June 2020

Multigrid Methods for Elliptic Optimal Control Problems

Sijing Liu
Louisiana State University and Agricultural and Mechanical College

MULTIGRID METHODS FOR ELLIPTIC OPTIMAL CONTROL PROBLEMS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Mathematics

by
Sijing Liu
B.S., Fujian Normal University, 2011
M.S., Xiamen University, 2014
August 2020

Acknowledgments

This dissertation would not be possible without several contributions. First I want to express my gratitude to my advisor Prof. Susanne C. Brenner, and Prof. Li-yeng Sung for the guidance and support. I also would like to thank Prof. Hongchao Zhang for many conversations about the optimization techniques related to this dissertation. I also want to thank Prof. Shawn Walker for providing help for the FELICITY package that I used in this dissertation.

I would like to thank the National Science Foundation for the support under Grant DMS-16-20273.

This dissertation is dedicated to my parents and Ting Lin for their endless love and encouragement.

Table of Contents

List of Figures

Abstract

In this dissertation we study multigrid methods for linear-quadratic elliptic distributed optimal control problems.

For optimal control problems constrained by general second order elliptic partial differential equations, we design and analyze a P_1 finite element method based on a saddle point formulation. We construct a W-cycle algorithm for the discrete problem and show that it is uniformly convergent in the energy norm for convex domains. Moreover, the contraction number decays at the optimal rate of m^{-1}, where m is the number of smoothing steps. We also prove that the convergence is robust with respect to a regularization parameter. The robust convergence of V-cycle and W-cycle algorithms on general domains are demonstrated by numerical results.

For optimal control problems constrained by symmetric second order elliptic partial differential equations together with pointwise constraints on the state variable, we design and analyze symmetric positive definite P_1 finite element methods based on a reformulation of the optimal control problem as a fourth order variational inequality. We develop a multigrid algorithm for the reduced systems that appear in a primal-dual active set method for the discrete variational inequalities. The performance of the algorithm is demonstrated by numerical results.

Chapter 1
Introduction

1.1 Elliptic Optimal Control Problems

Optimal control of systems governed by partial differential equations (PDEs) are optimization problems that are subject to constraints by partial differential equations. The essential features of an optimal control problem include a cost functional, a partial differential equation constraint, a state y, a control function u and other constraints. The problem is to minimize the cost functional under all the constraints. In many cases, optimal control of partial differential equations has to be considered. For example, heat conduction, diffusion, fluid flows and many other physical phenomena can be modeled by partial differential equations. We refer to [74, 97] for more details about such optimal control problems. In this dissertation, we focus on optimal control problems with quadratic cost functional while the state is governed by a linear elliptic partial differential equation. Such a problem is called a linear-quadratic elliptic control problem.

To start with, we consider a region $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ to be heated or cooled. We are given a desired state y_d which can be treated as the desired temperature distribution in Ω. The control u is a heat source that we want to choose such that the state y is the best possible approximation to y_d. Here we assume the temperature vanishes at the boundary. Figure 1.1 is an illustration of this process in $\mathbb{R}^2$. This problem can be modeled by the following optimal control problem,

$$\min_{(y,u)} \left[\frac{1}{2} \|y - y_d\|^2_{L^2(\Omega)} + \frac{\beta}{2} \|u\|^2_{L^2(\Omega)} \right], \tag{1.1.1}$$

subject to

$$\begin{aligned} -\Delta y &= u \quad \text{in} \quad \Omega, \\ y &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{1.1.2}$$

This is a typical example of a linear-quadratic elliptic control problem with distributed control. The constant $\beta > 0$ can be viewed as a measure of how much energy is needed to implement the control u. Mathematically, the number β can also be viewed as a regularization parameter. It is also natural to consider pointwise control and state constraints, since the available energy for heating or cooling is limited and the temperature should not exceed a certain range.

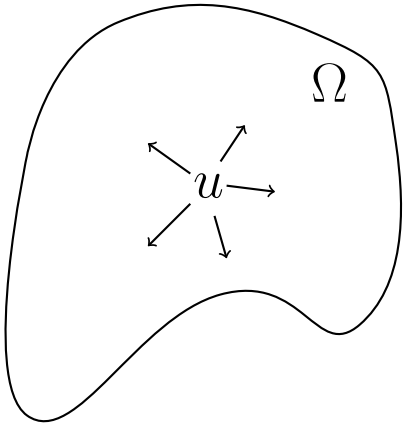

Figure 1.1. Distributed Control.

In this dissertation, we focus on the following general elliptic optimal control problem with pointwise state constraints. Let Ω be a bounded polygonal/polyhedral domain in $\mathbb{R}^n$ $(n = 2, 3)$, $y_d \in L^2(\Omega)$ and β be a positive constant, find

$$(\bar{y}, \bar{u}) = \underset{(y,u)\in K}{\operatorname{argmin}} \left[\frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L^2(\Omega)} \right], \tag{1.1.3}$$

where (y, u) belongs to $K \subset H^1_0(\Omega) \times L^2(\Omega)$ if and only if

$$a(y, v) = (u, v)_{L^2(\Omega)} \quad \forall v \in H^1_0(\Omega) \tag{1.1.4}$$

and

$$y \leq \psi \quad \text{a.e. in} \quad \Omega. \tag{1.1.5}$$

Here ψ belongs to $W^{2,\infty}(\Omega) \cap H^3(\Omega)$ and $\psi > 0$ on $\partial\Omega$. The bilinear form $a(\cdot,\cdot)$ is defined by

$$a(y,v) = \int_\Omega \nabla y \cdot \nabla v \; dx + \int_\Omega [(\boldsymbol{\zeta} \cdot \nabla y)v - (\boldsymbol{\zeta} \cdot \nabla v)y] \; dx + \int_\Omega \gamma y v \; dx, \qquad (1.1.6)$$

where the vector field $\boldsymbol{\zeta} \in [W^{1,\infty}(\Omega)]^n$ and the function $\gamma \in L^\infty(\Omega)$ is nonnegative. If $\boldsymbol{\zeta} \neq 0$ then the constraint (1.1.4) is the weak form of a general second order PDE with an advective/convective term.

1.2 Literature Review

In the absence of pointwise state constraints (1.1.5), the optimal control problem (1.1.3)-(1.1.4) can be characterized by the following first order optimality system (cf. [97, 74, 62]),

$$\begin{aligned} a(q,\bar{p}) &= (\bar{y} - y_d, q)_{L^2(\Omega)} \quad \forall q \in H_0^1(\Omega), && (1.2.1a)\\ \bar{p} + \beta\bar{u} &= 0, && (1.2.1b)\\ a(\bar{y},z) &= (\bar{u},z)_{L^2(\Omega)} \qquad \forall z \in H_0^1(\Omega). && (1.2.1c) \end{aligned}$$

Multigrid methods for the system (1.2.1) were well-studied and can be categorized into at least two main approaches. First, notice that (1.2.1b) is simple thus we can replace the control $\bar{u}$ in (1.2.1c) by $-\frac{1}{\beta}\bar{p}$, the resulting saddle point problem is

$$\begin{aligned} a(q,\bar{p}) - (q,\bar{y})_{L^2(\Omega)} &= -(q,y_d)_{L^2(\Omega)} \quad \forall q \in H_0^1(\Omega), && (1.2.2a)\\ -(\bar{p},z)_{L^2(\Omega)} - \beta a(\bar{y},z) &= 0 \qquad\qquad\quad \forall z \in H_0^1(\Omega). && (1.2.2b) \end{aligned}$$

Multigrid methods that are directly applied to (1.2.1) or (1.2.2) belong to the class of all-at-once methods where all the unknowns are solved simultaneously. This approach can be found in [17, 91, 96, 2, 95, 18, 94] and the references therein. Meanwhile, multigrid methods for general saddle point problems have been investigated in [28, 31, 29, 24, 92, 108, 99, 100]. However, the multigrid convergence

results in [17, 91, 96, 2, 95, 18, 94] are not established in the energy norm. This issue was addressed in [29, 28, 31] for general saddle point problems. Also, the analyses of the multigrid convergence results in [17, 91, 96, 2, 95, 18, 94] often require Ω to be convex. A recent result on arbitrary domains is established in [96] when $\boldsymbol{\zeta} = \mathbf{0}$. The other important feature of the multigrid methods for optimal control problems is the robustness with respect to the regularization parameter β. When β is small, the performance of multigrid methods often deteriorates. In the case when $\boldsymbol{\zeta} = \mathbf{0}$, multigrid methods that are robust with respect to β can be found in [91, 96]. However, the contraction numbers decay at the rate $O(m^{-\frac{1}{2}})$ in [91, 96] where m is the number of pre-smoothing steps and their results cannot be directly extended to the case when $\boldsymbol{\zeta} \neq \mathbf{0}$. Second, we can eliminate one more unknown in (1.2.2) resulting in a single equation which involves the control $\bar{u}$ or the state $\bar{y}$. This can be done for a large class of optimal control problems (cf. [74]). Multigrid methods that are applied to this single equation belong to the other approach [13, 14, 85, 93, 56]. The advantage of this approach is that we can exploit the well-known multigrid theory for elliptic PDEs and use this as building blocks for the outer iterative methods.

On the other hand, if (1.1.5) is present, the elliptic optimal control problem (1.1.3)-(1.1.5) is equivalent to a fourth order variational inequality. Multigrid methods for variational inequalities can be found in [70, 71, 65, 60, 6, 53]. We refer to [47, 17, 16] and the references therein for multigrid methods designed for constrained optimal control problems. In most cases (cf. [34, 47, 11]), an outer optimization method is needed to handle the constraints while a reduced system needs to be solved during each outer iteration. Several optimization methods were proven to be efficient for solving constrained optimal control problems, for example, primal-dual active set (PDAS) algorithms [63, 10] and interior-point methods

[79, 104]. The challenge of this approach is that the reduced system becomes harder to solve when the mesh size h of the discretization decreases, especially when a three dimensional problem is considered. In this situation, fast solvers were studied to remedy this issue, for example, multigrid methods were developed for general second order elliptic problems in [72, 69]. However, the reduced system we consider here is fourth-order and hence more difficult to analyze. The other issue to address is that the pointwise state constraint (1.1.5) imposes difficulties for constructing and analyzing numerical methods. Specifically, the Lagrange multiplier associated to (1.1.3) is only a measure in general (cf. [41, 33]). This low regularity of the Lagrange multiplier causes the difficulties. Recent results on the finite element methods of elliptic optimal control problems with pointwise state constraints can be found in [33, 42, 45, 34] and the references therein. In general, multigrid methods for state constrained optimal control problems are difficult to analyze hence not much work has been done.

1.3 Outline of the Dissertation

The main goal of this dissertation is to construct and analyze multigrid methods for (1.1.3)-(1.1.5).

In Chapter 2 we review some fundamentals that are needed in this dissertation. We briefly review the concept of Sobolev spaces, the theory of saddle point problems and the elliptic regularity for second order and fourth order PDEs. We also review the existence and uniqueness of solution for elliptic optimal control problems (1.1.3)-(1.1.5) and derive the first order optimality condition. Iterative methods including the Richardson iteration, the Gauss-Seidel iteration, the minimal residual (MINRES) algorithm and the generalized minimal residual (GMRES) algorithm are described. These are useful in the construction of the smoothing steps

of multigrid methods. We then briefly review the P_1 finite element methods. Lastly we review the structure and ingredients of basic multigrid algorithms.

In Chapter 3 we introduce P_1 finite element methods for (1.1.3)-(1.1.5). Two approaches are considered, namely, the saddle point problem (SPP) approach and the symmetric positive definite (SPD) approach. For SPD approach, a mass lumping mesh-dependent inner product is introduced to enable efficient implementation of multigrid solvers. We prove the convergence in energy norm for both approaches. For the SPP approach, we track the regularization parameter β in the error analysis. This is essential for the convergence analysis of the W-cycle algorithm in Chapter 4. Numerical results are presented to support the theoretical results.

In Chapter 4 we propose an all-at-once multigrid method for (1.1.3)-(1.1.4) based on the P_1 finite element methods that are introduced in Chapter 3. We prove that the W-cycle algorithms are uniformly convergent and robust with respect to β on convex domains while the contraction numbers decay at the rate $O(m^{-1})$, where m is the number of pre-smoothing steps and post-smoothing steps. This result is established in the energy norm. Numerical results are presented to illustrate the performance of W-cycle and V-cycle algorithms. We also compare the performance of our methods to preconditioned GMRES. As far as we know, this multigrid method is the first one that are provably robust with respect to the regularization parameter β when the elliptic PDE constraint (1.1.4) involves an advective/convective term. The materials in this chapter come from [30].

In Chapter 5 we propose a PDAS algorithm with multigrid solver for (1.1.3)-(1.1.5) when $\boldsymbol{\zeta} = 0$ and $\gamma = 0$. We briefly review the PDAS algorithm for the discretized problem. Then we construct a W-cycle multigrid algorithm to solve the reduced system that appears during each outer PDAS iteration. This multigrid solver is efficient since we utilize a mass-lumping technique. We observe the

convergence of our W-cycle multigrid algorithm numerically on arbitrary domains. We also compare the performance of our methods to preconditioned MINRES. Numerical results are shown for various examples.

Chapter 2
Preliminaries

2.1 Sobolev Spaces

In this section we briefly review the concept of Sobolev spaces. We refer to [32, 1, 55, 103] for more details.

Let f be a Lebesgue measurable real-valued function on a given domain $\Omega \in \mathbb{R}^d$, where d is a positive integer. We assume Ω is a Lebesgue measurable subset of $\mathbb{R}^d$ with non-empty interior. We denote the Lebesgue integral of f on Ω by

$$\int_\Omega f(x)\ dx.$$

For $1 \leq p < \infty$, let

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p\ dx\right)^{\frac{1}{p}},$$

and for $p = \infty$,

$$\|f\|_{L^\infty(\Omega)} = ess\ sup\{|f(x)| :\ x \in \Omega\}.$$

We define the Lebesgue spaces for $1 \leq p \leq \infty$,

$$L^p = \{f : \|f\|_{L^p(\Omega)} < \infty\}. \tag{2.1.1}$$

We identify two functions f and g in $L^p(\Omega)$ if they differ only on a set of measure zero, namely $\|f - g\|_{L^p(\Omega)} = 0$. For $1 \leq p \leq \infty$, $L^p(\Omega)$ is a Banach space.

Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ be a multi-index, where α_i is a non-negative integer, $i = 1, 2, \cdots, n$. We denote the length of α as

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

For $\phi(x) \in C^\infty(\Omega)$, we denote the usual partial derivative as

$$D^\alpha \phi(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \phi(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} \phi(x).$$

Definition 2.1. *We define the set of locally integrable functions as*

$$L^1_{loc}(\Omega) = \{f : f \in L^1(K) \quad \forall \text{ compact } K \subset \text{ interior } \Omega\}. \tag{2.1.2}$$

Definition 2.2. *We say that $f \in L^1_{loc}(\Omega)$ has a weak derivative $D^\alpha_w f$ if there exists a function $g \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega g(x)\phi(x)\ dx = (-1)^{|\alpha|}\int_\Omega f(x)D^\alpha\phi(x)\ dx \quad \forall \phi \in C^\infty_0(\Omega). \tag{2.1.3}$$

Here $C^\infty_0(\Omega)$ denotes the space of infinitely differentiable functions with compact support in Ω. We denote $D^\alpha_w f = g$.

For $\psi \in C^{|\alpha|}(\Omega)$, $D^\alpha_w\psi$ exists and coincides with $D^\alpha\psi$. Therefore we ignore the differences between $D^\alpha_w\psi$ and $D^\alpha\psi$ from now on.

Definition 2.3. *Let k be a non-negative integer, suppose $f \in L^1_{loc}(\Omega)$ and the weak derivative $D^\alpha f$ exists for all $|\alpha| \le k$. Define the Sobolev norm*

$$\|f\|_{W^k_p(\Omega)} := (\sum_{|\alpha|\le k} \|D^\alpha f\|^p_{L^p(\Omega)})^{\frac{1}{p}} \tag{2.1.4}$$

if $1 \le p < \infty$, and in the case $p = \infty$,

$$\|f\|_{W^k_\infty(\Omega)} := \max_{|\alpha|\le k} \|D^\alpha f\|_{L^\infty(\Omega)}. \tag{2.1.5}$$

We then define the Sobolev spaces as

$$W^k_p(\Omega) = \{f \in L^1_{loc}(\Omega) : \|f\|_{W^k_p(\Omega)} < \infty\}, \quad 1 \le p \le \infty. \tag{2.1.6}$$

Definition 2.4. *Let k be a non-negative integer, suppose $f \in W^k_p(\Omega)$, we define the Sobolev semi-norm as*

$$|f|_{W^k_p(\Omega)} := (\sum_{|\alpha|=k} \|D^\alpha f\|^p_{L^p(\Omega)})^{\frac{1}{p}} \tag{2.1.7}$$

if $1 \le p < \infty$, and in the case $p = \infty$,

$$|f|_{W^k_\infty(\Omega)} := \max_{|\alpha|=k} \|D^\alpha f\|_{L^\infty(\Omega)}. \tag{2.1.8}$$

Theorem 2.5. *The Sobolev space $W_p^k(\Omega)$ is a Banach space.*

One can easily see $W_p^0(\Omega) = L^p(\Omega)$. We denote $W_2^k(\Omega)$ by $H^k(\Omega)$ and denote its norm and semi-norm by $\|\cdot\|_{H^k(\Omega)}$ and $|\cdot|_{H^k(\Omega)}$. One can show that (cf. [32, 1]) $H^k(\Omega)$ is a Hilbert space under the following inner-product,

$$(v, w)_{H^k(\Omega)} = \sum_{|\alpha| \le k} \int_\Omega D^\alpha v \, D^\alpha w \, dx.$$

Definition 2.6. *Let k be a non-negative integer. We define $H_0^k(\Omega)$ to be the closure of $C_0^\infty(\Omega)$ under the norm $\|\cdot\|_{H^k(\Omega)}$.*

Definition 2.7. *Let k be a non-negative integer. We define $H^{-k}(\Omega)$ as the dual space of $H_0^k(\Omega)$. The norm in $H^{-k}(\Omega)$ is defined as*

$$\|u\|_{H^{-k}(\Omega)} = \sup_{v \in H_0^k(\Omega), v \neq 0} \frac{\langle u, v \rangle}{\|v\|_{H^k(\Omega)}}, \tag{2.1.9}$$

where $\langle \cdot, \cdot \rangle$ is the usual paring between $H_0^k(\Omega)$ and its dual space.

Definition 2.8. *Let Ω be an open set in $\mathbb{R}^d$. For a real number $l = k + \lambda$ where k is a nonnegative integer and $\lambda \in (0, 1)$. We define*

$$W_2^l(\Omega) = \{f \in L^2(\Omega) : D^s f \in L^2(\Omega) \ \ for \ \ |s| \le k \ \ and \ \ I_\lambda(D^s f) < \infty\}, \tag{2.1.10}$$

where

$$I_\lambda(D^s f) = \int_\Omega \int_\Omega \frac{|D^s f(x) - D^s f(y)|}{|x - y|^{d+2\lambda}} \, dxdy. \tag{2.1.11}$$

Remark 2.9. *$W_2^l(\Omega)$ is a Hilbert space with respect to the inner product*

$$\begin{aligned}(v, w)_{W_2^l(\Omega)} = &\sum_{|s| \le k} \int_\Omega D^s v D^s w \, dx \\ &+ \sum_{|s| \le k} \int_\Omega \int_\Omega \frac{(D^s v(x) - D^s v(y))(D^s w(x) - D^s w(y))}{|x - y|^{d+2\lambda}} \, dxdy.\end{aligned}$$

We denote the Hilbert space $W_2^l(\Omega)$ as $H^l(\Omega)$ (cf. [103, Theorem 5.3]).

We need the following density theorems (cf. [81, 1]) and embedding theorems to analyze the convergence of finite element methods in Chapter 3.

Theorem 2.10. *Let Ω be any open set. Then $C^\infty(\Omega) \cap W_p^k(\Omega)$ is dense in $W_p^k(\Omega)$ for $1 \leq p < \infty$.*

Theorem 2.11. *Let Ω be any Lipschitz open set. Then $C^\infty(\bar{\Omega})$ is dense in $W_p^k(\Omega)$ for $1 \leq p < \infty$.*

Theorem 2.12 (Sobolev Embedding [1, Theorem 4.12])**.** *Assume that Ω is a (bounded or unbounded) open set of $\mathbb{R}^d$ with a Lipschitz continuous boundary and $1 \leq p < \infty$. Then the following continuous embeddings hold:*

- *If $0 \leq kp < d$, then $W_p^k(\Omega) \subset L^{p^\star}(\Omega)$ for $p^\star = dp/(d-kp)$;*
- *If $kp = d$, then $W_p^k(\Omega) \subset L^q(\Omega)$ for $p \leq q < \infty$;*
- *If $kp > d$, then $W_p^k(\Omega) \subset C(\bar{\Omega})$.*

Notice that it is meaningless to write $v|_{\partial\Omega}$ for $v \in H^k(\Omega)$ since we cannot define functions in $H^k(\Omega)$ on a subset of measure zero. In order to remedy this issue, we introduce the concept of trace.

Theorem 2.13 (Trace Theorem [84, Theorem 1.3.1])**.** *Let Ω be a bounded open set of $\mathbb{R}^d$ with smooth boundary $\partial\Omega$ and let $k > \frac{1}{2}$.*

- *There exists a unique linear continuous map $\gamma_0 : \ H^k(\Omega) \to H^{k-\frac{1}{2}}(\partial\Omega)$ such that $\gamma_0 v = v|_{\partial\Omega}$ for each $v \in H^k(\Omega) \cap C(\bar{\Omega})$.*
- *There exists a linear continuous map $\mathcal{R}_0 : \ H^{k-\frac{1}{2}}(\partial\Omega) \to H^k(\Omega)$ such that $\gamma_0 \mathcal{R}_0 \psi = \psi$ for each $\psi \in H^{k-\frac{1}{2}}(\partial\Omega)$.*

Remark 2.14. *In particular, the following inequality holds,*

$$\|v\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^1(\Omega)}, \tag{2.1.12}$$

where $C > 0$ is a constant and v on the left-hand side is understood in the sense of $\gamma_0 v$. Note that this estimate is also valid for a bounded polygonal open subset Ω of $\mathbb{R}^d$. Moreover, we can characterize the space $H_0^1(\Omega)$ by using the trace operator,

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : \gamma_0 v = 0 \quad on \quad L^2(\partial\Omega)\}.$$

For a subtle description of trace on polygonal domains, we refer to [1, 55].

Theorem 2.15 (Poincaré Inequality [32, Theorem 5.3.5])**.** *Let Ω be a bounded connected open set of $\mathbb{R}^d$. Then there exists a constant $C_\Omega > 0$ such that*

$$\|v\|_{L^2(\Omega)} \leq C_\Omega |v|_{H^1(\Omega)} \tag{2.1.13}$$

for each $v \in H_0^1(\Omega)$.

Theorem 2.16 (Riesz Representation Theorem [32, Theorem 2.4.2])**.** *Any continuous linear functional L on a Hilbert space $(H, (\cdot, \cdot))$ can be represented uniquely as*

$$L(v) = (u, v) \tag{2.1.14}$$

for some $u \in H$. Furthermore, we have

$$\|L\|_{H'} = \|u\|_H. \tag{2.1.15}$$

For more general discussion on Riesz Representation Theorem, we refer to [86, Section 13, Theorem 25] and [87, Theorem 2.14]. Riesz Representation Theorem is often used to prove the existence and uniqueness of solutions for symmetric variational problems. For nonsymmetric variational problems, we need the following theorem.

Theorem 2.17 (Lax-Milgram [84, Theorem 5.1.1])**.** *Given a Hilbert space $(V, (\cdot, \cdot))$, a continuous, coercive bilinear form $a(\cdot, \cdot)$ and a continuous linear functional $F \in$*

V', there exists a unique $u \in V$ such that

$$a(u, v) = \langle F, v \rangle \qquad \forall v \in V. \tag{2.1.16}$$

Moreover we have

$$\|u\|_V \leq \frac{1}{\alpha} \|F\|_{V'}, \tag{2.1.17}$$

where α is the coercivity constant and V' is the dual space of V.

Remark 2.18. *A bilinear form $a(\cdot, \cdot)$ on a linear space V is a mapping $a : V \times V \to \mathbb{R}$ such that each of the maps $v \to a(v, w)$ and $w \to a(v, w)$ is a linear form on V.*

A bilinear form $a(\cdot, \cdot)$ on a normed linear space H is said to be bounded (or continuous) if $\exists\ C < \infty$ such that

$$|a(v, w)| \leq C \|v\|_H \|w\|_H \qquad \forall v, w \in H,$$

and coercive on $V \subset H$ if $\exists\ \alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|_H^2 \qquad \forall v \in V.$$

2.2 Saddle Point Problems

In this section we briefly discuss the saddle point problems. Saddle point problems arise in many areas including optimal control, constrained optimization, fluid dynamics etc. We refer to [8, 15] for a thorough discussion of saddle point problems and their numerical approximation.

We start with a simple example in algebraic setting (cf. [8]). Consider the following linear system

$$\begin{bmatrix} A & B^T \\ B & O \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{2.2.1}$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times n}$, O is zero matrix and A is symmetric positive semidefinite. This form of linear system arises as the first-order optimality condition of

the constrained optimization problem,

$$\min \ \frac{1}{2}x^T Ax - f^T x \tag{2.2.2}$$

$$\text{s.t.} \quad Bx = g. \tag{2.2.3}$$

Here y in (2.2.1) represents the vector of Lagrange multipliers. Any solution (x^*, y^*) of (2.2.1) is a saddle point of the Lagrangian

$$\mathcal{L}(x, y) = \frac{1}{2}x^T Ax - f^T x + (Bx - g)^T y. \tag{2.2.4}$$

Therefore we call (2.2.1) a "saddle point problem". Note that a saddle point is a point (x^*, y^*) that satisfies

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*) \quad \forall x \in \mathbb{R}^n, \ y \in \mathbb{R}^m. \tag{2.2.5}$$

A more general saddle point problem is of the form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{2.2.6}$$

where C is also symmetric positive semidefinite. This system often arises in stabilized finite element methods and PDE-constrained optimization problems.

On the other hand, consider the following mixed variational problem,

$$a(u, v) + b(v, p) = \langle f, v \rangle, \quad \forall v \in V, \tag{2.2.7a}$$

$$b(u, q) \qquad\quad = \langle g, q \rangle, \quad \forall q \in Q, \tag{2.2.7b}$$

where $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ and $b(\cdot, \cdot) : V \times Q \to \mathbb{R}$ are continuous bilinear form (cf. Remark 2.18)

$$a(u, v) \leq C_1 \|u\|_V \|v\|_V, \quad \forall u, v \in V,$$

$$b(v, q) \leq C_2 \|v\|_V \|q\|_Q, \quad \forall v \in V, q \in Q,$$

$f \in V'$ and $q \in Q'$. The system (2.2.7) is a general variational problem that arises in many areas, for instance, fluid dynamics. After we discretize (2.2.7) by certain numerical methods, the discretized system is of the form (2.2.1). Hence we also call (2.2.7) a "saddle point problem". Previous work on the existence and uniqueness of the solution of (2.2.7) dates back to 1970s. Brezzi proved the following theorem.

Theorem 2.19 (Brezzi [37]). *The variational problem* (2.2.7) *is well-posed if and only if the following conditions hold*

$$\inf_{u\in V_0}\sup_{v\in V_0}\frac{a(u,v)}{\|u\|_V\|v\|_V}=\inf_{v\in V_0}\sup_{u\in V_0}\frac{a(u,v)}{\|u\|_V\|v\|_V}\equiv\alpha>0, \tag{2.2.8a}$$

where $V_0=\{v\in V : b(v,q)=0 \quad \forall q\in Q\}$, *and*

$$\inf_{q\in Q}\sup_{v\in V}\frac{b(v,q)}{\|v\|_V\|q\|_Q}\equiv\beta>0. \tag{2.2.8b}$$

Conditions (2.2.8) are called Babuška-Brezzi conditions or BB conditions in short. Other names, for example, Ladyzhenskaya-Babuška-Brezzi conditions or LBB conditions and inf-sup conditions can be found in the literature.

Alternatively, Babuška proposed the following framework. Let $(U,(\cdot,\cdot)_U)$ and $(V,(\cdot,\cdot)_V)$ be two Hilbert spaces. Let $\mathcal{B}: U\times V\to\mathbb{R}$ be a continuous bilinear form. Consider the following variational problem: Find $u\in U$ such that

$$\mathcal{B}(u,v)=\langle f,v\rangle,\quad \forall v\in V, \tag{2.2.9}$$

where $f\in V'$.

Theorem 2.20 (Babuška [4, 5]). *The problem* (2.2.9) *is well-posed if and only if the following BB conditions hold:*

$$\inf_{u\in U}\sup_{v\in V}\frac{\mathcal{B}(u,v)}{\|u\|_U\|v\|_V}>0,\quad \inf_{v\in V}\sup_{u\in U}\frac{\mathcal{B}(u,v)}{\|u\|_U\|v\|_V}>0, \tag{2.2.10}$$

furthermore if (2.2.10) *hold, then*

$$\inf_{u\in U}\sup_{v\in V}\frac{\mathcal{B}(u,v)}{\|u\|_U\|v\|_V}=\inf_{v\in V}\sup_{u\in U}\frac{\mathcal{B}(u,v)}{\|u\|_U\|v\|_V}\equiv\alpha>0. \tag{2.2.11}$$

We connect Babuška theory with Brezzi theory by the following argument (cf. [106]). Setting $\mathcal{B}((u,p),(v,q)) = a(u,v)+b(v,p)+b(u,q)$ then (2.2.7) is equivalent to

$$\mathcal{B}((u,p),(v,q)) = \langle f,v\rangle + \langle g,q\rangle, \quad \forall (v,q)\in V\times Q. \tag{2.2.12}$$

Hence, by Babuška theory, (2.2.7) is well-posed if and only if

$$\begin{aligned} &\inf_{(u,p)\in V\times Q}\sup_{(v,q)\in V\times Q}\frac{\mathcal{B}((u,p),(v,q))}{\|(u,p)\|_{V\times Q}\|(v,q)\|_{V\times Q}} = \\ &\inf_{(v,q)\in V\times Q}\sup_{(u,p)\in V\times Q}\frac{\mathcal{B}((u,p),(v,q))}{\|(u,p)\|_{V\times Q}\|(v,q)\|_{V\times Q}} \equiv \gamma > 0, \end{aligned} \tag{2.2.13}$$

where $\|(v,q)\|^2_{V\times Q} = \|v\|^2_V + \|q\|^2_Q$ for all $(v,q)\in V\times Q$.

It can be shown that (2.2.13) is equivalent to (2.2.8). More details can be found in [106]. The concise formulation (2.2.12), which can be applied to problem (2.2.6), will be used throughout this dissertation.

2.3 Existence and Uniqueness of Solution for Optimal Control Problems

In this section, we discuss the existence and uniqueness of solution for optimal control problems and derive the first order optimality system. The references [74] and [97] contain a detailed discussion of this topic. Minimizing sequence technique is often used to prove the well-posedness (cf. [74, Theorem 1.1], [97, Theorem 2.14]) and the state is often eliminated by a control-to-state operator. We provide a simple proof which eliminates the control in this section (cf. [75]). It is natural to do so since the constraints are imposed on the state directly. The derivation of the first order optimality condition is subtle and we only provide enough details which enable us to proceed in the numerical analysis. We refer to [74, 97, 41, 42] for more general cases.

Let us consider problem (1.1.3)-(1.1.5). It is easy to see that

$$a(u,v) \le C_1 \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}, \tag{2.3.1}$$

$$a(v,v) \ge C_2 \|v\|_{H^1(\Omega)}, \tag{2.3.2}$$

where C_1 and C_2 are two positive constants. Here we use the fact $\boldsymbol{\zeta} \in [W^{1,\infty}(\Omega)]^n$ and the function $\gamma \in L^\infty(\Omega)$ is nonnegative. Hence (1.1.4) is well-posed by Lax-Milgram Theorem (cf. Theorem 2.17).

Let V be the subspace of $H_0^1(\Omega)$ defined by

$$V = \{y \in H_0^1(\Omega) : \mathcal{L}y \in L^2(\Omega)\}, \tag{2.3.3}$$

where $\mathcal{L}y = -\Delta y + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y$ and Δy is understood in the sense of weak derivative (cf. Definition 2.2).

Remark 2.21. *The reason why we define V is that, when Ω is a general domain, we can only conclude y belongs to V when y satisfies* (1.1.4)*. Indeed, we have*

$$\begin{aligned} \int_\Omega y(\Delta\phi)\, dx &= \int_\Omega (-\nabla y)\nabla\phi\, dx \\ &= \int_\Omega (-u + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y)\phi\, dx, \quad \forall \phi \in C_0^\infty(\Omega), \end{aligned} \tag{2.3.4}$$

which means that $\Delta y = -u + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y$ in the weak sense. Thus we can conclude $y \in V$. Conversely, if we have $y \in V$, we obtain

$$\int_\Omega (-\nabla y)\nabla\phi\, dx = \int_\Omega y(\Delta\phi)\, dx = \int_\Omega f_1 \phi\, dx, \quad \forall \phi \in C_0^\infty(\Omega), \tag{2.3.5}$$

where f_1 equals to Δy in the weak sense. Since $y \in V$ we know $f := [-f_1 + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y] \in L^2(\Omega)$. Combining with the fact that $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ (cf. Theorem 2.10), we know y satisfies a second order PDE of the form $-\Delta y + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y = f$ (in the weak sense).

Therefore we can see the space V is the natural space to look for the solution. Due to elliptic regularity for polygonal/polyhedral domains (cf. [44, 55]), V is a

subspace of $H^{1+\alpha}(\Omega) \cap H^2_{loc}(\Omega) \cap H^1_0(\Omega)$ for some $\alpha \in (\frac{1}{2}, 1]$. Then it follows from Theorem 2.12 that we can identify V with a subset of $C(\bar{\Omega})$ and ignore the a.e. in (1.1.5).

We rewrite (1.1.3) as

$$\bar{y} = \underset{y \in K}{\operatorname{argmin}} \left[\frac{1}{2}(y, y)_{L^2(\Omega)} + \frac{\beta}{2}(\mathcal{L}y, \mathcal{L}y)_{L^2(\Omega)} - (y, y_d)_{L^2(\Omega)} \right], \tag{2.3.6}$$

where

$$K = \{y \in V : y \leq \psi \quad \text{in } \Omega\}. \tag{2.3.7}$$

Let the inner product $(\!(\cdot, \cdot)\!)$ on V be defined by

$$(\!(v, w)\!) = (v, w)_{L^2(\Omega)} + \beta(\mathcal{L}v, \mathcal{L}w)_{L^2(\Omega)}. \tag{2.3.8}$$

Lemma 2.22. *The space $(V, (\!(\cdot, \cdot)\!))$ is a Hilbert space.*

Proof. It is trivial that $(\!(\cdot, \cdot)\!)$ is an inner product on V. We only need to prove that this space is complete. First denote the new norm by $|||v||| = \sqrt{(\!(v, v)\!)}$. Suppose $\{f_n\}$ is a Cauchy sequence in $(V, (\!(\cdot, \cdot)\!))$. Notice that

$$|||f_m - f_n|||^2 = \|f_m - f_n\|^2_{L^2(\Omega)} + \beta\|\mathcal{L}f_m - \mathcal{L}f_n\|^2_{L^2(\Omega)}. \tag{2.3.9}$$

This implies that $\{\mathcal{L}f_n\}$ is also a Cauchy sequence in standard $L^2(\Omega)$. Due to the completeness of standard $L^2(\Omega)$, there exists a $f \in L^2(\Omega)$ such that $\|\mathcal{L}f_n - f\|_{L^2(\Omega)} \to 0$ as $n \to \infty$. By Lax-Milgram (cf. Theorem 2.17), there exists a unique $g \in H^1_0(\Omega)$ such that $\mathcal{L}g = f$. Notice that $g \in V$ (cf. Remark 2.21) and

$$\|f_n - g\|_{L^2(\Omega)} \leq \|f_n - g\|_{H^1(\Omega)} \leq C\|\mathcal{L}f_n - f\|_{L^2(\Omega)}, \tag{2.3.10}$$

we conclude $|||f_n - g||| \to 0$ as $n \to \infty$.

This implies the completeness of $(V, (\!(\cdot, \cdot)\!))$. □

Lemma 2.23. *There exists a unique $\tilde{y} \in V$ such that*

$$(\tilde{y}, z)_{L^2(\Omega)} + \beta(\mathcal{L}\tilde{y}, \mathcal{L}z)_{L^2(\Omega)} = (y_d, z)_{L^2(\Omega)} \quad \forall\ z \in V. \tag{2.3.11}$$

Proof. Notice we can write (2.3.11) as

$$(\!(\tilde{y}, z)\!) = F(z) \tag{2.3.12}$$

where $F(z) = (y_d, z)_{L^2(\Omega)}$. F is linear and bounded. Then by Riesz Representation Theorem (cf. Theorem 2.16), there exists a unique $\tilde{y}$ that satisfies (2.3.11). □

We then need the following projection theorem to prove the well-posedness of (2.3.6)-(2.3.7).

Theorem 2.24 (Projection Theorem [32, Proposition 2.3.1])**.** *Let K be a non-empty closed convex subset of a Hilbert space H. Given any $v \in H$, there exists a unique $w^* \in K$ such that*

$$\|v - w^*\| = \min_{w \in K} \|v - w\|. \tag{2.3.13}$$

Proof. Let $\delta = \inf_{w \in K} \|v - w\|$. There exists a minimizing sequence $\{w_n\} \in K$ such that

$$\lim_{n \to \infty} \|v - w_n\| = \delta.$$

We want to show that $\{w_n\}$ is a Cauchy sequence, then there exists $w^* \in \bar{K} = K$ such that $w_n \to w^*$. Continuity of the norm implies that $\|v - w^*\| = \delta$.

To prove that $\{w_n\}$ is a Cauchy sequence, note that the following parallelogram law (cf. [32, (2.2.8)]) holds

$$\|a + b\|^2 + \|a - b\|^2 = 2\|a\|^2 + 2\|b\|^2. \tag{2.3.14}$$

Apply the parallelogram law to $a = v - w_m$ and $b = v - w_n$, we have

$$\begin{aligned} &\|(v - w_m) + (v - w_n)\|^2 + \|(v - w_m) - (v - w_n)\|^2 \\ &= 2\|v - w_m\|^2 + 2\|v - w_n\|^2, \end{aligned}$$

hence

$$\|w_m - w_n\|^2 + 4\|v - \frac{w_m + w_n}{2}\|^2$$
$$= 2\|v - w_m\|^2 + 2\|v - w_n\|^2.$$

Notice that $\frac{w_m+w_n}{2} \in K$ since K is convex, thus

$$\|w_m - w_n\|^2 \leq 2\|v - w_m\|^2 + 2\|v - w_n\|^2 - 4\delta^2.$$

Then $\|w_m - w_n\| \to 0$ as $m, n \to \infty$.

To prove the uniqueness, suppose $w^*, z^* \in K$ such that

$$\|v - w^*\| = \delta = \|v - z^*\|.$$

Apply the parallelogram law to $a = v - w^*$ and $b = v - z^*$ we have

$$\|w^* - z^*\|^2 \leq 2\|v - w^*\|^2 + 2\|v - z^*\|^2 - 4\delta^2$$
$$= 2\delta^2 + 2\delta^2 - 4\delta^2 = 0$$

which implies $w^* = z^*$. □

Theorem 2.25. *There exists a unique solution $\bar{y}$ of the problem* (2.3.6)-(2.3.7).

Proof. First we notice that K is closed and convex. Since $\psi > 0$ on $\partial\Omega$ and $y|_{\partial\Omega} = 0$, we can conclude that K is also nonempty. We can rewrite (2.3.6) as

$$\begin{aligned}
\bar{y} &= \operatorname*{argmin}_{y \in K} \left[\frac{1}{2}((y, y)) - ((\tilde{y}, y)) \right] \\
&= \operatorname*{argmin}_{y \in K} \left[\frac{1}{2}((y - \tilde{y}, y - \tilde{y})) - \frac{1}{2}((\tilde{y}, \tilde{y})) \right] \\
&= \operatorname*{argmin}_{y \in K} \frac{1}{2} |||y - \tilde{y}|||^2 .
\end{aligned}$$

Then the solution will be the projection of $\tilde{y}$ onto K. The existence and uniqueness of the projection is guaranteed by Theorem 2.24. □

2.3.1 First Order Optimality Condition

In the absence of the pointwise state constraints, for simplicity, we define a control-to-operator $S : L^2(\Omega) \to L^2(\Omega), u \mapsto y(u)$ (cf. [97, Section 2.5]) and rewrite (1.1.3) as

$$\bar{u} = \underset{u \in L^2(\Omega)}{\operatorname{argmin}} \left[\frac{1}{2} \|Su - y_d\|^2_{L^2(\Omega)} + \frac{\beta}{2} \|u\|^2_{L^2(\Omega)} \right]. \tag{2.3.15}$$

Follow along the same lines as the preceding argument, we can show that the unique solution $\bar{u}$ of (2.3.15) can be characterized by

$$(S\bar{u}, Sv)_{L^2(\Omega)} + \beta(\bar{u}, v)_{L^2(\Omega)} = (y_d, Sv)_{L^2(\Omega)} \quad \forall\ v \in L^2(\Omega). \tag{2.3.16}$$

Define the adjoint operator S^* as

$$(Su, v)_{L^2(\Omega)} = (u, S^*v)_{L^2(\Omega)} \quad \forall u, v \in L^2(\Omega). \tag{2.3.17}$$

Then define $\bar{p} \in H^1_0(\Omega)$ (adjoint state) by

$$\bar{p} = S^*(\bar{y} - y_d), \tag{2.3.18}$$

thus we can write (2.3.16) as

$$(\bar{p}, v)_{L^2(\Omega)} + \beta(\bar{u}, v)_{L^2(\Omega)} = 0 \quad \forall v \in L^2(\Omega), \tag{2.3.19}$$

which is

$$\bar{p} + \beta\bar{u} = 0. \tag{2.3.20}$$

To find the explicit expression of S^*, we need the following lemma.

Lemma 2.26 ([97, Lemma 2.23])**.** *Let function $z, u \in L^2(\Omega)$, let y and p denote, respectively, the weak solutions to the elliptic boundary value problems,*

$$\begin{aligned} -\Delta y + \boldsymbol{\zeta} \cdot \nabla y + \nabla \cdot (\boldsymbol{\zeta} y) + \gamma y &= u \quad in \quad \Omega \\ y &= 0 \quad on \quad \partial\Omega \end{aligned}$$

and

$$\begin{aligned} -\Delta p - \boldsymbol{\zeta} \cdot \nabla p - \nabla \cdot (\boldsymbol{\zeta} p) + \gamma p &= z \quad in \quad \Omega \\ p &= 0 \quad on \quad \partial\Omega. \end{aligned}$$

Then $(z, y)_{L^2(\Omega)} = (p, u)_{L^2(\Omega)}$.

Proof. We can write out the weak form of both equations using test functions p and y. By integration by parts, we can easily see that the left-hand side of both equation is $a(y, p)$. The result immediately follows. □

Now we can state the following lemma which gives an explicit expression for S^*.

Lemma 2.27 ([97, Lemma 2.24])**.** *For the problem (1.1.4), the adjoint operator* $S^* : L^2(\Omega) \to L^2(\Omega)$ *is given by*

$$S^* z = p, \tag{2.3.21}$$

where $p \in H_0^1(\Omega)$ *is the weak solution to the boundary value problem*

$$\begin{aligned} -\Delta p - \boldsymbol{\zeta} \cdot \nabla p - \nabla \cdot (\boldsymbol{\zeta} p) + \gamma p &= z \quad in \quad \Omega \\ p &= 0 \quad on \quad \partial\Omega. \end{aligned}$$

Proof. First we have

$$(z, Su) = (S^* z, u) \quad \forall z, u \in L^2(\Omega) \tag{2.3.22}$$

by the definition of the adjoint operator. Secondly, apply Lemma 2.26 to $y = Su$, we have

$$(z, Su) = (z, y) = (p, u). \tag{2.3.23}$$

□

Therefore by Lemma 2.27 the solution of (1.1.3)-(1.1.4) is determined by

$$a(q,\bar{p}) = (q,\bar{y}-y_d)_{L^2(\Omega)} \quad \forall q \in H_0^1(\Omega), \tag{2.3.24a}$$

$$\bar{p}+\beta\bar{u} = 0, \tag{2.3.24b}$$

$$a(\bar{y},z) = (\bar{u},z)_{L^2(\Omega)} \qquad \forall z \in H_0^1(\Omega). \tag{2.3.24c}$$

Remark 2.28. *The optimality system* (2.3.24) *is called the first order optimality condition. It can also be derived from [74, Theorem 1.4] and [97, Lemma 2.21]. It is a necessary and sufficient condition of* (1.1.3)-(1.1.4) *(cf. [97, Theorem 2.22]). The optimality condition* (2.3.24) *is the starting point of our multigrid methods in Chapter 4.*

If the pointwise state constraint (1.1.5) is imposed, the first order optimality condition becomes a variational inequality. For simplicity, we consider the case $\boldsymbol{\zeta} = \mathbf{0}$. Without loss of generality, we also assume $\gamma = 0$. We first prove the following general theorem.

Theorem 2.29 (Variational inequality [74, Theorem 1.2])**.** *Let $a(\cdot,\cdot)$ be a symmetric, continuous and coercive bilinear form defined on a Hilbert space V, $F \in V'$ and K be a convex subset of V. Then*

$$u = \operatorname*{argmin}_{v\in K}\left[\frac{1}{2}a(v,v) - \langle F, v\rangle\right] \tag{2.3.25}$$

if and only if

$$a(u, v-u) \geq \langle F, v-u\rangle \quad \forall v \in K. \tag{2.3.26}$$

Proof. Assume u is the solution of (2.3.25). Let $v \in K$ be arbitrary and $E(v) = \frac{1}{2}a(v,v) - \langle F, v\rangle$. Define $\phi(t) = E((1-t)u + tv)$ on $[0,1]$. Specifically, since $a(\cdot,\cdot)$ is symmetric, we have

$$\phi(t) = \frac{1}{2}a(u,u) - \langle F, u\rangle + t\left[a(u, v-u) - \langle F, v-u\rangle\right] + \frac{1}{2}t^2 a(v-u, v-u).$$

We know for $t \in [0,1]$, $(1-t)u + tv \in K$ since K is convex, thus by the fact u is the minimizer of $E(v)$ for all $v \in K$, we obtain

$$\phi(0) \leq \phi(t).$$

Hence $\phi'(0) \geq 0$, which is

$$a(u, v-u) \geq \langle F, v-u \rangle \quad \forall v \in K.$$

On the other hand, we know $\phi'(t) = a(u, v-u) - \langle F, v-u \rangle + ta(v-u, v-u)$. By the coercivity of $a(\cdot,\cdot)$, we have for $u \in K$ and all $v \in K$,

$$\phi'(t) \geq 0 \quad t \in [0,1].$$

Therefore we must have $\phi(0) \leq \phi(1)$ which means

$$u = \underset{v \in K}{\operatorname{argmin}} \left[\frac{1}{2} a(v,v) - \langle F, v \rangle \right].$$

□

Note that we can rewrite (2.3.6)-(2.3.7) as the following,

$$\bar{y} = \underset{y \in K}{\operatorname{argmin}} \left[\frac{1}{2} a(y,y) - \langle F, y \rangle \right], \tag{2.3.27}$$

where $a(y,z) = (y,z)_{L^2(\Omega)} + \beta(\Delta y, \Delta z)_{L^2(\Omega)}$, $\langle F, y \rangle = (y_d, y)_{L^2(\Omega)}$ and $K = \{y \in V : y \leq \psi \text{ in } \Omega\}$. It is easy to check that $a(\cdot,\cdot)$ and K satisfy the assumptions of Theorem 2.29. Therefore we can apply Theorem 2.29 to (2.3.27) and hence the solution to (2.3.6)-(2.3.7) can be characterized by the following variational inequality

$$a(\bar{y}, y - \bar{y}) \geq (y_d, y - \bar{y})_{L^2(\Omega)} \quad \forall y \in K,$$

which is

$$(\bar{y} - y_d, y - \bar{y})_{L^2(\Omega)} + \beta(\Delta \bar{y}, \Delta(y - \bar{y}))_{L^2(\Omega)} \geq 0 \quad \forall y \in K. \tag{2.3.28}$$

The variational inequality (2.3.28) is equivalent to the following system involving the adjoint state $\bar{p}$. If $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ is the solution of (1.1.3)-(1.1.5), then there exists $\bar{p} \in H_0^1(\Omega)$ such that

$$\begin{aligned} \Delta \bar{y} + \bar{u} &= 0, \\ \bar{p} + \beta \bar{u} &= 0, \\ \bar{y} &\leq \psi, \\ (\nabla \bar{p}, \nabla(y - \bar{y}))_{L^2(\Omega)} + (y_d - \bar{y}, y - \bar{y})_{L^2(\Omega)} &\leq 0 \quad \forall y \in K. \end{aligned} \tag{2.3.29}$$

Remark 2.30. *Details about the derivation of* (2.3.29) *are discussed in Section 2.4.3. Note that* (2.3.29) *is equivalent to* (2.3.24) *when the pointwise state constraints are absent under the assumption* $\boldsymbol{\zeta} = \mathbf{0}$ *and* $\gamma = 0$*. While the starting point of our multigrid methods for* (1.1.3)-(1.1.4) *is the saddle point problem (first order optimality condition)* (2.3.24)*, the variational inequality* (2.3.28) *is the starting point for our multigrid methods for* (1.1.3)-(1.1.5) *instead of* (2.3.29)*.*

2.4 Regularity Results

In this section we briefly review the elliptic regularity results for second order PDEs, fourth order PDEs and optimal control problems with pointwise state constraints. Since this is a broad subject we only present a few relevant results. More details can be found in [33, 50, 51, 55, 44, 41, 42, 82, 48].

2.4.1 Second Order Problems

Let Ω be a bounded polygonal/polyhedral domain of $\mathbb{R}^d$, $d = 2, 3$. We denote the second order elliptic operator $\mathcal{L}$ by

$$\mathcal{L}u := -\sum_{i=1}^{d} D_i^2 u + \sum_{i=1}^{d} [D_i(b_i u) + c_i D_i u] + a_0 u. \tag{2.4.1}$$

Here D_j is the jth partial derivative, $b_i, c_i, a_0 \in C^\infty(\bar{\Omega})$.

Consider the following problem

$$\begin{aligned} \mathcal{L}u &= f \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{2.4.2}$$

The variational formulation of (2.4.2) is, given $f \in L^2(\Omega)$, to find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = (f, v)_{L^2(\Omega)}, \tag{2.4.3}$$

where $a(u,v) = \int_\Omega \left[\sum_{i=1}^d D_i u D_i v - \sum_{i=1}^d (b_i u D_i v - c_i v D_i u) + a_0 uv\right] dx$.

Theorem 2.31 ([44, Section 6])**.** *Let $u \in H_0^1(\Omega)$ be the weak solution of (2.4.2), $f \in L^2(\Omega)$ and $b_i, c_i, a_0 \in C^\infty(\bar{\Omega})$. Then $u \in H^{1+\alpha}(\Omega)$ where $\alpha \in (\frac{1}{2}, 1]$ and*

$$\|u\|_{H^{1+\alpha}(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{2.4.4}$$

Remark 2.32. *If Ω is convex, we have $\alpha = 1$, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and*

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

For more general elliptic regularity result for second order problems, we refer to [55, 82, 44].

2.4.2 Fourth Order Problems

For fourth order problems we restrict ourselves to the following problem. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain, $f \in L^2(\Omega)$ and

$$\begin{aligned} \Delta^2 u + u &= f \quad \text{in} \quad \Omega, \\ u = \Delta u &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{2.4.5}$$

The weak formulation of (2.4.5) is to find $u \in V = H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in V, \tag{2.4.6}$$

where $a(u,v) = \int_\Omega [\Delta u \Delta v + uv]\, dx$.

Theorem 2.33. *Assume Ω is a polygonal domain and $f \in L^2(\Omega)$. Let u be the weak solution of (2.4.5), then $u \in H^{2+\alpha}(\Omega)$ for some $\alpha \in (0, 2]$. The number α can be close to 0 even when the domain is convex. If the largest interior angle of Ω is less than or equal to $\frac{\pi}{2}$ then $\alpha = 1$. Moreover, we have*

$$\|u\|_{H^{2+\alpha}(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{2.4.7}$$

More details of the elliptic regularity result for fourth order problems can be found in [55, 44, 78, 82].

2.4.3 Optimal Control Problems with Pointwise State Constraints

In this subsection we discuss the regularity results for (2.3.6)-(2.3.7) under the assumption $\boldsymbol{\zeta} = \mathbf{0}, \gamma = 0$ and the equivalent variational inequality (2.3.28). These results are useful in the convergence analysis of finite element methods for (2.3.28) (cf. [33]). In this subsection the space V is defined in (2.3.3).

Let us introduce the Lagrange multiplier μ. We take $y = -\phi + \bar{y} \in K$ in (2.3.28) where ϕ is a nonnegative function in $C_0^\infty(\Omega)$. Thus we have

$$\int_\Omega \Big[(\bar{y} - y_d)\phi + \beta(\Delta\bar{y})(\Delta\phi)\Big] dx \leq 0. \tag{2.4.8}$$

It follows from [86, Section 13, Theorem 25] or [87, Theorem 2.14] that

$$\int_\Omega \Big[(\bar{y} - y_d)z + \beta(\Delta\bar{y})(\Delta z)\Big] dx = \int_\Omega z \, d\mu \quad \forall z \in V, \tag{2.4.9}$$

where μ is a non-positive regular Borel measure.

Let $\mathfrak{C} = \{x \in \Omega : \bar{y}(x) = \psi(x)\}$ be the contact set. Note that $\mathfrak{C}$ is a compact set of Ω since $\bar{y} = 0$ on $\partial\Omega$ and $\psi > 0$ on $\partial\Omega$. Let $\phi \in C_0^\infty(\Omega)$ such that supp $\phi \cap \mathfrak{C} = \emptyset$. We consider the functions $(\pm\epsilon)\phi + \bar{y}$ with sufficiently small ϵ, we have

$$(\pm\epsilon)\phi + \bar{y} - \psi = \begin{cases} < 0 & \text{on supp } \phi, \\ \leq 0 & \text{otherwise.} \end{cases} \tag{2.4.10}$$

This implies $(\pm\epsilon)\phi + \bar{y} \in K$. Then we can substitute $y = (\pm\epsilon)\phi + \bar{y}$ in (2.3.28) to obtain

$$\int_\Omega \Big[(\bar{y} - y_d)\phi + \beta(\Delta\bar{y})(\Delta\phi)\Big]dx = 0. \tag{2.4.11}$$

Therefore by (2.4.9) and (2.4.11) we have

$$\int_\Omega \phi \; d\mu = 0 \tag{2.4.12}$$

for all $\phi \in C_0^\infty(\Omega)$ satisfying supp $\phi \cap \mathfrak{C} = \emptyset$. From (2.4.12) we can conclude (cf. [33, 34]) μ is supported on $\mathfrak{C}$, which is equivalent to the complementarity condition

$$\int_\Omega (\psi - \bar{y})\, d\mu = 0. \tag{2.4.13}$$

Moreover, we derive an important property of μ here. Assume $\bar{y}$ is the solution of (2.3.6)-(2.3.7), it is known that [33, 50, 51, 39, 40] the following interior regularity result of $\bar{y}$ holds,

$$\bar{y} \in H^3_{loc}(\Omega) \cap W^{2,\infty}_{loc}(\Omega). \tag{2.4.14}$$

Define a linear functional

$$\langle \mu, z \rangle = \int_\Omega z \; d\mu. \tag{2.4.15}$$

Take $\phi \in C_0^\infty(\Omega)$ such that $\phi = 1$ in an open neighborhood of $\mathfrak{C}$. We obtain, for all $z \in C_0^\infty(\Omega)$,

$$\int_\Omega z \; d\mu = \int_\Omega \phi z \; d\mu = \int_\Omega \Big[(\bar{y} - y_d)\phi z + \beta(\Delta\bar{y})\Delta(\phi z)\Big]dx. \tag{2.4.16}$$

Let a subset G of Ω satisfy $\mathfrak{C} \subset$ supp $\phi \subset G$. Then we can rewrite the last integral in (2.4.16) by simply replacing Ω with G. Since $\bar{y} \in H^3_{loc}(\Omega)$ (by (2.4.14)) we have

$$\begin{aligned} & \left|\int_G \Big[(\bar{y} - y_d)\phi z + \beta(\Delta\bar{y})\Delta(\phi z)\Big]dx\right| \qquad (2.4.17)\\ = \; & \left|\int_G \Big[(\bar{y} - y_d)\phi z - \beta\nabla(\Delta\bar{y})\nabla(\phi z)\Big]dx\right| \\ \leq \; & C\|z\|_{H^1(\Omega)} + \beta\|\bar{y}\|_{H^3(G)}\|\phi\|_{H^1(G)}\|z\|_{H^1(\Omega)} \\ \leq \; & C\|z\|_{H^1(\Omega)}. \end{aligned}$$

Hence we conclude,

$$\left|\int_\Omega z \, d\mu\right| \leq C\|z\|_{H^1(\Omega)} \quad \forall z \in C_0^\infty(\Omega). \tag{2.4.18}$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, (2.4.18) is also true in $H_0^1(\Omega)$. Therefore we conclude $\mu \in H^{-1}(\Omega)$.

Remark 2.34. *The fact that the Lagrange multiplier μ is a non-positive regular Borel measure and $\mu \in H^{-1}(\Omega)$ at the same time is crucial in the analysis of finite element methods in the following chapter. It can be shown that* (2.4.13), (2.4.9) *and μ is a non-positive regular Borel measure implies* (2.3.28).

We define the adjoint state $\bar{p} \in H_0^1(\Omega)$ by

$$(\nabla \bar{p}, \nabla v)_{L^2(\Omega)} = (\bar{y} - y_d, v)_{L^2(\Omega)} - \int_\Omega v \, d\mu \quad \forall v \in H_0^1(\Omega). \tag{2.4.19}$$

By integration by parts, we have

$$(\bar{p}, \Delta z)_{L^2(\Omega)} = (y_d - \bar{y}, z)_{L^2(\Omega)} + \int_\Omega z \, d\mu \quad \forall z \in V. \tag{2.4.20}$$

Comparing (2.4.20) and (2.4.9), we have

$$(\bar{p} - \beta\Delta\bar{y}, \Delta z)_{L^2(\Omega)} = 0 \quad \forall z \in V, \tag{2.4.21}$$

which implies $\beta\Delta\bar{y} = \bar{p}$ since Δ is a bijection from $V \to L^2(\Omega)$ (cf. Remark 2.21). Thus we have following regularity of $\bar{u}$

$$\bar{u} = -\Delta\bar{y} \in H_0^1(\Omega). \tag{2.4.22}$$

We refer to [42, 33, 64] for more discussion about regularity results of optimal control problems with pointwise state constraints.

2.5 Classical Iterative Methods

In this section we give a brief review of classical iterative methods which we use throughout this dissertation. Iterative methods are indirect methods for solving

linear systems by approximating the solutions iteratively. These methods are suitable for solving large sparse linear systems. We exploit classical iterative methods and utilize them as smoothers for multigrid methods. In this section $k = 0, 1, \cdots$ represents the number of iterations. For thorough reviews of iterative methods, we refer to [89, 52, 102, 46].

We consider the following linear system in this section,

$$A\mathbf{x} = \mathbf{b}, \tag{2.5.1}$$

where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $b \in \mathbb{R}^n$. Suppose we have an initial guess $\mathbf{x}_0$, we denote the error by $\mathbf{e} = \mathbf{x} - \mathbf{x}_0$. Note that

$$A\mathbf{e} = \mathbf{b} - A\mathbf{x}_0. \tag{2.5.2}$$

Here $\mathbf{r} = \mathbf{b} - A\mathbf{x}_0$ is called the residual. Hence we call (2.5.2) the residual equation. Assume we can solve (2.5.2), then we can recover the exact solution $\mathbf{x}$ by

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{e}.$$

Clearly we cannot solve (2.5.2) exactly (otherwise we are able to solve (2.5.1) and obtain the exact solution $\mathbf{x}$). Instead we approximate A^{-1} by some matrices B. Therefore we have

$$\mathbf{e}' = B(\mathbf{b} - A\mathbf{x}_0),$$

where $\mathbf{e}'$ is an approximation of $\mathbf{e}$. Hence we update our initial guess by

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{e}' = \mathbf{x}_0 + B(\mathbf{b} - A\mathbf{x}_0). \tag{2.5.3}$$

We hope that $\mathbf{x}_1$ is a better approximation of $\mathbf{x}$ than $\mathbf{x}_0$. Motivated by (2.5.3), we can repeat this process and write the general iterative methods as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + B(\mathbf{b} - A\mathbf{x}_k). \tag{2.5.4}$$

From (2.5.4) we see that we are free to choose B. In general, $B\mathbf{x}$ should be easy to compute and B should be a "good" approximation of A^{-1}.

Alternatively, the iterative scheme (2.5.4) can be written as

$$\mathbf{x}_{k+1} = M\mathbf{x}_k + \mathbf{g}, \tag{2.5.5}$$

where $M = (I - BA)$ and $\mathbf{g} = B\mathbf{b}$. Let us consider the iterative methods of the form (2.5.5).

Definition 2.35. *An iterative method is convergent if*

$$\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}$$

with any initial guess.

The following theorem is well-known (cf. [89, Theorem 4.1]).

Theorem 2.36. *The iterative method* (2.5.5) *is convergent if and only if* $\rho(M) < 1$ *where* $\rho(M)$ *is the spectral radius of* M*.*

Remark 2.37. *M is called the iteration matrix of* (2.5.5)*. It can be shown that* $\rho(M)$ *is the convergence factor which indicates how fast the iterative method* (2.5.5) *converges. The smaller the convergence factor* $\rho(M)$ *is, the faster the method converges.*

2.5.1 Richardson Iteration

Assuming A is SPD, an easy choice of B is γI where $\gamma > 0$ is a constant, hence we have

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma(\mathbf{b} - A\mathbf{x}_k). \tag{2.5.6}$$

The iterative method (2.5.6) is called the Richardson iteration. Clearly the choice of γ is crucial for Richardson iteration. According to Theorem 2.36, the spectral radius of the iteration matrix $I - \gamma A$ should be less than 1 for Richardson iteration

to converge. Assume λ_{min} and λ_{max} are the smallest and largest eigenvalues of A respectively, the smallest and largest eigenvalues of $I - \gamma A$ are $1 - \gamma\lambda_{max}$ and $1 - \gamma\lambda_{min}$. The following conditions must be satisfied in order for the method to converge,

$$1 - \gamma\lambda_{min} < 1,$$

$$1 - \gamma\lambda_{max} > -1.$$

These imply the Richardson iteration converges for any $\gamma > 0$ which satisfies

$$0 < \gamma < \frac{2}{\lambda_{max}}. \tag{2.5.7}$$

Moreover, it can be shown that (cf. [89, Example 4.1]) the optimal choice of γ in terms of minimizing the convergence factor $\rho(M)$ is

$$\gamma_{opt} = \frac{2}{\lambda_{max} + \lambda_{min}}, \tag{2.5.8}$$

while the optimal convergence factor is

$$\rho_{opt} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}. \tag{2.5.9}$$

2.5.2 Gauss-Seidel Iteration

Let us split the matrix A as $A = D + L + U$, here D is the diagonal part of the matrix A, L is the strictly lower triangular part and U is the strictly upper triangular part.

If we choose $B = (L + D)^{-1}$, the iterative method (2.5.4) becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k + (L + D)^{-1}(\mathbf{b} - A\mathbf{x}_k). \tag{2.5.10}$$

The method (2.5.10) is called the forward Gauss-Seidel iteration. Similarly, if we choose $B = (U + D)^{-1}$ we have

$$\mathbf{x}_{k+1} = \mathbf{x}_k + (U + D)^{-1}(\mathbf{b} - A\mathbf{x}_k). \tag{2.5.11}$$

The method (2.5.11) is called the backward Gauss-Seidel iteration.

Alternatively, we have

$$(L + D + U)\mathbf{x} = \mathbf{b} \Rightarrow (L + D)\mathbf{x} = \mathbf{b} - U\mathbf{x}. \tag{2.5.12}$$

Thus we can write down another iterative scheme motivated by (2.5.12),

$$(L + D)\mathbf{x}_{k+1} = \mathbf{b} - U\mathbf{x}_k. \tag{2.5.13}$$

It is trivial that (2.5.13) is equivalent to (2.5.10). Therefore forward Gauss-Seidel iteration can be viewed as a matrix splitting method. Similarly, backward Gauss-Seidel iteration is equivalent to the following matrix splitting method,

$$(U + D)\mathbf{x}_{k+1} = \mathbf{b} - L\mathbf{x}_k. \tag{2.5.14}$$

In applications, A is often symmetric positive definite. However, B in forward or backward Gauss-Seidel iteration is not symmetric. Hence it is preferable to use symmetric Gauss-Seidel iteration in some applications, namely

$$\begin{aligned} \mathbf{x}_{k+\frac{1}{2}} &= \mathbf{x}_k + (L + D)^{-1}(\mathbf{b} - A\mathbf{x}_k), \\ \mathbf{x}_{k+1} &= \mathbf{x}_{k+\frac{1}{2}} + (U + D)^{-1}(\mathbf{b} - A\mathbf{x}_{k+\frac{1}{2}}). \end{aligned} \tag{2.5.15}$$

Notice that the symmetric Gauss-Seidel iteration consists of a forward sweep followed by a backward sweep. It is easy to show that symmetric Gauss-Seidel iteration is of the following form,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + (U + D)^{-1}D(L + D)^{-1}(\mathbf{b} - A\mathbf{x}_k). \tag{2.5.16}$$

Here $B = (U + D)^{-1}D(L + D)^{-1}$ is symmetric. Furthermore, if A is SPD then B is also SPD. The following theorem guarantees the convergence of Gauss-Seidel iteration.

Theorem 2.38. *Assume A is SPD, then forward Gauss-Seidel* (2.5.10)*, backward Gauss-Seidel* (2.5.11)*, and symmetric Gauss-Seidel* (2.5.15) *converge for any initial guess.*

Remark 2.39. *The condition "A is SPD" is sufficient but not necessary. We refer to [89, 102, 46, 52] for more general results of Gauss-Seidel iterations.*

2.6 Projection Methods

In this section we briefly review the projection methods to solve the linear system (2.5.1) and discuss two important examples: the minimal residual method (MINRES) (cf. [83]) and the generalized minimal residual method (GMRES) (cf. [90]). Detailed discussion about this topic can be found in [89, Chapter 5-6].

In general, suppose $\mathcal{K}$ and $\mathcal{L}$ are two subspaces of $\mathbb{R}^d$ and $\mathbf{x}_0$ is given as an initial guess. A projection technique is to seek an approximate solution $\tilde{\mathbf{x}}$ in the space $\mathbf{x}_0 + \mathcal{K}$ such that the new residual vector be orthogonal to $\mathcal{L}$, namely

$$\text{Find} \quad \tilde{\mathbf{x}} \in \mathbf{x}_0 + \mathcal{K}, \quad \text{such that} \quad \mathbf{b} - A\tilde{\mathbf{x}} \perp \mathcal{L}. \tag{2.6.1}$$

For a specific choice of $\mathcal{L}$, we have the following important lemma (cf. [89, Proposition 5.3]).

Lemma 2.40. *Let A be an arbitrary square matrix and and assume that $\mathcal{L} = A\mathcal{K}$. Then a vectore $\tilde{\mathbf{x}}$ is the result of an projection method with respect to $\mathcal{K}$ and $\mathcal{L}$ with the starting vector $\mathbf{x}_0$ if and only if it minimizes the 2-norm of the residual $\mathbf{b} - A\mathbf{x}$ over $\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}$, i.e., if and only if*

$$R(\tilde{\mathbf{x}}) = \min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}} R(\mathbf{x}), \tag{2.6.2}$$

where $R(\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|_2$.

Remark 2.41. *Lemma 2.40 is the starting point of MINRES and GMRES. They all minimize the 2-norm of the residual over an affine space hence the methods are*

called "minimal residual methods" while MINRES aims for symmetric indefinite matrix and GMRES works for nonsymmetric matrix.

One of the most important choices for subspace $\mathcal{K}$ is so-called Krylov subspace.

Definition 2.42. *Define*

$$\mathcal{K}_m(A, \mathbf{v}) = span\{\mathbf{v}, A\mathbf{v}, \cdots, A^{m-1}\mathbf{v}\}. \tag{2.6.3}$$

$\mathcal{K}_m(A, \mathbf{v})$ *is called a Krylov subspace.* $\mathcal{K}_m(A, \mathbf{v})$ *is denoted by* $\mathcal{K}_m$ *if there is no ambiguity.*

2.6.1 MINRES

Assume A is symmetric, choose $\mathcal{K} = \mathcal{K}_m(A, \mathbf{v}_1)$ and $\mathcal{L} = A\mathcal{K}$ where $\mathbf{v}_1 = \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|_2}$ and $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ is the initial residual. This choice leads to MINRES. We briefly review the idea of MINRES without attempting to be exhaustive.

According to Lemma 2.40, we try to solve the following constrained optimization,

$$\min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_m} \|\mathbf{b} - A\mathbf{x}\|_2. \tag{2.6.4}$$

Here we exploit the famous Lanczos algorithm [73]. Lanczos algorithm is an algorithm for building an orthogonal basis of the Krylov subspace $\mathcal{K}_m$. Starting from $\mathbf{v}_1$, Lanczos algorithm generates vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m$ that form an orthonormal basis of the Krylov subspace (cf. [89, 98, 54]), i.e.,

$$\mathcal{K}_m(A, \mathbf{v}_1) = span\{\mathbf{v}_1, A\mathbf{v}_1, \cdots, A^{m-1}\mathbf{v}_1\}.$$

Moreover, denote by V_m, the $n \times m$ matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m$, we have the following,

$$AV_m = V_{m+1}T_m, \tag{2.6.5}$$

where T_m is a tridiagonal matrix with $m + 1$ rows and m columns. Note that any vector in $\mathcal{K}_m$ can be written as $V_m\mathbf{y}$ where $\mathbf{y} = (y_1, y_2, \cdots, y_m)$ is the vector that

contains the coefficients. Hence we rewrite (2.6.4) as

$$\min_{\mathbf{y}} \|\mathbf{b} - A(\mathbf{x}_0 + V_m\mathbf{y})\|_2.$$

By (2.6.5) and the orthonormality of V_{m+1}, we have

$$\begin{aligned}
&\|\mathbf{b} - A(\mathbf{x}_0 + V_m\mathbf{y})\|_2 \\
=&\|\mathbf{r}_0 - AV_m\mathbf{y}\|_2 \\
=&\|\mathbf{r}_0 - V_{m+1}T_m\mathbf{y}\|_2 \\
=&\|V_{m+1}(\|\mathbf{r}_0\|_2\mathbf{e}_1 - T_m\mathbf{y})\|_2 \\
=&\|\|\mathbf{r}_0\|_2\mathbf{e}_1 - T_m\mathbf{y}\|_2.
\end{aligned}$$

Therefore we need to solve

$$\min_{\mathbf{y}} \|\|\mathbf{r}_0\|_2\mathbf{e}_1 - T_m\mathbf{y}\|_2, \tag{2.6.6}$$

which is a least-squares problem of small size. We can apply Givens rotation to T_m hence the QR decomposition is obtained, then the least-squares problem can be solved efficiently. To summarize, MINRES approximation is the unique vector $\mathbf{x}_m$ of $\mathbf{x}_0 + \mathcal{K}_m$ which solves (2.6.4). It can be obtained by

$$\begin{aligned}
&\mathbf{x}_m = \mathbf{x}_0 + V_m\mathbf{y}, \\
&\text{where} \quad \mathbf{y} = \operatorname*{argmin}_{\mathbf{y}} \|\|\mathbf{r}_0\|_2\mathbf{e}_1 - T_m\mathbf{y}\|_2.
\end{aligned} \tag{2.6.7}$$

We include the following convergence theorem for MINRES under some special assumptions. We refer to [54, Chapter 3] for more general cases.

Theorem 2.43 ([54, Section 3.1])**.** *Assume A is symmetric and the eigenvalues of A are contained in two intervals $[a, b] \cup [c, d]$, where $a < b < 0 < c < d$ and $b - a = d - c$. We have*

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 2\left(\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right)^{[k/2]}, \tag{2.6.8}$$

where $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ *is the residual after* k *MINRES iterations and* $[k/2]$ *is the integer part of* $k/2$.

We omit the details of specific implementation. We refer to [98, Figure 6.9] and [15, Algorithm 2.4] for the implementation. If MINRES is applied to a preconditioned linear system, we have the preconditioned MINRES algorithm. This algorithm can be found, for example, in [15, Algorithm 4.1]. We refer to [83, 98, 54, 15, 89] for more details about MINRES.

2.6.2 GMRES

Let A be nonsymmetric, under the same choice of subspaces $\mathcal{K}$ and $\mathcal{L}$ in Section 2.6.1, we have the GMRES algorithm. According to Lemma 2.40, we also start with the following constrained optimization,

$$\min_{\mathbf{x}\in\mathbf{x}_0+\mathcal{K}_m} \|\mathbf{b} - A\mathbf{x}\|_2. \tag{2.6.9}$$

However, a general version of Lanczos algorithm, Arnoldi's procedure [3], is used. This procedure also generates an orthonormal basis of the Krylov subspace. Instead of (2.6.5), we have

$$AV_m = V_{m+1}H_m, \tag{2.6.10}$$

where H_m is a Hessenberg matrix with $m+1$ rows and m columns. Therefore we replace V_m in MINRES with H_m and the analysis holds. The GMRES approximation is the unique vector of $\mathbf{x}_m$ of $\mathbf{x}_0 + \mathcal{K}_m$ which solves (2.6.4). It can be obtained by

$$\begin{aligned} &\mathbf{x}_m = \mathbf{x}_0 + V_m\mathbf{y}, \\ &\text{where} \quad \mathbf{y} = \operatorname*{argmin}_{\mathbf{y}} \|\|\mathbf{r}_0\|_2\mathbf{e}_1 - H_m\mathbf{y}\|_2. \end{aligned} \tag{2.6.11}$$

A common technique to solve the least-squares problem in (2.6.11) is to transform the Hessenberg matrix into upper triangular form by using plane rotations.

We refer to [89] for more details. GMRES algorithm can be found in [89, Algorithm 6.9, 6.10].

One major issue of GMRES is that it requires the storage of V_m which can be large after several iterations. In order to avoid large storage requirements and computational costs for the orthogonalization, GMRES is usually restarted after each k iteration steps using the latest $\mathbf{x}_k$ as the initial guess. This algorithm is referred to as GMRES(k). GMRES(k) algorithm can be found in, for example, [98, Figure 6.2]. We refer to [90, 89, 54, 98] for more discussion of implementation of GMRES and preconditioned GMRES.

The convergence analysis of GMRES is subtle, we only include two convergence theorems for special cases. For more discussion about the convergence of GMRES, we refer to [89, 54].

Theorem 2.44 ([89, Theorem 6.30])**.** *If A is a positive definite matrix, then GMRES(k) converges for any $k \geq 1$.*

Theorem 2.45 ([89, Proposition 6.32])**.** *Assume A is a diagonalizable matrix and let $A = X\Lambda X^{-1}$ where $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the diagonal matrix of eigenvalues. Then the residual norm achieved by the k-th step of GMRES satisfies*

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq \kappa_2(X) \min_{p_k} \max_{i=1,\ldots,n} p_k(\lambda_i), \tag{2.6.12}$$

where $\kappa_2(X) = \|X\|_2\|X^{-1}\|_2$ and p_k is a polynomial of degree k or less with $p_k(0) = 1$.

2.7 P_1 Finite Element Methods

In this section we briefly review the construction and the error analysis of the P_1 finite element methods [32]. We discuss the methods for the following model

problem for simplicity,

$$-\Delta u = f \quad \text{in} \quad \Omega, \tag{2.7.1a}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{2.7.1b}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain and $f \in L^2(\Omega)$. The variational problem of (2.7.1) is to find $u \in V := H_0^1(\Omega)$ such that

$$a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in V, \tag{2.7.2}$$

where $a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx$. It is easy to see that $a(\cdot, \cdot)$ is a symmetric bilinear form and

$$a(u, v) \leq \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \quad \forall u, v \in V. \tag{2.7.3}$$

By Poincaré inequality (cf. Theorem 2.15), we also have

$$a(v, v) = |v|^2_{H^1(\Omega)} \geq C \|v\|^2_{H^1(\Omega)} \quad \forall v \in V. \tag{2.7.4}$$

It can be shown that $(V, a(\cdot, \cdot))$ is a Hilbert space (cf. [32, (2.5.3)]). Therefore by Riesz Representation Theorem (cf. Theorem 2.16) there exists a unique solution $u \in V$ solving (2.7.1).

Remark 2.46. *For nonsymmetric problems, Lax-Milgram (cf. Theorem 2.17) is often used to prove the well-posedness of the problems.*

Let V_h be a finite dimensional subspace of V, the Ritz-Galerkin approximation problem is to find $u_h \in V_h$ such that

$$a(u_h, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in V_h. \tag{2.7.5}$$

Proposition 2.47 (Galerkin Orthogonality)**.** *Let u and u_h be solutions to* (2.7.2) *and* (2.7.5) *respectively. Then*

$$a(u - u_h, v) = 0 \quad \forall v \in V_h. \tag{2.7.6}$$

Proof. By subtracting (2.7.5) from (2.7.2), we immediately obtain the result. □

Lemma 2.48 (Abstract Error Estimate)**.** *Let u and u_h be solutions to* (2.7.2) *and* (2.7.5) *respectively. Then*

$$\|u - u_h\|_a = \min_{v \in V_h} \|u - v\|_a, \tag{2.7.7}$$

where $\|\cdot\|_a^2 = a(\cdot, \cdot)$.

Proof. Let $v \in V_h$ be arbitrary, we have

$$\begin{aligned} \|u - v\|_a^2 &= a(u - v, u - v) \\ &= a(u - u_h + u_h - v, u - u_h + u_h - v) \\ &= a(u - u_h, u - u_h) + a(u_h - v, u_h - v) \\ &\geq \|u - u_h\|_a^2. \end{aligned}$$

Here we use the fact that $a(\cdot, \cdot)$ is symmetric and the Galerkin orthogonality $a(u - u_h, u_h - v) = 0$. □

Remark 2.49. *Lemma 2.48 shows that the Ritz-Galerkin method delivers the best approximation of u from V_h with respect to $\|\cdot\|_a$. If $a(\cdot, \cdot)$ is not symmetric, Lemma 2.48 is invalid. Alternatively, Céa's Theorem (cf. [32, Theorem 2.8.1]) provides a quasi-optimal error estimate in the sense that $\|u - u_h\|_{H^1(\Omega)}$ is proportional to the best it can be using the subspace V_h. We refer to [32] for more details.*

From Lemma 2.48 we can see that it is crucial to construct the finite dimensional space V_h. We begin with a triangulation $\mathcal{T}_h$ of Ω which is a collection of triangles that satisfies the following requirements,

- Ω is the union of the triangles in $\mathcal{T}_h$.
- Any two different triangles in $\mathcal{T}_h$ satisfy one of the following,

– Disjoint,

– Share a common vertex,

– Share a common edge.

See Figure 2.2 for an example of a triangulation of a square. The mesh size of the triangulation is $h = \max_{T\in\mathcal{T}_h} diam(T)$.

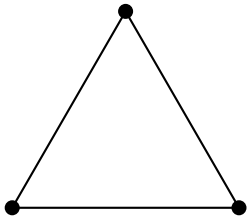

Figure 2.1. P_1 Finite Element.

The P_1 finite element space V_h associated with $\mathcal{T}_h$ is defined as,

$$V_h = \{v \in C(\bar{\Omega}) : v|_T \in \mathcal{P}_1 \text{ and } v = 0 \text{ on } \partial\Omega\}, \tag{2.7.8}$$

where $\mathcal{P}_1$ denote the set of all polynomials in two variables of degree less than or equal to 1. A polynomial of degree less than or equal to 1 defined on a triangle can be determined by its values at the three vertices hence the dimension of V_h is the number of interior vertices of $\mathcal{T}_h$. Figure 2.1 shows the P_1 finite element in two dimensions. Note that the dot indicates the nodal variable evaluation at the point where the dot is located. It can be shown that $V_h \subset V$. Therefore the P_1 finite element method for (2.7.2) is (2.7.5) with V_h defined in (2.7.8).

In order to obtain a concrete error estimate, we need to choose some $v \in V_h$ in (2.7.7). The interpolation operator $\Pi_h : V \to V_h$ is defined by $\Pi_h u = u$ at all vertices of $\mathcal{T}_h$. By Theorem 2.31, we know $u \in H^{1+\alpha}(\Omega) \cap H_0^1(\Omega)$. Moreover we have the following standard interpolation error estimate [32, Chapter 4],

$$\|u - \Pi_h u\|_{L^2(\Omega)} + h|u - \Pi_h u|_{H^1(\Omega)} \leq Ch^{1+\alpha}|u|_{H^{1+\alpha}(\Omega)}, \tag{2.7.9}$$

where the positive constant C is independent of h.

Theorem 2.50 (Concrete Error Estimate)**.** *Let u and u_h be solutions to* (2.7.2) *and* (2.7.5) *respectively. We have*

$$|u - u_h|_{H^1(\Omega)} \leq Ch^{\alpha}|u|_{H^{1+\alpha}(\Omega)}, \tag{2.7.10}$$

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{2\alpha}|u|_{H^{1+\alpha}(\Omega)}. \tag{2.7.11}$$

Proof. By Lemma 2.48 and (2.7.9) we have,

$$\begin{aligned} |u - u_h|_{H^1(\Omega)} &= \min_{v \in V_h} |u - v|_{H^1(\Omega)} \\ &\leq |u - \Pi_h u|_{H^1(\Omega)} \\ &\leq Ch^{\alpha}|u|_{H^{1+\alpha}(\Omega)}. \end{aligned}$$

The estimate (2.7.11) can be established by a standard duality argument. Let w be the solution of the following variational problem. Find $w \in V$ such that

$$a(w, v) = (u - u_h, v)_{L^2(\Omega)} \quad \forall v \in V. \tag{2.7.12}$$

The problem (2.7.12) is well-defined since $u - u_h \in L^2(\Omega)$. Therefore, by Galerkin orthogonality and (2.4.4),

$$\begin{aligned} \|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} \\ &= a(w, u - u_h) \\ &= a(u - u_h, w - \Pi_h w) \\ &\leq \|u - u_h\|_{H^1(\Omega)} \|w - \Pi_h w\|_{H^1(\Omega)} \\ &\leq Ch^{\alpha}\|u - u_h\|_{H^1(\Omega)}|w|_{H^{1+\alpha}(\Omega)} \\ &\leq Ch^{\alpha}\|u - u_h\|_{H^1(\Omega)}\|u - u_h\|_{L^2(\Omega)}. \end{aligned}$$

Therefore we conclude

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{\alpha}\|u - u_h\|_{H^1(\Omega)} \leq Ch^{2\alpha}|u|_{H^{1+\alpha}(\Omega)}. \tag{2.7.13}$$

□

In order to solve (2.7.5) with the P_1 finite element method, we introduce a basis of V_h. Let $\{p_i\}_{i=1}^n$ be the interior vertices of $\mathcal{T}_h$ where $n = dimV_h$. The natural nodal basis $\{\varphi_i\}_{i=1}^n$ of V_h is defined by

$$\varphi_j(p_k) = \delta_{jk} := \begin{cases} 1 & j = k, \\ 0 & j \neq k. \end{cases} \tag{2.7.14}$$

Here δ_{jk} is called the Kronecker delta. By using this basis of the P_1 finite element space V_h, the discrete problem (2.7.5) is equivalent to the following problems,

$$a(u_h, \varphi_i) = (f, \varphi_i)_{L^2(\Omega)} \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.7.15}$$

We can write $u_h = \sum_{j=1}^n x_j\varphi_j$ and (2.7.15) becomes

$$a(\sum_{j=1}^n x_j\varphi_j, \varphi_i) = (f, \varphi_i)_{L^2(\Omega)} \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Since $a(\cdot, \cdot)$ is symmetric and bilinear, we have

$$\sum_{j=1}^n a(\varphi_i, \varphi_j)x_j = (f, \varphi_i)_{L^2(\Omega)} \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.7.16}$$

We rewrite (2.7.16) in matrix-vector form

$$A\mathbf{x} = \mathbf{b}, \tag{2.7.17}$$

where $A \in \mathbb{R}^{n\times n}$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, $A(i, j) = a(\varphi_i, \varphi_j) = (\nabla\varphi_i, \nabla\varphi_i)_{L^2(\Omega)}$, $\mathbf{x}(i) = x_i$ and $\mathbf{b}(i) = (f, \varphi_i)_{L^2(\Omega)}$.

The matrix A is called the stiffness matrix. It is well-known that the condition number $\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = O(h^{-2})$. Therefore the system (2.7.16) becomes ill-conditioned when h is small.

Remark 2.51. *We only review some basic ingredients for the error analysis of the P_1 finite element methods in this section. However, even for a simple problem like* (2.7.2)*, we found that A is ill-conditioned when h is small. This is one of the difficulties to solving the problem* (2.7.17) *efficiently.*

2.8 Multigrid Algorithms

In this section we review the multigrid algorithm (cf. [32, Chapter 6]). The multigrid method was proposed in [49] and initially designed to solve elliptic boundary value problems. A large sparse linear system is obtained after one applies certain numerical methods (finite difference methods, finite element methods, etc.) to such a problem. As the mesh size h decreases, the problem usually becomes ill-conditioned hence classical iterative methods are not efficient (cf. Remark 2.51). Multigrid methods are multilevel methods that can overcome this issue. Moreover, the multigrid method is an optimal solver in the sense that the amount of computational work involved is only proportional to the number of unknowns in the discretized equations.

The multigrid method has two main features: smoothing on the current grid and error correction on a coarser grid. The smoothing step has the effect of damping out the oscillatory part of the error. Classical iterative methods are often used as smoothers, for example, Richardson iteration, Jacobi iteration and Gauss-Seidel iteration. The smooth part of the error can then be accurately corrected on the coarser grid.

We briefly illustrate the construction and the analysis of the multigrid methods by considering a simple model problem. We only consider finite element based multigrid methods in this dissertation. Let $\Omega \subset \mathbb{R}^2$ be a convex polygon and

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx. \tag{2.8.1}$$

Consider the Dirichlet problem, find $u \in V = H_0^1(\Omega)$ such that

$$a(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in V, \tag{2.8.2}$$

where $f \in L^2(\Omega)$.

2.8.1 The Algorithms

Let $\mathcal{T}_h$ be a shape regular simplicial triangulation of Ω and $V_h \subset V$ be the P_1 finite element space associated with $\mathcal{T}_h$. The diameter of $T \in \mathcal{T}_h$ is denoted by h_T and $h = \max_{T \in \mathcal{T}_h} h_T$ is the mesh size. Let the triangulation $\mathcal{T}_1, \mathcal{T}_2, ...$ be generated from the triangulation $\mathcal{T}_0$ through uniform subdivisions, and V_k be the P_1 finite element space associated with $\mathcal{T}_k$. Let h_k be the mesh size of $\mathcal{T}_k$. See Figure 2.2 for an example of the first three triangulations for the square. Note that we have $V_0 \subset V_1 \subset \cdots \subset V_k$.

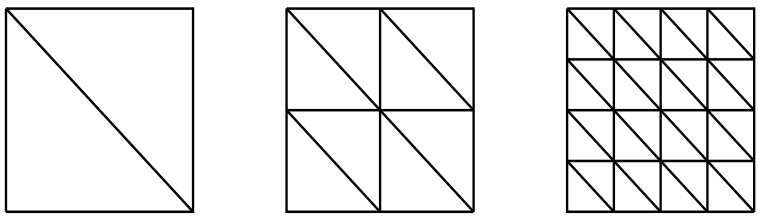

Figure 2.2. Triangulation.

The discretized problem is to find $u_k \in V_k$ such that

$$a(u_k, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in V_k. \tag{2.8.3}$$

Definition 2.52. *The mesh-dependent inner product $(\cdot, \cdot)_k$ on V_k is defined by*

$$(v, w)_k := h_k^2 \sum_{i=1}^{n_k} v(p_i) w(p_i), \tag{2.8.4}$$

where $\{p_i\}_{i=1}^{n_k}$ is the set of internal vertices of $\mathcal{T}_k$.

The operator $A_k : V_k \to V_k$ is defined by

$$(A_k v, w)_k = a(v, w) \quad \forall v, w \in V_k. \tag{2.8.5}$$

The coarse-to-fine operator $I_{k-1}^{k} : V_{k-1} \longrightarrow V_k$ is the natural injection and the fine-to-coarse operator $I_k^{k-1} : V_k \longrightarrow V_{k-1}$ is the transpose of I_{k-1}^k with respect to the mesh-dependent inner product, i.e.,

$$(I_k^{k-1} v_1, v_2)_{k-1} = (v_1, I_{k-1}^k v_2)_k \quad \forall v_1 \in V_k \text{ and } v_2 \in V_{k-1}.$$

In terms of the operator A_k, the general k-th equation is

$$A_k v = g. \tag{2.8.6}$$

W-cycle Algorithm: Let the output of the W-cycle algorithm for (2.8.6) with initial guess v_0 and m_1 (resp. m_2) pre-smoothing (resp. post-smoothing) steps be denoted by $MG_W(k, g, v_0, m_1, m_2)$. We use a direct solve for $k = 0$, i.e., we take $MG_W(0, g, v_0, m_1, m_2)$ to be $A_0^{-1} g$. For $k \geq 1$, we compute $MG_W(k, g, v_0, m_1, m_2)$ in three steps.

Pre-Smoothing The approximate solutions $v_1, \ldots, v_{m_1}$ are computed recursively by

$$v_j = v_{j-1} + \frac{1}{\Lambda_k}(g - A_k v_{j-1}) \tag{2.8.7}$$

for $1 \leq j \leq m_1$.

Coarse Grid Correction Let $g' = I_k^{k-1}(g - A_k v_{m_1})$ be the transferred residual of v_{m_1} and compute $v_1', v_2' \in V_{k-1}$ by

$$v_1' = MG_W(k-1, g', 0, m_1, m_2), \tag{2.8.8}$$

$$v_2' = MG_W(k-1, g', v_1', m_1, m_2). \tag{2.8.9}$$

We then take v_{m_1+1} to be $v_{m_1} + I_{k-1}^k v_2'$.

Post-Smoothing The approximate solutions $v_{m_1+2}, \ldots, v_{m_1+m_2+1}$ are computed recursively by

$$v_j = v_{j-1} + \frac{1}{\Lambda_k}(g - A_k v_{j-1}) \tag{2.8.10}$$

for $m_1 + 2 \leq j \leq m_1 + m_2 + 1$.

The final output is $MG_W(k, g, v_0, m_1, m_2) = v_{m_1+m_2+1}$.

Remark 2.53. *Notice that we use Richardson iteration as the smoothers hence the parameter Λ_k should satisfy* (2.5.7)*. Other classical iterative methods, for example, Jacobi iteration, Gauss-Seidel iteration can also be used in the pre-smoothing and post-smoothing steps.*

V-cycle Algorithm: Let the output of the V-cycle algorithm for (2.8.6) with initial guess v_0 and m_1 (resp. m_2) pre-smoothing (resp. post-smoothing) steps be denoted by $MG_V(k, g, v_0, m_1, m_2)$. The difference between the computations of $MG_V(k, g, v_0, m_1, m_2)$ and $MG_W(k, g, v_0, m_1, m_2)$ is only in the coarse grid correction step, where we compute

$$v_1' = MG_V(k-1, g', 0, m_1, m_2)$$

and take v_{m_1+1} to be $v_{m_1} + I_{k-1}^k v_1'$.

Two-Grid Algorithm: If we solve the coarse grid system exactly, we have so-called two-grid method, namely, we take

$$v_{m_1+1} = v_{m_1} + I_{k-1}^k A_{k-1}^{-1} g'. \tag{2.8.11}$$

Full Multigrid Algorithm: The full multigrid algorithm is the following,

- For $k = 0$, $v_0 = A_0^{-1} f$.
- For $k \geq 1$, the approximate solutions v_k are obtained recursively from
  - $v_0^k = I_{k-1}^k v_{k-1}$,

- $v_l^k = MG_W(k, v_{l-1}^k, g, m_1, m_2)$ or

  $v_l^k = MG_V(k, v_{l-1}^k, g, m_1, m_2)$, $1 \le l \le r$,

- $v_k = v_r^k$.

One can see Figure 2.3 for an illustration of V and W multigrid cycles at level 2. Every node in the graph represents a smoothing procedure or an exact solve (only at level 0). Every edge represents an inter-grid transfer procedure involving the operators I_{k-1}^k and I_k^{k-1}. Hence Figure 2.3 depicts the movement of the approximate solution among different levels. For V-cycle and W-cycle algorithms, the names come from the shape of the path among different levels.

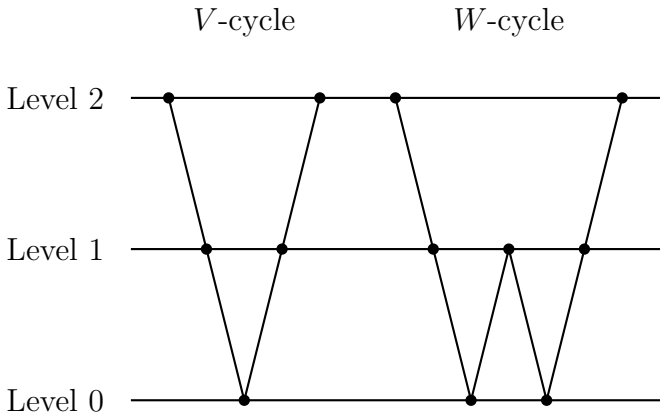

Figure 2.3. V-cycle and W-cycle.

2.8.2 Convergence Analysis

In this subsection we review the essential ingredients for analyzing multigrid methods. We refer to [59, 38, 23, 105] for thorough reviews of the convergence analysis of multigrid methods.

We define a mesh-dependent norm $|||\cdot|||_{s,k}$ as

$$|||v|||_{s,k} := \sqrt{(A_k^s v, v)_k}, \tag{2.8.12}$$

where A_k^s denotes the s power of the SPD operator A_k. Note that $|||v|||_{1,k} \approx |v|_{H^1(\Omega)}$ and $|||v|||_{0,k} \approx \|v\|_{L^2(\Omega)}$.

The operator $P_k^{k-1} : V_k \to V_{k-1}$ is defined by

$$a(P_k^{k-1}v, w) = a(v, I_{k-1}^k w) \quad \forall v \in V_k, w \in V_{k-1}. \tag{2.8.13}$$

Let us consider two-grid algorithm first. It can be shown that [32, Chapter 6] the error propagation operator E_k of two-grid algorithm for (2.8.6) is

$$E_k = R_k^{m_2}(I - I_{k-1}^k P_k^{k-1})R_k^{m_1}, \tag{2.8.14}$$

where $R_k = I - \frac{1}{\Lambda_k}A_k$ is the relaxation operator that represents a single step of Richardson iteration (2.8.7). As we can see in (2.8.14), there are two components to analyze, the smoothing part R_k and the approximation part $I - I_{k-1}^k P_k^{k-1}$. Hence, we have the following results (cf. [7]).

Lemma 2.54 (Smoothing Property)**.** *There exists a positive constant C such that*

$$|||R_k^m v|||_{t,k} \leq C h_k^{s-t} \left(\frac{t-s}{t-s+2m}\right)^{\frac{t-s}{2}} |||v|||_{s,k}. \tag{2.8.15}$$

Lemma 2.55 (Approximation Property)**.** *There exists a positive constant C such that*

$$|||(I - I_{k-1}^k P_k^{k-1})v|||_{0,k} \leq C h_k^2 |||v|||_{2,k}. \tag{2.8.16}$$

Combining the smoothing property and approximation property, we have the two-grid convergence.

Theorem 2.56 (Two-Grid Convergence)**.** *Given the initial error e_0 and $\delta \in (0,1)$. If $m_1 + m_2$ is large enough, then*

$$|E_k e_0|_{H^1(\Omega)} \leq \delta |e_0|_{H^1(\Omega)}. \tag{2.8.17}$$

Proof. According to Lemma 2.54 and Lemma 2.55, we have

$$\begin{aligned}
|||E_k e_0|||_{1,k} &= |||R_k^{m_2}(I - I_{k-1}^k P_k^{k-1}) R_k^{m_1} e_0|||_{1,k} \\
&\leq C h_k^{-1} (\frac{1}{1+2m_2})^{\frac{1}{2}} |||(I - I_{k-1}^k P_k^{k-1}) R_k^{m_1} e_0|||_{0,k} \\
&\leq C h_k^{-1} (\frac{1}{1+2m_2})^{\frac{1}{2}} h_k^2 \, |||R_k^{m_1} e_0|||_{2,k} \\
&\leq C h_k^{-1} (\frac{1}{1+2m_2})^{\frac{1}{2}} h_k^2 h_k^{-1} (\frac{1}{1+2m_1})^{\frac{1}{2}} \, |||e_0|||_{1,k} \\
&= C (\frac{1}{1+2m_2})^{\frac{1}{2}} (\frac{1}{1+2m_1})^{\frac{1}{2}} \, |||e_0|||_{1,k} \, .
\end{aligned}$$

This implies the result. □

Once we obtain the two-grid convergence, a standard perturbation argument can be used to prove the W-cycle convergence. We state the following theorem without proof.

Theorem 2.57 (W-cycle Convergence)**.** *Given any $\gamma \in (0,1)$, if $m_1 + m_2$ is large enough. Then*

$$|v - MG_W(k, v_0, g, m_1, m_2)|_{H^1(\Omega)} \leq \gamma |v - v_0|_{H^1(\Omega)}, \tag{2.8.18}$$

where v_0 is the initial guess and v is the exact solution of (2.8.6).

Remark 2.58. *Smoothing property and approximation property were introduced in [57, 58]. These two properties are essential to standard W-cycle convergence analysis. The convergence of W-cycle algorithm is a direct result of two-grid convergence as stated above.*

Remark 2.59. *We assume Ω is convex in this section. For nonconvex domain, the convergence of W-cycle can be obtained by altering the approximation property where partial elliptic regularity (cf. Theorem 2.31) is utilized.*

Remark 2.60. *The convergence of V-cycle algorithm is more delicate. For convex domain, we refer to [20, 19] for proofs. For nonconvex domain, the analysis is more*

difficult and the results in [20, 19] cannot be extended directly. Hence multiplicative theory [107, 21] and additive theory [26] were introduced to prove the convergence of V-cycle algorithm without full elliptic regularity.

Chapter 3

P_1 Finite Element Methods for Elliptic Optimal Control Problems

In this chapter we consider P_1 finite element methods for elliptic optimal control problems (1.1.3)-(1.1.5). We prove the convergence of these P_1 finite element methods and concrete error estimates are established. Numerical results are provided at the end of the chapter. We refer to [30, 34] for more details. Throughout this chapter, we use C to denote a generic constant which is independent of mesh size. Also to avoid the proliferation of the constants, we use the notation $A \lesssim B$ (or $A \gtrsim B$) to represent $A \leq (constant)B$. The notation $A \approx B$ means that $A \lesssim B$ and $B \lesssim A$.

3.1 Optimal Control Problems without Pointwise State Constraints

As stated in Section 2.3.1, the problem (1.1.3)-(1.1.4) is equivalent to (2.3.24). After eliminating $\bar{u}$, we obtain the following saddle point problem,

$$a(q,\bar{p}) - (q,\bar{y})_{L^2(\Omega)} = -(q,y_d)_{L^2(\Omega)} \quad \forall q \in H_0^1(\Omega), \tag{3.1.1a}$$

$$-(\bar{p},z)_{L^2(\Omega)} - \beta a(\bar{y},z) = 0 \qquad \forall z \in H_0^1(\Omega). \tag{3.1.1b}$$

Note that the system (3.1.1) is unbalanced with respect to β since it only appears in (3.1.1b). This can be remedied by the following change of variables:

$$\bar{p} = \beta^{\frac{1}{4}}\tilde{p} \quad \text{and} \quad \bar{y} = \beta^{-\frac{1}{4}}\tilde{y}. \tag{3.1.2}$$

The resulting saddle point problem is

$$\beta^{\frac{1}{2}} a(q,\tilde{p}) - (q,\tilde{y})_{L^2(\Omega)} = -\beta^{\frac{1}{4}}(q,y_d)_{L^2(\Omega)} \qquad \forall\, q \in H_0^1(\Omega), \tag{3.1.3a}$$

$$-(\tilde{p},z)_{L^2(\Omega)} - \beta^{\frac{1}{2}} a(\tilde{y},z) = 0 \qquad \forall\, z \in H_0^1(\Omega). \tag{3.1.3b}$$

We then use a P_1 finite element method to discretize (3.1.3) and follow Babuška's approach to analyze our finite element methods (cf. Section 2.2). We can write

(3.1.3) concisely as

$$\mathcal{B}((\tilde{p},\tilde{y}),(q,z)) = -\beta^{\frac{1}{4}}(y_d,q)_{L^2(\Omega)} \qquad \forall\,(q,z)\in H_0^1(\Omega)\times H_0^1(\Omega), \tag{3.1.4}$$

where

$$\mathcal{B}((p,y),(q,z)) = \beta^{\frac{1}{2}}a(q,p) - (q,y)_{L^2(\Omega)} - (p,z)_{L^2(\Omega)} - \beta^{\frac{1}{2}}a(y,z). \tag{3.1.5}$$

3.1.1 Continuous Problem

We will analyze the bilinear form $\mathcal{B}(\cdot,\cdot)$ in terms of the weighted H^1 norm $\|\cdot\|_{H^1_\beta(\Omega)}$ defined by

$$\|v\|^2_{H^1_\beta(\Omega)} = \beta^{\frac{1}{2}}|v|^2_{H^1(\Omega)} + \|v\|^2_{L^2(\Omega)} \qquad \forall\, v\in H^1(\Omega). \tag{3.1.6}$$

Lemma 3.1. *Let $(p,y),(q,z)\in H^1(\Omega)\times H^1(\Omega)$ be arbitrary. We have*

$$\mathcal{B}((p,y),(q,z)) \lesssim (\|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}}(\|q\|^2_{H^1_\beta(\Omega)} + \|z\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}}. \tag{3.1.7}$$

Proof. The result follows immediately from (1.1.6), (3.1.5), (3.1.6) and the Cauchy-Schwarz inequality. Note that the hidden constant here may depend on $\boldsymbol{\zeta}$ and γ. □

Lemma 3.2. *We have*

$$\sup_{(q,z)\in H^1_0(\Omega)\times H^1_0(\Omega)} \frac{\mathcal{B}((p,y),(q,z))}{(\|q\|^2_{H^1_\beta(\Omega)} + \|z\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}}} \geq 2^{-\frac{1}{2}}(\|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} \tag{3.1.8}$$

for any $(p,y)\in H^1_0(\Omega)\times H^1_0(\Omega)$.

Proof. We have

$$\begin{aligned}\mathcal{B}((p,y),(p-y,-p-y)) &= \beta^{\frac{1}{2}}a(p,p) + (p,p)_{L^2(\Omega)} + (y,y)_{L^2(\Omega)} + \beta^{\frac{1}{2}}a(y,y)\\ &\geq \|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)}.\end{aligned}$$

and

$$\begin{aligned}(\|p-y\|^2_{H^1_\beta(\Omega)} + \|-p-y\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} &= (\beta^{\frac{1}{2}}|p-y|^2_{H^1(\Omega)} + \|p-y\|^2_{L^2(\Omega)}\\ &\quad +\beta^{\frac{1}{2}}|p+y|^2_{H^1(\Omega)} + \|p+y\|^2_{L^2(\Omega)})^{\frac{1}{2}}\\ &= 2^{\frac{1}{2}}(\|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}}.\end{aligned}$$

The last equality is due to the parallelogram law (cf. [32, Theorem 2.2.8]). This immediately implies the result. □

It follows from Lemma 3.1 and 3.2 that

$$\|p\|_{H^1_\beta(\Omega)} + \|y\|_{H^1_\beta(\Omega)} \approx \sup_{(q,z)\in H^1_0(\Omega)\times H^1_0(\Omega)} \frac{\mathcal{B}((p,y),(q,z))}{\|q\|_{H^1_\beta(\Omega)} + \|z\|_{H^1_\beta(\Omega)}} \tag{3.1.9}$$

for all $(p, y) \in H^1_0(\Omega) \times H^1_0(\Omega)$. Hence (3.1.4) is well-posed (cf. Theorem 2.20). Similarly, we have

$$\|p\|_{H^1_\beta(\Omega)} + \|y\|_{H^1_\beta(\Omega)} \approx \sup_{(q,z)\in H^1_0(\Omega)\times H^1_0(\Omega)} \frac{\mathcal{B}((q,z),(p,y))}{\|q\|_{H^1_\beta(\Omega)} + \|z\|_{H^1_\beta(\Omega)}} \tag{3.1.10}$$

for all $(p, y) \in H^1_0(\Omega) \times H^1_0(\Omega)$.

3.1.2 Discrete Problem

Let $\mathcal{T}_h$ be a triangulation of Ω and $V_h \subset H^1_0(\Omega)$ be the P_1 finite element space associated with $\mathcal{T}_h$. The P_1 finite element method for (3.1.4) is to find $(\tilde{p}_h, \tilde{y}_h) \in V_h \times V_h$ such that

$$\mathcal{B}((\tilde{p}_h, \tilde{y}_h), (q_h, z_h)) = -\beta^{\frac{1}{4}}(y_d, q_h)_{L^2(\Omega)} \qquad \forall\, (q_h, z_h) \in V_h \times V_h. \tag{3.1.11}$$

For the convergence analysis of the multigrid algorithms, it is necessary to consider a more general problem: Find $(p, y) \in H^1_0(\Omega) \times H^1_0(\Omega)$ such that

$$\mathcal{B}((p,y),(q,z)) = (f,q)_{L^2(\Omega)} + (g,z)_{L^2(\Omega)} \qquad \forall\, (q,z) \in H^1_0(\Omega) \times H^1_0(\Omega), \tag{3.1.12}$$

where $f, g \in L^2(\Omega)$, together with the following dual problem: Find $(p, y) \in H^1_0(\Omega) \times H^1_0(\Omega)$ such that

$$\mathcal{B}((q,z),(p,y)) = (f,q)_{L^2(\Omega)} + (g,z)_{L^2(\Omega)} \qquad \forall\, (q,z) \in H^1_0(\Omega) \times H^1_0(\Omega). \tag{3.1.13}$$

The unique solvability of (3.1.12) (resp., (3.1.13)) follows immediately from (3.1.9) (resp., (3.1.10)) (cf. Theorem 2.20). The P_1 finite element method for (3.1.12) is to find $(p_h, y_h) \in V_h \times V_h$ such that

$$\mathcal{B}((p_h, y_h), (q_h, z_h)) = (f, q_h)_{L^2(\Omega)} + (g, z_h)_{L^2(\Omega)} \qquad \forall\, (q_h, z_h) \in V_h \times V_h, \tag{3.1.14}$$

and the P_1 finite element method for (3.1.13) is to find $(p_h, y_h) \in V_h \times V_h$ such that

$$\mathcal{B}((q_h, z_h), (p_h, y_h)) = (f, q_h)_{L^2(\Omega)} + (g, z_h)_{L^2(\Omega)} \qquad \forall\, (q_h, z_h) \in V_h \times V_h. \quad (3.1.15)$$

Note that Lemma 3.1 and 3.2 also yield the following analog of (3.1.9):

$$\|p_h\|_{H^1_\beta(\Omega)} + \|y_h\|_{H^1_\beta(\Omega)} \approx \sup_{(q_h,z_h)\in V_h\times V_h} \frac{\mathcal{B}((p_h, y_h), (q_h, z_h))}{\|q_h\|_{H^1_\beta(\Omega)} + \|z_h\|_{H^1_\beta(\Omega)}} \quad (3.1.16)$$

for all $(p_h, y_h) \in V_h \times V_h$. Similarly, we have

$$\|p_h\|_{H^1_\beta(\Omega)} + \|y_h\|_{H^1_\beta(\Omega)} \approx \sup_{(q_h,z_h)\in V_h\times V_h} \frac{\mathcal{B}((q_h, z_h), (p_h, y_h))}{\|q_h\|_{H^1_\beta(\Omega)} + \|z_h\|_{H^1_\beta(\Omega)}} \quad (3.1.17)$$

for all $(p_h, y_h) \in V_h \times V_h$. Therefore the discrete problems (3.1.14) and (3.1.15) are also uniquely solvable (cf. Theorem 2.20).

3.1.3 Error Estimates

We first have the following quasi-optimal error estimate.

Lemma 3.3. *Let (p, y) (resp., (p_h, y_h)) be the solution of* (3.1.12) *or* (3.1.13) (*resp.,* (3.1.14) *or* (3.1.15)). *We have*

$$\|p-p_h\|_{H^1_\beta(\Omega)}+\|y-y_h\|_{H^1_\beta(\Omega)} \lesssim \inf_{(q_h,z_h)\in V_h\times V_h} \big(\|p-q_h\|_{H^1_\beta(\Omega)}+\|y-z_h\|_{H^1_\beta(\Omega)}\big). \quad (3.1.18)$$

Proof. We only consider (3.1.12) since the arguments for (3.1.13) are similar. By (3.1.7), (3.1.16) and Galerkin orthogonality, we have for all $(q_h, z_h) \in V_h \times V_h$,

$$\begin{aligned}
\|p_h - q_h\|_{H^1_\beta(\Omega)} + \|y_h - z_h\|_{H^1_\beta(\Omega)} &\lesssim \sup_{(q_h,z_h)\in V_h\times V_h} \frac{\mathcal{B}((p_h - q_h, y_h - z_h), (q_h, z_h))}{\|q_h\|_{H^1_\beta(\Omega)} + \|z_h\|_{H^1_\beta(\Omega)}} \\
&= \sup_{(q_h,z_h)\in V_h\times V_h} \frac{\mathcal{B}((p - q_h, y - z_h), (q_h, z_h))}{\|q_h\|_{H^1_\beta(\Omega)} + \|z_h\|_{H^1_\beta(\Omega)}} \\
&\lesssim \|p - q_h\|_{H^1_\beta(\Omega)} + \|y - z_h\|_{H^1_\beta(\Omega)},
\end{aligned}$$

hence we obtain

$$\begin{aligned}
\|p - p_h\|_{H^1_\beta(\Omega)} + \|y - y_h\|_{H^1_\beta(\Omega)} &\lesssim \|p - q_h\|_{H^1_\beta(\Omega)} + \|y - z_h\|_{H^1_\beta(\Omega)} \\
&\quad + \|p_h - q_h\|_{H^1_\beta(\Omega)} + \|y_h - z_h\|_{H^1_\beta(\Omega)} \\
&\lesssim \|p - q_h\|_{H^1_\beta(\Omega)} + \|y - z_h\|_{H^1_\beta(\Omega)}.
\end{aligned}$$

This implies the result. □

In order to obtain a concrete estimate from the quasi-optimal estimate, we need the following regularity result.

Lemma 3.4. *The solution* (p, y) *of* (3.1.12) *or* (3.1.13) *belongs to* $H^{1+\alpha}(\Omega) \times H^{1+\alpha}(\Omega)$ *and we have*

$$\|\beta^{\frac{1}{2}} p\|_{H^{1+\alpha}(\Omega)} + \|\beta^{\frac{1}{2}} y\|_{H^{1+\alpha}(\Omega)} \leq C_\Omega(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}), \tag{3.1.19}$$

where $\alpha \in (\frac{1}{2}, 1]$ *is the index of elliptic regularity (cf. Theorem 2.31).*

Proof. We only consider (3.1.12) since the arguments for (3.1.13) are similar. We can write (3.1.12) as

$$\begin{aligned} a(q, \beta^{\frac{1}{2}} p) &= (y + f, q)_{L^2(\Omega)} \qquad && \forall q \in H_0^1(\Omega), \\ a(\beta^{\frac{1}{2}} y, z) &= (-p - g, z)_{L^2(\Omega)} \qquad && \forall z \in H_0^1(\Omega). \end{aligned}$$

Thus by the elliptic regularity (cf. Theorem 2.31) we have,

$$\begin{aligned} \|\beta^{\frac{1}{2}} p\|_{H^{1+\alpha}(\Omega)} &\lesssim \|y\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)}, \\ \|\beta^{\frac{1}{2}} y\|_{H^{1+\alpha}(\Omega)} &\lesssim \|p\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}. \end{aligned}$$

From (3.1.9), (3.1.12) and (3.1.6) we have

$$\begin{aligned} \|p\|_{L^2(\Omega)} + \|y\|_{L^2(\Omega)} &\leq \sup_{(q,z) \in H_0^1(\Omega) \times H_0^1(\Omega)} \frac{(f, q)_{L^2(\Omega)} + (g, z)_{L^2(\Omega)}}{\|q\|_{H^1_\beta(\Omega)} + \|z\|_{H^1_\beta(\Omega)}} \\ &\leq \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}. \end{aligned}$$

The estimate (3.1.19) immediately follows. □

We can now derive the concrete error estimates for the P_1 finite element methods.

Theorem 3.5. *Let* (p, y) *(resp.,* (p_h, y_h)*) be the solution of* (3.1.12) *or* (3.1.13) *(resp.,* (3.1.14) *or* (3.1.15)*). We have*

$$\|p - p_h\|_{H^1_\beta(\Omega)} + \|y - y_h\|_{H^1_\beta(\Omega)} \le C(1 + \beta^{\frac{1}{2}} h^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h^{1+\alpha} (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}),$$
$$\|p - p_h\|_{L^2(\Omega)} + \|y - y_h\|_{L^2(\Omega)} \le C(1 + \beta^{\frac{1}{2}} h^{-2}) \beta^{-1} h^{2+2\alpha} (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}),$$

where the positive constant C *is independent of* β *and* h *and* $\alpha \in (\frac{1}{2}, 1]$.

Proof. We only consider the case that involves (3.1.12) and (3.1.14). Let $\Pi_h : H^{1+\alpha}(\Omega) \cap H^1_0(\Omega) \to V_h$ be the nodal interpolation operator. We have the following standard interpolation error estimate [32, 43] for all $\zeta \in H^{1+\alpha}(\Omega) \cap H^1_0(\Omega)$:

$$\|\zeta - \Pi_h \zeta\|_{L^2(\Omega)} + h|\zeta - \Pi_h \zeta|_{H^1(\Omega)} \le C h^{1+\alpha} |\zeta|_{H^{1+\alpha}(\Omega)}, \tag{3.1.20}$$

where the positive constant C only depends on the shape regularity of $\mathcal{T}_h$.

The first estimate follows from (3.1.18), (3.1.19) and (3.1.20),

$$\begin{aligned}
\|p - p_h\|^2_{H^1_\beta(\Omega)} + \|y - y_h\|^2_{H^1_\beta(\Omega)} &\le \|p - \Pi_h p\|^2_{H^1_\beta(\Omega)} + \|y - \Pi_h y\|^2_{H^1_\beta(\Omega)} \\
&= \|p - \Pi_h p\|^2_{L^2(\Omega)} + \beta^{\frac{1}{2}} |p - \Pi_h p|^2_{H^1(\Omega)} + \|y - \Pi_h y\|^2_{L^2(\Omega)} + \beta^{\frac{1}{2}} |y - \Pi_h y|^2_{H^1(\Omega)} \\
&\le (\beta^{-1} h^{2+2\alpha} + \beta^{-\frac{1}{2}} h^{2\alpha}) \Big(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \Big) \\
&= (1 + \beta^{\frac{1}{2}} h^{-2}) \beta^{-1} h^{2+2\alpha} \Big(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)} \Big).
\end{aligned}$$

The second estimate is established by a duality argument. Let $(\xi, \theta) \in H^1_0(\Omega) \times H^1_0(\Omega)$ be defined by

$$\mathcal{B}((q, z), (\xi, \theta)) = (q, p - p_h)_{L^2(\Omega)} + (z, y - y_h)_{L^2(\Omega)}. \tag{3.1.21}$$

We have, by Galerkin orthogonality and Lemma 3.4,

$$\begin{aligned}
\|p-p_h\|^2_{L^2(\Omega)} + \|y-y_h\|^2_{L^2(\Omega)} &= \mathcal{B}((p-p_h, y-y_h),(\xi,\theta)) \\
&= \mathcal{B}((p-p_h, y-y_h),(\xi-\Pi_h\xi, \theta-\Pi_h\theta)) \\
&\lesssim (\|p-p_h\|^2_{H^1_\beta(\Omega)} + \|y-y_h\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} (\|\xi-\Pi_h\xi\|^2_{H^1_\beta(\Omega)} + \|\theta-\Pi_h\theta\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} \\
&\lesssim (1+\beta^{\frac{1}{2}}h^{-2})^{\frac{1}{2}}\beta^{-\frac{1}{2}}h^{1+\alpha}(\|p-p_h\|^2_{L^2(\Omega)} + \|y-y_h\|^2_{L^2(\Omega)})^{\frac{1}{2}} \\
&\quad \times (\|p-p_h\|^2_{H^1_\beta(\Omega)} + \|y-y_h\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}},
\end{aligned}$$

which implies the second estimate. □

Let us consider a P_1 finite element method for (3.1.1), to find $(\bar{p}_h, \bar{y}_h) \in V_h \times V_h$ such that

$$a(q_h, \bar{p}_h) - (q_h, \bar{y}_h)_{L^2(\Omega)} = -(q_h, y_d)_{L^2(\Omega)} \qquad \forall q_h \in V_h, \tag{3.1.22a}$$

$$-(\bar{p}_h, z_h)_{L^2(\Omega)} - \beta a(\bar{y}_h, z_h) = 0 \qquad \forall z \in V_h, \tag{3.1.22b}$$

which is equivalent to (3.1.11) under the change of variables,

$$\bar{p}_h = \beta^{\frac{1}{4}}\tilde{p}_h \quad \text{and} \quad \bar{y}_h = \beta^{-\frac{1}{4}}\tilde{y}_h. \tag{3.1.23}$$

Applying Theorem 3.5 to (3.1.22), we arrive at the following error estimates.

Lemma 3.6. *Let (p,y) (resp., (p_h, y_h)) be the solution of* (3.1.1) (*resp.,* (3.1.22)). *We have*

$$\|\bar{p}-\bar{p}_h\|_{H^1_\beta(\Omega)} + \beta^{\frac{1}{2}}\|\bar{y}-\bar{y}_h\|_{H^1_\beta(\Omega)} \le C(1+\beta^{\frac{1}{2}}h^{-2})^{\frac{1}{2}}h^{1+\alpha}\|y_d\|_{L^2(\Omega)}, \tag{3.1.24}$$

$$\|\bar{p}-\bar{p}_h\|_{L^2(\Omega)} + \beta^{\frac{1}{2}}\|\bar{y}-\bar{y}_h\|_{L^2(\Omega)} \le C(1+\beta^{\frac{1}{2}}h^{-2})\beta^{-\frac{1}{2}}h^{2+2\alpha}\|y_d\|_{L^2(\Omega)}, \tag{3.1.25}$$

where positive constant C is independent of β and h and $\alpha \in (\frac{1}{2}, 1]$.

Proof. Through the change of variables (3.1.2) and (3.1.23), we have

$$\|\tilde{p}-\tilde{p}_h\|_{H^1_\beta(\Omega)} = \beta^{-\frac{1}{4}}\|\bar{p}-\bar{p}_h\|_{H^1_\beta(\Omega)}, \quad \|\tilde{y}-\tilde{y}_h\|_{H^1_\beta(\Omega)} = \beta^{\frac{1}{4}}\|\bar{y}-\bar{y}_h\|_{H^1_\beta(\Omega)},$$

$$\|\tilde{p}-\tilde{p}_h\|_{L^2(\Omega)} = \beta^{-\frac{1}{4}}\|\bar{p}-\bar{p}_h\|_{L^2(\Omega)}, \quad \|\tilde{y}-\tilde{y}_h\|_{L^2(\Omega)} = \beta^{\frac{1}{4}}\|\bar{y}-\bar{y}_h\|_{L^2(\Omega)}.$$

Replace these norms in Theorem 3.5 and let $f = -\beta^{\frac{1}{4}} y_d$ and $g = 0$, we obtain the error estimates immediately. □

Remark 3.7. *According to Lemma 3.6, the performance of the P_1 finite element methods with respect to H^1 norm and L^2 norm will deteriorate as β goes to 0. Therefore it is necessary to use very fine mesh when β is small in which case it is important to have an efficient iterative solver, especially for three dimensional problems.*

The performance of the P_1 finite element method is illustrated in the following numerical example.

Example 3.8. *We solve* (3.1.1) *by the P_1 finite element method defined by* (3.1.22) *on the unit square $\Omega = (0,1)^2$ for $y_d = 1$ and $y_d = x_1(1-x_1)x_2(1-x_2)$. We take $\boldsymbol{\zeta} = \mathbf{0}$ and $\gamma = 0$ for simplicity. In both cases the exact solution can be found in the form of a double Fourier sine series. The relative errors for $h = 2^{-6}$ and various β together with the solution times are displayed in Table 3.1.*

Table 3.1. Relative Errors and Solution Times of the P_1 Finite Element Method Defined by (3.1.22) for $y_d = 1$ and $y_d = x_1(1-x_1)x_2(1-x_2)$, with $h = 2^{-6}$ and $\beta = 10^{-2}, 10^{-2}, 10^{-6}$.

β	$\frac{\lvert\bar{p}-\bar{p}_h\rvert_{H^1(\Omega)}}{\lvert\bar{p}\rvert_{H^1(\Omega)}}$	$\frac{\lVert\bar{p}-\bar{p}_h\rVert_{L^2(\Omega)}}{\lVert\bar{p}\rVert_{L^2(\Omega)}}$	$\frac{\lvert\bar{y}-\bar{y}_h\rvert_{H^1(\Omega)}}{\lvert\bar{y}\rvert_{H^1(\Omega)}}$	$\frac{\lVert\bar{y}-\bar{y}_h\rVert_{L^2(\Omega)}}{\lVert\bar{y}\rVert_{L^2(\Omega)}}$	Time (s)
$y_d = 1$					
10^{-2}	1.65e-02	6.92e-04	1.17e-02	6.31e-04	4.43e+00
10^{-4}	5.62e-02	4.64e-03	1.92e-02	9.29e-04	4.68e+00
10^{-6}	1.87e-01	3.99e-02	6.13e-01	4.51e-03	6.11e+00
$y_d = x_1(1-x_1)x_2(1-x_2)$					
10^{-2}	1.16e-02	2.65e-04	1.16e-02	4.26e-04	2.96e+00
10^{-4}	1.47e-02	1.88e-04	1.17e-02	1.92e-04	2.99e+00
10^{-6}	4.55e-02	8.82e-04	1.22e-02	1.86e-04	3.74e+00

3.2 Optimal Control Problems with Pointwise State Constraints

In this section we consider P_1 finite element methods for (1.1.3)–(1.1.5) under the assumption $\boldsymbol{\zeta} = \mathbf{0}$ and $\gamma = 0$.

According to (2.3.6)–(2.3.7) in Section 2.3, we consider the following problem,

$$\bar{y} = \underset{y \in K}{\operatorname{argmin}} \left[\frac{1}{2}(y - y_d, y - y_d)_{L^2(\Omega)} + \frac{\beta}{2}(\Delta y, \Delta y)_{L^2(\Omega)} \right], \tag{3.2.1}$$

where

$$K = \{ y \in V : y \leq \psi \quad \text{in } \Omega \}. \tag{3.2.2}$$

Let $\mathcal{T}_h$ be a shape regular simplicial triangulation of Ω and $V_h \subset H_0^1(\Omega)$ be the P_1 finite element space associated with $\mathcal{T}_h$. The diameter of $T \in \mathcal{T}_h$ is denoted by h_T and $h = \max_{T \in \mathcal{T}_h} h_T$ is the mesh diameter.

3.2.1 Discrete Laplace Operators

The discrete problems for (3.2.1) involve two discrete Laplace operators. The operator $\Delta_h : H_0^1(\Omega) \to V_h$ is defined by

$$(\Delta_h v, w)_{L^2(\Omega)} = -(\nabla v, \nabla w)_{L^2(\Omega)} \qquad \forall\ w \in V_h. \tag{3.2.3}$$

Let the inner product $(\cdot, \cdot)_h$ be defined by

$$(v, w)_h = \sum_{p \in \mathcal{V}_h} \left(\sum_{T \in \mathcal{T}_p} \frac{|T|}{d+1} \right) v(p) w(p) \quad \forall\ v, w \in V_h, \tag{3.2.4}$$

where $\mathcal{V}_h$ is the set of the vertices of $\mathcal{T}_h$, $\mathcal{T}_p$ is the set of the elements in $\mathcal{T}_h$ that share p as a common vertex, and $|T|$ is the area ($d = 2$) or volume ($d = 3$) of T. We have $(v, v)_h \approx \|v\|^2_{L^2(\Omega)}$ for all $v \in V_h$ (cf. [32]). The operator $\widetilde{\Delta}_h : H_0^1(\Omega) \to V_h$ is defined by

$$(\widetilde{\Delta}_h v, w)_h = -(\nabla v, \nabla w)_{L^2(\Omega)} \qquad \forall\ w \in V_h. \tag{3.2.5}$$

Remark 3.9. *The first discrete Laplace operator Δ_h is defined in terms of the L^2 inner product while the second one $\widetilde{\Delta}_h$ is defined in terms of the discrete inner product* (3.2.4) *related to mass lumping.*

3.2.2 Discrete Problems

We provide three P_1 finite element methods utilizing the discrete Laplace operators Δ_h and $\widetilde{\Delta}_h$.

The first discrete problem is to find

$$\bar{y}_h = \underset{y_h \in K_h}{\operatorname{argmin}} \left[\frac{1}{2}(y_h - y_d, y_h - y_d)_{L^2(\Omega)} + \frac{\beta}{2}(\Delta_h y_h, \Delta_h y_h)_{L^2(\Omega)} \right], \tag{3.2.6}$$

the second discrete problem is to find

$$\bar{y}_h = \underset{y_h \in K_h}{\operatorname{argmin}} \left[\frac{1}{2}(y_h - y_d, y_h - y_d)_{L^2(\Omega)} + \frac{\beta}{2}(\widetilde{\Delta}_h y_h, \widetilde{\Delta}_h y_h)_h \right], \tag{3.2.7}$$

the third discrete problem is to find

$$\bar{y}_h = \underset{y_h \in K_h}{\operatorname{argmin}} \left[\frac{1}{2}(y_h - y_d, y_h - y_d)_h + \frac{\beta}{2}(\widetilde{\Delta}_h y_h, \widetilde{\Delta}_h y_h)_h \right], \tag{3.2.8}$$

where

$$K_h = \{ y \in V_h : y_h \leq \psi \quad \text{at the vertices of } \mathcal{T}_h \}. \tag{3.2.9}$$

Remark 3.10. *The P_1 finite element method defined by* (3.2.6) *and* (3.2.9) *can be found in [34, 42, 80].*

We can derive the following analogs of (2.3.28) which are the first order optimality conditions for (3.2.6), (3.2.7) and (3.2.8),

$$(\bar{y}_h - y_d, y_h - \bar{y}_h)_{L^2(\Omega)} + \beta \Big(\Delta_h \bar{y}_h, \Delta_h (y_h - \bar{y}_h) \Big)_{L^2(\Omega)} \geq 0 \qquad \forall y_h \in K_h, \tag{3.2.10}$$

$$(\bar{y}_h - y_d, y_h - \bar{y}_h)_{L^2(\Omega)} + \beta \Big(\widetilde{\Delta}_h \bar{y}_h, \widetilde{\Delta}_h (y_h - \bar{y}_h) \Big)_h \geq 0 \qquad \forall y_h \in K_h, \tag{3.2.11}$$

$$(\bar{y}_h - y_d, y_h - \bar{y}_h)_h + \beta \Big(\widetilde{\Delta}_h \bar{y}_h, \widetilde{\Delta}_h (y_h - \bar{y}_h) \Big)_h \geq 0 \qquad \forall y_h \in K_h. \tag{3.2.12}$$

Remark 3.11. *Let $\mathbf{A}_h$ (resp., $\mathbf{M}_h$) be the stiffness (resp., mass) matrix represents the bilinear form $(\nabla \cdot, \nabla \cdot)_{L^2(\Omega)}$ (resp., $(\cdot, \cdot)_{L^2(\Omega)}$) with respect to the nodal basis of V_h. Then Δ_h can be represented as $-\mathbf{M}_h^{-1}\mathbf{A}_h$. On the other hand, the matrix representing $\widetilde{\Delta}_h$ is given by $-\widetilde{\mathbf{M}}_h^{-1}\mathbf{A}_h$, where $\widetilde{\mathbf{M}}_h$ is the diagonal matrix representing the bilinear form defined by* (3.2.4).

We are able to solve (3.2.7) *and* (3.2.8) *with a primal-dual active set algorithm since the system matrix* $\mathbf{M}_h + \beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h$ *or* $\widetilde{\mathbf{M}}_h + \beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h$ *is available. This also enables us to construct multigrid methods to solve the reduced system during each PDAS iteration. That is the main reason we utilize the mass lumping inner product* (3.2.4).

3.2.3 Error Estimates

We have the following error estimate for the P_1 finite element methods.

Theorem 3.12 ([34, Section 6])**.** *Let* $\bar{y}_h \in K_h$ *be the solution of* (3.2.6) (*or* (3.2.7), (3.2.8)) *and* $\bar{u}_h = -\Delta_h \bar{y}_h$ *(or* $\bar{u}_h = -\widetilde{\Delta}_h \bar{y}_h$*). We have*

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} + \|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} + |\bar{y} - \bar{y}_h|_{H^1(\Omega)} \leq C(|\ln h|^{\frac{1}{2}} h + h^{\tau}), \tag{3.2.13}$$

where

$$\tau = \begin{cases} \alpha & \text{if } d = 2 \text{ or } 3 \text{ and } \mathcal{T}_h \text{ is quasi-uniform,} \\ 1 & \text{if } d = 2 \text{ and } \mathcal{T}_h \text{ is graded around the reentrant corners.} \end{cases}$$

$\alpha \in (\frac{1}{2}, 1]$ *is the index of elliptic regularity.*

Remark 3.13. *We refer to [34] for a detailed proof of the theorem. Note that the constant* C *in* (3.2.13) *might depend on* β*. We do not attempt to explore the relation between* C *and* β *in this theorem.*

3.3 Numerical Results

In this section we present several numerical results to illustrate the behavior of the P_1 finite element method (3.1.14) and (3.2.8). We employed the MATLAB/C++ toolbox FELICITY [101] in our computations.

Example 3.14. *In this example we solve* (3.1.12) *on* $\Omega = (0,1)^2$ *with* $\beta = 1$, $\boldsymbol{\zeta} = \frac{1}{2}[1,0]^t$, $\gamma = 0$ *and exact solution*

$$(p, y) = (\sin(2\pi x_1)\sin(2\pi x_2), x_1(1-x_1)x_2(1-x_2)). \tag{3.3.1}$$

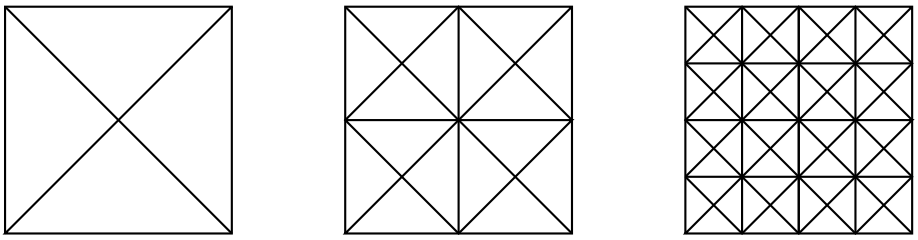

Figure 3.1. Criss-Cross Mesh.

There are 10 *degrees of freedom (dofs) for the* P_1 *finite element space associated with the first level mesh* $(k = 1)$*. After* 11 *uniform mesh refinements, the* P_1 *finite element space associated with the final mesh* $(k = 12)$ *has* 67092482 *dofs. See Figure 3.1 for the first three meshes in this example. The relative errors are displayed in Tables 3.2. We observe* $O(h)$ *convergence in* $|\cdot|_{H^1(\Omega)}$ *and* $O(h^2)$ *convergence in* $\|\cdot\|_{L^2(\Omega)}$*, which agrees with Theorem 3.5.*

Table 3.2. Convergence Results for Example 3.14.

k	$\frac{\lvert\bar{y}-y_h\rvert_{H^1(\Omega)}}{\lvert\bar{y}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lVert\bar{y}-y_h\rVert_{L^2(\Omega)}}{\lVert\bar{y}\rVert_{L^2(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{H^1(\Omega)}}{\lvert\bar{p}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{L^2(\Omega)}}{\lvert\bar{p}\rvert_{L^2(\Omega)}}$	Order
1	2.77e-01	-	1.96e-01	-	1.60e-01	-	1.45e-01	-
2	1.33e-01	1.06	6.97e-02	1.49	1.92e-01	-0.27	1.36e-01	0.09
3	5.83-e02	1.19	2.01e-02	1.79	9.54e-02	1.01	4.20e-02	1.70
4	2.67-e02	1.13	5.30e-03	1.92	4.67e-02	1.03	1.10e-02	1.93
5	1.27e-02	1.07	1.40e-03	1.92	2.32e-02	1.01	2.80e-03	1.97
6	6.20e-03	1.03	3.45e-04	2.02	1.16e-02	1.00	7.00e-04	2.00
7	3.10e-03	1.00	8.67e-05	1.99	5.80e-03	1.00	1.75e-04	2.00
8	1.50e-03	1.05	2.17e-05	2.00	2.90e-03	1.00	4.38e-05	2.00
9	7.65e-04	0.97	5.44e-06	2.00	1.40e-03	1.05	1.10e-05	1.99
10	3.82e-04	1.00	1.36e-06	2.00	7.23e-04	0.95	2.74e-06	2.01
11	1.91e-04	1.00	3.40e-07	2.00	3.62e-04	1.00	6.85e-07	2.00
12	9.55e-05	1.00	8.51e-08	2.00	1.81e-04	1.00	1.71e-07	2.00

Example 3.15. *In this example we solve* (3.1.12) *on a L-shaped domain* $\Omega = (0,1)^2 \setminus (0.5)^2$ *with* $\beta = 1$*,* $\boldsymbol{\zeta} = \frac{1}{2}[1,0]^t$*,* $\gamma = 0$ *and exact solution*

$$(p, y) = (\sin(2\pi x_1)\sin(2\pi x_2), x_1(1-x_1)(0.5-x_1)x_2(1-x_2)(0.5-x_2)). \quad (3.3.2)$$

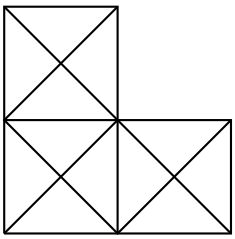

Figure 3.2. Initial Mesh for L-shaped Domain.

There are 34 *degrees of freedom (dofs) for the* P_1 *finite element space associated with the first level mesh* $(k = 1)$*. After* 9 *uniform mesh refinements, the* P_1 *finite element space associated with the final mesh* $(k = 10)$ *has* 12574722 *dofs. See Figure 3.2 for the initial mesh in this example. The relative errors are displayed in Table 3.3. We observe* $O(h)$ *convergence in* $|\cdot|_{H^1(\Omega)}$ *and* $O(h^2)$ *convergence in* $\|\cdot\|_{L^2(\Omega)}$*, which is higher than the convergence order in Theorem 3.5.*

Table 3.3. Convergence Results for Example 3.15.

k	$\frac{\lvert\bar{y}-y_h\rvert_{H^1(\Omega)}}{\lvert\bar{y}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lVert\bar{y}-y_h\rVert_{L^2(\Omega)}}{\lVert\bar{y}\rVert_{L^2(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{H^1(\Omega)}}{\lvert\bar{p}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{L^2(\Omega)}}{\lvert\bar{p}\rvert_{L^2(\Omega)}}$	Order
1	6.24e-01	-	3.17e-01	-	3.93e+00	-	3.28e-01	-
2	2.93e-01	1.09	1.46e-01	1.12	1.99e+00	0.98	9.86e-02	1.73
3	1.27e-01	1.20	4.36e-02	1.74	1.00e+00	0.99	2.61e-02	1.92
4	5.95e-02	1.10	1.15e-02	1.93	5.03e-01	1.00	6.63e-03	1.98
5	2.91e-02	1.03	2.91e-03	1.98	2.52e-01	1.00	1.67e-03	1.99
6	1.45e-02	1.01	7.31e-04	1.99	1.26e-01	1.00	4.17e-04	2.00
7	7.21e-03	1.00	1.83e-04	2.00	6.28e-02	1.00	1.04e-04	2.00
8	3.58e-03	1.01	4.59e-05	2.00	3.12e-02	1.01	2.59e-05	2.01
9	1.75e-03	1.04	1.16e-05	1.99	1.52e-02	1.04	6.34e-06	2.03
10	7.81e-04	1.16	3.01e-06	1.95	6.82e-03	1.16	1.44e-06	2.14

Example 3.16. *In this example we solve* (3.1.12) *on* $\Omega = (0,1)^3$ *with* $\beta = 1$*,* $\boldsymbol{\zeta} = \frac{1}{2}[1,0,0]^t$*,* $\gamma = 0$ *and exact solution*

$$(p, y) = (\sin(2\pi x_1)\sin(2\pi x_2)\sin(2\pi x_3), x_1(1-x_1)x_2(1-x_2)x_3(1-x_3)). \quad (3.3.3)$$

There are 54 *degrees of freedom (dofs) for the* P_1 *finite element space associated with the first level mesh* ($k = 1$). *After* 5 *uniform mesh refinements, the* P_1 *finite element space associated with the final mesh* ($k = 6$) *has* 4096766 *dofs. The relative errors are displayed in Table 3.4. We observe* $O(h)$ *convergence in* $|\cdot|_{H^1(\Omega)}$ *and* $O(h^2)$ *convergence in* $\|\cdot\|_{L^2(\Omega)}$, *which agrees with Theorem 3.5.*

Table 3.4. Convergence Results for Example 3.16.

k	$\frac{\lvert\bar{y}-y_h\rvert_{H^1(\Omega)}}{\lvert\bar{y}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lVert\bar{y}-y_h\rVert_{L^2(\Omega)}}{\lVert\bar{y}\rVert_{L^2(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{H^1(\Omega)}}{\lvert\bar{p}\rvert_{H^1(\Omega)}}$	Order	$\frac{\lvert\bar{p}-p_h\rvert_{L^2(\Omega)}}{\lvert\bar{p}\rvert_{L^2(\Omega)}}$	Order
1	5.57e-01	-	3.84e-01	-	8.04e-01	-	7.41e-01	-
2	3.07e-01	0.86	1.35e-01	1.51	4.86e-01	0.73	3.44e-01	1.11
3	1.43e-01	1.10	4.24e-02	1.67	2.51e-01	0.95	1.06e-01	1.70
4	6.68e-02	1.10	1.15e-02	1.89	1.26e-01	1.00	2.80e-02	1.92
5	3.18e-02	1.07	2.91e-03	1.98	6.14e-02	1.03	6.99e-03	2.00
6	1.41e-02	1.17	7.13e-04	2.03	2.74e-02	1.16	1.63e-03	2.10

Example 3.17 ([34, Example 7.1])**.** *In this example we take* $\Omega = [-4, 4]^2$, $\beta = 1$, $\psi = |x|^2 - 1$ *and*

$$y_d = \begin{cases} \Delta^2\bar{y} + \bar{y} & |x| > 1 \\ \Delta^2\bar{y} + \bar{y} + 2 & |x| \le 1 \end{cases}$$

in (3.2.8), *where*

$$\bar{y} = \begin{cases} |x|^2 - 1 & |x| \le 1 \\ v(|x|) + (1 - \phi(|x|))w(x) & 1 \le |x| \le 3 \\ w(x) & |x| \ge 3 \end{cases},$$

with

$$\begin{aligned} v(|x|) &= (|x|^2 - 1)(1 - \frac{|x| - 1}{2})^4 + \frac{1}{4}(|x| - 1)^2(|x| - 3)^4, \\ \phi(|x|) &= (1 + 4\frac{|x| - 1}{2} + 10(\frac{|x| - 1}{2})^2 + 20(\frac{|x| - 1}{2})^3)(1 - \frac{|x| - 1}{2})^4, \\ w(x) &= 2\sin(\frac{\pi}{8}(x_1 + 4))\sin(\frac{\pi}{8}(x_2 + 4)). \end{aligned}$$

We report the absolute errors in Table 3.5. We observe $O(h)$ *convergence for the state in* H^1 *norm which agrees with Theorem 3.12. The convergence rate of the state*

in L^2 norm is close to $O(h^2)$ in average and the convergence rate of the control in L^2 norm is close to $O(h^{1.5})$ which are better than the estimates in Theorem 3.12. These convergence rates also coincide with those of [34, Example 7.1].

Table 3.5. Convergence Results for Example 3.17.

k	$\|\bar{y} - y_h\|_{L^2(\Omega)}$	Order	$\|\bar{y} - y_h\|_{H^1(\Omega)}$	Order	$\|\bar{u} - u_h\|_{L^2(\Omega)}$	Order
1	1.97e+01	-	1.71e+01	-	1.44e+01	-
2	1.21e+00	4.03	3.94e+00	2.12	5.88e+00	1.29
3	8.60e-01	0.49	1.85e+00	1.09	4.18e+00	0.49
4	3.51e-01	1.29	8.58e-01	1.11	1.46e+00	1.52
5	1.10e-01	1.67	3.96e-01	1.12	4.83e-01	1.60
6	4.41e-02	1.32	1.94e-01	1.03	1.79e-01	1.43
7	1.31e-02	1.75	9.47e-02	1.03	6.12e-02	1.55
8	1.27e-03	3.37	4.57e-02	1.05	2.16e-02	1.50
9	1.12e-03	0.18	2.05e-02	1.16	7.61e-03	1.51

Chapter 4

Multigrid Methods for Elliptic Optimal Control Problems

In this chapter we focus on the multigrid method for (2.3.24). As we discussed in Chapter 3, our goal is to design multigrid methods for the general problem (3.1.14). We follow the approach introduced in Section 2.8 which involves smoothing property and approximation property. The main ingredient is to use a post-smoother that can be interpreted as a Richardson iteration (cf. (2.5.6)) for a SPD problem that has the same solution as the saddle point problem (3.1.14). We only focus on the convergence analysis of W-cycle algorithm. We refer to [30, 29, 28, 31] for more details about this approach. The materials in this chapter come from [30].

Throughout this chapter, $\Omega \in \mathbb{R}^d (d = 2, 3)$ is a convex domain. Let $\mathcal{T}_0$ be a triangulation of Ω and the the triangulations $\mathcal{T}_1, \mathcal{T}_2, ...$ be generated from $\mathcal{T}_0$ through a refinement process so that $h_k = h_{k-1}/2$. The P_1 finite element subspace of $H_0^1(\Omega)$ associated with $\mathcal{T}_k$ is denoted by V_k. We use C to denote a generic constant which is independent of mesh size. Also to avoid the proliferation of the constants, we use the notation $A \lesssim B$ (or $A \gtrsim B$) to represent $A \leq (constant)B$. The notation $A \approx B$ means that $A \lesssim B$ and $B \lesssim A$.

4.1 Multigrid Algorithm

We want to design multigrid methods for problems of finding $(p, y) \in V_k \times V_k$ such that

$$\mathcal{B}((p, y), (q, z)) = F(q) + G(z) \qquad \forall\, (q, z) \in V_k \times V_k, \tag{4.1.1}$$

where $F, G \in V_k'$, and for the dual problem of finding $(p, y) \in V_k \times V_k$ such that

$$\mathcal{B}((q, z), (p, y)) = F(q) + G(z) \qquad \forall\, (q, z) \in V_k \times V_k. \tag{4.1.2}$$

4.1.1 Mesh-dependent Inner Product

Similar to (2.8.4), we define a mesh-dependent inner product on V_k.

Definition 4.1. *The mesh-dependent inner product* $(\cdot,\cdot)_k$ *on* V_k *is defined by*

$$(v,w)_k := h_k^d \sum_{i=1}^{n_k} v(p_i)w(p_i), \tag{4.1.3}$$

where $\{p_i\}_{i=1}^{n_k}$ *is the set of internal vertices of* $\mathcal{T}_k$.

By a standard scaling argument [32, 43], we have

$$(v,v) \approx \|v\|^2_{L^2(\Omega)} \quad \forall v \in V_k, \tag{4.1.4}$$

where the hidden constants only depend on the shape regularity of $\mathcal{T}_0$.

Then we introduce a mesh dependent inner product on $V_k \times V_k$ to rewrite (4.1.1) in terms of an operator that maps $V_k \times V_k$ to $V_k \times V_k$. Define the mesh-dependent inner product $[\cdot,\cdot]_k$ on $V_k \times V_k$ by

$$[(p,y),(q,z)]_k = (p,q)_k + (y,z)_k. \tag{4.1.5}$$

Let the operator $\mathfrak{B}_k : V_k \times V_k \longrightarrow V_k \times V_k$ be defined by

$$[\mathfrak{B}_k(p,y),(q,z)]_k = \mathcal{B}((p,y),(q,z)) \qquad \forall\,(p,y),(q,z) \in V_k \times V_k. \tag{4.1.6}$$

We can then rewrite (4.1.1) in the form

$$\mathfrak{B}_k(p,y) = (f,g), \tag{4.1.7}$$

where $(f,g) \in V_k \times V_k$ is defined by

$$[(f,g),(q,z)]_k = F(q) + G(z) \qquad \forall\,(q,z) \in V_k \times V_k,$$

and (4.1.2) becomes

$$\mathfrak{B}_k^t(p,y) = (f,g), \tag{4.1.8}$$

where

$$\begin{aligned} [\mathfrak{B}_k^t(p,y),(q,z)]_k &= [(p,y),\mathfrak{B}_k(q,z)]_k \\ &= \mathcal{B}((q,z),(p,y)) \qquad \forall (p,y),(q,z) \in V_k \times V_k. \end{aligned} \tag{4.1.9}$$

Similar to Section 2.8, we take the coarse-to-fine operator $I_{k-1}^{k} : V_{k-1} \times V_{k-1} \longrightarrow V_k \times V_k$ to be the natural injection and define the fine-to-coarse operator $I_k^{k-1} : V_k \times V_k \longrightarrow V_{k-1} \times V_{k-1}$ to be the transpose of I_{k-1}^{k} with respect to the mesh-dependent inner products, i.e.,

$$[I_k^{k-1}(p,y),(q,z)]_{k-1} = [(p,y), I_{k-1}^{k}(q,z)]_k \quad \forall\, (p,y) \in V_k \times V_k, (q,z) \in V_{k-1} \times V_{k-1}. \tag{4.1.10}$$

4.1.2 A Block-diagonal Preconditioner

Let $L_k : V_k \longrightarrow V_k$ be a linear operator symmetric with respect to the inner product $(\cdot,\cdot)_k$ on V_k such that

$$(L_k v, v)_k \approx \|v\|^2_{H^1_\beta(\Omega)} = \|v\|^2_{L^2(\Omega)} + \beta^{\frac{1}{2}} |v|^2_{H^1(\Omega)} \qquad \forall\, v \in V_k. \tag{4.1.11}$$

Then the operator $\mathfrak{C}_k : V_k \times V_k \longrightarrow V_k \times V_k$ defined by

$$\mathfrak{C}_k(p,y) = (L_k p, L_k y) \tag{4.1.12}$$

is symmetric positive definite (SPD) with respect to $[\cdot,\cdot]_k$ and we have

$$[\mathfrak{C}_k(p,y),(p,y)]_k \approx \|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)} \qquad \forall\, (p,y) \in V_k \times V_k. \tag{4.1.13}$$

Remark 4.2. *We will use $\mathfrak{C}_k^{-1}$ as a preconditioner in the constructions of the smoothing operators. In practice we can take L_k^{-1} to be an approximate solve of the P_1 finite element discretization of the following boundary value problem:*

$$-\beta^{\frac{1}{2}} \Delta u + u = f \quad \text{in} \quad \Omega \quad \text{and} \quad u = 0 \quad \text{on} \quad \partial\Omega. \tag{4.1.14}$$

The multigrid algorithms in this section are $O(n)$ as long as L_k^{-1} is also an $O(n)$ algorithm. This can be done using the multigrid algorithm introduced in Section 2.8 (apply to (4.1.14)). In practice, a V-cycle algorithm with very few smoothing steps is sufficient. Numerical results are presented in Section 4.4 to illustrate the effects of the preconditioner.

We refer to [15, 77] for a general discussion on the construction of block-diagonal preconditioners for saddle point problems arising from the discretization of partial differential equations.

Lemma 4.3. *We have*

$$[\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y),(p,y)]_k \approx \|p\|^2_{H^1_{\vec{\beta}}(\Omega)} + \|y\|^2_{H^1_{\vec{\beta}}(\Omega)} \qquad \forall\,(p,y)\in V_k\times V_k, \quad (4.1.15)$$

$$[\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t(p,y),(p,y)]_k \approx \|p\|^2_{H^1_{\vec{\beta}}(\Omega)} + \|y\|^2_{H^1_{\vec{\beta}}(\Omega)} \qquad \forall\,(p,y)\in V_k\times V_k, \quad (4.1.16)$$

where the hidden constants are independent of k and β.

Proof. Let $(p,y)\in V_k\times V_k$ be arbitrary and $(r,x)=\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y)$. By (3.1.16), (4.1.6), (4.1.13) and duality, we derive (4.1.15) as follows,

$$\begin{aligned}
[\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y),(p,y)]_k &= [\mathfrak{C}_k(\mathfrak{C}_k^{-1}\mathfrak{B}_k)(p,y),\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y)]_k\\
&= [\mathfrak{C}_k(r,x),(r,x)]_k\\
&= \sup_{(q,z)\in V_k\times V_k}\frac{[\mathfrak{C}_k(r,x),(q,z)]_k^2}{[\mathfrak{C}_k(q,z),(q,z)]_k}\\
&\approx \sup_{(q,z)\in V_k\times V_k}\frac{[\mathfrak{B}_k(p,y),(q,z)]_k^2}{\|q\|^2_{H^1_{\vec{\beta}}(\Omega)}+\|z\|^2_{H^1_{\vec{\beta}}(\Omega)}}\\
&= \sup_{(q,z)\in V_k\times V_k}\frac{\mathcal{B}((p,y),(q,z))^2}{\|q\|^2_{H^1_{\vec{\beta}}(\Omega)}+\|z\|^2_{H^1_{\vec{\beta}}(\Omega)}} \approx \|p\|^2_{H^1_{\vec{\beta}}(\Omega)}+\|y\|^2_{H^1_{\vec{\beta}}(\Omega)}.
\end{aligned}$$

The derivation of (4.1.16) is analogous with (3.1.16) (resp., (4.1.6)) replaced by (3.1.17) (resp., (4.1.9)). □

Lemma 4.4. *We have the following bounds on the minimum and maximum eigenvalues of $\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k$ and $\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t$:*

$$\lambda_{\min}(\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k),\lambda_{\min}(\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t)\geq C_{\min}, \quad (4.1.17)$$

$$\lambda_{\max}(\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k),\lambda_{\max}(\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t)\leq C_{\max}(1+\beta^{\frac{1}{2}}h_k^{-2}), \quad (4.1.18)$$

where the positive constants $C_{\min}$ and $C_{\max}$ are independent of k and β.

Proof. We only derive the estimates for $\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k$ since the derivation of $\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t$ is similar. We have, from (4.1.4) and (4.1.5),

$$[(p,y),(p,y)]_k \approx \|p\|^2_{L^2(\Omega)} + \|y\|^2_{L^2(\Omega)} \qquad \forall\,(p,y)\in V_k\times V_k, \tag{4.1.19}$$

where the hidden constants only depend on the shape regularity of $\mathcal{T}_0$. It follows from (3.1.6), (4.1.15) and (4.1.19) that

$$[\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y),(p,y)]_k \geq C_{\min}[(p,y),(p,y)]_k \qquad \forall\,(p,y)\in V_k\times V_k, \tag{4.1.20}$$

which then implies (4.1.17) by the Rayleigh quotient formula.

By a standard inverse estimate [32, 43], we have

$$\|v\|^2_{H^1_{\beta}(\Omega)} = \|v\|^2_{L^2(\Omega)} + \beta^{\frac{1}{2}}|v|^2_{H^1(\Omega)} \leq (1+C\beta^{\frac{1}{2}}h_k^{-2})\|v\|^2_{L^2(\Omega)} \qquad \forall\, v\in V_k,$$

where the positive constant C depends only on the shape regularity of $\mathcal{T}_0$. It then follows from (3.1.6), (4.1.15) and (4.1.19) that

$$[\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k(p,y),(p,y)]_k \leq C_{\max}(1+\beta^{\frac{1}{2}}h_k^{-2})[(p,y),(p,y)]_k \qquad \forall\,(p,y)\in V_k\times V_k,$$

and hence (4.1.18) holds because of the Rayleigh quotient formula. □

Remark 4.5. *It follows from Theorem 4.4 that the operators* $\mathfrak{B}_k^t\mathfrak{C}_k^{-1}\mathfrak{B}_k$ *and* $\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t$ *are well-conditioned when* $\beta^{\frac{1}{2}}h_k^{-2}=O(1)$. *When* β *is fixed,* $\beta^{\frac{1}{2}}h_k^{-2}=O(1)$ *is satisfied at relative lower levels. We will exploit this fact and* (2.5.7) *for the choice of the damping parameter in the smoothers of our multigrid methods.*

4.1.3 A W-cycle Multigrid Algorithm for (4.1.7)

Let the output of the W-cycle algorithm for (4.1.7) with initial guess (p_0,y_0) and m_1 (resp., m_2) pre-smoothing (resp., post-smoothing) steps be denoted by $MG_W(k,(f,g),(p_0,y_0),m_1,m_2)$.

We use a direct solve for $k = 0$, i.e., we take $MG_W(0, (f, g), (p_0, y_0), m_1, m_2)$ to be $\mathfrak{B}_0^{-1}(f, g)$. For $k \geq 1$, we compute $MG_W(k, (f, g), (p_0, y_0), m_1, m_2)$ in three steps.

Pre-Smoothing The approximate solutions $(p_1, y_1), \ldots, (p_{m_1}, y_{m_1})$ are computed recursively by

$$(p_j, y_j) = (p_{j-1}, y_{j-1}) + \lambda_k \mathfrak{C}_k^{-1} \mathfrak{B}_k^t ((f, g) - \mathfrak{B}_k(p_{j-1}, y_{j-1})) \tag{4.1.21}$$

for $1 \leq j \leq m_1$. The choice of the damping factor λ_k will be given below in (4.1.24) and (4.1.25).

Coarse Grid Correction Let $(f', g') = I_k^{k-1}((f, g) - B_k(p_{m_1}, y_{m_1}))$ be the transferred residual of (p_{m_1}, y_{m_1}) and compute $(p_1', y_1'), (p_2', y_2') \in V_{k-1} \times V_{k-1}$ by

$$(p_1', y_1') = MG_W(k-1, (f', g'), (0, 0), m_1, m_2), \tag{4.1.22a}$$

$$(p_2', y_2') = MG_W(k-1, (f', g'), (p_1', y_1'), m_1, m_2). \tag{4.1.22b}$$

We then take (p_{m_1+1}, y_{m_1+1}) to be $(p_{m_1}, y_{m_1}) + I_{k-1}^k(p_2', y_2')$.

Post-Smoothing The approximate solutions $(p_{m_1+2}, y_{m_1+2}), \ldots, (p_{m_1+m_2+1}, y_{m_1+m_2+1})$ are computed recursively by

$$(p_j, y_j) = (p_{j-1}, y_{j-1}) + \lambda_k \mathfrak{B}_k^t \mathfrak{C}_k^{-1} ((f, g) - \mathfrak{B}_k(p_{j-1}, y_{j-1})) \tag{4.1.23}$$

for $m_1 + 2 \leq j \leq m_1 + m_2 + 1$.

The final output is $MG_W(k, (f, g), (p_0, y_0), m_1, m_2) = (p_{m_1+m_2+1}, y_{m_1+m_2+1})$.

To complete the description of the algorithm, we choose the damping factor λ_k as follows:

$$\lambda_k = \frac{2}{\lambda_{\min}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) + \lambda_{\max}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)} \qquad \text{if } \beta^{\frac{1}{2}} h_k^{-2} < 1, \tag{4.1.24}$$

and

$$\lambda_k = [C_\dagger(1 + \beta^{\frac{1}{2}} h_k^{-2})]^{-1} \qquad \text{if } \beta^{\frac{1}{2}} h_k^{-2} \geq 1, \tag{4.1.25}$$

where $C_\dagger$ is greater than or equal to the constant $C_{\max}$ in (4.1.18).

Remark 4.6. *Note that the post-smoothing step is exactly the Richardson iteration for the equation*

$$\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k (p, y) = \mathfrak{B}_k^t \mathfrak{C}_k^{-1} (f, g),$$

which is equivalent to (4.1.7).

Remark 4.7. *In the case where $\beta^{\frac{1}{2}} h_k^{-2} < 1$, the choice of λ_k is motivated by the well-conditioning of $\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k$ (cf. Remark 4.5) and the optimal choice of damping factor for the Richardson iteration (Section 2.5.1). In the case where $\beta^{\frac{1}{2}} h_k^{-2} \geq 1$, the choice of λ_k is motivated by the condition $\lambda_{\max}(\lambda_k \mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) \leq 1$ (cf.* (4.1.18)) *that will ensure the highly oscillatory part of the error is damped out when Richardson iteration is used as a smoother for an ill-conditioned system.*

Remark 4.8. *In practice, $\lambda_{\min}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)$ and $\lambda_{\max}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)$ are estimated by power iterations or Lanczos algorithm on coarse grids. We refer to [88, 52] for more details.*

4.1.4 A V-cycle Multigrid Algorithm for (4.1.7)

Let the output of the V-cycle algorithm for (4.1.7) with initial guess (p_0, y_0) and m_1 (resp., m_2) pre-smoothing (resp., post-smoothing) steps be denoted by $MG_V(k, (f, g), (p_0, y_0), m_1, m_2)$. The difference between the computations of $MG_V(k, (f, g), (p_0, y_0), m_1, m_2)$ and $MG_W(k, (f, g), (p_0, y_0), m_1, m_2)$ is only in the coarse grid correction step, where we compute

$$(p_1', y_1') = MG_V(k-1, (f', g'), (0, 0), m_1, m_2).$$

Remark 4.9. *We will focus on the analysis of the W-cycle algorithm in this dissertation. But numerical results indicate that the performance of the V-cycle algorithm is robust with respect to k and β (cf. Section 4.4).*

4.1.5 Multigrid Algorithms for (4.1.8)

We can define W-cycle and V-cycle algorithms for (4.1.8) by simply interchanging the operators $\mathfrak{B}_k$ and $\mathfrak{B}_k^t$ in Sections 4.1.3 and 4.1.4. In particular, the pre-smoothing step is given by

$$(p_j, y_j) = (p_{j-1}, y_{j-1}) + \lambda_k \mathfrak{C}_k^{-1} \mathfrak{B}_k((f, g) - \mathfrak{B}_k^t(p_{j-1}, y_{j-1})), \tag{4.1.26}$$

and the post-smoothing step is given by

$$(p_j, y_j) = (p_{j-1}, y_{j-1}) + \lambda_k \mathfrak{B}_k \mathfrak{C}_k^{-1}((f, g) - \mathfrak{B}_k^t(p_{j-1}, y_{j-1})). \tag{4.1.27}$$

4.2 Smoothing and Approximation Properties

As we discussed in Section 2.8, we will develop in this section two key ingredients for the convergence analysis of the W-cycle algorithm, namely, the smoothing and approximation properties. They will be expressed in terms of two scales of mesh-dependent norms defined by

$$|||(p, y)|||_{s,k} = [(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)^s (p, y), (p, y)]_k^{\frac{1}{2}} \qquad \forall (p, y) \in V_k \times V_k, \tag{4.2.1}$$

$$|||(p, y)|||_{s,k}^{\sim} = [(\mathfrak{B}_k \mathfrak{C}_k^{-1} \mathfrak{B}_k^t)^s (p, y), (p, y)]_k^{\frac{1}{2}} \qquad \forall (p, y) \in V_k \times V_k. \tag{4.2.2}$$

Note that

$$|||(p, y)|||_{0,k}^2 \approx \|p\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2 \approx (|||(p, y)|||_{0,k}^{\sim})^2 \qquad \forall\, (p, y) \in V_k \times V_k \tag{4.2.3}$$

by (4.1.19), and

$$|||(p, y)|||_{1,k}^2 \approx \|p\|_{H_\beta^1(\Omega)}^2 + \|y\|_{H_\beta^1(\Omega)}^2 \approx (|||(p, y)|||_{1,k}^{\sim})^2 \qquad \forall\, (p, y) \in V_k \times V_k \tag{4.2.4}$$

by (4.1.15) and (4.1.16).

4.2.1 Post-smoothing Properties

The error propagation operator for one post-smoothing step defined by (4.1.23) is given by

$$R_k = Id_k - \lambda_k \mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k, \tag{4.2.5}$$

where Id_k is the identity operator on $V_k \times V_k$. Similarly, the error propagation operator for one post-smoothing step defined by (4.1.27) is given by

$$\tilde{R}_k = Id_k - \lambda_k \mathfrak{B}_k \mathfrak{C}_k^{-1} \mathfrak{B}_k^t. \tag{4.2.6}$$

Lemma 4.10. *In the case where $\beta^{\frac{1}{2}} h_k^{-2} < 1$, we have*

$$|||R_k(p,y)|||_{1,k} \leq \tau \, |||(p,y)|||_{1,k} \qquad \forall\, (p,y) \in V_k \times V_k, \tag{4.2.7}$$

$$|||\tilde{R}_k(p,y)|||_{1,k}^{\sim} \leq \tau \, |||(p,y)|||_{1,k}^{\sim} \qquad \forall\, (p,y) \in V_k \times V_k, \tag{4.2.8}$$

where the constant $\tau \in (0,1)$ is independent of k and β.

Proof. In this case λ_k given by (4.1.24) is the optimal damping parameter for the Richardson iteration and we have

$$C_{\min} \leq \lambda_{\min}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) \leq \lambda_{\max}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) < 2C_{\max}$$

by Lemma 4.4. It follows that (cf. Section 2.5.1)

$$\begin{aligned} |||R_k(p,y)|||_{1,k} &= [\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k R_k(p,y), R_k(p,y)]_k^{\frac{1}{2}} \\ &\leq \left(\frac{\lambda_{\max}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) - \lambda_{\min}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)}{\lambda_{\max}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) + \lambda_{\min}(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)} \right) [\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k (p,y), (p,y)]_k^{\frac{1}{2}} \\ &\leq \left(\frac{2C_{\max} - C_{\min}}{2C_{\max} + C_{\min}} \right) |||(p,y)|||_{1,k} \, . \end{aligned}$$

Therefore (4.2.7) holds for $\tau = (2C_{\max} - C_{\min})/(2C_{\max} + C_{\min})$. The derivation of (4.2.8) is identical. □

Lemma 4.11. *In the case where $\beta^{\frac{1}{2}} h_k^{-2} \geq 1$, we have, for $0 \leq s \leq 1$,*

$$|||R_k^m(p,y)|||_{1,k} \leq C(1+\beta^{\frac{1}{2}}h_k^{-2})^{s/2} m^{-s/2} \, |||(p,y)|||_{1-s,k} \quad \forall (p,y) \in V_k \times V_k, \quad (4.2.9)$$

$$|||\tilde{R}_k^m(p,y)|||_{1,k}^{\sim} \leq C(1+\beta^{\frac{1}{2}}h_k^{-2})^{s/2} m^{-s/2} \, |||(p,y)|||_{1-s,k}^{\sim} \quad \forall (p,y) \in V_k \times V_k, \quad (4.2.10)$$

where the positive constant C is independent of k and β.

Proof. In this case λ_k is given by (4.1.25) and $\lambda_{\max}(\lambda_k \mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k) \leq 1$. It follows from (4.1.25), (4.2.1), (4.2.5), calculus and the spectral theorem that

$$\begin{aligned}
|||R_k^m(p,y)|||_{1,k}^2 &= [\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k R_k^m(p,y), R_k^m(p,y)]_k \\
&= \lambda_k^{-s} [(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)^{1-s} (\lambda_k \mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)^s R_k^m(p,y), R_k^m(p,y)]_k \\
&\leq C_\dagger^s (1+\beta^{\frac{1}{2}}h_k^{-2})^s \max_{0 \leq x \leq 1} [(1-x)^{2m} x^s] [(\mathfrak{B}_k^t \mathfrak{C}_k^{-1} \mathfrak{B}_k)^{1-s}(p,y),(p,y)]_k \\
&\leq C(1+\beta^{\frac{1}{2}}h_k^{-2})^s m^{-s} \, |||(p,y)|||_{1-s,k}^2 \, .
\end{aligned}$$

The proof for (4.2.10) is identical. □

Remark 4.12. *In the special case where $s = 0$, the calculation in the proof of Lemma 4.11 shows that*

$$\begin{aligned}
|||R_k(p,y)|||_{1,k} &\leq |||(p,y)|||_{1,k} \qquad \forall\, (p,y) \in V_k \times V_k, \\
|||\tilde{R}_k(p,y)|||_{1,k}^{\sim} &\leq |||(p,y)|||_{1,k}^{\sim} \qquad \forall\, (p,y) \in V_k \times V_k.
\end{aligned}$$

4.2.2 Approximation Properties

We define two Ritz projection operators $P_k^{k-1} : V_k \times V_k \to V_{k-1} \times V_{k-1}$ and $\tilde{P}_k^{k-1} : V_k \times V_k \to V_{k-1} \times V_{k-1}$ in terms of the bilinear form $\mathcal{B}(\cdot,\cdot)$ and the natural injection $I_{k-1}^k : V_{k-1} \times V_{k-1} \to V_k \times V_k$ as follows. For any $(p,y) \in V_k \times V_k$ and $(q,z) \in V_{k-1} \times V_{k-1}$,

$$\mathcal{B}(P_k^{k-1}(p,y),(q,z)) = \mathcal{B}((p,y), I_{k-1}^k(q,z)) = \mathcal{B}((p,y),(q,z)), \quad (4.2.11)$$

$$\mathcal{B}((q,z), \tilde{P}_k^{k-1}(p,y)) = \mathcal{B}(I_{k-1}^k(q,z),(p,y)) = \mathcal{B}((q,z),(p,y)). \quad (4.2.12)$$

It follows that

$$P_k^{k-1} I_{k-1}^k = Id_{k-1} = \tilde{P}_k^{k-1} I_{k-1}^k, \tag{4.2.13}$$

and hence

$$(I_{k-1}^k P_k^{k-1})^2 = I_{k-1}^k P_k^{k-1} \quad \text{and} \quad (Id_k - I_{k-1}^k P_k^{k-1})^2 = Id_k - I_{k-1}^k P_k^{k-1}, \tag{4.2.14}$$

$$(I_{k-1}^k \tilde{P}_k^{k-1})^2 = I_{k-1}^k \tilde{P}_k^{k-1} \quad \text{and} \quad (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})^2 = Id_k - I_{k-1}^k \tilde{P}_k^{k-1}. \tag{4.2.15}$$

Moreover we have the following Galerkin orthogonality relations for all $(p, y) \in V_k \times V_k, (q, z) \in V_{k-1} \times V_{k-1}$:

$$\mathcal{B}((Id_k - I_{k-1}^k P_k^{k-1})(p, y), I_{k-1}^k (q, z)) = 0, \tag{4.2.16}$$

$$\mathcal{B}(I_{k-1}^k (q, z), (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})(p, y)) = 0. \tag{4.2.17}$$

The effects of the operators $Id_k - I_{k-1}^k P_k^{k-1}$ and $Id_k - I_{k-1}^k \tilde{P}_k^{k-1}$ are measured by the following approximation properties.

Lemma 4.13. *For all $(p, y) \in V_k \times V_k$, there exists a positive constant C independent of k and β such that*

$$|||(Id_k - I_{k-1}^k P_k^{k-1})(p, y)|||_{0,k} \leq C(1 + \beta^{\frac{1}{2}} h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 \, |||(p, y)|||_{1,k}, \tag{4.2.18}$$

$$|||(Id_k - I_{k-1}^k \tilde{P}_k^{k-1})(p, y)|||_{0,k}^{\sim} \leq C(1 + \beta^{\frac{1}{2}} h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 \, |||(p, y)|||_{1,k}^{\sim}. \tag{4.2.19}$$

Proof. We only present the proof for (4.2.18). Let $(p, y) \in V_k \times V_k$ be arbitrary and

$$(\zeta, \mu) = (Id_k - I_{k-1}^k P_k^{k-1})(p, y). \tag{4.2.20}$$

In view of (4.2.3), it suffices to establish the estimate

$$\|\zeta\|_{L^2(\Omega)} + \|\mu\|_{L^2(\Omega)} \lesssim (1 + \beta^{\frac{1}{2}} h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 \, |||(p, y)|||_{1,k} \tag{4.2.21}$$

by a duality argument. Let $(\xi, \theta) \in H_0^1(\Omega) \times H_0^1(\Omega)$ be defined by

$$\mathcal{B}((q, z), (\xi, \theta)) = (\zeta, q)_{L^2(\Omega)} + (\mu, z)_{L^2(\Omega)} \qquad \forall\, (q, z) \in H_0^1(\Omega) \times H_0^1(\Omega), \tag{4.2.22}$$

and $(\xi_{k-1}, \theta_{k-1}) \in V_{k-1} \times V_{k-1}$ be defined by

$$\mathcal{B}((q,z),(\xi_{k-1},\theta_{k-1})) = (\zeta,q)_{L^2(\Omega)} + (\mu,z)_{L^2(\Omega)} \qquad \forall\, (q,z) \in V_{k-1} \times V_{k-1}. \tag{4.2.23}$$

Since $h_{k-1} = 2h_k$, we have, according to Theorem 3.5,

$$\|\xi - \xi_{k-1}\|_{H^1_\beta(\Omega)} + \|\theta - \theta_{k-1}\|_{H^1_\beta(\Omega)} \lesssim (1+\beta^{\frac{1}{2}} h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 (\|\zeta\|_{L^2(\Omega)} + \|\mu\|_{L^2(\Omega)}). \tag{4.2.24}$$

Putting (3.1.7), (4.2.4), (4.2.16), (4.2.20) and (4.2.22)–(4.2.24) together, we find

$$\begin{aligned}
\|\zeta\|^2_{L^2(\Omega)} + \|\mu\|^2_{L^2(\Omega)} &= \mathcal{B}((\zeta,\mu),(\xi,\theta)) \\
&= \mathcal{B}((Id_k - I^k_{k-1}P^{k-1}_k)(p,y),(\xi,\theta)) \\
&= \mathcal{B}((Id_k - I^k_{k-1}P^{k-1}_k)(p,y),(\xi,\theta) - (\xi_{k-1},\theta_{k-1})) \\
&= \mathcal{B}((p,y),(\xi,\theta) - (\xi_{k-1},\theta_{k-1})) \\
&\lesssim (\|\xi - \xi_{k-1}\|^2_{H^1_\beta(\Omega)} + \|\theta - \theta_{k-1}\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} (\|p\|^2_{H^1_\beta(\Omega)} + \|y\|^2_{H^1_\beta(\Omega)})^{\frac{1}{2}} \\
&\lesssim (1+\beta^{\frac{1}{2}} h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 (\|\zeta\|_{L^2(\Omega)} + \|\mu\|_{L^2(\Omega)})\, |||(p,y)|||_{1,k}\,,
\end{aligned}$$

which implies (4.2.18). The estimate (4.2.19) is established by similar arguments based on (4.2.17). □

We also need the following stability estimates.

Lemma 4.14. *We have*

$$|||I^k_{k-1}(q,z)|||_{1,k} \approx |||(q,z)|||_{1,k-1} \qquad \forall\, (q,z) \in V_{k-1} \times V_{k-1}, \tag{4.2.25}$$

$$|||P^{k-1}_k(p,y)|||_{1,k-1} \lesssim |||(p,y)|||_{1,k} \qquad \forall\, (p,y) \in V_k \times V_k, \tag{4.2.26}$$

$$|||\tilde{P}^{k-1}_k(p,y)|||^{\sim}_{1,k-1} \lesssim |||(p,y)|||^{\sim}_{1,k} \qquad \forall\, (p,y) \in V_k \times V_k, \tag{4.2.27}$$

where the hidden constants are independent of k and β.

Proof. The estimate (4.2.25) follows from (4.2.4) and the fact that I_{k-1}^k is the natural injection. The estimate (4.2.26) then follows from (3.1.16), (4.2.4), (4.2.11) and (4.2.25) :

$$\begin{aligned}
|||P_k^{k-1}(p,y)|||_{1,k-1} &\approx \sup_{(q,z)\in V_{k-1}\times V_{k-1}} \frac{\mathcal{B}\Big(P_k^{k-1}(p,y),(q,z)\Big)}{|||(q,z)|||_{1,k-1}} \\
&= \sup_{(q,z)\in V_{k-1}\times V_{k-1}} \frac{\mathcal{B}\Big((p,y),I_{k-1}^k(q,z)\Big)}{|||(q,z)|||_{1,k-1}} \lesssim |||(p,y)|||_{1,k}\,.
\end{aligned}$$

We can obtain (4.2.27) similarly by using (3.1.17), (4.2.4), (4.2.12) and (4.2.25). □

4.3 Convergence Analysis of the W-cycle Algorithms

In this section we will first establish the convergence of the two-grid algorithm and then prove the convergence of the W-cycle algorithm as we discussed in Section 2.8.

Let $E_k : V_k \times V_k \longrightarrow V_k \times V_k$ be the error propagation operator for the k-th level W-cycle algorithm. We have the following well-known recursive relation, (cf. [59, 76, 22]):

$$E_k = R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1} + I_{k-1}^k E_{k-1}^q P_k^{k-1}) S_k^{m_1}, \tag{4.3.1}$$

where

$$S_k = Id_k - \lambda_k \mathfrak{C}_k^{-1}\mathfrak{B}_k^t\mathfrak{B}_k \tag{4.3.2}$$

is the error propagation operator for one pre-smoothing step (cf. (4.1.21)).

Note that S_k is the transpose of $\tilde{R}_k$ (the error propagation operator for one post-smoothing step for the dual problem (4.1.8)) with respect to the variational form $\mathcal{B}(\cdot,\cdot)$ by (4.1.6) and (4.2.6):

$$\begin{aligned}
\mathcal{B}(S_k(p,y),(q,z)) &= [\mathfrak{B}_k(Id_k - \lambda_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t\mathfrak{B}_k)(p,y),(q,z)]_k \\
&= [\mathfrak{B}_k(p,y),(Id_k - \lambda_k\mathfrak{B}_k\mathfrak{C}_k^{-1}\mathfrak{B}_k^t)(q,z)]_k \\
&= \mathcal{B}((p,y),\tilde{R}_k(q,z)) \qquad \forall\,(p,y),(q,z)\in V_k\times V_k.
\end{aligned} \tag{4.3.3}$$

Remark 4.15. *The duality between S_k and $\tilde{R}_k$ is the reason why we consider the multigrid algorithms for* (4.1.7) *and* (4.1.8) *simultaneously.*

The relations (4.2.11) and (4.3.3) lead to the following useful result.

Lemma 4.16. *We have*

$$\|(Id_k - I_{k-1}^k P_k^{k-1}) S_k^m\| \approx \|\tilde{R}_k^m (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})\|, \tag{4.3.4}$$

where $\|\cdot\|$ denotes the operator norm with respect to $|||\cdot|||_{1,k}$ and the hidden constants are independent of k and β.

Proof. It follows from (3.1.16), (4.2.4), (4.2.11), (4.2.12) and (4.3.3) that

$$\begin{aligned}
&|||(Id_k - I_{k-1}^k P_k^{k-1}) S_k^m(p,y)|||_{1,k} \\
&\qquad \approx \sup_{(q,z)\in V_k\times V_k} \frac{\mathcal{B}\Big((Id_k - I_{k-1}^k P_k^{k-1}) S_k^m(p,y), (q,z)\Big)}{|||(q,z)|||_{1,k}} \\
&\qquad = \sup_{(q,z)\in V_k\times V_k} \frac{\mathcal{B}\Big((p,y), \tilde{R}_k^m (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})(q,z)\Big)}{|||(q,z)|||_{1,k}} \\
&\qquad \lesssim |||(p,y)|||_{1,k}\, \|\tilde{R}_k^m (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})\| \qquad\qquad \forall\, (p,y)\in V_k\times V_k,
\end{aligned}$$

and hence

$$\|(Id_k - I_{k-1}^k P_k^{k-1}) S_k^m\| \lesssim \|\tilde{R}_k^m (Id_k - I_{k-1}^k \tilde{P}_k^{k-1})\|.$$

The estimate in the other direction is established by a similar argument that uses (3.1.17) instead of (3.1.16). □

4.3.1 Convergence for the Two-grid Algorithm for (4.1.7)

In the two-grid algorithm the coarse grid residual equation is solved exactly. By setting $E_{k-1} = 0$ in (4.3.1), we obtain the error propagation operator $R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1}) S_k^{m_1}$ for the two-grid algorithm with m_1 (resp., m_2) pre-smoothing (resp., post-smoothing) steps.

We will separate the convergence analysis into two cases.

The case where $\beta^{\frac{1}{2}}h_k^{-2} < 1$. Here we can apply Lemma 4.10 which states that R_k (resp., $\tilde{R}_k$) is a contraction with respect to $|||\cdot|||_{1,k}$ (resp., $|||\cdot|||_{1,k}^{\sim}$) and the contraction number τ is independent of k and β.

Lemma 4.17. *In the case where* $\beta^{\frac{1}{2}}h_k^{-2} < 1$, *there exists a positive constant* $C_\sharp$ *independent of* k *and* β *such that*

$$\|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \leq C_\sharp \tau^{m_1+m_2}, \tag{4.3.5}$$

where $\|\cdot\|$ *is the operator norm with respect to* $|||\cdot|||_{1,k}$.

Proof. We have, from Lemma 4.10 and Lemma 4.14,

$$\begin{aligned}
&|||R_k^m(Id_k - I_{k-1}^k P_k^{k-1})(p,y)|||_{1,k} \\
&\qquad \leq \tau^m |||(Id_k - I_{k-1}^k P_k^{k-1})(p,y)|||_{1,k} \lesssim \tau^m \, |||(p,y)|||_{1,k} \quad \forall\, (p,y) \in V_k \times V_k,
\end{aligned}$$

and hence

$$\|R_k^m(Id_k - I_{k-1}^k P_k^{k-1})\| \lesssim \tau^m. \tag{4.3.6}$$

Similarly, we also have, by (4.2.4),(4.2.8) and Lemma 4.14,

$$\|\tilde{R}_k^m(Id_k - I_{k-1}^k \tilde{P}_k^{k-1})\| \lesssim \tau^m, \tag{4.3.7}$$

which together with Lemma 4.16 implies

$$\|(Id_k - I_{k-1}^k P_k^{k-1})S_k^m\| \lesssim \tau^m. \tag{4.3.8}$$

Finally we establish (4.3.5) by combining (4.2.14), (4.3.6) and (4.3.8):

$$\begin{aligned}
&\|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \\
&\qquad = \|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \\
&\qquad \leq \|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})\|\|(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \lesssim \tau^{m_1+m_2}.
\end{aligned}$$

□

The case where $\beta^{\frac{1}{2}}h_k^{-2} \geq 1$. Here we can apply Lemma 4.11.

Lemma 4.18. *In the case where* $\beta^{\frac{1}{2}}h_k^{-2} \geq 1$*, there exists a positive constant* $C_\flat$ *independent of* k *and* β *such that*

$$\|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \leq C_\flat[\max(1, m_1)\max(1, m_2)]^{\frac{1}{2}}, \tag{4.3.9}$$

where $\|\cdot\|$ *is the operator norm with respect to* $|||\cdot|||_{1,k}$.

Proof. Let m be any positive integer. We have, from (4.2.9) and (4.2.18),

$$\begin{aligned}
&|||R_k^m(Id_k - I_{k-1}^k P_k^{k-1})(p, y)|||_{1,k} \\
&\quad\lesssim (1 + \beta^{\frac{1}{2}}h_k^{-2})^{\frac{1}{2}} m^{-\frac{1}{2}} |||(Id_k - I_{k-1}^k P_k^{k-1})(p, y)|||_{1,k} \\
&\quad\lesssim (1 + \beta^{\frac{1}{2}}h_k^{-2})^{\frac{1}{2}} m^{-\frac{1}{2}} (1 + \beta^{\frac{1}{2}}h_k^{-2})^{\frac{1}{2}} \beta^{-\frac{1}{2}} h_k^2 \, |||(p, y)|||_{1,k} \\
&\quad= m^{-\frac{1}{2}} (\beta^{-\frac{1}{2}} h_k^2 + 1) \, |||(p, y)|||_{1,k} \\
&\quad\leq 2m^{-\frac{1}{2}} \, |||(p, y)|||_{1,k} \qquad\qquad \forall\, (p, y) \in V_k \times V_k,
\end{aligned}$$

and hence

$$\|R_k^m(Id_k - I_{k-1}^k P_k^{k-1})\| \lesssim m^{-\frac{1}{2}}. \tag{4.3.10}$$

Similarly, we have, by (4.2.4), (4.2.10) and (4.2.19),

$$\|\tilde{R}_k^m(Id_k - I_{k-1}^k \tilde{P}_k^{k-1})\| \lesssim m^{-\frac{1}{2}}, \tag{4.3.11}$$

which together with Lemma 4.16 implies

$$\|(Id_k - I_{k-1}^k P_k^{k-1})S_k^m\| \lesssim m^{-\frac{1}{2}}. \tag{4.3.12}$$

Combining (4.2.14), (4.3.10) and (4.3.12), we obtain for $m_1, m_2 \geq 1$,

$$\begin{aligned}
\|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| &= \|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \\
&\leq |R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1})\|\|(Id_k - I_{k-1}^k P_k^{k-1})S_k^{m_1}\| \\
&\lesssim (m_1 m_2)^{-\frac{1}{2}}.
\end{aligned}$$

The cases where $m_1 = 0$ or $m_2 = 0$ follow from (4.2.25) and (4.2.26). □

4.3.2 Convergence for the W-cycle Algorithm for (4.1.7)

We will derive error estimates for the W-cycle algorithm through (4.3.1) and the results for the two-grid algorithm in Section 4.3.1. We assume $m_1, m_2 \geq 1$.

According to Remark 4.12, there exists a positive constant C_1 independent of k and m such that

$$\|R_k^m\|, \|\tilde{R}_k^m\| \leq C_1 \qquad \text{for } m \geq 1, \tag{4.3.13}$$

where $\|\cdot\|$ is the operator norm with respect to $|||\cdot|||_{1,k}$. Moreover it follows from (3.1.17), (4.2.4) and (4.3.3) that for all $(p, y) \in V_k \times V_k$,

$$\begin{aligned} |||S_k^m(p,y)|||_{1,k} &\approx \sup_{(q,z)\in V_k\times V_k} \frac{\mathcal{B}\big(S_k^m(p,y),(q,z)\big)}{|||(q,z)|||_{1,k}} \\ &= \sup_{(q,z)\in V_k\times V_k} \frac{\mathcal{B}\big((p,y),\tilde{R}_k^m(q,z)\big)}{|||(q,z)|||_{1,k}} \lesssim |||(p,y)|||_{1,k}\,\|\tilde{R}_k^m\|, \end{aligned}$$

and hence, by (4.3.13),

$$\|S_k^m\| \leq C_2 \qquad \text{for } m \geq 1, \tag{4.3.14}$$

where the positive constant C_2 is independent of k and m.

Putting Lemma 4.14, (4.3.1), (4.3.13) and (4.3.14) together, we obtain the recursive estimate

$$\|E_k\| \leq \|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1}) S_k^{m_1}\| + C_*\|E_{k-1}\|^2 \qquad \text{for } k \geq 1, \tag{4.3.15}$$

where the positive constant C_* is independent of k and β. The behavior of $\|E_k\|$ is therefore determined by (4.3.15), the behavior of $\|R_k^{m_2}(Id_k - I_{k-1}^k P_k^{k-1}) S_k^{m_1}\|$, and the initial condition

$$\|E_0\| = 0. \tag{4.3.16}$$

Specifically, for $\beta^{\frac{1}{2}} h_k^{-2} < 1$, we have

$$\|E_k\| \leq C_\sharp \tau^{m_1+m_2} + C_*\|E_{k-1}\|^2 \tag{4.3.17}$$

by Lemma 4.17, and for $\beta^{\frac{1}{2}} h_k^{-2} \geq 1$, we have

$$\|E_k\| \leq C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}} + C_* \|E_{k-1}\|^2 \tag{4.3.18}$$

by Lemma 4.18.

The following result is useful for the analysis of (4.3.16)–(4.3.18).

Lemma 4.19. *Let α_k $(k = 0, 1, 2, \ldots)$ be a sequence of nonnegative numbers such that*

$$\alpha_k \leq 1 + \delta \alpha_{k-1}^2 \qquad \textit{for} \quad k \geq 1, \tag{4.3.19}$$

where the positive constant δ satisfies

$$\delta \leq \frac{1}{4(1+\alpha_0)}. \tag{4.3.20}$$

Then it holds that

$$\alpha_k \leq 2 + 4^{1-2^k} \alpha_0 \qquad \textit{for} \quad k \geq 0. \tag{4.3.21}$$

Proof. The bound (4.3.21) holds trivially for $k = 0$. Suppose it holds for $k \geq 0$. We have, by (4.3.19) and (4.3.20),

$$\begin{aligned}
\alpha_{k+1} &\leq 1 + \delta \alpha_k^2 \\
&\leq 1 + \delta (2 + 4^{1-2^k} \alpha_0)^2 \\
&= 1 + \delta (4 + 4^{1-2^k} 4\alpha_0) + (\delta \alpha_0) 4^{2-2^{k+1}} \alpha_0 \\
&\leq 1 + \delta (4 + 4\alpha_0) + \left(\frac{1}{4}\right) 4^{2-2^{k+1}} \alpha_0 \leq 2 + 4^{1-2^{k+1}} \alpha_0.
\end{aligned}$$

Therefore the bound (4.3.21) holds for $k \geq 0$ by mathematical induction. □

Theorem 4.20. *Let k_* be the largest positive integer such that $\beta^{\frac{1}{2}} h_{k_*}^{-2} < 1$. There exists positive integers m_1^* and m_2^* independent of k and β such that $m_1 \geq m_1^*$, $m_2 \geq m_2^*$ imply*

$$\|E_k\| \leq 2 C_\sharp \tau^{m_1+m_2} \qquad \forall\, 1 \leq k \leq k_*, \tag{4.3.22}$$

$$\|E_k\| \leq 2 C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}} + 4^{1-2^{k-k_*}} (2 C_\sharp \tau^{m_1+m_2}) \qquad \forall\, k \geq k_* + 1, \tag{4.3.23}$$

where $\|\cdot\|$ *is the operator norm with respect to* $|||\cdot|||_{1,k}$.

Proof. For $1 \leq k \leq k_*$, we take

$$\alpha_k = \|E_k\|/(C_\sharp \tau^{m_1+m_2})$$

and observe that

$$\alpha_k \leq 1 + (C_* C_\sharp \tau^{m_1+m_2})\alpha_{k-1}^2$$

by (4.3.17). It then follows from (4.3.16) and Lemma 4.19 that $\alpha_k \leq 2$, or equivalently

$$\|E_k\| \leq 2C_\sharp \tau^{m_1+m_2},$$

provided that

$$C_* C_\sharp \tau^{m_1+m_2} \leq \frac{1}{4}. \tag{4.3.24}$$

We now define

$$\mu_k = \|E_{k_*+k}\|/(C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}})$$

and observe that

$$\mu_k \leq 1 + (C_* C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}})\mu_{k-1}^2 \qquad \text{for} \quad k \geq 1$$

by (4.3.18). It then follows from Lemma 4.19 that

$$\mu_k \leq 2 + 4^{1-2^k}\mu_0 \qquad \text{for} \quad k \geq 1,$$

or equivalently

$$\|E_k\| \leq 2C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}} + 4^{1-2^{k-k_*}}\|E_{k_*}\| \qquad \text{for} \quad k \geq k_* + 1,$$

provided that

$$C_* C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}} \leq \frac{1}{4(1 + \|E_{k_*}\|/(C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}}))},$$

or equivalently

$$C_* C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}} + C_* \|E_{k_*}\| \leq \frac{1}{4}. \tag{4.3.25}$$

Finally we observe that if we choose m_1^* and m_2^* so that

$$C_* C_\flat (m_1^*)^{-\frac{1}{2}} (m_2^*)^{-\frac{1}{2}} + 2C_* C_\sharp \tau^{m_1^*+m_2^*} \leq \frac{1}{4},$$

then (4.3.24) and (4.3.25) are satisfied for $m_1 \geq m_1^*$, $m_2 \geq m_2^*$. □

Remark 4.21. *According to Theorem 4.20, the k-th level W-cycle algorithm is a contraction if the number of pre-smoothing and post-smoothing steps is sufficiently large and the contraction number is bounded away from 1 uniformly in k and β. Moreover, for the coarser levels where $\beta^{\frac{1}{2}} h_k^{-2} < 1$, the contraction number of the W-cycle algorithm will decrease exponentially with respect to the number of smoothing steps. After a few transition levels the dominant term on the right-hand side of* (4.3.23) *becomes $2C_\flat {m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}}$ and the contraction number will decrease at the rate of ${m_1}^{-\frac{1}{2}} {m_2}^{-\frac{1}{2}}$ for the finer levels where $\beta^{\frac{1}{2}} h_k^{-2} \geq 1$.*

4.3.3 Convergence for the W-cycle Algorithm for (4.1.8)

The error propagation operator $E_k : V_k \times V_k \longrightarrow V_k \times V_k$ for the W-cycle algorithm for (4.1.8) satisfies the following analog of (4.3.1):

$$\tilde{E}_k = \tilde{R}_k^{m_2} (Id_k - I_{k-1}^k \tilde{P}_k^{k-1} + I_{k-1}^k \tilde{E}_{k-1}^2 \tilde{P}_k^{k-1}) \tilde{S}_k^{m_1}, \tag{4.3.26}$$

where $\tilde{R}_k$ is giving by (4.2.6) and $\tilde{S}_k = Id_k - \lambda_k \mathfrak{C}_k^{-1} \mathfrak{B}_k \mathfrak{B}_k^t$ is the error propagation operator for one pre-smoothing step (cf. (4.1.26)), and we have the relations for $(p, y), (q, z) \in V_k \times V_k$,

$$\mathcal{B}((p, y), \tilde{S}_k(q, z)) = \mathcal{B}(R_k(p, y), (q, z)),$$

$$\|(Id_k - I_{k-1}^k \tilde{P}_k^{k-1}) \tilde{S}_k^m\| \approx \|R_k^m (Id_k - I_{k-1}^k P_k^{k-1})\|,$$

that are analogs of (4.3.3) and (4.3.4). The results for E_k in Section 4.1.3 holds for $\tilde{E}_k$ by essentially identical arguments based on Lemmas 4.10, 4.11, (4.2.15), Lemmas 4.13 and 4.14.

4.4 Numerical Results

In this section we report numerical results of symmetric W-cycle and V-cycle algorithms for (4.1.7) on two and three dimensional domains, where the preconditioner $\mathfrak{C}_k^{-1}$ is based on a V-cycle multigrid solve for (4.1.14). We employed the MATLAB/C++ toolbox FELICITY [101] in our computations.

The contraction numbers in energy norm that are presented in this section are computed using the following algorithm.

Algorithm 4.1 Contraction Numbers at Level k with m Smoothing Steps

Initialization: choose a random vector $\mathbf{q}$, calculate $\mathbf{b} = \mathfrak{B}_k\mathbf{q}$. $\mathbf{p} = \mathbf{p}_0, \mathbf{r}_0 = \mathbf{q}$.
repeat
  $\mathbf{p} \leftarrow MG_W(k, \mathbf{b}, \mathbf{p}, m, m)$ or $MG_V(k, \mathbf{b}, \mathbf{p}, m, m)$
  $\mathbf{r} \leftarrow \mathbf{q} - \mathbf{p}$
  $c \leftarrow \frac{\|\mathbf{r}\|_{H^1_\beta(\Omega)}}{\|\mathbf{r}_0\|_{H^1_\beta(\Omega)}}$
  $\mathbf{r}_0 \leftarrow \mathbf{r}$
until c converges

Example 4.22 (Unit Square)**.** *The domain Ω for this example is the unit square $(0,1)^2$. We take $\boldsymbol{\zeta} = \frac{1}{2}[1\ 0]^t$ and $\gamma = 0$ in* (1.1.4)*, and $C_\dagger = 5$ in* (4.1.25)*. Here $\mathfrak{C}_k^{-1}$ is based on a $V(4,4)$ multigrid solve for* (4.1.14)*. See Figure 3.1 for the meshes at the first three levels.*

The contraction numbers of the kth level symmetric W-cycle algorithm in the energy norm with $\beta = 10^{-2}$ (resp., $\beta = 10^{-4}$ and $\beta = 10^{-6}$) are presented in Table 4.1 (resp., Tables 4.2 and 4.3), while the number m of pre-smoothing and post-smoothing steps increases from 2^0 to 2^8.

Table 4.1. Contraction Numbers of W-cycle Algorithm for Example 4.22 ($\beta = 10^{-2}$).

m \ k	1	2	3	4	5	6	7	8
2^0	2.94e-01	6.01e-01	4.46e-01	3.81e-01	3.71e-01	3.68e-01	3.66e-01	3.65e-01
2^1	8.84e-02	3.88e-01	2.43e-01	2.15e-01	2.13e-01	2.11e-01	2.09e-01	2.09e-01
2^2	7.81e-03	1.85e-01	1.01e-01	8.66e-02	8.20e-02	8.12e-02	8.08e-02	8.03e-02
2^3	6.11e-05	4.89e-02	3.54e-02	3.73e-02	3.96e-02	3.95e-02	3.94e-02	3.92e-02
2^4	6.43e-08	1.82e-02	1.57e-02	2.04e-02	1.97e-02	2.00e-02	2.00e-02	2.00e-02
2^5	9.06e-17	2.70e-03	7.41e-03	7.47e-03	9.55e-03	1.02e-02	1.02e-02	1.02e-02
2^6	1.62e-16	5.95e-05	1.90e-03	3.40e-03	5.03e-03	4.94e-03	5.00e-03	5.03e-03
2^7	7.05e-17	3.70e-08	1.24e-04	1.57e-03	1.76e-03	2.38e-03	2.52e-03	2.54e-03
2^8	3.10e-17	3.06e-15	5.34e-07	3.45e-04	8.28e-04	1.23e-03	1.21e-03	1.23e-03

Table 4.2. Contraction Numbers of W-cycle Algorithm for Example 4.22 ($\beta = 10^{-4}$).

m \ k	1	2	3	4	5	6	7	8
2^0	1.21e-01	2.32e-01	4.88e-01	5.45e-01	4.13e-01	3.77e-01	3.68e-01	3.66e-01
2^1	1.47e-02	7.46e-02	2.58e-01	3.21e-01	2.42e-01	2.18e-01	2.12e-01	2.10e-01
2^2	2.17e-04	5.85e-03	7.11e-02	1.68e-01	1.02e-01	8.56e-02	8.17e-02	8.07e-02
2^3	5.06e-08	3.58e-05	6.99e-03	6.00e-02	4.75e-02	4.16e-02	4.00e-02	3.93e-02
2^4	1.00e-15	3.88e-07	2.88e-04	2.38e-02	2.42e-02	2.16e-02	2.05e-02	2.01e-02
2^5	1.37e-17	1.44e-16	5.53e-07	7.07e-03	1.19e-02	1.13e-02	1.06e-02	1.03e-02
2^6	1.45e-16	1.03e-16	9.37e-13	1.25e-03	3.70e-03	5.61e-03	5.30e-03	5.12e-03
2^7	1.13e-16	1.05e-16	2.43e-16	3.93e-05	1.08e-03	2.72e-03	2.75e-03	2.61e-03
2^8	3.14e-17	1.24e-16	2.18e-16	8.94e-08	1.20e-04	7.89e-04	1.34e-03	1.29e-03

Table 4.3. Contraction Numbers of W-cycle Algorithm for Example 4.22 ($\beta = 10^{-6}$).

m \ k	1	2	3	4	5	6	7	8
2^0	2.56e-01	3.89e-01	2.36e-01	3.77e-01	7.02e-01	4.90e-01	3.95e-01	3.72e-01
2^1	6.79e-02	1.71e-01	5.63e-02	1.38e-01	4.92e-01	2.90e-01	2.30e-01	2.14e-01
2^2	4.61e-03	2.96e-02	3.08e-03	2.29e-02	2.82e-01	1.38e-01	9.38e-02	8.35e-02
2^3	2.13e-05	8.88e-04	8.91e-06	1.04e-03	1.32e-01	6.00e-02	4.49e-02	4.06e-02
2^4	6.45e-11	9.79e-07	2.77e-11	2.43e-06	3.33e-02	2.84e-02	2.37e-02	2.11e-02
2^5	1.25e-16	8.18e-14	1.34e-16	3.26e-12	6.03e-03	1.06e-02	1.22e-02	1.11e-02
2^6	6.45e-17	2.29e-16	1.40e-16	1.63e-16	3.44e-04	2.54e-03	5.70e-03	5.69e-03
2^7	1.65e-16	1.29e-16	1.37e-16	1.59e-16	1.47e-06	2.29e-04	1.98e-03	2.91e-03
2^8	7.25e-17	1.31e-16	1.34e-16	1.52e-16	6.78e-12	2.58e-06	3.90e-04	1.31e-03

Table 4.4. The Times for One Iteration of the Symmetric W-cycle Algorithm with m Smoothing Steps at Level 7 (Unit Square).

m	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7
Times (s)	3.0e-01	5.4e-01	1.0e+00	2.0e+00	4.0e+00	7.9e+00	1.6e+01	3.1e+01

We observe that the symmetric W-cycle algorithm is a contraction with $m = 1$ for all three choices of β, and the behavior of the contraction numbers as k and m vary agree with Remark 4.21. The robustness with respect to β and k is also clearly observed.

The times for one iteration of the symmetric W-cycle algorithm at level 7 (where there are roughly 6×10^4 dofs) are reported in Table 4.4. They are proportional to the number of smoothing steps, which confirms that this is an $O(n)$ algorithm.

We have also computed the contraction numbers for the kth level symmetric V-cycle algorithm, which are similar to those of the W-cycle algorithm. We present the results for $k = 1, \cdots, 8$, $\beta = 10^{-2}, 10^{-4}, 10^{-6}$ and $m = 2^0, \cdots, 2^8$ in Tables 4.5, 4.6 and 4.7. Again we observe that the V-cycle algorithm is a contraction for $m = 1$ and the contraction numbers are robust with respect to both β and k.

The times for one iteration of the symmetric V-cycle algorithm at level 7 are reported in Table 4.8. They are proportional to the number of smoothing steps, which again confirms that this is an $O(n)$ algorithm.

Table 4.5. Contraction Numbers of V-cycle Algorithm for Example 4.22 ($\beta = 10^{-2}$).

m \ k	1	2	3	4	5	6	7	8
2^0	2.94e-01	6.01e-01	5.58e-01	5.38e-01	5.33e-01	5.28e-01	5.12e-01	5.03e-01
2^1	8.84e-02	3.87e-01	3.44e-01	3.31e-01	3.01e-01	2.93e-01	2.76e-01	2.71e-01
2^2	7.81e-03	1.86e-01	1.67e-01	1.55e-01	1.33e-01	1.31e-01	1.29e-01	1.26e-01
2^3	6.12e-05	5.03e-02	5.63e-02	5.59e-02	5.85e-02	5.96e-02	5.93e-02	5.83e-02
2^4	6.82e-08	1.82e-02	2.34e-02	2.48e-02	2.70e-02	2.83e-02	2.90e-02	2.92e-02
2^5	9.50e-18	2.70e-03	8.08e-03	1.03e-02	1.24e-02	1.38e-02	1.44e-02	1.47e-02
2^6	1.43e-16	5.96e-05	1.91e-03	4.24e-03	5.25e-03	6.59e-03	7.10e-03	7.21e-03
2^7	1.45e-17	3.45e-08	1.24e-04	1.60e-03	2.24e-03	3.00e-03	3.40e-03	3.58e-03
2^8	1.02e-16	2.83e-15	5.34e-07	3.45e-04	9.82e-04	1.29e-03	1.62e-03	1.75e-03

Table 4.6. Contraction Numbers of V-cycle Algorithm for Example 4.22 ($\beta = 10^{-4}$).

m \ k	1	2	3	4	5	6	7	8
2^0	1.21e-01	2.31e-01	4.88e-01	5.46e-01	4.94e-01	4.86e-01	4.85e-01	4.84e-01
2^1	1.47e-02	7.59e-02	2.55e-01	3.20e-01	3.18e-01	3.17e-01	3.16e-01	3.16e-01
2^2	2.17e-04	5.73e-03	7.18e-02	1.68e-01	1.73e-01	1.75e-01	1.75e-01	1.75e-01
2^3	4.76e-08	3.56e-05	6.90e-03	5.98e-02	7.23e-02	7.71e-02	7.80e-02	7.76e-02
2^4	1.12e-15	1.57e-07	2.88e-04	2.37e-02	3.38e-02	3.61e-02	3.61e-02	3.53e-02
2^5	4.11e-17	1.82e-16	6.45e-07	7.09e-03	1.28e-02	1.58e-02	1.67e-02	1.66e-02
2^6	7.54e-17	7.46e-17	1.50e-12	1.25e-03	3.77e-03	6.91e-03	7.94e-03	8.18e-03
2^7	6.34e-17	1.57e-16	2.31e-16	3.93e-05	1.09e-03	2.76e-03	3.64e-03	3.93e-03
2^8	4.93e-17	1.59e-16	2.30e-16	9.08e-08	1.20e-04	8.06e-04	1.62e-03	1.91e-03

Table 4.7. Contraction Numbers of V-cycle Algorithm for Example 4.22 ($\beta = 10^{-6}$).

m \ k	1	2	3	4	5	6	7	8
2^0	2.56e-01	3.91e-01	2.36e-01	3.71e-01	7.03e-01	6.31e-01	6.03e-01	5.86e-01
2^1	6.79e-02	1.68e-01	5.61e-02	1.42e-01	4.93e-01	4.12e-01	4.03e-01	4.01e-01
2^2	4.61e-03	3.09e-02	3.13e-03	2.35e-02	2.82e-01	2.54e-01	2.48e-01	2.47e-01
2^3	2.13e-05	8.86e-04	8.53e-06	1.06e-03	1.31e-01	1.17e-01	1.15e-01	1.15e-01
2^4	5.87e-11	7.40e-07	2.78e-11	2.47e-06	3.34e-02	3.80e-02	4.20e-02	4.28e-02
2^5	1.41e-16	1.26e-13	1.36e-16	3.57e-12	6.00e-03	1.15e-02	1.59e-02	1.75e-02
2^6	5.45e-17	1.72e-16	1.71e-16	1.58e-16	3.51e-04	2.55e-03	6.11e-03	8.01e-03
2^7	8.87e-17	1.58e-16	1.16e-16	1.69e-16	1.45e-06	2.29e-04	2.01e-03	3.44e-03
2^8	1.28e-16	1.34e-16	1.51e-16	1.62e-16	6.55e-12	2.65e-06	3.91e-04	1.36e-03

Table 4.8. The Times for One Iteration of the Symmetric V-cycle Algorithm with m Smoothing Steps at Level 7 (Unit Square).

m	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7
Times (s)	7.1e-02	1.3e-01	2.5e-01	4.9e-01	9.4e-01	1.9e+00	3.7e+00	7.4e+00

Example 4.23 (Unit Cube)**.** *The domain for this example is the unit cube* $(0,1)^3$*. We take* $\boldsymbol{\zeta} = \frac{1}{2}[1\ 1\ 1]^t$ *and* $\gamma = 0$ *in* (1.1.4)*, and* $C_\dagger = 4$ *in* (4.1.25)*. The number of grid points in all directions are doubled in each refinement and the triangulations inside the cubic subdomains at all levels are similar to one another. The triangulations* $\mathcal{T}_0$ *and* $\mathcal{T}_1$ *are depicted in Figure 4.1. Here* $\mathfrak{C}_k^{-1}$ *is based on a* $V(4,4)$ *multigrid solve for* (4.1.14)*.*

The contraction numbers of the kth level symmetric W-cycle algorithm in the energy norm with $\beta = 10^{-2}$ (resp., $\beta = 10^{-4}$ and $\beta = 10^{-6}$) are presented in Table 4.9 (resp., Tables 4.10 and 4.11), while the number m of pre-smoothing and post-smoothing steps increases from 2^0 to 2^8.

Table 4.9. Contraction Numbers of W-cycle Algorithm for Example 4.23 ($\beta = 10^{-2}$).

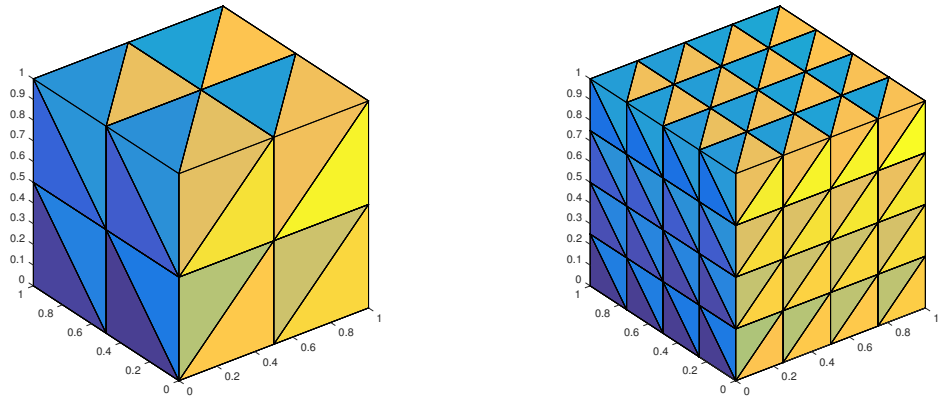

Figure 4.1. Triangulations $\mathcal{T}_0$ and $\mathcal{T}_1$ for the Unit Cube.

m \ k	1	2	3	4	5	6
2^0	4.74e-01	6.67e-01	5.99e-01	5.59e-01	5.56e-01	5.57e-01
2^1	2.24e-01	4.74e-01	4.23e-01	3.96e-01	3.91e-01	3.90e-01
2^2	5.18e-02	2.77e-01	2.51e-01	2.53e-01	2.47e-01	2.50e-01
2^3	4.98e-03	1.32e-01	1.46e-01	1.44e-01	1.44e-01	1.46e-01
2^4	4.62e-05	5.87e-02	7.09e-02	7.49e-02	7.71e-02	7.87e-02
2^5	7.20e-08	1.88e-02	2.79e-02	3.44e-02	3.73e-02	4.04e-02
2^6	2.04e-16	2.02e-03	1.08e-02	1.54e-02	1.41e-02	1.67e-02
2^7	1.76e-16	2.38e-05	2.51e-03	6.47e-03	5.92e-03	5.32e-03
2^8	1.60e-16	3.37e-07	1.42e-04	2.40e-03	2.80e-03	2.71e-03

Table 4.10. Contraction Numbers of W-cycle Algorithm for Example 4.23 ($\beta = 10^{-4}$).

m \ k	1	2	3	4	5	6
2^0	2.32e-01	4.77e-01	4.88e-01	6.51e-01	5.87e-01	5.63e-01
2^1	5.41e-02	2.57e-01	3.32e-01	4.80e-01	4.17e-01	3.97e-01
2^2	2.92e-03	9.33e-02	1.87e-01	3.11e-01	2.69e-01	2.55e-01
2^3	8.25e-06	1.47e-02	8.34e-02	1.85e-01	1.60e-01	1.49e-01
2^4	5.80e-12	4.70e-04	2.32e-02	9.63e-02	9.10e-02	8.18e-02
2^5	1.26e-16	4.91e-07	2.96e-03	4.31e-02	4.17e-02	4.26e-02
2^6	1.52e-16	2.46e-13	4.69e-05	1.43e-02	1.63e-02	1.84e-02
2^7	1.29e-16	1.39e-16	7.60e-08	3.03e-03	6.43e-03	6.10e-03
2^8	1.07e-16	1.43e-16	2.90e-15	1.91e-04	1.94e-03	3.04e-03

Table 4.11. Contraction Numbers of W-cycle Algorithm for Example 4.23 ($\beta = 10^{-6}$).

m \ k	1	2	3	4	5	6	7
2^0	2.88e-01	2.66e-01	4.36e-01	5.24e-01	7.64e-01	6.69e-01	5.79e-01
2^1	8.49e-02	7.23e-02	1.93e-01	3.39e-01	5.95e-01	4.86e-01	4.08e-01
2^2	7.37e-03	5.19e-03	4.36e-02	1.71e-01	4.28e-01	3.15e-01	2.63e-01
2^3	5.26e-05	2.27e-05	2.04e-03	5.08e-02	2.71e-01	1.94e-01	1.57e-01
2^4	4.50e-07	2.09e-11	5.25e-06	5.56e-03	1.43e-01	1.10e-01	8.91e-02
2^5	1.57e-16	1.92e-16	1.39e-12	8.55e-05	5.94e-02	5.46e-02	4.73e-02
2^6	1.80e-16	1.88e-16	1.62e-16	2.64e-07	1.44e-02	2.20e-02	2.38e-02
2^7	1.54e-16	1.82e-16	1.59e-16	2.47e-16	8.27e-04	4.84e-03	7.83e-03
2^8	1.26e-16	1.85e-16	1.60e-16	2.22e-16	6.14e-06	5.12e-04	2.92e-03

Table 4.12. The Times for One Iteration of the Symmetric W-cycle Algorithm with m Smoothing Steps at Level 5 (Unit Cube).

m	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7
Times (s)	8.3e-01	1.5e+00	2.8e+00	5.4e+00	1.1e+01	2.1e+01	4.2e+01	8.4e+01

We observe that the symmetric W-cycle algorithm is a contraction for $m = 1$. The behavior of the contraction numbers agrees with Remark 4.21, and the contraction numbers are robust with respect to both β and k. The times for one iteration of the W -cycle algorithm at level 5 (where there are roughly 5×10^5 dofs) are reported in Table 4.12. They are proportional to m, which confirms the $O(n)$ complexity of the algorithm.

The performance of the symmetric V-cycle algorithm is similar and we present the numerical results for $m = 2^0, \cdots, 2^8$ in Tables 4.13, 4.14 and 4.16. Again the symmetric V-cycle algorithm is a contraction for $m = 1$ and contraction numbers are robust with respect to both β and k. The times for one iteration of the V-cycle algorithm at level 5 are reported in Table 4.15.

Table 4.13. Contraction Numbers of V-cycle Algorithm for Example 4.23 ($\beta = 10^{-2}$).

m \ k	1	2	3	4	5	6
2^0	4.74e-01	6.71e-01	7.03e-01	7.11e-01	7.13e-01	7.13e-01
2^1	2.25e-01	4.76e-01	5.23e-01	5.39e-01	5.42e-01	5.43e-01
2^2	5.15e-02	2.76e-01	3.36e-01	3.58e-01	3.65e-01	3.64e-01
2^3	4.98e-03	1.33e-01	1.83e-01	2.04e-01	2.12e-01	2.10e-01
2^4	4.62e-05	5.88e-02	8.40e-02	1.03e-01	1.09e-01	1.08e-01
2^5	7.89e-08	1.88e-02	3.32e-02	4.60e-02	5.22e-02	5.35e-02
2^6	1.23e-16	2.02e-03	1.12e-02	1.91e-02	2.34e-02	2.21e-02
2^7	1.66e-16	2.41e-05	2.51e-03	7.14e-03	1.01e-02	1.10e-02
2^8	1.65e-16	3.93e-07	1.42e-04	2.43e-03	4.15e-03	4.49e-03

Table 4.14. Contraction Numbers of V-cycle Algorithm for Example 4.23 ($\beta = 10^{-4}$).

m \ k	1	2	3	4	5	6
2^0	2.32e-01	4.85e-01	5.54e-01	6.52e-01	7.00e-01	7.02e-01
2^1	5.41e-02	2.59e-01	3.61e-01	4.77e-01	5.37e-01	5.48e-01
2^2	2.92e-03	9.33e-02	2.00e-01	3.23e-01	3.72e-01	3.80e-01
2^3	8.16e-06	1.48e-02	8.37e-02	1.92e-01	2.28e-01	2.31e-01
2^4	1.03e-11	4.70e-04	2.33e-02	1.01e-01	1.16e-01	1.19e-01
2^5	1.36e-16	4.11e-07	2.95e-03	3.95e-02	5.48e-02	5.13e-02
2^6	9.35e-17	2.16e-13	4.39e-05	1.38e-02	2.08e-02	2.19e-02
2^7	1.07e-16	1.45e-16	7.65e-08	3.04e-03	6.22e-03	7.87e-03
2^8	1.03e-16	1.44e-16	3.18e-15	1.91e-04	2.01e-03	3.46e-03

Table 4.15. The Times for One Iteration of the Symmetric V-cycle Algorithm with m Smoothing Steps at Level 5 (Unit Cube).

m	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7
Times (s)	7.2e-01	1.3e+00	2.4e+00	4.6e+00	9.1e+00	1.8e+01	3.6e+01	7.2e+01

Table 4.16. Contraction Numbers of V-cycle Algorithm for Example 4.23 ($\beta = 10^{-6}$).

m \ k	1	2	3	4	5	6
2^0	2.91e-01	2.65e-01	4.35e-01	5.38e-01	7.64e-01	7.72e-01
2^1	8.51e-02	7.09e-02	1.97e-01	3.49e-01	5.97e-01	6.27e-01
2^2	4.29e-03	5.23e-03	4.37e-02	1.74e-01	4.31e-01	4.61e-01
2^3	4.98e-05	2.27e-05	2.25e-03	5.50e-02	2.69e-01	3.01e-01
2^4	2.70e-07	4.41e-11	5.39e-06	5.73e-03	1.46e-01	1.65e-01
2^5	1.36e-16	1.88e-16	1.37e-12	8.40e-05	5.88e-02	7.19e-02
2^6	1.43e-16	1.76e-16	1.65e-16	2.94e-07	1.43e-02	2.29e-02
2^7	1.41e-16	1.80e-16	1.60e-16	2.49e-16	8.63e-04	4.96e-03
2^8	1.78e-16	1.70e-16	1.61e-16	2.23e-16	6.15e-06	5.05e-04

Example 4.24 (L-shaped domain)**.** *In this example we consider the L-shaped domain* $(0,1)^2 \setminus (0.5,1)^2$. *We take* $\boldsymbol{\zeta} = \frac{1}{2}[1\ 0]^t$ *and* $\gamma = 0$ *in* (1.1.4)*, and* $C_\dagger = 5$ *in* (4.1.25). *See Figure 3.2 for the initial mesh* $\mathcal{T}_0$. *Here* $\mathfrak{C}_k^{-1}$ *is based on a* $V(1,1)$ *multigrid solve for* (4.1.14).

The contraction numbers for the symmetric V-cycle (resp., W-cycle) algorithm in the energy norm with 1 pre-smoothing step and 1 post-smoothing step can be found in Table 4.18 (resp., Table 4.17). The times for one iteration of the multigrid algorithms at level 6 (where there are roughly 5×10^4 dofs) are also included in Tables 4.18 and 4.17.

Table 4.17. The Contraction Numbers of the Symmetric W-cycle Algorithm with $m = 1$, Together with the Time (in Seconds) for One Iteration of the W-cycle Algorithm at Level 6 (L-shaped Domain).

β \ k	1	2	3	4	5	6	Time
10^{-2}	7.97e-01	7.04e-01	6.32e-01	6.07e-01	6.01e-01	5.92e-01	1.56e-01
10^{-4}	2.18e-01	4.64eâĹŠ01	7.54e-01	6.68e-01	6.18e-01	5.91e-01	1.57e-01
10^{-6}	4.02e-01	1.63e-01	4.06e-01	8.61e-01	7.67e-01	6.57e-01	1.59e-01

Table 4.18. The Contraction Numbers of the Symmetric V-cycle Algorithm with $m = 1$, Together with the Time (in Seconds) for One Iteration of the V-cycle Algorithm at Level 6 (L-shaped Domain).

β \ k	1	2	3	4	5	6	Time
10^{-2}	7.97e-01	7.85e-01	7.89e-01	7.93e-01	7.96e-01	7.99e-01	4.70e-02
10^{-4}	2.18e-01	4.67e-01	7.56e-01	7.57e-01	7.64e-01	7.71e-01	4.73e-02
10^{-6}	4.02e-01	1.62e-01	4.20e-01	8.62e-01	8.40e-01	8.36e-01	4.74e-02

Remark 4.25. *Numerical results indicate that our multigrid algorithms are robust for nonconvex domains. Moreover, our symmetric multigrid algorithm is a contraction with* $m = 1$.

Example 4.26 (Comparison with preconditioned GMRES)**.** *In this example we compare our* W*-cycle algorithm with preconditioned GMRES with restart after* 10 *iterations (PGMRES(10)). See Section 2.6.2 for details of GMRES.*

In the case of unit square $\Omega = (0,1)^2$*, we take* $\boldsymbol{\zeta} = \frac{1}{2}[1\ 0]^t$*,* $\gamma = 0$ *and* $y_d = x_1(1-x_1)x_2(1-x_2)$ *in* (1.1.4)*, and* $C_\dagger = 5$ *in* (4.1.25)*. In the case of unit cube* $\Omega = (0,1)^3$*, we take* $\boldsymbol{\zeta} = \frac{1}{2}[1\ 1\ 1]^t$*,* $\gamma = 0$ *and* $y_d = 1$ *in* (1.1.4)*, and* $C_\dagger = 4$ *in* (4.1.25)*. We set the tolerance in Euclidean norm to be* 10^{-8} *for both methods.*

Here the preconditioner $\mathfrak{C}_k^{-1}$ *in our multigrid method is based on a* $V(1,1)$ *or* $V(2,2)$ *multigrid solve for* (4.1.14)*. We used symmetric* V*-cycle algorithm with 4 pre-smoothing steps and 4 post-smoothing steps as the left preconditioner for GMRES.*

The computational times (in seconds) for unit square case are presented in Table 4.19 (where there are about 1.7×10^7 dofs). The computational times (in seconds) for unit cube case are presented in Table 4.20 (where there are about 4.0×10^6 dofs). We do not include GMRES and GMRES(k) (cf. Section 2.6.2) because they fail to solve the problems at fine levels where the condition number of the system

is large. Meanwhile, we observe that our W-cycle algorithm and PGMRES(10) are comparable with respect to computational times. These numerical results also indicate that our multigrid methods can serve as good preconditioners for other iterative methods.

Table 4.19. Computational Times of the Symmetric W-cycle Algorithm with m Smoothing Steps and PGMRES(10) at Level 11 (Unit Square).

Use V(1,1) for $\mathfrak{C}_k^{-1}$				
β \ m	1	2	4	PGMRES(10)
10^{-2}	5.0e+02	5.9e+02	8.0e+02	5.5e+02
10^{-4}	4.4e+02	5.6e+02	7.6e+02	5.2e+02
10^{-8}	3.5e+02	3.9e+02	5.4e+02	4.8e+02
Use V(2,2) for $\mathfrak{C}_k^{-1}$				
β \ m	1	2	4	PGMRES(10)
10^{-2}	4.3e+02	5.8e+02	6.8e+02	5.5e+02
10^{-4}	4.3e+02	5.5e+02	7.2e+02	5.1e+02
10^{-8}	3.2e+02	3.7e+02	4.8e+02	4.7e+02

Table 4.20. Computational Times of the Symmetric W-cycle Algorithm with m Smoothing Steps and PGMRES(10) at Level 6 (Unit Cube).

Use V(1,1) for $\mathfrak{C}_k^{-1}$				
β \ m	1	2	4	PGMRES(10)
10^{-2}	4.5e+02	4.5e+02	4.3e+02	4.4e+02
10^{-4}	4.5e+02	4.4e+02	4.3e+02	4.4e+02
10^{-8}	3.4e+02	3.4e+02	3.5e+02	2.4e+02
Use V(2,2) for $\mathfrak{C}_k^{-1}$				
β \ m	1	2	4	PGMRES(10)
10^{-2}	3.1e+02	3.1e+02	3.2e+02	3.5e+02
10^{-4}	3.1e+02	3.1e+02	3.5e+02	4.2e+02
10^{-8}	2.5e+02	2.5e+02	2.9e+02	2.5e+02

Chapter 5

Multigrid Methods for Elliptic Optimal Control Problems with Pointwise State Constraints

In this chapter we discuss multigrid methods for (1.1.3)–(1.1.5) based on the P_1 finite element methods introduced in Section 3.2. We assume $\boldsymbol{\zeta} = \mathbf{0}$ and $\gamma = 0$ for simplicity. Throughout the chapter we focus on the finite element method (3.2.8).

Recall the discrete problem (3.2.8) is ,

$$\bar{y}_h = \underset{y_h \in K_h}{\operatorname{argmin}} \left[\frac{1}{2}(y_h - y_d, y_h - y_d)_h + \frac{\beta}{2}(\widetilde{\Delta}_h y_h, \widetilde{\Delta}_h y_h)_h \right],$$

where

$$K_h = \{y \in V_h : y_h \leq \psi \quad \text{at the vertices of } \mathcal{T}_h\}.$$

Let $\mathbf{A}_h$ be the stiffness matrix representing the bilinear form $(\nabla\cdot, \nabla\cdot)_{L^2(\Omega)}$ with respect to the nodal basis of V_h and $\widetilde{\mathbf{M}}_h$ be the diagonal matrix representing the bilinear form defined by (3.2.4). Then the matrix representing $\widetilde{\Delta}_h$ is given by $-\widetilde{\mathbf{M}}_h^{-1}\mathbf{A}_h$.

We can rewrite (3.2.8) in matrix form, namely,

$$\begin{aligned} \bar{\mathbf{y}}_h &= \underset{\mathbf{y}_h \leq \boldsymbol{\psi}}{\operatorname{argmin}} \frac{1}{2}(\mathbf{y}_h - \mathbf{y}_d)^t \widetilde{\mathbf{M}}_h (\mathbf{y}_h - \mathbf{y}_d) + \frac{\beta}{2}\mathbf{y}_h^t \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \widetilde{\mathbf{M}}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h \mathbf{y}_h \\ &= \underset{\mathbf{y}_h \leq \boldsymbol{\psi}}{\operatorname{argmin}} \frac{1}{2}\mathbf{y}_h^t \left[\beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h + \widetilde{\mathbf{M}}_h \right] \mathbf{y}_h - \mathbf{y}_h^t (\widetilde{\mathbf{M}}_h \mathbf{y}_d) \\ &= \underset{\mathbf{y}_h \leq \boldsymbol{\psi}}{\operatorname{argmin}} \frac{1}{2}\mathbf{y}_h^t \mathbf{B}_h \mathbf{y}_h - \mathbf{y}_h^t \tilde{\mathbf{y}}_d, \end{aligned} \tag{5.0.1}$$

where $\mathbf{B}_h = \beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h + \widetilde{\mathbf{M}}_h$ and $\tilde{\mathbf{y}}_d = \widetilde{\mathbf{M}}_h \mathbf{y}_d$. Note that $\mathbf{B}_h$ is SPD. Moreover, $\mathbf{B}_h$ is available due to the mass lumping technique (3.2.4) (cf. Remark 3.11).

Remark 5.1. *Let $\mathbf{M}_h$ be the mass matrix represents the bilinear form $(\cdot,\cdot)_{L^2(\Omega)}$ with respect to the nodal basis of V_h. The matrix form of the discrete problem* (3.2.7) *is the following,*

$$\bar{\mathbf{y}}_h = \underset{\mathbf{y}_h \leq \boldsymbol{\psi}}{\operatorname{argmin}} \frac{1}{2}\mathbf{y}_h^t \mathbf{B}_h \mathbf{y}_h - \mathbf{y}_h^t \tilde{\mathbf{y}}_d, \tag{5.0.2}$$

where $\mathbf{B}_h = \beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h + \mathbf{M}_h$ *and* $\tilde{\mathbf{y}}_d = \mathbf{M}_h \mathbf{y}_d$.

5.1 Primal-dual Active Set Algorithm

Our goal is to solve (5.0.1) efficiently. One of the most efficient methods is a primal-dual active set strategy (PDAS) which was developed in [12, 66, 67, 9, 10]. PDAS is an active set strategy involving primal as well as dual variables. This method is related to the early work [61]. It was also shown that it is related to the semismooth Newton method [63]. For convergence properties, we refer to [12, 66, 67, 9, 10, 63, 68] and the references therein.

We illustrate the primal-dual active set strategy by a simple example (cf. [63]). Consider the following finite dimensional minimization problem,

$$\begin{aligned} &\min J(y) = \frac{1}{2}(y, Ay) - (f, y), \\ &\text{subject to } y \leq \psi, \end{aligned} \tag{5.1.1}$$

where $A \in \mathbb{R}^{n \times n}$ is SPD, $f, \psi \in \mathbb{R}^n$ and $(\cdot, \cdot)$ denotes the inner product in $\mathbb{R}^n$. The optimality system for (5.1.1) is

$$Ay + \lambda = f, \tag{5.1.2}$$

$$y \leq \psi,\ \lambda \geq 0,\ (\lambda, y - \psi) = 0, \tag{5.1.3}$$

where y is the primal variable and λ is the dual variable. The key observation [10, 12] is that (5.1.3) is equivalent to

$$\lambda = \max(0, \lambda + c(y - \psi)) \tag{5.1.4}$$

for any $c > 0$, where the max-operation is understood componentwise. Hence (5.1.2)–(5.1.3) is equivalent to

$$\begin{aligned} &Ay + \lambda = f, \\ &\lambda = \max(0, \lambda + c(y - \psi)). \end{aligned} \tag{5.1.5}$$

The primal-dual active set method is based on utilizing (5.1.4) as a prediction strategy. Given a current primal-dual pair (y, λ), the choice for the next active and inactive set is given by

$$\mathcal{I} = \{i : \lambda_i + c(y - \psi)_i \leq 0\} \quad \text{and} \quad \mathcal{A} = \{i : \lambda_i + c(y - \psi)_i > 0\}. \tag{5.1.6}$$

Note that by (5.1.3), the definition (5.1.6) is equivalent to

$$\mathcal{I} = \{i : \lambda_i = 0 \text{ and } y_i \leq \psi_i\} \quad \text{and} \quad \mathcal{A} = \{i : \lambda_i > 0 \text{ and } y_i = \psi_i\}. \tag{5.1.7}$$

Hence when a node is active, the primal variable equals to the obstacle. Using (5.1.6) we have the following algorithm [63].

Algorithm 5.1 Primal-Dual Active Set Algorithm

1: Initialize y^0, λ^0. Set $k = 0$.
2: Set $\mathcal{I}_k = \{i : \lambda_i^k + c(y^k - \psi)_i \leq 0\}$ and $\mathcal{A}_k = \{i : \lambda_i^k + c(y^k - \psi)_i > 0\}$.
3: Solve $Ay^{k+1} + \lambda^{k+1} = f$, $y^{k+1} = \psi$ on $\mathcal{A}_k$, $\lambda^{k+1} = 0$ on $\mathcal{I}_k$.
4: Stop, or set $k = k + 1$ and return to step 2.

Remark 5.2. *The primal-dual active set algorithm terminates if the active set and inactive set stop changing. In practice, we choose a large constant c, for example,* 10^8*, to get better prediction of the active/inactive sets. It is shown in [63], the primal-dual active set algorithm converges if the initial guess is sufficiently close to the true solution (with respect to usual Euclidean norm) and the convergence is superlinear. This is similar to the behavior of the classical Newton's method.*

5.2 Primal-dual Active Set Algorithm with Multigrid Solver

It is easy to see that (5.0.1) is of the form (5.1.1). Hence we apply Algorithm 5.1 to (5.0.1) and obtain the following algorithm.

1. Given an initial guess $(\mathbf{y}_0, \boldsymbol{\lambda}_0)$ where $\boldsymbol{\lambda}_0 \geq 0$, we define

$$\begin{aligned} \mathcal{A}_0 &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_0(j) + c(\mathbf{y}_0(j) - \boldsymbol{\psi}(j)) > 0\}, \\ \mathcal{I}_0 &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_0(j) + c(\mathbf{y}_0(j) - \boldsymbol{\psi}(j)) \leq 0\} = \mathfrak{n} \setminus \mathcal{A}_0. \end{aligned}$$

2. For $k \geq 1$ we recursively solve the system

$$\begin{aligned} \mathbf{B}_h \mathbf{y}_k + \boldsymbol{\lambda}_k &= \tilde{\mathbf{y}}_d, && \text{(5.2.1a)} \\ \mathbf{y}_k &= \boldsymbol{\psi} \quad \text{on} \quad \mathcal{A}_{k-1}, && \text{(5.2.1b)} \\ \boldsymbol{\lambda}_k &= 0 \quad \text{on} \quad \mathcal{I}_{k-1}. && \text{(5.2.1c)} \end{aligned}$$

3. Then update the active set and inactive set by

$$\begin{aligned} \mathcal{A}_k &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_k(j) + c(\mathbf{y}_k(j) - \boldsymbol{\psi}(j)) > 0\}, \\ \mathcal{I}_k &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_k(j) + c(\mathbf{y}_k(j) - \boldsymbol{\psi}(j)) \leq 0\} = \mathfrak{n} \setminus \mathcal{A}_k. \end{aligned}$$

Here $\mathfrak{n} = \{1, 2, \ldots, \dim V_h\}$.

Note that the most expensive step of the algorithm is to solve the equation (5.2.1a). Furthermore we can rewrite (5.2.1) by using $\mathcal{I}_k$ and $\mathcal{A}_k$ as index sets,

$$\begin{aligned} (\mathbf{y}_k)_{\mathcal{A}_{k-1}} &= \boldsymbol{\psi}_{\mathcal{A}_{k-1}}, \\ (\boldsymbol{\lambda}_k)_{\mathcal{I}_{k-1}} &= 0, \\ (\mathbf{B}_h)_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}} (\mathbf{y}_k)_{\mathcal{I}_{k-1}} &= (\tilde{\mathbf{y}}_d)_{\mathcal{I}_{k-1}} - (\mathbf{B}_h)_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} \boldsymbol{\psi}_{\mathcal{A}_{k-1}}, \qquad \text{(5.2.2)} \\ (\boldsymbol{\lambda}_k)_{\mathcal{A}_{k-1}} &= (\tilde{\mathbf{y}}_d)_{\mathcal{A}_{k-1}} - (\mathbf{B}_h \mathbf{y}_k)_{\mathcal{A}_{k-1}}. \end{aligned}$$

It is clear that (5.2.2) is a reduced system that we need to solve during each PDAS iteration. In practice, this system becomes harder to solve when the mesh size h is small, especially in three dimensions. Our goal is to design multigrid solver for the reduced system (5.2.2). Let V_h be the finite element space at refinement level r and $\mathcal{I}_r$ be the inactive set $\mathcal{I}_{k-1}$ for a particular k. The general system we want to solve is

$$\mathbf{B}_{\mathcal{I}_r \mathcal{I}_r} \mathbf{y}_{\mathcal{I}_r} = \mathbf{g}_{\mathcal{I}_r}. \qquad \text{(5.2.3)}$$

Let the output of the W-cycle multigrid method at level r be denoted by $MG_W(r, \mathbf{g}, \mathbf{y}_0, m_1, m_2, \mathcal{I}_r)$ where $\mathbf{g}$ is the right-hand side, $\mathbf{y}_0$ is the initial guess,

$m_1(resp., m_2)$ is the number of pre-smoothing (resp., post-smoothing) steps and $\mathcal{I}_r$ is the inactive set at level r. We use direct solve when $r \leq 2$ to avoid empty inactive set at initial level and level 1. It is possible to have empty inactive set at low levels since the problem size is very small. When $r > 2$ the W-cycle algorithm for the system (5.2.3) is as follows.

• **Pre-Smoothing.** For $1 \leq j \leq m_1$,

$$\mathbf{y}_{\mathcal{I}_r}^{j} = \mathbf{y}_{\mathcal{I}_r}^{j-1} + \gamma_r(\mathbf{g}_{\mathcal{I}_r} - \mathbf{B}_{\mathcal{I}_r\mathcal{I}_r}\mathbf{y}_{\mathcal{I}_r}^{j-1}), \tag{5.2.4}$$

where γ_r is the damping factor in Richardson iteration.

• **Coarse-Grid Correction.** First we calculate the residual of the system (5.2.3),

$$\mathbf{f}_{\mathcal{I}_r} = \mathbf{g}_{\mathcal{I}_r} - \mathbf{B}_{\mathcal{I}_r\mathcal{I}_r}\mathbf{y}_{\mathcal{I}_r}^{m_1}. \tag{5.2.5}$$

Then we extend the residual to all grid points, namely

$$\mathbf{f} = \mathbf{f}_{\mathcal{I}_r} + \mathbf{f}_{\mathcal{A}_r}, \tag{5.2.6}$$

where $\mathbf{f}_{\mathcal{A}_r} = \mathbf{0}$. After that we transfer the residual to level $(r-1)$. Suppose the matrix represents the coarse-to-fine operator I_{r-1}^{r} is $\mathbf{I}_{r-1}^{r}$, we have

$$\mathbf{g}' = (\mathbf{I}_{r-1}^{r})^t\mathbf{f}. \tag{5.2.7}$$

Then we generate the coarse level inactive set $\mathcal{I}_{r-1}$ from the current inactive set $\mathcal{I}_r$. We use the same procedure in [65, 69, 72]. Specifically, on level $r-1$ we only label a node as inactive when it is inactive on level r together with all its neighbors. See Figure 5.1 for a simple example. If the red node and all the black nodes are inactive at level r, then the red node is inactive at level $r-1$ since it is

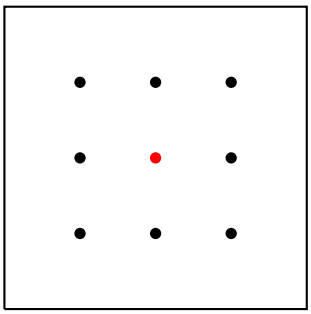

Figure 5.1. Inactive Nodes.

inactive at level r with all its neighbors. Once we obtain the inactive set $\mathcal{I}_{r-1}$ we perform the following,

$$\begin{aligned} \mathbf{q}'_{\mathcal{I}_{r-1}} &= MG_W(r-1, \mathbf{g}', 0, m_1, m_2, \mathcal{I}_{r-1}), \\ \mathbf{q}_{\mathcal{I}_{r-1}} &= MG_W(r-1, \mathbf{g}', \mathbf{q}'_{\mathcal{I}_{r-1}}, m_1, m_2, \mathcal{I}_{r-1}). \end{aligned}$$

Then we extend $\mathbf{q}_{\mathcal{I}_{r-1}}$ to all grid points by

$$\mathbf{q} = \mathbf{q}_{\mathcal{I}_{r-1}} + \mathbf{q}_{\mathcal{A}_{r-1}}, \tag{5.2.8}$$

where $\mathbf{q}_{\mathcal{A}_{r-1}} = \mathbf{0}$. At last let

$$\mathbf{y}_{\mathcal{I}_r}^{m_1+1} = \mathbf{y}_{\mathcal{I}_r}^{m_1} + (\mathbf{I}_{r-1}^r \mathbf{q})_{\mathcal{I}_r}. \tag{5.2.9}$$

• **Post-Smoothing.** For $m_1 + 2 \leq j \leq m_1 + m_2 + 1$,

$$\mathbf{y}_{\mathcal{I}_r}^{j} = \mathbf{y}_{\mathcal{I}_r}^{j-1} + \gamma_r(\mathbf{g}_{\mathcal{I}_r} - \mathbf{B}_{\mathcal{I}_r\mathcal{I}_r}\mathbf{y}_{\mathcal{I}_r}^{j-1}). \tag{5.2.10}$$

Remark 5.3. *In practice, in order to generate the inactive set $\mathcal{I}_{r-1}$ from $\mathcal{I}_r$, we assign a vector $\mathbf{v}$ with 1s at $\mathcal{A}_r$ and 0s at $\mathcal{I}_r$, calculate $\mathbf{v}' = (\mathbf{I}_{r-1}^r)^t\mathbf{v}$ and designate the nodes with zero entries in $\mathbf{v}'$ as inactive at level $r-1$.*

Remark 5.4. *We choose $\gamma_r = \frac{C}{\beta h_r^{-2} + h_r^2}$ in the pre-smoothing step* (5.2.4) *and the post-smoothing step* (5.2.10) *where C is a constant, h_r is the mesh size at level r. This is due to the fact $\rho(\gamma_r \mathbf{B}_h) \leq 1$. Alternatively, let $\mathbf{B}_{\mathcal{I}_r\mathcal{I}_r} = \mathbf{D} + \mathbf{L} + \mathbf{U}$, we can*

replace the pre-smoothing step (5.2.4) *and the post-smoothing step* (5.2.10) *by the following symmetric Gauss-Seidel iteration (cf. Section 2.5.2).*

$$\mathbf{y}_{\mathcal{I}_r}^{j} = \mathbf{y}_{\mathcal{I}_r}^{j-1} + (\mathbf{U}+\mathbf{D})^{-1}\mathbf{D}(\mathbf{L}+\mathbf{D})^{-1}(\mathbf{g}_{\mathcal{I}_r} - \mathbf{B}_{\mathcal{I}_r\mathcal{I}_r}\mathbf{y}_{\mathcal{I}_r}^{j-1}). \tag{5.2.11}$$

The advantage of using symmetric Gauss-Seidel iteration (SGS) as smoothers is that we do not need to choose the parameter γ_r.

Overall, the primal-dual active set algorithm with multigrid solver for (5.0.1) is described in Algorithm 5.2.

Algorithm 5.2 PDAS Algorithm with Multigrid Solver for (5.0.1) at Level r.

1: Initialize $(\mathbf{y}_0, \boldsymbol{\lambda}_0)$ where $\boldsymbol{\lambda}_0 \geq 0$, ε and c. Given $\boldsymbol{\psi}$. Compute

$$\begin{aligned} \mathcal{A}_0 &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_0(j) + c(\mathbf{y}_0(j) - \boldsymbol{\psi}(j)) > 0\}, \\ \mathcal{I}_0 &= \mathfrak{n} \setminus \mathcal{A}_0. \end{aligned}$$

2: For $k \geq 1$, $(\mathbf{y}_k)_{\mathcal{A}_{k-1}} \leftarrow \boldsymbol{\psi}_{\mathcal{A}_{k-1}}$, $(\boldsymbol{\lambda}_k)_{\mathcal{I}_{k-1}} \leftarrow 0$.
3: Let $\mathcal{I}_r = \mathcal{I}_{k-1}$ and $\mathcal{A}_r = \mathcal{A}_{k-1}$.
4: $\mathbf{f} \leftarrow (\tilde{\mathbf{y}}_d)_{\mathcal{I}_r} - \mathbf{B}_{\mathcal{I}_r\mathcal{A}_r}\boldsymbol{\psi}_{\mathcal{A}_r}$.
5: **repeat**
6: $\quad (\mathbf{y}_k)_{\mathcal{I}_r} \leftarrow MG_W(r, \mathbf{f}, (\mathbf{y}_k)_{\mathcal{I}_r}, m_1, m_2, \mathcal{I}_r)$
7: **until** $\frac{\|\mathbf{f} - \mathbf{B}_{\mathcal{I}_r\mathcal{I}_r}(\mathbf{y}_k)_{\mathcal{I}_r}\|}{\|\mathbf{f}\|} \leq \varepsilon$.
8: $(\boldsymbol{\lambda}_k)_{\mathcal{A}_{k-1}} \leftarrow (\tilde{\mathbf{y}}_d)_{\mathcal{A}_{k-1}} - (\mathbf{B}\mathbf{y}_k)_{\mathcal{A}_{k-1}}$.
9: Update the active and inactive sets

$$\begin{aligned} \mathcal{A}_k &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_k(j) + c(\mathbf{y}_k(j) - \boldsymbol{\psi}(j)) > 0\}, \\ \mathcal{I}_k &= \mathfrak{n} \setminus \mathcal{A}_0. \end{aligned}$$

10: Stop when $\mathcal{A}_k = \mathcal{A}_{k-1}$ or go to step 2.

Remark 5.5. *As we mentioned in Remark 5.2, PDAS converges if the initial guess is sufficiently close to the true solution. Hence the choice of the initial guess* $(\mathbf{y}_0, \boldsymbol{\lambda}_0)$ *in Algorithm 5.2 is important. In practice, we solve the problem at level* 0 *with zero initial guess and use that as the initial solution for level* 1*. In general, we use the solution at level* $r-1$ *as the initial guess for level* r*.*

5.3 Numerical Results

In this section we present the numerical results of the symmetric W-cycle algorithm for (5.0.1) on two and three dimensional domains. We compute the contraction numbers using similar strategy in Algorithm 4.1. We employed the MATLAB/C++ toolbox FELICITY [101] in our computations.

Example 5.6 (No State Constraints)**.** *In this example we consider an extreme case which no state constraints are imposed in* (3.2.8)*. Hence it is equivalent to solve the following system,*

$$\mathbf{B}_h \bar{\mathbf{y}}_h = \tilde{\mathbf{y}}_d, \tag{5.3.1}$$

where $\mathbf{B}_h = \beta \mathbf{A}_h \widetilde{\mathbf{M}}_h^{-1} \mathbf{A}_h + \widetilde{\mathbf{M}}_h$ *and* $\tilde{\mathbf{y}}_d = \widetilde{\mathbf{M}}_h \mathbf{y}_d$*. Our* W*-cycle algorithm can still apply to* (5.3.1) ($\mathcal{A}_r = \varnothing$)*. We take* $\Omega = (0,1)^2$ *and* $\beta = 1$*. We use symmetric Gauss-Seidel iteration* (5.2.11) *as smoother in this example. See Figure 2.2 for the meshes. In this case* $\widetilde{\mathbf{M}}_h$ *is a multiple of the identity matrix since every interior node has six triangles around it.*

We report the contraction numbers of our W-cycle algorithm in Table 5.1. It is clear that our W-cycle algorithm is a contraction when $m = 1$ and the behavior of the contraction numbers agrees with the standard $O(m^{-\frac{1}{2}})$ multigrid performance for fourth order problems [25].

Table 5.1. Contraction Numbers of W-cycle Algorithm for Example 5.6.

m \ k	1	2	3	4	5	6
2^0	2.85e-01	3.64e-01	4.10e-01	4.23e-01	4.24e-01	4.27e-01
2^1	8.92e-02	1.66e-01	2.05e-01	2.19e-01	2.26e-01	2.26e-01
2^2	1.51e-02	8.50e-02	1.27e-01	1.43e-01	1.47e-01	1.51e-01
2^3	1.84e-03	5.42e-02	8.57e-02	1.02e-01	1.04e-01	1.07e-01
2^4	4.10e-05	3.78e-02	5.19e-02	7.02e-02	7.18e-02	7.50e-02
2^5	2.98e-08	2.13e-02	3.69e-02	4.83e-02	4.89e-02	5.18e-02
2^6	1.81e-15	6.87e-03	2.13e-02	2.75e-02	3.70e-02	3.60e-02

Example 5.7 (Disk Active Set [34, Example 7.1])**.** *In this example we take* $\Omega = [-4, 4]^2$, $\beta = 1$ *and* $\psi = |x|^2 - 1$. *We use* y_d *in Example 3.17 here.*

We take $c = 10^8$, $\varepsilon = 10^{-8}$ *in Algorithm 5.2 and* $\gamma_r = 0.015h_r^2$ *in smoothing steps* (5.2.4) *and* (5.2.10). $\mathcal{T}_r$ *is a regular triangulation of the domain* Ω *(see Figure 2.2). In this example the resulting active set is a disk which is depicted in Figure 5.2.*

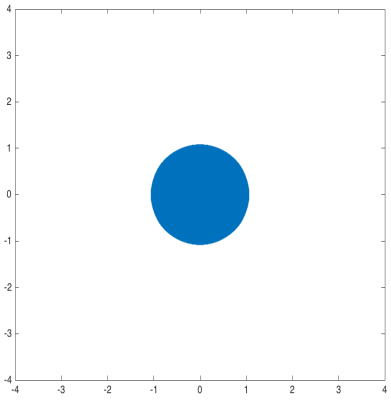

Figure 5.2. Disk Active Set.

We report the contraction numbers of the symmetric W-cycle algorithm in Tables 5.2–5.5. At each level, we compute the contraction numbers in each PDAS iterations with smoothing steps $m = 2^1, 2^2, 2^3$. We use Richardson iteration and

symmetric Gauss-Seidel iteration as smoothers. We need k PDAS iterations to obtain the solution at each level. The number k varies from level to level and hence the tables have different number of rows. The results start from level 4 since we take level 2 as the coarsest level. We observe that our symmetric W-cycle algorithm is a contraction with $m = 2$ for both smoothers. The performance of our multigrid algorithm with SGS is clearly better. See Figure 5.3 for an example of the active sets at different levels generated by the multigrid algorithm.

Table 5.2. Level 4 Contraction Numbers for Example 5.7.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	6.66e-01	4.69e-01	2.80e-01	4.15e-01	2.52e-01	1.89e-01
$k = 2$	8.32e-01	7.13e-01	5.72e-01	6.85e-01	5.17e-01	3.73e-01
$k = 3$	8.32e-01	7.13e-01	5.72e-01	6.88e-01	5.15e-01	3.73e-01

Table 5.3. Level 5 Contraction Numbers for Example 5.7.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	7.62e-01	5.98e-01	4.07e-01	4.78e-01	3.33e-01	2.68e-01
$k = 2$	8.96e-01	8.21e-01	7.29e-01	7.97e-01	6.76e-01	5.65e-01
$k = 3$	8.96e-01	8.21e-01	7.29e-01	7.98e-01	6.77e-01	5.65e-01
$k = 4$	8.96e-01	8.21e-01	7.29e-01	7.95e-01	6.77e-01	5.65e-01

Table 5.4. Level 6 Contraction Numbers for Example 5.7.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	9.21e-01	8.45e-01	7.34e-01	8.19e-01	6.83e-01	5.35e-01
$k = 2$	8.79e-01	7.83e-01	6.54e-01	7.65e-01	6.09e-01	4.61e-01
$k = 3$	8.89e-01	8.03e-01	6.98e-01	7.79e-01	6.44e-01	5.26e-01
$k = 4$	8.92e-01	8.08e-01	7.07e-01	7.80e-01	6.51e-01	5.36e-01

Table 5.5. Level 7 Contraction Numbers for Example 5.7.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k=1$	9.10e-01	8.31e-01	7.18e-01	8.06e-01	6.65e-01	5.30e-01
$k=2$	9.27e-01	8.66e-01	7.99e-01	8.53e-01	7.57e-01	6.60e-01
$k=3$	9.36e-01	8.83e-01	8.22e-01	8.69e-01	7.84e-01	6.96e-01
$k=4$	9.33e-01	8.83e-01	8.23e-01	8.69e-01	7.83e-01	6.94e-01
$k=5$	9.35e-01	8.83e-01	8.23e-01	8.69e-01	7.83e-01	6.95e-01
$k=6$	9.31e-01	8.85e-01	8.23e-01	8.68e-01	7.81e-01	6.93e-01
$k=7$	9.32e-01	8.85e-01	8.23e-01	8.68e-01	7.81e-01	6.92e-01
$k=8$	9.35e-01	8.87e-01	8.24e-01	8.69e-01	7.83e-01	6.94e-01
$k=9$	9.35e-01	8.87e-01	8.24e-01	8.68e-01	7.83e-01	6.95e-01

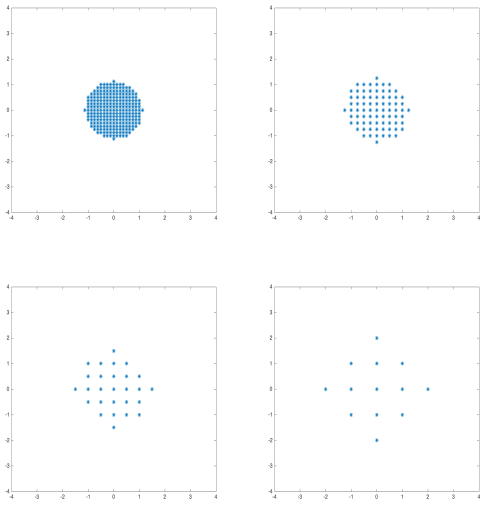

Figure 5.3. Active Sets for Example 5.7 at Different Levels.

Example 5.8 (Disjoint Active Set [36, Example 4])**.** *In this example we take* $\Omega = [0,1]^2$, $y_d = \sin(4\pi xy) + 1.5$, $\psi = 1$ *and* $\beta = 10^{-4}$ *in* (3.2.8). *Other parameters are identical as those of Example 5.7. In this example the resulting active set is disjoint which is shown in Figure 5.4.*

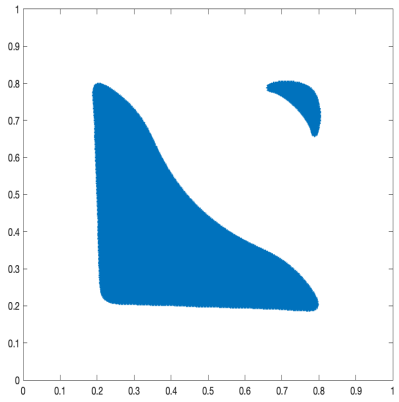

Figure 5.4. Disjoint Active Set.

We report the contraction numbers of the symmetric W-cycle algorithm for Example 5.8 in Tables 5.6–5.9. We observe that our symmetric W-cycle algorithm is a contraction with $m = 2$ for both smoothers. Again, our multigrid algorithm with SGS has better performance with respect to the contraction numbers. See Figure 5.5 for an example of the active sets at different levels generated by the multigrid algorithm.

Table 5.6. Level 4 Contraction Numbers for Example 5.8.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	7.74e-01	5.98e-01	3.83e-01	4.20e-01	2.18e-01	1.80e-01
$k = 2$	8.06e-01	6.74e-01	5.21e-01	6.39e-01	4.56e-01	3.24e-01
$k = 3$	8.64e-01	7.64e-01	6.30e-01	7.14e-01	5.58e-01	4.14e-01

Table 5.7. Level 5 Contraction Numbers for Example 5.8.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	7.77e-01	6.04e-01	4.49e-01	5.35e-01	3.98e-01	3.01e-01
$k = 2$	8.88e-01	7.91e-01	6.53e-01	7.57e-01	6.06e-01	4.59e-01
$k = 3$	8.88e-01	7.94e-01	6.72e-01	7.60e-01	6.15e-01	4.80e-01
$k = 4$	8.93e-01	8.07e-01	6.95e-01	7.71e-01	6.30e-01	5.07e-01

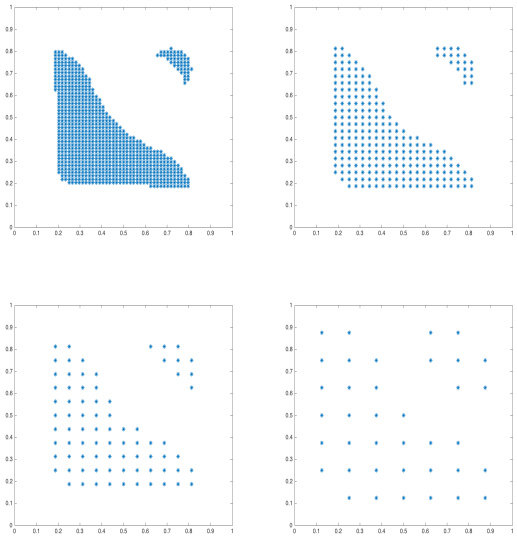

Figure 5.5. Active Sets for Example 5.8 at Different Levels.

Table 5.8. Level 6 Contraction Numbers for Example 5.8.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k=1$	8.07e-01	6.55e-01	4.52e-01	5.54e-01	3.73e-01	3.19e-01
$k=2$	9.08e-01	8.31e-01	7.09e-01	8.04e-01	6.62e-01	5.14e-01
$k=3$	9.06e-01	8.26e-01	7.15e-01	8.01e-01	6.73e-01	5.51e-01
$k=4$	9.08e-01	8.30e-01	7.16e-01	8.03e-01	6.71e-01	5.46e-01
$k=5$	9.08e-01	8.30e-01	7.16e-01	8.03e-01	6.72e-01	5.47e-01

Table 5.9. Level 7 Contraction Numbers for Example 5.8.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k=1$	8.17e-01	4.71e-01	5.82e-01	6.45e-01	5.36e-01	4.06e-01
$k=2$	9.16e-01	8.49e-01	7.40e-01	8.27e-01	6.96e-01	5.48e-01
$k=3$	9.17e-01	8.50e-01	7.43e-01	8.27e-01	6.95e-01	5.57e-01
$k=4$	9.16e-01	8.51e-01	7.43e-01	8.29e-01	7.00e-01	5.56e-01
$k=5$	9.15e-01	8.47e-01	7.37e-01	8.25e-01	6.92e-01	5.51e-01
$k=6$	9.15e-01	8.48e-01	7.38e-01	8.26e-01	6.94e-01	5.53e-01

Example 5.9 (Active Set with Empty Interior [35, Example 3]). *In this example we take* $\Omega = [0,1]^2$, $\beta = 1$, $\psi = 1 - 5|x|^2 - |x|^4$ *and* $y_d = 0$ *in the following problem which is very similar to* (3.2.8),

$$\bar{y}_h = \underset{y_h \in K_h}{\operatorname{argmin}} \left[\frac{1}{2}(y_h - y_d, y_h - y_d)_h + \frac{\beta}{2}(\widetilde{\Delta}_h y_h, \widetilde{\Delta}_h y_h)_h \right],$$

where

$$K_h = \{y \in V_h : y_h \geq \psi \quad \text{at the vertices of } \mathcal{T}_h\}.$$

Notice that the only difference is that ψ *is a lower bound of* y_h *instead of an upper bound. PDAS can be easily altered to handle this problem, specifically, use the following definition of active and inactive sets,*

$$\begin{aligned} \mathcal{A}_k &= \{j \in \mathfrak{n} : \boldsymbol{\lambda}_k(j) + c(\mathbf{y}_k(j) - \boldsymbol{\psi}(j)) < 0\}, \\ \mathcal{I}_k &= \mathfrak{n} \setminus \mathcal{A}_0. \end{aligned}$$

We refer to [35, Example 3] for more details about this example. In Figure 5.6, it shows that the active set in this example has an empty interior.

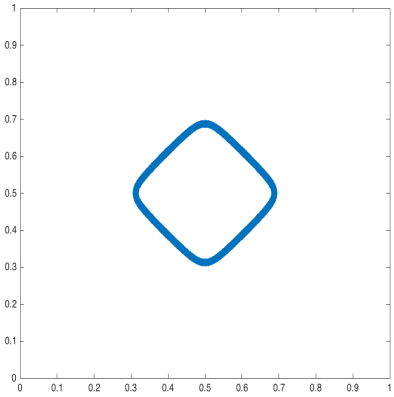

Figure 5.6. Active Set with Empty Interior.

We report the contraction numbers of the symmetric W-cycle algorithm at level 7 (where there are roughly 6.5×10^4 dofs) in Table 5.10. For simplicity, we omit

results from other levels. It takes 33 PDAS iterations to converge at level 7, we include the contraction numbers of first and last three iterations since other contraction numbers are similar. We observe that the symmetric W-cycle algorithm is a contraction with $m = 2$. See Figure 5.7 for an example of the active sets at different levels generated by the multigrid algorithm.

Table 5.10. Level 7 Contraction Numbers for Example 5.9.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k = 1$	9.50e-01	9.09e-01	8.63e-01	8.26e-01	7.58e-01	6.85e-01
$k = 2$	9.09e-01	8.44e-01	7.66e-01	7.10e-01	6.11e-01	4.99e-01
$k = 3$	9.52e-01	9.38e-01	9.37e-01	9.33e-01	8.98e-01	8.56e-01
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k = 31$	9.48e-01	9.36e-01	8.96e-01	8.59e-01	7.99e-01	7.34e-01
$k = 32$	9.39e-01	9.20e-01	8.95e-01	8.67e-01	8.05e-01	7.33e-01
$k = 33$	9.41e-01	9.17e-01	8.80e-01	8.46e-01	7.78e-01	7.00e-01

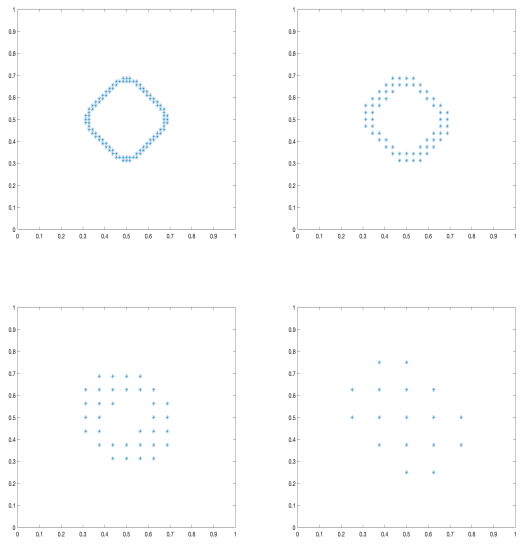

Figure 5.7. Active Sets for Example 5.9 at Different Levels.

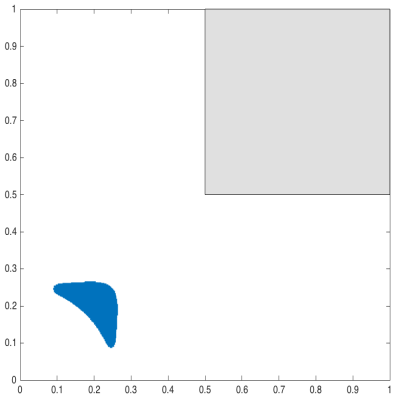

Figure 5.8. Active Set in a L-shaped Domain.

Example 5.10 (L-shaped Domain [27, Example 6.1])**.** *In this example we take* $\Omega = (0,1)^2 \setminus (0.5,1)^2$, $\beta = 1$, $\psi = [(\frac{2x-0.5}{0.48})^2 + (\frac{2y-0.5}{0.48})^2] - 1$ *and* $y_d = 0$ *in* (3.2.8). *Figure 5.8 shows the active set of this example.*

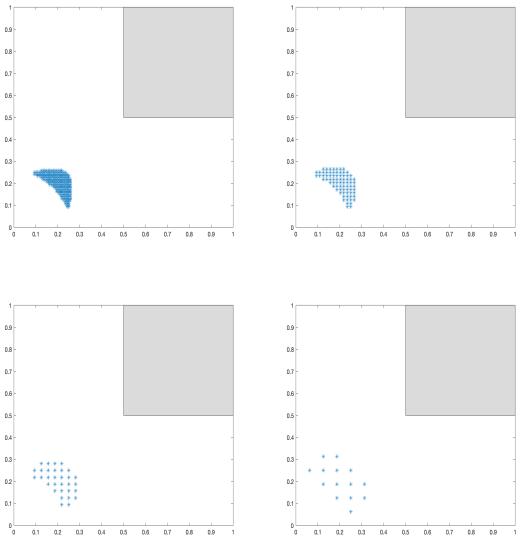

Figure 5.9. Active Sets for Example 5.10 at Different Levels.

We report the contraction numbers of the symmetric W-cycle algorithm at level 7 (where there are roughly 2×10^5 dofs) in Table 5.11. The numerical results indicate that our algorithm is a contraction with $m = 2$ on nonconvex domain.

See Figure 5.9 for an example of the active sets at different levels generated by the multigrid algorithm.

Table 5.11. Level 7 Contraction Numbers for Example 5.10.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k=1$	9.29e-01	8.71e-01	8.12e-01	8.54e-01	7.80e-01	6.94e-01
$k=2$	9.11e-01	8.20e-01	7.45e-01	7.76e-01	6.73e-01	5.67e-01
$k=3$	9.33e-01	8.71e-01	8.12e-01	8.46e-01	7.72e-01	6.86e-01
$k=4$	9.17e-01	8.32e-01	7.19e-01	7.66e-01	6.43e-01	5.13e-01
$k=5$	9.20e-01	8.49e-01	7.73e-01	8.04e-01	7.09e-01	6.15e-01
$k=6$	9.25e-01	8.47e-01	7.42e-01	7.86e-01	6.71e-01	5.58e-01
$k=7$	9.28e-01	8.64e-01	7.79e-01	8.15e-01	7.10e-01	6.01e-01
$k=8$	9.23e-01	8.43e-01	7.31e-01	7.64e-01	6.36e-01	5.15e-01
$k=9$	9.24e-01	8.51e-01	7.50e-01	7.83e-01	6.63e-01	5.45e-01
$k=10$	9.23e-01	8.46e-01	7.36e-01	7.69e-01	6.42e-01	5.20e-01
$k=11$	9.25e-01	8.53e-01	7.54e-01	7.93e-01	6.75e-01	5.54e-01
$k=12$	9.25e-01	8.52e-01	7.53e-01	7.93e-01	6.75e-01	5.54e-01
$k=13$	9.24e-01	8.51e-01	7.54e-01	7.93e-01	6.75e-01	5.54e-01

Example 5.11 (Cube [34, Example 7.5])**.** *In this example we take* $\Omega = [-4,4]^3$, $\beta = 1$, $\psi = |x|^2 - 1$ *and use identical* y_d *in Example 5.7 except replacing* $w(x)$ *with*

$$w(x) = 2\sin(\frac{\pi}{8}(x_1+4))\sin(\frac{\pi}{8}(x_2+4))\sin(\frac{\pi}{8}(x_3+4)).$$

This example is a three dimensional generalization of Example 5.7 (cf. [34, Example 7.5]). We set the coarsest level to be level 1 for this example. Figure 5.10 shows the ball-shaped active set.

We briefly report the contraction numbers of the symmetric W-cycle algorithm at level 5 (where there are roughly 2.5×10^5 dofs) in Table 5.11. It is clear that the algorithm is a contraction with $m = 2$. See Figure 5.11 for an example of the active sets at different levels generated by the multigrid algorithm.

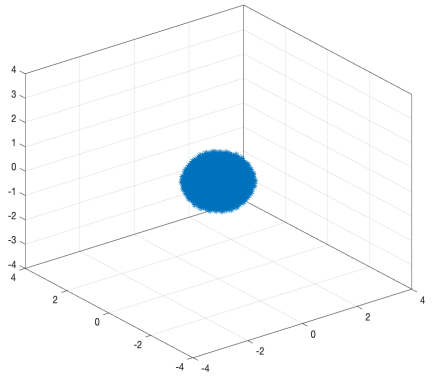

Figure 5.10. Active Set in a Cube.

Table 5.12. Level 5 Contraction Numbers for Example 5.11.

	Richardson			SGS		
m	2^1	2^2	2^3	2^1	2^2	2^3
$k=1$	9.60e-01	9.17e-01	8.46e-01	7.70e-01	6.72e-01	5.71e-01
$k=2$	9.60e-01	9.26e-01	8.66e-01	8.04e-01	7.17e-01	6.23e-01
$k=3$	9.45e-01	8.96e-01	8.18e-01	7.43e-01	6.47e-01	5.52e-01
$k=4$	9.45e-01	8.97e-01	8.19e-01	7.43e-01	6.46e-01	5.49e-01
$k=5$	9.45e-01	8.97e-01	8.19e-01	7.44e-01	6.46e-01	5.49e-01

Example 5.12 (Comparison with preconditioned MINRES)**.** *In this example we compare the computational time of our W-cycle algorithm with that of the preconditioned MINRES (cf. Section 2.6.1). We use $V(1,1)$ with SGS smoothers as the left preconditioner of MINRES in two dimensions and three dimensions.*

We report the computational times of Example 5.7 (resp., Examples 5.8) at level 8 (where there are roughly 2.6×10^5 dofs) in Table 5.13 (resp., Table 5.14) where $m = 2^1, 2^2$ for Richardson smoothers and $m = 2^0, 2^1$ for SGS smoothers.

We observe that our W-cycle algorithm with 2 SGS smoothing steps are faster than PMINRES with respect to the total computational time. For each PDAS iteration, our W-cycle algorithm becomes faster while PMINRES has similar per-

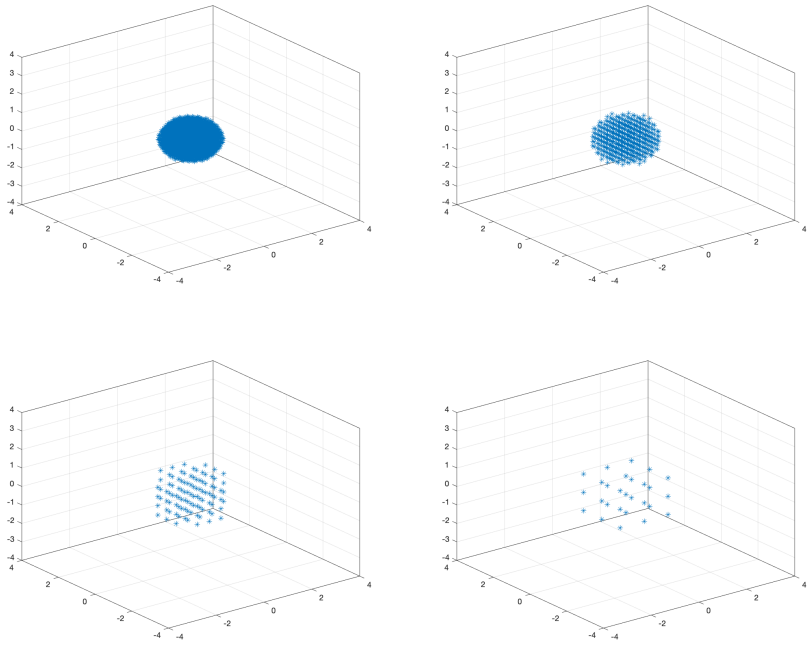

Figure 5.11. Active Sets for Example 5.11 at Different Levels.

formance. The numerical results also indicate that our multigrid algorithm with SGS smoothers can be used as preconditioners for other iterative methods.

Table 5.13. Level 8 Computational Times (in Seconds) for Example 5.7.

	Richardson		SGS		PMINRES
m	2^1	2^2	2^0	2^1	-
$k = 1$	48.0248	37.7692	32.4575	27.5180	18.3618
$k = 2$	47.1066	35.9143	31.4178	29.0540	18.0510
$k = 3$	34.4820	27.9079	25.8978	22.9460	17.7606
$k = 4$	27.5818	22.6635	22.1258	17.6000	18.2602
$k = 5$	25.9348	20.2082	19.6119	15.8561	18.1571
$k = 6$	18.6314	14.2765	13.3810	11.5135	19.2007
$k = 7$	14.2909	11.3900	10.9514	8.8073	22.0979
$k = 8$	13.3463	10.8952	10.6121	8.8232	20.8777

Table 5.14. Level 8 Computational Times (in Seconds) for Example 5.8.

	Richardson		SGS		PMINRES
m	2^1	2^2	2^0	2^1	-
$k=1$	19.9301	17.8965	8.8707	7.6867	8.6414
$k=2$	20.0603	17.3770	12.9998	10.7860	14.7845
$k=3$	7.8013	6.0977	5.4200	4.5432	14.7572
$k=4$	4.9935	3.9591	1.4424	3.0455	14.7656
$k=5$	1.6861	1.3349	0.9701	1.5330	14.8192
$k=6$	0.5818	0.8961	0.0334	0.7930	14.8007

We also provide the comparison results for Example 5.11 at level 5 and level 6 (about 2.05×10^6 dofs) in Tables 5.15 and 5.16. We also include the built-in function backslash in MATLAB which is a sparse direct solver. Backslash cannot solve the problem in a reasonable time at level 6 thus we ignore the results. As we can see in the numerical results, our W-cycle algorithm and the PMINRES have better performance than backslash at level 5 and level 6 which agree with the discussion in Section 5.2. Meanwhile, PMINRES is more efficient than W-cycle algorithm at level 5 and level 6. But since W-cycle algorithm is an $O(n)$ algorithm while MINRES is not, the gap between PMINRES and W-cycle algorithm will decrease as we increase levels.

Table 5.15. Level 5 Computational Times (in Seconds) for Example 5.11.

	Backslash	SGS ($m=2$)	PMINRES
$k=1$	109.4605	52.1178	19.4686
$k=2$	99.8426	57.7145	19.9472
$k=3$	106.9629	41.2576	17.9270
$k=4$	101.8495	34.5839	18.6486
$k=5$	101.2255	19.5259	18.6386

Table 5.16. Level 6 Computational Times (in Seconds) for Example 5.11.

	Backslash	SGS ($m = 2$)	PMINRES
$k = 1$	-	757.2665	248.4203
$k = 2$	-	817.2732	254.4706
$k = 3$	-	957.6223	270.2558
$k = 4$	-	558.7001	231.6902
$k = 5$	-	526.7740	246.9906
$k = 6$	-	417.7242	253.9759
$k = 7$	-	263.6513	253.8082

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140. Elsevier, 2003.

[2] E. Arian and S. Ta'asan. Multigrid one-shot methods for optimal control problems: Infinite dimensional control. ICASE-Report 94-52, NASA Langley Research Center, Hampton, VA, 1994.

[3] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9(1):17–29, 1951.

[4] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322–333, 1971.

[5] I. Babuška. The finite element method with Lagrangian multipliers. *Numerische Mathematik*, 20(3):179–192, 1973.

[6] L. Badea. Global convergence rate of a standard multigrid method for variational inequalities. *IMA Journal of Numerical Analysis*, 34(1):197–216, 2014.

[7] R. E. Bank and T. Dupont. An optimal order process for solving finite element equations. *Mathematics of Computation*, 36(153):35–51, 1981.

[8] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[9] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems. *SIAM Journal on Optimization*, 11(2):495–521, 2000.

[10] M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM Journal on Control and Optimization*, 37(4):1176–1194, 1999.

[11] M. Bergounioux and K. Kunisch. Augemented lagrangian techniques for elliptic state constrained optimal control problems. *SIAM Journal on Control and Optimization*, 35(5):1524–1543, 1997.

[12] M. Bergounioux and K. Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Computational Optimization and Applications*, 22(2):193–224, 2002.

[13] G. Biros and O. Ghattas. Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part I: The Krylov–Schur solver. *SIAM Journal on Scientific Computing*, 27(2):687–713, 2005.

[14] G. Biros and O. Ghattas. Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part II: The Lagrange–Newton solver and its application to optimal control of steady viscous flows. *SIAM Journal on Scientific Computing*, 27(2):714–739, 2005.

[15] D. Boffi, F. Brezzi, and M. Fortin. *Mixed Finite Element Methods and Applications*, volume 44. Springer, 2013.

[16] A. Borzì and K. Kunisch. A multigrid scheme for elliptic constrained optimal control problems. *Computational Optimization and Applications*, 31(3):309–333, 2005.

[17] A. Borzì and V. Schulz. Multigrid methods for PDE optimization. *SIAM Review*, 51(2):361–395, 2009.

[18] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*, volume 8. SIAM, 2011.

[19] D. Braess and W. Hackbusch. A new convergence proof for the multigrid method including the V-cycle. *SIAM Journal on Numerical Analysis*, 20(5):967–975, 1983.

[20] J. H. Bramble and J. E. Pasciak. New convergence estimates for multigrid algorithms. *Mathematics of Computation*, 49(180):311–329, 1987.

[21] J. H. Bramble and J. E. Pasciak. New estimates for multilevel algorithms including the V-cycle. *Mathematics of Computation*, 60(202):447–471, 1993.

[22] J. H. Bramble and X. Zhang. The analysis of multigrid methods. *Handbook of Numerical Analysis*, 7:173–415, 2000.

[23] J.H. Bramble. *Multigrid Methods*. Pitman Research Notes in Mathematics Series. John Wiley & Sons, 1993.

[24] S. C. Brenner. Multigrid methods for parameter dependent problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 30(3):265–297, 1996.

[25] S. C. Brenner. Convergence of nonconforming multigrid methods without full elliptic regularity. *Mathematics of Computation*, 68(225):25–53, 1999.

[26] S. C. Brenner. Convergence of the multigrid V-cycle algorithm for second-order boundary value problems without full elliptic regularity. *Mathematics of Computation*, 71(238):507–525, 2002.

[27] S. C. Brenner, J. Gedicke, and L.-Y. Sung. C^0 interior penalty methods for an elliptic distributed optimal control problem on nonconvex polygonal domains with pointwise state constraints. *SIAM Journal on Numerical Analysis*, 56(3):1758–1785, 2018.

[28] S. C. Brenner, H. Li, and L.-Y. Sung. Multigrid methods for saddle point problems: Stokes and Lamé systems. *Numerische Mathematik*, 128(2):193–216, 2014.

[29] S. C. Brenner, H. Li, and L.-Y. Sung. Multigrid methods for saddle point problems: Oseen system. *Computers & Mathematics with Applications*, 74(9):2056–2067, 2017.

[30] S. C. Brenner, S. Liu, and L.-Y. Sung. Multigrid methods for saddle point problems: Optimality systems. *Journal of Computational and Applied Mathematics*, 372, 2020.

[31] S. C. Brenner, D.-S. Oh, and L.-Y. Sung. Multigrid methods for saddle point problems: Darcy systems. *Numerische Mathematik*, 138(2):437–471, 2018.

[32] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15. Springer Science & Business Media, 2008.

[33] S. C. Brenner and L.-Y. Sung. A new convergence analysis of finite element methods for elliptic distributed optimal control problems with pointwise state constraints. *SIAM Journal on Control and Optimization*, 55(4):2289–2304, 2017.

[34] S. C Brenner, L.-Y. Sung, and J. Gedicke. P_1 finite element methods for an elliptic optimal control problem with pointwise state constraints. *IMA Journal of Numerical Analysis*, 11 2018.

[35] S. C. Brenner, L.-Y. Sung, H. Zhang, and Y. Zhang. A quadratic C^0 interior penalty method for the displacement obstacle problem of clamped kirchhoff plates. *SIAM Journal on Numerical Analysis*, 50(6):3329–3350, 2012.

[36] S. C. Brenner, L.-Y. Sung, and Y. Zhang. A quadratic C^0 interior penalty method for an elliptic optimal control problem with state constraints. In *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, pages 97–132. Springer, 2014.

[37] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 8(R2):129–151, 1974.

[38] W. L. Briggs, S. F. McCormick, and Hensen V. E. *A Multigrid Tutorial*, volume 72. SIAM, 2000.

[39] L. Caffarelli and A. Friedman. The obstacle problem for the biharmonic operator. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 6(1):151–184, 1979.

[40] L. Caffarelli, A. Friedman, and A. Torelli. The two-obstacle problem for the biharmonic operator. *Pacific Journal of Mathematics*, 103(2):325–335, 1982.

[41] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, 24(6):1309–1318, 1986.

[42] E. Casas, M. Mateos, and B. Vexler. New regularity results and improved error estimates for optimal control problems with state constraints. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(3):803–822, 2014.

[43] P. Ciarlet. The finite element method for elliptic problems. *North Holland-Elsevier Science Publishers, Amsterdam, New York, Oxford*, 19(7):8, 1978.

[44] M. Dauge. Elliptic boundary value problems on corner domains. *Lecture Notes in Mathematics*, 1341:1, 1988.

[45] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM Journal on Numerical Analysis*, 45(5):1937–1953, 2007.

[46] J. W. Demmel. *Applied Numerical Linear Algebra*, volume 56. SIAM, 1997.

[47] M. Engel and M. Griebel. A multigrid method for constrained optimal control problems. *Journal of Computational and Applied Mathematics*, 235(15):4368–4388, 2011.

[48] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, R.I., 2010.

[49] R. P. Fedorenko. A relaxation method for solving elliptic difference equations. *USSR Computational Mathematics and Mathematical Physics*, 1(4):1092–1096, 1962.

[50] J. Frehse. Zum differenzierbarkeitsproblem bei variationsungleichungen höherer ordnung. In *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, volume 36, pages 140–149. Springer, 1971.

[51] J. Frehse. On the regularity of the solution of the biharmonic variational inequality. *Manuscripta Mathematica*, 9(1):91–103, 1973.

[52] G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2013.

[53] C. Gräser and R. Kornhuber. Multigrid methods for obstacle problems. *Journal of Computational Mathematics*, pages 1–44, 2009.

[54] A. Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17. SIAM, 1997.

[55] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 69. SIAM, 1985.

[56] W. Hackbusch. Fast solution of elliptic control problems. *Journal of Optimization Theory and Applications*, 31(4):565–581, 1980.

[57] W. Hackbusch. Survey of convergence proofs for multi-grid iterations. *Special Topics of Applied Mathematics, North-Holland, Amsterdam*, pages 151–164, 1980.

[58] W. Hackbusch. On the convergence of multi-grid iterations. *Beiträge Numer. Math*, 9:213–239, 1981.

[59] W. Hackbusch. *Multi-grid Methods and Applications.* Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1985.

[60] W. Hackbusch and H. D. Mittelmann. On multi-grid methods for variational inequalities. *Numerische Mathematik*, 42(1):65–76, 1983.

[61] W. W. Hager and G. D. Ianculescu. Dual approximations in optimal control. *SIAM Journal on Control and Optimization*, 22(3):423–465, 1984.

[62] M. Heinkenschloss and D. Leykekhman. Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. *SIAM Journal on Numerical Analysis*, 47(6):4607–4638, 2010.

[63] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2002.

[64] M. Hintermüller and K. Kunisch. Stationary optimal control problems with pointwise state constraints. Numerical PDE Constrained Optimization. volume 72 of *Lecture Notes in Computational Science and Engineering.*

[65] R.H.W. Hoppe. Multigrid algorithms for variational inequalities. *SIAM Journal on Numerical Analysis*, 24(5):1046–1065, 1987.

[66] K. Ito and K. Kunisch. Augmented lagrangian formulation of nonsmooth, convex optimization in Hilbert spaces. *Lecture Notes in Pure and Applied Mathematics. Control of Partial Differential Equations and Applications*, 174:107–117, 1995.

[67] K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Applied Mathematics and Optimization*, 41(3):343–364, 2000.

[68] T. Kärkkäinen, K. Kunisch, and P. Tarvainen. Augmented lagrangian active set methods for obstacle problems. *Journal of Optimization Theory and Applications*, 119(3):499–533, 2003.

[69] T. Kärkkäinen and J. Toivanen. Building blocks for odd–even multigrid with applications to reduced systems. *Journal of Computational and Applied Mathematics*, 131(1-2):15–33, 2001.

[70] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities I. *Numerische Mathematik*, 69(2):167–184, 1994.

[71] R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities II. *Numerische Mathematik*, 72(4):481–499, 1996.

[72] R. Kornhuber and H. Yserentant. Multilevel methods for elliptic problems on domains not resolved by the coarse grid. *Contemporary Mathematics*, 180:49–49, 1994.

[73] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 1950.

[74] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer, 1971.

[75] W. Liu, W. Gong, and N. Yan. A new finite element approximation of a state-constrained optimal control problem. *Journal of Computational Mathematics*, pages 97–114, 2009.

[76] J. Mandel, S. McCormick, and R. Bank. Variational multigrid theory. In *Multigrid methods*, pages 131–177. SIAM, 1987.

[77] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[78] V. MazĐ"ya and J. Rossmann. *Elliptic Equations in Polyhedral Domains*. Mathematical Surveys and Monographs, 2010.

[79] S. Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4):575–601, 1992.

[80] C. Meyer. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. *Control and Cybernetics*, 37:51–83, 2008.

[81] N. Meyers and J. Serrin. H=W. In *Proceedings of the National Academies of Science USA*, volume 51, pages 1055–6, 1964.

[82] S. Nazarov and B. A. Plamenevsky. *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, volume 13. Walter de Gruyter, 2011.

[83] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[84] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23. Springer Science & Business Media, 2008.

[85] T. Rees, H. S. Dollar, and A. J. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010.

[86] H. L. Royden and P. Fitzpatrick. *Real Analysis*, volume 32. Macmillan New York, 1988.

[87] W. Rudin. *Real and Complex Analysis*. Tata McGraw-Hill Education, 2006.

[88] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, 1992.

[89] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[90] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[91] J. Schöberl, R. Simon, and W. Zulehner. A robust multigrid method for elliptic optimal control problems. *SIAM Journal on Numerical Analysis*, 49(4):1482–1503, 2011.

[92] J. Schöberl and W. Zulehner. On Schwarz-type smoothers for saddle point problems. *Numerische Mathematik*, 95(2):377–399, 2003.

[93] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3):752–773, 2007.

[94] R. Simon and W. Zulehner. On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems. *Numerische Mathematik*, 111(3):445–468, 2009.

[95] S. Ta'asan. "One-shot" methods for optimal control of distributed parameter systems 1: Finite dimensional control. ICASE-Report 91-2, NASA Langley Research Center, Hampton, VA, 1991.

[96] S. Takacs and W. Zulehner. Convergence analysis of all-at-once multigrid methods for elliptic control problems under partial elliptic regularity. *SIAM Journal on Numerical Analysis*, 51(3):1853–1874, 2013.

[97] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, volume 112. American Mathematical Soc., 2010.

[98] H. A. Van der Vorst. *Iterative Krylov Methods for Large Linear Systems*, volume 13. Cambridge University Press, 2003.

[99] R. Verfürth. A multilevel algorithm for mixed problems. *SIAM Journal on Numerical Analysis*, 21(2):264–271, 1984.

[100] R. Verfürth. Multilevel algorithms for mixed problems. II. treatment of the mini-element. *SIAM Journal on Numerical Analysis*, 25(2):285–293, 1988.

[101] S. W. Walker. FELICITY: A MATLAB/C++ toolbox for developing finite element methods and simulation modeling. *SIAM Journal on Scientific Computing*, 40(2):C234–C257, 2018.

[102] D. S. Watkins. *Fundamentals of Matrix Computations*, volume 64. John Wiley & Sons, 2004.

[103] J. Wloka. *Partial Differential Equations*. Cambridge University, 1987.

[104] S. J. Wright. *Primal-dual Interior-point Methods*, volume 54. SIAM, 1997.

[105] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34(4):581–613, 1992.

[106] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numerische Mathematik*, 94(1):195–202, 2003.

[107] X. Zhang. Multilevel Schwarz methods. *Numerische Mathematik*, 63(1):521–539, 1992.

[108] W. Zulehner. A class of smoothers for saddle point problems. *Computing*, 65(3):227–246, 2000.

Vita

Sijing Liu was born in 1988, in Fuzhou, Fujian, China. He finished his undergraduate studies at Fujian Normal University in June 2011. He earned a master of science degree in computational mathematics from Xiamen University in June 2014. In August 2014 he came to Louisiana State University to pursue graduate studies in mathematics. He is currently a candidate for the degree of Doctor of Philosophy in mathematics, which will be awarded in August 2020.