\
&\quad + 4\eta^2 \|\Sigma'_t (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)\| \operatorname{Tr}\left(\widetilde{\Sigma}_t \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \right) \Bigg] \\
&\leqslant -\eta \left[1 - \frac{\sum_{t=0}^{T-1} \|\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t\|}{\underline{\sigma}_X} - \eta \|\Sigma_{\widehat{K}'}\| \sum_{t=0}^{T-1} \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\| \right] \sum_{t=0}^{T-1} \left[\nabla_t C(\widehat{K})^\top \nabla_t C(\widehat{K}) \right].
\end{aligned}
\tag{3.2.52}
$$

By Lemma 3.4, we have

$$
\begin{aligned}
C(\widehat{K}') - C(\widehat{K}) \leqslant -\eta & \left[1 - \frac{\sum_{t=0}^{T-1} \|\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t\|}{\underline{\sigma}_X} - \eta \|\Sigma_{\widehat{K}'}\| \sum_{t=0}^{T-1} \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\| \right] \\
& \left(\frac{4\underline{\sigma}_X^2 \underline{\sigma}_R}{\|\Sigma_{\widehat{K}^*}\| + 1} \right) \left(C(\widehat{K}) - C(\widehat{K}^*) \right),
\end{aligned}
\tag{3.2.53}
$$

provided that

$$1-\frac{\sum_{t=0}^{T-1}\|\widetilde{\Sigma}_t'-\widetilde{\Sigma}_t\|}{\underline{\sigma}_X}-\eta\|\Sigma_{\widehat{K}'}\|\sum_{t=0}^{T-1}\|R_t+\widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t\|>0. \tag{3.2.54}$$

By (3.2.46), (3.2.47), and (3.2.50),

$$\begin{aligned}\sum_{t=0}^{T-1}\|\widetilde{\Sigma}_t'-\widetilde{\Sigma}_t\| &\leqslant \sum_{t=0}^{T-1}\|\Sigma_t'-\Sigma_t\|+\sum_{t=0}^{T-1}\|\mu_t'-\mu_t\| \\ &\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\frac{C(\widehat{K})}{\underline{\sigma}_Q}+T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\|+\|\widehat{D}\|^2\|\Sigma^w\|)\right) \\ &\quad\left(\eta(2\rho+1)\|\widehat{B}\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right)+2\eta\|\widehat{D}\|\|\bar{w}\|\|\mu\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\| \\ &\quad+\frac{\rho^T-1}{\rho-1}\left(\sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}}+T\|\widehat{D}\|\|\bar{w}\|\right)\left(\eta\|\widehat{B}\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right).\end{aligned}$$

Given the step-size condition in (3.2.48), we have

$$\begin{aligned}\eta(2\rho+1)\|\widehat{B}\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\| &\leqslant \eta(2\rho+1)\|\widehat{B}\|\left(T\cdot\max_t\left\{\|\nabla_t C(\widehat{K})\|\right\}\right) \\ &\leqslant \frac{(\rho^2-1)\,\underline{\sigma}_Q\underline{\sigma}_X}{2\,(\rho^{2T}-1)\,Dn_1}.\end{aligned} \tag{3.2.55}$$

Then, by Corollary 3.1, (3.2.46), (3.2.51) and (3.2.55),

$$\begin{aligned}
\frac{\|\Sigma_{\widehat{K}'} - \Sigma_{\widehat{K}}\|}{\underline{\sigma}_X} &\leqslant \frac{\rho^{2T}-1}{\rho^2-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_{\widehat{K}_t} - \mathcal{F}_{\widehat{K}'_t}\|\right)\frac{\|\Sigma_0\| + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)}{\underline{\sigma}_X} \\
&\quad + \sum_{t=0}^{T-1}\frac{2\|\widehat{D}\|\|\bar{w}\|\|\mu\|\|\widehat{K}_t - \widehat{K}'_t\|}{\underline{\sigma}_X} \\
&\leqslant \frac{\rho^{2T}-1}{\rho^2-1}(2\rho+1)\|\widehat{B}\|\left(\sum_{t=0}^{T-1}\eta\|\nabla_t C(K)\|\right)\frac{C(K) + \underline{\sigma}_Q T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)}{\underline{\sigma}_Q\underline{\sigma}_X} \\
&\quad + \frac{2\underline{\sigma}_Q\|\widehat{D}\|\|\bar{w}\|\|\mu\|}{\underline{\sigma}_Q\underline{\sigma}_X}\left(\sum_{t=0}^{T-1}\eta\|\nabla_t C(K)\|\right) \\
&\leqslant \frac{1}{2},
\end{aligned}$$

Therefore, the bound of $\|\Sigma_{\widehat{K}'}\|$ in (3.2.54) is given by

$$\|\Sigma_{\widehat{K}'}\| \leqslant \|\Sigma_{\widehat{K}'} - \Sigma_{\widehat{K}}\| + \|\Sigma_{\widehat{K}}\| \leqslant \frac{1}{2}\underline{\sigma}_X + \frac{C(\widehat{K})}{\underline{\sigma}_Q} \leqslant \frac{1}{2}\|\Sigma_{\widehat{K}'}\| + \frac{C(\widehat{K})}{\underline{\sigma}_Q}, \tag{3.2.56}$$

which indicates that $\|\Sigma_{\widehat{K}'}\| \leqslant \frac{2C(\widehat{K})}{\underline{\sigma}_Q}$. Therefore, (3.2.54) gives

$$\begin{aligned}
&1 - \frac{\sum_{t=0}^{T-1}\|\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t\|}{\underline{\sigma}_X} - \eta\|\Sigma_{\widehat{K}'}\|\sum_{t=0}^{T-1}\|R_t + \widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t\| \\
&\geqslant 1 - \frac{(\rho^{2T}-1)}{(\rho^2-1)\,\underline{\sigma}_X}\left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} + T(2\rho\|\widehat{D}\|\|\bar{w}\|\|\mu\| + \|\widehat{D}\|^2\|\Sigma^w\|)\right)\left(\eta(2\rho+1)\|\widehat{B}\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right) \\
&\quad - \frac{2\eta\|\widehat{D}\|\|\bar{w}\|\|\mu\|}{\underline{\sigma}_X}\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\| - \frac{\rho^T-1}{(\rho-1)\underline{\sigma}_X}\left(\sqrt{\frac{TC(\widehat{K})}{\underline{\sigma}_Q}} + T\|\widehat{D}\|\|\bar{w}\|\right)\left(\eta\|\widehat{B}\|\sum_{t=0}^{T-1}\|\nabla_t C(\widehat{K})\|\right) \\
&\quad - \eta\frac{2C(\widehat{K})}{\underline{\sigma}_Q}\sum_{t=0}^{T-1}\|R_t + \widehat{B}_t^\top\widehat{P}_{t+1}\widehat{B}_t\| \\
&= 1 - C_1\eta,
\end{aligned}$$

where C_1 is defined in (3.2.49). So if $\eta \leqslant \frac{1}{2C_1}$, then

$$1 - \frac{\sum_{t=0}^{T-1} \|\widetilde{\Sigma}'_t - \widetilde{\Sigma}_t\|}{\underline{\sigma}_X} - \eta \|\Sigma_{\widehat{K}'}\| \sum_{t=0}^{T-1} \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\| \geqslant 1 - C_1 \eta \geqslant \frac{1}{2} > 0.$$

Hence,

$$C(\widehat{K}') - C(\widehat{K}) \leqslant -\frac{\eta}{2} \left(\frac{4\underline{\sigma}_X^2 \underline{\sigma}_R}{\|\Sigma_{\widehat{K}^*}\| + 1} \right) \left(C(\widehat{K}) - C(\widehat{K}^*) \right),$$

and

$$\begin{aligned} C(\widehat{K}') - C(\widehat{K}^*) &= \left(C(\widehat{K}') - C(\widehat{K}) \right) + \left(C(\widehat{K}) - C(\widehat{K}^*) \right) \\ &\leqslant \left(1 - 2\eta \frac{\underline{\sigma}_X^2 \underline{\sigma}_R}{\|\Sigma_{\widehat{K}^*}\| + 1} \right) \left(C(\widehat{K}) - C(\widehat{K}^*) \right). \end{aligned}$$

Thus, this lemma has been proven. □

Lemma 3.11. *We have that*

$$\sum_{t=0}^{T-1} \|\nabla_t C(\widehat{K})\|^2 \leqslant 4 \left(\frac{TC(\widehat{K})}{\underline{\sigma}_Q} + \left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} \right)^2 \right) \frac{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|}{\underline{\sigma}_X + 1} \left(C(\widehat{K}) - C(\widehat{K}^*) \right),$$

and that

$$\sum_{t=0}^{T-1} \|\widehat{K}_t\| \leqslant \frac{1}{\underline{\sigma}_R} \left(\sqrt{T \cdot \frac{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|}{\underline{\sigma}_X + 1} \left(C(\widehat{K}) - C(\widehat{K}^*) \right)} + \sum_{t=0}^{T-1} \|\widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t\| \right).$$

Proof. Using Lemma 3.6, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \|\nabla_t C(\widehat{K})\|^2 &\leqslant 4 \sum_{t=0}^{T-1} \operatorname{Tr}\left(\widetilde{\Sigma}_t^\top \widetilde{E}_t^\top \widetilde{E}_t \widetilde{\Sigma}_t \right) \\ &\leqslant 4 \sum_{t=0}^{T-1} \|\widetilde{\Sigma}_t\|^2 \operatorname{Tr}(\widetilde{E}_t^\top \widetilde{E}_t) \\ &\leqslant 4 \left(\frac{TC(\widehat{K})}{\underline{\sigma}_Q} + \left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} \right)^2 \right) \sum_{t=0}^{T-1} \operatorname{Tr}(\widetilde{E}_t^\top \widetilde{E}_t). \end{aligned}$$

From Lemma 3.4, we have

$$\begin{aligned} C(\widehat{K}) - C(\widehat{K}^*) &\geqslant (\underline{\sigma}_X + 1) \sum_{t=0}^{T-1} \frac{1}{\|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|} \operatorname{Tr}(\widetilde{E}_t^\top \widetilde{E}_t) \\ &\geqslant \frac{\underline{\sigma}_X + 1}{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|} \sum_{t=0}^{T-1} \operatorname{Tr}(\widetilde{E}_t^\top \widetilde{E}_t), \end{aligned} \tag{3.2.57}$$

hence

$$\sum_{t=0}^{T-1} \|\nabla_t C(\widehat{K})\|^2 \leqslant 4 \left(\frac{T C(\widehat{K})}{\underline{\sigma}_Q} + \left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} \right)^2 \right) \frac{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|}{\underline{\sigma}_X + 1} \left(C(\widehat{K}) - C(\widehat{K}^*) \right).$$

For the second claim, using Lemma 3.4 again,

$$\begin{aligned} \sum_{t=0}^{T-1} \|\widehat{K}_t\| &= \sum_{t=0}^{T-1} \|(R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)^{-1} \widehat{K}_t (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)\| \\ &\leqslant \sum_{t=0}^{T-1} \frac{1}{\sigma_{\min}(R_t)} \|\widehat{K}_t (R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t)\| \\ &\leqslant \sum_{t=0}^{T-1} \frac{1}{\sigma_{\min}(R_t)} \left(\|E_t\| + \|\widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t\| \right) \\ &\leqslant \sum_{t=0}^{T-1} \left(\frac{\sqrt{\operatorname{Tr}(E_t^\top E_t) + F_t^\top F_t}}{\sigma_{\min}(R_t)} + \frac{\|\widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t\|}{\sigma_{\min}(R_t)} \right) \\ &\leqslant \frac{1}{\underline{\sigma}_R} \left(\sqrt{T \cdot \sum_{t=0}^{T-1} \operatorname{Tr}(E_t^\top E_t) + F_t^\top F_t} + \sum_{t=0}^{T-1} \|\widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t\| \right) \\ &\leqslant \frac{1}{\underline{\sigma}_R} \left(\sqrt{T \cdot \frac{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|}{\underline{\sigma}_X + 1} \left(C(\widehat{K}) - C(\widehat{K}^*) \right)} + \sum_{t=0}^{T-1} \|\widehat{B}_t^\top \widehat{P}_{t+1} \widehat{A}_t\| \right). \end{aligned}$$

The second inequality holds by the definition of E_t in (3.2.30), the second last step uses the Cauchy-Schwarz inequality, and the last inequality holds by (3.2.57). □

Theorem 3.4 (Global Convergence of the Gradient-Based Policy Update)**.** *Assume that* $\underline{\sigma}_X > 0$ *and the initial cost* $C(\widehat{K}^0)$ *is finite. Let* $\beta \in (0,1)$ *denote a prescribed confidence level and let* $\varepsilon_M(\beta)$ *be the radius of the Wasserstein ball* $\mathbb{B}_{\varepsilon_M(\beta)}(\widehat{\mathbb{P}}_M)$ *determined by Theorem 3.1. Then, there exists a constant stepsize* η *ensuring (3.2.48) that with probability at least* $1-\beta$ *over the sampling of* $\widehat{\mathbb{P}}_M$*, the gradient-based policy update rule (3.2.25) satisfies*

$$\mathbb{P}^M \left\{ C(\widehat{K}^N) - C(\widehat{K}^*) \leqslant \varepsilon \right\} > 1 - \beta$$

whenever

$$N \geqslant \frac{\|\Sigma_{\widehat{K}^*}\| + 1}{2\eta \underline{\sigma}_X^2 \underline{\sigma}_R} \log \frac{C(\widehat{K}^0) - C(\widehat{K}^*)}{\varepsilon}.$$

Proof. To establish the existence of a positive stepsize η ensuring condition (3.2.48), it suffices to demonstrate that the right-hand side of (3.2.48) admits a strictly positive lower bound. From Lemma 3.11 and by applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1} \|\nabla_t C(\widehat{K})\| \\
&\leqslant \sqrt{T \cdot \sum_{t=0}^{T-1} \|\nabla_t C(\widehat{K})\|^2} \\
&\leqslant \sqrt{4T \left(\frac{TC(\widehat{K})}{\underline{\sigma}_Q} + \left(\frac{C(\widehat{K})}{\underline{\sigma}_Q} \right)^2 \right) \frac{\max_t \|R_t + \widehat{B}_t^\top \widehat{P}_{t+1} \widehat{B}_t\|}{\underline{\sigma}_X + 1} \left(C(\widehat{K}) - C(\widehat{K}^*) \right)}.
\end{aligned} \tag{3.2.58}$$

Observe that for positive scalars $a, b, c, d > 0$, the inequality $d < ab + c$ implies $\frac{1}{d} > \frac{1}{(a+1)(b+1)(c+1)}$, and for $a > 0$ and $n \in \mathbb{N}^+$, we have $\frac{1}{a^n+1} > \frac{1}{(a+1)^n}$. Consequently,

using inequalities (3.2.49), we obtain a positive lower bound on $\frac{1}{C_1}$ that depends explicitly on the system parameters. Next, we show that $\frac{1}{\rho}$ also admits a positive lower bound in terms of the same quantities. From Lemma 3.10, under the condition $\|\widehat{B}_t\|\|\widehat{K}'_t - \widehat{K}_t\| \leqslant \frac{\sigma_Q \sigma_X}{4C(\widehat{K})} \leqslant \frac{1}{2}$, we have

$$\max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}'_t\| \leqslant \max_t \|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\| + \frac{1}{2}.$$

Hence,

$$\begin{aligned}
\rho &= \max\left\{\max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\|, \max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}'_t\|, 1+\xi\right\} \\
&\leqslant \max\left\{\max_{0\leqslant t\leqslant T-1} \|\widehat{A}_t - \widehat{B}_t\widehat{K}_t\| + \frac{1}{2}, 1+\xi\right\} \\
&\leqslant \max\left\{\|\widehat{A}_t\| + \|\widehat{B}_t\| \sum_{t=0}^{T-1} \|\widehat{K}_t\| + \frac{1}{2}, 1+\xi\right\}.
\end{aligned}$$

According to Lemma 3.11, $\sum_{t=0}^{T-1} \|\widehat{K}_t\|$ is uniformly bounded on the sublevel set $\{K : C(K) \leqslant C(\widehat{K}^0)\}$, and Lemma 3.6 yields $\|\widehat{P}_t\| \leqslant \frac{C(\widehat{K})}{\underline{\sigma}_X}$. Therefore, ρ is bounded above by an explicit function of $\|\widehat{A}\|, \|\widehat{B}\|, \|R\|, \underline{\sigma}_X, \underline{\sigma}_R$ and $C(\widehat{K}^0)$, and $\frac{1}{\rho}$ is bounded below by a positive quantity depending on the same terms. By combining the bounds on $\frac{1}{C_1}$ and $\frac{1}{\rho}$, we can define a positive constant $\eta_{\max}$ such that choosing any constant stepsize $0 < \eta \leqslant \eta_{\max}$ ensures condition (3.2.48). Moreover, since each gradient step satisfies $C(\widehat{K}^1) < C(\widehat{K}^0)$, the same $\eta_{\max}$ remains valid along the iterates. Applying Lemma 3.10, we obtain the descent inequality:

$$C(\widehat{K}^1) - C(\widehat{K}^*) \leqslant \left(1 - 2\eta \frac{\underline{\sigma}_X^2 \underline{\sigma}_R}{\|\Sigma_{\widehat{K}^*}\| + 1}\right)\left(C(\widehat{K}^0) - C(\widehat{K}^*)\right),$$

which ensures a strict reduction in cost after the first iteration. By induction, suppose $C(\widehat{K}^n) \leqslant C(\widehat{K}^0)$. The stepsize condition in (3.2.48) continues to hold, and applying

Lemma 3.10 again yields

$$C(\widehat{K}^{n+1}) - C(\widehat{K}^*) \leqslant \left(1 - 2\eta \frac{\underline{\sigma}_X^2 \underline{\sigma}_R}{\|\Sigma_{\widehat{K}^*}\| + 1}\right) \left(C(\widehat{K}^n) - C(\widehat{K}^*)\right).$$

Thus, the cost sequence $C(\widehat{K}^n)$ decreases geometrically, and for any $\varepsilon > 0$, it follows that

$$C(\widehat{K}^N) - C(\widehat{K}^*) \leqslant \varepsilon \text{ whenever } N \geqslant \frac{\|\Sigma_{\widehat{K}^*}\| + 1}{2\eta \underline{\sigma}_X^2 \underline{\sigma}_R} \log \frac{C(\widehat{K}^0) - C(\widehat{K}^*)}{\varepsilon}.$$

By the uniform performance guarantee (3.1.16), with probability at least $1 - \beta$, the DRO objective dominates the true objective simultaneously for all policies and for every iterate $\widehat{K}^i$ and for $\widehat{K}^*$. For this high probability event, the true suboptimality gap at any iterate is bounded by the corresponding empirical gap. Since the first part of the proof already establishes geometric decay of the empirical gap and yields the iteration bound N that makes the empirical gap less than or equal to ε, the same bound applies to the true gap on the same event. Therefore,

$$\mathbb{P}^M \left\{ C(\widehat{K}^N) - C(\widehat{K}^*) \leqslant \varepsilon \right\} \geqslant 1 - \beta,$$

which completes the proof. □

3.3 Numerical Examples

We present two numerical studies to illustrate the performance and properties of the proposed algorithm. The first examines the MV portfolio optimization problem, highlighting how control constraints influence the optimal policy and cost. The second study is a dynamic benchmark tracking problem with additive uncertainty under a DRO framework focusing exclusively on the unknown noise setting and providing a sensitivity analysis with respect to the parameters.

3.3.1 MV Portfolio Selection

This section applies the MV problem to illustrate its formulation as an LQ problem with a scalar state. This example pertains to the finite time horizon scenario, wherein the investor aims to minimize their risk over a span of T periods. As mentioned in the Section 2.3.1, the MV problem can be stated according to the model (2.3.14) as follows:

$$\min_{\{u_t\}_{t=0}^{T-1}} \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \left(u_t^\top R u_t \right) + x_T^2 \middle| x_0 = x \right], \tag{3.3.59}$$

which is subject to

$$x_{t+1} = r_t x_t + P_t u_t. \tag{3.3.60}$$

Suppose the investor requires the proportion of each risky asset to be between 10% and 20% of the total wealth throughout all periods. This constraint can be expressed as:

$$0.1|x_t| \leqslant u_t^i \leqslant 0.2|x_t|, \quad \forall i = 1, \cdots, n,$$

where x_t is the total wealth at time t, and u_i represents the amount invested in risky asset i at time t. The problem (MV) is evidently a specific instance of the problem presented in Section 3.1 by setting $x_t \in \mathbb{R}$, $u_t \in \mathbb{R}^n$, $Q_t \in \mathbb{R}$, $R_t \in \mathbb{R}^{n\times n}$, $A_t = r_t$, $B_t = P_t$, $Q_t = 0$ and $D_t = 0$ for $t = 0, \cdots, T-1$ and $Q_T = 1$. For this state-dependent upper and lower bounds constraint case, we have $H_t = \begin{bmatrix} \mathbf{I}_n \\ -\mathbf{I}_n \end{bmatrix}$ and $d_t = \begin{bmatrix} \bar{d}_t \\ -\underline{d}_t \end{bmatrix}$ with $\bar{d}_t = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}$ and $\underline{d}_t = \begin{bmatrix} -0.1 \\ -0.1 \end{bmatrix}$. To ensure that our selected parameters meet the criteria for our LQ framework, we randomly choose the following parameters: $T = 5$ and $n = 3$. For the state transition matrix A_t, we establish the risk-free rate 6%, resulting in $A_t = 1.06$. The expected excess returns vector B_t for the risky assets is expressed as $B_t = \begin{bmatrix} 14.8\% & 16\% & 9.8\% \end{bmatrix}$. At each time step, R_t is chosen as a diagonal positive definite matrix. The initial wealth x_0 follows normal distribution,

$x_0 \sim \mathcal{N}(10, 5 \times 10^{-2})$, with the currency expressed in thousands of dollars. All entries in the initial policy matrix $K^0 \in \mathbb{R}^n$ are 0.08.

Results and Analysis

We evaluate the performance of the proposed policy gradient algorithm under both known and unknown parameter settings. Performance is reported using the normalized error.

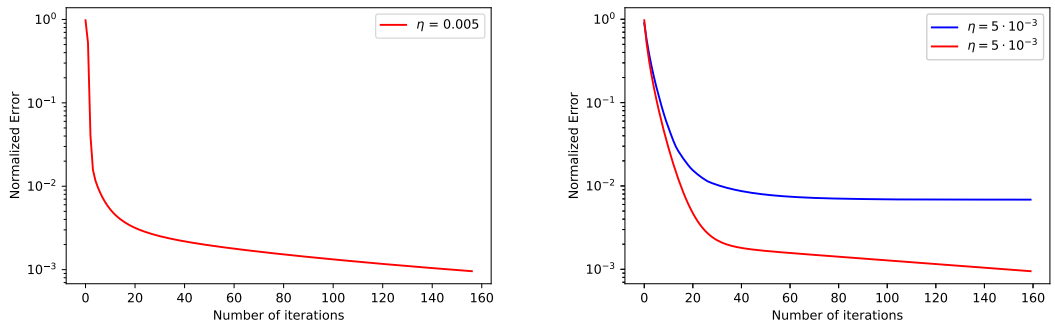


Figure 3.1: Performance of the policy gradient algorithm

Figure 3.1 summarizes two aspects: convergence and the impact of the control constraint. The left shows fast convergence—the normalized error drops below 10^{-3} in roughly 160 iterations. The right compares constrained (blue) and unconstrained (red) updates under the same backtest: the constrained variant converges more slowly and settles at a slightly higher error level that is consistent with a practical risk limit that tempers leverage and tail risk at a modest performance cost.

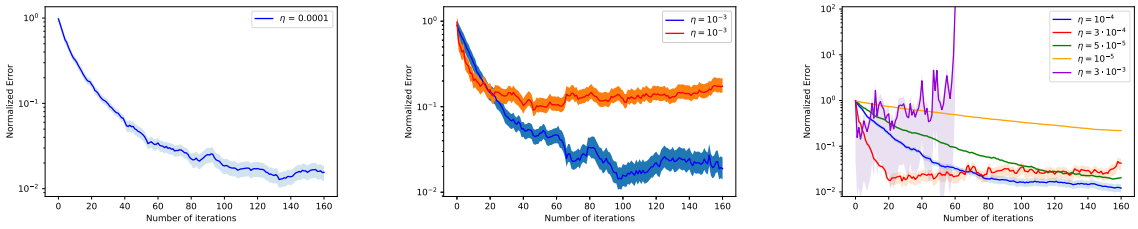


Figure 3.2: Comparison between constrained and unconstrained cases

When system parameters are unknown, Figure 3.2 collects the corresponding results: convergence, the constrained (orange) versus unconstrained (blue) comparison, and the step-size study. Convergence is noticeably slower because the algorithm must jointly estimate dynamics and optimize control; after a similar number of iterations, the error is around 10^{-2} (left). The gap between constrained and unconstrained cases widens relative to the known parameter case due to added estimation uncertainty (middle). The step-size analysis (right) shows that η in the range $[10^{-5}, 3 \times 10^{-3}]$ materially affects learning: very small η yields steady but slow progress with low fluctuation, while overly large η can cause divergence.

3.3.2 Dynamic Benchmark Tracking

Based on Section 2.3.3, with full state and no hard boxes, the classical LQ feedback is from the Riccati recursion. To hedge misspecified unknown or light-tailed w_t, we also use a Wasserstein penalized DRO-LQ (3.1.19) for which linear policies remain optimal and the backward recursion admits a simple modification via a positive definite factor Δ.

Data Processing

We use publicly available daily data from Yahoo Finance for the benchmark (e.g., SPY) and a set of sector ETFs (e.g., XLB, XLE, XLF, XLI, XLK, XLP, XLU, XLV, XLY, XLRE, XLC) from 2015/01/01 to 2025/09/30. See Figure 3.3, which plots normalized-adjusted close series for SPY and selected sectors over 2015–2025, thus illustrating broad co-movement with intermittent sector rotations for context only. Series are aligned on U.S. trading days; rows with a missing benchmark price or any missing sector price are dropped. Let P_t^{bench} denote the benchmark-adjusted close and $P_t^{(j)}$ the adjusted close of sector $j \in \{1, \cdots, m\}$. Returns are computed as log differences $r_t^{\text{bench}} = \log P_t^{\text{bench}} - \log P_{t-1}^{\text{bench}}$ and $r_t^{(j)} = \log P_t^{(j)} - \log P_{t-1}^{(j)}$.

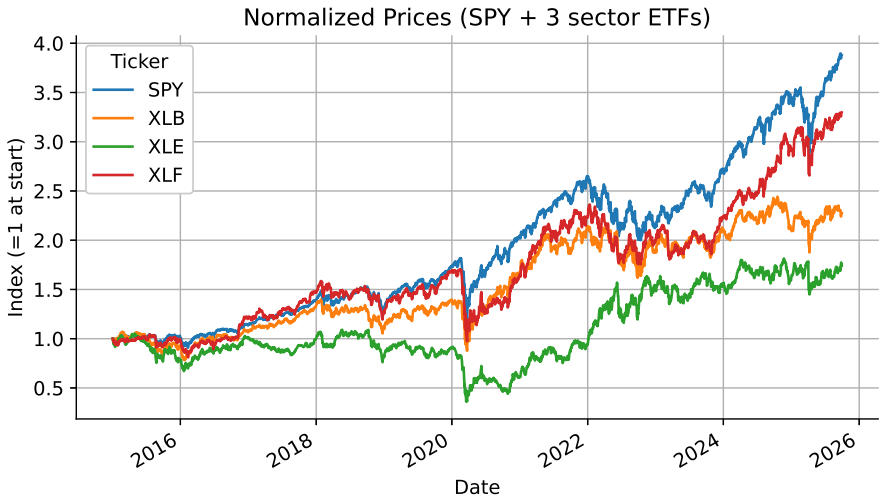


Figure 3.3: Normalized prices for ETFs and the benchmark

To stabilize later volatility estimates, we optionally trim each return series at the 1^{st} /99^{th} percentiles within the calibration sample only. Table 3.1 shows the log-return output for benchmark SPY and the other three sector ETFs.

Time-varying target exposures $a_t^* \in \mathbb{R}^m$ are estimated by rolling ordinary least squares over a fixed lookback length $L_{\text{ols}} = 60$ days. For each $t > L_{\text{ols}}$, we solve

$$a_t^* \in \arg\min_{a \in \mathbb{R}^m} \sum_{\tau=t-L_{\text{ols}}+1}^{t} \left(r_\tau^{\text{bench}} - r_\tau^{\text{sec}\top} a\right)^2, \quad r_\tau^{\text{sec}} = \left(r_\tau^{(1)}, \cdots, r_\tau^{(m)}\right)^\top,$$

without intercept so that a_t^* captures pure sector exposures (see Figure 3.4, left). If the normal matrix is ill-conditioned, a small ridge penalty $\alpha\|a\|_2^2$ (with $\alpha > 0$) can be added; by default, we take $\alpha = 0$. Long-only or budget constraints are optional post-projections and are not required for the LQ formulation.

The noise feeding the state equation is then taken directly from the data as the one-step drift in targets, $w_t = a_{t+1}^* - a_t^*$, which appears approximately zero-mean and independent of the initial state (see Figure 3.4, right). The tracking state x_t is the deviation between actual sector exposures and a_t^*; unless otherwise stated, we initialize $x_0 = \mathbf{0}_{m\times 1}$ at the start of the evaluation sample. Risk scaling uses the

Date	SPY	XLB	XLC	XLE
2025-09-23	-0.005458	-0.003214	-0.001435	0.017077
2025-09-24	-0.003187	-0.012287	-0.008311	0.012923
2025-09-25	-0.004624	-0.013122	-0.003412	0.008926
2025-09-26	0.005713	0.011548	0.009694	0.009173
2025-09-29	0.002806	0.003820	0.003211	-0.018542

Table 3.1: Daily log-returns for SPY and three ETFs over the latest five trading days

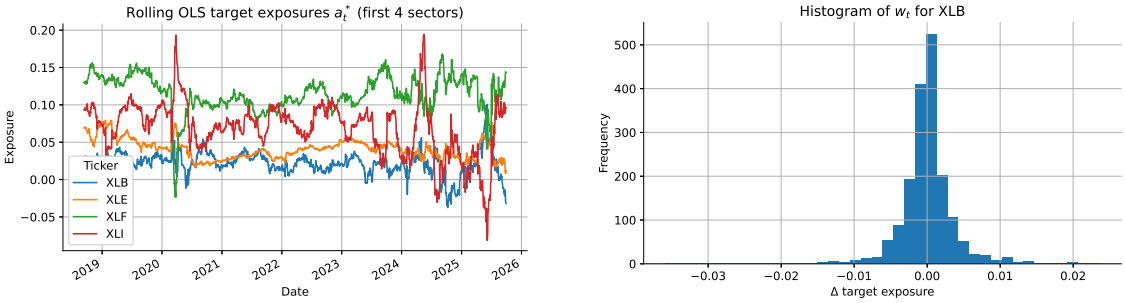


Figure 3.4: Empirical features for the tracking model

benchmark's realized volatility. Let $\widehat{\sigma}_{\text{daily}}$ be the rolling standard deviation of r_t^{bench} computed over L_{vol} trading days, which is the same length as the OLS window. For a horizon of T decision steps, the per-step scale is

$$\sigma_{\text{step}} = \widehat{\sigma}_{\text{daily}} / \sqrt{T},$$

where $T = 120$. We set $Q_t = \lambda_{\text{step}} \sigma_{\text{step}}^2 \mathbf{I}_m$ for all t with $\lambda_{\text{step}} = 10$ and $\bar{q} = 10^6$ for Q_T. Trading cost weights are diagonal, $R_t = \text{diag}(c_1, \cdots, c_m)$, with each c_j calibrated from observable liquidity via

$$c_j = \kappa_{\text{tc}} \frac{(S_0^{(j)})^2}{\text{ADV}_j},$$

where $S_0^{(j)}$ is the most recent adjusted close of sector j, ADV_j is the $L_{\text{adv}} = 60$ days moving average of its share volume, and $\kappa_{\text{tc}} > 0$ sets the overall trading effort scale (typical order $10^{-6} - 10^{-5}$). This specification is dimensionally consistent when x_t and u_t are measured in shares; if exposures are in dollars, the same form applies after

the natural rescaling by price (Benidis et al. (2018)).

To prevent look-ahead, $\widehat{\sigma}_{\text{daily}}$, a_t^*, and ADV_j are computed using only information available up to time t. Fixing Q_t, R_t, λ_{step} , κ_{tc}, and $\bar{q}$, we calibrate on a training period, then evaluate on a hold-out period using the realized noise path $\{w_t\}$; Q_t and R_t are kept at their last calibrated values. Dates with missing prices, hence undefined w_t, are omitted. The full pipeline is deterministic and thus fully reproducible.

Evaluation Results

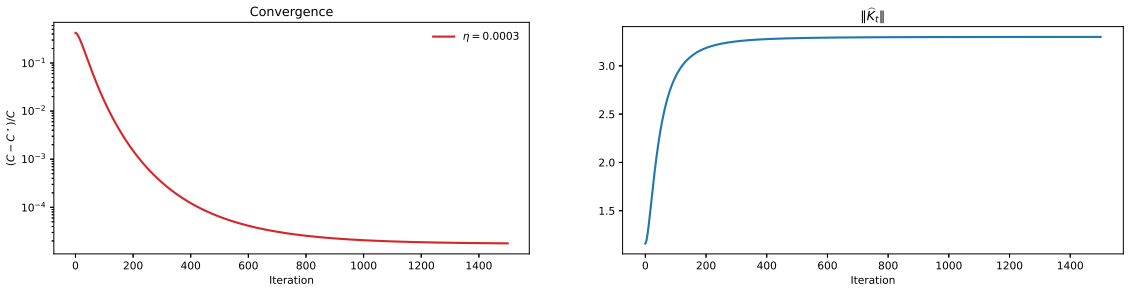


Figure 3.5: DRO-LQ: policy gradient convergence and gain stabilization

Two complementary views of the same learning run for the tracker in Figure 3.5. The left (normalized error) shows a smooth, monotone descent over roughly three orders of magnitude from about 10^0 down to the low 10^{-4} range before flattening into a stable plateau, which indicates the policy gradient has essentially reached the Riccati benchmark and further gains are marginal. The right one (gain size) starts with $\|\widehat{K}_t\|$ from initial $\|\widehat{K}_0\|$, then contracts steadily to a tight band a little above 3, where it remains flat; this stabilization lines up in time with the error plateau, confirming the controller has settled into a steady, well-conditioned feedback policy.

We plot the percentage difference versus the baseline because the raw convergence curves sit on very different scales, making tiny gaps visually indistinguishable. Normalizing to a relative metric makes the series dimensionless, isolating the marginal effect of each parameter from mere scale shifts. We then apply a running mean to

the difference series to suppress high-frequency noise, from sampling error and line search, while preserving the underlying trend.

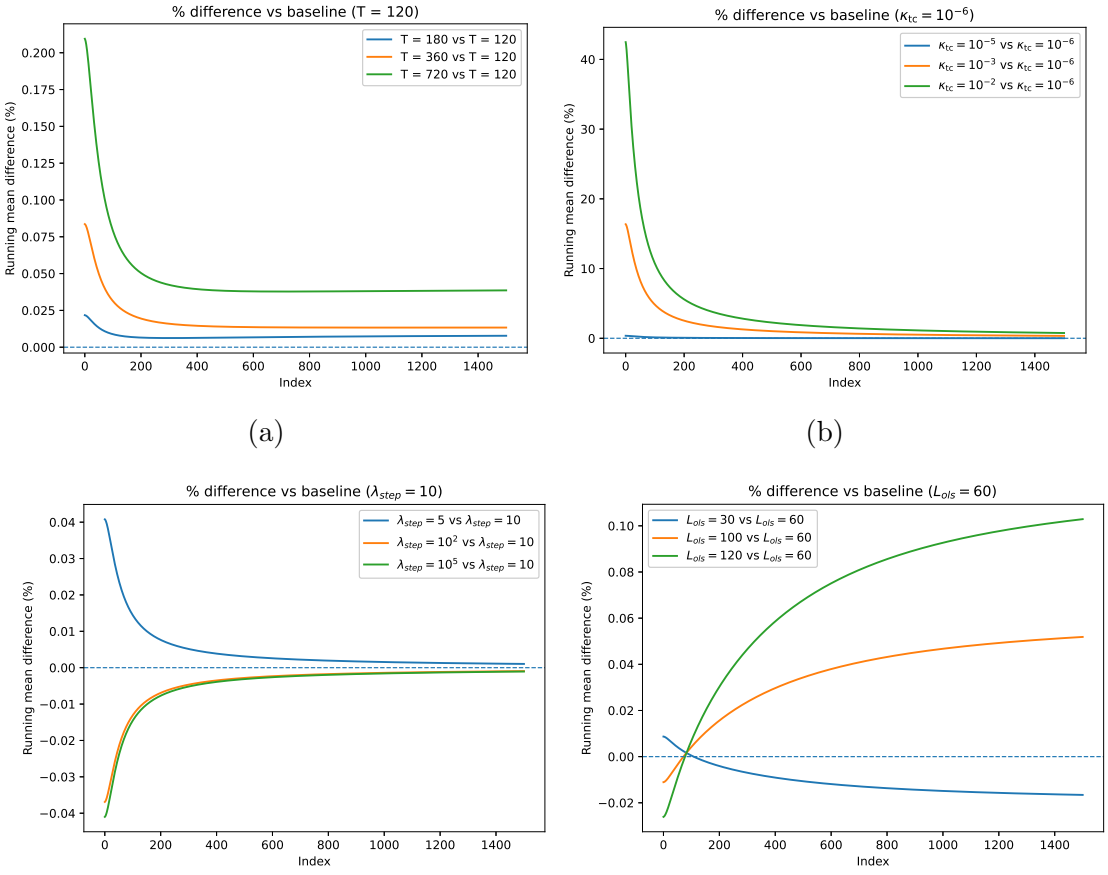


Figure 3.6: DRO-LQ: sensitivity analysis

T. In Figure 3.6(a), the running-mean gaps start large and positive and then converge toward a fixed level; moreover, a larger T produces a larger initial gap consistent with a modest but systematic horizon effect rather than instability, and with attenuating marginal benefit as T grows.

κ_{tc}. When κ_{tc} rises materially above baseline, the start gap can approach 40%, then declines and is near zero in the tail; larger κ_{tc} penalizes control effort more, shrinking

gains and slowing trading so curves separate early and flatten as the optimizer settles.

λ_{step} . Smaller λ yields a positive signed difference; much larger λ yields a negative one; the start gap is clear, but the tail drifts to 0, and pushing λ from 10 to 10^3–10^5 changes results only slightly.

$\mathbf{L_{ols}}$. The start is small, but the tail becomes larger; from Figure 3.6(d), we can find larger L produces a larger positive tail. Since L_{ols} sets the regression window for a_t^* with $w_t = \Delta a_t$, longer windows smooth exposures and shrink empirical Σ^w; shorter windows are more reactive and noisier, so the smoothed estimator induces a stable offset that persists.

Chapter 4

Decentralized Reinforcement Learning Policy for Linear Quadratic Problem with Entropy Regularization

In this chapter, we focus on the analysis and evaluation of the IPO method for discounted entropy-regularized multi-agent LQ problems over an infinite time horizon with multiplicative noise. Inspired by techniques from Fazel et al. (2018) and Guo et al. (2023), we rigorously analyze the IPO method from decentralized to centralized control and establish that it achieves a super-linear convergence rate if it enters a local region around the optimal policy. This convergence property is critical for ensuring efficient learning in a scalar state and high-dimensional and constrained scenarios, which is shown in the numerical example. We demonstrate the practicality of the IPO method through its application to the MV optimization and the optimal liquidation problem (Ni et al. (2014), Wu et al. (2018)), thus showing promising results for practical financial decision-making problems. These evaluations highlight the potential of IPO to improve decision-making processes in complex environments. By addressing these aspects, our work provides a comprehensive understanding of the IPO method for the scalability of LQ problems with multiplicative noise, thereby

offering theoretical insights and practical implications for more complex cases.

4.1 Decentralized Multi-Agent LQ Problem Setup and Solution

4.1.1 Problem Formulation

In the decentralized setting, the system consists of N independent agents. Each agent $i \in \{1, \ldots, N\}$ seeks to minimize its own long-term entropy-regularized cost. Thus, each admissible policy $\pi \in \Pi$ maps a state $x_i \in \mathcal{X}$ to a randomized action in $\mathcal{U}$. For a given admissible policy $\pi \in \Pi$, the corresponding Shannon entropy for agent i is changed to:

$$\mathcal{H}(\pi(\cdot|x_i)) = -\int_{\mathcal{U}} \pi(u_i|x_i) \log \pi(u_i|x_i) du_i \quad \forall x_i \in \mathcal{X},$$

where $\mathbb{E}[-\log \pi(u_i|x_i)]$ can be used to generalize the above. The decision-maker aims to find an optimal policy by minimizing the following objective function for each agent $i \in \{1, \ldots, N\}$:

$$\min_{\pi \in \Pi} \sum_{i=1}^{N} J_{i,\pi}(x_i), \tag{4.1.1}$$

where value function $J_{i,\pi}$ is given with negative entropy:

$$J_{i,\pi}(x_i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(x_{i,t}^\top Q_i x_{i,t} + u_{i,t}^\top R_i u_{i,t} + \tau \log \pi\left(u_{i,t}|x_{i,t}\right) \right) \middle| x_{i,0} = x_i \right], \tag{4.1.2}$$

and for $t = 0, 1, 2, \cdots$,

$$x_{i,t+1} = A_i x_{i,t} + B_i u_{i,t} + (C_i x_{i,t} + D_i u_{i,t}) w_t, x_{i,0} \sim \mathcal{D}_{i,0}. \tag{4.1.3}$$

The agents are decoupled in physical dynamics but subject to a common scalar noise. For each agent i, the basic settings in Section 2.1 for parameters still work. We introduce $\gamma \in (0, 1)$, which denotes the discount factor, and $\tau > 0$, which denotes the regularization parameter in addition.

4.1.2 Centralized Formulation for Dynamics

We stack all agents' states and controls into a single vector (x_t, u_t) only for notation and analysis. Execution is decentralized: agent i selects $u_{i,t}$ using only its local state $x_{i,t}$. Moreover, we focus on a common policy (shared parameter) class where all agents use the same policy π_{common} (equivalently the same feedback gain K_{common} in Section 4.1.4), which is natural for homogeneous agents and yields a scalable controller.

We define the aggregated state and control vectors:

$$x_t = \begin{bmatrix} x_{1,t} \\ \vdots \\ x_{N,t} \end{bmatrix} \in \mathbb{R}^{mN}, \quad u_t = \begin{bmatrix} u_{1,t} \\ \vdots \\ u_{N,t} \end{bmatrix} \in \mathbb{R}^{nN},$$

and construct the block matrices:

$$\begin{aligned} A &= \operatorname{diag}(A_1, \cdots, A_N), \\ B &= \operatorname{diag}(B_1, \cdots, B_N), \\ C &= \operatorname{diag}(C_1, \cdots, C_N), \\ D &= \operatorname{diag}(D_1, \cdots, D_N), \\ Q &= \operatorname{diag}(Q_1, \cdots, Q_N), \\ R &= \operatorname{diag}(R_1, \cdots, R_N). \end{aligned}$$

With these augmented matrices, the global dynamics are

$$x_{t+1} = Ax_t + Bu_t + (Cx_t + Du_t)\, w_t.$$

The global cost function is simply the sum of the individual agents' costs:

$$J_\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^\top Q x_t + u_t^\top R u_t + \tau \log \pi\left(u_t | x_t\right) \right) \right], \tag{4.1.4}$$

where $\pi(u_t|x_t) = \prod_{i=1}^{N} \pi(u_{i,t}|x_{i,t})$ is the product of decentralized policies. The global agent gathers the decoupled dynamics and additive costs of agents. Under the augmented system representation, we solve for such an equilibrium by treating the collection of agents as a single system with block-structured dynamics and cost. The optimal value function $J^* : \mathcal{X} \to \mathbb{R}$ is defined as

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x). \tag{4.1.5}$$

The present model is a decentralized shared policy benchmark with decoupled physical dynamics. Its multi-agent character comes from the information structure and the requirement of a common implementable controller, rather than from direct state coupling across agents.

4.1.3 Optimal Strategy

The following theorem establishes the explicit expression for the optimal control policy and the corresponding optimal value function: the optimal policy is characterized as a multivariate Gaussian distribution, with the mean linear in the state x and a constant covariance matrix.

Since the agents are decoupled, the problem reduces to solving N decoupled single-agent entropy-regularized LQ problems.

Theorem 4.1 (Optimal value function and optimal policy)**.** *The optimal value function $J^* : \mathcal{X} \to \mathbb{R}$ in (4.1.5) can be expressed as $J^*(x) = x^\top P x + q$ with $P \in \mathbb{R}^{mN \times mN}$ and a constant vector $q \in \mathbb{R}$ satisfying*

$$\begin{aligned} P = {} & Q + K_{global}^{*\top} R K_{global}^{*} + \gamma \left(A - B K_{global}^{*}\right)^\top P \left(A - B K_{global}^{*}\right) \\ & + \gamma \left(C - D K_{global}^{*}\right)^\top W P \left(C - D K_{global}^{*}\right), \\ q = {} & \frac{1}{1-\gamma} \left[\operatorname{Tr}\left(\Sigma_{global}^{*} \left(R + \gamma B^\top P B + \gamma D^\top W P D\right)\right) - \frac{\tau}{2} \left(nN + \log\left((2\pi)^{nN} \det \Sigma_{global}^{*}\right)\right) \right], \end{aligned} \tag{4.1.6}$$

where

$$\begin{aligned} K^*_{global} &= \gamma(R + \gamma B^\top P B + \gamma D^\top W P D)^{-1}(B^\top P A + D^\top W P C), \\ \Sigma^*_{global} &= \frac{\tau}{2}(R + \gamma B^\top P B + \gamma D^\top W P D)^{-1}. \end{aligned} \tag{4.1.7}$$

For any $x \in \mathcal{X}$, the corresponding optimal policy π^* to (4.1.5) is a Gaussian policy: $\pi^*(\cdot|x) = \mathcal{N}\left(-K^*_{global}\, x, \Sigma^*_{global}\right)$.

Proof of Theorem 4.1 relies on the following lemma, which establishes the optimal solution for the one-step reward function, given the presence of entropy regularization in the reward.

Lemma 4.1. *For any given symmetric positive definite matrix $M \in \mathbb{R}^{nN \times nN}$ and vector $b \in \mathbb{R}^{nN}$, the optimal solution $p^* \in \mathcal{P}(\mathcal{U})$ to the following optimization problem is a multivariate Gaussian distribution with covariance $\frac{\tau}{2}M^{-1}$ and mean $-\frac{1}{2}M^{-1}$ b:*

$$\begin{aligned} \min_{p \in \mathcal{P}(\mathcal{U})} \quad & \mathbb{E}_{u \sim p(\cdot)}\left[u^\top M u + b^\top u + \tau \log p(u)\right], \\ \text{s. t.} \quad & \int_{\mathcal{U}} p(u)du = 1, \\ & p(u) \geqslant 0, \quad \forall u \in \mathcal{U}. \end{aligned}$$

Proof of lemma 4.1. Let λ be a Lagrangian multiplier to the constraint $\int_{\mathcal{U}} p(u)du = 1$. Consider

$$\begin{aligned} F(p, \lambda) &= \int_{\mathcal{U}} \left(u^\top M u + b^\top u + \tau \log p(u)\right) p(u)du + \lambda\left(\int_{\mathcal{U}} p(u)du - 1\right) \\ &= \int_{\mathcal{U}} \left(u^\top M u + b^\top u + \tau \log p(u) + \lambda\right) p(u)du - \lambda \\ &= \int_{\mathcal{U}} \mathcal{L}(u, p(u), \lambda)du - \lambda, \end{aligned}$$

where $\mathcal{L}(u, v, \lambda) = \left(u^\top M u + b^\top u\right) v + \tau v \log v + \lambda v$. Since $\frac{\partial \mathcal{L}}{\partial v} = \lambda + u^\top M u + b^\top u + \tau + \tau \log v$, by setting $\frac{\partial \mathcal{L}}{\partial v} = 0$, we then get $v \propto e^{-\frac{1}{\tau}\left(u^\top M u + b^\top u\right)}$. Thus, the optimal solution p^* is a multivariate Gaussian distribution with $\mathcal{N}\left(-\frac{1}{2}M^{-1}b, \frac{\tau}{2}M^{-1}\right)$. □

Proof of Theorem 4.1. By definition of J^* in (4.1.5),

$$J^*(x) = \min_{\pi \in \Pi} \mathbb{E}_\pi \left\{ x^\top Q x + u^\top R u + \tau \log(\pi(u|x)) + \gamma J^*(Ax + Bu + (Cx + Du)(Cx + Du)w) \right\}, \tag{4.1.8}$$

where the expectation is taken with respect to $u \sim \pi(\cdot|x)$ and the noise term w, with mean 0 and covariance W. By stipulating

$$J^*(x) = x^\top P x + q, \tag{4.1.9}$$

for a positive definite $P \in \mathbb{R}^{mN \times mN}$ and $q \in \mathbb{R}$ and substituting into (4.1.8), we can obtain the optimal value function with a dynamic programming principle:

$$\begin{aligned}
& J^*(x) \\
= \; & x^\top Q x \\
& + \min_\pi \mathbb{E}_\pi \left\{ u^\top R u + \tau \log(\pi(u|x)) \right. \\
& \left. + \gamma \left[(Ax + Bu + (Cx + Du)w)^\top P (Ax + Bu + (Cx + Du)w) + q \right] \right\} \qquad (4.1.10) \\
= \; & x^\top Q x + \gamma x^\top (A^\top P A + C^\top W P C) x + \gamma q \\
& + \min_\pi \mathbb{E}_\pi \left\{ u^\top (R + \gamma B^\top P B + \gamma D^\top W P D) u \right. \\
& \left. + \tau \log(\pi(u|x)) + 2\gamma u^\top (B^\top P A + D^\top W P C) x \right\}.
\end{aligned}$$

By then applying Lemma 4.1 to (4.1.10) with $M = R + \gamma B^\top P B + \gamma D^\top W P D$ and

$b = 2\gamma(B^\top PA + D^\top WPC)x$, one can get the optimal policy at state x:

$$\begin{aligned}
&\pi^*(\cdot|x) \\
&= \mathcal{N}\left(-\gamma(R + \gamma B^\top PB + \gamma D^\top WPD)^{-1}(B^\top PA + D^\top WPC)x,\right. \\
&\left.\frac{\tau}{2}(R + \gamma B^\top PB + \gamma D^\top WPD)^{-1}\right) \\
&= \mathcal{N}\left(-K^*_{\text{global}}\, x, \Sigma^*_{\text{global}}\right),
\end{aligned} \tag{4.1.11}$$

where $K^*_{\text{global}}, \Sigma^*_{\text{global}}$ are defined in (4.1.7). To derive the associated optimal value function, we first calculate the entropy of policy π^* at any state $x \in \mathcal{X}$:

$$\mathbb{E}_{\pi^*}\left[-\log\left(\pi^*(u|x)\right)\right] = -\int_{\mathcal{U}} \log\left(\pi^*(u|x)\right)\pi^*(u|x)du = \frac{1}{2}\left(nN + \log\left((2\pi)^{nN}\det\Sigma^*_{\text{global}}\right)\right). \tag{4.1.12}$$

Substitute (4.1.11) and (4.1.12) into (4.1.10) to get

$$\begin{aligned}
J^*(x) = x^\top\Big(&Q + K^{*\top}_{\text{global}} RK^*_{\text{global}} + \gamma\left(A - BK^*_{\text{global}}\right)^\top P\left(A - BK^*_{\text{global}}\right) \\
&+ \gamma\left(C - DK^*_{\text{global}}\right)^\top WP\left(C - DK^*_{\text{global}}\right)\Big)x \\
&+ \text{Tr}\left(\Sigma^*_{\text{global}}\, R\right) - \frac{\tau}{2}\left(nN + \log\left((2\pi)^{nN}\det\Sigma^*_{\text{global}}\right)\right) \\
&+ \gamma\left(\text{Tr}(\Sigma^*_{\text{global}}\, B^\top PB) + \text{Tr}(\Sigma^*_{\text{global}}\, D^\top WPD) + q\right).
\end{aligned}$$

By combining this with (4.1.9), the proof is complete. □

Remark 4.1. *The constant q in $J^*(x) = x^\top Px + q$ collects state-independent terms induced by the entropy-regularized stochastic policy (i.e., $\Sigma^*_{global} \neq 0$) and the entropy contribution. In particular, when $\tau = 0$, the optimal policy becomes deterministic ($\Sigma^*_{global} = 0$) and the state-independent terms vanish; consequently, $q = 0$ and J^* reduces to a purely quadratic form.*

4.1.4 From Global Optimal Policy to a Decentralized Common Gain

Definition 5 (Decentralized system and common policy admissibility)**.** *We impose a common policy restriction: all agents use the same decision rule and shared parameters. In particular, within the linear Gaussian class, an admissible common decentralized policy takes the form*

$$\pi(u_t|x_t) = \mathcal{N}\big(-(\mathbf{I}_N \otimes K_{common})\, x_t,\ \mathbf{I}_N \otimes \Sigma_{common}\big),$$

for some shared gain $K_{common} \in \mathbb{R}^{n\times m}$ (and covariance $\Sigma_{common} \succeq 0$). Equivalently, each agent implements locally:

$$u_{i,t} \sim \mathcal{N}\big(-K_{common}\, x_{i,t},\ \Sigma_{common}\big), \qquad i = 1, \dots, N.$$

In the stacked representation of N agents, we derive the benchmark optimal policy by solving an entropy-regularized LQ control problem. The resulting optimal policy is Gaussian:

$$\pi^*(u_t|x_t) = \mathcal{N}(-K^*_{\text{global}}\, x_t, \Sigma^*_{\text{global}}),$$

where $K^*_{\text{global}} \in \mathbb{R}^{nN\times mN}$ is obtained from the ARE in Theorem 4.1. However, our execution constraint is more restrictive than this benchmark: we require a shared decentralized controller, that is, a feedback matrix of the scalable form K_{common} and Σ_{common}. We therefore seek K_{common} that minimizes the aggregate cost $\sum_{i=1}^{N} J_i$ within this implementable class. The multi-agent aspect here is the implementability constraint and the shared policy restriction rather than coupling in the physical dynamics. To find the best fit K^*_{common} that approximates the global policy, we match the affine term in the optimal policy

$$u^*_t = -(\mathbf{I}_N \otimes K^*_{\text{common}})\, x_t,$$

where $K^*_{\text{common}} \in \mathbb{R}^{n\times m}$ is a shared gain matrix across agents and satisfies (4.1.7):

$$\frac{1}{\gamma}(R+\gamma B^\top PB+\gamma D^\top WPD)(\mathbf{I}_N \otimes K^*_{\text{common}}) = B^\top PA + D^\top WPC.$$

To solve for K^*_{common} from the matrix equation, we employ the vectorization technique to convert the matrix equation into a linear system. Let

$$F = \frac{1}{\gamma}(R+\gamma B^\top PB+\gamma D^\top WPD) \in \mathbb{R}^{nN\times nN},$$
$$G = B^\top PA + D^\top WPC \in \mathbb{R}^{nN\times mN},$$

and denote their row-wise block structures as

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_N \end{bmatrix}, \quad G = \begin{bmatrix} G_1 \\ \vdots \\ G_N \end{bmatrix},$$

where each $F_j \in \mathbb{R}^{1\times nN}$ and $G_j \in \mathbb{R}^{1\times mN}$. Then the matrix equation is equivalent to a system of equations:

$$F_j(\mathbf{I}_N \otimes K^*_{\text{common}}) = G_j, \quad \forall j \in \{1,\cdots,N\}.$$

Since $\mathbf{I}_N \otimes K^*_{\text{common}}$ is a block-diagonal matrix, each vector F_j and G_j can be partitioned into subvectors compatible with the dimensions of K^*_{common}. Suppose $K^*_{\text{common}} \in \mathbb{R}^{n\times m}$ and $\mathbf{I}_N$ is the identity matrix; then we have

$$F_j = \begin{bmatrix} F_{j,1} & \cdots & F_{j,N} \end{bmatrix},$$
$$G_j = \begin{bmatrix} G_{j,1} & \cdots & G_{j,N} \end{bmatrix},$$

where each $F_{jk} \in \mathbb{R}^{1\times n}$ and $G_{jk} \in \mathbb{R}^{1\times m}$ for $k \in \{1,\cdots,N\}$. Then the system is equivalent to the set of equations in terms of K^*_{common}:

$$F_{j,k}K^*_{\text{common}} = G_{j,k}.$$

Since both F and G are block-diagonal matrices with agent-specific structure, the matrix product only yields nontrivial equations when j indexes the block rows that align with the diagonal blocks of F and G. Specifically, only a subset of rows in F and G are involved in multiplying K^*_{common} , namely those rows where the row index j lies in the interval:

$$j \in \{1 + (k-1)n, 2 + (k-1)n, \cdots, nk\}, \qquad \text{for each } k \in \{1, 2, \cdots, N\}.$$

By collecting these active block rows and stacking them, we construct a linear system of the form

$$F_{\text{reshaped}}\, K^*_{\text{common}} = G_{\text{reshaped}}\ ,$$

where

$$F_{\text{reshaped}} = \begin{bmatrix} F_{1,1} \\ \vdots \\ F_{n,1} \\ \vdots \\ F_{1+(k-1)n,k} \\ \vdots \\ F_{nk,k} \\ \vdots \\ F_{(1+(N-1)n),N} \\ \vdots \\ F_{nN,N} \end{bmatrix} \in \mathbb{R}^{nN \times n}, \quad G_{\text{reshaped}} = \begin{bmatrix} G_{1,1} \\ \vdots \\ G_{n,1} \\ \vdots \\ G_{1+(k-1)n,k} \\ \vdots \\ G_{nk,k} \\ \vdots \\ G_{(1+(N-1)n),N} \\ \vdots \\ G_{nN,N} \end{bmatrix} \in \mathbb{R}^{nN \times m}.$$

Since F_{reshaped} has full column rank, we obtain the optimal common gain K_{common} by solving the linear system via the least-squares method:

$$K^*_{\text{common}} = F^{\dagger}_{\text{reshaped}}\, G_{\text{reshaped}}\ , \tag{4.1.13}$$

where $F^{\dagger}_{\text{reshaped}} = \left(F^{\top}_{\text{reshaped}}\, F_{\text{reshaped}}\right)^{-1} F^{\top}_{\text{reshaped}}$, which denotes the Moore-Penrose pseudoinverse of F_{reshaped} .

Remark 4.2 (Common covariance)**.** *The common covariance Σ^*_{common} is defined as the Frobenius norm projection of Σ^*_{global} onto the implementable class $\{\mathbf{I}_N \otimes \Sigma\}$. This mirrors the construction of the common gain K^*_{common} and enforces a shared local exploration structure across agents.*

In the subsequent analysis, we denote the decentralized feedback gain matrix in the full system dimension as

$$K^* = \mathbf{I}_N \otimes K^*_{\text{common}}\,, \text{ and } \Sigma^* = \mathbf{I}_N \otimes \Sigma^*_{\text{common}}\,,$$

which is a block-diagonal matrix in which each agent shares the same local gain K^*_{common} and covariance Σ^*_{common}. This notation simplifies expressions and emphasizes the structure of the decentralized controller across the agent population.

4.2 Analysis of Value Function

Throughout the analysis, we assume that there exists $\rho_1, \rho_2 \in (0, \frac{1}{\sqrt{\gamma}})$ satisfying $\|A - BK^*\| \leqslant \rho_1$ and $\|C - DK^*\| \leqslant \rho_2$ where K^* is the optimal solution in Theorem 4.1. We consider the following domain Ω, that is, the admissible control set, for $(K, \Sigma) : \Omega = \left\{K \in \mathbb{R}^{nN\times mN}, \Sigma \in \mathbb{R}^{nN\times nN} : \|A - BK\| \leqslant \rho_1, \|C - DK\| \leqslant \rho_2, \Sigma \succ 0, \Sigma^\top = \Sigma\right\}$. For any $x_0 \in \mathcal{X}$ following the initial distribution $\mathcal{D}_0$, we assume that $\mathbb{E}_{x_0\sim\mathcal{D}_0}\left[x_0 x_0^\top\right]$ exists and $\mu = \sigma_{\min}\left(\mathbb{E}_{x_0\sim\mathcal{D}_0}\left[x_0 x_0^\top\right]\right) > 0$. For any $(K, \Sigma) \in \Omega$, we define $S_{K,\Sigma}$ as the discounted state correlation matrix, in other words,

$$S_{K,\Sigma} = \mathbb{E}_{\pi_{K,\Sigma}}\left[\sum_{t=0}^{\infty} \gamma^t x_t x_t^\top\right]. \tag{4.2.14}$$

According to Theorem 4.1, the optimal policy of (4.1.5) is a Gaussian policy with a mean following a linear function of the state and a constant covariance matrix. In the remainder of the chapter, we look for a parameterized policy of the form

$$\pi_{K,\Sigma}(\cdot|x) = \mathcal{N}(-Kx, \Sigma), \tag{4.2.15}$$

for any $x \in \mathcal{X}$. With a slight abuse of notation, we use $J_{K,\Sigma}$ to denote $J_{\pi_{K,\Sigma}}$ and denote the objective in (4.1.5) as a function of (K, Σ), given by

$$C(K, \Sigma) = \mathbb{E}_{x_0 \sim \mathcal{D}_0} \left[J_{K,\Sigma}(x) \right]. \tag{4.2.16}$$

To analyze the dependence of Σ in the objective function (4.2.16) for any fixed K and $\Sigma > 0$, we also define $f_K : \mathbb{R}^{nN \times nN} \to \mathbb{R}$ as

$$f_K(\Sigma) = \frac{\tau}{2(1-\gamma)} \log\det(\Sigma) - \frac{1}{1-\gamma} \operatorname{Tr}\left(\Sigma \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D \right) \right). \tag{4.2.17}$$

By applying the Bellman equation, we can get $J_{K,\Sigma}(x) = x^\top P_K x + q_{K,\Sigma}$, with some symmetric positive definite matrix $P_K \in \mathbb{R}^{mN \times mN}$ and $q_{K,\Sigma} \in \mathbb{R}$ satisfying

$$\begin{aligned} P_K &= Q + K^\top R K + \gamma (A - BK)^\top P_K (A - BK) + \gamma (C - DK)^\top W P_K (C - DK), \\ q_{K,\Sigma} &= \frac{1}{1-\gamma} \left[\operatorname{Tr}\left(\Sigma \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D \right) - \frac{\tau}{2} \left(nN + \log\left((2\pi)^{nN} \det \Sigma \right) \right) \right) \right]. \end{aligned} \tag{4.2.18}$$

We now provide an explicit form for the gradient of the cost function $C(K, \Sigma)$ with respect to K and Σ. This explicit form will be used to show the gradient dominance condition in Lemma 4.3 and also to analyze the algorithms in Section 4.3.

Lemma 4.2 (Explicit form for the policy gradient)**.** *If one takes a policy in the form of (4.2.15) with parameter $(K, \Sigma) \in \Omega$, then the policy gradient has the following explicit form:*

$$\begin{aligned} \nabla_K C(K, \Sigma) &= 2 E_K S_{K,\Sigma}, \\ \nabla_\Sigma C(K, \Sigma) &= (1-\gamma)^{-1} \left(R - \frac{\tau}{2} \Sigma^{-1} + \gamma B^\top P_K B + \gamma D^\top W P_K D \right), \end{aligned} \tag{4.2.19}$$

where $E_K = -\gamma B^\top P_K (A - BK) - \gamma D^\top W P_K (C - DK) + RK$ and $S_{K,\Sigma}$ is defined in (4.2.14).

Proof of Lemma 4.2. $\nabla_\Sigma C(K,\Sigma)$ in (4.2.19) can be checked by direct gradient calculation. To verify $\nabla_K C(K,\Sigma)$ in (4.2.19), first define $f : \mathcal{X} \times \mathbb{R}^{nN\times mN} \to \mathbb{R}$ by $f(y,K) = y^\top P_K y, \forall y \in \mathcal{X}$, and we aim to find the gradient of f with respect to K. For any $y \in \mathcal{X}$,

$$\nabla_K f(y,K) = 2\left(-\gamma B^\top P_K(A-BK) - \gamma D^\top W P_K(C-DK) + RK\right) \sum_{t=0}^{\infty} \gamma^t [(A-BK)(A-BK)^\top + (C-DK)W(C-DK)^\top]^t yy^\top. \tag{4.2.20}$$

Since $C(K,\Sigma) = \mathbb{E}_{x_0\sim\mathcal{D}_0}[x_0 P_K x_0] + q_{K,\Sigma}$ with P_K and $q_{K,\Sigma}$ satisfying (4.2.18), then the gradient of $C(K,\Sigma)$ with respect to K is

$$\begin{aligned}\nabla_K C(K,\Sigma) &= \mathbb{E}\left[\nabla_K f\left(x_0,K\right)\right] + \nabla_K q_{K,\Sigma} \\ &= 2\left(-\gamma B^\top P_K(A-BK) - \gamma D^\top W P_K(C-DK) + RK\right)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t x_t x_t^\top\right],\end{aligned}$$

which follows from applying (4.2.20) with $y = x_0$, taking the gradient of $q_{K,\Sigma}$ in (4.2.18) with respect to K and using recursion to get the second equation. □

Gradient dominance. The key idea is to show the gradient dominance condition, which states that $C(K,\Sigma) - C\left(K^*,\Sigma^*\right)$ can be bounded by $\|\nabla_K C(K,\Sigma)\|_F^2$ and $\|\nabla_\Sigma C(K,\Sigma)\|_F^2$ for any $(K,\Sigma) \in \Omega$. This result suggests that when the gradient norms are sufficiently small, the cost function of the given policy is sufficiently close to the optimal cost function.

Lemma 4.3 (Gradient dominance of $C(K,\Sigma)$)**.** *Let π^* be the optimal policy with parameters K^*,Σ^*. Suppose policy π with parameter $(K,\Sigma) \in \Omega$ satisfying $\Sigma \preceq I$ has a finite expected cost, in other words, $C(K,\Sigma) < \infty$. Then*

$$C(K,\Sigma) - C\left(K^*,\Sigma^*\right) \leqslant \frac{\|S_{K^*,\Sigma^*}\|}{4\mu^2\sigma_{\min}(R)}\|\nabla_K C(K,\Sigma)\|_F^2 + \frac{1-\gamma}{\sigma_{\min}(R)}\|\nabla_\Sigma C(K,\Sigma)\|_F^2. \tag{4.2.21}$$

For a lower bound, with E_K defined in Lemma 4.2,

$$C(K,\Sigma) - C\left(K^*,\Sigma^*\right) \geqslant \frac{\mu}{\|R+\gamma B^\top P_K B+\gamma D^\top W P_K D\|}\operatorname{Tr}\left(E_K^\top E_K\right).$$

By introducing the notations of the Q-function in (4.2.22) and the advantage function in (4.2.24), we first provide a convenient form for the difference of the cost functions with respect to two different policies in Lemma 4.4. It will then be used in the proof of Lemma 4.3.

For a given policy $\pi_{K,\Sigma}(\cdot|x) = \mathcal{N}(-Kx,\Sigma)$ with parameter K and Σ, we define the state-action value function (also known as Q-function) $Q_{K,\Sigma} : \mathcal{X}\times\mathcal{U}\to\mathbb{R}$ as the cost of the policy starting with $x_0 = x$, taking a fixed action $u_0 = u$, and then proceeding with $\pi_{K,\Sigma}$. The Q-function is related to the value function $J_{K,\Sigma}$ defined in (4.1.2) as

$$Q_{K,\Sigma}(x,u) = x^\top Q x + u^\top R u + \tau\log\pi_{K,\Sigma}(u|x) + \gamma\mathbb{E}\left[J_{K,\Sigma}(Ax+Bu+(Cx+Du)w)\right], \tag{4.2.22}$$

for any $(x,u)\in\mathcal{X}\times\mathcal{U}$. By definition of the Q-function, we also have the relationship:

$$\forall x\in\mathcal{X}: \quad J_{K,\Sigma}(x) = \mathbb{E}_{u\sim\pi(\cdot|x)}\left[Q_{K,\Sigma}(x,u)\right]. \tag{4.2.23}$$

We also introduce the advantage function $A_{K,\Sigma} : \mathcal{X}\times\mathcal{U}\to\mathbb{R}$ of the policy π :

$$A_{K,\Sigma}(x,u) = Q_{K,\Sigma}(x,u) - J_{K,\Sigma}(x), \tag{4.2.24}$$

which reflects the gain one can harvest by executing control u instead of following the policy $\pi_{K,\Sigma}$ in state x.

Lemma 4.4 (Cost difference)**.** *Suppose policies π and π' are in the form of (4.2.15) with parameters (K',Σ') and (K,Σ). Let $\{x'_t\}_{t=0}^\infty$ and $\{u'_t\}_{t=0}^\infty$ be state and action*

sequences generated by π' with noise sequence $\{w'_t\}_{t=0}^{\infty}$ (i.i.d. with mean 0 and covariance W), that is, $x'_{t+1} = Ax'_t + Bu'_t + (Cx'_t + Du'_t)w'_t$. Then for any $x \in \mathcal{X}$,

$$J_{K',\Sigma'}(x) - J_{K,\Sigma}(x) = \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A_{K,\Sigma}\left(x'_t, u'_t\right) \middle| x'_0 = x\right], \tag{4.2.25}$$

where the expectation is taken over $u'_t \sim \pi'\left(\cdot|x'_t\right)$ and w_t for all $t = 0, 1, 2, \cdots$.

The expected advantage for any $x \in \mathcal{X}$ by taking expectation over $u \sim \pi'(\cdot|x)$ is:

$$\begin{aligned}
&\mathbb{E}_{\pi'}\left[A_{K,\Sigma}(x,u)\right] \\
&= x^\top (K'-K)^\top (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)(K'-K)x \\
&\quad + 2x^\top (K'-K)^\top \left[(R + \gamma B^\top P_K B + \gamma D^\top W P_K D)K - \gamma(B^\top P_K A + D^\top W P_K C)\right] x \\
&\quad + (1-\gamma)\left(f_K(\Sigma) - f_K\left(\Sigma'\right)\right).
\end{aligned} \tag{4.2.26}$$

Proof of Lemma 4.4. Let $\{c'_t\}_{t=0}^{\infty}$ be the cost sequences generated by π', that is, $c'_t = x_t'^\top Q x'_t + u_t'^\top R u'_t + \tau \log \pi'\left(u'_t|x'_t\right)$, and $J_{K',\Sigma'}(x) = \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t c'_t\right]$. Then we have

$$\begin{aligned}
&J_{K',\Sigma'}(x) - J_{K,\Sigma}(x) \\
&= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t c'_t\right] - J_{K,\Sigma}(x) \\
&= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t \left(c'_t - J_{K,\Sigma}\left(x'_t\right) + J_{K,\Sigma}\left(x'_t\right)\right)\right] - J_{K,\Sigma}(x) \\
&\overset{(a)}{=} \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t \left(c'_t - J_{K,\Sigma}\left(x'_t\right)\right) + \sum_{t=1}^{\infty} \gamma^t J_{K,\Sigma}\left(x'_t\right)\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t \left(c'_t + \gamma J_{K,\Sigma}\left(x'_{t+1}\right) - J_{K,\Sigma}\left(x'_t\right)\right)\right] \\
&= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t \left(x_t'^\top Q x'_t + u_t'^\top R u'_t + \tau \log \pi'\left(u'_t|x'_t\right) + \gamma J_{K,\Sigma}\left(x'_{t+1}\right) - J_{K,\Sigma}\left(x'_t\right)\right)\right],
\end{aligned}$$

where (a) follows from $x_0' = x$. By (4.2.22) and (4.2.24), we can continue the above calculations by

$$J_{K',\Sigma'}(x) - J_{K,\Sigma}(x) = \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t\left(Q_{K,\Sigma}\left(x_t', u_t'\right) - J_{K,\Sigma}\left(x_t'\right)\right)\right]$$
$$= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty}\gamma^t A_{K,\Sigma}\left(x_t', u_t'\right)\right].$$

To derive the expectation of the advantage function, for any x in $\mathcal{X}$, take expectation over $u \sim \pi'(\cdot|x) = \mathcal{N}\left(-K'x, \Sigma'\right)$ to get the following:

$$\begin{aligned}
&\mathbb{E}_{\pi'}\left[A_{K,\Sigma}(x,u)\right] \\
&= \mathbb{E}_{\pi'}\left[Q_{K,\Sigma}(x,u)\right] - J_{K,\Sigma}(x) \\
&= x^\top(Q + K'^\top R K')x + \operatorname{Tr}\left(\Sigma' R\right) - \frac{\tau}{2}\left(nN + \log\left((2\pi)^{nN}\det\Sigma'\right)\right) \\
&\quad + \gamma\mathbb{E}\left[(Ax + Bu + (Cx + Du)w)^\top P_K\left(Ax + Bu + (Cx + Du)w\right) + q_{K,\Sigma}\right] \\
&\quad - x^\top P_K x - q_{K,\Sigma} \\
&= x^\top(Q + K'^\top R K')x + \operatorname{Tr}\left(\Sigma' R\right) - \frac{\tau}{2}\left(nN + \log\left((2\pi)^{nN}\det\Sigma'\right)\right) - x^\top P_K x - (1-\gamma)q_{K,\Sigma} \\
&\quad + \gamma\Big[x^\top\left(A - BK'\right)^\top P_K\left(A - BK'\right)x + x^\top\left(C - DK'\right)^\top W P_K\left(C - DK'\right)x \\
&\quad + \operatorname{Tr}\left(\Sigma'(B^\top P_K B + D^\top W P_K D)\right)\Big].
\end{aligned}$$

Substitute (4.2.18) into the above equation, then use (4.2.17) and write $K' = K +$

$K' - K$ to get

$$
\begin{aligned}
&\mathbb{E}_{\pi'}[A_{K,\Sigma}(x,u)] \\
&= (1-\gamma)\left(f_K(\Sigma) - f_K\left(\Sigma'\right)\right) + x^\top \Big(Q + \left(K + K' - K\right)^\top R \left(K + K' - K\right)\Big) x \\
&\quad + \gamma x^\top \left(A - BK - B\left(K' - K\right)\right)^\top P_K \left(A - BK - B\left(K' - K\right)\right) x \\
&\quad + \gamma x^\top \left(C - DK - D\left(K' - K\right)\right)^\top W P_K \left(C - DK - D\left(K' - K\right)\right) x - x^\top P_K x \\
&= (1-\gamma)\left(f_K(\Sigma) - f_K\left(\Sigma'\right)\right) \\
&\quad + x^\top \left(Q + K^\top R K + \gamma (A - BK)^\top P_K (A - BK) + \gamma (C - DK)^\top W P_K (C - DK)\right) x \\
&\quad + 2x^\top \left(K' - K\right)^\top \left[\left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right) K - \gamma (B^\top P_K A + D^\top W P_K C\right] x \\
&\quad + x^\top \left(K' - K\right)^\top \left(R + \gamma B^\top P_K B + \gamma D^\top P_K D\right) \left(K' - K\right) x - x^\top P_K x.
\end{aligned}
$$

Substituting (4.2.18) finishes the proof to verify (4.2.26). □

Proof of Lemma 4.3. By Lemma 4.4, for any π and π' in form of (4.2.15) with parameter (K, Σ) and (K', Σ') respectively, and for any $x \in \mathcal{X}$,

$$
\begin{aligned}
&\mathbb{E}_{\pi'}\left[A_{K,\Sigma}(x,u)\right] \\
&= (1-\gamma)\left(f_K(\Sigma) - f_K(\Sigma')\right) - \operatorname{Tr}\left(x x^\top E_K^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) \\
&\quad + \operatorname{Tr}\Big[x x^\top \left(K' - K + \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right)^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right. \\
&\quad \left. + \gamma D^\top W P_K D\right) \cdot \left(K' - K + \left(R + \gamma B^\top P_K B\right)^{-1} E_K\right)\Big] \\
&\geqslant - \operatorname{Tr}\left(x x^\top E_K^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) + (1-\gamma)\left(f_K(\Sigma) - f_K\left(\Sigma'\right)\right),
\end{aligned}
\tag{4.2.27}
$$

with equality when $K - (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K = K'$ holds. Let $\{x_t^*\}_{t=0}^\infty$ and $\{u_t^*\}_{t=0}^\infty$ be the state and control sequences generated under $\pi^*(\cdot|x) = \mathcal{N}\left(-K^* x, \Sigma^*\right)$ with noise sequence $\{w_t^*\}_{t=0}^\infty$ with mean 0 and covariance W. Apply

Lemma 4.4 and (4.2.27) with π and π^* to get:

$$\begin{aligned}
&C(K,\Sigma) - C\left(K^*,\Sigma^*\right) \\
&= -\mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty}\gamma^t A_{K,\Sigma}\left(x_t^*,u_t^*\right)\right] \\
&\leqslant \operatorname{Tr}\left(S_{K^*,\Sigma^*}E_K^\top\left(R+\gamma B^\top P_K B+\gamma D^\top W P_K D\right)^{-1}E_K\right) + f_K\left(\Sigma^*\right) - f_K(\Sigma),
\end{aligned} \tag{4.2.28}$$

where f_K is defined in (4.2.17). To analyze the first term in (4.2.28), note that

$$\begin{aligned}
&\operatorname{Tr}\left(S_{K^*,\Sigma^*}E_K^\top\left(R+\gamma B^\top P_K B+\gamma D^\top W P_K D\right)^{-1}E_K\right) \\
&\leqslant \frac{\|S_{K^*,\Sigma^*}\|}{\sigma_{\min}(R)}\operatorname{Tr}\left(E_K^\top E_K\right) \\
&\leqslant \frac{\|S_{K^*,\Sigma^*}\|}{\mu^2\sigma_{\min}(R)}\operatorname{Tr}\left(S_{K,\Sigma}E_K^\top E_K S_{K,\Sigma}\right) \\
&= \frac{\|S_{K^*,\Sigma^*}\|}{4\mu^2\sigma_{\min}(R)}\operatorname{Tr}\left(\nabla_K^\top C(K,\Sigma)\nabla_K C(K,\Sigma)\right),
\end{aligned} \tag{4.2.29}$$

where the last equation follows from (4.2.19).

To analyze the second terms in (4.2.28), note that f_K is a concave function; thus we can find its maximizer Σ_K^* by taking the gradient of f_K and setting it to 0, that

is, $\Sigma_K^* = \frac{\tau}{2}\left(R + \gamma B^\top P_K B + \gamma\cdot\ D^\top W P_K D\right)^{-1}$. Thus,

$$\begin{aligned}
&f_K\left(\Sigma^*\right) - f_K(\Sigma) \\
\leqslant\ & f_K\left(\Sigma_K^*\right) - f_K(\Sigma) \\
\overset{(a)}{\leqslant}\ & \mathrm{Tr}\left(\nabla f_K(\Sigma)^\top\left(\Sigma_K^* - \Sigma\right)\right) \\
=\ & \mathrm{Tr}\Big[\left(\nabla_\Sigma C(K,\Sigma)\cdot\left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1}\right. \\
& \left.\left(\left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right) - \frac{\tau}{2}\Sigma^{-1}\right)\Sigma\right] \\
\overset{(b)}{\leqslant}\ & (1-\gamma)\|(R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1}\|\,\|\nabla_\Sigma C(K,\Sigma)\|_F^2 \\
\leqslant\ & \frac{(1-\gamma)\,\|\nabla_\Sigma C(K,\Sigma)\|_F^2}{\sigma_{\min}(R)},
\end{aligned} \tag{4.2.30}$$

where (a) follows from the first order concavity condition for f_K and (b) is from $0 \prec \Sigma \preceq I$.

For the lower bound, consider $K' = K - (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K$ and $\Sigma' = \Sigma$, where equality holds in (4.2.27). Let $\{x_t'\}_{t=0}^\infty, \{u_t'\}_{t=0}^\infty$ be the sequence generated with K', Σ'. By $C\left(K^*, \Sigma^*\right) \leqslant C\left(K', \Sigma'\right)$; we have

$$\begin{aligned}
C(K,\Sigma) - C\left(K^*,\Sigma^*\right) &\geqslant C(K,\Sigma) - C\left(K',\Sigma'\right) \\
&= -\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t A_{K,\Sigma}\left(x_t', u_t'\right)\right] \\
&= \mathrm{Tr}\left(S_{K',\Sigma'} E_K^\top\left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) \\
&\geqslant \frac{\mu}{\|R + \gamma B^\top P_K B + \gamma D^\top W P_K D\|}\,\mathrm{Tr}\left(E_K^\top E_K\right).
\end{aligned}$$

□

"Almost smoothness". Next, we will develop a smoothness property for the cost objective $C(K,\Sigma)$, which is necessary for establishing the convergence algorithms

proposed in Section 4.3. A function $f : \mathbb{R}^m \to \mathbb{R}$ is considered smooth if the following condition is satisfied:

$$\left|f(x) - f(y) + \nabla f(x)^\top (y - x)\right| \leqslant \frac{m}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^m,$$

with m a finite constant. In general, characterizing the smoothness of $C(K, \Sigma)$ is challenging because it may become unbounded when the eigenvalues of $A - BK$ and $C - DK$ exceed $\frac{1}{\sqrt{\gamma}}$ or when $\sigma_{\min}(\Sigma)$ is close to 0. Nevertheless, in Lemma 4.6, we will see that if $C(K, \Sigma)$ is "almost" smooth, then the difference $C(K, \Sigma) - C\left(K', \Sigma'\right)$ can be bounded by the sum of linear and quadratic terms involving $K - K'$ and $\Sigma - \Sigma'$.

Before Lemma 4.6, we first introduce Lemma 4.5 to show that the cost objective is smooth in Σ when utilizing entropy regularization, given that Σ is bounded.

Lemma 4.5 (Smoothness of f_K)**.** *Let $K \in \mathbb{R}^{nN \times mN}$ be given and let f_K be defined in (4.2.17). Fix $0 < a \leqslant 1$. For any symmetric positive definite matrices $X \in \mathbb{R}^{n \times n}$ and $Y \in \mathbb{R}^{n \times n}$ satisfying $aI \preceq X \preceq I$ and $aI \preceq Y \preceq I$, the following inequality holds:*

$$\left|f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top (Y - X)\right)\right| \leqslant M_a \operatorname{Tr}\left((X^{-1}Y - I)^2\right),$$

where $M_a \in \mathbb{R}$ is defined in Lemma 4.6, and $M_a \geqslant \frac{\tau}{4(1-\gamma)}$.

Proof of Lemma 4.5. Fix symmetric positive definite matrices X and Y satisfying $aI \preceq X \preceq I$ and $aI \preceq Y \preceq I$. Then f_K being concave implies $f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top (Y - X)\right) \geqslant 0$. To find an upper bound, observe that

$$\begin{aligned} & f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top (Y - X)\right) \\ = & \frac{\tau}{2(1-\gamma)}\left(\log\det(X) - \log\det(Y) + \operatorname{Tr}\left(X^{-1}(Y - X)\right)\right). \end{aligned} \tag{4.2.31}$$

Since $X \succ 0$ and $Y \succ 0$, then all eigenvalues of $X^{-1}Y$ are real and positive and $\sigma_{\min}\left(X^{-1}Y\right) \geqslant \sigma_{\min}\left(X^{-1}\right)\sigma_{\min}(Y) \geqslant a$.

Now let us show that there exists a constant $m > 0$ (independent of K) such that for any $Z \in \mathbb{R}^{n\times n}$ with real positive eigenvalues $a \leqslant z_1 \leqslant \cdots \leqslant z_k$, the following holds:

$$-\log(\det(Z)) + \operatorname{Tr}(Z - I) \leqslant m \operatorname{Tr}\left((Z - I)^2\right). \tag{4.2.32}$$

Note that (4.2.32) is equivalent to $\sum_{i=1}^{k} -\log(z_i) + z_i - 1 \leqslant m \sum_{i=1}^{k} (z_i - 1)^2$. With elementary algebra, one can verify that when $m = (-\log(a) + a - 1) \cdot (a - 1)^{-2}$, it holds that $-\log(z) + z - 1 \leqslant m(z - 1)^2$ for all $z \geqslant a$ and such an m satisfies $m \geqslant \frac{1}{2}$. Therefore, (4.2.32) holds. Combining (4.2.31) and (4.2.32) with $Z = X^{-1}Y$, we see

$$f_K(X) - f_K(Y) + \operatorname{Tr}\left(\nabla f_K(X)^\top (Y - X)\right) \leqslant \frac{\tau m}{2(1-\gamma)} \operatorname{Tr}\left((X^{-1}Y - I)^2\right).$$

□

Lemma 4.6 ("Almost smoothness" of $C(K, \Sigma)$)**.** *Fix $0 < a \leqslant 1$ and define $M_a = \frac{\tau(-\log(a)+a-1)}{2(1-\gamma)(a-1)^2}$. For any K, Σ and K', Σ' satisfying $aI \preceq \Sigma \preceq I$ and $aI \preceq \Sigma' \preceq I$,*

$$\begin{aligned}
&C(K', \Sigma') - C(K, \Sigma) \\
&= \operatorname{Tr}\left(S_{K',\Sigma'} (K' - K)^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)(K' - K)\right) \\
&\quad + 2 \operatorname{Tr}\left(S_{K',\Sigma'} (K' - K)^\top E_K\right) + f_K(\Sigma) - f_K(\Sigma') \\
&\leqslant \operatorname{Tr}\left(S_{K',\Sigma'} (K' - K)^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)(K' - K)\right) \\
&\quad + 2 \operatorname{Tr}\left(S_{K',\Sigma'} (K' - K)^\top E_K\right) \\
&\quad + \frac{1}{1-\gamma} \operatorname{Tr}\left(\left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D - \frac{\tau}{2}\Sigma^{-1}\right)(\Sigma' - \Sigma)\right) \\
&\quad + M_a \operatorname{Tr}\left(\left(\Sigma^{-1}\Sigma' - I\right)^2\right),
\end{aligned}$$

where f_K is defined in (4.2.17).

Proof of Lemma 4.6. The first equality immediately results from Lemma 4.4, by taking an expectation with respect to x in (4.2.26). The last inequality follows directly from Lemma 4.5. □

4.3 Iterative Policy Optimization Method

For IPO, one can show both the global convergence with a linear rate and a local super-linear convergence when the initialization is close to the optimal policy.

By the Bellman equation for the value function $J_{K,\Sigma}$,

$$J_{K,\Sigma}(x) = \mathbb{E}_{u\sim\pi_{K,\Sigma}}\left[x^\top Qx + u^\top Ru + \tau\log\pi_{K,\Sigma}(u|x) + \gamma J_{K,\Sigma}(Ax+Bu+(Cx+Du)w)\right].$$

We assume the one-step update (K',Σ') satisfies:

$$K',\Sigma' = \arg\min_{\widetilde{K},\widetilde{\Sigma}} \mathbb{E}_{u\sim\pi_{\widetilde{K},\widetilde{\Sigma}}}\left[x^\top Qx + u^\top Ru + \tau\log\pi_{\widetilde{K},\widetilde{\Sigma}}(u|x) + \gamma J_{K,\Sigma}(Ax+Bu+(Cx+Du)w)\right].$$

By direct calculation, we have the following explicit forms for the updates:

$$\begin{aligned} K^{(t+1)} &= K^{(t)} - \left(R+\gamma B^\top P_{K^{(t)}}B + \gamma D^\top W P_{K^{(t)}}D\right)^{-1} E_{K^{(t)}}, \\ \Sigma^{(t+1)} &= \frac{\tau}{2}\left(R+\gamma B^\top P_{K^{(t)}}B + \gamma D^\top W P_{K^{(t)}}D\right)^{-1}, \end{aligned} \tag{IPO}$$

for $t = 1, 2, \cdots$.

4.3.1 Global Linear Convergence

In this section, we establish the global convergence for (IPO) with a linear rate.

Theorem 4.2 (Global convergence of (IPO))**.** *For*

$$N \geqslant \frac{\|S_{K^*,\Sigma^*}\|}{\mu}\log\frac{C\left(K^{(0)},\Sigma^{(0)}\right) - C\left(K^*,\Sigma^*\right)}{\varepsilon},$$

the iterative policy optimization algorithm (IPO) has the following performance bound:

$$C\left(K^{(N)},\Sigma^{(N)}\right) - C\left(K^*,\Sigma^*\right) \leqslant \varepsilon.$$

The proof of Theorem 4.2 is immediately given by the following lemma, which bounds the one-step progress of (IPO):

Lemma 4.7 ((Contraction of (IPO))**.** *Suppose* (K', Σ') *follows a one-step updating rule of (IPO) from* (K, Σ) :

$$K' = K - \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K, \quad \Sigma' = \frac{\tau}{2} \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1}. \tag{4.3.33}$$

Then

$$C\left(K', \Sigma'\right) - C\left(K^*, \Sigma^*\right) \leqslant \left(1 - \frac{\mu}{\|S_{K^*,\Sigma^*}\|}\right) \left(C(K, \Sigma) - C\left(K^*, \Sigma^*\right)\right),$$

with $0 < \frac{\mu}{\|S_{K^*,\Sigma^*}\|} \leqslant 1$.

Proof of Lemma 4.7. Let K, Σ be fixed and let f_K be defined in (4.2.17). Observe that Σ' defined in (4.3.33) is the maximizer of f_K. Then by Lemma 4.6,

$$\begin{aligned}
& C\left(K', \Sigma'\right) - C(K, \Sigma) \\
= & -\operatorname{Tr}\left(S_{K',\Sigma'} E_K^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) - f_K\left(\Sigma'\right) + f_K(\Sigma) \\
\overset{(a)}{\leqslant} & -\frac{\mu}{\|S_{K^*,\Sigma^*}\|} \operatorname{Tr}\left(S_{K^*,\Sigma^*} E_K^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) \\
& - \frac{\mu}{\|S_{K^*,\Sigma^*}\|} \left(f_K\left(\Sigma'\right) - f_K(\Sigma)\right) \\
\overset{(b)}{\leqslant} & -\frac{\mu}{\|S_{K^*,\Sigma^*}\|} \left(\operatorname{Tr}\left(S_{K^*,\Sigma^*} E_K^\top \left(R + \gamma B^\top P_K B + \gamma D^\top W P_K D\right)^{-1} E_K\right) + f_K\left(\Sigma^*\right) - f_K(\Sigma)\right),
\end{aligned}$$

where (a) follows the fact that $f_K\left(\Sigma'\right) - f_K(\Sigma) \geqslant 0$ and $\frac{\mu}{\|S_{K^*,\Sigma^*\|}} \leqslant 1$; and (b) follows from fact that $f_K\left(\Sigma'\right) \geqslant f_K\left(\Sigma^*\right)$. Finally, apply (4.2.28) to get

$$C\left(K', \Sigma'\right) - C(K, \Sigma) \leqslant -\frac{\mu}{\|S_{K^*,\Sigma^*\|}} \left(C(K, \Sigma) - C\left(K^*, \Sigma^*\right)\right).$$

Adding $C(K,\Sigma)-C\left(K^*,\Sigma^*\right)$ to both sides of the above inequality finishes the proof.

□

Theorem 4.2 suggests that (IPO) achieves a global linear convergence. Furthermore, the subsequent section demonstrates that (IPO) enjoys a super-linear local convergence when the initial policy parameter is within a neighborhood of the optimal policy parameter.

4.3.2 Super-Linear Local Convergence

This section establishes a super-linear local convergence for (IPO). We first introduce some constants used throughout this section:

$$\begin{aligned}
&\xi_{\gamma,\rho}=\frac{1-\gamma\rho^2+\gamma}{\left(1-\gamma\rho^2\right)^2},\quad \zeta_{\gamma,\rho}=\frac{2-\rho^2}{\left(1-\rho^2\right)^2\left(1-\gamma\right)}+\frac{1}{\left(1-\rho^2\right)^2\left(1-\gamma\rho^2\right)},\\
&\omega_{\gamma,\rho}=\frac{1}{\left(1-\rho^2\right)\left(1-\gamma\right)}+\frac{1}{\left(1-\rho^2\right)\left(1-\gamma\rho^2\right)}.
\end{aligned}\tag{4.3.34}$$

To simplify the exposition, we often make use of the notation $S_{K^{(t)},\Sigma^{(t)}}$ and S_{K^*,Σ^*}, which we abbreviate as $S^{(t)}$ and S^*, respectively, provided that the relevant parameter values are clear from the context. Then, we define

$$\begin{aligned}
\kappa &= \min\left\{\frac{\rho_1+\|A\|}{|\sigma_{\min}(B)|},\frac{\rho_2+\|C\|}{|\sigma_{\min}(D)|}\right\},\\
c &= 2\xi_{\gamma,\rho}\left(\rho_1\|B\|+\rho_2\|D\|\|W\|\right)\left(\|Q\|+\|R\|\kappa^2\right)+\frac{1}{\mu}\|S^*\|\,\|R\|\cdot\left(\kappa+\|K^*\|\right),\\
c_1 &= \left(\xi_{\gamma,\rho}\mathbb{E}[x_0x_0^\top]+\zeta_{\gamma,\rho}\|B\Sigma^*B^\top+DW\Sigma^*D^\top\|\right)\cdot 2(\rho_1\|B\|+\rho_2\|D\|\|W\|)\\
&\quad\cdot\Big(1+\sigma_{\min}(R)\cdot\|R+\gamma B^\top P_{K^*}B+\gamma D^\top WP_{K^*}D\|+c\gamma\sigma_{\min}(R)\\
&\quad\cdot\left[(\|B\|\|A\|-\|D\|\|W\|\|C\|)+(\|B\|^2+\|D\|^2\|W\|)\kappa\right]\Big),\\
c_2 &= \frac{c\tau\gamma\omega_{\gamma,\rho}(\|B\|^2+\|D\|\|W\|)^2}{2\sigma_{\min}(R)^2}.
\end{aligned}\tag{4.3.35}$$

Note that $\kappa \geqslant \|K\|$ for any admissible K, since

$$\begin{aligned}\|K\| &\leqslant \min\left\{\frac{\|BK\|}{|\sigma_{\min}(B)|}, \frac{\|DK\|}{|\sigma_{\min}(D)|}\right\} \\ &\leqslant \left\{\frac{\|A-BK\|+\|A\|}{|\sigma_{\min}(B)|}, \frac{\|C-DK\|+\|C\|}{|\sigma_{\min}(D)|}\right\} \\ &\leqslant \left\{\frac{\rho_1+\|A\|}{|\sigma_{\min}(B)|}, \frac{\rho_2+\|C\|}{|\sigma_{\min}(D)|}\right\} \\ &= \kappa,\end{aligned}$$

for any $K \in \Omega$.

We now show that (IPO) achieves a super-linear convergence rate if the policy parameter (K, Σ) enters a neighborhood of the optimal policy parameter (K^*, Σ^*).

Theorem 4.3 ((Super-linear local convergence of (IPO))**.** *Let c_1 and c_2 be defined in (4.3.35). Let $\delta = \min\left\{\frac{1}{c_1+c_2}\sigma_{\min}(S^*), \frac{\rho_1-\|A-BK^*\|}{\|B\|}, \frac{\rho_2-\|C-DK^*\|}{\|D\|}\right\}$. Suppose that the initial policy $\left(K^{(0)}, \Sigma^{(0)}\right)$ satisfies*

$$C\left(K^{(0)}, \Sigma^{(0)}\right) - C\left(K^*, \Sigma^*\right) \leqslant \left(\frac{1}{\mu} - \frac{1}{\|S^*\|}\right)^{-1} \sigma_{\min}\left(R + \gamma B P_{K^*} B + \gamma D W P_{K^*} D\right) \delta^2, \tag{4.3.36}$$

then the iterative policy optimization algorithm (IPO) has the following convergence rate: for $t = 0, 1, 2, \cdots$,

$$C\left(K^{(t+1)}, \Sigma^{(t+1)}\right) - C\left(K^*, \Sigma^*\right) \leqslant \frac{(c_1+c_2)\left(C\left(K^{(t)}, \Sigma^{(t)}\right) - C\left(K^*, \Sigma^*\right)\right)^{1.5}}{\sigma_{\min}(S^*)\sqrt{\mu\sigma_{\min}\left(R + \gamma B^\top P_{K^*} B + \gamma D W P_{K^*} D\right)}}.$$

The following Lemma 4.8 is critical for establishing this super-linear local convergence; it shows that there is a contraction if the differences between two discounted state correlation matrices $\left\|S^{(t+1)} - S^*\right\|$ are small enough. Then, by the perturbation analysis for $S_{K,\Sigma}$ (Lemma 4.9), one can bound $\left\|S^{(t+1)} - S^*\right\|$ by $\left\|K^{(t)} - K^*\right\|$

(Lemma 4.13). The proof of Theorem 4.3 follows by ensuring the admissibility of model parameters $\left\{K^{(t)}, \Sigma^{(t)}\right\}_{t=0}^{\infty}$ along all the updates.

Lemma 4.8. *Suppose that the updating rule (IPO) satisfies* $\left\|S^{(t+1)} - S^*\right\| \leqslant \sigma_{\min}(S^*)$ *for all* $t \geqslant 0$, *then*

$$C\left(K^{(t+1)}, \Sigma^{(t+1)}\right) - C\left(K^*, \Sigma^*\right) \leqslant \frac{\left\|S^{(t+1)} - S^*\right\|}{\sigma_{\min}(S^*)} \left(C(K^{(t)}, \Sigma^{(t)}) - C\left(K^*, \Sigma^*\right)\right).$$

Proof of Lemma 4.8. Let K' denote $K^{(t+1)}$, K denote $K^{(t)}$ and S' denote $S^{(t+1)}$. Apply Lemma 4.6 with f_K defined in (4.2.17) to get

$$\begin{aligned}
& C\left(K', \Sigma'\right) - C(K, \Sigma) \\
= & -\operatorname{Tr}\left(S^* E_K^\top (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K\right) \\
& -\operatorname{Tr}\left(\left(S' - S^*\right) E_K^\top (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K\right) + f_K(\Sigma) - f_K\left(\Sigma'\right) \\
\overset{(a)}{\leqslant} & \left(-1 + \left\|\left(S' - S^*\right) S^{*-1}\right\|\right) \left(\operatorname{Tr}\left(S^* E_K^\top (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K\right)\right. \\
& \left. + f_K\left(\Sigma^*\right) - f_K(\Sigma)\right),
\end{aligned}$$

where (a) follows from the assumption $\left\|S' - S^*\right\| \leqslant \sigma_{\min}(S^*)$, which implies $-1 + \left\|\left(S' - S^*\right) S^{*-1}\right\| \in [-1, 0]$ and the fact Σ' is the maximizer of f_K. Finally, apply (4.2.28) to get

$$\begin{aligned}
C\left(K', \Sigma'\right) - C(K, \Sigma) & \leqslant \left(-1 + \left\|\left(S' - S^*\right) S^{*-1}\right\|\right) \left(C(K, \Sigma) - C\left(K^*, \Sigma^*\right)\right) \\
& \leqslant \left(-1 + \frac{\left\|S' - S^*\right\|}{\sigma_{\min}(S^*)}\right) \left(C(K, \Sigma) - C\left(K^*, \Sigma^*\right)\right).
\end{aligned}$$

Adding $C(K, \Sigma) - C\left(K^*, \Sigma^*\right)$ to both sides of the above inequality finishes the proof.

□

Lemma 4.9 ($S_{K,\Sigma}$ perturbation)**.** *For any* (K_1, Σ_1) *and* (K_2, Σ_2) *satisfying* $\|A - BK_1\| \leqslant \rho_1$, $\|A - BK_2\| \leqslant \rho_1$, $\|C - DK_1\| \leqslant \rho_2$, *and* $\|C - DK_2\| \leqslant \rho_2$,

$$\begin{aligned}
&\|S_{K_1,\Sigma_1} - S_{K_2,\Sigma_2}\| \\
&\leqslant \left(\xi_{\gamma,\rho}\|\mathbb{E}[x_0 x_0^\top]\| + \zeta_{\gamma,\rho}\|B\Sigma_1 B^\top + DW\Sigma_1 D^\top\|\right) \cdot 2(\rho_1\|B\| + \rho_2\|D\|\|W\|)\,\|K_1 - K_2\| \\
&\quad + \omega_{\gamma,\rho}(\|B\|^2 + \|D\|^2\|W\|)\,\|\Sigma_1 - \Sigma_2\|\,.
\end{aligned}$$

The proof of Lemma 4.9 proceeds with a few technical lemmas. First, define the linear operators on symmetric matrices. For symmetric matrix $X \in \mathbb{R}^{mN\times mN}$, we set

$$\begin{aligned}
\mathcal{F}_K(X) &= [(A - BK)(A - BK)^\top + (C - DK)W(C - DK)^\top]X, \\
\mathcal{G}_t^K(X) &= [(A - BK)(A - BK)^\top + (C - DK)W(C - DK)^\top]^t X, \\
\mathcal{T}_K(X) &= \sum_{t=0}^{\infty} \gamma^t [(A - BK)(A - BK)^\top + (C - DK)(C - DK)^\top]^t X \\
&= \sum_{t=0}^{\infty} \gamma^t \mathcal{G}_t^K(X).
\end{aligned} \tag{4.3.37}$$

Note that when $\|A - BK\| < \frac{1}{\sqrt{\gamma}}$ and $\|C - DK\| < \frac{1}{\sqrt{\gamma}}$, we have

$$\mathcal{T}_K = (I - \gamma\mathcal{F}_K)^{-1}\,. \tag{4.3.38}$$

We also define the induced norm for these operators as $\|T\| = \sup_X \frac{\|T(X)\|}{\|X\|}$, where $T = \mathcal{F}_K, \mathcal{G}_t^K, \mathcal{T}_K$ and the supremum is over all symmetric matrix $X \in \mathbb{R}^{mN\times mN}$ with nonzero spectral norm.

Lemma 4.10. *For any* $K \in \mathbb{R}^{nN\times mN}$ *and any symmetric* $\Sigma \in \mathbb{R}^{nN\times nN}$,

$$S_{K,\Sigma} = \mathcal{T}_K(\mathbb{E}[x_0 x_0^\top]) + \sum_{t=0}^{\infty} \gamma^t \sum_{s=1}^{t} \mathcal{G}_{t-s}^K(B\Sigma B^\top + D^\top W\Sigma D^\top).$$

Proof of Lemma 4.10. Let us first show the following equation holds for any $t = 1, 2, 3, \cdots$ by induction:

$$\begin{aligned}\mathbb{E}[x_t x_t^\top] = &[(A-BK)(A-BK)^\top + (C-DK)W(C-DK)^\top]^t \mathbb{E}[x_0 x_0^\top] \\ &+ \sum_{s=1}^{t} [(A-BK)(A-BK)^\top + (C-DK)W(C-DK)^\top]^{t-s} \\ &(B\Sigma B^\top + DW\Sigma D^\top).\end{aligned}$$

When $t = 1$, it holds since

$$\begin{aligned}\mathbb{E}[x_1 x_1^\top] = &(A-BK)\mathbb{E}[x_0 x_0^\top](A-BK)^\top + (C-DK)W\mathbb{E}[x_0 x_0^\top](C-DK)^\top \\ &+ B\Sigma B^\top + DW\Sigma D^\top.\end{aligned}$$

Now assume it holds for t and consider $t+1$:

$$\begin{aligned}\mathbb{E}[x_{t+1} x_{t+1}^\top] &= (A-BK)\mathbb{E}[x_t x_t^\top](A-BK)^\top + (C-DK)\mathbb{E}[x_t x_t^\top](C-DK)^\top \\ &\quad + B\Sigma B^\top + DW\Sigma D^\top \\ &= [(A-BK)(A-BK)^\top + (C-DK)W(C-DK)^\top]^{t+1} \mathbb{E}[x_0 x_0^\top] \\ &\quad + \sum_{s=1}^{t+1} [(A-BK)(A-BK)^\top + (C-DK)(C-DK)^\top]^{t-s+1} \\ &\quad (B\Sigma B^\top + DW\Sigma D^\top).\end{aligned}$$

Thus, the induction is completed. Then by (4.3.37), we can write $S_{K,\Sigma}$ as:

$$S_{K,\Sigma} = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[x_t x_t^\top] = \mathcal{T}_K\left(\mathbb{E}[x_0 x_0^\top]\right) + \sum_{t=0}^{\infty} \gamma^t \sum_{s=1}^{t} \mathcal{G}_{t-s}^K \left(B\Sigma B^\top + DW\Sigma D^\top\right).$$

□

Lemma 4.11. *For any K_1 and K_2,*

$$\begin{aligned}\|\mathcal{F}_{K_1} - \mathcal{F}_{K_2}\| \leqslant &\left(\|A-BK_1\| + \|A-BK_2\|\right)\|B\|\,\|K_1 - K_2\| \\ &+ \left(\|C-DK_1\| + \|C-DK_2\|\right)\|D\|\|W\|\,\|K_1 - K_2\|.\end{aligned}$$

Lemma 4.11 follows a direct calculation with the triangle inequality.

Lemma 4.12. *For any K_1 and K_2 satisfying $\|A - BK_1\| \leqslant \rho_1$ and $\|A - BK_2\| \leqslant \rho_1$, $\|C - DK_1\| \leqslant \rho_2$, $\|C - DK_2\| \leqslant \rho_2$ with $\rho^2 = \rho_1^2 + \|W\|\rho_2^2$ and $\xi_{\gamma,\rho}$ defined in (4.3.34),*

$$\|(\mathcal{T}_{K_1} - \mathcal{T}_{K_2})(X)\| \leqslant \sum_{t=0}^{\infty} \gamma^t \left\|\left(\mathcal{G}_t^{K_1} - \mathcal{G}_t^{K_2}\right)(X)\right\| \leqslant \xi_{\gamma,\rho} \|\mathcal{F}_{K_1} - \mathcal{F}_{K_2}\| \|X\|, \quad (4.3.39a)$$

$$\sum_{t=0}^{T-1} \|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| \leqslant \frac{2 - \rho^2 - \rho^{2T}}{(1-\rho^2)^2} \|\mathcal{F}' - \mathcal{F}\| \|X\|, \quad \forall T \geqslant 1. \quad (4.3.39b)$$

Proof of Lemma 4.12. To ease the exposition, let us denote $\mathcal{G}_t = \mathcal{G}_t^{K_1}, \mathcal{G}_t' = \mathcal{G}_t^{K_2}, \mathcal{F} = \mathcal{F}_{K_1}$ and $\mathcal{F}' = \mathcal{F}_{K_2}$. Then for any symmetric matrix $X \in \mathbb{R}^{mN \times mN}$ and $t = 0, 1, 2, \cdots,$

$$\begin{aligned}
\left\|\left(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1}\right)(X)\right\| &= \|\mathcal{F}' \circ \mathcal{G}_t'(\Sigma) - \mathcal{F}' \circ \mathcal{G}_t(X) + \mathcal{F}' \circ \mathcal{G}_t(X) - \mathcal{F} \circ \mathcal{G}_t(X)\| \\
&\leqslant \|\mathcal{F}' \circ \mathcal{G}_t'(X) - \mathcal{F}' \circ \mathcal{G}_t(X)\| + \|\mathcal{F}' \circ \mathcal{G}_t(X) - \mathcal{F} \circ \mathcal{G}_t(X)\| \\
&= \|\mathcal{F}' \circ (\mathcal{G}_t' - \mathcal{G}_t)(X)\| + \|(\mathcal{F}' - \mathcal{F}) \circ \mathcal{G}_t(X)\| \\
&\leqslant \|\mathcal{F}'\| \|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| + \|\mathcal{G}_t\| \|\mathcal{F}' - \mathcal{F}\| \|X\| \\
&\leqslant \rho^2 \|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| + \rho^{2t} \|\mathcal{F}' - \mathcal{F}\| \|X\|.
\end{aligned} \quad (4.3.40)$$

Summing (4.3.40) for $t = 0, 1, 2, \cdots$ with a discount factor γ, we have

$$\sum_{t=0}^{\infty} \gamma^t \left\|\left(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1}\right)(X)\right\| \leqslant \rho^2 \sum_{t=0}^{\infty} \gamma^t \|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| + \frac{1}{1 - \gamma\rho^2} \|\mathcal{F}' - \mathcal{F}\| \|X\|.$$

Note that the left-hand side of the above inequality equals to

$$\frac{1}{\gamma} \sum_{t=0}^{\infty} \gamma^{t+1} \left\|\left(\mathcal{G}_{t+1}' - \mathcal{G}_{t+1}\right)(X)\right\| = \frac{1}{\gamma} \sum_{t=0}^{\infty} \gamma^t \|(\mathcal{G}_t' - \mathcal{G}_t)(X)\| - \frac{1}{\gamma} \|(\mathcal{G}_0' - \mathcal{G}_0)(X)\|.$$

Thus, by combining the above calculations, we obtain:

$$\left(\frac{1}{\gamma}-\rho^2\right)\sum_{t=0}^{\infty}\gamma^t\left\|\left(\mathcal{G}_t'-\mathcal{G}_t\right)(X)\right\|\leqslant\frac{1}{\gamma}\left\|\left(\mathcal{G}_0'-\mathcal{G}_0\right)(X)\right\|+\frac{1}{1-\gamma\rho^2}\left\|\mathcal{F}'-\mathcal{F}\right\|\|X\|.$$

The proof of (4.3.39a) is finished by noting $\mathcal{G}_0=\mathcal{F}$ and $\mathcal{G}_0'=\mathcal{F}'$. To show (4.3.39b), fix integer $T\geqslant 1$. Summing up (4.3.40), for $t=0,1,\cdots,T-1$,

$$\sum_{t=0}^{T-1}\left\|\left(\mathcal{G}_{t+1}'-\mathcal{G}_{t+1}\right)(X)\right\|\leqslant\rho^2\sum_{t=0}^{T-1}\left\|\left(\mathcal{G}_t'-\mathcal{G}_t\right)(X)\right\|+\frac{1-\rho^{2T}}{1-\rho^2}\left\|\mathcal{F}'-\mathcal{F}\right\|\|X\|. \quad (4.3.41)$$

Note that the left-hand side equals

$$\sum_{t=0}^{T-1}\left\|\left(\mathcal{G}_{t+1}'-\mathcal{G}_{t+1}\right)(X)\right\|=\sum_{t=0}^{T-1}\left\|\left(\mathcal{G}_t'-\mathcal{G}_t\right)(X)\right\|-\left\|\left(\mathcal{F}'-\mathcal{F}\right)(X)\right\|.$$

Substituting the above equation into (4.3.41) finishes the proof of (4.3.39b). □

Proof of Lemma 4.9. Denote $\mathcal{G}_t=\mathcal{G}_t^{K_1},\mathcal{G}_t'=\mathcal{G}_t^{K_2},\mathcal{T}=\mathcal{T}_{K_1},\mathcal{T}'=\mathcal{T}_{K_2},\mathcal{F}=\mathcal{F}_{K_1}$ and $\mathcal{F}'=\mathcal{F}_{K_2}$ to ease the exposition. First observe that for any $X,X'\in\mathbb{R}^{mN\times mN}$ and $t\geqslant 0$,

$$\begin{aligned}\left\|\mathcal{G}_t(X)-\mathcal{G}_t\left(X'\right)\right\|&=\left\|\left[(A-BK)(A-BK)^\top+(C-DK)W(C-DK)^\top\right]^t(X-X')\right\|\\&\leqslant\rho^{2t}\left\|X-X'\right\|.\end{aligned} \quad (4.3.42)$$

By Lemma 4.10,

$$
\begin{aligned}
&\|S_{K_1,\Sigma_1} - S_{K_2,\Sigma_2}\| \\
\leqslant\ & \left\|(\mathcal{T}-\mathcal{T}')\left(\mathbb{E}[x_0x_0^\top]\right)\right\| + \left\|\sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\left(\mathcal{G}'_{t-s}-\mathcal{G}_{t-s}\right)\left(B\Sigma_1B^\top+DW\Sigma_1D^\top\right)\right\| \\
&+\sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\left\|\mathcal{G}'_{t-s}\left(B\Sigma_1B^\top+DW\Sigma_1D^\top\right)-\mathcal{G}'_{t-s}\left(B\Sigma_2B^\top+DW\Sigma_2D^\top\right)\right\| \\
\overset{(a)}{\leqslant}\ & \left\|(\mathcal{T}-\mathcal{T}')\left(\mathbb{E}[x_0x_0^\top]\right)\right\| + \left\|\sum_{t=0}^{\infty}\gamma^t\sum_{s=0}^{t-1}\left(\mathcal{G}'_{s}-\mathcal{G}_{s}\right)\left(B\Sigma_1B^\top+DW\Sigma_1D^\top\right)\right\| \\
&+\sum_{t=0}^{\infty}\gamma^t\sum_{s=1}^{t}\rho^{2(t-s)}(\|B\|^2+\|D\|^2\|W\|)\,\|\Sigma_1-\Sigma_2\| \\
\overset{(b)}{\leqslant}\ & \xi_{\gamma,\rho}\,\|\mathcal{F}-\mathcal{F}'\|\left\|\mathbb{E}[x_0x_0^\top]\right\| + \sum_{t=0}^{\infty}\gamma^t\frac{2-\rho^2-\rho^{2t}}{(1-\rho^2)^2}\,\|\mathcal{F}'-\mathcal{F}\|\left\|B\Sigma_1B^\top+DW\Sigma_1D^\top\right\| \\
&+\omega_{\gamma,\rho}(\|B\|^2+\|D\|^2\|W\|)\,\|\Sigma_1-\Sigma_2\| \\
=\ & \xi_{\gamma,\rho}\,\|\mathcal{F}-\mathcal{F}'\|\left\|\mathbb{E}[x_0x_0^\top]\right\| + \zeta_{\gamma,\rho}\,\|\mathcal{F}'-\mathcal{F}\|\cdot\left\|B\Sigma_1B^\top+DW\Sigma_1D^\top\right\| \\
&+\omega_{\gamma,\rho}(\|B\|^2+\|D\|^2\|W\|)\,\|\Sigma_1-\Sigma_2\|\,,
\end{aligned}
$$

where (a) follows from (4.3.42); (b) follows from (4.3.39b) and Lemma 4.12. The constants $\xi_{\gamma,\rho}, \zeta_{\gamma,\rho}, \omega_{\gamma,\rho}$ are defined in (4.3.34). Finally, applying Lemma 4.11 finishes the proof. □

Lemma 4.13 (Bound of one-step update of $S^{(t)}$)**.** *Assume that the update of parameter (K,Σ) follows the updating rule in (IPO). Let c_1, c_2 be defined in (4.3.35). Then for $K^{(t+1)}$ satisfying $\|A-BK^{(t+1)}\| \leqslant \rho_1$ and $\|C-DK^{(t+1)}\| \leqslant \rho_2$, we have*

$$\|S^{(t+1)}-S^*\| \leqslant (c_1+c_2)\,\|K^{(t)}-K^*\|.$$

For the proof of Lemma 4.13, we derive the bounds for $\left\|K^{(t+1)} - K^*\right\|$ and $\left\|\Sigma^{(t+1)} - \Sigma^*\right\|$ in terms of $\left\|K^{(t)} - K^*\right\|$ and $\|P_{K^{(t)}} - P_{K^*}\|$(Lemma 4.14 and Lemma 4.15). Additionally, the perturbation analysis for P_K in Lemma 4.16 demonstrates $\|P_{K^{(t)}} - P_{K^*}\|$ can be bounded by $\left\|K^{(t)} - K^*\right\|$, which completes the proof for Lemma 4.13.

We first establish the bound for the one-step update of K and Σ in Lemma 4.14 and Lemma 4.15, respectively.

Lemma 4.14 (Bound of one-step update of K)**.** *Assume the update of parameter K follows the updating rule in (IPO). Then it holds that:*

$$\begin{aligned}\|K^{(t+1)} - K^*\| \leqslant & \left(1 + \sigma_{\min}(R)\left\|\gamma B^\top P_{K^*} B + \gamma D^\top W P_{K^*} D + R\right\|\right) \cdot \|K^{(t)} - K^*\| \\ & + \gamma\sigma_{\min}(R)\left[\left(\|B\|\|A\| - \|D\|\|W\|\|C\|\right) + \left(\|B\|^2 + \|D\|^2\|W\|\right)\kappa\right)\Big] \\ & \|P_{K^{(t)}} - P_{K^*}\|.\end{aligned}$$

Proof of Lemma 4.14. Let K' denote $K^{(t+1)}$ and K denote $K^{(t)}$ to ease the notation. Theorem 4.1 shows that for an optimal K^*,

$$K^* = \gamma(R + \gamma B^\top P_{K^*} B + \gamma D^\top W P_{K^*} D)^{-1} B^\top P_{K^*} A.$$

Then, by the definition of E_K in Lemma 4.2,

$$E_{K^*} = -\gamma(B^\top P_{K^*} A + D^\top W P_{K^*} C) + (\gamma B^\top P_{K^*} B + \gamma D^\top W P_{K^*} D + R)K^* = 0. \tag{4.3.43}$$

Now we bound the difference between $K' - K$:

$$\begin{aligned}& \|K' - K^*\| \\ = & \left\|K - (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_K - K^*\right. \\ & \left. + (R + \gamma B^\top P_K B + \gamma D^\top W P_K D)^{-1} E_{K^*}\right\| \qquad (4.3.44) \\ \leqslant & \|K - K^*\| + \|(R + \gamma B^\top P_{K^*} B + \gamma D^\top W P_{K^*} D)^{-1}\| \, \|E_{K^*} - E_K\| \\ \leqslant & \|K - K^*\| + \sigma_{\min}(R)\|E_K - E_{K^*}\|.\end{aligned}$$

To bound the difference between E_K and E_{K*}, observe:

$$\begin{aligned}\|E_K - E_{K*}\| \leqslant{}& \gamma(\|B\|\|A\| - \|D\|\|W\|\|C\|)\,\|P_K - P_{K*}\| \\ &+ \|(\gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D + R)K^* - (\gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D + R)K\| \\ &+ \|(\gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D + R)K - (\gamma B^\top P_K B + \gamma D^\top W P_K D + R)K\| \\ \leqslant{}& \gamma(\|B\|\|A\| - \|D\|\|W\|\|C\|)\,\|P_K - P_{K*}\| + \|\gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D + R\| \\ &\|K^* - K\| + \gamma(\|B\|^2 + \|D\|^2\|W\|)\,\|P_{K*} - P_K\|\,\|K\|.\end{aligned} \tag{4.3.45}$$

Combining (4.3.44) and (4.3.45) and then using $\kappa \geqslant \|K\|$ for any $K \in \Omega$ completes the proof. □

Lemma 4.15 (Bound of one-step update of Σ)**.** *Suppose* $\left\{K^{(t)}, \Sigma^{(t)}\right\}_{t=0}^{\infty}$ *follows the update rule in IPO. Then we have* $\left\|\Sigma^{(t+1)} - \Sigma^*\right\| \leqslant \frac{\tau\gamma(\|B\|^2+\|D\|^2\|W\|)}{\sigma_{\min}(R)}\,\|P_{K^{(t)}} - P_{K*}\|$.

Proof of Lemma 4.15. Observe from (IPO) and (4.1.7) that

$$\begin{aligned}&\left\|\Sigma^{(t+1)} - \Sigma^*\right\| \\ ={}& \frac{\tau}{2}\left\|(R + \gamma B^\top P_{K^{(t)}} B + \gamma D^\top W P_{K^{(t)}} D)^{-1} - (R + \gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D)^{-1}\right\| \\ ={}& \frac{\tau}{2}\left\|(R + \gamma B^\top P_{K^{(t)}} B + \gamma D^\top W P_{K^{(t)}} D)^{-1} \cdot \gamma\left(B^\top\left(P_{K^{(t)}} - P_{K*}\right)B + D^\top W\left(P_{K^{(t)}} - P_{K*}\right)D\right)\right. \\ &\left.(R + \gamma B^\top P_{K*} B + \gamma D^\top W P_{K*} D)^{-1}\right\| \\ \leqslant{}& \frac{\tau\gamma(\|B\|^2 + \|D\|^2\|W\|)}{2\sigma_{\min}(R)^2}\left\|P_{K^{(t)}} - P_{K*}\right\|.\end{aligned}$$

□

Lemma 4.16 performs perturbation analysis on P_K and establishes bounds for the difference in P_K with respect to the perturbation in K. Consequently, both $\left\|K^{(t+1)} - K^*\right\|$ and $\left\|\Sigma^{(t+1)} - \Sigma^*\right\|$ (in Lemma 4.14 and Lemma 4.15) can be bounded in terms of $\left\|K^{(t)} - K^*\right\|$.

Lemma 4.16 (P_K perturbation)**.** *For any $K \in \Omega$, with c defined in (4.3.35),* $\|P_K - P_{K*}\| \leqslant c\,\|K - K^*\|$.

Proof of Lemma 4.16. Fix $K \in \Omega$. By (4.3.38) and (4.2.18),

$$\begin{aligned}\mathcal{T}_K^{-1}(P_K) &= (I - \gamma\mathcal{F}_K)(P_K)\\ &= P_K - \gamma[(A-BK)^\top(A-BK) + (C-DK)^\top W(C-DK)]P_K\\ &= Q + K^\top RK,\end{aligned}$$

which immediately implies $P_K = \mathcal{T}_K\left(Q + K^\top RK\right)$. Similarly $P_{K*} = \mathcal{T}_{K*}\left(Q + K^{*\top} RK^*\right)$. To bound the difference between P_K and P_{K*}, observe that

$$\|P_K - P_{K*}\| \leqslant \left\|\mathcal{T}_K\left(Q + K^\top RK\right) - \mathcal{T}_{K*}\left(Q + K^\top RK\right)\right\| + \|\mathcal{T}_{K*}\|\,\|K^{*\top}RK^* - K^\top RK\|. \tag{4.3.46}$$

For the first term in (4.3.46), we can apply Lemma 4.12 and Lemma 4.11 to get

$$\begin{aligned}&\left\|\mathcal{T}_K\left(Q + K^\top RK\right) - \mathcal{T}_{K*}\left(Q + K^\top RK\right)\right\|\\ &\leqslant \xi_{\gamma,\rho}\,\|\mathcal{F}_{K*} - \mathcal{F}_K\|\,\|Q + K^\top RK\|\\ &\leqslant 2\xi_{\gamma,\rho}(\rho_1\|B\| + \rho_2\|D\|\|W\|)\,\|K^* - K\|\cdot\left(\|Q\| + \|R\|\|K\|^2\right).\end{aligned} \tag{4.3.47}$$

For the second term in (4.3.46), $\|\mathcal{T}_K\| \leqslant \frac{1}{\mu}\left\|\mathcal{T}_K\left(\mathbb{E}\left[x_0x_0^\top\right]\right)\right\|$. Since $S_{K,\Sigma} \succeq \mathcal{T}_K\left(\mathbb{E}\left[x_0x_0^\top\right]\right)$, $\|\mathcal{T}_K\| \leqslant \frac{1}{\mu}\sigma_{\max}\left(\mathcal{T}_K\left(\mathbb{E}\left[x_0x_0^\top\right]\right)\right) \leqslant \frac{1}{\mu}\|S_{K,\Sigma}\|$. Then we have

$$\begin{aligned}&\|\mathcal{T}_{K*}\|\,\|K^{*\top}RK^* - K^\top RK\|\\ &= \|\mathcal{T}_{K*}\|\,\|K^{*\top}RK^* - K^{*\top}RK + K^{*\top}RK - K^\top RK\|\\ &\leqslant \|\mathcal{T}_{K*}\|\,\|R\|\,\|K - K^*\|\,(\|K^*\| + \|K\|)\\ &\leqslant \frac{\|S_{K^*,\Sigma^*}\|}{\mu}\|R\|\,\|K - K^*\|\,(\|K^*\| + \|K\|).\end{aligned} \tag{4.3.48}$$

Substituting (4.3.47),(4.3.48), and (4.3.35) in (4.3.46) completes the proof. □

With these lemmas, the proof of Lemma 4.13 is completed as follows:

Proof of Lemma 4.13. Let K' denote $K^{(t+1)}$ and K denote $K^{(t)}$ to ease the notation. Apply Lemma 4.16 with K and K^* to get

$$\|P_{K^*} - P_K\| \leqslant c\,\|K - K^*\|\,. \tag{4.3.49}$$

Based on the assumption that $\|A - BK^*\| \leqslant \rho_1$, $\|A - BK'\| \leqslant \rho_1$, $\|C - DK^*\| \leqslant \rho_2$, and $\|C - DK'\| \leqslant \rho_2$, we can apply Lemma 4.9 to obtain:

$$\begin{aligned}
&\|S_{K^*,\Sigma^*} - S_{K',\Sigma'}\| \\
\leqslant\ &\left(\xi_{\gamma,\rho}\|\mathbb{E}[x_0 x_0^\top]\| + \zeta_{\gamma,\rho}\|B\Sigma^* B^\top + DW\Sigma^* D^\top\|\right)\cdot 2(\rho_1\|B\| + \rho_2\|D\|\|W\|)\,\|K^* - K'\| \\
&+ \omega_{\gamma,\rho}(\|B\|^2 + \|D\|^2\|W\|)\,\|\Sigma^* - \Sigma'\|\,.
\end{aligned} \tag{4.3.50}$$

Applying Lemma 4.14 with $\|K\| \leqslant \kappa$ and (4.3.49) yields

$$\begin{aligned}
&\|K' - K^*\| \\
\leqslant\ &\left(1 + \sigma_{\min}(R)\|\gamma B^\top P_{K^*} B + \gamma D^\top W P_{K^*} D + R\|\right)\|K - K^*\| \\
&+ \gamma\sigma_{\min}(R)\Big[(\|B\|\|A\| - \|D\|\|W\|\|C\|) + (\|B\|^2 + \|D\|^2\|W\|)\kappa\Big]\|P_K - P_{K^*}\| \\
\leqslant\ &\Big(1 + \sigma_{\min}(R)\|\gamma B^\top P_{K^*} B + D^\top W P_{K^*} D + R\| + c\gamma\sigma_{\min}(R) \\
&\Big[(\|B\|\|A\| - \|D\|\|W\|\|C\|) + (\|B\|^2 + \|D\|^2\|W\|)\kappa\Big]\Big)\,\|K - K^*\|\,.
\end{aligned}$$

Apply Lemma 4.15 with (4.3.49) to get

$$\|\Sigma' - \Sigma^*\| \leqslant \frac{\tau\gamma(\|B\|^2 + \|D\|^2\|W\|)}{2\sigma_{\min}(R)^2}\,\|P_K - P_{K^*}\| \leqslant c\frac{\tau\gamma(\|B\|^2 + \|D\|^2\|W\|)}{2\sigma_{\min}(R)^2}\,\|K - K^*\|\,.$$

Finally, substituting the above two inequalities into (4.3.50) finishes the proof. □

Proof of Theorem 4.3. Fix integer $t \geqslant 0$. Observe that

$$\begin{aligned}
&\left(1-\frac{\mu}{\|S_{K^*,\Sigma^*}\|}\right)\left(C\left(K^{(t)},\Sigma^{(t)}\right)-C\left(K^*,\Sigma^*\right)\right)\\
&\overset{(a)}{\geqslant} C(K^{(t+1)},\Sigma^{(t+1)})-C\left(K^*,\Sigma^*\right)\\
&\overset{(b)}{\geqslant} \mu\sigma_{\min}\left(R+\gamma BP_{K^*}B+\gamma DWP_{K^*}D\right)\|K^{(t+1)}-K^*\|^2+f_{K^*}\left(\Sigma^*\right)-f_{K^*}(\Sigma^{(t+1)})\\
&\overset{(c)}{\geqslant} \mu\sigma_{\min}\left(R+\gamma BP_{K^*}B+\gamma DWP_{K^*}D\right)\|K^{(t+1)}-K^*\|^2.
\end{aligned} \tag{4.3.51}$$

(a) is clear from the contraction property in Lemma 4.7; (b) follows from Lemma 4.6 and (4.3.43); to obtain (c), note that $f_{K^*}\left(\Sigma^*\right)-f_{K^*}(\Sigma^{(t+1)})\geqslant 0$, since Σ^* is the maximizer of f_{K^*}. Thus, (4.3.51) and (4.3.36) imply

$$\|K^{(t+1)}-K^*\|\leqslant \min\left\{\frac{1}{c_1+c_2}\sigma_{\min}\left(S^*\right),\frac{\rho_1-\|A-BK^*\|}{\|B\|},\frac{\rho_2-\|C-DK^*\|}{\|D\|}\right\},$$

which suggests that

$$\begin{aligned}
&\|A-BK^{(t+1)}\|\leqslant\|A-BK^*\|+\|B\|\|K^{(t+1)}-K^*\|\leqslant\rho_1,\ \text{and}\\
&\|C-DK^{(t+1)}\|\leqslant\|C-DK^*\|+\|D\|\|K^{(t+1)}-K^*\|\leqslant\rho_2.
\end{aligned}$$

The above inequality, along with Lemma 4.13, leads

$$\|S^{(t+1)}-S^*\|\leqslant(c_1+c_2)\|K^{(t)}-K^*\|\leqslant\sigma_{\min}\left(S^*\right). \tag{4.3.52}$$

Then, apply the Lemma 4.8 and (4.3.52) to get

$$\begin{aligned}
&C(K^{(t+1)},\Sigma^{(t+1)})-C\left(K^*,\Sigma^*\right)\\
\leqslant&\frac{(c_1+c_2)}{\sigma_{\min}\left(S^*\right)}\|K^{(t)}-K^*\|\left(C(K^{(t)},\Sigma^{(t)})-C\left(K^*,\Sigma^*\right)\right).
\end{aligned} \tag{4.3.53}$$

Using the same reasoning as in (4.3.51), we have

$$C(K^{(t)},\Sigma^{(t)})-C\left(K^*,\Sigma^*\right)\geqslant\mu\sigma_{\min}(R+\gamma B^\top P_{K^*}B+\gamma D^\top WP_{K^*}D)\|K^{(t)}-K^*\|^2.$$

Substituting the above results in (4.3.53) finishes the proof. □

4.4 Numerical Examples

We follow the basic setup of the MV problem and optimal liquidation problem in Chapter 2 with some modifications, while the same performance measure acts as a quantitative indicator.

4.4.1 Application: MV Portfolio Selection

Sometimes investment decisions are made not by a single centralized planner but by multiple agents, each focusing on a specific sector or investment theme. For instance, a university endowment or pension fund may divide its capital among specialized sub-funds (agents): one focused on medical technology innovation, another on renewable energy, and a third on emerging markets. These sub-funds may act autonomously under a shared global investment objective while managing distinct exposures and sectoral risks. This naturally motivates a decentralized formulation of the MV problem.

We extend the standard MV model to a decentralized multi-agent setting. Suppose there are N sub-funds for different investment themes, each with access to a local set of n risky assets and a common riskless asset. The variable x_t and u_t are global setting that includes all agent $j \in \{1, \cdots, N\}$. Consider the infinite time MV problem in this section with Shannon entropy in a global objective:

$$\min_{u_t} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^2 + u_t^\top R u_t + \tau \log \pi \left(u_t | x_t \right) \right) \middle| x_0 = x \right]. \tag{4.4.54}$$

Then, we shall transform (2.3.17) into a linear controlled system of the form (2.1.1) by which the general theory in the above sections will work. Precisely, we define system dynamics for each agent j:

$$\begin{cases} w_{j,t}^i = e_{j,t}^i - \mathbb{E}(e_{j,t}^i), \\ D_{j,t}^i = (0, \cdots, 0, 1, 0, \cdots, 0), \\ i = 1, \cdots, n, \ t = 0, 1, \cdots, \infty, j = 1, \cdots, N, \end{cases}$$

which leads to

$$x_{j,t+1} = (r_{j,t}x_{j,t} + \mathbb{E}(P_{j,t})u_{j,t}) + \sum_{i=1}^{n} D^i_{j,t}u_{j,t}w^i_{j,t}. \tag{4.4.55}$$

We now aggregate the agent-wise dynamics into a single system-level linear stochastic equation, following the recipe in Section 4.1.2:

$$x_{t+1} = (r_t x_t + \mathbb{E}(P_t)u_t) + \sum_{i=1}^{n} D^i_t u_t w^i_t. \tag{4.4.56}$$

It is not hard to see that the problem (MV) is just a special case of the problem in Section 2 by letting $x_t \in \mathbb{R}^N$, $u_t \in \mathbb{R}^{nN}$, $Q \in \mathbb{R}^N$, $R \in \mathbb{R}^{nN\times nN}$, $A_j = r_{j,t}$, $B_j = \mathbb{E}(P_{j,t})$, $Q = \mathbf{I}_N$ and $C = \mathbf{0}_{N\times N}$ for $t = 0, \cdots, \infty$. In performance measurement defined in Chapter 2, K^*, Σ^* is the optimal policy defined in Section 4.1.4 and Theorem 4.1. Then, the following provides numerical experiments using (IPO) to illustrate the theoretical results established in the sections before.

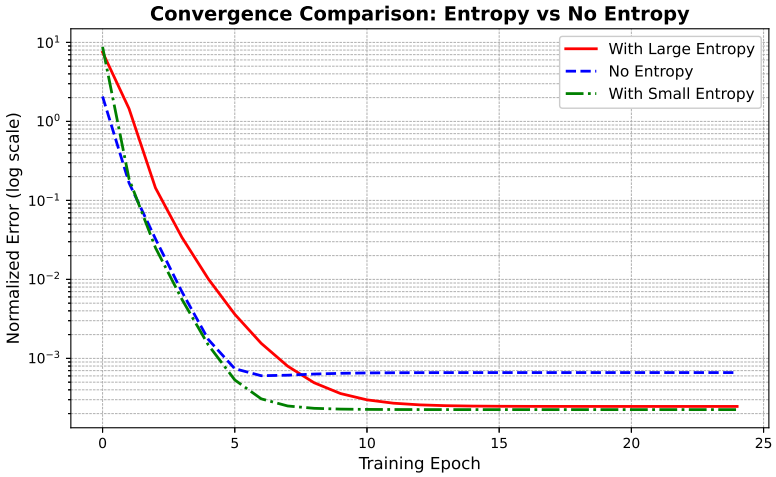


Figure 4.1: Convergence Comparison: Entropy vs. No Entropy

Setup. (1) Parameters: $m = 1$, $n = 3$, $N = 3$. We establish the risk-free rate at 6%; hence, $A_j = 1.06$ for all j. The scale of A is selected to ensure that A

is chosen satisfying $\sigma_{\max}(A) < \frac{1}{\sqrt{\gamma}}$, with γ set to 0.85. (2) Agent-specific parameters are as follows: Agent 1 (medical technology) has an expected excess return matrix $B_1 = \begin{bmatrix}19.5\% & 9.67\% & 27.9\%\end{bmatrix}$, a noise second moment $W_1 = 0.98$. Agent 2 (renewable energy) has $B_2 = \begin{bmatrix}30.7\% & 23.7\% & 33.5\%\end{bmatrix}$, $W_2 = 0.99$. Agent 3 (emerging markets) has $B_3 = \begin{bmatrix}18.85\% & 12\% & 19.4\%\end{bmatrix}$, $W_3 = 0.98$. For each agent, the penalty term R_j is chosen as a diagonal positive definite matrix whose diagonal entries are sampled from $(0, 10^{-3})$. (3) The regularization parameter is selected with $\tau = 0.9$. The starting policy is established as $K^{(0)}_{\text{common}} = 0.01$ for each entry with $K^{(0)} = \mathbf{I}_N \otimes K^{(0)}_{\text{common}}$, and $\Sigma^{(0)} = \mathbf{I}_{nN}$. (4) We assume that the initial wealth $x_{j,0}$ is drawn from $\mathcal{N}(1, 5 \times 10^{-4})$ for all j.

Convergence. The green curve in the convergence plot represents the training error under Shannon entropy regularization ($\tau = 0.9$). This setting shows a smooth decrease in normalized error throughout the training process. Figure 4.1 shows the superior convergence rate of (IPO). The normalized error falls below 10^{-3} within just 5 iterations, and from the first iteration, it enters a region of super-linear convergence in a log scale normalized error.

The effect of τ. The convergence plot illustrates the log scale normalized error over training epochs for three settings: no entropy term, small entropy ($\tau = 0.9$), and large entropy ($\tau = 10$). All curves exhibit a general downward trend, indicating successful learning, but they differ significantly in convergence speed and error decay patterns. The no entropy case (blue) converges the fastest in the early stages, with the error decreasing rapidly in the first few epochs. However, it plateaus early and fails to further reduce the error beyond a certain point. In contrast, the small entropy case (green), though slower initially, demonstrates a super-linear convergence

pattern. The error decreases gradually at first, then sharply accelerates in later epochs, ultimately reaching the lowest final error among all settings. This process reflects the effect of entropy regularization in smoothing early updates while enabling sharper optimization later on. The large entropy (red) setting converges the slowest and smoothest and flattens at a relatively higher error level than the smaller τ case, indicating that excessive regularization can hinder convergence precision. Overall, the results highlight the trade-off between initial speed and final accuracy, with small τ achieving the most favorable convergence profile through a clear super-linear improvement trajectory.

4.4.2 Application: Optimal Liquidation

In many institutional trading scenarios, liquidation tasks are delegated to multiple agents, each responsible for managing assets in a specific market. For instance, a global investment firm may allocate portions of a large equity position to N regional trading desks (agents)—one operating in Asia, another in Europe, and a third in North America. Each desk j executes trades according to its local market microstructure and cost sensitivities but still contributes to the overall liquidation goal. In the setting for the optimal liquidation problem in Chapter 2, joined with the framework in Section 4.1.2, we take the global cost function to be

$$TC_{\text{LQ}} = \min \left[\frac{1}{2}\beta^{pi} q^2 + \sum_{t=0}^{\infty} \gamma^t \left(\delta u_t^2 + \phi \sigma^2 q_t^2 + \xi S_t^2 + \tau \log \pi(u_t|x_t) \right) \right]. \tag{4.4.57}$$

Here, for each agent j, $x_{j,t} = \begin{bmatrix} S_{j,t} & q_{j,t} \end{bmatrix}^\top$, with the dynamics in (2.3.18). Note that based on the model in Hambly et al. (2021), where ξ_j is a very small positive number, we add the entropy regularization on it, with $\gamma \in (0,1)$ and $\tau > 0$ being the discount factor and entropy regularization parameter, respectively. We also have the

parameters for each agent j:

$$A_j = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B_j = \begin{bmatrix} -\beta_j^{pi} & -1 \end{bmatrix}^\top, \quad C_j = \mathbf{0}_{2\times 2}, \quad D_j = \begin{bmatrix} -1 & 0 \end{bmatrix}^\top \text{ and } \quad w_{j,t} = \varepsilon_{j,t},$$

and the objective function has

$$Q_j = \begin{bmatrix} \xi_j & 0 \\ 0 & \phi_j \sigma_j^2 \end{bmatrix} \quad \text{and} \quad R_j = \delta_j.$$

Following similar assumptions for u_t, stock price, inventory, and market impact, and using the projected method in Hambly et al. (2021), we will show super-linear convergence for IPO in the optimal liquidation problem.

Setup.(1) Parameters: We consider a decentralized setting with $N = 3$ agents. Each agent $j \in \{1, 2, 3\}$ is assigned coefficient ϕ_j, with $\phi_1 = 5 \times 10^{-6}$, $\phi_2 = 4.2 \times 10^{-6}$ and $\phi_3 = 3.5 \times 10^{-6}$ reflecting heterogeneous transaction cost sensitivities. All agents share the similar volatility parameter $\sigma_{S,1} = 0.107$, $\sigma_{S,2} = 0.108$, $\sigma_{S,3} = 0.109$, and $\xi_j = 10^{-8}$ for all j. The discount factor $\gamma = 0.94$. Initial policy $K^0_{\text{common}} = -0.2$ for all entries; step sizes are as indicated in the figures; β_j^{pi} and β_j^{ti} are randomly generated between $(10^{-6}, 5 \times 10^{-5})$ for the projection set and stay unchanged over t. (2) For initialization, assume the initial inventory $q_{j,0} = q_j$ follows $\mathcal{N}(500, 1)$. The small variance of the initial inventory distribution is used to guarantee that the initial state covariance matrix is positive definite. In practice, the model converges with deterministic initial inventories.

Convergence. Figure 4.2 illustrates the convergence behavior of the optimal liquidation algorithm in terms of three indices: normalized error, policy distance $\|K - K^*\|_F$, and final cost. The normalized error, shown on the left log-scale axis, exhibits a sharp and smooth decline across iterations, reaching below 10^{-10}, which indicates rapid and

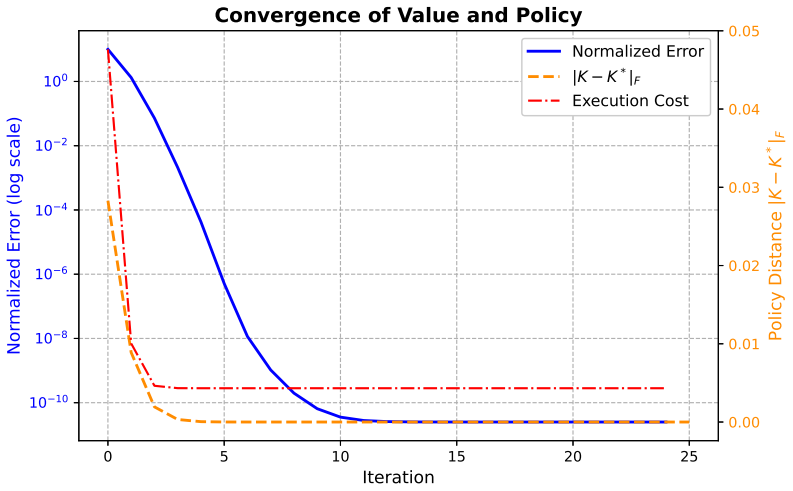


Figure 4.2: Convergence of Value and Policy

stable convergence of the value function. On the right axis, both the policy distance and execution cost decrease gradually. Initially, reductions in $\|K - K^*\|_F$ correspond closely with the drop in normalized error, showing that learning a better policy directly improves value accuracy. As training progresses, the policy distance stabilizes while the normalized error continues to decline, suggesting that even minor refinements in K can yield significant value improvements. Meanwhile, the execution cost steadily decreases, confirming that the learned policy becomes not only theoretically optimal but also practically more efficient.

Impact of parameter ξ. The convergence rate of the policy optimization algorithm is significantly influenced by the choice of the parameter ξ, which is related to the inclusion of an extra regularization component represented by $\sum_{t=0}^{\infty} \xi S_t^2$. The purpose of this term is to allow the problem to be formulated inside the LQ framework and to ensure the clarity and validity of the Riccati equation. As shown in Figure 4.3, we observe that smaller values of ξ (e.g., $\xi = 10^{-8.5}$) lead to faster error decay during the early convergence phase. However, larger ξ values (e.g., $\xi = 10^{-7.8}$) yield lower final error levels, with convergence reaching below 10^{-11} in comparison to a plateau

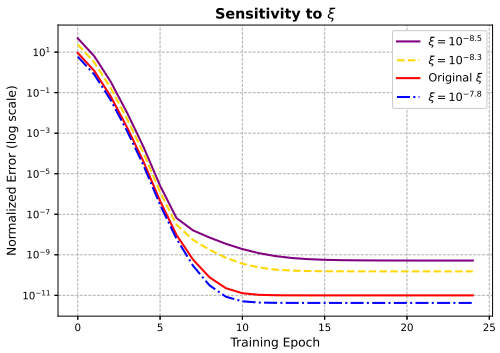


Figure 4.3: Sensitivity to ξ

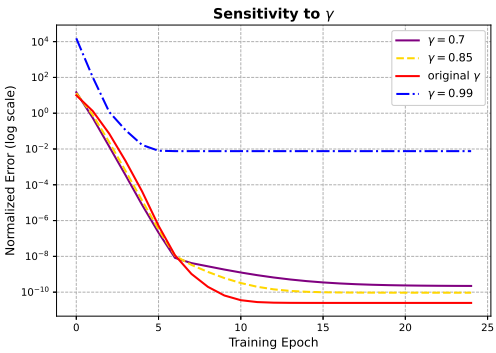


Figure 4.4: Sensitivity to γ

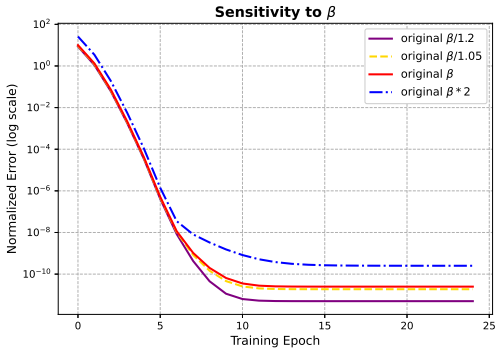


Figure 4.5: Sensitivity to β^{ti} (β on the figure is the old symbol)

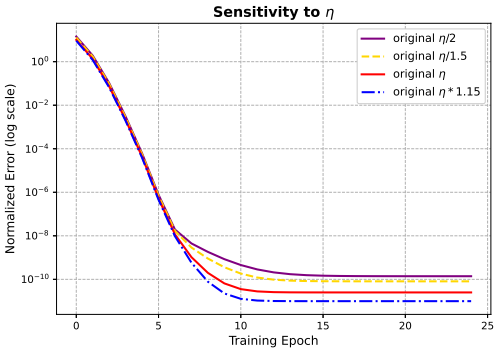


Figure 4.6: Sensitivity to β^{pi} (η on the figure is the old symbol)

around 10^{-9} for smaller ξ. This trade-off suggests that while weaker regularization accelerates early learning, stronger regularization results in more numerically stable and accurate final solutions. In practical terms, penalizing inventory more heavily prevents the policy from retaining residual positions and thus promotes more consistent liquidation, even under uncertainty.

Impact of parameter γ. The discount factor γ plays a crucial role in our LQ model with entropy regularization, influencing the speed and pattern of convergence during training. In Figure 4.4, we explore $\gamma \in \{0.7, 0.85, 0.94, 0.99\}$. The three smaller values all lead to stable convergence around a similar final error level of 10^{-10}, with

slightly lower final errors as γ increases. However, when $\gamma = 0.99$, the algorithm fails to converge properly, with the error stalling at approximately 10^{-2}. This phenomenon is likely due to the fact that when γ is too close to 1, it implies almost no discounting, which effectively transforms the problem into an undiscounted (or nearly infinite horizon) one. Such a setting can de-emphasize immediate inventory liquidation, leading to less aggressive policies and degraded numerical convergence.

Impact of parameter β^{ti}. Figure 4.5 presents the sensitivity to the temporary impact parameter β^{ti}, which captures the instantaneous cost of executing trades. We vary original β^{ti} within a practical range-specifically, using $\beta^{ti}/1.5$, $\beta^{ti}/1.2$, and $1.15\beta^{ti}$—to ensure the constraint $\delta = \beta^{ti} - \beta^{pi}/2 > 0$ is maintained. The results show that larger β^{ti} values lead to faster convergence, while smaller β^{ti} values achieve lower final error, thus indicating more accurate solutions. This makes intuitive sense: when the temporary impact is high, the algorithm is incentivized to slow down execution and make more conservative updates, which stabilizes training. In contrast, lower impact encourages more aggressive trading behavior, which, while harder to optimize numerically, results in sharper and more precise policies.

Impact of parameter β^{pi}. Figure 4.6 reports convergence under varying levels of permanent impact β^{pi}, again under the constraint that $\delta = \beta^{ti} - \beta^{pi}/2 > 0$. Unlike the case with β^{ti}, here we observe the opposite effect: increasing β^{pi} slightly leads to slower convergence but lower final error. This inverse relationship aligns with the form of the effective impact coefficient δ, where a larger β^{pi} reduces δ, similar to the effect of decreasing β^{ti}. From a financial perspective, greater permanent impact discourages aggressive liquidation, and the algorithm must carefully balance the long-term cost of price influence, which leads to slower but ultimately more refined policy updates.

Chapter 5

Data-Driven Long-Term Asset Allocation with Tsallis Entropy Regularization

This chapter explores the application of Tsallis entropy in multiplicative LQ models and introduces an innovative policy iteration approach. The Tsallis entropy framework, serving as a one-parameter extension of Shannon entropy, is employed to regularize the control problem, thereby providing a versatile method to balance exploration and sparsity in the control law. We obtain explicit solutions within this framework, illustrating its efficacy in improving robustness and adaptability for stochastic control systems characterized by multiplicative noise. We formulate a model-free, entirely data-driven policy iteration technique to tackle the issues of state- and input-multiplicative noise in LQ systems. This methodology utilizes Q-functions, estimated via a least-squares problem with instrumental variables, similar to least-squares temporal-difference (LSTD) methods. By including Tsallis entropy into this data-driven policy iteration framework, we present a holistic approach that enhances the efficiency of the RL-based policy. This methodology fills voids in current research and creates a solid basis for the application of entropy-based techniques to stochastic control in intricate settings.

5.1 Problem Setup

5.1.1 Model Formulation

On the basis of the LQ problem in Section 2.1, we add Tsallis entropy on (2.1.3) and the noise term of A and B on (2.1.1) as

$$J_\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^\top Q x_t + u_t^\top R u_t + \frac{\tau}{2-q} \left(\log_q \pi(u_t|x_t) - 1 \right) \right) \middle| x_0 = x \right], \quad (5.1.1)$$

such that $t = 0, 1, 2, \cdots,$

$$x_{t+1} = (A + \Delta A_t)x_t + (B + \Delta B_t)u_t + (Cx_t + Du_t)w_t, \ x_0 \sim \mathcal{D}_0, \quad (5.1.2)$$

where $\tau > 0$ is the regularization parameter, and the remaining coefficients and variables retain the same interpretation as in Section 2.1.

Assumption 5.1 (Noise Model)**.** *Let $\Delta A_t \in \mathbb{R}^{m\times m}$ and $\Delta B_t \in \mathbb{R}^{m\times n}$ denote matrix-valued parameter noises and let $w_t \in \mathbb{R}$ denote the system noise.*

(i) The sequences $\{\Delta A_t\}$, $\{\Delta B_t\}$, and $\{w_t\}$ are zero-mean, mutually i.i.d. in t, and independent of the initial state.

(ii) All noises have finite second moments. In particular, define the covariance matrices $\Sigma_A = \mathbb{E}\big[\Delta A_t \Delta A_t^\top\big]$, $\Sigma_B = \mathbb{E}\big[\Delta B_t \Delta B_t^\top\big]$, $W = \mathbb{E}\big[w_t w_t^\top\big]$ and assume $\mathrm{Tr}(\Sigma_A) < \infty$, $\mathrm{Tr}(\Sigma_B) < \infty$, and $W < \infty$.

(iii) For every symmetric matrix, for example, matrix G, the second-moment operators $\Sigma_A(G) = \mathbb{E}\big[\Delta A_t^\top G \Delta A_t\big]$ and $\Sigma_B(G) = \mathbb{E}\big[\Delta B_t^\top G \Delta B_t\big]$ are well defined with finite entries.

Remark 5.1. *The parameter noises ΔA_t, ΔB_t perturb the transition matrices A and B, whereas w_t is the process noise entering through $Cx_t + Du_t$. Zero means and*

mutual independence eliminate drift and mixed cross terms in the Bellman update, preserving a quadratic value form and linear state feedback structure. The identically distributed assumption yields time-homogeneous second moments, leading to algebraic time-invariant ARE and stationary optimal policies. Finite second moments ensure all trace terms are well defined.

5.1.2 Optimal Strategy and Algebraic Riccati Equation

The optimal value function $J^* : \mathcal{X} \to \mathbb{R}$ is defined as

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x). \tag{5.1.3}$$

This theorem provides the explicit formulation for the optimal control policy and the associated optimal value function: the optimal policy is defined as a multivariate q-Gaussian distribution, with the mean being linear in the state x and a constant covariance matrix.

Theorem 5.1 (Optimal value functions and optimal policy)**.** *Under the Assumption 5.1, the optimal value function $J^* : \mathcal{X} \to \mathbb{R}$ in (5.1.3) can be expressed as*

$$J^*(x) = x^\top P x + c, \tag{5.1.4}$$

where

$$\begin{aligned} P &= Q + K^{*\top} R K^* + \gamma \left(A - BK^*\right)^\top P \left(A - BK^*\right) \\ &\quad + \gamma \left(C - DK^*\right)^\top W P \left(C - DK^*\right) + \gamma \Sigma_A(P) + \gamma K^{*\top} \Sigma_B(P) K^*, \\ c &= \frac{1}{1-\gamma} \left[\mathrm{Tr}\left(\Sigma^* (R + \gamma B^\top P B + \gamma \Sigma_B(P) + \gamma W D^\top P D) \right) - \frac{\tau}{2-q} \right. \\ &\quad \left. \left(\log_q \left(Z_q\right) + \frac{n}{(n+4) - (n+2)q} + \frac{n(q-1)\log_q\left(Z_q\right)}{(n+4) - (n+2)q} + 1 \right) \right], \end{aligned} \tag{5.1.5}$$

with

$$
\begin{aligned}
K^* &= \gamma \left(R + \gamma B^\top P B + \gamma \Sigma_B(P) + \gamma W D^\top P D\right)^{-1} (B^\top P A + W D^\top P C), \\
\Sigma^* &= \frac{\tau}{(n+4) - (n+2)q} \left(R + \gamma B^\top P B + \gamma \Sigma_B(P) + \gamma W D^\top P D\right)^{-1}.
\end{aligned} \tag{5.1.6}
$$

For any $x \in \mathcal{X}$, the corresponding optimal policy π^* to (5.1.3) is a q-Gaussian policy

$$
\pi^*(\cdot|x) = \mathcal{N}_q \left(-K^* x, \Sigma^*\right), \tag{5.1.7}
$$

where $\mu^* = -K^* x$.

The constant c in $J^*(x) = x^\top P x + c$ collects all state-independent terms induced by the entropy-regularized stochastic policy and the entropy contribution. These terms appear because the optimal policy has nonzero covariance when $\tau > 0$. In the unregularized case $\tau = 0$, the optimal policy becomes deterministic and $\Sigma^* = 0$; therefore, the state-independent terms vanish, $c = 0$, and J^* reduces to a purely quadratic form.

Remark 5.2. *In our formulation, the optimal mean control remains linear, while the Tsallis parameters determine the dispersion and support through Σ^*. Therefore, the regularizer affects the stochasticity and boundedness of actions rather than changing the linear feedback structure.*

Proof. By the dynamic programming principle for the optimal value function, J^*

satisfies

$$
\begin{aligned}
J^*(x) = x^\top Q x + \min_\pi \mathbb{E}_\pi \Big\{ & u^\top R u + \frac{\tau}{2-q}\left(\log_q(\pi(u|x)) - 1\right) \\
& + \gamma \Big[((A+\Delta A)x + (B+\Delta B)u + (Cx+Du)w)^\top P \\
& ((A+\Delta A)x + (B+\Delta B)u + (Cx+Du)w) + c \Big] \Big\} \\
= x^\top Q x + \gamma x^\top & (A^\top P A + W C^\top P C + \Sigma_A(P)) x + \gamma c \\
& + \min_\pi \mathbb{E}_\pi \Big\{ u^\top (R + \gamma B^\top P B + \gamma W D^\top P D + \gamma \Sigma_B(P)) u \\
& + \frac{\tau}{2-q}\left(\log_q(\pi(u|x)) - 1\right) + 2\gamma u^\top (B^\top P A + W D^\top P C) x \Big\}.
\end{aligned}
\tag{5.1.8}
$$

We now minimize the expression inside the braces in (5.1.8) with respect to $\pi(u|x)$. Following the variational argument of Guo et al. (2023), for a fixed x, write $\phi = \pi(u|x)$ and denote M and b as the coefficients for quadratic term and linear term of u, respectively. Then consider the functional with the constraint $\int_{\mathcal{U}} \pi(u)du = 1$,

$$
\mathcal{J}(\phi) = \int_{\mathcal{U}} \Big[(u^\top M u + b^\top u)\phi + \frac{\tau}{2-q}\phi\big(\log_q \phi - 1\big) \Big] du,
$$

where λ is a Lagrangian multiplier connected with the constraint. Thus, its Lagrangian is

$$
\mathcal{L}(u, \phi, \lambda) = (u^\top M u + b^\top u)\phi + \frac{\tau}{2-q}\phi(\log_q \phi - 1) + \lambda\phi.
$$

Taking the first derivative of $\mathcal{L}$ with respect to ϕ and setting it to zero gives the stationarity condition $\lambda + u^\top M u + b^\top u + \frac{\tau}{1-q}(\phi^{1-q} - 1) = 0$. Solving for ϕ, we obtain $\phi \propto \left[1 - \frac{1-q}{\tau}(u^\top M u + b^\top u)\right]^{\frac{1}{1-q}}$, which shows that the minimizer is a multivariate q-Gaussian as

$$
\pi^*(\cdot|x) = \mathcal{N}_q\left(-K^* x, \Sigma^*\right).
$$

Since $b = 2\gamma(B^\top PA + WD^\top PC)x$, the above equation can be written as (5.1.7) with K^* and Σ^* in (5.1.6). This proves that the minimizer exists and is q-Gaussian under the Tsallis entropy formulation. To derive the associated optimal value function, we calculate $\mathbb{E}_{\pi^*}\left[\log_q\left(\pi^*(u|x)\right)\right]$ in the entropy of policy π^* at any state $x \in \mathcal{X}$:

$$\begin{aligned}\mathbb{E}_{\pi^*}\left[\log_q\left(\pi^*(u|x)\right)\right] &= \int_{\mathcal{U}} \pi^*(u|x)\log_q\left(\pi^*(u|x)\right)du \\ &= -\log_q\left(Z_q\right) - \alpha n + (1-q)\alpha n \log_q\left(Z_q\right),\end{aligned}$$

where $\alpha = \frac{1}{(n+4)-(n+2)q}$ and Z_q is defined in Definition 2. The calculation follows the product rule for q-Logarithm with $\log_q(a\cdot b) = \log_q(a)+\log_q(b)+(1-q)\log_q(a)\log_q(b)$. Thus, the entropy is

$$\begin{aligned}\mathcal{H}_q(\pi(\cdot|x)) &= -\frac{1}{2-q}\left(\mathbb{E}[\log_q \pi(u|x)] - 1\right) \\ &= \frac{1}{2-q}\left(\log_q\left(Z_q\right) + \alpha n + (q-1)\alpha n \log_q\left(Z_q\right) + 1\right).\end{aligned} \tag{5.1.9}$$

Finally, we substitute the entropy value of $\mathcal{H}_q(\pi(\cdot|x))$ from (5.1.9) and $u^* = -K^*x$ into (5.1.8). Straightforward algebra yields

$$\begin{aligned}J^*(x) = x^\top\Big(&Q + K^{*\top}RK^* + \gamma\left(A - BK^*\right)^\top P\left(A - BK^*\right) \\ &+ \gamma\left(C - DK^*\right)^\top WP\left(C - DK^*\right) + \gamma\Sigma_A(P) + \gamma K^{*\top}\Sigma_B(P)K^*\Big)x \\ &+ \mathrm{Tr}\left(\Sigma^* R\right) - \frac{\tau}{2-q}\left(\log_q\left(Z_q\right) + \alpha n + (q-1)\alpha n\log_q\left(Z_q\right) + 1\right) \\ &+ \gamma\left(\mathrm{Tr}(\Sigma^* B^\top PB) + \mathrm{Tr}\left(\Sigma^*\Sigma_B(P)\right) + \mathrm{Tr}(\Sigma^* WD^\top PD) + c\right).\end{aligned}$$

The equations in (5.1.5) can be observed and the proof is complete. □

5.2 Dynamic Programming for Tsallis Entropy LQ Problem

In this section, we study dynamic programming for the Tsallis entropy-regularized LQ control problem in operator form from the perspective of the Q-function and develop corresponding policy iteration and value iteration schemes. The algorithm computes the optimal value function and control policy, and their convergence is established in Theorems 5.2 and 5.3.

Following the dynamics specified in equation (5.1.2), we write x_+ and $u_+$for next-step state and action. Define

$$Z = \mathbb{E}\left[\begin{bmatrix} x \\ u \end{bmatrix}\begin{bmatrix} x^\top & u^\top \end{bmatrix}\right] = \begin{bmatrix} \mathbf{I}_m \\ -K \end{bmatrix} X \begin{bmatrix} \mathbf{I}_m & -K^\top \end{bmatrix} = \mathcal{M}(X), \tag{5.2.10}$$

where $X = \mathbb{E}[xx^\top]$. Z is the joint second-moment matrix in terms of mean state and control under the policy class with linear feedback $u = -Kx$. To express the second moment of state value of next time X_+ in terms of Z as a function $\mathcal{E}(Z)$:

$$\begin{aligned} X_+ &= \mathbb{E}[x_+ x_+^\top] \\ &= \begin{bmatrix} A & B \end{bmatrix} Z \begin{bmatrix} A & B \end{bmatrix}^\top + W \begin{bmatrix} C & D \end{bmatrix} Z \begin{bmatrix} C & D \end{bmatrix}^\top \\ &\quad + \Sigma_A^\top(\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \end{bmatrix} Z \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \end{bmatrix}^\top) \\ &\quad + \Sigma_B^\top(\begin{bmatrix} \mathbf{0}_{n\times m} & \mathbf{I}_n \end{bmatrix} Z \begin{bmatrix} \mathbf{0}_{n\times m} & \mathbf{I}_n \end{bmatrix}^\top), \end{aligned} \tag{5.2.11}$$

where $\Sigma_A^\top(\cdot)$ and $\Sigma_B^\top(\cdot)$ denote the adjoint operators of $\Sigma_A(\cdot)$ and $\Sigma_B(\cdot)$ in Assumption 5.1 defined by $\Sigma_A^\top(G) = \mathbb{E}[\Delta A_t G \Delta A_t^\top]$, and same is true for $\Sigma_B^\top(G)$ with given matrix G. The value function (5.1.1) is converted to

$$J_\pi(X) = \sum_{t=0}^{\infty} \gamma^t \left(\mathrm{Tr}\left[H\mathcal{M}(X_t)\right] + const'.\right), \tag{5.2.12}$$

where $H = \text{diag}(Q, R)$ and

$$const'. = -\frac{\tau}{2-q}\left(\log_q(Z_q) + \alpha n + (q-1)\alpha n \log_q(Z_q) + 1\right).$$

To formulate our policy iteration method, we examine the Bellman operator (Bertsekas (2022)), defined in accordance with the framework presented therein, as

$$(\mathcal{T}_\pi J)(X) = \text{Tr}\left[\mathcal{M}(X)H\right] + const'. + \gamma J\left(\mathcal{E}\left(\mathcal{M}(X)\right)\right), \tag{5.2.13}$$

and $\mathcal{T}J = \min_\pi \mathcal{T}_\pi J$, operating on $J : \mathcal{X} \to \mathbb{R}$. The optimal value of (5.1.3) is determined as the fixed-point of $\mathcal{T}$ if the dynamics can be stabilized, meaning that a mean-square stabilizing controller exists. The optimal controller, according to (5.2.13), is $\text{argmin}_\pi \mathcal{T}_\pi J^*$, where $J^* = \mathcal{T}J^*$.

We aim to determine K^* via policy iteration, which involves the iterative updating of K to identify a K^+ such that

$$K^+ = \underset{K'}{\text{argmin}} \mathcal{T}_{K'} J_K, \text{ with } J_K = \mathcal{T}_K J_K. \tag{5.2.14}$$

5.2.1 Q-Function

A Q-function in this context maps Z to some real number, and we define the Bellman operator with it as:

$$(\mathcal{F}_\pi \mathcal{Q})(Z) = \text{Tr}\left[ZH\right] + const'. + \gamma \mathcal{Q}\left(\mathcal{M}\left(\mathcal{E}(Z)\right)\right), \tag{5.2.15}$$

and $\mathcal{F}\mathcal{Q} = \min_\pi \mathcal{F}_\pi \mathcal{Q}$. For some policy π, we define $\mathcal{Q}_\pi$ as

$$\mathcal{Q}_\pi(Z) = \text{Tr}\left[ZH\right] + const'. + \gamma J_\pi\left(\mathcal{E}(Z)\right), \tag{5.2.16}$$

where J_π solves $J_\pi = \mathcal{T}_\pi J_\pi$.

To ensure that the value functions J_π and Q-function Q_π and the operators $\mathcal{T}_\pi$ and $\mathcal{F}_\pi$ are well defined and finite for all policies considered in this section, we impose the following standing condition on the policy class Π.

Assumption 5.2 (Admissible Tsallis-regularized policies)**.** *Let the discount factor satisfy* $\gamma \in (0,1)$*. We consider a class* Π *of stationary Tsallis-regularized policies* π *such that:*

(i) For each $\pi \in \Pi$*, the closed-loop system induced by (5.1.2) is mean-square stable, that is, letting* $X_{t+1} = \mathcal{E}(\mathcal{M}(X_t))$ *with admissible* $X_0 \succeq 0$*, we have* $\sup_{t\geqslant 0} \mathrm{Tr}(X_t) < \infty$*.*

(ii) For each $\pi \in \Pi$*, the discounted Tsallis-regularized cost* J_π *in (5.2.12) is finite for all admissible initial second moments* X_0*.*

Under Assumption 5.2, the Tsallis-regularized Bellman operators inherit the usual contraction property from discounted dynamic programming. We summarize this in the following Lemma, which will be used in the proofs of Theorems 5.2 and 5.3.

Lemma 5.1. *Under Assumption 5.2, the operators* $\mathcal{T}_\pi$ *and* $\mathcal{F}_\pi$ *are* γ*-contractions on the spaces of bounded functions defined on the corresponding sets of admissible second-moment matrices. In particular, for any bounded functions* J_1, J_2 *and* $\mathcal{Q}_1, \mathcal{Q}_2$*,*

$$\|\mathcal{T}_\pi J_1 - \mathcal{T}_\pi J_2\|_\infty \leqslant \gamma \|J_1 - J_2\|_\infty, \quad \|\mathcal{F}_\pi \mathcal{Q}_1 - \mathcal{F}_\pi \mathcal{Q}_2\|_\infty \leqslant \gamma \|\mathcal{Q}_1 - \mathcal{Q}_2\|_\infty.$$

Proof. We first show the contraction property for $\mathcal{T}_\pi$. Recall from (5.2.13) that for bounded function J_1, J_2 and any admissible second-moment matrix X, we have

$$(\mathcal{T}_\pi J_1 - \mathcal{T}_\pi J_2)(X) = \gamma \left(J_1(\mathcal{E}(\mathcal{M}(X))) - J_2(\mathcal{E}(\mathcal{M}(X)))\right),$$

where $\mathcal{M}(\cdot)$ and $\mathcal{E}(\cdot)$ are the linear maps defined in (5.2.10)-(5.2.11). Taking the supremum over all admissible X gives

$$\|\mathcal{T}_\pi J_1 - \mathcal{T}_\pi J_2\|_\infty = \gamma \sup_X |J_1(\mathcal{E}(\mathcal{M}(X))) - J_2(\mathcal{E}(\mathcal{M}(X)))|.$$

For each admissible X, the point $Y = \mathcal{E}(\mathcal{M}(X))$ lies in the domain of J_1 and J_2. Let $\mathcal{S} = \{\mathcal{E}(\mathcal{M}(X)) : X \text{ admissible}\}$ to denote the image of the admissible set under the map $\mathcal{E} \circ \mathcal{M}$. Then,

$$\sup_X \left|J_1(\mathcal{E}(\mathcal{M}(X))) - J_2(\mathcal{E}(\mathcal{M}(X)))\right| = \sup_{Y \in \mathcal{S}} \left|J_1(Y) - J_2(Y)\right|.$$

Since $\mathcal{S}$ is a subset of the full domain of J_1 and J_2, we obtain

$$\sup_{Y \in \mathcal{S}} \left|J_1(Y) - J_2(Y)\right| \leqslant \sup_Y \left|J_1(Y) - J_2(Y)\right| = \left\|J_1 - J_2\right\|_\infty.$$

Combining the last two displays yields $\|\mathcal{T}_\pi J_1 - \mathcal{T}_\pi J_2\|_\infty \leqslant \gamma \|J_1 - J_2\|_\infty$. Thus $\mathcal{T}_\pi$ is a γ-contraction on the space of bounded functions. An analogous computation using the definition of $\mathcal{F}_\pi$ in (5.2.15) yields

$$\left(\mathcal{F}_\pi \mathcal{Q}_1 - \mathcal{F}_\pi \mathcal{Q}_2\right)(Z) = \gamma \left(\mathcal{Q}_1(\mathcal{M}(\mathcal{E}(Z))) - \mathcal{Q}_2(\mathcal{M}(\mathcal{E}(Z)))\right),$$

so $\|\mathcal{F}_\pi \mathcal{Q}_1 - \mathcal{F}_\pi \mathcal{Q}_2\|_\infty \leqslant \gamma \|\mathcal{Q}_1 - \mathcal{Q}_2\|_\infty$. The Tsallis term "$const'$." cancels in both differences, so regularization does not change the contraction factor. □

5.2.2 Tsallis Policy Evaluation

Theorem 5.2 (Tsallis Policy Evaluation)**.** *Under Assumption 5.2, for fixed $\pi \in \Pi$ and $0 < q < 1$, consider the operator $\mathcal{F}_\pi$. Starting from any bounded initial guess $\mathcal{Q}^{(0)}$, define the iterates $\mathcal{Q}^{(k+1)} = \mathcal{F}_\pi \mathcal{Q}^{(k)}$. Then $\mathcal{Q}^{(k)}$ converges to the unique fixed point $\mathcal{Q}_\pi$, which satisfies the policy evaluation equation (5.2.15)–(5.2.16).*

Proof of Theorem 5.2. We can show that this choice of $\mathcal{Q}_\pi$ is a fixed-point of $\mathcal{F}_\pi$

:

$$\begin{aligned}
\left(\mathcal{F}_\pi \mathcal{Q}_\pi\right)(Z) &= \operatorname{Tr}\left[ZH\right] + const'. + \gamma\left[\operatorname{Tr}\left(\mathcal{M}\left(\mathcal{E}(Z)\right)H\right)\right. \\
&\quad \left. + const'. + \gamma J_\pi(\mathcal{E}\left(\mathcal{M}\left(\mathcal{E}(Z)\right)\right))\right] \\
&= \operatorname{Tr}\left[ZH\right] + const'. + \gamma\left(\mathcal{T}_\pi J_\pi\right)\left(\mathcal{E}(Z)\right) \\
&= \operatorname{Tr}\left[ZH\right] + const'. + \gamma J_\pi\left(\mathcal{E}(Z)\right) \\
&= \mathcal{Q}_\pi(Z),
\end{aligned}$$

where we utilize the definition of $\mathcal{T}_\pi$ for the second equality and apply $J_\pi = \mathcal{T}_\pi J_\pi$ for the third. By Lemma 5.1, $\mathcal{F}_\pi$ is a γ-contraction on the space of bounded functions, so this fixed point is unique, and the iterates $\mathcal{Q}^{(k+1)} = \mathcal{F}_\pi Q^{(k)}$ converge to $\mathcal{Q}_\pi$ for any bounded initial $\mathcal{Q}^{(0)}$. □

The value function evaluated from Tsallis policy evaluation can be employed to update the policy distribution. In the policy improvement step, given such a $\mathcal{Q}_K$, we implement policy iteration as:

$$K^+ = \operatorname{argmin}_{K'} \mathcal{F}_{K'} \mathcal{Q}_K. \tag{5.2.17}$$

According to Theorem 5.1, $J_\pi\left(\mathcal{E}(Z)\right)$ in (5.2.16) can be written as

$$\begin{aligned}
& J_\pi\left(\mathcal{E}(Z)\right) \\
&= \mathbb{E}_\pi\left[x_+^\top P x_+\right] + c \\
&= \mathbb{E}_\pi\Big[\left((A+\Delta A)x + (B+\Delta B)u + (Cx+Du)w\right)^\top P \\
&\quad \left((A+\Delta A)x + (B+\Delta B)u + (Cx+Du)w\right)\Big] + c \\
&= \mathbb{E}_\pi\left[x^\top(A^\top P A + \Sigma_A P + W C^\top P C)x + u^\top\left(B^\top P B + \Sigma_B P\right.\right. \\
&\quad \left.\left. + W D^\top P D\right)u + 2u^\top\left(B^\top P A + W D^\top P C\right)x\right] + c \\
&= \operatorname{Tr}\left[\Theta' Z\right] + c,
\end{aligned} \tag{5.2.18}$$

where

$$\Theta' = \begin{bmatrix} A^\top PA + \Sigma_A P + WC^\top PC & \gamma A^\top PB + WC^\top PD \\ B^T PA + WD^\top PC & B^\top PB + \Sigma_B P + WD^\top PD \end{bmatrix}.$$

When taking (5.2.18) into (5.2.16), we can use a linear parametrization (Coppens and Patrinos (2022); Wang et al. (2018)) regarding Θ to express $\mathcal{Q}_\pi$ as

$$\mathcal{Q}_\pi(x) = \mathbb{E}_\pi \left\{ \begin{bmatrix} x^\top & u^\top \end{bmatrix} \Theta \begin{bmatrix} x^\top & u^\top \end{bmatrix}^\top - \tau \mathcal{H}_q(\pi(\cdot|x)) \right\} + \gamma c,$$

where the Θ matrix is defined as follows

$$\begin{aligned} \Theta &= \begin{bmatrix} Q + \gamma A^\top PA + \gamma \Sigma_A P + \gamma WC^\top PC & \gamma A^\top PB + \gamma WC^\top PD \\ \gamma B^T PA + \gamma WD^\top PC & R + \gamma B^\top PB + \gamma \Sigma_B P + \gamma WD^\top PD \end{bmatrix} \\ &= \begin{bmatrix} \Theta_{xx} & \Theta_{xu} \\ \Theta_{ux} & \Theta_{uu} \end{bmatrix}. \end{aligned} \tag{5.2.19}$$

Thus, we can use a linear parametrization (Coppens and Patrinos (2022); Wang et al. (2018)) regarding Θ to express $(\mathcal{F}_\pi \mathcal{Q})$ and $\mathcal{Q}_\pi$ as

$$\mathcal{Q}_\pi(Z) = \mathrm{Tr}[ZH] + const'. + \gamma \left(\mathrm{Tr}\left[\Theta' Z\right] + c \right) = \mathrm{Tr}\left[\Theta Z\right] + const. \tag{5.2.20}$$

where

$$const. = const'. + \gamma c,$$

and $\Theta = H + \gamma \Theta'$ is defined by

$$\Theta = \begin{bmatrix} \Theta_{xx} & \Theta_{xu} \\ \Theta_{ux} & \Theta_{uu} \end{bmatrix}, \tag{5.2.21}$$

where

$$\begin{aligned} \Theta_{xx} &= Q + \gamma A^\top PA + \gamma \Sigma_A P + \gamma WC^\top PC, \\ \Theta_{xu} &= \gamma A^\top PB + \gamma WC^\top PD, \\ \Theta_{ux} &= \gamma B^T PA + \gamma WD^\top PC, \\ \Theta_{uu} &= R + \gamma B^\top PB + \gamma \Sigma_B P + \gamma WD^\top PD. \end{aligned}$$

Then we have

$$K^+ = (\Theta_\pi)_{uu}^{-1} (\Theta_\pi)_{ux} . \tag{5.2.22}$$

The comparison with K^* from the previous section further supports that the optimal gain can be attained by choosing the appropriate value for Θ_π.

Remark 5.3. *For any fixed Z, $\mathcal{Q}_\pi(Z)$ is a quadratic function of Z with Hessian Θ_{uu}, which is positive definite since $R \succ 0$ and the remaining terms are positive semidefinite; hence, the minimization in (5.2.17) is strongly convex and admits a unique minimizer of (5.2.22). For $0 < q < 1$, Tsallis entropy is concave in the policy distribution (see, e.g., Lee et al. (2019)); thus, $-\tau\mathcal{H}_q$ is convex. In the policy improvement step (5.2.17), all quantities obtained from the policy evaluation step are held fixed, so the objective minimized over K' remains quadratic. It implies that the Tsallis term does not affect the convexity of (5.2.17). See also (Furuichi et al. (2004)) for a convex analytic view of Tsallis-regularized objectives.*

5.2.3 Tsallis Policy Improvement and Convergence

Theorem 5.3 (Tsallis Policy Improvement)**.** *Under Assumption 5.2 and for $0 < q < 1$, let K^+ be the updated policy from (5.2.22) using $\mathcal{Q}_K$. For all $(x, u) \in \mathcal{X} \times \mathcal{U}$, $\mathcal{Q}_{K^+}(x, u)$ is less than or equal to $\mathcal{Q}_K(x, u)$.*

Proof of Theorem 5.3. For any gain K satisfying Assumption 5.2 and by Lemma 5.1 and Theorem 5.2, the Q-function $\mathcal{Q}_K$ is the unique fixed point of $\mathcal{F}_K$ and is well defined. As the Remark 5.3 states, K^+ is updated by equation and the minimization in (5.2.17) is convex, then the following inequality holds:

$$\begin{aligned}
&\mathbb{E}_{u\sim\pi_{K^+}(\cdot|x)} \left[\mathcal{Q}_K(x,u) - \tau\mathcal{H}_q\left(\pi_{K^+}(u|x)\right)\right] \\
\leqslant\ &\mathbb{E}_{u\sim\pi_K(\cdot|x)} \left[\mathcal{Q}_K(x,u) - \tau\mathcal{H}_q\left(\pi_K(u|x)\right)\right] \\
=\ &J_{\pi_K}(x).
\end{aligned}$$

This is the defining optimality of π_{K^+} as a greedy minimizer of the regularized one-step objective based on $\mathcal{Q}_K$. The equality holds when $\pi_{K^+} = \pi_K$. As noted in Lee et al. (2019), this inequality induces a performance improvement,

$$
\begin{aligned}
& \mathcal{Q}_K(x,u) \\
=\ & \mathbb{E}_\pi \left[x_0^\top Q x_0 + u_0^\top R u_0 - \tau \mathcal{H}_q(\pi_K(u_0|x_0)) \right. \\
& \left. + \gamma J_{\pi_K}(x_1) | x_0 = x \right] \\
\geqslant\ & \mathbb{E}_\pi \left[x_0^\top Q x_0 + u_0^\top R u_0 - \tau \mathcal{H}_q(\pi_K(u_0|x_0)) | x_0 = x \right] \\
& + \gamma \mathbb{E}_\pi \left[x_1^\top Q x_1 + u_1^\top R u_1 - \tau \mathcal{H}_q(\pi_{K^+}(u_1|x_1)) \right. \\
& \left. + \gamma J_K(x_2) | x_0 = x \right] \\
\geqslant\ & \mathbb{E}_\pi \left[x_0^\top Q x_0 + u_0^\top R u_0 - \tau \mathcal{H}_q(\pi_K(u_0|x_0)) | x_0 = x \right] \\
& + \gamma \mathbb{E}_\pi \left[\sum_{k=1}^{t} \gamma^{k-1} c\left(x_k, u_k\right) \middle| x_0 = x \right] \\
& + \gamma^{t+1} \mathbb{E}_\pi \left[J_{\pi_k}(x_{t+1}) | x_0 = x \right] \\
& \vdots \\
\geqslant\ & \mathbb{E}_\pi \left[x_0^\top Q x_0 + u_0^\top R u_0 - \tau \mathcal{H}_q(\pi_K(u_0|x_0)) | x_0 = x \right] \\
& + \gamma \mathbb{E}_\pi \left[J_{\pi_{K^+}}(x_1) \right] \\
=\ & \mathcal{Q}_{K^+}(x,u),
\end{aligned}
$$

where $\gamma^{t+1} \mathbb{E}_\pi \left[J_{\pi_K}(x_{t+1}) | x_0 = x \right] \to 0$ when $t \to \infty$. □

For the optimality of Tsallis entropy policy iteration, when $0 < q < 1$, we define the Tsallis policy iteration in an alternative way by applying (5.2.16) and (5.2.17); then π_K converges to the optimal policy.

5.3 Data-Driven Implementation

5.3.1 Data-Related Matrices Construction with Least-Squares

It is clear from (5.2.22) that the optimal policy depends exclusively on the matrix Θ, thereby avoiding any constraints imposed by the system parameters. The matrix Θ clearly encompasses the information regarding system dynamics. We work with sample transitions to estimate $\mathcal{Q}_\pi$ and write a sampled Bellman equation whose discrepancy is a zero-mean error term to regress the one-step target on features using observed pairs.

Data are generated by selecting states x_i and computing the successors x_{i+}, while the rollout construction is described in Section 5.3.2. For each transition, we form the single sample outer product moments,

$$Z_i = \begin{bmatrix} x_i \\ u_i \end{bmatrix} \begin{bmatrix} x_i^\top & u_i^\top \end{bmatrix} \text{ and } Z_{i+} = \begin{bmatrix} x_{i+} \\ u_{i+} \end{bmatrix} \begin{bmatrix} x_{i+}^\top & u_{i+}^\top \end{bmatrix},$$

and define $X_{i+} = x_{i+}x_{i+}^\top$. Collecting N samples yields $\{(Z_i, X_{i+})\}_{i=1}^N$. Unlike in the preceding section, where Z and X represent second moments defined by expectations, the terms Z_i and X_{i+} here are treated as realized values without any expectation taken. Using (5.1.2), X_{i+} can be written as

$$\begin{aligned} X_{i+} = &\begin{bmatrix} A + \Delta A_i + Cw_i & B + \Delta B_i + Dw_i \end{bmatrix} Z_i \\ &\begin{bmatrix} A + \Delta A_i + Cw_i & B + \Delta B_i + Dw_i \end{bmatrix}^\top. \end{aligned} \tag{5.3.23}$$

We want to find $\mathcal{Q}_\pi$ such that $\mathcal{Q}_\pi(Z) = (\mathcal{F}_\pi \mathcal{Q}_\pi)(Z)$ for all Z. We sample $\mathcal{F}_\pi^i \mathcal{Q}_\pi$ by:

$$\left(\mathcal{F}_\pi^i \mathcal{Q}_\pi\right)(Z_i) = \operatorname{Tr}(Z_i H) + const'. + \gamma \mathcal{Q}_\pi(Z_{i+}), \tag{5.3.24}$$

which is used to construct a model equation:

$$\mathcal{Q}_\pi(Z_i) = \left(\mathcal{F}_\pi^i \mathcal{Q}_\pi\right)(Z_i) + \left((\mathcal{F}_\pi \mathcal{Q}_\pi)(Z_i) - \left(\mathcal{F}_\pi^i \mathcal{Q}_\pi\right)(Z_i)\right). \tag{5.3.25}$$

Using the parametrization in (5.2.15), (5.2.20) and (5.3.24) to (5.3.25), we obtain

$$\begin{aligned}
&\operatorname{Tr}[\Theta_\pi Z_i] + const. \\
&= \operatorname{Tr}[Z_i H] + const'. + \gamma \mathcal{Q}_\pi(Z_{i+}) + [\operatorname{Tr}[Z_i H] + const'. + \gamma \mathcal{Q}_\pi(\mathcal{M}(\mathcal{E}(Z_i))) \\
&\quad - (\operatorname{Tr}(Z_i H) + const'. + \gamma \mathcal{Q}_\pi(Z_{i+}))] \\
&= \operatorname{Tr}[Z_i H] + const'. + \gamma(\operatorname{Tr}[\Theta_\pi Z_{i+}] + const.) + \gamma \operatorname{Tr}[\Theta_\pi(\mathcal{M}(\mathcal{E}(Z_i)) - Z_{i+})],
\end{aligned} \tag{5.3.26}$$

which is linear in the parameter Θ_π with zero-mean error. It is important to observe that (5.3.26) incorporates temporal differences, as discussed in Bradtke and Barto (1996).

We reformulate the model utilizing the symmetric vectorization operator $\operatorname{vec_s} : \mathbb{R}^{m\times m} \to \mathbb{R}^{\frac{m(m+1)}{2}}$, which satisfies $\operatorname{Tr}[XY] = \operatorname{vec_s}(X)^\top \operatorname{vec_s}(Y)$. We rewrite (5.3.26) as:

$$b_i = (a_i + e_i)^\top \theta_\pi, \tag{5.3.27}$$

with

$$\begin{aligned}
\theta_\pi &= \operatorname{vec_s}(\Theta_\pi), \\
b_i &= \operatorname{Tr}(H Z_i) + const'. + (\gamma - 1) const., \\
a_i &= \operatorname{vec_s}(Z_i - \gamma Z_{i+}), \\
e_i &= \gamma \operatorname{vec_s}(Z_{i+} - \mathcal{M}(\mathcal{E}(Z_i))).
\end{aligned} \tag{5.3.28}$$

This step is referred to as the error-in-variables setting. The problem arises from the linear dependence of both a_i and e_i on $w_i w_i^\top$ and w_i, as expressed through Z_{i+}. Therefore, $\mathbb{E}\left[a_i e_i^\top\right] \neq 0$, resulting in an inconsistency in the classical least-squares estimate.

This is the traditional rationale for the use of instrumental variables (Young (2011)). We employ $g_i = \operatorname{vec_s}(Z_i) = \operatorname{vec_s}\left(\mathcal{M}(x_i x_i^\top)\right)$ as an instrumental variable, following the approach outlined in (Bradtke and Barto (1996)). This selection is

independent of the error e_i (i.e., $\mathbb{E}\left[g_i e_i^\top\right] = 0$) and exhibits correlation with $a_i + e_i$, which is a requisite characteristic of an instrumental variable. Conceptually, g_i represents the optimal estimate in the absence of supplementary information.

The estimate using the instrumental variables is

$$\hat{\theta}_\pi = \left(\sum_{i=1}^{N} g_i a_i^\top\right)^{-1} \left(\sum_{i=1}^{N} g_i b_i\right). \tag{5.3.29}$$

Alternatively, we can use the recursive algorithm

$$\begin{aligned} \hat{\theta}_{\pi,i} &= \hat{\theta}_{\pi,i-1} + L_i \left[b_i - a_i^\top \hat{\theta}_{\pi,i-1}\right], \\ L_i &= S_{\pi,i-1} g_i / \left(1 + a_i^\top S_{\pi,i-1} g_i\right), \\ S_{\pi,i} &= S_{\pi,i-1} - \left(L_i a_i^\top\right) S_{\pi,i-1}, \end{aligned} \tag{5.3.30}$$

for $i = 1, \cdots, N$ and where the initialization $\hat{\theta}_{\pi,0} \in \mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$ and

$S_{\pi,0} \in \mathbb{R}^{\frac{(m+n)(m+n+1)}{2} \times \frac{(m+n)(m+n+1)}{2}}$ represent an initial estimate for θ_π associated confidence in that estimate, respectively.

5.3.2 Dataset Persistent Excitation Condition

The data are generated iteratively along policy updates. In each iteration, we roll out M trajectories of length T (total $N = MT$ samples) that satisfy (5.3.23). For $j = 1, \cdots, M$ and $t = 0, \cdots, T-1$,

$$x_{t+1}^j = (A + \Delta A_t^j) x_t^j + (B + \Delta B_t^j) u_t^j + (C x_t^j + D u_t^j) w_t^j,$$

and we set $z_t^j = \begin{bmatrix} x_t^j \\ u_t^j \end{bmatrix}$ with input

$$u_t^j = -K x_t^j + \nu_t^j, \quad \nu_t^j \sim \mathcal{N}_q(0, \Sigma),$$

where ν_t^j is used only to perturb the applied control u_t^j so as to ensure sufficient excitation of the data. In the Bellman update, the Q-function at the next control

u_{t+1}^j is evaluated at the noise-free policy action $-Kx_{t+1}^j$. Here ν_t^j is assumed to be uniformly bounded over $\{\nu : \|\nu\| \leqslant r_\nu\}$. This step is the standard persistent excitation requirement ensuring invertibility of the normal matrix in LSTD (Bradtke and Barto (1996)).

Assumption 5.3. *(Persistent excitation and bounded moments) Across the N collected samples $\{(Z_i, X_{i+})\}_{i=1}^N$, we have*

(i) ν_t^j is independent of $\left\{w_t^j, \Delta A_t^j, \Delta B_t^j\right\}$ and the initial states, with $\mathbb{E}\left[\nu_t^j\right] = 0$ and $\operatorname{cov}\left(\nu_t^j\right) = \Sigma \succ 0$;

(ii) As g_i and a_i are defined in Section 5.3.1, the empirical cross-moment $\sum_{i=1}^N g_i a_i^\top$ has full column rank, with its smallest singular value bounded away from zero as N grows, so the instrumental-variable normal matrix is invertible;

(iii) Second moments along rollouts are uniformly bounded, and either resets or independent initializations are used so that empirical averages approximate expectations, that is, $\frac{1}{N}\sum_i g_i a_i^\top \to \mathbb{E}\left[g_i a_i^\top\right]$ and $\frac{1}{N}\sum_i g_i b_i \to \mathbb{E}\left[g_i b_i\right]$.

Remark 5.4. *The exploration covariance Σ only influences empirical covariances used to estimate Θ. It does not alter the population Bellman operators in Section 5.2. A bounded support choice for ν_t^j may be used in practice to avoid actuator saturation while preserving Assumption 5.3 (i)-(ii).*

The functional form of the dynamics is unnecessary since we only collect states and inputs, thereby enabling model-free implementation from either the simulator or experiments. Each trajectory is initialized at the beginning of a controller iteration. Initial states are obtained either by continuing from the terminal state of the previous iteration or by sampling $\left\{x_0^j\right\}_{j=1}^M$ uniformly from $\{x : \|x\| \leqslant r_x\}$. Snapshots X_t^i and Z_t^i are estimated by outer products. When the particular trajectory index is immaterial

for policy iteration and identification, we write the pairs $\left(Z_t^j, X_{t+1}^j\right)$ as (Z_i, X_{i+}), with i reindexing (t, j).

5.3.3 Policy Iteration Algorithm

We can now explain the complete approximate policy iteration algorithm. The term *generate pair* pertains to the rollouts outlined in the preceding section.

Algorithm 1 Approximate Policy Iteration

Input: Initial guess $\hat{\theta}_{\pi_0}$, number of samples N, and confidence parameter β_0.
Initialize: $S_{\pi_0} = \beta_0 I$.

1: **for** $t = 1$ to ∞ **do**
2: $K_t = \Theta_{uu}^{-1}\Theta_{ux}$, with $\Theta = \text{unvec}_{\text{s}}\left(\hat{\theta}_{\pi_{t-1}}\right)$.
3: Let $\hat{\theta}_{\pi_t,0} = \hat{\theta}_{\pi_{t-1}}$ and $S_{\pi_t,0} = S_{\pi_{t-1}}$.
4: **for** $i = 1$ to N **do**
5: Generate pair (Z_i, X_{i+}) and find Z_{i+}.
6: Update $\hat{\theta}_{\pi_t,i}$ and $S_{\pi_t,i}$ using Equation (5.3.30).
7: Set $\hat{\theta}_{\pi_t} = \hat{\theta}_{\pi_t,N}$ and $S_{\pi_t} = S_{\pi_t,N}$.

Our approach allows the scheme to continuously enhance its estimation of the Q-function, as evidenced by the numerical experiments.

Lemma 5.1 and Theorem 5.2–5.3 establish contraction and convergence for the population operators; Algorithm 1 is an approximate policy iteration whose convergence follows under consistent LSTD estimation and sufficient excitation, which we assume here.

5.4 Numerical Examples

In this section, we illustrate the proposed Tsallis-regularized LQ policy iteration on the MV portfolio selection problem. This example fits exactly into the multiplicative noise LQ framework developed in the sections before, and the wealth dynamics can be written in the form of (5.1.2). Following common practice in empirical finance using standard portfolio metrics, we: i) construct a reproducible ETF dataset; ii)

calibrate a stationary multiplicative noise LQ model from a training window; iii) learn a stabilizing policy via Algorithm 1; and iv) report backtesting performance on an out-of-sample test window that is not used for calibration.

5.4.1 Dataset and Backtesting Protocol

Assets and sampling. We use three liquid ETFs as risky assets: SPY (U.S. equities), IEF (U.S. Treasury bonds), and GLD (gold). Daily adjusted close prices are obtained from Yahoo Finance and converted to month-end prices; monthly gross returns are compounded to yearly gross returns $e_t = (e_t^1, e_t^2, e_t^3)^\top$. Partial calendar years are removed to avoid incomplete compounding. For the risk-free asset, we use the 3-month U.S. Treasury bill rate (FRED series TB3MS), convert it to monthly gross returns, and compound to obtain yearly gross risk-free returns r_t. We form yearly excess returns $P_t = e_t - r_t\mathbf{1}$.

Test split. We use years 2006–2016 as the training window and 2017–2024 as the out-of-sample test window.

Backtesting rule. Starting from $x_0 = 1$, we apply each learned policy over the test window and compute the realized wealth sequence using the true historical returns. For stochastic policies, results are averaged over $N_{\mathrm{MC}} = 300$ Monte Carlo runs with different random seeds.

Baselines and metrics. We compare no entropy ($\tau = 0$), Shannon entropy (Gaussian, corresponding to $q \to 1$), and Tsallis entropy ($q < 1$). We report CAGR, annualized volatility, Sharpe ratio, maximum drawdown (MaxDD), and turnover.

Metric definitions. CAGR denotes the compound annual growth rate $(x_T/x_0)^{1/T} - 1$ on the test window of T years. Vol is the standard deviation of annual portfolio returns, and the Sharpe ratio is the mean excess return over Vol. MaxDD is the maximum peak-to-trough drawdown of the wealth curve, and turnover is the average ℓ_1 change in risky weights between consecutive rebalances.

5.4.2 Market Model Calibration and LQ Mapping

Consider the model of the MV problem in Chapter 2 and the infinite time case in section 4.4.1 but with Tsallis entropy rather than Shannon entropy:

$$\min_{u_t} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^2 + u_t^\top R u_t + \frac{\tau}{2-q} \left(\log_q \pi(u_t|x_t) - 1 \right) \right) \middle| x_0 = x \right]. \tag{5.4.31}$$

The discount factor γ reflects the time value of money or utility and keeps the total cost finite. Then, we shall transform (2.3.17) into a linear controlled system of form (2.1.1) by which the general theory in the above sections will work. Precisely, we define

$$\begin{cases} w_{i,t} = e_t^i - \mathbb{E}\left(e_t^i\right), \\ D_{i,t} = (0, \cdots, 0, 1, 0, \cdots, 0), \\ i = 1, \cdots, n, \ t = 0, 1, \cdots, \infty, \end{cases}$$

which leads to

$$x_{t+1} = (r_t x_t + \mathbb{E}(P_t) u_t) + \sum_{i=1}^{n} D_{i,t} u_t w_{i,t}. \tag{5.4.32}$$

It is not hard to see that problem (MV) is just a special case of the problem in section 5.1 by letting $x_t \in \mathbb{R}$, $u_t \in \mathbb{R}^n$, $R \in \mathbb{R}^{n\times n}$, $Q = 1$, $A = r_t$, $B = \mathbb{E}(P_t)$. It is the special case of Section 2.1 for $\Delta A_t = 0$, $\Delta B_t = \mathbf{0}_{1\times n}$ and $C = 0$ with multiplicative noise entering only through the control channel via $D_t u_t w_t$. Thus, the example still lies within the multiplicative noise LQ framework.

Calibration: We estimate parameters from the training data in Section 5.4.1. The quadratic penalty $R > 0$ is selected to scale risky exposures and ensure well-posed Q-function updates. Unless otherwise stated, we use $(q, \tau, \gamma) = (0.5, 0.7, 0.85)$.

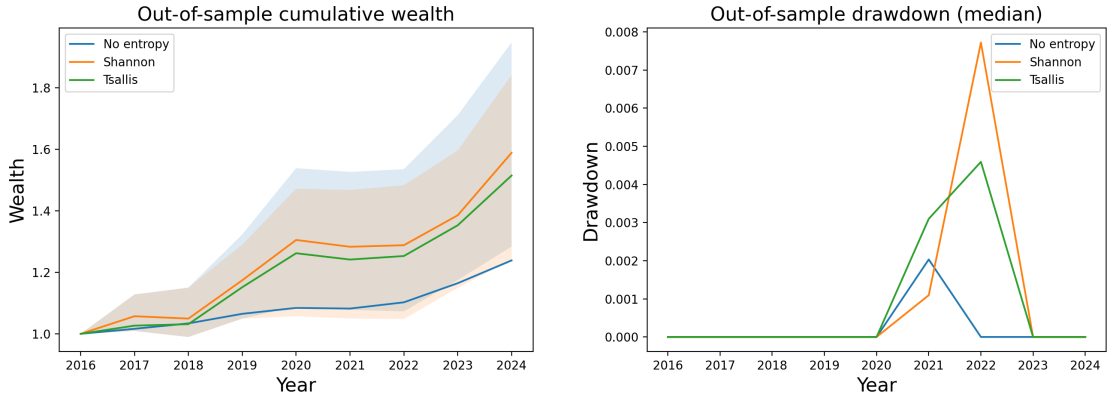
5.4.3 Backtesting Results

Table 5.1 summarizes out-of-sample performance for no entropy, Shannon, and Tsallis. All methods share the same mean feedback gain K; Shannon or Tsallis differ

Table 5.1: Out-of-sample backtesting performance on the ETF dataset (test window). For stochastic policies, results are mean±std over N_{MC} runs.

Metric	No entropy	Shannon	Tsallis
CAGR	0.0271 ± 0.0000	0.0589 ± 0.0215	0.0527 ± 0.0193
Vol	0.0221 ± 0.0000	0.0841 ± 0.0235	0.0737 ± 0.0235
Sharpe	0.2346 ± 0.0000	0.4464 ± 0.2189	0.4070 ± 0.2088
MaxDD	0.0020 ± 0.0000	0.0303 ± 0.0344	0.0202 ± 0.0177
Turnover	0.0000 ± 0.0000	0.6173 ± 0.2228	0.4715 ± 0.1745

only in the exploration distribution. Shannon attains the highest mean CAGR and Sharpe ratio but also has the highest volatility, MaxDD, and turnover. Tsallis delivers comparable growth with lower risk and trading intensity, consistent with the risk-limited exploration induced by the bounded support q-Gaussian policy class. The deterministic baseline is the most conservative, with the lowest growth and near-zero turnover.



(a) Cumulative wealth on the test window (b) Median drawdown on the test window

Figure 5.1: Out-of-sample backtesting results on the ETF dataset

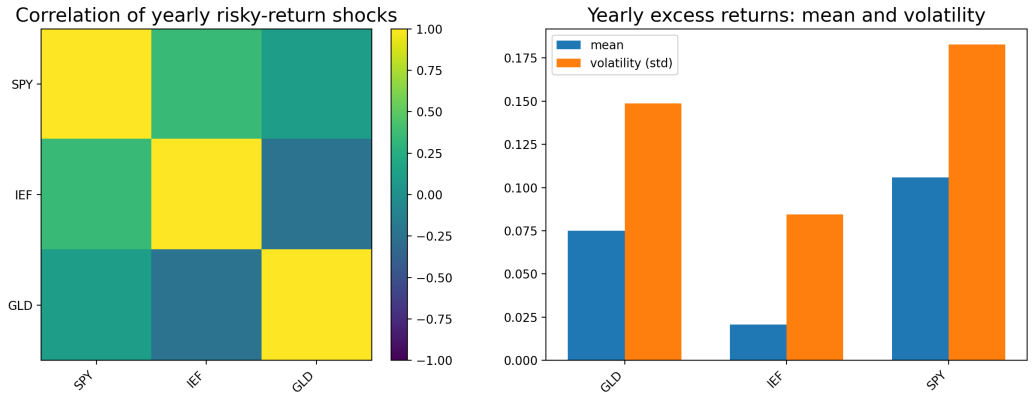
Fig. 5.1(a) plots cumulative wealth on the test window. For stochastic policies, the solid line is the median over N_{MC} runs, and the shaded band is the 10–90% range. Shannon achieves the highest median terminal wealth, while Tsallis is close with a narrower band, consistent with its lower volatility in Table 5.1. Fig. 5.1(b) reports the median drawdown trajectory. Shannon exhibits deeper drawdowns during

adverse periods, whereas Tsallis reduces drawdown severity, matching the MaxDD ranking in Table 5.1.

5.4.4 Learning Behavior and Sensitivity

To isolate the algorithmic effect of Tsallis regularization on sample-based policy iteration, we also evaluate learning performance under rollouts generated by Algorithm 1 from the stationary calibrated model.

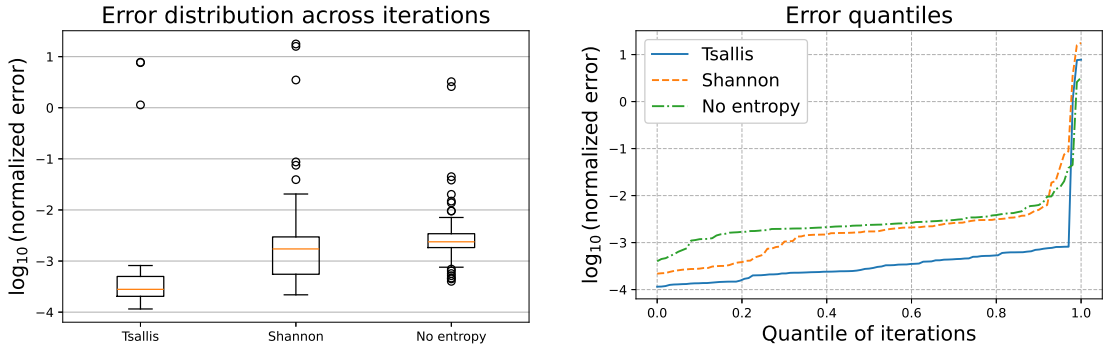
We initially examine the influence of the number of rollouts on the effectiveness of our policy iteration method. According to the notation defined in Section 5.3.2, the additive noise radius is set to $r_\nu = 0.8$. Each policy iteration update employs $N = 1200$ samples initiated from states with a distribution parameter of $r_x = 1$. A constant, stabilizing control gain K_0 is utilized to generate the samples. We set the initial value of policy iteration as $\hat{\theta}_{\pi_0} \in \mathbb{R}^{(1+n)(1+n+1)/2}$, with all entries equal to 0.8 and $\beta_0 = 20$, unless specified differently. The initial Q-function's optimal controller is represented by K_0, which ensures stability. Here we use the normalized training error $(J_K - J_{K*})/J_{K*}$ as a learning diagnostic, where K^* is the oracle stabilizing gain obtained from the Riccati solution of the calibrated model.



(a) Correlation of yearly return shocks (b) Mean and volatility of yearly excess returns

Figure 5.2: Empirical characteristics of data

Fig. 5.2(a) shows the correlation of yearly risky-return shocks, indicating that asset return uncertainties are correlated rather than independent. Fig. 5.2(b) reports the empirical mean and volatility of yearly excess returns. The sample means determine the drift parameter B, while the heterogeneity in volatilities motivates scaling the quadratic penalty matrix R. Overall, the figure shows that the parameters are directly estimated from observed market data.



(a) Boxplot of normalized errors (b) Quantile plot of normalized errors

Figure 5.3: Empirical distribution of normalized errors

Figure 5.3 (a) and (b) show the empirical distribution of normalized errors over training iterations for the three regularizations. The boxplot 5.3(a) shows that the Tsallis-regularized model has the lowest median error and relatively compact interquartile range together with the fewest and smallest outliers, thus presenting more stable and less heavy-tailed training behavior. The quantile curve in 5.3(b) indicates that the Tsallis curve lies below the others for almost all quantiles, implying that Tsallis has smaller errors for most iterations and better convergence performance.

Figure 5.4 compares the convergence behavior of the three policy iteration variants. The no entropy case descends fastest initially, yet exhibits relatively large fluctuations in the later stages. Shannon entropy regularization dampens some of this variability but stabilizes at a comparatively higher error level. In contrast, the Tsallis-regularized method displays the most consistent decay and achieves the low-

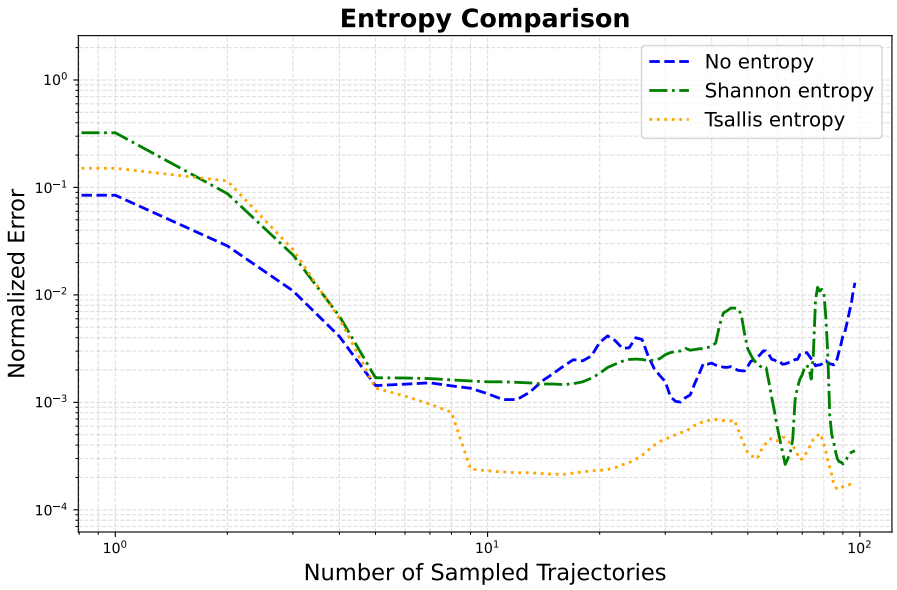
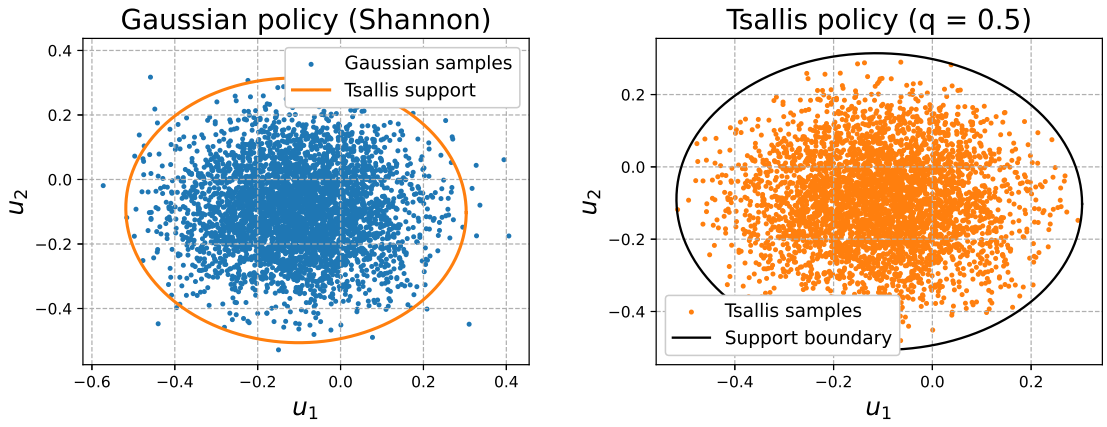


Figure 5.4: Performance of policy iteration

est error plateau, near 10^{-4}, indicating a more favorable trade-off between accuracy and stability.



(a) Shannon entropy (Gaussian policy). (b) Tsallis entropy ($q < 1$) with bounded support.

Figure 5.5: Gaussian vs. Tsallis policy ellipse

Figure 5.5 illustrates the sparsity effect of Tsallis entropy. For a fixed state x, we plot control samples in the (u_1, u_2)-plane obtained from the Gaussian policy under Shannon entropy (left) and from the Tsallis-regularized policy with $q < 1$ (right) us-

ing the learned parameters (K^*, Σ^*). The ellipse indicates the Tsallis support region $(u + K^*x)^\top \Sigma^{*-1} (u + K^*x) \preccurlyeq R_q^2$; Tsallis samples lie entirely inside this ellipsoid, while Gaussian samples extend beyond it.

This sparsity-like truncation of extreme actions reduces the variance of sampled costs in the multiplicative noise system and is consistent with the empirical behavior in Figures 5.3 and 5.4, where the Tsallis entropy exhibits smoother error decay, fewer large error spikes, and the lowest asymptotic error level.

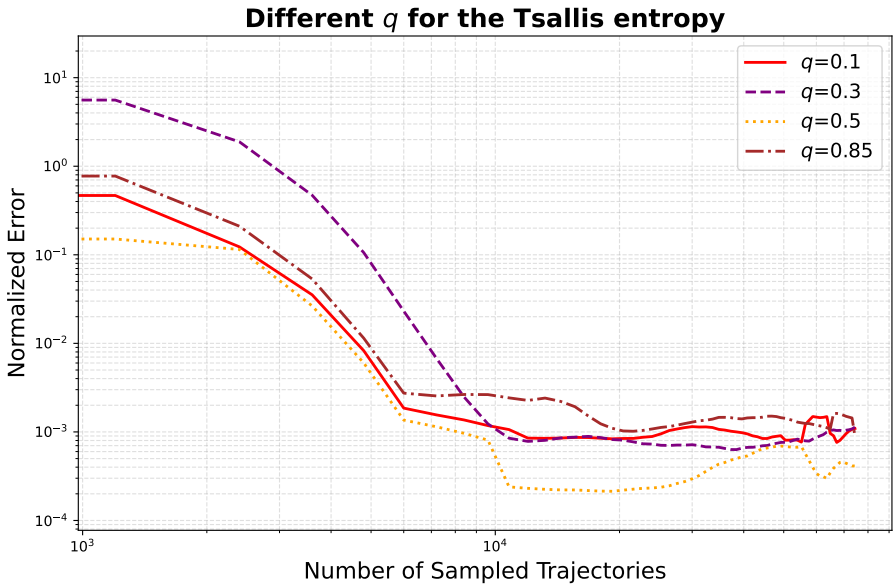


Figure 5.6: Different q values

Figure 5.6 explores the impact of different Tsallis entropy parameters $q \in [0.1, 0.85]$. No strict monotonic relationship emerges: the intermediate value $q = 0.5$ achieves the fastest decay and the lowest error, while both smaller $q = 0.1$ and larger $q = 0.85$ converge to slightly higher error levels. The choice $q = 0.3$ results in slower initial progress and a less favorable asymptotic error.

Figure 5.7 presents convergence under Tsallis entropy with varying discount factors. Smaller values of γ (e.g., 0.75) yield a slightly faster initial decrease in the normalized error, reflecting the shorter effective horizon of the underlying control

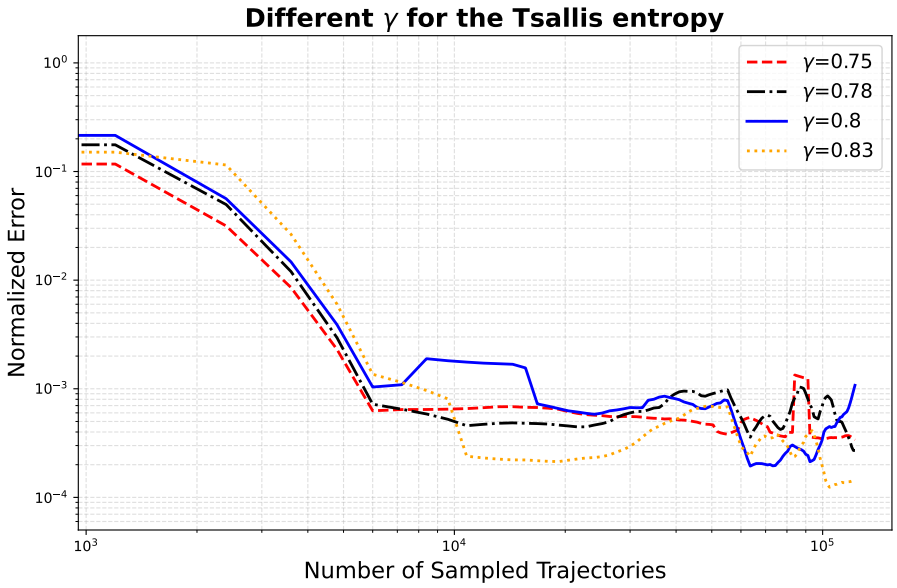


Figure 5.7: Different γ values

problem. However, all choices of γ eventually reach a similar error range, with the baseline $\gamma = 0.83$ attaining the lowest plateau and comparable or lower variance than the more myopic settings.

Chapter 6

Conclusion

This thesis investigates RL and data-driven methods for discrete time stochastic LQ control, with an emphasis on features that frequently arise in practice: multiplicative noise, implementability constraints, decentralized information structures, and entropy-regularized objectives. Across the three parts, the thesis develops a theoretical understanding of policy optimization methods, including policy gradient and policy iteration, and proposes computational approaches that remain applicable when model knowledge is incomplete. Dynamic MV portfolio optimization, benchmark tracking, and optimal liquidation are used as representative case studies to connect the theoretical developments with financially motivated decision-making problems.

Chapter 3 studies policy gradient methods for finite horizon, discrete time LQ control under multiplicative noise, with particular attention to how constraints and uncertainty in the noise statistics affect convergence and robustness. The chapter is organized around three progressively more practical settings: (i) a baseline LQ formulation that clarifies the policy gradient mechanism and provides a reference point; (ii) a constrained scalar state LQ model with multiplicative noise, which highlights how feasibility requirements interact with gradient-based policy updates; and (iii) an uncertain noise setting in which the noise has nonzero mean and unknown covariance, leading to a DRO formulation and a corresponding policy gradient procedure

that explicitly accounts for these uncertain moments. The chapter illustrates the implications of these results using MV and benchmark tracking examples, thereby linking the convergence analysis to two canonical financial control tasks.

Despite these contributions, several limitations suggest natural extensions. First, the constrained analysis is developed in a scalar state setting and under a restricted constraint structure; extending the results to higher-dimensional systems and more general convex constraints would substantially broaden applicability. Second, the DRO analysis is based on moment uncertainty, that is, mean and covariance. Future work can explore richer uncertainty sets and study how their geometry affects both computational tractability and convergence behavior. Finally, a systematic finite sample analysis, for example, sample complexity and stability under noisy gradients, would strengthen the practical guidance for policy gradient implementations in stochastic control.

Chapter 4 analyzes the multi-agent LQ problem characterized by multiplicative noise, by emphasizing the IPO technique. This work meticulously examines IPO and establishes a super-linear local convergence rate near the optimal policy. Practical applications of MV optimization and optimal liquidation problems are illustrated, underscoring their significance in financial decision-making scenarios. The approach presumes a static parameter framework and fails to consider situations characterized by dynamically evolving system dynamics or parameter uncertainty. Moreover, the framework can be extended from decentralized control to distributed control. Future studies may investigate the modification of IPO methodologies to accommodate time-varying or uncertain parameters on the distributed multi-agent LQ problem, hence enhancing the robustness and scalability of the methodology in dynamic environments. Furthermore, creating hybrid algorithms that combine IPO with RL methods can resolve practical problems in more complicated contexts.

Chapter 5 discusses the implementation of Tsallis entropy inside LQ models

and presents a model-free, data-driven policy iteration methodology. This chapter presents an innovative solution framework for RL-based policy by integrating Q-functions computed by a least-squares method with instrumental variables. The incorporation of Tsallis entropy provides a measured method to augment exploration and sparsity in control laws, illustrating its capacity to boost robustness and adaptability in stochastic control systems. Nonetheless, the analysis does not provide theoretical assurances regarding the convergence of the least-squares approach within this framework. This gap raises inquiries regarding the algorithm's efficacy under diverse data situations and parameter configurations. Future research may focus on rigorously establishing and examining the convergence characteristics of the least-squares approach in this case to guarantee reliability. Moreover, investigating alternative or adaptable estimating methods may enhance the algorithm's efficacy. Expanding the approach to assess its applicability using real-world data will augment its practical significance.

Bibliography

Abeille, M., Lazaric, A., Brokmann, X. et al. (2016) Lqg for portfolio optimization. *arXiv preprint arXiv:1611.00997*.

Abouheaf, M. I., Lewis, F. L., Vamvoudakis, K. G., Haesaert, S. and Babuska, R. (2014) Multi-agent discrete-time graphical games and reinforcement learning solutions. *Automatica*, **50**, 3038–3053.

Ackermann, J., Kruse, T. and Urusov, M. (2021) Càdlàg semimartingale strategies for optimal trade execution in stochastic order book models. *Finance and Stochastics*, **25**, 757–810.

Agarwal, N., Hazan, E. and Singh, K. (2019) Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, **32**.

Alfonsi, A., Fruth, A. and Schied, A. (2008) Constrained portfolio liquidation in a limit order book model. *Banach Center Publ*, **83**, 9–25.

Almgren, R. and Chriss, N. (2001) Optimal execution of portfolio transactions. *Journal of Risk*, **3**, 5–40.

Anderson, B. D. O. and Moore, J. B. (2007) *Optimal control: linear quadratic methods*. Courier Corporation.

Auer, P. and Ortner, R. (2006) Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, **19**.

Balakrishnan, V. and Vandenberghe, L. (2003) Semidefinite programming duality and linear time-invariant systems. *IEEE Transactions on Automatic Control*, **48**, 30–41.

Bao, H. and Sakaue, S. (2022) Sparse regularized optimal transport with deformed q-entropy. *Entropy*, **24**, 1634.

Basei, M., Guo, X., Hu, A. and Zhang, Y. (2022) Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, **23**, 1–34.

Basin, M., Perez, J. and Skliar, M. (2006) Optimal filtering for polynomial system states with polynomial multiplicative noise. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, **16**, 303–314.

Baz, J. and Chacko, G. (2004) *Financial derivatives: pricing, applications, and mathematics*. Cambridge University Press.

Becherer, D., Bilarev, T. and Frentrup, P. (2018) Optimal liquidation under stochastic liquidity. *Finance and Stochastics*, **22**, 39–68.

Beiglböck, M. and Pratelli, A. (2012) Duality for rectified cost functions. *Calculus of Variations and Partial Differential Equations*, **45**, 27–41.

Bemporad, A., Morari, M., Dua, V. and Pistikopoulos, E. N. (2002) The explicit linear quadratic regulator for constrained systems. *Automatica*, **38**, 3–20.

Benidis, K., Feng, Y., Palomar, D. P. et al. (2018) Optimization methods for financial index tracking: From theory to practice. *Foundations and Trends® in Optimization*, **3**, 171–279.

Benigno, P. and Woodford, M. (2012) Linear-quadratic approximation of optimal policy problems. *Journal of Economic Theory*, **147**, 1–42.

Bertsekas, D. (2022) *Abstract dynamic programming*. Athena Scientific.

Bhandari, J. and Russo, D. (2024) Global optimality guarantees for policy gradient methods. *Operations Research*.

Bjork, T. and Murgoci, A. (2010) A general theory of markovian time inconsistent stochastic control problems. *Available at SSRN 1694759*.

Borges, E. P. (2004) A possible deformed algebra and calculus inspired in nonextensive thermostatistics. *Physica A: Statistical Mechanics and its Applications*, **340**, 95–101.

Bradtke, S. J. and Barto, A. G. (1996) Linear least-squares algorithms for temporal difference learning. *Machine learning*, **22**, 33–57.

Bu, J., Mesbahi, A. and Mesbahi, M. (2020) Policy gradient-based algorithms for continuous-time linear quadratic control. *arXiv preprint arXiv:2006.09178*.

Bu, J., Ratliff, L. J. and Mesbahi, M. (2019) Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*.

Choy, J., Lee, K. and Oh, S. (2020) Sparse actor-critic: Sparse tsallis entropy regularized reinforcement learning in a continuous action space. 68–73.

Coppens, P. and Patrinos, P. (2022) Policy iteration using q-functions: Linear dynamics with multiplicative noise. *arXiv preprint arXiv:2212.01192*.

Costa, O. L. V., Fragoso, M. D. and Marques, R. P. (2005) *Discrete-time Markov jump linear systems*. Springer Science & Business Media.

Cui, X., Gao, J., Li, X. and Li, D. (2014a) Optimal multi-period mean–variance policy under no-shorting constraint. *European Journal of Operational Research*, **234**, 459–468.

Cui, X., Li, X. and Li, D. (2014b) Unified framework of mean-field formulations for optimal multi-period mean-variance portfolio selection. *IEEE Transactions on Automatic Control*, **59**, 1833–1844.

Czichowsky, C. and Schweizer, M. (2013) Cone-constrained continuous-time markowitz problems. *The Annals of Applied Probability*.

Donnelly, R. and Jaimungal, S. (2024) Exploratory control with tsallis entropy for latent factor models. *SIAM Journal on Financial Mathematics*, **15**, 26–53.

Elliott, R., Li, X. and Ni, Y.-H. (2013) Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, **49**, 3222–3233.

Fazel, M., Ge, R., Kakade, S. and Mesbahi, M. (2018) Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, 1467–1476. PMLR.

Fournier, N. and Guillin, A. (2015) On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, **162**, 707–738.

Fu, C., Lari-Lavassani, A. and Li, X. (2010) Dynamic mean–variance portfolio selection with borrowing constraint. *European Journal of Operational Research*, **200**, 312–319.

Furuichi, S., Yanagi, K. and Kuriyama, K. (2004) Fundamental properties of tsallis relative entropy. *Journal of Mathematical Physics*, **45**, 4868–4877.

Gao, J., Li, D., Cui, X. and Wang, S. (2015) Time cardinality constrained mean–variance dynamic portfolio selection and market timing: A stochastic control approach. *Automatica*, **54**, 91–99.

Gârleanu, N. and Pedersen, L. H. (2013) Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, **68**, 2309–2340.

Gell-Mann, M. and Tsallis, C. (2004) *Nonextensive entropy: interdisciplinary applications*. Oxford University Press, USA.

Gershon, E. and Shaked, U. (2006) Static H_2 and H_∞ output-feedback of discrete-time lti systems with state multiplicative noise. *Systems & Control Letters*, **55**, 232–239.

Gravell, B., Esfahani, P. M. and Summers, T. (2020) Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, **66**, 5283–5298.

Gu, C. (2011) Qlmor: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **30**, 1307–1320.

Guo, X., Lai, T. L., Shek, H. and Wong, S. P.-S. (2017) *Quantitative trading: algorithms, analytics, data, models, optimization*. Chapman and Hall/CRC.

Guo, X., Li, X. and Xu, R. (2023) Fast policy learning for linear-quadratic control with entropy regularization. *arXiv preprint arXiv:2311.14168*.

Haarnoja, T., Tang, H., Abbeel, P. and Levine, S. (2017) Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hambly, B., Xu, R. and Yang, H. (2021) Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, **59**, 3359–3391.

Hashizume, Y., Oishi, K. and Kashima, K. (2024) Tsallis entropy regularization for linearly solvable mdp and linear quadratic regulator. *arXiv preprint arXiv:2403.01805*.

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

Jin, Z., Schmitt, J. M. and Wen, Z. (2020) On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*.

Kalman, R. E. et al. (1960) Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, **5**, 102–119.

Kim, K. and Yang, I. (2023) Distributional robustness in minimax linear quadratic control with wasserstein distance. *SIAM Journal on Control and Optimization*, **61**, 458–483.

Lavagno, A. and Swamy, P. N. (2002) Non-extensive entropy in q-deformed quantum groups. *Chaos, Solitons & Fractals*, **13**, 437–444.

Lee, D. and Hu, J. (2018) Primal-dual q-learning framework for lqr design. *IEEE Transactions on Automatic Control*, **64**, 3756–3763.

Lee, J. Y., Park, J. B. and Choi, Y. H. (2012) Integral q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica*, **48**, 2850–2859.

Lee, K., Choi, S. and Oh, S. (2018) Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, **3**, 1466–1473.

Lee, K., Kim, S., Lim, S., Choi, S. and Oh, S. (2019) Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv preprint arXiv:1902.00137*.

Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020) Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, D. and Ng, W.-L. (2000) Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical finance*, **10**, 387–406.

Li, W. and Todorov, E. (2004) Iterative linear quadratic regulator design for nonlinear biological movement systems. In *First International Conference on Informatics in Control, Automation and Robotics*, vol. 2, 222–229. SciTePress.

Li, X., Zhou, X. Y. and Lim, A. E. (2002) Dynamic mean-variance portfolio selection with no-shorting constraints. *SIAM Journal on Control and Optimization*, **40**, 1540–1555.

Littlestone, N. and Warmuth, M. K. (1994) The weighted majority algorithm. *Information and computation*, **108**, 212–261.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P. and Wainwright, M. (2019) Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2916–2925. PMLR.

Markowitz, H. (1952) Portfolio selection. *Journal of Finance*, **7**, 77–91.

Mayne, D. Q. (2014) Model predictive control: Recent developments and future promise. *Automatica*, **50**, 2967–2986.

McKean, Henry P., J. (1966) A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, **56**, 1907–1911.

Mesbah, A. (2016) Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, **36**, 30–44.

Moallemi, C. C. and Sağlam, M. (2017) Dynamic portfolio choice with linear rebalancing rules. *Journal of Financial and Quantitative Analysis*, **52**, 1247–1278.

Mohajerin Esfahani, P. and Kuhn, D. (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, **171**, 115–166.

Mukaidani, H. and Xu, H. (2017) Infinite horizon linear-quadratic stackelberg games for discrete-time stochastic systems. *Automatica*, **76**, 301–308.

Neu, G., Jonsson, A. and Gómez, V. (2017) A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*.

Ni, Y.-H., Elliott, R. and Li, X. (2015) Discrete-time mean-field stochastic linear–quadratic optimal control problems, ii: Infinite horizon case. *Automatica*, **57**, 65–77.

Ni, Y.-H., Zhang, J.-F. and Krstic, M. (2017) Time-inconsistent mean-field stochastic lq problem: Open-loop time-consistent control. *IEEE Transactions on Automatic Control*, **63**, 2771–2786.

Ni, Y.-H., Zhang, J.-F. and Li, X. (2014) Indefinite mean-field stochastic linear-quadratic optimal control. *IEEE Transactions on Automatic Control*, **60**, 1786–1800.

Pang, B. and Jiang, Z.-P. (2022) Robust reinforcement learning for stochastic linear quadratic control with multiplicative noise. *Trends in Nonlinear and Adaptive Control: A Tribute to Laurent Praly for his 65th Birthday*, 249–277.

Primbs, J. A. and Sung, C. H. (2009) Stochastic receding horizon control of constrained linear systems with state and control multiplicative noise. *IEEE Transactions on Automatic Control*, **54**, 221–230.

Puterman, M. L. (2014) *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rami, M. A., Chen, X., Moore, J. B. and Zhou, X. Y. (2001) Solvability and asymptotic behavior of generalized riccati equations arising in indefinite stochastic lq controls. *IEEE Transactions on Automatic Control*, **46**, 428–440.

Rubinstein, M. (2002) Markowitz's "portfolio selection": A fifty-year retrospective. *The Journal of finance*, **57**, 1041–1045.

Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. et al. (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Sun, W. G. and Wang, C. F. (2006) The mean-variance investment problem in a constrained financial market. *Journal of Mathematical Economics*, **42**, 885–895.

Sutton, R. S. and Barto, A. G. (2018) *Reinforcement learning: An introduction*. MIT press.

Tsallis, C. (1988) Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, **52**, 479–487.

Tsallis, C. (1994) What are the numbers that experiments provide. *Quimica Nova*, **17**, 468–471.

Tsallis, C. (2011) The nonadditive entropy s_q and its applications in physics and elsewhere: Some remarks. *Entropy*, **13**, 1765–1804.

Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wang, H., Zariphopoulou, T. and Zhou, X. Y. (2020) Reinforcement learning in continuous time and space: A stochastic control approach. *The Journal of Machine Learning Research*, **21**, 8145–8178.

Wang, T., Zhang, H. and Luo, Y. (2018) Stochastic linear quadratic optimal control for model-free discrete-time systems based on q-learning algorithm. *Neurocomputing*, **312**, 1–8.

Wang, W., Zhang, F. and Han, C. (2017) Distributed linear–quadratic regulator control for discrete-time multi-agent systems. *IET Control Theory & Applications*, **11**, 2279–2287.

Wilk, P., Wang, N. and Li, J. (2024) Multi-agent reinforcement learning for smart community energy management. *Energies*, **17**, 5211.

Wu, W., Gao, J., Li, D. and Shi, Y. (2018) Explicit solution for constrained scalar-state stochastic linear-quadratic control with multiplicative noise. *IEEE Transactions on Automatic Control*, **64**, 1999–2012.

Wulfmeier, M., Ondruska, P. and Posner, I. (2015) Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

Yao, D. D., Zhang, S. and Zhou, X. Y. (2006) Tracking a financial benchmark using a few assets. *Operations Research*, **54**, 232–246.

Young, P. C. (2011) *Recursive estimation and time-series analysis: An introduction for the student and practitioner*. Springer science & business media.

Zhang, Y., Qu, G., Xu, P., Lin, Y., Chen, Z. and Wierman, A. (2023) Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **7**, 1–51.

Zhao, R., Sun, X. and Tresp, V. (2019) Maximum entropy-regularized multi-goal reinforcement learning. In *International Conference on Machine Learning*, 7553–7562. PMLR.

Zhou, X. Y. and Li, D. (2000) Continuous-time mean-variance portfolio selection: A stochastic lq framework. *Applied Mathematics and Optimization*, **42**, 19–33.

Zuhlsdorff, C. (2001) The pricing of derivatives on assets with quadratic volatility. *Applied Mathematical Finance*, **8**, 235–262.