DISCONTINUOUS GALERKIN METHODS FOR TIME-DEPENDENT CONVECTION DOMINATED OPTIMAL CONTROL PROBLEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TUĞBA AKMAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
SCIENTIFIC COMPUTING

JULY 2011

Approval of the thesis:

DISCONTINUOUS GALERKIN METHODS FOR TIME-DEPENDENT CONVECTION DOMINATED OPTIMAL CONTROL PROBLEMS

submitted by **TUĞBA AKMAN** in partial fulfillment of the requirements for the degree of **Master of Science in Department of Scientific Computing, Middle East Technical University** by,

Prof. Dr. Ersan Akyıldız _____
Director, Graduate School of **Applied Mathematics**

Prof. Dr. Bülent Karasözen _____
Head of Department, **Scientific Computing**

Prof. Dr. Bülent Karasözen _____
Supervisor, **Department of Mathematics, METU**

Examining Committee Members:

Assoc. Prof. Dr. Yusuf Uludağ _____
Department of Chemical Engineering, METU

Prof. Dr. Bülent Karasözen _____
Department of Mathematics & Institute of Applied Mathematics, METU

Assoc. Prof. Dr. Ömür Uğur _____
Institute of Applied Mathematics, METU

Date: _____

* Write the country name for the foreign committee member.

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TUĞBA AKMAN

Signature :

ABSTRACT

DISCONTINUOUS GALERKIN METHODS FOR TIME-DEPENDENT CONVECTION DOMINATED OPTIMAL CONTROL PROBLEMS

Akman, Tuğba

M.S., Department of Scientific Computing

Supervisor : Prof. Dr. Bülent Karasözen

July 2011, 89 pages

Distributed optimal control problems with transient convection dominated diffusion convection reaction equations are considered. The problem is discretized in space by using three types of discontinuous Galerkin (DG) method: symmetric interior penalty Galerkin (SIPG), nonsymmetric interior penalty Galerkin (NIPG), incomplete interior penalty Galerkin (IIPG). For time discretization, Crank-Nicolson and backward Euler methods are used. The discretize-then-optimize approach is used to obtain the finite dimensional problem. For one-dimensional unconstrained problem, Newton-Conjugate Gradient method with Armijo line-search. For two-dimensional control constrained problem, active-set method is applied. A priori error estimates are derived for full discretized optimal control problem. Numerical results for one and two-dimensional distributed optimal control problems for diffusion convection equations with boundary layers confirm the predicted orders derived by a priori error estimates.

Keywords: Transient diffusion convection reaction equation, optimal control, discontinuous Galerkin method, Crank-Nicolson, a priori error estimates

ÖZ

ZAMANA BAĞLI KONVEKSİYON AĞIRLIKLI ENİYİLEMELİ KONTROL PROBLEMLERİNİN KESİNTİLİ GALERKİN YÖNTEMLERİ

Akman, Tuğba
Yüksek Lisans, Bilimsel Hesaplama Bölümü
Tez Yöneticisi : Prof. Dr. Bülent Karasözen

Temmuz 2011, 89 sayfa

Zamana bağlı konveksiyon ağırlıklı konveksiyon-difüzyon-reaksiyon denklemlerin dağıtık eniyileme kontrol problemi ele alındı. Problem uzayda üç farklı sürekli olmayan Galerkin yöntemiyle ayrıklaştırıldı: Simetrik iç ceza Galerkin yöntemi (SIPG), simetrik olmayan iç ceza Galerkin yöntemi (NIPG), eksik iç ceza Galerkin yöntemi (IIPG). Zaman değişkeninin ayrıklaştırılmasında ise Crank-Nicolson ve geriye dönük Euler yöntemi kullanıldı. Sonlu boyutlu problem, ayrıklaştır-eniyile yaklaşımı ile elde edildi. Tek boyutlu kısıtlı olmayan problem, Newton eşlenik gradyan yöntemi ve Armijo doğru arama yöntemi ile çözüldü. İki boyutlu kısıtlı problem için, aktif kümeler yöntemi uygulandı Tam ayrıklaştırılmış eniyileme kontrol problemi için, a priori hata tahminleri elde edildi. Çözümü katmanlar içeren, tek ve iki boyutlu dağıtık adveksiyon-difüzyon-reaksiyon denkleminin eniyileme kontrol problemi için elde edilen sayısal sonuçlar, a priori hata tahminleriyle uyuşmaktadır.

Anahtar Kelimeler: Zamana bağlı konveksiyon-difüzyon problemleri, dağıtık kontrol problemi, süreksiz Galerkin yöntemi, Crank-Nicolson, önceden hata tahminleri

To my family

ACKNOWLEDGMENTS

I would like to express my gratitude to all those people who have helped and supported me during my studies. I am grateful to my supervisor Prof. Dr. Bülent Karasözen who has introduced me the world of applied mathematics and I thank to him for patiently guiding, motivation and encouraging me throughout this study.

I would like to give special thanks to Hamdullah Yücel to whom MATLAB programs for steady diffusion convection reaction equation belong. In addition, I am grateful to Fikriye Yılmaz for her support.

I especially thank to my family who have provided me love, support, encourage and motivation from the first day of my education. Deepest thanks to Burak Yıldız who has always been with me during my tough times. Without them, I could't have been such strong and hopeful.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Optimal control theory has gained importance during the last few decades. Some applications can be observed both in the science and the daily life. Optimal control of ordinary differential equations may be confronted in aviation. In space technology, robotics, movement sequences in sports, and the control of chemical processes and power plants, while the control of partial differential equations are required to investigate heat conduction, diffusion, electromagnetic waves, fluid flows, freezing processes, elastic deformation, option prices, the design of an airplane wing to achieve the optimal performance, control of pollution in a river, wave propagation, elastic deformation and other phenomena [17, 33, 49, 60].

A optimal control problem consists of an objective function, an ordinary/partial differential equation, named the state equation, and control constraints depending on the problem. The state equation establish the relation between the control and the state. The optimal control problem concerns to find the optimal control so that an adequate state close to the observation or target is obtained by minimizing the cost or objective function [49].

During this study, we have considered distributed unconstrained and control constrained optimal control problem governed by time dependent diffusion convection reaction equation. Finite element method(FEM) is a tool to obtain the approximate solution of the partial differential equations by using a variational formulation. The variational formulation is an integral over the time-space domain. From this point of view, it differs from the finite difference scheme. By FEM, the domain is divided into elements, which are usually triangles in 2D and the PDE is approximated on each subdomain [27]. For diffusion convection equations, the problem with small diffusion terms is *singularly perturbed*. This small term has a great influence on the problem that would change the character of the problem. Boundary/interior

layers, which are observed in such problems, results rapid changes of the solution on the boundary/interior with large derivatives. The region where the boundary/interior layers occur cannot be determined a priori, this increases the difficulty in estimating the solution. Singularly perturbed problems can be analyzed by perturbation theory and the regions of the boundary/interior layers [42]. In case of convection dominated problems, the spatial discretization by Galerkin FEM is not efficient unless the mesh size h is not sufficiently small compared to $\varepsilon/\|c\|$. Node-to-node oscillations can be confronted in practice and by mesh and time-step refinement, this can be dealt with. The scale between the diffusion and the convection term results in a constant so called the mesh Peclet number $Pe = \frac{ch}{2\varepsilon}$. It enables us to extract some information about the problem [49]. It can be noted that the large Peclet number results in the non-physical oscillations. For such cases, stabilization techniques have to be used [9, 10, 22, 49].

In 1973, the discontinuous Galerkin (DG) methods proposed for hyperbolic problem by Reed and Hill by [51] and then the solution of hyperbolic and nearly hyperbolic problems by DG became the main concern. In addition, purely elliptic problems have been tried to be solved by DG. For singularly perturbed problems, application do DG to elliptic problems has become very popular [2, 3, 31, 38, 48, 58].

During this study, interior penalty Galerkin methods have been used to perform spatial distribution and we have used the definitions and theorem given [53]. DGFEMs results in high-order and stable solutions spite of the boundary or interior layers, and discontinuous parts of the solution [12]. DG methods are highly preferable due to their locally conservative, stable, and high-order accurate nature. By DG, irregular meshes, complex geometries can be handled and basis polynomials of several degrees can be used. However, the degrees of freedom is increased. Thus, many problems of fluid dynamics and Hamilton-Jacobi equations, second-order elliptic problems, elasticity are the application areas of the method although the latter ones is not directly to be the purpose of the action [13].

There are two different approaches to solve the optimal control problems: discretize-then-optimize and optimize-then discretize. We have used the former one to obtain an approximate solution. For spatial discretization, interior penalty Galerkin methods, SIPG, NIPG and IIPG have been used, while temporal discretization has been performed by backward Euler and Crank-Nicolson methods.

In contrast to the continuous FEM, DG does not insert a continuity requirement between the neighboring elements while the solution is approximated by piecewise polynomials on the mesh. Similar to the finite volume method, a numerical flux is used to obtain the discontinuous approximations and boundary fluxes. Discontinuities and steep gradients can be caught. As a common property, it is necessary to add the fact that higher degree polynomial approximations of the sought solution is used to increase the accuracy of the solution [20].

The outline of the thesis is as follows: In the next Chapter, we provide the optimal control problems governed by steady and unsteady diffusion convection reaction equation, discuss the existence and the uniqueness of the solution. Then, we state the optimality system for each of the problems. In addition, steady and unsteady diffusion convection reaction equations are discussed by underlining the necessary conditions for the existence and the uniqueness of the solution. Chapter 3 is devoted to discontinuous Galerkin method by which the space variable is discretized. Some definitions, properties and DG (bi)linear forms are introduced. In Chapter 4, two approaches, optimize-then-discretize and discretize-then-optimize are discussed. Then, we proceed by temporal discretization by backward Euler and Crank-Nicolson methods. Optimization methods used during this study is described and some remarks related to the implementation of the model problems are given in Chapter 5. In Chapter 6, we discuss the consistency of the DG method. Then, we provide stability and convergence estimates for semidiscrete state equation. In addition, these estimates are proved for full-discrete state, adjoint and the control, separately. Chapter 7 is devoted to numerical results. We have obtained approximate solution to distributed optimal control problems: One-dimensional unconstrained optimal control problem and two-dimensional control constrained optimal control problem. Then, we have calculated the numerical order for both of the problems which confirm theoretically obtained a priori error estimates.

CHAPTER 2

OPTIMAL CONTROL PROBLEM

Optimal control problems arises in science, technology and daily life. While the problem of obtaining the optimal trajectory of an aircraft is an example of the optimal control, the problem of roasting a potato up to a pleasant temperature is an optimal control problem, too [60]. Let us introduce the basic elements of an optimal control problem [23]: A control u of which minimum is the solution of the problem and it satisfies the constraints of the problem. A state equation by which the relation between the state y and the control u is constructed. Its solution uniquely determined. The state of the system y is the solution of the state equation and depends on the control. Thus, any change in the control causes a change in the state. Lastly, an objective function is needed and it depends on the state and control variables. The aim of the optimal control problem is to obtain an admissible control so that a desired state can be obtained. At the same time, the value of the objective function is minimized. There are different types of optimal control problems such as distributed control, boundary control and Dirichlet control problem. For the first one, the control is looked for on the whole domain, while the control acts on the boundary for the latter one. Dirichlet boundary control problems are not of variational type. In case of a steady PDE, if the control space is H^2 for $s \geq 1/2$, very weak form of the state equation must be considered [36, 41]. We present general form of the optimal control problems governed by steady and unsteady PDE in this Chapter.

It is beneficial to mention the definitions of the vector spaces that we have used during this study. The vector space $L^2(\Omega)$ is the space of square-integrable functions on $\Omega \subset \mathbb{R}^n$:

$$L^2(\Omega) = \{f : \Omega \longmapsto \mathbb{R} \quad \text{s.t.} \quad \int_\Omega (f(x))^2 d\Omega \leq +\infty\}.$$

Indeed, $L^2(\Omega)$ is a space of equivalence of measurable functions. Indeed, $L^2(\Omega)$ is a Hilbert

space with respect to the following inner product and norm:

$$(u, v)_\Omega = \int_\Omega uv, \qquad \|v\|_{L^2(\Omega)} = \left(\int_\Omega v^2\right)^{1/2}.$$

The space $L^\infty(\Omega)$ is the space of bounded functions:

$$L^\infty(\Omega) = \{v : \|v\|_{L^\infty(\Omega)} < \infty\},$$

with

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

We introduce the Sobolev space

$$H^1(\Omega) = \left\{v \in L^2(\Omega) : \frac{\partial v}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, d\right\}.$$

Similarly, we denote $H^s(\Omega)$ for integer s:

$$H^s(\Omega) = \{v \in L^2(\Omega) : \forall 0 \le |\alpha| \le s, D^\alpha v \in L^2(\Omega)\}.$$

The Sobolev norm associated with $H^s(\Omega)$ is

$$\|v\|_{H^s(\Omega)} = \left(\sum_{0 \le |\alpha| \le s} \|D^\alpha v\|^2_{L^2(\Omega)}\right)^{1/2}.$$

The Sobolev seminorm associated with $H^s(\Omega)$ is

$$|v|_{H^s(\Omega)} = \|\nabla^s v\|_{L^2(\Omega)} = \left(\sum_{|\alpha| = s} \|D^\alpha v\|^2_{L^2(\Omega)}\right)^{1/2}.$$

We now introduce the space

$$W(0, T) = \left\{f | f \in L^2(0, T; V), \frac{df}{dt} \in L^2(0, T; V')\right\}.$$

We define the Sobolev spaces with fractional indices. By an interpolation between $H^s(\Omega)$ and $H^{s+1}(\Omega)$, we obtain the space $H^{s+1/2}(\Omega)$ with s integer. In [53], the K-interpolation is given as follows: Given $v \in H^s(\Omega)$, the following splitting is defined:

$$v = v_1 + v_2,$$

for $v_1 \in H^s(\Omega)$ and $v_2 \in H^{s+1}(\Omega)$. Then, for a given $t \in \mathbb{R}$, the kernel is defined as

$$K(v, t) = \left(\inf_{v_1 + v_2} = v(\|v_1\|^2_{H^s(\Omega)} + t^2\|v_2\|^2_{H^{s+1}(\Omega)})\right)^{1/2}.$$

The space $H^{s+1/2}(\Omega)$ is defined as the completion of all functions in $H^{s+1}(\Omega)$ with respect to the norm:

$$\|v\|_{H^{s+1/2}(\Omega)} = \left(\int_0^\infty t^{-2} K^2(v, t) dt\right)^{1/2}.$$

Indeed, $H^{s+1}(\Omega) \subset H^{s+1/2}(\Omega) \subset H^s(\Omega)$.

2.1 Diffusion-Convection-Reaction Equation

Diffusion convection reaction equations can be used to model so many physical problems related to the transport of air, flow in oil reservoir, ground water pollutants, air pollution, heat dissipation [4, 18, 25, 26, 50, 49, 61]. The problematic nature of this problem arises from the multiscale between the diffusion and convection term. In many practical applications, diffusion term ε is too small compared to convection term. Then, the problem is called a singularly perturbed equation. Although the perturbation is too small, the nature of the problem changes completely and boundary layers, which are rapid changes of the solution close to the boundary, are observed. It becomes harder to obtain stable solutions and more grid points are required to resolve the boundary layer [42]. Apart from the boundary layers, interior layers where a rapid change is observed interior of the domain. To overcome this difficulty, perturbation theory can be handled singularly perturbed problems and it facilitates the determination of the place and width of the layers [42]. Now, let us give some examples of steady and unsteady diffusion convection reaction equations.

2.1.1 Steady Diffusion-Convection-Reaction Equation

Consider the steady diffusion convection reaction equation with Dirichlet boundary condition [49])

$$
\begin{aligned}
-\nabla \cdot (\varepsilon \nabla y(x)) + c(x) \cdot \nabla y(x) + r(x) y(x) &= f(x) \quad x \in \Omega, && (2.1)\\
y(x) &= g_D \quad x \in \partial\Omega. && (2.2)
\end{aligned}
$$

where ε, r, c, f and c are given functions. In general, it is assumed that $\varepsilon > 0$, $r \in L^{\infty}(\Omega)$ and $r \geq 0$, $c \in (W(^{1,\infty}(\Omega)))^n$ and $f \in L^2(\Omega)$. The weak form of the problem can be stated as follows:

$$
\begin{aligned}
\text{Find} \quad y \in V, \quad a(y, \upsilon) &= F(\upsilon), \quad \forall \upsilon \in V, && (2.3)\\
\text{where} \quad a(y, \upsilon) &= \int_{\Omega} \varepsilon \nabla y \cdot \nabla \upsilon + c \cdot \nabla y \upsilon + r y \upsilon dx, \quad \forall y, \upsilon \in V, && (2.4)\\
(f, \upsilon) &= \int_{\Omega} f \upsilon dx, \quad \forall \upsilon \in V. && (2.5)
\end{aligned}
$$

2.1.1.1 Existence and Uniqueness of The Solution

We have to guarantee the existence of the solution before trying to find it. Thus, the conditions of the Lax Milgram lemma must be satisfied. Let me mention the Lax Milgram lemma.

Lemma 2.1.1 *Assume that V is a Hilbert space, $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ is a continuous and coercive bilinear form, $F(\cdot) : V \longrightarrow \mathbb{R}$ a linear and continuous functional. Then, the following problem admits a unique solution*

$$Find \quad y \in V, \quad a(y,v) = F(v), \quad \forall v \in V.$$

By [49], for $a(\cdot,\cdot)$ to be coercive,

$$-\frac{1}{2}div(c) + r \geq 0, a.e. \in \Omega,$$

with the coercivity constant $C = \frac{\varepsilon_0}{1+\tilde{C}}$ where $\tilde{C}$ is the constant coming from *Poincaré* inequality applied to determine a bound for $\|v\|_{L^2(\Omega)}$. In addition, $a(\cdot,\cdot)$ is continuous with the constant $C = \|\varepsilon\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)} + \|r\|_{L^2(\Omega)}$.

2.1.2 Unsteady Diffusion-Convection-Reaction Equation

The convection dominated diffusion convection equation gain great importance. Due to the multiscale between the diffusion and the convection term, to obtain an accurate and effective numerical approximation becomes a difficult process [39]. Thus, the numerical methods such as finite difference or finite element method would result in oscillations that are not observed in the exact solution. Let us mention the general unsteady diffusion convection reaction equation

$$\frac{\partial y}{\partial t} - \nabla \cdot (\varepsilon \nabla y(x,t)) + c(x) \cdot \nabla y(x,t) + r(x)y(x,t) = f(x,t) \quad (x,t) \in \Omega \times (0,T], \tag{2.6}$$

$$y(x,t) = g_D \quad (x,t) \in \partial\Omega \times [0,T], \tag{2.7}$$

$$y(x,0) = y_0 \quad x \in \Omega, \tag{2.8}$$

where $f \in L^2(0,T;L^2(\Omega))$, $g_D \in L^2(0,T;H^{\frac{1}{2}}(\partial\Omega))$, $y_0 \in L^2(\Omega)$.

Let me give some applications related to the diffusion convection reaction equation. The

general form of the problem can be written in terms of temperature as

$$\begin{aligned} \frac{\partial T}{\partial t} - \nabla \cdot (\varepsilon \cdot \nabla T) + c \cdot \nabla T = S, \quad \text{in} \quad Q_T, \\ T = T_D \quad \text{on} \quad (0,T) \times \partial\Omega^{in}, \\ T(x,0) = T_0(x), \quad \text{on}\Omega, \quad t = 0. \end{aligned}$$

The term $c \cdot \nabla T$ corresponds to convection. The temperature is transferred by the velocity field c. Accumulation for the non-steady processes is represented by the term $\frac{\partial T}{\partial t}$. The diffusion term is related to $\nabla \cdot (\varepsilon \cdot \nabla T)$. In general, ε can be a full (but symmetric and positive definite) second order tensor. Indeed, ε can be a diagonal matrix or a scalar [28]. By [50], temporal changes of the temperature and the speed of the propagation can be modeled the diffusion convection reaction equation, too. By the unsteady heat equation, temporal changes of the temperature y of an isotropic and homogenous instrument in $\bar{\Omega}$ under the heat source f is modeled. Firstly, let us consider the following heat equation

$$\begin{aligned} \frac{\partial y}{\partial t} - \Delta y &= f \quad \text{in} \quad Q_T, \\ u &= 0 \quad \text{on} \quad (0,T) \times \partial\Omega, \\ u &= u_0, \quad \text{on} \quad \Omega, \quad t = 0. \end{aligned}$$

Secondly, we consider the speed of propagation denoted by y at any point be a. By the linear transport equation, the transport of q quantity y is modelled.

$$\begin{aligned} \frac{\partial y}{\partial t} + a \cdot \nabla y &= f \quad \text{in} \quad Q_T, \\ u &= l \quad \text{on} \quad (0,T) \times \partial\Omega, \\ u &= u_0, \quad \text{on} \quad \Omega, \quad t = 0. \end{aligned}$$

2.1.2.1 Existence and Uniqueness of The Solution

Let us consider the unsteady diffusion convection reaction equation that we have mentioned. The weak form of the problem can be written as Find $y \in L^2(0,T;L^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$,

$$\begin{aligned} \left(\frac{\partial y}{\partial t}, \upsilon\right) + a(y,\upsilon) &= F(\upsilon), \quad \forall t > 0, \quad \forall \upsilon \in V, \\ (y(0),\upsilon) &= (y_0,\upsilon), \quad \forall \upsilon \in V. \end{aligned}$$

To guarantee the existence and uniqueness of the solution, the following sufficient condition must hold [49]. That is, the bilinear form $a(\cdot,\cdot)$ must be continuous and weakly coercive:

$$\text{For} \quad \lambda \geq 0, \quad \alpha > 0, \quad a(\upsilon,\upsilon) + \lambda \|\upsilon\|^2_{L^2(\omega)} \geq \alpha \|\upsilon\|^2_V, \quad \forall \upsilon \in V.$$

2.2 Optimal Control Problem For Steady Diffusion-Convection-Reaction Equation

The distributed unconstrained optimal control problem governed by steady diffusion convection equation is as follows [31]:

$$\min J(y,u) := \frac{1}{2}\|y-\hat{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \tag{2.9a}$$

$$\text{subject to} \quad -\nabla\cdot(\varepsilon\nabla y(x)) + c(x)\cdot\nabla y(x) + r(x)y(x) = f(x)+u(x) \quad x\in\Omega, \tag{2.9b}$$

$$y(x) = g_D \quad x\in\partial\Omega. \tag{2.9c}$$

It is a generalized version of the one given for the heat equation at [60].

The weak form of the state equation (2.9c) is given by

$$a(y,\upsilon) + b(u,\upsilon) \quad = \quad (f,\upsilon), \quad \forall t>0, \quad \forall\upsilon\in V = H^1_0(\Omega) \tag{2.10}$$

with

$$a(y,\upsilon) = \int_\Omega \varepsilon\nabla y\cdot\nabla\upsilon + c(x)\cdot\nabla y\upsilon + r(x)y\upsilon dx, \tag{2.11a}$$

$$b(u,\upsilon) = -\int_\Omega u\upsilon dx, \tag{2.11b}$$

$$(f,\upsilon) = \int_\Omega f\upsilon dx, \quad \forall\upsilon\in V. \tag{2.11c}$$

We are interested in the solution of the optimal control problem in variational form

$$\min J(y,u) := \frac{1}{2}\|y-\hat{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \tag{2.12a}$$

$$\text{subject to} \quad a(y,\upsilon) + b(u,\upsilon) = (f,\upsilon), \quad \forall\upsilon\in V, \tag{2.12b}$$

$$y\in Y, u\in U. \tag{2.12c}$$

where $y\in Y = H^1_0(\Omega)$, $u\in U = L^2(\Omega)$, $V = \{\upsilon\in H^1(\Omega) : \upsilon = 0 \quad \text{on} \quad \Gamma_D\}$.
Suppose that Ω is a bounded domain, $\alpha > 0$, $\varepsilon > 0$, $c(x)\in(W^{1,\infty}(\Omega))^n$, $r\in L^\infty(\Omega)$, f, $\hat{y}\in L^2(\Omega)$, $g_D\in H^{\frac{3}{2}}(\partial\Omega)$, $r(x)-\frac{1}{2}\nabla\cdot c(x)\geq r_0\geq 0 \quad a.e. \quad \text{in} \quad \Omega$. With these assumptions, the bilinear form $a(\cdot,\cdot)$ is continuous and coercive. In addition, the vector c is a given divergence-free velocity field, that is,

$$\nabla\cdot(cy) = (\nabla\cdot c)y + c\cdot\nabla y = c\cdot\nabla y.$$

2.3 Optimal Control Problem For Unsteady Diffusion-Convection-Reaction Equation

In the literature, there are many examples and ways to solve the optimal control problems governed by steady diffusion convection equation. For example, at [19], optimal control problem with linear advection-diffusion equation is solved by using a stabilization method. Instead of stabilizing the state and the adjoint separately, a stabilization method is applied to the Lagrangian. As a different point of view, an optimal control problem governed by an elliptic PDE can be viewed as a parameter estimation problem in case that the control can enter the state equation as a variable [40]. In addition, the edge stabilization Galerkin method is used to solve the state equation of the optimal control problem [35]. In case of continuous Galerkin method, the stabilization techniques are suggested [6, 16]. One can find some solution suggestions for the optimal control problems governed by parabolic PDEs, such as [5, 25, 26]. While continuous finite element discretization are preferable for much of the studies, studies by DG methods are not very common, although it is known that they are more suitable for these kind of problem.

The problem we are interested in is the optimal control problem governed by the unsteady convection diffusion equations. There have been extensive theoretical and numerical studies for the finite element approximation of various optimal control problems [7, 8, 26, 32, 35, 44]. We discuss the discontinuous Galerkin finite element(DGFE) approximation of optimal control problem governed by convection-diffusion equations.

Firstly, we consider the unconstrained distributed linear-quadratic optimal control problem governed by the time dependent diffusion convection reaction equation, formally defined by

$$\min J(y,u) := \frac{1}{2}\int_0^T \|y-\hat{y}\|^2_{L^2(\Omega)}dt + \frac{\alpha}{2}\int_0^T \|u\|^2_{L^2(\Omega)}dt \tag{2.13}$$

subject to

$$\frac{\partial y}{\partial t} - \nabla\cdot(\varepsilon\nabla y(x,t)) + c(x)\cdot\nabla y(x,t) + r(x)y(x,t) = f(x,t)+u(x,t) \quad (x,t)\in\Omega\times(0,T], \tag{2.14a}$$

$$y(x,t) = g_D \quad (x,t)\in\partial\Omega\times[0,T], \tag{2.14b}$$

$$y(x,0) = y_0 \quad x\in\Omega. \tag{2.14c}$$

We define

$$\Omega = (0,1)\times(0,1), \quad Q = (0,T]\times\Omega \quad \text{and} \quad \Sigma = [0,T]\times\partial\Omega \quad \text{for} \quad T > 0 \quad \text{and} \quad \text{it is fixed.}$$

We assume that

$$\alpha > 0, \quad \varepsilon > 0, \quad c(x) = (c_1(x), c_2(x)) \in C(0,T; C_0^1(\bar{\Omega})^2)), \quad r(x) - \frac{1}{2}\nabla\cdot c(x) \geq r_0 \geq 0 \quad a.e. \quad \text{in} \quad \Omega$$
$$f \in L^2(0,T; L^2(\Omega)), \quad \hat{y} \in H^1(0,T; L^2(\Omega)), \quad y_0 \in H_0^1(\Omega), \quad g_D \in L^2(0,T; H^{\frac{1}{2}}(\partial\Omega)),$$
$$y \in Y = L^2(0,T;V), \quad u \in U = L^2(0,T; L^2(\Omega)), \quad V = \{\upsilon \in H^1(\Omega) : \upsilon = 0 \quad \text{on} \quad \Gamma_D\}.$$

The vector c is a given divergence-free velocity field, that is,

$$\nabla\cdot(cy) = (\nabla\cdot c)y + c\cdot\nabla y = c\cdot\nabla y.$$

The weak form of the state equation (2.14) is given by

$$\left(\frac{\partial y}{\partial t}, \upsilon\right) + a(y,\upsilon) + b(u,\upsilon) = (f,\upsilon), \quad \forall t > 0, \quad \forall \upsilon \in V = H_0^1(\Omega) \tag{2.15}$$
$$(y(0),\upsilon) = (y_0,\upsilon), \quad \forall \upsilon \in H_1^0(\Omega), \tag{2.16}$$

with

$$\left(\frac{\partial y}{\partial t}, \upsilon\right) = \int_\Omega \frac{\partial y}{\partial t}\upsilon dx, \tag{2.17a}$$
$$a(y,\upsilon) = \int_\Omega \varepsilon\nabla y\cdot\nabla\upsilon + c(x)\cdot\nabla y\upsilon + r(x)y\upsilon dx, \tag{2.17b}$$
$$b(u,\upsilon) = -\int_\Omega u\upsilon dx, \tag{2.17c}$$
$$(f,\upsilon) = \int_\Omega f\upsilon dx, \quad \forall \upsilon \in V. \tag{2.17d}$$

We are interested in the solution of the optimal control problem in variational form

$$\min J(y,u) := \frac{1}{2}\int_0^T \|y - \hat{y}\|^2_{L^2(\Omega)} dt + \frac{\alpha}{2}\int_0^T \|u\|^2_{L^2(\Omega)} dt \tag{2.18a}$$
$$\text{subject to} \quad \left(\frac{\partial y}{\partial t}, \upsilon\right) + a(y,\upsilon) + b(u,\upsilon) = (f,\upsilon), \quad \forall \upsilon \in V, \quad t \in (0,T], \tag{2.18b}$$
$$(y(0),\upsilon) = (y_0,\upsilon), \quad \forall \upsilon \in V. \tag{2.18c}$$

2.4 Existence and Uniqueness of The Solution of The Optimal Control Problem

Let me define $\mathcal{U}_{ad} = U = L^2(0,T;L^2(\Omega))$. Consider the operator $\mathcal{B} \in \mathcal{L}(U; L^2(0,T;V'))$. By $y(\upsilon)$, we denote a solution of

$$\frac{dy(\upsilon)}{dt} + \mathcal{A}(t)y(\upsilon) = f + B\upsilon, \tag{2.19a}$$

$$y(\upsilon)_{|_{t=0}} = y_0, \tag{2.19b}$$

$$y(\upsilon) \in L^2(0,T;V). \tag{2.19c}$$

Here, $y(\upsilon)$ is a function $t \longrightarrow y(\upsilon)(t)$. For simplicity, write it as $y(t;\upsilon)$. Hence,(2.19b) corresponds to $y(0;\upsilon) = y_0$. The state of the system is $y(\upsilon)$. The cost functional is given by

$$J(v) = \|Cy(\upsilon) - z_d\|_H^2 + (\mathcal{N}\upsilon, v)_U.$$

where the observation $z(\upsilon) = Cy(\upsilon)$, $\quad C \in \mathcal{L}(w(0,T);\mathcal{H})$, $\quad \mathcal{N} \in \mathcal{L}(U,U)$, $\quad (\mathcal{N}u,u)_U \geq \nu\|u\|_U^2$, $\quad \nu > 0.$, $\mathcal{U}_{ad}$ is a closed, convex subset of U and we are interested in

$$\mathrm{Inf}_{\upsilon\in\mathcal{U}_{ad}} J(\upsilon).$$

Theorem 2.4.1 *[43] Assume that $a(\cdot,\cdot)$ is coercive and $(\mathcal{N}u,u)_U \geq \nu\|u\|_U^2, for \quad \nu > 0$ is satisfied. Let me write the state of the system as $\mathcal{A}y(u) = f + \mathcal{B}u, \quad y(u) \in V$. Then, there exist a unique element $u \in \mathcal{U}_{ad}$ such that $J(u) = inf_{\upsilon\in\mathcal{U}_{ad}} J(\upsilon)$.*

According to [43], this theorem can be applied to unsteady PDE constraint, too. For our case, we are interested in the continuity of the affine map $\upsilon \longmapsto y(\upsilon)$ of $u \longrightarrow W(0,T)$. Hence, there exist a unique control. Indeed, incase of $\mathcal{N} = 0$ and $\mathcal{U}_{ad}$ be bounded, there exist a non-empty, closed, convex set consisting of optimal controls [43].

2.5 Optimality System For Unconstrained Problems

Let me rewrite the optimization problem as

$$\begin{aligned} &\min \quad J(y,u) \quad \text{over} \quad (y,u) \in Y \times U, \\ &\text{subject} \quad \text{to} \quad e(y,u) = 0. \end{aligned}$$

Now, we have denoted the PDE constraint $e(y,u) = 0$ in a weak from. By [55], the spaces where we seek the state and control variables y and u of the problem are some Banach spaces Y and U or often even Hilbert spaces. From the optimal control point of view, in case of the distributed control, Y consists of functions defined on Ω or a part thereof positive measure and in case of the boundary control, a space of functions defined on the boundary $\partial\Omega$ or a part thereof (boundary control) are considered. A solution operator for $e(y,u) = 0$ is

$$U \ni u \longmapsto S(u) \in Y.$$

We have a reduced problem

$$\min \quad J(S(u),u) \quad \text{where} \quad u \in U.$$

In order to obtain the Lagrangian, we mention the original problem formulation. Denote a Lagrange multiplier by p. It is named as the adjoint state which we need to determine during the optimization part. We define the Lagrangian for the problem above as

$$L(y,u,p) = f(y,u) + \langle e(y,u), p\rangle_{Y^*,Y}. \tag{2.20}$$

By setting the partial *Fréchet*-derivatives of (2.20) with respect to the unknowns state y, control u and adjoint p and equating to zero, we determine the necessary, and for our problem, sufficient optimality conditions. Now, we have the following adjoint, gradient and the state equations, respectively:

$$L_y(y,u,p)(\delta y) = f_y(y,u)\delta y + \langle e_y(y,u)\delta y, p\rangle = 0, \quad \forall \delta y \in Y, \tag{2.21a}$$

$$L_u(y,u,p)(\delta u) = f_u(y,u)\delta u + \langle e_u(y,u)\delta y, p\rangle = 0, \quad \forall \delta u \in U, \tag{2.21b}$$

$$L_p(y,u,p)(\delta \mu) = \langle e(y,u), \delta p\rangle = 0, \quad \forall \delta p \in Y. \tag{2.21c}$$

As in [55], the optimality system corresponds to

$$f_y(y,u) + e_y(y,u)^* p = 0, \tag{2.22}$$

$$f_u(y,u) + e_u(y,u)^* p = 0, \tag{2.23}$$

$$e(y,u) = 0. \tag{2.24}$$

Optimality system for the optimal control problem subject to the steady diffusion-convection-reaction equation is as follows [55, 30]

$$-\varepsilon\Delta y(x) + c(x)\cdot\nabla y(x) + r(x)y(x) = f(x) + u(x) \quad \text{in} \quad \Omega, \tag{2.25a}$$

$$y(x) = g_D \quad \text{on} \quad \partial\Omega, \tag{2.25b}$$

$$-\varepsilon\Delta p(x) - c(x)\cdot\nabla p(x) + (r(x) - \nabla\cdot c(x))p(x) = -(y-\hat{y}) \quad \text{in} \quad \Omega, \tag{2.26a}$$

$$p(x) = 0 \quad \text{on} \quad \Sigma, \tag{2.26b}$$

$$p(x) = \alpha u(x), \quad \text{in} \quad \Omega. \tag{2.27a}$$

Optimality system for the optimal control problem subject to the unsteady diffusion-convection-reaction equation is as follows [26, 55]

$$\frac{\partial y}{\partial t} - \varepsilon\Delta y(x,t) + c(x)\cdot\nabla y(x,t) + r(x)y(x,t) = f(x,t) + u(x,t) \quad \text{in} \quad Q, \tag{2.28a}$$

$$y(x,t) = g_D \quad \text{on} \quad \Sigma, \tag{2.28b}$$

$$y(\cdot,0) = y_0 \quad \text{in} \quad \Omega. \tag{2.28c}$$

$$-\frac{\partial p}{\partial t} - \varepsilon\Delta p(x,t) - c(x)\cdot\nabla p(x,t) + (r(x) - \nabla\cdot c(x))p(x,t) = -(y-\hat{y}) \quad \text{in} \quad Q, \tag{2.29a}$$

$$p(x,t) = 0 \quad \text{on} \quad \Sigma, \tag{2.29b}$$

$$p(\cdot,T) = 0 \quad \text{in} \quad \Omega. \tag{2.29c}$$

$$\alpha u = p, \quad \text{in} \quad Q. \tag{2.30a}$$

The adjoint equations (2.26) and (2.29) are also diffusion convection equations with the convection term $-c$.

2.6 Optimality System For Constrained Problems

Let me rewrite the constrained optimal control problem governed by steady PDE as follows:

$$\min \quad J(y,u) \quad \text{over} \quad (y,u) \in Y \times U, \tag{2.31}$$

$$\text{subject to} \quad e(y,u) = 0, \tag{2.32}$$

$$u_a \le u \le u_b \quad \text{on} \quad \Omega. \tag{2.33}$$

For pointwise control constraints, the set of admissible control constraints can be written as

$$U_{ad} = \{u \in L^2(\Omega) : u_a \leq u \leq u_b \quad \text{a.e.} \quad \text{on} \quad \Omega\}.$$

Here, the only difference from the unconstrained problem arise from the variational inequality. This can be state in two different ways [55]: As the first choice,

$$(\alpha u - p, \upsilon - u) \geq 0 \quad \forall \upsilon \in U_{ad}.$$

As the second choice, we define the additional Lagrange multipliers ξ^a, ξ^b for the inequality constraints and the complementarity conditions. ξ^a, ξ^b corresponds to the Lagrange multipliers for the control constraints $u_a - u \leq 0$ and $u - u_b \leq 0$ for the local optimal solution. Then we have,

$$\xi^a \geq 0, \qquad u_a - u \leq 0, \qquad \xi^a(u - u_a) = 0, \tag{2.34}$$

$$\xi^b \geq 0, \qquad u - u_b \leq 0, \qquad \xi^b(u_b - u) = 0. \tag{2.35}$$

It can be noted that the variational inequality of the control constrained problem can also be written as

$$u = \mathbb{P}_{[u_a(x), u_b(x)]} \left\{ \frac{1}{\alpha} p \right\}.$$

Let me rewrite the constrained optimal control problem governed by unsteady PDE as follows:

$$\min \quad J(y, u) \quad \text{over} \quad (y, u) \in Y \times U, \tag{2.36}$$

$$\text{subject to} \qquad e(y, u) = 0, \tag{2.37}$$

$$u_a \leq u \leq u_b \quad \text{on} \quad Q. \tag{2.38}$$

The only difference from the steady case arises from the time variable. Thus, U_{ad} must be defined differently. For pointwise control constraints, the set of admissible control constraints can be written as

$$U_{ad} = \{u \in L^2(0, T; L^2(\Omega)) : u_a \leq u \leq u_b \quad \text{a.e.} \quad \text{on} \quad \Omega \times (0, T]\}.$$

Here, the only difference from the unconstrained problem arise from the variational inequality. This can be state in two different ways [55]: As the first choice, the variational inequality is written as a projection

$$\int_0^T (\alpha u - p, \upsilon - u) dt \geq 0 \quad \forall \upsilon \in U_{ad}.$$

As the second choice, we define the additional Lagrange multipliers ξ^a, ξ^b for the inequality constraints and the complementarity conditions. ξ^a, ξ^b corresponds to the Lagrange multipliers for the control constraints $u_a - u \leq 0$ and $u - u_b \leq 0$ for the local optimal solution. Then we have,

$$\xi^a \geq 0, \qquad u_a - u \leq 0, \qquad \xi^a(u - u_a) = 0, \tag{2.39}$$

$$\xi^b \geq 0, \qquad u - u_b \leq 0, \qquad \xi^b(u_b - u) = 0. \tag{2.40}$$

It can be noted that the variational inequality of the control constrained problem can also be written as

$$u = \mathbb{P}_{[u_a(x,t),u_b(x,t)]}\left\{\frac{1}{\alpha}p\right\}.$$

CHAPTER 3

DISCONTINUOUS GALERKIN METHODS

Discontinuous Galerkin(DG) methods was introduced by Reed and Hill [51] in 1973 to solve the hyperbolic problems. Then, the methods has been applied to hyperbolic, nearly hyperbolic, elliptic and parabolic problems [3]. DG methods are highly preferable because stabilization techniques are not needed. Different degrees of polynomials can be used on different elements and they are mass-conservative and it is allowed to use non-conforming or unstructured meshes [48, 53, 54]. Up to now, different variants of DG has been improved, such as Runga-Kutta discontinuous Galerkin (RKDG), local discontinuous Galerkin (LDG), compact discontinuous Galerkin (CDG) and interior penalty discontinuous Galerkin. Nonlinear diffusion-convection equation are solved by RKDG in [15] by conducting spatial discretization by DG and temporal discretization by Runge-Kutta method. The development of RKDG method is discussed, too. As an extension of RKDG, LDG method is introduced in [14] for nonlinear convection diffusion systems and common properties with RKDG and advantages of LDG are provided. A variation of LDG, CG, is discussed in [48] in order to eliminate the distant connections between the nonneighboring elements for multiple dimension. Apart from these, some variations of DG can be found in the literature [3]. During the study, we have used the interior penalty DG methods such as NIPG, SIPG and IIPG to discretize the diffusion part and the upwind method has been used to discretize the convection term. We have used the definitions and properties given in [53].

3.1 1-D Discontinuous Galerkin Methods

3.1.1 Model Problem

Consider the one-dimensional steady diffusion-convection-reaction equation

$$-\varepsilon y''(x) + cy'(x) + ry(x) = f(x) \qquad \text{in} \quad \Omega = (0,1), \tag{3.1a}$$

$$y(0) = y_0, \tag{3.1b}$$

$$y(1) = y_1. \tag{3.1c}$$

where $f \in C^0(0,1)$. Suppose that μ lies between two constants μ_0 and μ_1.

If $y \in C^2(0,1)$ and (3.1a-3.1c) is satisfied pointwisely, then y is a solution of (3.1a-3.1c).

3.1.2 DG Scheme

Let $(0,1)$ be divided into N subintervals as $0 = x_0 < x_1 < \cdots < x_n = 1$. Let me denote each partition by ξ_h, each subinterval by $I_n = (x_n, x_{n+1})$ and the length of these subintervals by

$$h_n = x_{n+1} - x_n, \quad h_{n-1,n} = \max(h_{n-1}, h_n), \quad h = \max_{0 \le n \le N-1} h_n \ .$$

The space of piecewise discontinuous polynomials of degree k is

$$D_k(\xi_h) = \{\upsilon : \upsilon|_{I_n} \in P_k(I_n) \quad \forall j = 0, \ldots, N-1\}.$$

Here, $P_k(I_n)$ corresponds to the space of polynomials of degree k on the interval I_n.

The jump and average of υ can be defined for the endpoints of the subintervals as

$$[\upsilon(x_n)] = \upsilon(x_n^-) - \upsilon(x_n^+), \qquad \{\upsilon(x_n)\} = \frac{1}{2}(\upsilon(x_n^-) + \upsilon(x_n^+)) \quad \forall n = 1, \ldots, N-1,$$

where $\upsilon(x_n^+) = \lim_{\epsilon \to 0,\, \epsilon > 0} \upsilon(x_n + \epsilon)$ and $\upsilon(x_n^-) = \lim_{\epsilon \to 0,\, \epsilon > 0} \upsilon(x_n - \epsilon)$.

These definitions can be extended to the end points of $(0,1)$ as

$$[\upsilon(x_0)] = -\upsilon(x_0^+), \qquad \{\upsilon(x_0)\} = \upsilon(x_0^+), \qquad [\upsilon(x_N)] = \upsilon(x_N^-), \qquad \{\upsilon(x_N)\} = \upsilon(x_N^-).$$

Consider any υ in $D_k(\xi_h)$. (3.1a) is multiplied by υ and the integration by parts is performed on each interval I_n:

$$\begin{aligned}
&\int_{x_n}^{x_{n+1}} \varepsilon y^{'}(x)\upsilon^{'}(x)dx \quad - \quad \varepsilon y^{'}(x_{n+1})\upsilon(x_{n+1}^{-}) + \varepsilon y^{'}(x_n)\upsilon(x_n^{+}) \\
&+ \int_{x_n}^{x_{n+1}} y'(x)\upsilon(x)dx \quad + \quad \int_{x_n}^{x_{n+1}} ry(x)\upsilon(x)dx = \int_{x_n}^{x_{n+1}} f(x)\upsilon(x)dx, \quad n = 0, \ldots, N-1.
\end{aligned}$$

We sum all N equations above and get

$$\begin{aligned}
&\sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varepsilon y^{'}(x)\upsilon^{'}(x)dx \quad - \quad \sum_{n=0}^{N} [\varepsilon y^{'}(x_n)\upsilon(x_n)] \\
&+ \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} y'(x)\upsilon(x)dx \quad + \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} ry(x)\upsilon(x)dx = \int_0^1 f(x)\upsilon(x)dx.
\end{aligned}$$

By the definition of the jump and the average, the following equality holds

$$[\varepsilon y^{'}(x_n)\upsilon(x_n)] = \{\varepsilon y^{'}(x_n)\}[\upsilon(x_n)] + \{\upsilon(x_n)\}[\varepsilon y^{'}(x_n)], \quad 1 \le n \le N-1.$$

$[\varepsilon y^{'}(x_n)] = 0$ for $1 \le n \le N-1$, for the exact solution y. Then by substituting (3.1.2) into the equation above

$$\begin{aligned}
&\sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varepsilon y^{'}(x)\upsilon^{'}(x)dx \quad - \quad \sum_{n=0}^{N} \{\varepsilon y^{'}(x_n)\}[\upsilon(x_n)] \\
&+ \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} y'(x)\upsilon(x)dx \quad + \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} ry(x)\upsilon(x)dx = \int_0^1 f(x)\upsilon(x)dx.
\end{aligned}$$

The exact solution y satisfies $[y(x_n)] = 0$, since it is continuous. Therefore, if y is a solution of (3.1a-3.1c), then we obtain the following equality satisfied by y

$$\begin{aligned}
&\sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varepsilon y^{'}(x)\upsilon^{'}(x)dx \quad - \quad \sum_{n=0}^{N} \{\varepsilon y^{'}(x_n)\}[\upsilon(x_n)] + \gamma \sum_{n=0}^{N-1} \{\varepsilon \upsilon^{'}(x_n)\}[y(x_n)] \\
&+ \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} y'(x)\upsilon(x)dx \quad + \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} ry(x)\upsilon(x)dx = \int_0^1 f(x)\upsilon(x)dx - \gamma\varepsilon\upsilon^{'}(x_0)y(x_0) + \gamma\varepsilon\upsilon^{'}(x_N)y(x_N).
\end{aligned}$$

Here, γ is any real number. However, we restrict ourselves to the case $\gamma \in \{-1, 0, 1\}$ denoting which primal DG methods is considered.

3.1.2.1 DG (Bi)linear Forms

Let me define the DG bilinear form $a_\epsilon : D_k(\xi_h)\ X\ \ D_k(\xi_h) \to \mathbb{R}$:

$$a_\epsilon(y,\upsilon) \quad = \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \varepsilon y^{'}(x)\upsilon^{'}(x)dx - \sum_{n=0}^{N} \{\varepsilon y^{'}(x_n)\}[\upsilon(x_n)] + \gamma \sum_{n=0}^{N-1} \{\varepsilon \upsilon^{'}(x_n)\}[y(x_n)] + J_0(y,\upsilon) + J_1(y,\upsilon) \tag{3.2}$$

The terms $J_0(y,\upsilon)$ and $J_1(y,\upsilon)$ penalizes the jump of the solution and its derivative:

$$J_0(y,\upsilon) = \sum_{n=0}^{N} \frac{\sigma^0}{h}[y(x_n)][\upsilon(x_n)],$$

$$J_1(y,\upsilon) = \sum_{n=0}^{N} \frac{\sigma^1}{h}[y^{'}(x_n)][\upsilon^{'}(x_n)]$$

where σ^0 and σ^1 are two real nonnegative numbers.

Let me define the term coming from the convection part and the reaction part, respectively,

$$b(y,v) \quad = \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} y'(x)\upsilon(x)dx, \tag{3.3}$$

$$r(y,v) \quad = \quad \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} y(x)\upsilon(x)dx. \tag{3.4}$$

and $L(\upsilon) : D_k(\xi_h) \longrightarrow \mathbb{R}$ is the linear form

$$L(\upsilon) \quad = \quad \int_0^1 f(x)\upsilon(x)dx - \gamma\varepsilon\upsilon^{'}(x_0)y(x_0) + \gamma\varepsilon\upsilon^{'}(x_N)y(x_N) + \frac{\sigma^0}{h_{0,1}}\upsilon(x_0)y(x_0) + \frac{\sigma^0}{h_{N-1,N}}\upsilon(x_N)y(x_N) \tag{3.5}$$

The problem has been converted into the following one:

$$\text{Find} \quad y \in D_k(\xi_h) \quad \text{such that} \quad \forall \upsilon \in D_k(\xi_h), \tag{3.6}$$

$$a_\epsilon(y,\upsilon) + b(y,\upsilon) + r(y,\upsilon) = L(\upsilon). \tag{3.7}$$

3.1.3 Existence and uniqueness of the DG solution

By [53], for NIPG with $\sigma^0 > 0$, the solution exists and it is unique. When SIPG and IIPG methods is preferred, (uniqueness) existence of the solution is guaranteed if some conditions on the penalty parameters are imposed [53]. Thus, during the study, we have chosen σ^0 as 1 for NIPG and IIPG, σ^0 as 2 for SIPG, as in [53].

3.1.4 The Linear System

Let $\sigma^1 = 0$. The discontinuous piecewise quadratic polynomials are preferred. We used $\mathbb{P}_2(I_n)$ the monomial basis functions for the local basis functions. They have been translated from the interval $(-1, 1)$:

$$P_2(I_n) = span\{\phi_0^n, \phi_1^n, \phi_2^n\}$$

with

$$\phi_0^n(x) = 1, \qquad \phi_1^n = 2\frac{x - x_{n+1/2}}{x_{n+1} - x_n}, \qquad \phi_2^n = 4\frac{(x - x_{n+1/2})^2}{(x_{n+1} - x_n)^2}.$$

Define $x_{n+1/2} = \frac{1}{2}(x_n + x_{n+1})$ is the midpoint of the interval I_n. To facilitate the computation, let each subinterval have the same length:

$$x_n = x_0 + nh, \qquad h = \frac{1}{N}$$

Now, the local basis and their derivatives can be rewritten as

$$\phi_0^n(x) = 1, \qquad \phi_1^n(x) = \frac{2}{h}(x - (n + 1/2)h), \qquad \phi_2^n(x) = \frac{4}{h^2}(x - (n + 1/2)h)^2$$

$$(\phi_0^n)'(x) = 0, \qquad (\phi_1^n)'(x) = \frac{2}{h}, \qquad (\phi_2^n)'(x) = \frac{8}{h^2}(x - (n + 1/2)h).$$

The global basis functions $\{\Phi_i^n\}$ for the space $D_2(\xi_h)$ can be extended as follows::

$$\Phi_i^n(x) = \begin{cases} \phi_i^n(x), & x \in I_n \\ 0, & \text{otherwise.} \end{cases}$$

Computing Local Matrices

Computing local matrices arising from the diffusion part

The bilinear form a_ϵ consists of three kinds of terms: Involving integrals over I_n, involving the interior nodes x_n, and involving the boundary nodes x_0, x_n. Now, let me obtain the local matrices and then arrange them to obtain the global matrix.

First, we focus on the term arising from the integrals over the intervals I_n. Since we have preferred to use the quadratic polynomials as the basis functions, the solution y to (3.1a-3.1c) is a quadratic polynomial on each element I_n:

$$\forall x \in I_n \qquad y(x) = \alpha_0^n\phi_0^n + \alpha_1^n\phi_1^n + \alpha_2^n\phi_2^n. \tag{3.8}$$

By choosing $\upsilon = \phi_i^n$ for $i = 0, 1, 2,$, we obtain

$$\int_{I_n} y^{'}(x)\upsilon^{'}(x)dx = \int_{I_n} y^{'}(x)(\phi_i^n)^{'}(x)dx = \sum_{j=0}^{2} \alpha_j^n \int_{I_n} (\phi_j^n)^{'}(x)(\phi_i^n)^{'}(x)dx. \tag{3.9}$$

Hence, this linear system can be written as $A_n\alpha^n$, where

$$\alpha^n = \begin{pmatrix} \alpha_0^n \\ \alpha_1^n \\ \alpha_2^n \end{pmatrix} \quad (A_n)_{ij} = \int_{I_n} (\phi_j^n)^{'}(x)(\phi_i^n)^{'}(x)dx.$$

Hence,

$$\mathbf{A}_n = \frac{1}{h}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & \frac{16}{3} \end{pmatrix}$$

Second, the terms involving the interior nodes x_n is considered.

$$-\{y^{'}(x_n)\}[\upsilon(x_n)] + \gamma\{\upsilon^{'}(x_n)\}[y(x_n)] + \frac{\sigma^0}{h}[y(x_n)][\upsilon(x_n)] = b_n + c_n + d_n + e_n, \tag{3.10}$$

The terms are defined as follows:

$$b_n = \frac{1}{2}y^{'}(x_n^+)\upsilon(x_n^+) - \frac{\gamma}{2}y(x_n^+)\upsilon^{'}(x_n^+) + \frac{\sigma^0}{h}y(x_n^+)\upsilon(x_n^+),$$

$$c_n = -\frac{1}{2}y^{'}(x_n^-)\upsilon(x_n^-) + \frac{\gamma}{2}y(x_n^-)\upsilon^{'}(x_n^-) + \frac{\sigma^0}{h}y(x_n^-)\upsilon(x_n^-),$$

$$d_n = -\frac{1}{2}y^{'}(x_n^+)\upsilon(x_n^-) - \frac{\gamma}{2}y(x_n^+)\upsilon^{'}(x_n^-) - \frac{\sigma^0}{h}y(x_n^+)\upsilon(x_n^-),$$

$$e_n = \frac{1}{2}y^{'}(x_n^-)\upsilon(x_n^+) + \frac{\gamma}{2}y(x_n^-)\upsilon^{'}(x_n^+) - \frac{\sigma^0}{h}y(x_n^-)\upsilon(x_n^+).$$

We use the definition of the DG solution $y(x)$ and choose $\upsilon = \phi_i^n$, the local matrices B_n, C_n, D_n, and E_n, can be rewritten respectively:

$$\begin{aligned}
(B_n)_{ij} &= \frac{1}{2}(\phi_j^n)^{'}(x_n^+)(\phi_i^n)x_n^+) - \frac{\gamma}{2}(\phi_j^n)(x_n^+)(\phi_i^n)^{'}(x_n^+) + \frac{\sigma^0}{h}(\phi_j^n)(x_n^+)(\phi_i^n)(x_n^+), \\
(C_n)_{ij} &= -\frac{1}{2}(\phi_j^{n-1})^{'}(x_n^-)(\phi_i^{n-1})x_n^-) + \frac{\gamma}{2}(\phi_j^{n-1})(x_n^-)(\phi_i^n)^{'}(x_{n-1}^-) + \frac{\sigma^0}{h}(\phi_j^{n-1})(x_n^-)(\phi_i^{n-1})(x_n^-), \\
(D_n)_{ij} &= -\frac{1}{2}(\phi_j^n)^{'}(x_n^+)(\phi_i^{n-1})x_n^-) - \frac{\gamma}{2}(\phi_j^n)(x_n^+)(\phi_i^{n-1})^{'}(x_n^-) - \frac{\sigma^0}{h}(\phi_j^n)(x_n^+)(\phi_i^{n-1})(x_n^-), \\
(E_n)_{ij} &= \frac{1}{2}(\phi_j^{n-1})^{'}(x_n^-)(\phi_i^n)x_n^+) + \frac{\gamma}{2}(\phi_j^{n-1})(x_n^-)(\phi_i^n)^{'}(x_n^+) - \frac{\sigma^0}{h}(\phi_j^{n-1})(x_n^-)(\phi_i^n)(x_n^+).
\end{aligned}$$

Hence, we obtain the local matrices

$$\mathbf{B}_n = \frac{1}{h}\begin{pmatrix} \sigma^0 & 1-\sigma^0 & -2+\sigma^0 \\ -\gamma-\sigma^0 & -1+\gamma+\sigma^0 & 2-\gamma-\sigma^0 \\ 2\gamma+\sigma^0 & 1-2\gamma-\sigma^0 & -2+2\gamma+\sigma^0 \end{pmatrix},$$

$$\mathbf{C}_n = \frac{1}{h}\begin{pmatrix} \sigma^0 & -1+\sigma^0 & -2+\sigma^0 \\ \gamma+\sigma^0 & -1+\gamma+\sigma^0 & -2+\gamma+\sigma^0 \\ 2\gamma+\sigma^0 & 1+2\gamma+\sigma^0 & -2+2\gamma+\sigma^0 \end{pmatrix},$$

$$\mathbf{D}_n = \frac{1}{h}\begin{pmatrix} -\sigma^0 & -1+\sigma^0 & 2-\sigma^0 \\ -\gamma-\sigma^0 & -1+\gamma+\sigma^0 & 2-\gamma-\sigma^0 \\ -2\gamma-\sigma^0 & 1+2\gamma+\sigma^0 & 2-2\gamma-\sigma^0 \end{pmatrix},$$

$$\mathbf{E}_n = \frac{1}{h}\begin{pmatrix} -\sigma^0 & 1-\sigma^0 & 2-\sigma^0 \\ \gamma+\sigma^0 & -1+\gamma+\sigma^0 & -2+\gamma+\sigma^0 \\ -2\gamma-\sigma^0 & 1-2\gamma-\sigma^0 & 2-2\gamma-\sigma^0 \end{pmatrix}.$$

Finally, the local matrices obtained from the boundary nodes x_0 and x_N can be computed as:

$$f_0 = y^{'}(x_0)\upsilon(x_0) - \gamma\upsilon^{'}(x_0)y(x_0) + \frac{\sigma}{h}y(x_0)\upsilon(x_0), \tag{3.11}$$

$$f_N = -y^{'}(x_N)\upsilon(x_N) + \gamma\upsilon^{'}(x_N)y(x_N) + \frac{\sigma}{h}y(x_N)\upsilon(x_N). \tag{3.12}$$

We substitute the DG solution y^{DG} and $\upsilon = \phi_i^n$, the terms F_0 and F_n are of the form:

$$F_0 = \phi^{'}(x_0)\upsilon(x_0) - \gamma\upsilon^{'}(x_0)y(x_0) + \frac{\sigma^0}{h}y(x_0)\upsilon(x_0),$$

$$F_N = -\phi^{'}(x_N)\upsilon(x_N) + \gamma\upsilon^{'}(x_N)y(x_N) + \frac{\sigma^0}{h}y(x_N)\upsilon(x_N).$$

Then, the matrices F_0 and F_N are obtained:

$$\mathbf{F}_0 = \frac{1}{h}\begin{pmatrix} \sigma^0 & 2-\sigma^0 & -4+\sigma^0 \\ -2\gamma-\sigma^0 & -2+2\gamma+\sigma^0 & 4-2\gamma-\sigma^0 \\ 4\gamma+\sigma^0 & 2-4\gamma-\sigma^0 & -4+4\gamma+\sigma^0 \end{pmatrix},$$

$$\mathbf{F}_N = \frac{1}{h}\begin{pmatrix} \sigma^0 & -2+\sigma^0 & -4+\sigma^0 \\ 2\gamma+\sigma^0 & -2+2\gamma+\sigma^0 & -4+2\gamma+\sigma^0 \\ 4\gamma+\sigma^0 & 2+4\gamma+\sigma^0 & -4+4\gamma+\sigma^0 \end{pmatrix}.$$

Computing local matrices arising from the convection part

By using the DG solution y^{DG} and choosing $\upsilon = \phi_i^n$ for $i = 0, 1, 2,$, we obtain

$$\int_{I_n} y'(x)\upsilon(x)dx = \int_{I_n} y'(x)(\phi_i^n)(x)dx = \sum_{j=0}^{2} \alpha_j^n \int_{I_n} (\phi_j^n)'(x)(\phi_i^n)(x)dx. \tag{3.13}$$

Hence, this linear system can be written as $C_n\alpha^n$, where

$$\alpha^n = \begin{pmatrix} \alpha_0^n \\ \alpha_1^n \\ \alpha_2^n \end{pmatrix} \quad (\boldsymbol{c}^n)_{ij} = \int_{I_n} (\phi_j^n)'(x)(\phi_i^n)(x)dx.$$

Hence,

$$\boldsymbol{c}^n = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & \frac{16}{3} \\ 0 & \frac{8}{3} & 0 \end{pmatrix}.$$

Computing local matrices arising from the reaction part

Similar to the other cases, use y^{DG} and choose $\upsilon = \phi_i^n$ for $i = 0, 1, 2,$, we have

$$\int_{I_n} y(x)\upsilon(x)dx = \int_{I_n} y(x)(\phi_i^n)(x)dx = \sum_{j=0}^{2} \alpha_j^n \int_{I_n} (\phi_j^n)(x)(\phi_i^n)(x)dx. \tag{3.14}$$

Hence, this linear system can be written as $\mathcal{R}\alpha^n$, where

$$\alpha^n = \begin{pmatrix} \alpha_0^n \\ \alpha_1^n \\ \alpha_2^n \end{pmatrix} \quad (\boldsymbol{r}^n)_{ij} = \int_{I_n} (\phi_j^n)(x)(\phi_i^n)(x)dx.$$

Hence,

$$\boldsymbol{r}^n = h\begin{pmatrix} 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{5} \end{pmatrix}$$

Computing the right-side

Let me mention the linear form obtained from the right-hand side:

$$\mathcal{L}(\upsilon) = \int_0^1 f(x)\upsilon(x)dx - \gamma\varepsilon\upsilon'(x_0)y(x_0) + \gamma\varepsilon\upsilon'(x_N)y(x_N) \tag{3.15}$$

$$+ \frac{\sigma^0}{h_{0,1}}\upsilon(x_0)y(x_0) + \frac{\sigma^0}{h_{N-1,N}}\upsilon(x_N)y(x_N). \tag{3.16}$$

We choose $\upsilon = \Phi_n^i$ and use the given boundary conditions to obtain:

$$\begin{aligned}\mathcal{L}(\Phi_n^i) &= \int_0^1 f(x)\Phi_n^i dx - \gamma\varepsilon(\Phi_n^i)'(x_0)y_0 + \gamma\varepsilon(\Phi_n^i)'(x_N)y_1 \\ &+ \frac{\sigma^0}{h_{0,1}}\Phi_n^i(x_0)y_0 + \frac{\sigma^0}{h_{N-1,N}}\Phi_n^i(x_N)y_1\end{aligned}$$

By the definition of the global basis functionsΦ_i^n, the first term can be rewritten as

$$\int_0^1 f(x)\Phi_i^n dx = \int_{x_n}^{x_{n+1}} f(x)\phi_i^n dx$$

If we perform a change of variable, then we obtain

$$\int_0^1 f(x)\Phi_i^n dx = \frac{h}{2}\int_0^1 f(\frac{h}{2}t + (n+1/2)h)t^i dt$$

In order to facilitate the computation of the integral,the Gauss quadrature rule is used. Define a set of weights $(w_j)_{1\le j\le Q_G}$ and a set of nodes $(s_j)_{1\le j\le Q_G}$. Then,

$$\int_{-1}^1 \upsilon(t)dt \approx \sum_{j=1}^{Q_G} w_j\upsilon(s_j).$$

By using above equality, we have

$$\int_0^1 f(x)\Phi_i^n dx \approx \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j^i dt$$

Then, the vector L can be constructed by taking into account of the order of α_i^n:

$$(l_0^0, l_1^0, l_2^0, l_0^1, l_1^1, l_2^1, \ldots, l_0^{N-1}, l_0^{N-1}, l_0^{N-1})$$

where the first three components are

$$
\begin{aligned}
l_0^0 &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h) + \frac{\sigma^0}{h}y_0 \\
l_1^0 &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j - \epsilon\mu\frac{2}{h}y_0 - \frac{\sigma^0}{h}y_0 \\
l_2^0 &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j^2 + \epsilon\mu\frac{4}{h}y_0 + \frac{\sigma^0}{h}y_0
\end{aligned}
$$

the last three components are

$$
\begin{aligned}
l_0^{N-1} &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h) + \frac{\sigma^0}{h}y_1 \\
l_1^{N-1} &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j + \epsilon\mu\frac{2}{h}y_1 - \frac{\sigma^0}{h}y_1 \\
l_1^{N-1} &= \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j^2 - \epsilon\mu\frac{4}{h}y_1 + \frac{\sigma^0}{h}y_1
\end{aligned}
$$

and the other components are

$$
\forall 1 \le n \le N-1, \quad \forall 0 \le i \le 2, \quad l_i^n = \frac{h}{2}\sum_0^1 f(\frac{h}{2}s_j + (n+1/2)h)s_j^i.
$$

Up to now, the boundary conditions have been imposed weakly, by the terms $-\gamma\upsilon'(x_0)y(x_0) + \frac{\sigma}{h}y(x_0)\upsilon(x_0)$ and $\gamma\upsilon'(x_N)y(x_N) + \frac{\sigma}{h}y(x_N)\upsilon(x_N)$. Indeed, boundary conditions can be imposed strongly by defining

$$
D_k^0(\xi_h) = \{\upsilon \in D_k(\xi_h) : \upsilon(0) = 0, \upsilon(1) = 0\}.
$$

Then, define the DG solution as $y^{DG} = y_0^{DG} + \tilde{y}$ where $\tilde{y}$ is a continuous piecewise polynomial of degree k such that $\tilde{y}(0) = y_0$ and $\tilde{y}(1) = y_1$. In addition, the modified scheme is satisfied by $y_0^{DG} \in D_k^0(\xi_h)$ [53].

Global Matrices For Steady Diffusion-Convection-Reaction Equation

Up to now, we have introduced how to construct the local matrices on each subinterval I_n. Now, these matrices need to be assembled by keeping the order of α_i^n to obtain the global matrices. Let me order α_i^n as follows:

$$\alpha_0^0, \alpha_1^0, \alpha_2^0, \cdot, \alpha_0^{N-1}, \alpha_1^{N-1}, \alpha_2^{N-2}.$$

Then the global matrices arising from the diffusion, convection and reaction terms are obtained, respectively:

$$S = \begin{pmatrix} R_0 & D_1 & & & & \\ E_1 & R & D_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & E_{N-2} & R & D_{N-1} \\ & & & & E_{N-1} & R_N \end{pmatrix}$$

$$C = \begin{pmatrix} \boldsymbol{c}^0 & & & & & \\ & \boldsymbol{c}^1 & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \boldsymbol{c}^{N-2} & \\ & & & & & \boldsymbol{c}^{N-1} \end{pmatrix} \mathcal{R} = \begin{pmatrix} \boldsymbol{r}^0 & & & & & \\ & \boldsymbol{r}^1 & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \boldsymbol{r}^{N-2} & \\ & & & & & \boldsymbol{r}^{N-1} \end{pmatrix}$$

where

$$R = A_n + B_n + C_{n+1}, \qquad R_0 = A_0 + F_0 + C_1, \qquad R_N = A_{N-1} + F_N + B_{N-1}.$$

Then (3.1a-3.1c) is converted into a linear system

$$(\mathcal{S} + C + \mathcal{R})y = L \tag{3.17}$$

where y consisting of the coefficients α_i^n, $\quad \forall i = 0, 1, 2$ and $\forall \quad 0 \leq n \leq N-1$.

3.2 2-D Discontinuous Galerkin Methods

3.2.1 Model Problem

We consider the time dependent advection-diffusion-reaction problem with Dirichlet boundary conditions. Let Ω be a bounded polygonal in $\mathbb{R}^d$, $d = 2$ or 3, $f \in L^2(\Omega)$, $g_D \in H^{\frac{1}{2}}(\partial\Omega)$ and $y_0 \in L^2(\Omega)$,

$$-\varepsilon\Delta y(x) + c(x) \cdot \nabla y(x) + r(x)y(x) = f(x) \quad \text{in} \quad \Omega, \tag{3.18a}$$

$$y(x) = g_D \quad \text{on} \quad \partial\Omega. \tag{3.18b}$$

If f and g_D are smooth, then $y \in C^2(\bar{\Omega})$ is a strong solution of (3.18a-3.18b).

Suppose that we have a mesh ξ_h. Let me call the set of interior edges (or faces) as Γ_h. A unit normal vector n_e is considered with each edge (or face) e. In case of a boundary edge, unit outward vector normal to the boundary is used.

For the trace of υ along any side of one element E be well-defined, υ must be an element of $H^1(\xi_h)$. Two traces of υ along e are observed when two elements E_1^e and E_2^e are neighbors and have a common side e. Consider the normal vector n_e in the direction from E_1^e to E_2^e. Then, a jump and an average are as follows:

$$[\upsilon] = (\upsilon|_{E_1^e}) - (\upsilon|_{E_2^e}) \qquad \{\upsilon\} = \frac{1}{2}(\upsilon|_{E_1^e}) + \frac{1}{2}(\upsilon|_{E_2^e}) \quad \forall e = \partial E_1^e \cap \partial E_2^e.$$

In case of a boundary edge, a jump and an average is given, by convention,:

$$\{\upsilon\} = [\upsilon] = (\upsilon|_{E_1^e}) \qquad \forall e = \partial E_1^e \cap \partial\Omega.$$

Define the inflow and outflow boundaries Γ_-, Γ_+ such that

$$\Gamma_- = \{x \in \Gamma : \quad c(x) \cdot n(x) < 0\}, \qquad \Gamma_+ = \{x \in \Gamma : \quad c(x) \cdot n(x) \geq 0\}.$$

For any element E, the inflow and outflow parts of ∂E are defined by

$$\partial_- E = \{x \in \partial E : \quad c(x) \cdot n_E(x) < 0\}, \qquad \partial_+ E = \{x \in \partial E : \quad c(x) \cdot n_E(x) \geq 0\},$$

respectively, where $n_E(x)$ denotes the unit outward vector to ∂E at $x \in \partial E$. For each E and $\upsilon \in H^1(E)$, define υ_E^+ the interior trace and υ_E^- the exterior trace of $\upsilon_{|_E}$ on ∂E. Here, jump depends on the direction of the flow.

Let us subdivide Ω into elements E. Here, E can be a triangle or a quadrilateral in 2D, or a tetrahedron or hexahedron in 3D. A conforming mesh is preferred, that is, the intersection of two elements in the mesh is either empty, a vertex, an edge, or a face. Let me call the mesh and the maximum element diameter, respectively, as ξ_h and h. We are interested in a positive constant ρ defined as follows:

$$\forall E \in \xi_h, \quad \frac{h_E}{\rho_E} \leq \rho,$$

where h_E is the diameter of E and ρ_E is the maximum diameter of a ball inscribed in E. Then, we obtain a regular mesh.

In order apply the DG methods, the broken Sobolev spaces which depend on the partition of the domain are used. The broken Sobolev space can be defined for any real number s,

$$H^s(\xi_h) = \{v \in L^2(\Omega) : \forall E \in \xi_h, v|_E \in H^s(E)\},$$

where the broken Sobolev norm:

$$|||v|||_{H^s(\xi_h)} = \left(\sum_{E \in \xi_h} \|v\|^2_{H^s(E)} \right)^{1/2},$$

and the broken gradient seminorm:

$$|||\nabla v|||_{H^s(\xi_h)} = \left(\sum_{E \in \xi_h} \|\nabla v\|^2_{H^s(E)} \right)^{1/2}.$$

Then,

$$H^s(\Omega) \subset H^s(\xi_h) \quad \text{and} \quad H^{s+1}(\xi_h) \subset H^s(\xi_h).$$

3.2.1.1 The DG (Bi)linear forms

We now define the DG bilinear forms $a_\epsilon : H^s(\xi_h) \times H^s(\xi_h) \to \mathbb{R}$:

$$a_\epsilon(y, v) = \sum_{E \in \xi_h} \int_E \varepsilon \nabla y \cdot \nabla v dx - \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{\varepsilon \nabla y \cdot n_e\}[v] ds \tag{3.19}$$

$$+ \gamma \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{\varepsilon \nabla v \cdot n_e\}[y] ds + J_0^{\sigma_0, \beta_0}(y, v). \tag{3.20}$$

Regarding to [53], for $s > 3/2$, a bilinear form $J_0^{\sigma_0, \beta_0}(y, v) : H^s(\xi_h) \times H^s(\xi_h) \to \mathbb{R}$ that penalizes the jump of the function value:

$$J_0^{\sigma_0,\beta_0}(y,\upsilon) = \sum_{e\in\Gamma_h\cup\Gamma_D} \frac{\sigma_\epsilon^0}{|e|^{\beta_0}} \int_e [y][\upsilon] ds.$$

σ_e^0, a penalty parameter, is a nonnegative real number and β_0 is a positive number depending on the dimension d,

$$\forall e \subset \partial E, \quad |e| \leq h_E^{d-1} \leq h^{d-1}.$$

The DG forms of convection and the reaction parts are, respectively, [38]:

$$c(y,\upsilon) = \sum_{E\in\xi_h} \left(\int_E c\nabla y \cdot \upsilon dx - \int_{\partial_- E\setminus\Gamma} (c\cdot n_e)(y^+ - y^-)\upsilon^+ ds - \int_{\partial_- E\cap\Gamma_-} (c\cdot n_e) y^+\upsilon^+ ds \right), \tag{3.21}$$

$$r(y,\upsilon) = \sum_{E\in\xi_h} \int_E ry\upsilon, \tag{3.22}$$

such that

$$y^+ = \begin{cases} y_{|_{E_e^1}} & \text{if} \quad c\cdot n \geq 0 \\ y_{|_{E_e^2}} & \text{if} \quad c\cdot n < 0, \end{cases} \qquad y^- = \begin{cases} y_{|_{E_e^2}} & \text{if} \quad c\cdot n \geq 0 \\ y_{|_{E_e^1}} & \text{if} \quad c\cdot n < 0. \end{cases}$$

Define the following linear form:

$$L(v) = \sum_{E\in\xi_h} \left(\int_E f\upsilon dx - \int_{\partial_- E\cap\Gamma_-} (c\cdot n_e) g_D \upsilon^+ ds \right) + \sum_{e\in\Gamma_D} \int_e (\gamma\varepsilon\nabla\upsilon\cdot n_e + \frac{\sigma_e^0}{|e|^{\beta_0}}\upsilon) g_D ds + \sum_{e\in\Gamma_N} \int_e \upsilon g_N ds. \tag{3.23}$$

If the functions, in the above forms, belong to $H^s(\xi_h)$ for any $s > 3/2$, then the fact that the integrals in these forms are suitable to use Cauchy-Schwartz's inequality and trace inequalities [53]. The DG variational form of (3.18a-3.18b) is as follows: Find $p \in H^s(\xi_h)$ for any $s > 3/2$ such that

$$a_\epsilon(y,\upsilon) + c(y,\upsilon) + r(y,\upsilon) \quad = \quad L(\upsilon). \tag{3.24}$$

3.2.2 DG scheme

The general DG finite element method is as follows: Find $y_h \in D_k(\xi_h)$ such that

$$a_\epsilon(y_h,\upsilon) + c(y_h,\upsilon) + r(y_h,\upsilon) \quad = \quad L(\upsilon), \quad \forall \quad \upsilon \in D_k(\xi_h). \tag{3.25}$$

As in the one-dimensional case, the parameter ϵ in the bilinear form a_ϵ can be any real number. However, we have chosen ϵ as -1, 0, or 1.

- For ϵ=-1, the method is called symmetric interior penalty Galerkin (SIPG). If a large penalty term σ_e^0 is used, then the method is convergent.

- For ϵ=1, the method is called nonsymmetric interior penalty Galerkin (NIPG). If a nonnegative value is chosen for the penalty parameter σ_e^0, the method is convergent.

- For ϵ=0, the method is called incomplete interior penalty Galerkin (IIPG). For this method to be convergent, we choose σ_e^0 large enough.

- An extra stabilization term $J_1^{\sigma_1,\beta_1}(\upsilon,\omega)$ can be added to the bilinear form a_ϵ. The jump of the derivative is penalized by this term.

$$J_1^{\sigma_1,\beta_1}(\upsilon,\omega) = \sum_{e\in\Gamma_h} \frac{\sigma_e^1}{|e|^{\beta_1}} \int_e [\varepsilon\nabla\upsilon\cdot n_e][\varepsilon\nabla\omega\cdot n_e]ds.$$

  During the study, σ_e^1 has been taken zero, for simplicity.

Definition 3.2.1 *A bilinear form defined on a normed linear space V with norm* $\|\cdot\|_V$ *is coercive if there exists a positive constant* κ *such that* $\forall \quad \upsilon \in V, \quad \kappa\|\upsilon\|_V^2 \le a(\upsilon,\upsilon).$ ■

By [53], a_{+1} is coercive for any choice of σ_e^0. However, for a_{-1} and a_0 to be coercive, $\beta_0(d-1)$ must be larger than 1 and σ_e^0 must bounded below by a constant σ_e^* that depends only on the bounds ε_0 and ε_1 where $\varepsilon_0 \le \varepsilon \le \varepsilon_1$; and the constant in the following trace inequality [53]:

$$\forall \quad \upsilon \in \mathbb{P}_k(E), \forall \quad e \subset \partial E, \quad \|\nabla\upsilon\cdot n_e\|_{L^2(e)} \le Ch_E^{-1/2}\|\nabla\upsilon\|_{L^2(E)}.$$

Definition 3.2.2 *A bilinear form defined on a normed linear space V with norm* $\|\cdot\|_V$ *is continuous if there exists a positive constant M such that* $\forall \quad \upsilon,\omega \in V, \quad a(\upsilon,\omega) \le M\|\upsilon\|_V\|\omega\|_V.$

■

Continuity of a_ϵ on $D_k(\xi_h)$ depends on σ_e^0. The bilinear form is continuous for positive σ_e^0 for all e with the energy norm $\|\cdot\|_\epsilon$ [53] :

$$\forall \quad \upsilon \in D_k(\xi_h), \quad a_\epsilon(\upsilon,\omega) \le M\|\upsilon\|_\epsilon\|\omega\|_\epsilon.$$

3.2.3 Existence and Uniqueness of The DG solution

Lemma 3.2.3 *Let the following conditions be true:*

- *In the NIPG case, $k \geq 1$ and either $r > 0$ or $\sigma_e^0 > 0$ for all e;*
- *In the SIPG case or IIPG case, $k \geq 1$ and σ_e^0 is bounded below by a large constant for all e;*
- *In the NIPG case, $k \geq 2$ and $\sigma_e^0 = 0$ for all e and $r = 0$.*

Then, the DG solution y_h exist and unique [53].

3.2.3.1 Basic Definitions

We consider the finite element space

$$D_k(\xi_h) = \{\upsilon \in L^2(\Omega) : \forall E \in \xi_h, \upsilon|_E \in \mathbb{P}_k(E)\},$$

which is a subspace of $H^s(\xi_h)$ for $s > 3/2$. $P_k(E)$ denotes the space of polynomials of total degree less than or equal to k, which is a positive integer. The test functions are chosen from $D_k(\xi_h)$ and they are discontinuous along the edges of the mesh. Mesh elements, that is triangles or quadrilateral for 2D, are named as physical elements. To facilitate the computation, we map the physical elements to the reference elements $\hat{E}$ and perform all computations on the reference element.

Reference triangular element: Consider a triangle $\hat{E}$ of which vertices are $\hat{A1}(0,0)$, $\hat{A2}(1,0)$, $\hat{A3}(0,1)$. An affine map F_E can be defined from the reference element to the physical one. Suppose that the vertices $A_i(x_i, y_i)$ for $i = 1, 2, 3$ belongs to a physical element E. Then the map F_E can be written as

$$F_E \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad x = \sum_{i=1}^{3} x_i \hat{\phi}_i(\hat{x}, \hat{y}), \quad y = \sum_{i=1}^{3} x_i \hat{\phi}_i(\hat{x}, \hat{y}),$$

where

$$\begin{aligned} \hat{\phi}_1(\hat{x}, \hat{y}) &= 1 - \hat{x} - \hat{y}, \\ \hat{\phi}_2(\hat{x}, \hat{y}) &= \hat{x}, \\ \hat{\phi}_2(\hat{x}, \hat{y}) &= \hat{y}, \end{aligned}$$

If we arrange the terms, then

$$\begin{pmatrix} x \\ y \end{pmatrix} = F_E \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = B_E \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + b_E$$

where

$$B_E = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad b_E = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

In order to compute the integrals, the determinant of B_E, which is twice of the area of an element, is needed. Then, the fact that B_E is invertible is obvious. In addition, $\|B_E\| \leq \frac{h_E}{\hat{\rho}}, \quad \|B_E^{-1}\| \leq \frac{\hat{h}}{\rho_E}$ where $\hat{h}$, $\hat{\rho}$ and ρ_E refer to the diameter of $\hat{E}$, diameter of the largest circle inscribed in $\hat{E}$ and the diameter of the largest circle inscribed in E, respectively. Thus, we derive that the matrix norm (induced by the Euclidean norm) of B_E and B_E^{-1} are bounded.

Passing to the reference element from the physical elements via the mapping F_E can be seen as a change of variables. By $\hat{\upsilon} = \upsilon \circ F_E$, we have $\hat{\upsilon}(\hat{x}, \hat{y}) = \upsilon(x, y)$. Since the gradients of the functions are seen in the bilinear forms, they must be defined in terms of the affine map. Then, $\hat{\nabla}\hat{\upsilon}$ the gradient of $\hat{\upsilon}$ with respect to $\hat{x}$ and $\hat{y}$ is:

$$\hat{\nabla}\hat{\upsilon} = \begin{pmatrix} \frac{\partial \hat{\upsilon}}{\partial \hat{x}} \\ \frac{\partial \hat{\upsilon}}{\partial \hat{y}} \end{pmatrix}.$$

It can be written as $\hat{\nabla}\hat{\upsilon} = B_E^T \nabla \upsilon \circ F_E$ in terms of F_E.

Reference triangular element: We have mentioned that the test functions are discontinuous along the edges. Then, the support of the basis functions of $D_k(\xi_h)$ is contained in one element. Consider

$$D_k(\xi_h) = span\{\phi_i^E : 1 \leq i \leq N_{loc}, E \in \xi_h\}$$

with

$$\phi_i^E(x) = \begin{cases} \hat{\phi}_i o F_E(x), & x \in E \\ 0, & x \notin E. \end{cases}$$

We define $(\hat{\phi}_i)_{1 \leq i \leq N_{loc}}$ on the reference element. If we prefer to use the monomials; for 2D, we have

$$\hat{\phi}_i(\hat{x}, \hat{y}) = \hat{x}^I \hat{y}^J, \quad I + J = i, \quad 0 \leq i \leq k.$$

$$N_{loc} = \frac{(k+1)(k+2)}{2}$$

refers to the local dimension.

- Piecewise linear monomials:

$$\hat{\phi_0} = 1, \quad \hat{\phi_1} = \hat{x}, \quad \hat{\phi_2} = \hat{y}.$$

- Piecewise quadratic monomials:

$$\hat{\phi_0} = 1, \quad \hat{\phi_1} = \hat{x}, \quad \hat{\phi_2} = \hat{y},$$
$$\hat{\phi_3} = \hat{x}^2, \quad \hat{\phi_4} = \hat{x}\hat{y}, \quad \hat{\phi_5} = \hat{y}^2.$$

Numerical quadrature-2D: As we have mentioned before, the integrals obtained by the (bi)linear forms are computed on the reference element. As in the one-dimensional case, a quadrature rule can be used. Consider the following approximation:

$$\int_{\hat{E}} \hat{v} dx \approx \sum_{j=1}^{Q_D} \omega_j \hat{v}(s_{x,j}, s_{y,j}).$$

A set of weights ω_j and nodes $(s_{x,j}, s_{y,j}) \in \hat{E}$ for different values of Q_D can be found [53]. To obtain a better approximation, high order quadrature rule must be used. By the following equality, we can observe how the map F_E is used to pass to the reference element from the physical element:

$$\int_E v dx = \int_{\hat{E}} v \circ F_E det(B_E) dx = 2|E| \int_{\hat{E}} \hat{v} dx.$$

By a quadrature rule, we can approximate this integral as

$$\int_E v dx \approx 2|E| \sum_{j=1}^{Q_D} \omega_j \hat{v}(s_{x,j}, s_{y,j}).$$

In case of a vector function and the gradient, this approximation can be written as

$$\int_E \nabla v \cdot \omega dx = 2|E| \int_{\hat{E}} (B_e^T)^{-1} \hat{\nabla} \hat{v} \cdot \hat{\omega} dx \approx 2|E| \sum_{j=1}^{Q_D} \omega_j (B_e^T)^{-1} \hat{\nabla} \hat{v}(s_{x,j}, s_{y,j}) \cdot \hat{\omega}(s_{x,j}, s_{y,j}).$$

In addition, if the integral of the gradient of two functions is needed, then we have

$$\int_E \nabla v \cdot \nabla \omega dx \approx 2|E| \sum_{j=1}^{Q_D} \omega_j (B_e^T)^{-1} \hat{\nabla} \hat{v}(s_{x,j}, s_{y,j}) \cdot (B_e^T)^{-1} \hat{\nabla} \hat{\omega}(s_{x,j}, s_{y,j}).$$

CHAPTER 4

SPACE AND TIME DISCRETIZATION

After having discretized the problem in space by DG method, the problem must be discretized with respect to time variable. There are some different ways to handle this. In [15], non-linear convection dominated problems are discretized by a DG method in space and discretized in time by explicit-high order accurate Runge-Kutta methods. In addition, DG can be used to perform both of spatial and temporal discretization as in [22, 24]. As a different approach, the problem, firstly, can be discretized in time by Crank-Nicolson and then spatial discretization can be performed by usual conforming finite elements [45]. During the study, we have discretized the problem by DG method in space. Then, we have derived the full discrete problem by backward Euler and Crank-Nicolson. Since the problem we are interested in is stiff, these implicit methods are preferable [39]. Consider the following problem

$$\min_{u \in U} \widehat{J}(u), \tag{4.1}$$

where U is a closed convex subset of $\mathbb{R}^{n_u}$, such as $U = \mathbb{R}^{n_u}$ or $U = [-1, 1]^{n_u}$, and $\widehat{J} : U \longrightarrow \mathbb{R}$ is a smooth function. To assess the value of $\widehat{J}$, it is required to solve a system of linear equations. For given J and e, the problem can be written in detail as follows [17]:

$$\widehat{J}(u) = J(y(u), u), \tag{4.2}$$

where $y(u) \in \mathbb{R}^{n_y}$ is the solution of an equation

$$e(y, u) = 0. \tag{4.3}$$

We denote the solution of (4.1) as an implicit function $y(\cdot)$ and a vector y in $\mathbb{R}^{n_y}$. In addition, the partial Jacobian of the function e with respect to y and the partial gradient of the function J with respect to u are notated as $e_y(y, u) \in \mathbb{R}^{n_y \times n_y}$ and $\nabla_u J(y, u)\mathbb{R}^{n_u}$, respectively.

We have to insert some conditions in order to guarantee the existence of a differentiable function

$$y : \mathbb{R}^{n_u} \longrightarrow \mathbb{R}^{n_y} \quad \text{defined by} \quad e(y,u) = 0 \tag{4.4}$$

by the implicit function theorem [17].

Assumption

- e(y,u)=0 for all $u \in U$ and a unique $y \in \mathbb{R}^{n_y}$ corresponding to that $u \in U$.
- J and e are twice continuously differentiable on D, where D is an open set of $\mathbb{R}^{n_y \times n_u}$ with $\{(y,u) : u \in U, e(y,u) = 0\} \subset D$.
- For all $(y,u) \in \{(y,u) : u \in U, e(y,u) = 0\}$, existence of $e_y(y,u)^{-1}$ is guaranteed.

Let me mention the unconstrained optimal control problem:

$$\min J(y,u) := \frac{1}{2}\int_0^T \|y - \hat{y}\|^2_{L^2(\Omega)} dt + \frac{\alpha}{2}\int_0^T \|u\|^2_{L^2(\Omega)} dt \tag{4.5}$$

subject to

$$\frac{\partial y}{\partial t} - \varepsilon \Delta y(x,t) + c(x,t)\cdot\nabla y(x,t) + r(x,t)y(x,t) = f(x,t) + u(x,t) \quad \text{in} \quad \Omega \times (0,T], \tag{4.6a}$$

$$y(x,t) = g_D \quad \text{on} \quad \partial\Omega \times [0,T], \tag{4.6b}$$

$$y(x,0) = y_0 \quad \text{in} \quad \Omega. \tag{4.6c}$$

4.1 Discretize Then Optimize and Optimize Then Discretize Approaches

Optimal control problems can be solved in two different ways [33]: Discretize Then Optimize and Optimize Then Discretize. If discretize then optimize approach is preferred, then the full discrete state equation is written and all functions spaces are substituted by the finite dimensional function spaces. Then, Lagrangian for the full discrete problem has been constructed to extract the optimality system. For optimize-then-discretize approach, we set continuous necessary optimality conditions which consists of the state, adjoint and the gradient equation. Then, a finite element method is used to discretize the optimality system [16, 17, 18, 33].

4.2 Variational Formulation

A strong solution of the PDE lies in $C([0,T]\times\Omega)$. A weak solution of the parabolic problem is a function of $Y = H^1((0,T);H^{-1}(\Omega)) \cap L^2((0,T);H^1(\Omega))$ [53]. A possible weak formula for the state equation can be written as follows: For given f, u and y_0, find $y(u) \in Y$ such that

$$\begin{aligned} \left(\frac{\partial y}{\partial t}, \upsilon\right) + a(y,\upsilon) &= \int_\Omega (f+u)\upsilon dx \\ (y(\cdot,0),\upsilon) &= (y_0,\upsilon) \quad \forall \upsilon \in V, \end{aligned} \tag{4.7}$$

where

$$a(y,\upsilon) = \int_\Omega (\varepsilon \nabla y \cdot \nabla \upsilon + c(x)\cdot \nabla y \upsilon + r(x) y \upsilon) dx, \quad \forall \upsilon \in V$$

and

$$V = \{\upsilon \in H^1(\Omega) : \upsilon = 0 \quad on \quad \Gamma_D\}.$$

4.2.1 Optimize Then Discretize

For this approach, the Lagrangian of the problem is set. Then, by obtaining the partial *Fréchet* derivatives of

$$\mathcal{L} = J(y,u) + a(y,p) + b(u,p) - (f,p)$$

with respect yo y, p, u to zero, we get the necessary and sufficient optimality conditions. We have stated that the optimal control problem has a unique solution $(y,u) \in Y \times U$ [43]. The functions $(y,u) \in Y \times U$ solve the optimal control problem if and only if there exist an adjoint $p \in Y$ such that (y,u,p) satisfies the following optimality conditions [26, 31]:

$$\left(\frac{\partial y}{\partial t}, \upsilon\right) + a(y,\upsilon) + b(u,\upsilon) = (f,\upsilon), \quad \forall \upsilon \in V, \tag{4.8}$$

$$y(x,0) = y_0; \tag{4.9}$$

$$-\left(\frac{\partial p}{\partial t}, \psi\right) + a(\psi,p) = -(y-\hat{y},\psi), \quad \forall \psi \in V, \tag{4.10}$$

$$p(T) = 0; \tag{4.11}$$

$$\int_0^T (\alpha u + p, \omega - u) = 0, \quad \forall \omega \in K, \tag{4.12}$$

where $K = \{u \in L^2(0,T;U) : u_a \le u \le u_b \quad a.e.Q\}$ and inequality is substituted by the equality in case of pointwise control constraints. Indeed, (4.8)-(4.9) is the weak form the state equation, (4.10)-(4.11) is the weak form the adjoint equation with with the convection term $-c$. (4.12) corresponds to the gradient equation.

4.2.2 Discretize Then Optimize

Instead of obtaining the solution the the infinite dimensional problem (4.6), we are interested in the solution of the discretized (4.6).

As we mentioned before, we use monomial basis functions to perform the spatial discretization of (4.6) by Discontinuous Galerkin Method.

The state y and the control u are approximated by functions of the form

$$y_h(x,t) = \sum_{E\in\varepsilon_h}\sum_{i=1}^{Nloc} y_i^E(t)\phi_i^E(x), \quad \forall x \in \Omega, \quad \forall t \in (0,T), \tag{4.13a}$$

$$u_h(x,t) = \sum_{E\in\varepsilon_h}\sum_{i=1}^{Nloc} u_i^E(t)\phi_i^E(x), \quad \forall x \in \Omega, \quad \forall t \in (0,T). \tag{4.13b}$$

y_i^E's are called as the degrees of freedom which are functions of time [53]. The number of elements in the mesh is denoted by N_{el}. Indeed, the basis functions can be interpreted in detail as follows:

$$\{\phi_i^E : 1 \le i \le N_{loc}, \quad E \in \varepsilon_h\} = \{\tilde{\phi}_i^E : 1 \le j \le N_{loc}N_{el}, \quad E \in \varepsilon_h\},$$
$$\{y_i^E : 1 \le i \le N_{loc}, \quad E \in \varepsilon_h\} = \{\tilde{y}_i{}^E : 1 \le j \le N_{loc}N_{el}, \quad E \in \varepsilon_h\}.$$

We set $\bar{y}(t) = (y_0(t), \ldots, y_N(t))$ and $\bar{u}(t) = (u_0(t), \ldots, u_N(t))$.

4.2.2.1 Semi-discretization

The approximate solution $Y_h(t)$ lie in the finite-dimensional space $D^k(\xi_h)$ for all $t \ge 0$. The solution Y_h is called as the *semidiscreet* solution, or sometimes as the *continuous in time* solution [53]. The weak form of the state is discretized by DG method in space and then by θ-method in time. Let

$$D^k(\xi_h) = \{\upsilon \in L^2(\Omega) : \forall E \in \varepsilon_h, \quad \forall \upsilon_{|_E} \in P_k(E)\}$$

be the DG conforming finite elements with respect to a triangulation ε_h of the computational domain Ω and let $M_h \in \mathbb{R}^{N_h\times N_h}$ and $A_h(t) \in \mathbb{R}^{N_h\times N_h}$, $t \in [0,T]$ be the associated mass matrix, the stiffness matrix and the time interval.

The semi-discreet variational formulation of the state equation is as follows: For all $t \geq 0$, find $Y_h(t) \in D^k(\varepsilon_h)$ such that

$$(\frac{\partial Y_h}{\partial t}, \upsilon)_\Omega + a_\epsilon(Y_h(t), \upsilon) + c(Y_h(t), \upsilon)_\Omega = \quad L(t; \upsilon) + (U_h(t), \upsilon)_\Omega, \quad \forall t > 0, \quad \forall \upsilon \in D^k(\xi_h), \tag{4.14}$$

$$(Y(0), \upsilon)_\Omega = \quad (\tilde{y}_0, \upsilon)_\Omega, \quad \forall \upsilon \in D^k(\xi_h). \tag{4.15}$$

Depending on the value of the parameter γ, the method is called SIPG (γ=-1), NIPG (γ=1), or IIPG (γ=0). The initial condition $\tilde{y}_0$ can be chosen to be y_0 if y_0 belongs to the discrete space $D^k(\xi_h)$, or it can be chosen to be $\tilde{y}_0$, where $\tilde{y}$ is an approximation to y to be specified later.

Let me insert (4.13a), (4.13b) into the DG (bi)linear forms, then we obtain the system of ordinary differential equations

$$M\frac{d}{dt}\bar{y}(t) + A\bar{y}(t) + M\bar{u}(t) = F(t) + M\bar{y}(t), \quad t \in (0, T), \tag{4.16}$$

$$M\bar{y}(0) = \bar{Y}_0. \tag{4.17}$$

If we insert (4.13a), (4.13b) into (4.6), a semi-discretization of the optimal control problem (4.6) is given by

$$\int_0^T (\frac{1}{2}\bar{y}(t)^T M\bar{y}(t) - (Y_d(t))^T\bar{y}(t) + \frac{\alpha}{2}\bar{u}(t)^T M\bar{u}(t)dt)dt + \int_0^T\int_0^1 \frac{1}{2}\hat{y}^2(x,t)dx, \tag{4.18a}$$

$$M\frac{d}{dt}\bar{y}(t) + A\bar{y}(t) + M\bar{u}(t) = F(t) + M\bar{y}(t), \quad t \in (0, T), \tag{4.18b}$$

$$M\bar{y}(0) = \bar{Y}_0. \tag{4.18c}$$

The matrices $M = (M_{ij})_{ij}$, $A = (A_{ij})_{ij}$ are named as the mass and the stiffness matrices, respectively; $\forall \quad 1 \leq i, j \leq N_{loc}N_{el}$, they are defined by

$$\begin{aligned}
M = (M_{ij})_{ij} &= (\phi_j, \phi_i)_\Omega, \\
A = (A_{ij})_{ij} &= D + C + R = a_\epsilon(\phi_j, \phi_i) + c(\phi_j, \phi_i) + r(\phi_j, \phi_i), \\
(F(t))_i &= (L(t; \phi_i))_i, \\
\bar{Y}_0 &= ((\bar{y}_0, \phi_i)_\Omega)_i, \\
Y_d = (Y_d(t))_i &= (y_d(x,t), \phi_i)_\Omega.
\end{aligned}$$

The matrix M is block diagonal, symmetric positive definite, and thus invertible. The existence and uniqueness of $\vec{y}$ is obtained from the theory of ordinary differential equations [53].

4.2.2.2 Full Discretization

We now discretize the time derivative by the θ-method with respect to a partition

$$0 = t_0 < t_1 < \cdots < t_N = T$$

of the time interval $[0, T]$ with time step $\Delta t := T/N$, $N \in \mathbb{N}$. We also use the following notation for any function $y = y(t, x)$:

$$\forall \quad n \geq 0, \quad t^n = n\Delta t, \quad y^n(x) = y(t^n)(x) = y(t^n, x).$$

The full discretized state equation is as follows:

$$(\frac{Y_h^{n+1} - Y_h^n}{\Delta t}, \upsilon)_\Omega + a_\epsilon(\theta Y_h^{n+1} + (1-\theta)Y_h^n, \upsilon) + c((\theta Y_h^{n+1} + (1-\theta)Y_h^n), \upsilon)_\Omega \tag{4.19}$$

$$= \theta L(t^{n+1}; \upsilon) + (1-\theta)L(t^n; \upsilon) + \theta(U_h^{n+1}, \upsilon)_\Omega + (1-\theta)(U_h^n, \upsilon)_\Omega \quad \forall n \geq 0, \quad \forall \upsilon \in D^k(\varepsilon_h), \tag{4.20}$$

$$M\bar{y}(0) = \bar{Y}_0. \tag{4.21}$$

We expand the full discreet adjoint solution Y_h^n using the basis functions of $D^k(\xi_h)$

$$\forall \quad n \geq 0, \qquad Y_h^n = \sum_{j=1}^{N_{loc}N_{el}} \tilde{y}_j^n \tilde{\phi}_j^n.$$

The full discretized state equation is equivalent to a linear system with vector of unknowns $\tilde{y}_n = (\bar{y}_n^i)_i$. The full discrete optimal control problem:

$$\min_{\tilde{u}_0, \ldots, \tilde{u}_N} \sum_{i=0}^{N} \Delta t(\frac{1}{2}\tilde{y}_i^T M \tilde{y}_i - (Y_d(t_i))^T \tilde{y}_i + \frac{\alpha}{2}\tilde{u}_i^T M \tilde{u}_i), \tag{4.22a}$$

$$(M + \Delta t\theta A)\tilde{y}_{i+1} = (M - \Delta t(1-\theta)A)\tilde{y}_i$$

$$+\Delta t(\theta F(t_{i+1}) + (1-\theta)F(t_i)) + \Delta t(\theta M\tilde{u}_{i+1} + (1-\theta)M\tilde{u}_i), \tag{4.22b}$$

$$M\bar{y}(0) = \bar{Y}_0. \tag{4.22c}$$

$i = 0, \ldots, N$ and $y(0)$ is given. We construct the Lagrangian corresponding to (4.22a) to obtain

the optimality system:

$$\mathcal{L}(\tilde{y}_0,\ldots,\tilde{y}_N,\tilde{u}_0,\ldots,\tilde{u}_N,\tilde{p}_0,\ldots,\tilde{p}_N) = \sum_{i=0}^{N}\Delta t\left(\frac{1}{2}\tilde{y}_i^T M\tilde{y}_i - (Y_d(t_i))^T\tilde{y}_i + \frac{\alpha}{2}\tilde{u}_i^T M\tilde{u}_i\right)$$
$$+\sum_{i=0}^{N-1}\tilde{p}_{i+1}^T[(M+\Delta t\theta A)\tilde{y}_{i+1} - (M-\Delta t(1-\theta)A)\tilde{y}_i$$
$$-\Delta t(\theta F(t_{i+1}) + (1-\theta)F(t_i)) - \Delta t(\theta M\tilde{u}_{i+1} + (1-\theta)M\tilde{u}_i)]. \tag{4.23}$$

We obtain the adjoint equations by setting the partial derivatives with respect to y_i of the Lagrangian to zero:

$$\begin{aligned}(M+\Delta t\theta A)^T\tilde{p}_N &= -\frac{\Delta t}{2}(M\tilde{y}_N - (Y_d(t))_N),\\ (M+\Delta t\theta A)^T\tilde{p}_N &= (M-\Delta t(1-\theta)A)^T\tilde{p}_{N+1} - \Delta t(M\tilde{y}_i - Y_d(t_i), \quad i = N-1,\ldots,0.\end{aligned}$$

Optimize-then-discretize and discretize-then-optimize approaches differ in terms of some aspects. Although there is no basic difference between these two approaches, one of them is applied depending on the application and computational requirements related to the problem [33]. As we have mentioned, the optimize-then-discretize approach leads to an adjoint equation with a convection term $-c$. For the distributed optimal control problems governed by a steady diffusion convection reaction equation, if symmetric interior penalty Galerkin method (SIPG) is used for spatial discretization, then the optimize-then-discretize and discretize-then-optimize approaches results in equivalent formulations. However, this is not the case for NIPG and IIPG. Indeed, for the distributed optimal control problems governed by an unsteady diffusion convection reaction equation, two approaches do not commute. In addition to the difference arising from convection term of the adjoint equation, there is a difference for the adjoint evaluated at the final time. For optimize-then-discretize approach, the adjoint equation is equal to zero, that is, $p(\cdot,T) = 0$, while discretize-then-optimize approach results in the following equality for the final time

$$(M+\Delta t\theta A)^T\tilde{p}_N = -\frac{\Delta t}{2}(M\tilde{y}_N - (Y_d(t))_N).$$

However, in the literature there are some ways to make these two approaches commutative. In [1], a variant for Crank-Nicolson scheme is suggested for temporal discretization of the optimal control problem governed by parabolic PDEs. In addition to this, backward Euler is

used firstly, and then finite element space discretization is applied in [57]. thus, optimize-ten-discretize and discretize-then-optimize approaches are coincide.

During the study, we have solved the state equation forward in time, and solved the adjoint adjoint equation backward in time. As we have decrease the mesh size to approximate the solution accurately, a memory problem occurs if all data are stored. For large problems, storing all necessary data is impossible. Thus, as suggested in [17, 21], storage reducing techniques like checkpointing can be used. Checkpointing suggested in [17] is as follows: Instead of storing the coefficients for the state equation for all time steps N, store M of them. Here, $N+1$ is assumed to be a constant multiple of M. It corresponds to storing $y_0, y_M, \ldots, y_{N+1}$. To compute the coefficients of the adjoint, the coefficients of the state are needed. Thus, to compute coefficients of the adjoint p_i for $i \in \{kM+1, \ldots, (k+1)M-1\}$ and some $k \in \{0, \ldots, (N+1)/M\}$, y_i is needed which is not stored. Thus, y_{kM} is used to recompute $y_{kM}+1, \ldots, y_{k+1}M-1$. In addition, for the state equation, we have the following system of equations

$$
\begin{pmatrix}
M & & & & & \\
-D^0 & C^1 & & & & \\
 & -D^1 & C^2 & & & \\
 & & \ddots & \ddots & & \\
 & & & -D^{N-2} & C^{N-1} & \\
 & & & & -D^{N-1} & C^N
\end{pmatrix}
\begin{pmatrix}
y_0 \\ \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_{N-1} \\ \tilde{y}_N
\end{pmatrix}
$$

$$
= \begin{pmatrix}
\tilde{Y}_0 \\
\Delta t(\theta F_1 + (1-\theta)F_0) \\
\Delta t(\theta F_2 + (1-\theta)F_1) \\
\vdots \\
\Delta t(\theta F_{N-1} + (1-\theta)F_{N-2}) \\
\Delta t(\theta F_N + (1-\theta)F_{N-1})
\end{pmatrix}
+ \begin{pmatrix}
0 \\
\Delta t M(\theta \tilde{u}_1 + (1-\theta)\tilde{u}_0) \\
\Delta t M(\theta \tilde{u}_2 + (1-\theta)\tilde{u}_1) \\
\vdots \\
\Delta t M(\theta \tilde{u}_{N-1} + (1-\theta)\tilde{u}_{N-2}) \\
\Delta t M(\theta \tilde{u}_N + (1-\theta)\tilde{u}_{N-1})
\end{pmatrix}.
$$

where $D^j = M - \Delta t(1-\theta)A$ and $C^j = M + \Delta t\theta A$.

Equivalently, the adjoint equation can be written a linear system

$$\begin{pmatrix} C_0 & -D_1 & & & & & \\ & C_1 & -D_2 & & & & \\ & & C_2 & & & & \\ & & & \ddots & \ddots & & \\ & & & & C_{N-2} & -D_{N-1} & \\ & & & & & D_N & \end{pmatrix} \begin{pmatrix} \tilde{p}_0 \\ \tilde{p}_1 \\ \tilde{p}_2 \\ \vdots \\ \tilde{p}_{N-1} \\ \tilde{p}_N \end{pmatrix}$$

$$= -\Delta t \begin{pmatrix} M & & & & & & \\ & M & & & & & \\ & & M & & & & \\ & & & \ddots & & & \\ & & & & M & & \\ & & & & & M & \end{pmatrix} \begin{pmatrix} \tilde{y}_0 \\ \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_{N-1} \\ \frac{1}{2}\tilde{y}_N \end{pmatrix} + \Delta t \begin{pmatrix} (Y_d)_0 \\ (Y_d)_1 \\ (Y_d)_2 \\ \vdots \\ (Y_d)_{N-1} \\ \frac{1}{2}(Y_d)_N \end{pmatrix}.$$

where $C_j = M - \Delta t(1-\theta)A$ and $D_j = M + \Delta t\theta A$. The matrices at the right-hand side is of size $N_{loc}N_{el}(N_T+1) \times N_{loc}N_{el}(N_T+1)$.

Apart from this, if one uses discretize-then-optimize approach, then the full discrete optimal control problem can be solved by all-at-once method. Discretization of the problem and the solution via first order necessary optimality condition on a Lagrangian results in a linear system. the system is in the form of a saddle point problem. Due to the high dimension of this system, iterative methods are preferable. To accelerate the convergence of the method, a preconditioner is used. Details related to the all-at-once can be found in [52, 56] for PDE constrained optimization problems.

CHAPTER 5

OPTIMIZATION METHODS

Consider the following problem

$$\min_{u \in U} \widehat{J}(u), \text{s.t.} \quad e(y,u) = 0, u \in U. \tag{5.1}$$

No bounds are inserted on the control u, so this problem is an unconstrained optimization problem. A general framework of the optimization algorithms is as follows by [47]: Firstly, an initial value is set. Then, a sequence of iterations is generated. Depending on the method, the value of the objective function or the previous iterations are needed to obtain the mew iterate. Then we decide whether accurate solution is reached or not to terminate the algorithm. Unconstrained optimization methods are line search and trust region method. During the study, we have used line search methods. Thus, we give detail only about line search methods. Idea behind the line search methods lies in obtaining a search direction p_k and a step length α_k. Values of them are critical for the efficiency of the method. The search direction can be obtained by using different schemes, but most of the algorithms requires p_k to be a descent direction. For example, p_k must be chosen such a way that $p_k^T \nabla \widehat{J}_k < 0$. Another way to choose it is $p_k = -B_k^{-1} \nabla \widehat{J}_k$. In the steepest descent method, $B_k = I$, in the Newton's method $B_k = \nabla^2 \widehat{J}_k$. It is critical to choose the step size efficiently. Thus, the following Armijo condition can be used to determine α_k

$$\widehat{J}(u_k + \alpha p_k) \leq \widehat{J}(u_k) + c_1 \alpha \nabla \widehat{J}_k^T p_k, \tag{5.2}$$

where $c \in (0,1)$. By this formula, we can say that the step length α_k and the directional derivative $\nabla \widehat{J}_k^T p_k$ plays an important role in the reduction of $\widehat{J}_k$. However, the sufficient decrease cannot guarantee the steps to be acceptable. Thus, curvature condition must be satisfied in order to eliminate the too short step sizes by the following inequality

$$\nabla \widehat{J}(u_k + \alpha_k p_k)^T p_k \geq c_2 \nabla \widehat{J}_k^T p_k, \tag{5.3}$$

where $c_2 \in (c_1, 1)$. Apart from the line search method, Newton's method can be sued in order to obtain the search direction p_k by

$$\nabla^2 \widehat{J_k} s_k = -\nabla \widehat{J_k}. \tag{5.4}$$

We have followed the idea in [17]. Thus, the Newton equation is solved by conjugate gradient(CG) method. Let me state basics of (CG) method. We consider the linear system of equations $Ax = b$ with a symmetric positive-definite matrix A. Then the linear system can be written as a minimization problem of $\frac{1}{2}x^T Ax - b^T x$. Solution of two problems are the same and unique. It is necessary to mention a theorem given in [17].

Theorem 5.0.1 *Assume that $e_y(y(u), u)$ be invertible and $\nabla^2 \widehat{J}(u)$ be symmetric positive semidefinite. The solution of $\nabla^2 \widehat{J}(u_k)s_k = -\nabla \widehat{J}(u_k)$ is the vector s_k if and only if*

$$\min \begin{pmatrix} \nabla_y \widehat{J}(y,u)^T \\ \nabla_u \widehat{J}(y,u)^T \end{pmatrix}^T \begin{pmatrix} s_y \\ s_u \end{pmatrix} + \frac{1}{2} \begin{pmatrix} s_y \\ s_u \end{pmatrix}^T \begin{pmatrix} \nabla_{yy} L(y,u,p) & \nabla_{yu} L(y,u,p) \\ \nabla_{uy} L(y,u,p) & \nabla_{uu} L(y,u,p) \end{pmatrix} \begin{pmatrix} s_y \\ s_u \end{pmatrix}$$

is solved for (s_y, s_u) with $s_y = e_y(y(u), u)^{-1} c_u(y(u), u)s_u$ such that
$e_y(y, u)s_y + c_u(y, u)s_u = 0$, where $y = y(u)$ and $p = p(u)$.

Thus, the minimization problem in the theorem is equivalent to $\frac{1}{2}x^T Ax - b^T x$ with $x_k = s_u$, $b = \nabla \widehat{J_k}$ and $A = \nabla^2 \widehat{J_k}$.

In case of boundaries inserted on the control u, then the space of controls are defined as U_{ad}. Then, the problem is converted into an inequality constrained optimization problem. Active-set methods, gradient based methods and interior-point methods can be used to solve the inequality constraint problems. Active-set method is an efficient tool to solve small-to medium scale problems. The method is started by initiating the optimal active set. If this guess is wrong, then gradient and Lagrange multiplies information is used to updating the active set by dropping an index from the chosen active-set and by inserting a new index.

5.1 Unconstrained Optimal Control Problem

For this study, we have followed the idea and use the MATLAB programs given [17]. One-dimensional unconstrained optimal control problem has been solved by Newton-Conjugate

Gradient method with Armijo line-search. Two-dimensional constrained optimal control problem has been solved by active-set method.

The Newton-CG Algorithm with Armijo-Line Search

The conjugate gradient (CG) method is used to approximate the Newton equation

$$\nabla^2 \widehat{J}(u_k) s_k = -\nabla \widehat{J}(u_k). \tag{5.5}$$

If the Newton system is close to zero, which means that

$$\|\nabla^2 \widehat{J}(u_k) s_k + \nabla \widehat{J}(u_k)\|_2 \leq \eta_k \|\nabla \widehat{J}(u_k)\|_2, \tag{5.6}$$

$\eta_k \in (0, 1)$, or in case of a direction of negative curvature, the CG method is terminated. After having computed the direction s_k, we apply a simple Armijo line-search procedure to determine the step-size α_k. The Newton-CG algorithm is as follows [17, 47]:

Algorithm

1. Given u_0 and gtol > 0. Initialize $k = 0$.
2. Evaluate $\nabla \widehat{J}(u_k)$.
3. If $\|\nabla \widehat{J}(u_k)\| <$ gtol, stop.
4. Evaluate $\nabla^2 \widehat{J}(u_k)$.
5. To determine an approximate solution of the Newton equation $\nabla^2 \widehat{J}(u_k) s_k = -\nabla \widehat{J}(u_k)$, perform the CG method to compute :
   (a) Choose $\eta_k \in (0, 1)$, $s_k = 0$ and $p_{k,0} = r_{k,0} = -\nabla \widehat{J}(u_k)$.
   (b) For $i = 0, 1, 2, \ldots$ do
      i. If $\|r_{k,i}\|_2 < \eta_k \|r_{k,0}\|_2$ go to 5.3.
      ii. Evaluate $q_{k,i} = \nabla^2 \widehat{J}(u_k) p_i$.
      iii. If $p_{k,i}^T q_{k,i} < 0$ go to 5.3.
      iv. $\gamma_{k,i} = \|r_{k,i}\|^2 / p_{k,i}^T q_{k,i}$.
      v. $s_k = s_k + \gamma_{k,i} p_{k,i}$.
      vi. $r_{k,i+1} = r_{k,i} - \gamma_{k,i} q_{k,i}$.
      vii. $\beta_{k,i} = \|r_{k,i+1}\|^2 / \|r_{k,i}\|^2$.
      viii. $p_{k,i+1} = r_{k,i+1} + \beta_{k,i} p_{k,i}$.

(c) If $i = 0$, set $s_k = -\nabla\widehat{J}(u_k)$.

6. Perform Armijo line-search.

   (a) Choose $\alpha_k = 1$ and evaluate $f(u_k + \alpha_k s_k)$.

   (b) While $f(u_k + \alpha_k s_k) > f(u_k) + 10^{-4}\alpha_k s_k^T \nabla\widehat{J}(u_k)$ do

      i. Set $\alpha_k = \alpha_k/2$ and evaluate $f(u_k + \alpha_k s_k)$.

   (c) Set $u_{k+1} = u_k + \alpha_k s_k$, $k \longleftarrow k + 1$. Go to 2.

Gradient Computation Using Adjoints

Algorithm

1. Given u, determine the solution of $e(y, u) = 0$ for y.

2. Determine the solution, $p(u)$, of the adjoint equation $e_y(y(u), u)^T p = -\nabla_y J(y(u), u)$ for p.

3. Evaluate $\nabla\widehat{J}(u) = \nabla_u J(y(u), u) + e_u(y(u), u)^T p(u)$.

Hessian-Times-Vector Computation

We compute the Hessian of $\widehat{J}$ because J and e has been assumed to be twice continuously differentiable. In order to facilitate the computation of the Hessian, the Newton-CG algorithm has been suggested. Therefore, Hessian-times-vector products $\nabla^2\widehat{J}(u)v$ can be evaluated to lessen the disadvantage of the expensive nature of the Hessian computation.

Algorithm

1. Given u, obtain the solution of $e(y, u) = 0$ for y. Let me call the solution as $y(u)$.

2. Determine the solution of the adjoint equation $e_y(y(u), u)^T p = -\nabla_y J(y(u), u)$ for p. Let me call the solution as $p(u)$.

3. Determine the solution $e_y(y(u), u)w = e_u(y, u)v$.

4. Determine the solution $e_y(y(u), u)^T p = \nabla_{yy}\mathcal{L}(y(u), u, p(u))w - \nabla_{yu}\mathcal{L}(y(u), u, p(u))v$.

5. Evaluate $\nabla^2\widehat{J}(u)v = e_u(y(u), u)^T p - \nabla_{uy}\mathcal{L}(y(u), u, p(u))w + \nabla_{uu}\mathcal{L}(y(u), u, p(u))v$.

Given $\tilde{u}_0, \ldots, \tilde{u}_N$ and $\tilde{y}_0$, compute $\tilde{y}_1, \ldots, \tilde{y}_N$ by solving the full discrete state equation. Then, compute $\tilde{p}_N, \ldots, \tilde{p}_0$ by solving the full discrete adjoint equation. From step 3, the above algorithm can be adapted to our problem as

1. Compute $\tilde{w}_1, \ldots, \tilde{w}_N$ from
   $(M + \Delta t\theta(A)^T \tilde{w}_{i+1} = (M - \Delta t(1-\theta)A)^T \tilde{w}_i - \Delta t M(\theta\tilde{v}_{i+1} + (1-\theta)\tilde{v}_i)$,
   $i = 0, \cdots, N-1$, where $\tilde{w}_0 = 0$.

2. Compute $\tilde{p}_N, \ldots, \tilde{p}_0$ by solving
   $(M + \Delta t\theta A)^T \tilde{p}_{i+1} = \frac{\Delta t}{2} M\tilde{w}_{i+1}$,
   $(M + \Delta t\theta A)^T \tilde{p}_i = (M - \Delta t(1-\theta)A)^T \tilde{p}_{i+1} + \Delta t M\tilde{w}_i$,
   for $i = N-1, \cdots, 0$.

3. Compute $\nabla_u^2 \widehat{J}(v)$.

5.2 Constrained Optimal Control Problem

For given J and e, the problem can be formulated as follows:

$$\widehat{J}(u) = J(y(u), u), \tag{5.7}$$

where $y(u) \in \mathbb{R}^{n_y}$ is the solution of an equation

$$e(y, u) = 0, \tag{5.8}$$

$$u_a \leq u \leq u_b. \tag{5.9}$$

We have motivated to apply the active set strategy given in [46]. The algorithm that we have used is as follows. The vectors $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{u}}$ are obtained by listing the solution of $\bar{p}$ and $\bar{u}$ for each time steps column by column.

Algorithm

1. Initialize $u = 0$.

2. Set $A^- = A^+ = \emptyset$.

3. Set $\mathcal{J} = \Omega$.

4. Set $n = 0$.

5. While $n < n_{max}$

   - Determine the solution of $e(y, u) = 0$ for y.
   - Determine the solution, $p(u)$, of the adjoint equation $e_y(y(u), u)^T p = -\nabla_y J(y(u), u)$ for p.
   - Determine the solution of

$$\alpha \Delta t C_{1/2} \tilde{\mathbf{u}} + \Delta t \text{diag}(\mathcal{X}_{\mathcal{J}}) C_\theta \tilde{\mathbf{p}} = \alpha C_{1/2} (\Delta t \mathcal{X}_{\mathcal{A}^-} \cdot \tilde{\mathbf{u}}_a + \Delta t \mathcal{X}_{\mathcal{A}^+} \cdot \tilde{\mathbf{u}}_b),$$

     where

$$C_{1/2} = \begin{pmatrix} 1/2 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1/2 \end{pmatrix},$$

$$C_\theta = \begin{pmatrix} \theta & & & & & \\ (1-\theta) & \theta & & & & \\ & (1-\theta) & \theta & & & \\ & & \ddots & \ddots & & \\ & & & & (1-\theta) & \theta \end{pmatrix}.$$

   - Define $\mathcal{A}^- = \{x \in \Omega : -\tilde{\mathbf{p}} - \alpha \tilde{\mathbf{u}}_a < 0\}$.
   - Define $\mathcal{A}^+ = \{x \in \Omega : -\tilde{\mathbf{p}} - \alpha \tilde{\mathbf{u}}_b > 0\}$.
   - Define $\mathcal{J} = \Omega \setminus (\mathcal{A}^- \cap \mathcal{A}^+)$.
   - Define $\delta_n = \alpha^2 \|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_a\|^2_{L^2(\mathcal{A}^-)} + \alpha^2 \|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_b\|^2_{L^2(\mathcal{A}^+)} + \|\tilde{\mathbf{p}} - \alpha \tilde{\mathbf{u}}\|^2_{L^2(\mathcal{J})}$.
   - Stop if $\delta_n < \sqrt{\varepsilon}$ and $\delta_n = \delta_{n-1}$.

CHAPTER 6

A PRIORI ERROR ANALYSIS

The state, adjoint and the control can be discretized by using different functions to improve the accuracy of the method. One way is to use the same order of basis polynomial to approximate the state, adjoint and the control. The control can be approximated by piecewise constants as in [26, 34]. The other ways is to use the control u without discretization in order to get rid of the discretization error and increase the order of accuracy [34]. For the steady-state optimal control problem, [34] gives the following orders for different approximations

$$\alpha\|u-u_h\|_U + \|y-y_h\|_{L^2(\Omega)} \leq \begin{cases} Ch & \text{for piecewise constants,} \\ Ch^{3/2} & \text{for continuous and piecewise linear } u_h, \\ Ch^2 & \text{for variational discretization.} \end{cases}$$

In the literature, there are several ways to define DG bilinear and linear forms. As we mentioned before, the bilinear form coming from the diffusion part is as follows by [53]:

$$a_\epsilon(y,\upsilon) = \sum_{E\in\xi_h}\int_E \varepsilon\nabla y\cdot\nabla\upsilon dx - \sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{\varepsilon\nabla y\cdot n_e\}[\upsilon]ds \tag{6.1}$$

$$+ \gamma\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{\varepsilon\nabla\upsilon\cdot n_e\}[y]ds + J_0^{\sigma_0,\beta_0}(y,\upsilon). \tag{6.2}$$

where

$$J_0^{\sigma_0,\beta_0}(y,\upsilon) = \sum_{e\in\Gamma_h\cup\Gamma_D}\frac{\sigma_\epsilon^0}{|e|^{\beta_0}}\int_e [y][\upsilon]ds.$$

In addition to this, this part can be defined as [38]

$$\tilde{a}(\omega,\upsilon) = \sum_{E\in\xi_h}\int_E \varepsilon\nabla y\cdot\nabla\upsilon dx + \int_{\Gamma_D}(\omega((\varepsilon\nabla\upsilon)\cdot\mu) - ((\varepsilon\nabla\upsilon)\cdot\mu)\upsilon)ds \tag{6.3}$$

$$+ \int_{\Gamma_{int}}([\omega]\{(\varepsilon\nabla\upsilon)\cdot\nu\} - \{(\varepsilon\nabla\omega)\cdot\nu\}[\upsilon])ds \tag{6.4}$$

$$+ \int_{\Gamma_D}\sigma\omega\upsilon ds + \int_{\Gamma_{int}}\sigma[\omega][\upsilon]ds, \tag{6.5}$$

where σ is the discontinuity penalization parameter and $\nu = \mu$ for boundary edges and for given Dirichlet and Neumann boundary conditions.

For the convection term, we have used the definition in [38]:

$$c(y,\upsilon) = \sum_{E\in\xi_h}\left(\int_E c\nabla y\cdot\upsilon dx - \int_{\partial_- E\setminus\Gamma}(c\cdot n_e)(y^+ - y^-)\upsilon^+ ds - \int_{\partial_- E\cap\Gamma_-}(c\cdot n_e)y^+\upsilon^+ ds\right), \tag{6.6}$$

The convection term is approximated by an upwind discretization by [53]

$$c(y,\upsilon) = -\sum_{E\in\xi_h}\int_E cy\cdot\nabla\upsilon dx + \int_{\Gamma_h}(c\cdot n_e)y^{up}[\upsilon]ds + \int_{\Gamma_{out}}(c\cdot n_e)y\upsilon ds, \tag{6.7}$$

where ω^{up} which is the upwind value of the function is written $\forall e = \partial E_e^1 \cap E_e^2$

$$\omega^{up} = \begin{cases} \omega_{|_{E_e^1}} & \text{if} \quad u\cdot n_e \geq 0 \\ \omega_{|_{E_e^2}} & \text{if} \quad u\cdot n_e < 0 \end{cases}$$

.

With this definitions, the method is consistent, which will be explained next section, with a suitable right-hand side

$$L(\upsilon) = \sum_{E\in\xi_h}\left(\int_E f\upsilon dx - \int_{\partial_- E\cap\Gamma_-}(c\cdot n_e)g_D\upsilon^+ ds\right) + \sum_{e\in\Gamma_D}\int_e(\gamma\varepsilon\nabla\upsilon\cdot n_e + \frac{\sigma_e^0}{|e|^{\beta_0}}\upsilon)g_D ds + \sum_{e\in\Gamma_N}\int_e \upsilon g_N ds, \tag{6.8}$$

while in [59]

$$L(\upsilon) = \int_E f\upsilon dx. \tag{6.9}$$

For the stability and the convergence estimates, the reaction term can be inserted into the term coming from the convection part as in [38]. As different from [38, 59], for our case, $(c_0(x))^2 = r(x) - \frac{1}{2}\nabla\cdot c(x) \geq 0$ in Ω as in [49], which is the coercivity condition. $(c_0(x))^2$ is zero, automatically, because r and c are constants.

6.1 Consistency of DG method

The interior penalty DG method inserts some terms coming from the penalty parameters, jump terms and Dirichlet boundary conditions. Thus, each term at the right-hand side of the DG variational formulation must be matched to the terms the left-hand side. In other words, we need to confirm the equivalence of the weak form and the PDE. The terms coming from the $a_\epsilon(\cdot,\cdot)$ are matches the ones at the right-hand side $L(t;\upsilon)$ by [53]. In addition, the terms of the convection parts are consistent with the ones at the right-hand side by [38]. Thus, the method is consistent.

6.2 Error Analysis For The State Equation

6.2.1 Stability Estimates For The Semi-discrete State

To make the error analysis easy to understand, it is beneficial to proceed term by term. Firstly, a priori error analysis for the diffusion equation with Dirichlet boundary condition can be performed as in [53] in detail. Secondly, a priori error analysis for the convection equation with Dirichlet boundary condition is derived as in [38].

6.2.1.1 Stability Estimates For Diffusion Equation

Consider the semi-discrete DG variational formulation of the diffusion equation:

$$\forall t>0, \quad \forall \upsilon \in D_k(\xi_h), \quad \left(\frac{\partial Y_h}{\partial t},\upsilon\right)_\Omega + a_\epsilon(Y_h,\upsilon) = L(t;v), \tag{6.10}$$

$$\forall \upsilon \in D_k(\xi_h), \quad (Y_h(0),\upsilon)_\Omega = (y_0,\upsilon)_\Omega; \tag{6.11}$$

where the DG bilinear from and the right-hand side are as follows:

$$a_\epsilon(\omega,\upsilon) = \sum_{E\in\xi_h}\int_E \varepsilon\nabla\omega\cdot\nabla\upsilon dx - \sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{\varepsilon\nabla\omega\cdot n_e\}[\upsilon]ds \tag{6.12}$$

$$+ \gamma\sum_{e\in\Gamma_h\cup\Gamma_D}\int_e \{\varepsilon\nabla\upsilon\cdot n_e\}[\omega]ds + \sum_{e\in\Gamma_h\cup\Gamma_D}\frac{\sigma_\epsilon^0}{|e|^{\beta_0}}\int_e [\omega][\upsilon]ds, \tag{6.13}$$

$$L(\upsilon) = \sum_{E\in\xi_h}\int_E f(t)\upsilon dx + \sum_{e\in\Gamma_D}\int_e (\gamma\varepsilon\nabla\upsilon\cdot n_e + \frac{\sigma_e^0}{|e|^{\beta_0}}\upsilon)g_D(t)ds. \tag{6.14}$$

We choose $\upsilon = Y_h(t)$ at (6.11). If $\beta_0(d-1) \geq 1$ and and σ_e^0 is bounded below by a constant, then a_ϵ is coercive by [53]. We use coercivity for the left-hand side of (6.11), apply the trace inequality and the Cauchy-Schwarz's inequality for the right-hand side of (6.11) to obtain the following inequality for a constant C independent of h:

$$\frac{1}{2}\frac{d}{dt}\|Y_h\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|Y_h\|^2_\epsilon \leq \|f(t)\|^2_{L^2(\Omega)}\|Y_h(t)\|^2_{L^2(\Omega)} + C\sum_{e\in\Gamma_D}\frac{1}{|e|^{\beta_0}}\|g_D(t)\|^2_{L^2(e)}.$$

Then, we apply Young's inequality to $\|f(t)\|^2_{L^2(\Omega)}\|Y_h(t)\|^2_{L^2(\Omega)}$ by choosing the Young's inequality constant as 1. Then, we multiply the equation by 2 and integrate from 0 to t. In order to eliminate the term $\int\|Y_h(s)\|$ at the right-hand side, we apply the continuous Gronwall's inequality to obtain

$$\|Y_h\|^2_{L^2(\Omega)} + \kappa\int_0^t\|Y_h\|^2_\epsilon \leq C\left(\int_0^t\|f(s)\|^2_{L^2(\Omega)} + \|Y_h(0)\|^2_{L^2(\Omega)} + \sum_{e\in\Gamma_D}\frac{1}{|e|^{\beta_0}}\|g_D(t)\|^2_{0,e}\right),$$

C increases exponentially in time. The final result follows for a positive constant C independent of h:

$$\|Y_h\|^2_{L^\infty(0,T;L^2(\Omega))} + \int_0^T\|Y_h\|^2_\epsilon \leq C\|y_0\|^2_{L^2(\Omega)} + C\|f(s)\|^2_{L^2(0,T;L^2(\Omega))} + C\sum_{e\in\Gamma_D}\frac{1}{|e|^{\beta_0}}\|g_D(t)\|^2_{L^2(0,T;L^2(e))}.$$

6.2.1.2 Stability Estimates For Convection-Reaction Equation

Secondly, we obtain the a priori error estimates for the convection-reaction equation with Dirichlet boundary condition. Consider the following DG variational formulation

$$\forall t>0,\quad \forall\upsilon\in D_k(\xi_h),\quad \left(\frac{\partial Y_h}{\partial t},,\upsilon\right)_\Omega + c(Y_h,\upsilon) = L(t;v), \tag{6.15}$$

$$\forall\upsilon\in D_k(\xi_h),\quad (Y_h(0),\upsilon)_\Omega = (y_0,\upsilon)_\Omega; \tag{6.16}$$

where the DG bilinear form and the right-hand side are defined as [38]:

$$c(\omega,\upsilon) = \sum_{E\in\xi_h}\left(\int_E\mathcal{L}_0\omega\cdot\upsilon dx - \int_{\partial_- E\setminus\Gamma}(c\cdot n_e)(\omega^+-\omega^-)\upsilon^+ds - \int_{\partial_- E\cap\Gamma_-}(c\cdot n_e)\omega^+\upsilon^+ds\right), \tag{6.17}$$

$$L(t;v) = \sum_{E\in\xi_h}\left(\int_E f\upsilon dx - \int_{\partial_- E\cap\Gamma_-}(c\cdot n_e)g_D\upsilon^+\right)ds, \tag{6.18}$$

where $\mathcal{L}_0 = c\cdot\nabla y + ry$. A bound for the semi-discrete solution of the steady convection equation is given by [38]

$$\sum_{E\in\xi_h}\left(\|c_0Y_h\|^2_{L^2(E)} + \frac{1}{2}\|Y_h^+\|^2_{\partial_- E\cap\Gamma_-} + \|Y_h^+ - Y_h^+\|^2_{\partial_- E\setminus\Gamma} + \|Y_h^+\|^2_{\partial_+ E\cap\Gamma}\right) \leq \sum_{E\in\xi_h}\left(\|c_0^{-1}f\|^2_{L^2(E)} + 2\|g_D\|^2_{\partial_- E\cap\Gamma_-}\right),$$

where $\|\cdot\|_E$ corresponds to the (semi)norm for the following (semi)inner-product

$$(\upsilon,\omega)_E = \int_E |c\cdot n_e|\upsilon\omega ds$$

and $c_0 \geq 0$. A similar process is observed at [59],too. For our case, $c_0 = 0$. Thus, the bound above must be modified to make the fraction $1/c_0$ well-defined. In addition, instead of using the (semi)norm defined above, $L2$-norm is preferred to be able use the error estimates in given in [53] with respect to L^2-norm.

We choose $\upsilon = Y_h(t)$ in (6.16). There is no need to change anything at right-hand side. The first difference from [38] arises from the right-hand side. We apply Cauchy-Schwarz's and Young's inequality to the integrals

$$\int fY_h(t) \quad \text{and} \quad \int (c\cdot n_e)g_D Y_h(t)^+.$$

The constants used in Young's inequality are 1 and 2 for these integrals, respectively. Then, we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|Y_h\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\|c_0(x)Y_h\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)\\
&+\sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+ - Y_h^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma)}\right)\\
&\leq \sum_{E\in\xi_h}\left(\frac{1}{2}\|f\|^2_{L^2(E)} + \frac{1}{2}\|Y_h\|^2_{L^2(E)} + \frac{1}{4}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} + \||c\cdot n_e|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right).
\end{aligned}$$

If we arrange the terms, then we obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|Y_h\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\|c_0(x)Y_h\|^2_{L^2(E)} + \frac{1}{4}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)\\
&+\sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+ - Y_h^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma)}\right)\\
&\leq \sum_{E\in\xi_h}\left(\frac{1}{2}\|f\|^2_{L^2(E)} + \frac{1}{2}\|Y_h\|^2_{L^2(E)} + \||c\cdot n_e|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right).
\end{aligned}$$

Then, we multiply the inequality by 2 and integrate from 0 to t. In order to eliminate the term $\int\|Y_h(s)\|$ arising from the right-hand side, continuous Gronwall's inequality can be applied

with a constant C increasing exponentially in time:

$$\begin{aligned}
&\|Y_h\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\int_0^T \|c_0(x)Y_h\|^2_{L^2(\Omega)} + \frac{1}{2}\int_0^T \||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)\\
&+\sum_{E\in\xi_h}\left(\int_0^T \||c\cdot n_e|^{1/2}Y_h^+ - Y_h^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \int_0^T \||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma)}\right)\\
&\le C\left(\|Y_h(0)\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\int_0^T \|f\|^2_{L^2(\Omega)} + 2\int_0^T \||c\cdot n_e|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)\right).
\end{aligned}$$

Similar to the diffusion equation, we obtain

$$\begin{aligned}
&\|Y_h\|^2_{L^\infty(0,T;L^2(\Omega))} + \sum_{E\in\xi_h}\left(\|c_0(x)Y_h\|^2_{L^2(0,T;L^2(\Omega))} + \frac{1}{2}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\Omega)(0,T;L^2(\partial_- E\cap\Gamma_-))}\right)\\
&+\sum_{E\in\xi_h}\left(\||c\cdot n_e|^{1/2}Y_h^+ - Y_h^-\|^2_{L^2(0,T;L^2(\partial_- E\setminus\Gamma))} + \||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(0,T;L^2(\partial_- E\cap\Gamma))}\right)\\
&\le C\|y_0\|^2_{L^2(\Omega)} + C\sum_{E\in\xi_h}\left(\|f\|^2_{L^2(0,T;L^2(\Omega))} + 2\||c\cdot n_e|^{1/2}g_D\|^2_{L^2(0,T;L^2(\partial_- E\cap\Gamma_-))}\right).
\end{aligned}$$

Stability Estimates For The Semi-discrete State

Let us provide stability estimates for the semi-discrete state equation by combining the bounds that we have obtained up to now.

Lemma 6.2.1 *Assume that $\beta_0 \ge (d-1)^{-1}$. There exists a constant $C > 0$ independent of h such that*

$$\begin{aligned}
\|Y_h\|^2_{L^\infty(0,T;L^2(\Omega))} + \kappa\int_0^T \|Y_h\|^2_\epsilon \le \tilde{C}&\left(\|\tilde{y}_0\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\|f\|^2_{L^2(0,T;L^2(E))}\right)\\
&+\tilde{C}\left(\sum_{E\in\xi_h}\||c\cdot n|^{1/2}g_D\|^2_{L^2(0,T;L^2(\partial_- E\cap\Gamma_-))} + \sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D\|^2_{L^2(0,T;L^2(e))}\right).
\end{aligned} \tag{6.19}$$

where C increases exponentially in time.

Proof. Choose $\upsilon = Y_h(t)$ in the semi-discrete variational formulation of the state equation to obtain

$$\left(\frac{\partial Y_h}{\partial t}, Y_h(t)\right)_\Omega + a_\epsilon(Y_h(t), Y_h(t)) + c(Y_h(t), Y_h(t)) = L(t; Y_h(t)).$$

Let me insert $r(\cdot,\cdot)$ into $c(\cdot,\cdot)$ to facilitate the procedure. Then, by [38, 53], the following bounds can be obtained by using the coercivity of a_ϵ, Young's and Cauchy-Schwartz's in-

equality:

$$\frac{1}{2}\frac{d}{dt}\|Y_h\|^2_{L^2(\Omega)} + \kappa\|Y_h(t)\|^2_\epsilon + \sum_{E\in\xi_h}\left(\|c_0 Y_h(t)\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)$$
$$+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}(Y_h^+ - Y_h^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right)$$
$$\leq \sum_{E\in\xi_h}\left(\frac{1}{2}\|f\|^2_{L^2(E)} + \frac{1}{2}\|Y_h(t)\|^2_{L^2(E)} + \frac{1}{4}\||c\cdot n_e|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} + \||c\cdot n|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)$$
$$+ \frac{\kappa}{2}\|Y_h\|^2_\epsilon + 2C\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D\|^2_{L^2(e)}.$$

We arrange the terms and multiply the above inequality by 2 and integrate from 0 to t:

$$\|Y_h\|^2_{L^2(\Omega)} + \kappa\int_0^t\|Y_h(s)\|^2_\epsilon + \sum_{E\in\xi_h}\left(2\int_0^t\|c_0 Y_h\|^2_{L^2(E)} + \frac{1}{2}\int_0^t\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)$$
$$+ \sum_{E\in\xi_h}\left(\int_0^t\||c\cdot n|^{1/2}(Y_h^+ - Y_h^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \int_0^t\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right)$$
$$\leq \|y_0\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\int_0^t\|f\|^2_{L^2(E)} + \int_0^t\|Y_h(t)\|^2_{L^2(E)} + 2\int_0^t\||c\cdot n|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)$$
$$+ 4C\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\int_0^t\|g_D\|^2_{L^2(e)}.$$

We arrange the terms and use the continuous Gronwall's inequality to get rid of $\int_0^t\|Y_h(t)\|^2_{L^2(E)}$ with a constant $\tilde{C}$ increasing exponentially in time.

$$\|Y_h\|^2_{L^2(\Omega)} + \kappa\int_0^t\|Y_h(s)\|^2_\epsilon + \sum_{E\in\xi_h}\left(2\int_0^t\|c_0 Y_h\|^2_{L^2(E)} + \frac{1}{2}\int_0^t\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)$$
$$+ \sum_{E\in\xi_h}\left(\int_0^t\||c\cdot n|^{1/2}(Y_h^+ - Y_h^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \int_0^t\||c\cdot n|^{1/2}Y_h^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right)$$
$$\leq \tilde{C}\left(\|y_0\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\int_0^t\|f\|^2_{L^2(E)} + 2\int_0^t\||c\cdot n|^{1/2}g_D\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) + 4C\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\int_0^t\|g_D\|^2_{L^2(e)}\right).$$

Remark: As we decrease the mesh size h, the last term at the right-hand side of the above inequality increases rapidly. This occurs because we have inserted the Dirichlet boundary conditions to the variational from weakly. If the boundary conditions are imposed strongly by setting the space of the test functions as

$$D_k^0(\xi_h) = \{\upsilon\in D_k(\xi_h) : \upsilon = 0 \quad \text{on} \quad \partial\Omega\},$$

then, the stability bound can be rewritten as follows [53]:

$$\|Y_h\|^2_{L^\infty(0,T;L^2(\Omega))} + \int_0^T\|Y_h\|^2_\epsilon \leq \tilde{C}\|\tilde{y}_0\|^2_{L^2(\Omega)} + \tilde{C}\sum_{E\in\xi_h}\|f\|^2_{L^2(0,T;L^2(E))}.$$

6.2.2 Error Estimates For The Semi-discrete State

6.2.2.1 Error Estimates For Diffusion Equation

For diffusion equation, let me mention how to obtain error bounds as in [53]. The global error $y - Y_h$ can be analyzed as

$$y - Y_h = (y - \tilde{y}) - (Y_h - \tilde{y}) = \eta + \xi$$

. $\eta = y - \tilde{y}$ is the elliptic projection, that is,

$$a_\epsilon(y - \tilde{y}, \upsilon) = 0, \quad \forall t \geq 0, \quad \forall \upsilon \in D_k(\xi_h).$$

The scheme is consistent, then we can write

$$\left(\frac{\partial \xi}{\partial t}, \upsilon\right)_\Omega + a_\epsilon(\xi, \upsilon) = \left(\frac{\partial \eta}{\partial t}, \upsilon\right)_\Omega + a_\epsilon(\eta, \upsilon), \quad \forall t \geq 0, \quad \forall \upsilon \in D_k(\xi_h). \tag{6.20}$$

By the elliptic projection, this equation can be rewritten as

$$\left(\frac{\partial \xi}{\partial t}, \upsilon\right)_\Omega + a_\epsilon(\xi, \upsilon) = \left(\frac{\partial \eta}{\partial t}, \upsilon\right)_\Omega, \quad \forall t \geq 0, \quad \forall \upsilon \in D_k(\xi_h). \tag{6.21}$$

We choose $\upsilon = \xi$. We find a bound for the left-hand side by using the coercivity of a_ϵ with the condition $\beta_0(d-1) \geq 1$. The left-hand side is bounded as in [53] by Cauchy-Schwartz's and Young's inequalities for positive penalty parameters σ_0^e:

$$\frac{1}{2}\frac{d}{dt}\|\xi\|^2_{L^2(\Omega)} + \kappa\|\xi\|^2_\epsilon \leq \frac{\kappa}{2}\|\xi\|^2_\epsilon + \frac{1}{2\kappa}\left\|\frac{\partial(y-\tilde{y})}{\partial t}\right\|^2_{L^2(\Omega)}.$$

For the right-hand side, we use the error estimates for the elliptic projection given by [53]: If $y \in L^2(0,T;H^s(\xi_h))$ for $s > 3/2$, then

$$\forall t \geq 0, \quad \|y(t) - \tilde{y}(t)\|_\epsilon \leq Ch^{\min(k+1,s)-1}|||y(t)|||_{H^s(\xi_h)}.$$

If Ω is convex, then the following error estimates are valid

$$\forall t \geq 0, \quad \|y(t) - \tilde{y}(t)\|_{L^2(\Omega)} \leq Ch^{\min(k+1,s)}|||y(t)|||_{H^s(\xi_h)}, \quad \text{for} \quad \text{SIPG},$$

$$\forall t \geq 0, \quad \|y(t) - \tilde{y}(t)\|_{L^2(\Omega)} \leq Ch^{\min(k+1,s)-1}|||y(t)|||_{H^s(\xi_h)}, \quad \text{for} \quad \text{NIPG/IIPG}.$$

Now, we can use the error estimates to bound the right-hand side

$$\frac{1}{2}\frac{d}{dt}\|\xi\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|\xi\|^2_\epsilon \leq Ch^{2\min(k+1,s)-2\delta}\left\||\frac{\partial y}{\partial t}\right\|^2_{H^s(\xi_h)}.$$

We multiply the inequality by 2 and integrate from 0 to t

$$\|\xi\|^2_{L^2(\Omega)} + \kappa \int_0^t \|\xi\|^2_\epsilon \leq \|\xi(0)\|^2_{L^2(\Omega)} + Ch^{2\min(k+1,s)-2\delta} \left\| \frac{\partial y}{\partial t} \right\|^2_{L^2(0,t;H^s(\xi_h))}.$$

Then, the final result is obtained by the triangle inequality in L^2 and the energy norm are respectively,

$$\|y(t) - Y_h(t)\|_{L^2(\Omega)} \leq \|Y_h(t) - \tilde{y}(t)\|_{L^2(\Omega)} + \|y(t) - \tilde{y}(t)\|_{L^2(\Omega)},$$

$$\left(\int_0^T \|y(t) - Y_h(t)\|^2_\epsilon \right)^{1/2} \leq \left(\int_0^T \|Y_h(t) - \tilde{y}(t)\|^2_\epsilon \right)^{1/2} + \left(\int_0^T \|y(t) - \tilde{y}(t)\|^2_\epsilon \right)^{1/2}.$$

6.2.2.2 Error Estimates For Convection-Reaction Equation

Let me mention that $\eta = y - \tilde{y}$, $\xi = Y_h - \tilde{y}$. The scheme is consistent, then we can write

$$\left(\frac{\partial \xi}{\partial t}, \upsilon \right)_\Omega + c(\xi, \upsilon) = \left(\frac{\partial \eta}{\partial t}, \upsilon \right)_\Omega + c(\eta, \upsilon), \quad \forall t \geq 0, \quad \forall \upsilon \in D_k(\xi_h). \tag{6.22}$$

Then, we choose $\upsilon = \xi$ and substitute the bounds given by [38] and use L^2-norm

$$\begin{aligned}
&\sum_{E\in\xi_h} \|c_0\xi\|^2_{L^2(E)} + \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} \\
&+ \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}(\xi^+ - \xi^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+ E\cap\Gamma)} \\
&\leq 2 \sum_{E\in\xi_h} \int_E (c_0)^2 \xi\eta - \sum_{E\in\xi_h} \int_E \eta \mathcal{L}_0 \xi + \sum_{E\in\xi_h} \int_{\partial_+ E\cap\Gamma} (c\cdot n)\xi^+\eta^+ \\
&+ \sum_{E\in\xi_h} \int_{\partial_+ E\setminus\Gamma} (c\cdot n)\xi^+\eta^+ + \sum_{E\in\xi_h} \int_{\partial_- E\setminus\Gamma} (c\cdot n)\xi^+\eta^-.
\end{aligned}$$

The first term at the right-hand side can be bounded by using the Cauchy-Schwartz's and Young's inequality with the constant 1 while an upper bound for the second term at the right-hand side can be obtained by [53], and the bounds for the other terms can be found at [38]:

$$\begin{aligned}
&\sum_{E\in\xi_h} \|c_0\xi\|^2_{L^2(E)} + \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} \\
&+ \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}(\xi^+ - \xi^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+ E\cap\Gamma)} \\
&\leq \frac{1}{2} \sum_{E\in\xi_h} \|c_0\xi\|^2_{L^2(E)} + 2 \sum_{E\in\xi_h} \|c_0\eta\|^2_{L^2(E)} + \frac{\kappa}{8}\|\xi\|^2_\epsilon + C\|\eta\|^2_{L^2(\Omega)} \\
&+ \frac{1}{4} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+ E\cap\Gamma))} + \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\eta^+\|^2_{L^2(\partial_+ E\cap\Gamma)} \\
&+ \frac{1}{4} \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\xi^+ - \xi^-\|_{L^2(\partial_- E\setminus\Gamma)} + \sum_{E\in\xi_h} \||c\cdot n|^{1/2}\eta^-\|_{L^2(\partial_- E\setminus\Gamma)}.
\end{aligned}$$

We arrange the terms to obtain the following:

$$\begin{aligned}
&\frac{1}{2}\sum_{E\in\xi_h}\|c_0\xi\|^2_{L^2(E)}+\frac{1}{2}\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_-E\cap\Gamma_-)}\\
&+\frac{1}{4}\sum_{E\in\xi_h}\||c\cdot n|^{1/2}(\xi^+-\xi^-)\|^2_{L^2(\partial_-E\setminus\Gamma)}+\frac{1}{4}\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+E\cap\Gamma)}\\
&\leq 2\sum_{E\in\xi_h}\|c_0\eta\|^2_{L^2(E)}+\frac{\kappa}{8}\|\xi\|^2_\epsilon+C\|\eta\|^2_{L^2(\Omega)}\\
&+\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\eta^+\|^2_{L^2(\partial_+E\cap\Gamma)}+\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\eta^-\|_{L^2(\partial_-E\setminus\Gamma)}.
\end{aligned}$$

As one can realize, the term $\frac{\kappa}{8}\|\xi\|^2_\epsilon$ must be eliminated. At [38], this term can be got rid of by using a special interpolation. For our case, the terms coming from the diffusion equation are needed. However, it is beneficial to obtain the bounds for the rest for the terms at the right-hand side by [53] as follows: We have

$$\frac{1}{2}\sum_{E\in\xi_h}\|c_0\eta\|^2_{L^2(E)}\leq\frac{1}{2}\sum_{E\in\xi_h}c_0^2Ch_E^{\min(k+1,s)}|y|_{H^s(E)}.$$

We can bound the third term as follows by the error estimates given previously,

$$C\|\eta\|^2_{L^2(\Omega)}\leq C\tilde{C}h^{\min(k+1,s)-\delta}\|\|y(t)\|\|_{H^s(\xi_h)}.$$

Lastly, we consider $\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\eta^+\|^2_{L^2(\partial_+E\cap\Gamma)}$ and $\sum_{E\in\xi_h}\||c\cdot n|^{1/2}\eta^-\|_{L^2(\partial_-E\setminus\Gamma)}$.

In general, for each element E and $\upsilon\in H^1(E)$, the trace of υ along any side of one element E is well-defined [53]. In case of a common side e for the elements E_1^e and E_2^e, two traces of υ along e are considered, that is, $\upsilon_{|_{E_1^e}}$ and $\upsilon_{|_{E_2^e}}$. By [38], υ_E^+ is defined as the interior trace of υ_E on ∂E.The trace is taken from within E [37]. Let me mention the definition of υ_E^+ and υ_E^-:

$$y^+=\begin{cases}y_{|_{E_e^1}} & \text{if} \quad c\cdot n\geq 0\\ y_{|_{E_e^2}} & \text{if} \quad c\cdot n<0,\end{cases}\qquad y^-=\begin{cases}y_{|_{E_e^2}} & \text{if} \quad c\cdot n\geq 0\\ y_{|_{E_e^1}} & \text{if} \quad c\cdot n<0.\end{cases}$$

Firstly, we consider $\||c\cdot n|^{1/2}\eta^+\|_{L^2(\partial_+E\cap\Gamma)}$. By the definition of the interior trace, this term can be written as

$$\||c\cdot n|^{1/2}\eta^+\|_{L^2(\partial_+E\cap\Gamma)}=\begin{cases}\||c\cdot n|^{1/2}\eta\|_{L^2(E_e^1)} & \text{if} \quad c\cdot n\geq 0\\ \||c\cdot n|^{1/2}\eta\|_{L^2(E_e^2)} & \text{if} \quad c\cdot n<0.\end{cases}$$

Then, by the trace inequality as in [53], it can be bounded by

$$\leq C|e|^{1/2}\|\begin{cases}|E_e^1|^{-1/2}\||c\cdot n|^{1/2}\eta_{|_{E_e^1}}\|_{L^2(E_e^1)}+h_{|_{E_e^1}}\||c\cdot n|^{1/2}\nabla\eta_{|_{E_e^1}}\|_{L^2(E_e^1)} & \text{if} \quad c\cdot n\geq 0\\ |E_e^2|^{-1/2}\||c\cdot n|^{1/2}\eta_{|_{E_e^2}}\|_{L^2(E_e^2)}+h_{|_{E_e^2}}\||c\cdot n|^{1/2}\nabla\eta_{|_{E_e^2}}\|_{L^2(E_e^2)} & \text{if} \quad c\cdot n<0.\end{cases}$$

By using the previous theorem and the fact that for $i = 1, 2$ $|e|^{1/2}|E_e^1|^{-1/2}$ is bounded below by a constant C in $2D$, we obtain

$$\leq Ch^{\min(k+1,s)-1} \begin{cases} \|c \cdot n|^{1/2} y\|_{H^s(E_e^1)} & \text{if} \quad c \cdot n \geq 0 \\ \|c \cdot n|^{1/2} y\|_{H^s(E_e^2)} & \text{if} \quad c \cdot n < 0. \end{cases}$$

If we add them up, we obtain

$$\sum_{E \in \xi_h} \| |c \cdot n|^{1/2} \eta^+ \|^2_{L^2(\partial_+ E \cap \Gamma)} \leq Ch^{2\min(k+1,s)-2} \|c \cdot n|^{1/2} y\|^2_{H^s(\xi_h)}.$$

For the term, $\| |c \cdot n|^{1/2} \eta^- \|_{L^2(\partial_- E \setminus \Gamma)}$, a similar argument is applied. For this case, we are interested in η^- which is the exterior trace. Then, we obtain

$$\sum_{E \in \xi_h} \| |c \cdot n|^{1/2} \eta^- \|^2_{L^2(\partial_- E \setminus \Gamma)} \leq Ch^{2\min(k+1,s)-2} \|c \cdot n|^{1/2} y\|^2_{H^s(\xi_h)}.$$

Error Estimates For The Semi-discrete State

At this part, we determine the bounds for the state equation of which consists of the bounds that were found for the diffusion and convection-reaction equation separately. In order not to lose the connection between the terms, the properties and the definitions are mentioned one more time.

Lemma 6.2.2 *Suppose that $y \in L^2(0, T; H^s(\xi_h))$ and that y_0 belongs to $H^s(\xi_h)$ for $s > 3/2$ and $\beta_0(d-1) \geq 1$. Let σ_e^0 is sufficiently large for all e if SIPG and IIPG is preferred. Then, there exist a constant C independent of h such that*

$$\|y(t) - Y_h(t)\|_{L^2(\Omega)} \leq \quad Ch^{\min(k+1,s)-\delta} \left(\|y(0)\|_{L^2(\Omega)} + \|y(t)\|_{L^2(0,T;H^s(\xi_h))} + \left\| \frac{\partial y}{\partial t} \right\|_{L^2(0,T;H^s(\xi_h))} \right). \tag{6.23}$$

where $\delta = 0$ for SIPG and $\delta = 0$ for NIPG and IIPG if $\beta_0 \geq 3(d-1)^{-1}$ and $g_D \in D_k(\xi_h)$. Otherwise, $\delta = 1$ for NIPG and IIPG.

Proof. We set $\eta = y - \tilde{y}$, $\xi = Y_h - \tilde{y}$ where $\tilde{y} \in D_k(\xi_h)$ can be chosen as an elliptic projection of y defined below:

$$\forall t \geq 0, \quad \forall \upsilon \in D_k(\xi_h), \qquad a_\epsilon(y(t) - \tilde{y}(t), \upsilon) = 0.$$

Now, we obtain the following for all $\upsilon \in D_k(\xi_h)$ due to the consistent scheme:

$$\left(\frac{\partial \xi}{\partial t}, \upsilon \right)_\Omega + a_\epsilon(\xi, \eta) + c(\xi, \eta) = \left(\frac{\partial \eta}{\partial t}, \upsilon \right)_\Omega + a_\epsilon(\eta, \eta) + c(\eta, \eta). \tag{6.24}$$

We choose $\upsilon = \xi$ and use coercivity and continuity of a_ϵ, bounds given in [38, 37, 53] to obtain:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\xi\|^2_{L^2(\Omega)} + \kappa\|\xi\|^2_\epsilon + \sum_{E\in\xi_h}\left(\|c_0\xi\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right)\\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}(\xi^+ - \xi^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right)\\
&\leq \frac{\kappa}{2}\|\xi\|^2_\epsilon + \frac{1}{2\kappa}\|\frac{\partial\eta}{\partial t}\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(2\|c_0\xi|^2_{L^2(E)} + \frac{1}{2}\|c_0\eta|^2_{L^2(E)}\right)\\
&+ \frac{\kappa}{8}\|\xi\|^2_\epsilon + C\|\eta\|^2_{L^2(\Omega)}\\
&+ \sum_{E\in\xi_h}\left(\frac{1}{4}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(\partial_+ E\cap\Gamma)} + \||c\cdot n|^{1/2}\eta^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right)\\
&+ \sum_{E\in\xi_h}\left(\frac{1}{4}\||c\cdot n|^{1/2}(\xi^+ - \xi^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \||c\cdot n|^{1/2}\eta^-\|^2_{L^2(\partial_- E\setminus\Gamma)}\right).
\end{aligned}$$

We use the error estimates satisfied by the elliptic projection [53] and combine the terms that we found before. Then, we multiply the inequality by 2 and integrate from 0 to t to obtain:

$$\begin{aligned}
&\|\xi\|^2_{L^2(\Omega)} + \frac{3\kappa}{4}\int_0^t \|\xi(s)\|^2_\epsilon + \sum_{E\in\xi_h}\left(\|c_0\xi\|^2_{L^2(0,T;L^2(E))} + \||c\cdot n|^{1/2}\xi^+\|^2_{L^2(0,T;L^2(\partial_- E\cap\Gamma_-))}\right)\\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}(\xi^+ - \xi^-)\|^2_{L^2(0,T;L^2(\partial_- E\setminus\Gamma))} + \frac{1}{2}\||c\cdot n|^{1/2}\xi^+\|^2_{L^2(0,T;L^2(\partial_+ E\cap\Gamma))}\right)\\
&\leq Ch^{2\min(k+1,s)-2\delta}\|\|y(0)\|\|^2_{L^2(\Omega)} + Ch^{2\min(k+1,s)-2\delta}\left\|\frac{\partial y}{\partial t}\right\|^2_{L^2(0,T;H^s(\xi_h))}\\
&+ 2Cc_0^2h^{2\min(k+1,s)-2\delta}\|y\|^2_{L^2(0,T;H^s(\xi_h))} + Ch^{2\min(k+1,s)-2\delta}\|y\|^2_{L^2(0,T;H^s(\xi_h))}.
\end{aligned}$$

The final result can be reached by using the triangle inequality as follows:

$$\|y(t) - Y_h(t)\|_{L^2(\Omega)} \leq \|Y_h(t) - \tilde{y}(t)\|_{L^2(\Omega)} + \|y(t) - \tilde{y}(t)\|_{L^2(\Omega)}.$$

6.2.3 Stability Estimates For The Full-discrete State (Backward Euler)

For the stability estimates of the full-discrete state, we need to use the full-discrete DG variational formulation of the state equation. At [37, 38], the problems that are considered are not time-dependent, while error analysis has been performed for only semi-discrete problem at [59] after having applied optimize-then-discretize approach. As we have mentioned, we have used discretize-then-optimize approach. Due to the transient nature of the problem, we need

to conduct error analysis for the full-discrete problem, too. We start by determining a bound for $\|Y_h^m\|$. The only difference arises from $\left(\frac{Y_h^{n+1}-Y_h^n}{\Delta t}, Y_h^{n+1}\right)_\Omega$.

Lemma 6.2.3 *Suppose that there exist a constant C independent of h and Δt such that for all $m > 0$,*

$$\|Y_h^m\|^2_{L^2(\Omega)} + \kappa\Delta t \sum_{n=1}^{m} \|Y_h^n\|^2_\epsilon \leq C\left(\|y_0\|^2_{L^2(\Omega))} + \Delta t \sum_{n=1}^{m} \|f^n\|^2_{L^2(\Omega)}\right)$$
$$+ C\Delta t\left(\sum_{n=1}^{m}\sum_{E\in\xi_h} \||c\cdot n|^{1/2} g_D^n\|^2_{L^2(\partial_- E\cap\Gamma_-)} + \sum_{n=1}^{m}\sum_{e\in\partial\Omega} \frac{1}{|e|^{\beta_0}}\|g_D^n\|^2_{L^2(e)}\right). \tag{6.25}$$

Proof. We obtain the full discrete state equation by backward euler method and choose $\upsilon = Y_h^{n+1}$:

$$\left(\frac{Y_h^{n+1}-Y_h^n}{\Delta t}, Y_h^{n+1}\right)_\Omega + a_\epsilon(Y_h^{n+1}, Y_h^{n+1}) + c(Y_h^{n+1}, Y_h^{n+1}) = L(t^{n+1}; Y_h^{n+1}), \quad \forall n > 0, \quad \forall \upsilon \in D_k(\xi_h), \tag{6.26}$$

$$Y_h^0 = y_0. \tag{6.27}$$

Use coercivity of a_ϵ, the inequality "$\frac{1}{2}(x^2 - y^2) \leq (x-y)x$", bounds for $c(\cdot,\cdot)$ and $L(\cdot,\cdot)$ on [38] and [53], respectively, to obtain:

$$\begin{aligned}
&\frac{1}{2\Delta t}(\|Y_h^{n+1}\|^2_{L^2(\Omega)} - \|Y_h^n\|^2_{L^2(\Omega))}) + \kappa\|Y_h^{n+1}\|^2_\epsilon \\
&+ \sum_{E\in\xi_h}\left(\|c_0 Y_h^{n+1}\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}(Y_h^{n+1})^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}(Y_h^{n+1})^+ - (Y_h^{n+1})^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}(Y_h^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\leq \frac{1}{2}\|f^{n+1}\|^2_{L^2(\Omega)} + \frac{1}{2}\|Y_h^{n+1}\|^2_{L^2(\Omega)} \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{4}\||c\cdot n|^{1/2}(Y_h^{n+1})^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} + \||c\cdot n|^{1/2} g_D^{n+1}\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \frac{\kappa}{2}\|Y_h^{n+1}\|^2_\epsilon + 2C\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D^{n+1}\|^2_{L^2(e)}.
\end{aligned}$$

We arrange the terms to obtain the following:

$$
\begin{aligned}
&\frac{1}{2\Delta t}(\|Y_h^{n+1}\|^2_{L^2(\Omega)} - \|Y_h^n\|^2_{L^2(\Omega))}) + \frac{\kappa}{2}\|Y_h^{n+1}\|^2_\epsilon \\
&+ \sum_{E\in\xi_h}\left(\|c_0 Y_h^{n+1}\|^2_{L^2(E)} + \frac{1}{4}\||c\cdot n|^{1/2}(Y_h^{n+1})^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}(Y_h^{n+1})^+ - (Y_h^{n+1})^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}(Y_h^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\leq \frac{1}{2}\|f^{n+1}\|^2_{L^2(\Omega)} + \frac{1}{2}\|Y_h^{n+1}\|^2_{L^2(\Omega)} \\
&+ \sum_{E\in\xi_h}\left(\||c\cdot n|^{1/2} g_D^{n+1}\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) + 2C\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D^{n+1}\|^2_{L^2(e)}.
\end{aligned}
$$

We multiply the inequality by $2\Delta t$ and sum from $n = 0$ to $n = m - 1$:

$$
\begin{aligned}
&\|Y_h^m\|^2_{L^2(\Omega)} - \|Y_h^0\|^2_{L^2(\Omega))} + \kappa\Delta t\sum_{n=1}^m\|Y_h^n\|^2_\epsilon \\
&+ 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\|c_0 Y_h^n\|^2_{L^2(E)} + \frac{1}{2}\Delta t\sum_{n=1}^m\||c\cdot n|^{1/2}(Y_h^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\||c\cdot n|^{1/2}(Y_h^n)^+ - (Y_h^n)^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \Delta t\sum_{n=1}^m\||c\cdot n|^{1/2}(Y_h^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\leq \Delta t\sum_{n=1}^m\|f^n\|^2_{L^2(\Omega)} + \Delta t\sum_{n=1}^m\|Y_h^n\|^2_{L^2(\Omega)} \\
&+ 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\||c\cdot n|^{1/2} g_D^n\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) + 4C\Delta t\sum_{n=1}^m\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D^n\|^2_{L^2(e)}.
\end{aligned}
$$

We substitute $\|Y_h^0\|^2_{L^2(\Omega))}$ by the approximate solution $\|y_0\|^2_{L^2(\Omega))}$. In addition, the term $\Delta t\sum_{n=1}^m\|Y_h^n\|^2_{L^2(\Omega)}$ at the right-hand side must be eliminated to obtain a stability bound. To do this, discreet Gronwall's lemma can be used for a constant C which increases exponentially in time to obtain the final result.

$$
\begin{aligned}
&\|Y_h^m\|^2_{L^2(\Omega)} + \kappa\Delta t\sum_{n=1}^m\|Y_h^n\|^2_\epsilon \\
&+ 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\|c_0 Y_h^n\|^2_{L^2(E)} + \frac{1}{2}\Delta t\sum_{n=1}^m\||c\cdot n|^{1/2}(Y_h^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\||c\cdot n|^{1/2}(Y_h^n)^+ - (Y_h^n)^-\|^2_{L^2(\partial_- E\setminus\Gamma)} + \Delta t\sum_{n=1}^m\||c\cdot n|^{1/2}(Y_h^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\leq C\left(\|y_0\|^2_{L^2(\Omega))} + \Delta t\sum_{n=1}^m\|f^n\|^2_{L^2(\Omega)} + 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\||c\cdot n|^{1/2} g_D^n\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ 4C\left(\tilde{C}\Delta t\sum_{n=1}^m\sum_{e\in\partial\Omega}\frac{1}{|e|^{\beta_0}}\|g_D^n\|^2_{L^2(e)}\right).
\end{aligned}
$$

6.2.4 Error Estimates For The Full-discrete State (Backward Euler)

This part is similar to the error estimates for the semi-discrete state equation. Instead of $y - \tilde{y}$ and $Y_h - \tilde{y}$, we use $\eta = y^n - \tilde{y}^n$ and $\xi = Y_h^n - \tilde{y}^n$, respectively.

Lemma 6.2.4 *(Backward Euler) For $s > 3/2$, assume that the exact solution $y \in H^1(0,T;H^s(\xi_h))$, $\frac{\partial^2 y}{\partial t^2} \in L^2(0,T;L^2(\Omega))$. There exist a constant C independent of h and Δt such that for all $m > 0$*

$$\|Y_h^m - y^m\|_{L^2(\Omega)}^2 \le Ch^{2\min(k+1,s)-2\delta}\left(|||y(0)|||_{H^s(\xi_h)}^2 + \left\|\frac{\partial y}{\partial t}\right\|_{H^1(0,T;H^s(\xi_h))}^2\right) + C\Delta t^2\left\|\frac{\partial^2 y}{\partial t^2}\right\|_{L^2(0,T;L^2(\Omega))}^2$$
$$+ C\Delta t h^{2\min(k+1,s)-2\delta}\sum_{n=1}^{m}\left(|||y^{n+1}|||_{H^s(\xi_h)}^2 + \sum_{E\in\xi_h}|||c_0 y^{n+1}|||_{H^s(\xi_h)}^2 + |||c\cdot n|^{1/2}(y^{n+1})\|_{H^s(\xi_h)}^2\right). \tag{6.28}$$

In general, for SIPG, $\delta = 0$ while for NIPG and IIPG, $\delta = 1$.

Proof. Let $\tilde{y}$ be the elliptic projection of y, as we mentioned before. Let me write $y^n = y(t^n)$, $\tilde{y}^n = \tilde{y}(t^n)$. Define $\xi^n = Y_h^n - \tilde{y}^n$ and $\eta^n = y^n - \tilde{y}^n$. As in [53], we subtract the semi-discrete variational form from the full discrete one to obtain:

$$\left(\frac{\xi^{n+1}-\xi^n}{\Delta t}, \upsilon\right)_\Omega + a_\epsilon(\xi^{n+1},\upsilon) + c(\xi^{n+1},\upsilon) \tag{6.29}$$
$$= \left(\frac{\partial y^{n+1}}{\partial t} - \frac{y^{n+1}-y^n}{\Delta t}, \upsilon\right)_\Omega + \left(\frac{\eta^{n+1}-\eta^n}{\Delta t}, \upsilon\right)_\Omega + a_\epsilon(\eta^{n+1},\upsilon) + c(\eta^{n+1},\upsilon). \tag{6.30}$$

We begin with a definition $\theta^{n+1} = \frac{\partial z^{n+1}}{\partial t} - \frac{z^{n+1}-z^n}{\Delta t}$. We choose $\upsilon = \xi^{n+1}$. For the left-hand side, we use coercivity of a_ϵ, the elliptic projection $\tilde{y}$. Cauchy-Schwarz's and *Poincaré*'s inequalities are used for the first product at the right-hand side. Then, we proceed similar to

the semi-discrete case to obtain:

$$
\begin{aligned}
&\frac{1}{2\Delta t}(\|\xi^{n+1}\|^2_{L^2(\Omega)} - \|\xi^n\|^2_{L^2(\Omega)}) + \kappa\|\xi^{n+1}\|^2_\epsilon \\
&+ \sum_{E\in\xi_h}\left(\|c_0\xi^{n+1}\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}(\xi^{n+1})^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}((\xi^{n+1})^+ - (\xi^{n+1})^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}(\xi^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\le C\|\xi^{n+1}\|_\epsilon\left(\|\theta^{n+1}\|_{L^2(\Omega)} + \left\|\frac{\eta^{n+1}-\eta^n}{\Delta t}\right\|_{L^2(\Omega)}\right) \\
&+ \frac{\kappa}{8}\|\xi^{n+1}\|^2_\epsilon + C\|\eta^{n+1}\|^2_{L^2(\Omega)} + \sum_{E\in\xi_h}\left(\frac{1}{2}\|c_0\xi^{n+1}\|^2_{L^2(E)} + 2\|c_0\eta^{n+1}\|^2_{L^2(E)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{4}\||c\cdot n|^{1/2}(\xi^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)} + \||c\cdot n|^{1/2}(\eta^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{4}\||c\cdot n|^{1/2}((\xi^{n+1})^+ - (\xi^{n+1})^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \||c\cdot n|^{1/2}(\eta^{n+1})^-\|^2_{L^2(\partial_- E\setminus\Gamma)}\right).
\end{aligned}
$$

Let me consider the term $C\|\xi^{n+1}\|_\epsilon\left(\|\theta^{n+1}\|_{L^2(\Omega)} + \left\|\frac{\eta^{n+1}-\eta^n}{\Delta t}\right\|_{L^2(\Omega)}\right)$ separately to facilitate the procedure. As in [53], by Young's inequality,

$$
C\|\xi^{n+1}\|_\epsilon\left(\|\theta^{n+1}\|_{L^2(\Omega)} + \left\|\frac{\eta^{n+1}-\eta^n}{\Delta t}\right\|_{L^2(\Omega)}\right) \le \frac{\kappa}{2}\|\xi^{n+1}\|^2_\epsilon + \left(\|\theta^{n+1}\|^2_{L^2(\Omega)} + \left\|\frac{\eta^{n+1}-\eta^n}{\Delta t}\right\|^2_{L^2(\Omega)}\right).
$$

By Taylor expansion, we have

$$
\theta^{n+1} = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}(t-t_n)\frac{\partial^2 y}{\partial t^2}dt, \qquad \eta^{n+1}-\eta^n = \int_{t^n}^{t^{n+1}}\frac{\partial\eta}{\partial t}dt.
$$

Using Cauchy-Schwarz's inequality, they can be written as

$$
\|\theta^{n+1}\|^2_{L^2(\Omega)} \le \frac{\Delta t}{3}\int_{t^n}^{t^{n+1}}\left\|\frac{\partial^2 y}{\partial t^2}\right\|^2_{L^2(\Omega)}dt, \qquad \|\eta^{n+1}-\eta^n\|^2_{L^2(\Omega)} \le \Delta t\int_{t^n}^{t^{n+1}}\left\|\frac{\partial\eta}{\partial t}\right\|^2_{L^2(\Omega)}dt.
$$

We substitute these into the inequality, arrange the terms and multiple the inequality by $2\Delta t$ and sum from $n = 0$ to $n = m-1$.

$$
\begin{aligned}
&\|\xi^m\|^2_{L^2(\Omega)} - \|\xi^0\|^2_{L^2(\Omega)} + \frac{3}{4}\kappa\Delta t \sum_{n=1}^{m} \|\xi^n\|^2_{\epsilon} \\
&+ 2\Delta t \sum_{n=1}^{m} \sum_{E\in\xi_h} \left(\frac{1}{2}\|c_0\xi^n\|^2_{L^2(E)} + \||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} \right) \\
&+ 2\Delta t \sum_{n=1}^{m} \sum_{E\in\xi_h} \left(\frac{1}{4}\||c\cdot n|^{1/2}((\xi^n)^+ - (\xi^n)^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{4}\||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)} \right) \\
&\leq C\Delta t^2 \int_0^T \left\| \frac{\partial^2 y}{\partial t^2} \right\|^2_{L^2(\Omega)} dt + C\int_0^T \left\| \frac{\partial \eta}{\partial t} \right\|^2_{L^2(\Omega)} dt + 2C\Delta t \sum_{n=1}^{m} \|\eta^{n+1}\|^2_{L^2(\Omega)} \\
&+ 2\Delta t \sum_{n=1}^{m} \sum_{E\in\xi_h} \left(2\|c_0\eta^{n+1}|^2_{L^2(E)} + \||c\cdot n|^{1/2}(\eta^{n+1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)} + \||c\cdot n|^{1/2}(\eta^{n+1})^-\|^2_{L^2(\partial_- E\setminus\Gamma)} \right).
\end{aligned}
$$

Then, we use error bounds given in [53] similar to the semi-discrete case to obtain the following:

$$
\begin{aligned}
&\|\xi^m\|^2_{L^2(\Omega)} + \frac{3}{4}\kappa\Delta t \sum_{n=1}^{m} \|\xi^n\|^2_{\epsilon} \\
&+ 2\Delta t \sum_{n=1}^{m} \sum_{E\in\xi_h} \left(\frac{1}{2}\|c_0\xi^n\|^2_{L^2(E)} + \||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)} \right) \\
&+ 2\Delta t \sum_{n=1}^{m} \sum_{E\in\xi_h} \left(\frac{1}{4}\||c\cdot n|^{1/2}((\xi^n)^+ - (\xi^n)^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{4}\||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)} \right) \\
&\leq Ch^{2\min(k+1,s)-2\delta} \left(\|\|y(0)\|\|^2_{H^s(\xi_h)} + \left\| \frac{\partial y}{\partial t} \right\|^2_{H^1(0,T;H^s(\xi_h))} \right) + C\Delta t^2 \left\| \frac{\partial^2 y}{\partial t^2} \right\|^2_{L^2(0,T;L^2(\Omega))} \\
&+ 2\Delta t Ch^{2\min(k+1,s)-2\delta} \sum_{n=1}^{m} (\|\|y^{n+1}\|\|^2_{H^s(\xi_h)} + \sum_{E\in\xi_h} \|\|c_0 y^{n+1}\|\|^2_{H^s(\xi_h)} + \||c\cdot n|^{1/2}(y^{n+1})\|^2_{H^s(\xi_h)}).
\end{aligned}
$$

The final result is obtained by the triangle inequality.

6.2.5 Stability/Convergence Estimates For The Full-discrete State (Crank-Nicolson)

Without going into details, the full discrete variational formulation by Crank Nicolson is A-stable [29]. Under the smoothness assumption for the solution, we deduce that [53]

$$
\|Y_h^n - y^n\|_{L^2(\Omega)} \quad \leq \quad O(h^{\min(k+1,s)-\delta} + \Delta t^2) \tag{6.31}
$$

6.3 Error Analysis For The Adjoint Equation

6.3.1 Stability Estimates For The Adjoint (Backward Euler)

We have discretized the optimal control by DG in time and by θ-method as follows:

$$\min_{\tilde{u}_0,\ldots,\tilde{u}_N} \sum_{i=0}^{N} \Delta t(\frac{1}{2}\tilde{y}_i^T M\tilde{y}_i - (Y_d(t))^T\tilde{y}_i + \frac{\alpha}{2}\tilde{u}_i^T M\tilde{u}_i), \tag{6.32}$$

where $\tilde{y}_0,\ldots,\tilde{y}_N$ is the solution of

$$(M + \Delta t\theta(D + C + R))\tilde{y}_{i+1} = (M - \Delta t(1-\theta)(D + C + R))\tilde{y}_i$$

$$+ \Delta t(\theta F(t_{i+1}) + (1-\theta)F(t_i)) + \Delta t(\theta M\tilde{u}_{i+1} + (1-\theta)M\tilde{u}_i), \tag{6.33}$$

$$i = 0,\ldots,N-1 \quad \text{and} \quad M\bar{y}(0) = \bar{Y}_0. \tag{6.34}$$

The corresponding adjoint equation by Backward Euler has been written as

$$(M + \Delta t\theta(D + C + R))^T\tilde{p}_N = -\frac{\Delta t}{2}(M\tilde{y}_N - (Y_d(t))_N), \tag{6.35}$$

$$(M + \Delta t\theta(D + C + R))^T\tilde{p}_N = (M - \Delta t(1-\theta)(D + C + R))^T\tilde{p}_{N+1} - \Delta t(M\tilde{y}_i - (Y_d(t))_i), \tag{6.36}$$

$$i = N-1,\ldots,0.$$

Then, we can obtain the variational from corresponding to the full discrete adjoint equation:

$$\left(\frac{P_h^n - P_h^{n+1}}{\Delta t}, \upsilon\right)_\Omega + a_\epsilon(P_h^n,\upsilon) + c(P_h^n,\upsilon) = -(Y_h^n - Y_d^n,\upsilon)_\Omega, \quad \text{for} \quad 0 \le n < N_T - 1, \quad \forall\upsilon \in D_k(\xi_h),$$

$$p_h^{N_T} = p_T.$$

Lemma 6.3.1 *There exist constants C independent of h and Δt such that for $0 \le m < N_T$*

$$\|P_h^m\|^2_{L^2(\Omega)} \le C\left(\sum_{n=0}^{m-1}\|Y_h^n - Y_d^n\|^2_{L^2(\Omega)} + \|p_T\|^2_{L^2(\Omega)}\right), \tag{6.37}$$

where C increases exponentially in time.

Proof. We consider the full discrete adjoint equation and choose $\upsilon = P_h^n$.

$$\left(\frac{P_h^n - P_h^{n+1}}{\Delta t}, P_h^n\right)_\Omega + a_\epsilon(P_h^n,P_h^n) + c(P_h^n,P_h^n) = (Y_h^n - Y_d^n, P_h^n)_\Omega, \quad \text{for} \quad 0 \le n < N_T - 1, \quad \forall\upsilon \in D_k(\xi_h), \tag{6.38}$$

$$p_h^{N_T} = p_T. \tag{6.39}$$

As in the full discrete state equation, we use the bounds for $c(\cdot,\cdot)$, ellipticity of a_ϵ, Cauchy-Schwartz and Young's inequality to proceed as follows:

$$\begin{aligned}
&\frac{1}{2\Delta t}(\|P_h^n\|^2_{L^2(\Omega)} - \|P_h^{n+1}\|^2_{L^2(\Omega)}) + \kappa\|P_h^n\|^2_\epsilon \\
&+ \sum_{E\in\xi_h}\left(\|c_0 P_h^n\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}(P_h^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ \sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}((P_h^n)^+ - (P_h^n)^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}(P_h^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\le \frac{1}{2}\|Y_h^n - Y_d^n\|^2_{L^2(\Omega)} + \frac{1}{2}\|P_h^n\|^2_{L^2(\Omega)}.
\end{aligned}$$

We multiply the inequality by $2\Delta t$ and sum from $n = m$ to $n = 1$ to obtain:

$$\begin{aligned}
&\|P_h^{m-1}\|^2_{L^2(\Omega)} - \|p_T\|^2_{L^2(\Omega)} + 2\Delta t\kappa\sum_{n=1}^m \|P_h^{n-1}\|^2_\epsilon \\
&+ 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\|c_0 P_h^{n-1}\|^2_{L^2(E)} + \frac{1}{2}\||c\cdot n|^{1/2}(P_h^{n-1})^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ 2\Delta t\sum_{n=1}^m\sum_{E\in\xi_h}\left(\frac{1}{2}\||c\cdot n|^{1/2}((P_h^{n-1})^+ - (P_h^{n-1})^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{2}\||c\cdot n|^{1/2}(P_h^{n-1})^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\le \Delta t\sum_{n=1}^m \|Y_h^{n-1} - Y_d^{n-1}\|^2_{L^2(\Omega)} + \Delta t\sum_{n=1}^m \|P_h^{n-1}\|^2_{L^2(\Omega)}.
\end{aligned}$$

We need to eliminate $\Delta t\sum_{n=1}^m \|P_h^{n-1}\|^2_{L^2(\Omega)}$ at the right-hand side. Thus, we apply the discrete Gronwall inequality to obtain the final result.

6.3.2 Error Estimates For Adjoint (Backward Euler)

Lemma 6.3.2 *(Backward Euler) For $s > 3/2$, suppose that $p \in H^1(0,T;H^s(\xi_h))$, $\frac{\partial^2 p}{\partial t^2} \in L^2(0,T;L^2(\Omega))$. There exist a constant C independent of h and Δt such that for all $0 \le m < N_T$*

$$\begin{aligned}
\|P_h^m - p^m\|^2_{L^2(\Omega)} \le Ch^{2\min(k+1,s)-2\delta}&\left(\|\|p(T)\|\|^2_{H^s(\xi_h)} + \left\|\frac{\partial p}{\partial t}\right\|^2_{H^1(0,T;H^s(\xi_h))}\right) + C\Delta t^2\left\|\frac{\partial^2 p}{\partial t^2}\right\|^2_{L^2(0,T;L^2(\Omega))} \\
&+ 2\Delta t Ch^{2\min(k+1,s)-2\delta}\sum_{n=0}^m(\|\|p^n\|\|^2_{H^s(\xi_h)} + \sum_{E\in\xi_h}\|\|c_0 p^n\|\|^2_{H^s(\xi_h)} + \||c\cdot n|^{1/2}(p^n)\|^2_{H^s(\xi_h)})
\end{aligned} \tag{6.40}$$

Proof. Let $\tilde{p}$ be the elliptic projection of p. Denote $p^n = p(t^n)$, $\tilde{p}^n = \tilde{p}(t^n)$. Let $\xi^n = P_h^n - \tilde{p}^n$

and $\eta = p^n - \tilde{p}^n$. Similar to the convergence proof of the full discrete state equation, we obtain

$$\left(\frac{\xi^n - \xi^{n+1}}{\Delta t}, \upsilon\right)_\Omega + a_\epsilon(\xi^n, \upsilon) + c(\xi^n, \upsilon) \tag{6.41}$$

$$= \left(\frac{\partial p^n}{\Delta t} - \frac{p^n - p^{n+1}}{\Delta t}, \upsilon\right)_\Omega + \left(\frac{\eta^n - \eta^{n+1}}{\Delta t}, \upsilon\right)_\Omega + a_\epsilon(\eta^n, \upsilon) + c(\eta^n, \upsilon). \tag{6.42}$$

We choose $\upsilon = \xi^n$ and proceed similar to the full-discreet state equation. At the end, we multiple the inequality by $2\Delta t$ and sum from $n = m$ to $n = 0$.

$$\begin{aligned}
&\|\xi^m\|^2_{L^2(\Omega)} + \frac{3}{4}\kappa\Delta t \sum_{n=0}^{m} \|\xi^n\|^2_\epsilon \\
&+ 2\Delta t \sum_{n=0}^{m} \sum_{E\in\xi_h} \left(\frac{1}{2}\|c_0\xi^n\|^2_{L^2(E)} + \||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_- E\cap\Gamma_-)}\right) \\
&+ 2\Delta t \sum_{n=0}^{m} \sum_{E\in\xi_h} \left(\frac{1}{4}\||c\cdot n|^{1/2}((\xi^n)^+ - (\xi^n)^-)\|^2_{L^2(\partial_- E\setminus\Gamma)} + \frac{1}{4}\||c\cdot n|^{1/2}(\xi^n)^+\|^2_{L^2(\partial_+ E\cap\Gamma)}\right) \\
&\leq Ch^{2\min(k+1,s)-2\delta}\left(\|\|p(T)\|\|^2_{H^s(\xi_h)} + \left\|\frac{\partial p}{\partial t}\right\|^2_{H^1(0,T;H^s(\xi_h))}\right) + C\Delta t^2 \left\|\frac{\partial^2 p}{\partial t^2}\right\|^2_{L^2(0,T;L^2(\Omega))} \\
&+ 2\Delta t Ch^{2\min(k+1,s)-2\delta} \sum_{n=0}^{m} (\|\|p^n\|\|^2_{H^s(\xi_h)} + \sum_{E\in\xi_h} \|\|c_0 p^n\|\|^2_{H^s(\xi_h)} + \||c\cdot n|^{1/2}(p^n)\|^2_{H^s(\xi_h)}).
\end{aligned}$$

The final result is obtained by the triangle inequality.

6.3.3 Stability/Convergence Estimates For The Full-discrete Adjoint (Crank-Nicolson)

Crank-Nicolson method is A-stable [50]. For the convergence, we consider the weak form of the adjoint equation. Under the smoothness assumption for the solution, we deduce that [53]

$$\|P_h^n - p^n\|_{L^2(\Omega)} \quad \leq \quad O(h^{\min(k+1,s)-\delta} + \Delta t^2). \tag{6.43}$$

6.4 Error Estimates For The Control

6.4.1 Error Estimates For The Unconstrained Optimal Control

Lemma 6.4.1 *(Backward Euler) The solutions to continuous and the discrete optimal control problem satisfy*

$$\|\bar{u} - \bar{u}_h^n\|_{L^2(0,T;U)} \leq \frac{1}{\alpha}\|p(\bar{u}) - p_h^n(\bar{u})\|_{L^2(0,T;U)} + \|\bar{u} - q\|_{L^2(0,T;U)}. \tag{6.44}$$

Proof. The PDE constraint of the optimal control problem, $e(y,u) = 0$ give rise to a solution operator $U \ni u \longmapsto S(u) \in Y$. The reduced cost functional can be written as

$$j(u) = j(S(u),u) \qquad u \in U.$$

We can rewrite the optimal control problem as one piece

$$\min j(u) \qquad u \in U_{ad} = U.$$

The fact that the reduced cost functional is continuously differentiable enables us to note the derivatives as [6]

$$j^{'}(u)(\delta u) = \int_0^T (p, \delta u) - \alpha \int_0^T (u, \delta u).$$

Since U_{ad} is convex, we state the necessary optimality condition for this problem as follows:

$$j^{'}(\bar{u})(\delta u - \bar{u}) = \int_0^T (\bar{p} - \alpha\bar{u}, \delta u - \bar{u}) = 0 \qquad \forall \delta u \in U_{ad},$$

where $\bar{u}$ is the optimal control. The sufficient optimality condition give rise to

$$j^{''}(\bar{u})(\delta u, \delta u) \geq \alpha \|\delta u\|^2_{L^2(0,T;U)}, \quad \forall \delta u \in U.$$

Similar the continuous case, the discrete solution operator can be defined as [6] $S^n_h : U \longmapsto D_k(\xi_h)$ to state the cost functional

$$j^n_h(u) = j(S^n_h(u), u).$$

We obtain the necessary and sufficient optimality condition for the discretized problem as follows:

$$j^{n^{'}}_h(\bar{u}^n_h)(\delta u^n_h - \bar{u}^n_h) = 0 \qquad \forall \delta \bar{u}^n_h \in U_{ad_h},$$

$$j^{n^{''}}_h(\bar{u}^n_h)(\delta u^n_h, \delta u^n_h) \geq \alpha \|\delta u^n_h\|^2_{L^2(0,T;U)}, \quad \forall \delta u^n_h \in U_{ad_h}.$$

Now let me start the proof by choosing any $q \in U_{ad,h} = D_k(\xi_h)$. Consider the following

$$\begin{aligned} \alpha\|q - \bar{u}^n_h\|^2_{L^2(0,T;U)} &\leq j^{n^{''}}_h(\bar{u}^n_h)(q - \bar{u}^n_h, q - \bar{u}^n_h) \\ &= j^{n^{'}}_h(q)(q - \bar{u}^n_h) - j^{n^{'}}_h(\bar{u})(q - \bar{u}^n_h). \end{aligned}$$

Since $U_{ad} = U$ and $U_{ad,h} = U_h$, we obtain by [6]

$$j^{n^{'}}_h(\bar{u})(q - \bar{u}^n_h) = 0 = j^{n^{'}}_h(\bar{u}^n_h)(q - \bar{u}^n_h).$$

Then,

$$\alpha\|q-\bar{u}_h^n\|^2_{L^2(0,T;U)} \leq j_h^{n'}(q)(q-\bar{u}_h^n) - j'(\bar{u})(q-\bar{u}_h^n).$$

Now, we use the relation between the solutions of the continuous and the discrete optimal control problem [21]. Consider

$$j'(u)(\phi) = \int_0^T (p(u)-\alpha u,\phi), \qquad j_h^{n'}(u)(\phi) = \int_0^T (p_h^n(u)-\alpha u,\phi), \qquad \forall \phi \in U_{ad}.$$

Then, we obtain

$$\|j'(u)(\phi) - j_h^{n'}(u)(\phi)\|_{L^2(0,T;U)} = \|(p(u)-p_h^n(u),\phi)\|_{L^2(0,T;U)} \leq \|(p(u)-p_h^n(u)\|_{L^2(0,T;U)}\|\phi\|_{L^2(0,T;U)}.$$

which enables us to write

$$\alpha\|q-\bar{u}_h^n\|^2_{L^2(0,T;U)} \leq \|p(\bar{u})-p_h^n(\bar{u})\|_{L^2(0,T;U)}\|q-\bar{u}_h^n\|_{L^2(0,T;U)}.$$

By cancelation,

$$\|q-\bar{u}_h^n\|_{L^2(0,T;U)} \leq \frac{1}{\alpha}\|p(\bar{u})-p_h^n(\bar{u})\|_{L^2(0,T;U)}.$$

Let q be the pointwise interpolant of $\bar{u}$ [6]. Then, by the triangle inequality, we have

$$\|\bar{u}-\bar{u}_h^n\|_{L^2(0,T;U)} \leq \frac{1}{\alpha}\|p(\bar{u})-p_h^n(\bar{u})\|_{L^2(0,T;U)} + \|\bar{u}-q\|_{L^2(0,T;U)}.$$

Solutions to continuous and the discrete optimal control problem satisfy for backward Euler and Crank-Nicolson, respectively

$$\|\bar{u}-\bar{u}_h^n\|_{L^2(0,T;U)} \quad \leq \quad O(h^{\min(k+1,s)-\delta}+\Delta t), \tag{6.45}$$

$$\|\bar{u}-\bar{u}_h^n\|_{L^2(0,T;U)} \quad \leq \quad O(h^{\min(k+1,s)-\delta}+\Delta t^2). \tag{6.46}$$

6.4.2 Error Estimates For The Constrained Optimal Control

Lemma 6.4.2 *(Backward Euler) The solutions to continuous and the discrete optimal control problem satisfy*

$$\|\bar{u}-\bar{u}_h^n\|_{L^2(0,T;U)} \leq \frac{1}{\alpha}\|p(\bar{u})-p_h^n(\bar{u})\|_{L^2(0,T;U)}. \tag{6.47}$$

Proof. By [60], for $\bar{u}$ and p to be an optimal control and a weak solution of the adjoint equation, respectively; the following variational inequality is satisfied

$$\int_0^T (p - \alpha\bar{u})(u - \bar{u})dt \geq 0, \quad \forall u \in U_{ad}.$$

A necessary and sufficient conditions for the variational equation to be satisfied for almost every $(x, t) \in \Omega \times [0, T]$ is given by [60]

$$\bar{u}(x,t) = \begin{cases} u_a(x,t) & \text{if} \quad p - \alpha\bar{u} \geq 0 \\ [u_a(x,t), u_a(x,t)] & \text{if} \quad p - \alpha\bar{u} = 0 \\ u_b(x,t) & \text{if} \quad p - \alpha\bar{u} \leq 0. \end{cases}$$

An equivalent condition can be written pointwisely in $\mathbb{R}$:

$$\int_0^T (p - \alpha\bar{u})(\upsilon - \bar{u}) \geq 0, \quad \forall \upsilon \in [u_a(x,t), u_b(x,t)], \quad \text{for} \quad \text{a.e.} \quad x \in \Omega \times [0,T].$$

Then, one can obtain the weak minimum principle [60]

$$\min_{\upsilon \in [u_a(x,t), u_b(x,t)]} \{(p - \alpha\bar{u})\upsilon\} = (p - \alpha\bar{u})\bar{u},$$

or the minimum principle

$$\min_{\upsilon \in [u_a(x,t), u_b(x,t)]} \left\{(p\upsilon - \frac{\alpha}{2}\upsilon^2\right\} = p\bar{u} - \frac{\alpha}{2}\bar{u}^2.$$

For $\alpha > 0$ and $\bar{u}$ is an optimal control of the problem if and only if $\bar{u} = \mathbb{P}_{[u_a(x,t), u_b(x,t)]}\left\{\frac{1}{\alpha}p\right\}$, is satisfied for a.e. $x \in \Omega \times [0, T]$. Indeed, for real $a \leq b$, $\mathbb{P}_{[u_a(x,t), u_b(x,t)]}$ corresponds to the projection of $\mathbb{R}$ onto $[a, b]$,

$$\mathbb{P}_{[a,b]}(u) := \min\{b, \max\{a, u\}\}.$$

Then, by [33]

$$\left\| \mathbb{P}_{[u_a(x,t), u_b(x,t)]}(\frac{1}{\alpha}p) - \mathbb{P}_{[u_a(x,t), u_b(x,t)]}(\frac{1}{\alpha}p_h^n) \right\|_{L^2(0,T;U)} \leq \frac{1}{\alpha}\|p(\bar{u}) - p(\bar{u})_h^n\|_{L^2(0,T;U)},$$

we obtain the desired inequality.

In addition, as in the unconstrained case, solutions to continuous and the discrete optimal control problem satisfy for Backward Euler and Crank-Nicolson, respectively,

$$\|\bar{u} - \bar{u}_h^n\|_{L^2(0,T;U)} \quad \leq \quad O(h^{\min(k+1,s)-\delta} + \Delta t), \tag{6.48}$$

$$\|\bar{u} - \bar{u}_h^n\|_{L^2(0,T;U)} \quad \leq \quad O(h^{\min(k+1,s)-\delta} + \Delta t^2). \tag{6.49}$$

CHAPTER 7

NUMERICAL RESULTS

7.1 Unconstrained optimal control problem

We consider the unconstrained optimal control problem on $\Omega = (0, 1)$. This problem is a modified version of a steady diffusion convection equation given by [31] which is solved by SIPG. At [30], a similar problem with $\varepsilon = 10^{-4}$, $\alpha = 10^{-2}$ is solved by SUPG. If Dirichlet boundary conditions are imposed strongly, a boundary layer at $x = 1$ is observed for the solution of the state equation with $u = 0$. The weak treatment of the boundary conditions are suggested. In addition, the error between the exact solution, which is obtained by a mesh size $h = 1/(5 \cdot 2^{10})$, and the approximate solution is computed by narrowing the spatial interval in order to eliminate the boundary layer. We have constructed the following problem, an unsteady diffusion convection equation. The only difference is the diffusion parameter which is 10^{-9} at [30].

Example 7.1.1 *We specify the source function f, the desired state y_d and the data* $f = 1,\ y_d = 1,\ \varepsilon = 0.01,\ c = [1, 1],\ r = 0,\ \alpha = 0.1.$

We don't know analytical solution of the optimal state, adjoint and control of this problem. Thus, we have just analyzed the order of the optimal control problem for decreasing time subintervals by fixing $\Delta x = 1/400$. We have used piecewise quadratic polynomials. We show the evolution of the values of the cost functional $J(\bar{y}_h, \bar{u}_h)$ for a sequence of uniformly refined temporal interval $\mathcal{T}_h$, Δt tending to zero. From this sequence, we compute the approximative order of convergence with respect to time by the formula

$$\text{order} = \frac{\log \frac{|J(\bar{y}_{2h},\bar{u}_{2h}) - J(\bar{y}_h,\bar{u}_h)|}{|J(\bar{y}_{4h},\bar{u}_{4h}) - J(\bar{y}_{2h},\bar{u}_{2h})|}}{\log 2}.$$

The maximum number of Newton iterations and tolerance have been set as 20 and $1e-8$, respectively. The penalty parameters have been chosen as in the [53]: 1 for NIPG and IIPG, 2 for SIPG. At the tables, I_N and I_C denotes the number of Newton iterations and the maximum number of CG iterations, respectively.

	SIPG				NIPG				IIPG			
Δt	I_N	I_C	$J_h(y_h,u_h)$	order	I_N	I_C	$J_h(y_h,u_h)$	order	I_N	I_C	$J_h(y_h,u_h)$	order
1/100	3	39	0.1716519	-	3	39	0.1716666	-	3	39	0.1716639	-
1/200	3	32	0.1708871	-	3	32	0.1709018	-	3	32	0.1708991	-
1/400	3	35	0.1705129	1.03	3	34	0.1705276	1.03	3	34	0.1705249	1.03
1/800	2	31	0.1703279	1.02	2	31	0.1703426	1.02	2	31	0.1703400	1.02
1/1600	3	37	0.1702359	1.02	4	37	0.1702506	1.01	4	37	0.1702480	1.01

Table 7.1: Piecewise Quadratic Elements - Backward Euler - Newton-CG

	SIPG				NIPG				IIPG			
Δt	I_N	I_C	$J_h(y_h,u_h)$	order	I_N	I_C	$J_h(y_h,u_h)$	order	I_N	I_C	$J_h(y_h,u_h)$	order
1/100	2	24	0.1701756	-	2	24	0.1701902	-	2	23	0.1701876	-
1/200	2	28	0.1701523	-	2	28	0.1701669	-	2	27	0.1701643	-
1/400	2	27	0.1701464	1.98	2	27	0.1701611	2.01	2	27	0.1701584	1.98
1/800	2	28	0.1701449	1.98	2	27	0.1701596	1.95	2	27	0.1701569	1.98
1/1600	2	28	0.1701446	2.32	2	24	0.1701592	1.91	2	24	0.1701566	2.32

Table 7.2: Piecewise Quadratic Elements - Crank Nicolson - Newton-CG

For this problem, we have observed the temporal changes of the optimal control problem. As we decrease the size of the temporal subinterval, we could have obtained a smaller value of the optimal control problem as expected. We have used piecewise quadratic polynomials. The solution profiles show that the gradient equation $\alpha u = p$ is satisfied for both of the solutions with $\alpha = 0.1$. The boundary layer at $x = 1$ is properly resolved. For backward Euler and Crank-Nicolson, we have obtained the solution profiles. As one can observe by the solution profiles, the results for SIPG, NIPG and IIPG are almost the same. These DG methods are different from each other in terms of order and symmetry. Indeed, SIPG gives the optimal solution, while NIPG and IIPG are the suboptimal methods. After having observe the approximate solutions for 2D problem, some details are given related to the efficiency of the DG methods.

The numerical order of the objective function is related to the order of the state and the control, because the objective function is the sum of two terms: The difference between the state and the desired state, and the control. The average order of backward Euler is 1.02 for for SIPG, NIPG and IIPG. In case of Crank-Nicolson, average order is 2.09 for NIPG and IIPG, 1.96 for

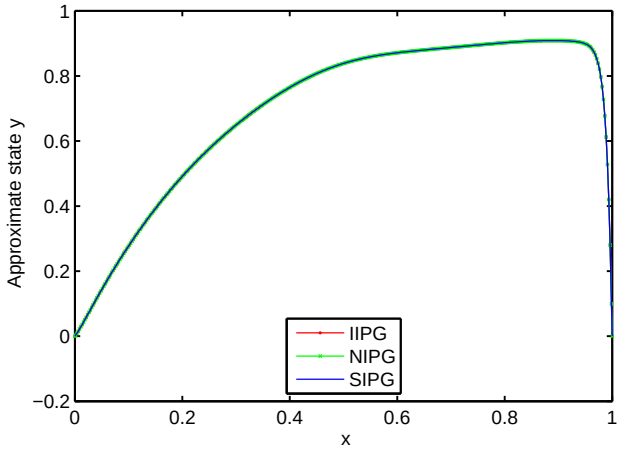

Figure 7.1: State solution, t=0.5, $\Delta x = \Delta t = 1/400$, Backward Euler

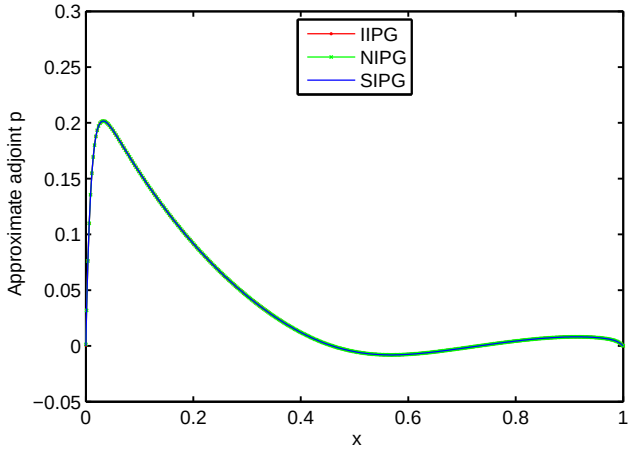

Figure 7.2: Adjoint solution, t=0.5, $\Delta x = \Delta t = 1/400$, Backward Euler

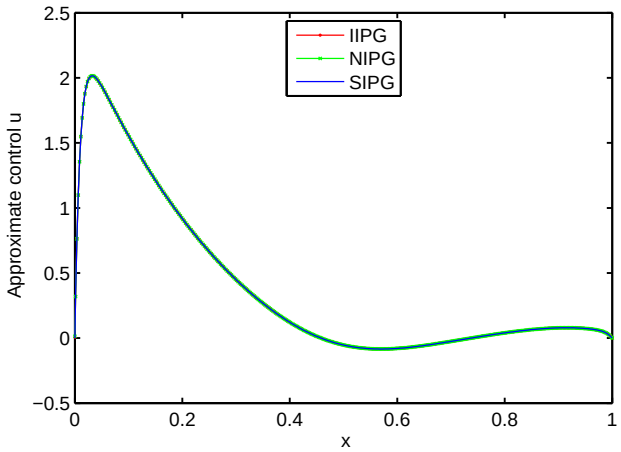

Figure 7.3: Control solution, t=0.5, $\Delta x = \Delta t = 1/400$, Backward Euler

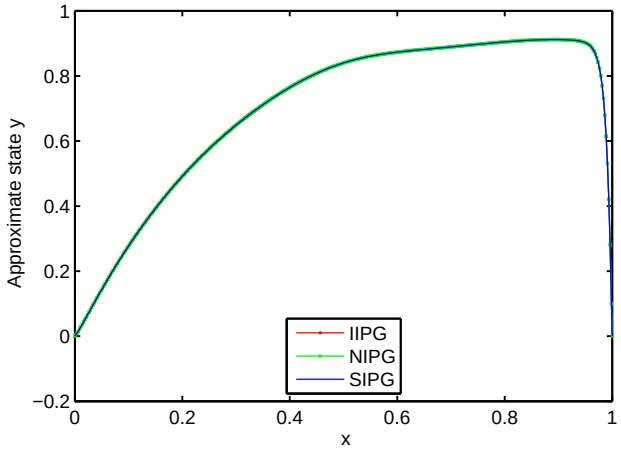

Figure 7.4: State solution, t=0.5, $\Delta x = \Delta t = 1/400$, Crank-Nisolson

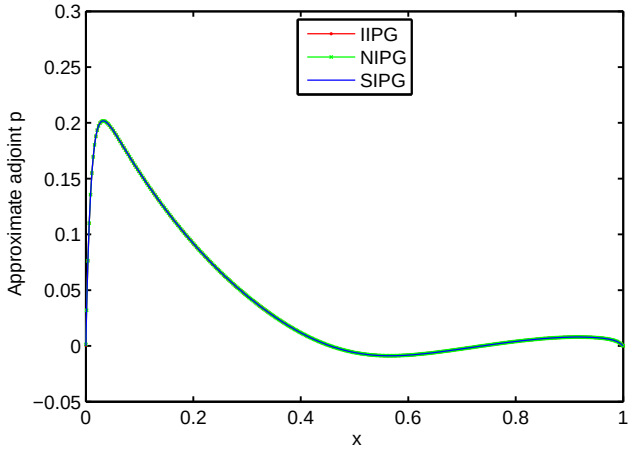

Figure 7.5: Adjoint solution, t=0.5, $\Delta x = \Delta t = 1/400$, Crank-Nisolson

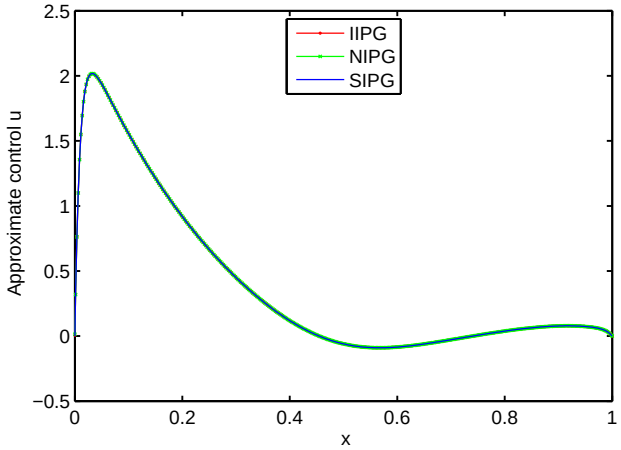

Figure 7.6: Control solution, t=0.5, $\Delta x = \Delta t = 1/400$, Crank-Nisolson

SIPG which match with the orders of backward Euler and Crank-Nicolson. In addition, the number of Newton and conjugate gradient equation for the backward Euler is more than the ones for Crank-Nicolson. In Chapter 5, we have given details how to obtain Hessian-times-vector Computation. To obtain the Hessian-Times-Vector, one needs to compute the state and the adjoint. As we check the condition number of the matrices at the right-hand side of the state and the adjoint, we see that the condition number for backward Euler methods is larger than the one for Crank-Nicolson. Since, the number of iterations for CG method is related to the condition number of the system matrix [47], we can deduce that the number of CG iterations are affected by the mentioned condition numbers.

7.2 Control constrained optimal control problem

We consider the constrained optimal control problem which is given [26]. At the article, optimize-then-discretize approach has been preferred. The problem has been discretized by characteristic finite element method in space and by backward Euler method in time in [26].

Example 7.2.1 *The data has been set as* $c = [1, 0], r = 0, \alpha = 1$*,* $u \geq 0$*. The source function* $f(x, t)$ *and* $y_d(x, t)$ *has been chosen to satisfy the* optimize-then-discretize *scheme given in the [26]. The analytical solutions of state, adjoint and control solutions are as follows:*

$$y(x, t) = \exp(-t)\sin(2\pi x_1)\sin(2\pi x_2),$$
$$p(x, t) = \exp(-t)(1 - t)\sin(2\pi x_1)\sin(2\pi x_2),$$
$$u(x, t) = \max(-p, 0).$$

The maximum number of iterations and the tolerance have been set as 100 and $1e - 16$. We have fixed $\Delta x = 1/40$ and the temporal subintervals have been divided by half successively. We show the profiles of the solutions for $\varepsilon = 0.001$ for linear piecewise basis polynomials. The penalty term is defined as $2\sigma_e^0$ SIPG and IIPG, while it is used as σ for NIPG. Indeed, σ has been chosen as 1 for NIPG, while $3k(k + 1)$ has been used for SIPG and IIPG for. The following solution profiles have been obtain by SIPG method in space and backward Euler(left-hand side) or Crank-Nicolson(right-hand side) in time. In addition, we provide the error between the exact and the numerical solutions for the state, adjoint and the control. The solutions obtained by NIPG and IIPG are similar to the ones that we have attached to the following pages. Although SIPG, NIPG and IIPG gives similar results, they differ in some respects. In case of an steady diffusion convection reaction equation, if one discretizes the problem in space by SIPG, then optimize-then-discretize and discretize-then-optimize approaches commute. To explain this, let me mention the discretized weak formulation of steady diffusion convection reaction equation for $y_h \in Y_h$ and $u_h \in U_h$.

$$a_\epsilon(y_h, \upsilon) + c(y_h, \upsilon) + r(y_h, \upsilon) + b(u_h, \upsilon) \quad = \quad L(\upsilon).$$

To simplify the notation, let me insert the convection and the reaction term into the bilinear form $\tilde{a}_h(\cdot, \cdot)$. By discretize-then-optimize approach, the necessary and sufficient optimality conditions are given as follows.

$$\tilde{a}_h^s(y_h, \upsilon) \quad + \quad b_h(u, \upsilon) = l_h^s(\upsilon), \qquad \forall \upsilon \in V_h.$$

$$\tilde{a}_h^s(\psi_h, p_h) = -(y_h - y_d, \psi_h) \qquad \forall \psi_h \in V_h,$$

$$b_h(w_h, p_h) + \omega(u_h, w_h) = 0 \qquad \forall w_h \in U_h,$$

By optimize-then-discretize approach,

$$\tilde{a}_h^s(y_h, \upsilon) + b_h(u_h, \upsilon_h) = l_h^s(\upsilon), \qquad \forall \upsilon \in V_h.$$

$$\tilde{a}_h^a(p_h, \psi_h) = -(y_h - y_d, \psi_h)_h \qquad \forall \psi_h \in \Lambda_h,$$

$$b(w_h, p_h) + \omega\langle u_h, w_h\rangle = 0 \qquad \forall w_h \in U_h$$

For SIPG, $a_h^s(\upsilon_h, p_h)$ is equal to $a_h^a(p_h, \upsilon_h)$. However, for NIPG and IIPG, two methods are not equivalent since

$$a_h^s(\upsilon_h, p_h) \neq a_h^a(p_h, \upsilon_h).$$

If the problem is analyzed in space, then it can be seen that SIPG results in optimal convergence, while suboptimal convergence is attained by NIPG and IIPG. To deal with this suboptimal nature of the methods, superconvergence can be used. Inconsistency of the adjoint can restated by using large penalty parameter. But then the condition number of the DG matrices increases. By the superpenalization, optimize-then-discretize and discretize-then-optimize approach lead to similar results when compared to the standart penalization. Thus, the approach in [11], different types of DG methods are compared for the elliptic model problem with Dirichlet boundary conditions in terms of the spectral condition number of the stiffness matrix, cost of storage, convergence rates and accuracy. In the article, traces are changed by the fluxes and different numerical fluxes results in variations of DG methods. Thus, the approach in the article differs from our approach since we have used the trace values to connect the neighboring elements in the mesh.

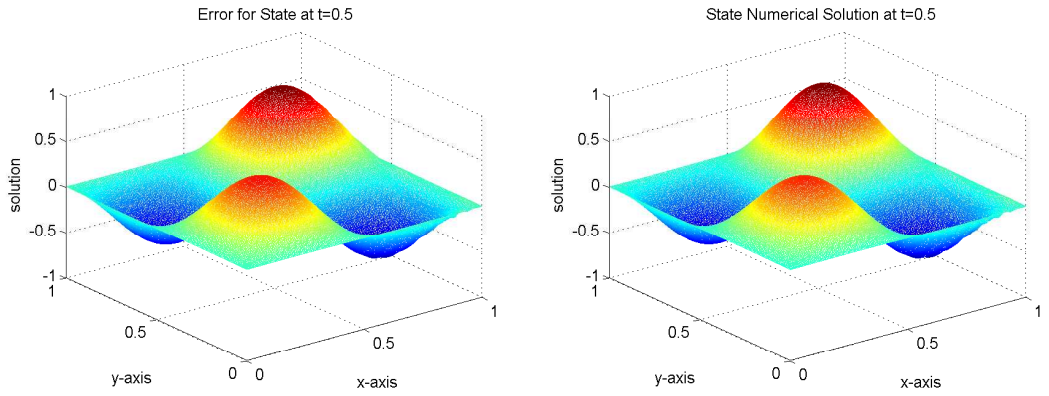

Figure 7.7: State solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

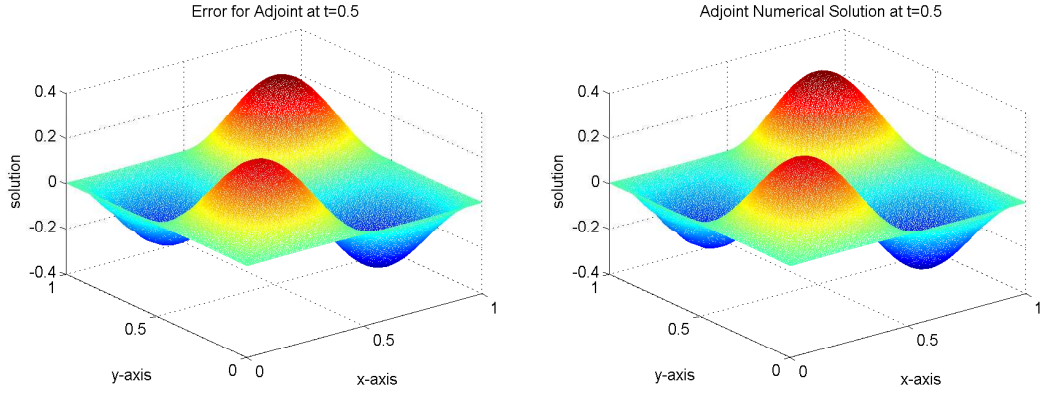

Figure 7.8: Adjoint solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

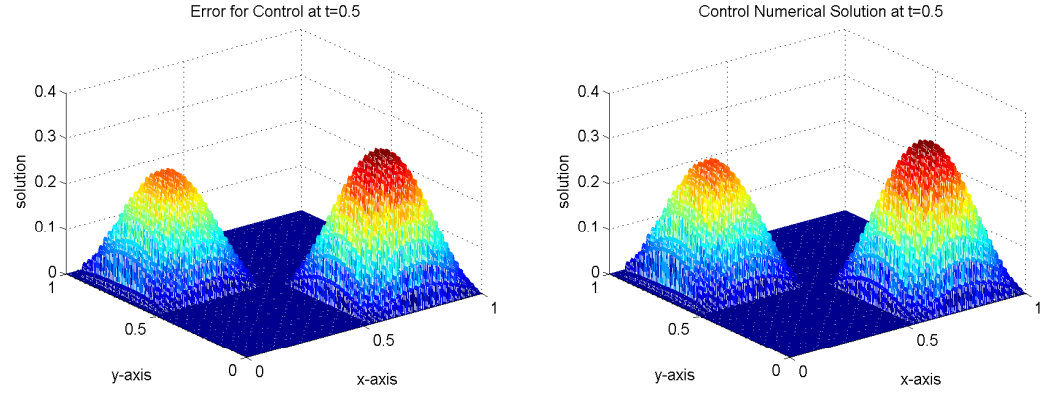

Figure 7.9: Control solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

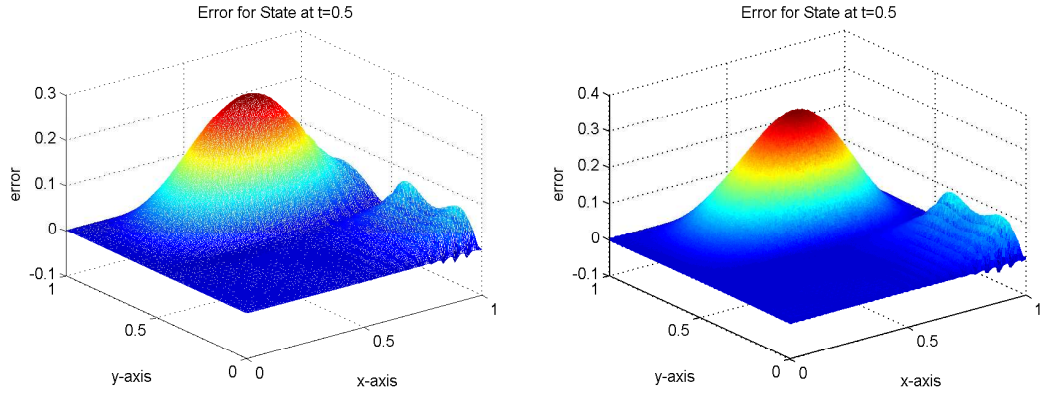

Figure 7.10: Error in the state solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

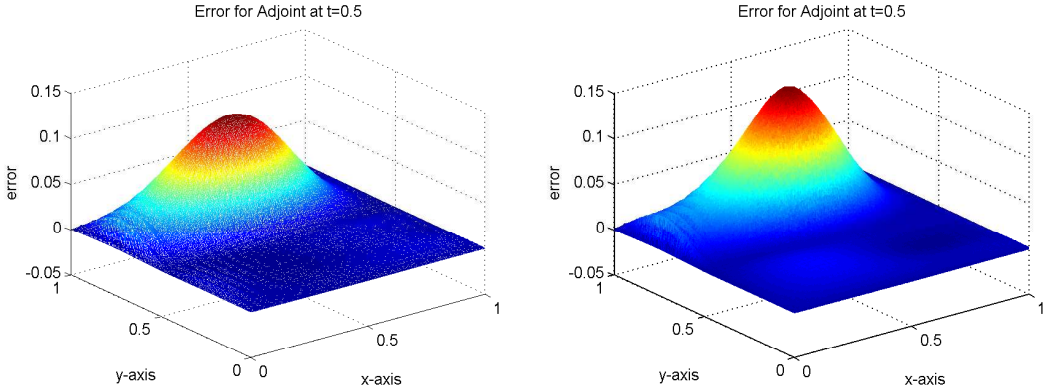

Figure 7.11: Error in the adjoint solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

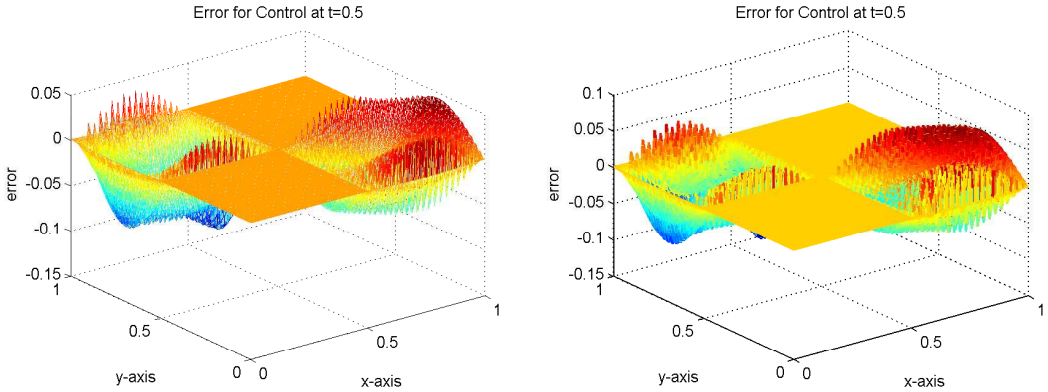

Figure 7.12: Error in the control solution at t=0.5 with $\Delta x = \Delta t = 1/40$, Piecewise Linear Elements, Backward Euler versus Crank-Nicolson

	SIPG			NIPG			IIPG		
Δt	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order
1/10	7	1.243731	-	8	1.243724	-	6	1.243730	-
1/20	8	1.226632	-	7	1.226623	-	7	1.226632	-
1/40	7	1.221406	1.710	8	1.221395	1.708	8	1.221406	1.710
1/80	8	1.219645	1.569	8	1.219632	1.569	8	1.219645	1.569
1/160	8	1.218983	1.410	7	1.218968	1.410	9	1.218983	1.410

Table 7.3: Piecewise Linear Elements - Backward Euler - Active Set

	SIPG			NIPG			IIPG		
Δt	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order
1/10	7	1.237547	-	9	1.237530	-	7	1.237547	-
1/20	6	1.223427	-	7	1.223412	-	7	1.223428	-
1/40	8	1.219801	1.961	9	1.219785	1.961	7	1.219801	1.961
1/80	8	1.218848	1.928	9	1.218832	1.928	8	1.218848	1.928
1/160	8	1.218586	1.866	8	1.218571	1.867	10	1.218586	1.867

Table 7.4: Piecewise Linear Elements - Crank-Nicolson - Active Set

	SIPG			NIPG			IIPG		
Δt	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order
1/10	7	1.243718	-	7	1.243718	-	7	1.243718	-
1/20	7	1.226617	-	7	1.226615	-	6	1.226617	-
1/40	7	1.221386	1.709	8	1.221384	1.709	7	1.221386	1.708
1/80	8	1.219620	1.567	7	1.219618	1.567	8	1.219620	1.565
1/160	7	1.218954	1.408	9	1.218952	1.407	10	1.218954	1.408

Table 7.5: Piecewise Quadratic Elements - Backward Euler - Active Set

	SIPG			NIPG			IIPG		
Δt	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order	I_A	$J_h(y_h,u_h)$	order
1/10	8	1.237501	-	8	1.237498	-	7	1.237501	-
1/20	6	1.223386	-	9	1.223384	-	6	1.223386	-
1/40	7	1.219764	1.962	8	1.219761	1.962	7	1.219764	1.962
1/80	8	1.218813	1.930	9	1.218810	1.930	8	1.218813	1.930
1/160	8	1.218553	1.868	9	1.218550	1.868	10	1.218553	1.868

Table 7.6: Piecewise Quadratic Elements - Crank-Nicolson - Active Set

For this problem, we have observed the temporal changes of the optimal control problem,too. As we decrease the size of the temporal subinterval, we could have obtained a smaller value of the optimal control problem as expected. We have used piecewise linear and quadratic polynomials. The numerical order of the objective function is related to the order of the state and the control, because the objective function is the sum of two terms: *The difference between the state and the desired state*, and *the control*. For the results obtained by using piecewise linear polynomials, average order of backward Euler is approximately 1.563 for SIPG and IIPG, 1.562 for NIPG. These orders are affected by the temporal subinterval Δt and the mesh size h. For Crank-Nicolson, orders are approximately 1.918 for SIPG and 1.919 for NIPG and IIPG. As one can observe that, the numerical order of backward Euler is little larger than the expected one. Actually, the a priori error analysis we have derived is valid for a general unsteady diffusion convection reaction equation. For the convection dominated problems, the solution contains boundary or interior layers and this may lead to pollute solution. However, if we decrease the length of the temporal subinterval, then the orders tend to decrease as expected. Our theoretical results confirm the numerical orders. The orders for piecewise quadratic polynomials are similar to the ones obtained by the piecewise linear polynomials. Indeed, the solution profiles are more similar to the exact solution in terms of the smoothness as expected. The accuracy of this kind of problems can be increased by hp-adaptivity. We have obtained valid numerical orders for both of the problem. The solution procedure with hp-adaptivity is more meaningful and there are many examples in the literature that are solved by hp-adaptivity.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this work, we have considered the linear-quadratic distributed optimal control problem governed by the unsteady diffusion convection reaction equation. We have discussed the existence and uniqueness of the optimal control problem and the diffusion convection reaction equation. Discontinuous Galerkin methods for one and two-dimensional problems have been introduced. We have performed spatial discretization by three types of discontinuous Galerkin method: nonsymmetric interior penalty Galerkin (NIPG) method, symmetric interior penalty Galerkin (SIPG) method and incomplete interior penalty Galerkin (IIPG) method. For temporal discretization, two implicit methods, backward Euler and Crank-Nicolson have been used to discretize the semi-discrete problem in time. Then, we have converted the infinite-dimensional problem into a finite-dimensional one.

We have analyzed the the problem by conducting the stability and convergence estimates for the semi-discrete state equation. In addition, these estimates are provided for the full-discrete state, adjoint equation and the control. We have determined the order of the methods in space and in time.

We have solved a one-dimensional unconstrained distributed optimal control problem and a two-dimensional constrained distributed optimal control problem. Solution profiles of the state, adjoint and the control have been given and numerical orders of the optimal control problem have been computed for both of the problems and they are confirmed by a priori error analysis.

As a future work, we are going to focus on optimize-then-discretize and discretize-then-optimize approaches and try to make these two approaches commutative. Two studies for parabolic PDEs and Stokes flow problem are recently online [1, 57]. In these articles, a vari-

ant of time integration techniques can enable these approaches to commute. Nicolson scheme is suggested in [1] for temporal discretization of the optimal control problem governed by parabolic PDEs.

REFERENCES

[1] T. Apel and T.G. Flaig. Crank-Nicolson schemes for optimal control problems with evolution equations. December 10, 2010. http://www.am.uni-erlangen.de/home/spp1253.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Discontinuous Galerkin methods for elliptic problems. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, pages 89–101. Springer, 2000.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39:1749–1779, 2001/02.

[4] D. A. Beard and J. B. Bassingthwaighte. Modelling advection and diffusion of oxygen in complex vascular networks. *Annals of Biomedical Engineering*, 29:298–310, 2001.

[5] R. Becker, D. Meidner, and B. Vexler. Efficient numerical solution of parabolic optimization problems by finite element methods. *Optim. Methods Softw.*, 22:813–833, 2007.

[6] R. Becker and B. Vexler. Optimal control of the convection-diffusion equation using stabilized finite element methods. *Numer. Math.*, 106:349–367, 2007.

[7] M. Braack. Optimal control in fluid mechanics by finite elements with symmetric stabilization. *SIAM J. Control Optim.*, 48:672–687, 2009.

[8] E. Burman. Crank-nicolson finite element methods using symmetric stabilization with an application to optimal control problems subject to transient advection-diffusion equations. *Communications in Mathematical Sciences*, 9:319–329, 2011.

[9] E. Burman and M. A. Fernández. Finite element methods with symmetric stabilization for the transient convection-diffusion-reaction equation. *Comput. Methods Appl. Mech. Engrg.*, 198:2508–2519, 2009.

[10] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1437–1453, 2004.

[11] P. Castillo. Performance of discontinuous Galerkin methods for elliptic PDEs. *SIAM J. Sci. Comput.*, 24:524–547, 2002.

[12] X. Chunguang and L. Yuan. Error analysis for optimal control problem governed by convection diffusion equations: Dg method. *J. Computational Applied Mathematics*, 235:3163–3177, 2011.

[13] B. Cockburn. Discontinuous Galerkin methods. *ZAMM Z. Angew. Math. Mech.*, 83:731–754, 2003.

[14] B. Cockburn and C. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35:2440–2463 (electronic), 1998.

[15] B. Cockburn and C. Shu. Runge-kutta discontinuous galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16:173–261, 2001.

[16] S.S. Collis and M. Heinkenschloss. Analysis of the streamline upwind/ petrov galerkin method applied to the solution of optimal control problems. Technical Report Technical Report TR02/01, Department of Computational and Applied Mathematics, Rice University, Houston, 2002. http://www.caam.rice.edu/h̃einken.

[17] S.S. Collis and M. Heinkenschloss. Numerical solution of implicitly constrained optimization problems. Technical Report Technical Report TR08/05, Department of Computational and Applied Mathematics, Rice University, Houston, 2008. http://www.caam.rice.edu/h̃einken.

[18] L. Dede. *Adaptive and reduced basis methods for optimal control problems in environmental applications*. PhD thesis, Mathematical Engineering, Politechnico Di Milano, Milano, 2008.

[19] L. Dede and A. Quarteroni. Optimal control and numerical adaptivity for advection-diffusion equation. *Mathematical Modelling and Numerical Analysis*, 39:1019–1040, 2005.

[20] V. Dolejší, M. Feistauer, and J. Hozman. Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems on nonconforming meshes. *Comput. Methods Appl. Mech. Engrg.*, 196:2813–2827, 2007.

[21] M. Dominik. *Adaptive Space-Time Finite element Methods For Optimization Problems Governed By Nonlinear Parabolic Systems*. PhD thesis, University of Heidelberg, 2008. http://archiv.ub.uni-heidelberg.de/volltextserver/volltexte/2008/8272/.

[22] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. Wiley, 2003.

[23] C. Eduardo. *Optimal Control of PDE Theory and Numerical Analysis, Wuhan University*. June, 2007. www.gmcnetwork.org/files/summerschool/2007/edcasas.full.pdf.

[24] M. Feistauer, J. Hájek, and K. Svadlenka. Space-time discontinuous Galerkin method for solving nonstationary convection-diffusion-reaction problems. *Appl. Math.*, 52:197–233, 2007.

[25] H. Fu. A characteristic finite element method for optimal control problems governed by convection diffusion equations. *J. Comput. Appl. Math.*, 235:825–836, 2010.

[26] H. Fu and H. Rui. A priori error estimates for optimal control problems governed by transient advection-diffusion equations. *J. Sci. Comput.*, 38:290–315, 2009.

[27] M. S. Gockenbach. *Understanding and implementing the finite element method*. Society for Industrial and Applied Mathematics (SIAM), 2006.

[28] P.M. Gresho and R.L. Sani. *Incompressible Flow and the Finite Element Element Method, Vol 1:Advection-Diffusion*. Wiley, 2000.

[29] D. F. Griffiths and D. J. Higham. *Numerical methods for ordinary differential equations*. Springer-Verlag London Ltd., 2010.

[30] M. Heinkenschloss and D. Leykekhman. Local error analysis of discontinuous galerkin methods for advection-dominated elliptic linear-quadratic optimal control problems. Technical Report Technical Report TR10-11, Department of Computational and Applied Mathematics, Rice University, Houston, 2010. http://www.caam.rice.edu/~heinken.

[31] M. Heinkenschloss and D. Leykekhman. Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. *SIAM J. Numer. Anal.*, 47:4607–4638, 2010.

[32] M. Hintermüller and R. H. W. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM J. Control Optim.*, 47:1721–1743, 2008.

[33] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Springer, 2009.

[34] M. Hinze and F. Tröltzsch. Discrete concepts versus error analysis in pde constrained optimization. *GAMM-Mitt*, 33:148–162, 2010.

[35] M. Hinze, N. Yan, and Z. Zhou. Variational discretization for optimal control governed by convection dominated diffusion equations. *J. Comput. Math.*, 27:237–253, 2009.

[36] R. H.W. Hoppe. Numerical solution of parabolic optimal control problems, Summer school optimal control of pdes, Cortona, Italy, 2010.

[37] P. Houston, C. Schwab, and E. Süli. Stabilized *hp*-finite element methods for first-order hyperbolic problems. *SIAM J. Numer. Anal.*, 37:1618–1643 (electronic), 2000.

[38] P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39:2133–2163 (electronic), 2002.

[39] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*. Springer-Verlag, 2003.

[40] A. Kröner and B. Vexler. A priori error estimates for elliptic optimal control problems with a bilinear state equation. *J. Comput. Appl. Math.*, 230:781–802, 2009.

[41] K. Kunisch and B. Vexler. Constrained Dirichlet boundary control in L^2 for a class of evolution equations. *SIAM J. Control Optim.*, 46:1726–1753 (electronic), 2007.

[42] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations*. Society for Industrial and Applied Mathematics (SIAM), 2007.

[43] J.-L. Lions. *Optimal control of systems governed by partial differential equations.* Springer-Verlag, 1971.

[44] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. ii. problems with control constraints. *SIAM J. Control Optim.*, 47:1301–1329, 2008.

[45] D. Meidner and B. Vexler. A priori error analysis of the petrov Galerkin Crank-Nicolson scheme for parabolic optimal control problems. submitted,2010. http://www-m1.ma.tum.de/bin/view/Lehrstuhl/DominikMeidner.

[46] C. Meyer. Lecture notes on "Optimal Control with PDE's, TU Darmstadt, 2010. http://www.graduate-school-ce.de/index.php?id=131.

[47] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer-Verlag, 1999.

[48] J. Peraire and P.-O. Persson. The compact discontinuous Galerkin (CDG) method for elliptic problems. *SIAM J. Sci. Comput.*, 30:1806–1824, 2008.

[49] A. Quarteroni. *Numerical models for differential problems*. Springer-Verlag Italia, Milan, 2009.

[50] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*. Springer-Verlag, 1994.

[51] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[52] T. Rees, M. Stoll, and A. Wathen. All-at-once preconditioning in PDE-constrained optimization. *Kybernetika (Prague)*, 46:341–360, 2010.

[53] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*. Society for Industrial and Applied Mathematics (SIAM), 2008. Theory and implementation.

[54] B. Rivière and M. F. Wheeler. Non conforming methods for transport with nonlinear reaction. In *Fluid flow and transport in porous media: mathematical and numerical treatment (South Hadley, MA, 2001)*, pages 421–432. Amer. Math. Soc., 2002.

[55] G. Roland. Short course on infinite-dimensional optimization, University of Bremen, January, 2006. http://www.ricam.oeaw.ac.at/people/page/griesse/.

[56] M. Stoll and A. Wathen. All-at-once solution of time-dependent pde-constrained optimization problems. Technical Report Report Number 10/47, Oxford Center For Collaborative Applied Mathematics, University of Oxford, Oxford, 2010.

[57] M. Stoll and A. Wathen. All-at-once solution of time-dependent stokes control. June 8, 2011. http://www.mpi-magdeburg.mpg.de/preprints/.

[58] J. J. Sudirham, J. J. W. van der Vegt, and R. M. J. van Damme. Space-time discontinuous Galerkin method for advection-diffusion problems on time-dependent domains. *Appl. Numer. Math.*, 56:1491–1518, 2006.

[59] T. Sun. Discontinuous Galerkin finite element method with interior penalties for convection diffusion optimal control problem. *Int. J. Numer. Anal. Model.*, 7:87–107, 2010.

[60] F. Tröltzsch. *Optimal control of partial differential equations*. American Mathematical Society, 2010.

[61] A. Yazdani and L. Shojai. Solution of a scalar convection-diffusion equation using FEMLAB. *ArXiv e-prints*, 2011. http://adsabs.harvard.edu/abs/2011arXiv1101.1809Y.