$\mathcal{H}_2$ LINEAR CONTROL WITH $\mathcal{H}_\infty$ ROBUSTNESS GUARANTEE: A GAME-THEORETIC APPROACH

BY

XIANGYUAN ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Electrical and Computer Engineering
in the Grainger College of Engineering of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Adviser:

Professor Tamer Başar

Abstract

In recent years, reinforcement learning (RL) has shown promising developments in solving sequential decision-making problems as well as handling continuous control tasks. Among the success stories, many are related to policy optimization (PO) algorithms, developed in the context of constrained optimization. To address the stability and robustness of the controller as the algorithm iterates, constraints such as the $\mathcal{H}_\infty$-norm one need to be enforced on-the-fly. Recently, Zhang et al. (2019) showed the implicit regularization and the global convergence property of PO methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem, a classic problem in the robust control literature. Despite the non-convex, non-coercive optimization landscape of the problem, iterates of PO methods are guaranteed to preserve the $\mathcal{H}_\infty$-norm constraint without explicit encoding, while converging to the global optimizer.

In this thesis, we demonstrate that the solution of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem can also be obtained through solving the Nash equilibrium (NE) of a sequential zero-sum linear-quadratic (LQ) game via double-loop PO methods. Specifically, we first show that the natural policy gradient algorithm can be applied to solve the inner loop problem with a fixed outer loop control policy. Then, we establish the desired stability and global convergence properties despite the non-coercive nature of the inner loop cost function. Subsequently, the outer loop problem can also be solved using the natural policy gradient algorithm, similar to the techniques presented in Zhang et al. (2019). The connection between the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem and the zero-sum LQ game provides a path to investigate model-free PO methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem.

To my parents, for their love and support.

Acknowledgments

This thesis is the product of the tremendous support that I have received from my adviser, mentors, collaborators, friends, and family throughout my undergraduate career. I am incredibly grateful and lucky that Prof. Tamer Başar took a chance on me when I shifted my research interest towards control at the beginning of my junior year. The two years that I worked in his group have been extremely rewarding, and I do not know how I could have learned so much if I had not been working in his group. I want to acknowledge the tremendously important role that Kaiqing Zhang, Erik Miehling, Weichao Mao, Muhammed Sayin, Aneeq Zaman, and Khalid Alshehri played as lab mates. Specifically, I would like to thank Kaiqing Zhang for not only being a role model in research but also being a mentor for many real-life problems.

I want to thank all my mentors and collaborators at UIUC, including Prof. Timothy Bretl, David Hanley, and Chuanzheng Li for mentoring during my sophomore year, Prof. Daniel Liberzon for many helpful discussions and his fabulous optimal control class, Prof. Seth Hutchinson for his warmest support during my graduate application season, and Peixin Chang for working together for many course projects. I would also like to thank all my friends for all their support during my undergraduate life. Special thanks to Xinyi Ren for her support and understanding.

In the summer of 2019, I had the fantastic experience of visiting CMU Biorobotics Lab under the mentorship of Prof. Howie Choset. I want to thank all the members in the lab, especially Shuo Yang, Zhaoyuan Gu, Hans Kumar, and Lu Li, for showing me what a great roboticist should be.

Lastly, I would like to thank my parents for their unconditional support and guidance throughout my life.

Table of Contents

Chapter 1 Introduction **1**
1.1 Literature Review 3
1.2 Notation 4
1.3 Organization 4

Chapter 2 Problem Formulation **5**
2.1 Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design Problem 5
2.2 Zero-Sum Linear Quadratic Game 7
2.3 Connection Between $\mathcal{H}_2/\mathcal{H}_\infty$ and Zero-Sum LQ Game 9

Chapter 3 Main Results **11**
3.1 The Inner Loop Problem 11
3.2 The Outer Loop Problem 17
3.3 Model-free Natural Policy Gradient Design 21

Chapter 4 Numerical Experiments **23**

Chapter 5 Conclusion **26**

References **27**

Chapter 1

Introduction

In recent years, reinforcement learning (RL) has shown promising developments in solving sequential decision-making problems [1] as well as handling continuous control tasks [2, 3, 4]. Although achieving astonishing empirical performance [5], theoretical analysis of RL algorithms has received relatively less attention. Lack of theoretical footing and performance guarantees raises concerns when applying RL algorithms to safety-critical systems such as robotics and self-driving cars, as failures of those systems can cause catastrophic social and economic impacts.

Among the success stories in RL, many of them are related to policy optimization (PO) algorithms, in the context of constrained optimization. Within the realm of PO methods, policy gradient methods [6, 7], actor-critic methods [8, 9] and other PO methods [10, 11] play important roles in modern RL. See [12] for a detailed review. Moreover, PO methods are natural choices for learning controller designs, as the stability and robustness concerns of the controller designs can be translated into optimization constraints as PO methods proceed. For example, a recent analysis of the LQR problem [13] has shown that PO methods can preserve stability along iterates, despite the non-convex optimization landscape of the LQR problem. It has been further demonstrated [12] that PO methods can preserve both stability and an $\mathcal{H}_\infty$-norm constraint for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem, even with lack of convexity and coercivity (i.e., the cost may have finite cost around the boundary of the robustness constraint set). Progress reported in the recent literature has clearly shown increased interest in analyzing the convergence property (i.e., guarantees for converging to (global) optimizer) of PO methods for RL [14, 15, 16, 17, 18], as well as continuous control [12, 13, 19, 20, 21, 22].

As an initial step towards developing provably robust RL algorithms for safety-critical automated systems such as robotics [23, 24], it is natural to consider first analyzing RL algorithms under the context of robust control tasks. This connection comes from the fact that the disturbance input in the robust control setting generalizes the model misspecification and uncertainty when applying RL algorithms to real-world systems. Therefore,

ensuring the worst-case performance of the learning-based controllers, using techniques from the robust control literature, is crucial for satisfying the safety concerns of RL.

In this thesis, we study properties of PO methods, specifically the natural policy gradient algorithm, for the discrete-time $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$-norm robustness guarantee problem, a classic robust control task. In contrast to the approach taken in [12], we demonstrate that the solution of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem could also be obtained through solving the Nash equilibrium (NE) of a corresponding sequential zero-sum linear-quadratic (LQ) game. Despite the nonconvex-nonconcave nature of the zero-sum LQ game, we show that the stability and robustness can be implicitly preserved for a carefully chosen stepsize, and the natural policy gradient algorithm converges to the NE of the zero-sum LQ game. Moreover, this convergent NE recovers the global optimum of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem.

In chapter 3, we first show that for a fixed outer loop control policy, applying the natural policy gradient algorithm can solve the inner loop problem of the zero-sum LQ game. The main challenge for the inner loop problem lies in the indefiniteness of the state weight matrix and no coercive property for the cost function, in contrast to the standard LQR formulation. Therefore, techniques in [13] fail in this "non-standard" LQR problem. Interestingly, inspired by the ideas from [25], we can avoid the stability issue without the additional projection step that was required in [22]. Hence, we can prove the desired convergence result for the inner loop problem of the proposed zero-sum LQ game. Similarly, we can also establish the stability guarantee and convergence property for the outer loop problem, using techniques similar to the one proposed in [12]. Combining the two results, we provide the convergence to the NE of the game. This NE recovers the global optimum of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ under certain conditions, as will be justified in this thesis.

The connection between the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem and the zero-sum LQ game provides a path to avoid the technical difficulties when designing model-free PO methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem, as discussed in [12]. At the end of this thesis, we provide some discussion and an algorithmic sketch on how to design a model-free algorithm for solving the zero-sum LQ game. Equivalently, we obtain the solution of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem using only a simulator. Rigorous analyses are being partly covered in the author's ongoing research.

1.1 Literature Review

The concept of mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem was first brought up in [26, 27] for the continuous-time setting and [28, 29] for the discrete-time setting. This problem can be viewed as a simplification for the $\mathcal{H}_\infty$ control problem [30], which aims to find the optimal controller that minimizes the $\mathcal{H}_\infty$-norm. Later, [31] solved the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem via a constrained optimization approach, enforcing the $\mathcal{H}_\infty$ constraints explicitly. The mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem also has a close relationship to other classes of control problems including risk-sensitive control [32, 33], maximum-entropy $\mathcal{H}_\infty$ control [34, 35], and zero-sum dynamic games [36].

Recent advances in the theoretical analysis of PO methods for the LQR problem [13, 20] have drawn new attention to the classic mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem. Specifically, [12] first proposed PO methods for solving the mixed design problem with robustness and convergence guarantees, despite the challenging optimization landscape of the problem. Due to the close relationship between the mixed design problem and the risk-sensitive control, the mixed design problem is also connected to the literature of zero-sum LQ game, a classic setting in the multi-agent RL domain [18, 37, 38, 39]. PO methods for the zero-sum LQ game with a global convergence guarantee to the NE were first shown by [22] but required an additional projection step. Very recently, [25] provided an alternative proof technique to remove the projection requirements in [22]. In contrast to the above model-based analyses, several works also studied model-free PO methods and their convergence property for the LQR problem. Notably, [13] developed a model-free policy gradient method for the LQR with global convergence. [21] adapted a two-point zeroth-order method to improve the sample complexity of the model-free policy gradient algorithm. Very recently, the sample complexity was enhanced by [40], but for the continuous-time setting. To the best of the author's knowledge, model-free PO methods for the mixed design problem have not received much attention yet, except for a brief discussion in [12].

Lastly, the robustness of RL algorithms with respect to model misspecification and uncertainty in the sense of $\mathcal{H}_\infty$-norm was first studied in [41]. [41] modeled the uncertainty as an adversarial controller that plays against the nominal controller. [42] also adapted a similar game-theoretic approach. Empirically, a technique known as domain randomization is widely used to deal with model uncertainty and misspecification by slightly perturbing the simulation environment between different trails [43]. However, this is

instead an engineering solution and thus not the interest of this thesis.

1.2 Notation

For a square matrix A of proper dimension, we use $Tr(A)$ to denote its trace. We also use $\rho(A)$ to denote the spectral radius of A. $\|A\|$ and $\|A\|_F$ are defined to be, respectively, the Euclidean norm and the Frobenius norm of A. If A is further symmetric, we use $A > 0$ to denote that A is positive definite. Similarly, $A \geq 0$, $A \leq 0$, and $A < 0$ are used to denote A positive semi-definite, negative semi-definite, and negative definite, respectively. Further, again for a symmetric matrix A, $\lambda_{\min}$ and $\lambda_{\max}$ are used to denote, respectively, the smallest and the largest eigenvalues of A. Lastly, for the discrete-time state space system

$$x_{t+1} = Ax_t + Bu_t, \quad z_t = Cx_t + Du_t, \tag{1.1}$$

we use $G = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ to denote the transfer function G from input u to output z, which can also be written as $G(z) = C(zI - A)^{-1}B + D$. The corresponding $\mathcal{H}_\infty$-norm is then defined as

$$\|G\|_\infty := \sup_\theta \lambda_{\max}^{1/2}\big[G(e^{-j\theta})^T G(e^{j\theta}))\big]. \tag{1.2}$$

1.3 Organization

We organize the rest of the thesis as follows. In Chapter 2, we review the classic mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem and discuss its connection to the zero-sum LQ game. We also provide a detailed formulation of the corresponding zero-sum LQ game. In Chapter 3, we propose a double-loop natural policy gradient algorithm for solving the zero-sum LQ game. We first provide a detailed stability and convergence analysis for the inner loop problem. Subsequently, we analyze the outer loop problem of the game, which when combined with the inner loop results, yields the solution to the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem. A model-free implementation sketch of the double-loop PO method is also provided in that chapter. In Chapter 4, we present simulation results to support our theory. Lastly, we provide concluding remarks in Chapter 5.

Chapter 2

Problem Formulation

This chapter is structured as follows. In section 2.1, we review the classic mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem. Then, we propose a formulation of the zero-sum LQ game in section 2.2. Lastly, we show the connection between the mixed design problem and zero-sum LQ game and discuss why the NE of the zero-sum LQ game is equivalent to the optimum solution of the mixed design problem; see section 2.3 for details.

2.1 Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Design Problem

Consider the linear dynamical system

$$x_{t+1} = Ax_t + Bu_t + Dw_t, \quad z_t = Cx_t + Eu_t, \tag{2.1}$$

where $x_t \in \mathbb{R}^m$ is the system state, $u_t \in \mathbb{R}^d$ is the control input, $w_t \in \mathbb{R}^n$ is the disturbance, and $z_t \in \mathbb{R}^l$ is the system output. A, B, C, D, and E are system matrices with appropriate dimensions. Additionally, we assume that $E^T[C\ E] = [0\ R]$ for some $R > 0$, which is a common assumption in the robust control literature [36]. According to [12, 29], LTI state-feedback controller can achieve optimal performance for the mixed design problem. Therefore, it is sufficient to consider state-feedback controllers restricted to the form $u_t = -Kx_t$, and the corresponding transfer function from disturbance w to the output z is $(C - EK)(zI - A + BK)^{-1}D$, which, in view of the assumption $E^T[C\ E] = [0\ R]$, can be represented as

$$\mathcal{T}(K) := \left[\begin{array}{c|c} A - BK & D \\ \hline (C^TC + K^TRK)^{\frac{1}{2}} & 0 \end{array}\right]. \tag{2.2}$$

Moreover, robustness of the controller can be ensured if the $\mathcal{H}_\infty$-norm of the transfer function matrix satisfies $\|\mathcal{T}(K)\|_\infty < \gamma$ for some $\gamma > 0$, according to the celebrated small

gain theorem [44]. We denote the set of control gain matrices that are both stabilizing and satisfying the $\mathcal{H}_\infty$-norm constraint as

$$\mathcal{K} := \Big\{ K \,||\, \rho(A - BK) < 1, \text{and } \|\mathcal{T}(K)\|_\infty < \gamma \Big\}. \tag{2.3}$$

The objective of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem is to minimize the $\mathcal{H}_2$ norm of the transfer function matrix with the $\mathcal{H}_\infty$-norm constraint, which guarantees the robustness of the system, enforced. This objective function is introduced as

$$\min_{K\in\mathcal{K}} \mathcal{J}(K) = \min_{K\in\mathcal{K}} \Big\{ -\gamma^2 \log\det(I - \gamma^{-2} P_K D D^T) \Big\}, \tag{2.4}$$

where P_K solves the Riccati equation

$$(A - BK)^T \widetilde{P_K} (A - BK) + C^T C + K^T R K - P_K = 0, \tag{2.5}$$

where $\widetilde{P_K} := P_K + P_K D(\gamma^2 I - D^T P_K D)^{-1} D^T P_K$. Note that the optimization landscape of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem is nonconvex and the problem does not have the coercivity property as in the LQR problem [12]. Next, we introduce the following corollary for the policy gradient and the stationary point of the mixed design problem.

Corollary 2.1 *The policy gradient of the mixed design problem is represented as*

$$\forall K \in \mathcal{K}, \quad \nabla \mathcal{J}(K) = 2\Big[(R + B^T \widetilde{P_K} B)K - B^T \widetilde{P_K} A\Big]\Delta_K, \tag{2.6}$$

where

$$\Delta_K := \sum_{t=0}^{\infty} \Big[(I - \gamma^{-2} P_K D D^T)^{-T} (A - BK)\Big]^t D(I - \gamma^{-2} D^T P_K D)^{-1} D^T \Big[(A - BK)^T (I - \gamma^{-2} P_K D D^T)^{-1}\Big]^t.$$

Moreover, suppose that the mixed design problem has a solution $K^* \in \mathcal{K}$, and for any stationary point $K \in \mathcal{K}$ such that $\nabla \mathcal{C}(K) = 0$, the pair $\big((I - \gamma^{-2} P_K D D^T)^{-T}(A - BK), D\big)$ is controllable. Then K^* is unique and has the form of

$$K^* = (R + B^T \widetilde{P_{K^*}} B)^{-1} B^T \widetilde{P_{K^*}} A. \tag{2.7}$$

Proof The differentiability of the objective function $\mathcal{J}(K)$ and the exact form of the policy gradient follow from Lemmas 3.3 and 3.4 of [12]. Then, for $K \in \mathcal{K}$, the Bounded Real Lemma [36, 45] suggests that $I - \gamma^{-2} D^T P_K D > 0$, which implies $\Delta_K \geq 0$. Since

$\big((I-\gamma^{-2}P_K DD^T)^{-T}(A-BK), D\big)$ is controllable, Δ_K is full-rank. Hence, the necessary optimality condition $\nabla \mathcal{J}(K) = 0$ leads to the unique stationary point that has the form of $K^* = (R + B^T\widetilde{P_{K^*}}B)^{-1}B^T\widetilde{P_{K^*}}A$. This completes the proof. ■

2.2 Zero-Sum Linear Quadratic Game

In this section, we consider the zero-sum LQ game represented by a linear dynamical system

$$x_{t+1} = Ax_t + Bu_t + Dw_t, \tag{2.8}$$

where $x_t \in \mathbb{R}^m$ is the system state, $u_t \in \mathbb{R}^d$ and $w_t \in \mathbb{R}^n$ are the control inputs of player 1 and player 2, respectively. A, B, D are matrices with proper dimensions such that $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times d}$, and $D \in \mathbb{R}^{m\times n}$. The objective of player 1 (player 2) is to minimize (maximize) the function

$$\inf_{\{u_t\}_{t\geq 0}} \sup_{\{w_t\}_{t\geq 0}} \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}(x_t^T Q x_t + u_t^T R^u u_t - w_t^T R^w w_t)\right], \tag{2.9}$$

where $x_0 \sim \mathcal{D}$ is the initial state sampled from a certain distribution $\mathcal{D}$, $Q \in \mathbb{R}^{m\times m}$, $R^u \in \mathbb{R}^{d\times d}$, and $R^w \in \mathbb{R}^{n\times n}$ are all positive definite. We introduce the generalized algebraic Riccati equation (GARE):

$$P = A^TPA + Q - \begin{bmatrix} A^TPB & A^TPD \end{bmatrix}\begin{bmatrix} R^u + B^TPB & B^TPD \\ D^TPB & -R^w + D^TPD \end{bmatrix}^{-1}\begin{bmatrix} B^TPA \\ D^TPA \end{bmatrix}. \tag{2.10}$$

Consider now the following assumption to ensure the existence of a value for the game [22, 36].

Assumption 2.2 *There exists a minimal positive definite solution P^* to the GARE* (2.10) *such that $R^w - D^TP^*D > 0$.*

Under the above assumption, and for each fixed x_0, the game has a value and can be characterized by the matrix P^* as follows:

$$\forall x_0 \in \mathbb{R}^m, \quad x_0^T P^* x_0 = \inf_{\{u_t\}_{t\geq 0}} \sup_{\{w_t\}_{t\geq 0}} \sum_{t=0}^{\infty} c_t(x_t, u_t, w_t) = \sup_{\{w_t\}_{t\geq 0}} \inf_{\{u_t\}_{t\geq 0}} \sum_{t=0}^{\infty} c_t(x_t, u_t, w_t), \tag{2.11}$$

where $c_t(x_t, u_t, w_t) = x_t^T Q x_t + u_t^T R^u u_t - w_t^T R^w w_t$. Moreover, there exists a pair of linear feedback stabilizing policies $\{u_t^*\}_{t\geq 0}$ and $\{w_t^*\}_{t\geq 0}$ such that the equality in the above equation is obtained. The optimal policies can be explicitly written as

$$u_t^* = -K^* x_t, \qquad w_t^* = -L^* x_t, \tag{2.12}$$

where $K^* \in \mathbb{R}^{d\times m}$ and $L^* \in \mathbb{R}^{n\times m}$ are the control gain matrices for the minimizer and the maximizer, respectively. These gain matrices are given by

$$\begin{aligned} K^* &= [R^u + B^T P^* B - B^T P^* D(-R^w + D^T P^* D)^{-1} D^T P^* B]^{-1} \\ &\quad \times [B^T P^* A - B^T P^* D(-R^w + D^T P^* D)^{-1} D^T P^* A] \qquad (2.13) \\ L^* &= [-R^w + D^T P^* D - D^T P^* B(R^u + B^T P^* B)^{-1} B^T P^* D]^{-1} \\ &\quad \times [D^T P^* A - D^T P^* B(R^u + B^T P^* B)^{-1} B^T P^* A]. \qquad (2.14) \end{aligned}$$

The value of the game with random x_0 is therefore $\mathbb{E}_{x_0\sim\mathcal{D}}(x_0^T P^* x_0)$. This implies that searching the pair $(K^*, L^*) \in \mathbb{R}^{d\times m} \times \mathbb{R}^{n\times m}$ leads to the solution of (2.9) (with inf and sup operations interchanged), which is further the NE of the game. To develop PO methods that provably converge to the NE, (K^*, L^*), we focus on finding the state feedback policies of the players parameterized by $u_t = -Kx_t$, and $w_t = -Lx_t$, such that $\rho(A - BK - DL) < 1$. Subsequently, we denote the expected value in (2.9) as

$$\mathcal{C}(K, L) := \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} (x_t^T Q x_t + (Kx_t)^T R^u (Kx_t) - (Lx_t)^T R^w (Lx_t))\right], \tag{2.15}$$

and define $P_{K,L}$ as the unique solution to the Lyapunov equation

$$P_{K,L} = (A - BK - DL)^T P_{K,L}(A - BK - DL) + Q + K^T R^u K - L^T R^w L. \tag{2.16}$$

Then for any stabilizing control pair (K, L), we have

$$\mathcal{C}(K, L) = \mathbb{E}_{x_0 \sim \mathcal{D}}(x_0^T P_{K,L} x_0). \tag{2.17}$$

Also, we define $\Sigma_{K,L}$ as the state correlation matrix, i.e., $\Sigma_{K,L} := \mathbb{E}_{x_0 \sim \mathcal{D}}\{\sum_{t=0}^{\infty} x_t x_t^T\}$. The goal is to solve the following minimax problem

$$\min_K \max_L \ \mathcal{C}(K, L), \tag{2.18}$$

such that for any $K \in \mathbb{R}^{d \times m}$ and $L \in \mathbb{R}^{n \times m}$, $\mathcal{C}(K^*, L) \leq \mathcal{C}(K^*, L^*) \leq \mathcal{C}(K, L^*)$. Note that $\mathcal{C}(K, L)$ is not convex-concave [22]. The explicit forms of the policy gradients of $\mathcal{C}(K, L)$ are represented as

$$\nabla_K \mathcal{C}(K, L) = 2\big[(R^u + B^T P_{K,L} B)K - B^T P_{K,L}(A - DL)\big]\Sigma_{K,L} \tag{2.19}$$

$$\nabla_L \mathcal{C}(K, L) = 2\big[(-R^w + D^T P_{K,L} D)L - D^T P_{K,L}(A - BK)\big]\Sigma_{K,L}. \tag{2.20}$$

Moreover, the stationary point of $\mathcal{C}(K, L)$ (i.e., $\nabla_K \mathcal{C}(K, L) = \nabla_L \mathcal{C}(K, L) = 0$) captures the NE of the game if the control pair (K, L) is stabilizing, $\Sigma_{K,L}$ is full-rank, and $(-R^w + D^T P_{K,L} D)$ invertible. A detailed discussion of the stationary point property can be found in Lemma 3.3 of [22].

2.3 Connection Between $\mathcal{H}_2/\mathcal{H}_\infty$ and Zero-Sum LQ Game

Notice that under the assumption 2.2, the pair of optimal control gain matrices at the NE of the game, denoted as (K^*, L^*), are stabilizing (i.e., $\rho(A - BK^* - DL^*) < 1$). An explicit form of the optimal control gain matrices can be seen in (2.13) and (2.14). Therefore, the NE of the game can be obtained by searching for only the stabilizing control gain matrices (K, L) that solve

$$\begin{aligned} &\min_K \max_L \ \mathbb{E}_{x_0 \in \mathcal{D}}\Big\{ \sum_{t=0}^{\infty} c_t(x_t, u_t, w_t) \Big\} \\ =&\min_K \max_L \ \mathbb{E}_{x_0 \in \mathcal{D}}\Big\{ \sum_{t=0}^{\infty} \Big[x_t^T Q x_t + (Kx_t)^T R^u (Kx_t) - (Lx_t)^T R^w (Lx_t) \Big] \Big\}. \end{aligned} \tag{2.21}$$

According to (2.17), we have

$$\mathcal{C}(K, L) = Tr(\Sigma_0 P_{K,L}), \tag{2.22}$$

where $\Sigma_0 = \mathbb{E}_{x_0 \in \mathcal{D}}(x_0^T x_0)$ and $P_{K,L}$ solves the Lyapunov equation

$$(A - BK - DL)^T P_{K,L}(A - BK - DL) - P_{K,L} + Q + K^T R^u K - L^T R^w L. \tag{2.23}$$

For a given outer loop control gain matrix K, the inner loop player can maximize the value function over all L such that the control gain pair (K, L) preserves the stability. The optimal control gain matrix of the inner loop player has the form of

$$L(K) = (-R^w + D^T P_{K,L(K)} D)^{-1} D^T P_{K,L(K)}(A - BK). \tag{2.24}$$

Therefore, substituting (2.24) into (2.23) will give us

$$(A - BK)^T \overline{P_{K,L(K)}}(A - BK) - P_{K,L(K)} + Q + K^T R^u K = 0, \tag{2.25}$$

where $\overline{P_{K,L(K)}} := P_{K,L(K)} + P_{K,L(K)} D(R^w - D^T P_{K,L(K)} D)^{-1} D^T P_{K,L(K)}$. Furthermore, the optimal K satisfies the equation

$$K^* = (R^u + B^T \overline{P_{K,L(K)}} B)^{-1} B^T \overline{P_{K,L(K)}} A. \tag{2.26}$$

Based on the above formulations, we formally establish the connection between the mixed design problem and the zero-sum LQ game in the following corollary.

Corollary 2.3 *The zero-sum LQ game proposed in section 2.2 is equivalent to the mixed design problem introduced in section 2.1.*

Proof One can see that by replacing $R^u = R$, and $P_{K,L(K)} = P_K$, the optimal control gain solution (2.26) of the zero-sum LQ game becomes equivalent to (2.7), which is the equation of the solution of the mixed design problem. Also, by further replacing $R^w = \gamma^2 I$, and $Q = C^T C$, the Riccati equation (2.25) will have the same form as (2.5). Therefore, solving the zero-sum LQ game proposed in section 2.2 is equivalent to solving the mixed design problem in section 2.1. This completes the proof. ■

Chapter 3

Main Results

In this chapter, we introduce the solution to the inner loop problem in section 3.1 and the solution to the outer loop problem in section 3.2, for the zero-sum LQ game proposed in section 2.2. The stability and convergence analyses along with a sketch of the algorithm is also provided. In section 3.3, we provide an algorithmic sketch of the model-free double-loop natural policy gradient method.

3.1 The Inner Loop Problem

In this section, we focus on solving the inner loop problem for the zero-sum LQ game. First, we provide the following remark:

Remark 3.1 *Note that for a fixed K, the inner loop player solves*

$$\begin{aligned}\underset{L}{\text{argmax}}\ \mathcal{C}(K,L) &= \underset{L}{\text{argmax}}\ \mathbb{E}_{x_0\in\mathcal{D}}\left[\sum_{t=0}^{\infty}(x^T(Q+K^TR^uK)x-(Lx_t)^TR^w(Lx_t))\right]\\ &= \underset{L}{\text{argmin}}\ \mathbb{E}_{x_0\in\mathcal{D}}\left[\sum_{t=0}^{\infty}(x^T(-Q-K^TR^uK)x+(Lx_t)^TR^w(Lx_t))\right]\\ &= \underset{L}{\text{argmin}}\ \mathcal{C}_m(K,L),\end{aligned}$$

after converting to a standard minimization problem. The above problem can be viewed as a non-standard LQR problem with $-Q-K^TR^uK$ being the state weighting matrix. Unlike the Q matrix in the standard LQR problem, we do not have positive definiteness for $-Q-K^TR^uK$ (it is in fact negative definite), resulting in a landscape of the cost function, which is no longer coercive. In the following paragraphs of this section, we show that without the positive definiteness assumption, the pair (K,L) will remain stabilizing along the natural policy gradient iterates, and the global convergence to the optimal L given a fixed K can still be guaranteed.

Then, we introduce the natural policy gradient algorithm [7], which enjoys an improved convergence rate compared to the vanilla policy gradient algorithm, due to the dependence on the information geometry [13]. To solve the inner loop minimization problem for a fixed outer loop control policy K, the natural policy gradient algorithm has the form of

$$L' = L - \eta \nabla_L \mathcal{C}_m(K, L) \Sigma_{K,L}^{-1} = L - 2\eta \big[(R^w + D^T P_{K,L} D) L - D^T P_{K,L} (A - BK) \big], \tag{3.1}$$

where $P_{K,L}$ is the unique solution to the Lyapunov equation

$$P_{K,L} = (A - BK - DL)^T P_{K,L} (A - BK - DL) - Q - K^T R^u K + L^T R^w L. \tag{3.2}$$

For simplicity, we introduce

$$A_{K,L} = A - BK - DL, \tag{3.3}$$
$$E_{K,L} = R^w L - D^T P_{K,L} A_{K,L}, \tag{3.4}$$
$$F_{K,L} = R^w - D^T P_{K,L} D, \tag{3.5}$$
$$\text{value:} \quad V_{K,L}(x) = x^T P_{K,L} x, \tag{3.6}$$
$$\text{action-value:} \quad Q_{K,L}(x, u, w) = -x^T Q x - u^T R^u u + w^T R^w w + V_{K,L}(Ax + Bu + Dw), \tag{3.7}$$
$$\text{advantage:} \quad \mathcal{A}_{K,L}(x, u, w) = Q_{K,L}(x, u, w) - V_{K,L}(x). \tag{3.8}$$

Then, we provide the following three technical lemmas.

Lemma 3.2 ***(Inner Loop Comparison Lemma)*** *For the stabilizing pairs (K, L) and $(K, \widetilde{L})$, let $P_{K,L}$ and $P_{K,\widetilde{L}}$ be the corresponding solution of the Lyapunov equation* (3.2)*. We have*

$$\begin{aligned} P_{K,L} - P_{K,\widetilde{L}} =& A_{K,\widetilde{L}}^T (P_{K,L} - P_{K,\widetilde{L}}) A_{K,\widetilde{L}} + (L - \widetilde{L})^T E_{K,L} \\ &+ E_{K,L}^T (L - \widetilde{L}) + (L - \widetilde{L})^T F_{K,L} (L - \widetilde{L}). \end{aligned} \tag{3.9}$$

Proof By definition,

$$\begin{aligned} P_{K,L} &= (A - BK - DL)^T P_{K,L} (A - BK - DL) - Q - K^T R^u K + L^T R^w L \\ P_{K,\widetilde{L}} &= (A - BK - D\widetilde{L})^T P_{K,\widetilde{L}} (A - BK - D\widetilde{L}) - Q - K^T R^u K + \widetilde{L}^T R^w \widetilde{L}. \end{aligned}$$

Therefore, we have

$$\begin{aligned}
P_{K,L} - P_{K,\widetilde{L}} &= A_{K,\widetilde{L}}^T (P_{K,L} - P_{K,\widetilde{L}}) A_{K,\widetilde{L}} - (L - \widetilde{L})^T D^T P_{K,L} A_{K,L} - (D^T P_{K,L} A_{K,L})^T (L - \widetilde{L}) \\
&\quad - (L - \widetilde{L})^T D^T P_{K,L} D (L - \widetilde{L}) + (L - \widetilde{L})^T (R^w L) \\
&\quad + (R^w L)^T (L - \widetilde{L}) + (L - \widetilde{L})^T R^w (L - \widetilde{L}) \\
&= A_{K,\widetilde{L}}^T (P_{K,L} - P_{K,\widetilde{L}}) A_{K,\widetilde{L}} + (L - \widetilde{L})^T (R^w L - D^T P_{K,L} A_{K,L}) \\
&\quad + (R^w L - D^T P_{K,L} A_{K,L})^T (L - \widetilde{L}) + (L - \widetilde{L})^T (R^w - D^T P_{K,L} D)(L - \widetilde{L}) \\
&= A_{K,\widetilde{L}}^T (P_{K,L} - P_{K,\widetilde{L}}) A_{K,\widetilde{L}} + (L - \widetilde{L})^T E_{K,L} + E_{K,L}^T (L - \widetilde{L}) + (L - \widetilde{L})^T F_{K,L} (L - \widetilde{L}).
\end{aligned}$$

This completes the proof. ■

Lemma 3.3 ***(Inner Loop Cost Difference Lemma)*** *Suppose that both (K, L) and (K, L') are stabilizing, and let $\{x'_t\}_{t\geq 0}$ and $\{u_t, w'_t\}$ be the sequences of state and action pairs generated by (K, L'). Then, we have*

$$V_{K,L'}(x) - V_{K,L}(x) = \sum_{t\geq 0} \mathcal{A}_{K,L}(x'_t, u_t, w'_t) \tag{3.10}$$

$$\mathcal{A}_{K,L}(x, -Kx, -L'x) = 2x^T (L' - L)^T E_{K,L} x - x^T (L' - L)^T F_{K,L} (L' - L) x. \tag{3.11}$$

Proof Let c'_t be the finite cost generated by (K, L'). Then

$$\begin{aligned}
V_{K,L'}(x) - V_{K,L}(x) &= \sum_{t\geq 0} c'_t - V_{K,L}(x) = \sum_{t\geq 0} \Big[c'_t + V_{K,L}(x'_t) - V_{K,L}(x'_t) \Big] - V_{K,L}(x) \\
&= \sum_{t\geq 0} \Big[c'_t + V_{K,L}(x'_{t+1}) - V_{K,L}(x'_t) \Big] = \sum_{t\geq 0} \mathcal{A}_{K,L}(x'_t, u_t, w'_t).
\end{aligned}$$

Moreover, letting $u = -Kx$ and $w = -L'x$, we have

$$\begin{aligned}
\mathcal{A}_{K,L}(x, u, w) &= Q_{K,L}(x, u, w) - V_{K,L}(x) \\
&= x^T \Big[-Q - K^T R^u K + (L')^T R^w L' \Big] x + x^T A_{K,L'}^T P_{K,L} A_{K,L'} x - V_{K,L}(x) \\
&= 2x^T (L' - L)^T \Big[R^w - D^T P_{K,L} A_{K,L} \Big] x - x^T (L' - L) \Big[R^w - D^T P_{K,L} D \Big] (L' - L) x \\
&= 2x^T (L' - L)^T E_{K,L} x - x^T (L' - L)^T F_{K,L} (L' - L) x,
\end{aligned}$$

which completes the proof. ■

Lemma 3.4 ***(Inner Loop Gradient Domination Lemma)*** *Let $L(K)$ be optimal for a given fixed K, Suppose that (K, L) leads to finite cost. Then, the following upper bound holds:*

$$\mathcal{C}_m(K,L) - \mathcal{C}_m(K,L(K)) \leq \frac{\|\Sigma_{K,L(K)}\|}{\mu^2 \sigma_{min}(R^w)}\, Tr(\nabla_L \mathcal{C}_m(K,L)^T \nabla_L \mathcal{C}_m(K,L)), \tag{3.12}$$

where $\mu := \sigma_{min}(\mathbb{E}_{x_0 \in \mathcal{D}}[x_0 x_0^T])$.

Proof From (3.11), we have

$$\begin{aligned}
&\mathcal{A}_{K,L}(x, -Kx, -L'x) \\
&= 2x^T(L'-L)^T E_{K,L} x - x^T (L'-L)^T F_{K,L}(L'-L)x \\
&= 2Tr(xx^T(L'-L)^T E_{K,L}) + Tr(xx^T(L'-L)^T(D^T P_{K,L} D - R^w)(L'-L)) \\
&= Tr\Big[xx^T(L'-L+(D^T P_{K,L}D - R^w)^{-1}E_{K,L})^T (D^T P_{K,L}D - R^w)(L'-L+(D^T P_{K,L}D-R^w)^{-1}E_{K,L})\Big] \\
&\quad - Tr(xx^T E_{K,L}^T (D^T P_{K,L} D - R^w)^{-1} E_{K,L}) \\
&\geq -Tr(xx^T E_{K,L}^T (D^T P_{K,L} D - R^w)^{-1} E_{K,L}).
\end{aligned}$$

The third step is obtained by completing the square while the last step is due to the fact that $Tr(M) \geq 0$ if $M \geq 0$. Now, let $\{x_t^*\}_{t\geq 0}$, $\{u_t\}_{t\geq 0}$, and $\{w_t^*\}_{t\geq 0}$ be the sequences generated by the pair $(K, L(K))$. Then, the following upper bound can be derived:

$$\begin{aligned}
\mathcal{C}_m(K,L) - \mathcal{C}_m(K,L(K)) &= -\mathbb{E}_{x_0^* \in \mathcal{D}} \sum_{t\geq 0} \mathcal{A}_{K,L}(x_t^*, u_t, w_t^*) \\
&\leq \mathbb{E}_{x_0^* \in \mathcal{D}} \sum_{t\geq 0} Tr(x_t^* (x_t^*)^T E_{K,L}^T (D^T P_{K,L} D - R^w)^{-1} E_{K,L}) \\
&= Tr(\Sigma_{K,L(K)} E_{K,L}^T (D^T P_{K,L} D - R^w)^{-1} E_{K,L}) \\
&\leq \|\Sigma_{K,L(K)}\|\, \|(D^T P_{K,L} D - R^w)^{-1}\|\, Tr(E_{K,L}^T E_{K,L}) \\
&\leq \frac{\|\Sigma_{K,L(K)}\|}{\sigma_{\min}(R^w)}\, Tr(\Sigma_{K,L}^{-1} \nabla_L \mathcal{C}_m(K,L)^T \nabla_L \mathcal{C}_m(K,L) \Sigma_{K,L}^{-1}) \\
&\leq \frac{\|\Sigma_{K,L(K)}\|}{\sigma_{\min}(\Sigma_{K,L})^2 \sigma_{\min}(R^w)}\, Tr(\nabla_L \mathcal{C}_m(K,L)^T \nabla_L \mathcal{C}_m(K,L)) \\
&\leq \frac{\|\Sigma_{K,L(K)}\|}{\mu^2 \sigma_{\min}(R^w)}\, Tr(\nabla_L \mathcal{C}_m(K,L)^T \nabla_L \mathcal{C}_m(K,L)),
\end{aligned}$$

where $\mu := \sigma_{\min}(\mathbb{E}_{x_0 \in \mathcal{D}}[x_0 x_0^T])$. Here, the third step is by the definition of $\Sigma_{K,L(K)}$ and the last step is because $\Sigma_{K,L(K)} \geq \mathbb{E}_{x_0 \in \mathcal{D}}[x_0 x_0^T]$. This completes the proof. ■

With the technical lemmas stated above, we now have the stability and convergence results for the inner loop problem in the following theorem.

Theorem 3.5 *Suppose that for a fixed K, the ARE*

$$(A - BK)^T P (A - BK) - P - Q - K^T R^u K - (A - BK)^T P D (R^w + D^T P D)^{-1} D^T P (A - BK) = 0$$

has a stabilizing solution P^+. Then the update rule

$$L' = L - \eta \nabla_L \mathcal{C}_m(K, L) \Sigma_{K,L}^{-1} = L - 2\eta E_{K,L},$$

where $\eta = \frac{1}{2\|R^w - D^T P_{K,L} D\|}$ and $P_{K,L}$ is the solution of the Lyapunov equation

$$A_{K,L}^T P_{K,L} A_{K,L} - Q - K^T R^u K + L^T R^w L - P_{K,L} = 0,$$

converges linearly to $L(K)$ given (K, L_0) is a stabilizing pair. Equivalently,

$$\mathcal{C}_m(K, L_t) - \mathcal{C}_m(K, L(K)) \leq q^t (\mathcal{C}_m(K, L_0) - \mathcal{C}_m(K, L(K))),$$

for some constant $q \in (0, 1)$.

Proof We first show that natural policy gradient updates preserve stability when the stepsize is chosen to be sufficiently small. That is, starting with a stabilizing pair (K, L), the one-step natural policy gradient update of L, denoted by L′, will still guarantee that the updated pair, (K, L′), is stabilizing. Suppose (K, L) is a stabilizing pair and the natural gradient update of L is given by $L' = L - \nu \nabla_L \mathcal{C}_m(K, L) \Sigma_{K,L}^{-1}$. Then, for a sufficiently small ν, the pair (K, L') is stabilizing because of the continuity of the eigenvalues. Then, suppose $\nu \in [0, \xi)$ is the maximum interval that can ensure (K, L') stabilizing and $\nu \to \xi$ leads to (K, L') marginally stabilizing (i.e., $\rho(A - BK - DL') = 1$). Then, we know that such a marginally stable pair (K, L') has the corresponding $P_{K,L'}$ solving the Lyapunov equation

$$A_{K,L'}^T P_{K,L'} A_{K,L'} - Q - K^T R^u K + (L')^T R^w (L') - P_{K,L'} = 0. \tag{3.13}$$

Note that because $P_{K,L'}$ is bounded by $P_{K,L(K)} \leq P_{K,L'} \leq P_{K,L}$, there exists such a marginally stabilizing pair (K, L'). Then, by lemma 3.2, we have

$$\begin{aligned} A_{K,L'}^T(P_{K,L(K)} - P_{K,L'})A_{K,L'} - (P_{K,L(K)} - P_{K,L'}) + (L(K) - L')^T E_{K,L(K)} \\ + E_{K,L(K)}^T(L(K) - L') + (L(K) - L')^T F_{K,L(K)}(L(K) - L') = 0. \end{aligned} \tag{3.14}$$

The above equation suggests that if (K, L') is marginally stabilizing, then we will have $L(K)v = L'v$, where v is the corresponding eigenvector that satisfies $A_{K,L'}v = \lambda v$ and $|\lambda| = 1$. This is a contradiction as $(K, L(K))$ was assumed to be stabilizing. Therefore, $\{P_{K,L_t}\}_{t\geq 0}$ following natural policy gradient update is a monotonic non-increasing sequence lower bounded by $P_{K,L(K)}$, concluding that for a sufficiently small stepsize, the updates preserve the stability of the control gain pair.

Next, we prove that when $\eta \in \left[0, \frac{1}{\|R^w - D^T P_{K,L} D\|}\right]$, the above stability property indeed can be ensured. Starting from the inner loop comparison Lemma 3.2 and plugging in (3.1), we have

$$\begin{aligned} P_{K,L'} - P_{K,L} - A_{K,L'}^T(P_{K,L'} - P_{K,L})A_{K,L'} &= -4\eta E_{K,L}^T E_{K,L} + 4\eta^2 E_{K,L}^T F_{K,L} E_{K,L} \\ &= -4\eta E_{K,L}^T(I - \eta F_{K,L})E_{K,L}. \end{aligned} \tag{3.15}$$

Therefore, we have $P_{K,L'} \leq P_{K,L}$ if the stepsize η satisfies $\eta \in \left[0, \frac{1}{\|R^w - D^T P_{K,L} D\|}\right]$. Moreover, the optimal stepsize can be found by minimizing the RHS of (3.15), resulting in the minimizer $\eta = \frac{1}{2\|R^w - D^T P_{K,L} D\|}$.

Lastly, we prove that with $\eta = \frac{1}{2\|R^w - D^T P_{K,L} D\|}$, the natural policy gradient converges linearly to $L(K)$ given that the initial control input pair, (K, L_0), is stabilizing. By Lemma 3.3 and substituting in (3.1), the one-step difference between $\mathcal{C}_m(K, L')$ and $\mathcal{C}_m(K, L)$ can be computed as

$$\begin{aligned} \mathcal{C}_m(K, L') - \mathcal{C}_m(K, L) &= -2Tr(\Sigma_{K,L'}(L - L')^T E_{K,L}) + Tr(\Sigma_{K,L'}(L - L')^T(D^T P_{K,L} D - R^w)(L - L')) \\ &= -4\eta Tr(\Sigma_{K,L'} E_{K,L}^T E_{K,L}) + 4\eta^2 Tr(\Sigma_{K,L'} E_{K,L}^T (D^T P_{K,L} D - R^w) E_{K,L}) \\ &\leq -4\eta Tr(\Sigma_{K,L'} E_{K,L}^T E_{K,L}) + 4\eta^2 \|D^T P_{K,L} D - R^w\| Tr(\Sigma_{K,L'} E_{K,L}^T E_{K,L}) \\ &= -2\eta Tr(\Sigma_{K,L'} E_{K,L}^T E_{K,L}) \\ &\leq -2\eta\mu Tr(E_{K,L}^T E_{K,L}) \\ &\leq -2\eta \frac{\mu \sigma_{\min}(R^w)}{\|\Sigma_{K,L(K)}\|}(\mathcal{C}_m(K, L) - \mathcal{C}_m(K, L(K))), \end{aligned}$$

where the last inequality comes from Lemma 3.4. Then, the convergence rate can be computed as

$$\mathcal{C}_m(K, L') - \mathcal{C}_m(K, L(K)) \leq \left(1 - \frac{\mu\sigma_{\min}(R^w)}{\|R^w - D^T P_{K,L} D\| \, \|\Sigma_{K,L(K)}\|}\right)(\mathcal{C}_m(K, L) - \mathcal{C}_m(K, L(K))). \tag{3.16}$$

Applying the process iteratively, we can conclude that

$$\mathcal{C}_m(K, L_t) - \mathcal{C}_m(K, L(K)) \leq q^t(\mathcal{C}_m(K, L_0) - \mathcal{C}_m(K, L(K)), \tag{3.17}$$

where $q = \left(1 - \frac{\mu\sigma_{\min}(R^w)}{\|R^w - D^T P_{K,L} D\| \, \|\Sigma_{K,L(K)}\|}\right)$. This completes the proof. ■

To conclude, for a fixed outer loop control policy K, the natural policy gradient algorithm with a specific stepsize requirement is guaranteed to converge to the optimal solution $L(K)$ while preserving the stability of the control pair (K, L) along the iterates. In the next section, we focus on how the natural policy gradient algorithm can be applied to solve the outer loop problem of the zero-sum LQ game.

3.2 The Outer Loop Problem

The objective of the outer loop problem is to solve

$$\min_K \ \mathcal{C}(K, L(K)) \tag{3.18}$$

$$= \min_K \ \mathbb{E}_{x_0 \in \mathcal{D}}\Big[\sum_{t=0}^{\infty}(x^T Q x + (Kx_t)^T R^u (Kx_t) - (L(K)x_t)^T R^w (L(K)x_t))\Big] \tag{3.19}$$

$$= \min_K \ \mathbb{E}_{x_0 \in \mathcal{D}}\Big[\sum_{t=0}^{\infty}(x^T (Q - L(K)^T R^w L(K)) x + (Kx_t)^T R^u (Kx_t))\Big], \tag{3.20}$$

where $L(K)$ is the optimal (best-response) policy of the inner loop player for a fixed K, and the updates of K need to preserve the stability and robustness constraints defined in (2.3). Note that similar to the inner loop problem presented in section 3.1, the state weighting matrix $Q - L(K)^T R^w L(K)$ is also not positive definite. Therefore, the outer loop problem does not have the coercive property either, requiring extra care to address the stability and robustness issues along the iterates. We first present the following explicit form of

the natural policy gradient updates for the outer loop problem.

Lemma 3.6 *The natural policy gradient for the outer loop problem has the form of*

$$K' = K - 2\alpha\big[(R^u + B^T\overline{P_{K,L(K)}}B)K - B^T\overline{P_{K,L(K)}}A\big], \tag{3.21}$$

where

$$\overline{P_{K,L(K)}} = P_{K,L(K)} + P_{K,L(K)}D(R^w - D^T P_{K,L(K)}D)^{-1}D^T P_{K,L(K)}, \tag{3.22}$$

and $P_{K,L(K)}$ is the solution to the ARE

$$(A - BK)^T\overline{P_{K,L(K)}}(A - BK) + Q + K^T R^u K - P_{K,L(K)} = 0. \tag{3.23}$$

The term α is the stepsize to be determined later.

Proof According to (2.19), the one-step natural policy gradient update for the outer loop problem is represented as

$$\begin{aligned} K' &= K - \alpha\nabla_K \mathcal{C}(K, L(K))\Sigma^{-1}_{K,L(K)} \\ &= K - 2\alpha[(R^u + B^T P_{K,L(K)}B)K - B^T P_{K,L(K)}(A - DL(K))]. \end{aligned}$$

Substituting $L(K) = (-R^w + D^T P_{K,L(K)}D)^{-1}D^T P_{K,L(K)}(A - BK)$ into the above equation, we have

$$\begin{aligned} K' &= K - 2\alpha\big[(R^u + B^T P_{K,L(K)}B)K - B^T P_{K,L(K)}A \\ &\qquad + B^T P_{K,L(K)}D(-R^w + D^T P_{K,L(K)}D)^{-1}D^T P_{K,L(K)}(A - BK)\big] \\ &= K - 2\alpha[(R^u + B^T\overline{P_{K,L(K)}}B)K - B^T\overline{P_{K,L(K)}}A], \end{aligned}$$

where $\overline{P_{K,L(K)}}$ follows the definition in (3.22). This completes the proof. ■

For simplicity in notation, we let $Z_K = (R^u + B^T\overline{P_{K,L(K)}}B)K - B^T\overline{P_{K,L(K)}}A$. Then, we have $K' = K - 2\alpha Z_K$. With the above natural policy update rule, the desired regularization and convergence property can be shown following the proof in [12]. We provide the result in the following theorem, whose proof is sketched.

Theorem 3.7 *If the stepsize α satisfies*

$$\alpha \leq \frac{1}{2\|R^u + B^T\overline{P_{K,L(K)}}B\|}, \tag{3.24}$$

the natural policy gradient update (3.21) *converges to the optimal solution, which is the NE of the game, with globally sublinear rate. Moreover, the natural policy gradient update enjoys the implicit regularization property, that is, for $K \in \mathcal{K}$, we have $K' \in \mathcal{K}$.*

Proof The proof follows the proofs of Theorems 4.3 and 4.4 in [12]. We provide a sketch of the proof here. According to the Bounded Real Lemma [36, 45], the following two conditions are equivalent under the assumption that K is stabilizing: i) The control gain matrix K satisfies $K \in \mathcal{K}$ as defined in (2.3); ii) There exists some $P > 0$, such that $\gamma^{-2}(R^w - D^T P D) > 0$ and $(A - BK)^T\overline{P}(A - BK) - P + Q + K^T R^u K < 0$, where $\overline{P}$ is defined as $\overline{P} = P + PD(R^w - D^T P D)^{-1}D^T P$. Therefore, we focus on proving that K' after one-step natural policy gradient update (3.21) satisfies the second equivalent condition.

We first prove the implicit regularization property. Choosing $P = P_{K,L(K)}$ and substituting in (3.21), we have

$$\begin{aligned}
&(A - BK')^T\overline{P_{K,L(K)}}(A - BK') - P_{K,L(K)} + Q + (K')^T R^u (K') \\
&\quad = (K' - K)^T(R^u + B^T\overline{P_{K,L(K)}}B)\big[K' - (R^u + B^T\overline{P_{K,L(K)}})^{-1}B^T\overline{P_{K,L(K)}}A\big] \\
&\qquad + \big[K - (R^u + B^T\overline{P_{K,L(K)}}B)^{-1}B^T\overline{P_{K,L(K)}}A\big]^T(R^u + B^T\overline{P_{K,L(K)}}B)(K' - K) \\
&\quad = -4\alpha Z_K^T Z_K + 4\alpha^2 Z_K^T(R^u + B^T\overline{P_{K,L(K)}}B)Z_K.
\end{aligned}$$

If the stepsize α satisfies

$$\alpha \leq \frac{1}{2\|R^u + B^T\overline{P_{K,L(K)}}B\|},$$

then we can have

$$(A - BK')^T\overline{P_{K,L(K)}}(A - BK') - P_{K,L(K)} + Q + (K')^T R^u (K') \leq 0.$$

Now suppose that $P = P_{K,L(K)} + \beta\widehat{P}$ for some $\beta > 0$ and $\widehat{P}$ is the solution to the Lyapunov equation

$$(A - BK)^T(\gamma^{-2}R^w - \gamma^{-2}DD^T P_{K,L(K)})^{-T}\widehat{P}(\gamma^{-2}R^w - \gamma^{-2}DD^T P_{K,L(K)})(A - BK) - \widehat{P} = -I.$$

We have

$$
\begin{aligned}
&(A-BK')^T\overline{P}(A-BK') - P + Q + (K')^T R^u (K') \\
&= \left[(A-BK')^T\overline{P}(A-BK') - (A-BK)^T\overline{P}(A-BK)\right] + (K')^T R^u (K') - K^T R^u K \\
&\quad + (A-BK)^T\overline{P}(A-BK) - P + Q + K^T R^u K.
\end{aligned}
$$

One can observe that for a small β, the above equation is in $-\beta I + o(\beta)$. Hence, for a sufficiently small β, we have $(A-BK')^T\overline{P}(A-BK') - P + Q + (K')^T R^u (K') < 0$. From the equivalent conditions suggested by the Bounded Real Lemma, the one-step natural policy gradient update, K', will stay in the feasible set $\mathcal{K}$.

With the regularization property above, we focus on proving the global convergence property. According to Lemma 5.1 in [12] and the stepsize chosen to satisfy (3.24), we have the following upper bound

$$
\begin{aligned}
P_{K',L(K')} - P_{K,L(K)} \leq \sum_{t\geq 0} &\left[(A-BK')^T(\gamma^{-2}R^w - \gamma^{-2}P_{K',L(K')}DD^T)^{-1}\right]^t \left[-2\eta Z_K^T Z_K\right] \\
&\left[(\gamma^{-2}R^w - \gamma^{-2}P_{K',L(K')}DD^T)^{-T}(A-BK')\right]^t \\
\leq\ & 0.
\end{aligned}
$$

Therefore, $P_{K,L(K)}$ is monotonic decreasing and will converge to a matrix $P_{K_\infty,L(K_\infty)}$. Following the proof of [12], a globally sublinear convergence rate can be established. This completes the proof. ■

To conclude, the natural policy gradient algorithm with the stepsize satisfying (3.24) is guaranteed to converge to the global optimum with a globally sublinear rate, while preserving the feasibility of the control gain K along the iterates. According to Corollary 2.3, the convergent point recovers the NE of the zero-sum LQ game, which is also the global optimum solution of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem according to section 2.3. Lastly, we offer the pseudocode for the double-loop natural policy gradient algorithm we have proposed.

Algorithm 1 Double-Loop Natural Policy Gradient for Zero-Sum LQ Game

1: Initialize $K_0 \in \mathcal{K}$.
2: **for** $i = 0, 1, \cdots, N$ **do**
3: $\quad L(K_i) \leftarrow \arg\max_L \mathcal{C}(K_i, L)$,
4: $\quad K_{i+1} = K_i - \alpha \nabla_K \mathcal{C}(K_i, L_i) \Sigma_{K_i, L_i}$.
5: **end for**

3.3 Model-free Natural Policy Gradient Design

In this section, we provide the pseudocode of the model-free natural policy gradient algorithm. In contrast to the one suggested in [12], we use a two-point zeroth-order optimization technique when estimating the gradient information to improve the sample complexity of the algorithm. Particularly, Algorithm 2 estimates the policy gradient $\nabla_K \mathcal{C}(K, L)$ and the correlation matrix $\Sigma_{K,L}$ for any stabilizing (K, L); Algorithm 3 finds an estimate of the maximizer $L(K)$ for a given K; Algorithm 4 describes the updates of K for finding an estimate of K^*. Rigorous analysis of the model-free natural policy gradient algorithm is a topic of the author's ongoing research.

Algorithm 2 Est(L;K): Estimating $\nabla_L \mathcal{C}(K, L)$ and $\Sigma_{K,L}$ at L for given K

1: Input: K, L, number of trajectories N, rollout length $\mathcal{R}$, smooth parameter r, dimension $\widetilde{d} = mn$.
2: **for** $i = 1, \cdots N$ **do**
3: $\quad$ Sample policies $\widehat{L}_{i,0} = L + U_i$ and $\widehat{L}_{i,1} = L - U_i$, with U_i drawn uniformly over matrices with $\|U_i\|_F = r$.
4: $\quad$ For $j = 0, 1$, simulate $(K, \widehat{L}_{i,j})$ for $\mathcal{R}$ steps starting from $x_0 \sim \mathcal{D}$, and collect the empirical estimates $\widehat{\mathcal{C}}_{i,j}$ and $\widehat{\Sigma}_{i,j}$ as:

$$\widehat{\mathcal{C}}_{i,j} = \sum_{t=1}^{\mathcal{R}} c_t\,, \quad \widehat{\Sigma}_{i,j} = \sum_{t=1}^{\mathcal{R}} x_t x_t^\top,$$

$\quad$ where c_t and x_t are the costs and states following this trajectory.
5: **end for**
6: Return the estimates:

$$\widehat{\nabla}_L \mathcal{C}(K, L) = \frac{1}{2N} \sum_{i=1}^{N} \frac{\widetilde{d}}{r^2} (\widehat{\mathcal{C}}_{i,0} - \widehat{\mathcal{C}}_{i,1}) U_i\,, \quad \widehat{\Sigma}_{K,L} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\Sigma}_{i,0}.$$

Algorithm 3 Inner-NPG(K): Model-free updates for estimating $L(K)$

1: Input: K, number of iterations $\mathcal{T}$, initialization L_0 such that (K, L_0) is stabilizing.
2: **for** $\tau = 0, \cdots, \mathcal{T}-1$ **do**
3: Call **Est(L_τ;K)** to obtain the gradient and the correlation matrix estimates:

$$[\widehat{\nabla}_L \mathcal{C}(K, L_\tau), \widehat{\Sigma}_{K,L_\tau}] = \textbf{Est}(L_\tau; K).$$

4: Natural policy gradient update: $L_{\tau+1} = L_\tau + \eta \widehat{\nabla}_L \mathcal{C}(K, L_\tau) \cdot \widehat{\Sigma}_{K,L_\tau}^{-1}$.
5: **end for**
6: Return the iterate $L_{\mathcal{T}}$.

Algorithm 4 Outer-NPG: Model-free updates for estimating K^*

1: Input: K_0, number of trajectories N, number of iterations T, rollout length $\mathcal{R}$, parameter r, dimension $\widetilde{d} = md$.
2: **for** $t = 0, \cdots, T-1$ **do**
3: **for** $i = 1, \cdots N$ **do**
4: Sample two policies $\widehat{K}_{i,0} = K_t + V_i$ and $\widehat{K}_{i,1} = K_t - V_i$, with V_i drawn uniformly over matrices with $\|V_i\|_F = r$.
5: For $j = 0, 1$, Call **Inner-NPG**($\widehat{K}_{i,j}$) to obtain the estimate of $L(\widehat{K}_{i,j})$:

$$\widehat{L(\widehat{K}_{i,j})} = \textbf{Inner-NPG}(\widehat{K}_{i,j}).$$

6: For $j = 0, 1$, Simulate $\big(\widehat{K}_{i,j}, \widehat{L(\widehat{K}_{i,j})}\big)$ for $\mathcal{R}$ steps starting from $x_0 \sim \mathcal{D}$, and collect the empirical estimates $\widehat{C}_{i,j}$ and $\widehat{\Sigma}_{i,j}$ as:

$$\widehat{\mathcal{C}}_{i,j} = \sum_{t=1}^{\mathcal{R}} c_t\,, \quad \widehat{\Sigma}_{i,j} = \sum_{t=1}^{\mathcal{R}} x_t x_t^\top\,,$$

where c_t and x_t are the costs and states following this trajectory.
7: **end for**
8: Obtain the estimates of the gradient and the correlation matrix:

$$\widehat{\nabla}_K \mathcal{C}(K_t, \widehat{L(K_t)}) = \frac{1}{2N} \sum_{i=1}^{N} \frac{\widetilde{d}}{r^2} (\widehat{\mathcal{C}}_{i,0} - \widehat{\mathcal{C}}_{i,1}) V_i\,, \quad \widehat{\Sigma}_{K_t, \widehat{L(K_t)}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\Sigma}_{i,0}.$$

9: Natural policy gradient update: $K_{t+1} = K_t - \alpha \widehat{\nabla}_K \mathcal{C}(K_t, \widehat{L(K_t)}) \cdot \widehat{\Sigma}^{-1}_{K_t, \widehat{L(K_t)}}$.
10: **end for**
11: Return the iterate K_T.

Chapter 4

Numerical Experiments

In this chapter, we provide two cases of simulation results to support the theory stated in the preceding chapters. The specification of the first setting is chosen as

$$A = \begin{bmatrix} 0.99 & 0.01 & 0 \\ 0.01 & 0.98 & 0.01 \\ 0.5 & 0.12 & 0.97 \end{bmatrix}, B = \begin{bmatrix} 1 & 0.1 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0.1 & 1 \end{bmatrix}, Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and $D = I$. We randomly initialize the control gain matrix K_0 such that K_0 lies inside the feasible region $\mathcal{K}$, but near the boundary. The value of γ for the $\mathcal{H}_\infty$-norm constraint is chosen to be $1.0541 \cdot \|\mathcal{T}_{K_0}\| = 5.5908$. Then, we run the natural policy gradient algorithm 1 with stepsize chosen to be $\eta = 1 \times 10^{-3}$. The convergences of the objective function $\mathcal{J}$ and the gradient norm square are shown in Figure 4.1.

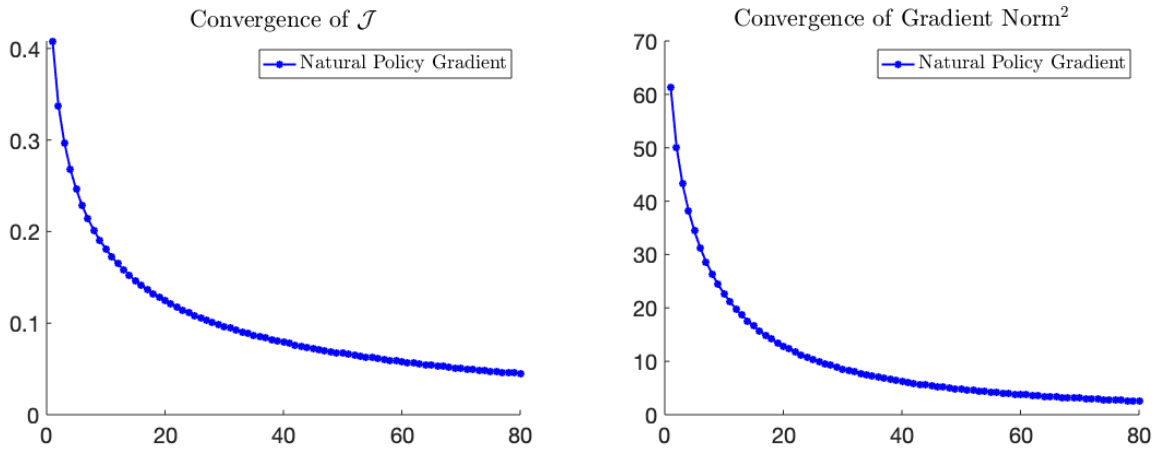

Figure 4.1: Left: Convergence of the objective function $\mathcal{J}(K)$ for the first case. Right: Convergence of gradient norm square for the first case.

The second setting we simulated is described by the matrices

$$A = \begin{bmatrix} 1 & 0 & -10 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, R = \begin{bmatrix} 5 & -3 & 0 \\ -3 & 5 & -2 \\ 0 & -2 & 5 \end{bmatrix},$$

and $D = I$. Similarly, the initial control gain matrix K_0 is randomly generated such that K_0 lies inside the feasible region $\mathcal{K}$, but near the boundary. The value of γ for the $\mathcal{H}_\infty$-norm constraint is chosen to be $1.0541 \cdot \|\mathcal{T}_{K_0}\| = 16.2855$. Then, we run the natural policy gradient algorithm 1 with stepsize chosen to be $\eta = 6 \times 10^{-6}$. In Figure 4.2, the desired convergences of objective function $\mathcal{J}$ and the gradient norm square are shown.

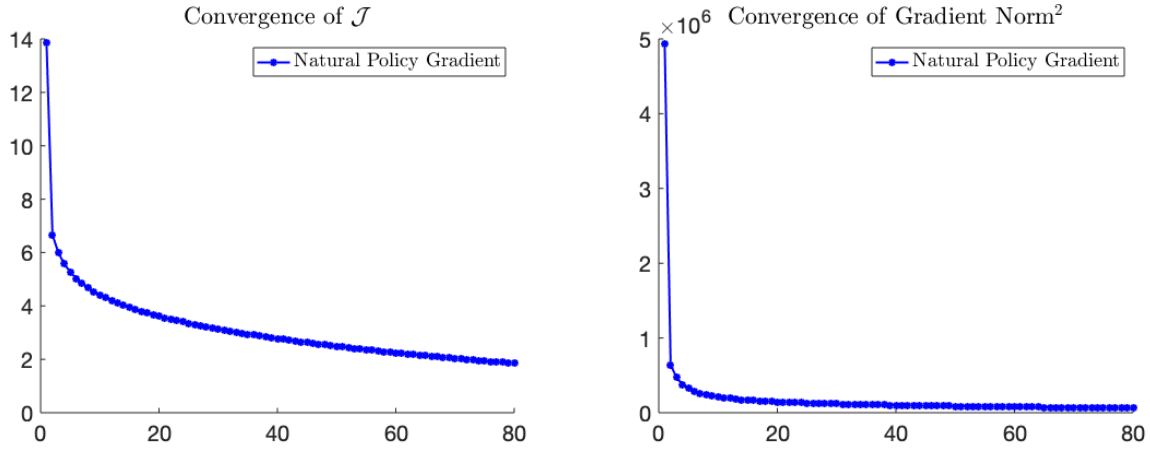

Figure 4.2: Left: Convergence of the objective function $\mathcal{J}(K)$ for the second case. Right: Convergence of gradient norm square for the second case.

Lastly, we show that for both of the settings we presented, the $\mathcal{H}_\infty$-norm constraints are implicitly regularized, in Figure 4.3. That is, for K_0 near but within the boundary of the $\mathcal{H}_\infty$-norm constraint, natural policy gradient updates with the stepsize chosen according to Chapter 3, will not violate robustness constraint along the updates.

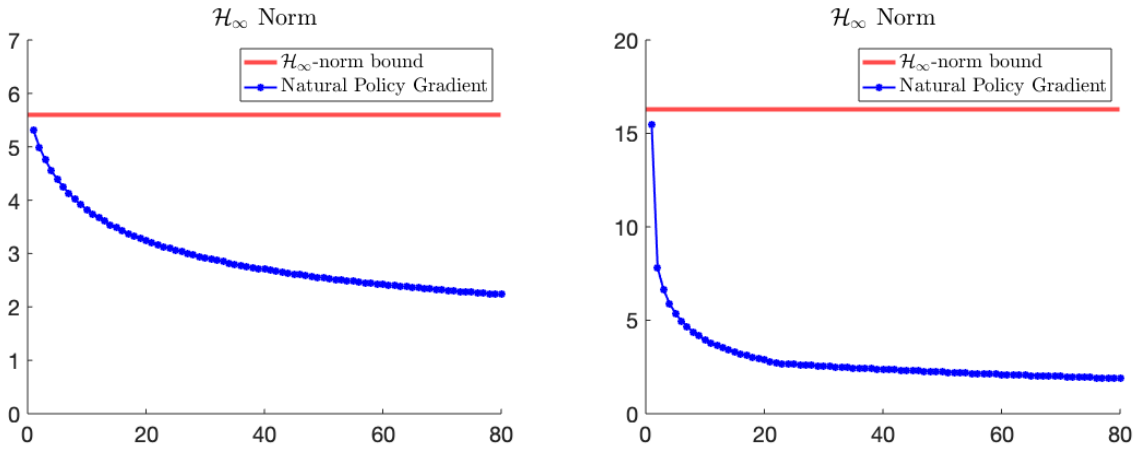

Figure 4.3: Left: $\mathcal{H}_\infty$-norm $\|\mathcal{T}(K)\|_\infty$ along the natural policy gradient updates for the first case, with the stepsize chosen to be $\eta = 1 \times 10^{-3}$. Right: $\mathcal{H}_\infty$-norm $\|\mathcal{T}(K)\|_\infty$ along the natural policy gradient updates for the second case with the stepsize chosen to be $\eta = 6 \times 10^{-6}$.

Chapter 5

Conclusion

In this thesis, we have demonstrated the close connection between the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem and the infinite-horizon zero-sum dynamic LQ game. Based on this connection, we proposed a double-loop natural policy gradient algorithm to solve the NE of the zero-sum LQ game, which recovers the global optimum of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ design problem. Despite the non-convex and non-coercive optimization landscape, we established that for carefully chosen stepsizes, both the inner loop and the outer loop updates of our algorithm converge with desired rates. Moreover, we showed that the stability and robustness along the updates are also regularized. The double-loop PO method we proposed facilitates the development of model-free PO methods for the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problem. An algorithmic sketch for the model-free natural policy gradient algorithm was also provided. This serves as a starting point for the author's future work.

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[4] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[6] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, pp. 1057–1063, 2000.

[7] S. M. Kakade, "A natural policy gradient," in *Advances in Neural Information Processing Systems*, pp. 1531–1538, 2002.

[8] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems*, pp. 1008–1014, 2000.

[9] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor-critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[10] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, pp. 1889–1897, 2015.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] K. Zhang, B. Hu, and T. Başar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," *arXiv preprint arXiv:1910.09496*, 2019.

[13] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*, pp. 1467–1476, 2018.

[14] K. Zhang, A. Koppel, H. Zhu, and T. Başar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *arXiv preprint arXiv:1906.08383*, 2019.

[15] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.

[16] N. Agarwal, E. Hazan, and K. Singh, "Logarithmic regret for online control," *arXiv preprint arXiv:1909.05062*, 2019.

[17] T. Chen, K. Zhang, G. B. Giannakis, and T. Başar, "Communication-efficient distributed reinforcement learning," *arXiv preprint arXiv:1812.03239*, 2018.

[18] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*, pp. 5872–5881, 2018.

[19] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," *arXiv preprint arXiv:1812.03565*, 2018.

[20] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, "LQR through the lens of first order methods: Discrete-time case," *arXiv preprint arXiv:1907.08921*, 2019.

[21] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *International Conference on Artificial Intelligence and Statistics*, pp. 2916–2925, 2019.

[22] K. Zhang, Z. Yang, and T. Başar, "Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games," in *Advances in Neural Information Processing Systems*, pp. 11598–11610, 2019.

[23] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *IEEE International Conference on Robotics and Automation*, pp. 2619–2624, 2004.

[24] S. Yang, H. Kumar, Z. Gu, X. Zhang, M. Travers, and H. Choset, "State estimation for legged robots using contact-centric leg odometry," *arXiv preprint arXiv:1911.05176*, 2019.

[25] J. Bu, L. J. Ratliff, and M. Mesbahi, "Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games," *arXiv preprint arXiv:1911.04672*, 2019.

[26] D. S. Bernstein and W. M. Haddad, "LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 293–305, 1989.

[27] P. P. Khargonekar and M. A. Rotea, "Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control: A convex optimization approach," *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 824–837, 1991.

[28] D. Mustafa and D. S. Bernstein, "LQG cost bounds in discrete-time $\mathcal{H}_2/\mathcal{H}_\infty$ control," *Transactions of the Institute of Measurement and Control*, vol. 13, no. 5, pp. 269–275, 1991.

[29] I. Kaminer, P. P. Khargonekar, and M. A. Rotea, "Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete-time systems via convex optimization," *Automatica*, vol. 29, no. 1, pp. 57–70, 1993.

[30] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems," *IEEE Transactions on Automatic Control*, vol. 34, pp. 831–847, 1989.

[31] P. Apkarian, D. Noll, and A. Rondepierre, "Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control via nonsmooth optimization," *SIAM Journal on Control and Optimization*, vol. 47, no. 3, pp. 1516–1546, 2008.

[32] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 124–131, 1973.

[33] V. S. Borkar and S. P. Meyn, "Risk-sensitive optimal control for Markov decision processes with monotone cost," *Mathematics of Operations Research*, vol. 27, no. 1, pp. 192–209, 2002.

[34] K. Glover and J. C. Doyle, "State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to relations to risk sensitivity," *Systems & Control Letters*, vol. 11, no. 3, pp. 167–172, 1988.

[35] D. Mustafa, "Relations between maximum-entropy/$\mathcal{H}_\infty$ control and combined $\mathcal{H}_\infty$/LQG control," *Systems & Control Letters*, vol. 12, no. 3, pp. 193–203, 1989.

[36] T. Başar and P. Bernhard, *$\mathcal{H}_\infty$-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer Science & Business Media, 2008.

[37] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv preprint arXiv:1911.10635*, 2019.

[38] X. Zhang, K. Zhang, E. Miehling, and T. Basar, "Non-cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 9482–9493, 2019.

[39] X. Sha, J. Zhang, K. Zhang, K. You, and T. Başar, "Asynchronous policy evaluation in distributed reinforcement learning over networks," *arXiv preprint arXiv:2003.00433*, 2020.

[40] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem," *arXiv preprint arXiv:1912.11899*, 2019.

[41] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural Computation*, vol. 17, no. 2, pp. 335–359, 2005.

[42] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*, pp. 2817–2826, 2017.

[43] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017.

[44] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity," *IEEE Transactions on Automatic Control*, vol. 11, no. 2, pp. 228–238, 1966.

[45] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall New Jersey, 1996.