Multiplicative Control of Second Order Partial Differential Equations on Reaction and Advection Terms

by

Yixing (Roger) Gu

A research report
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Waterloo, Ontario, Canada, 2025

Author's Declaration

I hereby declare that I am the sole author of this research report. This is a true copy of the report, including any required final revisions, as accepted by my examiners.

I understand that my report may be made electronically available to the public.

Abstract

This report investigates the approximate controllability of parabolic partial differential equations, focusing on the one-dimensional heat equation and the Fokker-Planck equation with multiplicative (bilinear) control. Unlike classical additive control, a multiplicative control acts on the system by multiplying the system's state variable in either the reaction or the advection terms.

We will first prove several results about the density of a certain class of functions and refine certain useful tools about weak derivatives; for instance, the chain rule and quotient rule.

We will follow a paper to study the heat equation under multiplicative control acting through the reaction term in detail, establishing sufficient conditions for approximate controllability. We will present code and plots from Python notebook as an intuition.

We will then extend the analysis to multiplicative control of the Fokker-Planck equation through the advection term, and establish a relationship to the control through the reaction term, presenting algorithms and conjectures about the approximate controllability of the Fokker-Planck equation by means of a multiplicative control.

Acknowledgements

I sincerely thank my supervisor, Professor Roberto Guglielmi, for his invaluable guidance, encouragement, and support throughout the research process. His insightful observations and expertise were crucial in shaping the direction of this project and helping me grow into a researcher.

I would like to extend my heartfelt thanks to my family for their unwavering support and understanding throughout my studies.

Finally, I want to express my deepest gratitude to my wife, Tina Xu, who first encouraged me to pursue graduate studies and who has provided me with love and strength throughout my academic journey. Without her, this work would not have been possible.

Table of Contents

List of Figures

List of Abbreviations

ODE Ordinary Differential Equation 5, 49

PDE Partial Differential Equation 1, 2

List of Symbols

$\mathcal{H}$ A real Hilbert Space. 7

$L^p(U)$ The p^{th} Lebesgue Space. 9

$H^k(U)$ The k^{th} Sobolev Space with respect to the second Lebesgue Space. 10

$W^{k,p}(U)$ The k^{th} Sobolev Space with respect to the p^{th} Lebesgue Space. 10

∂U The boundary of the given domain U. Namely, the difference between its closure and interior. 3

U Spatial Domain: An open and Lebesgue measurable subset of the Euclidean space $\mathbb{R}^n$, which represents the spatial domain of the problem. 3

$\langle\phi|$ A linear functional on the Hilbert space, in ket-bra notation. 7

$\bar{S}$ Closure of a subset $S \subseteq X$. 6

$V \subset\subset U$ V is compactly contained in U 10

$\delta(x - r)$ Dirac-delta function: An abuse of notation for the point measure at r. 13

$\mathcal{H}^*$ Dual Space: Bounded continuous linear operators from $\mathcal{H} \to \mathbb{R}$. 7

$[f]$ The equivalent function class of a Lebesgue measurable function f. 3, 9

f A Lebesgue measurable function from $U \to \mathbb{R}^n$. 3

$\int_U f(x)dx$ The Lebesgue integral of f over U. 8

$|\phi\rangle$ An element in the Hilbert space in ket-bra notation. 7

$L^p_{loc}(U)$ The p^{th} locally summable space. 10

u_x Spatial derivative: The first order (weak) derivative of u, with respect to the spatial variable x in 1 dimension. 3

$y(\cdot, t)$ State: A state at time t. 3

$C_c^\infty(U)$ The smooth functions that are compactly supported in U. 10

$f_{\partial U}$ Trace of a function f on the boundary of U. 3

t Time: A real number $t > 0$ that represents the time elapsed after the initialization of the system. 3

x Spatial Position: A point $x \in \mathbb{R}^n$ in the spatial domain. 3

Chapter 1

Introduction

Numerous fundamental physical phenomena are modelled by Partial Differential Equation (PDE), including "quantum mechanics, relativity, electromagnetism, optics, fluid mechanics, superconductivity, magneto-hydrodynamics, elasticity, thermodynamics, chemical reactions, finance, neuroscience, and many, many more" [20]. For instance, the distribution of heat may be described by a PDE, in particular, by the heat equation [28]

$$\frac{\partial u}{\partial t} = \Delta u + f,$$

where $u(x,t)$ is the heat distribution, and $f(x,t)$ represents an extra source term providing (or draining) energy from the environment.

The controllability theory of PDE aims to solve a fundamental problem: Can a system that is governed by a PDE be guided from a given initial state to a desired target state within a finite time by appropriately selected control inputs? This problem has both theoretical significance and important practical application value, due to the fact that so many phenomena in life are modelled by PDEs. For instance, one may ask the following question: Is it possible to cool down my bedroom within 5 minutes after I come back from work in summer? Such a question is eventually a controllability problem of the heat equation.

Classical results in PDE control theory mainly focus on additive controls, where the control term is added to the system as an external source or force. In the case of the classical heat equation, an additive control of the form $f(x,t)$ that acts either internally or on the boundary is relatively well-studied [28].

In contrast, multiplicative (or bilinear) control has gained increasing attention in recent years. In multiplicative control, the control term is introduced by multiplying the state (and/or its derivative) in the partial differential equation. Notice that in this case, the control effect is linear to both the control and the state itself, while in the additive case, the control term is independent of the state, thus the name bilinear control.

This type of control is motivated by physical scenarios where control is achieved by adjusting a parameter of the system (e.g., reaction rate coefficient, diffusion coefficient, or drift velocity field). For instance, consider the famous Schrödinger equation that governs the movement of a particle in one dimension [12]:

$$i\hbar\frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \Psi}{\partial x^2} + V\Psi,$$

where $\Psi(x,t)$ represents the wave function, $\hbar$ is a physical constant, m is the mass of the particle, and $V(x,t)$ represents the potential energy given by the surrounding environment. In this case, we can see that the only control with a physical meaning would have to appear in the V term, which multiplies to the state itself.

However, from a mathematical perspective, multiplicative control is far more complex: the control term and the state are non-linearly coupled, and the superposition principle no longer applies. Indeed, it is known that exact global controllability is impossible when the state space has infinite dimension in [1]. In particular, exact global controllability is impossible for states in the space of square-integrable functions $L^2(U)$, which is natural to a lot of Physics problems; for instance, the wave function in the Schrödinger equation.

With this in mind, it is natural to focus on the approximate controllability rather than the exact controllability. The term approximate controllability, as its name suggests, aims to steer the system arbitrarily close to a desired final state, while not necessarily reaching it exactly.

In particular, we are interested in the following problem:

> Suppose we are given a system governed by a PDE and a pair of initial state and target state, can we find a bilinear control and a final time such that the final state of the system is arbitrarily close to the target state?

One of the earliest results of multiplicative controls of parabolic equations through the reaction term was established by Alexander Khapalov [17, 16]. In [17], it is shown that global approximate controllability may be achieved for any non-negative initial and target states in a one-dimensional heat equation system, under homogeneous Dirichlet boundary conditions. In [16], the result is generalized to multi-dimensional semi-linear reaction-diffusion equations, where a multiplicative control through the reaction term brings the state arbitrarily close to a target state in arbitrarily small time intervals. For more works about the multiplicative controls through the reaction term, see [15, 18, 8, 21].

On the other hand, another recent trend of study in multiplicative controls considers the case where the control may appear in the advection term (first-order differentiation); in particular, a system governed by the Fokker–Planck equation (also known as the Kolmogorov forward equation in some different contexts):

$$\frac{\partial u}{\partial t} = -\nabla(\mu u) + \Delta(\mathcal{D}u),$$

where $\mu(x,t)$ is the drift coefficient and $\mathcal{D}(x,t)$ is the diffusion coefficient.

For example, in [6], it is shown that if the target state satisfies certain conditions, it may be reached by a multiplicative control on the Kolmogorov forward equation in finite time. In [5], the author showed that a version of the Fokker-Planck equation is locally controllable, given that the trajectory is regular enough.

Since the Fokker–Planck equation naturally arises from stochastic processes [22], such a control may be useful in many situations from physics to engineering; for instance, the Schrodinger-Bridge problem as in [3], or the density control of large-scale particle swarms as in [24], etc.

In this report, we will consider two specific problems in one-dimensional space and establish a relationship between them.

1.1 Notations

In this report, we will use $U \subseteq \mathbb{R}^n$ to denote the spatial domain, which will be open and Lebesgue measurable. It is not necessarily bounded, unless explicitly specified. We will use ∂U to denote its boundary.

We will use $x \in U$ to denote a spatial position in the domain U, and $t > 0$ to denote the time elapsed after the initialization of the system.

We will use $(L^2(U), \|\cdot\|_{L^2(U)})$ to denote the space of square-integrable functions defined on U with its norm. A function $f \in L^2(U)$ is any representative of the equivalent function class $[f] \in L^2(U)$, unless otherwise specified.

All the spatial derivatives are treated as weak derivatives, and all the equalities are understood as equalities of equivalent function class, unless otherwise specified. In one dimension, we will use u_x to denote the first (weak) derivative of u, and u_{xx} to denote the second (weak) derivative of u. See section 2.1.3 for details about the weak derivatives.

We will use $f_{\partial U}$ to denote the trace of f on the boundary of U.

For a function $y : U \times [0, T] \to \mathbb{R}$, we will use $y(\cdot, t) : U \to \mathbb{R}$ to denote the state at a time $t \in [0, T]$, given by $y(\cdot, t)(x) := y(x, t)$ for all $x \in U$. Also, we will use $y|_{\partial U \times (0,T)}$ to denote the trace of y on the boundary of U, given by $y|_{\partial U \times (0,T)}(t) := y(\cdot, t)|_{\partial U}$ for all $t \in (0, T)$.

1.2 Problem Formulation

Problem 1: Approximate Multiplicative Controllability through Reaction Term

Let $U \subseteq \mathbb{R}^n$. Given a pair of initial and target states $y_0, y_d \in L^2(U)$, and $\epsilon > 0$, find $T > 0$ and $\alpha : U \times [0, T] \to \mathbb{R}$, such that the solution of

$$\begin{aligned} \frac{\partial y}{\partial t} &= \Delta y + \alpha y, \quad \forall t \in (0, T) \\ y|_{\partial U \times (0,T)} &= 0, \\ y(\cdot, 0) &= y_0, \end{aligned} \tag{1.1}$$

satisfies $\|y(\cdot,T)-y_d\|_{L^2(U)} < \epsilon$.

Specifically, in one-dimension, we consider the following problem restricted to the bounded spatial domain $U=(0,1)$:

Problem 2: Approximate Multiplicative Controllability through Reaction Term in 1D

Given a pair of initial and target states $y_0, y_d \in L^2(0,1)$, and $\epsilon > 0$, find $T > 0$ and $\alpha : (0,1)\times[0,T] \to \mathbb{R}$, such that the solution of

$$\begin{aligned} \frac{\partial y}{\partial t} &= y_{xx} + \alpha y, \quad \forall t \in (0,T), \\ y(0,t) = y(1,t) &= 0, \quad \forall t \in (0,T), \\ y(\cdot,0) &= y_0, \end{aligned} \tag{1.2}$$

satisfies $\|y(\cdot,T)-y_d\|_{L^2(0,1)} < \epsilon$.

On the other hand, we consider a special form of the Fokker-Planck equation in one-dimension:

Problem 3: Approximate Multiplicative Controllability through Advection Term in 1D

Given a pair of initial and target states $y_0, y_d \in L^2(0,1)$, and $\epsilon > 0$, find $T > 0$ and $\alpha : (0,1)\times[0,T] \to \mathbb{R}$, such that the solution of

$$\begin{aligned} \frac{\partial y}{\partial t} &= y_{xx} + \alpha y_x, \quad \forall t \in (0,T) \\ y(0,t) = y(1,t) &= 0, \quad \forall t \in (0,T) \\ y(\cdot,0) &= y_0, \end{aligned} \tag{1.3}$$

satisfies $\|y(\cdot,T)-y_d\|_{L^2(0,1)} < \epsilon$.

1.3 Outline of the Report

In this report, we aim to:

1. In Chapter 2, we will quickly introduce some mathematical preliminaries involved in this report, including Lebesgue Spaces, weak derivatives, Sobolev Spaces, and some spectral theory. We will also prove some variants of the product rule (corollary 2.5.3), chain rule (corollary 2.5.9), and quotient rule (corollary 2.5.5), which will be used later.

2. In Chapter 3, we will follow Khapalov's paper [17] and do a more detailed analysis of the approximate controllability through the reaction term (Problem 2) on nonnegative states.

3. In Chapter 4, we will establish a relationship between control on reaction and the control through the advection term. We will firstly establish a necessary and sufficient condition for the transformation (theorem 4.1.8) without the boundary condition, then apply it to get several sufficient conditions of transforming Problem 3 to Problem 2. A recursive pseudo-algorithm 2 is given with the result in view. Also, we will apply the result on a specific control given by Khapalov [16], and show that the approximate controllability through the advection term reduces to an Ordinary Differential Equation (ODE) problem.

4. We also prove an explicit construction of uniform step functions approximation in $L^p(a, b)$ (theorem A.0.8) in the Appendix, which might be useful in other scenarios.

Chapter 2

Mathematical Tools

2.1 Preliminary

2.1.1 Hilbert space

Definition 2.1.1. A normed vector space is a vector space $(X, \|\cdot\|)$ over a field $\mathbb{F}$ endowed with a norm (length) function: $\|\cdot\| : X \to [0, \infty)$, such that $\forall x, y \in X, a \in \mathbb{F}$, it satisfies

1. subadditivity (triangular inequality); i.e. $\|x + y\| \leq \|x\| + \|y\|$,
2. absolute homogeneity; i.e. $\|a \cdot x\| = |a|\|x\|$, and
3. positive definiteness; i.e. if $x \neq 0$, we must have $\|x\| > 0$.

Definition 2.1.2. Let $(X, \|\cdot\|)$ be a normed vector space. A subset $S \subseteq X$ is called dense in X if

$$\bar{S} = X,$$

where $\bar{S}$ is the closure of S with respect to the norm $\|\cdot\|$. Namely, for all $x \in X$, $\epsilon > 0$, there is some $y \in S$, such that $\|x - y\| < \epsilon$.

Definition 2.1.3. A real inner product space is a vector space H over $\mathbb{R}$ endowed with an inner product: $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$, such that $\forall u, v, w \in H, a, b \in \mathbb{R}$, it satisfies

1. symmetry; i.e. $\langle v, w \rangle = \langle w, v \rangle$,
2. bi-linearity; i.e. $\langle au + bw, v \rangle = a\langle u, v \rangle + b\langle w, v \rangle$, and
3. positive definiteness; i.e. $\langle v, v \rangle \geq 0$, and if $v \neq 0$, we must have $\langle v, v \rangle > 0$.

Proposition 2.1.1. *For every inner product space with $\langle \cdot, \cdot \rangle$, there is a norm $\|x\| = \sqrt{\langle x, x \rangle}$.*

Definition 2.1.4. Let H be an inner product space. Two vectors $u, v \in H$ are called **orthogonal** if $\langle u, v \rangle = 0$.

Definition 2.1.5. Let H be an inner product space. A set $\{e_i\}_{i\in I} \subseteq H$ is called an **orthonormal set** if

$$\forall i, j \in I, \langle e_i, e_j \rangle = \delta_{ij} := \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}.$$

Definition 2.1.6. Let H be an inner product space. An orthonormal set$\{e_i\}_{i\in I} \subseteq H$ is called a **maximal orthonormal set / orthonormal basis / total orthonormal set** if $\mathrm{Span}(\{e_i\}_{i\in I})$ is dense in H. Namely,

$$H = \overline{\mathrm{Span}(\{e_i\}_{i\in I})}.$$

Definition 2.1.7. A normed vector space X is complete if every Cauchy sequence converges to an element in X with respect to its norm. An inner product space $\mathcal{H}$ is called a Hilbert Space if it is complete with respect to the induced norm.

Theorem 2.1.2 (generalized Fourier series)**.** *Let $\mathcal{H}$ be a Hilbert space, and $\{e_i\}_{i\in I} \subseteq \mathcal{H}$ be an orthonormal set, then the following are equivalent:*

1. *$\{e_i\}_{i\in I}$ is an orthonormal basis.*
2. *If $\forall i \in I, \langle x, e_i \rangle = 0$, then $x = 0$.*
3. *$\forall x \in \mathcal{H}, x = \sum_{i\in I} \langle e_i, x \rangle e_i$. (Fourier series)*
4. *$\forall x \in \mathcal{H}, \|x\|^2 = \sum_{i\in I} |\langle e_i, x \rangle|^2$. (Parseval Identity)*

Definition 2.1.8 (ket-bra notation)**.** For a Hilbert space $\mathcal{H}$ over $\mathbb{R}$, we use $|\phi\rangle$ to represent an element $\phi \in \mathcal{H}$, and use $\langle\phi|$ to represent the element $\langle \phi, \cdot \rangle \in \mathcal{H}^*$, where $\mathcal{H}^*$ is the set of bounded (or equivalently, continuous) linear operators from $\mathcal{H} \to \mathbb{R}$.

Theorem 2.1.3 (Riesz-Frechet Representation theorem)**.** *Let $\mathcal{H}$ be a real Hilbert space, then $\mathcal{H} \cong^* \mathcal{H}$, where the map $\Phi : \mathcal{H} \to \mathcal{H}^*;\ |\phi\rangle \mapsto \langle\phi|$ is the **canonical bijective isometric linear isomorphism**.*

Remark. Riesz-Frechet Representation theorem says that every element in $\mathcal{H}^*$ is actually in the form $\langle\phi|$ for some $\phi \in \mathcal{H}^*$.

Corollary 2.1.4. *Let $\mathcal{H}$ be a Hilbert space, for any (countable) orthornormal basis $\{\omega_k\}_{k\in I}$ of $\mathcal{H}$, we have that $\sum_{k\in I} |\omega_k\rangle \langle\omega_k| = \mathbb{1}$, where $\mathbb{1}$ is the identity map on $\mathcal{H}$.*

Proof. Consider any $|\phi\rangle \in \mathcal{H}$,

$$\begin{aligned}\left(\sum_{k\in I} |\omega_k\rangle \langle\omega_k|\right)|\phi\rangle &= \sum_{k\in I} |\omega_k\rangle\langle\omega_k, \phi\rangle \\ &= \sum_{k\in I} \langle\omega_k, \phi\rangle|\omega_k\rangle \\ &= |\phi\rangle\end{aligned}$$

□

2.1.2 Lebesgue Spaces

Definition 2.1.9. We denote the by λ Lebesgue measure on $\mathbb{R}^n$. We denote $\int_U f d\lambda$ by $\int_U f(x)dx$ for any measurable function f on any measurable set $U \subseteq \mathbb{R}^n$.

We will here define the Lebesgue Spaces and state some useful theorems. Detailed treatment can be found in [9, 23].

Definition 2.1.10. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable. We define

$$\mathcal{L}^1(U) := \left\{ f : U \to \mathbb{R} | \int_U |f(x)|dx < \infty \right\}.$$

Definition 2.1.11. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, and $1 \leq p < \infty$. We define

$$\mathcal{L}^p(U) := \left\{ f : U \to \mathbb{R} | f^p \in \mathcal{L}^1(U) \right\}.$$

In addition, we define the functional

$$\|f\|_{\mathcal{L}^p(U)} := \left(\int_U |f^p(x)|dx \right)^{\frac{1}{p}}$$

for any measurable function $f : U \to \mathbb{R}$.

Remark. We see that

$$\begin{aligned}\mathcal{L}^p(U) &= \left\{ f : U \to \mathbb{R} | f^p \in \mathcal{L}^1(U) \right\} \\ &= \left\{ f : U \to \mathbb{R} | \int_U |f^p(x)|dx < \infty \right\} \\ &= \left\{ f : U \to \mathbb{R} \Big| \|f\|_{\mathcal{L}^p(U)} < \infty \right\}.\end{aligned}$$

Definition 2.1.12. The **essential supremum** of a function $u : U \to \mathbb{R}$ is

$$\operatorname{ess\,sup} f := \inf \left\{ M \in \mathbb{R} : |\{x : f(x) > M\}| = 0 \right\}.$$

Definition 2.1.13. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable. We define

$$\mathcal{L}^\infty(U) := \{ f : U \to \mathbb{R} | \operatorname{ess\,sup} f < \infty \}.$$

In addition, we define the functional

$$\|f\|_{\mathcal{L}^\infty(U)} := \operatorname{ess\,sup} f$$

for any measurable function $f : U \to \mathbb{R}$.

Definition 2.1.14. Two measurable functions $f, g : U \to \mathbb{R}$ are said to be equal almost everywhere if $\{x \in U : f(x) \neq g(x)\}$ has measure zero.

Proposition 2.1.5. *For any $1 \leq p \leq \infty$, we have $\|f-g\|_{\mathcal{L}^p(U)} = 0 \iff f = g$ almost everywhere.*

Definition 2.1.15. For any $1 \leq p \leq \infty$, if we identify $f, g \in \mathcal{L}^p(U)$ by $f \sim g \iff f = g$ almost everywhere, we get the quotient space

$$L^p(U) := \mathcal{L}^p(U)/_\sim = \{[f] : f \in \mathcal{L}^p(U)\}$$

to be the collection of all equivalent function classes $[f]$ in $\mathcal{L}^p(U)$.

Theorem 2.1.6 (Fischer-Riesz)**.** *For any $1 \leq p \leq \infty$, we have the space $(L^p, \|\cdot\|_{L^p(U)})$ is a Banach space, where $\|[f]\|_{L^p(U)} := \|f\|_{\mathcal{L}^p(U)}$ for any representative $f \in [f]$. One can check this norm is well-defined.*

Theorem 2.1.7 (inner product on space of square-integrable functions)**.** *The space $L^2(U)$ is a Hilbert space, where $\|f\|_{L^2(U)}$ is induced by the inner product*

$$\langle [f], [g] \rangle_{L^2(U)} := \int_U f(x)g(x)dx,$$

for any representative $f \in [f], g \in [g]$. One can check this inner-product is well-defined.

In the following sections, we will abuse the notation in a standard way and use $f \in \mathcal{L}^p(U)$ and $[f] \in L^p(U)$ interchangeably. In particular, we write $f \in L^p(U)$ to mean f is a representative of $[f] \in L^p(U)$, namely, $f \in \mathcal{L}^p(U)$. This is due to the fact that whenever we have $f \in \mathcal{L}^p(U)$, we will have a unique $[f] \in L^p(U)$. On the other hand, whenever we have $[f] \in L^p(U)$ and any two representatives $f_1, f_2 \in \mathcal{L}^p(U)$ of the same equivalent class $[f]$, they are identical almost everywhere, and thus the difference is not detectable if we are working with integrals. Also, we may just write $\|\cdot\|_p$ for $\|\cdot\|_{L^p(U)}, \|\cdot\|_{\mathcal{L}^p(U)}$ when the context is clear. We will be more careful when this is not the case. Please refer to chapter 7 of [23] for a more detailed explanation.

Theorem 2.1.8 (Holder's Inequality)**.** *Let $1 \leq p \leq \infty$. Suppose $\frac{1}{p} + \frac{1}{q} = 1$, then $\forall f \in L^p(U), g \in L^q(U),\ fg \in L^1(U)$ and*

$$\|fg\|_{L^1(U)} \leq \|f\|_{L^p(U)} \|g\|_{L^q(U)}.$$

We will also use the following result in [23, Theorem 6.10 and Corollary 6.12], which is sometimes known as the "Fundamental Theorem of Lebesgue Integral Calculus", a generalization of the usual Fundamental Theorem of Calculus.

Theorem 2.1.9 (Fundamental Theorem of Lebesgue Integral Calculus)**.** *A function $f : [a, b] \to \mathbb{R}$ is absolutely continuous if and only if there is a Lebesgue integrable function g, such that $\forall x \in [a, b],\ f(x) = f(a) + \int_a^x g(t)dt$. In this case, f is differentiable a.e., and $f'(x) = g(x)$ for a.e. $x \in [a, b]$.*

2.1.3 Sobolev Spaces

We will here define the weak derivatives, Sobolev Spaces and state some useful theorems. Detailed treatment can be found in [27, 7].

Definition 2.1.16. Let $U, V \subseteq \mathbb{R}^n$ be open, we say that V is **compactly contained** in U if $V \subseteq \bar{V} \subseteq U$, and $\bar{V}$ is compact. We write this as $V \subset\subset U$.

Definition 2.1.17. The **locally summable spaces** are

$$L^p_{loc}(U) := \{u : U \to \mathbb{R} : \forall V \subset\subset U, u \in L^p(V)\}.$$

Definition 2.1.18. We say some property holds in $L^p_{loc}(U)$, if $\forall V \subset\subset U$, it holds in $L^p(V)$.

For instance, let $(f_n)_{n=1}^{\infty} \subseteq L^p_{loc}(U)$ and $f \in L^p_{loc}(U)$, then $f_n \to f$ in $L^p_{loc}(U)$ if $f_n \to f$ in $L^p(V)$, $\forall V \subset\subset U$.

Definition 2.1.19. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable. We define the set of **test functions**

$$C_c^{\infty}(U) := \{\phi \in C^{\infty}(U) : Supp(f) \subseteq U \text{ and is compact}\},$$

where

$$Supp(\phi) := \overline{\{x \in U : \phi(x) \neq 0\}}.$$

Definition 2.1.20. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, and $\alpha \in \mathbb{N}^n$ be an n tuple. For $u, v \in L^1_{loc}(U)$, we say v is the α^{th}-weak derivative of u if

$$\forall \phi \in C_c^{\infty}(U), \int_U u D^{\alpha}\phi dx = (-1)^{|\alpha|} \int_U v\phi dx,$$

where $D^{\alpha}\phi := \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n}\phi$, and $|\alpha| := \sum_{i=1}^n \alpha_i$.

If such a v exists, we say that $D^{\alpha}u = v$ or $u_{\alpha} = v$ in the weak sense. Otherwise, u does not possess a α^{th} weak derivative. One can check that if a weak derivative exists, it is unique almost everywhere.

Definition 2.1.21. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, and $k \in \mathbb{N}$. We define $W^k(U)$ to be the set of functions whose α^{th} weak derivatives exist for all $|\alpha| \leq k$.

Let $1 \leq p \leq \infty$. We define

$$W^{k,p}(U) := \left\{u \in W^k(U) : \forall |\alpha| \leq k,\ D^{\alpha}u \in L^p(U)\right\},$$

where $D^{\alpha}u \in L^p(U)$ is the α^{th} weak derivative of u.

In particular, we define $H^k(U) := W^{k,2}(U)$.

Definition 2.1.22. Let $k \in \mathbb{N}, 1 \leq p \leq \infty, u \in W^k(U)$. The **Sobolev norm** of u is

$$\|u\|_{W^{k,p}(U)} := \begin{cases} \left(\sum_{|\alpha| \leq k} \|D^{\alpha}u\|^p_{L^p(U)}\right)^{1/p}, & 1 \leq p < \infty \\ \sum_{|\alpha| \leq k} \operatorname{ess\,sup}_{x \in U} |D^{\alpha}u(x)| \cong \max_{|\alpha| \leq k} \|D^{\alpha}u\|_{L^{\infty}(U)}, & p = \infty \end{cases}.$$

One can check that these are well-defined norms when restricted to the space $W^{k,p}(U)$, if we identify $u, v \in L^1_{loc}(U)$ by $u \sim v \iff u = v$ almost everywhere, similar to above. Indeed, $u \in W^{k,p}(U) \iff \|u\|_{W^{k,p}(U)} < \infty$.

Theorem 2.1.10 (Completeness of Sobolev Spaces)**.** *Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, and $1 \le p \le \infty$, we have that $W^{1,p}(U)$ is a Banach space with respect to the Sobolev norm. In particular, $H^1(U)$ is a Hilbert space.*

Theorem 2.1.11 (Characterization of Sobolev Spaces using Absolute Continuity)**.** *Let $p \in [0, \infty)$, and $U \subseteq \mathbb{R}^n$ be Lebesgue measurable. Suppose $u \in L^p(U)$, then $u \in W^{1,p}(U)$, if and only if u has a representative $\bar{u}$ that is absolutely continuous on almost all line segments in U parallel to the coordinate axes and whose (classical) partial derivatives belong to $L^p(U)$. [27, Theorem 2.1.4]*

The proof of the above theorem implies that

Corollary 2.1.12. *Let $p \in [0, \infty)$, and $U = (a, b) \subseteq \mathbb{R}$. Suppose $u \in L^p(a, b)$, then $u \in W^{1,p}(a, b)$, if and only if u has a representative $\bar{u}$ that is absolutely continuous, whose (classical) derivative $\frac{d}{dx}\bar{u}$ belongs to $L^p(a, b)$. In this case, $\frac{d}{dx}\bar{u}$ is a representative of the weak derivative u_x.*

The above corollary, in addition to the Fundamental Theorem of Lebesgue Integral Calculus 2.1.9, provides a way of identifying the normal and weak derivatives in the case of 1 dimension. In view of the following result, for any $u \in W^{1,p}(a, b)$, we will always consider the continuous representative of it. We will provide the proof for completeness.

Corollary 2.1.13 (Fundamental Theorem of Lebesgue Integral and Weak Derivative)**.** *Let $p \in [0, \infty]$, and $U = (a, b) \subseteq \mathbb{R}^1$. Suppose $u \in L^p(a, b)$, then the following are equivalent:*

1. *$u \in W^{1,p}(a, b)$,*
2. *u has a representative $\bar{u}$ that is absolutely continuous, whose (classical) derivative $\frac{d}{dx}\bar{u}$ (defined almost everywhere) belongs to $L^p(a, b)$,*
3. *u has a representative $\bar{u}$, and there is a Lebesgue integrable function $g \in L^p(a, b)$, such that $\forall x \in [a, b]$, $\bar{u}(x) = \bar{u}(a) + \int_a^x g(t)dt$.*

In this case, $\frac{d}{dx}\bar{u} = g$ almost everywhere, and is a representative of u_x.

Proof. For $p \in [1, \infty)$, the equivalence of 1. and 2. directly follows from the previous result.

Now assume $p = \infty$.

Suppose 1. holds, since U is bounded, we have $u, u_x \in L^\infty(a, b) \subset L^1(a, b)$, which means $u \in W^{1,1}(a, b)$. Thus, u has a representative $\bar{u}$ that is absolutely continuous, whose (classical) derivative $\frac{d}{dx}\bar{u}$ belongs to $L^1(a, b)$. Also, it is a representative of $u_x \in L^\infty(a, b)$. Thus $\frac{d}{dx}\bar{u}$ belongs to $L^\infty(a, b)$ as well.

On the other hand, suppose 2. holds, then since U is bounded, $\partial_x\bar{u}$, belongs to $L^\infty(a, b) \subseteq L^1(a, b)$. Thus, $u \in W^{1,1}(a, b)$, its weak derivative u_x exists, with $\frac{d}{dx}\bar{u}$ being a representative of u_x. Since the weak derivative is unique, $u_x \in L^\infty(a, b)$, so $u \in H^{1,\infty}(U)$.

2. $\iff$ 3. is always true by the Fundamental Theorem of Lebesgue Integral Calculus (theorem 2.1.9). □

Definition 2.1.23. Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, $H_0^1(U) \subseteq H^1(U)$ is the closure (with respect to $\|\cdot\|_{H^1}$) of $H^1(U) \cap C_c^\infty(U)$.

Remark. $H_0^1(U)$ are the functions in $H^1(U)$ that vanishes at boundaries.

Proposition 2.1.14. *Let $U = (0,1) \subseteq \mathbb{R}^1$, in this case,*

$$H_0^1(0,1) = \left\{ u : u, u_x \in L^2(U) : \lim_{x\to 0} u(x) = \lim_{x \to 1} u(x) = 0 \right\}.$$

Definition 2.1.24. We define $Q_T := (0,1) \times (0,T)$ to be the time-space domain.

2.2 Maximum Principle

In this section, we will state a variation of the maximum principle for the heat equation that would be used in Chapter 3. We will here sketch a proof for the continuous case following Theorem 7.1.9 of [7], and one may follow the steps of Theorem 8.1 in [11] to see the general case for weak solutions.

Theorem 2.2.1 (Maximum principle)**.** *[7] Let $\alpha \in L^\infty(a,b)$ be a non-positive function. For a continuously twice differentiable solution $y(x,t)$ to $y_t = y_{xx} + \alpha y$ with initial state $y(\cdot, 0) = y_0 \geq 0$, we have that*

$$\forall x \in U, t \in (0,T),\ 0 \leq y(x,t) \leq \|y_0\|_\infty.$$

Proof. We know y attains a minimum by the Extreme Value Theorem in $\bar{U} \times [0,T]$ (since y is continuous).

First, we show that $\forall x \in U, t \in (0,T), 0 \leq y(x,t)$. Suppose not, then since $y(\cdot,0) = y_0(\cdot) \geq 0$, and $y_{\partial U} = 0$, the minimum must be attained in $U \times (0,T]$. At that minimum, $y_{xx} > 0, y < 0$, since it must be a local minimum of $y(\cdot,t)$ in the open interior U. Since it is a global minimum of $y(x,\cdot)$, we must have $y_t = 0$. Thus we have $0 = y_t = y_{xx} + \alpha(x)y > 0$, which is a contradiction.

Similarly, if it attains a maximum in the interior or at $t = T$, we will have a contradiction of $0 \geq y_t = y_{xx} + \alpha(x)y < 0$. Thus the maximum must be on the boundary. Since $y_0 \geq 0$ and $y_{\partial U} = 0$, we must have $\forall x \in U, t \in (0,T), y(x,t) \leq \|y_0\|_\infty$. □

Corollary 2.2.2. *Let $\alpha \in L^\infty(a,b)$ be a function, with $C := \|\alpha\|_\infty$. For a continuously twice differentiable solution $y(x,t)$ to $y_t = y_{xx} + \alpha y$ with initial state $y(\cdot,0) = y_0 \geq 0$, we have that*

$$\forall x \in U, t \in (0,T),\ 0 \leq y(x,t) \leq e^{Ct}\|y_0\|_\infty. \tag{2.1}$$

Proof. If $\forall x, \alpha(x) \leq 0$, then this is trivially true from the above theorem, since $1 \leq e^{Ct}$.

Otherwise, consider $z(x,t) := e^{-Ct}y(x,t)$. Note $z_0 = y_0$

$$\begin{aligned} z_t(x,t) &= -Ce^{-Ct}y(x,t) + e^{-Ct}y_t(x,t) \\ &= -Ce^{-Ct}y(x,t) + e^{-Ct}y_{xx}(x,t) + e^{-Ct}\alpha(x)y(x,t) \\ &= z_{xx}(x,t) + (\alpha(x) - C)z(x,t) \end{aligned}$$

Since $\alpha(x) - C \leq 0$ a.e., we can apply the above theorem, and have $\forall x \in U, t \in (0,T), 0 \leq z(x,t) \leq \|z_0\|_\infty = \|y_0\|_\infty$, thus $\forall x \in U, t \in (0,T), 0 \leq y(x,t) \leq e^{Ct}\|y_0\|_\infty$. □

2.3 Ket-Bra Notation on function Spaces

We will now extend the ket-bra notation to our function spaces. Note that it is an abuse of notation to simplify some later calculations in chapter 3, and would require more careful treatment in a formal proof.

Definition 2.3.1. For a function $\omega \in L^2(a,b), x \in (a,b)$, we will abuse the notation $\langle x|\omega\rangle = \langle \omega|x\rangle$ to represent the evaluation map $\omega(x)$.

Remark. Notice that this is *not* the inner product on $L^2(U)$, since the evaluation map, though linear, is not bounded. Thus, it does not live in the dual space, and the Riesz-Frechet Representation theorem does not apply.

Definition 2.3.2. We will use the **Dirac-delta** function $\delta(x-r)$ to represent the Dirac point measure δ_r. Namely, for any function $f : (a,b) \to \mathbb{R}$, we have

$$\int_c^d f(x)\delta(x-r)dx := \int_c^d f(x)d\delta_r = \begin{cases} f(r) & r \in (c,d) \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.3.1. *For functions $\{\omega_k\}_{k=1}^\infty \subset L^2(a,b)$ that form an (countable) orthonormal basis, we (by abusing the notation) have*

$$\forall x, r \in (a,b), \sum_{k=1}^\infty \omega_k(r)\omega_k(x) = \delta(r-x).$$

Proof. Consider any function $u = \sum_{k=1}^\infty \langle \omega_k, u\rangle \omega_k \in L^2(0,1)$, we have that

$$\begin{aligned}
\int_a^b \left(\sum_{k=1}^\infty \omega_k(r)\omega_k(x)\right)u(x)dx &= \sum_{k=1}^\infty \left(\int_a^b \omega_k(x)u(x)dx\right)\omega_k(r) \\
&= \sum_{k=1}^\infty \langle \omega_k, u\rangle \omega_k(r) \\
&= \left\langle \sum_{k=1}^\infty \langle \omega_k, u\rangle \omega_k \middle| r \right\rangle \\
&= \langle u|r\rangle \\
&= u(r) \\
&= \int_a^b \delta(r-x)u(x)dx
\end{aligned}$$

□

Proposition 2.3.2. *$\forall x, r \in (a,b)$, we have (by abusing notation) $\langle r|\mathbb{1}|x\rangle = \delta(r-x)$. We thus write $\langle r|x\rangle := \delta(r-x)$.*

Proof. Consider any (countable) orthonormal basis $\{\omega_k\}_{k=1}^{\infty} \in L^2(0,1)$, we have

$$\begin{aligned}
\delta(r-x) &= \sum_{k=1}^{\infty} \omega_k(r)\omega_k(x) \\
&= \sum_{k=1}^{\infty} \langle r|\omega_k\rangle\langle\omega_k|x\rangle \\
&= \langle r| \sum_{k=1}^{\infty} |\omega_k\rangle \, \langle\omega_k||x\rangle \\
&= \langle r|\mathbb{1}|x\rangle
\end{aligned}$$

□

2.4 Approximation Theorems

This section will aim to prove a result (proposition 2.4.3) about the density of strictly positive step functions in nonnegative functions in $L^2(a,b)$, in a constructive way. The result will be useful later in Chapter 3.

Remark. The density of step functions in $L^p(a,b)$ is relatively well-known, by the regularity of Lebesgue measure. Although that could also be used to prove the result we want, it is not constructive, and the step sizes are not necessarily uniform. Also, it needs to be more carefully treated when we add the restriction that the approximation functions shall be strictly positive, which is essential to a later proof.

With that in mind, we proved an explicit formula (theorem A.0.8) to approximate any $f \in L^p(a,b)$ arbitrarily well, where the step sizes are also uniform. We will not distract the reader here, and we will directly use the result in proving the following lemmas. We have included it in Appendix A for the sake of completeness. Interested readers may find the proof there.

Lemma 2.4.1. *Given any non negative function $g \neq 0 \in L^2(0,1)$, we can find a sequence of positive functions $\{g_k\} \in L^2(0,1)$ such that $\forall K \geq 1$, $\inf_{x\in(0,1)} g_k(x) > 0$, and $\lim_{k\to\infty} g_k = g$ in $L^2(0,1)$.*

Proof. Let $g_k := g + \frac{1}{k}$. It is easy to see that $\inf_{x\in(0,1)} g_k(x) = \inf_{x\in(0,1)} g(x) + \frac{1}{k} \geq \frac{1}{k} > 0$.

In addition,

$$\begin{aligned}\lim_{k\to\infty} \|g - g_k\|_2 &= \left(\int_0^1 \left(\frac{1}{k}\right)^2 dx\right)^{\frac{1}{2}} \\ &= \lim_{k\to\infty} \left(\left(\frac{1}{k}\right)^2\right)^{\frac{1}{2}} \\ &= \lim_{k\to\infty} \frac{1}{k} \\ &= 0\end{aligned}$$

□

Lemma 2.4.2. *Given any positive function $g \in L^2(0,1)$ such that $c := \inf_{x\in(0,1)} g(x) > 0$, we can find a sequence of piecewise constant positive functions $\{g_k\}_{k=1}^{\infty} \in L^2(0,1)$ such that $\lim_{k\to\infty} g_k = g$ in $L^2(0,1)$.*

Proof. Consider $g_k(x) := \sum_{j=1}^{k} \chi_{[x_{j-1},x_j)}(x) k \int_{x_{j-1}}^{x_j} g(x')dx'$, where $x_j := j/k$.

$$\begin{aligned}g_k(x) &\geq k * \int_{x_{j-1}}^{x_j} \inf_{x'\in(0,1)} g(x')dx \\ &= k * \frac{1}{k} \inf_{x'\in(0,1)} g(x') \\ &= c \\ &> 0\end{aligned}$$

Thus g_k is positive. The convergence is by theorem A.0.8. □

Proposition 2.4.3. *Given any non-negative function $g \neq 0 \in L^2(0,1)$, we can find sequence of piecewise constant positive functions $\{g_k\}_{k=1}^{\infty} \in L^2(0,1)$ such that $\lim_{k\to\infty} g_k = g$ in $L^2(0,1)$.*

Proof. By the above lemmas. □

2.5 Weak Derivative Calculus

We here aim to list and prove certain results for weak derivatives that will be used in Chapter 4:

1. Product rule and Quotient rule.

2. Chain rule.

3. Fundamental Theorem of Calculus.

Proposition 2.5.1 (Local Product Rule for Weak Derivatives)**.** *Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, given any $p \in [1, \infty]$, and any $u \in W^{1,p}_{loc}(U)$, $v \in W^{1,\infty}_{loc}(U)$, then $uv \in W^{1,p}_{loc}(U)$ with the i^{th} weak derivative $\partial_i(uv) = u\partial_i v + v\partial_i u$. [25, Proposition 4.1.17]*

Proposition 2.5.2. *Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, $|\alpha| \le n$. For $u, v \in L^1_{loc}(U)$, suppose $v|_V = D^\alpha(u|_V)$ on every $V \subset\subset U$, then v is the α^{th} weak derivative of u on U globally. i.e. $v = D^\alpha u$.*

Proof. Consider any $\phi \in C_c(U)$, we have that $\text{Supp}(\phi) \subset\subset U$, so we can find a $\text{Supp}(\phi) \subset\subset V \subset\subset U$. Now,

$$\int_U \phi v dx = \int_V \phi v dx = (-1)^{|\alpha|} \int_V D^\alpha \phi v dx = (-1)^{|\alpha|} \int_U D^\alpha \phi v dx,$$

since $v|_V = D^\alpha(u|_V)$ and ϕ is constantly 0 outside of V. □

Corollary 2.5.3 (Product Rule for Weak Derivatives)**.** *Let $U \subseteq \mathbb{R}^n$ be Lebesgue measurable, given any $p \in [1, \infty]$, and any $u \in W^{1,p}(U)$, $v \in W^{1,\infty}(U)$, then $uv \in W^{1,p}(U)$ with the i^{th} weak derivative $\partial_i(uv) = u\partial_i v + v\partial_i u$. Also,*

$$\|uv\|_{W^{1,p}(U)} \le \left(1 + n^{\frac{p-1}{p}}\right) \|u\|_{W^{1,p}(U)} \|v\|_{W^{1,\infty}(U)}.$$

[25, Corollary 4.1.18]

Remark. The original statement in [25] claims that $\|uv\|_{W^{1,p}(U)} \le \|u\|_{W^{1,p}(U)} \|v\|_{W^{1,\infty}(U)}$, which is in general not true.

Proof. We firstly note that $\|uv\|_{L^p(U)} \le \|u\|_{L^p(U)} \|v\|_{L^\infty(U)}$. Also, for all $i \in [n]$, by the above proposition 2.5.2, we have that $\partial_i(uv) = u\partial_i v + v\partial_i u$ exists globally in U. In addition,

$$\begin{aligned}
\|(uv)_i\|_{L^p(U)} &\le \|u\partial_i v + v\partial_i u\|_{L^p(U)} \\
&\le \|u\partial_i v\|_{L^p(U)} + \|v\partial_i u\|_{L^p(U)} \\
&\le \|u\|_{L^p(U)} \|\partial_i v\|_{L^\infty(U)} + \|\partial_i u\|_{L^p(U)} \|v\|_{L^\infty(U)} \\
&< \infty.
\end{aligned}$$

This shows that $(uv)_i \in L^p(U)$. Since this holds for all $i \in [n]$, we have $uv \in W^{1,p}(U)$. In addition,

$$
\begin{aligned}
\|uv\|_{W^{1,p}(U)} &= \left(\|uv\|^p_{L^p(U)} + \sum_{i=1}^n \|(uv)_i\|^p_{L^p(U)} \right)^{\frac{1}{p}} \\
&\leq \|uv\|_{L^p(U)} + \sum_{i=1}^n \|(uv)_i\|_{L^p(U)} \\
&\leq \|u\|_{L^p(U)}\|v\|_{L^\infty(U)} + \sum_{i=1}^n \left(\|u\|_{L^p(U)}\|\partial_i v\|_{L^\infty(U)} + \|\partial_i u\|_{L^p(U)}\|v\|_{L^\infty(U)} \right) \\
&\leq \|u\|_{L^p(U)} \left(\|v\|_{L^\infty(U)} + \sum_{i=1}^n \|\partial_i v\|_{L^\infty(U)} \right) + \left(\sum_{i=1}^n \|\partial_i u\|_{L^p(U)} \right) \|v\|_{L^\infty(U)} \\
&\leq \|u\|_{L^p(U)}\|v\|_{W^{1,\infty}(U)} + n^{\frac{p-1}{p}} \left(\sum_{i=1}^n \|\partial_i u\|^p_{L^p(U)} \right)^{\frac{1}{p}} \|v\|_{L^\infty(U)} \\
&\leq \|u\|_{W^{1,p}(U)}\|v\|_{W^{1,\infty}(U)} + n^{\frac{p-1}{p}} \|u\|_{W^{1,p}(U)}\|v\|_{W^{1,\infty}(U)} \\
&= \left(1 + n^{\frac{p-1}{p}}\right) \|u\|_{W^{1,p}(U)}\|v\|_{W^{1,\infty}(U)}.
\end{aligned}
$$

□

Proposition 2.5.4 (Product Rule for Weak Derivatives of Bounded Functions)**.** *Given any $p \in [1, \infty]$, and any $u, v \in W^{1,p}(U) \cap L^\infty(U)$, then $uv \in W^{1,p}(U) \cap L^\infty(U)$ with the i^{th} weak derivative $\partial_i(uv) = u\partial_i v + v\partial_i u$. [2, Proposition 9.4])*

Corollary 2.5.5 (Quotient Rule for Weak Derivatives of Bounded Functions - Necessary Condition)**.** *Given any $p \in [1, \infty]$, and any $u, v \in W^{1,p}(U) \cap L^\infty(U)$, **such that** $\frac{u}{v} \in W^{1,p}(U) \cap L^\infty(U)$, then we have the i^{th} weak derivative is $\partial_i\left(\frac{u}{v}\right) = \frac{\partial_i uv - u\partial_i v}{v^2}$.*

Proof. Since $v\frac{u}{v} = u$, we have by product rule that

$$
\begin{aligned}
\partial_i u &= \partial_i\left(\frac{u}{v}\right)v + \frac{u}{v}\partial_i v \\
\partial_i\left(\frac{u}{v}\right)v &= \frac{\partial_i uv - u\partial_i v}{v} \\
\partial_i\left(\frac{u}{v}\right) &= \frac{\partial_i uv - u\partial_i v}{v^2}
\end{aligned}
$$

□

Remark. Notice that in the above proof, it is very important that we know $\frac{u}{v} \in W^{1,p}(U) \cap L^\infty(U)$ to begin with. i.e. The weak derivative already exists and is in the required space. Thus, this is only a necessary condition. Indeed, consider the simple counter-example given in [4], where two functions can both be weakly differentiable, but their quotient is not.

Proposition 2.5.6 (Local Chain Rule for Weak Derivatives)**.** *Given any $p \in [1, \infty]$. Suppose $u \in W^{1,p}_{loc}(U)$. Suppose $F \in C^1(\mathbb{R})$ has bounded derivative F', then the post-composition $F \circ u$ lies in $W^{1,p}_{loc}(U)$ and its i^{th} weak derivative is given by $\partial_i(F \circ u) = (F' \circ u) \cdot \partial_i u$. [25, Proposition 4.1.21]*

Remark. Since $\frac{1}{x}$ is not $C^1(\mathbb{R})$ (around 0), the quotient rule does not hold in general. Also, although $e^x, x^2 \in C^1(\mathbb{R})$, they don't have a bounded derivative. To this end, we will modify the previous chain rule to get a variation that will weaken the assumption of the behaviour of F over the entire real line.

Proposition 2.5.7 (Local Chain Rule Variation for Weak Derivatives)**.** *Given any $p \in [1, \infty]$. Suppose $u \in W^{1,p}_{loc}(U)$. Suppose $F \in C^1(I)$ has bounded derivative F', where $I \subseteq \mathbb{R}$ is an open interval that contains the closure of u's essential image. i.e.*

$$\overline{\bigcap \{J \subseteq \mathbb{R} : (U \setminus u^{-1}(J)) \text{ has measure } 0\}} \subset I.$$

The post-composition $F \circ u$ lies in $W^{1,p}_{loc}(U)$ and its i^{th} weak derivative is given by $\partial_i(F \circ u) = (F' \circ u) \cdot \partial_i u$.

Proof. Write $I = (a, b)$, where $a < b \in \mathbb{R} \cup \{\pm\infty\}$. WLOG, by redefining u on a measure zero set, we can assume $u(x) \in J$ for all $x \in U$ for some $\overline{J} \subset I$.

Since $\overline{J} \subset I$, we can find some $\delta > 0$, such that $J \subset [a+\delta, b-\delta]$. Indeed, if $a \in \mathbb{R}$, then J is bounded below, and $\inf_{x \in \bar{J}} x$ is achieved by some x_1, since $\bar{J}$ is closed. Since I is open, we can find $\delta_1 > 0$, such that $(x_1 - \delta_1, x_1 + \delta_1) \subset I$. Thus $\forall x \in J,\ x - \delta_1 \geq x_1 - \delta_1 \geq a$. If $a = -\infty$, we can pick any $\delta_2 > 0$, and we always have $\forall x \in J, x - \delta_1 \geq a$. Similarly, we can pick $\delta_1 > 0$, such that $\forall x \in J,\ x + \delta_2 < b$. Now pick $\delta := \min(\delta_1, \delta_2) > 0$, we have $J \subset [a + \delta, b - \delta]$.

Consider the function

$$g(x) := \begin{cases} x, & a + \delta \leq x \leq b - \delta \\ b - \delta e^{\frac{1}{\delta}(b-x-\delta)}, & x \geq b - \delta \\ a + \delta e^{\frac{1}{\delta}(x-a-\delta)}, & x \leq a + \delta \end{cases}.$$

We can check that g is well defined and continuous, and

$$g'(x) = \begin{cases} 1, & a + \delta \leq x \leq b - \delta \\ e^{\frac{1}{\delta}(b-x-\delta)}, & x > b - \delta \\ e^{\frac{1}{\delta}(x-a-\delta)}, & x < a + \delta \end{cases}$$

is also continuous and bounded by 1. Thus $g \in C^1(\mathbb{R})$. In addition, g is monotone increasing and $\forall x \in \mathbb{R},\ g(x) \in (a, b) = I$.

Now consider the composition $G := F \circ g$, which is well-defined. Since $F \in C^1(I)$ has bounded derivative F', we have some $M > 0$, such that $\forall s \in I, |F'(s)| \leq M$. We have by calculus chain rule that $G \in C^1(\mathbb{R})$, with $|G'(s)| = |F'(g(s))g'(s)| = |F'(g(s))||g'(s)| \leq M$, which is also bounded. Thus by the above chain rule, we have $G \circ u \in W^{1,p}_{loc}(U)$, and its i^{th} weak derivative id given by

$$\partial_i(G \circ u) = (G' \circ u) \cdot \partial_i u.$$

However, for any $x \in U$, we have $u(x) \in J \subset [a + \delta, b - \delta]$, so $g(u(x)) = u(x)$, and

$$(G \circ u)(x) = F(g(u(x))) = F(u(x)) = (F \circ u)(x).$$

Also,

$$(G' \circ u)(x) = G'(u(x)) = F'(g(u(x)))\cancelto{1}{g'(u(x))} = F'(u(x)) = (F' \circ u)(x).$$

Thus $G \circ u = F \circ u$, and $G' \circ u = F' \circ u$.

This proves $F \circ u \in W^{1,p}_{loc}(U)$, and its i^{th} weak derivative id given by

$$\partial_i(F \circ u) = \partial_i(G \circ u) = (G' \circ u) \cdot \partial_i u = (F' \circ u) \cdot \partial_i u.$$

□

Remark. Notice that the above proof requires that I is a connected interval, and $\bar{J} \subseteq I$ to find such a $\delta > 0$.

Corollary 2.5.8 (Chain Rule Variation for Weak Derivatives)**.** *Given any $p \in [1, \infty]$. Suppose $u \in W^{1,p}(U)$. Suppose $F \in C^1(I)$ has bounded derivative F', where $I \subseteq \mathbb{R}$ is an open interval that contains the closure of u's essential image, and the post-composition $F \circ u \in L^p(U)$. In this case, $F \circ u$ lies in $W^{1,p}(U)$ and its i^{th} weak derivative is given by $\partial_i(F \circ u) = (F' \circ u) \cdot \partial_i u$.*

Proof. Since $u \in W^{1,p}(U) \subseteq W^{1,p}_{loc}(U)$, we have that $F \circ u \in W^{1,p}_{loc}(U)$, with $\partial_i(F \circ u) = (F' \circ u) \cdot \partial_i u$. Thus, $(F' \circ u) \cdot \partial_i u$ is the global i^{th} weak derivative of $F \circ u$ by proposition 2.5.2.

Now let $M > 0$ be such that $\forall s \in I,\ F'(s) \leq M$. Suppose $p < \infty$,

$$\begin{aligned}
\|F \circ u\|^p_{W^{1,p}(U)} &= \|F \circ u\|^p_{L^p(U)} + \sum_{i \in [n]} \|(F' \circ u) \cdot \partial_i u\|^p_{L^p(U)} \\
&\leq \|F \circ u\|^p_{L^p(U)} + \sum_{i \in [n]} M^p \|\partial_i u\|^p_{L^p(U)} \\
&\leq \|F \circ u\|^p_{L^p(U)} + M^p \|u\|^p_{W^{1,p}(U)} \\
&< \infty.
\end{aligned}$$

Suppose $p = \infty$, we have

$$\begin{aligned}
\|F \circ u\|_{W^{1,\infty}(U)} &= \|F \circ u\|_{L^\infty(U)} + \sum_{i \in [n]} \|(F' \circ u) \cdot \partial_i u\|_{L^\infty(U)} \\
&\leq \|F \circ u\|_{L^\infty(U)} + \sum_{i \in [n]} M \|\partial_i u\|_{L^\infty(U)} \\
&\leq \|F \circ u\|_{L^\infty(U)} + M \|u\|_{W^{1,\infty}(U)} \\
&< \infty.
\end{aligned}$$

Thus, $F \circ u \in W^{1,p}(U)$. □

Notice that we always get $F \circ u \in L^p(U)$ if we assume U is bounded and other assumptions hold. Indeed, we have the following corollary.

Corollary 2.5.9 (Chain Rule Variation for Weak Derivatives on Bounded Domain)**.** *Given any $p \in [1,\infty]$. Suppose $u \in W^{1,p}(U)$, where $U \subset \mathbb{R}^n$ is a bounded open set. Suppose $F \in C^1(I)$ has bounded derivative F', where $I \subseteq \mathbb{R}$ is an open interval that contains the closure of u's essential image. In this case, $F \circ u$ lies in $W^{1,p}(U)$ and its i^{th} weak derivative is given by $\partial_i(F \circ u) = (F' \circ u) \cdot \partial_i u$. In addition, when $p = \infty$, the above holds true even when U is not assumed to be bounded.*

Proof. Let $M > 0$ be such that $\forall s \in I,\ |F'(s)| < M$. Fix any $s \in I$, and consider any $t \neq s \in I$. By the Mean Value Theorem, there is $r \in (\min(s,t), \max(s,t))$ such that $F'(r) = \frac{F(t)-F(s)}{t-s}$. Thus $F(t) = F(s) + F'(r)(t-s)$, and

$$\begin{aligned} |F(t)| &\leq |F(s)| + |F'(r)(t-s)| \\ &\leq |F(s)| + M|t-s| \\ &\leq |F(s)| + M|t| + M|s|. \end{aligned}$$

Let $C := |F(s)| + M|s| < \infty$, we have $\forall t \in I,\ |F(t)| \leq C + M|t|$.

Suppose $p < \infty$, we have

$$\begin{aligned} \|F \circ u\|_{L^p(U)}^p &= \int_U |F(u(x))|^p dx \\ &\leq \int_U |C + M|u(x)||^p dx \\ &\leq \int_U 2^{p-1}(C^p + M^p|u(x)|^p) dx \\ &= \int_U 2^{p-1}C^p dx + \int_U 2^{p-1}M^p|u(x)|^p dx \\ &= 2^{p-1}C^p|U| + 2^{p-1}M^p\|u\|_{L^p(U)}^p \\ &< \infty, \end{aligned}$$

since $u \in W^{1,p}(U) \subset L^p(U)$, and U is bounded.

Now suppose $p = \infty$, we have

$$\begin{aligned} \|F \circ u\|_{L^\infty(U)} &= \operatorname*{ess\,sup}_{x \in U} |F(u(x))| \\ &\leq \operatorname*{ess\,sup}_{x \in U} C + M|u(x)| \\ &= C + M \operatorname*{ess\,sup}_{x \in U} |u(x)| \\ &= C + M\|u\|_{L^\infty(U)} \\ &< \infty. \end{aligned}$$

Thus, $F \circ u \in L^p(U)$ and satisfies the above corollary. $\square$

Corollary 2.5.10. *Given any $p \in [1,\infty]$. Suppose $u \in W^{1,\infty}(U)$. In this case, u^2, e^u both lie in $W^{1,\infty}(U)$ and their i^{th} weak derivative is given by $\partial_i(u^2) = 2u\partial_i u,\ \partial_i(e^u) = e^u \partial_i u$.*

Proof. Let $M := \|u\|_\infty < \infty$. Notice that the essential image of u lies in $[-M, M]$, which is contained in the open interval $I := (-2M, 2M) \subset \mathbb{R}$.

Both $F(s) := s^2, G(s) := e^s$ are in $C^1(I)$, and $F'(s) = 2s$ and $G'(s) = e^s$ are bounded by $4M, e^{2M}$ on I, respectively. $\square$

Chapter 3

Global Nonnegative Controllability through the Reaction Term in 1D

In this chapter, we will follow the idea in [17] to show nonnegative approximate controllability for problem 2. We will first restrict the target states to a dense subset $S \subseteq L^2(0,1)$, and then use the Maximum Principle and some spectral analysis to acquire some estimates on the solution to eq. (1.2). Lastly, we will design a control and show exponential convergence using the previous results.

3.1 Approximation

In this section, we aim to show that the following set S is dense in the set of non-negative functions in space of square-integrable functions.

Definition 3.1.1. Consider the set S of functions $g \in L^2(0,1)$ that satisfies the following:

1. nonzero non-negative continuously differentiable,
2. vanish at $x = 0, 1$,
3. whose second derivatives are piecewise continuous with finitely many discontinuities of the first kind (jump or removable), and
4. the function α_* defined by $\alpha_*(x) := \begin{cases} \frac{g_{xx}(x)}{g(x)} & \text{if } g(x) \neq 0 \\ 0 & \text{if } g(x) = 0 \end{cases}$ is in $L^\infty(0,1)$.

Notice that such a g that satisfies 1. and 3. is always in $W^1(0,1)$ by corollary 2.1.13.

Lemma 3.1.1. *Given any piecewise constant positive function $g = \sum_{j=1}^{n} \alpha_j \chi_{[x_{j-1}, x_j)} \in L^2(0,1)$ with jumps at $0 = x_1 < x_2 < \cdots < x_n = 1$ and $\alpha_j > 0$ being constants, we can find a sequence of continuous bounded piecewise linear positive functions $\{g_k\}_1^\infty \in L^2(0,1)$ such that*

1. $\lim_{k\to\infty} g_k = g$ *in* $L^2(0,1)$, *and*

2. for all $k \geq 1$, $\lim_{x\to 0} g_k(x) = \lim_{x\to 1} g_k(x) = 0$.

Proof. Notice that $\inf_{x\in(0,1)} g(x) = \min_{j\in[n]} \alpha_j > 0$, and $\sup_{x\in(0,1)} g(x) = \max_{j\in[n]} \alpha_j < \infty$.

For each $k \geq 1$, let

$$\epsilon_k := \min(\frac{1}{k}, \frac{1}{4} \min_{1<j\leq n} (x_j - x_{j-1})) > 0.$$

Consider

$$g_k(x) := \begin{cases} g(x_j - \epsilon_k) + \frac{g(x_j+\epsilon_k)-g(x_j-\epsilon_k)}{2\epsilon_k}(x - x_j + \epsilon_k) & \text{if } \exists 1 < j < n, |x - x_j| \leq \epsilon_k \\ \frac{g(\epsilon_k)}{\epsilon_k} x & \text{if } x_0 = 0 < x \leq \epsilon_k \\ \frac{g(1-\epsilon_k)}{\epsilon_k}(1-x) & \text{if } 1 - \epsilon_k \leq x < 1 = x_1 \\ g(x) & \text{otherwise} \end{cases}.$$

Namely, it is the function where we shrink each constant part by ϵ_k at the beginning and end, and then we connect them linearly.

Suppose $\exists 1 \leq j < i \leq n, |x - x_i|, |x - x_j| \leq \epsilon_k$, then

$$\begin{aligned} x_i - x_j &= |x_i - x_j| \\ &\leq |x - x_i| + |x - x_j| \\ &\leq 2\epsilon_k \\ &\leq \frac{2}{4} \min_{1<l\leq n} (x_l - x_{l-1}) \\ &< x_i - x_{i-1}, \end{aligned}$$

a contradiction. Thus, g_k is well-defined. Also, for each $1 < j < n$, we have that

$$\begin{aligned} &g(x_j - \epsilon_k) + \frac{g(x_j + \epsilon_k) - g(x_j - \epsilon_k)}{2\epsilon_k}((x_j - \epsilon_k) - x_j + \epsilon_k) \\ &= g(x_j - \epsilon_k) + \frac{g(x_j + \epsilon_k) - g(x_j - \epsilon_k)}{2\epsilon_k} 0 \\ &= g(x_j - \epsilon_k), \\ &g(x_j - \epsilon_k) + \frac{g(x_j + \epsilon_k) - g(x_j - \epsilon_k)}{2\epsilon_k}((x_j + \epsilon_k) - x_j + \epsilon_k) \\ &= g(x_j - \epsilon_k) + \frac{g(x_j + \epsilon_k) - g(x_j - \epsilon_k)}{2\epsilon_k} 2\epsilon_k \\ &= g(x_j - \epsilon_k) + g(x_j + \epsilon_k) - g(x_j - \epsilon_k) \\ &= g(x_j + \epsilon_k), \end{aligned}$$

Also,

$$\frac{g(\epsilon_k)}{\epsilon_k}\epsilon_k = g(\epsilon_k),$$

$$\frac{g(1-\epsilon_k)}{\epsilon_k}(1-(1-\epsilon_k)) = g(1-\epsilon_k).$$

By gluing lemma, g_k is continuous. It is also easy to see that g_k vanishes at $0, 1$.

In the first case where $\exists 1 < j < n, |x - x_j| < \epsilon_k$, we have that

$$\begin{aligned} g_k(x) &= (1 - \frac{x - x_j + \epsilon_k}{2\epsilon_k})g(x_j - \epsilon_k) + \frac{x - x_j + \epsilon_k}{2\epsilon_k}g(x_j + \epsilon_k) \\ &\geq (1 - \frac{x - x_j + \epsilon_k}{2\epsilon_k}) \inf_{\tilde{x}\in(0,1)} g(\tilde{x}) + \frac{x - x_j + \epsilon_k}{2\epsilon_k} \inf_{\tilde{x}\in(0,1)} g(\tilde{x}) \\ &= \inf_{\tilde{x}\in(0,1)} g(\tilde{x}) \\ &> 0. \end{aligned}$$

Thus, it is positive and bounded below by $\inf_{x\in(0,1)} g(x)$. Similarly, it is bounded above by $\sup_{x\in(0,1)} g(x) < \infty$.

In the second case, if $0 < x < \epsilon_k$, since $g(\epsilon_k) > 0$, we have that

$$0 < \frac{g(\epsilon_k)}{\epsilon_k}x = g_k(x) < \frac{g(\epsilon_k)}{\epsilon_k}\epsilon_k = g(\epsilon_k) \leq \sup_{x\in(0,1)} g(x) < \infty.$$

Thus it is positive and bounded.

Similarly, in the third case, we can see that if $1 - \epsilon_k < x < 1$, we still have $g_k(x)$ being positive and bounded.

In the forth case, we have $g_k(x) = g(x)$, so $g_k(x)$ is bounded above by $\sup_{x\in(0,1)} g(x) < \infty$, and below by $\inf_{x\in(0,1)} g(x) > 0$.

Thus we have that g_k is positive and bounded above by $\sup_{x\in(0,1)} g(x) < \infty$.

In addition, since $\lim_{k\to\infty} \epsilon_k = 0$, point-wise convergence is also easy to see. By A.0.5, it also converges in $L^2(0, 1)$. □

Remark. The density of the continuous functions that satisfies 1. and 2. is a well-known fact in measure theory (Lusin's Theorem [9]). What we really need from this construction is property 3., which will be crucial later.

Lemma 3.1.2. *Given any continuous bounded piecewise linear positive functions $g \in W_0^1(0,1) \subseteq L^2(0,1)$ with "turning points" at $0 = x_1 < x_2 < \cdots < x_n = 1$ such that $\lim_{x\to 0} g(x) = \lim_{x\to 1} g(x) = 0$, we can find a sequence of continuously differentiable bounded positive functions $\{g_k\}_1^\infty \subseteq S$ such that $\lim_{k\to\infty} g_k = g$ in $L^2(0,1)$.*

Proof. Define g_k by replacing each angle generated by the corresponding adjacent straight lines of the graphs around the "turning point" x_j (other than $x_1 = 0, x_n = 1$), by an arc with radius $r_k := \min\left(\frac{1}{k}, \frac{C}{2}, \frac{1}{4}\min_{1<j\le n}(x_j - x_{j-1})\right)$, and tangent to the lines. For each angle, this arc is unique, with its centre located on the bisector of the angle.

Notice that g_k still vanishes at $0, 1$, since we did not change anything from g around $0, 1$. In addition, it is still bounded below and above by the bounds of the original function, thus positive and bounded. It is continuously differentiable because now it is formed only by straight lines and their tangent arcs. Its second derivative is either 0 (on the straight line) or finite and continuous (on the arc), with finitely many discontinuities of the first kind only at the $2(n-1)$ points where the lines and arcs connect.

Lastly, $\sup_{x\in[x_2-r_k, x_{n-1}+r_k]} \frac{\frac{d^2}{dx^2}g_k(x)}{g_k(x)} < \infty$, since it is piecewise continuous on closed (thus compact) intervals $[x_2 - r_k, x_2 + r_k] \cup [x_2 + r_k, x_3 - r_k] \cup \cdots [x_{n-1} - r_k, x_{n-1} + r_k]$. By construction, for all $x \in (0, x_1 - r_k] \cup [x_{n-1} + r_k, 1)$, $g_k(x) = g(x)$ is linear, so $\frac{\frac{d^2}{dx^2}g_k(x)}{g_k(x)} = \frac{0}{g(x)} = 0$. Thus, $\inf_{x\in(0,x_1-r_k]\cup[x_{n-1}+r_k,1)} \frac{\frac{d^2}{dx^2}g_k(x)}{g_k(x)} = 0 < \infty$. This shows $\frac{\frac{d^2}{dx^2}g_k(x)}{g_k(x)} \in L^\infty(0,1)$. Thus we have $\forall k \ge 1, g_k \in S$.

It is easy to see that each g_k is bound above by $\sup_{x\in(0,1)} g(x) < \infty$, and $g_k(x) \to g(x)$ point-wise, since $\lim_{k\to\infty} r_k = 0$, and thus by A.0.5, it also converges in $L^2(0,1)$. □

Theorem 3.1.3. *Any non-negative element $g \in L^2(0,1)$ can be approximated by a sequence of functions $\{g_k\}_{k=1}^\infty \subset S$. Namely, S is dense in $L^2(0,1)$.*

Proof. This directly follows from proposition 2.4.3 and the previous lemmas. □

Remark. Thus in the following sections, we can always assume that the target function y_d is in S as defined in 3.1.1.

3.2 Operator and Eigenfunctions

Here we will state some results about the Spectral problem associated to the differential operator. A more detailed analysis can be found in [7, 11, 26]. In particular, in view of [7, Theorem 6.3.4], we will always consider a weak solution $\omega \in H_0^1(0,1)$ of $\omega_{xx} + (\alpha(x) - \lambda)\omega = 0$ to be in $H^2(0,1)$, since the constant coefficient function $1 \in C^\infty([0,1]) \subset C^1([0,1])$. Namely, ω_{xx} make sense as the second weak derivative.

Proposition 3.2.1. *Let $\alpha \in L^\infty(0,1)$, consider the eigenvalues λ_k and orthonormalized eigenfunctions $\omega_k \in H_0^1(0,1)$ of the spectral problem $\omega_{xx} + \alpha(x)\omega = \lambda\omega$, we have that $\|\alpha\|_{L^\infty(0,1)} \ge \lambda_1 > \lambda_2 > \ldots$, and $\lim_{k\to\infty} \lambda_k = -\infty$. Also, $\{\omega_k\}_{k=1}^\infty$ forms an orthornormal basis for $L^2(U)$.*

Proof. It is known that if we order the eigenvalues by their absolute value, we have $\lim_{k\to\infty} |\lambda_k| = \infty$ for any second-order linear spectral problem. See Theorem 6.5.1 of [7] for the case α is smooth, and the general case may be found in Theorem 8.37 of [11].

In addition, suppose for contradiction that there is some eigenvalue $\lambda_k > \|\alpha\|_{L^\infty(0,1)}$, we must have that $(\omega_k)_{xx} = (\lambda_k - \alpha)\omega_k > 0$ a.e.. By Theorem 4.1.6 in [14], ω_k will be (weakly) convex and thus non-decreasing, and thus we cannot have the boundary condition $\omega_k(0) = \omega_k(1) = 0$ unless $\omega_k = 0$. Thus a contradiction with ω_k being an eigen-function, and we have $\forall k, \|\alpha\|_{L^\infty(0,1)} \geq \lambda_k$.

Since the absolute value of $\omega_k \to \infty$, but it is bounded above, we have that it must tend to negative infinity, thus completing the proof. □

Proposition 3.2.2. *Let $\omega \in H_0^1(0,1), \alpha \in L^\infty(0,1)$, consider the orthonormalized eigenfunctions $\omega_k(x)$ of the spectral problem $\omega_{xx} + \alpha(x)\omega = \lambda\omega$, then $\omega_k(x)$ are also the orthonormalized eigenfunctions of $\omega_{xx} + \beta(x)\omega = \lambda'\omega$, where $\beta = \alpha + c$ for some constant c. In addition, the corresponding eigenvalues are $\lambda'_k = \lambda_k + c$.*

Proof.

$$\begin{aligned}(\omega_k)_{xx} + \alpha(x)\omega_k &= \lambda_k\omega_k \\ (\omega_k)_{xx} + \alpha(x)\omega_k + c\omega_k &= \lambda_k\omega_k + c\omega_k \\ (\omega_k)_{xx} + \beta(x)\omega_k &= (\lambda_k + c)\omega_k\end{aligned}$$

Thus ω_k is still an eigenfunction, with corresponding eigenvalue $\lambda'_k = \lambda_k + c$. □

3.3 Solution to the Initial Boundary Problem

We will here establish an explicit form of the solution to eq. (1.2), and provide some estimates on some of its terms. For the simplicity of calculation, we will use the ket-bra notation with the Dirac-delta function. Please refer to chapter 5 and 6 of [7] to see how this could be rigorously done. Similar as above, we will always consider the weak solution $y(x,t)$ of eq. (1.2) to be in $L^2(H^2(0,1))$. Namely, y_{xx} makes sense as second weak derivative for the spatial variable.

Proposition 3.3.1. *Let $\omega \in H_0^1(0,1)$, consider the eigenvalues λ_k and orthonormalized eigenfunctions $\omega_k(x)$ of the spectral problem $\omega_{xx} + \alpha(x)\omega = \lambda\omega$, where $\lambda_1 > \lambda_2 > \dots$, then*

$$\begin{aligned}y(x,t) :&= \sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x) \\ &= \sum_{k=1}^{\infty} e^{\lambda_k t}\langle\omega_k, y_0\rangle_{L^2(0,1)}\omega_k(x) \\ &= \sum_{k=1}^{\infty} e^{\lambda_k t}|\omega_k\rangle\langle\omega_k|y_0\rangle(x)\end{aligned} \tag{3.1}$$

is a solution to eq. (1.2).

Proof. Since $\omega_k \in H_0^1$, we can directly see that

$$\begin{aligned}
y(0,t) &= \sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(0)\\
&= \sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)0\\
&= 0,\\
y(1,t) &= \sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(1)\\
&= 0.
\end{aligned}$$

In addition,

$$\begin{aligned}
y(x,0) &= \sum_{k=1}^{\infty} \left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\\
&= \int_0^1 y_0(r)\sum_{k=1}^{\infty}\omega_k(r)\omega_k(x)dr\\
&= \int_0^1 y_0(r)\delta(r-x)dr\\
&= y_0(x).
\end{aligned}$$

In addition,

$$\begin{aligned}
y_t &= \frac{d}{dt}\left(\sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\right)\\
&= \sum_{k=1}^{\infty} \left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\frac{d}{dt}e^{\lambda_k t}\\
&= \sum_{k=1}^{\infty} \left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\lambda_k e^{\lambda_k t}\\
&= \sum_{k=1}^{\infty} \lambda_k e^{\lambda_k t}\langle \omega_k, y_0\rangle_{L^2(0,1)}\omega_k(x)\\
y_x &= \frac{d}{dx}\left(\sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\right)\\
&= \sum_{k=1}^{\infty} e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\frac{d}{dx}\omega_k(x)\\
&= \sum_{k=1}^{\infty} e^{\lambda_k t}\langle \omega_k | y_0\rangle_{L^2(0,1)}\frac{d}{dx}\omega_k(x).
\end{aligned}$$

Thus we have

$$
\begin{aligned}
y_{xx} + \alpha y_x &= \sum_{k=1}^{\infty} e^{\lambda_k t} \langle \omega_k, y_0 \rangle_{L^2(0,1)} \frac{d^2}{dx^2} \omega_k(x) + \sum_{k=1}^{\infty} e^{\lambda_k t} \langle \omega_k, y_0 \rangle_{L^2(0,1)} \alpha \frac{d}{dx} \omega_k(x) \\
&= \sum_{k=1}^{\infty} e^{\lambda_k t} \langle \omega_k, y_0 \rangle_{L^2(0,1)} \left(\frac{d^2}{dx^2} \omega_k(x) + \alpha \frac{d}{dx} \omega_k(x) \right) \\
&= \sum_{k=1}^{\infty} e^{\lambda_k t} \langle \omega_k, y_0 \rangle_{L^2(0,1)} \lambda_k \omega_k(x) \\
&= y_t.
\end{aligned}
$$

Thus $y(x,t) := \sum_{k=1}^{\infty} e^{\lambda_k t} \left(\int_0^1 y_0(r) \omega_k(r) dr \right) \omega_k(x)$ is a solution to eq. (1.2). □

Proposition 3.3.2. *For any $n \geq 1$, where $\{\omega_k\}_1^{\infty}$ forms an orthonormal basis, let*

$$
r(x,t) := \sum_{k=n}^{\infty} e^{\lambda_k t} \left(\int_0^1 y_0(r) \omega_k(r) dr \right) \omega_k(x). \tag{3.2}
$$

then

$$
\forall t \geq 0, \ \|r(\cdot,t)\|_2 \leq e^{\lambda_n t} \|y_0\|_2.
$$

Proof.

$$
\begin{aligned}
\|r(\cdot,t)\|_2^2 &= \langle r(\cdot,t), r(\cdot,t) \rangle_{L^2(0,1)} \\
&= \left\langle \sum_{k=n}^{\infty} e^{\lambda_k t} \langle \omega_k | y_0 \rangle \omega_k, \sum_{j=n}^{\infty} e^{\lambda_j t} \langle \omega_j | y_0 \rangle \omega_j \right\rangle_{L^2(0,1)} \\
&= \sum_{k=n}^{\infty} e^{\lambda_k t} \langle \omega_k | y_0 \rangle \sum_{j=n}^{\infty} e^{\lambda_j t} \langle \omega_j | y_0 \rangle \underbrace{\langle \omega_k, \omega_j \rangle_{L^2(0,1)}}_{\delta_{jk}} \\
&= \sum_{k=n}^{\infty} e^{2\lambda_k t} \langle \omega_k | y_0 \rangle^2 \\
&\leq e^{2\lambda_n t} \sum_{k=1}^{\infty} \langle \omega_k | y_0 \rangle^2 \\
&= e^{2\lambda_n t} \sum_{k=1}^{\infty} \langle y_0 | \omega_k \rangle \langle \omega_k | y_0 \rangle \\
&= e^{2\lambda_n t} \langle y_0 | \underbrace{\sum_{k=1}^{\infty} |\omega_k\rangle \langle \omega_k|}_{\mathbb{1}} | y_0 \rangle \\
&= e^{2\lambda_n t} \langle y_0 | y_0 \rangle \\
&= \left(e^{\lambda_n t} \|y_0\|_2 \right)^2
\end{aligned}
$$

□

Corollary 3.3.3. *Let* $y(x,t) := \sum_{k=1}^{\infty} e^{\lambda_k t}\Big(\int_0^1 y_0(r)\omega_k(r)dr\Big)\omega_k(x)$, *where* $\{\omega_k\}_1^{\infty}$ *forms an orthonormal basis, then* $\|y(\cdot,t)\|_2 \leq e^{\lambda_1 t}\|y_0\|_2$

Definition 3.3.1. Given $\alpha \in L^{\infty}(0,1)$, consider the following rescaled norm on $H_0^1(0,1)$:

$$\|u\|_{H_0^1(0,1)} := \left(\int_0^1 (u_x^2(x) + (-\alpha(x)+c)u^2(x))dx\right)^{\frac{1}{2}},$$

where $c > \|\alpha\|_{L^{\infty}(0,1)}$ is a positive number.

Remark. One can show that the above norm is equivalent to $\|u\|_{H^1(0,1)}$. Namely, there is $C_1, C_2 > 0$, such that

$$\forall u \in H_0^1(0,1),\ C_1\|u\|_{H^1(0,1)} \leq \|u\|_{H_0^1(0,1)} \leq C_2\|u\|_{H^1(0,1)}.$$

Proposition 3.3.4. *Let* $y(x,t) := \sum_{k=1}^{\infty} e^{\lambda_k t}\Big(\int_0^1 y_0(r)\omega_k(r)dr\Big)\omega_k(x)$ *as in eq.* (3.1), *then* $\|y(\cdot,t)\|_{C[0,1]} \leq C_*\|y(\cdot,t)\|_{H_0^1(0,1)} \leq C(t)\|y_0\|_2$, *where* C_* *is a positive constant associated with the continuous embedding* $H_0^1(0,1) \subset C[0,1]$ *and the function* $C(t)$ *is nondecreasing.*

Proof. The first inequality follows from Morrey's inequality [7, Theorem 5.7.4] and the definition of Holder norms. We now focus on the second inequality.

Notice that $\lim_{k\to\infty}(c-\lambda_k)e^{2\lambda_k} = 0$, so it is bounded. Namely, there is some

$$C(t) \in \mathbb{R},\ \text{such that}\ \forall k \geq 0, (c-\lambda_k)e^{2\lambda_k t} \leq C^2(t).$$

Now we have

$$
\begin{aligned}
&\int_0^1 cy^2(x,t) - y_t(x,t)y(x,t)dx \\
=&\int_0^1 (cy(x,t) - y_t(x,t))y(x,t)dx \\
=&\int_0^1 \left(\sum_{k=1}^{\infty}(c-\lambda_k)e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)\right)y(x,t)dx \\
=&\int_0^1 \sum_{k,j=1}^{\infty}(c-\lambda_k)e^{\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\omega_k(x)e^{\lambda_j t}\left(\int_0^1 y_0(r)\omega_j(r)dr\right)\omega_j(x)dx \\
=&\sum_{k,j=1}^{\infty}\underbrace{\int_0^1 \omega_k(x)\omega_j(x)dx}_{\delta_{jk}}(c-\lambda_k)e^{\lambda_k t+\lambda_j t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\left(\int_0^1 y_0(r)\omega_j(r)dr\right) \\
=&\sum_{k=1}^{\infty}(c-\lambda_k)e^{2\lambda_k t}\left(\int_0^1 y_0(r)\omega_k(r)dr\right)\left(\int_0^1 y_0(r)\omega_k(r)dr\right) \\
=&\sum_{k=1}^{\infty}(c-\lambda_k)e^{2\lambda_k t}\int_0^1\int_0^1 y_0(r)y_0(r')\omega_k(r)\omega_k(r')drdr' \\
\leq&C^2(t)\sum_{k=1}^{\infty}\int_0^1\int_0^1 y_0(r)y_0(r')\omega_k(r)\omega_k(r')drdr' \\
=&C^2(t)\int_0^1\int_0^1 y_0(r)y_0(r')\underbrace{\sum_{k=1}^{\infty}\omega_k(r)\omega_k(r')}_{\delta(r-r')}drdr' \\
=&C^2(t)\int_0^1 y_0(r)y_0(r)dr \\
=&C^2(t)\|y_0\|_2^2.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\|y(\cdot,t)\|_{H_0^1(0,1)}^2 &= \int_0^1 (y_x^2(x,t) + (-\alpha(x)+c)y^2(x))dx \\
&= \int_0^1 -y(x,t)y_{xx}(x,t) + (-\alpha(x)+c)y^2(x,t)dx \\
&= \int_0^1 y(x,t)(\alpha(x)y(x,t) - y_t(x,t)) + (-\alpha(x)+c)y^2(x,t)dx \\
&= \int_0^1 \alpha(x)(y^2(x,t) - y^2(x,t))dx + \int_0^1 cy^2(x,t) - y_t(x,t)y(x,t)dx \\
&= \int_0^1 cy^2(x,t) - y_t(x,t)y(x,t)dx \\
&= C^2(t)\|y_0\|_2^2
\end{aligned}
$$

Thus $\|y(\cdot,t)\|_{H_0^1(0,1)} \leq C(t)\|y_0\|_2$ □

3.4 Designing the Control

Lemma 3.4.1. *Consider $y_d \in S$ as defined in 3.1.1, and* $\alpha_* := \begin{cases} -\frac{y_{d_{xx}}}{y_d} & \text{if } y_d \neq 0 \\ 0 & \text{if } y_d = 0 \end{cases} \in L^\infty(0,1)$. *We have that $\omega_{k_*} := \frac{y_d(x)}{\|y_d\|_2}$ is an eigenfunction for the spectral problem $y_{xx} + \alpha_* y$, with associated eigenvalue $\lambda_{k_*} = 0$.*

Proof. $\omega_{k_* xx} + \alpha_* \omega_{k_*} = \frac{y_{d_{xx}}(x)}{\|y_d\|_2} - \frac{y_{d_{xx}}(x)}{y_d(x)} \frac{y_d(x)}{\|y_d\|_2} = 0$ if $y_{d_x}(x) \neq 0$, which is everywhere on $(0,1)$.
In addition, $\|\omega_{k_*}\|_2 = \left\| \frac{y_d(x)}{\|y_d\|_2} \right\|_2 = \frac{\|y_d\|_2}{\|y_d\|_2} = 1$ □

Lemma 3.4.2. *Let $\omega \in H_0^1(0,1)$, consider the orthonormalized eigenfunctions $\omega_k(x)$ of the spectral problem $\omega_{xx} + \alpha_*(x)\omega = \lambda\omega$. We have that $\forall m \neq k_*, \omega_m$ changes sign in $(0,1)$, where ω_{k_*} is as in the above lemma.*

Proof. Suppose we have ω_m being non-negative. Then by orthonormality, we have $\int_0^1 \omega_m(x)\omega_{k_*}(x) = 0$. However, $\omega_{k_*}(x)$ is by construction positive everywhere in $(0,1)$, thus $\omega_m(x) = 0$ almost everywhere, which is not an eigenfunction, thus a contradiction.

Similarly, ω_m cannot be non-positive. □

Theorem 3.4.3. *Let $\omega \in H_0^1(0,1)$, consider the orthonormalized eigenfunctions $\omega_k(x)$ of the spectral problem $\omega_{xx} + \alpha_*(x)\omega = \lambda\omega$. We have that $k_* = 1$ and thus $\lambda_1 = 0$, where ω_{k_*} is as in above, and $\lambda_1 > \lambda_2 > \dots$.*

Proof. Suppose $k* > 1$, then we have ω_1 changes sign in $(0,1)$, and thus we can find a positive $y_0 \in L^2(0,1)$, such that $\int_0^1 y_0(x)\omega_1(x)dx < 0$.

Consider $\alpha := \alpha_* - \lambda_1$, then we know that $\omega_k(x)$ are still the eigenfunctions of $\omega_{xx} + \alpha(x)\omega_x = \lambda\omega$, and the corresponding eigenvalues are $\lambda_k - \lambda_1$.

Thus the solution for problem 1.2 is

$$\begin{aligned} y(x,t) &= \sum_{k=1}^{\infty} e^{(\lambda_k - \lambda_1)t} \left(\int_0^1 y_0(r)\omega_k(r)dr \right) \omega_k(x) \\ &= \int_0^1 y_0(r)\omega_1(r)dr\omega_1(x) + \sum_{k=2}^{\infty} e^{(\lambda_k - \lambda_1)t} \left(\int_0^1 y_0(r)\omega_k(r)dr \right) \omega_k(x) \end{aligned}$$

Since $\forall k \geq 2, \lambda_1 > \lambda_1$, we have $\lim_{t\to\infty} y(x,t) = \int_0^1 y_0(r)\omega_1(r)dr\omega_1(x) < 0$ for some $x \in (0,1)$ such that $\omega_1(x) > 0$.

However, this contradicts the maximum principle eq. (2.1).

Thus we must have $k* = 1$ and thus $\lambda_1 = 0$. □

Lemma 3.4.4. *Consider $\alpha := \alpha_* + a$, with $\alpha_*, \omega_k, \lambda_k$ as in above, and $a \in \mathbb{R}$. Let $y(x,t)$ be the solution as in eq. (3.1), we have that*

$$\|y(\cdot,t) - y_d\|_2 \leq \left| e^{at} \int_0^1 y_0(x)\omega_1(x)dx - \|y_d\|_2 \right| + e^{(\lambda_2+a)t}\|y_0\|_2$$

Proof. Notice that from above, we have that for all $k \geq 2, \lambda_k < \lambda_1 = 0$

$$\begin{aligned} y(x,t) &= \sum_{k=1}^{\infty} e^{(\lambda_k+a)t}\Big(\int_0^1 y_0(r)\omega_k(r)dr\Big)\omega_k(x) \\ &= e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big)\omega_1(x) + \sum_{k=2}^{\infty} e^{(\lambda_k+a)t}\Big(\int_0^1 y_0(r)\omega_k(r)dr\Big)\omega_k(x) \\ &= e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big)\omega_1(x) + r(x,t), \\ r(x,t) &:= \sum_{k=2}^{\infty} e^{(\lambda_k+a)t}\Big(\int_0^1 y_0(r)\omega_k(r)dr\Big)\omega_k(x) \end{aligned}$$

In addition,

$$\begin{aligned} \|y(\cdot,t) - y_d\|_2 &= \left\| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big)\omega_1 + r(\cdot,t) - y_d \right\|_2 \\ &\leq \left\| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big)\omega_1 - y_d \right\|_2 + \|r(\cdot,t)\|_2 \\ &= \left\| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big)\frac{y_d}{\|y_d\|_2} - y_d \right\|_2 + \|r(\cdot,t)\|_2 \\ &= \left| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big) - \|y_d\|_2 \right| \left\| \frac{y_d}{\|y_d\|_2} \right\|_2 + \|r(\cdot,t)\|_2 \\ &= \left| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big) - \|y_d\|_2 \right| + \|r(\cdot,t)\|_2 \\ &\leq \left| e^{at}\Big(\int_0^1 y_0(r)\omega_1(r)dr\Big) - \|y_d\|_2 \right| + e^{(\lambda_2+a)t}\|y_0\|_2 \quad 3.3.2 \end{aligned}$$

□

Lemma 3.4.5. *Let $\omega \in H_0^1(0,1)$, consider the orthonormalized eigenfunctions $\omega_k(x)$ of the spectral problem $\omega_{xx} + \alpha_*(x)\omega = \lambda\omega$. Consider any $y_0 \in S$ as defined in 3.1.1, we have $\int_0^1 y_0(x)\omega_1(x)dx > 0$, where $\omega_1 = \omega_{k*}$ is as in the above.*

Proof. This is easy since both $y_0(x)$ and $\omega_{k*}(x)$ are positive in $(0,1)$. □

Theorem 3.4.6. *Consider the setting as above, let $a := \frac{1}{T}\ln\left(\frac{\|y_d\|_2}{\int_0^1 y_0\omega_1 dx}\right)$, we have that*

$$\lim_{T\to\infty} \|y(\cdot,T) - y_d\|_2 = 0$$

Proof. By the above lemma, $a > 0$ is well-defined.

$$\begin{aligned}
\|y(\cdot, T) - y_d\|_2 &\leq \left| e^{aT} \int_0^1 y_0(x)\omega_1(x)dx - \|y_d\|_2 \right| + e^{(\lambda_2 + a)T} \|y_0\|_2 \\
&= \left| \frac{\|y_d\|_2}{\int_0^1 y_0\omega_1 dx} \int_0^1 y_0\omega_1 dx - \|y_d\|_2 \right| + e^{aT} e^{\lambda_2 T} \|y_0\|_2 \\
&= 0 + \frac{\|y_d\|_2}{\int_0^1 y_0\omega_1 dx} e^{\lambda_2 T} \|y_0\|_2 \\
&= \left(\frac{\|y_d\|_2 \|y_0\|_2}{\int_0^1 y_0\omega_1 dx} \right) e^{\lambda_2 T}
\end{aligned}$$

Notice that we have $\lambda_2 < \lambda_1 = 0$, so

$$\lim_{T\to\infty} \|y(\cdot, T) - y_d\|_2 = \lim_{T\to\infty} \left(\frac{\|y_d\|_2 \|y_0\|_2}{\int_0^1 y_0\omega_1 dx} \right) e^{\lambda_2 T} = 0.$$

□

Theorem 3.4.7 (Approximate Controllability of the Heat Equation in 1D)**.** *For any non-negative $u_0, u_d \in L^2(0,1)$ with $u_0 \neq 0$, and $\epsilon > 0$, there is a $T(\epsilon, u_0, u_d) > 0$ and multiplicative control $\alpha \in L^\infty(Q_T)$, such that for the solution 3.1 to the problem 1.2, we have $\|u(\cdot, T) - u_d\|_2 \leq \epsilon$. [15]*

Proof. Since the set S as in definition 3.1.1 is dense in the non-negative functions in $L^2(0,1)$ by theorem 3.1.3, we can approximate u_d by $y_d \in S$ arbitrarily close with $\|u_d - y_d\|_2 < \frac{1}{2}\epsilon$. Pick $\alpha(x,t) := \alpha_*(x) + a(T)$ as from above, we have by previous result that

$$\lim_{T\to\infty} \|u(\cdot, T) - y_d\|_2 = 0,$$

and thus we can find a T big enough so that $\|u(\cdot, T) - y_d\|_2 \leq \frac{1}{2}\epsilon$, so $\|u(\cdot, T) - u_d\|_2 \leq \epsilon$ holds by triangular inequality. □

This also suggests an easy numerical algorithm:

Algorithm 1: Approximate Multiplicative Control through Reaction Term

Data: Initial state $y_0 \in L^2(0,1)$, target state $y_d \in L^2(0,1)$, both positive, and final time T

Result: Multiplicative control α for Problem 2

$x \leftarrow$ partition of $[0,1]$;

$f(x) \leftarrow$ numerical approximation of $y_{d_{xx}}(x)$;

if $y_d(x) \neq 0$ **then**

  $\alpha(x) \leftarrow -\frac{f(x)}{y_d(x)}$;

else

  $\alpha(x) \leftarrow 0$;

$a \leftarrow \frac{1}{T} \ln \left(\frac{\|y_d\|_2^2}{\int_0^1 y_0 y_d dx} \right)$;

$\alpha(x) \leftarrow \alpha(x) + a$;

return α

To serve as an intuitive picture, we have implemented this algorithm in Python, and plotted the following easy example, with final time $T = 2$. See Appendix B.

$$\begin{aligned} y_0(x) &= 8x^2(1-x), \\ y_d(x) &= \sin(\pi x). \end{aligned} \tag{3.3}$$

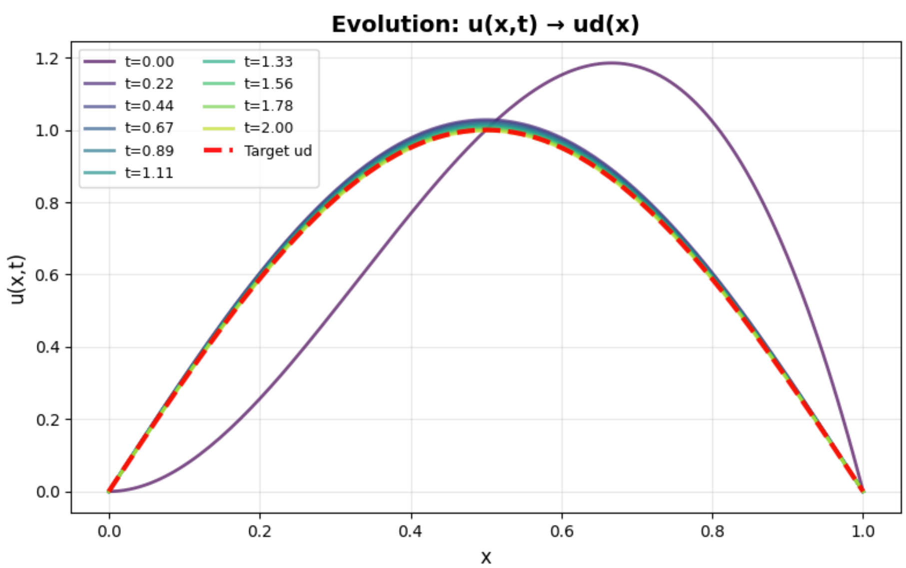

Figure 3.1: Evolution of control through the reaction term, method 1

Also, we plotted the $L^2(0,1)$ difference towards the target state.

Figure 3.2: Difference towards the target state, method 1

Chapter 4

Control through Advection Term via Transformation to Reaction Term

We now want to consider the multiplicative controllability through the Advection Term, as defined in problem 3.

4.1 Transformation between control through Advection Term and Reaction Term

4.1.1 Sturm-Liouville Theory for Continuous Functions

We will here show the motivation of our approach by examining a transformation for continuously differentiable functions. All the partial derivatives in this section shall be understood as normal calculus derivatives, and the equalities of functions shall be understood as pointwise.

Proposition 4.1.1. *Given $y_0 \in C^2(0,1)$, and a control $\alpha \in C^1(0,1)$ for the advection term linear initial problem $\forall x \in (0,1), t \in (0,T)$,*

$$\begin{aligned} \frac{\partial y}{\partial t}(x,t) &= y_{xx}(x,t) + \alpha(x) y_x(x,t) \\ y(x,0) &= y_0(x) \end{aligned}$$

then there is a control $\beta := -\frac{1}{4}\alpha^2 - \frac{1}{2}\alpha_x \in C(0,1)$ that solves the the reaction term linear initial problem $\forall x \in (0,1), t \in (0,T)$,

$$\begin{aligned} \frac{\partial y}{\partial t}(x,t) &= u_{xx}(x,t) + \beta(x) u(x,t) \\ u(x,0) &= y_0(x) e^{\frac{1}{2}\int_c^x \alpha(\tilde{x}) d\tilde{x}}, \end{aligned}$$

where $c \in (0,1)$, and y satisfies for any $(x,t) \in (0,1) \times (0,T)$,

$$u(x,t) = y(x,t) e^{\frac{1}{2}\int_c^x \alpha(z) dz}.$$

Proof. Consider $w(x) := e^{-\frac{1}{2}\int_c^x \alpha(\tilde{x})d\tilde{x}}$, then we have

$$\begin{aligned}
y &= uw \\
w_x(x) &= -\frac{1}{2}\alpha(x)e^{-\frac{1}{2}\int_c^x \alpha(\tilde{x})d\tilde{x}} \\
&= -\frac{1}{2}\alpha(x)w(x) \\
w_{xx}(x) &= \frac{1}{4}\alpha^2(x)e^{-\frac{1}{2}\int_c^x \alpha(z)dz} - \frac{1}{2}\alpha_x(x)e^{-\frac{1}{2}\int_c^x \alpha(\tilde{x})d\tilde{x}} \\
&= \frac{1}{4}\alpha^2(x)w(x) - \frac{1}{2}\alpha_x(x)w(x) \\
\frac{\partial y}{\partial t} &= y_{xx} + \alpha y_x \\
&= (uw_x + u_x w)_x + \alpha(uw_x + u_x w) \\
&= uw_{xx} + u_x w_x + u_x w_x + u_{xx}w + \alpha(x)(uw_x + u_x w) \\
&= wu_{xx} + (2w_x + \alpha w)u_x + (w_{xx} + \alpha w_x)u \\
&= wu_{xx} + \left(2\left(-\frac{1}{2}\alpha w\right) + \alpha w\right)u_x + \left(\frac{1}{4}\alpha^2 w - \frac{1}{2}\alpha_x w - \frac{1}{2}\alpha\alpha w\right)u \\
&= wu_{xx} + 0 \cdot u_x + \left(-\frac{1}{4}\alpha^2 - \frac{1}{2}\alpha_x\right)wu \\
&= w(u_{xx} + \beta u).
\end{aligned}$$

Notice that $y_t = u_t w$ since w is independent of t. Since $w > 0$, dividing both sides by w gives

$$u_t = u_{xx} + \beta u$$

as desired. In addition, the boundary conditions follow by that $u(x,t) = y(x,t)e^{\frac{1}{2}\int_c^x \alpha(z)dz}$. ☐

Proposition 4.1.2. *Given any $\beta : (0,1) \to \mathbb{R}$, we have that $\alpha \in C^1(0,1)$ is a solution to the Riccati equation*

$$\beta(x) = -\frac{1}{4}\alpha^2(x) - \frac{1}{2}\alpha_x(x)$$

if and only if $q(x) \in C^2(0,1)$ is a nonzero solution to

$$q_{xx}(x) + \beta(x)q(x) = 0,$$

where $\alpha = 2\frac{q_x}{q}$.

Proof. Suppose $\beta(x) = -\frac{1}{4}\alpha^2(x) - \frac{1}{2}\alpha_x(x)$. Define $q(x) := e^{\frac{1}{2}\int_c^x \alpha(\tilde{x})d\tilde{x}}$ for any $c \in (0,1)$. Similar to above, we can see that $q_x(x) = \frac{1}{2}\alpha(x)q(x)$, and $q_{xx}(x) = \frac{1}{4}\alpha^2(x)q(x) + \frac{1}{2}\alpha_x(x)q(x)$.

Thus, $\frac{q_x(x)}{q(x)} = \frac{1}{2}\alpha(x)$. Also,

$$
\begin{aligned}
q_{xx}(x) + \beta(x)q(x) &= \frac{1}{4}\alpha^2(x)q(x) + \frac{1}{2}\alpha_x(x)q(x) + \beta(x)q(x) \\
&= q(x)\left(\frac{1}{4}\alpha^2(x) + \frac{1}{2}\alpha_x(x) + \beta(x)\right) \\
&= 0.
\end{aligned}
$$

On the other hand, suppose $q_{xx}(x)+\beta(x)q(x) = 0$, with $q(x) \neq 0$. Define $\alpha(x) := 2\frac{q_x(x)}{q(x)}$, then

$$
\begin{aligned}
-\frac{1}{4}\alpha^2(x) - \frac{1}{2}\alpha_x(x) &= -\frac{1}{4}4\frac{q_x^2(x)}{q^2(x)} - \frac{1}{2}2\frac{q_{xx}(x)q(x) - q_x^2(x)}{q^2(x)} \\
&= -\frac{q_x^2(x) + q_{xx}(x)q(x) - q_x^2(x)}{q^2(x)} \\
&= -\frac{q_{xx}(x)q(x)}{q^2(x)} \\
&= -\frac{-\beta(x)q(x)q(x)}{q^2(x)} \\
&= \beta(x).
\end{aligned}
$$

□

We notice that the proofs in both of the above propositions require

1. Product rule and Quotient rule;
2. Chain rule, so that we have $\frac{d}{dx}e^{f(x)} = e^{f(x)}\frac{d}{dx}f$;
3. Fundamental Theorem of Calculus, so that we have $\frac{d}{dx}\left(\int_c^x \alpha(\tilde{x})d\tilde{x}\right) = \alpha(x)$.

We thus invoke results discussed in Section 2.5 in the next section to get a result for weak derivatives.

4.1.2 Transformation of Control

Theorem 4.1.3. *Given $y_0, y_d \in L^2(0,1)$, and $T > 0$. Consider all spatial derivatives below as weak derivatives, and all equalities as equalities of equivalent function classes in space of square-integrable functions, we have*

1. *Suppose there is a control $\alpha \in W^{1,\infty}(0,1)$ such that $\forall t \in (0,T)$,*

$$
\begin{aligned}
y(\cdot,t) &\in H^2(0,1), \\
y_t(\cdot,t) &= y_{xx}(\cdot,t) + \alpha y_x(\cdot,t), \\
y(\cdot,0) &= y_0 \\
y(\cdot,T) &= y_d,
\end{aligned}
\tag{4.1}
$$

then there are $\beta := -\frac{1}{4}\alpha^2 - \frac{1}{2}\alpha_x \in L^\infty(0,1)$, *and* $q := \left(x \mapsto e^{\frac{1}{2}\int_c^x \alpha(\tilde{x})d\tilde{x}}\right) \in W^{2,\infty}(0,1)$, *such that*

(a) β *is a control for*

$$\begin{aligned} u(\cdot,t) &\in H^2(0,1), \\ u_t(\cdot,t) &= u_{xx}(\cdot,t) + \beta u(\cdot,t), \\ u(\cdot,0) &= y_0 q, \\ u(\cdot,T) &= y_d q, \end{aligned} \tag{4.2}$$

(b)

$$q_{xx} + \beta q = 0, \tag{4.3}$$

and

(c) $\frac{1}{q} \in W^{2,\infty}(0,1)$.

In this case, we note $\alpha = 2\frac{q_x}{q}$.

2. *Conversely, suppose there is* $q \in W^{2,\infty}(0,1), \beta \in L^\infty(0,1)$ *such that*

   (a) β *is a control for eq.* (4.2)

   (b) *eq.* (4.3) $q_{xx} + \beta q = 0$ *holds*

   (c) $\frac{1}{q} \in W^{2,\infty}(0,1)$,

   then $\alpha := 2\frac{q_x}{q} \in W^{1,\infty}(0,1)$ *is a control for eq.* (4.1)

When 1 or 2 holds, $u(x,t) = q(x)y(x,t)$ *a.e..*

Proof. 1 is true by following the earlier proof for $\alpha \in C^1(0,1)$, and noticing that the conditions for product rule 2.5.3, chain rule 2.5.9, and the Fundamental Theorem of Lebesgue Integral Calculus with Weak Derivative 2.1.13 are all satisfied when we want to use them. We will focus on the proof of 2.

Let

$$y(x,t) := u(x,t)\frac{1}{q(x)}.$$

By construction, $y(\cdot,0) = y_0, y(\cdot,T) = y_d$. By linearity of the partial derivative, we have

$$y_t(\cdot,t) = u_t(\cdot,t)\frac{1}{q},$$

since $\frac{1}{q}$ does not depend on t. Since $1, q, \frac{1}{q} \in W^{1,\infty}(0,1) \cap L^\infty(0,1)$, by the necessary condition of the quotient rule 2.5.5, we have

$$\left(\frac{1}{q}\right)_x = \frac{0 - q_x}{q^2} = -\frac{q_x}{q^2}.$$

Also, by product rule 2.5.3 on $q_x, \frac{1}{q} \in W^{1,\infty}(0,1)$, we have $\frac{q_x}{q} \in W^{1,\infty}(0,1)$, and

$$\left(\frac{q_x}{q}\right)_x = \frac{q_{xx}}{q} - q_x\left(\frac{1}{q}\right)_x = \frac{q_{xx}}{q} + \frac{q_x^2}{q^2}.$$

Notice that this show $\alpha := 2\frac{q_x}{q} \in W^{1,\infty}(0,1)$ as well.

In addition, since $\frac{q_x}{q}, q, \left(\frac{1}{q}\right)_x = \frac{-\frac{q_x}{q}}{q} \in W^{1,\infty}(0,1) \cap L^\infty(0,1)$, by the necessary condition of the quotient rule 2.5.5, we have

$$\begin{aligned}
\left(\frac{1}{q}\right)_{xx} &= \frac{q\left(\frac{-q_x}{q}\right)_x - \frac{-q_x}{q} q_x}{q^2} \\
&= \frac{-q\left(\frac{q_{xx}}{q} - \frac{q_x^2}{q^2}\right) + \frac{q_x}{q} q_x}{q^2} \\
&= \frac{2q_x^2 - q_{xx} q}{q^3}.
\end{aligned}$$

Now, since $u(\cdot,t) \in H^2(0,1) \subset H^1(0,1)$, and $\frac{1}{q} \in W^{1,\infty}(0,1)$, by the product rule 2.5.3, we have $y(\cdot,t) \in H^1(0,1)$, and

$$\begin{aligned}
y_x(\cdot,t) &= \left(\frac{1}{q}\right)_x u(\cdot,t) + \frac{1}{q} u_x(\cdot,t) \\
&= \frac{u_x(\cdot,t) q - u(\cdot,t) q_x}{q^2}.
\end{aligned}$$

Since $u_x(\cdot,t), u(\cdot,t) \in W^1(0,1)$, and $\frac{1}{q}, \left(\frac{1}{q}\right)_x \in W^{1,\infty}(0,1)$, again by product rule 2.5.3 and linearity of weak derivatives, we have $y_x(\cdot,t) \in H^1(0,1)$, and

$$\begin{aligned}
y_{xx}(\cdot,t) &= \left(\frac{1}{q}\right)_{xx} u(\cdot,t) + \left(\frac{1}{q}\right)_x u_x(\cdot,t) + \left(\frac{1}{q}\right)_x u_x(\cdot,t) + \frac{1}{q} u_{xx}(\cdot,t) \\
&= \frac{2q_x^2 - q_{xx} q}{q^3} u(\cdot,t) - 2\frac{q_x}{q^2} u_x(\cdot,t) + \frac{1}{q} u_{xx}(\cdot,t) \\
&= (u_{xx}(\cdot,t) q - u(\cdot,t) q_{xx}) \frac{1}{q^2} - 2\frac{q_x^2}{q^3} u(\cdot,t) - 2\frac{q_x}{q^2} u_x(\cdot,t).
\end{aligned}$$

This shows $y(\cdot,t) \in H^2(0,1)$.

Now take $\alpha = 2\frac{q_x}{q}$, we have

$$\begin{aligned}
&y_{xx} + \alpha y_x \\
&= (u_{xx}(\cdot,t) q - u(\cdot,t) q_{xx}) \frac{1}{q^2} - 2\frac{q_x^2}{q^3} u(\cdot,t) - 2\frac{q_x}{q^2} u_x(\cdot,t) + 2\frac{q_x}{q} \frac{u_x(\cdot,t) q - u(\cdot,t) q_x}{q^2} \\
&= \frac{u_{xx}(\cdot,t) q - u(\cdot,t) q_{xx}}{q^2}.
\end{aligned}$$

Now applying the fact that $q_{xx} + \beta q = 0$ and $u_t(\cdot,t) = u_{xx}(\cdot,t) + \beta u(\cdot,t)$, we have

$$\begin{aligned} y_{xx} + \alpha y_x &= \frac{u_{xx}(\cdot,t)q - u(\cdot,t)q_{xx}}{q^2} \\ &= \frac{u_{xx}(\cdot,t)q + \beta u(\cdot,t)q}{q^2} \\ &= \frac{u_{xx}(\cdot,t) + \beta u(\cdot,t)}{q} \\ &= \frac{u_t(\cdot,t)}{q} \\ &= y_t(\cdot,t). \end{aligned}$$

This completes the proof of 2. □

The above theorem establishes the if and only if relationship between the control of the advection term with (a), (b), and (c). We will now investigate when the latter holds true.

Khapalov's result 3.4.7 gives an approximate control of (a) eq. (4.2), so we will examine the conditions (b) eq. (4.3), and (c) first.

Theorem 4.1.4. *Consider a linear system*

$$\dot{x} = A(t)x + h(t),$$

where $A(t)$ is an $n \times n$ matrix, $h(t)$ is an n-vector, whose elements are integrable on every finite interval. For any initial condition x_0, there is a unique absolutely continuous solution $x(t)$, such that $x(0) = x_0$, and $x(t) = \int_0^t A(s)x(s) + h(s)ds$. (This is a corollary of Carathéodory's Existence Theorem (5.1) and Theorem 5.3, listed on page 30 of [13])

Corollary 4.1.5. *Suppose $\beta \in L^\infty(0,1)$, then for any $q(0), q_x(0) \in \mathbb{R}$, there is always a unique solution $q \in W^{2,\infty}(0,1)$ satisfying*

$$q_{xx} + \beta q = 0 \quad \textit{eq. } (4.3)$$

in the weak sense, where we consider $q, q_x \in W^{1,\infty}(0,1)$ to be their absolute continuous representatives as in corollary 2.1.13.

Proof. Extend β by $\hat{\beta}(x) := \begin{cases} \beta(x), & x \in (0,1) \\ 0, & o.w. \end{cases}$.

Consider $A(x) := \begin{pmatrix} 0 & 1 \\ -\hat{\beta}(x) & 0 \end{pmatrix}$, $h(x) := 0$, and $y_0 = \begin{pmatrix} q(0) \\ q_x(0) \end{pmatrix} \in \mathbb{R}^2$.

Notice that for any finite interval $[0,1] \subset I \subset \mathbb{R}$, we have

$$\int_I \left|-\hat{\beta}\right| dx \leq \int_{\mathbb{R}} \left|\hat{\beta}\right| dx = \int_0^1 |\beta| dx \leq \|\beta\|_{L^\infty(0,1)} < \infty,$$

and $\int_I 1dx = |I| < \infty$, $\int_I 0dx = 0 < \infty$, so it satisfies the condition in the above theorem. We have that there is a unique absolutely continuous solution $y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}$, such that $y(0) = y_0$, and $\partial_x y = A(x)y$ a.e..

Notice that for any $\epsilon > 0$, y_1, y_2 are both absolutely continuous on $(-\epsilon, 1+\epsilon)$. In particular, y_1, y_2 are both continuous on the compact set $[0,1] \subset (-\epsilon, 1+\epsilon)$, so they achieve their minimums and maximums in $[0,1]$. Thus, they are bounded on $(0,1) \subset [0,1]$, which means $y_1, y_2 \in L^\infty(0,1)$.

Now take $q := y_1$, we have that q is absolutely continuous, and its normal derivative $\frac{d}{dx}q(x) = y_2(x)$ a.e.. By the Fundamental Theorem of Lebesgue Integral Calculus with Weak Derivative 2.1.13, we have that $q \in W^{1,\infty}(0,1)$, with the weak derivative $q_x = y_2|_{(0,1)}$.

Similarly, we have $q_x = y_2|_{(0,1)}$ is absolutely continuous, and its normal derivative $\frac{d}{dx}q_x(x) = -\beta(x)y_1(x) = -\beta(x)q(x)$ a.e., so $q_x \in W^{1,\infty}(0,1)$, with the weak derivative $(q_x)_x = -\beta q$.

This shows $q \in W^{2,\infty}(0,1)$, and satisfies $q_{xx} + \beta q = 0$. □

This guarantees that for any initial condition that we would like, and any control β we have for (a) eq. (4.2), we can always find a solution q to (b) eq. (4.3). The problem would be if there is any initial condition that yields a q that satisfies (c). The current condition $\frac{1}{q} \in W^{2,\infty}(0,1)$ is relatively hard to check, so we aim for an easier representation of the condition.

Proposition 4.1.6. *Given any $p \in [1,\infty]$, $u \in W^{1,\infty}(U)$. Suppose $\frac{1}{u} \in L^\infty(U)$; namely, u is bounded away from 0, then $\frac{1}{u}$ lies in $W^{1,\infty}(U)$, and its i^{th} weak derivative is given by $\partial_i(\frac{1}{u}) = -\frac{\partial_i u}{u^2}$. Also, in this case, u does not change sign.*

Proof. Suppose, for contradiction, that u changes sign. By corollary 2.1.13, $u \in W^{1,\infty}(U)$ has a continuous representative.

By the Intermediate Value Theorem, there must be some $x_0 \in U$, such that $u(x_0) = 0$. Now $\lim_{x\to x_0} \frac{1}{u} = \infty$, which contradicts that $\frac{1}{u} \in L^\infty(U)$. This shows u does not change sign.

Let $M := \left\|\frac{1}{u}\right\|_{L^\infty(U)} < \infty, N := \|u\|_{L^\infty(U)}$. Suppose first u is always positive, then $u(x) \in [\frac{1}{M}, N]$ for a.e. $x \in U$. Let $I := (\frac{1}{2M}, 2N) \supset [\frac{1}{M}, N]$, which is an open interval in $\mathbb{R}$ that contains the closure of u's essential image. Since $F(s) := \frac{1}{s} \in C^1(I)$ and has a bounded derivative $|F'(s)| = \left|-\frac{1}{s^2}\right| \leq 4M^2 < \infty$, by the modified chain rule 2.5.9, we have the result.

If u is always negative, we have $u(x) \in [-N, -\frac{1}{M}]$ for a.e. $x \in U$, and we get our result similarly. □

Corollary 4.1.7. *Given any $p \in [1,\infty]$, $u \in W^{2,\infty}(U)$. We have that $\frac{1}{u}$ lies in $W^{2,\infty}(U)$ if and only if $\frac{1}{u} \in L^\infty(U)$; namely, u is bounded away from 0. In this case, u does not change sign.*

Proof. ($\Longrightarrow$): Clearly $\frac{1}{u} \in W^{2,\infty}(U)$ implies $\frac{1}{u} \in L^\infty(U)$ by definition.

($\Longleftarrow$): By the above proposition, we have that $\frac{1}{u} \in W^{1,\infty}(U)$ with $\partial_i(\frac{1}{u}) = -\frac{\partial_i u}{u^2}$. By the corollary of the modified chain rule 2.5.9, we have $\frac{1}{u^2} \in W^{1,\infty}(U)$ as well. Since each $\partial_i u \in W^{1,\infty}(U)$, by the product rule 2.5.3, we have that $\partial_i(\frac{1}{u}) \in W^{1,\infty}(U)$.

Since this hold for all $i \in [n]$, we have that $\frac{1}{u} \in W^{2,\infty}(U)$. $\square$

The above corollary provides a sufficient and necessary condition for (c), and we have the following result.

Theorem 4.1.8 (Necessary and Sufficient Condition for Transformation)**.** *Given $y_0, y_d \in L^2(0,1)$, and $T > 0$. Consider all spatial derivatives below as weak derivatives, and all equalities equivalent classes in $L^2(0,1)$, we have that there is a control $\alpha \in W^{1,\infty}(0,1)$ for*

$$\begin{aligned} y(\cdot,t) &\in H^2(0,1), \\ y_t(\cdot,t) &= y_{xx}(\cdot,t) + \alpha y_x(\cdot,t), \\ y(\cdot,0) &= y_0 \\ y(\cdot,T) &= y_d, \end{aligned} \qquad (4.1)$$

if and only if there are $\beta \in L^\infty(0,1), q(0), q_x(0) \in \mathbb{R}$ such that β is a control for

$$\begin{aligned} u(\cdot,t) &\in H^2(0,1), \\ u_t(\cdot,t) &= u_{xx}(\cdot,t) + \beta u(\cdot,t), \\ u(\cdot,0) &= y_0 q, \\ u(\cdot,T) &= y_d q, \end{aligned} \qquad (4.2)$$

where $q \in W^{2,\infty}(0,1)$ is the unique solution (given by Carathéodory's Existence Theorem) to

$$q_{xx} + \beta q = 0 \quad (4.3),$$

satisfying

$$\frac{1}{q} \in L^\infty(0,1). \tag{4.4}$$

In this case, $u(x,t) = y(x,t)q(x)$ a.e., and $\alpha = 2\frac{q_x}{q}$.

In particular, for homogeneous Dirichlet boundary conditions,

$$u(x,t) = q(x)y(x,t) = 0 \iff y(x) = 0$$

when $\frac{1}{q} \in L^\infty(0,1)$ eq. (4.4) holds. Thus, the above result establishes a transformation between Equation (1.3) and Equation (1.2).

Corollary 4.1.9 (Sufficient Condition for Transformation of Approximate Control)**.** *Given $y_0, y_d \in L^2(0,1)$, and $\epsilon > 0$, $T > 0$. Consider all spatial derivatives below as weak derivatives, and all equalities equivalent classes in $L^2(0,1)$. Suppose for all $\epsilon > 0$, there are*

$\beta \in L^\infty(0,1), q(0), q_x(0) \in \mathbb{R}, M > 0$ such that β is a control for

$$\begin{aligned} u(\cdot,t) &\in H^2(0,1), \\ u_t(\cdot,t) &= u_{xx}(\cdot,t) + \beta u(\cdot,t), \\ u(\cdot,0) &= y_0 q, \\ \|u(\cdot,T) - y_d q\|_{L^2(0,1)} &< \frac{\epsilon}{M}, \end{aligned} \tag{4.5}$$

where $q \in W^{2,\infty}(0,1)$ is the unique solution (given by Carathéodory's Existence Theorem) to

$$q_{xx} + \beta q = 0,$$

satisfying

$$q_{xx} + \beta q = 0 \quad (4.3),$$

and

$$\left\| \frac{1}{q} \right\|_\infty < M, \tag{4.6}$$

then $\alpha := 2\frac{q_x}{q} \in W^{1,\infty}(0,1)$ is a control for

$$\begin{aligned} y(\cdot,t) &\in H^2(0,1), \\ y_t(\cdot,t) &= y_{xx}(\cdot,t) + \alpha y_x(\cdot,t), \\ y(\cdot,0) &= y_0 \\ \|y(\cdot,T) - y_d\| &< \epsilon. \end{aligned} \tag{4.7}$$

Proof. This follows directly from the above result, and the only thing we need to check is the approximate part.

Indeed, suppose $\|u(\cdot,T) - y_d q\|_{L^2(0,1)} < \frac{\epsilon}{M}$. Take $y(x,t) := u(x,t)\frac{1}{q(x)}$ as before.

$$\begin{aligned} \|y(\cdot,T) - y_d\|^2_{L^2(0,1)} &= \int_0^1 \left| u(\cdot,T)\frac{1}{q} - \frac{y_d q}{q} \right|^2 dx \\ &\leq \int_0^1 \left| \frac{1}{q}(u(\cdot,T) - y_d q) \right|^2 dx \\ &\leq \left\| \frac{1}{q} \right\|^2_{L^\infty(0,1)} \int_0^1 |u(\cdot,T) - y_d q|^2 dx \\ &= \left\| \frac{1}{q} \right\|^2_{L^\infty(0,1)} \|u(\cdot,T) - y_d q\|^2_{L^2(0,1)} \\ &< \left\| \frac{1}{q} \right\|^2_{L^\infty(0,1)} \left(\frac{\epsilon}{M}\right)^2 \\ &\leq \epsilon^2. \end{aligned}$$

□

Remark. The above Corollary reduces Problem 3 to Problem 2 and finding appropriate initial conditions for q. One may design an algorithm that picks a certain $q_0 \in W^{2,\infty}(0,1)$ that is bounded away from 0, find a control β_0 for eq. (4.5) with q_0 (which is possible for any M if Problem 2 is solvable as in 3.4.7). Find q_1 such that eq. (4.3) hold with β_0, and then find β_1 that solves eq. (4.5) with q_1. Recursively, one have a sequence of (q_i, β_i) that solves eq. (4.5), with (q_{i+1}, β_i) solves eq. (4.3).

Whether this sequence will eventually converge to a desired (q, β) or not is worth further studying.

Algorithm 2: Approximate Multiplicative Control through Advection Term

Data: Initial state $y_0 \in L^2(0,1)$, target state $y_d \in L^2(0,1)$, both non-negative, tolerance ϵ

Result: Multiplicative control α for Problem 3

$q \leftarrow$ a function in $W^{2,\infty}(0,1)$ that is bounded way from 0;

$M \leftarrow \left\|\frac{1}{q}\right\|_{L^\infty(0,1)}$;

$u_0 \leftarrow y_0 q$;

$u_d \leftarrow y_d q$;

$\epsilon_0 \leftarrow \frac{\epsilon}{M}$;

repeat

- $\beta \leftarrow$ control of Problem 2 with $\|u(\cdot,T) - u_d\|_{L^2(0,1)} < \epsilon_0$;
- $q(0), q'(0) \leftarrow$ appropriate value;
- $q \leftarrow$ solution to eq. (4.3);
- $M \leftarrow \left\|\frac{1}{q}\right\|_{L^\infty(0,1)}$;
- $u_0 \leftarrow y_0 q$;
- $\epsilon_0 \leftarrow \frac{\epsilon}{M}$;
- $u_d \leftarrow y_d q$;
- $u \leftarrow$ solution to eq. (1.2);

until $\|u(\cdot,T) - u_d\|_{L^2(0,1)} < \epsilon_0$;

$\alpha \leftarrow \frac{q_x}{q}$;

return α

An interesting study would be on for what β such a q that satisfies eq. (4.3) and eq. (4.4) exists. That would give us a criterion on what kind of control β through the reaction Term may be transformed to a control α through the Advection Term using the above algorithm.

In particular, it would be nice if the q_i's are uniformly bounded away from 0 by a fixed $\frac{1}{M}$ for the corresponding family of β_i, which will give the approximate controllability that we want.

We will here state a sufficient but not necessary condition on β:

Proposition 4.1.10. *Suppose $\beta \in L^\infty(0,1)$ satisfies $\beta(x) \le 0$ a.e., then for any $q(0) > 0, q_x(0) \ge 0$, the unique (continuous in view of corollary 2.1.13) solution to*

$$q_{xx} + \beta q = 0 \quad (4.3)$$

will satisfy

$$\left\|\frac{1}{q}\right\|_{L^\infty(U)} \le \frac{1}{q(0)} < \infty.$$

Proof. WLOG, we can redefine β on a measure null set so that for all $x \in (0,1), \beta(x) \le 0$.

Since q is continuous and $q(0) > 0$, there must be some neighbourhood $(0, x)$ such that $q > 0$. Let $x_0 := \sup\{x \in (0,1), \text{ such that } q(x) > 0\}$. Since $\beta \le 0$, we must have $q_{xx} \ge 0$ on the entire $(0, x_0)$.

By Theorem 4.1.6 in [14], q will be (weakly) convex and thus non-decreasing and strictly positive on $(0, x_0)$.

Suppose for contradiction that $x_0 \neq 1$, then we must have $\forall \delta > 0, \exists x \in (x_0, x_0 + \delta)$, such that $q(x) \le 0$ by choice of x_0. By continuity of q, we must have $q(x_0) \le 0$, which contradicts q being non-decreasing and strictly positive on $(0, x_0)$.

Thus we have shown $x_0 = 1$, and q is non-decreasing and strictly positive on $(0,1)$. Namely, $\left\|\frac{1}{q}\right\|_{L^\infty(U)} \le \frac{1}{q(0)} < \infty$. □

Corollary 4.1.11 (Transformation of Non-positive Control)**.** *Given $y_0, y_d \in L^2(0,1)$, and $\epsilon > 0$, $T > 0$. Consider all spatial derivatives below as weak derivatives, and all equalities equivalent classes in $L^2(0,1)$. Suppose there is $q(0) > 0$, such that for any $\epsilon > 0$ there is some $q_x(0) \ge 0$, and a non-positive $\beta \in L^\infty(0,1)$ such that β is a control for*

$$\begin{aligned} u(\cdot,t) &\in H^2(0,1), \\ u_t(\cdot,t) &= u_{xx}(\cdot,t) + \beta u(\cdot,t), \\ u(\cdot,0) &= y_0 q, \\ \|u(\cdot,T) - y_d q\|_{L^2(0,1)} &< q(0)\epsilon, \end{aligned} \tag{4.8}$$

where $q \in W^{2,\infty}(0,1)$ is the unique solution (given by Carathéodory's Existence Theorem) to

$$q_{xx} + \beta q = 0 \quad (4.3),$$

then there is a control $\alpha := 2\frac{q_x}{q} \in W^{1,\infty}(0,1)$ for

$$\begin{aligned} y(\cdot,t) &\in H^2(0,1), \\ y_t(\cdot,t) &= y_{xx}(\cdot,t) + \alpha y_x(\cdot,t), \\ y(\cdot,0) &= y_0 \\ \|y(\cdot,T) - y_d\| &< \epsilon. \end{aligned} \qquad (4.7)$$

Proof. Take $M = \frac{1}{q(0)} > 0$, and apply corollary 4.1.9. □

4.2 A Control on Nonnegative States

In this section, we will apply the above result to transform a particular control through the reaction Term to the advection term using corollary 4.1.11. Notice that the results stated in the above section are more general, and may be used to transform other types of controls through the reaction Term.

Although Khapalov's result 3.4.7 guarantees approximate controllability of the reaction term problem in (a), it is not necessarily non-positive. However, another more result from Khapalov may be used to evaluate this problem. Also, in this case, we may be released from the recursive algorithm described in Algorithm 2.

Theorem 4.2.1. *[16, Theorem 3.2] Given $T > 0$ and bounded $U \subsetneq \mathbb{R}^n$. Let $v_* \in W^{2,\infty}(U) \cap H^2(U)$ be such that $v_*(x) \leq L < 0$ a.e. in U for some constant L. Then for any $u_0 \in H_0^1(U), u_d \in L^2(U)$ and $\epsilon > 0$, we can find a $T_* \in (0, T)$, such that all solutions to*

$$\begin{aligned} u_t &= \Delta u + vu, \\ u|_{\partial U \times (0,T)} &= 0, \qquad (1.1) \\ u(\cdot, 0) &= u_0, \end{aligned}$$

with the control

$$v(x) := \frac{1}{T_*} v_*(x)$$

satisfy

$$\|u(\cdot, T_*) - e^{v_*} u_0\|_{L^2(U)} \leq \epsilon.$$

Notice that the control v here will always be negative, which allows us to apply corollary 4.1.11.

In addition, when u_0, u_d are such that $\frac{u_d(x)}{u_0(x)} \leq C < 1$ a.e. for some constant C, it immediate that $v_*(x) := \ln\left(\frac{u_d(x)}{u_0(x)}\right) \leq \ln(C) < \ln(1) = 0$, and the final estimate becomes

$$\|u(\cdot, T_*) - e^{v_*} u_0\|_{L^2(U)} = \left\| u(\cdot, T_*) - \frac{u_d}{u_0} u_0 \right\| = \|u(\cdot, T_*) - u_d\| \leq \epsilon,$$

which is the condition of approximate controllability of Problem 1. Indeed, as a direct result, we have the following Theorem:

Theorem 4.2.2. *[16, Theorem 3.1] Given $T > 0$ and bounded $U \subsetneq \mathbb{R}^n$. Consider any pair of initial and target states $u_0 \in H_0^1(U), u_d \in L^2(U)$ such that*

$$0 < c_1 \leq \frac{u_d(x)}{u_0(x)} \leq c_2 < 1 \text{ a.e.}$$

for some positive constants c_1, c_2. In addition, suppose $\frac{u_d}{u_0} \in W^{2,\infty}(U) \cap H^2(U)$, then for any $\epsilon > 0$, we can find a $T_* \in (0, T)$, such that all solutions to eq. (1.1) with the control

$$v(x) := \frac{1}{T_*} \ln\left(\frac{u_d(x)}{u_0(x)}\right) \qquad (4.9)$$

satisfy

$$\|u(\cdot, T_*) - u_d\|_{L^2(U)} \leq \epsilon.$$

To serve as an intuitive picture, we also implemented this control in Python, for the simple example eq. (3.3). Notice that for this control, Khapalov's analysis suggests the final error will be proportional to T_* [16, equation 3.32 + 3.33], so we selected a final time $T_* = 0.0001$. See Appendix B.

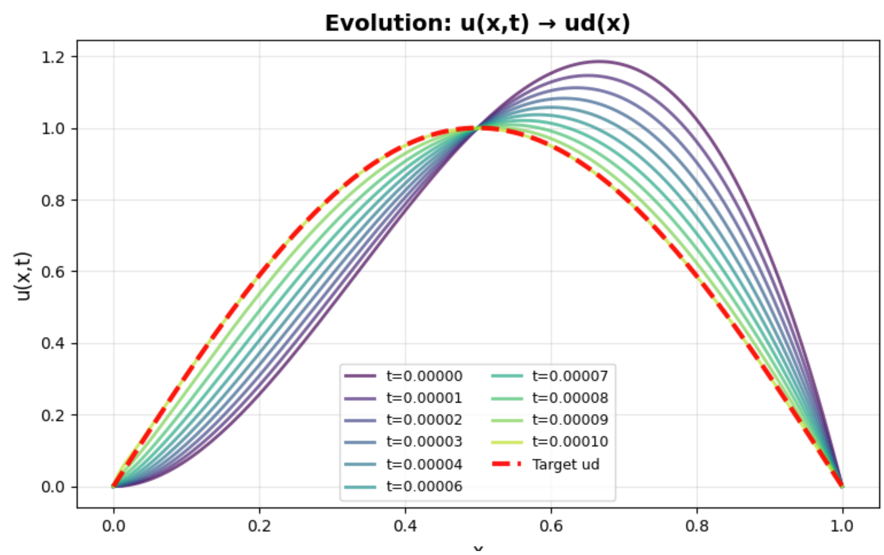

Figure 4.1: Evolution of control through the reaction term, method 2

Also, we plotted the $L^2(0,1)$ difference towards the target state.

Figure 4.2: Difference towards the target state, method 2

In view of the above theorem, we can apply the transformation 4.1.8 and aim to get the following result for the simple case $U = (0, 1)$.

Conjecture 1. *Given $T > 0$ and $U = (0, 1) \subsetneq \mathbb{R}$. Consider any pair of initial and target states $y_0 \in H_0^1(U), y_d \in L^2(U)$ such that*

$$0 < c_1 \leq \frac{y_d(x)}{y_0(x)} \leq c_2 < 1 \text{ a.e.} \tag{4.10}$$

for some positive constants c_1, c_2. For any $\epsilon > 0$, we can find a $T_* \in (0, T)$ and a control $\alpha \in W^{1,\infty}(U)$, such that all solutions to

$$\begin{aligned} y_t &= y_{xx} + \alpha y_x, \\ y(\cdot, 0) &= y_0 \qquad\qquad (1.3) \\ y(0, t) = y(1, t) &= 0, \end{aligned}$$

satisfy

$$\|y(\cdot, T) - y_d\|_{L^2(U)} \leq \epsilon.$$

Proof. We here sketch a proof for the claimed result, but a more detailed analysis on Khapalov's construction and the Comparison Theorems for ODEs is desired to complete the last step.

Consider $v_*(x) := \begin{cases} \ln\left(\frac{y_d(x)}{y_0(x)}\right), & \frac{y_d(x)}{y_0(x)} < c_2, \\ 0, & o.w. \end{cases}$, which by assumption, is zero on a measure null set, and no greater than $\ln c_2 < 0$ elsewhere.

Notice that for any constant $s > 0$, we still have $s v_*(x) < 0$ a.e.. Also, for any $q \in L^\infty(U)$, if we again take $u = yq$, we have that $\frac{u_d}{u_0} = \frac{y_d}{y_0}$ a.e., so the assumption $0 < c_1 \leq \frac{u_d(x)}{u_0(x)} \leq c_2 < 1$ a.e. always holds, and v_* defined above satisfies $v_* = \ln\left(\frac{u_d(x)}{u_0(x)}\right)$a.e., independent of the choice of q.

From the proof of the above theorem [16, equation 3.32 + 3.33], any solution u to

$$\begin{aligned} u_t &= \Delta u + v u, \\ u|_{\partial U \times (0,T)} &= 0, \\ u(\cdot, 0) &= u_0, \end{aligned}$$

with the control

$$v(x) := s v_*(x)$$

satisfies

$$\left\| u(\cdot, \frac{1}{s}) - u_d \right\|_{L^2(U)}^2 = \left\| u(\cdot, \frac{1}{s}) - e^{v_*} \right\|_{L^2(U)}^2 \leq C \frac{1}{s} \|u_0\|_{H^1(0,1)}^2$$

for some constant $C > 0$ independent of s and q.

In addition, if $\left\| \frac{1}{q} \right\|_{L^\infty(0,1)} < \infty$, i.e. q is bounded away from 0, we have $y(x, t) = 0 \iff u(x, t) = y(x, t) q(x) = 0$, so the boundary conditions are preserved.

In view of corollary 4.1.11, it suffices to find $s > 0, q(0) > 0, q_x(0) \geq 0$, such that

$$C\frac{1}{s}\|u_0\|^2_{H^1(0,1)} < (q(0)\epsilon)^2.$$

Or, equivalently,

$$\frac{1}{s}\left(\frac{\|y_0 q\|_{H^1(0,1)}}{q(0)}\right)^2 < \frac{\epsilon^2}{C}, \tag{4.11}$$

where q is the unique solution to

$$q_{xx} + sv_* q = 0,\ q(0) = q(0),\ q'(0) = q_x(0).$$

We note that by the product rule (corollary 2.5.3),

$$\begin{aligned}\frac{1}{s}\left(\frac{\|y_0 q\|_{H^1(0,1)}}{q(0)}\right)^2 &\leq \frac{1}{s}\left(\frac{2\|y_0\|_{H^1(0,1)}\|q\|_{W^{1,\infty}(0,1)}}{q(0)}\right)^2 \\ &= \frac{2}{s}\|y_0\|^2_{H^1(0,1)}\left(\frac{\|q\|_{W^{1,\infty}(0,1)}}{q(0)}\right)^2 \\ &= \frac{2}{s}\|y_0\|^2_{H^1(0,1)}\left\|\frac{q}{q(0)}\right\|^2_{W^{1,\infty}(0,1)},\end{aligned}$$

where $\|y_0\|^2_{H^1(0,1)}$ is a constant independent of q, and $q_{xx} + sv_* q = 0$ if and only if $\left(\frac{q}{q(0)}\right)_{xx} + sv_*\frac{q}{q(0)} = 0$. Thus, it suffices to show that for all $\epsilon_0 > 0$, there is $s > 0, q_x(0) \geq 0$, such that

$$\frac{1}{s}\|q\|^2_{W^{1,\infty}(0,1)} \leq \epsilon_0, \tag{4.12}$$

where q is the unique solution to the ODE

$$q_{xx} + sv_* q = 0,\ q(0) = 1,\ q'(0) = q_x(0). \tag{4.13}$$

Take $\epsilon_0 := \frac{\epsilon^2}{2C\|y_0\|^2_{H^1(0,1)}}$ will give us the original result we want. □

Remark. Studying eq. (4.13) and showing eq. (4.12) by studying the ODE will complete the proof of the above conjecture, and it is the most straightforward and easiest approach. In the case that it is actually not true, one may go back to eq. (4.11) and study that problem.

The following more general result may be found in more detail in section 3.3 of [16].

Theorem 4.2.3 (Non-negative Controllability through Reaction Term)**.** *Given any $T > 0$ and bounded $U \subsetneq \mathbb{R}^n$ such that ∂U is $C^{3+\lfloor \frac{n}{2} \rfloor}$ smooth. For any $\epsilon > 0$, and any pair of initial and target states $h_0, \tilde{h}_d \in L^2(U)$, which are nonnegative (almost everywhere) in U and $h_0 \neq 0$. There is a control*

$$v(\cdot, t) = \begin{cases} 0, & t \in (0, t_1) \\ \frac{\ln\gamma}{t_2 - t_1}, & t \in [t_1, t_2) \\ \frac{v_\sigma}{t_3 - t_2}, & t \in [t_2, t_3], \end{cases}$$

to

$$h_t = \Delta h + vh,$$
$$h|_{\partial U\times(0,T)} = 0, \qquad (1.1)$$
$$h(\cdot,0) = h_0,$$

where $h_d \in C_c^\infty(U)$ *is such that* $h_d \neq 0,\ \forall x \in U,\ h_d(x) \geq 0,$ *and* $\left\|h_d - \tilde{h}_d\right\|_{L^2(U)} < \epsilon_1$. *Also,*

1. $t_1 \in (0,T)$ *is arbitrary,* $h \in C^2\big(\bar{U}\times(\beta,t_1]\big)$ *for any* $\beta \in (0,t_1)$*, and*
$$h(\cdot,t_1) > 0,\ \ a.e.\ in\ U^o,$$
$$h(\cdot,t_1)|_{\partial U} = 0.$$
2. *There is* $\gamma > 1$ *such that* $\forall x \in \mathrm{Supp}(h_d),\ \gamma h(x,t_1) \geq h_d(x) + 1.$
3. *There is* $t_2 \in (t_1,T)$ *such that* $\|h(\cdot,t_2) - \gamma h(\cdot,t_1)\|_{C(\bar{U})} < \frac{\sigma}{2}$ *for some* $\sigma \in (0,1)$*, which results in*
$$\forall x \in \mathrm{Supp}(h_d),\ h(x,t_2) \geq h_d(x).$$
4. *Define*
$$v_\sigma := \ln\left(\frac{h_d + \frac{1}{2}\sigma^2}{h(\cdot,t_2)+\sigma}\right) \in C^2\big(\bar{U}\big),$$
then
$$\|h_d - e^{v_\sigma}h(\cdot,t_2)\|_{C(\bar{U})} < \sigma^2 + \sigma + \sigma\gamma\|h(\cdot,t_1)\|_{C^2(\bar{U})},$$
and $v_\sigma < 0$ *for all* $x \in U$.
5. *There is* $t_3 \in (t_2,T)$ *that satisfies*
$$\|h(\cdot,t_3) - e^{v_\sigma}h(\cdot,t_2)\|_{L^2(U)} < \sigma.$$

By taking any ϵ_1, σ *so small such that*

$$\epsilon_1 + \sigma + |U|^{\frac{1}{2}}\Big(\sigma^2 + \sigma + \sigma\gamma\|h(\cdot,t_1)\|_{C^2(\bar{U})}\Big) < \epsilon,$$

we have that the control v *satisfies*

$$\|h(\cdot,t_3) - h_d\|_{L^2(U)} < \epsilon.$$

Remark. If we again take $y = \frac{h}{q}$ for some $q \in L^\infty(U)$ that is positively bounded away from 0, when $\sigma \to 0$, we have

$$\begin{aligned} v_\sigma &= \ln\left(\frac{h_d + \frac{1}{2}\sigma^2}{h(\cdot,t_2)+\sigma}\right) \\ &\to \ln\left(\frac{h_d}{h(\cdot,t_2)}\right) \\ &= \ln\left(\frac{y_d}{y(\cdot,t_2)}\right). \end{aligned}$$

With this in mind, we may define $v_\sigma := \ln\left(\frac{y_d+\frac{1}{2}\sigma^2}{y(\cdot,t_2)+\sigma}\right)$ instead, which will be independent of the choice of q.

In addition, if we take $\gamma > 1$ such that $\forall x \in \text{Supp}(y_d),\ \gamma y(x,t_1) \geq y_d(x) + \left\|\frac{1}{q}\right\|_{L^\infty(U)}$, we have

$$\forall x \in \text{Supp}(h_d),\ \gamma h(x,t_1) \geq h_d(x) + q(x)\left\|\frac{1}{q}\right\|_{L^\infty(U)} \geq h_d(x) + 1.$$

Thus, the choice of v can be made only dependent on y_0, y_d, and not on q.

Conjecture 2. *Given any nonnegative $y_0, y_d \in L^2(0,1)$, and $T > 0$. For any $\epsilon > 0$, there is $T^* \in (0,T)$ and a multiplicative control $\alpha \in L^\infty((0,1)\times(0,T^*))$ on the system*

$$\begin{aligned} y_t &= y_{xx} + \alpha y_x, \\ y(\cdot,0) &= y_0 \\ y(0,y) = y(1,t) &= 0, \end{aligned} \tag{1.3}$$

such that

$$\|y(\cdot,T^*) - y_d\|_{L^2(0,1)} < \epsilon.$$

We will propose such a control α through the Advection Term, although a more detailed analysis on Khapalov's construction would be necessary to address some concerns that will be discussed below.

1. Pick any $t_1 \in (0,T)$. Let
$$\alpha(\cdot,t) := 0$$
be constantly zero for $t \in (0,t_1)$. With the same argument as in Khapalov's construction [16], by the smoothing effect and a strong maximum principle, $y \in C^2([0,1]\times(\beta,t_1])$ for any $\beta \in (0,t_1)$, and
$$\begin{aligned} y(\cdot,t_1) &> 0, \text{ a.e. in } (0,1), \\ y(0,t_1) = y(1,t_1) &= 0. \end{aligned}$$

2. Pick $\gamma > 1$ such that $\forall x \in \text{Supp}(y_d),\ \gamma y(x,t_1) \geq y_d(x) + M$. This is possible for any $M > 0$, since $y(\cdot,t_1) > 0$, a.e. in $(0,1)$.

   We note that for any $t_2 \in (t_1,T)$ such that $v = \frac{\ln\gamma}{t_2-t_1} < \pi^2$, we have that $\omega = \sqrt{\frac{\ln\gamma}{t_2-t_1}} < \pi$. Let
$$q(x) := \frac{\cos\left(\omega x - \frac{\omega}{2}\right)}{M\cos\left(\frac{\omega}{2}\right)},$$
   we have $q_{xx}(x) = -\omega^2\frac{\cos\left(\omega x-\frac{\omega}{2}\right)}{M\cos\left(\frac{\omega}{2}\right)} = -vq(x)$. Namely, $q_x x + vq = 0$ in U.

   In addition,
$$\inf_{x\in(0,1)} q(x) = \frac{\cos\left(\omega - \frac{\omega}{2}\right)}{M\cos\left(\frac{\omega}{2}\right)} = \frac{1}{M}.$$

Namely, $\left\|\frac{1}{q}\right\|_{L^\infty(U)} \leq M$. Thus, q satisfies the requirement of the transformation 4.1.8.

Taking $u(x,t) = y(x,t)q(x)$, we recover step 2 of Khapalov's construction.

3. Pick $t_2 \in (t_1, T)$ as in step 3 of Khapalov's, and define

$$\alpha(x,t) := 2\frac{q_x(x)}{q(x)} = -2\omega \frac{\sin\left(\omega x - \frac{\omega}{2}\right)}{\cos\left(\omega x - \frac{\omega}{2}\right)}$$

on $t \in [t_1, t_2)$.

We have that $y(\cdot,t)q = u(\cdot,t)$ on $t \in [t_1, t_2]$ by the transformation 4.1.8.

4. Define

$$v_\sigma := \ln\left(\frac{y_d + \frac{1}{2}\sigma^2}{y(\cdot,t_2) + \sigma}\right),$$

which will be negative.

5. Select a $t_3 \in (t_2, T)$ by modifying the argument before.

Notice that in the case that we can pick $\gamma > 1$ such that $\gamma y(x,t_1) > y_d(x)$ and $\frac{\ln\gamma}{t_2 - t_1} < \pi^2$, we can always apply our transformation. However, this is only a sufficient condition for the transformation. When $\frac{\ln\gamma}{t_2 - t_1} \geq \pi^2$, that does not mean the transformation does not work, and it might be worth further studying. One may apply some results form ODE theory to achieve a better result.

Chapter 5

Conclusion

As a conclusion, we studied the multiplicative controls through the reaction term or through the advection term. Following Khapalov's approach, we have seen that approximate (multiplicative) controllability through the reaction term may be achieved for any pair of nonnegative initial and final states, with more than one possible control.

We also established a transform between the multiplicative control through the reaction term and the multiplicative control through the advection term, using some variation of useful results on weak derivatives. With the transformation, we proposed an algorithm that will exploit the approximate controllability through the reaction term to achieve control through the advection term.

There are several possible directions that may follow this research. First, we can study the ODE eq. (4.3) more carefully to conclude about the convergence of the algorithm 2. Second, we aim to study eq. (4.12) better, to complete the proof of the two conjectures in Section 4.2. Third, it would be interesting to generalize the transformation to higher dimensions and see if similar results will apply. Fourth, we can further weaken the requirement, and only require the solution to be a weak solution to the system (instead of a strong solution in terms of weak derivatives, as is currently done). Lastly, we can try to impose different boundary conditions and analyze how the transform would behave in those cases.

References

[1] J. M. Ball, J. E. Marsden, and M. Slemrod. Controllability for Distributed Bilinear Systems. *SIAM Journal on Control and Optimization*, 20(4):575–597, July 1982.

[2] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, New York, NY, 2011.

[3] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. On the Relation Between Optimal Transport and Schrödinger Bridges: A Stochastic Control Viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, May 2016.

[4] daw. Answer to "Weak derivative of a quotient", August 2024. Available from https://math.stackexchange.com/a/4960688/1411179.

[5] Michel Duprez and Pierre Lissy. Bilinear local controllability to the trajectories of the Fokker–Planck equation with a localized control. *Annales de l'Institut Fourier*, 72(4):1621–1659, September 2022.

[6] Karthik Elamvazhuthi, Hendrik Kuiper, Matthias Kawski, and Spring Berman. Bilinear Controllability of a Class of Advection–Diffusion–Reaction Systems. *IEEE Transactions on Automatic Control*, 64(6):2282–2297, June 2019.

[7] Lawrence C. Evans. *Partial differential equations*. Number volume 19 in Graduate studies in mathematics. American mathematical society, Providence (R.I.), 2nd ed edition, 2010.

[8] G. Floridia, C. Nitsch, and C. Trombetti. Multiplicative controllability for nonlinear degenerate parabolic equations between sign-changing states. *ESAIM: Control, Optimisation and Calculus of Variations*, 26:18, 2020.

[9] Gerald B. Folland. *Real analysis: modern techniques and their applications*. Pure and applied mathematics. New York J. Wiley & sons, Chichester Weinheim [etc.], 2nd ed edition, 1999.

[10] Frigyes Riesz. Sur la convergence en moyenne. 1928.

[11] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Classics in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[12] David J. Griffiths. *Introduction to quantum mechanics*. Pearson international edition. Pearson Prentice Hall, Upper Saddle River, NJ London, 2. ed edition, 2005.

[13] Jack K. Hale. *Ordinary differential equations*. Number 21 in Pure and applied Mathematics. Krieger, Malabar, Fla, 2. ed edition, 1980.

[14] Lars Hörmander. *The Analysis of Linear Partial Differential Operators I*. Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[15] Alexander Y. Khapalov. Global Nonnegative Controllability of the 1-D Semilinear Parabolic Equation. In *Controllability of Partial Differential Equations Governed by Multiplicative Controls*, volume 1995, pages 15–31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. Series Title: Lecture Notes in Mathematics.

[16] Alexander Y. Khapalov. Multiplicative Controllability of the Semilinear Parabolic Equation: A Qualitative Approach. In *Controllability of Partial Differential Equations Governed by Multiplicative Controls*, volume 1995, pages 33–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. Series Title: Lecture Notes in Mathematics.

[17] Khapalov, Alexander Y. Global non-negative controllability of the semilinear parabolic equation governed by bilinear control. *ESAIM: COCV*, 7:269–283, 2002.

[18] M. Jidou Khayar. Controllability of some bilinear and semilinear parabolic problems. *Moroccan Journal of Pure and Applied Analysis*, 5(2):222–234, December 2019.

[19] Norbert Kusolitsch. Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica mathematica Hungarica*, 61(1-2):225–229, 2010. Place: Dordrecht Publisher: Springer Netherlands.

[20] Peter J. Olver. *Introduction to Partial Differential Equations*. Undergraduate Texts in Mathematics. Springer International Publishing, Cham, 2014.

[21] Mohamed Ouzahra. Approximate controllability of the semilinear reaction-diffusion equation governed by a multiplicative control. *Discrete & Continuous Dynamical Systems - B*, 27(2):1075, 2022.

[22] Hannes Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*, volume 18 of *Springer Series in Synergetics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996.

[23] Halsey L. Royden. *Real analysis*. Macmillan [u.a.], New York, 3. ed., 8. [print] edition, 1993.

[24] Carlo Sinigaglia, Andrea Manzoni, and Francesco Braghin. Density Control of Large-Scale Particles Swarm Through PDE-Constrained Optimization. *IEEE Transactions on Robotics*, 38(6):3530–3549, December 2022.

[25] Joa Weber. Introduction to Sobolev Spaces, 2018. Available from https://www.math.stonybrook.edu/~joa/PUBLICATIONS/SOBOLEV.pdf.

[26] Anton Zettl. *Recent Developments in Sturm-Liouville Theory*. De Gruyter, March 2021.

[27] William P. Ziemer. *Weakly Differentiable Functions*, volume 120 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1989.

[28] Enrique Zuazua. Controllability and Observability of Partial Differential Equations: Some Results and Open Problems. In *Handbook of Differential Equations: Evolutionary Equations*, volume 3, pages 527–621. Elsevier, 2007.

APPENDICES

Appendix A

Density of Uniform Step Functions - A constructive proof

Theorem A.0.1 (Lebesgue Monotone Convergence). *Let $(X, \mathcal{A}, \mu)$ be a measure space. Let $f_n : X \to [0, \infty]$ be measurable functions with $0 \leq f_1 \leq f_2 \leq \cdots \leq \infty$. Let $f(x) := \lim_{n\to\infty} f_n(x)$, then $f : X \to [0, \infty]$ is measurable, and*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu$$

Lemma A.0.2 (Fatou's). *Let $(X, \mathcal{A}, \mu)$ be a measure space. Let $f_n : X \to [0, \infty]$ be measurable functions. Then*

$$\int_X (\liminf f_n) d\mu \leq \liminf \int_X f_n d\mu$$

Theorem A.0.3 (Lebesgue Dominated Convergence). *Let $(X, \mathcal{A}, \mu)$ be a measure space. Let $f_n : X \to \mathbb{C}$ be measurable functions, defined almost everywhere on X, such that $f(x) := \lim_{n\to\infty} f_n(x)$ is defined almost everywhere for $x \in X$. If there is $0 \leq g(x) \in \mathcal{L}^1(X)$, such that for almost everywhere $x \in X, \forall n \in \mathbb{N}$, $|f_n(x)| \leq g(x)$, then $f \in \mathcal{L}^1(X)$, and*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu, \lim_{n\to\infty} \int_X |f - f_n| d\mu = 0.$$

Theorem A.0.4 (Lebesgue differentiation). *Let λ denote the Lebesgue measure, let $|A| = \lambda^n(A)$ denote the volume of $A \subseteq \mathbb{R}^n$ under the Lebesgue measure. Consider a family of subsets $\mathcal{V}$ such that*

$$\exists c > 0, \text{ such that } \forall A \in \mathcal{V}, \exists B \text{ a ball, such that } A \subseteq B, |A| \geq c|B|.$$

For $A \ni x$, let $A \to x$ denote the limit where the diameter of A shrinks to 0; namely, A "shrinks" to x. If f is Lebesgue integrable, we have

$$\lim_{A\to x, A\in\mathcal{V}} \frac{1}{|A|} \int_A f(x) dx = f(x), a.e..$$

Proposition A.0.5. *Given any bounded domain U, any function $g \in L^\infty(U)$, and a sequence of functions $\{g_k\}_{k=1}^\infty$ such that it converges point-wise to g almost everywhere. If $\exists M > 0$, such that $\forall k \geq 1, \operatorname{ess\,sup}_U g_k \leq M, \operatorname{ess\,sup}_U g \leq M$, then we have that it also converges in $\|\cdot\|_{L^p(U)}$, for any $1 \leq p < \infty$.*

Proof. We have that $\lim_{k\to\infty} g_k(x) = g(x)$ a.e.. Consider $f_k(x) := |g(x) - g_k(x)|^p$, we have that $\lim_{k\to\infty} f_k(x) = 0$, a.e., and

$$\forall k \geq 1, 0 \leq f_k(x) \leq (|g(x)| + |g_k(x)|)^p \leq (M + M)^p = (2M)^p, a.e.,$$

Notice that $\int_U (2M)^p dx = (2M)^p |U| < \infty$, so $(2M)^p \in \mathcal{L}^1(U)$.
Thus $f_k \to 0$ satisfies Lebesgue's dominated convergence theorem.

$$\begin{aligned}
\lim_{k\to\infty} \|g - g_k\|_{L^p(U)} &= \lim_{k\to\infty} \left(\int_U |g(x) - g_k(x)|^p dx \right)^{\frac{1}{p}} \\
&= \lim_{k\to\infty} \left(\int_U f_k(x) dx \right)^{\frac{1}{p}} \\
&= \left(\lim_{k\to\infty} \int_U f_k(x) dx \right)^{\frac{1}{p}} \\
&= \left(\int_U \lim_{k\to\infty} f_k(x) dx \right)^{\frac{1}{p}} \\
&= \left(\int_U 0 dx \right)^{\frac{1}{p}} \\
&= 0.
\end{aligned}$$

□

Corollary A.0.6. *Consider any $1 \leq p < \infty$. Given any function $g \in L^p(U)$, and a sequence of functions $\{g_k\}_{k=1}^\infty$ such that it converges point-wise to g a.e., and $\forall k \geq 1, |g_k(x)| \leq |g(x)|$ a.e., then we have that it also converges in $\|\cdot\|_{L^p(U)}$.*

Proof. Consider $f_k(x) := |g(x) - g_k(x)|^p$, we have that $\lim_{k\to\infty} f_k(x) = 0$, a.e.and $\forall k \geq 1, 0 \leq f_k(x) \leq 2^p |g(x)|^p$. Notice that $\int_U 2^p |g(x)|^p dx = 2^p \|g\|_{L^p(U)} < \infty$, and thus $f_k \to 0$ satisfies Lebesgue's dominated convergence theorem as above. □

We will prove a lemma that would be useful to show convergence of functions in $\|\cdot\|_{L^p(U)}$. It is a generalization of the Riesz's Lemma in [10].

Lemma A.0.7. *Consider any $1 \leq p < \infty$. Given any function $g \in L^p(U)$, and a sequence of functions $\{g_k\}_{k=1}^\infty$ in $L^p(U)$, such that it converges point-wise to g almost everywhere, and $\liminf_k \|g_k\|_{L^p(U)} \leq \|g\|_{L^p(U)}$, then we have that it also converges in $\|\cdot\|_{L^p(U)}$.*

Proof. Consider

$$h_k(x) := 2^{p-1}(|g(x)|^p + |g_k(x)|^p) - |g(x) - g_k(x)|^p.$$

By triangle inequality, $0 \leq |g(x) - g_k(x)| \leq |g(x)| + |g_k(x)|$, and $(\cdot)^p$ is increasing and convex in $\mathbb{R}^+$, so

$$\begin{aligned} |g(x) - g_k(x)|^p &\leq (|g(x)| + |g_k(x)|)^p \\ &= 2^p\left(\frac{1}{2}|g(x)| + \frac{1}{2}|g_k(x)|\right)^p \\ &\leq 2^p\left(\frac{1}{2}|g(x)|^p + \frac{1}{2}|g_k(x)|^p\right) \\ &= 2^{p-1}(|g(x)|^p + |g_k(x)|^p). \end{aligned}$$

Thus $h_k \geq 0$ for any $k \geq 0$. In addition,

$$\begin{aligned} \liminf_k h_k(x) &= \lim_{k\to\infty} h_k(x) \\ &= 2^{p-1}\left(\lim_{k\to\infty} |g(x)|^p + \lim_{k\to\infty} |g_k(x)|^p\right) - \lim_{k\to\infty} |g(x) - g_k(x)|^p \\ &= 2^{p-1}(|g(x)|^p + |g(x)|^p) - 0 \\ &= 2^p|g(x)|^p, \end{aligned}$$

almost everywhere. By Fatou's Lemma, we have that

$$\begin{aligned} 2^p\|g\|^p_{L^p(U)} &= \int_U 2^p|g(x)|^p dx \\ &= \int_U \liminf_k h_k(x) dx \\ &\leq \liminf_k \int_U h_k(x) dx \\ &= \liminf_k \int_U 2^{p-1}(|g(x)|^p + |g_k(x)|^p) - |g(x) - g_k(x)|^p dx \\ &= 2^{p-1}\int_U |g(x)|^p dx + 2^{p-1}\liminf_k \int_U |g_k(x)|^p dx + \liminf_k \int_U (-|g(x) - g_k(x)|^p) dx \\ &= 2^{p-1}\|g\|^p_{L^p(U)} + 2^{p-1}\liminf_k \|g_k\|^p_{L^p(U)} + \liminf\left(-\|g - g_k\|^p_{L^p(U)}\right) \\ &\leq 2^{p-1}\|g\|^p_{L^p(U)} + 2^{p-1}\|g\|^p_{L^p(U)} + \liminf_k\left(-\|g - g_k\|^p_{L^p(U)}\right) \\ &= 2^p\|g\|^p_{L^p(U)} - \limsup_k \|g - g_k\|^p_{L^p(U)}. \end{aligned}$$

Subtracting $2^p\|g\|_{L^p(U)}$ from both sides, we get

$$\begin{aligned} 0 &\leq -\limsup_k \|g - g_k\|^p_{L^p(U)} \\ 0 &\geq \limsup_k \|g - g_k\|^p_{L^p(U)} \\ &\geq \liminf_k \|g - g_k\|^p_{L^p(U)} \\ &\geq 0. \end{aligned}$$

Thus $\lim_{k\to\infty} \|g - g_k\|^p_{L^p(U)} = \limsup_k \|g - g_k\|^p_{L^p(U)} = \liminf_k \|g - g_k\|^p_{L^p(U)} = 0.$ $\square$

Remark. The proof is inspired by the proof of Scheffé's Lemma, which is a special case of Riesz's Lemma.[19]

Theorem A.0.8. *Given any $a < b \in \mathbb{R}$ and any function $g \in L^p(a,b)$, we can find a sequence of piecewise constant (step) functions $\{g_k\}_{k=1}^{\infty} \in L^p(a,b)$ such that $\lim_{k\to\infty} g_k = g$ in $L^p(a,b)$. In addition, we can let them have uniform step size $h_k = \frac{b-a}{k}$, with the explicit form*

$$g_k(x) := \sum_{j=1}^{k} \chi_{[x_{j-1},x_j)}(x) \frac{1}{h_k} \int_{x_{j-1}}^{x_j} g(x')dx',$$

where $x_j := a + jh_k = a + \frac{(b-a)j}{k}$.

Proof. For any $x \in (a,b), k \geq 0$, we can find some $U_k := [x_j^{(k)}, x_{j+1}^{(k)}] \ni x$. Notice that

$$\forall k \geq 1, \begin{cases} U_k \subseteq B(x, \max(x - x_j^{(k)}, x_{j+1}^{(k)} - x)), \\ |U_k| = h_k \geq \frac{1}{2}|B(x,h_k)| = \frac{1}{2}\left|B(x, x_{j+1}^{(k)} - x_j^{(k)})\right| \geq \frac{1}{2}\left|B(x, \max(x - x_j^{(k)}, x_{j+1}^{(k)} - x))\right| \end{cases}.$$

Thus it satisfies the Lebesgue differentiation theorem, and we have that

$$\begin{aligned} \lim_{k\to\infty} g_k(x) &= \lim_{k\to\infty} \frac{1}{h_k} \int_{x_j^{(k)}}^{x_{j+1}^{(k)}} g(x')dx' \\ &= \lim_{U_k \to x} \frac{1}{|U_k|} \int_{U_k} g(x')dx' \\ &= g(x) \text{ a.e.} \end{aligned}$$

Thus $g_k(x) \to g(x)$ point-wise almost everywhere, and it suffices to show that $\liminf_k \|g_k\|_{L^p(a,b)} \leq$

$\|g\|_{L^p(a,b)}$, and then apply lemma A.0.7. Notice that for any k, we have that

$$
\begin{aligned}
\|g_k\|^p_{L^p(a,b)} &= \int_a^b \left| \sum_{j=1}^{k} \chi_{[x_{j-1},x_j)}(x) \frac{1}{h_k} \int_{x_{j-1}}^{x_j} g(x')dx' \right|^p dx \\
&= \sum_{j=1}^{k} \int_{x_{j-1}}^{x_j} \left| \frac{1}{h_k} \int_{x_{j-1}}^{x_j} g(x')dx' \right|^p dx \\
&= \sum_{j=1}^{k} (x_j - x_{j-1}) \left(\frac{1}{h_k}\right)^p \left| \int_{x_{j-1}}^{x_j} g(x')dx' \right|^p \\
&= \sum_{j=1}^{k} h_k \left(\frac{1}{h_k}\right)^p \left| \int_{x_{j-1}}^{x_j} g(x') \cdot 1 dx' \right|^p \\
&\leq \sum_{j=1}^{k} \left(\frac{1}{h_k}\right)^{p-1} \left(\int_{x_{j-1}}^{x_j} |g(x') \cdot 1| dx' \right)^p \\
&\leq \sum_{j=1}^{k} \left(\frac{1}{h_k}\right)^{p-1} \left(\left(\int_{x_{j-1}}^{x_j} |g(x')|^p dx' \right)^{1/p} \left(\int_{x_{j-1}}^{x_j} |1|^{\frac{p}{p-1}} dx' \right)^{\frac{p-1}{p}} \right)^p \qquad 2.1.8 \\
&= \sum_{j=1}^{k} \left(\frac{1}{h_k}\right)^{p-1} \left(\int_{x_{j-1}}^{x_j} |g(x')|^p dx' \right) (x_j - x_{j-1})^{p-1} \\
&= \sum_{j=1}^{k} \left(\frac{1}{h_k}\right)^{p-1} \left(\int_{x_{j-1}}^{x_j} |g(x')|^p dx' \right) h_k^{p-1} \\
&= \sum_{j=1}^{k} \int_{x_{j-1}}^{x_j} |g(x')|^p dx' \\
&= \int_a^b |g(x')|^p dx' \\
&= \|g\|^p_{L^p(a,b)}.
\end{aligned}
$$

Thus $\liminf \|g_k\|_{L^2(a,b)} \leq \|g\|_{L^2(a,b)}$, and by lemma A.0.7, we have that $\lim_{k\to\infty} g_k = g$. □

Remark. The first claim in theorem A.0.8 that step functions are dense in $L^p(U)$ is relatively well-known, and is usually proven by the regularity of the Lebesgue measure. However, the proof is not constructive, and the step sizes are not necessarily uniform. Here, we provide a version that achieves uniform steps, and give an explicit formulation of them.

Appendix B

Python Code for Control Through Reaction Term

Code for tools.py that implements some basic tools:

```
import numpy as np
import logging


logger = logging.getLogger("pde_simulator")
logger.setLevel(logging.DEBUG)

import numpy as np
from scipy.integrate import solve_ivp
from scipy.interpolate import interp1d

def compute_derivatives(y: np.ndarray, dx: float) -> tuple[np.ndarray,
↪ np.ndarray]:
    """
    Compute y_x and y_xx using central differences.

    Parameters:
    -----------
    y : ndarray
        Function values on grid
    dx : float
        Grid spacing

    Returns:
    --------
    y_x, y_xx : ndarray
        First and second derivatives
```

```
    """
    y_x = np.gradient(y, dx)
    y_xx = np.gradient(y_x, dx)
    return y_x, y_xx

def compute_L2_norm(y: np.ndarray, x: np.ndarray) -> float:
    """
    Compute L2 norm: ||y||_2 = sqrt(int_0^1 y^2 dx)
    """
    return float(np.sqrt(np.trapezoid(y**2, x)))

def compute_L2_inner_product(y1: np.ndarray, y2: np.ndarray, x:
↪   np.ndarray) -> float:
    """
    Compute L2 inner product: <y1, y2> = int_0^1 y1*y2 dx
    """
    return float(np.trapezoid(y1*y2, x))
```

Code for pde_simulator.py that implements the simulator of the pde:

```
import numpy as np
from scipy.integrate import solve_ivp
from scipy.integrate._ivp.ivp import OdeResult
from typing import Callable
import logging
import fipy as fp
import matplotlib.pyplot as plt

from tools import *


logger = logging.getLogger("pde_simulator")
logger.setLevel(logging.DEBUG)

def plot_evolution(sol, x: np.ndarray, ud_vals: np.ndarray, T: float,
↪   plot_times=None):
    """
    Plot the evolution of u(x,t) and compare with target ud.
    """

    # Determine times to plot
    if plot_times is None:
        plot_times = np.linspace(0, T, 10)

    # Create figure with subplots
```

```python
fig = plt.figure(figsize=(16, 10))

# Subplot 1: Snapshots at different times
ax1 = plt.subplot(2, 2, 1)
colors = plt.cm.viridis(np.linspace(0, 0.9, len(plot_times)))

for i, t in enumerate(plot_times):
    u_t = sol.sol(t)
    ax1.plot(x, u_t, color=colors[i], linewidth=2,
            label=f't={t:.5f}', alpha=0.7)

ax1.plot(x, ud_vals, 'r--', linewidth=3, label='Target ud', alpha=0.9)
ax1.set_xlabel('x', fontsize=12)
ax1.set_ylabel('u(x,t)', fontsize=12)
ax1.set_title('Evolution: u(x,t) → ud(x)', fontsize=14,
↪   fontweight='bold')
ax1.legend(fontsize=9, ncol=2)
ax1.grid(True, alpha=0.3)


# Subplot 2: Error over time
ax2 = plt.subplot(2, 2, 2)
t_error = np.linspace(0, T, 100)
errors = []

for t in t_error:
    u_t = sol.sol(t)
    error = compute_L2_norm(u_t - ud_vals, x)
    errors.append(error)

ax2.plot(t_error, errors, 'b-', linewidth=2.5)
ax2.set_xlabel('Time t', fontsize=12)
ax2.set_ylabel('||u(·,t) - ud||_2', fontsize=12)
ax2.set_title('L2 Error to Target over Time', fontsize=14,
↪   fontweight='bold')
ax2.grid(True, alpha=0.3)
ax2.axhline(y=0, color='r', linestyle='--', alpha=0.5)

# Subplot 3: Initial vs Final vs Target
ax3 = plt.subplot(2, 2, 3)
u_initial = sol.sol(0)
u_final = sol.sol(T)

ax3.plot(x, u_initial, 'b-', linewidth=2.5, label='Initial u0',
↪   alpha=0.8)
```

```
    ax3.plot(x, u_final, 'g-', linewidth=2.5, label=f'Final u(·,T)',
    ↪  alpha=0.8)
    ax3.plot(x, ud_vals, 'r--', linewidth=2.5, label='Target ud',
    ↪  alpha=0.8)
    ax3.set_xlabel('x', fontsize=12)
    ax3.set_ylabel('u(x)', fontsize=12)
    ax3.set_title('Initial → Final vs Target', fontsize=14,
    ↪  fontweight='bold')
    ax3.legend(fontsize=11)
    ax3.grid(True, alpha=0.3)


    # Subplot 4: Heatmap
    ax4 = plt.subplot(2, 2, 4)
    t_plot = np.linspace(0, T, 200)
    u_plot = np.array([sol.sol(t) for t in t_plot])

    im = ax4.contourf(x, t_plot, u_plot, levels=50, cmap='RdBu_r')
    ax4.set_xlabel('x', fontsize=12)
    ax4.set_ylabel('t', fontsize=12)
    ax4.set_title('Space-Time Evolution', fontsize=14, fontweight='bold')
    plt.colorbar(im, ax=ax4, label='u(x,t)')


    plt.tight_layout()
    plt.show()

    print(f"\nFinal L2 error: ||u(·,T) - ud||_2 = {errors[-1]:.6e}")



class PDESolver:
    """
    Simulates the problem:
    u_t = u_xx + a(x)u_x + b(x)u
    """

    def __init__(self, nx=100, l=0, r=1):
        """
        Initialize solver on domain [l, r].

        Parameters:
        -----------
        nx : int
            Number of spatial grid points
```

```
        l, r : float
            Left and right endpoints, default [0, 1]
        """
        self.l = l
        self.r = r
        self.x = np.linspace(l, r, nx)
        self.dx = self.x[1] - self.x[0]
        self.nx = nx

    def solve(self, u0: Callable | np.ndarray, a: Callable | np.ndarray,
    ↪   b: Callable | np.ndarray, T: float,
                method='BDF', rtol=1e-6, atol=1e-8):
        """
        Solve u_t = u_xx + a(x)u_x + b(x)u using Method of Lines with
        ↪   scipy.

        Parameters:
        -----------
        u0 : ndarray or callable
            Initial condition
        a : ndarray or callable
            Coefficient a(x)
        b : ndarray or callable
            Coefficient b(x)
        T : float
            Final time
        method : str
            Integration method ('BDF', 'RK45', 'Radau')
            BDF is best for stiff problems
        rtol, atol : float
            Relative and absolute tolerances

        Returns:
        --------
        sol : OdeSolution object
            Solution with sol.t (times) and sol.y (solution at each
            ↪   time)
        """
        # Convert inputs to arrays
        if callable(u0):
            u0_vals = u0(self.x)
        else:
            u0_vals = np.asarray(u0)
            if len(u0_vals) != self.nx:
                x_temp = np.linspace(0, 1, len(u0_vals))
```

```
            u0_vals = np.interp(self.x, x_temp, u0_vals)

    if callable(b):
        b_vals = b(self.x)
    else:
        b_vals = np.asarray(b)
        if len(b_vals) != self.nx:
            x_temp = np.linspace(0, 1, len(b_vals))
            b_vals = np.interp(self.x, x_temp, b_vals)

    if callable(a):
        a_vals = a(self.x)
    else:
        a_vals = np.asarray(a)
        if len(a_vals) != self.nx:
            x_temp = np.linspace(0, 1, len(a_vals))
            a_vals = np.interp(self.x, x_temp, a_vals)

    # Define RHS function for Method of Lines
    def rhs(t, u):
        """Compute du/dt = u_xx + a*u_x + b*u with zero Dirichlet
         ↪  BC"""
        dudt = np.zeros_like(u)

        # If the reaction term b(x) dominates, use central
         ↪  differences
        if np.max(np.abs(b_vals)) > np.max(np.abs(a_vals)):

            # Interior points: central differences
            for i in range(1, self.nx - 1):
                u_xx = (u[i+1] - 2*u[i] + u[i-1]) / self.dx**2
                u_x = (u[i+1] - u[i-1]) / (2*self.dx)
                dudt[i] = u_xx + a_vals[i]*u_x + b_vals[i]*u[i]

        # If the advection term a(x) dominates, use upwind scheme
        else:
            for i in range(1, self.nx - 1):
                u_xx = (u[i+1] - 2*u[i] + u[i-1]) / self.dx**2
                # Upwind scheme for u_x based on sign of a
                if a_vals[i] >= 0:
                    u_x = (u[i] - u[i-1]) / self.dx  # Backward
                     ↪  difference
                else:
                    u_x = (u[i+1] - u[i]) / self.dx  # Forward
                     ↪  difference
```

```
                dudt[i] = u_xx + a_vals[i]*u_x + b_vals[i]*u[i]

        # Boundary conditions: zero Dirichlet (u = 0 at boundaries)
        dudt[0] = 0
        dudt[-1] = 0

        return dudt

    # Solve using scipy's solve_ivp
    sol = solve_ivp(
        rhs,
        [0, T],
        u0_vals,
        method=method,
        rtol=rtol,
        atol=atol,
        dense_output=True
    )

    return sol

def solve_with_fipy(self, u0, a, b, T, nt=100):
    """
    Solve u_t = u_xx + a(x)u_x + b(x)u with FiPy
    Returns an OdeResult object compatible with solve_ivp
    """
    # Create mesh
    mesh = fp.Grid1D(nx=self.nx, Lx=1.0)
    x = mesh.cellCenters[0].value

    # Create variable
    u = fp.CellVariable(mesh=mesh, name="u")

    # Set initial condition
    if callable(u0):
        u.setValue(u0(x))
    else:
        if len(u0) != len(x):
            u.setValue(np.interp(x, self.x, u0))
        else:
            u.setValue(u0)

    # Get coefficients - must match mesh size
    if callable(a):
```

```
    a_vals = a(x)
else:
    if len(a) != len(x):
        a_vals = np.interp(x, self.x, a)
    else:
        a_vals = a

if callable(b):
    b_vals = b(x)
else:
    if len(b) != len(x):
        b_vals = np.interp(x, self.x, b)
    else:
        b_vals = b

# Create CellVariable for b coefficient
b_cell = fp.CellVariable(mesh=mesh, value=b_vals)

# Create FaceVariable for convection coefficient
a_face = fp.FaceVariable(mesh=mesh, value=a_vals)

# Define PDE: u_t = u_xx + a*u_x + b*u
eq = (fp.TransientTerm() ==
    fp.DiffusionTerm(coeff=1.0) +
    fp.ConvectionTerm(coeff=a_face) +
    fp.ImplicitSourceTerm(coeff=b_cell))

# Apply boundary conditions
u.constrain(0, mesh.facesLeft)
u.constrain(0, mesh.facesRight)

# Time stepping
dt = T / nt
times = [0]
solutions = [u.value.copy()]

for step in range(nt):
    eq.solve(var=u, dt=dt)
    times.append((step + 1) * dt)
    solutions.append(u.value.copy())

# Convert to numpy arrays
t = np.array(times)
y = np.array(solutions).T  # Transpose to match solve_ivp format
↪  (nx, nt)
```

```
        # Create OdeResult object (compatible with solve_ivp)
        sol = OdeResult(
            t=t,
            y=y,
            sol=None,  # No dense output for now
            t_events=None,
            y_events=None,
            nfev=nt,
            njev=0,
            nlu=nt,
            status=0,
            message='Integration successful.',
            success=True
        )

        # Add dense_output interpolation function
        from scipy.interpolate import interp1d
        sol.sol = interp1d(t, y, kind='cubic', axis=1,
                        fill_value='extrapolate', assume_sorted=True)

        return sol
```

Code for reaction_contorl.py that implements the two control methods through reaction term:

```
import matplotlib.pyplot as plt
from pde_simulator import *


logger = logging.getLogger("pde_simulator")
logger.setLevel(logging.DEBUG)

import numpy as np

class ReactionControlSolver(PDESolver):
    """
    Solves the control problem: steer u0 to ud via the PDE
    u_t = u_xx + a(x)u
    """

    def compute_alpha(self, ud: Callable | np.ndarray, u0: Callable |
    ↪ np.ndarray, T: float) -> tuple[np.ndarray, np.ndarray, dict]:
        """
        Compute a(x) = -(ud)_xx/ud + ln(||ud||_2 / int_0^1 u0*omega dx) /
        ↪ T
```

```
where omega = ud / ||ud||_2

Parameters:
-----------
ud : ndarray or callable
    Target function
u0 : ndarray or callable
    Initial condition
T : float
    Final time

Returns:
--------
alpha : ndarray
    Coefficient alpha(x)
omega : ndarray
    Normalized target
info : dict
    Diagnostic information
"""
# Convert to arrays if needed
if callable(ud):
    ud_vals = ud(self.x)
else:
    ud_vals = np.asarray(ud)
    if len(ud_vals) != self.nx:
        x_temp = np.linspace(0, 1, len(ud_vals))
        ud_vals = np.interp(self.x, x_temp, ud_vals)

if callable(u0):
    u0_vals = u0(self.x)
else:
    u0_vals = np.asarray(u0)
    if len(u0_vals) != self.nx:
        x_temp = np.linspace(0, 1, len(u0_vals))
        u0_vals = np.interp(self.x, x_temp, u0_vals)

# Compute L2 norm of ud
norm_ud = compute_L2_norm(ud_vals, self.x)

# Compute omega = ud / ||ud||_2
omega = ud_vals / norm_ud

# Compute integral int_0^1 u0 * omega dx
```

```python
        integral_u0_omega = compute_L2_inner_product(u0_vals, omega,
         ↪  self.x)

        # Compute derivatives of ud
        ud_x, ud_xx = compute_derivatives(ud_vals, self.dx)

        # Compute -ud_xx/ud (handle division by zero)
        with np.errstate(divide='ignore', invalid='ignore'):
            neg_ud_xx_over_ud = -ud_xx / ud_vals
            neg_ud_xx_over_ud = np.nan_to_num(neg_ud_xx_over_ud, nan=0.0,
                                               posinf=0.0, neginf=0.0)

        # Compute logarithmic term
        if integral_u0_omega <= 0:
            print(f"Warning: integral u0*omega = {integral_u0_omega:.6f}
             ↪  <= 0")
            print("This may cause issues. Setting log term to 0.")
            log_term = 0.0
        else:
            log_term = np.log(norm_ud / integral_u0_omega) / T

        # Compute a(x)
        alpha = neg_ud_xx_over_ud + log_term

        # Store diagnostic info
        info = {
            'norm_ud': norm_ud,
            'integral_u0_omega': integral_u0_omega,
            'log_term': log_term,
            'ud_vals': ud_vals,
            'omega': omega
        }

        return alpha, omega, info

    def compute_v(self, ud: Callable | np.ndarray, u0: Callable |
     ↪  np.ndarray, T: float) -> tuple[np.ndarray, dict]:
        """
        Compute v(x) = ln(ud/u0) / T

        Parameters:
        -----------
        ud : ndarray or callable
            Target function
        u0 : ndarray or callable
```

```
        Initial condition
    T : float
        Final time

    Returns:
    --------
    v : ndarray
        Coefficient v(x)
    info : dict
        Diagnostic information
    """
    # Convert to arrays if needed
    if callable(ud):
        ud_vals = ud(self.x)
    else:
        ud_vals = np.asarray(ud)
        if len(ud_vals) != self.nx:
            x_temp = np.linspace(0, 1, len(ud_vals))
            ud_vals = np.interp(self.x, x_temp, ud_vals)

    if callable(u0):
        u0_vals = u0(self.x)
    else:
        u0_vals = np.asarray(u0)
        if len(u0_vals) != self.nx:
            x_temp = np.linspace(0, 1, len(u0_vals))
            u0_vals = np.interp(self.x, x_temp, u0_vals)

    # Compute v(x) = ln(ud/u0) / T, adding a small positive offset to
    ↪ avoid log(0) or division by 0
    # Avoid first and last values to avoid 0 boundary conditions
    v = np.zeros(self.nx)
    v[1:-1] = np.log(ud_vals[1:-1] / u0_vals[1:-1]) / T

    # Approximate boundary values by linear interpolation
    v[0] = v[1] - (v[2] - v[1]) * (self.x[0] - self.x[1]) / (self.x[2]
    ↪ - self.x[1])
    v[-1] = v[-2] + (v[-2] - v[-3]) * (self.x[-1] - self.x[-2]) /
    ↪ (self.x[-2] - self.x[-3])

    # Handle any remaining NaN or Inf values
    v = np.nan_to_num(v, nan=-1e-10, posinf=0.0, neginf=-1e-10)

    max_v = np.max(v[1:-1])
    if max_v >= 0:
```

```python
            print("Error: Method 'compute_v' requires v(x) < 0, but v(x) =
             ↪  ", max_v, " >= 0")

        print(f"Max value of v(x) = {max_v:.6f}")
        # Store diagnostic info
        info = {
            'ud_vals': ud_vals,
            'u0_vals': u0_vals,
            'max_v': max_v
        }

        return v, info


def run_reaction_control_problem(u0: Callable | np.ndarray, ud: Callable |
↪  np.ndarray,
                                 T: float, nx=100, plot_times=None,
                                 ↪  control_method='1', method='BDF'):
    """
    Complete workflow: compute b(x), solve PDE, and visualize.

    Parameters:
    -----------
    u0 : callable or array
        Initial condition
    ud : callable or array
        Target condition
    T : float
        Final time
    nx : int
        Number of spatial grid points
    plot_times : array-like or None
        Specific times to plot. If None, uses 10 evenly spaced times.
    """
    # Initialize solver
    solver = ReactionControlSolver(nx=nx)
    x = solver.x

    if control_method == '1':
        # Compute coefficient b(x)
        print("Computing coefficient b(x) from target ud...")
        b, __, info = solver.compute_alpha(ud, u0, T)

        # Print diagnostic info
```

```
        print(f"\nDiagnostic Information:")
        print(f"  ||ud||_2 = {info['norm_ud']:.6f}")
        print(f"  int_0^1 u0*omega dx = {info['integral_u0_omega']:.6f}")
        print(f"  ln(||ud||_2 / int_0^1 u0*omega dx) / T =
         ↪  {info['log_term']:.6f}")
        print(f"  max|b(x)| = {np.max(np.abs(b)):.6f}")
    elif control_method == '2':
        # Compute coefficient b(x)
        print("Computing coefficient b(x) from target ud...")
        b, info = solver.compute_v(ud, u0, T)

        # Print diagnostic info
        print(f"\nDiagnostic Information:")
        print(f"  max b(x) = {info['max_v']:.6f}")
        print(f"  max|b(x)| = {np.max(np.abs(b)):.6f}")
    else:
        raise ValueError(f"Invalid method: {method}")

    # Set a(x) = 0
    a = np.zeros_like(x)

    # Solve PDE
    print(f"\nSolving PDE from t=0 to t={T}...")
    sol = solver.solve(u0, a, b, T, method=method)

    print(f"  Integration successful!")
    print(f"  Number of time steps taken: {len(sol.t)}")
    print(f"  Final time reached: {sol.t[-1]:.6f}")

    # Prepare target values
    if callable(ud):
        ud_vals = ud(x)
    else:
        ud_vals = info['ud_vals']

    # Create visualizations
    plot_evolution(sol, x, ud_vals, T, plot_times)

    # Plot coefficient b(x)
    plt.plot(x, b, 'b-', linewidth=2.5, label='b(x)')
    plt.axhline(y=0, color='k', linestyle='--', alpha=0.3)
    plt.xlabel('x', fontsize=12)
    plt.title('Coefficient b(x)', fontsize=13, fontweight='bold')
    plt.legend(fontsize=9, ncol=2)
    plt.grid(True, alpha=0.3)
```

```
    return solver, sol, b, info

# ==================== EXAMPLE USAGE ====================

if __name__ == "__main__":
    print("="*60)
    print("PDE Control Problem: Steering u0 → ud")
    print("="*60)

    # Example 1:
    print("\n" + "="*60)
    print("Example 1: Parabola initial → Sine target")
    print("="*60)

    u0 = lambda x: x*(1-x)  # Parabola
    ud = lambda x: np.sin(np.pi*x)
    T = 1.0

    solver, sol, a, info = run_reaction_control_problem(u0, ud, T, nx=150)

    # Example 2: Different target
    print("\n" + "="*60)
    print("Example 2: Asymmetric initial → Sine target")
    print("="*60)

    u0 = lambda x: 8*x*x*(1-x)  # Asymmetric
    ud = lambda x: np.sin(np.pi*x)
    T = 2

    solver, sol, a, info = run_reaction_control_problem(u0, ud, T, nx=150)

    print("\n" + "="*60)
    print("Complete! Modify u0, ud, and T for your specific problem.")
    print("="*60)

    # Example 3: Different method
    print("\n" + "="*60)
    print("Example 3: Asymmetric initial → Sine target, method 2")
    print("="*60)

    u0 = lambda x: 8*x*x*(1-x)
    ud = lambda x: np.sin(np.pi*x)

    # T = 1
```

```
# solver, sol, a, info = run_reaction_control_problem(u0, ud, T,
↪   nx=150, method='2')

T = 0.01 # want very small T for this method

solver, sol, a, info = run_reaction_control_problem(u0, ud, T, nx=150,
↪   control_method='2')

T = 0.0001 # want very small T for this method

solver, sol, a, info = run_reaction_control_problem(u0, ud, T, nx=150,
↪   control_method='2')
```

Glossary

advection term The term with first-order spatial derivative in a second-order parabolic partial differential equation 2

control The part of a system which has the freedom to be designed so that the system behaves in a desired way 1, 2

dense A property of a subset in a space, where any element in the space can be approximated by a element in the set with arbitrarily small difference 6, 22

equivalent function class A set of functions that have the same value almost everywhere with respect to the Lebesgue measure ix, 3, 9, 38, 80

final state The final state that the system ends with, at time T 2

Hilbert Space A complete inner product space ix, 7

initial state The initial state that the system starts from, at time 0 1, 2

inner product space A vector space that is equipped with an inner product 6, 80

ket-bra notation A notation for elements and linear functionals 7, 13, 26

Lebesgue integrable Measurable functions with finite integral on a given domain, with respect to the Lebesgue measure 80

normed vector space A vector space that is equipped with a norm 6

reaction term The term with no derivative in a second-order parabolic partial differential equation 2

space of square-integrable functions The collection of equivalent function classes that are square Lebesgue integrable in the given domain 2, 3, 9, 22, 38

target state The target state that we wish to send the system to 1, 2

weak derivative A representative of an equivalent function class that satisfies intergration by parts formula for all test functions 3, 10, 38