Coefficient Optimal Control Problems for Elliptic PDE

by

Peter Umberto Vinella

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Mathematics
in the
Graduate Division
of the
University of California, Berkeley

Committee in Charge:
Professor Lawrence Craig Evans, Chair
Professor Alexandre Chorin
Professor Daniel Tataru

Spring 2021

Abstract

Coefficient Optimal Control Problems for Elliptic PDE

by

Peter U. Vinella

Doctor of Philosophy in Mathematics
University of California, Berkeley
Professor Lawrence Craig Evans, Chair

In this paper, we consider a class of optimal control problems known as *coefficient control problems.* Such problems are constrained by uniformly elliptic PDE in which the controls appear as some of the coefficients in the differential operator. We begin with a brief review of standard optimal control theory and show that it does not apply to control coefficient problems generally by means of two counterexamples. We then present existence results for the solution to such problems under two different assumptions regarding the controls: Lipschitz continuity with a bounded Lipschitz constant and the case in which the admissible set of controls is closed under H-convergence. We then present a new maximum principle for the latter class of problems which we subsequently use to characterize the nature and behavior of optimal solutions.

Dedication

To my wife, Jeanette, who has supported me for the last twenty years, to my daughter, Avalon, who encouraged me to go back to graduate school, and to my mother, Mary, who waited quite a long time for this.

Acknowledgments

I would like to thank Mitchell Berns of the law firm, Fox Rothschild, and the Federal Home Loan Bank of New York who gave me the opportunity to investigate new ways to estimate the coefficients of interest rate processes. I also would like to thank my fellow graduate students, especially Woo-Hyun Cook, Peyam Tabrizian, and Chris Miller, who welcomed me with open arms and gave me a lot of support. Additionally, I would like to thank Professor George Bergman who encouraged me to come back to Cal and Daniel Tataru who helped guide me through the unfathomable bureaucracy of the Graduate Division to get me readmitted after thirty years. Further, I would like to thank my professors from my first tenure at Cal, especially Professors Alexandre Chorin and Ole Hald as well as Professors Xin Guo and James Pitman who helped make my second tenure just as fun and fulfilling. Lastly, I would like to thank my advisor Craig Evans who made this all possible. Craig helped me to see and appreciate mathematics in new ways and was ever patient with my stumbling along the way. Craig also introduced to me to number of great books and music outside of mathematics and became a true friend. Who else but Craig would make reading Phillip K. Dick and listening to Warren Zevon part of the degree requirement. I will truly miss working as closely with him as I have over the past few years.

Preface

The motivation behind this thesis was a problem that I was given arising from the turmoil in the financial markets following the collapse of Lehman Brothers in September 2008. My task was no price hundreds of millions of dollars of interest swaps in connection with a lawsuit brought against Lehman by one of its clients. Such an exercise would typically involve discounting the various cash flows using some benchmark interest rate that was modeled using an Itô diffusion. This, in turn, requires calibrating the coefficients the SDE on each valuation date to reflect prevailing market conditions. However, interest rates are not directly observable in the market and the standard calibration methods rely on heuristics rather than empirical data. Given the market volatility at the time, these models were found wanting for this particular exercise.

While interest rates themselves are not directly observable in the market, prices of various classes financial instruments that are derived from them are directly observable, most notably zero-coupon bonds issued by the U.S. Treasury Department. Moreover, the price of such a financial instrument can be shown to be the solution to a PDE whose coefficients are the same as those of underlying interest rate process (*i.e.* the well-known Black-Scholes PDE). Hence, the problem of estimating the coefficients of the stochastic interest rate process can be transformed into estimating the coefficients of a PDE. Additionally, since these prices are observable in market, we can formulate an estimation scheme with the goal of minimizing the difference between computed and observed prices (in other words, a norm-minimization problem). This suggests that we can pose our calibration problem as one in which the coefficients of the PDE act as *controls* governing the evolution of the price of zero-coupon bonds whose values are chosen to achieve some optimal result (*i.e.* minimizing the estimation error between computed and observed prices) – in other words, an optimal control problem, albeit one which is infinite dimensional with PDE-constraints.

As is the the case with this calibration problem, optimal control problems generally involve "controlling" the evolution of the state of a given system in order to achieve some prescribed objective. Often, this evolution is described by a differential equation which is subject to various constraints imposed on permissible states and controls such as in the case of classical calculus of variation problems. The theory establishing the existence of optimal solution to such problems along with necessary conditions for optimality is well understood in a finite dimensional setting. Further, this theory has been extended to infinite dimensional problems generally and, in particular, those which are constrained by a PDE that is linear with respect to the controls. However, there are still many open problems regarding the existence and characterization of optimal solutions in the case of nonlinear problems generally.

In this thesis, we consider two special cases of such nonlinear problems. First, in the case of estimating the coefficients of a SDE such as the calibration problem above, we can rely on the fact that these coefficients must be Lipschitz continuous with bounded Lipschitz constants. This allows

us to pass the limits without having to rely on linearity. Consequently, under this assumption, we can demonstrate both the existence of optimal controls as well as characterize those controls using standard methods. In the second case, we show that solutions exist and be characterized by a maximization principle for a broader class of problems in which the diffusion coefficient is closed under H-convergence. However, it should be noted that these results are only small step toward resolving the problem estimating unobservable coefficients of Itô diffusions.

This thesis is organized as follows.

In Chapters 1 and 2, we introduce the basic optimal control problem and we review some key results from general optimal control theory. In particular, review standard Lagrange optimization methods in infinite-dimensions which can be applied to special classes of optimal control problems.

In Chapter 3, we introduced optimal control problems constrained by an elliptic PDE. We begin with basic results from PDE theory which establish the existence and uniqueness of the *weak form* of feasible solutions. We then review a special class of such problems which are linear with respect to the controls. In this case, we can apply the general theory developed in Part 2 following the pioneering work J.L. Lions in the 1960's.

In Chapter 4, we explore optimal control problems constrained by an elliptic PDE that are nonlinear with respect to the controls. Here, we introduce two counterexamples to show that solutions to such problems do not exist generally. In light of these counterexamples, we then turn our focus to a special class of nonlinear optimal control problems whose controls appear as coefficients of the PDE constraint. We begin with establishing the existence of solutions to such problems beginning with a discussion of *H-convergence*. We also show that the solutions to such problems exist if the controls are assumed to be Lipschitz continuous with the same bounded Lipschitz constant. Following this, we introduce a maximum principal for such problem which essentially replaces the variational inequality used to characterize solutions. Finally, we analyze the nature and behavior of optimal solutions under a variety of simplifying assumptions.

In developing this material, I generally follow J.L. Lions [**60**], F. Tröltzsch [**84**], and M.Hinze, *et al.* [**40**] in laying out the basic optimal control theory as it applies to PDE-constrained problems. Additionally, I have generally followed Murat and Tartar in developing the concept of H-convergence as it applies to PDE-constrained optimal control problems. Further, I have generally relied on Evans [**27**] for much of the material regarding PDE theory. Other authors as well as references to specific results are cited throughout the text. Lastly, well-known results are generally presented with only a reference to an authoritative source rather than a proof itself. Results presented with a proof are generally mine, that of my of advisor, the proof is original, or some combination of the above unless noted otherwise.

Contents

CHAPTER 1

INTRODUCTION

1.1. The basic optimal control problem

In the abstract, an optimal control problem involves finding a ***state-control pair*** (u^0, α^0) contained in a given ***admissible set*** $\mathcal{A}$ which minimizes a given ***objective*** (or ***cost***) ***functional*** $I : \mathcal{A} \to \mathbb{R}$ such that

$$I\left[u^0(\alpha), \alpha^0\right] \leq I[u, \alpha] \quad \text{for all } (u, \alpha) \in \mathcal{A},$$

where u represents the ***state*** of a given system while α represents the ***control*** (***policy***).

NOTATION. Throughout this paper, $\mathcal{A}$ denotes the admissible set and the superscript "0" indicates an optimal solution. □

An additional term is often included in the objective functional to account for the cost (or penalty) of using a specific control. As such, the objective functional often takes the form

$$I[u, \alpha] = J[u, \alpha] + K[\alpha],$$

where K represents the cost of using a particular control α. Additionally, we generally consider problems that are subject to equality and inequality constraints on the choice of states and/or controls.

Given this, we formally state such problems as follows:

Problem A. (The general optimal control problem)

$$\begin{cases} \min\limits_{(u,\alpha)\in\mathcal{A}} & I[u, \alpha], \\ \text{subject to} & e(u, \alpha) = 0 \\ & c(u, \alpha) \in \mathcal{K} \end{cases}$$

for a given admissible set $\mathcal{A} \subset X \times Y$, where X and Y are given real Banach spaces known as the ***state space*** and the ***control space***, respectively. The ***control operators*** $e : \mathcal{A} \to Z$ and $c : \mathcal{A} \to W$ are also given, where Z and W are given real Banach spaces known as the ***value spaces***. In particular, the operator identity

$$e(u, \alpha) = 0$$

represents ***equality constraints*** and is called the ***state equation***. Similarly the operator c represents ***inequality constraints*** which are defined as

$$c(u, \alpha) \in \mathcal{K},$$

where $\mathcal{K}$ is a convex cone in W.

A state-control pair $(u, \alpha) \in \mathcal{A}$ is said to be ***feasible*** if it satisfies the equality and inequality constraints,

$$e(u, \alpha) = 0 \text{ and } c(u, \alpha) \in \mathcal{K},$$

where Z^* is the dual of the value space Z. The ***feasible set*** is the collection of feasible solutions and is denoted as $\mathcal{F}$.

For the purposes of this paper, we only consider problems which have equality constraints. However, the results presented herein can be extended to those problems which also have inequality constraints using an appropriate extension of the *Karush-Kuhn-Tucker conditions* from standard optimization theory (*cf.* [**84**, ch. 6]). Additionally, we only consider problems in which any state $u \in X$ is admissible baring any explicit constraints set forth in the problem statement. Lastly, we ignore any penalty on the choice of control as this only constrains the regularity of any solution and does impact its existence. Consequently, our canonical problem under consideration in this paper is of the form:

Problem B. (The general equality constrained optimal control problem)

$$\begin{cases} \min\limits_{u \in X, \alpha \in \mathcal{A}} & I[u, \alpha], \\ \text{subject to} & e(u, \alpha) = 0, \end{cases}$$

where X is the given state space, Y is the given control space such that $\mathcal{A} \subset Y$, and e represents the equality constraints.

NOTATION. Unless otherwise noted, we follow PDE convention that the state of a given system is denoted by the lower case letters u, v, and w while controls are denoted by the lower case Greek letters α and β. □

1.2. The reduced form

Faced with optimal control problems such as Problem B, in many cases, it possible to express the state u as an explicit function of the control α such that the mapping $\alpha \mapsto u(\alpha)$ is unique. If such a mapping exists, we say that it is a ***control-to-state operator*** and we denote it as $S : \mathcal{A} \to X$. Additionally, we denote the state as u^α to explicitly indicate its dependence on the control α, where

$$u^\alpha := S(\alpha).$$

REMARK. In general, a control-to-state operator may need not be onto and it might not be possible to attain a particular state regardless of the choice of controls. Consequently, we say that the state u is ***attainable*** if there exists a control $\alpha \in \mathcal{A}$ such that $u = S(\alpha)$. □

Given the existence of an appropriate control-to-state operator, we can express optimal control problems of the form Problem B in ***reduced form*** which we define as:

Problem C. (The general optimal control problem in reduced form)

$$\begin{cases} \min\limits_{\alpha \in \mathcal{A}} & \hat{I}[\alpha], \\ \text{subject to} & \hat{e}(\alpha) = 0, \end{cases}$$

where $\mathcal{A} \subset Y$ is the admissible set of controls, $\hat{I} : \mathcal{A} \to \mathbb{R}$ is the ***reduced objective functional*** defined as

$$\hat{I}[\alpha] := I[S(\alpha), \alpha] = I[u^\alpha, \alpha] \quad (\alpha \in \mathcal{A}),$$

and $\hat{e} : \mathcal{A} \to Z$ is the ***reduced constraint operator***, defined as

$$\hat{e}(\alpha) := e(S(\alpha), \alpha) = e(u^\alpha, \alpha) \quad (\alpha \in \mathcal{A}).$$

Additionally, we say the collection of all feasible solutions of such problems, $\hat{\mathcal{F}}$, defined as

$$\hat{\mathcal{F}} = \{\alpha \in \mathcal{A} \mid \hat{e}(\alpha) = 0\} = \{\alpha \in \mathcal{A} \mid e(S(\alpha), \alpha) = 0\},$$

is the ***reduced feasible set***.

NOTATION. For the remainder of the paper, we will use the same notation for full and reduced form of the objective functional and constraint operator unless otherwise noted. □

CHAPTER 2

REVIEW OF OPTIMIZATION AND OPTIMAL CONTROL THEORY

In this chapter we review some key results from standard optimization theory and its application to equality constrained optimal control problems in infinite dimensions. Many of the classical results are provided without proof which can be found in the cited references.

In this chapter, assume that X, Y, and Z are given real Banach spaces with dual spaces X^*, Y^*, and Z^*, respectively (*i.e.* the space of bounded linear functionals on each of those spaces).

NOTATION. Given a Banach space W and its dual space W^*, we denote the dual pairing of $w \in W$ and $z \in W^*$ as

$$\langle z, w \rangle_{W^*,W} = z(w) \in \mathbb{R}.$$

If it is obvious from the context, we often drop the subscript and simply denote the dual pairing as

$$\langle z, w \rangle$$

Additionally, if the mapping A is linear, we often use the "multiplication" notation and for $w \in W$,we write

$$Aw := A(w).$$

Lastly, we use x, y, and z to represent elements of general Banach spaces and u,v, and w to represent elements of Banach spaces of functions to emphasize the fact that they are functions. □

2.1. Existence of optimal solutions

We now identify conditions under which solutions to optimal control problems that can be expressed in reduced form exist.

2.1.1. Under the weak topology. Recall that closed, bounded sets in infinite dimensional spaces are not compact generally. Therefore, we turn to alternative topologies on general Banach spaces from which we can derive the necessary compactness properties. Here, we generally follow Evans [**25**] and [**84**, ch.2.4].

DEFINITION. For a normed linear space W, we say the topology induced by its norm is the ***strong*** (***norm***) ***topology***. Additionally, we say the topology on W induced by W^*, referred to as the ***weak topology***, is the weakest topology on W under which every element of W^* remains continuous and is denoted as $\sigma(W, W^*)$.

We now review some properties of the weak topology.

2.1.1.1. *Weak convergence.*

DEFINITION. Suppose that $\{x_j\}_{j\in\mathbb{N}} \subset X$.

(i) We say that $\{x_j\}_{j\in\mathbb{N}}$ ***converges strongly*** *to* $x \in X$ if it converges in the norm-topology, *i.e.*

$$\lim_{j\to\infty} \|x_j - x\| = 0.$$

(ii) We say that $\{x_j\}_{j\in\mathbb{N}}$ ***converges weakly*** *to* $x \in X$, denoted as

$$x_j \rightharpoonup x,$$

if for all $x^* \in X^*$

$$\langle x^*, x_j\rangle \to \langle x^*, x\rangle \text{ as } j \to \infty.$$

REMARK. If X is a Hilbert space with the inner product $(\cdot\,,\cdot)$, then by the Riesz Representation Theorem, a sequence $\{x_j\}_{j\in\mathbb{N}} \subset X$ converges weakly to $x \in X$ provided

$$(y, x_j) \to (y, x) \text{ as } j \to \infty$$

for all $y \in X$. □

THEOREM 2.1.1. *Assume that* $\{x_j\}_{j\in\mathbb{N}} \subset X$ *and* $x \in X$.

(i) *If* $x_j \to x$, *then* $x_j \rightharpoonup x$.
(ii) *If* $x_j \rightharpoonup x$, *then the* x_j *are bounded in* X *and* $\|x\| \le \liminf_{j\to\infty} \|x_j\|$.

PROOF. *See* Brezis [**15**, p.58]. □

2.1.1.2. *Weak continuity.* Here, we offer a number of continuity statements under the weak topology. Equivalent statements exist under the weak* topology unless otherwise noted.

DEFINITION. A mapping f from a topological space (Ω, τ_Ω) to another topological space (Ψ, τ_Ψ) is said to be ***continuous at a point*** $x \in \Omega$ if and only if the preimage of every open neighborhood of $f(x)$ is an open neighborhood of x. Further, f is said to be ***continuous*** on $V \in \Omega$ if it is continuous at each $x \in V$. In particular, we say that a function f from one normed linear space to another is (***strongly***) ***continuous*** if it is continuous under the norm topology and ***weakly continuous*** if it is continuous under the weak topology.

THEOREM 2.1.2. *If* $f : X \to Y$ *is strongly continuous at* $x \in X$, *then* f *is weakly continuous at* x.

PROOF. *See* Aliprantis [**3**, p.233]. □

DEFINITION. Suppose $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \to x \in X$. We say the function $f : X \to Y$ is ***sequentially continuous at*** x if $f(x_j) \to f(x) \in Y$. Similarly, we say that f is ***sequentially weakly continuous at*** x if $x_j \rightharpoonup x \in X$ and $f(x_j) \rightharpoonup f(x) \in Y$.

THEOREM 2.1.3. *If* $f : X \to Y$ *is* (*weakly*) *continuous, then is it sequentially* (*weakly*) *continuous. Further, if* f *is sequentially continuous, then it is continuous.*

PROOF. *See* Rudin [**76**, p.395]. □

REMARK. The second statement follows from the fact that a Banach space is metrizable under the norm topology. This does not necessarily hold under the weak topology. □

Definitions.

(i) The function $f : X \to \mathbb{R}$ is said to be (***strongly***) ***lower semicontinuous*** if for every sequence $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \to x \in X$, we have

$$\liminf_{j\to\infty} f(x_j) \geq f(x).$$

(ii) f is said to be (***strongly***) ***upper semicontinuous*** if for every sequence $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \to x \in X$,

$$\limsup_{j\to\infty} f(x_j) \leq f(x).$$

(iii) f is said to be ***weakly lower semicontinuous*** if for every sequence $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \rightharpoonup x \in X$,

$$\liminf_{j\to\infty} f(x_j) \geq f(x).$$

(iv) f is said to be ***weakly upper semicontinuous*** if for every sequence $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \rightharpoonup x \in X$,

$$\limsup_{j\to\infty} f(x_j) \leq f(x).$$

The next result follows immediately from the definition above.

THEOREM 2.1.4. *The function $f : X \to \mathbb{R}$ is (weakly) continuous if and only if it is (weakly) lower and upper semicontinuous and*

$$f(x) = \liminf_{j\to\infty} f(x_j) = \limsup_{j\to\infty} f(x_j),$$

for every $\{x_j\}_{j\in\mathbb{N}} \subset X$ such that $x_j \to x \in X$ ($x_j \rightharpoonup x \in X$).

The next result emphasizes the key role that convexity plays in infinite dimensional optimization.

THEOREM 2.1.5. *Assuming $f : X \to \mathbb{R}$ is convex, then it is strongly lower semicontinuous if and only it is weakly lower semicontinuous.*

PROOF. *See* Clarke [**20**, p.52]. □

2.1.1.3. *Weak compactness.* We now offer a number of statements regarding compactness under the weak topology. Equivalent statements exist under the weak* topology unless otherwise noted.

Definitions. Let $V \subset X$.

(i) V is said to be ***weakly closed*** if its complement is open under the weak topology.
(ii) V is said to be ***weakly sequentially closed*** if for any sequence $\{x_j\}_{j\in\mathbb{N}} \subset V$ such that $x_j \rightharpoonup x$, we have $x \in V$.

THEOREM 2.1.6. *Let V be a nonempty subset of X. Consider the following statements*:

(i) *V is weakly closed.*

(ii) *V is weakly sequentially closed.*
(iii) *V is strongly sequentially closed.*
(iv) *V is strongly closed.*

Then (i)$\Rightarrow$ (ii)$\Rightarrow$ (iii)$\Leftrightarrow$ (iv). *Further, if V is convex, then these statements are equivalent.*

PROOF. *See* Peypouquet [**70**, p.12]. □

Definitions. Let $V \subset X$. We say

(i) V is ***weakly compact*** if it is compact under the weak topology.
(ii) V is ***weakly sequentially precompact*** (a.k.a. ***weakly sequentially relatively compact***) if for any sequence $\{x_j\}_{j\in\mathbb{N}} \subset V$, there exists a subsequence $\{x_{j_k}\}_{k\in\mathbb{N}} \subseteq \{x_j\}_{j\in\mathbb{N}}$ such that $x_{j_k} \rightharpoonup x \in \overline{V}$.
(iii) V is ***weakly sequentially compact*** if it is weakly sequentially closed and precompact (*i.e.* every sequence in V contains a subsequence which converges weakly to a point in V).

THEOREM 2.1.7 (Eberlein-Šmulian Theorem). *A subset of a normed linear space is weakly compact if and only if it is weakly sequential compact.*

PROOF. *See* Holmes [**43**, p.147]. □

The next result shows that a bounded, closed set in a reflexive Banach space is sequentially compact under the weak topology. It follows immediately from the Banach-Alaoglu Theorem which implies that a weakly closed unit ball in a reflexive Banach is weakly compact (and hence, weakly sequentially compact)..

THEOREM 2.1.8. (i) *Every bounded, weakly closed subset of a reflexive Banach space is weakly sequentially compact.*

(ii) *Every bounded sequence in a reflexive Banach space has a weakly convergent subsequence.*

PROOF. *See* Royden [**74**, p. 284] □

2.1.1.4. *Existence results.* The following result, known as James' Theorem, establishes an analog to the extreme value theorem under the weak topology.

THEOREM 2.1.9 (James' Theorem). *A nonempty, bounded, weakly closed subset V of a real Banach space X is weakly compact if and only if every continuous linear real-valued function f on X attains its extrema on V.*

PROOF. *See* Holmes [**43**, p.157]. □

The next result shows follows immediately from Theorems 2.1.6 and 2.1.8.

THEOREM 2.1.10. *Every bounded, closed, convex subset of a reflexive Banach space is weakly sequentially compact.*

PROOF. *See* Clarke [**20**, p.101]. □

Given these results, we state an alternative form of James' Theorem in the special case of reflexive Banach spaces.

THEOREM 2.1.11. *A nonempty, bounded, closed, convex subset V of a reflexive Banach space X is weakly sequentially compact if and only if every continuous linear real-valued function f attains its extrema on V.*

Given the properties of the following weak topology, we can express a more general theorem for the existence of optimal solutions which we use extensively in the sequel.

2.1.2. Under the norm topology. Although James' Theorem establishes conditions under which extrema exist for certain classes of optimization problems, it relies on the weak topology which is difficult from a computational perspective. We now state an existence result under the norm topology which uses convexity to provide the necessary compactness properties. We begin by considering unconstrained optimization problems of the following form:

Problem D. (The general unconstrained optimization problem)

$$\min_{u \in V} f(u),$$

where $f : V \to \mathbb{R}$ is lower semicontinuous and V is a nonempty, closed, bounded, and convex subset of a reflexive Banach space X.

We now present one of the key results of this chapter. The proof presented below is due to Tonelli and is often referred to as the ***direct method***.

THEOREM 2.1.12 (Infinite dimensional form of the extreme value theorem). *Assume we are given Problem D, where X is reflexive and the admissible set $V \subset X$ is nonempty, closed, bounded, and convex. If $f : V \to \mathbb{R}$ is lower semicontinuous and convex on V, then f achieves its minimum on V.*

PROOF. Assume that $f : V \to \mathbb{R}$ is lower semicontinuous and convex and define M as

$$M := \inf_{u \in V} f(u) \geq -\infty. \tag{2.1.1}$$

Under this assumption there is a minimizing sequence $\{u_j\}_{j \in \mathbb{N}} \subset V$ such that $f(u_j) \to M$ as $j \to \infty$. Also by assumption, $V \subset X$ is nonempty, closed, bounded, and convex and X is reflexive. Therefore V is weakly sequentially compact by Theorem 2.1.10. Thus there is a subsequence $\{u_{j_k}\}_{k \in \mathbb{N}} \subseteq \{u_j\}_{j \in \mathbb{N}} \subset V$ that converges weakly to some $u^0 \in V$. Since f is real-valued,

$$f\left(u^0\right) > -\infty.$$

Moreover, $\{u_{j_k}\}_{k \in \mathbb{N}}$ is also minimizing sequence of f and thus $f(u_{j_k}) \to M$ as $k \to \infty$. Since f is lower semicontinuity and convex, it weakly lower semicontinuous by Theorem 2.1.5. Consequently, we have

$$-\infty < f\left(u^0\right) \leq \liminf_{k \to \infty} f(u_{j_k}) = M.$$

Therefore M is finite and, f attains its minimum on V. □

REMARK. Recall that a function $g : X \to \mathbb{R} \cup \{\infty\}$, where X is a normed linear space X is ***coercive*** if

$$\lim_{\|x\| \to \infty} g(x) = \infty.$$

Under the assumption the objective functional f is coercive, we no longer need the requirement that the admissible set V is bounded to establish the existence of a minimizer (*cf.* [**20**, p.102]). □

We now turn our attention to equality constrained optimization problems of the form:

Problem E. (The equality constrained optimization problem)

$$\begin{cases} \min\limits_{u \in V} & f(u), \\ \text{subject to} & e(u) = 0, \end{cases}$$

where $f : V \to \mathbb{R}$ is lower semicontinuous, V is a nonempty, closed, bounded, and convex subset of reflexive Banach space X, and the constraint operator $e : V \to Z$ is given.

By Theorem 2.1.12 such problems have solutions provided that there is at least one feasible solution. This is capture in the statement below.

THEOREM 2.1.13. *Assume that we are given an unconstrained optimization problem of the form Problem D* with a solution x^0. *Then for a given constraint operator $e : V \to Z$, the constrained optimization problem of the form Problem E has a solution provided that the feasible set is nonempty.*

2.2. Characterization of optimal solutions

In this section, we derive several necessary optimality conditions for general optimal control problems with equality constraints of the form Problem E. In doing so, we rely on infinite analogs of the finite-dimensional derivative which are discussed in Appendix A. Here, we generally follow Hinze *et al.* [**40**].

2.2.1. The co-state equation.

DEFINITION. Suppose we are given a problem of the form Problem E. We say the function $L : X \times Y \times Z^* \to \mathbb{R}$ defined as

$$L(u, \alpha, p) = f(u, \alpha) - \langle p, e(u, \alpha) \rangle_{Z^*, Z} \quad (u \in X, \alpha \in \mathcal{A}, p \in Z^*) \tag{2.2.1}$$

is the ***augmented objective functional***.

REMARK. We use the term "augmented objective functional" herein to avoid confusion with the term "Lagrangian" which is also used in the literature to describe the integrand of the cost functional in standard calculus of variations theory. □

Above we used the the stationary points of the objective functional to identify possible minimizers to a constrained optimization problem. Since the objective functional f and the constraint function $e(u, \alpha)$ in this case are assumed to be Fréchet differentiable at every $(u, \alpha) \in X \times \mathcal{A}$, we can compute the partial derivatives of L from its definition directly.

Therefore fix $(u, \alpha) \in X \times \mathcal{A}$. Then for any $v \in X$, we have

$$\begin{aligned}\langle D_u L(u, \alpha, p), v\rangle_{X^*,X} &= \langle D_u f(u, \alpha), v\rangle_{X^*,X} - \langle p, D_u e(u, \alpha) v\rangle_{Z^*,Z} \\ &= \langle D_u f(u, \alpha) - D_u e(u, \alpha)^* p, v\rangle_{X^*,X},\end{aligned}$$

where $D_u e(u, \alpha)$ is the Fréchet derivative with respect to u and $D_u e(u, \alpha)^*$ is its adjoint operator. Since this is true for any $v \in X$, we have the operator identity

$$D_u L(u, \alpha, p) = D_u f(u, \alpha) - D_u e(u, \alpha)^* p. \tag{2.2.2}$$

Hence, if there exists $p \in Z^*$ for a particular $(u, \alpha) \in X \times \mathcal{A}$ such that

$$D_u e(u, \alpha)^* p = D_u f(u, \alpha)$$

then

$$D_u L(u, \alpha, p) = 0$$

and thus $(u, \alpha, p) \in X \times \mathcal{A} \times Z^*$ is a stationary point of L. This gives rise to the following definition.

DEFINITION. Assume that we are given a problem of the form Problem E. Then for a given $(u, \alpha) \in X \times \mathcal{A}$, we say the following equation for an unknown $p \in Z^*$

$$D_u e(u, \alpha)^* p = D_u f(u, \alpha) \tag{2.2.3}$$

is the ***co-state equation***. If a solution p exists, it is called a ***co-state***.

REMARK. Equation (2.2.3) is also commonly referred to as the ***adjoint equation***. Similarly, a solution p to this equation is often called the ***adjoint state***. We use the term, "co-state", herein to avoid confusion with the adjoint operator and, in particular, the adjoint PDE which is discussed extensively in the sequel. □

REMARK. Recall that finite dimensional case, we can solve for a co-state (known as a Lagrange multiplier in this case) using (2.2.3) if a minimizer is a regular point. There is a analogous result in infinite dimensions which shows that we can solve for the co-state if a minimizer is a regular point in the infinite dimensional sense (*cf.* [**47**, p.28]). However, in our case, we address the issue of the existence and uniqueness of the co-state using results from standard PDE theory. Consequently, we do not need to rely on this result and it is outside the scope of this paper. □

2.2.2. The variational inequality. We now derive a maximum principal that is an infinite dimensional analog to Fermat's stationary point theorem from ordinary calculus. Ideally, we would like to express this result in terms of the gradient of the objective functional. However, we cannot define a true derivative in a general Banach space. Instead, we turn to the *first variation*, as an analog to the standard directional derivative, to derive the result.

THEOREM 2.2.1. *Suppose $u^0 \in V$ is a solution to Problem D, where V is a nonempty subset of a real Banach space X. Further assume that the first variation of the objective functional f at u^0 in the direction $u - u^0$, denoted $\delta f(u^0)$, exists, where $u \in X$. Then u^0 satisfies the* **variational inequality**

$$\delta f\left(u^0\right)\left(u - u^0\right) \geq 0.$$

PROOF. Let $u^0 \in V$ be a solution to Problem D, where $V \subset X$ is nonempty and fix $u \in X$. By assumption $u - u^0$ is feasible direction since $\delta f(u^0)$ is assumed to exist. Then by A.4.1, there exists

a real, positive sequence $\{\tau_j\}_{j\in\mathbb{N}} \downarrow 0$ such that

$$u^0 + \tau_j \left(u - u^0\right) \in V$$

for all j. Since u^0 is a minimizer, we get

$$\lim_{j\to\infty} \frac{f\left(u^0 + \tau_j \left(u - u^0\right)\right) - f\left(u^0\right)}{\tau_j} \geq 0.$$

Upon inspection, the left-hand side is the definition of the first variation (*cf.* A.4.2) and by assumption, the limit exists. Thus, we have

$$\delta f\left(u^0\right)\left(u - u^0\right) \geq 0.$$

□

In the case that f is convex and Gâteaux differentiable, we can state the following necessary optimality condition which is analogous to Fermat's well-known result from ordinary calculus.

THEOREM 2.2.2. *Suppose $u^0 \in V$ is a solution to Problem D, where V is a convex subset of a real Banach space X. Additionally assume that f is Gâteaux differentiable at u^0 and denoted the Gâteaux derivative as $Gf\left(u^0\right)$. Then*

$$\left\langle Gf\left(u^0\right), u - u^0\right\rangle \geq 0 \quad (\textit{for all } u \in X). \tag{2.2.4}$$

If u^0 is in the interior of V, then $Gf\left(u^0\right) = 0$.

PROOF. Fix $x \in V$. Since f is Gâteaux differentiable at u^0 by assumption, the directional derivative exists at u^0 and is equal to the Gâteaux derivative. Since V is convex by assumption, every direction in V is feasible. Therefore we can apply the Theorem 2.2.1 above to all $u \in V$ and thus we have

$$Gf\left(u^0\right)\left(u - u^0\right) \geq 0 \quad (u \in V).$$

Since $Gf\left(u^0\right) \in X^*$, we can write the above expression as

$$\left\langle Gf\left(u^0\right), u - u^0\right\rangle \geq 0.$$

Now assume that V is open. Then we can chose $\lambda > 0$ small enough such that $u^0 + \lambda u, u^0 - \lambda u \in V$ for any $u \in V$. Therefore fix $u \in V$ and choose such a λ and define v as

$$v := u^0 + \lambda u.$$

Under this definition, v is admissible and we can apply the above inequality to get

$$0 \leq Gf\left(u^0\right)\left(y - u^0\right) = Gf\left(u^0\right)\left(\left(u^0 + \lambda u\right) - u^0\right) = \lambda Gf\left(u^0\right) u$$

by the linearity of the Gâteaux derivative. Similarly, we can define w as

$$w := u^0 - \lambda u$$

and thus we have

$$0 \leq Gf\left(u^0\right)\left(u - u^0\right) = -\lambda Gf\left(u^0\right) u.$$

Combining the two results implies

$$\lambda Gf\left(u^0\right) u = 0.$$

Since $\lambda > 0$ and the choice of x was arbitrary, we have

$$Gf\left(u^0\right) = 0.$$

□

The next result follows from that fact that if a function is Fréchet differentiable, then it is Gâteaux differentiable (*cf.* Theorem A.3.2).

COROLLARY 2.2.3. *Suppose $u^0 \in V$ is a solution to Problem D, where V is a convex subset of a real Banach space X. Additionally assume that f is Fréchet differentiable at u^0 and denoted the Fréchet derivative as $Df\left(u^0\right)$. Then*

$$\left\langle Df\left(u^0\right), u - u^0\right\rangle \geq 0 \quad \text{for all } x \in \mathcal{A}. \tag{2.2.5}$$

If u^0 is in the interior of V, then $Df\left(u^0\right) = 0$.

We now formulate the variational inequality for optimal control problems which can be expressed in reduced form.

THEOREM 2.2.4 (The variational inequality for optimal control problems). *Suppose $\left(u^0, \alpha^0\right) \in X \times \mathcal{A}$ is a solution to Problem B such that the admissible set $\mathcal{A}$ is convex subset of a real Banach space Y. Additionally, assume that the constraint operator $e : X \times \mathcal{A} \to Z$ is Fréchet differentiable at $\left(u^0, \alpha^0\right)$ and that $D_u e\left(u^0, \alpha^0\right)$ is invertible. Also assume that a control-to-state operator S defined as $\alpha \mapsto u^\alpha$ for $\alpha \in \mathcal{A}$ exists and that it is Fréchet differentiable at α^0. Lastly assume that the co-state $p^0 \in Z^*$ exists. Then for any $\alpha \in Y$, we have*

$$\left\langle D_\alpha f\left(u^0, \alpha^0\right) - D_\alpha e\left(u^0, \alpha^0\right)^* p^0, \alpha - \alpha^0\right\rangle_{Y^*, Y} \geq 0, \tag{2.2.6}$$

where $D_\alpha e\left(u^0, \alpha^0\right)^*$ is the adjoint operator of $D_\alpha e\left(u^0, \alpha^0\right)$.

PROOF. Let $\left(u^0, \alpha^0\right) \in X \times \mathcal{A}$ be a solution to the given optimal control problem and let S be the control-to-state operator defined as $\alpha \mapsto u^\alpha$ for $\alpha \in \mathcal{A}$. Using this operator, we can state our optimal control problem in reduced form and we have

$$\left(u^0, \alpha^0\right) = \left(S\left(\alpha^0\right), \alpha^0\right),$$

where $u^0 := u^{\alpha^0}$ such that α^0 is a solution to the problem in reduced form.

Now let $\hat{f} : \mathcal{A} \to \mathbb{R}$ be the reduced objective functional and let $\hat{e} : \mathcal{A} \to Z$ be reduced constraint operator, respectively, where

$$\hat{f}(\alpha) := f(S(\alpha), \alpha) \text{ and } \hat{e}(\alpha) := e(S(\alpha), \alpha).$$

Then by the chain rule for Fréchet differentiation (*cf.* Theorem A.3.3), we have

$$D\hat{f}\left(\alpha^0\right) = D_u f\left(S\left(\alpha^0\right), \alpha^0\right) DS\left(\alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right), \tag{2.2.7}$$

where $DS\left(\alpha^0\right) \in \mathscr{L}(Y, X)$ is the Fréchet derivative of the control-to-state operator at α^0 which exists by assumption. Consequently, for $\alpha \in Y$, we have

$$D\hat{f}\left(\alpha^0\right)\alpha = \left(D_u f\left(S\left(\alpha^0\right), \alpha^0\right) DS\left(\alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right)\right)\alpha$$

$$
\begin{aligned}
&= D_u f\left(S\left(\alpha^0\right), \alpha^0\right) DS\left(\alpha^0\right) \alpha + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right) \alpha \\
&= \left\langle D_u f\left(S\left(\alpha^0\right), \alpha^0\right), DS\left(\alpha^0\right) \alpha \right\rangle_{X^*, X} + \left\langle D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right), \alpha \right\rangle_{Y^*, Y} \\
&= \left\langle DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right), \alpha \right\rangle_{Y^*, Y} + \left\langle D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right), \alpha \right\rangle_{Y^*, Y} \\
&= \left\langle DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right), \alpha \right\rangle_{Y^*, Y} \\
&= \left(DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right)\right) \alpha.
\end{aligned}
$$

Since this true for all $\alpha \in Y$, we have the following operator identity

$$
D\hat{f}\left(\alpha^0\right) = DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right), \alpha^0\right). \tag{2.2.8}
$$

Note that the first term on the right-hand side is problematic since we must also consider the sensitivity of the objective functional to perturbations in the state variable as well as those in the control. Thus we would like to remove this in order to have a problem truly in reduced form dependent only on the control.

Consequently, we turn to the constraint operator. Recall that in reduced form, the state equation is

$$
\hat{e}\left(\alpha\right) = 0 \quad \left(\text{for all } \alpha \in \hat{\mathcal{F}}\right),
$$

where $\hat{\mathcal{F}}$ is the reduced feasible set which is nonempty since α^0 is a solution to the reduced problem by assumption. Differentiating both sides of this equation, we get

$$
D\hat{e}\left(\alpha\right) = 0 \quad \left(\alpha \in \hat{\mathcal{F}}\right).
$$

Therefore, using the chain rule once again, we have

$$
D_u e\left(S\left(\alpha\right), \alpha\right) DS\left(\alpha\right) + D_\alpha e\left(S\left(\alpha\right), \alpha\right) = 0 \quad \left(\alpha \in \hat{\mathcal{F}}\right).
$$

By assumption $D_u e\left(u^0, \alpha^0\right) = D_u e\left(S\left(\alpha^0\right), \alpha^0\right)$ is invertible. Thus, we have

$$
DS\left(\alpha^0\right) = -D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^{-1} D_\alpha e\left(S\left(\alpha^0\right), \alpha^0\right). \tag{2.2.9}
$$

Hence the adjoint p^0 state exists and is given by

$$
\begin{aligned}
p^0 &= \left(D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^*\right)^{-1} D_u f\left(S\left(\alpha^0\right), \alpha^0\right) \\
&= \left(D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^{-1}\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right).
\end{aligned} \tag{2.2.10}
$$

Since X is reflexive, (2.2.9) implies

$$
\begin{aligned}
DS\left(\alpha^0\right)^* &= -\left(D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^{-1} D_\alpha e\left(S\left(\alpha^0\right), \alpha^0\right)\right)^* \\
&= -D_\alpha e\left(S\left(\alpha^0\right), \alpha^0\right)^* \left(D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^*\right)^{-1}
\end{aligned}
$$

Recalling $DS\left(\alpha^0\right)^* : X^* \to Y^*$, we have

$$
\begin{aligned}
DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right), \alpha^0\right) &= -D_\alpha e\left(S\left(\alpha^0\right), \alpha^0\right)^* \left(D_u e\left(S\left(\alpha^0\right), \alpha^0\right)^*\right)^{-1} D_u f\left(S\left(\alpha^0\right), \alpha^0\right) \\
&= -D_\alpha e\left(S\left(\alpha^0\right), \alpha^0\right)^* p^0
\end{aligned}
$$

by (2.2.10).

Substituting this into (2.2.8), we get

$$\begin{aligned} D\hat{f}\left(\alpha^0\right) &= DS\left(\alpha^0\right)^* D_u f\left(S\left(\alpha^0\right),\alpha^0\right) + D_\alpha f\left(S\left(\alpha^0\right),\alpha^0\right) \\ &= -D_\alpha\left(S\left(\alpha^0\right),\alpha^0\right)^* p^0 + D_\alpha f\left(S\left(\alpha^0\right),\alpha^0\right). \end{aligned} \tag{2.2.11}$$

Since α^0 is an optimal solution, it also satisfies the variational inequality by Corollary 2.2.3 and thus

$$\left\langle D\hat{f}\left(\alpha^0\right), \alpha - \alpha^0 \right\rangle_{Y^*,Y} \geq 0 \quad (\alpha \in Y).$$

Substituting 2.2.11 into above expression, we have

$$\left\langle D_\alpha f\left(u^0,\alpha^0\right) - D_\alpha e\left(u^0,\alpha^0\right)^* p^0, \alpha - \alpha^* \right\rangle_{Y^*,Y} \geq 0 \quad (\alpha \in Y).$$

□

2.2.3. The optimality system. We now summarize the results derived above as a single statement describing first-order necessary optimality conditions for optimal control problems that can be expressed in reduced form.

DEFINITION 2.2.5. Assume that we are given a problem of the form Problem B, where f is the objective functional, e is the constraint operator, and $\mathcal{A} \subset Y$ is the admissible set. Then assuming all the derivatives exist, we call the system of equations

$$\begin{cases} e(u,\alpha) = 0 & & \text{(the state equation)} \\ D_u e(u,\alpha)^* p - D_u f(u,\alpha) = 0 & & \text{(the co-state equation)} \\ \langle D_\alpha f(u,\alpha) - D_\alpha e(u,\alpha)^* p, \alpha - \beta\rangle_{Y^*,Y} \geq 0 & (\beta \in Y) & \text{(the variational inequality)} \end{cases} \tag{2.2.12}$$

the problem's ***optimality system***, where $(u,\alpha,p) \in X \times \mathcal{A} \times Z^*$. Here X is the state space, Y is the control space, and Z^* is the dual of the value space Z.

Alternatively, we can state this system in terms of the partial derivatives augmented objective functional L.

THEOREM 2.2.6. *Assume that we are given a problem of the form Problem B where f is the objective functional, e is the constraint operator, and the admissible set $\mathcal{A}$ is convex. Then its optimality system is given by*

$$\begin{cases} D_p L(u,\alpha,p) = 0 & & \textit{(the state equation)} \\ D_u L(u,\alpha,p) = 0 & & \textit{(the adjoint equation)} \\ \langle D_\alpha L(u,\alpha,p), \alpha - \beta\rangle_{Y^*,Y} \geq 0 & (\beta \in \mathcal{A}). & \textit{(the variational inequality)} \end{cases}$$

for any $(u,\alpha,p) \in X \times \mathcal{A} \times Z^*$, where $L : X \times \mathcal{A} \times Z^* \to \mathbb{R}$ is the augmented objective functional.

PROOF. Assume that we are given such an optimal control problem let L be the augmented objective functional. Then L is given by

$$L(u,\alpha,p) = f(u,\alpha) - \langle p, e(u,\alpha)\rangle_{Z^*,Z} \quad ((u,\alpha,p) \in X \times \mathcal{A} \times Z^*).$$

A simple calculation shows that the partial derivative of L with respect to the co-state state p at any $(u, \alpha, p) \in X \times \mathcal{A} \times Z^*$ is given by

$$D_p L(u, \alpha, p) = e(u, \alpha).$$

Hence we can express the state equation as

$$D_p L(u, \alpha, p) = 0 \quad (\text{for any } p \in Z^*).$$

Additionally, recall that in deriving the co-state equation we showed at (2.2.2) that

$$D_u L(u, \alpha, p) = D_u e(u, \alpha)^* p - D_u f(u, \alpha) \quad (\text{for any } (u, \alpha, p) \in X \times \mathcal{A} \times Z^*).$$

Consequently, we can express the co-state equation in terms of L as

$$D_u L(u, \alpha, p) = 0.$$

Lastly, from same computation, we have

$$\langle D_\alpha L(u, \alpha, p), \beta \rangle_{Y^*, Y} = \langle D_\alpha f(u, \alpha) - D_\alpha e(u, \alpha)^* p, \beta \rangle_{Y^*, Y} \quad (\beta \in \mathcal{A}).$$

Thus we can express the variation inequality as

$$\langle D_\alpha L(u, \alpha, p), \alpha - \beta \rangle_{Y^*, Y} \geq 0 \quad (\beta \in \mathcal{A}).$$

□

Summarizing the above calculations, we state first-order optimality conditions for problems of the form Problem B that can be expressed in reduced form.

THEOREM 2.2.7. *Let (u^0, α^0) be a solution to a problem of the form Problem B* and assume that there exists a $p^0 \in Z^*$ which is a solution to the co-state equation. Then (u^0, α^0, p^0) *satisfies the optimality condition assuming that the Lagrangian in Fréchet differentiable at that point.*

2.3. Application to optimal control problems

In the two proceeding sections of this paper, we showed that if a suitable control-to-state operator S exists, we can restate a general optimal control problem of the form Problem B as a standard optimization problem. We then showed solutions to such problems exist by an extension of the extreme value theorem under the following hypotheses:

ASSUMPTION 2.3.1 (Hypotheses for the existence of solutions to a general optimal control problem).

(i) The state, control, and value spaces (X, Y, and Z, respectively) are real Banach spaces and X and Z are reflexive.
(ii) The admissible set $\mathcal{A} \subset Y$ is nonempty, closed, and bounded, convex (or simply closed and convex if the reduced objective functional is coercive).
(iii) The control-to-state operator S is injective and weakly continuous.
(iv) There exists at least one feasible solution to the constrained problem.
(v) The reduced objective functional is lower semicontinuous and convex.

Subsequently, we established three first-order necessary optimal conditions for such problems under the following hypotheses:

ASSUMPTION 2.3.2 (Hypotheses for the existence of an optimality system).

(i) The control-to-state operator S is Fréchet differentiable at α^0 in any direction, where α^0 is a solution to the reduced problem.

(ii) The full objective functional f and constraint operator e are Fréchet differentiable at any solution $(u^0, \alpha^0) = (S(\alpha^0), \alpha^0)$.

(iii) The co-state equation has a solution as in the case where $D_u e(u^0, \alpha^0)$ is invertible.

Given these hypotheses, we state the main result of this chapter.

THEOREM 2.3.3. *Assume we are given an optimal control problem of the form Problem B, then a solution exists under the assumptions* (2.3.1). *Further, assuming that such a solution exists, it satisfies the optimality system provided that the assumptions* (2.3.2) *are met.*

CHAPTER 3

ELLIPTIC PDE OPTIMAL CONTROL PROBLEMS

For the remainder of this paper, we consider optimal control problems which are constrained by uniformly elliptic PDE of the form

$$\begin{cases} Lu = f & \text{in } U \\ u = g & \text{on } \partial U, \end{cases} \tag{3.0.1}$$

where U is a given open subset of $\mathbb{R}^n$ with C^1 boundary, the boundary condition g is given, and u is unknown. The given nonhomogeneous term f is assumed to be in $L^2(U)$ and the given differential operator L is linear, second order, and elliptic written in divergence form as

$$Lu(x) = -\sum_{i,j=1}^{n} \left(a^{ij}(x)\, u_{x_i}\right)_{x_j} + \sum_{i=1}^{n} b^i(x)\, u_{x_i} + c(x)\, u \tag{3.0.2}$$

for given coefficients a^{ij}, b^i, c $(i,j = 1,\ldots,n)$. We assume there exists a real constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a^{ij}\xi_i\xi_j \geq \theta |\xi|^2 > 0 \quad \text{(a.e. } x \in U)$$

for all nonzero $\xi \in \mathbb{R}^n$. We note that without loss of generality, we need only consider problems in which the boundary condition g is identically zero in the trace sense (*cf.* [**27**, pp. 271-275]).

DEFINITION. Given such a constraint, we consider the following two types of problems defined below:

(i) ***Linear problems***, in which the control on appears solely in the nonhomogeneous term of the constraint PDE and as such, a linear (or affine) with respect to the controls. They are typically of the form

$$Lu = f(\alpha), \tag{3.0.3}$$

where the given nonhomogeneous term f is an affine mapping from the admissible set $\mathcal{A}$ to the value space Z of the form

$$f(\alpha) = A\alpha + g \quad (\alpha \in \mathcal{A}), \tag{3.0.4}$$

such that $A : Y \to Z$ is a given linear operator and g is given.

(ii) ***Coefficient control problems***, in which the controls appear as one of more coefficients in the differential operator and are generally nonlinear with respect to the controls.

We begin our study of elliptic PDE optimal control problems with a brief review of linear problems primarily to introduce the necessary PDE theory to establish the existence of feasible solutions. However, such problems have the additional benefit that much of the general optimal control theory

developed in the previous chapter can be used to both establish the existence of optimal solutions as well as characterize them using standard Lagrange methods. In presenting optimal control results, I generally follow J.L. Lions [**60**], F. Tröltzsch [**84**], and M.Hinze, *et al.* [**40**] while I generally follow Evans [**27**, ch.6] with regard to the PDE theory.

ASSUMPTION. For the remainder of this chapter, we define the admissible set $\mathcal{A} \subset L^2(U)$ as

$$\mathcal{A} := \left\{ \alpha \in L^2(U) \mid \underline{\alpha} \leq \alpha(x) \leq \overline{\alpha} \text{ for } x \in U \right\}, \tag{3.0.5}$$

where $0 < \underline{\alpha} < \overline{\alpha} < \infty$. We note that under this definition, $\mathscr{A}$ is clearly nonempty, closed, bounded, and convex.

Further, we assume that the objective functional $I : H_0^1(U) \times \mathcal{A} \to \mathbb{R}$ is of the form

$$I[u, \alpha] = \frac{1}{2} \|u - \hat{u}\|_{L^2(U)}^2 + \frac{\mu}{2} \|\alpha\|_{L^2(U)}^2, \tag{3.0.6}$$

where $\hat{u}$ is the given target function and $\mu \geq 0$ is the given penalty associated with control $\alpha \in \mathcal{A}$. □

REMARK. Optimization and optimal control problems which have an objective functional of the form (3.0.6) are often referred to as ***norm minimizing problems***. In such problems, the objective function is bounded from below and given our particular objective function as defined at (3.0.6), it bounded by zero. □

NOTATION. For the remainder of the the paper, we drop the subscript from the L^2-norm and inner product and simply use

$$\|\cdot\| := \|\cdot\|_{L^2(U)} \text{ and } (\cdot, \cdot) := (\cdot, \cdot)_{L^2(U)}$$

unless otherwise noted. □

3.1. The existence of feasible solutions

It is well-know that classical solutions to uniformly elliptic PDE of the form (3.0.1) may not exist generally. Therefore, in the section, we introduce the concept of weak solutions to such problems, as well the Lax-Milgram Theorem, which establishes the conditions under which such solutions exist and are unique.

3.1.1. The Sobolev spaces $H_0^1(U)$ and $H^{-1}(U)$.

DEFINITION. Assume that $u, v \in L^1_{\text{loc}}(U)$ and α is a multi-index. We say that v is the α^{th}-***weak derivative of*** u, denoted as

$$D^\alpha u = v,$$

if

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

for all test functions $\phi \in C_c^\infty(U)$, where $C_c^\infty(U)$ is the space of smooth functions with compact support.

DEFINITION. We define the ***Sobolev space***

$$W^{k,p}(U)$$

as the collection of locally summable functions $u : U \to \mathbb{R}$ such that for each multi-index α with $|\alpha| \leq k$, $D^\alpha u \in L^p(U)$. In particular, if $p = 2$, we write

$$H^k(U) = W^{k,2}(U) \quad (k = 0, 1, \ldots).$$

Further, we write

$$W_0^{k,p}(U)$$

to denote the closure of $C_c^\infty(U)$ in $W^{k,p}(U)$ and, in particular,

$$H_0^k(U) = W_0^{k,2}(U) \quad (k = 0, 1, \ldots),$$

REMARK. Under this definition, we can interpret $W_0^{k,p}(U)$ as the collection of functions $u \in W^{k,p}(U)$ such that

$$D^\alpha u = 0 \text{ on } \partial U \text{ in the trace sense for all } |\alpha| \leq k - 1.$$

□

It is well-known that equipped with the following norm,

$$\|\cdot\|_{W^{k,p}(U)} = \begin{cases} \left(\sum_{|\alpha|\leq k} \int_U |D^\alpha u|^p \, dx\right)^{1/p} & (1 \leq p < \infty) \\ \sum_{|\alpha|\leq k} \text{ess sup}_U |D^\alpha u| & (p = \infty), \end{cases}$$

$W^{k,p}(U)$ is a Banach space for $1 \leq p \leq \infty$. It is also well-known that equipped with the following inner product,

$$(u, v)_{H^k(U)} = \sum_{|\alpha|\leq k} \int_U Du \cdot Dv dx \quad \left(u, v \in H^k(U)\right),$$

$H^k(U)$ is a separable Hilbert space for any k. In particular, this is true for $H_0^1(U)$, where the norm and inner product are given by

$$\|\cdot\|_{H_0^1(U)} = \left(\|u\|_{L^2(U)}^2 + \|Du\|_{L^2(U)}^2\right)^{1/2}$$

and

$$(u, v)_{H_0^1(U)} = (u, v)_{L^2(U)} + (Du, Dv)_{L^2(U)},$$

respectively. And as a Hilbert space, $H_0^1(U)$ is reflexive and also isomorphic to its dual by the Riesz representation theorem. However, for the purposes of this paper, we identify its dual with the space $H^{-1}(U)$ endowed with the norm

$$\|f\|_{H^{-1}(U)} := \sup\left\{\langle f, u\rangle_{H^{-1}(U), H_0^1(U)} \mid u \in H_0^1(U), \|u\|_{H_0^1(U)} \leq 1\right\}$$

where f is a bounded linear functional on $H_0^1(U)$.

$H^{-1}(U)$ can be categorized as follows:

THEOREM 3.1.1. *Assume $f \in H^{-1}(U)$. Then there exists functions $f^0, f^1, \ldots, f^n$ in $L^2(U)$ such that*

$$\langle f, v\rangle = \int_U f^0 v + \sum_{j=1}^n f^j v_{x_j} dx \quad \left(v \in H_0^1(U)\right). \tag{3.1.1}$$

and

$$\|f\| = \inf\left\{ \int_U \sum_{j=1}^{n} \left|f^j\right|^2 \mid f \text{ satisfies (3.1.1) for } f^0, \ldots, f^n \in L^2(U) \right\}.$$

We note that by definition, we have $H_0^1(U) \subset L^2(U)$. In fact, by the Rellich-Kondrachov Theorem the embedding is compact and thus $H_0^1(U)$ is a dense subspace of $L^2(U)$. Further, following Evans (*cf.* [**27**, pp.299-300]), it can be shown that $L^2(U) \subset H^{-1}(U)$ and that the embedding is continuous. Consequently, for any $u \in H_0^1(U)$ we have

$$\langle v, u \rangle_{H^{-1}(U), H_0^1(U)} = (v, u)_{L^2(U)} \quad \left(v \in L^2(U) \subset H^{-1}(U)\right). \tag{3.1.2}$$

by the Riesz Representation Theorem. These embeddings will be used extensively in the sequel.

NOTATION. For the remainder of this thesis, we drop subscript from the $H_0^1(U)$ dual pairing and simply denote it as

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{H^{-1}(U), H_0^1(U)}$$

unless noted otherwise. □

3.1.2. Weak solutions to elliptic PDE.

DEFINITION. Given a uniformly elliptic PDE of the form (3.0.2), define the bilinear form $B : H_0^1(U) \times H_0^1(U) \to \mathbb{R}$ as

$$B[u, v] = \int_U \sum_{i,j=1}^{n} a^{ij} u_{x_i} v_{x_j} + \sum_{i=1}^{n} b^i u_{x_i} v + cuvdx \quad \left(\text{for all } u, v \in H_0^1(U)\right), \tag{3.1.3}$$

where $a^{ij}, b^i, c \ (i, j = 1, \ldots, n)$ are the coefficients of L dependent on $x \in U$. We say $u \in H_0^1(U)$ is a ***weak solution to*** the PDE if it satisfies

$$B[u, v] = (f, v) \tag{3.1.4}$$

for all $v \in H_0^1(U)$.

We note that we can define the weak of PDE of the form (3.0.2) in which the nonhomogeneous term $f \in H^{-1}(U)$ as

$$B[u, v] = \langle f, v \rangle. \tag{3.1.5}$$

THEOREM 3.1.2 (Lax-Milgram Theorem). *Assume that H is a Hilbert space and that $B : H \times H \to \mathbb{R}$ is a bilinear mapping for which there exists constants $a, b > 0$ such that*

$$|B[u, v]| \leq a \|u\| \|v\| \quad (u, v \in H) \quad (\textit{boundedness})$$

and

$$b \|u\|^2 \leq B[u, u] \quad (u \in H) \quad (\textit{coercivity}).$$

Lastly, let $f : H \to \mathbb{R}$ be a bounded linear functional on H. Then there exists a unique element $u \in H$ such that

$$B[u, v] = \langle f, v \rangle \quad (\textit{for all } v \in H).$$

PROOF. *See* Evans [**27**, p. 316]. □

THEOREM 3.1.3. *Assume that we are given a bilinear form B of the form* (3.1.3) *in which the coefficients are in $L^{\infty}(U)$. Then there are constants $\alpha, \beta > 0$ and $\gamma \geq 0$ such that*

$$|B[u,v]| \leq \alpha \|u\|_{H_0^1(U)} \|v\|_{H_0^1(U)}$$

and

$$\beta \|u\|^2_{H_0^1(U)} \leq B[u,u] + \gamma \|u\|^2_{L^2(U)}. \tag{3.1.6}$$

PROOF. *See* Evans [**27**, p. 318]. □

Consequently, if $\gamma = 0$, then we then we apply the Lax-Milgram Theorem to show that a solution to a given uniformly elliptic PDE exists and is unique such as in the cases in which (i) each of the b^i are identically zero and $c \geq 0$ or (ii) c is sufficient large (*cf.* [**27**, p.320]).

3.2. The existence of optimal solutions

To ensure the existence of feasible solutions in the sequel, we assume the state space is $H_0^1(U)$, the control space Y is $L^2(U)$, and the value space Z is $H^{-1}(U)$,the dual of $H_0^1(U)$. Additionally, we assume that the coefficients of the differential operator L admit weak solutions. Given this, we are concerned in this chapter with linear problems of the form:

Problem F. (The elliptic linear problem)

$$\begin{cases} \min\limits_{u \in H_0^1(U), \alpha \in \mathcal{A}} & I[u,\alpha], \\ \text{subject to} & Lu = f(\alpha). \end{cases}$$

Now recall that in Section 2.1 above, we established the existence of solutions to general optimal control problems under assumptions set forth in Assumptions 2.3.1. Specifically,

(i) The state, control, and value spaces are real Banach spaces and the state and control spaces are reflexive.
(ii) The admissible set is a nonempty, closed, bounded, convex subset of a reflexive Banach space (or in the case that the objective functional is coercive, merely nonempty, closed, and convex).
(iii) The control-to-state operator is injective and weakly continuous (this ensures that (a) the problem can be expressed in reduced form and (b) a solution to the reduced problem can be appropriately extended to the full optimal control problem).
(iv) There is least one feasible solution to the reduced problem.
(v) The reduced objective functional is lower semicontinuous and convex.

Under the problem statement above, hypotheses (i) and (ii) are satisfied. Under the assumption that the coefficients of L admit a unique solution, hypothesis (iv) is met. This, in turn, ensures that the mapping $\alpha \to u^{\alpha}$ is injective for any admissible control $\alpha \in \mathcal{A}$ and thus it is a candidate for a control-to-state operator. Hypothesis (v) follows from the fact our problem is norm minimizing and the norm is both continuous and convex. Additionally, S is affine and bounded by (3.1.2) and thus it is continuous. Consequently, it is weakly continuous and therefore hypothesis (iii) is satisfied. This is captured in the following result.

THEOREM 3.2.1. *Assume we are given a optimal control problem of the form Problem F in which the PDE constraint is of the form 3.0.3. Then it satisfies Assumptions 2.3.1 and thus there exists at least one optimal solution.*

3.3. Characterization of optimal solutions

For this section, assume that we are given a linear problem of the form Problem F.

3.3.1. The state equation. In the Introduction to this thesis, we defined the state equation for a general optimal control problem as the operator identity

$$e(u, \alpha) = 0.$$

for a state-control pair $(u, \alpha) \in X \times \mathcal{A}$, where X is the state space, the admissible set $\mathcal{A}$ is a subset of the control space Y, and $e : X \times \mathcal{A} \to Z$ is the given constraint operator taking values in the value space Z. Consequently, the strong form of the state equation in the case of a linear problem under consideration in this chapter is given by

$$Lu - f(\alpha) = 0 \quad \left((u, \alpha) \in H_0^1(U) \times \mathcal{A}\right),$$

where L is a uniformly elliptic differential operator and the nonhomogeneous term $f(\alpha) \in L^2(U)$ for all $\alpha \in \mathcal{A}$.

As for the weak form of the state equation, by the embedding $H_0^1(U) \hookrightarrow L^2(U) \hookrightarrow H^{-1}(U)$ we have

$$\begin{aligned} 0 &= \langle e(u, \alpha), v \rangle \\ &= \langle Lu - f(\alpha), v \rangle \\ &= \langle Lu, v \rangle - \langle f(\alpha), v \rangle \\ &= \langle Lu, v \rangle - B[u, v], \end{aligned}$$

for any $v \in H_0^1(U)$ by (3.1.5), where B is the bilinear form associated with L defined at (3.1.3). This implies

$$B[u, v] = \langle Lu, v \rangle,$$

and thus the weak form of the state equation is given by

$$B[u, v] - (f(\alpha), v) = 0 \quad \left(\text{for all } v \in H_0^1(U)\right).$$

3.3.2. The co-state equation. At (2.2.3), we defined the abstract form of the co-state equation as

$$D_u e(u, \alpha)^* p = D_u I(u, \alpha),$$

where I is the objective functional, e is constraint operator, $D_u e(u, \alpha)^*$ is the adjoint operator of $D_u e(u, \alpha)$, the partial Fréchet derivative of e with respect to the state and a solution p is the co-state assuming it exists. We now state a concrete form of this equation for Problem F.

LEMMA 3.3.1 (Gradient of the objective functional). *Let $I : H_0^1(U) \times \mathcal{A} \to \mathbb{R}$ be a functional defined as,*

$$I[u, \alpha] = \frac{1}{2} \|u - \hat{u}\|_{L^2(U)}^2 + \frac{\mu}{2} \|\alpha\|_{L^2(U)}^2,$$

where $\hat{u}$ and μ are given. Then I is Fréchet differentiable at any $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$ and its Fréchet derivative is given by

$$DI[u,\alpha](v,\beta) = D_u I[u,\alpha] v + D_\alpha I[u,\alpha] \beta,$$

for any $(v,\beta) \in H_0^1(U) \times \mathcal{A}$, where

$$D_u I[u,\alpha] = u - \hat{u} \quad \text{and} \quad D_\alpha I[u,\alpha] = \mu\alpha. \tag{3.3.1}$$

PROOF. We begin by explicitly computing the partial derivative of $I[u,\alpha]$ with respect to the state u. Therefore fix $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$ and $v \in H_0^1(U)$ and define the function $i : \mathbb{R} \to \mathbb{R}$ as

$$i(\tau; u, v) := I[u + \tau v, \alpha] = \frac{1}{2}\int_U |u + \tau v - \hat{u}|^2 \, dx + \frac{\mu}{2}\int_U |\alpha|^2 \, dx.$$

for any $\tau \in \mathbb{R}$. Taking the derivative of i with respect to τ, we have

$$\begin{aligned} i'(0; u, v) &= \frac{1}{2}\frac{d}{d\tau}\int_U |u + \tau v - \hat{u}|^2 \, dx \,|_{\tau=0} \\ &= (u - \hat{u}, v). \end{aligned}$$

On the other hand, by (A.1.2) we have

$$i'(0; u, v) = (G_u I[u,\alpha], v)$$

for any $v \in H_0^1(U)$, where $G_u I[u,\alpha]$ is the Gâteaux derivative with respect to u. Combining these two results we have

$$(G_u I[u,\alpha], v) = (u - \hat{u}, v) \quad \left(\text{for all } v \in H_0^1(U)\right)$$

and thus we have the operator identity

$$G_u I[u,\alpha] = u - \hat{u} \tag{3.3.2}$$

for a given $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$.

I now claim that I is Fréchet differentiable. Therefore consider the following expression

$$I[u+v,\alpha] - I[u,\alpha] = (G_u I[u,\alpha], v) + r(u; v) \quad \left(v \in H_0^1(U)\right),$$

where $r : H_0^1(U) \to \mathbb{R}$ and fix $v \in H_0^1(U)$. After rearranging terms, we have

$$\begin{aligned} r(u; v) &= I[u+v,\alpha] - I[u,\alpha] - (G_u I[u,\alpha], v) \\ &= \frac{1}{2}\int_U |(u+v) - \hat{u}|^2 - |u - \hat{u}|^2 - 2(u - \hat{u}) v dx \\ &= \frac{1}{2}\int_U |v|^2 \, dx = \frac{1}{2}\|v\|^2. \end{aligned} \tag{3.3.3}$$

Dividing both sides by $\|v\|$, we see that

$$\frac{|r(u; v)|}{\|v\|} \to 0 \text{ as } \|v\| \to 0$$

and thus I is Fréchet differentiable with respect to u at any $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$.

We can similarly compute $D_\alpha I$. For this, fix $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$ and $\beta \in \mathcal{A}$ and define $i : \mathbb{R} \to \mathbb{R}$ as

$$\begin{aligned} i(\tau;\alpha,\beta) &= I[u,\alpha+\tau\beta] \\ &= \frac{1}{2}\int_U |u-\hat{u}|^2\,dx + \frac{\mu}{2}\int_U |\alpha+\tau\beta|^2\,dx. \end{aligned}$$

Again taking the derivative of i with respect to τ and evaluating it at $\tau = 0$, we have

$$\begin{aligned} i'(0;\alpha,\beta) &= \frac{\mu}{2}\frac{d}{d\tau}\int_U |\alpha+\tau\beta|^2\,dx\,|_{\tau=0} \\ &= (\mu\alpha,\beta). \end{aligned}$$

Using (A.1.2) once more, we have

$$(G_\alpha I[u,\alpha],\beta) = i'(0;\alpha,\beta) = (\mu\alpha,\beta)$$

and thus

$$G_\alpha I[u,\alpha] = \mu\alpha \tag{3.3.4}$$

for any $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$.

The fact I is actually Fréchet differentiable follows for a calculation nearly identical to the one presented above and thus

$$D_\alpha I[u,\alpha] = \mu\alpha \tag{3.3.5}$$

for any $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$. □

We now compute the strong and weak forms of the partial derivative of the state operator with respect to the state.

LEMMA 3.3.2. *Assume that we are given a function $e : H_0^1(U) \times \mathcal{A} \to H^{-1}(U)$ defined as*

$$e(u,\alpha) = Lu + f(\alpha),$$

where L is uniformly elliptic operator of the form (3.0.2). Then for $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$, the strong form its partial derivative with respect to the state u at $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$ is given by operator identity

$$D_u e(u,\alpha) = L,$$

and its weak form by

$$D_u e(u,\alpha) = B[u,v] \quad \big(\textit{for any } v \in H_0^1(U)\big),$$

where B is weak bilinear form associated with L.

Additionally, its partial derivative with respect to control α

$$D_\alpha e(u,\alpha) = -A. \tag{3.3.6}$$

PROOF. Assume the set up above. We begin with computing $D_u e(u,\alpha)$. For this, fix $(u,\alpha) \in H_0^1(U) \times \mathcal{A}$ and $v \in H_0^1(U)$ such that u and v are sufficiently smooth and define the function $i : \mathbb{R} \to \mathbb{R}$ as

$$i(\tau;u,\alpha) = e(u+\tau v,\alpha) = L(u+\tau v) + f(\alpha) \quad (\tau \in \mathbb{R}).$$

This implies

$$i'(\tau; u, \alpha) = \frac{d}{d\tau} L(u + \tau v) = Lv \quad (\tau \in \mathbb{R}).$$

On the other hand, by (A.1.2) we have

$$G_u e(u, \alpha) v = i'(0; u, v) = Lv,$$

and thus the strong and weak forms of the co-state equation are given by

$$G_u e(u, \alpha) = L \quad \text{and} \quad G_u e(u, \alpha) v = B[u, v], \text{ respectively.}$$

We now show that the weak form $e(u, \alpha)$ is Fréchet differentiable with respect to u. Since the strong form is obvious, we only need to show that the weak form Fréchet differentiable. Again fix $(u, \alpha) \in H_0^1(U) \times \mathcal{A}$ and $v \in H_0^1(U)$. Following the definition Fréchet derivative at A.3.1, for any $w \in H_0^1(U)$, the residual is given by

$$\begin{aligned} \langle r(v; u), w \rangle &= \langle e(u + v, \alpha), w \rangle - \langle e(u, \alpha), w \rangle - (G_u e(u, \alpha) v, w) \\ &= B[u + v, w] - B[u, w] - B[v, w] = 0. \end{aligned}$$

Hence $e(u, \alpha)$ is Fréchet differentiable with respect to u at any $(u, \alpha) \in H_0^1(U) \times \mathcal{A}$.

Following the same steps as above, we can similarly compute $D_\alpha e(u, \alpha)$. Fix $(u, \alpha) \in H_0^1(U) \times \mathcal{A}$ and $\beta \in \mathcal{A}$ and define the function $i : \mathbb{R} \to \mathbb{R}$ this time as

$$i(\tau; u, \alpha) = e(u, \alpha) = L(u) - f(\alpha + \tau\beta).$$

This implies

$$i'(\tau; u, 0) = -\frac{d}{d\tau} f(\alpha + \tau\beta) \mid_{\tau=0} = -A\beta \quad \Rightarrow \quad G_\alpha e(u, \alpha) \beta = -A\beta,$$

and thus we have

$$G_\alpha e(u, \alpha) = -A.$$

A simple calculation nearly identical to the immediately shows that residual term is also zero and thus $e(u, \alpha)$ is Fréchet differentiable with respect to the control α at any $(u, \alpha) \in H_0^1(U) \times \mathcal{A}$ and is given by

$$D_\alpha e(u, \alpha) = -A.$$

□

The next results follows immediately from the lemmas above.

THEOREM 3.3.3. *Assume that we are given an linear problem of the form Problem F, where L is the elliptic operator. Then the strong form co-state equation is given by*

$$L^* p = u - \hat{u},$$

$\hat{u}$ is the given target function and L^* is the adjoint of L and its weak form is given by

$$B[p, v] = (u - \hat{u}, v) \quad \left(v \in H_0^1(U)\right),$$

where B is the bilinear form associated with L.

3.3.3. The variational inequality. At 2.2.6, we we defined the variational inequality for a general optimal control problem in which I is the objective functional, e is the constraint operator, and $\mathcal{A}$ is the admissible set, as

$$\left\langle D_\alpha I\left[u^0,\alpha^0\right]-D_\alpha e\left(u^0,\alpha^0\right)^* p^0,\beta-\alpha^0\right\rangle_{Y^*,Y}\geq 0\quad\left(\alpha\in\mathcal{A}\subset Y\right), \tag{3.3.7}$$

where $(u^0,\alpha^0)\in X\times\mathcal{A}$ is an optimal state-control pair and $p^0\in Z^*$ is the associated co-state assuming it exists. Here we assume state space X, the control space , and the value space Z are real, reflexive Banach spaces with dual spaces X^*,Y^* and Z^*. We now provide a concrete statement of this inequality for linear problems such as Problem F.

THEOREM 3.3.4. *Assume that $(u^0,\alpha^0)\in H_0^1(U)\times\mathcal{A}$ is a solution to a given linear problem of the form Problem F.Then the variational inequality associated with the problem is given by*

$$\left(\mu\alpha^0-A^*p^0,\alpha-\alpha^0\right)\geq 0\quad\left(\alpha^0\in\mathcal{A}\right),$$

where μ is the control penalty constant and $p^0\in H_0^1(U)$ is the co-state assuming it exists.

PROOF. Assume the set up above. By (3.3.1) and (3.3.6), we have

$$D_\alpha I\left[u^0,\alpha^0\right]=\mu\alpha^0$$

and

$$D_\alpha e\left(u^0,\alpha^0\right)^*=-A^*,$$

where $A^*:H_0^1(U)\to L^2(U)$ is the dual of the linear operator $A:L^2(U)\to H^{-1}(U)$. Substituting these into the above expression of the variational inequality, we have

$$\left(\mu\alpha^0-A^*p^0,\alpha-\alpha^0\right)\geq 0\quad\left(\alpha^0\in\mathcal{A}\right).$$

□

3.3.4. The optimality system. Given the results above, we state the strong and weak forms of the optimality system for linear problems of the form Problem F. Specifically, for $(u,\alpha)\in H_0^1(U)\times\mathcal{A}$, the strong form is given by

$$\begin{cases} Lu-f(\alpha)=0 & \\ L^*p-(u-\hat{u})=0 & \\ (\mu\alpha-A^*p,\beta-\alpha)\geq 0 & (\beta\in\mathcal{A}), \end{cases} \tag{3.3.8}$$

where $p\in H_0^1(U)$. Its weak form is

$$\begin{cases} B[u,v]-(f(\alpha),v)=0 & \text{for all } v\in H_0^1(U) \\ B[p,v]-(u-\hat{u},v)=0 & \text{for all } v\in H_0^1(U) \\ (\mu\alpha-A^*p,\beta-\alpha)\geq 0 & \beta\in\mathcal{A}, \end{cases} \tag{3.3.9}$$

where $p\in H_0^1(U)$.

CHAPTER 4

COEFFICIENT CONTROL PROBLEMS

We now turn our attention to a special class of optimal control problems constrained by elliptic PDE in which the controls appear in some of the coefficients of the second-order term of the differential operator. Again, our goal is to identify conditions that establish the existence of optimal solutions to such problems, as well as characterize them using a maximum principle. Unlike linear problems studied in the previous chapter, however, we cannot simply apply the extension of standard Lagrange methods to do this, as we show through two counterexamples below.

In this chapter, we again assume that U is an bounded, open subset of $\mathbb{R}^n$ with C^1 boundary and that the state, control, and value spaces are $H_0^1(U), L^2(U)$, and $H^{-1}(U)$, respectively. Additionally, we assume that the differential operator has the form:

$$L^\alpha u := -\text{div}(\alpha Du) = -\sum_{i,j=1}^{n} \left(\alpha^{ij}(x)\, u_{x_i}\right)_{x_j} \quad \left(u \in H_0^1(U)\right), \tag{4.0.1}$$

where the superscript α indicates that one or more of the operator's coefficients are controls. Additionally, we ignore the control penalty and simply set the parameter μ to zero. Consequently, we are interested in optimal control problems of the form:

Problem G. (The elliptic coefficient control problem)

$$\begin{cases} \min\limits_{u \in H_0^1(U), \alpha \in \mathcal{A}} & \frac{1}{2}\|u - \hat{u}\|^2, \\ \text{subject to} & -\sum\limits_{i,j=1}^{n} \left(\alpha^{ij}(x)\, u_{x_i}\right)_{x_j} = f \text{ in } U \\ & u = 0 \text{ on } \partial U, \end{cases}$$

where $\hat{u} \in L^2(U)$ is the given target function and $f \in L^2(U)$ is given. To ensure that feasible solutions exist, we assume that

$$\alpha^{ij} \in L^\infty(U) \quad (\text{for all } i, j = 1, \ldots, n)$$

and that there exists $\theta > 0$ such that

$$\sum_{i,j}^{n} \alpha^{ij} \xi_i \xi_j \geq \theta |\xi| > 0 \quad (\text{a.e. } x \in U)$$

for all $\xi \in \mathbb{R}^n$. However, we leave the choice of the admissible set $\mathcal{A}$ until later in the chapter.

4.1. Counterexamples to the existence of solutions to nonlinear optimal control problems

We begin this chapter by presenting two counterexamples which show optimal solutions to general non-linear PDE-constrained optimal control problems may not exist.

4.1.1. Problems with a nonlinear nonhomogeneous term. While we are generally interested in coefficient control problems for the remainder of this paper, the following counterexample found in Jahn [**48**, p.24-26] shows that even in the case of a linear problem which were discussed in the previous chapter, optimal solutions may not exist in the case in which nonhomogeneous term is nonlinear with respect to the controls.

Example 1. Let $U = [0,1] \subset \mathbb{R}$ and consider the following ODE

$$\begin{cases} u_x(x) + \alpha(x)\, u(x) = 0 \text{ in } U \text{ a.e.} \\ u(0) = 1, u(1) = 0, \end{cases} \tag{4.1.1}$$

where $\alpha \in L^2(U)$ is the control. A simple calculation confirms that the solution this ODE is given by

$$u(x) = C - \int_0^x |\alpha(y)|^2 \, dy \quad (x \in U).$$

Applying the given the boundary conditions, we also have

$$C = 1$$

and

$$1 - \int_0^1 \alpha(y)^2 \, dy = 0,$$

which implies

$$\|\alpha\|^2 = 1.$$

Hence $\alpha \in L^2(U)$ is feasible if it lies in the admissible set

$$\mathcal{A} := \left\{ \alpha \in L^2(U) \mid \|\alpha\|^2 = 1 \right\},$$

which we note is closed and bounded.

Given this, we consider optimal control problems of the form

$$\begin{cases} \min\limits_{u \in C^1(U), \alpha \in \mathcal{A}} & I[\alpha], \\ \text{subject to} & u_x(x) + \alpha(x)^2 = 0 \text{ in } U \\ & u(0) = 1, u(1) = 0, \end{cases}$$

where the objective functional $I : \mathcal{A} \to \mathbb{R}$ is defined as

$$I[\alpha] = \int_U x^2 \alpha(x)^2 \, dx \quad (\alpha \in A). \tag{4.1.2}$$

Clearly,

$$\inf_{\alpha \in \mathcal{A}} I[\alpha] \geq 0$$

and we now show that in fact

$$\inf_{\alpha\in\mathcal{A}} I[\alpha] = 0.$$

To see this, define

$$\alpha_j(x) = \begin{cases} j & \text{for } x \in \left[0, \frac{1}{j^2}\right) \\ \\ 0 & \text{for } x \in \left[\frac{1}{j^2}, 1\right] \end{cases}$$

for all j. Then for each j, we have

$$\|\alpha_j\|^2 = \int_0^1 |\alpha(y)|^2 dy = \int_0^{1/j^2} j^2 dt = 1$$

and thus $\alpha_j \in \mathcal{A}$ for all j. Moreover, by (4.1.2) we have

$$I[\alpha_j] = \int_0^1 x^2 \alpha_j(x)^2 dx = \int_0^{1/j^2} x^2 j^2 dx = \frac{1}{3j^4}$$

for each j and hence we have

$$\inf_{\alpha\in\mathcal{A}} I[\alpha] = 0.$$

Consequently, our optimal control has solution if there exists $\alpha \in \mathcal{A}$ such that $I[\alpha] = 0$. Therefore, assume such a control $\alpha \in \mathcal{A}$ exists. Then by (4.1.2), we have

$$\int_0^1 x^2 \alpha(x)^2 dx = 0.$$

However, the integrand is non-negative and thus α must be identically zero almost everywhere on U. But this is implies that $\alpha^0 \notin \mathcal{A}$ which is a contradiction. Hence no optimal solution exists. □

4.1.2. Controls that appear as coefficients of the differential operator. In this counterexample, first constructed by Murat (*cf.* [**82**, p.49]), we consider a one-dimensional coefficient control problem to show that solutions to such problems do not exist generally.

Example 2. Let $U = (0,1)$ and consider the following ODE

$$\begin{cases} -\left(\alpha(x) u_x^\alpha(x)\right)_x + \alpha(x) u^\alpha(x) = 0 \quad \text{in } U \\ u^\alpha(0) = 1, u^\alpha(1) = 2, \end{cases} \tag{4.1.3}$$

where the admissible set $\mathcal{A}$ defined as

$$\mathcal{A} = \left\{ \alpha \in L^\infty(U) \mid \text{a} := \frac{\sqrt{2}-1}{\sqrt{2}} \le \alpha(x) \le \frac{\sqrt{2}+1}{\sqrt{2}} =: \text{b} \ \text{ a.e. } x \in U \right\}.$$

Further, define the objective functional $I : \mathcal{A} \to \mathbb{R}$ as

$$I[\alpha] = \int_U \left| u^\alpha - \left(1 + x^2\right) \right|^2 dx. \tag{4.1.4}$$

As in the counterexample above, we wish to show that

$$\inf_{\alpha\in\mathcal{A}} I[\alpha] = 0$$

and yet, there is no $\alpha^0 \in \mathcal{A}$ such that

$$I\left[\alpha^0\right] = 0.$$

Therefore consider the following sequence $\{\alpha_j\}_{j\in\mathbb{N}} \subset \mathcal{A}$, defined for each j by

$$\alpha_j(x) = \begin{cases} 1 - \left(\frac{1}{2} - \frac{x^2}{6}\right)^{1/2} & \text{for } \frac{m}{j} < x \leq \frac{2m+1}{2j} \\ 1 + \left(\frac{1}{2} - \frac{x^2}{6}\right)^{1/2} & \text{for } \frac{2m+1}{2j} < x \leq \frac{m+1}{j} \end{cases}$$

for $m = 0, 1, \ldots, j-1$. Clearly, $\alpha_j(x) \in \mathcal{A}$ for all j, since

$$\frac{\sqrt{2}-1}{\sqrt{2}} \leq \alpha_j(x) \leq \frac{\sqrt{2}+1}{\sqrt{2}} \text{ for all } x \in U.$$

Recall that sequence $\{f_j\}_{j\in\mathbb{N}} \subset L^\infty(U)$ is said to converge weakly* in $L^\infty(U)$ if there exists $f \subset L^\infty(U)$ such that

$$\lim_{j\to\infty} \int_U f_j \phi dx = \int_U f \phi dx,$$

where for all test functions $\phi \in L^1(U)$.

CLAIM. Given the sequence $\{\alpha_j\}_{j\in\mathbb{N}}$ defined above, we have

$$\alpha_j \overset{*}{\rightharpoonup} \alpha_0 = 1 \text{ and } \frac{1}{\alpha_j} \overset{*}{\rightharpoonup} \frac{1}{\omega} \text{ in } L^\infty(U), \text{ where } \omega = \frac{1}{2} + \frac{x^2}{6}. \tag{4.1.5}$$

PROOF. Define the function f as

$$f(x) = \left(\frac{1}{2} - \frac{x^2}{6}\right)^{1/2} \quad (x \in (0,1)). \tag{4.1.6}$$

Additionally, define the sequence of functions $\{\psi_j\}_{j\in\mathbb{N}}$ as

$$\psi_j(x) = \begin{cases} -f(x) & \text{for } \frac{m}{j} < x \leq \frac{2m+1}{2j} \\ f(x) & \text{for } \frac{2m+1}{2j} < x \leq \frac{m+1}{j} \end{cases}$$

for each j and $m = 0, 1, \ldots, j-1$, respectively. Then for any j, we have

$$\begin{aligned} \int_0^1 \alpha_j \phi dx &= \int_0^1 \phi dx + \int_0^1 \psi_j \phi dx \\ &= \int_0^1 \phi dx + \sum_{m=0}^{j-1} \left(\int_{m/j}^{(2m+1)/2j} \psi_j \phi dx + \int_{(2m+1)/2j}^{(m+1)/j} \psi_j \phi dx \right) \end{aligned} \tag{4.1.7}$$

Now we can approximate each of the integrals above using trapezoidal approximation as

$$\begin{aligned}\int_{m/j}^{(2m+1)/2j} \psi_j \phi dx &= \frac{1}{2j}\left(\psi_j\left(\frac{2m+1}{2j}\right)\phi\left(\frac{2m+1}{2j}\right)+\psi_j\left(\frac{m}{j}\right)\phi\left(\frac{m}{j}\right)\right)+o\left(j^{-2}\right)\\ &= -\frac{1}{2j}\left(f\left(\frac{2m+1}{2j}\right)\phi\left(\frac{2m+1}{2j}\right)+f\left(\frac{m}{j}\right)\phi\left(\frac{m}{j}\right)\right)+o\left(j^{-2}\right)\end{aligned}$$

and

$$\begin{aligned}\int_{(2m+1)/2j}^{(m+1)/j} \psi_j \phi dx &= \frac{1}{2j}\left(\psi_j\left(\frac{m+1}{j}\right)\phi\left(\frac{m+1}{j}\right)+\psi_j\left(\frac{2m+1}{2j}\right)\phi\left(\frac{2m+1}{2j}\right)\right)+o\left(j^{-2}\right)\\ &= \frac{1}{2j}\left(f\left(\frac{m+1}{j}\right)\phi\left(\frac{m+1}{j}\right)+f\left(\frac{2m+1}{2j}\right)\phi\left(\frac{2m+1}{2j}\right)\right)+o\left(j^{-2}\right).\end{aligned}$$

Therefore, for each j we have

$$\begin{aligned}\sum_{m=0}^{j-1}\left(\int_{m/j}^{(2m+1)/2j}\psi_j\phi dx+\int_{(2m+1)/2j}^{(m+1)/j}\psi_j\phi dx\right) &= \frac{1}{2j}\sum_{m=0}^{j-1}\left(f\left(\frac{m+1}{j}\right)\phi\left(\frac{m+1}{j}\right)-f\left(\frac{m}{j}\right)\phi\left(\frac{m}{j}\right)\right)\\ &\quad +o\left(j^{-2}\right)\\ &= \frac{1}{2j}\left(\sum_{m=1}^{j}f\left(\frac{m}{j}\right)\phi\left(\frac{m}{j}\right)-\sum_{m=0}^{j-1}f\left(\frac{m}{j}\right)\phi\left(\frac{m}{j}\right)\right)\\ &\quad +o\left(j^{-2}\right)\\ &= \frac{1}{2j}\left(f(1)\phi(1)-f(0)\phi(0)\right)+o\left(j^{-2}\right).\end{aligned}$$

Substituting this into (4.1.7) and taking the limit, we have

$$\lim_{j\to\infty}\int_0^1 \alpha_j\phi dx = \int_0^1 \phi dx$$

and thus

$$\alpha_j \overset{*}{\rightharpoonup} \alpha_0 = 1 \text{ in } L^\infty(U).$$

Additionally, we have

$$\begin{aligned}\int_0^1 \frac{1}{\alpha_j}\phi dx = \sum_{m=0}^{j-1}\Bigg(&\int_{m/j}^{(2m+1)/2j}\frac{1}{1-f}\phi dx\\ &+\int_{(2m+1)/2j}^{(m+1)/j}\frac{1}{1+f}\phi dx\Bigg).\end{aligned}$$

Expressing the fractions under a common denominator and using trapezoidal approximation again, we have

$$\begin{aligned}\int_{m/j}^{(2m+1)/2j}\frac{1+f}{1-f^2}\phi dx = \frac{1}{2j}&\left(\frac{1+f\left(\frac{2m+1}{2j}\right)}{1-f^2\left(\frac{2m+1}{2j}\right)}\phi\left(\frac{2m+1}{2j}\right)+\frac{1+f\left(\frac{m}{j}\right)}{1-f^2\left(\frac{m}{j}\right)}\phi\left(\frac{m}{j}\right)\right)\\ &+o\left(j^{-2}\right)\end{aligned}$$

and

$$\int_{(2m+1)/2j}^{(m+1)/j} \frac{1-f}{1-f^2}\phi dx = \frac{1}{2j}\left(\frac{1-f\left(\frac{2m+1}{2j}\right)}{1-f^2\left(\frac{2m+1}{2j}\right)}\phi\left(\frac{2m+1}{2j}\right)+\frac{1-f\left(\frac{m}{j}\right)}{1+f^2\left(\frac{m}{j}\right)}\phi\left(\frac{m}{j}\right)\right) + o\left(j^{-2}\right).$$

Consequently, for each j we have

$$\begin{aligned}\int_0^1 \frac{1}{\alpha_j}\phi dx &= \frac{1}{2j}\left[\sum_{m=1}^{j}\frac{1}{1-f^2\left(\frac{m}{j}\right)}\phi\left(\frac{m}{j}\right)+\sum_{m=0}^{j-1}\frac{1}{1+f^2\left(\frac{m}{j}\right)}\phi\left(\frac{m}{j}\right)\right]+o\left(j^{-2}\right)\\ &= \frac{1}{2j}\left(\frac{1}{1-f^2(1)}\phi(1)+\frac{1}{1-f^2(0)}\phi(0)+\sum_{m=1}^{j-1}\frac{2}{1-f^2\left(\frac{m}{j}\right)}\phi\left(\frac{m}{j}\right)\right)+o\left(j^{-2}\right).\end{aligned}$$

Hence by (4.1.6) and applying trapezoidal approximation again, we have

$$\lim_{j\to\infty}\int_0^1 \frac{1}{\alpha_j}\phi dx = \int_0^1\left(\frac{1}{1-f(x)^2}\right)\phi dx = \int_0^1\left(\frac{1}{\frac{1}{2}+\frac{x^2}{6}}\right)\phi dx$$

and thus the claim is proved. □

Now since $\alpha_j > 0$ for all j, we can apply the Lax-Milgram Theorem to show that there exists a unique solution to (4.1.3) $u^j \in H^1(U)$ for each j.

Claim. The sequence $\{u^j\}_{j\in\mathbb{N}}$ is bounded in $H^1(U)$.

Proof. Fix j and consider the ODE (4.1.3) where the coefficient is α_j. After integrating by parts, we have

$$\int_U \left(\alpha_j u_x^{\alpha_j}\right) v_x + \alpha_j u^{\alpha_j} v = 0 \tag{4.1.8}$$

for any $v \in H_0^1(U)$. In particular, select v such that

$$v(x) = u^{\alpha_j}(x) - 1 - x.$$

Since $v(0) = 0$ and $v(1) = 0$, $v \in H_0^1(U)$. Substituting this v into (4.1.8), we have

$$\begin{aligned}0 &= \int_U \left(\alpha_j u_x^{\alpha_j}\right)\left(u_x^{\alpha_j}-1\right)+\alpha_j u^{\alpha_j}\left(u^{\alpha_j}-1-x\right)dx\\ &= \int_U \alpha_j\left(\left(u_x^{\alpha_j}\right)^2+\left(u^{\alpha_j}\right)^2\right)-\alpha_j\left(u_x^{\alpha_j}+u^{\alpha_j}(1+x)\right)dx.\\ &\geq \mathrm{a}\left(\int_U \left(u_x^{\alpha_j}\right)^2+\left(u^{\alpha_j}\right)^2 dx\right)-\mathrm{b}\int_U \left|u_x^{\alpha_j}\right|+\left|u^{\alpha_j}\right|\left|(1+x)\right|dx.\end{aligned}$$

Rearranging terms, we get

$$\mathrm{a}\int_U (u_x^{\alpha_j})^2 + (u^{\alpha_j})^2\,dx \le \mathrm{b}\int_U \left(|u_x^{\alpha_j}| + |u^{\alpha_j}|\,|(1+x)|\right)dx$$

which implies by Cauchy's inequality with ϵ

$$ab \le \varepsilon a^2 + \frac{1}{4\varepsilon}b^2,$$

that

$$\begin{aligned}
\mathrm{a}\int_U (u_x^{\alpha_j})^2 + (u^{\alpha_j})^2\,dx &\le \mathrm{b}\int_U |u_x^{\alpha_j}| + |u^{\alpha_j}|\,|(1+x)|\,dx \\
&\le \mathrm{b}\int_U \left(\frac{1}{4\varepsilon} + \varepsilon\,(u_x^{\alpha_j})^2\right) + \left(\varepsilon\,(u^{\alpha_j})^2 + \frac{1}{4\varepsilon}\,(1+x)^2\right)dx \\
&\le \mathrm{b}\varepsilon\int_U (u_x^{\alpha_j})^2 + (u^{\alpha_j})^2\,dx + \frac{\mathrm{b}}{4\varepsilon}\int_U 1 + (1+x)^2\,dx \\
&= b\varepsilon\int_U (u_x^{\alpha_j})^2 + (u^{\alpha_j})^2\,dx + C.
\end{aligned}$$

for an appropriate constant C. Gathering like terms and picking ε small enough such that

$$\frac{\mathrm{a}}{2} \ge b\epsilon$$

we have

$$\|u^{\alpha_j}\|_{H^1(U)} \le C$$

for all j and thus $\{u^j\}_{j\in\mathbb{N}}$ is bounded in $H^1(U)$. □

Consequently, passing to subsequence if necessary, there exist $u^0 \in H^1(U)$ such that $u^j \rightharpoonup u^0$ in $H^1(U)$ by Theorem 2.1.8. Additionally, by Rellich-Kondrachov Theorem, the embedding $H^1(U) \hookrightarrow L^2(U)$ is compact. Thus following Evans [**27**, p.287-289] we have

$$u^j \to u^0 \text{ in } L^2(U). \tag{4.1.9}$$

Now by (4.1.3), we have

$$\left(\alpha_j u_x^j\right)_x = \alpha_j u^j \quad (j = 1, 2, \ldots) \tag{4.1.10}$$

and thus the sequence $\left\{(\alpha_j u_x^j)_x\right\}_{j\in\mathbb{N}}$ is bounded in $L^2(U)$ since $\alpha_j u^j$ is bounded in $L^2(U)$. Hence after passing to subsequence if necessary, there exists $v \in L^2(U)$ such that

$$\alpha_j u_x^j \to v \text{ in } L^2(U) \tag{4.1.11}$$

and by (4.1.10)

$$v_x = \alpha_0 u^0, \tag{4.1.12}$$

since

$$\alpha_j u^j \to \alpha_0 u^0 \text{ in } L^2(U).$$

Recall that by (4.1.5), we have

$$\frac{1}{\alpha_j} \overset{*}{\rightharpoonup} \frac{1}{\omega} \text{ in } L^\infty(U) \tag{4.1.13}$$

and thus by (4.1.11) we have

$$u_x^j = \frac{1}{\alpha_j}\left(\alpha_j u_x^j\right) \rightharpoonup \frac{1}{\omega} v \text{ in } L^2(U).$$

On the other hand, by (4.1.9) we have

$$u_x^j \rightharpoonup u_x^0 \text{ in } H^{-1}(U).$$

Combining these results we have

$$u_x^0 = \frac{1}{\omega} v \;\Rightarrow\; v = \omega u_x^0.$$

Hence by (4.1.12) we have the following ODE

$$\begin{cases} -\left(\omega u_x^0\right)_x + \alpha_0 u^0 = 0 \quad \text{in } U \\ u^0(0) = 1, u^0(1) = 2, \end{cases}$$

where α_0 and ω are given by (4.1.5) and the boundary conditions follow from (4.1.9). A simple calculation shows that the unique solution to this ODE is given by

$$u^0(x) = 1 + x^2$$

and thus by (4.1.4), we have

$$I[\alpha_0] = \int_U \left|u^0 - \left(1 + x^2\right)\right|^2 dx = 0.$$

Now we show that no such $\alpha_0 \in \mathcal{A}$ exists. Therefore assume there exist $\alpha \in \mathcal{A}$ such that $I[\alpha] = 0$ which implies

$$u^\alpha(x) = (1 + x)^2 \quad (x \in U).$$

Additionally, the state-control pair (α, u^α) is feasible, hence we

$$\begin{aligned} 0 &= -\left(\alpha u_x^\alpha\right)_x + \alpha u^\alpha \\ &= -2\left(\alpha x\right)_x + \alpha(1 + x)^2. \end{aligned} \tag{4.1.14}$$

A simple calculations show that the solution to this ODE is

$$\alpha(x) = C x^{-1/2} \exp\left(\frac{x^2}{4}\right) \quad (x \in U),$$

where C is a constant. However α^0 is not bounded from below on U and Hence it is not admissible which is a contradiction. Therefore no optimal solution exists. □

4.2. The existence of optimal solutions

We now show that it is possible to show existence of optimal solutions coefficient control problems under two different assumptions. In the first, we consider problems in which the coefficients of

the differential operator are Lipschitz continuous with Lipschitz constant bounded by a finite real constant. In the second, we consider problems in which the admissible set is closed under *H-convergence*, a property first proposed by Murat and Tartar in connection with homogenization theory.

4.2.1. Existence under the assumption of Lipschitz continuous coefficients. Given a coefficient control problem of the form Problem G, we now assume that the coefficients of the differential operator are Lipschitz continuous and whose Lipschitz constant is bounded by a given finite constant. We note that this special class of problems is important given our original motivation of estimating the coefficients of Itô diffusions which are, in fact, Lipschitz continuous. Establishing the existence of solutions to such problems relies on the Arzela-Ascoli Theorem which provides a necessary property of a convergent minimizing subsequence.

NOTATION. Given a function $f : V \to Y$, we denote the Lipschitz constant, *i.e.* the smallest constant M such that

$$\|f(x) - f(y)\|_Y \leq M \|x - y\|_X \quad (\text{for all } x, y \in V),$$

as $\text{Lip}[f]$ and say that f is M-Lipschitz continuous. Additionally, we denote the space of all real-valued Lipschitz continuous functions on $V \subseteq X$ as $C^{0,1}(V)$. Lastly, for each finite M, we define the set $C^{0,1}_M(V)$ as

$$C^{0,1}_M(V) := \left\{ f \in C^{0,1}(V) \mid \text{Lip}[f] \leq M \right\}.$$

□

For this subsection, assume X and Y are two normed linear spaces and $V \subseteq X$ unless otherwise noted. Additionally, we assume that the admissible set is given by

$$\mathcal{A} := \left\{ \left(\left(a^{ij} \right) \right) \in C^{0,1}_M(U, \mathbb{M}^n) \mid \underline{a} |\xi|^2 \leq \sum_{i,j=1}^{n} a^{ij} \xi_i \xi_j \leq \overline{a} |\xi|^2 \text{ for all } x \in U \right\},$$

where $\underline{a}$ and $\overline{a}$ given real constants such that $0 < \underline{a} < \overline{a} < \infty$.

THEOREM 4.2.1 (The Arzela-Ascoli Theorem). *Let (X, d_X) be a compact metric space. If $\{f_j\}_{j \in \mathbb{N}}$ is a family of real-valued functions on X that are equicontinuous at each $x \in X$ and pointwise bounded on X, then $\{f_j\}_{j \in \mathbb{N}}$ is* (i) *uniformly equicontinuous and uniformly bounded and* (ii) *it has a uniformly convergent subsequence.*

PROOF. *See* Knapp ([**52**, p.121]). □

The following theorem follow directly from the definitions.

THEOREM 4.2.2. (i) *Let $V \subseteq X$ be bounded and suppose that $\{f_j\}_{j \in \mathbb{N}}$ is a family of Lipschitz continuous functions defined on V. Then $\{f_j\}_{j \in \mathbb{N}}$ is pointwise bounded on V and is equicontinuous at each $x \in V$.*

(ii) *Let $\{f_j\}_{j \in \mathbb{N}} \subset C^{0,1}_M(V)$ for a given M such that they are uniformly Lipschitz continuous on $\mathbb{R}^n$. Further, assume they converge uniformly to $f : \mathbb{R}^n \to \mathbb{R}$. Then f is M-Lipschitz continuous.*

Given the above results, we can now apply the Arzela-Ascoli Theorem to identify a convergent subsequence of a family of uniformly Lipschitz continuous functions in $C^{0,1}_M(V)$ for a given M.

THEOREM 4.2.3. *Let $\{f_j\}_{j\in\mathbb{N}} \subset C_M^{0,1}(V)$ be a uniformly Lipschitz continuous family of real-valued functions on an open, bounded set $V \subset \mathbb{R}^n$. Then there exists a M-Lipschitz continuous function $f : \bar{V} \to \mathbb{R}$ such that a subsequence $\{f_{j_k}\}_{k\in\mathbb{N}} \subset \{f_j\}_{j\in\mathbb{N}}$ such that $\{f_{j_k}\}_{k\in\mathbb{N}} \to f$ uniformly.*

Given the above theorem, we now are ready to show the existence of optimal solutions beginning with the next result.

THEOREM 4.2.4. *Assume that $\left\{\left(\left(a_k^{ij}\right)\right)\right\}_{k\in\mathbb{N}} \subset \mathcal{A}$ satisfies*

$$a_k^{ij} \to a_0^{ij} \text{ uniformly for } i,j = 1,\dots,n.$$

For each k, let $u^k \in H_0^1(U)$ be the weak solution to the PDE

$$\begin{cases} -\sum_{i,j=1}^{n} \left(a_k^{ij} u_{x_i}\right)_{x_j} = f & \text{in } U \\ u = 0 & \text{on } \partial U. \end{cases} \tag{4.2.1}$$

Then

$$u^k \to u^0 \text{ strongly in } H_0^1(U),$$

where $u^0 \in H_0^1(U)$ that is the weak solution to the PDE

$$\begin{cases} -\sum_{i,j=1}^{n} \left(a_0^{ij} u_{x_i}\right)_{x_j} = f & \text{in } U \\ u = 0 & \text{on } \partial U. \end{cases} \tag{4.2.2}$$

PROOF. Assume the setup above and fix $((a^{ij})) \in \mathcal{A}$. Any Lipschitz continuous function is uniformly continuous and thus bounded. Consequently, $a^{ij} \in L^\infty(U)$ for $i,j = 1,\dots,n$ and thus we can apply the Lax-Milgram Theorem to show there exists $u \in H_0^1(U)$ such that u is the unique weak solution to (4.2.2) with coefficients $((a^{ij}))$.

Consequently, for any $v \in H_0^1(U)$ we have

$$\int_U \sum_{i,j=1}^{n} a_0^{ij} u_{x_i}^0 v_{x_j} dx = (f, v)$$

and

$$\int_U \sum_{i,j=1}^{n} a_k^{ij} u_{x_i}^k v_{x_j} dx = (f, v)$$

for each k. Fixing both $v \in H_0^1(U)$ and $k \in \mathbb{N}$ and subtracting the latter equation above from the former we get

$$\int_U \sum_{i,j=1}^{n} \left(a_k^{ij} u_{x_i}^k - a_0^{ij} u_{x_i}\right) v_{x_j} dx = 0.$$

This implies

$$0 = \int_U \sum_{i,j=1}^{n} \left(\left(a_k^{ij} u_{x_i}^k - a_0^{ij} u_{x_i}^k\right) + \left(a_0^{ij} u_{x_i}^k - a_0^{ij} u_{x_i}\right)\right) v_{x_j} dx$$

and thus after rearranging terms, we have

$$\int_U \sum_{i,j=1}^n a_0^{ij} \left(u^k - u\right)_{x_i} v_{x_j} dx_{x_j} = \int_U \sum_{i,j=1}^n \left(a_0^{ij} - a_k^{ij}\right) u_{x_i}^k v_{x_j} dx.$$

Setting $v = u^k - u$, we have

$$\int_U \sum_{i,j=1}^n a_0^{ij} \left(u^k - u\right)_{x_i} \left(u^k - u\right)_{x_j} dx = \int_U \sum_{i,j=1}^n \left(a_0^{ij} - a_k^{ij}\right) u_{x_i}^k \left(u^k - u\right)_{x_j} dx. \tag{4.2.3}$$

Now by the uniform ellipticity condition,

$$\sum_{i,j=1}^n a_k^{ij} \xi_i \xi_j \geq \underline{a} \left|\xi\right|^2 > 0 \quad \text{(a.e. } x \in U)$$

for all nonzero $\xi \in \mathbb{R}^n$. Therefore we have

$$\sum_{i,j=1}^n a_0^{ij} \left(u^k - u^0\right)_{x_i} \left(u^k - u^0\right)_{x_j} \geq \underline{a} \left|D\left(u^k - u^0\right)\right|^2. \tag{4.2.4}$$

On the other hand, by the Poincaré inequality (*cf.* Evans [**27**, p.280]), we have

$$\left\|u^k - u^0\right\|^2 \leq C \int_U \left|D\left(u^k - u^0\right)\right|^2 dx.$$

for some positive constant C and thus by (4.2.4), we have

$$\left\|u^k - u^0\right\|_{H_0^1(U)}^2 \leq C \int_U \sum_{i,j=1}^n a_0^{ij} \left(u^k - u^0\right)_{x_i} \left(u^k - u^0\right)_{x_j} dx. \tag{4.2.5}$$

for a possibly different constant C. Hence by (4.2.3)

$$\left\|u^k - u^0\right\|_{H_0^1(U)}^2 \leq C \int_U \sum_{i,j=1}^n \left(a_0^{ij} - a_k^{ij}\right) u_{x_i}^k \left(u^k - u^0\right)_{x_j} dx.$$

Therefore, using Cauchy's inequality we have

$$\begin{aligned} \lim_{k\to\infty} \left\|u^k - u^0\right\|_{H_0^1(U)}^2 &\leq C \lim_{k\to\infty} \int_U \sum_{i,j=1}^n \left(a_0^{ij} - a_k^{ij}\right) u_{x_i}^k \left(u^k - u^0\right)_{x_j} dx \\ &\leq C \lim_{k\to\infty} \max_{x\in U} \left|a_0^{ij} - a_k^{ij}\right| \left\|u^k - u^0\right\|_{H_0^1(U)}^2 \end{aligned}$$

and thus

$$u^k \to u^0 \text{ strongly in } H_0^1\left(U\right).$$

□

We now state the principal existence result of this section.

THEOREM 4.2.5. *An optimal control problem of the form Problem G admits a solution.*

PROOF. Recall that such a problem is norm-minimizing and thus, we have

$$\inf_{\alpha\in\mathcal{A}} I\left[u^{\alpha},\alpha\right]=M\geq 0$$

for some positive constant M. Therefore, let $\left\{\left(\left(\alpha_k^{ij}\right)\right)\right\}_{k\in\mathbb{N}}\subset\mathcal{A}$ be a minimizing sequence. We begin showing that for each $i,j=1,\ldots,n$, there exists α_0^{ij} such that $\alpha_k^{ij}\to\alpha_0^{ij}\in C_M^{0,1}(U)$. Therefore, fix i and j. By assumption $\left\{\alpha_k^{ij}\right\}_{k\in\mathbb{N}}\subset C_M^{0,1}(U)$ for any k. Hence by Theorem 4.2.3 $\left\{\alpha_k^{ij}\right\}_{k\in\mathbb{N}}$ is uniformly Lipschitz continuous and thus moving to a subsequence if necessary, $\left\{\alpha_k^{ij}\right\}_{k\in\mathbb{N}}$ converges uniformly to some α_0^{ij} such that $\operatorname{Lip}\left[\alpha_0^{ij}\right]=M$. Hence $\alpha_0^{ij}\in C_M^{0,1}(U)$. Additionally, since

$$\underline{a}\left|\xi\right|^2\leq\sum_{i,j=1}^{n}\alpha^{ij}(x)\,\xi_i\xi_j\leq\overline{a}\left|\xi\right|^2\quad(\text{for all } x\in U)$$

for all k, in the limit we have

$$\underline{a}\left|\xi\right|^2\leq\sum_{i,j=1}^{n}\alpha_0^{ij}(x)\,\xi_i\xi_j\leq\overline{a}\left|\xi\right|^2\quad(\text{for all } x\in U)$$

and thus we see that $\left(\left(\alpha_0^{ij}\right)\right)\in\mathcal{A}$. Consequently, we can use the direct method as set forth in the proof to Theorem 2.1.12 to show that in fact

$$I\left[u^0,\alpha^0\right]=M,$$

where $u^0=u^{\alpha^0}\in H_0^1(U)$ that is the unique weak solution to state equation associated with a minimizer $\alpha^0=\left(\left(\alpha_0^{ij}\right)\right)$.

Since $a_k^{ij}\to a_0^{ij}$ uniformly as $k\to\infty$ for $i,j=1,\ldots,n$, we can apply by Theorem 4.2.4 to show that there is $u^k\in H_0^1(U)$ that is the unique weak solution to state equation associated with $\left(\left(\alpha_k^{ij}\right)\right)$ for each k. Moreover, $u^k\to u^0$ in $H_0^1(U)$. Hence u^0 is feasible and thus (u^0,α^0) is an optimal state-control pair. □

4.2.2. Existence under the assumption of H-convergence. We now show the existence of optimal solutions to coefficient control problems of the form Problem G in which the admissible set is closed under the property, H-convergence. This is based on work by Murat and Tartar that generalized of earlier results by Spagnolo in connection with *homogenization*, a field of PDE theory associated with the study of homogeneous materials comprising a composite material. From our perspective, this work is important because we can use H-convergence to show that a minimizing sequence of admissible controls will converge to an admissible control and critically, in limit the associated state will be a solution to the proper PDE constraint.

DEFINITION 4.2.6. For real constants a and b define the set of admissible matrices as

$$\begin{aligned}\mathcal{M}_{\mathrm{a,b}}:=\big\{A\in\mathbb{M}^n\mid{}&\xi^TA\xi\geq\mathrm{a}\left|\xi\right|^2>0\text{ and}\\&\text{and }\xi^TA^{-1}\xi\geq\mathrm{b}\left|\xi\right|^2>0\text{ for all nonzero }\xi\in\mathbb{R}^n\big\}\end{aligned}\tag{4.2.6}$$

LEMMA 4.2.7. *Assume that $A\in\mathcal{M}_{\mathrm{a,b}}$. Then we have*

$$\mathrm{a}\left|\xi\right|^2\leq\xi^TA\xi\leq\frac{1}{\mathrm{b}}\left|\xi\right|^2$$

for any $\xi \in \mathbb{R}^n$. *Further,* $\mathcal{M}_{\mathrm{a,b}}$ *is nonempty if and only if* $\mathrm{ab} \leq 1$.

PROOF. Assume that $A \in \mathcal{M}_{\mathrm{a,b}}$. Select $\xi \in \mathbb{R}^n$ and define $\eta \in \mathbb{R}^n$ as

$$\eta := A\xi. \tag{4.2.7}$$

Then by the definition of $\mathcal{M}_{\mathrm{a,b}}$, we have

$$\mathrm{b}\left|\eta\right|^2 \leq \eta^T A^{-1}\eta = \left(\eta, A^{-1}\eta\right).$$

Thus by (4.2.7), we get

$$\mathrm{b}\left|A\xi\right|^2 \leq \left|\left(A\xi, A^{-1}\left(A\xi\right)\right)\right| \leq \left|A\xi\right|\left|\xi\right| \tag{4.2.8}$$

Thus after rearranging terms, we have

$$\left|A\xi\right| \leq \frac{1}{\mathrm{b}}\left|\xi\right|.$$

Multiplying both sides of the above equation by $|\xi|$ and applying the Cauchy-Schwartz inequality once again, we have

$$\frac{1}{\mathrm{b}}\left|\xi\right|^2 \geq \left|A\xi\right|\left|\xi\right| \geq \left|\left(A\xi,\xi\right)\right| \geq \xi^T A\xi.$$

On the other hand, since $A \in \mathcal{M}_{\mathrm{a,b}}$, we also have

$$\mathrm{a}\left|\xi\right|^2 \leq \xi^T A\xi$$

and thus we have

$$\mathrm{a}\left|\xi\right|^2 \leq \xi^T A\xi \leq \frac{1}{\mathrm{b}}\left|\xi\right|^2$$

which is true for any $\xi \in \mathbb{R}^n$ for suitable constants a and b. Additionally, the above expression implies

$$\mathrm{a}\left|\eta\right|^2 \leq \frac{1}{\mathrm{b}}\left|\eta\right|^2 \;\Rightarrow\; \mathrm{ab} \leq 1.$$

Consequently, $\mathcal{M}_{\mathrm{a,b}}$ is nonempty if and only if $\mathrm{ab} \leq 1$. □

LEMMA 4.2.8. *Assume that* $\alpha \in L^\infty\left(U, \mathcal{M}_{\mathrm{a,b}}\right)$. *Then there exists* $u \in H_0^1\left(U\right)$ *such that* u *is the unique solution PDE*

$$\begin{cases} -\mathrm{div}\left(\alpha Du\right) = f & \text{in } U \\ u = 0 & \text{on } \partial U, \end{cases} \tag{4.2.9}$$

where $f \in L^2(U)$ *is given.*

PROOF. By assumption,

$$\alpha\left(x\right) = \left(\left(\alpha^{ij}\left(x\right)\right)\right) \in \mathcal{M}_{\mathrm{a,b}} \quad \left(\text{a.e. } x \in U\right)$$

and thus for any $\xi \in \mathbb{R}^n$

$$0 < \mathrm{a}\left|\xi\right|^2 \leq \xi^T \alpha\left(x\right)\xi \quad \left(\text{a.e. } x \in U\right).$$

Consequently, (4.2.9) is uniformly elliptic and $\alpha^{ij}\left(x\right) \in L^\infty\left(U\right)$ for $i, j = 1, \ldots, n$. Then by the Lax-Milgram Theorem, there exists $u \in H_0^1\left(U\right)$ which is a unique solution to (4.2.9). □

Given the above result, the following definition is meaningful.

DEFINITION. We say that a sequence $\{\alpha^j\}_{j\in\mathbb{N}} \subset L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$ ***H-converges*** to $\alpha^0 \in L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$, denoted as

$$\alpha^j \overset{H}{\rightharpoonup} \alpha^0,$$

if and only if for any $f \in L^2(U)$, the sequence of solutions $u^j \in H_0^1(U)$ of the PDE

$$\begin{cases} -\operatorname{div}\left(\alpha^j Du^j\right) = f & \text{in } U \\ u^j = 0 & \text{on } \partial U, \end{cases}$$

exists and satisfies

$$\begin{cases} u^j \rightharpoonup u^0 \text{ in } H_0^1(U) \\ \alpha^j Du^j \rightharpoonup \alpha^0 Du^0 \text{ in } \left(L^2(U)\right)^n, \end{cases}$$

where $u^0 \in H_0^1(U)$ is the solution to the PDE

$$\begin{cases} -\operatorname{div}\left(\alpha^0 Du^0\right) = f & \text{in } U \\ u^0 = 0 & \text{on } \partial U. \end{cases}$$

The following is a well-known result due to Murat and Tartar [**67**, Section 9].

THEOREM 4.2.9 (Sequential compactness under H-convergence). *For every sequence $\{\alpha^j\}_{j\in\mathbb{N}} \subset L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$, there exists a subsequence (which we also denote as $\{\alpha^j\}$) such that*

$$\alpha^j \overset{H}{\rightharpoonup} \alpha^0,$$

where $\alpha^0 \in L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$.

The next result follows immediately by applying the above theorem to a minimizing sequence to prove the existence of an solution to our canonical optimal control problem.

THEOREM 4.2.10. *Assume that we given an optimal control problem of the form Problem G, where the admissible set is $\mathcal{A} := L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$. Then it admits a solution.*

PROOF. Let $\mathcal{A} := L^\infty(U, \mathcal{M}_{\mathrm{a,b}})$ be the admissible set and again recall that given our objective functional as defined at (3.0.6), we have

$$\inf_{\alpha\in\mathcal{A}} I[u^\alpha, \alpha] = M \geq 0,$$

for a positive constant M. Therefore, let $\{\alpha^j\}_{j\in\mathbb{N}} \subset \mathcal{A}$ be a minimizing sequence. Then by Lemma 4.2.8, there exists $u^j \in H_0^1(U)$ which is the unique solution to the PDE

$$\begin{cases} -\operatorname{div}\left(\alpha^j Du\right) = f & \text{in } U \\ u = 0 & \text{on } \partial U, \end{cases}$$

for each j. Additionally, by Theorem 4.2.9, there exists $\alpha^0 \in \mathcal{A}$ such that, passing to a subsequence if necessary,

$$\alpha^j \overset{H}{\rightharpoonup} \alpha^0$$

and thus

$$u^j \rightharpoonup u^0 \text{ weakly in } H_0^1(U), \tag{4.2.10}$$

where $u^0 \in H_0^1(U)$ is the solution to the PDE

$$\begin{cases} -\text{div}\left(\alpha^0 D u^0\right) = f & \text{in } U \\ u^0 = 0 & \text{on } \partial U. \end{cases}$$

Since $H_0^1(U)$ is compactly embedded in $L^2(U)$ by the Rellich-Kondrachov Theorem, $u^j \to u^0$ in $L^2(U)$ and thus u^0 is a feasible solution. Additionally, we can apply the direct method once again (*cf.* Theorem 2.1.12) to show that

$$I\left[u^0, \alpha^0\right] = M,$$

and thus (u^0, α^0) is an optimal state-control pair. □

4.2.2.1. *Behavior of optimal solutions - a one dimensional example.* We now explore the behavior of optimal solutions to a special class one dimensional coefficient control problems under the assumption that the control does not hit the boundary of the admissible set over some open interval. Here, we assume that the constraint is an ODE of the form:

$$\begin{cases} -\left(\alpha(x) u_x\right)_x = f & \text{in } U = (0, L) \\ u(0) = C_1, u(L) = C_2, \end{cases} \tag{4.2.11}$$

where $f \in C(0,1)$ is a given and C_1 and C_2 are given real constants. The control α taken from the admissible set $\mathcal{A}$ defined as

$$\mathcal{A} := \left\{\alpha \in C^1(0, L) \mid \underline{\alpha} \leq \alpha(x) \leq \overline{\alpha} \text{ for } x \in (0, L)\right\}, \tag{4.2.12}$$

where $\underline{\alpha}$ and $\overline{\alpha}$ are given real constants such that $0 < \underline{\alpha} \leq \overline{\alpha} < \infty$.

Given such a constraint, we are specifically concerned with optimal control problems whose reduced form is given by:

Problem 3.

$$\begin{cases} \min\limits_{a \in \tilde{\mathcal{A}}} & I[\alpha], \\ \text{subject to} & u^\alpha \in \mathcal{F}_f, \end{cases}$$

where the objective functional $I : \mathcal{A} \to \mathbb{R}$ is given by

$$I[\alpha] = \frac{1}{2} \|u^\alpha - \hat{u}\|^2 \quad (\alpha \in \mathcal{A}), \tag{4.2.13}$$

such that $\hat{u} \in L^2(U)$ is the given target function and $\mathcal{F}_f$ is the set of feasible states defined as

$$\mathcal{F}_f := \left\{v^\alpha \in C^2(0, L) \mid v^\alpha \text{ is a solution to (4.2.11) for } \alpha \in \mathcal{A}\right\}.$$

Existence of feasible solutions

Given the set up above, we know from Section 4.2.2 that optimal solutions to the unconstrained problem exists. We now show that a feasible solution exists for each control $\alpha \in \mathcal{A}$. We begin with the following lemma from standard ODE theory.

LEMMA 4.2.11. *Assume that we are given a first-order ODE of the form*

$$v_x + fv = g, \tag{4.2.14}$$

such that $f, g \in C(0, L)$. *Then for any* $x_0 \in [0, L]$ *and any real constant* K, (4.2.14) *has a solution* $v \in C^1(0, L)$ *given by*

$$v(x) = \exp(-F(x)) \left(\int_{x_0}^{x} \exp(F(y)) g(y) \, dy + K \right) \quad (x_0 \leq x \leq L), \tag{4.2.15}$$

where

$$F(x) = \int_{x_0}^{x} f(y) \, dy.$$

PROOF. Assume that we are given such an ODE of the form (4.2.14), where $F, g \in C(0, L)$. First note that since f is continuous, by the fundamental theorem of calculus we can define $F : (0, L) \to \mathbb{R}$ as

$$F(x) = \int_{x_0}^{x} f(y) \, dy$$

for any fixed $x_0 \in [0, L]$. Noting that

$$F_x(x) = f(x) \quad (x \in [0, L]), \tag{4.2.16}$$

differentiating both sides of (4.2.15) gives

$$\begin{aligned} v_x &= -F_x \exp(-F) \left(\int_{x_0}^{x} \exp(F) g dy + K \right) + \exp(-F) \exp(F) g \\ &= -fv + g \end{aligned}$$

by (4.2.16) and (4.2.15). □

We now use the above lemma to show that (4.2.11) has a unique solution for each $\alpha \in \mathcal{A}$.

THEOREM 4.2.12. *Assume that we are given an ODE of the form* (4.2.11). *Then it has a unique solution* $u \in C^2(0, L)$ *given by*

$$u(x) = \int_0^x \frac{1}{\alpha(y)} \left(K_\alpha - \int_0^y f(z) \, dz \right) dy + C_1 \quad (x \in (0, L)), \tag{4.2.17}$$

where

$$K_\alpha = \left(\int_0^L \frac{1}{\alpha(y)} dy \right)^{-1} \left((C_2 - C_1) + \int_0^L \frac{1}{\alpha(y)} \left(\int_0^y f(z) \, dz \right) dy \right).$$

PROOF. Assume the set up above. Expanding (4.2.11), we have

$$f = -(\alpha u_x)_x = -\alpha_x u_x - \alpha u_{xx}, \tag{4.2.18}$$

where $\alpha \in \mathcal{A}$. Noting that $\alpha > 0$ for any $\alpha \in \mathcal{A}$ and letting $v = u_x$, we can rewrite this equation as the first-order ODE

$$-v_x - \frac{\alpha_x}{\alpha} v = \frac{f}{\alpha} \quad (\alpha \in \mathcal{A}). \tag{4.2.19}$$

By assumption, f is continuous and $\alpha_x \in C(0,L)$ and thus we can apply Lemma 4.2.11. Therefore fix $\alpha \in \mathcal{A}$ and after setting $x_0 = 0$, the solution to (4.2.19) is given by

$$v(x) = -\exp(-A)\left(\int_0^x \exp(A)\,\tilde{f}dy + K\right) \quad (x \in (0,L)) \tag{4.2.20}$$

for any constant K, where

$$A(x) = \int_0^x \frac{\alpha_y(y)}{\alpha(y)}dy = \log\alpha(x) - \log\alpha(0) \text{ and } \tilde{f}(x) = \frac{f(x)}{\alpha(x)}$$

for any $x \in (0,L)$. Noting that

$$\exp(A(x)) = \frac{\alpha(x)}{\log\alpha(0)} \text{ and } \exp(-A(x)) = \frac{\log\alpha(0)}{\alpha(x)}$$

for any $x \in (0,L)$ and substituting these identities into (4.2.20), we have

$$\begin{aligned} v(x) &= -\exp(-A(x))\left(\int_0^x \exp(A(y))\,\tilde{f}(y)\,dy + K_\alpha\right) \\ &= -\frac{\log\alpha(0)}{\alpha(x)}\left(\int_0^x \left(\frac{\alpha(y)}{\log\alpha(0)}\right)\left(\frac{f(y)}{\alpha(y)}\right)dy + K_\alpha\right) \\ &= -\frac{1}{\alpha(x)}\left(\int_0^x f(y)\,dy + K_\alpha\right), \end{aligned}$$

for any $x \in U$, where K_α is any real constant for each choice of $\alpha \in \mathcal{A}$.

Recalling $v = u_x$, after integrating both sides of the above equation, we have

$$u(x) = -\int_0^x \frac{1}{\alpha(y)}\left(\int_0^y f(z)\,dz + K_\alpha\right)dy + K \quad (x \in U), \tag{4.2.21}$$

where K is any real constant. We can determine the constants K_α and K using the boundary conditions,

$$u(L) = C_2 \text{ and } u(0) = C_1.$$

Then by (4.2.21), we have

$$K = u(0) = C_1$$

and thus

$$-\int_0^L \frac{1}{\alpha(y)}\left(\int_0^y f(z)\,dz + K_\alpha\right)dy + C_1 = C_2.$$

Rearranging terms, we have

$$C_2 - C_1 = -\int_0^L \frac{1}{\alpha(y)}\left(\int_0^y f(z)\,dz\right)dy - K_\alpha\int_0^L \frac{1}{\alpha(y)}dy.$$

Rearranging terms once more, we get

$$K_\alpha = \left(\int_0^L \frac{1}{\alpha(y)}dy\right)^{-1}\left((C_1 - C_2) - \int_0^L \frac{1}{\alpha(y)}\left(\int_0^y f(z)\,dz\right)dy\right)$$

and thus u defined as (4.2.21) is the unique solution to (4.2.11). □

REMARK. Given the fact that a unique state exists for each admissible control, there exists a bijective control-to-state operator $S : \mathcal{A} \to C^2(0, L)$ defined as

$$S(\alpha) = u^\alpha \quad (a \in \mathcal{A}).$$

Hence we can express any optimal control problem with an ODE constraint of the form 4.2.11 in reduced form given by Problem 3 without loss of generality. □

Behavior of the control

Using the concrete solution to ODE constraints above, we now examine the behavior of a solution to an optimal control problem of the form of Problem 3 under the assumption that that the control does not hit the boundary of the admissible set over some given open interval. To facilitate the discussion, we slightly restate the optimal control problem under consideration. Specifically, define $F : (0, L) \to \mathbb{R}$ as

$$F(x) := \int_0^x f(y)\, dy \quad (x \in (0, L)),$$

where $f \in C(0, L)$ is the given nonhomogeneous term and define $a : (0, L) \to \mathbb{R}$ as

$$a(x) := \frac{1}{\alpha(x)} \quad (x \in (0, L))$$

for each $\alpha \in \mathcal{A}$ noting that $\alpha > 0$ on $(0, L)$. Additionally define the set $\tilde{\mathcal{A}}$ as

$$\tilde{\mathcal{A}} := \left\{ a \mid a = \frac{1}{\alpha} \text{ for each } \alpha \in \mathcal{A} \right\}$$

and define the constants $\underline{a}$ and $\overline{a}$ as

$$\underline{a} = \frac{1}{\overline{\alpha}} \text{ and } \overline{a} = \frac{1}{\underline{\alpha}},$$

where $\underline{\alpha}$ and $\overline{\alpha}$ are the given real bounds on the admissible set $\mathcal{A}$. Here we note that $\tilde{\mathcal{A}}$ can be expressed as

$$\tilde{\mathcal{A}} = \left\{ a \in C^1(0, L) \mid \underline{a} \leq a(x) \leq \overline{a} \text{ for } x \in (0, L) \right\},$$

such that $0 < \underline{a} \leq \overline{a} < \infty$.

Given this set up, we can express the ODE constraint as

$$\begin{cases} -(a(x)\, u_x)_x = f \quad \text{in } (0, L) \\ u(0) = C_1, u(L) = C_2, \end{cases} \tag{4.2.22}$$

where $a \in \tilde{\mathcal{A}}$. Then by Theorem 4.2.12, the solution this ODE for any $a \in \tilde{\mathcal{A}}$ is given by as

$$u^a(x) = \int_0^x a(y)\,(K_a - F(y))\, dy + C_1 \quad (x \in (0, L)), \tag{4.2.23}$$

where

$$K_a = \left(\int_0^L a(y)\, dy \right)^{-1} \left((C_2 - C_1) + \int_0^L a(y)\, F(y)\, dy \right). \tag{4.2.24}$$

Additionally we note that for our analysis, we only need consider the behavior of control (and thus the ODE constraint) over the sub-interval $(A, B) \subset (0, L)$. To ensure is continuous, we set the

boundary conditions of the ODE constraint over this sub-interval as

$$C_A := u^a(A) \text{ and } C_B := u^a(B),$$

where u^a is the solution to 4.2.22 for $a \in \tilde{A}$. In particular, we note that if $(A, B) = (0, L)$, then the boundary conditions are simply

$$C_A = C_1 = u^\alpha(A) \text{ and } C_B = C_2.$$

Therefore without loss of generality, we assume that the boundary conditions in the calculations below are given by

$$C_A = u^\alpha(A) \text{ and } C_B = u^\alpha(B)$$

and we define the real constant $C_\triangle$ as

$$C_\triangle := C_B - C_A.$$

The next result shows that if the control does not hit the boundary of the admissible set over an open interval, the computed state is equal to the target function pointwise over the entire interval.

THEOREM 4.2.13. *Assume we are given a problem of the form of Problem 3 with an ODE constraint of the form* (4.2.22). *Further assume that* $a^0 \in \tilde{\mathcal{A}}$ *is a solution such that*

$$\underline{a} < a^0(x) < \overline{a} \text{ for } x \in (A, B), \tag{4.2.25}$$

where $0 \leq A < B \leq L$. *Then we have*

$$u^0(x) = \hat{u}(x) \quad (x \in (A, B)),$$

where $u^0 = u^{a^0}$ *is the solution to the ODE with the coefficient* a^0 *and* $\hat{u}$ *is the given target function.*

PROOF. Assume that $a^0 \in \tilde{\mathcal{A}}$ is a solution to Problem 3 under the assumptions above and $u^0 \in C^2(0, L)$ is the solution to the ODE constraint given by (4.2.23). Since the admissible set $\mathcal{A}$ is closed, bounded, and convex, so is $\tilde{\mathcal{A}}$ and by (4.2.25), $a^0 \in \operatorname{int}\left(\tilde{A}\right)$. Hence any $\beta \in L^2(U)$ is a feasible direction (*cf.* (A.4.1)) and by (4.2.13), we have

$$\begin{aligned}
0 = \left\langle DI\left[a^0\right], \beta \right\rangle &= \frac{d}{d\tau} I\left[a^0 + \tau\beta\right] |_{\tau=0} \\
&= \frac{1}{2}\frac{d}{d\tau}\int_A^B \left\| u^{a^0+\tau\beta} - \hat{u} \right\|^2 dx \, |_{\tau=0} \\
&= \int_A^B \left(u^{a^0+\tau\beta} - \hat{u}\right) \left\langle D_a u^{a^0+\tau\beta}, \beta \right\rangle dx \, |_{\tau=0}, \\
&= \int_A^B \left(u^{a^0+\tau\beta} - \hat{u}\right) \left\langle D_a u^{a^0+\tau\beta}, \beta \right\rangle dx
\end{aligned} \tag{4.2.26}$$

using the chain rule for the Fréchet derivative (*cf.* (A.3.3)).

Computing $\langle D_a u^0, \beta \rangle$ directly using (4.2.23), we have

$$\begin{aligned}
\left\langle D_a u^0, \beta \right\rangle &= \frac{d}{d\tau} u^{a^0+\tau\beta} |_{\tau=0} \\
&= \frac{d}{d\tau}\left[\int_A^x \left(a^0 + \tau\beta\right)\left(K_{a^0+\tau\beta} - F\right) dy + C_A\right] |_{\tau=0} \quad (x \in U)
\end{aligned}$$

$$= \left(K_{a^0} \int_A^x \beta dy - \int_A^x \beta F dy \right) + \langle DK_{a^0}, \beta \rangle \int_A^x a^0 dy. \tag{4.2.27}$$

Substituting this into (4.2.26), we have

$$\begin{aligned} 0 &= \int_A^B \left(u^0 - \hat{u}\right) \left\langle Du^0, \beta \right\rangle dx \\ &= \int_A^B \left(u^0 - \hat{u}\right) \left[K_{a^0} \int_A^x \beta dy - \int_A^x \beta F dy + \langle DK_{a^0}, \beta \rangle \int_A^x a^0 dy \right] dx. \end{aligned} \tag{4.2.28}$$

Now this is true for any $\beta \in L^2(U)$. Therefore, fix $x_0 \in U$ and chose positive constants ε_j such that

$$\lim_{j \to \infty} \varepsilon_j = 0$$

and define

$$\beta_j(x) := \mu_{\varepsilon_j}\left(x - x^0\right) \quad (x \in \mathbb{R})$$

for each $j \in \mathbb{N}$, where

$$\mu_{\varepsilon_j}(x) := \varepsilon^{-1} \mu\left(\frac{x}{\varepsilon_j}\right) \quad (x \in \mathbb{R}).$$

Here μ is the standard mollifier, defined as

$$\mu(x) := \begin{cases} C \exp\left(\frac{1}{|x|^2 - 1}\right) & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1, \end{cases}$$

where the real constant C is chosen such that

$$\int_{\mathbb{R}} \mu(x)\, dx = 1.$$

Then following Evans (*cf.* [**27**, p.713-714]), we know that for any $\varepsilon > 0$,

(a) $\mu, \mu_\varepsilon \in C^\infty(\mathbb{R})$.

(b) $\operatorname{spt}(\mu_\varepsilon) \subset B(0, \varepsilon)$.

(c) $\int_{\mathbb{R}} \mu_\varepsilon(x)\, dx = 1$.

Further, given any locally integrable function $f : U \to \mathbb{R}$, we know that for any $\varepsilon > 0$, the function f^ε, known as its mollification and defined as

$$f^\varepsilon(x) := \int_U \mu_\varepsilon(y - x) f(y)\, dy = \int_{B(0,\varepsilon)} \mu_\varepsilon(x) f(y - x)\, dy \quad (x \in U_\varepsilon), \tag{4.2.29}$$

where

$$U_\varepsilon := \{x \in U \mid \operatorname{dist}(x, \partial U) < \varepsilon\},$$

has the following properties:

(a) $f^\varepsilon \in C^\infty(U_\varepsilon)$.

(b) $f^\varepsilon \to f$ a.e. as $\varepsilon \to 0$.

(c) if $f \in C(U)$, then $f^\varepsilon \to f$ uniformly on compact subsets of U.

(d) if $f \in L^p_{\text{loc}}(U)$, then $f^\varepsilon \to f$ in $L^p_{\text{loc}}(U)$ for $1 \leq p < \infty$.

Moreover, by (4.2.29) we know

$$\lim_{\varepsilon \to 0} \mu_\varepsilon(x) = \delta_0(x) \text{ a.e.},$$

where δ_0 is the Dirac measure giving unit mass at 0 defined by

$$f(0) = \int_U f(y)\, d\delta_0(y)$$

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

Given this, we have

$$\lim_{j \to \infty} \beta_j(x) = \lim_{\varepsilon \to 0} \mu_\varepsilon(x - x_0) = \delta_0(x - x_0) = \delta_{x_0}(x),$$

where δ_{x^0} is the Dirac measure giving unit mass at x_0.

Then by property (c) above, we can pass the limit which results in the following identities:

(a) $\lim_{j \to \infty} \int_A^x \beta_j dy = \begin{cases} 0, & x < x_0 \\ 1, & x_0 \geq x \end{cases} =: H(x - x_0)$ (*i.e.*, the Heavyside function).

(b) $\lim_{j \to \infty} \int_A^B \beta_j dy = 1$.

(c) $\lim_{j \to \infty} \int_A^x \beta_j F dy = \begin{cases} 0, & x < x_0 \\ F(x_0), & x_0 \geq x \end{cases} = F(x_0) H(x - x_0)$.

(d) $\lim_{j \to \infty} \int_A^B \beta_j F dy = F(x_0)$.

Substituting these into (4.2.28) and passing the limit using property (c) above, we have

$$\begin{aligned} 0 &= \lim_{j \to \infty} \int_A^B \left(u^0 - \hat{u}\right) \left[K_{a^0} \int_A^x \beta_j dy - \int_A^x \beta_j F dy + \langle DK_{a^0}, \beta_j \rangle \int_A^x a^0 dy \right] dx \\ &= \int_A^B \left(u^0 - \hat{u}\right) \left[\left(K_{a^0} - F(x_0)\right) H(x - x_0) + \lim_{j \to \infty} \langle DK_{a^0}, \beta_j \rangle \int_A^x a^0 dy \right] dx \\ &= \left(K_{a^0} - F(x_0)\right) \int_{x_0}^B \left(u^0 - \hat{u}\right) dx + \lim_{j \to \infty} \langle DK_{a^0}, \beta_j \rangle \int_A^B \left(u^0 - \hat{u}\right) \left(\int_A^x a^0 dy \right) dx. \end{aligned} \tag{4.2.30}$$

Now for any β_j, $\beta_j \in L^2(U)$ and thus is a feasible direction and we have

$$\begin{aligned} \langle DK_{a^0}, \beta_j \rangle &= \frac{d}{d\tau} K_{a^0 + \tau \beta_j} \mid_{\tau = 0} \\ &= \frac{d}{d\tau} \left[\left(\int_A^B \left(a^0 + \tau \beta_j\right) dy \right)^{-1} \left(C_\triangle + \int_A^B \left(a^0 + \tau \beta_j\right) F dy \right) \right] \mid_{\tau = 0} \\ &= - \left(\int_A^B a^0 dy \right)^{-2} \int_A^B \beta_j dy \left(C_\triangle + \int_A^B a^0 F dy \right) + \left(\int_A^B a^0 dy \right)^{-1} \int_A^B \beta_j F dy. \\ &= \left(\int_A^B a^0 dy \right)^{-2} \left[\int_A^B a^0 dy \int_A^B \beta_j F dy - \left(C_\triangle \int_A^B \beta_j dy + \int_A^B \beta_j dy \int_A^B a^0 F dy \right) \right]. \end{aligned}$$

Using the identities above, we have

$$\begin{aligned}\lim_{j\to\infty}\langle DK_{a^0},\beta_j\rangle &= \lim_{j\to\infty}\left(\int_A^B a^0dy\right)^{-2}\left[\int_A^B a^0dy\int_A^B \beta_jFdy-\left(C_\triangle\int_A^B\beta_jdy+\int_A^B\beta_jdy\int_A^B a^0Fdy\right)\right]\\ &=\left(\int_A^B a^0dy\right)^{-2}\left[F(x_0)\int_A^B a^0dy-\left(C_\triangle+\int_A^B a^0Fdy\right)\right]\\ &=\left(\int_A^B a^0dy\right)^{-1}\left[F(x_0)-\left(\int_A^B a^0dy\right)^{-1}\left(C_\triangle+\int_A^B a^0Fdy\right)\right]\\ &=\left(\int_A^B a^0dy\right)^{-1}(F(x_0)-K_a)\end{aligned}$$

the last equality following from (4.2.24).

Substituting the above expression into (4.2.30), we have

$$\begin{aligned}0&=(K_{a^0}-F(x_0))\int_{x_0}^B(u^0-\hat u)\,dx+\lim_{j\to\infty}\langle DK_{a^0},\beta_j\rangle\int_A^B(u^0-\hat u)\left(\int_A^x a^0dy\right)dx\\ &=(K_{a^0}-F(x_0))\int_{x_0}^B(u^0-\hat u)\,dx+\left(\left(\int_A^B a^0dy\right)^{-1}(F(x_0)-K_{a^0})\right)\int_A^B(u^0-\hat u)\left(\int_A^x a^0dy\right)dx\end{aligned}$$

(4.2.31)

$$=(K_{a^0}-F(x_0))\left[\int_{x_0}^B(u^0-\hat u)\,dx-\left(\int_A^B a^0dy\right)^{-1}\int_A^B(u^0-\hat u)\left(\int_A^x a^0dy\right)dx\right].$$

Now K_{a^0} is independent of x_0 and since (4.2.31) is true for all $x_0\in U$, $K_{a^0}-F(x_0)$ cannot be identically zero. Thus, we have

$$0=\int_{x_0}^B(u^0-\hat u)\,dx-\left(\int_A^B a^0dy\right)^{-1}\int_A^B(u^0-\hat u)\left(\int_A^x a^0dy\right)dx.$$

Hence we can view the right-hand side of the above equation as a function of x_0. Therefore taking the derivative of both sides with respect to x_0, we have

$$0=\hat u(x_0)-u^0(x_0)\quad(\text{for all } x_0\in(A,B))$$

by the fundamental theorem of calculus and thus $u^{a^0}=\hat u$ pointwise on (A,B). □

4.3. Characterization of optimal solutions

We now characterize solutions to optimal control problems of the form Problem G under the assumption that the coefficients of the differential operator are symmetric. As noted in the previous section, H-convergence is a generalization of an earlier result by Spagnolo (known G-convergence) precisely under the assumption of symmetric elliptic operators (*cf.* [**4**, sec. 1.3.2]). In this case, we define the admissible set under G-convergence as follows:

DEFINITION. For real constants a and b such that $0 < \mathrm{a} \leq \mathrm{b} < \infty$, define the set of admissible matrices under G-convergence as

$$\mathcal{G}_{\mathrm{a,b}} := \left\{ A \in \mathbb{S}^n \mid \xi^T A \xi \geq \mathrm{a} \left|\xi\right|^2 > 0 \text{ and } \right. \\ \left. \text{and } \left|A\xi\right| \leq \mathrm{b} \left|\xi\right| \text{ for all nonzero } \xi \in \mathbb{R}^n \right\}, \tag{4.3.1}$$

where $\mathbb{S}^n$ is the space of symmetric $n \times n$ matrices.

Importantly, we can establish an analogous existence result to that of Theorem thm: op control H existence for coefficient control problems of the form Problem G in which the admissible set is defined as

$$\mathcal{A} := L^\infty \left(U, \mathcal{G}_{\mathrm{a,b}}\right).$$

We label such problems as Problem H.

We also note that if $A \in \mathcal{G}_{\mathrm{a,b}}$, then

$$\mathrm{a} \left|\xi\right|^2 \leq \xi^T A \xi \leq \mathrm{b} \left|\xi\right|^2, \tag{4.3.2}$$

an inequality which we will use extensively in the sequel.

4.3.1. The state and co-state equations. Given a coefficient control problem of the form Problem H, the strong form of the state equation is given by

$$L^\alpha u = f, \tag{4.3.3}$$

where L^α is a uniformly elliptic operator of the form (4.0.1) such that one or more of its coefficients depend on the choice of control $\alpha \in \mathcal{A}$. Consequently, by (3.3.9) the weak form of the state equation is given by

$$B^\alpha \left[u, v\right] = \left(f, v\right) \quad \left(\text{for all } v \in H_0^1 \left(U\right)\right),$$

where B^α is the bilinear form associated with the differential operator L^α for given control $\alpha \in \mathcal{A}$.

Also by (3.3.9), the strong form of the co-state equation is

$$\left(L^\alpha\right)^* p = u^\alpha - \hat{u}$$

for any choice of control $\alpha \in \mathcal{A}$, where $u^\alpha \in H_0^1 \left(U\right)$ and $p \in H_0^1 \left(U\right)$ and are the solutions to the state and co-state equations associated with α, respectively. Since L^α is self-adjoint, it follows that we can state the strong form of the co-state equation as

$$L^\alpha p = u^\alpha - \hat{u} \tag{4.3.4}$$

and thus state its weak form as

$$B^\alpha \left[p, v\right] = \left(u^\alpha - \hat{u}, v\right) \quad \left(v \in H_0^1 \left(U\right)\right) \tag{4.3.5}$$

for any choice of control $\alpha \in \mathcal{A}$.

4.3.2. A maximum principle for coefficient control problems. Rather than stating a variation inequality similar to that presented in Section 3.3.3, we present the following pointwise maximum principle governing optimal control problems of the form of Problem H.

THEOREM 4.3.1 (Maximum principle). *Assume that we are given a problem of the form Problem H. Further, assume that u^0 is a solution and that α^0 and p^0 are the corresponding optimal control policy and co-state state, respectively. Then*

$$\left(Dp^0(x)\right)^T \alpha^0(x) Du^0(x) = \max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{ \left(Dp^0(x)\right)^T ADu^0(x) \right\} \quad (\textit{for a.e. } x \in U).$$

REMARK. This is an analog of the Pontryagin Maximum Principle from standard optimal control theory. □

PROOF. Given such a problem of the form Problem H, assume that u^0 is a solution and $\alpha^0 = \left(\left(\alpha_0^{ij}\right)\right) \in \mathcal{A}$ is the corresponding optimal control policy. Then by (4.3.4), the weak form of the co-state state equation is given

$$B[p,v] = -\int_U \sum_{i,j=1}^n \left(\alpha_0^{ij} p_{x_i}\right) v_{x_j} = \int_u \left(u^0 - \hat{u}\right) v dx \quad \left(\text{for any } v \in H_0^1(U)\right), \tag{4.3.6}$$

where $p \in H_0^1(U)$. Since $\alpha^0 \in \mathcal{A}$ and $u^0 - \hat{u} \in L^2(U)$,the co-state equation satisfies the hypotheses of the Lax-Milgram Theorem and, thus, a unique co-state state $p^0 \in H_0^1(U)$ does indeed exist.

For $i, j = 1, \ldots, n$, defined the function $\tilde{\alpha}^{ij} : [0,1] \to L^\infty(U)$ as

$$\tilde{\alpha}^{ij}(\tau) = \tau a^{ij} + (1-\tau)\alpha_0^{ij} \quad (\tau \in [0,1]) \tag{4.3.7}$$

for some $((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$ and note that

$$\frac{d}{d\tau}\left(\tilde{\alpha}^{ij}(\tau)\right) = a^{ij} - \alpha_0^{ij} \quad (\tau \in [0,1]). \tag{4.3.8}$$

I claim that for any $((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$, $\tilde{\alpha}(\tau) = \left(\left(\tilde{\alpha}^{ij}(\tau)\right)\right) \in \mathcal{A}$ for all $\tau \in [0,1]$. Therefore, assume that $((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$. Then for any $\tau \in [0,1]$ and $\xi \in \mathbb{R}^n$, we have

$$\sum_{i,j}^n \tilde{\alpha}^{ij}(\tau)(x)\xi_i\xi_j = \tau \sum_{i,j}^n a^{ij}\xi_i\xi_j + (1-\tau)\sum_{i,j}^n \alpha_0^{ij}(x)\xi_i\xi_j \quad (\text{a.e. } x \in U).$$

Since $((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$, clearly $\tilde{\alpha}(\tau)$ is symmetric for all τ. Further, we have

$$\sum_{i,j}^n \tilde{\alpha}^{ij}(\tau)\xi_i\xi_j \geq \tau \mathrm{a}|\xi|^2 + (1-\tau)\mathrm{a}|\xi|^2$$

and thus

$$\sum_{i,j}^n \tilde{\alpha}^{ij}(\tau)\xi_i\xi_j \geq \mathrm{a}|\xi|^2 > 0.$$

By a similar calculation, we have

$$|\tilde{\alpha}(\tau)\xi| \leq \mathrm{b}|\xi|$$

and thus $\tilde{\alpha}(\tau) \in \mathcal{G}_{\mathrm{a,b}}$ for all $\tau \in [0,1]$ as claimed.

Given this, we can use the Lax-Milgram Theorem to define an injection $u^A : [0,1] \to H_0^1(U)$ for any $A = ((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$ such that for each $\tau \in [0,1]$, $u^A(\tau) \in H_0^1(U)$ is the unique solution to the

PDE

$$\begin{cases} -\sum_{i,j=1}^{n} \left(\tilde{\alpha}^{ij}(\tau)\, u^{\tilde{\alpha}}(\tau)_{x_i}\right)_{x_j} = f \text{ in } U \\ u^A = 0 \text{ on } \partial U. \end{cases} \tag{4.3.9}$$

In particular, when $\tau = 0$, we have

$$f = -\sum_{i,j=1}^{n} \left(\tilde{\alpha}^{ij}(0)\, u^A(0)_{x_i}\right)_{x_j} = \sum_{i,j=1}^{n} \left(\alpha_0^{ij} u^A(0)_{x_i}\right)_{x_j}$$

and thus

$$u^A(0) = u^0 \quad \text{for all } A \in \mathcal{G}_{\mathrm{a,b}}. \tag{4.3.10}$$

Taking the derivative of both sides (4.3.9) with respect to τ, we have the following PDE

$$\begin{cases} -\sum_{i,j=1}^{n} \left(\tilde{\alpha}^{ij}(\tau) \left(u^A(\tau)\right)'_{x_i}\right)_{x_j} = \sum_{i,j=1}^{n} \left(\left(\tilde{\alpha}^{ij}(\tau)\right)' u^A(\tau)_{x_i}\right)_{x_j} & \text{in } U \\ \tilde{u} = 0 & \text{on } \partial U \end{cases} \tag{4.3.11}$$

for any $\tau \in [0,1]$, where $' = {}^{d}\!/_{dt}$.

Now define the function $i : [0,1] \to \mathbb{R}$ as

$$i(\tau; A) = \left\| u^A(\tau) - \hat{u} \right\|^2 = \frac{1}{2} \int_U \left(u^A(\tau) - \hat{u}\right)^2 dx.$$

Taking the derivative of both sides with respect to τ, we have

$$i'(\tau; A) = \int_U \left(u^A(\tau) - \hat{u}\right) \left(u^A(\tau)\right)' dx \quad (\tau \in [0,1]).$$

By (4.3.10), i has a minimum at $\tau = 0$ for any $A = ((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$. Hence we have

$$0 \le i'(0; A) = \int_U \left(u^0 - \hat{u}\right) \tilde{u} dx, \tag{4.3.12}$$

where

$$\tilde{u} = \left(u^A(0)\right)'.$$

Since (4.3.12) holds for any $A \in \mathcal{G}_{\mathrm{a,b}}$, by the weak form of co-state equation (4.3.5) we have

$$\begin{aligned} 0 \le \int_U \left(u^0 - \hat{u}\right) \tilde{u} dx &= B\left[p^0, \tilde{u}\right] \\ &= \int_U \sum_{i,j=1}^{n} \left(\alpha_0^{ij} p^0_{x_i}\right) \tilde{u}_{x_j} dx. \end{aligned}$$

Additionally, by the weak form of (4.3.11) we have

$$\int_U \sum_{i,j=1}^{n} \left(\alpha_0^{ij} \tilde{u}_{x_i}\right) v_{x_j} dx = -\int_U \sum_{i,j=1}^{n} \left(\left(a^{ij} - \alpha_0^{ij}\right) u^0_{x_i}\right) v_{x_j} dx \quad \text{for any } v \in H_0^1(U).$$

Setting $v = p^0$ and combining these two equations, we have

$$0 \leq \int_U \sum_{i,j=1}^n \left(\alpha_0^{ij} p_{x_i}^0\right) \tilde{u}_{x_j} dx = -\int_U \sum_{i,j=1}^n \left(\left(a^{ij} - \alpha_0^{ij}\right) u_{x_i}^0\right) p_{x_j} dx.$$

Rearranging terms and noting that A and α_0 are symmetric, we have

$$\int_U \sum_{i,j=1}^n p_{x_i} a^{ij} u_{x_j}^0 dx \leq \int_U \sum_{i,j=1}^n p_{x_i} \alpha_0^{ij} u_{x_j}^0 dx \tag{4.3.13}$$

for any $A = ((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$. Now select a particular $((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$ and for a given $E \subset U$, define $\hat{\alpha}_E^{ij}$ for $i, j = 1, \ldots, n$ as

$$\hat{\alpha}_E^{ij} = \begin{cases} a^{ij} & \text{on } E \\ \alpha_0^{ij} & \text{on } U \setminus E. \end{cases}$$

Clearly, $\hat{\alpha}_E := \left(\left(\hat{a}_E^{ij}\right)\right) \in \mathcal{G}_{\mathrm{a,b}}$ on E. Hence by (4.3.13), we have

$$\int_E \sum_{i,j=1}^n p_{x_i}^0 a^{ij} u_{x_j}^0 dx \leq \int_E \sum_{i,j=1}^n p_{x_i}^0 \alpha_0^{ij} u_{x_j}^0 dx$$

and thus

$$\frac{1}{\mu(E)} \int_E \sum_{i,j=1}^n p_{x_i}^0 a^{ij} u_{x_j}^0 dx \leq \frac{1}{\mu(E)} \int_E \sum_{i,j=1}^n p_{x_i}^0 \alpha_0^{ij} u_{x_j}^0 dx,$$

where $\mu(\cdot)$ is the Lebesgue measure.

Since U is open and bounded, we can set $E = B(x, r)$ for r small enough such that

$$B(x, r) \subset U,$$

where $B(x, r)$ is the ball of radius r centered at x. By Lebesque's Differentiate Theorem, taking the limit as $r \to 0$, we have

$$\sum_{i,j=1}^n p_{x_i}^0(x) a^{ij} u_{x_j}^0(x) \leq \sum_{i,j=1}^n p_{x_i}^0(x) \alpha_0^{ij}(x) u_{x_j}^0(x)$$

for almost every $x \in U$ and every $A = ((a^{ij})) \in \mathcal{G}_{\mathrm{a,b}}$. □

4.3.3. Deriving an optimal control policy given an optimal state and co-state. We now derive an optimal control policy given an optimal state and co-state using the pointwise properties of the maximum principle derived immediately above to transform optimal control problem into a finite dimensional optimization problem. Specifically, we recognize that given an optimal state-control pair (u^0, p^0) and setting

$$x = Dp^0(z) \text{ and } y = Du^0(z)$$

at any fixed point $z \in U$, this maximum principle is equivalent to the following optimization problem:

Problem 1. Given $x, y \in \mathbb{R}^n$, find a matrix $A^0 \in \mathcal{G}_{\mathrm{a,b}}$ such that

$$x^T A^0 y = \max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{x^T A y\right\}.$$

Consequently, given $u^0(z)$ and $p^0(z)$ at each point $z \in U$, we use solutions to this optimization problem to determine an optimal control policy of our optimal control problem simply by setting $\alpha^0(z) := A^0$ at each $z \in U$. This is captured in the following results.

THEOREM 4.3.2. *For given $x, y \in \mathbb{R}^n$, assume that $A^0 \in \mathcal{G}_{\mathrm{a,b}}$ is a solution to Problem 1. Then*

$$x^T A^0 y = \frac{\mathrm{b}+\mathrm{a}}{2}(x \cdot y) + \frac{\mathrm{b}-\mathrm{a}}{2}|x|\,|y|.$$

PROOF. Assume that A^0 is a solution to Problem 1. If $x = 0$ or $y = 0$, the proof is trivial. Therefore, select $x, y \in \mathbb{R}^n$ such that $|x| = |y| = 1$. A simple calculation shows that

$$\max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{x^T A y\right\} = \max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{\frac{(x+y)^T A (x+y) - (x-y)^T A (x-y)}{4}\right\}. \tag{4.3.14}$$

Additionally, for any $A \in \mathcal{A}$, we have

$$\mathrm{a}\,|x+y|^2 \leq (x+y)^T A (x+y) \leq \mathrm{b}\,|(x+y)|^2$$

and

$$\mathrm{a}\,|x-y|^2 \leq (x-y)^T A (x-y) \leq \mathrm{b}\,|(x-y)|^2$$

by (4.3.2).

Hence we have

$$\max_{A \in \mathbb{A}} \left\{\frac{(x+y)^T A (x+y) - (x+y)^T A (x+y)}{4}\right\} = \frac{1}{4}\left(\mathrm{b}\,|(x+y)|^2 - \mathrm{a}\,|(x-y)|^2\right).$$

Substituting this into (4.3.14) and recalling that $|x| = |y| = 1$, we have

$$\begin{aligned}
\max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{x^T A y\right\} &= \max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{\frac{(x+y)^T A (x+y) - (x+y)^T A (x+y)}{4}\right\} \\
&= \frac{1}{4}\left(\mathrm{b}\,|(x+y)|^2 - \mathrm{a}\,|(x-y)|^2\right) \\
&= \frac{1}{4}\left(\mathrm{b}\left(|x|^2 + 2(x \cdot y) + |y|^2\right) - \underline{a}\left(|x|^2 - 2(x \cdot y) + |y|^2\right)\right) \\
&= \frac{1}{2}\left(\mathrm{b}\,(1 + (x \cdot y)) - \mathrm{a}\,(1 - (x \cdot y))\right) \\
&= \frac{\mathrm{b}+\mathrm{a}}{2}(x \cdot y) + \frac{\mathrm{b}-\mathrm{a}}{2}
\end{aligned}$$

and since A^0 is assumed to be an optimal solution, we have

$$x^T A^0 y = \frac{\mathrm{b}+\mathrm{a}}{2}(x \cdot y) + \frac{\mathrm{b}-\mathrm{a}}{2}. \tag{4.3.15}$$

Now assume that $|x|, |y| \neq 1$ and both are non-zero. It follows that

$$\frac{x^T}{|x|} A^0 \frac{y^T}{|y|} = \frac{\mathrm{b}+\mathrm{a}}{2}\frac{(x \cdot y)}{|x|\,|y|} + \frac{\mathrm{b}-\mathrm{a}}{2}$$

and hence

$$x^T A^0 y = \frac{\mathrm{b}-\mathrm{a}}{2}\,|x|\,|y| + \frac{\mathrm{b}+\mathrm{a}}{2}\,(x\cdot y)$$

for any $x, y \in \mathbb{R}^n$. □

The next results follows immediately from the above result and Theorem 4.3.1.

THEOREM 4.3.3. *Assume that we are given an optimal control problem of the form Problem H. If (u^0, p^0) is an optimal state / co-state pair, then corresponding optimal control policy α^0 satisfies the following equation:*

$$\begin{aligned}\left(Dp^0(x)\right)^T \alpha^0(x)\, Du^0(x) = &\left(\frac{\mathrm{b}+\mathrm{a}}{2}\right) Dp^0(x)\cdot Du^0(x)\\ &+\left(\frac{\mathrm{b}-\mathrm{a}}{2}\right)\left|Dp^0(x)\right|\left|Du^0(x)\right|\end{aligned}$$

for a.e. $x \in U$.

4.3.4. Deriving the optimal state and co-state state given an optimal control policy. We now derive the optimal state and co-state state given an optimal control policy for a problems of the form Problem H. We again consider an optimization problem of the form Problem 1 and some preliminary results from matrix algebra.

LEMMA 4.3.4. *Assume that $M \in \mathbb{M}^n$ and define the matrix N as*

$$N := P^T M P \tag{4.3.16}$$

where $P \in \mathbb{M}^n$ is orthogonal. Then $M \in \mathcal{G}_{\mathrm{a,b}}$ if and only if $N \in \mathcal{G}_{\mathrm{a,b}}$. Consequently, $\mathcal{G}_{\mathrm{a,b}}$ is closed under orthogonal rotations.

PROOF. Assume $P \in \mathbb{M}^n$ is orthogonal. We begin by showing that if $M \in \mathcal{G}_{\mathrm{a,b}}$, then $N \in \mathcal{G}_{\mathrm{a,b}}$,where N is defined at (4.3.16). Since $M \in \mathcal{G}_{\mathrm{a,b}}$, we have

$$N^T = \left(P^T M P\right)^T = (MP)^T P = P^T M P = N. \tag{4.3.17}$$

and thus $N \in \mathbb{S}^n$.

By assumption, P is orthogonal and thus it is an isometry on $\mathbb{R}^n$. Hence $|Px| = |x|$ for all $x \in \mathbb{R}^n$ and for any $x \in \mathbb{R}^n$, there exists a unique $y \in \mathbb{R}^n$ such that

$$y = Px \text{ and } x = P^T y.$$

Consequently, since $M \in \mathcal{G}_{\mathrm{a,b}}$ by assumption we have

$$\mathrm{a}\,|\xi|^2 = \mathrm{a}\,|P\xi|^2 \leq (P\xi)^T M\,(P\xi) \tag{4.3.18}$$

for any $\xi \in \mathbb{R}^n$. However, by (4.3.17) we have

$$(P\xi)^T M\,(P\xi) = \xi^T \left(P^T M P\right)\xi = \xi^T N \xi$$

and thus

$$\mathrm{a}\,|\xi|^2 \leq \xi^T N \xi.$$

Additionally, we have

$$|M\eta| \leq \mathrm{b}\,|\eta|$$

for any $\eta \in \mathbb{R}^n$. Setting $\eta = P\xi$, we have

$$|M\eta| \leq \mathrm{b}\,|\eta| = \mathrm{b}\,|P\xi| = \mathrm{b}\,|\xi|\,.$$

We also have

$$|M\eta| = \left|\left(PNP^T\right)P\xi\right| = |P\,(N\xi)| = |N\xi|$$

and therefore

$$|N\xi| \leq \mathrm{b}\,|\xi|\,.$$

Since this is true for any $\xi \in \mathbb{R}^n$, then $N \in \mathcal{G}_{\mathrm{a,b}}$. The converse follows from a similar calculation. □

LEMMA 4.3.5 (Spectral decomposition). *Assume that $M \in \mathbb{S}^n$. Then*

$$M = O\Lambda O^T, \tag{4.3.19}$$

where Λ is the diagonal matrix of the eigenvalues of B and O is the matrix whose columns are the eigenvectors of M.

PROOF. *See* Harville [**39**, pp. 537-539]. □

For a given $x, y \in \mathbb{R}^n$, we define the matrix $B_{x,y}$ that will be used extensively in subsequent analysis.

DEFINITION. For a given $x, y \in \mathbb{R}^n$, define the matrix $B_{x,y} \in \mathbb{M}^n$ as

$$B_{x,y} := (x \otimes y) + (y \otimes x)\,, \tag{4.3.20}$$

where "$\otimes$" is the tensor product defined as

$$s \otimes t = \begin{pmatrix} s_1t_1 & \cdots & s_1t_n \\ \vdots & \ddots & \cdots \\ s_nt_1 & \cdots & s_nt_n \end{pmatrix} \quad (s, t \in \mathbb{R}^n)\,.$$

We note that under this definition, $B_{x,y} = ((b^{ij})) \in \mathbb{S}^n$ since

$$b^{ij} = x_iy_j + y_ix_j \quad (\text{for } i, j = 1, \ldots, n)\,.$$

We also note that since $B \in \mathbb{S}^n$, by Lemma 4.3.5 there exists matrices D and O (also dependent on the choice of x and y) such that

$$B_{x,y} = ODO^T,$$

where D is the diagonal matrix of the eigenvalues of $B_{x,y}$ and O is the matrix whose columns are its corresponding eigenvectors. This justifies the following definition.

DEFINITION. Let $B_{x,y}$ the matrix defined at (4.3.20) for a given $x, y \in \mathbb{R}^n$. Define the matrix $D_{x,y}$ as

$$D_{x,y} := \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \tag{4.3.21}$$

where $\{\lambda_1, \ldots, \lambda_n\}$ are the (possibly non-unique) eigenvalues of $B_{x,y}$ and define

$$O_{x,y} := (z_1 \mid \cdots \mid z_n)$$

such that its columns $z_1, \ldots, z_n$ are the corresponding eigenvectors.

NOTATION. In the sequel, we often drop the subscript from the matrices B, D, and O if the meaning is clear from the context in which they are used. □

DEFINITION. Fix $x, y \in \mathbb{R}^n$. Given a matrix $A \in \mathcal{G}_{\text{a,b}}$, we define the matrix $\tilde{A}$ as

$$\tilde{A} = O^T A O,$$

where O is the matrix of the eigenvectors of the matrix $B_{x,y}$.

We now restate Problem 1 in terms of the matrix $B_{x,y}$ and its spectral decomposition

$$B_{x,y} = ODO^T.$$

LEMMA 4.3.6. *Fix $x, y \in \mathbb{R}^n$ and assume that we are given an optimization problem of the form Problem 1. Then we have*

$$\max_{A \in \mathcal{G}_{\text{a,b}}} \left\{x^T A y\right\} = \max_{A \in \mathcal{G}_{\text{a,b}}} \left\{\frac{1}{2}\text{tr}\left(AD_{x,y}\right)\right\}.$$

Additionally, $A^0 \in \underset{A \in \mathcal{G}_{\text{a,b}}}{arg\,max}\left(x^T A y\right)$ if and only if $\tilde{A}^0 = O^T A^0 O \in \underset{A \in \mathcal{G}_{\text{a,b}}}{arg\,max}\left(\frac{1}{2}\text{tr}\left(AD_{x,y}\right)\right)$.

PROOF. Fix $x, y \in \mathbb{R}^n$. Then for any $A = ((a^{ij})) \in \mathcal{G}_{\text{a,b}}$, we have

$$x^T A y = y^T A x$$

and

$$\tilde{A} = O^T A O \in \mathcal{G}_{\text{a,b}} \tag{4.3.22}$$

by Lemma 4.3.4 and

By the definition of $B_{x,y}$, we have

$$\begin{aligned}
x^T A y &= \frac{1}{2}\left(\sum_{i,j=1}^{n} a^{ij} x_i y_j + \sum_{i,j=1}^{n} a^{ij} y_i x_j\right) \\
&= \frac{1}{2}\left(\sum_{i,j=1}^{n} a^{ij}\left(x_i y_j + y_i x_j\right)\right) \\
&= \frac{1}{2}\left(\sum_{i,j=1}^{n} a^{ij} b^{ij}\right) \\
&= \frac{1}{2}\text{tr}\left(AB\right) \\
&= \frac{1}{2}\text{tr}\left(A\left(ODO^T\right)\right) \\
&= \frac{1}{2}\text{tr}\left(O^T A O D\right)
\end{aligned}$$

$$= \frac{1}{2}\mathrm{tr}\left(\tilde{A}D\right), \tag{4.3.23}$$

where "tr" is the trace of a square matrix. Consequently, by (4.3.22), we have

$$\max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left\{x^TAy\right\} = \max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left\{\frac{1}{2}\mathrm{tr}\left(AD_{x,y}\right)\right\}. \tag{4.3.24}$$

Moreover, since (4.3.23) is true for any $A \in \mathcal{G}_{\mathrm{a,b}}$, it is true for any $A^0 \in \arg\max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left(x^TAy\right)$. Hence we have

$$\frac{1}{2}\mathrm{tr}\left(\tilde{A}^0D\right) = x^TA^0y = \max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left\{x^TAy\right\} = \max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left\{\frac{1}{2}\mathrm{tr}\left(\tilde{A}D\right)\right\}$$

and thus $\tilde{A}^0 \in \arg\max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left(\frac{1}{2}\mathrm{tr}\left(\tilde{A}D\right)\right)$. The converse follows from the same calculation. □

Given this result, we now consider optimization problems of the following form:

Problem 2. Given $x, y \in \mathbb{R}^n$, find a matrix $A^0 \in \mathcal{G}_{\mathrm{a,b}}$ such that

$$\mathrm{tr}\left(A^0D_{x,y}\right) = \max_{A\in\mathcal{G}_{\mathrm{a,b}}}\left\{\mathrm{tr}\left(AD_{x,y}\right)\right\}.$$

In other words, we can identify solutions to Problem 1 provided that we can identify the eigenvalues and eigenvectors of the matrix $B_{x,y}$ for a given $x, y \in \mathbb{R}^n$. The next result, in fact, allows us to explicitly compute them directly.

LEMMA 4.3.7. *Select $x, y \in \mathbb{R}^n$ such that $x \neq y$ and $x, y \neq 0$. Then the eigenvalues $\{\lambda_1, \ldots \lambda_n\}$ of $B_{x,y}$ are given by*

$$\begin{cases} \lambda_1 & = x\cdot y + |x|\,|y| \geq 0 \\ \lambda_2 & = x\cdot y - |x|\,|y| \leq 0 \\ \lambda_i & = 0 \text{ for } i = 3, \ldots, n \end{cases} \tag{4.3.25}$$

and its corresponding eigenvectors $\{z_1, \ldots z_n\}$ are given by

$$\begin{cases} z_1 & = \dfrac{s}{|s|} \\ z_2 & = \dfrac{t}{|t|} \end{cases} \tag{4.3.26}$$

where

$$s = \frac{x}{|x|} + \frac{y}{|y|} \text{ and } t = \frac{x}{|x|} - \frac{y}{|y|}$$

and

$$\{z_3, \ldots, z_n\} \text{ is an orthonormal basis of } \left(\mathrm{span}\,\{x, y\}\right)^{\perp}.$$

PROOF. Select $x, y \in \mathbb{R}^n$ such that $x \neq y$ and $x, y \neq 0$ and assume that B is a matrix of the form of (4.3.20). Further, z_1 and z_2 as

$$z_1 = \frac{s}{|s|} \text{ and } z_2 = \frac{t}{|t|},$$

where

$$s = \frac{x}{|x|} + \frac{y}{|y|} \text{ and } t = \frac{x}{|x|} - \frac{y}{|y|}.$$

We first note that for any $w \in \mathbb{R}^n$, by the definition we have

$$\begin{aligned}(Bw)_i &= \sum_{j=1}^{n} b^{ij} w_j \\ &= \sum_{j=1}^{n} (x_i y_j + y_i x_j) w_j \\ &= x_i \sum_{j=1}^{n} y_j w_j + y_i \sum_{j=1}^{n} x_j w_j \\ &= (w \cdot y) x_i + (w \cdot x) y_i\end{aligned}$$

and thus

$$Bw = (w \cdot y) x + (w \cdot x) y \quad (w \in \mathbb{R}^n). \tag{4.3.27}$$

Therefore, setting $w = z_1$ we have

$$\begin{aligned}Bz_1 &= (z_1 \cdot y) x + (z_1 \cdot x) y \\ &= \left(\frac{s}{|s|} \cdot y\right) x + \left(\frac{s}{|s|} \cdot x\right) y \\ &= \frac{1}{|s|}\left[\left(\left(\frac{x}{|x|} + \frac{y}{|y|}\right) \cdot y\right) x + \left(\left(\frac{x}{|x|} + \frac{y}{|y|}\right) \cdot x\right) y\right] \\ &= \frac{1}{|s|}\left[\left(\frac{x \cdot y}{|x|} + |y|\right) x + \left(|x| + \frac{y \cdot x}{|y|}\right) y\right] \\ &= \frac{1}{|s|}\left[\left(\frac{(x \cdot y) x}{|x|} + |y| x\right) + \left(|x| y + \frac{(y \cdot x)}{|y|} y\right)\right] \\ &= \frac{1}{|s|}\left[\left(|y| x + |x| y + (x \cdot y)\left(\frac{x}{|x|} + \frac{y}{|y|}\right)\right)\right] \\ &= \frac{1}{|s|}\left[\left(|y| |x| \left(\frac{x}{|x|} + \frac{y}{|y|}\right) + (x \cdot y)\left(\frac{x}{|x|} + \frac{y}{|y|}\right)\right)\right] \\ &= (|y| |x| + (x \cdot y)) \frac{s}{|s|}.\end{aligned}$$

Consequently, we have

$$Bz_1 = (x \cdot y + |x| |y|) z_1$$

and thus z_1 is an eigenvector of B and its corresponding eigenvalue λ_1 is given by

$$\lambda_1 = x \cdot y + |x| |y|.$$

A similar calculation shows that

$$Bz_2 = (x \cdot y - |x| |y|) z_2$$

and thus z_2 is also an eigenvector of B and its corresponding eigenvalue λ_2 is given by

$$\lambda_2 = x \cdot y - |x|\,|y|\,.$$

As nonzero eigenvectors, $|z_1| = |z_2| = 1$. Moreover, since $\lambda_1 \neq \lambda_2$, then $z_1 \cdot z_2 = 0$. Additionally, by (4.3.27), we have

$$\ker(B) = (\operatorname{span}\{x, y\})^{\perp}$$

and thus $\lambda_3 = \cdots = \lambda_n = 0$ and $\{z_3, \ldots, z_n\}$ is an orthonormal basis of $(\operatorname{span}\{x, y\})^{\perp}$. □

We now define a class of matrices that we subsequently show are solutions optimization problems of the form 4.3.4.

Definition. We define $\mathbb{A}$ as the set of all matrices $M = ((m^{ij})) \in \mathbb{M}^n$ such that

$$m^{ij} := \begin{cases} \mathrm{b} & \text{for } i, j = 1 \\ \mathrm{a} & \text{for } i, j = 2 \\ c_{ij} & \text{for all other } i, j = 1, \ldots, n, \end{cases}$$

where $\mathrm{a} \leq c_{ij} \leq \mathrm{b}$ for $i, j = 1, \ldots, n$.

The next results show that every matrix $A \in \mathbb{A}$ is, in fact, a solution to optimization problems of the form Problem 2.

Lemma 4.3.8. *Fix $x, y \in \mathbb{R}^n$ and let $\mathbb{A}$ be the set defined above. Then*

$$\mathbb{A} = \underset{A \in \mathcal{G}_{\mathrm{a,b}}}{arg\,max} \left(\operatorname{tr}(AD_{x,y})\right).$$

Proof. Fix $x, y \in \mathbb{R}^n$ and let $A \in \mathcal{G}_{\mathrm{a,b}}$. By (4.3.25), we have

$$\begin{aligned} \operatorname{tr}(AD) &= \sum_{i=1}^{n} a^{ij} D^{ii} \\ &= a^{11}\lambda_1 + a^{22}\lambda_2 \\ &= \mathrm{b}\lambda_1 + \mathrm{a}\lambda_2, \end{aligned}$$

where

$$\begin{cases} \lambda_1 &= x \cdot y + |x|\,|y| \\ \lambda_2 &= x \cdot y - |x|\,|y|\,. \end{cases}$$

Hence we have

$$\operatorname{tr}(AD) = (\mathrm{b} + \mathrm{a})(x \cdot y) + (\mathrm{b} - \mathrm{a})\,|x|\,|y|$$

and thus by Lemma 4.3.2, we have

$$\operatorname{tr}(AD_{x,y}) = \max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{\operatorname{tr}(AD_{x,y})\right\}.$$

The converse follows from the same calculation. □

Theorem 4.3.9. *Assume that $\alpha^0 \in L^{\infty}(U, \mathcal{G}_{\mathrm{a,b}})$ is a solution to a given optimal control problem of the form Problem H. Then the corresponding optimal state $u^0 \in H_0^1(U)$ and co-state state $p^0 \in$*

$H_0^1(U)$ satisfy the following PDE

$$\begin{cases} -\beta_1 \triangle u^0 - \beta_2 \mathrm{div}\left(\dfrac{|Du^0|}{|Dp^0|}(Dp^0)\right) &= f \\ -\beta_2 \triangle p^0 - \beta_2 \mathrm{div}\left(\dfrac{|Dp^0|}{|Du^0|}(Du^0)\right) &= u^0 - \hat{u}, \end{cases} \quad \text{in } U \tag{4.3.28}$$

where

$$\beta_1 = \frac{\mathrm{b}+\mathrm{a}}{2} \text{ and } \beta_2 = \frac{\mathrm{b}-\mathrm{a}}{2}$$

assuming $Du^0, Dp^0 \neq 0$ in U.

REMARK. It is an open question to understand (4.3.28) if $Du^0 = 0$ and/or $Dp^0 = 0$ at some points in U. □

PROOF. Assume the setup above. Since α^0 is assumed to be an optimal solution, by the maximum principle Theorem 4.3.1, we have

$$\max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{ \left(Dp^0(w)\right)^T A Du^0(w) \right\} = \left(Dp^0(w)\right)^T \alpha^0(w) Du^0(w),$$

for any $w \in U$, where u^0 and p^0 are the unknown state and co-state, respectively. Fixing $w \in U$ and setting

$$x = Dp^0(w), y = Du^0(w), A = \alpha(w), \text{ and } A^0 = \alpha^0(w), \tag{4.3.29}$$

we see that the optimization problem above can be written as

$$\max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left\{ x^T A y \right\}.$$

Now define the matrix $\tilde{A}^0$ by

$$A^0 := O \tilde{A}^0 O^T. \tag{4.3.30}$$

Then by Lemma 4.3.8,

$$\tilde{A}^0 \in \arg\max_{A \in \mathcal{G}_{\mathrm{a,b}}} \left(\mathrm{tr}\left(A D_{x,y}\right) \right),$$

Therefore, multiplying both sides of (4.3.30) by z_1, we have

$$A z_1 = O \tilde{A}^0 O^T z_1.$$

Additionally, by the definition of O, we have

$$O e_i = z_i \;\Rightarrow\; O^T z_i = e_i,$$

for $i = 1, \ldots, n$, where e_i is the i^{th} unit vector. Substituting this into the above expression, we have

$$\begin{aligned} A^0 z_1 &= O \tilde{A}^0 O^T z_1 \\ &= O \tilde{A}^0 e_1 \\ &= \mathrm{b} O e_1 \\ &= \mathrm{b} z_1 \end{aligned}$$

By (4.3.26), we have

$$z_1 = \frac{x}{|x|} + \frac{y}{|y|}$$

and thus

$$A^0\left(\frac{x}{|x|} + \frac{y}{|y|}\right) = \mathrm{b}\left(\frac{x}{|x|} + \frac{y}{|y|}\right).$$

A similar calculation using z_2 shows that

$$A^0\left(\frac{x}{|x|} - \frac{y}{|y|}\right) = \mathrm{a}\left(\frac{x}{|x|} - \frac{y}{|y|}\right).$$

Then by (4.3.29), we have

$$\alpha^0(w)\left(\frac{Du^0(w)}{|Du^0(w)|} + \frac{Dp^0(w)}{|Dp^0(w)|}\right) = \mathrm{b}\left(\frac{Du^0(w)}{|Du^0(w)|} + \frac{Dp^0(w)}{|Dp^0(w)|}\right)$$

and

$$\alpha^0(w)\left(\frac{Du^0(w)}{|Du^0(w)|} - \frac{Dp^0(w)}{|Dp^0(w)|}\right) = \mathrm{a}\left(\frac{Du^0(w)}{|Du^0(w)|} - \frac{Dp^0(w)}{|Dp^0(w)|}\right)$$

for any $w \in U$.

Summing these two identities, we have

$$2\alpha^0(w)\frac{Du^0(w)}{|Du^0(w)|} = (\mathrm{b}+\mathrm{a})\frac{Du^0(w)}{|Du^0(w)|} + (\mathrm{b}-\mathrm{a})\frac{Dp^0(w)}{|Dp^0(w)|}$$

and thus

$$\alpha^0(w)\,Du^0(w) = \beta_1 Du^0(w) + \beta_2\frac{|Du^0(w)|}{|Dp^0(w)|}Dp^0(w),$$

where

$$\beta_1 = \frac{\mathrm{b}+\mathrm{a}}{2} \text{ and } \beta_2 = \frac{\mathrm{b}-\mathrm{a}}{2}.$$

A similar calculation shows

$$\alpha^0(w)\,Dp^0(w) = \beta_1 Dp^0(w) + \beta_2\frac{|Dp^0(w)|}{|Du^0(w)|}Du^0(w).$$

Since this is true for all $w \in U$, we have

$$\begin{cases} \alpha^0 Du^0 &= \beta_1 Du^0 + \beta_2\dfrac{|Du^0|}{|Dp^0|}Dp^0 \\ \alpha^0 Dp^0 &= \beta_1 Dp^0 + \beta_2\dfrac{|Dp^0|}{|Du^0|}Du^0. \end{cases} \text{ in } U \tag{4.3.31}$$

Now recall that the state and co-state equations for Problem H are given by

$$-\alpha Du = f \text{ in } U$$

and

$$-\alpha Dp = u^0 - \hat{u} \text{ in } U,$$

respectively. Consequently, if $\alpha^0 = \left(\left(\alpha_0^{ij}\right)\right)$ is an optimal control policy, then by (4.3.31) u^0 and p^0 satisfy the follow coupled-system PDE

$$\begin{cases} -\beta_1 \triangle u^0 - \beta_2 \mathrm{div}\left(\dfrac{|Du^0|}{|Dp^0|}\left(Dp^0\right)\right) &= f \\ -\beta_1 \triangle p^0 - \beta_2 \mathrm{div}\left(\dfrac{|Dp^0|}{|Du^0|}\left(Du^0\right)\right) &= u^0 - \hat{u}, \end{cases} \quad \text{in } U$$

where

$$u^0 = p^0 = 0 \text{ on } \partial U.$$

□

REMARK. It remains a very interesting open problem to study directly the coupled system of PDE above. □

APPENDIX A

DIFFERENTIATION IN FUNCTION SPACES

In the body of this paper, we employ a number of extensions of differentiation to infinite-dimensional functional spaces which we formally defined in this appendix. In developing this material, I generally rely on Luenberger [**61**] and Ekeland [**24**].

For this appendix, assume that X and Y are real Banach spaces and that $U \subseteq X$ is open and nonempty unless otherwise noted. Further, assume that

$$f : V \to Y$$

is a given.

A.1. The directional derivative and the first variation

Recall that in finite dimensions, the directional derivative of a function f at a point x in an open set $U \subset \mathbb{R}^n$ in the direction of a nonzero point y in $\mathbb{R}^n$ is defined as

$$\frac{\partial f}{\partial y}(x) := \lim_{\tau \downarrow 0} \frac{f(x + \tau y) - f(x)}{\tau}$$

provided it exists, where $\tau \in \mathbb{R}$ is small enough such that $x + \tau y \in U$.

We extend this definition to functions defined on a general Banach space in the natural way.

DEFINITION. Let $u \in U$ and $v \in X$. If the limit

$$\delta f(u)(v) := \lim_{\tau \downarrow 0} \frac{f(u + \tau v) - f(u)}{\tau} \tag{A.1.1}$$

exists, we call the limit the (***one-sided***) ***directional derivative of*** f ***at*** u ***in the direction*** v. If the limit exists for all $v \in X$, we call $\delta f(u)$ the the ***first variation of*** f ***at*** u.

We can compute the directional derivative using ordinary calculus as follows. Assume that for a fixed $u \in U$ and $v \in X$, define the function $i : \mathbb{R} \to Y$ as

$$i(\tau; u, v) := f(u + \tau v)$$

where $\tau \in \mathbb{R}$ is small enough such that $u + \tau v \in U$. Ignoring the parameters u and v for the moment, consider the following expression

$$i(\tau) - i(0) = f(u + \tau y) - f(u) \approx \tau \delta f(u)(v).$$

Dividing the above expression by $\tau > 0$ and taking the limit as $\tau \downarrow 0$, we have

$$\lim_{\tau \downarrow 0} \frac{i(\tau) - i(0)}{\tau} = \lim_{\tau \downarrow 0} \frac{f(u + \tau v) - f(u)}{\tau} = \delta f(u)(v).$$

Clearly, the left-hand side is simply the ordinary derivative of i evaluated at zero while the right-hand side is the directional derivative of f at u in the direction v assuming that the limits exist. In other words, we have

$$i'(0;u,v) = \delta f(u)(v). \tag{A.1.2}$$

This fact will be used throughout this paper.

A.2. The Gâteaux derivative

DEFINITION. Suppose that the first variation of f exists at a given point $u \in U$ and assume $Gf(u) : U \to Y$ is a bounded linear operator dependent on u such that

$$Gf(u)(v) = \delta f(u)(v)$$

exists for all for all $v \in X$. Then we say that f is said to be ***Gâteaux differentiable at*** u. Additionally, we say the operator $Gf(u)$ is the ***Gâteaux derivative of*** f ***at*** u and $Gf(u)(v)$ is said to be the ***Gâteaux derivative of*** f ***at*** u ***in the direction*** v. If f is Gâteaux differentiable at every $u \in U$, we say that f is ***Gâteaux differentiable in*** U.

NOTATION. If the Gâteaux derivative exists, then $Gf(u) \in \mathscr{L}(X,Y)$. Consequently we will generally use the standard notation for linear operators and express Gâteaux derivative of f at u acting on $v \in X$ as

$$Gf(u)\,v := Gf(u)(v).$$

If f is real-valued, then $Gf(u) \in \mathscr{L}(X,\mathbb{R})$ and thus it is an element of X^*. Consequently, we often write the Gâteaux derivative of f at u acting on $v \in X$ as the dual pairing

$$Gf(u)\,v = \langle Gf(u), v\rangle_{X^*,X}. \tag{A.2.1}$$

□

DEFINITION (Partial derivatives). Let X, Y and Z be real Banach spaces and assume $U \subset X$ and $V \subset Y$ are open and nonempty and assume we are given a function $f : U \times V \to Z$. Further, for any fixed $v \in Y$, define $g(v) : U \to Z$ as

$$g(v)(u) := f(u,v) \quad (u \in U).$$

Similarly, for any fixed $u \in U$, define $h(u) : V \to Z$ as

$$h(u)(v) := f(u,v) \quad (v \in V).$$

Then for $(u,v) \in U \times V$, we say that g and h are the ***partial Gâteaux derivatives of*** f ***at*** u and v provided they exist and denoted them as

$$G_u f(u,v) := Gg(v) \text{ and } G_v f(u,v) := Gh(u).$$

Further, we say that $Gf(u,v)$ is the ***Gâteaux gradient of*** f ***at*** (u,v), where

$$Gf(u,v)(w_1,w_2) = G_u f(u,v)\,w_1 + G_v f(u,v)\,w_2,$$

for any $(w_1,w_2) \in X \times Y$.

REMARK. Under this definition, $G_u f(u,v) \in \mathscr{L}(X,Z)$ and $G_v f(u,v) \in \mathscr{L}(Y,Z)$ for $(u,v) \in U \times V$. □

A.3. The Fréchet derivative

We now consider the analog of the gradient for functions of functions.

DEFINITION. Suppose we are given a function $f : U \to Y$. If there exists a bounded linear operator $Df(u) : X \to Y$ and a mapping $r(\cdot\,; u) : X \to Y$ (referred to as the ***remainder***) for a fixed $u \in U$ such that

$$f(u+v) - f(u) = Df(u)(v) + r(v;u) \tag{A.3.1}$$

for all $v \in X$, where $u+v \in U$ and $\|v\|_X > 0$ and $r(\cdot\,; u) : X \to Y$ satisfies the following condition

$$\frac{\|r(v;u)\|_Y}{\|v\|_X} \to 0 \text{ as } \|v\|_X \to 0.$$

Then f is said to be ***Fréchet differentiable at*** u. Additionally, the operator $Df(u)$ is said to be the ***Fréchet derivative of*** f ***at*** u and $Df(u)(v)$ is said to be the ***Fréchet derivative of*** f ***at*** u ***in the direction*** v. If f is Fréchet differentiable at every $u \in U$, we say that f is ***Fréchet differentiable in*** U.

REMARK. If it exists, $Df(u) \in \mathscr{L}(X,Y)$ by definition. □

We now show that if the Fréchet derivative exists, it is equal to the Gâteaux derivative which we can readily compute via A.1.2.

LEMMA A.3.1. *Let $u \in U$ and assume that $v \in X$ such that $u+v \in U$ and $\|v\|_X > 0$. If f is Fréchet differentiable at u, then*

$$\lim_{\|v\|_X \to 0} \frac{\|f(u+v) - f(v) - Df(u)v\|_Y}{\|v\|_X} = 0.$$

THEOREM A.3.2. *Assume $u \in U$. If f is Fréchet differentiable at u, then it also Gâteaux differentiable at u and the operators are the same.*

PROOF. Assume $u \in U$ and fix $v \in X$ such that $\|v\|_X > 0$. By assumption f is Fréchet differentiable at u. Therefore select $\tau > 0$ small enough such that $u + \tau v \in V$. By the above lemma, we have

$$\lim_{\|\tau v\|_X \to 0} \frac{\|f(u+\tau v) - f(u) - Df(u)(\tau v)\|_Y}{\|\tau v\|_X} = 0 \Rightarrow \lim_{\tau \to 0} \frac{\|f(u+\tau v) - f(u) - Df(u)(\tau v)\|_Y}{\tau} = 0.$$

Further, by the linearity of the Fréchet derivative, we have

$$0 = \lim_{\tau \downarrow 0} \frac{f(u+\tau v) - f(u) - Df(u)(\tau v)}{\tau} = \lim_{\tau \downarrow 0} \frac{f(u+\tau v) - f(u)}{\tau} - Df(u)v.$$

Hence it follows that

$$Df(u)v = \lim_{\tau \downarrow 0} \frac{f(u+\tau v) - f(u)}{\tau} = Gf(u)v.$$

This is holds for all $v \in X$ such that $\|v\|_X > 0$, f is Gâteaux differentiable and the two operators are the same. □

REMARK. If f is real-valued and it is Fréchet differentiable at $u \in U$, then $Df(u) \in X^*$ and thus

$$Df(u)\, v = \langle Df(u), v \rangle_{X^*,X} \quad (v \in X). \tag{A.3.2}$$

□

REMARK. Since Fréchet derivative is equal to the Gâteaux derivative, by A.1.2 we have the following identity

$$Df(u)\, v = i'(0; u, v),$$

where $i : \mathbb{R} \to Y$ such that

$$i(\tau; u, v) := f(u + \tau v).$$

Hence, we can compute the Fréchet derivative at u in the direction v as

$$Df(u)\, v = \frac{d}{d\tau} f(u + \tau v) \mid_{\tau=0}. \tag{A.3.3}$$

We use both (A.3.2) and (A.3.3)to express the Fréchet derivative extensively in the main body of the paper. □

THEOREM A.3.3 (Chain rule). *Let X, Y, and Z be real Banach spaces. Assume $U \subset X$ and $V \subset Y$ are open and that $f : U \to V \subset Y$ and $g : V \to Z$ are Fréchet differentiable at $u \in U$ and $f(u) \in V$, respectively. Then the composite mapping $h : U \to Z$ defined as*

$$h(\cdot) := g \circ f(\cdot) = g(f(\cdot))$$

is Fréchet differentiable at u and is given by

$$Dh(u) = Dg(f(u))\, Df(u).$$

PROOF. Assume that $f : U \to V \subset Y$ and $g : V \to Z$ are Fréchet differentiable at $u \in U$ and $f(u) \in V$, respectively. Further, define $h : U \to Z$ as

$$h(\cdot) := g(f(\cdot)).$$

We begin by showing that h is Fréchet differentiable at u. Hence we need to show that for any $v \in X$ such that $u + v \in U$ and $\|v\|_X > 0$, we have

$$\frac{\|r(v; u)\|_Y}{\|v\|_X} \to 0 \text{ as } \|v\|_X \to 0$$

or equivalently

$$\|h(u + v) - h(u) - Dh(u)(v)\| = o(\|v\|).$$

Therefore select $v \in X$ such that $u + v \in U$ and $\|v\|_X > 0$ and define $w, z \in V$ as

$$w := f(u) \text{ and } z := f(u + v) - f(u).$$

Then by assumption

$$h(u + v) - h(u) = g(f(u + v)) - g(f(u)) = g(w + z) - g(w).$$

Since g is assumed to be Fréchet differentiable at $f(u) = w$, using the definition of the Fréchet derivative, we have

$$o(\|z\|) = \|g(w+z) - g(w) - Dg(w)\, z\| = \|h(u+v) - h(u) - Dg(w)\, z\|. \tag{A.3.4}$$

On the other hand, f is also assumed to be Fréchet differentiable at u and thus by , we have

$$\|z - Df(u)\, v\| = o(\|v\|).$$

Substituting these two expressions into A.3.4, we get

$$\|h(u+v) - h(u) - Dg(w)\, Df(u)\, v\| = o(\|z\|) + o(\|v\|).$$

Now consider the right-hand side of the above expression. By assumption, f is Fréchet differentiable at u and thus $Df(u) \in \mathscr{L}(X, Y)$ and it is continuous at u. Hence

$$O(\|v\|) = \|f(u+v) - f(u)\| = \|z\|.$$

Consequently, the right-hand side is simply $o(v)$ and h is Fréchet differentiable at u. Further, by the definition of the Fréchet derivative, we have

$$Dh(u)\, v = Dg(w)\, Df(u)\, v.$$

Since v is arbitrary, we deduce the following operator identity,

$$Dh(u) = Dg(f(u))\, Df(u).$$

□

DEFINITION (Partial derivatives). Let X, Y and Z be real Banach spaces and assume $U \subset X$ and $V \subset Y$ are open and nonempty and assume we are given a function $f : U \times V \to Z$. Additionally fix $v \in V$ and define $g(v) : U \to Z$ as

$$g(v)(u) := f(u, v) \quad (u \in U).$$

Similarly, fix $u \in U$ and define $h(u) : V \to Z$ as

$$h(u)(v) := f(u, v) \quad (v \in V).$$

Then for $(u, v) \in U \times V$, we say that g and h are the ***partial Fréchet derivatives of*** f ***at*** u and v and denoted them as

$$D_u f(u, v) := Dg(v) \in \mathscr{L}(X, Z) \text{ and } D_v f(u, v) := Dh(u) \in \mathscr{L}(Y, Z)$$

provided they exist. Further, we say that $Df(u, v)$ is the ***Fréchet gradient of*** f ***at*** (u, v), where

$$Df(u, v)(w_1, w_2) = D_u f(u, v)\, w_1 + D_v f(u, v)\, w_2, \tag{A.3.5}$$

for any $(w_1, w_2) \in X \times Y$.

A.4. Differentiation on convex sets

Above, we defined several forms differentiation where we restricted the domain of the function in question to open sets. We now consider differentiation on convex sets.

DEFINITION. For a given $u \in U$, we say that for $v \in X$, the direction $v - u$ is ***feasible*** if there exists a real, positive sequence $\{\tau_j\}_{j\in\mathbb{N}} \downarrow 0$ such that

$$u + \tau_j (v - u) \in U \quad (\text{for all } j). \tag{A.4.1}$$

Further, we denote the set of all feasible directions from u as

$$\mathcal{D}_u = \{v \in X \mid v \text{ is a feasible direction from } u\}.$$

THEOREM A.4.1. *Assume that U is nonempty, closed, bounded, convex subset of a real Banach space X, then every direction in X is feasible.*

PROOF. Assume that $U \subset X$ is nonempty, closed, bounded, and convex and fix $u \in U$. Then by the definition of convexity, for any $v \in U$

$$\tau_j v + (1 - \tau_j) u = u + \tau_j (v - u) \in U$$

for any real constant $\tau_j \in [0, 1]$. Therefore taking a sequence of such τ_j such that $\{\tau_j\}_{j\in\mathbb{N}} \downarrow 0$, we see that $v - u$ is feasible. □

DEFINITION. Let $u \in U$, where $U \subseteq X$ is a convex. Given a function $f : U \to Y$, if a bounded linear operator $G_U(u) : \mathcal{D}_u \to Y$ exists such that

$$G_U f(u)(v - u) = \lim_{\tau\downarrow 0} \frac{f(u + \tau(v - u)) - f(u)}{\tau} \tag{A.4.2}$$

exists for all $v \in U$, we say that f is ***Gâteaux differentiable at*** u and we call the operator $G_U(u)$ the ***Gâteaux derivative of*** f ***at*** u ***with respect to the convex set*** U. Similarly, if there a function $r(\cdot\,; u) : X \to Y$ satisfying

$$f(u + v) - f(u) = G_U f(u)(v - u) + r(v; u) \tag{A.4.3}$$

for all $v \in U$ such that

$$\frac{\|r(v; u)\|_Y}{\|v\|_X} \to 0 \text{ as } \|v\|_X \to 0,$$

we say that f is ***Fréchet differentiable at*** u and we call the operator $G_U f(u)$, which denote $D_U f(u)$, as is the ***Fréchet derivative of*** f ***at*** u ***with respect to the convex set*** U.

Importantly, we see from the definition that if a function f is Gâteaux or Fréchet differentiable, then it agrees with the that derivative on a convex set.

THEOREM A.4.2. *Assume u^0 is a minimizer of a real-valued function $f : U \to \mathbb{R}$, where $U \in X$ is convex. If $D_U f(u^0)$ exists, then the following inequality holds*

$$\left\langle D_U f\left(u^0\right), u - u^0 \right\rangle \geq 0 \quad (u \in U).$$

PROOF. Since u^0 is a minimizer, then for any $\tau \in [0, 1]$,we have

$$f\left(u^0 + \tau\left(u - u^0\right)\right) - f\left(u^0\right) \geq 0 \quad (u \in U).$$

Taking a sequence of such τ such that $\{\tau_j\}_{j\in\mathbb{N}} \downarrow 0$, we have

$$D_U f\left(u^0\right)\left(u - u^0\right) = \lim_{\tau\downarrow 0} \frac{f(u^0 + \tau(u - u^0)) - f(u^0)}{\tau} \geq 0.$$

□

Given this result, we will not differential a derivative on a convex set from a derivative on an open set in the body of the paper.

Bibliography

1. R.A. Adams and J.J.F. Fournier, *Sobolev Spaces*, Pure and Applied Mathematics, Elsevier Science, 2003.
2. N.I. Akhiezer and I.M. Glazman, *Theory of Linear Operators in Hilbert Space*, Dover Books on Mathematics, Dover Publications, 2013.
3. C.D. Aliprantis and K.C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 2007.
4. G. Allaire, *Shape Optimization by the Homogenization Method*, Applied Mathematical Sciences, Springer New York, 2012.
5. J.P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Pure and applied mathematics, J. Wiley, 1984.
6. G. Bachman and L. Narici, *Functional Analysis*, Dover Books on Mathematics, Dover Publications, 2012.
7. V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Mathematics in Science and Engineering, Elsevier Science, 1992.
8. ———, *Controllability and Stabilization of Parabolic Equations*, Progress in Nonlinear Differential Equations and Their Applications, Springer International Publishing, 2018.
9. V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Springer Monographs in Mathematics, Springer Netherlands, 2012.
10. M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Modern Birkhäuser Classics, Birkhäuser Boston, 2008.
11. A. Bensoussan, G. Da Prato, M.C. Delfour, and S.K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Systems & Control: Foundations & Applications, Birkhäuser Boston, 2006.
12. S.K. Berberian, *Introduction to Hilbert Space*, AMS Chelsea Publishing Series, AMS Chelsea, 1999.
13. D.P. Bertsekas and W. Rheinboldt, *Constrained optimization and lagrange multiplier methods*, Computer science and applied mathematics, Elsevier Science, 2014.
14. L.T. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. Waanders, *Real-Time PDE-Constrained Optimization*, Computational Science and Engineering, Society for Industrial and Applied Mathematics, 2007.
15. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer New York, 2010.
16. J. Brinkhuis and V. Tikhomirov, *Optimization: Insights and Applications*, Princeton Series in Applied Mathematics, Princeton University Press, 2011.
17. R.W. Brockett, *Finite Dimensional Linear Systems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2015.
18. E. Casas, J.-P. Raymond, and H. Zidani, *Optimal Control Problem Governed by Semilinear Elliptic Equations with Integral Control Constraints and Pointwise State Constraints*, Control and Estimation of Distributed Parameter Systems (Basel) (W. Desch, F. Kappel, and K. Kunisch, eds.), Birkhäuser Basel, 1998, pp. 89–102.
19. W. Cheney, *Analysis for Applied Mathematics*, Graduate Texts in Mathematics, Springer New York, 2013.
20. F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics, Springer London, 2013.
21. J.C. De los Reyes, *Numerical PDE-Constrained Optimization*, SpringerBriefs in Optimization, Springer International Publishing, 2015.
22. W. Desch, F. Kappel, and K. Kunisch, *Control and Estimation of Distributed Parameter Systems*, International Series of Numerical Mathematics, Birkhäuser Basel, 2012.
23. I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1999.

24. I. Ekeland and T. Turnbull, *Infinite-Dimensional Optimization and Convexity*, Chicago Lectures in Mathematics, University of Chicago Press, 1983.
25. L.C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Regional Conference Series, no. no. 74, Conference Board of the Mathematical Sciences, 1990.
26. _____, *Measure Theory and Fine Properties of Functions*, CRC Press, 1992.
27. _____, *Partial Differential Equations*, second ed., American Mathematical Society, 2010.
28. H.O. Fattorini, G.C. Rota, B. Doran, F.H. O, P. Flajolet, M. Ismail, T.Y. Lam, and E. Lutwak, *Infinite Dimensional Optimization and Control Theory*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 1999.
29. P. Fitzpatrick, *Advanced Calculus*, Pure and applied undergraduate texts, American Mathematical Society, 2009.
30. W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Stochastic Modelling and Applied Probability, Springer New York, 2012.
31. W. Forst and D. Hoffmann, *Optimization - Theory and Practice*, Springer Undergraduate Texts in Mathematics and Technology, Springer New York, 2010.
32. A.R. Forsyth, *Calculus of Variations*, Dover books on advanced mathematics, Dover Publications, 1960.
33. J.N. Franklin, *Methods of Mathematical Economics: Linear and Nonlinear Programming, Fixed-Point Theorems*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2002.
34. P.A. Fuhrmann, *Linear Systems and Operators in Hilbert Space*, Dover Books on Mathematics, Dover Publications, 2014.
35. A.V. Fursikov, *Optimal control of distributed systems. theory and applications*, American Mathematical Soc., 1999.
36. I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Dover Books on Mathematics, Dover Publications, 2012.
37. P.R. Halmos, *Finite-Dimensional Vector Spaces*, Undergraduate Texts in Mathematics, Springer New York, 2012.
38. P. Hamill, *A Student's Guide to Lagrangians and Hamiltonians*, Student's Guides, Cambridge University Press, 2014.
39. D.A. Harville, *Matrix Algebra From a Statistician's Perspective*, Springer New York, 2008.
40. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, Springer Netherlands, 2008.
41. K. Hoffman and R.A. Kunze, *Linear Algebra*, Prentice-Hall, 1971.
42. K.H. Hoffmann, G. Leugering, and F. Tröltzsch, *Optimal Control of Partial Differential Equations*, International Series of Numerical Mathematics, Birkhäuser Basel, 2012.
43. R.B. Holmes, *Geometric Functional Analysis and its Applications*, Graduate Texts in Mathematics, Springer New York, 2012.
44. R. Hoppe, *Optimization with PDE Constraints: ESF Networking Program 'OPTPDE'*, Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2014.
45. A.D. Ioffe, V.M. Tikhomirov, and K. Makowski, *Theory of Extremal Problems*, Studies in Logic and the Foundations of Mathematics, North-Holland Publishing Company, 1979.
46. V. Isakov, *Inverse Problems for Partial Differential Equations*, Applied Mathematical Sciences, Springer International Publishing, 2017.
47. K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2008.
48. J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer Berlin Heidelberg, 2013.
49. J.L. Kelley, *General Topology*, Graduate Texts in Mathematics, Springer New York, 1975.
50. J.L. Kelley and I. Namioka, *Linear Topological Spaces*, Graduate Texts in Mathematics, Springer Berlin Heidelberg, 2013.
51. D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 2000.
52. A.W. Knapp, *Basic real analysis*, Cornerstones, Birkhäuser Boston, 2007.

53. I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Volume 1, Abstract Parabolic Systems: Continuous and Approximation Theories*, Control Theory for Partial Differential Equations: Continuous and Approximation Theories, Cambridge University Press, 2000.
54. U. Ledzewicz and A. Nowakowski, *Optimality Conditions for Control Problems Governed by Abstract Semilinear Differential Equations in Complex Banach Spaces*, Journal of Applied Analysis **3** (01 Jun. 1997), no. 1, 67 – 90.
55. G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, *Trends in PDE Constrained Optimization*, International Series of Numerical Mathematics, Springer International Publishing, 2014.
56. X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Systems & Control: Foundations & Applications, Birkhäuser Boston, 2012.
57. D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*, Princeton University Press, 2012.
58. G.M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, 1996.
59. J.L. Lions, *Some Aspects of the Optimal Control of Distributed Parameter Systems*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, 1972.
60. ———, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2011.
61. D.G. Luenberger, *Optimization by Vector Space Methods*, Professional Series, Wiley, 1997.
62. D.G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, International Series in Operations Research & Management Science, Springer US, 2008.
63. R.E. Moore and M.J. Cloud, *Computational Functional Analysis*, Ellis Horwood Series in Mathematics and Its Applications, Elsevier Science, 2007.
64. B.S. Mordukhovich, *Variational analysis and generalized differentiation i: Basic theory*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2006.
65. F. Morgan, *Real Analysis*, American Mathematical Society, 2005.
66. François Murat and Luc Tartar, *On the Control of Coefficients in Partial Differential Equations*, pp. 1–8, Birkhäuser Boston, Boston, MA, 1997.
67. ———, *H-Convergence*, pp. 21–43, Springer International Publishing, Cham, 2018.
68. G. Pavliotis and A. Stuart, *Multiscale Methods: Averaging and Homogenization*, Texts in Applied Mathematics, Springer New York, 2008.
69. L. Perko, *Differential Equations and Dynamical Systems*, Texts in Applied Mathematics, Springer New York, 2013.
70. J. Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, SpringerBriefs in Optimization, Springer International Publishing, 2015.
71. M. Reed and B. Simon, *I: Functional Analysis*, Methods of Modern Mathematical Physics, Elsevier Science, 1981.
72. F. Riesz and B.S. Nagy, *Functional Analysis*, Dover Books on Mathematics, Dover Publications, 2012.
73. R.T. Rockafellar, *Convex Analysis*, Princeton Landmarks in Mathematics and Physics, Princeton University Press, 2015.
74. H. Royden and P. Fitzpatrick, *Real Analysis*, Pearson Modern Classics for Advanced Mathematics Series, Pearson, 2017.
75. W. Rudin, *Principles of Mathematical Analysis*, third ed., Springer, 1976.
76. ———, *Functional Analysis*, International series in pure and applied mathematics, McGraw-Hill, 2006.
77. S. Serovajsky, *Optimization and Differentiation*, Chapman & Hall/CRC Monographs and Research Notes in Mathematics, CRC Press, 2017.
78. D.R. Smith, *Variational Methods in Optimization*, Prentice-Hall, 1974.
79. E.D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Texts in Applied Mathematics, Springer New York, 2013.
80. M. Spivak, *Calculus*, Publish or Perish, 2008.
81. R.K. Sundaram and R. K, *A First Course in Optimization Theory*, Cambridge University Press, 1996.

82. Luc Tartar, *An introduction to the homogenization method in optimal design*, pp. 47–156, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.
83. ———, *Estimations of homogenized coefficients*, pp. 9–20, Springer International Publishing, Cham, 2018.
84. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, Graduate studies in mathematics, American Mathematical Society, 2010.
85. J.L. Troutman, *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*, Undergraduate Texts in Mathematics, Springer New York, 2012.
86. K. Yosida, *Functional Analysis*, Classics in Mathematics, Cambridge University Press, 1995.
87. I. Yousept, *Optimal Control of Partial Differential Equations Involving Pointwise State Constraints: Regularization and Applications*, Cuvillier, 2008.
88. E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, Applied Mathematical Sciences, Springer New York, 2012.