On Approximation and Optimal Control of Nonnormal Distributed Parameter Systems

Eric D. Vugrin

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

John Burns, Chair
Jeff Borggaard
Gene Cliff
Terry Herdman
Robert Rogers

April 26, 2004
Blacksburg, Virginia

Keywords: Linear Quadratic Regulator Problems, Nonnormal, Convection-Diffusion Equation

On Approximation and Optimal Control of Nonnormal Distributed Parameter Systems

Eric D. Vugrin

ABSTRACT

For more than 100 years, the Navier-Stokes equations and various linearizations have been used as a model to study fluid dynamics. Recently, attention has been directed toward studying the nonnormality of linearized problems and developing convergent numerical schemes for simulation of these sytems. Numerical schemes for optimal control problems often require additional properties that may not be necessary for simulation; these properties can be critical when studying nonnormal problems. This research is concerned with approximating infinite dimensional optimal control problems with nonnormal system operators. We examine three different finite element methods for a specific convection-diffusion equation and prove convergence of the infinitesimal generators. Additionally, for two of these schemes, we prove convergence of the associated feedback gains. We apply these three schemes to control problems and compare the performance of all three methods.

This work was funded in part by the Air Force Office of Scientific Research under grant number F49620-03-1-0243.

Acknowledgments

I would like to thank my advisor, John Burns, for his guidance, encouragement, and expertise during my time as a student at Virginia Tech. I would also like to thank my committee members Jeff Borggaard, Gene Cliff, Terry Herdman, and Bob Rogers for all of their help. Completing the last four years of school and this dissertation has truly been a team effort by all of my committee members, and I am grateful to all of them. Additionally, thanks are in order for all of the people in the Mathematics Department and ICAM who have made my time at Virginia Tech very enjoyable. Finally, I thank my wife, Kay, for her never-ending support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Fluid dynamics has been the subject of mathematical research for almost 200 years. The emergence of high performance computing has provided the foundation for many of the recent advances in computational fluid dynamics. However, most of these advances are limited to simulation, and much remains to be done in the area of computational methods for optimal design and control of fluid flows.

Classical linear stability analysis applies eigenvalue analysis to evaluate the stability of flows [15]. Although this classical method has been useful in certain cases, it is known to fail for many flows. For example, the linearized Navier-Stokes equations for channel flow have been analyzed using these techniques. The application of classical linear theory implies that plane Poiseuille flow is stable if the Reynolds number is less than 5772 [41]. Also, the same analysis suggests that plane Couette flow should be stable for all Reynolds numbers [29]. However, experiments have shown that Poiseuille flow becomes unstable when the Reynolds number is approximately 1000 [42], and Couette flow can become unstable at a Reynolds number as low as 360 [40], [48]. Thus, classical linear eigenvalue analysis appears to be insufficient to describe and explain the transition to unstable flows. Non-linear theories attempting to explain this transition have been developed and have had limited success in matching experimental results [15].

Recently, non-classical linear methods (based on concepts from robust control theory) have been applied to the linearized Navier-Stokes equations with increasing success [2], [14], [17], [27], [32], [31], [33], [44], [50]. It is well known that the linearized Navier-Stokes equations can lead to nonnormal linear operators. Baggett, Driscoll, and Trefethen hypothesize in [1] that the nonnormality of the system leads to transient growth and, coupled with the non-linearity, small solutions can lead to large amplitude solutions. Several investigations indicate that the nonnormality of the system may be a crucial mechanism in the transition to instability [2], [44], [46], [50]. Since the linearized equation is highly nonnormal, it can be highly sensitive to parameter variations and inputs. Numerical simulations supporting this proposal can be found in [2] and [28].

Consistency and stability of an approximating system is of fundamental importance when developing numerical techniques to simulate the dynamics of a system. In particular, the Lax Equivalence Theorem and the Trotter-Kato Theorem establish that convergence is equivalent to consistency and stability. However, in order to develop good numerical schemes for certain optimal control problems, one needs to consider additional requirements for the approximation schemes. Dual convergence and preservation of exponential stabilizability (POES) are additional properties that must be considered when developing approximation schemes for optimization-based control design techniques. When approximating a normal system, dual convergence is automatic for a convergent approximation scheme. However, for non-normal systems, this may not be the case.

Differential-delay equations provide simple examples of highly nonnormal systems. In [4], [5], [30], and [34], numerical schemes are developed for control and identification of differential-delay equations. In [8], numerical results demonstrate that the feedback gain operators do not converge uniformly for the so-called Banks-Kappel finite element scheme. However, the averaging scheme from [4] and [34] produces gains that converge uniformly. One problem with the spline scheme in [5] is that the approximation does not satisfy the property of dual convergence. Thus, dual convergence can have a major impact on convergence in optimal control problems.

In [23], [24], and [25], Gibson uses dual convergence to establish a framework for developing approximation schemes for regulator problem theory. Although Jetto et al. have implied that dual convergence is not necessary for convergence [45], the recent paper [11] suggests that dual convergence is necessary.

This dissertation is concerned with approximation techniques for nonnormal distributed parameter systems. Our goal is to develop methods that will ensure uniform convergence of the feedback gain operators. Although we focus on linear quadratic regulator problems, these results can be extended to linear quadratic Gaussian and Min-Max optimal control problems.

1.1 Notation

Let $\Omega \subseteq \mathbb{R}^3$ be a connected, open set. We always assume that Ω is bounded in at least one direction. That is, for $i = 1, 2$, or 3, there exists a positive, finite constant M such that $|x_i| < M$ for any $(x_1, x_2, x_3) \in \Omega$. We denote the usual space of Lebesgue square integrable functions defined on Ω by $L^2(\Omega)$. This space is a Hilbert space when endowed with the standard $L^2(\Omega)$ inner product

$$\langle f, g \rangle = \int_\Omega f(x)g(x)dx, \quad \forall f, g \in L^2(\Omega).$$

Unless otherwise noted, we use $\langle\cdot,\cdot\rangle$ to denote the $L^2(\Omega)$ inner product and $\|\cdot\|$ to denote the $L^2(\Omega)$ norm. We denote the space of infinitely differentiable functions on Ω by $C^\infty(\Omega)$. The subset of these functions that have compact support in Ω is denoted by $C_0^\infty(\Omega)$. If α is the multi-index $\alpha = (\alpha_1, ..., \alpha_N)$ such that α_i are positive integers, then

$$|\alpha| = \sum_{i=1}^{N} \alpha_i.$$

For u in $C_0^\infty(\Omega)$,

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} ... \partial x_N^{\alpha_N}}$$

denotes the usual differentiation operator.

For nonnegative integers m and p such that $1 \leq p < \infty$, $W^{m,p}(\Omega)$ denotes the standard Sobolev space

$$W^{m,p}(\Omega) = \{f \in L^p(\Omega) | \partial^\alpha f \in L^p(\Omega) \ \ \forall \ \ |\alpha| \leq m\},$$

with norm

$$\|u\|_{m,p} = \left(\sum_{|\alpha| \leq m} \int_\Omega |\partial^\alpha u(x)|^p \, dx \right)^{1/p}, \quad 1 \leq p < \infty.$$

When $p = 2$, $W^{m,2}(\Omega)$ is a Hilbert space with the inner product

$$\langle f, g\rangle_m = \sum_{|\alpha| \leq m} \int_\Omega \partial^\alpha f(x) \partial^\alpha g(x) dx$$

and is denoted by $H^m(\Omega)$. It is common to drop the subscript 2 and denote the norm by $\|\cdot\|_m$.

The closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$ is a Banach space, and the Hilbert space $W_0^{m,2}(\Omega)$ is denoted by $H_0^m(\Omega)$. The following result can be found in [26].

Corollary 1.1.1 *When Ω is connected and bounded in at least one direction, then for each integer $m \geq 0$, there exists a constant $K = K(m, \Omega) > 0$ such that*

$$\|u\|_m \leq K \, |u|_m = K \left(\sum_{|\alpha| = m} \int_\Omega |\partial^\alpha u(x)|^p \, dx \right)^{1/p} \quad \textit{for all} \ \ u \in H_0^m(\Omega).$$

Thus, $|\cdot|_m$ is an equivalent norm to $\|\cdot\|_m$ on $H_0^m(\Omega)$.

Finally, if X and Y are Banach spaces, we introduce the notation $\mathcal{L}(X, Y)$ to denote the space of bounded linear operators from X into Y. The space $\mathcal{L}(X, Y)$ is a Banach space with the norm

$$\|L\|_{\mathcal{L}(X,Y)} = \sup_{x \in X, \ \ \|x\|_X = 1} \|Lx\|_Y .$$

The space of bounded linear operators from X into X is denoted by $\mathcal{L}(X)$.

1.2 General Theory For The Infinite Dimensional Linear Quadratic Regulator Problem

Let H and U be Hilbert spaces endowed with the inner products $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_U$, respectively. These inner products induce the norms

$$\|v\|_H^2 = \langle v, v\rangle_H\,, \quad \text{for all} \quad v \in H, \tag{1.1}$$

$$\|u\|_U^2 = \langle u, u\rangle_U\,, \quad \text{for all} \quad u \in U. \tag{1.2}$$

Let $\mathcal{Q} : H \to H$ and $\mathcal{R} : U \to U$ be bounded linear operators with the following properties:

(i) The operator $\mathcal{R}$ is a positive definite operator on U.

(ii) The operator $\mathcal{Q}$ is a positive semidefinite operator on H.

Consider the following optimal control problem **(LQR)**:

Find $u \in L^2(0,\infty;U)$ such that u minimizes

$$\mathcal{J}(z_0, u) = \int_0^\infty \{\langle \mathcal{Q}z(t), z(t)\rangle_H + \langle \mathcal{R}u(t), u(t)\rangle_U\}\, dt, \tag{1.3}$$

subject to

$$\frac{dz(t)}{dt} = \mathcal{A}_0 z(t) + \mathcal{B}u(t) \quad \text{on} \quad H, \tag{1.4}$$

$$z(0) = z_0 \text{ in } \; H. \tag{1.5}$$

In general, this is a problem where the state evolves in an infinite dimensional space. The theory for existence and uniqueness of an optimal control $\hat{u}$ for this class of problems is well established in [10]. For a more general discussion, see [37], [38], [39].

In order to discuss whether a solution to **(LQR)** exists, we need to introduce some definitions about the dynamic equation (1.4) and the cost function (1.3). The following definitions concerning semigroups are found in [43].

Definition 1.2.1 *Let X be a Banach space. A one parameter family $T(t)$, $0 \le t < \infty$, of bounded linear operators from X into X is a semigroup of bounded linear operators on X if*

i. $T(0) = I$, *where I denotes the identity operator on X, and*

ii. $T(t+s) = T(t)T(s)$ *for every* $t, s \ge 0$.

Furthermore, the linear operator A defined on the domain

$$\mathcal{D}(A) = \left\{ x \in X : \lim_{t \downarrow 0} \frac{T(t)x - x}{t} \quad \textit{exists} \right\}$$

by

$$Ax = \lim_{t \downarrow 0} \frac{T(t)x - x}{t}, \quad x \in \mathcal{D}(A)$$

is defined to be the infinitesimal generator of the semigroup $T(t)$.

Definition 1.2.2 *Let X be a Banach space, and let $T(t)$ be a semigroup of bounded linear operators on X. $T(t)$ is called uniformly continuous if*

$$\lim_{t \downarrow 0} \|T(t) - I\|_{\mathcal{L}(X)} = 0.$$

Definition 1.2.3 *Let H be a Hilbert space. The infinitesimal generator A is called dissipative if*

$$Re\left(\langle Ax, x\rangle_H\right) \leq 0, \quad \textit{for all} \;\; x \in \mathcal{D}(A).$$

The operator A is maximally dissipative if both A and A^* are dissipative.

Definition 1.2.4 *A semigroup $T(t)$, $0 \leq t < \infty$, of bounded linear operators on the Banach space X is strongly continuous on X if*

$$\lim_{t \downarrow 0} T(t)x = x \textit{ for all } x \in X.$$

A strongly continuous semigroup is called a C_0-semigroup.

Definition 1.2.5 *A C_0-semigroup $T(t)$, $\;0 \leq t < \infty$, of bounded linear operators on the Banach space X is exponentially stable if there exist constants $M \geq 1$ and $\omega > 0$ such that*

$$\|T(t)x\|_X \leq Me^{-wt} \|x\|_X, \quad \textit{for all} \;\; x \in X, t \geq 0.$$

We frequently use the following theorems from [43] to show that an operator generates a semigroup.

Theorem 1.2.1 *A linear operator A is the infinitesimal generator of a uniformly continuous semigroup if and only if A is a bounded linear operator.*

Theorem 1.2.2 *Let H be a Hilbert space and assume A is a densely defined, closed linear operator on H. If both A and A^* are dissipative, then A is the infinitesimal generator of a C_0-semigroup of contractions on H. That is, if $S(t)$ is the C_0-semigroup generated by A, then $\|S(t)\|_{\mathcal{L}(H)} \leq 1$.*

The following definition may be found in [10], [37], [38], [39].

Definition 1.2.6 *The pair* $(\mathcal{A}_0, \mathcal{B})$ *in* ***(LQR)*** *is said to be stabilizable if there exists a bounded linear operator* $\mathcal{K} : H \to U$ *such that* $\mathcal{A}_0 - \mathcal{B}\mathcal{K}$ *generates an exponentially stable semigroup on* H.

We are now ready to introduce a significant result found in [23].

Theorem 1.2.3 *Assume that* $\mathcal{A}_0$ *generates a* C_0-*semigroup on* H, *and* $\mathcal{B}$ *is a bounded linear operator from* U *into* H. *If for each* $z_0 \in H$, *there exists* $u \in L^2((0,\infty);U)$ *that drives the state of* ***(LQR)*** *to zero asymptotically so that* $\mathcal{J}(z_0, u) < \infty$, *then there exists a unique optimal pair* $(\hat{u}, \hat{z})$ *for* ***(LQR)***. *Furthermore,*

$$\hat{u}(t) = -\mathcal{R}^{-1}\mathcal{B}^*\Pi\hat{z}(t), \quad 0 < t < \infty,$$

where Π *is the self-adjoint, non-negative definite solution in* $\mathcal{L}(H)$ *of the Algebraic Riccati equation (ARE)*

$$\langle \Pi x, \mathcal{A}_0 y \rangle_H + \langle \Pi \mathcal{A}_0 x, y \rangle_H + \langle \mathcal{Q}x, y \rangle_H - \left\langle \mathcal{R}^{-1}\mathcal{B}^*\Pi x, \mathcal{B}^*\Pi y \right\rangle_U = 0 \quad \textit{for all} \quad x, y \in \mathcal{D}(A_0). \tag{1.6}$$

Additionally,

$$\mathcal{J}(z_0, \hat{u}) = \langle \Pi z_0, z_0 \rangle_H .$$

For future reference, when we write that Π solves the ARE

$$\mathcal{A}_0^*\Pi + \Pi\mathcal{A}_0 - \Pi\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\Pi + \mathcal{Q} = 0,$$

we mean that Π is a solution of (1.6).

1.3 The Approximate Linear Quadratic Regulator Problem

The goal of this section is to generate an approximating, finite dimensional, problem $(\mathbf{LQR}_N)$ such that the optimal control for the approximate problem $(\mathbf{LQR}_N)$ converges to the optimal control of the infinite dimensional problem **(LQR)**. In addition, we focus on the convergence of the resulting feedback operators. We begin by reviewing some standard notation and results found in [7] and [9].

Let V be a subspace of H such that

$$\mathcal{D}(\mathcal{A}_0) \subseteq V \subseteq H = H' \subseteq V' \subseteq \mathcal{D}(\mathcal{A}_0)'.$$

Furthermore, let V be continuously imbedded in H.

Definition 1.3.1 *A mapping $a : V \times V \to \mathbb{R}$ is called a bilinear form if it is linear in each variable. That is, $a(\cdot,\cdot)$ satisfies*

$$a(x, \alpha y + \beta z) = \alpha a(x,y) + \beta a(x,z) \tag{1.7}$$

and

$$a(\alpha y + \beta z, x) = \alpha a(y,x) + \beta a(z,x) \tag{1.8}$$

for all x, y, z in V and α, β in $\mathbb{R}$.

Definition 1.3.2 *The bilinear form $a(\cdot,\cdot)$ is continuous if the following condition holds: There exists a positive constant c_1 such that*

$$a(x,y) \le c_1 \|x\|_V \|y\|_V \quad \text{for all} \quad x, y \in V. \tag{1.9}$$

Furthermore, if $a(\cdot,\cdot)$ is continuous and there exists a positive constant c_2 such that

$$|a(x,x)| \ge c_2 \|x\|_V^2 \quad \text{for all} \quad x \in V, \tag{1.10}$$

we say that $a(\cdot,\cdot)$ is $V-$elliptic.

If $a : V \times V \to \mathbb{R}$ is $V-$elliptic, then there exists a unique continuous linear mapping $L : V \to V'$ such that

$$a(x,y) = \langle Lx, y\rangle_{V' \times V}, \quad \text{for all} \quad x, y \in V.$$

This statement is a direct result of the Lax-Milgram Lemma [26].

Lemma 1.3.1 (Lax-Milgram) *Let V be a Hilbert space and let*

$$B : V \times V \to \mathbb{R}$$

be a continuous, $V-$elliptic bilinear mapping. Then for every f in V', there exists a unique $y \in V$ such that

$$B(x,y) = f(x)$$

for all $x \in V$.

We now outline a standard procedure for approximating the infinite dimensional optimal control problem **(LQR)**. Assume that **(LQR)** satisfies the hypotheses of Theorem 1.2.3. Introduce a bilinear form $a : V \times V \to \mathbb{R}$ such that

$$a(x,y) = -\langle \mathcal{A}_0 x, y\rangle_H, \quad \forall x \in \mathcal{D}(\mathcal{A}_0), y \in V. \tag{1.11}$$

We restrict ourselves to the case where a is $V-$elliptic. Thus by the Lax-Milgram Lemma, there exists an operator $\mathcal{A} : V \to V'$ such that

$$a(x,y) = -\langle \mathcal{A}x, y\rangle_{V' \times V}, \quad \forall x, y \in V. \tag{1.12}$$

For a discussion of how to proceed when a is not $V-$elliptic, see [7].

Since $\mathcal{B} : U \to H \subseteq V'$, the weak form of equation (1.4) is given by

$$\frac{dz(t)}{dt} = \mathcal{A}z(t) + \mathcal{B}u(t) \text{ on } V'. \tag{1.13}$$

1.4 Approximations

The next step in the approximation procedure is to choose a finite dimensional subspace $V_N \subseteq V$. Since V_N is finite dimensional, it is the span of a finite number of basis functions in V.

We now define operators that approximate $\mathcal{A}, \mathcal{B}, \mathcal{Q}$, and $\mathcal{R}$. The following properties for the approximate operators $\mathcal{A}_N, \mathcal{B}_N, \mathcal{Q}_N$, and $\mathcal{R}_N$ are important in establishing convergence of feedback control laws.

1.) The operator $\mathcal{A}_N : V_N \to V_N$ generates a C_0-semigroup on V_N.

2.) The operator $\mathcal{B}_N : U \to V_N$ is bounded.

3.) The operator $\mathcal{Q}_N : V_N \to V_N$ is bounded, self-adjoint, and positive semi-definite.

4.) The operator $\mathcal{R}_N : U \to U$ is bounded and positive definite.

Let $\mathcal{P}_N$ denote the orthogonal projection from H into V_N. By restricting $a(\cdot, \cdot)$ to V_N, we define the operator $\mathcal{A}_N : V_N \to V_N$ such that

$$a(x, y) = \langle\!\langle \mathcal{A}_N x, y \rangle\!\rangle , \quad \text{for all } x, y \in V_N.$$

In order to define the operators $\mathcal{B}_N, \mathcal{Q}_N$, and $\mathcal{R}_N$, we utilize the projection operator $\mathcal{P}_N$. Thus,

$$\mathcal{B}_N : U \to V_N, \quad \mathcal{B}_N = \mathcal{P}_N \mathcal{B},$$

and

$$\mathcal{Q}_N : V_N \to V_N, \quad \mathcal{Q}_N = \mathcal{P}_N \mathcal{Q}.$$

We consider problems where $U = \mathbb{R}^m$ and set $\mathcal{R}_N = \mathcal{R}$. Based on the previous notation, we define an approximate problem $(\mathbf{LQR}_N)$.

$(\mathbf{LQR}_N)$ *Find* $u_N \in L^2((0, \infty); U)$ *so that* u_N *minimizes*

$$J_N(z_{0,N}, u_N(\cdot)) = \int_0^\infty \{< \mathcal{Q}_N z_N(t), z_N(t) >_H + < \mathcal{R}_N u_N(t), u_N(t) >_U\} dt \tag{1.14}$$

subject to

$$\frac{dz_N(t)}{dt} = \mathcal{A}_N z_N(t) + \mathcal{B}_N u_N(t), \tag{1.15}$$
$$z_N(0) = z_{0,N}. \tag{1.16}$$

If the system $(\mathbf{LQR}_N)$ satisfies the hypotheses of Theorem 1.2.3, we find the unique positive definite solution Π_N of the finite dimensional ARE

$$\mathcal{A}_N^*\Pi_N + \Pi_N\mathcal{A}_N - \Pi_N\mathcal{B}_N\mathcal{R}_N^{-1}\mathcal{B}_N^*\Pi_N + \mathcal{Q}_N = 0, \tag{1.17}$$

and the optimal control $\hat{u}$ is defined by

$$\hat{u}_N(t) = -\mathcal{K}_N\hat{z}_N(t) = -\mathcal{R}_N^{-1}\mathcal{B}_N^*\Pi_N z(t).$$

1.4.1 Convergence of the Riccati Operators

In this section, we provide the framework for proving the convergence of the Riccati operators Π_N to Π.

Let $H = L^2(0,1)$, $V = H_0^1(0,1)$, and $U = \mathbb{R}^m$. Let $\mathcal{A}, \mathcal{B}, \mathcal{Q}, \mathcal{R}$ satisfy the hypotheses in Theorem 1.2.3, and consider the following equation

$$\frac{dy(t)}{dt} = \mathcal{A}y(t) + \mathcal{B}u(t) \quad \text{on} \quad H \tag{1.18}$$
$$y(0) = y_0 \in H \tag{1.19}$$

together with the associated performance measure

$$\mathcal{J}(y_0, u) = \int_0^\infty \{\langle\!\langle \mathcal{Q}y(t), y(t)\rangle\!\rangle_H + \langle\!\langle \mathcal{R}u(t), u(t)\rangle\!\rangle_U\}dt. \tag{1.20}$$

Consider the following optimal control problem $(\mathbf{\Gamma})$:

Minimize $\mathcal{J}(y_0, u)$ over $u \in L^2((0,\infty);U)$ subject to (1.18) and (1.19).

The problem $(\mathbf{\Gamma})$ is approximated by a sequence of approximate problems. Let V_N, $N = 1, 2, \ldots$ be a sequence of finite dimensional subspaces of $V \subseteq H$ and $\mathcal{P}_N$ be the orthogonal projections from H into V_N under the H inner product. Let $\mathcal{S}_N(t)$ be the sequence of C_0-semigroups defined on V_N with infinitesimal generators $\mathcal{A}_N \in \mathcal{L}(V_N)$, and let $\mathcal{S}_N(t)^*$ denote their respective adjoint semigroups. Additionally, let $\mathcal{B}_N \in \mathcal{L}(\mathcal{U}, V_N)$ and $\mathcal{Q}_N \in \mathcal{L}(V_N)$.

We consider the following approximate optimal control problem $(\mathbf{\Gamma^N})$.

($\mathbf{\Gamma^N}$): Minimize

$$\mathcal{J}_N(y_N(0), u_N) = \int_0^\infty \{\langle\!\langle \mathcal{Q}_N y_N(t), y_N(t)\rangle\!\rangle_H + \langle\!\langle \mathcal{R}_N u_N(t), u_N(t)\rangle\!\rangle_U\} dt \tag{1.21}$$

over $u_N \in L^2((0,\infty);U)$ where

$$\frac{dy_N(t)}{dt} = \mathcal{A}_N y^N(t) + \mathcal{B}_N u_N(t) \tag{1.22}$$

$$y_N(0) = \mathcal{P}_N y_0. \tag{1.23}$$

We shall say that a function $u_N \in L^2((0,\infty);U)$ is an admissible control for the initial state $y_N(0)$ if $\mathcal{J}_N(y_N(0), u_N)$ is finite. Banks and Kunisch provide a fundamental theorem in [7] that establishes sufficiency for the convergence of Π_N to Π. They make the following assumptions:

(H1) For each $y_N(0) \in V_N$, there exists a control $u_N \in L^2((0,\infty);U)$ for ($\mathbf{\Gamma^N}$), and any admissable control for ($\mathbf{\Gamma^N}$) drives the state to zero asymptotically.

(H2)
- (i) For each $z \in H$, $\mathcal{S}_N(t)\mathcal{P}_N z \to \mathcal{S}(t)z$ with convergence uniform in t on bounded subsets of $[0,\infty)$.
- (ii) For each $z \in H$, $\mathcal{S}_N^*(t)\mathcal{P}_N z \to \mathcal{S}^*(t)z$ with convergence uniform in t on bounded subsets of $[0,\infty)$.
- (iii) For each $u \in U$, $\mathcal{B}_N u \to \mathcal{B}u$ and for each $z \in H$, $\mathcal{B}_N^* z \to \mathcal{B}^* z$.
- (iv) For each $z \in H$, $\mathcal{Q}_N \mathcal{P}_N z \to \mathcal{Q}z$.

(C1) For each $z \in V$, there exists an element $z_N \in V_N$ such that $\|z - z_N\|_V \leq \epsilon(N)$, where $\epsilon(N) \to 0$ as $N \to \infty$.

(C2) The pair $(\mathcal{A}, \mathcal{B})$ is stabilizable.

(C3) **(POES)** There exists a bounded linear operator $\mathcal{K} : H \to U$ that stabilizes $(\mathcal{A}, \mathcal{B})$. Furthermore, there exists an integer N_0 such that for all $N \geq N_0$, the pairs $(\mathcal{A}_N, \mathcal{B}_N)$ are stabilized by $\mathcal{K}$.

The following result may be found in [7].

Theorem 1.4.1 (Banks-Kunisch) *Suppose (H1), (H2), and (C1)-(C3) hold, and that* $\mathcal{R} > 0, \mathcal{Q} > 0, \mathcal{Q}_N \geq 0$. *Then there exist unique Riccati operators* Π *and* Π_N *associated with* ($\mathbf{\Gamma}$) *and* ($\mathbf{\Gamma^N}$) *on* $\mathcal{H}$ *and* V_N, *respectively, such that if* $\mathcal{P}_N$ *denotes the orthogonal projection into* V_N, *then*

$$\Pi_N \mathcal{P}_N z \to \Pi z, \forall z \in \mathcal{H} \tag{1.24}$$

$$T_N(t)\mathcal{P}_N z \to T(t)z, \quad \forall z \in \mathcal{H} \tag{1.25}$$

$$\tag{1.26}$$

and

$$\hat{u}^N(t) \to \hat{u}(t), \tag{1.27}$$

with these last two statements holding uniformly in t on compact subsets of $[0,\infty)$. Here, $T_N(t)$, and $T(t)$ are the semigroups generated by $\mathcal{A}_N - \mathcal{P}_N\mathcal{B}_N\mathcal{R}^{-1}\mathcal{B}_N^*\mathcal{P}_N\Pi_N$ and $\mathcal{A}-\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\Pi$, and $\hat{u}(t)$ and $\hat{u}_N(t)$ are the optimal feedback controls for $(\boldsymbol{\Gamma})$ and $(\boldsymbol{\Gamma^N})$.

Π_N is the positive definite Riccati operator that solves the ARE

$$\mathcal{A}_N^*\Pi_N + \Pi_N\mathcal{A}_N - \Pi_N\mathcal{B}_N\mathcal{R}^{-1}\mathcal{B}_N^*\Pi_N + \mathcal{Q}_N = 0. \tag{1.28}$$

1.4.2 Issues Approximating Nonnormal Operators

Our goal when constructing $(\mathbf{LQR}_N)$ is that the feedback gain $\mathcal{K}_N = \mathcal{R}^{-1}\mathcal{B}_N^*\mathcal{P}_N\Pi_N$ converges to $\mathcal{K} = \mathcal{R}^{-1}\mathcal{B}^*\Pi$, the resulting feedback gain for **(LQR)**. Specifically, we want

$$\|\mathcal{K} - \mathcal{K}_N\|_{\mathcal{L}(H,U)} \to 0. \tag{1.29}$$

To facilitate discussion of convergence of the feedback gains, we state the Trotter-Kato Theorem from [43].

Theorem 1.4.2 *Let $X_n, n = 1,2,\ldots$ denote a set of finite dimensional subspaces of the Banach space X such that $X_n \subseteq X_{n+1} \subseteq \ldots \subseteq X$. Moreover, if $\mathcal{P}_n$ denotes the sequence of orthogonal projections from X into X_n, we assume that $\|\mathcal{P}_n\| \leq 1$ and*

$$\|\mathcal{P}_n x - x\| \to 0 \quad \forall x \in X. \tag{1.30}$$

Let $S(t)$ be the C_0-semigroup on X generated by $\mathcal{A}$, and let $S_n(t)$ be the sequence of C_0-semigroups on X_n generated by $\mathcal{A}_n$, respectively. If there exist real constants $M \geq 1$, and ω such that

$$\|S(t)x\|_X \leq Me^{\omega t}\|x\|_X, \quad \textit{for all } \ x \in X$$
$$\|S_n(t)x_n\|_X \leq Me^{\omega t}\|x_n\|_X, \quad \textit{for all } \ x_n \in X_n,$$

and there exists λ_0 with $Re(\lambda_0 > \omega$ and a core D of $(\lambda_0 I - \mathcal{A})$ such that as $n \to \infty$,

$$\mathcal{A}_n x \to \mathcal{A}x, \quad \textit{for all } \ x \in D,$$

then for every $x \in X$ and $t \geq 0$,

$$\|S_n(t)P_n x - S(t)x\|_X \to 0$$

as $n \to \infty$, and the convergence is uniform on bounded $t-$intervals.

The following theorem is equivalent to the Trotter-Kato Theorem [43].

Theorem 1.4.3 *Let $X_n, n = 1, 2, \ldots$ denote a set of finite dimensional subspaces of the Banach space X such that $X_n \subseteq X_{n+1} \subseteq \ldots \subseteq X$. Moreover, if $\mathcal{P}_n$ denotes the sequence of orthogonal projections from X into X_n, we assume that $\|\mathcal{P}_n\| \leq 1$ and*

$$\|\mathcal{P}_n x - x\| \to 0 \quad \forall x \in X. \tag{1.31}$$

Let $S(t)$ be the C_0-semigroup on X generated by $\mathcal{A}$, and let $S_n(t)$ be the sequence of C_0-semigroups on X_n generated by $\mathcal{A}_n$, respectively. If there exist real constants $M \geq 1$ and ω such that

$$\|S(t)x\|_X \leq Me^{\omega t} \|x\|_X, \quad \forall x \in X$$
$$\|S_n(t)x_n\|_X \leq Me^{\omega t} \|x_n\|_X, \quad \forall x_n \in X_n,$$

then the following are equivalent:

(a) For every $x \in X$ and λ with $Re(\lambda) > \omega$,

$$\left\|(\lambda I - \mathcal{A}_n)^{-1}\mathcal{P}_n x - (\lambda I - \mathcal{A})^{-1}x\right\|_X \to 0$$

as $n \to \infty$.

(b) For every $x \in X$ and $t \geq 0$,

$$\|S_n(t)\mathcal{P}_n x - S(t)x\|_X \to 0$$

as $n \to \infty$, and the convergence is uniform on bounded $t-$intervals.

We also use the following Theorem from [43].

Theorem 1.4.4 *Let X be a reflexive Banach space and let $S(t)$ be a C_0-semigroup on X with infinitesimal generator $\mathcal{A}$. The adjoint semigroup $S(t)^*$ is a C_0-semigroup on X^* with infinitesimal generator $\mathcal{A}^*$.*

For future reference, when we say that $\mathcal{A}_N$ converges to $\mathcal{A}$ in the Trotter-Kato sense, we mean that $\mathcal{A}_N$ and $\mathcal{A}$ satisfy the hypothesis of the Trotter-Kato theorem.

In order to ensure condition (1.29), we need compactness of the gain operator $\mathcal{K}$ and dual convergence (in the Trotter-Kato sense) of the approximating operators, i.e.,

$$\mathcal{A}_N^* \to \mathcal{A}^*$$

in the Trotter-Kato sense [6],[23],[25].

Certain applications to structural and flow control lead to problems where the operator $\mathcal{A}$ is nonnormal, that is,

$$\mathcal{A}\mathcal{A}^* \neq \mathcal{A}^*\mathcal{A}.$$

This observation has three important implications [12].

1) If $\mathcal{A}$ is nonnormal, the system

$$\dot{z}(t) = \mathcal{A}z(t) + \mathcal{B}u(t)$$

can be highly sensitive to parameter variations and inputs. Small variations in the operator $\mathcal{A}$ can produce changes in the stability of the system, and small inputs u can be greatly magnified.

2) Riccati equations involving a nonnormal operator $\mathcal{A}$ can be "stiff" (see [13]).

3) If $\mathcal{A}$ is nonnormal, then we must be careful when developing approximations for optimal control of these systems in order to ensure uniform convergence of the feedback operators.

Issue 1) has been extensively studied in [2], [3],[18], [19], [49], [50]. The following simple problem from [13] illustrates issue 2).

Consider the linear system

$$\frac{d}{dt}\begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = \begin{bmatrix} \mu & R \\ 0 & -\mu \end{bmatrix}\begin{bmatrix} v(t) \\ w(t) \end{bmatrix} + \begin{bmatrix} p \\ q \end{bmatrix} u(t), \tag{1.32}$$

where $R > 0,\quad \mu > 0$, and p and q are not both zero. If we define $A = \begin{bmatrix} \mu & R \\ 0 & -\mu \end{bmatrix}$ and $B = \begin{bmatrix} p \\ q \end{bmatrix}$, then (1.32) can be written as

$$\frac{d}{dt}\begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = A\begin{bmatrix} v(t) \\ w(t) \end{bmatrix} + Bu(t), \tag{1.33}$$

When $p \neq 0$, $q = 0$, and

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

then the positive definite solution of the ARE

$$A^T\Pi + \Pi A - \Pi BB^T\Pi + Q = 0$$

is given by

$$\Pi(\mu, p, R) = \begin{bmatrix} \frac{\mu+\sqrt{\mu^2+p^2}}{p^2} & \frac{R}{p^2} \\ \frac{R}{p^2} & \frac{R^2+p^2}{2\mu p^2} \end{bmatrix}.$$

As R becomes large, the matrix A becomes "more nonnormal." Moreover, when $p >> 1$,

$$\lim_{R\to+\infty} \|\Pi(\mu, p, R)\| = \infty. \tag{1.34}$$

Thus, the ARE can become ill-conditioned for large values of R.

However, when $R \to 0$ the matrix A becomes normal (self-adjoint), and when $R = 0$,

$$\Pi(\mu, p, 0) = \begin{bmatrix} \frac{\mu+\sqrt{\mu^2+p^2}}{p^2} & 0 \\ 0 & \frac{p^2}{2\mu p^2} \end{bmatrix}.$$

This example shows that as A becomes more nonnormal, the computational problem associated with the ARE can become ill-conditioned.

Issue 3) is of considerable interest and is a main focus of this study. We address convergence of the feedback operators for a specific problem in the following chapter.

1.5 The Resulting Matrix LQR Problem

To obtain an approximate solution for $(\mathbf{LQR}_N)$ by computational methods, we first find the matrix representations of the approximate operators. To do so, we establish the notation $[T]$ to denote the matrix representation of the operator T. We present two equivalent approaches to find the matrix representations of the approximate operators.

1.5.1 Approach 1

Let $\{b_i\}_{i=1}^N$ be a basis for V_N. We define the $N \times N$ mass matrix to be

$$[M_N] = [\langle b_i, b_j \rangle_H]_{i,j=1}^N.$$

Note that $[M_N]$ is a symmetric matrix.

Define the M_N-inner product on $\mathbb{R}^N$ by

$$\left\langle \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{M_N} \equiv \begin{bmatrix} a_1 & \dots & a_N \end{bmatrix} [M_N]^T \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix}, \quad \text{for all} \quad \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N.$$

The matrix representation of the operator $\mathcal{A}_N$ with respect to the basis vectors $\{b_i\}_{i=1}^N$ and the M_N-inner product is the matrix $[\mathcal{A}_N]$ defined by

$$\left\langle \mathcal{A}_N \sum_{i=1}^N a_i b_i, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\mathcal{A}_N] \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{M_N}, \quad \text{for all} \quad \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N. \quad (1.35)$$

As in [25] and [8], the matrix representation of $(\mathcal{A}_N)^*$, the Hilbert adjoint of $\mathcal{A}_N$, is given by

$$[(\mathcal{A}_N)^*] = [M_N]^{-1}[\mathcal{A}_N]^T[M_N].$$

The matrix representation of the operator $\mathcal{Q}_N$ with respect to the basis vectors $\{b_i\}_{i=1}^N$ and the M_N-inner product is the matrix $[\mathcal{Q}_N]$ such that

$$\left\langle \mathcal{Q}_N \sum_{i=1}^N a_i b_i, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\mathcal{Q}_N] \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{M_N}, \quad \text{for all} \quad \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N. \quad (1.36)$$

We restrict our attention to the case when the space U is finite dimensional. Thus, the operator $\mathcal{B} : U \to H$ is an operator of finite rank. This fact implies that there exist vectors $f_i \in H, i = 1..., m$ such that

$$\mathcal{B} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} = \sum_{i=1}^m f_i u_i, \quad \text{for all} \quad \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \in \mathbb{R}^m.$$

In this case, $[\mathcal{B}_N] \in \mathbb{R}^{m\times N}$, the matrix representation of $\mathcal{B}_N$, is the matrix given by

$$\left\langle \mathcal{B}_N \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\mathcal{B}_N] \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{M_N} \quad \text{for all} \quad \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \in \mathbb{R}^m, \quad \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N.$$

Furthermore, $[(\mathcal{B}_N)^*] = [\mathcal{B}_N]^T[M_N]$.

Using these matrices, we write the matrix linear quadratic regulator problem **([LQR$_N$])** associated with **(LQR$_N$)**.

([LQR$_N$]) *Find* $[u_N(t)]$ *in* $L^2((0,\infty);\mathbb{R})$ *such that* $[u_N(t)]$ *minimizes*

$$J([z_{0N}],[u_N(\cdot)]) = \int_0^\infty \{< [\mathcal{Q}_N][z_N(t)],[z_N(t)] >_{M_N} + < [\mathcal{R}_N][u_N(t)],[u_N(t)] >_U\}dt \quad (1.37)$$

subject to

$$\frac{d[z_N(t)]}{dt} = [\mathcal{A}_N][z_N(t)] + [\mathcal{B}_N][u_N(t)], \quad \text{on} \quad \mathbb{R}^N, \quad (1.38)$$

$$[z_N](0) = [z_{0N}], \quad \text{in} \quad \mathbb{R}^N. \quad (1.39)$$

Again, when **([LQR$_N$])** satisfies the hypotheses of Theorem 1.2.3, we can solve the matrix ARE

$$[M_N]^{-1}[\mathcal{A}_N]^T[M_N][\Pi_N] + [\Pi_N][\mathcal{A}_N] - [\Pi_N][\mathcal{B}_N][\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[M_N][\Pi_N] + [\mathcal{Q}_N] = 0. \quad (1.40)$$

We arrive at this equation by substituting the matrix representations of $\mathcal{A}_N^*, \mathcal{A}_N, \mathcal{B}_N, \mathcal{B}_N^*, \mathcal{R}_N, \mathcal{Q}_N$, and Π_N into (1.28). The optimal control for **([LQR$_N$])** is

$$[\hat{u}_N(t)] = -[\mathcal{K}_N][\hat{z}_N(t)] = -[\mathcal{R}_N]^{-1}[(\mathcal{B}_N)^*][\Pi_N][\hat{z}(t)] = -[\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[M_N][\Pi_N][\hat{z}_N(t)].$$

For most numerical ARE solvers, (1.40) is not in standard form. We remedy this issue by premultiplying (1.40) by $[M_N]$ and introducing the change of variables

$$[P_N] = [M_N][\Pi_N].$$

To find $[\Pi_N]$, we first solve the equation

$$[\mathcal{A}_N]^T[P_N] + [P_N][\mathcal{A}_N] - [P_N][\mathcal{B}_N][\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[P_N] + [M][\mathcal{Q}_N] = 0 \tag{1.41}$$

for $[P_N]$ and recover $[\Pi_N]$ since $[\Pi_N] = [M]^{-1}[P_N]$. Finally, we write

$$[\hat{u}_N(t)] = -[\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[P_N][\hat{z}_N(t)].$$

1.5.2 Approach 2

Alternatively, we can develop **([LQR$_N$])** in the following manner. Find the matrices, $[\tilde{A}_N], [\tilde{Q}_N]$ and $[\tilde{B}_N]$ such that

$$\left\langle \mathcal{A}_N \sum_{i=1}^N a_i b_i, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\tilde{A}_N] \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{\mathbb{R}^N}, \quad \text{for all} \quad \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N, \tag{1.42}$$

$$\left\langle \mathcal{Q}_N \sum_{i=1}^N a_i b_i, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\tilde{Q}_N] \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{\mathbb{R}^N}, \quad \text{for all} \quad \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N, \tag{1.43}$$

and

$$\left\langle \mathcal{B}_N \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \sum_{j=1}^N c_j b_j \right\rangle_H = \left\langle [\tilde{B}_N] \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \right\rangle_{\mathbb{R}^N} \quad \text{for all} \quad \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \in \mathbb{R}^m, \quad \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N.$$

Note that

$$[\mathcal{A}_N] = [M_N]^{-1}[\tilde{A}_N] \tag{1.44}$$

$$[M_N][\mathcal{Q}_N] = [\tilde{Q}_N] \tag{1.45}$$

$$[\mathcal{B}_N] = [M_N]^{-1}[\tilde{B}_N]. \tag{1.46}$$

Thus, we have a second linear quadratic regulator problem.

([LQR$_{N2}$]) *Find* $[u_N(t)]$ *in* $L^2((0,\infty);\mathbb{R})$ *such that* $[u_N(t)]$ *minimizes*

$$J([z_{0,N}],[u_N(\cdot)]) = \int_0^\infty \{< [\tilde{Q}_N][z_N(t)],[z_N(t)] >_{\mathbb{R}^N} + < [\mathcal{R}_N][u_N(t)],[u_N(t)] >_U\}dt \quad (1.47)$$

subject to

$$[M_N]\frac{d[z_N(t)]}{dt} = [\tilde{A}_N][z_N(t)] + [\tilde{B}_N][u_N(t)], \quad \text{on} \quad \mathbb{R}^N, \quad (1.48)$$

$$[z_N](0) = [z_{0,N}], \quad \text{in} \quad \mathbb{R}^N. \quad (1.49)$$

By premultiplying (1.48) by $[M_N]^{-1}$, we see that (1.48) is the same equation as (1.38). The matrix ARE corresponding to **([LQR$_{N2}$])** is

$$[\mathcal{A}_N]^T[P_N] + [P_N][\mathcal{A}_N] - [P_N][\mathcal{B}_N][\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[P_N] + [\tilde{Q}_N] = 0. \quad (1.50)$$

Note that this is the same ARE as in (1.40) since $[M_N][\mathcal{Q}_N] = [\tilde{Q}_N]$. Furthermore, the optimal control is given by

$$[\hat{u}_N(t)] = -[\mathcal{R}_N]^{-1}[\mathcal{B}_N]^T[P_N][\hat{z}_N(t)].$$

Thus, the optimal control computed for ([LQR$_{N2}$]) is the same optimal control for ([LQR$_N$]).

Remark: Although Approaches 1 and 2 are equivalent, there is one key difference. Note that $[P_N]$ is not the matrix respresentation of Π_N. To find $[\Pi_N]$, we must premultiply $[P_N]$ by $[M]^{-1}$.

Chapter 2

Optimal Control of the Convection-Diffusion Equation

Having established the general framework for approximating infinite dimensional linear quadratic regulator problems, we investigate the approximation of a specific linear quadratic regulator problem in which the constraint equation is the one-dimensional convection-diffusion equation. The fundamental issue is concerned with approximation of the convection-diffusion operator. We present three Galerkin finite element methods. These schemes differ in choice of V_N and inner products. We use the Trotter-Kato Theorem to prove convergence of the approximations and their adjoints.

2.1 The Convection-Diffusion Equation

Consider the following partial differential equation where $\mu > 0, \quad v > 0$:

$$\frac{\partial w(t,x)}{\partial t} = \mu \frac{\partial^2 w(t,x)}{\partial x^2} - v \frac{\partial w(t,x)}{\partial x} + b(x)u(t) \quad t > 0, \quad 0 < x < 1, \tag{2.1}$$

with boundary conditions

$$w(t,0) = 0, \quad w(t,1) = 0 \tag{2.2}$$

and initial condition

$$w(0,x) = w_0(x). \tag{2.3}$$

When μ is large in comparison to v, the equation is diffusion dominated. In general, standard Galerkin finite element techniques that simulate the diffusion dominated equation are convergent. However, when μ is small in comparison to v, the equation is convection dominated. When (2.1), is convection dominated, the open loop form of (2.1) - (2.3) becomes highly nonnormal. Standard Galerkin approximation techniques for this equation can fail to be convergent. Stabilized, or upwinding, methods have been developed to overcome this shortcoming [20]. However, in an optimal control setting, the stabilized methods may have limitations [36]. Our goal is to develop numerical optimal control techniques to approximate (2.1) - (2.3) for a wide range of values of μ.

The convection-diffusion operator is denoted by A_0 and defined on $H = L^2(0,1)$. The domain is

$$\mathcal{D}(A_0) = H_0^1(0,1) \bigcap H^2(0,1),$$

and

$$A_0 w = \mu \frac{\partial^2 w}{\partial x^2} - v \frac{\partial w}{\partial x}, \quad \text{for all} \quad w \in \mathcal{D}(A_0). \tag{2.4}$$

Let A_H denote the diffusion operator defined on

$$\mathcal{D}(A_H) = H_0^1(0,1) \bigcap H^2(0,1)$$

by

$$A_H w = \mu \frac{\partial^2 w}{\partial x^2}.$$

It is well known that the eigenvalues of A_H are

$$\lambda_n = -n^2 \pi^2 \mu, \quad n = 1, 2, \ldots,$$

and the eigenvalues of A_0 are

$$\lambda_n = -n^2 \pi^2 \mu - \frac{v^2}{4\mu^2}, \quad n = 1, 2, \ldots. \tag{2.5}$$

Note that the eigenvalues of the convection-diffusion operator are the eigenvalues of the diffusion operator shifted by $-\frac{v^2}{4\mu^2}$.

In this chapter, we investigate approximation and optimal control of the following distributed parameter optimal control problem. Let $\mu > 0, \quad v > 0, \quad q(x) \geq 0, q(\cdot) \in L^\infty(0,1)$, and constant $r > 0$. For each $u \in L^2((0,\infty); \mathbb{R})$, define the cost function

$$J(w_0, u) = \int_0^\infty \left(\left(\int_0^1 q(x) w^2(t,x) dx \right) + r u^2(t) \right) dt, \tag{2.6}$$

where $w(t,x)$ is the solution to (2.1)-(2.3). The linear quadratic regulator problem is to minimize (2.6) subject to (2.1)-(2.3).

Define the operators $B : \mathbb{R} \to H, \quad Q : H \to H$, and $R : \mathbb{R} \to \mathbb{R}$ by

$$Bu = b(x)u, \quad \text{for all} \quad u \in \mathbb{R}, \tag{2.7}$$

$$[Qw](x) = q(x)w, \quad \text{for all} \quad w \in H, \tag{2.8}$$

$$Ru = ru, \quad \text{for all} \quad u \in \mathbb{R}. \tag{2.9}$$

Using this framework, the control problem (2.1)-(2.6) becomes the infinite dimensional linear quadratic regulator problem $(\mathbf{\Sigma})$.

$(\mathbf{\Sigma})$ *Minimize* $J(z_0, u)$ *over* $u \in L^2((0,\infty); \mathbb{R})$ *where*

$$\frac{dz(t)}{dt} = A_0 z(t) + Bu(t) \tag{2.10}$$

$$z(0) = z_0 \tag{2.11}$$

and

$$J(z_0, u) = \int_0^\infty \{\langle\langle Qz(t), z(t)\rangle\rangle + \langle\langle Ru(t), u(t)\rangle\rangle_{\mathbb{R}}\}dt. \tag{2.12}$$

2.2 The Linear Quadratic Regulator Problem $(\mathbf{\Sigma})$

In this section, we apply Theorem 1.2.3 to the optimal control problem $(\mathbf{\Sigma})$. Clearly, R is a positive definite operator on $\mathbb{R}$, Q is positive semidefinite, and B is bounded. We must show that A_0 generates a C_0-semigroup on H, the pair (A_0, B) is stabilizable, and there exists a control $u \in L^2((0,\infty); \mathbb{R})$ that drives the state z to zero in order to apply Theorem 1.2.3 to $(\mathbf{\Sigma})$.

Although it is well known that A_0 generates a C_0-semigroup on H, we present a proof based on Theorem 1.2.2. The estimates used in the proof will be useful later.

If w in $\mathcal{D}(A_0)$, then

$$\langle\langle A_0 w, w\rangle\rangle = \left\langle \mu \frac{\partial^2 w}{\partial x^2} - v\frac{\partial w}{\partial x}, w\right\rangle \tag{2.13}$$

$$= \int_0^1 \left\{\mu \frac{\partial^2 w(x)}{\partial x^2} w(x) - v\frac{\partial w(x)}{\partial x} w(x)\right\} dx \tag{2.14}$$

$$= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x}\right)^2 dx - v/2 \int_0^1 \frac{\partial}{\partial x}\left(w^2(x)\right) dx \tag{2.15}$$

$$= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x}\right)^2 dx - v/2\left(w^2(1) - w^2(0)\right) \tag{2.16}$$

$$= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x}\right)^2 dx \tag{2.17}$$

$$\le 0. \tag{2.18}$$

Thus, A_0 is dissipative, and we turn to A_0^*. It is not difficult to show that under the $L^2(0,1)$ inner product, the domain of A_0^* is

$$\mathcal{D}(A_0^*) = H_0^1(0,1) \bigcap H^2(0,1),$$

and

$$A_0^* w = \mu \frac{\partial^2 w}{\partial x^2} + v \frac{\partial w}{\partial x}, \quad \text{for all} \quad w \in \mathcal{D}(A_0^*).$$

Also, if w in $\mathcal{D}(A_0^*)$, then

$$\begin{aligned}
\langle A_0^* w, w \rangle &= \left\langle \mu \frac{\partial^2 w}{\partial x^2} + v \frac{\partial w}{\partial x}, w \right\rangle && (2.19)\\
&= \int_0^1 \left\{ \mu \frac{\partial^2 w(x)}{\partial x^2} w(x) + v \frac{\partial w(x)}{\partial x} w(x) \right\} dx && (2.20)\\
&= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x} \right)^2 dx + v/2 \int_0^1 \frac{\partial}{\partial x} \left(w^2(x) \right) dx && (2.21)\\
&= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x} \right)^2 dx + v/2 \left(w^2(1) - w^2(0) \right) && (2.22)\\
&= -\int_0^1 \mu \left(\frac{\partial w(x)}{\partial x} \right)^2 dx && (2.23)\\
&\leq 0. && (2.24)
\end{aligned}$$

Therefore, A_0^* is dissipative. Note that $C_0^\infty(0,1) \subseteq H_0^1(0,1) \bigcap H^2(0,1)$ is dense in H. Thus, $\mathcal{D}(A_0) = \mathcal{D}(A_0^*) = H_0^1(0,1) \bigcap H^2(0,1)$ is dense in H. Moreover, it is well known that both A_0 and A_0^* are closed operators. Since both A_0 and A_0^* are densely defined, closed, dissipative operators, by Theorem 1.2.2, A_0 generates the C_0-semigroup $S_1(T)$ on H.

Now, we must show that the pair (A_0, B) is stabilizable. To do this, we prove the following result.

Lemma 2.2.1 *The convection-diffusion operator A_0 defined in* (2.4) *generates an exponentially stable semigroup on H, i.e. if $S_1(t)$ is the C_0-semigroup on H generated by A_0, then there exist positive constants $M \geq 1$ and $\omega \geq 0$ such that*

$$\|S_1(t)z\| \leq Me^{-\omega t} \|z\|, \quad \textit{for all} \quad z \in H.$$

Proof: We start by recalling some well known facts.

(i) The eigenvalues of A_0 are

$$\lambda_n = -n^2\pi^2\mu - \frac{v^2}{4\mu^2}, \quad n = 1, 2, \ldots.$$

Furthermore, the corresponding eigenvectors are given by

$$z_n(x) = e^{vx/2\mu} \sin(n\pi x) / \left\| e^{v(\cdot)/2\mu} \sin(n\pi(\cdot)) \right\|, \quad n = 1, 2, \ldots$$

and form a basis for H.

(ii) For any positive constant ϵ_1 and element z in H, there exist a finite set of eigenvectors, $\{z_{n,i}\}_{i=1}^{k}$ and constants $c_i, i = 1, ..., k$, such that

$$\left\| z - \sum_{i=1}^{k} c_i z_{n,i} \right\| < \epsilon_1. \tag{2.25}$$

It follows that

$$\left\| \sum_{i=1}^{k} c_i z_{n,i} \right\| < \epsilon_1 + \|z\| . \tag{2.26}$$

(iii) From [43], we know that

$$S_1(t)z = \lim_{n\to\infty} \left(I - \frac{t}{n} A_0 \right)^{-n} z. \tag{2.27}$$

(iv) It follows from (2.27) that

$$S_1(t)z_n = e^{\lambda_n t} z_n, \quad n = 1, 2, \ldots \tag{2.28}$$

for any eigenvector z_n.

We now proceed to prove the desired result. Let $\epsilon > 0$ be arbitrary, and let z be any vector in H such that $\|z\| = 1$. Take $\epsilon_1 < \epsilon / \|S_1(t)\|_{\mathcal{L}(H)}$. By (2.25),(2.26) , and (2.28)

$$\begin{aligned}
\|S_1(t)z\| &= \left\| S_1(t)z - S_1(t)\left(\sum_{i=1}^{k} c_i z_{n,i}\right) + S_1(t)\left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| && (2.29)\\
&\le \left\| S_1(t)z - S_1(t)\left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| + \left\| S_1(t)\left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| && (2.30)\\
&\le \|S_1(t)\|_{\mathcal{L}(H)} \left\| z - \left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| + \left\| S_1(t)\left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| && (2.31)\\
&\le \|S_1(t)\|_{\mathcal{L}(H)} \epsilon_1 + \left\| \left(\sum_{i=1}^{k} e^{\lambda_n t} c_i z_{n,i}\right) \right\| && (2.32)\\
&\le \|S_1(t)\|_{\mathcal{L}(H)} \epsilon_1 + e^{\lambda_1 t} \left\| \left(\sum_{i=1}^{k} c_i z_{n,i}\right) \right\| && (2.33)\\
& && (2.34)\\
&< \epsilon + e^{\lambda_1 t}(\epsilon + \|z\|) && (2.35)\\
&= \epsilon(e^{\lambda_1 t} + 1) + e^{\lambda_1 t}. && (2.36)
\end{aligned}$$

To summarize, for arbitrary $\epsilon > 0$,

$$\|S_1(t)z\| < \epsilon(e^{\lambda_1 t} + 1) + e^{\lambda_1 t}. \tag{2.37}$$

for all z in H such that $\|z\| = 1$.

Therefore,

$$\|S_1(t)\| = \sup_{\|z\|=1} \|S_1(t)z\| \leq e^{\lambda_1 t}. \tag{2.38}$$

As a result, for any z in H,

$$\|S_1(t)z\| < e^{-(\pi^2\mu+(v^2)/(4\mu^2)t} \|z\|. \quad \square \tag{2.39}$$

Note that $\|S_1(t)\|_{\mathcal{L}(H)} \leq e^{-(\pi^2\mu+(v^2)/(4\mu^2)t}$, so $S_1(t)$ is a semigroup of contractions. The stabilizabilty of (A_0, B) is a direct consequence of Lemma 2.2.1 when $K = 0$ in Definition 1.2.6. Moreover, $u = 0$ is a control that drives the state of $(\mathbf{\Sigma})$ to zero exponentially so $J(z_0, 0) < \infty$ for any initial condition z_0. Thus, by Theorem 1.2.3, $(\mathbf{\Sigma})$ has a unique optimal control.

2.3 Approximations

We develop three different approximations of A_0. These approximations are described in subsequent sections.

2.3.1 A Traditional Approximation Scheme

In order to develop the approximation A_1^N of A_0, we define a bilinear operator and show that it is elliptic. We call this approximation scheme a traditional approximation scheme because it uses the standard $L^2(0,1)$ spaces.

Let $V = H_0^1(0,1) \subseteq H = L^2(0,1)$. Define the bilinear operator $a_1(\cdot,\cdot)$ on $V \times V$ by

$$a_1(w,z) = \int_0^1 \mu \frac{\partial w(x)}{\partial x}\frac{\partial z(x)}{\partial x} + v\frac{\partial w(x)}{\partial x} z(x)dx \quad \text{for all} \quad w, z \in V. \tag{2.40}$$

Let $\|\cdot\|$ denote the $L^2(0,1)$ norm, and let $\|\cdot\|_1$ and $|\cdot|_1$ denote the equivalent norms from Section 1.1 on V. By the equivalence of the $\|\cdot\|_1$ and $|\cdot|_1$ norms on V, there exists a positive constant b_1 such that

$$a_1(w,w) = \mu|w|_1^2 \geq b_1^2 \|w\|_1^2, \quad \text{for all} \quad w \in V. \tag{2.41}$$

Moreover, if w, z in V, then

$$a_1(w,z) = \int_0^1 \mu \frac{\partial w(x)}{\partial x}\frac{\partial z(x)}{\partial x} + v\frac{\partial w(x)}{\partial x} z(x)dx \tag{2.42}$$
$$\leq \mu \|w_x\| \|z_x\| + v \|w_x\| \|z\| \tag{2.43}$$
$$\leq (\mu + v) \|w\|_1 \|z\|_1 . \tag{2.44}$$

Thus, if $b_2 = \mu + v$,

$$a_1(w,z) \leq b_2 \|w\|_1 \|z\|_1 , \quad \text{for all} \quad w, z \in V. \tag{2.45}$$

From (2.41) and (2.45), it is clear that a_1 is $V-$elliptic. By the Lax-Milgram Lemma, there exists an operator $A : V \to V'$ such that

$$a_1(w,z) = -\langle Aw, z\rangle_{V' \times V} , \quad \text{for all} \quad w, z \in V. \tag{2.46}$$

We now proceed to approximate A.

The first step in approximating A is to choose a finite dimensional subspace of V. Let $x_i = \frac{i}{N}$, $i = 0, 1, \ldots, N$. Define the piecewise linear, continuous splines h_i^N, $i = 1, ..., N-1$, on the interval $[0, 1]$ by

$$h_i^N(x) = \begin{cases} N(x - x_{i-1}^N), & x \in [x_{i-1}^N, x_i^N], \\ N(x_{i+1}^N - x), & x \in [x_i^N, x_{i+1}^N], \\ 0 & \text{otherwise.} \end{cases} \tag{2.47}$$

Define $V^N \subseteq V$ as

$$V^N \equiv span\{h_i^N\}_{i=1}^{N-1}.$$

Note that $V^N \subseteq V \subseteq H$.

When $a_1(\cdot,\cdot)$ is restricted to $V^N \times V^N$, the inequalities (2.41) and (2.45) still hold. Thus, by the Lax Milgram Lemma, there exists a unique operator $A_1^N : V^N \to V^N$ such that

$$a_1(w,z) = < -A_1^N w, z >, \quad \text{for all} \quad w, z \in V^N. \tag{2.48}$$

V^N is a finite dimensional space. Since A_1^N is an operator of finite rank defined on all of V^N, by Theorem 1.2.1, A_1^N generates a uniformly continuous semigroup on V^N.

Furthermore, if $w \in V^N$, then by (2.41),

$$< A_1^N w, w > = -a_1(w,w) \tag{2.49}$$
$$\leq -b_1^2 \|w\|^2 \tag{2.50}$$
$$\leq 0. \tag{2.51}$$
$$\tag{2.52}$$

and

$$
\begin{aligned}
< A_1^{N*} w, w > &= -a_1(w,w) && (2.53)\\
&\leq -b_1^2 \|w\|^2 && (2.54)\\
&\leq 0. && (2.55)\\
& && (2.56)
\end{aligned}
$$

2.3.2 A Second Approximation Scheme

We introduce a second approach to approximating the convection-diffusion operator. To do so, we introduce another inner product and norm for $L^2(0,1)$. The technique of renorming is not uncommon. Several examples are provided in [16] where renorming is advantageous.

2.3.3 The $\sigma-$Inner Product

Let σ be the function

$$\sigma(x) = e^{-vx/\mu}/\mu,$$

and consider the following inner product on $L^2(0,1)$. Define $<\cdot,\cdot>_\sigma$ by

$$< z, w >_\sigma = \int_0^1 \sigma(x) z(x) w(x) dx, \quad \text{for all} \quad z, w \in L^2(0,1).$$

The $\sigma-$inner product induces a norm, denoted $\|\cdot\|_\sigma$, that is equivalent to the standard $L^2(0,1)$ norm, $\|\cdot\|$, since

$$e^{-\frac{v}{2\mu}}/\sqrt{\mu}\, \|z\| \leq \|z\|_\sigma \leq \frac{1}{\sqrt{\mu}} \|z\| \quad \text{for all} \quad z \in L^2(0,1). \tag{2.57}$$

Let H_σ denote the Hilbert space of $L^2(0,1)$ functions with the $\sigma-$inner product. The convection-diffusion operator, A_0, is self-adjoint on H_σ. Since A_0 generates a C_0-semigroup on H and the $\sigma-$norm is equivalent to the norm $\|\cdot\|$, A_0 generates a C_0-semigroup on H_σ.

Additionally, let H_σ^1 denote the Hilbert space $H^1(0,1)$ with the norm

$$\|z\|_{1,\sigma}^2 = \|z\|_\sigma^2 + \left\| \frac{dz}{dx} \right\|_\sigma^2, \quad \text{for all} \quad z \in H_\sigma^1(0,1). \tag{2.58}$$

2.3.4 Approximating A_0

Changing the inner product on $L^2(0,1)$ leads to development of another approximation scheme. We define another bilinear form to proceed with this approximation scheme.

Let V_σ denote the space $H_0^1(0,1)$ when it is endowed with the $\|\cdot\|_{1,\sigma}$ norm. Define the bilinear operator $a_\sigma(\cdot,\cdot) : V_\sigma \times V_\sigma$ by

$$a_\sigma(z,w) \equiv \int_0^1 \sigma(x)\frac{dz(x)}{dx}\frac{dw(x)}{dx}dx, \quad \text{for all} \quad z,w \in V_\sigma. \tag{2.59}$$

We show that $a_\sigma(\cdot,\cdot)$ is elliptic on V_σ with respect to the norm $\|\cdot\|_{1,\sigma}$. Before doing so, we develop an inequality that is a specific case of Poincare's Inequality.

For $z \in H_0^1(0,1)$,

$$\begin{aligned}
\|z\|^2 &= \int_0^1 z^2(x)dx && (2.60)\\
&= \int_0^1 (1)z^2(x)dx && (2.61)\\
&= -\int_0^1 x\frac{d}{dx}\left(z^2(x)\right)dx && (2.62)\\
&= -2\int_0^1 xz(x)\frac{dz(x)}{dx}dx && (2.63)\\
&\leq 2\int_0^1 |x||z(x)||\frac{dz(x)}{dx}|dx && (2.64)\\
&\leq 2\int_0^1 |z(x)||\frac{dz(x)}{dx}|dx && (2.65)\\
&\leq 2\|z\|\left\|\frac{dz}{dx}\right\|. && (2.66)
\end{aligned}$$

Thus,

$$\|z\| \leq 2\left\|\frac{dz}{dx}\right\|, \quad \text{for all} \quad z \in H_0^1(0,1), \tag{2.67}$$

and it follows from the Cauchy-Schwartz inequality and (2.67) that $a_\sigma(\cdot,\cdot)$ is elliptic, as demonstrated below.

For all z in V_σ, we have

$$2 \;\; a_\sigma(z,z) = 2\left\|\frac{dz}{dx}\right\|_\sigma^2 \tag{2.68}$$

$$= \left\|\frac{dz}{dx}\right\|_\sigma^2 + \left\|\frac{dz}{dx}\right\|_\sigma^2 \tag{2.69}$$

$$\geq \left\|\frac{dz}{dx}\right\|_\sigma^2 + \left\|\frac{dz}{dx}\right\|^2 \left(\frac{e^{-v/\mu}}{\mu}\right) \tag{2.70}$$

$$\geq \left\|\frac{dz}{dx}\right\|_\sigma^2 + \|z\|^2 \left(\frac{e^{-v/\mu}}{4\mu}\right) \tag{2.71}$$

$$\geq \left\|\frac{dz}{dx}\right\|_\sigma^2 + \|z\|_\sigma^2 \left(\frac{e^{-v/\mu}}{4}\right) \tag{2.72}$$

$$\geq \frac{e^{-v/\mu}}{4}\left(\|z\|_\sigma^2 + \left\|\frac{dz}{dx}\right\|_\sigma^2\right)$$

$$= \frac{e^{-v/\mu}}{4}\|z\|_{1,\sigma}^2 . \tag{2.73}$$

Hence, for all z in V_σ,

$$a_\sigma(z,z) \geq c_1 \|z\|_{1,\sigma}^2 , \tag{2.74}$$

where $c_1 = \frac{e^{-v/\mu}}{4}$.

Moreover, for all z, w in V_σ,

$$|\, a_\sigma(z,w)| = \left|\int_0^1 \sigma(x)\frac{dz(x)}{dx}\frac{dw(x)}{dx}dx\right| \tag{2.75}$$

$$\leq \frac{1}{\mu}\int_0^1 \left|\frac{dz(x)}{dx}\right|\left|\frac{dw(x)}{dx}\right| dx \tag{2.76}$$

$$\leq \frac{1}{\mu}\left\|\frac{dz}{dx}\right\|\left\|\frac{dw}{dx}\right\| \tag{2.77}$$

$$\leq \frac{1}{\mu}\left(\|z\|^2 + \left\|\frac{dz}{dx}\right\|^2\right)^{1/2}\left(\|w\|^2 + \left\|\frac{dw}{dx}\right\|^2\right)^{1/2} \tag{2.78}$$

$$\leq \frac{1}{e^{-v/\mu}}\left(\|z\|_\sigma^2 + \left\|\frac{dz}{dx}\right\|_\sigma^2\right)^{1/2}\left(\|w\|_\sigma^2 + \left\|\frac{dw}{dx}\right\|_\sigma^2\right)^{1/2} . \tag{2.79}$$

Thus, for all z, w in V_σ,

$$|a_\sigma(z,w)| \leq c_2 \|z\|_{1,\sigma} \|w\|_{1,\sigma} , \tag{2.80}$$

where $c_2 = e^{v/\mu}$. From (2.80) and (2.74), we see that $a_\sigma(\cdot,\cdot)$ is elliptic on V_σ with respect to the $\|\cdot\|_{1,\sigma}$ norm. By the Lax-Milgram Lemma, there exists an operator $A_{\sigma,2} : V_\sigma \to V'_\sigma$ such that for all z, w in V_σ,

$$a_\sigma(z,w) = - < A_{\sigma,2} z, w >_{V_\sigma \times V'_\sigma} . \tag{2.81}$$

We define the finite dimensional subspace $V^N_{\sigma,1} \subseteq V_\sigma \subseteq H_\sigma$ by

$$V^N_{\sigma,1} \equiv span\{h^N_i\}^{N-1}_{i=1},$$

where h^N_i are the splines defined in (2.47).

Restricting $a_\sigma(\cdot,\cdot)$ to $V^N_{\sigma,1} \times V^N_{\sigma,1}$, still yields the inequalities (2.74) and (2.80). Thus, application of the Lax-Milgram Lemma defines a unique A^N_2 on $V^N_{\sigma,1}$ such that

$$a_\sigma(w,z) = < -A^N_2 w, z >_\sigma, \quad \text{for all} \quad w, z \in V^N_{\sigma,1}. \tag{2.82}$$

Note that $A^N_2 = (A^N_2)^*$.

Since A^N_2 is an operator of finite rank defined on all of $V^N_{\sigma,1}$, A^N_2 is a bounded operator on $V^N_{\sigma,1}$. By Theorem 1.2.1, A^N_2 generates a uniformly continuous semigroup on $V^N_{\sigma,1}$.

If $w \in V^N_{\sigma,1}$, then by (2.74),

$$\begin{aligned} < A^N_2 w, w >_\sigma &= -a_\sigma(w,w) \\ &\le -c_1 \|w\|^2_{1,\sigma} \\ &\le 0. \end{aligned}$$

Since A^N_2 is self-adjoint on $V^N_{\sigma,1}$, A^N_2 is a maximally dissipative.

2.3.5 A Third Approximation Scheme

For this approximation scheme, we use the inner product and bilinear forms defined in Sections 2.3.3 and 2.3.4, respectively. However, we choose a different finite dimensional subspace of V_σ.

Define the basis functions

$$\phi^N_i \equiv \frac{1}{\sqrt{\sigma}} h^N_i, i = 1, 2, ..., N-1,$$

and let $V^N_\sigma \subset V_\sigma$ be defined by

$$V^N_\sigma \equiv span\{\phi^N_i\}^{N-1}_{i=1}.$$

We use the bilinear form $a_\sigma(\cdot,\cdot)$ defined in (2.59), but we restrict $a_\sigma(\cdot,\cdot)$ to $V_\sigma^N \times V_\sigma^N$. Despite a change of basis, inequalities (2.74) and (2.80) still hold, so we can apply the Lax-Milgram Lemma to define a unique A_σ^N on V_σ^N such that

$$a_\sigma(w,z) = - < A_\sigma^N w, z >_\sigma, \quad \forall w, z \in V_\sigma^N. \tag{2.83}$$

The space V_σ^N is finite dimensional, and A_σ^N is an operator of finite rank defined on all of A_σ^N. By Theorem 1.2.1, A_σ^N generates a uniformly continuous semigroup on V_σ^N. Also, we have $A_\sigma^N = (A_\sigma^N)^*$.

If $w \in V_\sigma^N$, then by (2.74),

$$\begin{aligned} < A_\sigma^N w, w >_\sigma &= -a_\sigma(w,w) \\ &\leq -c_1 \|w\|_{1,\sigma}^2 \\ &\leq 0. \end{aligned} \tag{2.84}$$

Furthermore, since A_σ^N is self-adjoint on V_σ^N, A_σ^N is a maximally dissipative.

2.4 Proof of Convergence of the Infinitesimal Generators

In this section, we discuss the convergence of the approximate infinitesimal generators. Again, when we discuss convergence of the infinitesimal generators, we mean convergence in the sense of the Trotter-Kato Theorem.

2.4.1 Convergence of the First Scheme

Let A_0 be the convection-diffusion operator defined in (2.4), and let A_1^N be the sequence of approximating operators defined in (2.48). Convergence of A_1^N to A_0 is provided by the following theorem.

Theorem 2.4.1 *Let $S_1(t)$ be the C_0-semigroup on H generated by A_0, and let $S_1^N(t)$ be the C_0-semigroups on V^N generated by A_1^N. If P^N denotes the orthogonal projection from H into V^N, then, for every $x \in H$ and $t \geq 0$,*

$$\left\| S_1^N(t) P^N z - S_1(t) z \right\| \to 0$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Furthermore, for every $z \in H$ and $t \geq 0$,

$$\left\| \left(S_1^N\right)(t)^* P^N z - \left(S_1\right)(t)^* z \right\| \to 0$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Proof: The proof is a specific case of Lemma 3.2 from [7].□

The proof of Lemma 3.2 in [7] uses Theorem 1.4.3. In order to satisfy condition (1.31) in Theorem 1.4.3, the following lemma from [21] and [47] is needed.

Lemma 2.4.1 *For each $z \in V$, there exists an element $z^N \in V^N$ such that $\left\|z - z^N\right\|_1 \leq \epsilon_1(N)$, where $\epsilon_1(N) \to 0$ as $N \to \infty$.*

2.4.2 Convergence of the Second Scheme

In order to prove convergence of A_2^N, we use the following result concerning $V_{\sigma,1}^N$.

Lemma 2.4.2 *For each $z \in V_\sigma$, there exists an element $z^N \in V_{\sigma,1}^N$ such that $\left\|z - z^N\right\|_{1,\sigma} \leq \epsilon_2(N)$, where $\epsilon_2(N) \to 0$ as $N \to \infty$.*

Proof: We recall that the $L^2(0,1)$ norm and the σ-norm are equivalent on $L^2(0,1)$ since

$$e^{-\frac{v}{2\mu}}/\sqrt{\mu}\,\|z\|_2 \leq \|z\|_\sigma \leq \frac{1}{\sqrt{\mu}}\|z\|_2 \quad \text{for all} \quad z \in L^2(0,1). \tag{2.85}$$

Let $z \in V_\sigma$ be arbitrarily chosen. Thus, z is also in V. By Lemma 2.4.1, there exists some $z^N \in V^N$ such that

$$\sqrt{\mu}\left\|z - z^N\right\|_{1,\sigma} \leq \left\|z - z^N\right\|_1 \leq \epsilon_1(N). \tag{2.86}$$

Since $z^N \in V^N$ implies that $z^N \in V_{\sigma,1}^N$, the lemma is proven with $\epsilon_2(N) = 1/\sqrt{\mu}\epsilon_1(N)$.□

This lemma is useful when proving the convergence (in the Trotter-Kato sense) of A_2^N to A_0. Since A_2^N and A_0 are self-adjoint on H_σ, dual convergence follows automatically when we show A_2^N converges to A_0.

The proof that A_2^N converges to A_0 follows similar reasoning to arguments presented in [7], and we use similar notation. However, we present this proof because of some fundamental differences. We use the $\sigma-$inner product and analyze a specific method of approximating the convection-diffusion operator. In [7], Banks et al. provide a much more general framework for a large class of operators and use the $\|\cdot\|_1$ norm. Since their framework is more general, they must rely on a result from [22] that we do not use.

Theorem 2.4.2 *Let $S_1(t)$ be the C_0-semigroup on H_σ generated by A_0. Let $S_2^N(t)$ be the sequence of C_0-semigroups on $V_{\sigma,1}^N$ generated by A_2^N. If $P_{\sigma,1}^N$ denotes the orthogonal projection from H_σ into $V_{\sigma,1}^N$, then, for every $z \in H_\sigma$ and $t \geq 0$,*

$$\left\|S_2^N(t)P_{\sigma,1}^N z - S_1(t)z\right\|_\sigma \to 0 \tag{2.87}$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Furthermore, for every $z \in H_\sigma$ and $t \geq 0$,

$$\left\| \left(S_2^N\right)(t)^* P_{\sigma,1}^N z - \left(S_1\right)(t)^* z \right\|_\sigma \to 0$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Proof: We note that

$$\|z\|_\sigma^2 + \|z\|_{1,\sigma}^2 \leq 2|\ \|z\|_\sigma^2 + a(z,z)|, \quad \text{for all} \quad z \in V_\sigma. \tag{2.88}$$

For $z \in H_\sigma$, define $w = (I - A_0)^{-1} z$ and $w_N = (I - A_2^N)^{-1} P_{\sigma,1}^N z$. Thus, for all z_N in $V_{\sigma,1}^N$,

$$< w, z_N >_\sigma + a_\sigma(w, z_N) = < w, z_N >_\sigma - < A_0 w, z_N >_\sigma = < z, z_N >_\sigma \tag{2.89}$$

$$< w_N, z_N >_\sigma + a_\sigma(w_N, z_N) = < w_N, z_N >_\sigma - < A_0 w_N, z_N >_\sigma = < z, z_N >_\sigma . \tag{2.90}$$

When $e_N = w - w_N$, it follows that

$$< e_N, z_N >_\sigma + a_\sigma(e_N, z_n) = 0, \quad \text{for all} \quad z_N \in V_{\sigma,1}^N. \tag{2.91}$$

If we take $z = e_N$ in (2.88) and note that $e_N \in V_\sigma$, (2.91) implies that for all $z_N \in V_{\sigma,1}^N$,

$$\|e_N\|_\sigma^2 + \|e_N\|_{1,\sigma}^2 \leq 2|\ \|e_N\|_\sigma^2 + a_\sigma(e_N, e_N)| \tag{2.92}$$

$$= 2|\ \|e_N\|_\sigma^2 + a_\sigma(e_N, e_N) + < e_N, z_N >_\sigma + a_\sigma(e_N, z_n)| \tag{2.93}$$

$$= 2| < e_N, e_N + z_N >_\sigma + a_\sigma(e_N, e_N + z_n)|. \tag{2.94}$$

Let $z_N = w_N - P_{\sigma,1}^N w$. Then

$$\|e_N\|_\sigma^2 + \|e_N\|_{1,\sigma}^2 \leq 2| < e_N, e_N + w_N - P_{\sigma,1}^N w >_\sigma + a_\sigma(e_N, e_N + w_N - P_{\sigma,1}^N w)| \tag{2.95}$$

$$= 2| < e_N, w - P_{\sigma,1}^N w >_\sigma + a_\sigma(e_N, w - P_{\sigma,1}^N w)| \tag{2.96}$$

$$\leq 2\|e_N\|_\sigma \left\|w - P_{\sigma,1}^N w\right\|_\sigma + 2|a_\sigma(e_N, w - P_{\sigma,1}^N w)| \tag{2.97}$$

$$\leq 2\|e_N\|_\sigma \left\|w - P_{\sigma,1}^N w\right\|_\sigma + 2c_2 \|e_N\|_{1,\sigma} \left\|w - P_{\sigma,1}^N w\right\|_{1,\sigma} \tag{2.98}$$

$$\leq 2(1 + c_2)\epsilon_2(N)(\|e_N\|_\sigma + \|e_N\|_{1,\sigma}), \tag{2.99}$$

where c_2 is the constant defined in (2.80). Hence,

$$2\left(\|e_N\|_\sigma^2 + \|e_N\|_{1,\sigma}^2\right) \leq 4(1 + c_2)\epsilon_2(N)(\|e_N\|_\sigma + \|e_N\|_{1,\sigma}),$$

which implies

$$\|e_N\|_\sigma + \|e_N\|_{1,\sigma} \leq \frac{2\left(\|e_N\|_\sigma^2 + \|e_N\|_{1,\sigma}^2\right)}{\|e_N\|_\sigma + \|e_N\|_{1,\sigma}} \leq 4(1 + c_2)\epsilon_2(N).$$

Thus, we have

$$(I - A_2^N)^{-1} P_{\sigma,1}^N z \to (I - A_0)^{-1} z, \quad \text{for all} \quad z \in H_\sigma. \tag{2.100}$$

Equation (2.87) follows from Theorem 1.4.3. Since A_0 and A_2^N are self adjoint on H_σ, dual convergence follows from Theorem 1.4.4. □

2.4.3 Convergence of the Third Scheme

The following result can be found in [21].

Lemma 2.4.3 *Define* $x_i = \frac{i}{N}, i = 0, \ldots, N$. *Let* z *be a function in* $H_0^1(0,1)$, *and let* $\hat{z}_N$ *denote the linear interpolate of* z *in* V^N. *Then* $\hat{z}_N(x_i) = z(x_i)$, *for all* $i = 0, 1, ..., N$. *Additionally, there exists a positive constant* $C = C(z)$ *such that*

(i)

$$\|z - \hat{z}_N\| \leq \frac{C}{N^2}, \tag{2.101}$$

(ii)

$$\|z' - \hat{z}_N'\| \leq \frac{C}{N}, \tag{2.102}$$

and

(iii)

$$\|z - \hat{z}_N\|_1^2 \leq C^2 \left(\frac{1}{N^2} + \frac{1}{N^4}\right). \tag{2.103}$$

(iv) If z^N *denotes the orthogonal projection of* z *into* V^N, *then*

$$\left\|z - z^N\right\|_1^2 \leq C^2 \left(\frac{1}{N^2} + \frac{1}{N^4}\right). \tag{2.104}$$

To prove the convergence of A_σ^N, the operator defined in (2.83), we need the following result.

Lemma 2.4.4 *For each* $z \in V_\sigma$, *there exists an element* $z^N \in V_\sigma^N$ *such that* $\left\|z - z^N\right\|_{1,\sigma} \leq \epsilon(N)$, *where* $\epsilon(N) \to 0$ *as* $N \to \infty$.

Proof: We proceed in the spirit of the proof of Lemma 2.4.1 from [21].

Let z be any function in V_σ. By the density of $V_\sigma \bigcap H^3(0,1)$ in V_σ, there exists z_σ in $V_\sigma \bigcap H^3(0,1)$ such that

$$\|z - z_\sigma\|_{1,\sigma} \leq \frac{\tilde{C}}{2N}, \tag{2.105}$$

where $\tilde{C}$ is a constant to be defined later and N is a positive integer.

Define

$$\tilde{z} \equiv \sqrt{\sigma} z_\sigma = \left(e^{-\frac{v}{2\mu}(\cdot)} / \sqrt{\mu}\right) z_\sigma.$$

Note that $\tilde{z} \in V_\sigma \bigcap H^3(0,1)$ and $\tilde{z}$ is also a function in $V = H_0^1(0,1)$. Let $\hat{z}_N \in V^N$ denote the linear interpolate of $\tilde{z}$. It follows from Lemma 2.4.3 that there exists a constant $C = C(\tilde{z})$ such that

$$\|\tilde{z} - \hat{z}_N\| \leq \frac{C}{N^2} \tag{2.106}$$

and

$$\|\tilde{z}' - \hat{z}_N'\| \leq \frac{C}{N}. \tag{2.107}$$

It follows from (2.106) and (2.107) that

$$\begin{aligned}
\left\| z_\sigma - \frac{1}{\sqrt{\sigma}}\hat{z}_N \right\|_{1,\sigma}^2 &= \left\| \frac{1}{\sqrt{\sigma}}\tilde{z} - \frac{1}{\sqrt{\sigma}}\hat{z}_N \right\|_{1,\sigma}^2 \\
&= \sum_{i=0}^{N} \int_{x_{i-1}}^{x_i} e^{-\frac{v}{\mu}x}/\mu \left\{ \frac{d}{dx}\left(\sqrt{\mu}e^{\frac{v}{2\mu}x}\tilde{z}(x)\right) - \frac{d}{dx}\left(\sqrt{\mu}e^{\frac{v}{2\mu}x}\hat{z}_N(x)\right) \right\}^2 dx + \|\tilde{z} - \hat{z}_N\|^2 \\
&= \sum_{i=0}^{N} \left\{ \int_{x_{i-1}}^{x_i} \left\{ \left(\frac{d\tilde{z}(x)}{dx} + \frac{v}{2\mu}\tilde{z}(x) \right) - \left(\frac{d\hat{z}_N(x)}{dx} + \frac{v}{2\mu}\hat{z}_N(x) \right) \right\}^2 dx \right\} + \|\tilde{z} - \hat{z}_N\|^2 \\
&= \sum_{i=0}^{N} \left\{ \int_{x_{i-1}}^{x_i} \left\{ \left(\frac{d\tilde{z}(x)}{dx} - \frac{d\hat{z}_N(x)}{dx} \right) + \frac{v}{2\mu}\left(\tilde{z}(x) - \hat{z}_N(x)\right) \right\}^2 dx \right\} + \|\tilde{z} - \hat{z}_N\|^2 \\
&= \sum_{i=0}^{N} \left\{ \int_{x_{i-1}}^{x_i} \left(\frac{d\tilde{z}(x)}{dx} - \frac{d\hat{z}_N(x)}{dx} \right)^2 dx + \frac{v}{\mu}\left(\frac{d\tilde{z}(x)}{dx} - \frac{d\hat{z}_N(x)}{dx} \right)\left\{\tilde{z}(x) - \hat{z}_N(x)\right\} dx \right\} \\
&+ \left(\frac{v^2}{4\mu} + 1 \right) \|\tilde{z} - \hat{z}_N\|^2 \\
&\leq \|\tilde{z}' - \hat{z}_N'\|^2 + \frac{v}{\mu}\|\tilde{z}' - \hat{z}_N'\| \|\tilde{z} - \hat{z}_N\| + \left(\frac{v^2}{4\mu} + 1 \right)\|\tilde{z} - \hat{z}_N\|^2 \\
&\leq C^2 \left(\frac{1}{N^2} + \frac{v}{\mu}\frac{1}{N^3} + \frac{v^2+1}{4\mu}\frac{1}{N^4} \right).
\end{aligned}$$

To summarize,

$$\left\| z_\sigma - \frac{1}{\sqrt{\sigma}}\hat{z}_N \right\|_{1,\sigma}^2 \leq \frac{C^2}{N^2}\left(1 + \frac{v}{\mu} + \frac{v^2+1}{4\mu} \right).$$

Define the constant $\tilde{C}$ by

$$\frac{\tilde{C}^2}{4} = C^2\left(1 + \frac{v}{\mu} + \frac{v^2+1}{4\mu} \right).$$

Then, we see that

$$\left\| z_\sigma - \frac{1}{\sqrt{\sigma}}\hat{z}_N \right\|_{1,\sigma} \leq \frac{\tilde{C}}{2N}.$$

If z^N denotes the orthogonal projection of z into V_σ^N, then

$$\left\| z - z^N \right\|_{1,\sigma} \leq \left\| z - \frac{1}{\sqrt{\sigma}}\hat{z}_N \right\|_{1,\sigma} \leq \|z - z_\sigma\|_{1,\sigma} + \|z_\sigma - \hat{z}_N\|_{1,\sigma} \leq \frac{\tilde{C}}{N}$$

since $\sqrt{\sigma}\hat{z}_N \in V_\sigma^N$. Thus, we have

$$\left\| z - z^N \right\|_{1,\sigma} \leq \epsilon(N), \tag{2.108}$$

where $\epsilon(N) = \frac{\tilde{C}}{N}$. □

Let A_0 and A_σ^N be the operators defined in (2.4) and (2.83), respectively. We have the following result.

Theorem 2.4.3 *Let $S_1(t)$ be the C_0-semigroup on H_σ generated by A_0, and let $S_\sigma^N(t)$ be the sequence of C_0-semigroups on V_σ^N generated by A_σ^N. If P_σ^N denotes the orthogonal projection from H_σ into V_σ^N, then, for every $z \in H_\sigma$ and $t \geq 0$,*

$$\left\| S^N(t) P_\sigma^N z - S_1(t) z \right\|_\sigma \to 0 \tag{2.109}$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Furthermore, for every $z \in H_\sigma$ and $t \geq 0$,

$$\left\| \left(S^N\right)(t)^* P_\sigma^N z - \left(S_1\right)(t)^* z \right\|_\sigma \to 0$$

as $N \to \infty$, and the convergence is uniform on bounded $t-$intervals.

Proof: The proof is the same as the proof for Theorem 2.4.2 with the substitution of $\epsilon(N)$ for $\epsilon_2(N)$.

2.5 Optimal Control of the Approximate Equations

In this section, we employ the approximation schemes from Sections 2.3.1 and 2.3.5 to compute feedback gains for the linear quadratic regulator problem $(\mathbf{\Sigma})$ defined in Section 2.1.

We first define a linear quadratic regulator problem that is equivalent to $(\mathbf{\Sigma})$. In order to do this, we define the following operators.

Consider the operators $B_\sigma : \mathbb{R} \to H_\sigma$, $Q_\sigma : H_\sigma \to H_\sigma$, $R_\sigma : \mathbb{R} \to \mathbb{R}$ where

$$B_\sigma u = b(x)u, \quad \forall u \in \mathbb{R} \tag{2.110}$$

$$Q_\sigma w = q\sigma^{-1} w, \quad \forall w \in H_\sigma \tag{2.111}$$

$$R_\sigma u = ru, \quad \forall u \in \mathbb{R}. \tag{2.112}$$

Recall that the convection-diffusion operator is the operator A_0 with domain

$$\mathcal{D}(A_0) = H_0^1(0,1) \bigcap H^2(0,1)$$

so that

$$A_0 w = \mu \frac{\partial^2 w}{\partial x^2} - v \frac{\partial w}{\partial x}, \quad \forall w \in \mathcal{D}(A_0). \tag{2.113}$$

We introduce the notation $A_{\sigma,0}$ to denote the convection-diffusion operator on $H_\sigma = L^2_\sigma(0,1)$. Define $A_{\sigma,0}$ on

$$\mathcal{D}(A_{\sigma,0}) = H^1_{\sigma,0}(0,1) \bigcap H^2_\sigma(0,1)$$

by

$$A_{\sigma,0} w = \mu \frac{\partial^2 w}{\partial x^2} - v \frac{\partial w}{\partial x}, \quad \forall w \in \mathcal{D}(A_{\sigma,0}). \tag{2.114}$$

Using this framework, we pose the following optimal control problem, $(\mathbf{\Sigma}_\sigma)$, that is equivalent to $(\mathbf{\Sigma})$.

$(\mathbf{\Sigma}_\sigma)$ *Minimize* $J_\sigma(z_0, u_\sigma)$ over $u_\sigma \in L^2((0,\infty); \mathbb{R})$ *where*

$$\frac{dz_\sigma(t)}{dt} = A_{\sigma,0} z_\sigma(t) + B_\sigma u_\sigma(t), \tag{2.115}$$

$$z_\sigma(0) = z_0 \tag{2.116}$$

and

$$J_\sigma(z_0, u_\sigma) = \int_0^\infty \{ < Q_\sigma z_\sigma(t), z_\sigma(t) >_\sigma + \langle R_\sigma u_\sigma(t), u_\sigma(t) \rangle_{\mathbb{R}} \} \, dt. \tag{2.117}$$

We approximate $(\mathbf{\Sigma})$ and $(\mathbf{\Sigma}_\sigma)$ in two different ways. The first approximation technique uses the traditional approximation scheme discussed in Section 2.3.1 to approximate A_0. Orthogonal projections are used to approximate B and Q in the traditional approximation scheme. We also use the third approximation scheme outlined in Section 2.3.5 to approximate $A_{\sigma,0}$. Here, orthogonal projections into V^N_σ are used to approximate B_σ and Q_σ. We do not use the second approximation scheme discussed in Section 2.3.2 because of numerical difficulties that we discuss in the next chapter.

2.5.1 Approximating the Linear Quadratic Regulator Problem $(\mathbf{\Sigma})$

In this section, we define a linear quadratic regulator problem to approximate $(\mathbf{\Sigma})$. Let $P^N : H \to V^N$ denote the orthogonal projection from H into V^N. Let A^N_1 be the operator on V^N defined in (2.48). Furthermore, define the operators $B^N_1 : \mathbb{R} \to V^N$, $Q^N_1 : V^N \to V^N$, $R^N_1 : \mathbb{R} \to \mathbb{R}$ by

$$B^N_1 = P^N B, \tag{2.118}$$

$$Q^N_1 = P^N Q P^N, \tag{2.119}$$

$$R^N_1 = R. \tag{2.120}$$

Consider the following approximate optimal control problem $(\mathbf{\Sigma_1^N})$:

$(\mathbf{\Sigma_1^N})$ *Minimize* $J_1^N(z_1^N(0), u_1^N)$ *over* $u_1^N \in L^2((0,\infty); \mathbb{R})$ *where*

$$\frac{dz_1^N(t)}{dt} = A_1^N z_1^N(t) + B_1^N u_1^N(t), \tag{2.121}$$

$$z_1^N(0) = P^N z_0 \tag{2.122}$$

and

$$J_1^N(z_1^N(0), u) = \int_0^\infty \left\{ \langle Q_1^N z_1^N(t), z_1^N(t) \rangle + \langle R_1^N u_1^N(t), u_1^N(t) \rangle_{\mathbb{R}} \right\} dt. \tag{2.123}$$

We use $(\mathbf{\Sigma_1^N})$ to approximate $(\mathbf{\Sigma})$.

2.5.2 Approximating the Linear Quadratic Regulator Problem $(\mathbf{\Sigma}_\sigma)$

The objective of this section is to approximate the linear quadratic regulator problem $(\mathbf{\Sigma}_\sigma)$. Let $P_\sigma^N : H_\sigma \to V_\sigma^N$ denote the orthogonal projection from H_σ into V_σ^N. Let A_σ^N be the operator on V_σ^N defined in (2.83). Moreover, let the operators $B_\sigma^N : \mathbb{R} \to V_\sigma^N$, $Q_\sigma^N : V_\sigma^N \to V_\sigma^N$, $R_\sigma^N : \mathbb{R} \to \mathbb{R}$ be defined such that

$$B_\sigma^N = P_\sigma^N B_\sigma, \tag{2.124}$$

$$Q_\sigma^N = P_\sigma^N Q_\sigma P_\sigma^N, \tag{2.125}$$

$$R_\sigma^N = R_\sigma. \tag{2.126}$$

We approximate the linear quadratic regulator problem $(\mathbf{\Sigma}_\sigma)$ in the following manner.

$(\mathbf{\Sigma_\sigma^N})$ *Minimize* $J_\sigma^N(z_\sigma^N(0), u_\sigma^N)$ *over* $u_\sigma^N \in L^2((0,\infty); \mathbb{R})$ *where*

$$\frac{dz_\sigma^N(t)}{dt} = A_\sigma^N z_\sigma^N(t) + B_\sigma^N u_\sigma^N(t), \tag{2.127}$$

$$z_\sigma^N(0) = P_\sigma^N z_0 \tag{2.128}$$

and

$$J_\sigma^N(z_\sigma^N(0), u_\sigma^N) = \int_0^\infty \left\{ < Q_\sigma^N z_\sigma^N(t), z_\sigma^N(t) >_\sigma + \langle R_\sigma^N u_\sigma^N(t), u_\sigma^N(t) \rangle_{\mathbb{R}} \right\} dt. \tag{2.129}$$

2.5.3 Applying Theorem 1.4.1 to $(\mathbf{\Sigma_1})$ and $(\mathbf{\Sigma_1^N})$

The following result is a specific case of Theorem 3.1 in [7].

Theorem 2.5.1 *For problems* $(\mathbf{\Sigma})$ *and* $(\mathbf{\Sigma_1^N})$*, there exist unique Riccati operators* Π_1 *and* Π_1^N *such that if* P^N *denotes the orthogonal projection into* V^N*,*

$$\Pi_1^N P^N z \to \Pi_1 z, \forall z \in H \tag{2.130}$$

$$T_1^N(t) P^N z \to T_1(t) z, \quad \forall z \in H \tag{2.131}$$

$$\text{and} \qquad \hat{u}_1^N(t) \to \hat{u}_1(t), \tag{2.132}$$

with these last two statements holding uniformly in t *on compact subsets of* $[0, \infty)$*.* $T_1^N(t)$ *and* $T_1(t)$ *are the semigroups generated by* $A_1^N - B_1^N (R_1^N)^{-1} B_1^{N*} \Pi_1^N$ *and* $A - BR^{-1}B^*\Pi$*, and* $\hat{u}_1(t)$ *and* $\hat{u}_1^N(t)$ *are the optimal feedback controls for* $(\mathbf{\Sigma})$ *and* $(\mathbf{\Sigma_1^N})$*.*

2.5.4 Applying Theorem 1.4.1 to $(\mathbf{\Sigma}_\sigma)$ and $(\mathbf{\Sigma_\sigma^N})$

In this section, we apply Theorem 1.4.1 to prove a result concerning $(\mathbf{\Sigma}_\sigma)$ and $(\mathbf{\Sigma_\sigma^N})$.

Theorem 2.5.2 *For problems* $(\mathbf{\Sigma}_\sigma)$ *and* $(\mathbf{\Sigma_\sigma^N})$*, there exist unique Riccati operators* Π_σ *and* Π_σ^N *such that if* P_σ^N *denotes the orthogonal projection into* V_σ^N*,*

$$\Pi_\sigma^N P_\sigma^N z \to \Pi_\sigma z, \forall z \in H_\sigma \tag{2.133}$$

$$T_\sigma^N(t) P_\sigma^N z \to T_\sigma(t) z, \quad \forall z \in H_\sigma \tag{2.134}$$

$$\text{and} \qquad \hat{u}_\sigma^N(t) \to \hat{u}_\sigma(t), \tag{2.135}$$

with these last two statements holding uniformly in t *on compact subsets of* $[0, \infty)$*.* $T_\sigma^N(t)$ *and* $T_\sigma(t)$ *are the semigroups generated by* $A_\sigma^N - B_\sigma^N (R_\sigma^N)^{-1} B_\sigma^{N*} \Pi_\sigma^N$ *and* $A_\sigma - B_\sigma {R_\sigma}^{-1} {B_\sigma}^* \Pi_\sigma$*, and* $\hat{u}_\sigma(t)$ *and* $\hat{u}_\sigma^N(t)$ *are the optimal feedback controls for* $(\mathbf{\Sigma}_\sigma)$ *and* $(\mathbf{\Sigma_\sigma^N})$*.*

Proof: To prove this theorem, we show that the systems $(\mathbf{\Sigma}_\sigma)$ and $(\mathbf{\Sigma_\sigma^N})$ satisfy conditions (H1), (H2), and (C1)-(C3) from Section 1.4.1. Once these conditions are satisfied, the proof follows as a direct result of Theorem 1.4.1.

In Section 2.3.5, we show that A_σ^N generates a uniformly continuous semigroup on V_σ^N. Thus, A_σ^N generates a C_0-semigroup on V_σ^N. We first show that (H1) is satisfied.

The eigenvalues of A_σ^N are real, negative, and bounded away from zero. For all eigenvalues λ_i^N of A_σ^N, there exists a positive constant M such that

$$\lambda_i^N < -M < 0, \quad i, N = 1, 2, \ldots.$$

Thus, A_σ^N generates an exponentially stable semigroup. Hence, for any z_0 in V_σ^N, there exists a control $u_\sigma^N = 0$ such that u_σ^N drives the state of $(\mathbf{\Sigma_\sigma^N})$ to zero asymptotically, and condition (H1) is satisfied for $(\mathbf{\Sigma_\sigma^N})$.

Lemma 2.4.4 shows that condition (C1) holds for this scheme. In Section 2.4.3, we prove that this approximation scheme satisfies parts (i) and (ii) of (H2). In addition, by defining B_σ^N and Q_σ^N in (2.124) and (2.125), (H2) (iii) and (H2) (iv) hold because of condition (C1).

Since $A_{0,\sigma}$ generates an exponentially stable semigroup on H_σ, the semigroup generated by $(A_{\sigma,0} - B_\sigma K)$ is stable when $K = 0$. Thus, $(A_{\sigma,0}, B_\sigma)$ is stabilizable. Moreover, (A_σ^N, B_σ^N) is stabilized by $K = 0$. Thus, conditions (C2) and (C3) hold.

Since we have shown that $(\mathbf{\Sigma}_\sigma)$ and $(\mathbf{\Sigma}_\sigma^\mathbf{N})$ satisfy conditions (H1), (H2), and (C1)-(C3) in Theorem 1.4.1, application of Theorem 1.4.1 completes the proof.□

2.6 Convergence of the Feedback Gains

In this section, we explain how the results from Sections 2.5.3 and 2.5.4 imply convergence of the resulting functional gains.

Theorem 1.2.3 states that if $\hat{u}_\sigma$ and $\hat{z}_\sigma$ denote the optimal control and resulting state for $(\mathbf{\Sigma}_\sigma)$, respectively, then

$$\hat{u}_\sigma(t) = -K_\sigma \hat{z}_\sigma(t) = -R_\sigma^{-1} B_\sigma^* \Pi_\sigma \hat{z}_\sigma(t) \tag{2.136}$$

where the Riccati operator Π_σ is the unique positive definite solution of the ARE

$$A_{\sigma,0}^* \Pi_\sigma + \Pi_\sigma A_{\sigma,0} - \Pi_\sigma B_\sigma R_\sigma^{-1} B_\sigma^* \Pi_\sigma + Q_\sigma = 0.$$

Since Π_σ is a bounded linear operator on H_σ and B_σ^* is a bounded operator of finite rank from H_σ into $\mathbb{R}$, we have that K_σ is a bounded operator of finite rank. Thus, K_σ is a compact operator.

Moreover, if $\hat{u}_\sigma^N$ and $\hat{z}_\sigma^N$ denote the optimal control and resulting state for $(\mathbf{\Sigma}_\sigma^\mathbf{N})$, respectively, then

$$\hat{u}_\sigma^N(t) = -K_\sigma^N \hat{z}_\sigma^N(t) = -(R_\sigma^N)^{-1} B_\sigma^{N*} \Pi_\sigma^N \hat{z}_\sigma^N(t) \tag{2.137}$$

where Π_σ^N is the solution to the ARE

$$(A_\sigma^N)^* \Pi_\sigma^N + \Pi_\sigma^N A_\sigma^N - \Pi_\sigma^N B_\sigma^N (R_\sigma^N)^{-1} B_\sigma^{N*} \Pi_\sigma^N + Q_\sigma^N = 0. \tag{2.138}$$

Additionally, since Π_σ^N is a bounded linear operator on H_σ^N and B_σ^{N*} is a bounded operator of finite rank from H_σ^N into $\mathbb{R}$, it follows that K_σ^N is a compact operator.

From Theorem 2.5.2, we know that for all z in H_σ,

$$\Pi_\sigma^N P_\sigma^N z \to \Pi_\sigma z. \tag{2.139}$$

It follows from (2.136), (2.137), and (2.139) that for all z in H_σ,

$$K_\sigma^N P_\sigma^N z \to K_\sigma z. \tag{2.140}$$

Hence, we have pointwise convergence of the sequence of compact operators $K_\sigma^N P_\sigma^N$ to the compact operator K_σ. The convergence is uniform, that is,

$$\left\|K_\sigma^N P_\sigma^N - K_\sigma\right\|_{\mathcal{L}(H_\sigma,\mathrm{I\!R})} \to 0. \tag{2.141}$$

Since K_σ is a bounded linear functional on H_σ and K_σ^N is a bounded linear functional on V_σ^N, there exist functional gains k_σ in H_σ and k_σ^N in V_σ^N such that

$$K_\sigma z = < k_\sigma, z >_\sigma, \quad \text{for all} \quad z \in H_\sigma \tag{2.142}$$

$$K_\sigma^N P_\sigma^N z = < k_\sigma^N, z >_\sigma, \quad \text{for all} \quad z \in H_\sigma. \tag{2.143}$$

We use the following result to show that

$$\left\|k_\sigma^N - k_\sigma\right\|_\sigma \to 0.$$

Lemma 2.6.1 *Let K_σ be a bounded linear operator from H_σ into $\mathrm{I\!R}$, and let $K_\sigma^N P_\sigma^N$ be a sequence of of bounded linear operators from H_σ into $\mathrm{I\!R}$. Let k_σ in H_σ and k_σ^N in V_σ^N be defined as in* (2.142), *and* (2.143). *Furthermore, assume* (2.141) *holds. Then, k_σ^N converges in norm to k_σ i.e.,*

$$\left\|k_\sigma^N - k_\sigma\right\|_\sigma \to 0.$$

Proof: If $\left\|k_\sigma^N - k_\sigma\right\|_\sigma \neq 0$, then

$$\begin{aligned}
\left\|k_\sigma^N - k_\sigma\right\|_\sigma &= \frac{\left\|k_\sigma^N - k_\sigma\right\|_\sigma^2}{\left\|k_\sigma^N - k_\sigma\right\|_\sigma} \\
&= \frac{\left|\left(K_\sigma^N P_\sigma^N - K_\sigma\right)\left(k_\sigma^N - k_\sigma\right)\right|}{\left\|k_\sigma^N - k_\sigma\right\|_\sigma} \\
&\leq \sup_{z\in H_\sigma} \frac{\left|\left(K_\sigma^N P_\sigma^N - K_\sigma\right) z\right|}{\left\|z\right\|_\sigma} \\
&\leq \left\|K_\sigma^N P_\sigma^N - K_\sigma\right\|_{\mathcal{L}(H_\sigma,\mathrm{I\!R})}.
\end{aligned}$$

Thus, we have the estimate

$$\left\|k_\sigma^N - k_\sigma\right\|_\sigma \leq \left\|K_\sigma^N P_\sigma^N - K_\sigma\right\|_{\mathcal{L}(H_\sigma,\mathrm{I\!R})} \to 0.\Box$$

We conclude that the functional gains for $(\mathbf{\Sigma_\sigma^N})$ converge in norm.

Let K_1 denote the feedback operator for $(\mathbf{\Sigma})$, and let K_1^N be the feedback opertors for $(\mathbf{\Sigma_1^N})$. By the same reasoning as above, $K_1^N P^N$ converges uniformly to K_1 and there exist functions k_1 in H and k_1^N in V^N such that

$$K_1 z = < k, z >_\sigma, \quad \text{for all} \quad z \in H \tag{2.144}$$

$$K_1^N P^N z = < k_1^N, P^N z >_\sigma, \quad \text{for all} \quad z \in H. \tag{2.145}$$

and

$$\left\| k_1^N - k_1 \right\| \to 0.$$

Thus, the functional gains for $(\mathbf{\Sigma^N})$ converge in norm.

Chapter 3

Numerical Results

In this chapter, we discuss numerical issues associated with computation of eigenvalues for all three approximation schemes. Eigenvalues are computed and presented for specific examples. Furthermore, we address methods to compute the solutions to the matrix AREs associated with $(\mathbf{\Sigma_1^N})$ and $(\mathbf{\Sigma_\sigma^N})$. We present two constant scaling techniques and a new preconditioning method, all of which lead to improvements when computing solutions to the matrix AREs.

3.1 Matrix Representations of the Three Approximation Schemes

Let A_1^N, A_2^N, and A_σ^N be the operators defined by the first, second, and third approximation schemes in Sections 2.3.1, 2.3.2, and 2.3.5, respectively. In this section, we present methods to compute the matrix representations of the operators A_1^N, A_2^N, and A_σ^N. Furthermore, we demonstrate methods to compute the matrix representations of the operators B_1^N, B_σ^N, Q_1^N and Q_σ^N from the previous chapter.

The mass matrix $[M_1^N]$ for the first approximation scheme is

$$[M_1^N] = [< h_i, h_j >]_{i,j=1}^{N-1} = \frac{1}{6N} \begin{bmatrix} 4 & 1 & 0 & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots \\ & 0 & 1 & 4 \end{bmatrix}. \tag{3.1}$$

The stiffness matrix $[K_1^N]$ is

$$\begin{aligned}[K_1^N] &= [-\mu < h_i', h_j' > - v < h_i', h_j >]_{i,j=1}^{N-1} \\ &= -\frac{\mu}{N}\begin{bmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & 0 & -1 & 2 & \end{bmatrix} + \frac{v}{2}\begin{bmatrix} 0 & -1 & 0 & & \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & 0 & 1 & 0 & \end{bmatrix}. \end{aligned} \tag{3.2}$$

The matrix representation of A_1^N is given by

$$[A_1^N] = [M_1^N]^{-1}[K_1^N].$$

For the second approximation scheme,

$$[M_2^N] = [< h_i, h_j >_\sigma]_{i,j=1}^{N-1}, \tag{3.3}$$

and

$$[K_2^N] = [- < h_i', h_j' >_\sigma]_{i,j=1}^{N-1} = [< \sigma h_i, h_j >]_{i,j=1}^{N-1}. \tag{3.4}$$

The matrix representation of A_2^N is

$$[A_2^N] = [M_2^N]^{-1}[K_2^N].$$

Recall that σ is a rapidly decreasing function when μ is small in comparison to v. Hence, it is difficult to accurately compute $[A_2^N]$ and $[M_2^N]$ when μ is small in comparison to v. Furthermore, $[M_2^N]$ is nearly singular and ill-conditioned when μ is small compared to v. For example, when $v = 1, \mu = 1/100$, and $N = 64$, MATLAB computes that $[M_2^N]$ has a 2-norm condition number on the order of 10^{42}. Thus, computing $[M_2^N]^{-1}$ to find $[A_2^N]$ can introduce additional large errors.

For the last approximation scheme,

$$[M_3^N] = [< \phi_i, \phi_j >_\sigma]_{i,j=1}^{N-1} = [< h_i, h_j >]_{i,j=1}^{N-1}. \tag{3.5}$$

Thus, $[M_3^N] = [M_1^N]$, and

$$[K_3^N] = [- < \phi_i', \phi_j' >_\sigma]_{i,j=1}^{N-1} \tag{3.6}$$

$$= [-\mu < h_i', h_j' > - \frac{v^2}{4\mu} < h_i, h_j >]_{i,j=1}^{N-1} \tag{3.7}$$

$$= -\frac{\mu}{N}\begin{bmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & 0 & -1 & 2 & \end{bmatrix} - \frac{v^2}{(6N)(4\mu)}\begin{bmatrix} 4 & 1 & 0 & & \\ 1 & 4 & 1 & \ddots & \\ & \ddots & \ddots & \ddots & \\ & 0 & 1 & 4 & \end{bmatrix}. \tag{3.8}$$

Hence,

$$[A_\sigma^N] = [M_3^N]^{-1}[K_3^N].$$

The matrices $[\hat{B}_1^N]$ and $[\hat{B}_\sigma^N]$ are defined by

$$[\hat{B}_1^N] = [< b, h_i >]_{i=1}^{N-1}$$

and

$$[\hat{B}_\sigma^N] = [< b(, \phi_j >_\sigma]_{j=1}^{N} = [< \sqrt{\sigma} b, h_j >]_{j=1}^{N-1}.$$

Thus,

$$[B_1^N] = [M_1^N]^{-1}[\hat{B}_1^N],$$

and

$$[B_\sigma^N] = [M_3^N]^{-1}[\hat{B}_\sigma^N].$$

The matrix $[Q_1^N]$ is defined by

$$[Q_1^N] = q[M_1^N]$$

and

$$[Q_\sigma^N] = (q\mu)[< \sigma^{-1} h_i, h_j >]_{i,j=1}^{N-1}.$$

Computation of the above matrices requires integration. We first evaluate all the integrals by hand in order to eliminate any error introduced by numerical quadrature.

3.2 Open Loop Eigenvalues of the Matrix Representations

For all the matrix representations in this section, we let $v = 1$. We vary the size of μ and N.

Figures 3.1 - 3.5 show the eigenvalues for $[A_1^N], [A_2^N]$ and $[A_\sigma^N]$ when $\mu = 1$. For all values of N, all the matrices have only real, negative eigenvalues.

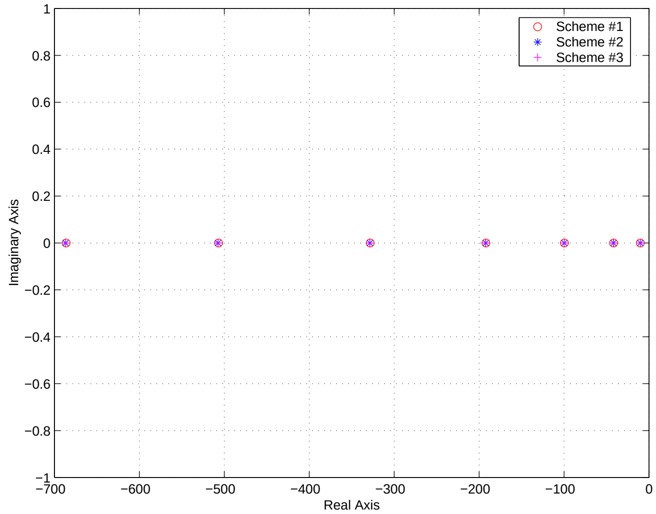

Figure 3.1: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1, \quad N = 8$.

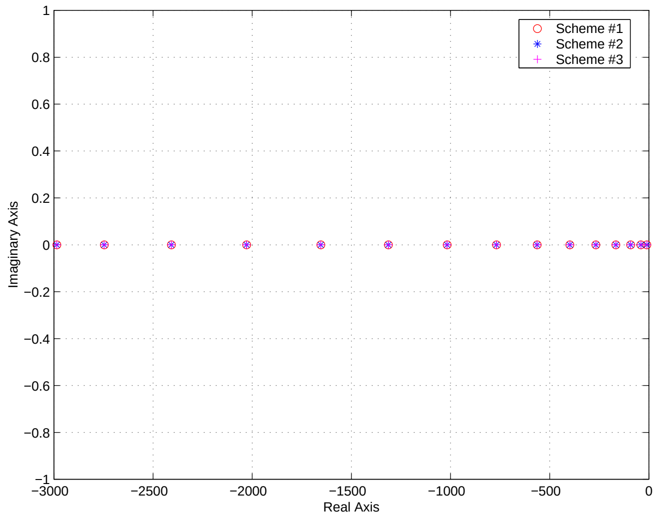

Figure 3.2: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1, \quad N = 16$.

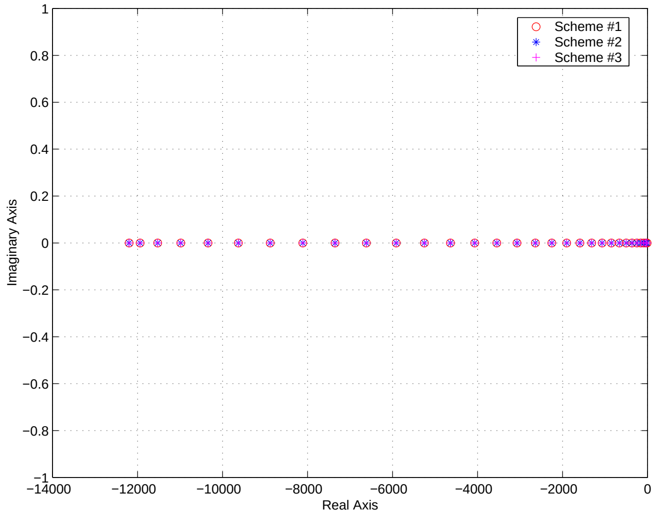

Figure 3.3: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1, \quad N = 32$.

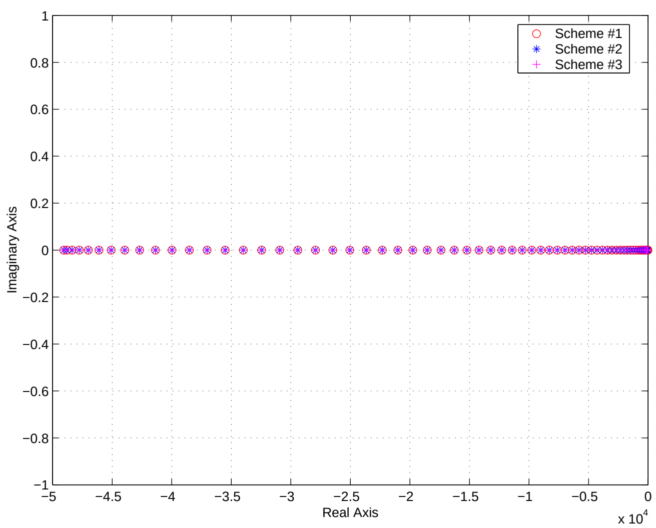

Figure 3.4: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1, \quad N = 64$.

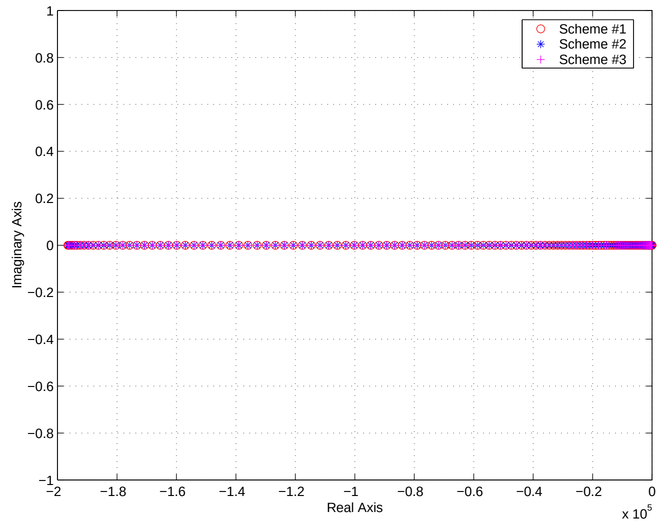

Figure 3.5: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1, \quad N = 128$.

Table 3.1: The Ten Largest Eigenvalues of $[A_1^N], [A_2^N]$ and $[A_\sigma^N]$: $\mu = 1, \quad N = 128$

$[A_1^N]$	$[A_2^N]$	$[A_\sigma^N]$
$-0.00010120024882 \times 10^5$	$-0.00010120125276 \times 10^5$	$-0.00010120099859 \times 10^5$
$-0.00039736044595 \times 10^5$	$-0.00039736446134 \times 10^5$	$-0.00039736345413 \times 10^5$
$-0.00089115901099 \times 10^5$	$-0.00089116804402 \times 10^5$	$-0.00089116578167 \times 10^5$
$-0.00158289342380 \times 10^5$	$-0.00158290947858 \times 10^5$	$-0.00158290545881 \times 10^5$
$-0.00247298040596 \times 10^5$	$-0.00247300548383 \times 10^5$	$-0.00247299920396 \times 10^5$
$-0.00356195616959 \times 10^5$	$-0.00356199226731 \times 10^5$	$-0.00356198322515 \times 10^5$
$-0.00485047673560 \times 10^5$	$-0.00485052584606 \times 10^5$	$-0.00485051353833 \times 10^5$
$-0.00633931832053 \times 10^5$	$-0.00633938243094 \times 10^5$	$-0.00633936635418 \times 10^5$
$-0.00802937779033 \times 10^5$	$-0.00802945888180 \times 10^5$	$-0.00802943853185 \times 10^5$
$-0.00992167317941 \times 10^5$	$-0.00992177322491 \times 10^5$	$-0.00992174809813 \times 10^5$

Table 3.1 lists the 10 largest eigenvalues for each scheme when $N = 128$. These eigenvalues agree very closely in magnitude. When $\mu = 1$ and $v = 1$, the approximation schemes yield nearly identical sets of eigenvalues. However, when $\mu = 1/75$, the approximation schemes do not all have real eigenvalues.

In Figures 3.6 - 3.8, $[A_1^N]$ has complex eigenvalues, but $[A_2^N]$ and $[A_\sigma^N]$ have only real eigenvalues. When N increases, the more negative eigenvalues of $[A_1^N]$ approach the real axis. When $N = 64$ in Figure 3.9, all three matrices have only real eigenvalues. However, when $N = 128$

in Figure 3.10, the second approximation scheme has complex eigenvalues. This result is counterintuitive because we expect that as we increase N, the accuracy of computing the eigenvalues would increase.

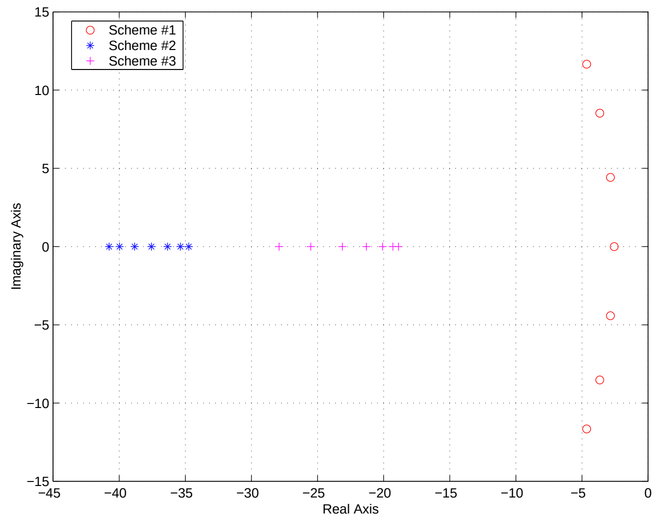

Figure 3.6: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/75, \quad N = 8$.

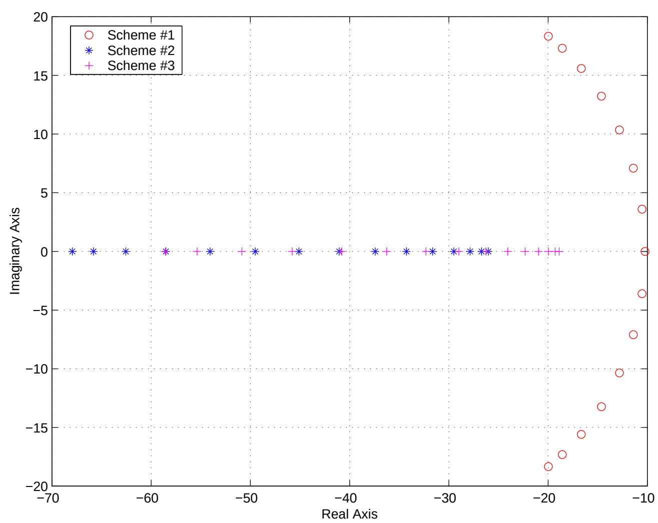

Figure 3.7: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/75, \quad N = 16$.

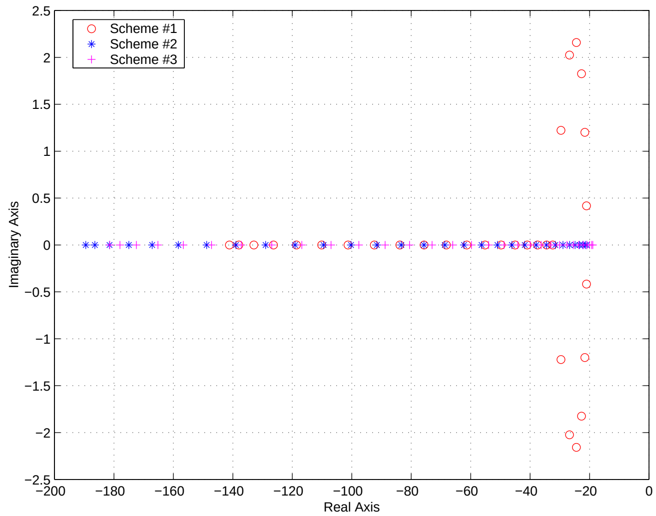

Figure 3.8: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/75, \quad N = 32$.

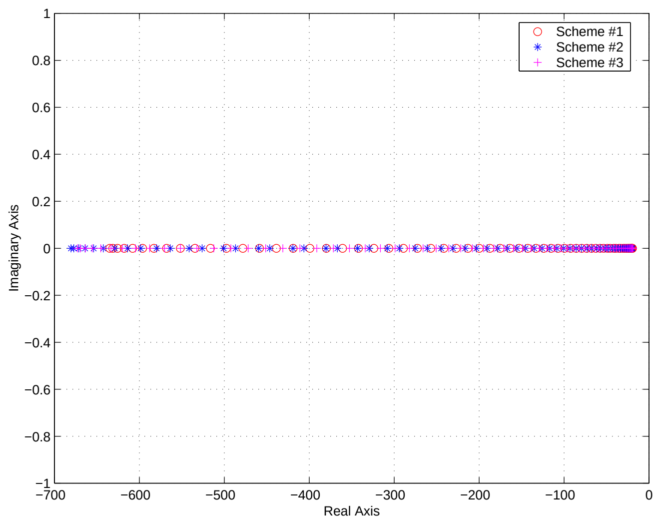

Figure 3.9: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/75, \quad N = 64$.

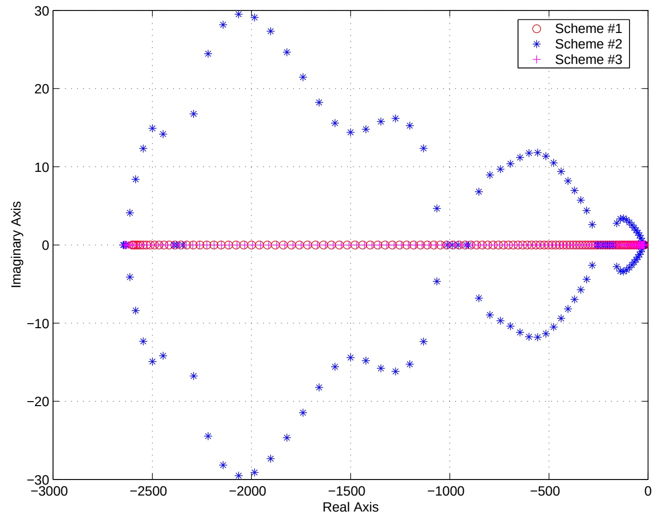

Figure 3.10: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/75, \quad N = 128$.

Figures 3.11 - 3.13 display the eigenvalues of $[A_1^N], [A_2^N]$ and $[A_\sigma^N]$ when $\mu = 1/100$ and $N = 8, 16, 32$. Again, $[A_1^N]$ has complex eigenvalues, but $[A_2^N]$ and $[A_\sigma^N]$ have only real eigenvalues. However, when N is increased to 64 and 128 in Figures 3.14 and 3.15, the third approximation scheme is the only scheme that does not have complex eigenvalues. As N increases, the eigenvalues for the first scheme move closer to the real axis, but the second scheme has eigenvalues that move farther and farther from the real axis.

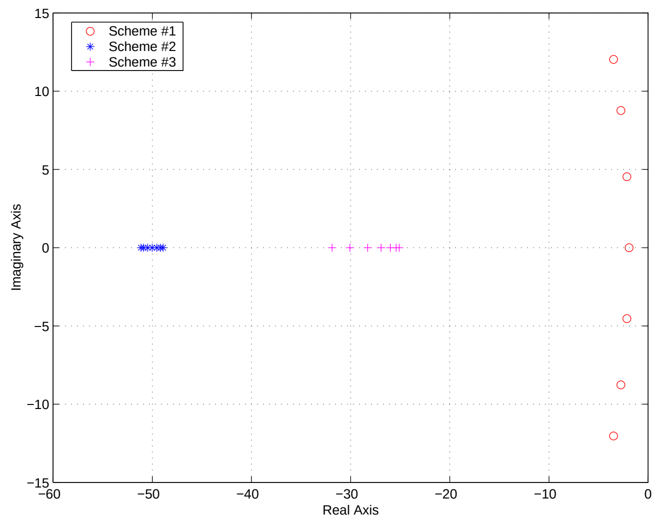

Figure 3.11: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/100, \quad N = 8$.

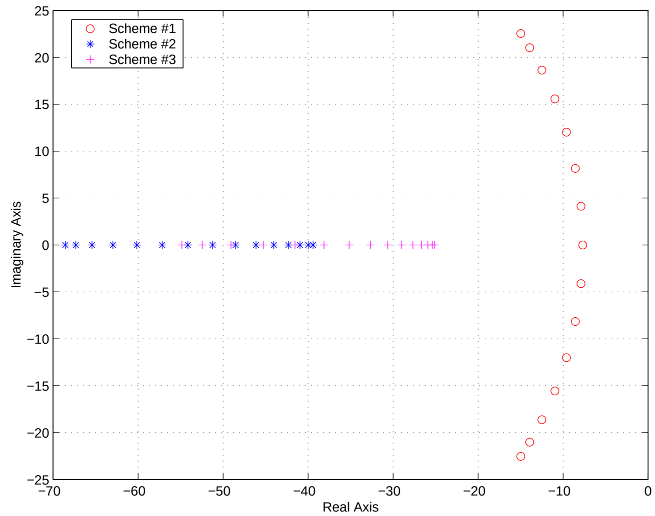

Figure 3.12: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/100, \quad N = 16$.

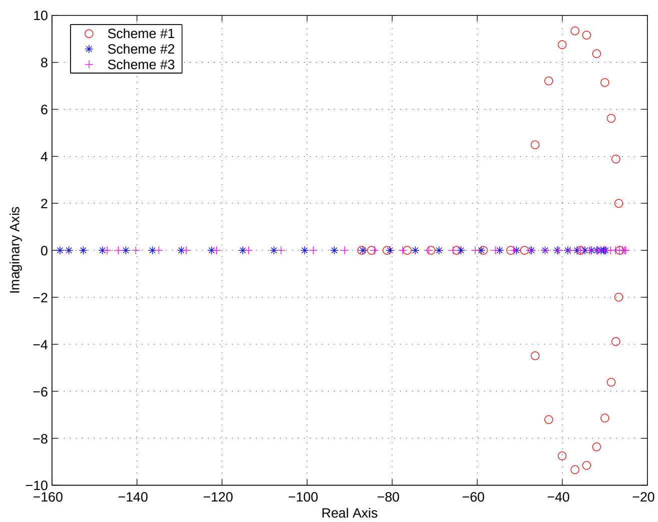

Figure 3.13: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/100, \quad N = 32$.

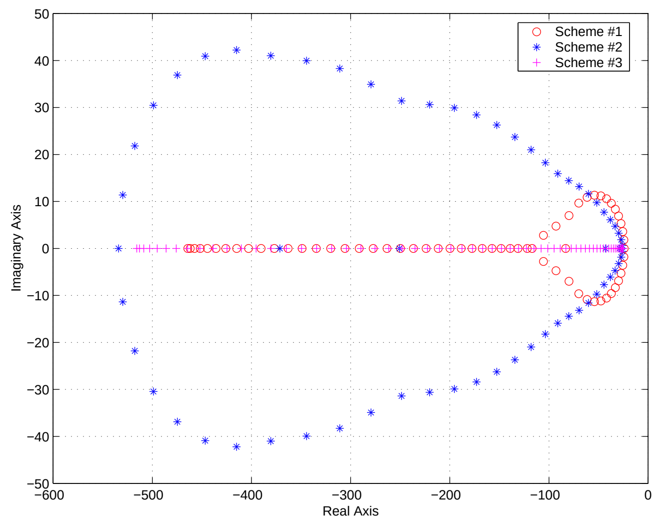

Figure 3.14: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/100, \quad N = 64$.

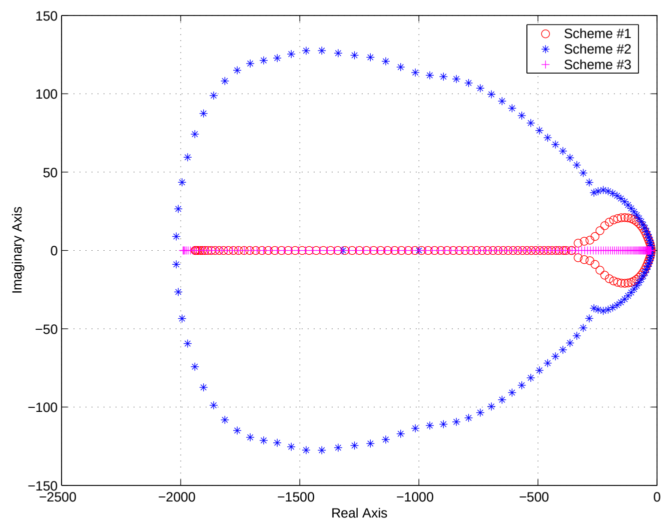

Figure 3.15: Eigenvalues of $[A_1^N], [A_2^N]$, and $[A_\sigma^N]$: $\mu = 1/100, \quad N = 128$.

As mentioned previously, it is difficult to accurately compute the mass and stiffness matrices for the second scheme when μ is small in comparison to v. Furthermore, $[A_2^N]$ is ill-conditioned. When $\mu = 1/100$ and $N = 64$, $[A_2^N]$ has 1-norm condition number on the order of 10^{43}. As a result, the eigenvalue problem is ill-conditioned. Thus, the eigenvalues computed for $[A_2^N]$ may not be close to the real eigenvalues of $[A_2^N]$.

However, $[A_1^N]$ is a much better conditioned matrix, and the associated eigenvalue problem is better conditioned. Thus, we have confidence that the eigenvalues that we compute are close to the real eigenvalues of $[A_1^N]$. $[A_1^N]$ has complex eigenvalues because the stiffness matrix $[K_1^N]$ is not symmetric.

For all values of μ and N, the third approximation scheme always has real eigenvalues. This property can be explained by the structure of the matrix representation of $[A_\sigma^N]$.

Recall that

$$[A_\sigma^N] = [M_3^N]^{-1}[K_3^N] \tag{3.9}$$

$$= [M_3^N]^{-1}\left([-\mu < h_i', h_j' >]_{i,j=1}^{N-1} - \frac{v^2}{4\mu}[< h_i, h_j >]_{i,j=1}^{N-1}\right) \tag{3.10}$$

$$= [M_3^N]^{-1}[-\mu < h_i', h_j' >]_{i,j=1}^{N-1} - \frac{v^2}{4\mu}[I_N]. \tag{3.11}$$

Thus,

$$[A_\sigma^N] = [H_N] - v^2/(4\mu)[I_N],$$

where $[I_N]$ is the $N-1 \times N-1$ identity matrix and

$$[H_N] = [M_3^N]^{-1}[-\mu < h_i', h_j' >]_{i,j=1}^{N-1}].$$

Note that $[H_N]$ is the matrix representation of a symmetric bilinear operator. Furthermore, $[H_N]$ is a symmetric, real matrix. As a result, this matrix has only real eigenvalues. The eigenvalues of $[A_N]$ are the eigenvalues of $[H_N]$, but translated by $-v^2/(4\mu)$. As mentioned at the beginning of Chapter 2, the eigenvalues of the convection-diffusion operator are the eigenvalues of the heat equation operator translated by $-v^2/(4\mu)$. For this reason, the third approximation scheme may be a more desirable approximation scheme when computing eigenvalues.

3.3 Preconditioning the Algebraic Riccati Equations

Due to the difficulty of computing $[A_2^N]$ accurately, we do not use the second approximation scheme to formulate a matrix ARE. We do, however, use the first and third approximation schemes to develop matrix AREs.

Consider the matrix ARE.

$$[A_1^N]^T[\Pi_1^N] + [\Pi_1^N][A_1^N] - [\Pi_1^N][B_1^N][R_1^N]^{-1}[B_1^N]^T[\Pi_1^N] + [Q_1^N] = 0, \tag{3.12}$$

where the matrices $[A_1^N], [B_1^N], [Q_1^N]$, and $[R_1^N]$ are the matrices defined in Section 3.1. These matrices are developed using the first approximation method outlined in Section 2.3.1. The ARE command in Matlab returns a solution to (3.12) for a large range of values of μ and N. This is not the case for the following ARE:

$$[A_\sigma^N]^T[\Pi_\sigma^N] + [\Pi_\sigma^N][A_\sigma^N] - [\Pi_\sigma^N][B_\sigma^N][R_\sigma^N]^{-1}[B_\sigma^N]^T[\Pi_\sigma^N] + [Q_\sigma^N] = 0. \tag{3.13}$$

$[A_\sigma^N], [B_\sigma^N], [Q_\sigma^N]$, and $[R_\sigma^N]$ are the matrices developed using the third approximation scheme. These matrices are defined in Section 3.1. In this section, we introduce preconditioing methods to facilitate computation of solutions to (3.13).

3.3.1 Solving the Algebraic Riccati Equation

Recall for positive μ and v, we define

$$\sigma(x) = e^{-vx/\mu}/\mu.$$

When $x \approx 0$, then $\sigma(x) \approx 1/\mu$. When μ is small in comparison to v, $\sigma(x)$ decays rapidly as x increases from 0.

The vector $[\hat{B}_\sigma^N]$ has dimension $(N-1 \times 1)$. When μ is small in comparison to v, $[\hat{B}_\sigma^N]_i$ is approximately $1/\mu$ for small integers i. For large values of i, $[\hat{B}_\sigma^N]_i$ is very close to zero. However, $[Q_\sigma^N]_{i,i}$ and $[Q_\sigma^N]_{i,i\pm 1}$ are approximately $1/\mu$ for small integers i, but $[Q_\sigma^N]_{i,i}$ and $[Q_\sigma^N]_{i,i\pm 1}$ are very large when $i \approx N-1$. This imbalance can cause (3.13) to be ill conditioned and difficult to solve numerically.

Consider the case where $b(x) = e^x$, $\mu \leq 1/45$, and $v = 1$ when defining (2.1). Additionally, let $q = 1$ and $r = 0.1$. When we attempt to solve the ARE using MATLAB, it returns an error stating that $\big([A_\sigma^N], [B_\sigma^N)]\big)$is uncontrollable or that no solution exists.

3.3.2 Scaling by a Constant

Let $A, Q \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $R \in \mathbb{R}$. We want to solve the following ARE for $P \in \mathbb{R}^{N\times N}$:

$$A^T P + PA - PBR^{-1}B^T P + Q = 0. \tag{3.14}$$

According to the procedure in [35], we develop the following approach to compute a solution to (3.13).

Let s be a positive scalar, and introduce the following change of variables:

$$\tilde{P} = P/s \tag{3.15}$$

$$\tilde{A} = sA \tag{3.16}$$

$$\tilde{B} = sB. \tag{3.17}$$

Thus, instead of solving (3.14), solve the transformed ARE

$$\tilde{A}^T \tilde{P} + \tilde{P}\tilde{A} - \tilde{P}\tilde{B}R^{-1}\tilde{B}^T \tilde{P} + Q = 0. \tag{3.18}$$

We consider another scaling approach, as well. If $(\hat{z}, \hat{u})$ minimizes the cost function

$$J = \int_0^\infty \left(\int_0^1 qz^2(t,x)dx \right) + ru^2(t)dt, \tag{3.19}$$

then $(\hat{z}, \hat{u})$ also minimizes the cost function

$$J_0 = \int_0^\infty \left(\int_0^1 \frac{q}{s} z^2(t,x)dx \right) + \frac{r}{s} u^2(t)dt.$$

Thus, instead of solving the ARE (3.14), we can solve the equivalent ARE

$$A^T P + PA - s(PBR^{-1}B^T P) + \frac{1}{s}(Q) = 0.$$

These two modifcations lead to significant improvements when solving (3.13). Numerical implementation of these two modifications is relatively simple. Furthermore, when we use a fine mesh, i.e. $N \geq 128$, we can compute a solution to (3.13) for $\mu \geq 1/80$ when either scaling modification is applied.

However, there are some major drawbacks. We could not compute a solution for $\mu < 1/80$ for any combination of scaling factors and meshes. Additionally, a very fine mesh is necessary to solve the ARE (3.18) for small values of μ. Furthermore, it is not clear how to choose the optimal, or even a good, scaling factor.

3.3.3 Scaling by a Matrix

In (3.13), $[Q_\sigma^N]$ and $[B_\sigma^N]$ have specific structural characteristics that we exploit in order to develop a preconditioning technique to solve (3.13) numerically. We outline the structure below.

The matrix $[Q_\sigma^N] \in R^{N-1 \times N-1}$ is tridiagonal. Also

1) the matrix entry $[Q_\sigma^N]_{1,1} \approx 1/\mu$,

2) the matrix entry $[Q_\sigma^N]_{N,N}$ is very large,

3) the matrix entry $[Q_\sigma^N]_{i,i} \approx [Q_\sigma^N]_{i,i\pm 1}$, and

4) the non-zero sequence $\{|[Q_\sigma^N]_{i,i}|\}_{i=1}^{N-1}$ is monotonically increasing.

Moreover, $[B_\sigma^N]$ is a vector in $\in R^{N-1}$such that

1) the vector entry $[B_\sigma^N]_1 \approx 1/\mu$,

2) the vector entry $[B_\sigma^N]_N$ is very small, and

3) the non-zero sequence $\{|[B_N]_i|\}_{i=1}^{N-1}$ is monotonically decreasing.

We define the diagonal matrix $[D_N]$ such that

$$[D_N]_{i,i} \equiv |[R_\sigma^N]^{-1/2}[B_\sigma^N]_i|^{1/2}/|[Q_\sigma^N]_{i,i}|^{1/4}.$$

It follows that

$$[D_N]_{i,i}^2[Q_\sigma^N]_{i,i} = \left(|[R_\sigma^N]^{-1}[B_\sigma^N]_i^2|\right) / \left([D_N]_{i,i}^2\right).$$

We use the matrix $[D_N]$ to precondition the ARE (3.13). When we pre- and post-multiply (3.13) with $[D_N]$, the resulting equation is

$$\begin{aligned}&[D_N][A_\sigma^N]^T[\Pi_\sigma^N][D_N] + [D_N][\Pi_\sigma^N][A_\sigma^N][D_N] - \\ &\qquad [D_N][\Pi_\sigma^N][B_\sigma^N][R_\sigma^N]^{-1}[B_\sigma^N]^T[\Pi_\sigma^N][D_N] + [D_N][Q_\sigma^N][D_N] = 0. \qquad (3.20)\end{aligned}$$

Consider the following change of variables:

$$[\tilde{A}_N] = [D_N]^{-1}[A_\sigma^N][D_N] \qquad (3.21)$$

$$[\tilde{B}_N] = [D_N]^{-1}[B_\sigma^N] \qquad (3.22)$$

$$[\tilde{Q}_N] = [D_N][Q_\sigma^N][D_N] \qquad (3.23)$$

$$[\tilde{\Pi}_N] = [D_N][\Pi_\sigma^N][D_N]. \qquad (3.24)$$

Noting that $[D_N]^{-1}[D_N]$ is the identity matrix, $[D_N] = [D_N]^T$, and $[D_N]^{-1} = [D_N]^{-T}$, we transform (3.20) into

$$[\tilde{A}_N]^T[\tilde{\Pi}_N] + [\tilde{\Pi}_N][\tilde{A}_N] - [\tilde{\Pi}_N][\tilde{B}_N][R_N]^{-1}[\tilde{B}_N]^T[\tilde{\Pi}_N] + [\tilde{Q}_N] = 0. \qquad (3.25)$$

Instead of solving (3.13) numerically, we solve (3.25). We compute $[D_N]^{-1}[\tilde{\Pi}_N][D_N]^{-1}$ to recover $[\Pi_\sigma^N]$. The diagonal entries of $[\tilde{B}_N][R_N]^{-1}[\tilde{B}_N]^T$ are equal to the diagonal elements of $[\tilde{Q}_N]$. As a result, the ARE is more balanced. Hence, MATLAB can solve (3.25) for a wider range of values of μ. For $N = 256$, (3.25) can be solved for $\mu = 1/650$, and we can solve the ARE when $\mu = 1/850$ if $N = 512$. When (3.25) is evaluated with the numerical solution $[\tilde{\Pi}_N]$, the error has a norm of approximately 10^{-11}.

3.3.4 Residual Errors

Let $[\tilde{P}_N]$ denote the numerical solution that MATLAB computes for (3.13) . We define the residual to be the matrix 2-norm of

$$[\tilde{A}_N]^T[\tilde{P}_N] + [\tilde{P}_N][\tilde{A}_N] - [\tilde{P}_N][\tilde{B}_N][R_N]^{-1}[\tilde{B}_N]^T[\tilde{P}_N] + [\tilde{Q}_N]. \qquad (3.26)$$

In this section, we compare the residual errors when using the different scaling methods to compute a solution to the ARE (3.13).

Table 3.2 displays the residual error from solving the ARE when $\mu = 1/50$. For this table, we vary the value of s when we use the constant scaling technique. We use $s = 10^{11}, 10^{13}, 10^{17}$. Among these three scaling factors, the residual error is lowest when $s = 10^{17}$. However, for such a large scaling factor, we are unable to solve the ARE when $N = 128$. Thus, it is imperative to find a scaling factor that is large enough to keep the error low, but not so large that we cannot numerically solve the ARE. For this problem, $|[Q_\sigma^{64}]_{63,63}|/|[B_\sigma^{64}]|_{63,1} \approx 10^{11}$. Choosing a scaling factor $s = |[Q_\sigma^{64}]_{63,63}|/|[B_\sigma^{64}]|_{63,1}$ worked well for other simulations. Thus, for all further computations when we scale by a constant, we scale the ARE by setting $s = |[Q_\sigma^{64}]_{63,63}|/|[B_\sigma^{64}]_{63,1}$.

Table 3.2 also shows the residual error when we use the matrix scaling technique. With the exception of $N = 8$, the matrix scaling technique has superior performance when compared to the constant scaling attempts.

Table 3.2: Residual Error From Solving the ARE (3.13): $\mu = 1/50$

N	$s = 10^{11}$	$s = 10^{13}$	$s = 10^{17}$	Matrix Scaling
8	1.76×10^{-4}	9.04×10^{-5}	2.95×10^{-5}	1.79×10^{-4}
16	1.16×10^{-5}	3.98×10^{-7}	6.23×10^{-8}	9.42×10^{-13}
32	2.3×10^{-3}	6.95×10^{-6}	1.17×10^{-7}	7.41×10^{-13}
64	8.5×10^{-3}	1.1×10^{-3}	$--$	3.55×10^{-12}
128	1.70×10^{-1}	3.27×10^{-2}	$--$	1.83×10^{-11}

Table 3.3 shows the residual errors for $\mu = 1/75$. Scaling the ARE by a constant results in errors on the order of 10^{10} or larger. Additionally, when $N = 128$, we cannot numerically solve the ARE using this scaling factor. However, despite an error of 7.26×10^7 when $N = 8$, scaling by a matrix leads to an error no larger than 1.22×10^{-11} when $N = 32, 64$, or 128.

Table 3.3: Residual Error From Solving the ARE (3.13): $\mu = 1/75$

N	$s = 10^{12}$	Matrix Scaling
8	3.67×10^{10}	7.26×10^{7}
16	6.80×10^{10}	5.05×10^{0}
32	8.74×10^{13}	9.11×10^{-13}
64	5.78×10^{15}	2.23×10^{-12}
128	$--$	1.22×10^{-11}

Table 3.4 displays the error when we use the matrix scaling technique and $\mu = 1/100$ and $\mu = 1/200$. For both values of μ, the constant scaling technique does not allow us to compute a solution for any value of N. Although we try various values of s, we cannot compute a solution. Additionally, when $\mu = 1/100$ and $N = 8$ and 16, the matrix scaling technique produces large errors. Yet, when the mesh is refined, the errors decrease drastically. When $\mu = 1/200$, we cannot solve the ARE until $N = 64$. However, the error is very small when $N = 128$.

Table 3.4: Residual Error From Solving the ARE (3.13): $\mu = 1/100, \quad 1/200$

N	$\mu = 1/100$	$\mu = 1/200$
8	5.94×10^{12}	$--$
16	6.81×10^{10}	$--$
32	1.49×10^{-9}	$--$
64	1.72×10^{-12}	58.4×10^{0}
128	1.11×10^{-11}	7.14×10^{-12}

We conclude this section by mentioning the following. Not only is the matrix scaling method an improvement for computing solutions to (3.13), but this method can also be used to compute solutions to (3.12). The matrix scaling method performs at least as well solving the ARE (3.12) as solving (3.12) without preconditioning. That is, the matrix scaling method has residual errors on the order of the residual errors from solving the ARE without preconditioning. In fact, in several cases, the matrix scaling method outperforms the unconditioned method.

Chapter 4

Computing the Functional Gains

Recall that the functional gains k_σ^N defined in Section 2.6 are associated with the linear quadratic regulator problem $(\mathbf{\Sigma_\sigma^N})$. In this Chapter, we compute k_σ^N when $b(x) = e^x, q = 1, v = 1$ and $r = 0.1$. We vary the values μ and N. We must compute a solution to the matrix ARE (3.13) in order to compute k_σ^N. Whenever we need to solve (3.13), we use the matrix scaling technique described in Section 3.3.3.

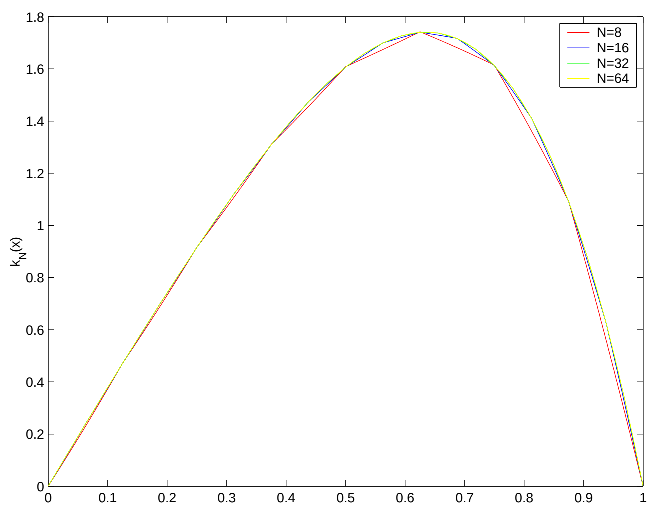

Figure 4.1: Functional Gains $k_\sigma^N(x)$: $\mu = 1, \quad N = 8...64$.

Figure 4.1 displays the functional gains $k_\sigma^N(x)$ when $\mu = 1$. As N increases, the functional gains converge rapidly.

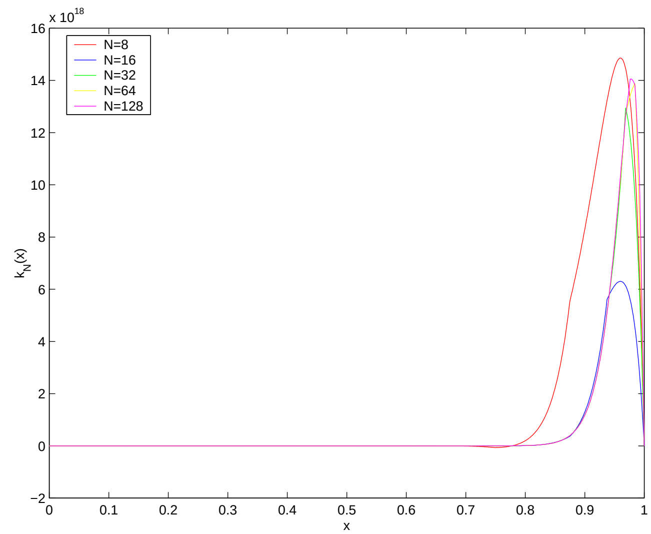

Figure 4.2: Functional Gains $k_\sigma^N(x)$: $\mu = 1/50$, $N = 8...128$.

In Figures 4.2 and 4.3, we plot the functional gains when $\mu = 1/50$ and $\mu = 1/75$. When $\mu = 1/50$, the gains converge but not as rapidly as in Figure 4.1. Additionally, the gains take on values ranging approximately from 0 to 10^{19}. In Figure 4.3, we see that $k_\sigma^N(x)$ can reach values on the order of 10^{29} when $\mu = 1/75$. The functional gains converge very slowly. Viewing this figure on such a large scale does not reveal the full dynamics of the functional gains. Consider Figure 4.4, wherein, we plot the functional gains for $\mu = 1/100$ and $r = 10^6$. When r increases from 0.1, the functional gains are are smaller, and we can plot them on a smaller scale. On this scale, $k_\sigma^{16}(x)$ appears to be approximately 0. However, this is clearly not the case if we examine $k_\sigma^{16}(x)$ by itself on in Figure 4.5. Figure 4.5 displays $k_\sigma^{16}(x)$ when x is near zero. In this plot, we see that $k_\sigma^{16}(x)$ is increasing and actually takes on values as large as 10^9. Figure 4.6 plots this same function over a larger interval. In this figure it is obvious that $k_\sigma^{16}(x)$ takes on negative values. Figure 4.7 plots $k_\sigma^{16}(x)$ over an even larger interval, and once again, $k_{16}(x)$ takes on positive values. When $\mu = 1/100$ and $N = 16$, the gain actually oscillates from positive to negative values. However, when we increase N to 32 or larger, we can verify that the gains do not oscillate for $\mu = 1/100$. We have also verified that the gains in Figures 4.1 - 4.3 do not have these oscillations.

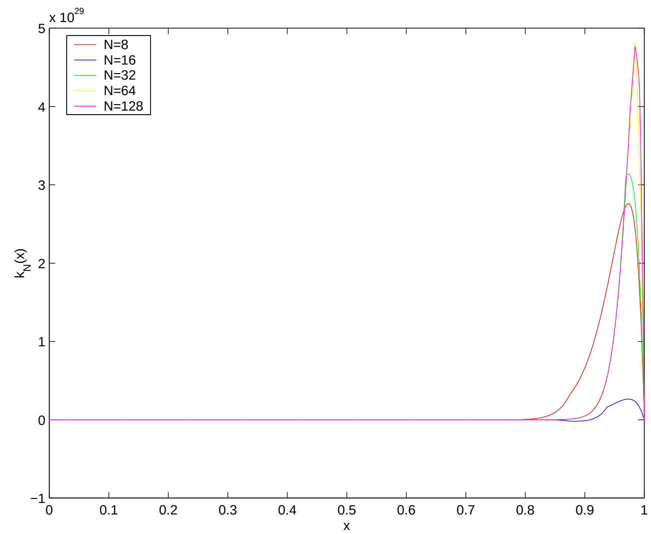

Figure 4.3: Functional Gains $k_\sigma^N(x)$: $\mu = 1/75, \quad N = 8...128$.

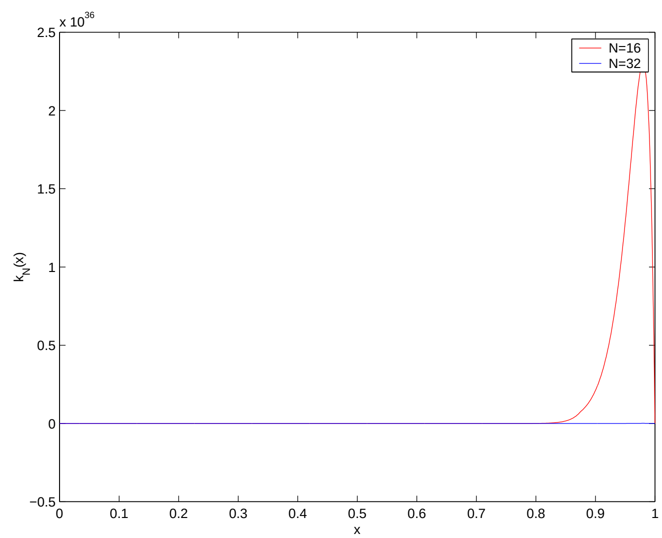

Figure 4.4: Functional Gains $k_\sigma^N(x)$: $\mu = 1/100, \quad r = 10^6, \quad N = 16, 32$.

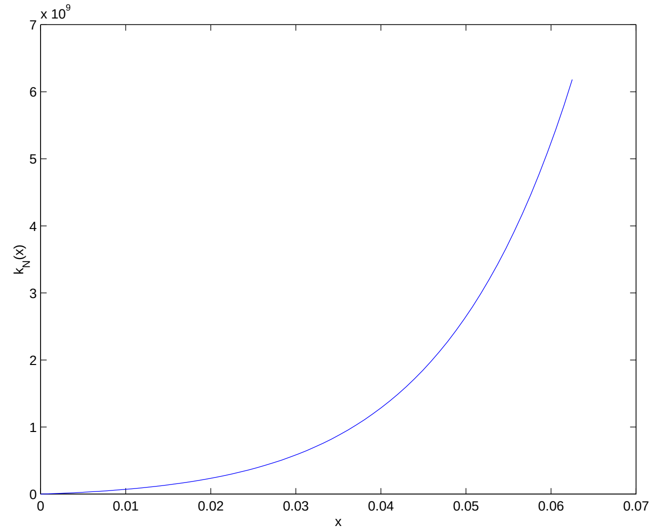

Figure 4.5: The Functional Gain $k_\sigma^N(x)$, $0 \leq x \leq .07$: $\mu = 1/100$, $r = 10^6$, $N = 16$.

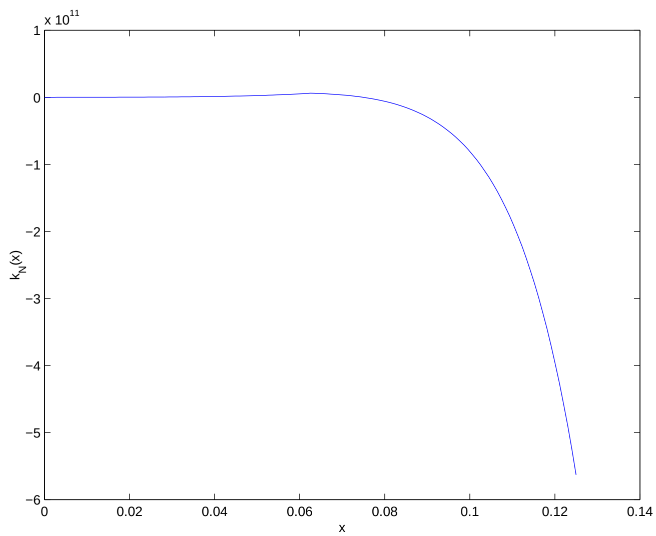

Figure 4.6: The Functional Gain $k_\sigma^N(x)$, $0 \leq x \leq .14$: $\mu = 1/100$, $r = 10^6$, $N = 16$.

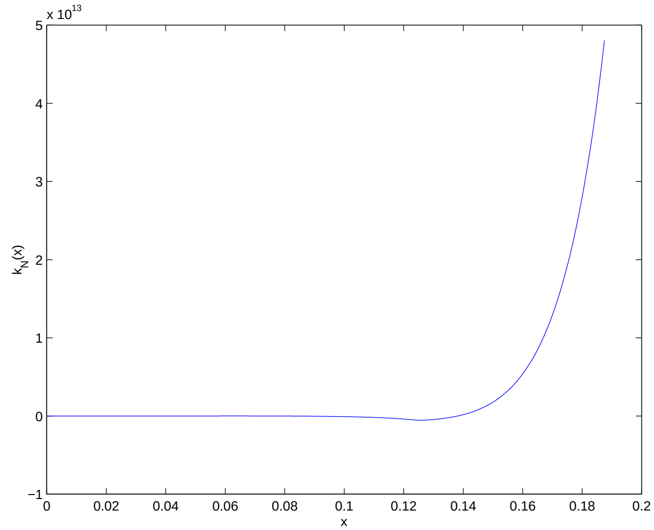

Figure 4.7: The Functional Gain $k_\sigma^N(x)$, $0 \leq x \leq .21$: $\mu = 1/100$, $r = 10^6$, $N = 16$.

4.1 Comparing Interpolants of the Functional Gains k_σ^N

In Section 2.6, we discussed the convergence of the functional gains for $(\mathbf{\Sigma}_\sigma)$ and $(\mathbf{\Sigma_\sigma^N})$. However, due to the large values that these gains achieve, it is difficult to see convergence in Figures 4.2 and 4.3. We use an interpolation technique to provide numerical support that the gains k_σ^N are converging. We outline this technique below.

In Section 2.6 we show that k_σ^N in V_σ^N has the basis representation

$$k_\sigma^N(x) = \sigma^{-1/2}(x) \sum_{i=1}^{N-1} b_i^N h_i^N(x),$$

where $\sigma = e^{-v(\cdot)/\mu}/\mu$ and h_i^N are the hat functions introduced in Section 2.3.1 . We describe how to find the coefficients b_i^N.

Since the operators feedback K_σ and K_σ^N defined in Section 2.6 are bounded linear functionals on H_σ and V_σ^N, respectively, we can use the Riesz Representation Theorem to find unique functions k_σ in H_σ and k_σ^N in V_σ^N such that

$$K_\sigma z = < k_\sigma, z >_\sigma, \quad \text{for all} \quad z \in H_\sigma$$

and

$$K_\sigma^N z_N = < k_\sigma^N, z_N >_\sigma, \quad \text{for all} \quad z_N \in V_\sigma^N.$$

Let $[K_\sigma^N] = [\kappa_1 \cdots \kappa_{N-1}]$ be the matrix representation of K_σ^N resulting from the numerical solution of (3.13). Let $F_N = \sigma^{-1/2} \sum_{i=1}^{N-1} f_i h_i^N$ be a function in V_σ^N. Then

$$K_\sigma^N F_N = [\kappa_1 \cdots \kappa_{N-1}] \begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \end{bmatrix}. \tag{4.1}$$

Moreover, since k_σ^N is in V_σ^N, k_σ^N has the basis representation $k_\sigma^N = \sigma^{-1/2} \sum_{i=1}^{N-1} c_i h_i^N$, and

$$K_\sigma^N F_N = < k_N, f_N >_\sigma = [c_1 \cdots c_{N-1}][M_3^N] \begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \end{bmatrix}, \tag{4.2}$$

where $[M_3^N]$ denotes the mass matrix defined in Section 3.1. It follows from (4.1) and (4.2) that for all $\begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \end{bmatrix} \in \mathbb{R}^{N-1}$,

$$[\kappa_1 \cdots \kappa_N] \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} = [c_1 \cdots c_{N-1}][M_3^N] \begin{bmatrix} f_1 \\ \vdots \\ f_{N-1} \end{bmatrix}.$$

Therefore, the coefficients for k_σ^N are

$$[c_1 \cdots c_{N-1}] = [\kappa_1 \cdots \kappa_{N-1}][M_3^N]^{-1}.$$

To find the $\sigma-$norm of k_σ^N, we use the relationship

$$\left\|k_\sigma^N\right\|_\sigma^2 = \begin{bmatrix} c_1^N & \cdots & c_{N-1}^N \end{bmatrix} [M_N^3] \begin{bmatrix} c_1^N \\ \vdots \\ c_{N-1}^N \end{bmatrix}.$$

However, when $M \neq N$, we can only approximate $\left\|k_\sigma^N - k_\sigma^M\right\|_\sigma$. The following interpolation technique approximates $\left\|k_\sigma^N - k_\sigma^M\right\|_\sigma$.

Let

$$k_\sigma^N = \sigma^{-1/2} \sum_{i=1}^{N-1} c_i^N h_i^N \in V_\sigma^N$$

and

$$k_\sigma^{2N} = \sigma^{-1/2} \sum_{i=1}^{2N-1} c_i^{2N} h_i^{2N} \in V_\sigma^{2N}.$$

We denote the the interpolant of k_σ^N into V_σ^{2N} by k_N^{2N}. The interpolant k_N^{2N} is defined by

$$k_N^{2N} = (\sigma(\cdot))^{-1/2} \sum_{i=1}^{2N} c_{i,N}^{2N} h_i^{2N}(\cdot) \in V_\sigma^{2N},$$

where

$$c_{i,N}^{2N} = \begin{cases} c_{i,N}^{2N} = c_i^N/2, & i = 1 \quad \text{or} \quad 2N-1 \\ c_{2i,N}^{2N} = c_i^N, & i = 1, 2, ..., N-1 \\ c_{2i+1,N}^{2N} = (c_i^N + c_{i+1}^N)/2, & i = 1, 2, ..., N-2. \end{cases} \tag{4.3}$$

To compute the interpolant of k_σ^N into V_σ^{4N}, we first interpolate into V_σ^{2N} and then interpolate the result into V_σ^{4N}. By this iterative process, we compute the interpolants of $k_\sigma^{16}, k_\sigma^{32}, k_\sigma^{64}, k_\sigma^{128}$ into V_σ^{256}.

Once again, let $b(x) = e^x, q = 1$, and $r = 0.1$ in (2.1). Table 4.1 lists the norms of the differences between k_{256} and the interpolants of $k_\sigma^{16}, k_\sigma^{32}, k_\sigma^{64}$, and k_σ^{128} when $\mu = 1$. The differences and relative differences are no larger than 10^{-5} and decrease as N increases. This numerical data supports the conclusion that the gains are converging.

Table 4.1: Norms of the Functional Gains: $\mu = 1$, $\|k_\sigma^{256}\|_\sigma^2 = 8.675 \times 10^{-1}$

N	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2$	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2 \,/\, \|k_\sigma^{256}\|_\sigma^2$
16	1.72×10^{-5}	1.98×10^{-5}
32	1.05×10^{-6}	1.22×10^{-6}
64	6.07×10^{-8}	7.00×10^{-8}
128	2.66×10^{-9}	3.07×10^{-9}

Table 4.2: Norms of the Functional Gains: $\mu = 1/50$, $\|k_\sigma^{256}\|_\sigma^2 = 5.28 \times 10^{17}$

N	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2$	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2 \,/\, \|k_\sigma^{256}\|_\sigma^2$
16	6.87×10^{16}	1.30×10^{-1}
32	6.21×10^{15}	1.18×10^{-2}
64	3.83×10^{14}	7.26×10^{-4}
128	1.70×10^{13}	3.22×10^{-5}

In Table 4.2, we see the norms of the differences between k_σ^{256} and the interpolants of $k_\sigma^{16}, k_\sigma^{32}, k_\sigma^{64}$, and k_σ^{128} when $\mu = 1/50$. The differences are all very large. Even when $N = 128$, the norm of the difference is on the order of 10^{13}. However, we note that $\|k_\sigma^{256}\|_\sigma^2 = 5.28 \times 10^{17}$. Thus, we expect the differences between k_σ^{256} and the interpolants to be significantly larger when $\mu = 1/50$ than when $\mu = 1$. Examination of the relative differences reveals that they are all on the order of 10^{-1} or smaller. In fact, when $N = 128$, the relative error is only 3.22×10^{-5}. Again, the differences decrease as N increases.

Table 4.3: Norms of the Functional Gains: $\mu = 1/75$, $\|k_\sigma^{256}\|_\sigma^2 = 8.87 \times 10^{27}$

N	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2$	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2 \,/\, \|k_\sigma^{256}\|_\sigma^2$
16	7.51×10^{27}	8.47×10^{-1}
32	4.12×10^{26}	4.64×10^{-2}
64	2.85×10^{25}	3.20×10^{-3}
128	1.28×10^{24}	1.45×10^{-4}

Table 4.3 displays this same pattern when $\mu = 1/75$. Even though the differences are as large as 7.51×10^{27}, the relative differences are not larger than 8.47×10^{-1}.

Table 4.4: Norms of the Functional Gains: $\mu = 1/100$, $\|k_\sigma^{256}\|_\sigma^2 = 2.29 \times 10^{38}$

N	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2$	$\|k_\sigma^{256} - k_N^{256}\|_\sigma^2 / \|k_\sigma^{256}\|_\sigma^2$
16	2.98×10^{40}	1.35×10^{2}
32	3.14×10^{37}	1.42×10^{-1}
64	2.04×10^{36}	9.30×10^{-3}
128	9.48×10^{34}	4.29×10^{-4}

In Table 4.4, we list the norms of the differences between k_σ^{256} and the interpolants of $k_\sigma^{16}, k_\sigma^{32}, k_\sigma^{64}$, and k_σ^{128} when $\mu = 1/100$. Notice that when $N = 16$, the norm of the difference is 2.98×10^{40}, and the relative difference is 1.35×10^2. This large difference can be explained by the oscillations described in Section 4 and pictured in Figures 4.5 - 4.7.

4.2 Comparing the Gains for (Σ_1^N) and (Σ_σ^N)

In this section, we compare the functional gains k_σ^N with the functional gains k_1^N. Recall that k_1^N is defined in Section 2.6.

Since the bases and the inner products used to find k_σ^N and k_1^N are different, visual examination of the gains is not an effective means of comparison. However, once we transform k_σ^N, we can compare k_σ^N and k_1^N.

The approximate control for $(^{\mathbf{N}})$ is of the form

$$\hat{u}_\sigma^N(t) = -K_\sigma^N \hat{z}_\sigma^N(\cdot)(t).$$

In Section 4.1, we show that

$$\hat{u}_\sigma^N(t) = -\int_0^1 \sigma(x) k_\sigma^N(x) \hat{z}_\sigma^N(x)(t)dx.$$

Since k_σ^N and $\hat{z}_\sigma^N$ are in V_σ^N, we can express them as linear combinations of the basis elements of V_σ^N. Thus, there exist constants $\{b_i\}_{i=1}^{N-1}$ such that

$$k_\sigma^N = \sum_{i=1}^{N-1} b_i \phi_i^N.$$

Thus,

$$\hat{u}_\sigma^N(t) = -\int_0^1 (\sigma(x))^{1/2} \left(\sum_{i=1}^{N-1} c_i h_i^N(x) \right) \hat{z}_\sigma^N(x)(t) dx.$$

The optimal control for $(\mathbf{\Sigma_1^N})$ is of the form

$$\hat{u}_1^N(t) = -K_1^N \hat{z}_1^N(\cdot)(t)$$

where K_1^N is the feedback operator on V^N. In a similar process to the one outlined above, we can show there exist constants $\tilde{c}_i, i = 1, ..., N-1$ with

$$\hat{u}_1^N(t) = -\int_0^1 \left(\sum_{i}^{N-1} \tilde{c}_i h_i^N \right) \hat{z}_1^N(x)(t) dx.$$

Since $(\mathbf{\Sigma_\sigma^N})$ and $(\mathbf{\Sigma_1^N})$ are both developed to approximate the optimal control (2.1)-(2.6), we expect that for large values of N, $\hat{u}_1^N(t) \approx \hat{u}_\sigma^N(t)$. Additionally, we expect that for $t \geq 0, x \in [0,1]$, $\hat{z}_\sigma^N(t,x) \approx \hat{z}_1^N(t,x)$, and

$$\int_0^1 \sum_{i}^{N-1} \tilde{c}_i h_i^N \tilde{z}_1^N(x) dx \approx \int_0^1 (\sigma(x))^{1/2} \sum_{i}^{N-1} c_i h_i^N \hat{z}_\sigma^N(x) dx.$$

Furthermore, at meshpoints $x_i = i/N, i = 0, 1, ..., N$, $h_i^N(x_i) = 1$ and $h_j^N(x_i) = 0, \quad j \neq i$. Thus, we have

$$(\sigma(x_i))^{1/2} c_i \approx \tilde{c}_i.$$

Define

$$\tilde{k}_\sigma^N(x) = \sum_{i}^{N-1} (\sigma(x_i))^{1/2} c_i h_i(x).$$

For the figures in this section, we plot $\tilde{k}_\sigma^N(x)$ and $k_1^N(x)$ as a means of comparing the gains $k_\sigma^N(x)$ and $k_1^N(x)$.

Figure 4.8 shows $\tilde{k}_\sigma^N(x)$ and $k_1^N(x)$ when $\mu = 1$ and $N = 128$. The two functions are identical.

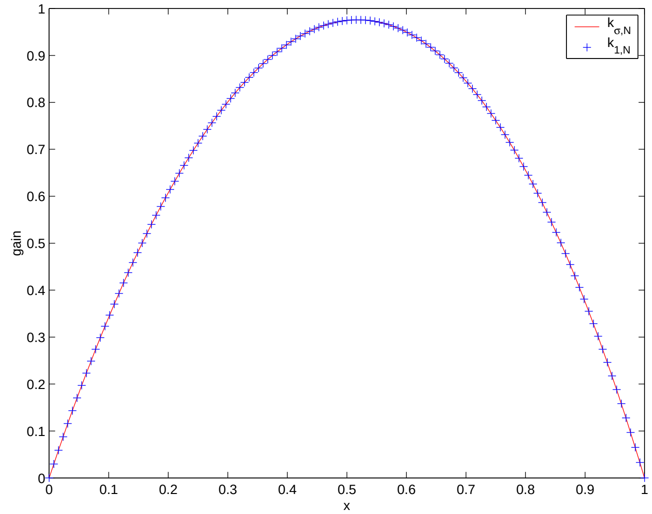

Figure 4.8: The Functions $\tilde{k}_\sigma^N(x)$ and $k_1^N(x)$: $\mu = 1, \quad N = 256$.

Figures 4.9 and 4.10 show $\tilde{k}_\sigma^N(x)$ and $k_1^N(x)$ when μ is decreased to 1/50 and 1/75. When $\mu = 1/50$, the two functions have precisely the same values. When $\mu = 1/75$, the shapes of the gains are similar, but $k_1^N(x)$ attains larger values than $\tilde{k}_\sigma^N(x)$ when x is between 0.3 and 0.7.

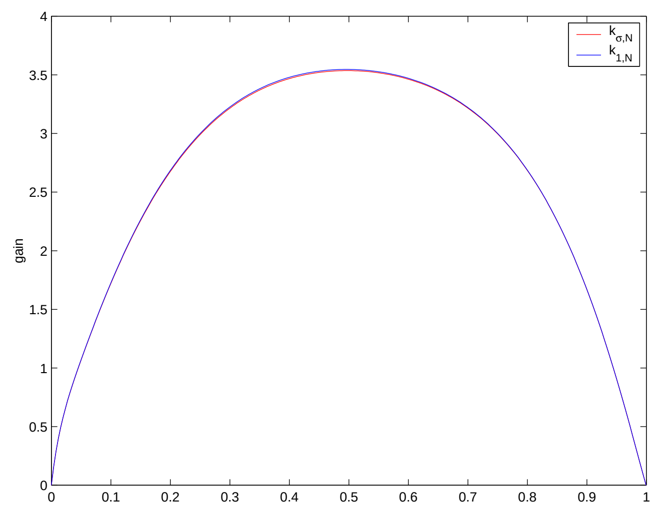

Figure 4.9: The Functions $\tilde{k}_\sigma^N(x)$ and $k_1^N(x)$: $\mu = 1/50, \quad N = 256$.

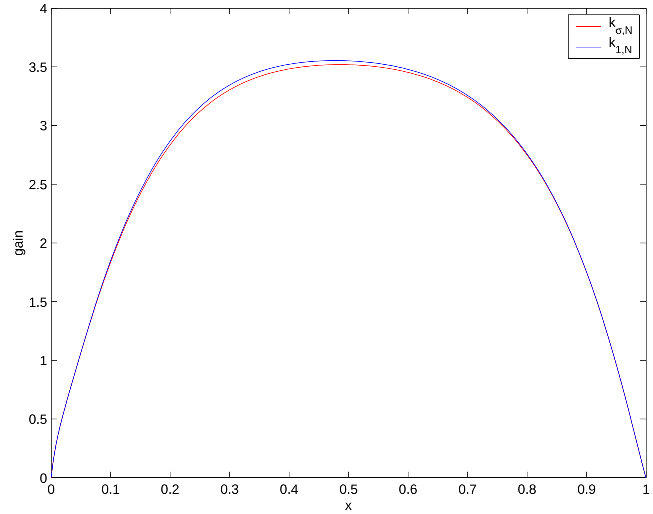

Figure 4.10: The Functions $\tilde{k}_{\sigma}^{N}(x)$ and $k_{1}^{N}(x)$: $\mu = 1/75, \quad N = 256$.

Finally, Figure 4.11 shows $\tilde{k}_{\sigma}^{N}(x)$ and $k_{1}^{N}(x)$ when $\mu = 1/850$ and $N = 512$. For this small value of μ, there are significant differences between $\tilde{k}_{\sigma}^{N}(x)$ and $k_{1}^{N}(x)$. Initially, $k_{1}^{N}(x)$ decreases and then increases. The function $\tilde{k}_{\sigma}^{N}(x)$ increases immediately. However, $k_{1}^{N}(x)$ still reaches a maximum of about 3.5. The other function only reaches a maximum of approximately 0.5

Several explanations can make clear the differences between the two gains. The two schemes may have different rates of convergence. Additionally, when we transform k_{σ}^{N}, we may introduce large numerical error into our plots since we are working with very large numbers.

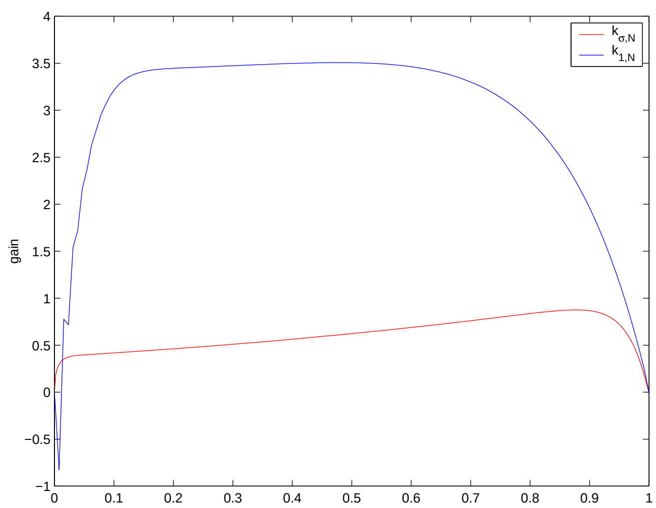

Figure 4.11: The Functions $\tilde{k}_{\sigma}^{N}(x)$ and $k_{1}^{N}(x)$: $\mu = 1/850, \quad N = 512$.

Chapter 5

Conclusion

The main focus of this dissertation is the development and understanding of approximation techniques for infinite dimensional linear quadratic regulator problems. Specifically, we consider these problems when their constraints are equations with nonnormal linearizations. The general framework and theory for approximating infinite dimensional linear quadratic regulator problems indicates that dual convergence is necessary to ensure convergence of the Riccati operators and feedback gains.

We describe three different approximation schemes to approximate a linear quadratic regulator problem governed by the one dimensional convection-diffusion equation. All three schemes are convergent, but each scheme has its drawbacks.

The first is a traditional Galerkin finite element method. When the convection-diffusion equation is diffusion dominated, this method is good for computing eigenvalues of the open loop problem. However, when the equation is convection dominated, this method produces complex eigenvalues. Still, computing a solution to algebraic Riccati equations developed using this scheme is possible for both convection and diffusion dominated equations.

The second method renorms the problem. Under the new norm and inner product, the convection-diffusion operator is self-adjoint. Since the second approximation scheme is convergent, dual convergence follows automatically. This method is difficult to implement accurately in a numerical setting, so we do not use this scheme to compute feedback gains.

The third scheme uses the renorming approach, but it has a different basis to overcome some of the numerical difficulties of the second scheme. This is a good method for computing eigenvalues, but computing a solution to the resulting algebraic Riccati equations requires preconditioning when the convection-diffusion equation is convection dominated. We develop a preconditioning scheme that allows for solution of the algebraic Riccati equations. Finally, we present and compare functional computed with the first and the third approximation schemes. For a convection dominated equation, the functional gains have some significant differences.

For the convection-diffusion equation, we are able to take advantage of renorming to develop approximation schemes. However, it is not clear if it is possible to do this for the Navier-Stokes equations. Without going into the details, the Navier-Stokes equations can be written as

$$\dot{z}(t) = Az(t) + F(z(t)),$$

where A is a linear operator and F is a nonlinear operator. If we linearize the equation about plane Poiseuille flow, we can discard F. It is well known that the operator A is highly nonnormal. Our initial efforts to develop an inner product on which A is self-adjoint have not been successful, so approximation techniques for examining the Navier-Stokes equations in an optimal control setting remains an open issue. We need to develop robust numerical techniques that consider dual convergence and POES. We plan to investigate approximation of linear quadratic regulator problems involving the Navier-Stokes equations in the future.

Bibliography

[1] J. S. Baggett, T. A. Driscoll, and L. N. Trefethen. A mostly linear model of transition to turbulence. *Phys. Fluids*, 7:833–838, 1995.

[2] B. Bamieh and M. Dahleh. Disturbance energy amplification in three-dimensional channel flows. *Proceedings of the American Control Conference*, pages 4532–4537, 1998.

[3] B. Bamieh and M. Dahleh. Energy amplification in channel flow with stochastic excitation. *Physics of Fluids*, 13:3258–3269, 2001.

[4] H. T. Banks and J. A. Burns. Hereditary control problems: numerical methods based on averaging approximations. *SIAM J. Control and Optimization*, 16:169–208, 1978.

[5] H. T. Banks and F. Kappel. Spline approximations for functional differential equations. *J. Differential Equations*, 34:469–522, 1979.

[6] H. T. Banks and K. Kunisch. An approximation theory for nonlinear partial differential equations with applications to identification and control. *SIAM J. Control and Optimization*, 16:815–849, 1982.

[7] H. T. Banks and K. Kunisch. The linear regulator problem for parabolic systems. *SIAM J. Control and Optimization*, 22:684–698, 1984.

[8] H. T. Banks, G. I. Rosen, and K. Ito. A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations. *SIAM J. Sci. Statist. Comput*, 5:830–855, 1984.

[9] H. T. Banks, R. C. Smith, and Y. Wang. *Smart Material Structures: Modeling, Estimation, and Control.* Wiley, New York, NY, 1996.

[10] A. Bensoussan, G. Da Prato, M. Delfour, and S. Mitter. *Representation and Control of Infinite Dimensional Systems.* Birkhauser, Boston, MA, 1993.

[11] J. T. Borggaard, J. A. Burns, E. D. Vugrin, and L. Zietsman. On strong convergence of feedback operators for non-normal distributed parameter systems. *submitted to The Proceedings of the Conference on Decision and Control*, 2004.

[12] J. A. Burns. Non-linear distributed parameter systems with non-normal linearizations: Applications and approximations. *Research Directions in Distributed Parameter Systems*, R. C. Smith and M. A. Demetriou, Eds., SIAM Publications, Philadelphia:17–53, 2003.

[13] J. A. Burns, B. B. King, D. Krueger, and L. Zietsman. Computation of feedback operators for distributed parameter systems with non-normal linearizations. *Proceedings of the American Control Conference*, pages 2083–2088, 2003.

[14] K. M. Butler and B. F. Farrell. Three dimensional disturbances in plane Poiseuille flow. *Physics of Fluids A*, 4:1637, 1992.

[15] P. G. Drazin and W. H. Reid. *Hydrodynamic Stability*. Cambridge University Press, New York, NY, 1981.

[16] R. H. Fabiano. Renorming for stability and approximation of linear systems: examples. *Mathematical and Computer Modelling*, 33:159–172, 2001.

[17] B. F. Farrell and P. J. Ioannou. Stochastic forcing of the linearized Navier-Stokes equations. *Physics of Fluids A*, 5:2600–2609, 1993.

[18] B. F. Farrell and P. J. Ioannou. Generalized stability theory. part i: autonomous operators. *Journal of the Atmospheric Sciences*, 53:2025–2040, 1996.

[19] B. F. Farrell and P. J. Ioannou. Generalized stability theory. part ii: nonautonomous operators. *Journal of the Atmospheric Sciences*, 53:2041–2053, 1996.

[20] L. P. Franca, S. L. Frey, and T. J. R. Hughes. Stabilized finite element methods: I. application to the advective-diffusive model. *Computer Methods in Applied Mechanics and Engineering*, 95:252–276, 1992.

[21] Isaac Fried. *Numerical Solution of Differential Equations*. Academic Press, New York, NY, 1979.

[22] H. Fujita and A. Mizutani. On the finite element method for parabolic equations, i: Approximation of holomorphic semigroups. *J. Math. Soc. Japan*, 28:749–771, 1976.

[23] J. S. Gibson. The Riccati integral equations for optimal control problems on Hilbert spaces. *SIAM J. Control and Optimization*, 17:537–565, 1979.

[24] J. S. Gibson. An analysis of optimal model regulation: convergence and stability. *SIAM J. Control and Optimization*, 19:686–707, 1981.

[25] J. S. Gibson. Linear quadratic optimal control of hereditary differential systems: inifinite dimensional Riccati equations and numerical approximations. *SIAM J. Control and Optimization*, 21:95–139, 1983.

[26] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations.* Springer-Verlag, New York, NY, 1980.

[27] L. H. Gustavsson. Energy growth of three-dimmensional disturbances in plane Poiseuille flow. *J. Fluid Mech.*, 224:241–260, 1991.

[28] D. S. Henninngson, A. Lundbladh, and A. V. Johansson. A mechanism for bypass transition from localized disturbances in wall bounded shear flows. *J. Fluid Mech.*, 250:169–207, 1993.

[29] I. H. Herron. Observations on the role of the vorticity in the stability of wall-bounded flows. *Stud. Appl. Math*, 85:169–286, 1991.

[30] K. Ito and F. Kappel. A uniformly differentiable approximation scheme for delay systems using splines. *Appl. Math. Optim.*, 23:217–262, 1991.

[31] M. R. Jovanovich and B. Bamieh. Modeling flow statistics using the linearized Navier-Stokes equations. *Proceedings of the 40th IEEE Conference on Decision and Control*, pages 4944–4949, 2001.

[32] M. R. Jovanovich and B. Bamieh. The spatio-temporal impulse response of the linearized Navier-Stokes equations. *Proceedings of the 2001 American Control Conference*, pages 1948–1953, 2001.

[33] M. R. Jovanovich and B. Bamieh. Input-output analysis of the linearized Navier-Stokes equations in channel flows. *Under consideration in J. Fluid Mech*, 2002.

[34] F. Kappel and D. Salamon. Spline approximations for retarded systems and the Riccati equation. *SIAM J. Control and Optimization*, 25:1082–1117, 1987.

[35] C. Kenney, A. Laub, and M. Wette. A stability-enhancing scaling procedure for Schur-Riccati solvers. *systems and Control Letters*, 12:499–516, 1989.

[36] D. Krueger. *Stabilized Finite Element Methods and Feedback Control.* PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2004.

[37] I. Lasiecka and R. Triggiani. *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, volume 164 of *Lecture Notes in Control and Information Sciences.* Springer-Verlag, New York, NY, 1991.

[38] I. Lasiecka and R. Triggiani. *Control Theory for Partial Differential Equations: Continuous and Approximation Theories I.* Cambridge University Press, New York, NY, 2000.

[39] I. Lasiecka and R. Triggiani. *Control Theory for Partial Differential Equations: Continuous and Approximation Theories II.* Cambridge University Press, New York, NY, 2000.

[40] A. Lundbladh and A. V. Johnsson. Direct simulation of turbulent spots in plane Couette flow. *J. Fluid Mech.*, 229:499–516, 1991.

[41] S. A. Orszag. Accurate solutions of the Orr-Sommerfeld equation. *J. Fluid Mech.*, 38:689–703, 1971.

[42] V. C. Patel and M. R. Head. Some observations on skin friction and velocity profiles in fully developed pipe and channel flows. *J. Fluid Mech.*, 38:181–201, 1969.

[43] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations.* Springer-Verlag, New York, NY, 1983.

[44] S. C. Reddy and D. S. Henningson. Energy growth in viscous channel flow. *J. Fluid Mech.*, 252:209–238, 1993.

[45] A. De Santis, A. Germani, and L. Jetto. Approximation of the algebraic Riccati equation in the hilbert space of hilbert-schmidt operators. *SIAM J. Control and Optimization*, 31:847–874, 1993.

[46] P. J. Schmid and D. S. Henningson. Optimal energy growth in Hagen-Poiseuille flow. *J. Fluid Mech.*, 277:197–225, 1994.

[47] G. Strang and G. Fix. *An Analysis of the Finite Element Method.* Wellesley-Cambridge Press, Massachusetts, 1988.

[48] N. Tillmark and H. Alfredsson. Experiments on transition in plane Couette flow. *J. Fluid Mech.*, 235:89–102, 1992.

[49] L. N. Trefethen. Pseudospectra of linear operators. *SIAM Rev.*, 39:383–406, 1997.

[50] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll. Hydrodynamic stability without eigenvalues. *Science*, 261:578–584, 1993.

Vita

Eric Davor Vugrin was born in White Plains, New York to Margaret and Davor Vugrin on July 20, 1976. He graduated with honors from Lubbock High School in 1994. Eric earned a Bachelor of Science Degree in Mathematics and graduated Summa Cum Laude from Texas Tech University in 1998, and in 2000, he earned a Master of Science Degree in Mathematics from Texas Tech University. This thesis is the final requirement for his Doctoral Degree in Mathematics from Virginia Polytechnic Institute and State University.