A New Class of Discontinuous Galerkin Methods for Wave Equations in Second-Order Form

Lu Zhang
luzhang@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_mathematics_etds

Part of the Numerical Analysis and Computation Commons, and the Partial Differential Equations Commons

Recommended Citation
https://scholar.smu.edu/hum_sci_mathematics_etds/5

This Dissertation is brought to you for free and open access by the Mathematics at SMU Scholar. It has been accepted for inclusion in Mathematics Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
A NEW CLASS OF DISCONTINUOUS GALERKIN METHODS FOR WAVE EQUATIONS IN SECOND-ORDER FORM

Approved by:

Dr. Thomas Hagstrom
Professor of Mathematics

Dr. Daniel Appelö
Associate Professor of Mathematics

Dr. Daniel Reynolds
Professor of Mathematics

Dr. Johannes Tausch
Professor of Mathematics
A NEW CLASS OF DISCONTINUOUS GALERKIN METHODS FOR WAVE EQUATIONS IN SECOND-ORDER FORM

A Dissertation Presented to the Graduate Faculty of the

Dedman College
Southern Methodist University

in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy

with a
Major in Computational and Applied Mathematics

by
Lu Zhang

B.S., Applied Mathematics, Southwest University of Finance and Economics
M.S., Computational and Applied Mathematics, Southern Methodist University

May 16, 2020
ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisor Professor Thomas Hagstrom for his guidance, support, encouragement, and friendship during my study at Southern Methodist University. All my achievements together with this thesis would have not been possible without his time and efforts.

I would like to thank Professor Daniel Appelö for the guidance and support, and I have learned a lot from him during our collaboration. Thank Dr. Anders Petersson, Professors Yu Du, Siyang Wang, Qi Wang, and Zhimin Zhang with whom I have developed joyful and fruitful collaborations over the years. My gratitude also goes to Professors Alejandro Aceves, Wei Cai, Daniel Reynolds, Johannes Tausch and so many others who generously offer their help and support over the years. Thank Liz Berry for keeping SMU Mathematics Department so well organized and I will deeply miss it.

This work was supported by NSF Grant DMS-1418871 and by the Dean’s Dissertation Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Finally, I devote this thesis to my family, for understanding, support and endless love.
Discontinuous Galerkin methods are widely used in many practical fields. In this thesis, we focus on a new class of discontinuous Galerkin methods for second-order wave equations. This thesis is constructed by three main parts. In the first part, we study the convergence properties of the energy-based discontinuous Galerkin proposed in [3] for wave equations. We improve the existing suboptimal error estimates to an optimal convergence rate in the energy norm. In the second part, we generalize the energy-based discontinuous Galerkin method proposed in [3] to the advective wave equation and semilinear wave equation in second-order form. Energy-conserving or energy-dissipating methods follow from simple, mesh-independent choices of the interelement fluxes. Error estimates in the energy norm are established. In the third part, we focus on establishing methods to overcome the computational stiffness from the high-order piecewise polynomial approximations in the energy-based discontinuous Galerkin methods and reduce the computational cost of the inversion of the stiffness matrix.
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................. x

LIST OF TABLES ................................................................. xiv

CHAPTER

1. INTRODUCTION ................................................................. 1

2. CONVERGENCE ANALYSIS ................................................. 5
   2.1. Introduction ........................................................... 5
   2.2. General formulation .................................................. 6
      2.2.1. Fluxes ............................................................ 8
   2.3. Cartesian meshes ..................................................... 9
      2.3.1. Problem statement and elliptic projection .................... 10
      2.3.2. Supercloseness and optimal convergence analysis .......... 11
   2.4. Quadrilateral meshes ................................................. 20
      2.4.1. Elliptic projection and penalty term ........................ 20
      2.4.2. Supercloseness and optimal convergence analysis .......... 22
   2.5. Numerical examples .................................................. 30
      2.5.1. Verification on Cartesian meshes ......................... 31
      2.5.2. Verification on quadrilateral meshes ...................... 31

3. APPLICATION TO CONVECTIVE WAVE EQUATIONS ......................... 36
   3.1. Introduction ........................................................... 36
   3.2. DG formulation ....................................................... 37
      3.2.1. Fluxes ............................................................ 40
      3.2.2. Boundary conditions .......................................... 43
         3.2.2.1. Inflow boundary conditions ......................... 44
         3.2.2.2. Outflow boundary conditions ...................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Error estimates in the energy norm</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>3.3.1. General case</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.3.2. Improved estimates for one dimension</td>
<td>53</td>
</tr>
<tr>
<td>3.4</td>
<td>Numerical experiments</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>3.4.1. Periodic boundary conditions in one space dimension</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>3.4.2. Periodic boundary conditions in two space dimensions</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>3.4.3. Dirichlet and radiation boundary conditions in two space dimensions</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>APPLICATION TO SEMI-LINEAR WAVE EQUATIONS</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>4.1. Introduction</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>4.2. Problem formulation</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>4.2.1. Semi-discrete DG formulation</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>4.2.2. Fluxes</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>4.2.2.1. Inter-element boundaries</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>4.2.2.2. Physical boundaries</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>4.2.3. Stability of the scheme</td>
<td>73</td>
</tr>
<tr>
<td>4.3</td>
<td>Error estimates</td>
<td>73</td>
</tr>
<tr>
<td>4.4</td>
<td>Numerical experiments</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>4.4.1. Convergence in 1D</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>4.4.2. Soliton solutions of the sine-Gordon equation in 1D</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>4.4.2.1. Standing breather soliton</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>4.4.2.2. Kink soliton and antikink soliton</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>4.4.2.3. Kink-kink collision and kink-antikink collision</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>4.4.3. Convergence in 2D</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>4.4.4. Time history of the numerical energy in 2D</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>4.4.5. Focusing equation</td>
<td>98</td>
</tr>
</tbody>
</table>
5. STAGGERED GRIDS

5.1. Introduction

5.2. Energy Based Discontinuous Galerkin Method for the Scalar Wave Equation

5.2.1. Staggered Formulation

5.3. Operator Bounds in One Dimension on Periodic Domains

5.3.0.1. Operator Bounds for the Non-staggered Scheme

5.3.1. Staggered Scheme

5.3.1.1. Operator Bounds for the Staggered Scheme

5.4. Local Timestepping

5.5. Numerical Experiments

6. GALERKIN DIFFERENCE BASIS

6.1. Introduction

6.2. Preliminaries

6.2.1. The Galerkin Difference Basis

6.2.1.1. Modification Near Boundaries

6.2.1.2. Extension to Higher Dimensions

6.2.2. Energy Based Discontinuous Galerkin Method for the Wave Equation

6.3. Efficient Formulation on Cartesian Grids

6.3.1. Complexity with Galerkin Difference Basis

6.3.2. Optimal Computational Complexity by Simultaneous Diagonalization
6.3.3. Numerical Verification of the Computational Complexity .......... 133
6.4. Dispersion Analysis .................................................. 134
6.5. Spectral Radius ...................................................... 136
6.6. Numerical experiments .............................................. 137
   6.6.1. Periodic boundary conditions in one dimension ................. 138
   6.6.2. Dirichlet and Neumann boundary conditions in one dimension .... 139
   6.6.3. Periodic boundary conditions in two dimensions ............... 139
7. CONCLUSION ............................................................... 149
BIBLIOGRAPHY ............................................................. 151


LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The $L^2$-norm errors of $v_h$ on Cartesian meshes for different $p$ and $\epsilon$.</td>
<td>32</td>
</tr>
<tr>
<td>2.2</td>
<td>The broken $H^1$-seminorm errors of $u_h$ and the errors between $I_h u$ and $u_h$ on Cartesian meshes for different $p$ and $\epsilon$.</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>The errors of numerical solutions on perturbed grids with $\alpha = 2$ for $\epsilon = 0.5$.</td>
<td>34</td>
</tr>
<tr>
<td>2.4</td>
<td>The errors of numerical solutions on perturbed grids with $\alpha = 0.5$ for $\epsilon = 1$.</td>
<td>35</td>
</tr>
<tr>
<td>3.1</td>
<td>Plots of the errors in $u$ as a function of $h$ in 1d with upwind flux for periodic boundary condition. In the legend, $q$ is the maximum degree of the approximation of $u$, solid lines represent the case of $u^h$ and $v^h$ in the same approximation space, dotted lines represent the case of $v^h$ one degree lower than $u^h$.</td>
<td>57</td>
</tr>
<tr>
<td>3.2</td>
<td>Plots of the errors in $u$ as a function of grid spacing $h$ in 1d with central flux for periodic boundary condition. In the legend, $q$ is the maximum degree of the approximation space of $u$, solid lines represent the case of $u^h$ and $v^h$ in the same space, dotted lines represent the case of $v^h$ one degree lower than $u^h$.</td>
<td>59</td>
</tr>
<tr>
<td>3.3</td>
<td>Plots of the error in $u$ as a function of spacing-grid $h$ in 2d with the upwind flux and periodic boundary conditions. In the legend, $q$ is the degree of the approximation of $u$ for both $x$ and $y$ directions.</td>
<td>61</td>
</tr>
<tr>
<td>3.4</td>
<td>Plots of the error in $u$ as a function of $h$ in 2d with the central flux and periodic boundary conditions. In the legend, $q$ is the degree of approximation to $u$ and $v$ for both $x$ and $y$ directions.</td>
<td>62</td>
</tr>
<tr>
<td>3.5</td>
<td>Plots of the error in $u$ as a function of $h$ in 2d with upwind (left) and central (right) fluxes for Dirichlet boundary condition on inflow boundaries and a radiation boundary condition on outflow boundaries. In the legend, $q$ is the degree of the approximation to $u$ for both $x$ and $y$ directions.</td>
<td>63</td>
</tr>
<tr>
<td>4.1</td>
<td>Plots of the history of the numerical energy for the standing breather solution.</td>
<td>88</td>
</tr>
</tbody>
</table>
4.2 Plots of the standing breather with the degree of approximation space \( q = s = 4 \). The S.-flux is used in the simulation. From the top to the bottom, the left to the right, the numerical and exact breather solutions at \( t = 0, 45, 90, 120 \) are plotted.

4.3 Space-time plots of the standing breather with the degree of approximation space \( q = s = 4 \). The S.-flux is used in the simulation.

4.4 Plots of the history of the \( L^2 \) errors for \( u \), standing breather. The first row is for energy-conserving schemes, from the left to right the A.-flux and the C.-flux respectively. The second row is for energy-dissipating schemes, from the left to right the A.S.-flux and the S.-flux respectively. The degree of the approximation space for \( u \) is \( q = 4 \) and for \( v \) is \( s \).

4.5 From the left to the right, plots for the kink and antikink solitons respectively. The degree of approximation space for \( u \) is \( q = 4 \) and for \( v \) is \( s = 3 \). An energy-based DG scheme with the A.-flux is used in the simulation of kink soliton and the C.-flux is used for the simulation of antikink soliton.

4.6 From the left to the right plots of the kink-kink and kink-antikink collisions respectively: the degree of approximation space for \( u \) is \( q = 4 \) and for \( v \) is \( s = 3 \). An energy-based DG scheme with A.S.-flux is used in the simulation of kink-kink collision and the S.-flux is used for the simulation of the kink-antikink collision.

4.7 The plot of \( L^2 \) errors for \( u \) for the 2D convergence test: from left to right, top to bottom A.-flux, C.-flux, A.S.-flux and S.-flux respectively. The approximation degrees for \( u^h, v^h \) are \( q_x = q_y = s_x = s_y = q \).

4.8 The plots of energy for DG solutions of (4.42) with four different fluxes. For the left graph, the dissipating term is considered, i.e., \( \theta = 1 \); while the right graph does not contain the dissipating term, i.e., \( \theta = 0 \).

4.9 Plots of \( u \) for the focusing equation (4.43) at times \( t = 0, 0.69, 6, 10 \) with the S.-flux and \( q_x = q_y = s_x = s_y = q = 4 \). For the left column, \( \theta = 0 \); for the right column, \( \theta = 1 \).

5.1 To the left are the \( L^2 \)-errors in \( u^h \) for \( q_u = 2, 6, \ldots \) and to the right for \( q_u = 3, 7, \ldots \). The dashed lines are indicates rates of convergence \( 2, 6, 10, \ldots \) to the left and \( 4, 8, 12, \ldots \) to the right.

5.2 The figure displays the spectral radii of the time stepping matrix \( A \) scaled by the element size \( h \) and the reciprocal of \( q_u \) as a function of \( q_u \).

5.3 To the left are the spectrum for \( q_u = 14 \) and \( q_v = 13 \) with NOG = 2 and to the right are the spectrum for \( q_u = 14 \) with NOG = 3.
5.4 From the left to the right are the spectrum for \( q_u = 14 \) and \( q_v = 13 \) with NOG = 3 and number of DG elements for \( u \) are \( n = 20, 40, 80 \), respectively.

5.5 Staggered grids in two dimensions. Blue boxes are DG elements for \( u \) and red boxes are DG elements for \( v \). Here, we have \( 3 \times 3 = 9 \) DG elements for \( u \) and \( 4 \times 4 = 16 \) DG elements for \( v \).

5.6 The plot of \( L^2 \) errors for \( u \) : from left to right are \( q_u = q_v = 2, 3 \) and \( q_u = q_v = 6, 7 \), respectively.

5.7 The top panel, from the left to the right are the spectral radii when \( q = 2, 3 \), respectively. The bottom panel, from the left to the right are the spectral radii when \( q = 6, 7 \), respectively.

6.1 To the left, the generating basis function \( \Phi_p(x) \) for \( p = 1, 3, 5, 7, 9 \). Note that \( \Phi_p(x) = \Phi_p(-x) \). Note that each function is vertically offset by \( q - 1 \). In the middle, the “right” four degree 7 Lagrange polynomials that make up \( \Phi_7(x) \), the part of \( L_j \) that is used is in bold. To the right, a three dimensional representation of the generating basis function \( \Phi_p(x) \) for \( p \) up to 39.

6.2 Plots of CPU time for the standard Galerkin difference basis (left column) and the Galerkin difference basis by simultaneous diagonalization (right column) as a function of degrees of freedom (DOF) in two dimensions. The first row is for \( n = 1 \) and the second row is for \( n = 6 \). The sound wave speed \( c = 1 \), and the splitting parameter for upwind flux is \( \xi = c \).

6.3 On the top, we show the numerical dispersion relation for the linear acoustic operator with the Sommerfeld flux. The black dashed lines represent the exact case. ‘p.-mode’ represents the physical mode and ‘s.-mode’ represent the spurious mode. On the bottom, we illustrate the dissipation associated with the eight modes. The sound wave speed is \( c = 1 \) and the degree of the approximation spaces for both \( u \) and \( v \) is \( p = 3 \). The flux splitting parameter \( \xi = c = 1 \).

6.4 On the top, we show the numerical dispersion relation for the linear acoustic operator with the Sommerfeld flux. The black dashed line represents the exact case. We present the numerical dispersion relation for the physical mode at different orders \( p = (1, 3, 5, 7, 9) \). On the bottom, we illustrate the dissipation associated with the different orders \( p = (1, 3, 5, 7, 9) \). The sound wave speed \( c \) and the flux splitting parameter \( \xi \) are chosen to be same \( c = \xi = 1 \).
6.5 On the top, we show the numerical dispersion relation for the linear acoustic operator with the central flux. On the bottom, we show the numerical dispersion relation for the linear acoustic operator with the alternating flux. The black dashed lines represent the exact case. ‘p.-mode’ represents the physical mode and ‘s.-mode’ represent the spurious mode. The sound wave speed is $c = 1$ and the degree of the approximation spaces for both $u$ and $v$ is $p = 3$. .................................................. 144

6.6 On the left, we show the numerical dissipation for the linear acoustic operator with the central flux for all modes. On the right, we show the numerical dissipation for the linear acoustic operator with the alternating flux for all modes. The sound wave speed is $c = 1$ and the degree of the approximation spaces for both $u$ and $v$ is $p = 3$. ‘p.-mode’ represents physical mode and ‘s.-mode’ represents the spurious mode. .................. 145

6.7 Spectral radii (scaled by $h = 1/N$) as a function of the degree of approximation ($p$) for the three different fluxes. The sound wave speed is $c = 1$, and the splitting parameter $\xi$ for upwind flux is $\xi = c$. .................. 145

6.8 Plots of the error in $u$ as a function of $h$ in one dimension with upwind flux, central flux and alternating flux for periodic boundary conditions. In the legend, $p$ is the degree of the approximation space of $u$. .................. 146

6.9 Plots of the error in $u$ as a function of $h$ in one dimension with upwind flux, central flux and alternating flux for Dirichlet boundary condition on the left and Neumann boundary condition on the right. In the legend, $p$ is the degree of the approximation space of $u$. .................. 147

6.10 Plots of the error in $u$ as a function of $h$ in two dimensions with upwind flux, central flux and alternating flux for periodic boundary conditions. In the legend, $p$ is the degree of the approximation space of $u$. .................. 148
LIST OF TABLES

Table | Page
-----|-----
2.1  The convergence rates of various errors on Cartesian meshes for different $p$ and $\epsilon$. | 31
2.2  The convergence rates of various errors on perturbed grids with $\alpha = 2$ for different $p$ and $\epsilon$. | 35
2.3  The convergence rates of various errors on perturbed grids with $\alpha = 0, 0.5, 1$ for fixed $\epsilon = 1$ and different $p$. | 35
3.1  Linear regression estimates of the convergence rate of $u$ in 1$d$ with upwind flux for periodic boundary condition, approximation for $v$ is one degree lower than $u$. | 58
3.2  Linear regression estimates of the convergence rate of $u$ in 1$d$ with upwind flux for periodic boundary condition, the approximation for $v$ is one degree lower than $u$. | 58
3.3  Linear regression estimates of the convergence rate for $u$ and $v$ in 1$d$ with central flux for periodic boundary condition, the approximation for $v$ is one degree lower than $u$. | 60
3.4  Linear regression estimates of the convergence rate of $u$ and $v$ in 1$d$ with central flux for periodic boundary condition, $u$ and $v$ are in the same approximation space. | 60
3.5  Linear regression estimates of the convergence rate of $u$ and $v$ in 2$d$ with upwind flux for periodic boundary condition and $q_x = q_y = q$. | 63
3.6  Linear regression estimates of the convergence rate of $u$ and $v$ in 2$d$ with central flux for periodic boundary condition and $q_x = q_y = q$. | 64
3.7  Linear regression estimates of the convergence rate for $u$ in 2$d$ with Dirichlet boundary condition on inflow boundaries, radiation boundary condition on outflow boundaries and $q_x = q_y = q$. Here the first two rows correspond to the upwind flux and the last two to the central flux. | 65
4.1 $L^2$ errors in $u$ for problem 1 (4.35) when the A.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$, and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.2 $L^2$ errors in $u$ for problem 2 (4.36) when the A.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.3 $L^2$ errors in $u$ for problem 1 (4.35) when the C.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.4 $L^2$ errors in $u$ for problem 2 (4.36) when the C.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.5 $L^2$ errors in $u$ for problem 1 (4.35) when the S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.6 $L^2$ errors in $u$ for problem 2 (4.36) when the S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.7 $L^2$ errors in $u$ for problem 1 (4.35) when the A.S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.8 $L^2$ errors in $u$ for problem 2 (4.36) when the A.S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

4.9 Linear regression estimates of the convergence rate for $u$ with C.-flux, A.-flux, S.-flux and A.S.-flux for the 2D test problem. The approximation degrees for $u^h, v^h$ are $q_x = q_y = s_x = s_y = q$.

5.1 Linear regression estimates of the convergence rate for $u$ with central flux in two dimensions. The degree of the approximation space for $u$ and $v$ are $q$ for both $x$ and $y$ directions.

6.1 Linear regression estimates of the convergence rate of $u$ in one dimensional with periodic boundary condition.

6.2 Linear regression estimates of the convergence rate of $u$ in one dimension with Dirichlet boundary condition on the left and Neumann boundary condition on the right. first 14 points for $q = 9$. 

xv
6.3 Linear regression estimates of the convergence rate of \( u \) in two dimensions with periodic boundary conditions. ........................................ 141
This thesis is dedicated to my family.
Chapter 1

INTRODUCTION

The discontinuous Galerkin (DG) method is a combination of finite element and finite volume methods, using piecewise polynomial basis functions. The DG method was firstly introduced in 1973 to solve the steady-state neutron transport equation by Reed and Hill in [69]. In 1974, LeSaint and Raviart [56] conducted the first analysis of this DG method. Let $h$ be the cell size and $q$ be the degree of the polynomial approximation. They demonstrated $O(h^q)$ convergence rate for a general mesh and $O(h^{q+1})$ convergence rate for a Cartesian mesh. In 1986, Johnson and Pitkäranta [50] improved the result to $O(h^{q+\frac{1}{2}})$ for a general mesh. Later, the DG methods were used to solve both first-order hyperbolic equations [30,31,33,41,49,71,72] and second-order elliptic equations [8,21,55,68,77,78,84,86].

In recent years, DG methods have become very popular in many fields of science, engineering and industry because of their nice properties: (1) easily handle complex geometries; (2) arbitrary high-order accuracy; (3) hp-adaptivity; (4) explicit semi-discrete form for time dependent problems; (5) conservative element-wise; (6) effectively capture the discontinuity of the solution; (7) allow unstructured grids; (8) highly parallelizable because of local data communication.

DG methods have proven to be robust, high-order, and geometrically flexible when used to solve first-order systems in Friedrichs form [46]. But the basic wave equations arising in physical theories are expressed as action principles for a Lagrangian, leading directly to second-order equations. Even though it is possible to write second-order hyperbolic equations in the first-order form, the first-order formulation is quite different. It may need more boundary conditions and it is only equivalent to the original second-order equation for constrained data. Besides, not all second-order hyperbolic equations can be rewritten as a first-order system which is in Friedrichs form. A wide variety of other DG methods
have been proposed to solve second order wave equations, like local discontinuous Galerkin method (LDG), interior penalty parameter discontinuous Galerkin method (IPDG).

The LDG method was firstly proposed by Cockburn and Shu in [34] to solve time dependent convection-diffusion systems. This work was motivated by the work of Bassi and Rebay in [13] for the compressible Navier-Stokes equations. The idea of LDG methods is to introduce auxiliary variables, spatial derivatives of the solution, to rewrite a high order partial differential equation (PDE) into a first-order system. And then use the idea of DG methods to solve the system. The numerical fluxes for both interfaces of interelements and physical boundaries are key points for the method, which are used to guarantee the stability of the method. In [25], Cockburn and Dawson extended the original work in [34] and further analyzed the problem with non-constant diffusion coefficient in multi-dimensions and non-periodic boundary conditions. Later, Castillo et. al [19] presented the first a priori error analysis for the LDG method for a model elliptic problem. In [27], Cockburn et. al showed a superconvergence result for the LDG method for a model elliptic problem on Cartesian grids. Castillo et. al further explored the convergence properties of the $hp$-version of the LDG method for convection-diffusion problems in [20]. We refer to [9, 10, 18, 26, 29, 31, 32, 35, 36, 58, 62, 82, 87–91] and the references therein for more details and applications of LDG methods.

Symmetric interior penalty parameter discontinuous Galerkin (SIPDG) methods were firstly introduced by Wheeler [84], which used a discontinuous collocation-finite element method with interior penalties to solve second-order elliptic problems. Later, Arnold [8] extended the method to second-order nonlinear parabolic boundary value problems. Another IPDG method is the non-symmetric interior penalty discontinuous Galerkin (NIPDG) method. In [78], Rivière et. al antisymmetrized the bilinear form of the interior penalty Galerkin method proposed in [8, 37, 84, 85] to solve second-order elliptic problems. Houston et. al [47] analyzed the $hp$-version of the method for second-order PDEs with nonnegative characteristic form. The essential idea for IPDG methods is to add two stabilization terms. One is a symmetrizing term corresponding to fluxes on the interfaces of DG elements; the
other is a penalty term that imposes a weak continuity of the numerical solution. The difference between SIPDG and NIPDG is only one sign: the symmetrizing term is added to SIPDG and subtracted from NIPDG. For more details and applications of IPDG methods for PDEs, we refer to [11, 14, 28, 42, 51, 61, 67, 73, 75, 78, 79, 81] and the references therein.

Both LDG methods and IPDG methods admit many attractive properties when solving a wide range of PDEs. But there are small drawbacks. For LDG methods, we need to introduce extra fields which are spatial derivatives. This makes computation inefficient, especially for problems in high dimensions (it already quadrupling the number of fields in three space dimensions). For IPDG methods, the stability of the schemes depends on the mesh-dependent and order dependent penalty parameters. In this thesis, we propose a new class of DG methods to solve various wave equations in second-order form. One goal of this thesis is to minimize the extra auxiliary fields and use simple numerical fluxes to guarantee the stability of the scheme.

Another challenge is that high order accurate methods are superior for propagating waves over many periods [53]. Since 1972 much research has been devoted to spectral and high order accurate methods and as a result many highly accurate finite-difference, finite-element, spectral element and discontinuous Galerkin methods have been developed. Of course, the past half-century also saw rapid progress not only in the applied sciences, with their need for numerical methods for modeling and design, but also in computational hardware with its ever-increasing level of parallelism. These changes have favored methods that are robustly stable, geometrically flexible and suitable to implement on parallel computers. The discontinuous Galerkin method possesses all these qualities and has become popular among practitioners, for example in computational electromagnetics where it is gradually replacing the Yee - FDTD scheme. Thus it is essential to overcome numerical stiffness introduced by the high order piecewise polynomial approximations in the discontinuous Galerkin framework.

In 2015, Appelö and Hagstrom [3] firstly used an energy-based DG method to solve acoustic wave equations in second-order form. The method features direct, mesh-independent
inter-element fluxes and allows both energy conserving and energy dissipating discretizations. They further presented a sub-optimal convergence rate in the energy norm for a general mesh and numerically observed an optimal convergence rate in $L^2$ norm with Cartesian grids. Later, the methods were extended to solve the elastic wave equation in second-order form [4], and the coupled elasto-acoustic wave equation in second-order form [7]. In this thesis, we extend the energy-based DG methods beyond the original formulation in [3], especially for advective wave equations and semi-linear wave equations; all methods in this thesis possess favorable properties: minimal extra auxiliary fields and stability determined by simple numerical fluxes. We also propose two potential ways to overcome the numerical stiffness associated with high order polynomial approximations in the energy-based discontinuous Galerkin methods: the staggered formulation and special basis functions. The special basis functions proposed in this thesis also help reduce the computational cost of inversion of stiffness matrix from $O(N^2)$ to $O(N)$ with $N$ to be the degrees of the freedom. The thesis is organized as follows.

In Chapter 2, sharper error estimates for the original scheme in [3] are presented. The generalization of the energy-based DG methods for wave equations in a second-order form with advection is shown in Chapter 3. In Chapter 4, we present an application of energy-based DG methods to semi-linear wave equations in second-order form. A staggered formulation of energy-based DG methods which improves the efficiency of the method on regular meshes is shown in Chapter 5. In Chapter 6, the energy-based DG method in a difference-based formulation is presented. A conclusion is drawn in Chapter 7.
Chapter 2

CONVERGENCE ANALYSIS

The content in this chapter has been published in SIAM J. Numer. Anal., 57(1), 2019, pp. 238-265 under the name "Convergence analysis of a discontinuous Galerkin method for wave equations in second-order form" [39].

In this chapter, the convergence properties of the original energy-based DG formulation for wave equations in second-order form [3] is investigated. We prove an optimal convergence in the energy-norm and also obtain supercloseness results between the finite element and the interpolation solutions.

2.1. Introduction

In 2015, Appelö and Hagstrom [3] proposed a new DG method, energy-based DG method, to solve the scalar wave equations in second-order form. In their work, they proved a sub-optimal convergence in energy-norm and observed an optimal convergence in $L^2$ norm without constructing a special projection of the initial data. The optimal convergence in the energy norm is proven in one space dimension. Compared with other DG methods, the method is either energy-conserving or energy-dissipating depending only on simple numerical fluxes; also it minimizes the auxiliary fields that need to be solved since it only introduces one extra velocity field no matter the dimensions of the problem. Finally, the DG discretization they proposed arises naturally from a general formulation based directly on the Lagrangian, which is central to the formulation of wave equations in most physical settings.

To obtain a supercloseness result, we have to overcome the difficulty which is from the gradient of the interpolation of the velocity field minus the DG solution of the velocity, that is $\nabla(I_h v - v_h)$ with $I_h$ to be the interpolation operator. By defining a special elliptic projection operator for displacement field $u$ and combining with Galerkin orthogonality, the
The term $\nabla (I_h v - v_h)$ is eliminated. In the analysis of convergence, we add a penalty term to penalize the term which contains the jump of the discrete solution of $u$ at mesh interfaces. This technique is usually used to improve the convergence and stabilization of the methods [38, 74]. Note that the new formulation after adding penalty term is energy-conserving and the mass matrix in the discrete version of the scheme is block diagonal in this work.

The outline of this chapter is as follows. In Section 2.2 we recall the energy-based DG method in [3]. The supercloseness of the DG method on Cartesian meshes and optimal convergence of the DG method on a general mesh is given in Section 2.3. The supercloseness property of the DG method on quadrilateral meshes is proved in Section 2.4. Section 2.5 presents numerical examples to show the theoretical convergence rate and supercloseness result.

2.2. General formulation

In this section, we recall the energy-based DG discretization for the scalar wave equation in second-order form (2.1) which is proposed in [3],

$$\frac{\partial^2 u}{\partial t^2} - \nabla \cdot A \nabla u = f.$$  \hspace{1cm} (2.1)

And it is equivalent to the first-order in time system as follows

$$\frac{\partial u}{\partial t} - v = 0, \quad \frac{\partial v}{\partial t} - \nabla \cdot A \nabla u = f.$$  \hspace{1cm} (2.2)

To state the method, we firstly introduce some notations. Let $\mathcal{T}_h$ be a quasi-uniform partition of $\Omega$ (cf. [63–65]), $\forall K \in \mathcal{T}_h$, it is an image of the reference element $[-1, 1] \times [-1, 1]$. Denote $\mathcal{E}_h$ to be the set of all edges of $\mathcal{T}_h$. Further, define $h_K := \text{diam}(K), \forall K \in \mathcal{T}_h$. Let $e$ be the edge/face of element $K$, and define $h_e := \text{diam}(e)$. Denote $h = \max_{K \in \mathcal{T}_h} h_K$, the boundary edges by $\mathcal{E}_h^B := \{ e \in \mathcal{E}_h : e \subset \Gamma \}$ and the interior edges by $\mathcal{E}_h^I := \mathcal{E}_h \setminus \mathcal{E}_h^B$. Last, we adopt the standard space, norm and inner product notations and definitions as in [16, 24] and define the broken $H^1$-seminorm $|\cdot|_{H^1(\Omega)} = \left( \sum_{K \in \mathcal{T}_h} |\cdot|_{H^1(K)}^2 \right)^{\frac{1}{2}}$. 

6
Let $V_h$ be the piecewise $p$-th order polynomial approximation space, and $\mathcal{N}$ be the null space of $\frac{1}{2}A\nabla u \cdot \nabla u$. It is obvious that piecewise constant polynomials belong to $\mathcal{N}$. The energy-based DG method then reads as: $\forall \Phi = (\phi_u, \phi_v, \tilde{\phi}_u) \in V_h \times V_h \times \mathcal{N}$, find $U_h = (u_h, v_h) \in V_h \times V_h$ satisfying

\[
\int_K A\nabla \phi_u \cdot \nabla \left( \frac{\partial u_h}{\partial t} - v_h \right) = \int_{\partial K} n \cdot A\nabla \phi_u (v^* - v_h), \quad (2.3)
\]

\[
\int_K \phi_v \frac{\partial v_h}{\partial t} + A\nabla \phi_v \cdot \nabla u_h = \int_K \phi_v (n \cdot A\nabla u)^* + \int_K \phi_v f, \quad (2.4)
\]

\[
\int_K \tilde{\phi}_u \left( \frac{\partial u_h}{\partial t} - v_h \right) = 0. \quad (2.5)
\]

In a simpler form, we can write it as: find $U_h = (u_h, v_h) \in V_h \times V_h$ satisfying

\[
\mathcal{B}(\Phi, U_h) = \langle \phi_v, f \rangle \quad \forall \Phi = (\phi_u, \phi_v, \tilde{\phi}_u) \in V_h \times V_h \times \mathcal{N}. \quad (2.6)
\]

Here, $\mathcal{B}(\cdot, \cdot)$ is a bilinear form satisfies

\[
\mathcal{B}(\Phi, U^h) = \sum_{K \in T_h} \int_K \left( A\nabla \phi_u \cdot \nabla + \tilde{\phi}_u \right) \left( \frac{\partial u_h}{\partial t} - v_h \right)
+ \sum_{K \in T_h} \int_K \phi_v \frac{\partial v_h}{\partial t} + A\nabla \phi_v \cdot \nabla u_h
- \sum_{K \in T_h} \int_{\partial K} n \cdot A\nabla \phi_u (v^* - v_h) + \phi_v (n \cdot A\nabla u)^*,
\]

where $(\cdot)^*$ is the numerical flux operator which is defined later and $n$ denotes the unit outward normal to $\partial K$. Recall that the discrete energy associated with (2.2)

\[
E_h(t) = \sum_{K \in T_h} \frac{1}{2} \int_K |v_h|^2 + \frac{1}{2} A\nabla u_h \cdot \nabla u_h \quad (2.7)
\]

satisfies [3, Theorem 1]

\[
\frac{\partial E_h(t)}{\partial t} = \sum_{K \in T_h} \int_K u_h f(x, t) + \int_{\partial K} n \cdot A\nabla u_h (v^* - v_h) + v_h (n \cdot A\nabla u)^*. \quad (2.8)
\]
2.2.1. Fluxes

Now, let’s introduce the numerical fluxes $v^*$ and $(\mathbf{n} \cdot A\nabla u)^*$ for both inter-element and physical boundaries. Denote superscripts “±” to be the data from outside and inside of the element, respectively. And introduce the common notations

$$
\{v_h\} = \frac{1}{2}(v_h^+ + v_h^-), \quad [v_h] = v_h^- - v_h^+,
$$

$$
\{\mathbf{n} \cdot A\nabla u_h\} = \frac{1}{2}\left(\mathbf{n} \cdot A\nabla u_h^- + \mathbf{n} \cdot A\nabla u_h^+\right),
$$

$$
[\mathbf{n} \cdot A\nabla u_h] = \mathbf{n} \cdot A\nabla u_h^- - \mathbf{n} \cdot A\nabla u_h^+.
$$

Then a general form of the numerical fluxes on the inter-element boundaries reads as

$$
v^* = (\theta v_h^+ + (1 - \theta)v_h^-) - \tau [\mathbf{n} \cdot A\nabla u_h], \\
(\mathbf{n} \cdot A\nabla u)^* = -\beta [v_h] + \left(\theta \mathbf{n} \cdot A\nabla u_h^- + (1 - \theta)\mathbf{n} \cdot A\nabla u_h^+\right),
$$

where $\theta \in [0, 1]$, $\beta, \tau \geq 0$. Specifically, $\beta = \tau = 0$ corresponds to an energy-conserving scheme, and $\beta \neq 0$ or $\tau \neq 0$ corresponds to an energy-dissipating scheme:

alternating flux : $\theta = 0$, $\beta = \tau = 0$; $\theta = 1$, $\beta = \tau = 0$,

central flux : $\theta = \frac{1}{2}$, $\beta = \tau = 0$;

Sommerfeld flux : $\theta = \frac{1}{2}$, $\beta = \frac{1}{2\xi}$, $\tau = \frac{\xi}{2}$,

where $\xi$ is a flux splitting parameter with the same units as the elements in $A$.

On the other hand, assume that the physical boundary condition is given by

$$
a(x) \frac{\partial u}{\partial t} + b(x)\mathbf{n} \cdot A\nabla u = 0,
$$

where $a^2 + b^2 = 1, a, b \geq 0$. It is clear that $a(x) = 0, b(x) \neq 0$ corresponds to a homogeneous Neumann boundary condition and $a(x) \neq 0, b(x) = 0$ corresponds to a homogeneous Dirichlet
boundary condition. Further, the numerical fluxes at physical boundaries are given by

\[
v^* = v_h - (a - \eta b)\rho, \quad (2.14)
\]

\[
(n \cdot A \nabla u)^* = n \cdot A \nabla u_h - (b + \eta a)\rho, \quad (2.15)
\]

where \(\rho = a(x)v_h + b(x)n \cdot A \nabla u_h\) and \(\eta\) is a penalty parameter to be specified later based on different cases.

For the rest of the section, to obtain an optimal convergence and a supercloseness property, we set

\[
\theta = \frac{1}{2}, \quad \tau = 0, \quad \beta = h^{-(1+\varepsilon)} \quad (2.16)
\]

with \(\varepsilon \geq 0\) for numerical fluxes (2.9)–(2.10) at inter-element boundaries, and

\[
\eta = \begin{cases} 
-h^{1+\varepsilon}, & a = 0, \ b = 1, \\
h^{-(1+\varepsilon)}, & a = 1, \ b = 0, \\
\frac{a}{b}, & a, b > 0
\end{cases} \quad (2.17)
\]

with \(\varepsilon \geq 0\) for the numerical fluxes (2.14)–(2.15) at the physical boundaries.

### 2.3. Cartesian meshes

In this section, we assume that \(\mathcal{T}_h\) is a Cartesian mesh and its elements \(K\) are regular rectangles. \(A\) in (2.1) is set to be a diagonal matrix \(A = (c^2(x), 0; 0, c^2(x))\) with \(c(x) \in W^{1,\infty}(\Omega)\) and \(c_0 \leq c(x) \leq c_\infty\), both \(c_0\) and \(c_\infty\) are positive constants, i.e, the wave equation is given by

\[
\frac{\partial^2 u}{\partial t^2} - \nabla \cdot (c^2(x)\nabla u) = f. \quad (2.18)
\]
2.3.1. Problem statement and elliptic projection

Define the approximation space \( V_h \) as

\[
V_h := \{ v_h \in L^2(\Omega) : v_h|_K \in Q_k(K), \forall K \in \mathcal{T}_h \},
\]

where \( Q_k(K) \) is the set of polynomials of degree up to \( k \) in each variable on \( K \). For simplicity of analysis, we rewrite the formulation (2.6) as: find \((u_h, v_h) \in V_h \times V_h \) such that

\[
\forall (\phi_u, \phi_v, \tilde{\phi}_u) \in V_h \times V_h \times N
\]

\[
a_1(u_h, v_h; \phi_u, \tilde{\phi}_u) = 0,
\]

\[
a_2(u_h, v_h; \phi_v) = (f, \phi_v),
\]

where

\[
a_1(u_h, v_h; \phi_u, \tilde{\phi}_u) := \sum_K \left( c^2 \nabla \frac{\partial u_h}{\partial t} \cdot \nabla \phi_u \right)_K - (c^2 v_h, \nabla \phi_u)_K
\]

\[
- \left( c^2 (v^* - v_h), \frac{\partial \phi_u}{\partial \mathbf{n}} \right)_{\partial K} + \left( \frac{\partial u_h}{\partial t} - v_h, \tilde{\phi}_u \right)_K,
\]

\[
a_2(u_h, v_h; \phi_v) := \sum_K \left( \frac{\partial v_h}{\partial t}, \phi_v \right)_K + (c^2 \nabla u_h, \nabla \phi_v)_K - \left( c^2 \frac{\partial u}{\partial \mathbf{n}}^*, \phi_v \right)_{\partial K}.
\]

We denote by \( a_1(u_h, v_h; \phi_u) = a_1(u_h, v_h; \phi_u, 0) \). In addition, for (2.18), the discrete energy (2.7) has the following form

\[
E_h = \frac{1}{2} \left( \sum_K \| c(x) \nabla u_h \|_{L^2(K)}^2 + \| v_h \|_{L^2(K)}^2 \right).
\]

Thus, from (2.8), we have the following identity

\[
\frac{\partial E_h}{\partial t} = \sum_K (f, v_h)_K - \sum_{e \in \mathcal{E}_h^I} \beta \| [v_h] \|_{L^2(e)}^2
\]

\[
- \sum_{e \in \mathcal{E}_h^B} \left[ ab \left( \| v^* \|_{L^2(e)}^2 + \left( c^2 \frac{\partial u}{\partial \mathbf{n}} \right)^* \right)_e^2 \right] + \lambda \left| a v_h + b c^2 \frac{\partial u_h}{\partial \mathbf{n}} \right|_{L^2(e)}^2,
\]

where \( \lambda = (1 - \eta^2)ab + \eta(a^2 - b^2) \). It is obvious to see that \( \frac{\partial E_h}{\partial t} \leq 0 \) if \( f = 0 \) and \( \beta, \lambda \geq 0 \).
In addition, we define an elliptic projection \( Q_h : H^1(\Omega) \rightarrow V_h \) satisfies \( \forall \phi_h \in V_h, \)
\[
\sum_{K \in T_h} (c^2 \nabla (\psi - Q_h \psi), \nabla \phi_h)_K - \left\langle c^2 \left( \frac{\partial(\psi - Q_h \psi)}{\partial n} \right)^{**}, \phi_h \right\rangle_{\partial K} = 0, \tag{2.22}
\]
where the flux operator \( (\cdot)^{**} \) is defined as
\[
\left( \frac{\partial(\psi - Q_h \psi)}{\partial n} \right)^{**} = \begin{cases} 
\left\{ \frac{\partial(\psi - Q_h \psi)}{\partial n} \right\}_e - \bar{\beta} [\psi - Q_h \psi]_e, & e \in E^I_h, \\
0, & e \in E^B_h.
\end{cases} \tag{2.23}
\]
Here, \( \bar{\beta} = \sigma h^{-(1+\varepsilon)} \) with \( \sigma \) to be a positive constant and satisfy some conditions in Lemma 2.2 and \( Q_h \) is independent of the time \( t \).

2.3.2. Super closeness and optimal convergence analysis

Next, we present the super closeness property between the DG solution \((u_h,v_h)\) and the Lagrange interpolation \((I_h u, I_h v)\) which is based on Gauss-Lobatto points [93] and belongs to \( V_h \). For the following analysis, \( C \) is a positive constant independent of \( h, u, v \) and flux parameters \( \theta, \tau, \beta, \eta \). Denote \( e_u = u - u_h, \ e_v = v - v_h \) and
\[
\xi = I_h u - u_h, \ \xi_v = I_h v - v_h, \tag{2.24}
\]
\[
\delta = u - I_h u, \ \delta_v = u - I_h v. \tag{2.25}
\]
It is easy to see that \( e_u = \xi + \delta_u \) and \( e_v = \xi_v + \delta_v \). By a similar analysis as (2.21), we have
\[
a_1(\xi, \xi_v; \xi_u) + a_2(\xi, \xi_v; \xi_v) := \frac{\partial F(t)}{\partial t} + J(t),
\]
where
\[
F(t) = \frac{1}{2} \left( \sum_K \|c \nabla \xi_u\|^2_{L^2(K)} + \|\xi_v\|^2_{L^2(K)} \right),
\]
\[
J(t) = \sum_{e \in E^I_h} \beta \|\xi\|^2_{L^2(e)} + \sum_{e \in E^B_h} \left[ ab \left( \|\xi\|^2_{L^2(e)} + \left\| (c^2 \frac{\partial \xi}{\partial n})^* \right\|^2_{L^2(e)} \right) \right. \\
+ \left. \lambda \left\| a\xi + bc^2 \frac{\partial \xi}{\partial n} \right\|^2_{L^2(e)} \right). \tag{2.26}
\]
Through the Galerkin orthogonality, we obtain

\[ a_1(\xi_u, \xi_v; \phi_h) = -a_1(\delta_u, \delta_v; \phi_h), \quad a_2(\xi_u, \xi_v; \psi_h) = -a_2(\delta_u, \delta_v; \psi_h), \]

Therefore, we conclude that

\[
\frac{\partial F(t)}{\partial t} + J(t) = \sum_{K \in T_h} \left[ - \left( c^2 \nabla \frac{\partial \delta_u}{\partial t}, \nabla \xi_u \right)_K + \left( c^2 \nabla \delta_v, \nabla \xi_u \right)_K \right.
\]

\[
- \left( \frac{\partial \delta_v}{\partial t}, \xi_v \right)_K - \left( c^2 \nabla \delta_u, \nabla \xi_v \right)_K \right]
\]

\[
+ \sum_{K \in T_h} \left[ \left( c^2((\delta_u)^* - \delta_u), \frac{\partial \xi_u}{\partial n} \right)_K + \left( c^2 \frac{\partial \delta_u}{\partial n}, \xi_v \right) \right] \partial K \right].
\]  \tag{2.27}

Then by the first equation in (2.19) and the definition of \( a_1 \) in (2.20), we have

\[
- \sum_{K \in T_h} \left( c^2 \nabla \delta_u, \nabla \xi_v \right)_K
\]

\[
= a_1(\xi_u, \xi_v; \delta_u) - \sum_{K \in T_h} \left[ \left( c^2 \nabla \delta_u, \nabla \frac{\partial \xi_u}{\partial t} \right)_K - \left( c^2 \frac{\partial \delta_u}{\partial n}, (\xi_v)^* - \xi_v \right) \right] \partial K \]

\[
= a_1(\xi_u, \xi_v; u - Q_h u) + a_1(\delta_u, \delta_v; I_h u - Q_h u)
\]

\[
- \sum_{K \in T_h} \left[ \left( c^2 \nabla \delta_u, \nabla \frac{\partial \xi_u}{\partial t} \right)_K - \left( c^2 \frac{\partial \delta_u}{\partial n}, (\xi_v)^* - \xi_v \right) \right] \partial K \]

\[
= \sum_{K \in T_h} \left[ \left( c^2 \nabla \delta_u, \nabla (u - Q_h u) \right)_K - \left( c^2 \frac{\partial (u - Q_h u)^*}{\partial n}, \xi_v \right) \right] \partial K
\]

\[
- \left( c^2((\xi_v)^* - \xi_v), \frac{\partial (u - Q_h u)}{\partial n} \right) \partial K \right] + a_1(\delta_u, \delta_v; I_h u - Q_h u)
\]

\[
- \sum_{K \in T_h} \left[ \left( c^2 \nabla \delta_u, \nabla \frac{\partial \xi_u}{\partial t} \right)_K - \left( c^2 \frac{\partial \delta_u}{\partial n}, (\xi_v)^* - \xi_v \right) \right] \partial K \right].
\]

Substitute this into (2.27), we generate

\[
\frac{\partial F(t)}{\partial t} + J(t) = I_1 + I_2, \tag{2.28}
\]

12
\[ I_1 := \sum_{K \in \mathcal{T}_h} \left[ -\frac{\partial}{\partial t} (c^2 \nabla \delta u, \nabla \xi u)_K + (c^2 \nabla \delta v, \nabla \xi u)_K - (\frac{\partial \delta v}{\partial t}, \xi v)_K \right. \\
\left. + (c^2 \nabla \frac{\partial \xi u}{\partial t}, \nabla (u - Q_h u))_K + (c^2 \nabla \frac{\partial \delta v}{\partial t}, \nabla (I_h u - Q_h u))_K \\
- (c^2 \nabla \delta v, \nabla (I_h u - Q_h u))_K \right], \tag{2.29} \]

\[ I_2 := \sum_{K \in \mathcal{T}_h} \left[ \left\langle c^2 (\delta v)^* - \delta v, \frac{\partial \xi u}{\partial n} \right\rangle_\partial K + \left\langle c^2 \frac{\partial \delta u}{\partial n}, \xi v \right\rangle_\partial K \\
- \left\langle c^2 (\frac{\partial (u - Q_h u)}{\partial n})^*, \xi v \right\rangle_\partial K \right. \\
- \left\langle c^2 (\xi v)^* - \xi v, \frac{\partial (I_h u - Q_h u)}{\partial n} \right\rangle_\partial K + \left\langle c^2 \frac{\partial \delta u}{\partial n}, (\xi v)^* - \xi v \right\rangle_\partial K \right]. \tag{2.30} \]

We want to mention that the fundamental equality (2.28) is essential to the rest of the analysis.

**Lemma 2.1.** For any \( K \in \mathcal{T}_h \) and \( \psi \in H^{k+2}(K) \), there holds

\[ (c^2 \nabla (\psi - I_h \psi), \nabla \phi h)_K \leq C h^{k+1} |\psi|_{H^{k+2}(K)} |\phi h|_{H^1(K)} \quad \forall \phi h \in V_h. \]

The proof is given in [93].

**Lemma 2.2.** Let \( \bar{\varepsilon} = \min(1, \varepsilon/2) \). Set the parameter \( \bar{\beta} \) in (2.23) to be \( \bar{\beta} = \sigma h^{-(1+\bar{\varepsilon})} \). Then, there exists a positive constant \( \sigma \) such that if \( \sigma \leq \sigma \), there holds for all \( \psi \in H^{k+2}(\Omega) \)

\[ h^{-1} |\psi - Q_h \psi|_{L^2(\Omega)} + |\psi - Q_h \psi|_{H^{1}(\Omega)} + J^{1/2}(\psi) \leq C h^{k} |\psi|_{H^{k+1}(\Omega)}, \tag{2.31} \]

\[ |I_h \psi - Q_h \psi|_{H^{1}(\Omega)} + J^{1/2}(\psi) \leq C h^{k+\bar{\varepsilon}} |\psi|_{H^{k+2}(\Omega)}, \tag{2.32} \]

where \( J(\psi) = \sum_{e \in \mathcal{E}_h} \bar{\beta} \|[Q_h \psi]\|_{L^2(e)}^2 \)

**Proof.** The first inequality (2.31) can be proved by the same arguments as those in [74, Subsection 2.8]. We refer [93, Theorem 3.2] for the proof of the second inequality (2.32). \( \square \)

Furthermore, we assume that the elliptic projection (2.22) satisfies the conditions in Lemma 2.2 for the rest of this section.

**Theorem 2.3.** Let \( \bar{\varepsilon} = \min(1, \varepsilon/2) \). Assume the parameters \( \theta, \tau, \beta \) and \( \eta \) for fluxes (2.9)-(2.10) and (2.14)-(2.15) are defined by (2.16)-(2.17). Then for solutions \( u \in L^\infty(0,T;H^{k+2}(\Omega)) \),
where $L$ and (2.15), we get

$$v \in L^\infty(0, T; H^{k+2}(\Omega)) \text{ and } v_t \in L^\infty(0, T; H^{k+1}(\Omega)) \text{ we have}$$

$$\mathcal{F}(T) + \int_0^T J(t) dt \leq C(T + 1)(\mathcal{F}(0) - \mathcal{L}(0))$$

$$+ CT(T + 1)^2 h^{2k+2} \max_{t \leq T} \left( \|v\|^2_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|^2_{H^{k+1}(\Omega)} \right)$$

$$+ CT(T + 1) h^{2k+2} \max_{t \leq T} \left( \|u\|^2_{H^{k+2}(\Omega)} + \|v\|^2_{H^{k+2}(\Omega)} \right),$$

where $\mathcal{L}(t) = \sum_{K \in \mathcal{T}_h} (c^2 \nabla (I_h u - Q_h u), \nabla \xi_u)_K$.

**Proof.** Combine (2.29), Lemma 2.2 and Lemma 2.1, we get

$$I_1 \leq \sum_{K \in \mathcal{T}_h} -\frac{\partial}{\partial t} (c^2 \nabla \delta_u, \nabla \xi_u)_K + \frac{\partial}{\partial t} (c^2 \nabla \xi_u, \nabla (u - Q_h u))_K$$

$$+ Ch^{k+1} \|v\|_{H^{k+2}(\Omega)} \|\xi_u\|_{H^1(\Omega)} + Ch^{k+1} \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)} \|\xi_v\|_{L^2(\Omega)}$$

$$+ Ch^{k+\varepsilon} \|v\|_{H^{k+2}(\Omega)} \|\xi_u\|_{H^1(\Omega)} + Ch^{2k+1+\varepsilon} \|v\|_{H^{k+2}(\Omega)} \|u\|_{H^{k+2}(\Omega)}$$

$$\leq \sum_{K \in \mathcal{T}_h} -\frac{\partial}{\partial t} (c^2 \nabla \delta_u, \nabla \xi_u)_K + \frac{\partial}{\partial t} (c^2 \nabla \xi_u, \nabla (u - Q_h u))_K + Ch^{k+\varepsilon} \sqrt{\mathcal{F}(t)}$$

$$\cdot \left( \|v\|_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)} \right) + Ch^{2k+1+\varepsilon} \|u\|_{H^{k+2}(\Omega)} \|v\|_{H^{k+2}(\Omega)}.$$

As for $I_2$, there are three different cases based on different physical boundary conditions.

**Case 1.** Homogeneous Dirichlet boundary condition, $a = 1, b = 0$. In this case, by (2.14) and (2.15), we get $v^* = 0$ and $(c^2 \nabla u \cdot \mathbf{n})^* = c^2 \nabla u_h \cdot \mathbf{n} - \eta v_h$ with $\eta = h^{-(1+\varepsilon)}$ on the physical boundary. Then by (2.26), we obtain

$$J(t) = \sum_{e \in \mathcal{E}_h^I} \beta \|\xi_v\|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^B} \eta \|\xi_v\|_{L^2(e)}^2.$$ (2.34)

Since $\delta_v$ is continuous and vanishes on the physical boundaries, we have

$$\sum_{K \in \mathcal{T}_h} \left< c^2((\delta_v)^* - \delta_v), \frac{\partial \xi_u}{\partial \mathbf{n}} \right>_{\partial K} = 0,$$

$$\sum_{K \in \mathcal{T}_h} \left< c^2((\delta_v)^* - \delta_v), \frac{\partial (I_h u - Q_h u)}{\partial \mathbf{n}} \right>_{\partial K} = 0.$$ (2.35)

14
and
\[
\sum_{K \in T_h} \left< c^2 \left( \frac{\partial (u - Q_h u) *}{\partial n} \right)^*, \xi_v \right>_{\partial K} = \sum_{e \in E_h^l} \left( \left< c^2 \left( \frac{\partial u}{\partial n} \right)^*, \xi_v \right> - \beta \left< \delta_v, \xi_v \right>_e \right) + \sum_{e \in E_h^B} \left( \left< c^2 \left( \frac{\partial u}{\partial n} \right)^*, \xi_v \right> - \eta \left< \delta_v, \xi_v \right>_e \right) \\
= \sum_{e \in E_h^l} \left< c^2 \left( \frac{\partial u}{\partial n} \right)^*, \xi_v \right> + \sum_{e \in E_h^B} \left< c^2 \left( \frac{\partial u}{\partial n} \right)^*, \xi_v \right>_e \\
\leq C \left( \sum_{e \in E_h^l} \beta^{-1} \left\| \frac{\partial u}{\partial n} \right\|^2_{L^2(e)} + \sum_{e \in E_h^B} \eta^{-1} \left\| \frac{\partial u}{\partial n} \right\|^2_{L^2(e)} \right)^{1/2} \\
\cdot \left( \sum_{e \in E_h^l} \beta \left\| \xi_v \right\|^2_{L^2(e)} + \sum_{e \in E_h^B} \eta \left\| \xi_v \right\|^2_{L^2(e)} \right)^{1/2} \\
\leq Ch^{k+\epsilon/2} \sqrt{J(t)} \left\| u \right\|_{H^{k+1}(\Omega)}.
\]
(2.14) and (2.15), we get \( \nu^* = \nu_h + \eta c^2 \frac{\partial u_h}{\partial n} \) with \( \eta = -h^{1+\epsilon} \) and \((c^2 \frac{\partial u}{\partial n})^* = 0\) on the physical boundaries. Then by (2.26), we obtain

\[
J(t) = \sum_{e \in \mathcal{E}_h^I} \beta \|\{\xi_v\}\|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^B} \zeta \left\| c^2 \frac{\partial \xi_u}{\partial n} \right\|_{L^2(e)}^2.
\] (2.40)

Since \( \delta v \) is continuous and \( \frac{\partial \delta u}{\partial n} \) vanishes on the physical boundaries, we have

\[
\sum_{K \in \mathcal{T}_h} \left\langle c^2 ((\delta v)^* - \delta v), \frac{\partial (I_h u - Q_h u)}{\partial n} \right\rangle_{\partial K} + \sum_{K \in \mathcal{T}_h} \left\langle (c^2 \frac{\partial \delta u}{\partial n})^*, \xi_v \right\rangle_{\partial K} = - \sum_{e \in \mathcal{E}_h^B} \zeta \left\langle c^2 \frac{\partial \delta u}{\partial n}, c^2 \frac{\partial (I_h u - Q_h u)}{\partial n} \right\rangle_e + \sum_{e \in \mathcal{E}_h^I} \left\langle \left\{ c^2 \frac{\partial \delta u}{\partial n} \right\}, [\xi_v] \right\rangle_e \leq Ch^{k+\epsilon/2} \sqrt{J(t)} |u|_{H^{k+1}(\Omega)},
\] (2.41)

and

\[
\sum_{K \in \mathcal{T}_h} \left\langle c^2 ((\delta v)^* - \delta v), \frac{\partial (I_h u - Q_h u)}{\partial n} \right\rangle_{\partial K} = - \sum_{e \in \mathcal{E}_h^B} \zeta \left\langle c^2 \frac{\partial \delta u}{\partial n}, c^2 \frac{\partial (I_h u - Q_h u)}{\partial n} \right\rangle_e \leq Ch^{2k+\epsilon/2+\epsilon} |u|_{H^{k+1}(\Omega)} \|u\|_{H^{k+2}(\Omega)} \leq Ch^{2k+\epsilon/2+\epsilon} \|u\|_{H^{k+2}(\Omega)}^2.
\] (2.42)

By Lemma 2.2, we derive in analogy to (2.37) and (2.38),

\[
\sum_{K \in \mathcal{T}_h} \left\langle c^2 \frac{\partial(u - Q_h u)}{\partial n}^{**}, \xi_v \right\rangle_{\partial K} + \sum_{K \in \mathcal{T}_h} \left\langle c^2 \frac{\partial \delta u}{\partial n} (\xi_v)^* - \xi_v \right\rangle_{\partial K} - \left\langle c^2 (\xi_v)^*, \frac{\partial (u - Q_h u)}{\partial n} \right\rangle_{\partial K} \leq Ch^{k+\epsilon} \sqrt{J(t)} \|u\|_{H^{k+2}(\Omega)}.
\] (2.43)

Plugging the inequalities (2.41)–(2.43) into (2.30) yields

\[
I_2 \leq Ch^{k+\epsilon} \sqrt{J(t)} \|u\|_{H^{k+2}(\Omega)} + Ch^{2k+\epsilon/2+\epsilon} \|u\|_{H^{k+2}(\Omega)}^2.
\] (2.44)

Case 3. Robin boundary conditions with \( a, b > 0 \). In this case, by (2.14) and (2.15), we have \( \nu^* = \nu_h \) and \((c^2 \frac{\partial u}{\partial n})^* = -\frac{a}{b} \nu_h \) on the physical boundaries. Then by (2.26), we get

\[
J(t) = \sum_{e \in \mathcal{E}_h^I} \beta \|\{\xi_v\}\|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^B} \frac{a}{b} \|\xi_v\|_{L^2(e)}^2.
\] (2.45)
Since \( \delta_v \) is continuous, we have
\[
\sum_{K \in \mathcal{T}_h} \langle c^2((\delta_v)^* - \delta_v), \frac{\partial \xi_u}{\partial \mathbf{n}} \rangle_{\partial K} + \sum_{K \in \mathcal{T}_h} \langle (c^2 \frac{\partial \delta_u}{\partial \mathbf{n}})^*, \xi_v \rangle_{\partial K} \\
= \sum_{e \in \mathcal{E}_h^I} \langle \{ c^2 \frac{\partial \delta_u}{\partial \mathbf{n}} \}, \{ \xi_v \} \rangle_e - \sum_{e \in \mathcal{E}_h^B} \frac{a}{b} \langle \delta_v, \xi_u \rangle_e \\
\leq C \sqrt{\mathcal{J}(t)} \left( h^{k+\varepsilon/2} |u|_{H^{k+1}(\Omega)} + h^{k+1} |v|_{H^{k+1}(\Gamma)} \right),
\]
and
\[
\sum_{K \in \mathcal{T}_h} \langle c^2((\delta_v)^* - \delta_v), \frac{\partial (I_h u - Q_h u)}{\partial \mathbf{n}} \rangle_{\partial K} = 0.
\] (2.47)

By Lemma 2.2, we derive in analogy to (2.37),
\[
\sum_{K \in \mathcal{T}_h} \langle c^2 \left( \frac{\partial (u - Q_h u)}{\partial \mathbf{n}} \right)^*, \xi_v \rangle_{\partial K} \leq C h^{k+\varepsilon} \sqrt{\mathcal{J}(t)} \| u \|_{H^{k+2}(\Omega)},
\] (2.48)
and
\[
\sum_{K \in \mathcal{T}_h} \left[ \langle c^2 \frac{\partial \delta_u}{\partial \mathbf{n}}, (\xi_v^*) - \xi_v \rangle_{\partial K} - \langle c^2 ((\xi_v)^*) - \xi_v), \frac{\partial (u - Q_h u)}{\partial \mathbf{n}} \rangle_{\partial K} \right] \\
= \sum_{e \in \mathcal{E}_h^I} \langle c^2 [\xi_v], \left\{ \frac{\partial (Q_h u - I_h u)}{\partial \mathbf{n}} \right\} \rangle_e \leq C h^{k+\varepsilon+\varepsilon/2} \sqrt{\mathcal{J}(t)} \| u \|_{H^{k+2}(\Omega)}.
\] (2.49)

Plugging the inequalities (2.46)–(2.49) into (2.30) yields
\[
I_2 \leq C \sqrt{\mathcal{J}(t)} \left( h^{k+\varepsilon} \| u \|_{H^{k+2}(\Omega)} + h^{k+1} |v|_{H^{k+1}(\Gamma)} \right),
\] (2.50)

By combining the inequalities (2.33), (2.39), (2.44) and (2.50), we have
\[
\frac{\partial \mathcal{F}(t)}{\partial t} + \mathcal{J}(t) \leq \frac{\partial \mathcal{L}}{\partial t}(t) + C h^{k+\varepsilon} \sqrt{\mathcal{J}(t)} \left( \| v \|_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)} \right) \\
+ C h^{2k+1+\varepsilon} \| u \|_{H^{k+2}(\Omega)} \| v \|_{H^{k+2}(\Omega)} \\
+ C \sqrt{\mathcal{J}(t)} \left( h^{k+\varepsilon} \| u \|_{H^{k+2}(\Omega)} + s(ab) h^{k+1} |v|_{H^{k+1}(\Gamma)} \right), \\
+ C h^{2k+\varepsilon/2+\varepsilon} \| u \|^2_{H^{k+2}(\Omega)} ,
\]
where \( s(0) = 0 \) and \( s(x) = 1 \) if \( x \neq 0 \), and \( \mathcal{L}(t) \) has the following definition
\[
\mathcal{L}(t) := \sum_{K \in \mathcal{T}_h} - (c^2 \nabla \delta_u, \nabla \xi_u)_K + (c^2 \nabla \xi_u, \nabla (u - Q_h u))_K.
\]
Integrating in time from 0 to $T$ gives

$$
\mathcal{F}(T) + \frac{1}{2} \int_0^T \mathcal{J}(t) \, dt \leq \mathcal{F}(0) + \mathcal{L}(T) - \mathcal{L}(0)
$$

$$
+ C T (T + 1) h^{2k + 2\varepsilon} \max_{t \leq T} \left( \|v\|^2_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|^2_{H^{k+1}(\Omega)} \right)
$$

$$
+ C T \max_{t \leq T} \left( h^{2k+1+\varepsilon} \|u\|_{H^{k+2}(\Omega)} + h^{2k+2\varepsilon} \|v\|^2_{H^{k+2}(\Omega)} + h^{2k+2\varepsilon} \|u\|^2_{H^{k+2}(\Omega)} \right)
$$

$$
+ h^{2k+2} \|v\|^2_{H^{k+1}(\Gamma)} + \frac{1}{2} \int_0^T t + 1 \mathcal{F}(t) \, dt.
$$

(2.51)

Based on Lemmas 2.1 and 2.2, we have

$$
\mathcal{L}(t) = \sum_{K \in T_h} \left( c^2 \nabla(I_h u - Q_h u, \nabla \xi)_{K} \right) \leq C h^{k+\varepsilon} \sqrt{\mathcal{F}(t)} \|u\|_{H^{k+2}(\Omega)}.
$$

(2.52)

Then combining (2.51) and (2.52) yields,

$$
\mathcal{F}(T) + \int_0^T \mathcal{J}(t) \, dt \leq 2 (\mathcal{F}(0) - \mathcal{L}(0))
$$

$$
+ C T (T + 1) h^{2k + 2\varepsilon} \max_{t \leq T} \left( \|v\|^2_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|^2_{H^{k+1}(\Omega)} \right)
$$

$$
+ C T h^{2k+2\varepsilon} \max_{t \leq T} \left( \|u\|^2_{H^{k+2}(\Omega)} + \|v\|^2_{H^{k+2}(\Omega)} \right)
$$

$$
+ \int_0^T \frac{1}{t + 1} \mathcal{F}(t) \, dt.
$$

Finally through the Grönwall inequality [15], we obtain

$$
\mathcal{F}(T) + \int_0^T \mathcal{J}(t) \, dt \leq C (T + 1) (\mathcal{F}(0) - \mathcal{L}(0))
$$

$$
+ C T (T + 1)^2 h^{2k + 2\varepsilon} \max_{t \leq T} \left( \|v\|^2_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|^2_{H^{k+1}(\Omega)} \right)
$$

$$
+ C T (T + 1) h^{2k+2\varepsilon} \max_{t \leq T} \left( \|u\|^2_{H^{k+2}(\Omega)} + \|v\|^2_{H^{k+2}(\Omega)} \right).
$$

Remark 2.1.

(a) Letting $(u_h, v_h) = (I_h u, I_h v)$, we can obtain the following error estimate

$$
\mathcal{F}(T) + \int_0^T \mathcal{J}(t) \, dt \leq C (T, u, v) h^{2(k+\varepsilon)}.
$$
(b) The Authors in [3] proved a suboptimal error estimate,

\[ \| v - v_h \|_{L^2(\Omega)}^2 + \| \nabla (u - u_h) \|_{L^2(\Omega)}^2 \leq C h^{2\sigma}, \sigma = \begin{cases} k - 1, & \beta, \tau, \lambda \geq 0, \\ k - \frac{1}{2}, & \beta, \tau, \lambda > 0, \end{cases} \]

where \( C \) is a constant which depends only on the polynomial degree \( k \), the shape-regularity of the mesh, and a smooth solution \( u \) at time \( T \).

(c) Since Lemma 2.2 holds in three dimensions (cf. [93]), we can obtain the same superclose property for three dimensions.

(d) On the shape-regular mesh \( T_h \), we have the following optimal error estimate for the interpolant of \( \psi \in H^{k+1}(\Omega) \),

\[ h^{-1} \| \delta \psi \|_{L^2(\Omega)} + \| \nabla \delta \psi \|_{L^2(\Omega)}^2 + \left( \sum_{K \in T_h} h \| \delta \psi \|_{H^{2}(\partial K)}^2 + h \left\| \frac{\partial \psi}{\partial n} \right\|_{L^2(\partial K)}^2 \right)^{1/2} \leq C h^k |\psi|_{H^{k+1}(\Omega)}. \]

Thus, Lemma 2.2 still holds when \( \varepsilon = 0, \bar{\varepsilon} = 0 \) on a regular and quasi-uniform mesh. Then even on regular and quasi-uniform quadrilateral and triangular meshes, we can derive optimal error estimates. By the same arguments as those in Theorem 2.3, we have the following corollary.

**Corollary 2.4.** Assume that the parameters are set to be the same as those in Theorem 2.3 and the solution \( (u, v_t) \) is in \( L^\infty(0, T; H^{k+1}(\Omega)) \times L^\infty(0, T; H^{k+1}(\Omega)) \), the partition \( T_h \) of \( \Omega \) is quasi-uniform and shape-regular and the approximation space \( V_h \) consists of piecewise polynomials of order \( k \). Let \( (u_h, v_h) \in V_h \) be the DG solution. Assume further (for simplicity) that \( (u_h, v_h) = (I_h u, I_h v) \) at \( t = 0 \). Then we have

\[ \| I_h v - v_h \|_{L^2(\Omega)}^2 + \| \nabla (I_h u - u_h) \|_{L^2(\Omega)}^2 \leq C(T + 1)^2 h^{2k} \max_{t \leq T} \left( \| u \|_{H^{k+1}(\Omega)}^2 + \| v \|_{H^{k+1}(\Omega)}^2 + \left\| \frac{\partial v}{\partial t} \right\|_{H^{k+1}(\Omega)}^2 \right). \]

We want to mention that combining Corollary 2.4 with (2.53) gives

\[ \| v - v_h \|_{L^2(\Omega)}^2 + \| \nabla (u - u_h) \|_{L^2(\Omega)}^2 \leq C(T + 1)^2 h^{2k} \max_{t \leq T} \left( \| u \|_{H^{k+1}(\Omega)}^2 + \| v \|_{H^{k+1}(\Omega)}^2 + \left\| \frac{\partial v}{\partial t} \right\|_{H^{k+1}(\Omega)}^2 \right). \]
which is an optimal error estimate.

2.4. Quadrilateral meshes

In this section, we study the supercloseness property for the bilinear element on quadrilateral meshes. Assume $\Omega$ is a polygonal domain and $\mathcal{T}_h$ contains regular and quasi-uniform quadrilateral elements $K$. Let $\hat{K}$ be the reference element $[-1,1] \times [-1,1]$ with four vertices $\hat{Z}_i, i = 1, 2, 3, 4$, counterclockwisely starting from $\hat{Z}_1$. Further, let $Z^K_i, i = 1, 2, 3, 4$ be the four vertices of $K \in \mathcal{T}_h$ and $\hat{e}_i, i = 1, 2, 3, 4$ be the four edges of reference element $\hat{K}$. Denote $F_K$ to be the unique bilinear mapping satisfying $F_K(\hat{K}) = K$ and $F_K(\hat{Z}_i) = Z^K_i$, and define the finite element approximation space

$$V_h := \left\{ v \in L^2(\Omega) : v \circ F_K \in Q_k(\hat{K}), \ K \in \mathcal{T}_h \right\}.$$

In addition, $A$ is assumed to be a 2-by-2 positive definite matrix whose elements are sufficiently smooth functions. Particularly, on each $K \in \mathcal{T}_h$, the following equality holds

$$\|A - A_0\|_{0,\infty,K} = O(h^K_\alpha), \quad (2.54)$$

where $A_0$ is a piecewise constant matrix-valued function. For example, $A_0$ can be set as $A_0 = \frac{1}{|K|} \int_K A dx$ on each $K \in \mathcal{T}_h$ for some constant $\alpha \geq 0$.

2.4.1. Elliptic projection and penalty term

To obtain the supercloseness property of DG solution $(u_h, v_h)$ and Lagrange interpolant $(I^h_u, I^h_v)$ on quadrilateral meshes, we introduce a new bilinear form $B_J(\cdot, \cdot)$,

$$B_J(\Phi, U^h) = B(\Phi, U^h) + J(u_h, \phi_u),$$

where $J(u_h, \phi_u)$ is a penalty term which has the following form

$$J(u_h, \phi_u) = \sum_{e \in \mathcal{E}^I_h} \gamma^I_e \langle [u_h], [\phi_u] \rangle_e + \sum_{e \in \mathcal{E}^B_h} \gamma^B_e \langle u_h, \phi_u \rangle_e. \quad (2.55)$$

Here, $\gamma^I_e, \gamma^B_e \geq 0$ are penalty parameters that are used to enhance the stability of the DG
scheme and hence improve the accuracy of the scheme. In particular, \( \forall e \in \mathcal{E}^B_h \) and \( \forall e \in \mathcal{E}^I_h \), we define
\[
\gamma^I_e = h^{-(1+\varepsilon)}, \quad \gamma^B_e = \begin{cases} h^{-(1+\varepsilon)}, & b(x) = 0, \text{i.e.,} \, a(x) = 1, \\ 0, & b(x) \neq 0. \end{cases}
\] (2.56)

Let \( U^h = (u_h, v_h) \in V_h \times V_h \) be the energy-conserving DG solution which satisfies
\[
\mathcal{B}_J(\Phi, U^h) = \langle \phi, f \rangle, \quad \forall \Phi \in V_h \times V_h \times \mathcal{N},
\] (2.57)
with the numerical fluxes in (2.9)–(2.15). And the corresponding discrete energy for the scheme (2.57) satisfies
\[
\frac{\partial E^h(t)}{\partial t} = \sum_{K \in T_h} \int_K v^*_h f(x, t) + \int_{\partial K} \mathbf{n} \cdot A \nabla u_h(v^* - v_h) + v_h(\mathbf{n} \cdot A \nabla u_h)^*_e - J(u_h, u_h).
\]

In this section, the interpolation points for Lagrange interpolation are not Gauss-Lobatto points as in Section 2.3. We use the interpolation in [59, 60]: let \( I_h \hat{\Phi} = I_h \hat{\phi} \circ F_K \) and \( \hat{\phi} = \phi \circ F_K \), which satisfy
\[
I_h \hat{\phi} I_h \hat{\phi}(\hat{Z}_i) = \hat{\phi}(\hat{Z}_i), \quad i = 1, 2, 3, 4,
\]
\[
\int_{\hat{e}_i} (I_h \hat{\phi} I_h \hat{\phi} - \hat{\phi}) \hat{w}_h ds = 0, \quad \forall \hat{w}_h \in P_{k-2}(\hat{e}_i), \quad i = 1, 2, 3, 4,
\]
\[
\int_K (I_h \hat{\phi} I_h \hat{\phi} - \hat{\phi}) \hat{w}_h d\xi d\eta = 0, \quad \forall \hat{w}_h \in Q_{k-2}(K).
\]

Further, \( \forall \phi_h \in V_h \), we define an elliptic projection \( Q_h : H^1(\Omega) \rightarrow V_h \) by
\[
\sum_{K \in T_h} (A \nabla (\phi - Q_h \phi), \nabla \phi_h)_K - \langle (\mathbf{n} \cdot A \nabla (\phi - Q_h \phi))^{**}, \phi_h \rangle_{\partial K} = 0,
\] (2.58)
where the flux \((\cdot)^{**}\) has different definitions for different physical boundary conditions. Specifically, when \( b > 0 \),
\[
(n \cdot A \nabla (\phi - Q_h \phi))^{**} = \begin{cases} \{n \cdot A \nabla (\phi - Q_h \phi)\}_e - \tilde{\beta} [\phi - Q_h \phi]_e, & e \in \mathcal{E}^I_h, \\ 0, & e \in \mathcal{E}^B_h. \end{cases}
\] (2.59)
and when $b = 0$, 

$$
(n \cdot A \nabla (\phi - Q_h \phi))_{**}^{e} = \begin{cases} 
(n \cdot A \nabla (\phi - Q_h \phi))_{e} - \tilde{\beta} [\phi - Q_h \phi]_{e}, & e \in \mathcal{E}_h^I, \\
(n \cdot A \nabla (\phi - Q_h \phi))_{e} - \tilde{\beta} (\phi - Q_h \phi), & e \in \mathcal{E}_h^B. 
\end{cases}
$$

(2.60)

Here, $\tilde{\beta} = \omega h^{-(1+\epsilon)}$ with $\omega$ is a positive constant and satisfy the condition in Lemma 2.8.

2.4.2. Supercloseness and optimal convergence analysis

In the following analysis, we use the same notations (2.24)–(2.25) as in Section 2.3, and $C$ is a positive constant independent of $u, v$, mesh size and flux parameters $\theta, \tau, \beta, \eta$. By a similar calculation for (2.27)–(2.30), we obtain

$$
\frac{\partial \mathcal{F}(t)}{\partial t} + \mathcal{G}(t) + \mathcal{J}(t) = I_1 + I_2 + I_3
$$

(2.61)

where

$$
\mathcal{F}(t) = \frac{1}{2} \left( \sum_K \| A^{1/2} \nabla \xi_u \|_{L^2(K)}^2 + \| \xi_v \|_{L^2(K)}^2 \right),
$$

$$
\mathcal{G}(t) = \sum_{e \in \mathcal{E}_h^I} \gamma_e^I \| [\xi_u] \|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^B} \gamma_e^B \| \xi_u \|_{L^2(e)}^2,
$$

$$
\mathcal{J}(t) = \sum_{e \in \mathcal{E}_h^I} \beta \| [\xi_v] \|_{L^2(e)}^2 + \sum_{e \in \mathcal{E}_h^B} \left[ ab \left( \| \xi_v \|_{L^2(e)}^2 + \left\| (n \cdot A \nabla \xi_u)^* \right\|_{L^2(e)}^2 \right) \right. + \lambda \| a \xi + b n \cdot A \nabla \xi_u \|_{L^2(e)}^2 \right].
$$
and

\[ I_1 := \sum_{K \in T_h} \left[ -\frac{\partial}{\partial t} (A \nabla \delta_u, \nabla \xi_u)_K + (A \nabla \delta_v, \nabla \xi_u)_K - (\frac{\partial \delta_v}{\partial t}, \xi_v)_K \right. \]

\[ + (A \nabla \frac{\partial \xi_u}{\partial t}, \nabla (u - Q_h u))_K + (A \nabla \frac{\partial \delta_u}{\partial t}, \nabla (I_h u - Q_h u))_K \]

\[ - (A \nabla \delta_v, \nabla (I_h u - Q_h u))_K \right], \]

\[ I_2 := \sum_{K \in T_h} \left[ \langle (\delta_v)^* - \delta_v, n \cdot A \nabla \xi_u \rangle_{\partial K} + \langle (n \cdot A \nabla \delta_u)^*, \xi_v \rangle_{\partial K} \right. \]

\[ - \langle (n \cdot A \nabla (u - Q_h u))^*, \xi_v \rangle_{\partial K} - \langle ((\xi_v)^* - \xi_v), n \cdot A \nabla (u - Q_h u) \rangle_{\partial K} \]

\[ - \langle ((\delta_v)^* - \delta_v), n \cdot A \nabla (I_h u - Q_h u) \rangle_{\partial K} + \langle n \cdot A \nabla \delta_u, (\xi_v)^* - \xi_v \rangle_{\partial K} \right], \]

\[ I_3 := -J(\xi_u, \delta_u). \]

We first define the partition \( T_h \) which will be used in the rest of analysis and a basic result for the interpolation on the partition \( T_h \).

**Definition 2.1.** The partition \( T_h \) is said to satisfy *Condition (\( \alpha \)) if there exists \( \alpha > 0 \) such that

(a) any \( K \in T_h \) satisfies the diagonal condition, that is, the distance between the two diagonal midpoints \( O_1 \) and \( O_2, |O_1O_2| \), is \( O(h_1 + \alpha) \);

(b) any two \( K_1, K_2 \) in \( T_h \) that share a common edge satisfy the *neighbouring condition*, for \( j = 1, 2, \)

\[ a_j^{K_i} = a_j^{K_2}(1 + O(h_{K_1}^\alpha + h_{K_2}^\alpha)), \quad b_j^{K_i} = b_j^{K_2}(1 + O(h_{K_1}^\alpha + h_{K_2}^\alpha)), \]  

(2.63)

where \( a_1^K = \frac{1}{3}(-x_1^K + x_2^K + x_3^K - x_4^K) \), \( a_2^K = \frac{1}{4}(-x_1^K - x_2^K + x_3^K + x_4^K) \) and \( b_1^K = \frac{1}{4}(-y_1^K + y_2^K + y_3^K - 4y_4^K) \), \( b_2^K = \frac{1}{4}(-y_1^K - y_2^K + y_3^K + y_4^K) \) for all \( K \in T_h \) with \( Z_i^K = (x_i^K, y_i^K) \), i = 1,2,3,4, be its four vertices.
Lemma 2.5. Assume that \( \mathcal{T}_h \) satisfy Condition (\( \alpha \)). Then for any \( \phi_h \in V_h \), we have

\[
\sum_{K \in \mathcal{T}_h} (A \nabla (\phi - I_h \phi), \nabla \phi_h)_K \leq C \left( h^{k+\tilde{a}} \|\phi\|_{H^{k+2}(\Omega)} |\phi_h|_{H^1(\Omega)} + h^k \gamma^{-1/2} \|\phi\|_{H^{k+2}(\Omega)} J(\phi_h, \phi_h)^{1/2} \right),
\]

where

\[
\tilde{a} = \begin{cases} 
\min(1/2, \alpha) & b(x) \neq 0 \\
\min(1, \alpha) & b(x) = 0,
\end{cases} \quad (2.64)
\]

and \( J(\cdot, \cdot) \) is defined in (2.57), \( \gamma \) is given in (2.56).

Proof. We first introduce notations for the unique bilinear mapping \( F_K \) (cf. [94]). For any \( K \in \mathcal{T}_h \), let \( Z^K = (x^K, y^K) \), \( i = 1, 2, 3, 4 \), be its four vertices. Then the mapping \( F_K \) is given by

\[
x = a_0^K + a_1^K \xi + a_2^K \eta + a_3^K \xi \eta, \quad y = b_0^K + b_1^K \xi + b_2^K \eta + b_3^K \xi \eta,
\]

with \( \xi \) and \( \eta \) are variables for the reference domain \([-1, 1] \times [-1, 1]\) and

\[
a_0^K = \frac{1}{4}(x_1^K + x_2^K + x_3^K + x_4^K), \quad b_0^K = \frac{1}{4}(y_1^K + y_2^K + y_3^K + y_4^K);
\]

\[
a_1^K = \frac{1}{4}(-x_1^K + x_2^K + x_3^K - x_4^K), \quad b_1^K = \frac{1}{4}(-y_1^K + y_2^K + y_3^K - y_4^K);
\]

\[
a_2^K = \frac{1}{4}(-x_1^K - x_2^K + x_3^K + x_4^K), \quad b_2^K = \frac{1}{4}(-y_1^K - y_2^K + y_3^K + y_4^K);
\]

\[
a_3^K = \frac{1}{4}(x_1^K - x_2^K + x_3^K - x_4^K), \quad b_3^K = \frac{1}{4}(y_1^K - y_2^K + y_3^K - y_4^K).
\]

Then, the Jacobi matrix of the mapping \( F_K \) has the following form

\[
(DF_K)(\xi, \eta) = \begin{pmatrix} a_1^K + a_3^K \eta & b_1^K + b_3^K \eta \\
 a_2^K + a_3^K \xi & b_2^K + b_3^K \xi \end{pmatrix},
\]

and the corresponding determinant is

\[
J_K = J_K(\xi, \eta) = J_0^K + J_1^K \xi + J_2^K \eta,
\]

where

\[
J_0^K = a_1^K b_2^K - a_2^K b_1^K, \quad J_1^K = a_1^K b_3^K - a_3^K b_1^K, \quad J_2^K = b_2^K a_3^K - a_2^K b_3^K.
\]

In addition, the inverse of the Jacobi matrix is given by

\[
(DF_K)^{-1} = X = X_0 + X_1.
\]
Here,

\[ X_0 = \begin{pmatrix} b_2^K & -b_1^K \\ -a_2^K & a_1^K \end{pmatrix}, \quad X_1 = \begin{pmatrix} b_3^K \\ -a_3^K \end{pmatrix} (\xi, -\eta). \]

For any function \( \phi \) on \( K \), there is a unique function \( \hat{\phi}(\xi, \eta) = \phi \circ F_K \). Let \( \hat{\nabla} \) be the gradient operator for the reference element \([-1, 1] \times [-1, 1]\). We obtain

\[ (A\nabla \phi, \nabla \psi)_K = \int_K (\nabla \phi)^T A \nabla \psi \, dx \, dy = \int_K \frac{1}{J_K} (\hat{\nabla} \hat{\phi})^T X^T A X (\hat{\nabla} \hat{\psi}) d\xi d\eta. \]

Then, define

\[ (A\nabla \phi, \nabla \psi)_0^K = \int_{K_0} \frac{1}{J_{K_0}} (\hat{\nabla} \hat{\phi})^T X_0^T A_0 X_0 (\hat{\nabla} \hat{\psi}) d\xi d\eta = \int_K (\hat{\nabla} \hat{\psi})^T B^K \hat{\nabla} \hat{\psi} d\xi d\eta. \]

For convenience, we set \( w = \phi - I_h \phi \) for the rest of proof.

**Lemma 2.6.** Assume that (2.54) is satisfied and \( K \) satisfies Condition (\( \alpha \)). Then there exists a constant \( C \) depending only on the shape-regularity of \( K \), such that

\[ \left| (A\nabla w, \nabla \phi_h)_K - (A\nabla w, \nabla \phi_h)_0^K \right| \leq C h^\alpha \| \nabla w \|_{L^2(K)} \| \nabla \phi_h \|_{L^2(K)}. \]

The proof for the above lemma can be found in [94, Lemma 3.1]. Therefore, we only need to estimate \( (A\nabla w, \nabla \phi_h)_0^K \) which contains the terms

\[ b_{11}^K \int_K \partial_\xi \hat{w} \partial_\xi \hat{\phi}_h, \quad b_{12}^K \int_K \partial_\xi \hat{w} \partial_\eta \hat{\phi}_h, \quad b_{21}^K \int_K \partial_\eta \hat{w} \partial_\xi \hat{\phi}_h, \quad b_{22}^K \int_K \partial_\eta \hat{w} \partial_\eta \hat{\phi}_h, \quad (2.65) \]

where

\[ b_{11}^K = \frac{A_0}{J_0} ((b_2^K)^2 + (a_2^K)^2), \quad b_{12}^K = \frac{A_0}{J_0} ((b_1^K)^2 + (a_1^K)^2), \quad b_{21}^K = b_{22}^K = -2 \frac{A_0}{J_0} (b_1^K b_2^K + a_1^K a_2^K). \]

From [60], we have the following lemma,

**Lemma 2.7.** Under the same assumption as in Lemma 2.6, there holds

\[ \left| \int_K \partial_\xi \hat{w} \partial_\xi \hat{\phi}_h \right| + \left| \int_K \partial_\eta \hat{w} \partial_\eta \hat{\phi}_h \right| \leq C h^{k+1} \| \phi \|_{H^{k+2}(K)} \| \nabla \phi_h \|_{L^2(K)}, \quad (2.66) \]
and
\[
\int_K \partial_\xi \hat{w} \partial_\eta \hat{\phi}_h = O(h^{k+1}) |\phi|_{H^{k+2}(K)} \|\nabla \phi_h\|_{L^2(K)} \\
+ (-1)^k (\int_{\hat{e}_1} + \int_{\hat{e}_3}) (\xi^2 - 1)^k \partial_\xi^{k+1} \hat{\phi}_h \cdot \partial_\xi^{k+1} \phi_h d\xi,
\]
\[
\int_K \partial_\eta \hat{w} \partial_\xi \hat{\phi}_h = O(h^{k+1}) |\phi|_{H^{k+2}(K)} \|\nabla \phi_h\|_{L^2(K)} \\
+ (-1)^k (\int_{\hat{e}_2} + \int_{\hat{e}_4}) (\eta^2 - 1)^k \partial_\eta^{k+1} \hat{\phi}_h \cdot \partial_\eta^{k+1} \phi_h d\eta.
\]

Now, we only need to estimate the second term and the third term in (2.65). Combine Lemma 2.7 and \(b_{12}^K = b_{21}^K\), we have
\[
b_{12}^K \int_K \partial_\xi \hat{w} \partial_\eta \hat{\phi}_h + b_{21}^K \int_K \partial_\eta \hat{w} \partial_\xi \hat{\phi}_h = O(h^{k+1}) |\phi|_{H^{k+2}(K)} \|\nabla \phi_h\|_{L^2(K)} \\
+ (-1)^k \sum_{j=1}^4 b_{12}^K \frac{|e_j|}{2} 2k \int_{e_j} \left( \frac{2s}{|e_j|} - 1 \right)^k \partial_\xi^{k+1} \phi_h \cdot \partial_\xi^{k+1} \phi_h ds,
\]
which gives
\[
\sum_{K \in \mathcal{T}_h} b_{12}^K \int_K \partial_\xi \hat{w} \partial_\eta \hat{\phi}_h + b_{21}^K \int_K \partial_\eta \hat{w} \partial_\xi \hat{\phi}_h = O(h^{k+1}) |\phi|_{H^{k+2}(K)} \|\nabla \phi_h\|_{L^2(\Omega)} \\
+ (-1)^k \sum_{e \in \mathcal{E}_h^I} \left( \frac{|e|}{2} \right)^{2k} \int_e \left( \frac{2s}{|e|} - 1 \right)^k \partial_\xi^{k+1} \phi_h \cdot \partial_\xi^{k+1} \phi_h ds \\
+ (-1)^k \sum_{e \in \mathcal{E}_h^B} b_{12}^K \frac{|e|}{2} 2k \int_e \left( \frac{2s}{|e|} - 1 \right)^k \partial_\xi^{k+1} \phi_h \cdot \partial_\xi^{k+1} \phi_h ds.
\]

Here, \(K_e\) is the quadrilateral element which contains \(e\). Let \(K_1\) and \(K_2\) be two adjacent elements sharing the common edge \(e\) and the jump of \(b_{12}^K \phi_h\) on \(e \in \mathcal{E}_h^I\) be \(b_{12}^K |\phi_h|_{e} = b_{12}^K |\phi_h|_{K_1} - b_{12}^K |\phi_h|_{K_2}\). Based on the neighboring condition (2.63), the matrix \(B^K\) satisfies
\[
\|B^{K_1} - B^{K_2}\| = O(h^\alpha).
\]
Then by the trace theory and the inverse inequality, we obtain

\[
\sum_{K \in \mathcal{T}_h} b_{12}^K \int_K \partial_k \hat{w} \partial_q \hat{\phi}_h + b_{21}^K \int_K \partial_q \hat{w} \partial_k \hat{\phi}_h = O(h^{k+1}) \|\phi\|_{H^{k+2}(\Omega)} \|\nabla \phi_h\|_{L^2(\Omega)} \\
+ (-1)^k \sum_{e \in \mathcal{E}_h^I} (b_{12}^I - b_{12}^B) \left( \frac{|e|}{2} \right)^{2k} \int_e \left( \frac{2s}{|e|} - 1 \right)^k \partial_s^{k+1} \phi \cdot \partial_s^k \phi_h |_{K_i} ds \\
+ (-1)^k \sum_{e \in \mathcal{E}_h^B} b_{12}^B \left( \frac{|e|}{2} \right)^{2k} \int_e \left( \frac{2s}{|e|} - 1 \right)^k \partial_s^{k+1} \phi \cdot \partial_s^k \phi_h ds \\
= O\left(h^{k+\min\left(1, \frac{1}{2}+\alpha\right)} \|\phi\|_{H^{k+2}(\Omega)} \|\nabla \phi_h\|_{L^2(\Omega)} \right) \\
+ O(h^k (\gamma^I)^{-1/2}) \|\phi\|_{H^{k+2}(\Omega)} \left( \sum_{e \in \mathcal{E}_h^I} \gamma_I e ||[\phi_h]||_{L^2(e)}^2 \right)^{1/2} \\
+ (-1)^k \sum_{e \in \mathcal{E}_h^B} b_{12}^B \left( \frac{|e|}{2} \right)^{2k} \int_e \left( \frac{2s}{|e|} - 1 \right)^k \partial_s^{k+1} \phi \cdot \partial_s^k \phi_h ds.
\]

Now, let’s consider different physical boundary conditions.
Case I $\gamma_e^B \neq 0$, we set $\gamma = \gamma_e^I = \gamma_e^B$ and get

\[
\sum_{K \in \mathcal{T}_h} b_{12}^K \int_K \partial_k \hat{w} \partial_q \hat{\phi}_h + b_{21}^K \int_K \partial_q \hat{w} \partial_k \hat{\phi}_h = O(h^{k+\min\left(1, \frac{1}{2}+\alpha\right)}) \|\phi\|_{H^{k+2}(\Omega)} \\
\cdot |\phi_h|_{H^1(\Omega)} + O(h^k \gamma^{-1/2}) \|\phi\|_{H^{k+2}(\Omega)} J(\phi_h, \phi_h)^{1/2}.
\]

Case II $\gamma_e^B = 0$, we set $\gamma = \gamma_e^I$ and get

\[
\sum_{K \in \mathcal{T}_h} b_{12}^K \int_K \partial_k \hat{w} \partial_q \hat{\phi}_h + b_{21}^K \int_K \partial_q \hat{w} \partial_k \hat{\phi}_h = O(h^{k+\min\left(1, \frac{1}{2}+\alpha\right)}) \|\phi\|_{H^{k+2}(\Omega)} \\
\cdot |\phi_h|_{H^1(\Omega)} + O(h^k \gamma^{-1/2}) \|\phi\|_{H^{k+2}(\Omega)} J(\phi_h, \phi_h)^{1/2}.
\]

Finally, combine Lemma 2.6 with (2.65)–(2.69), we obtain Lemma 2.5.

By a similar analysis for the derivation of Lemma 2.2, we have

**Lemma 2.8.** Let $\varepsilon = \min(1, \varepsilon/2)$. Assume that $\mathcal{T}_h$ satisfies Condition (α) and the elliptic projection $Q_h$ satisfies the conditions in (2.59)–(2.60). Then there exists a positive constant
such that $\forall \phi \in H^{k+2}(\Omega)$, if $\omega \leq \omega$, there holds

$$h^{-1} \|\phi - Q_h \phi\|_{L^2(\Omega)} + |\phi - Q_h \phi|_{H^1(\Omega)} + J_{Q}^{1/2}(\phi) \leq Ch^{k} \|\phi\|_{H^{k+1}(\Omega)};$$

$$|I_h \phi - Q_h \phi|_{H^1(\Omega)} + J_{Q}^{1/2}(\phi) \leq Ch^{k+\min(\alpha,\varepsilon)} \|\phi\|_{H^{k+2}(\Omega)},$$

where $J_{Q}(\phi) = \sum_{e \in e_h} \tilde{\beta} \|\phi\|_{L^2(e)}^2 + \bar{s}(b) \sum_{e \in e_h} \tilde{\beta} \|\phi\|_{L^2(e)}^2$ ($\bar{s}(0) = 1$ and $\bar{s}(b) = 0$ when $b \neq 0$) with $\tilde{\beta} = \omega h^{1+\varepsilon}$.

In the rest of analysis, we assume that the conditions in Lemma 2.8 are satisfied when the elliptic projection (2.58) is used. Then we have the following theorem.

**Theorem 2.9.** Let $\varepsilon = \min(1,\varepsilon/2)$ and $\tilde{\alpha}$ be define in (2.64). Assume that the parameters $\theta, \tau, \beta$ and $\eta$ for (2.9)–(2.10) and (2.14)–(2.15) are defined by (2.16)–(2.17), and the penalty parameter $\gamma_{\varepsilon}^I$ and $\gamma_{\varepsilon}^B$ in (2.55) are defined in (2.56). Then for smooth solutions $u \in L^\infty(0, T; H^{k+2}(\Omega)), v \in L^\infty(0, T; H^{k+2}(\Omega)), v_t \in L^\infty(0, T; H^{k+1}(\Omega))$, we have

$$\mathcal{F}(T) + \int_0^T \left( \mathcal{G}(t) + \mathcal{J}(t) \right) dt \leq C(T + 1)(\mathcal{F}(0) - \mathcal{L}(0))$$

$$+ CT(T + 1)^2 h^{2k+2\min(\tilde{\alpha}, \varepsilon)} \max_{t \leq T} \left( \|v\|_{H^{k+2}(\Omega)}^2 + \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)}^2 \right)$$

$$+ CT(T + 1) h^{2k+2\min(\tilde{\alpha}, \varepsilon)} \max_{t \leq T} \left( \|u\|_{H^{k+2}(\Omega)}^2 + \|v\|_{H^{k+2}(\Omega)}^2 \right),$$

where $\mathcal{L}(t) = \sum_{K \in T_h} (A \nabla (I_h u - Q_h u), \nabla \xi_u)_K$.

**Proof.** Based on (2.62) and Lemmas 2.5–2.8, we obtain

$$I_1 \leq \sum_{K \in T_h} -\frac{\partial}{\partial t} (A \nabla \xi_u, \nabla \xi_u)_K + \frac{\partial}{\partial t} (A \nabla \xi_u, \nabla (u - Q_h u))_K$$

$$+ Ch^{k+\tilde{\alpha}} \|v\|_{H^{k+2}(\Omega)} \|\xi_u\|_{H^{k+1}(\Omega)} + Ch^{k+\frac{1}{2}\varepsilon} \sqrt{\mathcal{G}(t)} \|v\|_{H^{k+2}(\Omega)}$$

$$+ Ch^{k+1} \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)} \|\xi_u\|_{L^2(\Omega)} + Ch^{k+\min(\tilde{\alpha}, \varepsilon)} \|v\|_{H^{k+2}(\Omega)} \|\xi_u\|_{H^1(\Omega)}$$

$$+ Ch^{2k+2\min(\tilde{\alpha}, \varepsilon)} \|v\|_{H^{k+2}(\Omega)} \|u\|_{H^{k+2}(\Omega)}$$

$$\leq \sum_{K \in T_h} -\frac{\partial}{\partial t} (A \nabla \delta_u, \nabla \xi_u)_K + \frac{\partial}{\partial t} (A \nabla \xi_u, \nabla (u - Q_h u))_K$$

$$+ Ch^{k+\min(\tilde{\alpha}, \varepsilon)} \sqrt{\mathcal{F}(t)} \left( \|v\|_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)} \right)$$

$$+ Ch^{k+\frac{1}{2}\varepsilon} \sqrt{\mathcal{G}(t)} \|v\|_{H^{k+2}(\Omega)} + Ch^{2k+2\min(\tilde{\alpha}, \varepsilon)} \|u\|_{H^{k+2}(\Omega)} \|v\|_{H^{k+2}(\Omega)}.$$
By the same analysis as in (2.34)–(2.50) for different cases, we get
\[ I_2 \leq C \sqrt{J(t) \left( h^{k+\min(\bar{\alpha}, \bar{\epsilon})} \left\| u \right\|_{H^{k+2}(\Omega)} + h^{k+1} \left| v \right|_{H^{k+1}(\Gamma)} \right) + h^{2k+2\min(\bar{\alpha}, \bar{\epsilon})} \left\| u \right\|^2_{H^{k+2}(\Omega)}}. \] (2.72)

When \( b = 0 \), we have \([\delta_u] = 0, e \in \mathcal{E}_h^I\) and \([\delta_u] = 0, e \in \mathcal{E}_h^B\), then it is clear that
\[ I_3 = -J(\xi_u, \delta_u) = 0. \] (2.73)

Combine (2.71)–(2.73) with (2.61), we obtain
\[
\frac{\partial F(t)}{\partial t} + G(t) + J(t) + \frac{\partial L}{\partial t} \leq \left( \frac{\partial F}{\partial t} \right)_{H^{k+1}(\Omega)} + \frac{\partial v}{\partial t} \left( \frac{\partial F}{\partial t} \right) \left( \frac{\partial F}{\partial t} \right)_{H^{k+1}(\Omega)} + C h^{k+\min(\bar{\alpha}, \bar{\epsilon})} \sqrt{F(t)} \left( \left\| u \right\|_{H^{k+2}(\Omega)} + \left\| v \right\|_{H^{k+2}(\Omega)} \right) + C h^{2k+2\min(\bar{\alpha}, \bar{\epsilon})} \left( \left\| u \right\|^2_{H^{k+2}(\Omega)} + \left\| v \right\|^2_{H^{k+2}(\Omega)} \right),
\]

where
\[ L(t) := \sum_{K \in \mathcal{T}_h} - (A \nabla \delta_u, \nabla \xi_u)_K + (A \nabla \xi_u, \nabla (u - Q_h u))_K. \]

Integration in time from 0 to \( T \) yields
\[ F(T) + \frac{1}{2} \int_0^T \left( G(t) + J(t) \right) dt \leq L(T) - L(0) + F(0) + CT(T + 1) h^{2k+2\min(\bar{\alpha}, \bar{\epsilon})} \max_{t \leq T} \left( \left\| u \right\|^2_{H^{k+2}(\Omega)} + \left| \frac{\partial v}{\partial t} \right|^2_{H^{k+1}(\Omega)} \right) + C T h^{2k+2\min(\bar{\alpha}, \bar{\epsilon})} \max_{t \leq T} \left( \left\| u \right\|^2_{H^{k+2}(\Omega)} + \left\| v \right\|^2_{H^{k+2}(\Omega)} \right) + \frac{1}{2} \int_0^T \frac{1}{1 + t} F(t) dt. \]

Finally, by combining
\[ L(t) = \sum_{K \in \mathcal{T}_h} (A \nabla (I_h u - Q_h u), \nabla \xi_u)_K \leq C h^{k+\min(\bar{\alpha}, \bar{\epsilon})} \sqrt{F(t)} \left\| u \right\|_{H^{k+2}(\Omega)} \]
and the integral form of the Grönwall inequality [15], the proof is completed.

\[ \square \]
Remark 3.1

(a) On each element \( K \), using one dimension as an example, we approximate \( u_h \) and \( v_h \) by

\[
\begin{align*}
  u_h &= \sum_{l=0}^{k} \hat{u}_l h^l, \\
  v_h &= \sum_{l=0}^{k} \hat{v}_l h^l,
\end{align*}
\]

respectively. Note that for two dimensions, the basis functions are the tensor product of basis functions in one dimension; but for three dimensions, the analysis in this section fails; actually, we don’t have a supercloseeness result for the problem in three dimensions. Let \( \hat{u} = [\hat{u}_0^h, \hat{u}_1^h, \cdots, \hat{u}_k^h]^T \) and \( \hat{v} = [\hat{v}_0^h, \hat{v}_1^h, \cdots, \hat{v}_k^h]^T \), then the discrete version of (2.57) can be written as

\[
M^u \frac{\partial \hat{u}_h}{\partial t} + S^v \hat{v}_h = F^u, \quad M^v \frac{\partial \hat{v}_h}{\partial t} + S^u \hat{u}_h = F^v. \tag{2.74}
\]

Here, the mass matrix \( M^u \) is a block-diagonal matrix. Therefore, the system (2.74) can be solved by explicit methods, such as explicit Runge–Kutta methods.

(b) Setting \((u_h, v_h) = (I_h u, I_h v)\) at \( t = 0 \), the following estimate can be derived from (2.70)

\[
\mathcal{F}(T) + \int_0^T \left( \mathcal{G}(t) + \mathcal{J}(t) \right) dt \leq C(T, u, v) h^{2k+2 \min(\tilde{\alpha}, \tilde{\varepsilon})}.
\]

(c) The optimal error estimate on shape-regular and quasi-uniform meshes for the scheme (2.57) can be derived by the same arguments as those in this section. Particularly, we have the following corollary which is similar to Corollary 2.4.

**Corollary 2.10.** The assumptions are the same as those in Corollary 2.4 except that \((u_h, v_h)\) is the numerical solution of the scheme (2.57). Then for smooth \( u \in L^\infty(0, T; H^{k+1}(\Omega)) \), \( v_t \in L^\infty(0, T; H^{k+1}(\Omega)) \) we have

\[
\|I_h v - v_h\|_{L^2(\Omega)}^2 + \|\nabla(I_h u - u_h)\|_{L^2(\Omega)}^2 \leq C T (T + 1)^2 h^{2k} \max_{t \leq T} (\|u\|_{H^{k+1}(\Omega)}^2 + \|v\|_{H^{k+1}(\Omega)}^2 + \left| \frac{\partial v}{\partial t} \right|_{H^{k+1}(\Omega)}^2).
\]

2.5. Numerical examples

To verify the convergence rates of the methods, we solve \( u_{tt} - \Delta u + u = f \) on \( \Omega = [-6, 6]^2 \) with homogeneous Dirichlet boundary conditions \( v|_{\partial \Omega} = 0 \) and the initial data chosen
such that the solution is \( u(x, y, t) = e^{-(x^2+y^2)} \sin(1 + t) \). Specifically, we show cases with different numerical fluxes and present results for both \( L^2 \)-norm and \( H^1 \)-seminorm. The domain \( \Omega \) is discretized with both Cartesian mesh and the quadrilateral mesh which satisfies the Condition \((\alpha)\). Finally, we evolve the solution until \( T = 0.1 \) with the classic 4-th order accurate Runge-Kutta method and \( \delta t = \frac{1}{500} \), where \( \delta t \) is the time step size. In addition, the degrees of the approximation space are set to be \( p = 1, 2, 3 \).

2.5.1. Verification on Cartesian meshes

We first consider regular Cartesian grids with elements whose edge sizes are \( h_x = h_y = h = 12/n \). Figure 2.1 shows the \( L^2 \) errors for \( v_h \) with \( p = 1, 2, 3 \) and \( \epsilon = 0, 0.5, 1 \). Both the \( H^1 \) errors for \( u_h \) and the \( H^1 \) errors between \( u_h \) and \( I_h u \) are shown in Figure 2.2 with \( p = 1, 2, 3 \) and \( \epsilon = 0, 0.5, 1 \). Least squares fits for the rates of convergence can be found in Table 2.1.

We observe that the convergence rates of \( \|v - v_h\|_{L^2(\Omega)} \) are \( p+1 \) for \( p = 1, 3 \) with different \( \epsilon \), while for \( p = 2 \) the convergence rates increase to \( p+1 \) as \( \epsilon \) becomes larger. For \( |u - u_h|_{H^1(\Omega)} \), the convergence rates are \( p \) for \( p = 1, 2, 3 \) and \( \epsilon = 0, 0.5, 1 \). For the supercloseness between \( I_h u \) and \( u_h \), the convergence rates are \( p+1 \) for \( p = 1, 3 \) as \( \epsilon \) becomes larger for \( p = 2 \).

2.5.2. Verification on quadrilateral meshes

| \( p \) | \( \epsilon \) | \( \|v - v_h\|_{L^2} \) | \( |u - u_h|_{H^1} \) | \( |I_h u - u_h|_{H^1} \) |
|-----|-------|-----------------|-----------------|-----------------|
| \( \epsilon \) | 0 0.5 1 | 0 0.5 1 | 0 0.5 1 |
| 1   | 2.00 1.97 1.97 | 0.99 0.99 0.99 | 1.97 2.00 1.99 |
| 2   | 2.29 2.77 2.97 | 1.98 1.98 1.98 | 2.45 3.03 3.46 |
| 3   | 4.02 4.13 4.12 | 2.98 2.98 2.97 | 4.10 4.27 3.97 |

Table 2.1: The convergence rates of various errors on Cartesian meshes for different \( p \) and \( \epsilon \).
Figure 2.1: The $L^2$-norm errors of $v_h$ on Cartesian meshes for different $p$ and $\epsilon$.

Now, we consider the quadrilateral grids obtained by perturbing the $x$ and $y$ coordinates of the interior nodes of the Cartesian grid by a uniformly distributed random perturbation $(\delta x, \delta y)$. For given $\alpha$ and the mesh size $h_x = h_y = h = 12/n$, the perturbation $(\delta x, \delta y)$ is set to be $\frac{1}{4}(h_x^{1+\alpha}, h_y^{1+\alpha})$. It can be verified that the quadrilateral grid satisfies the Condition $(\alpha)$. We evolve the solution until $T = 0.1$ for different $\alpha'$ and $\epsilon'$, and the degree of approximation space is $p = 1, 2, 3$.

We first compute the errors for fixed $\alpha = 2$ and different $\epsilon = 0, 0.5, 1$. We only show the $L^2$-norm errors of $v_h$, the broken $H^1$-norm errors of $u_h$ in Figure 2.3 since the errors are quite similar to each other for the same $p$ and $\alpha$. The least squares estimates of the convergence rates can be found in Table 2.2. We observe that the rates of $\|v - v_h\|_{L^2(\Omega)}$, and
Figure 2.2: The broken $H^1$-seminorm errors of $u_h$ and the errors between $I_h u$ and $u_h$ on Cartesian meshes for different $p$ and $\epsilon$. 
$|I_h u - u_h|_{H^1(\Omega)}$ increase as $\epsilon$ becomes larger for fixed $p$, while the rates of $|u - u_h|_{H^1(\Omega)}$ are almost equal to $p$.

To investigate the order of accuracy for different $\alpha$, we compute the errors for the fixed $\epsilon = 1$ and the different $\alpha = 0, 0.5, 1$. Figure 2.4 displays the errors for $\epsilon = 1, \alpha = 0.5$ and $p = 1, 2, 3$, respectively. The reader can find the convergence determined by least squares for the different $\alpha$ and $p$ in Table 2.3. We see the optimal convergence rates $p$ of the broken $H^1$-seminorm errors of $u_h$, and the convergence rates increase as $\alpha$ increases for the errors of $v_h$ in $L^2$-norms, respectively.

Figure 2.3: The errors of numerical solutions on perturbed grids with $\alpha = 2$ for $\epsilon = 0.5$.  

34
Table 2.2: The convergence rates of various errors on perturbed grids with $\alpha = 2$ for different $p$ and $\epsilon$.

| $p$/rate | $\|v - v_h\|_{L^2}$ | $|u - u_h|_{H^1}$ | $|I_hu - u_h|_{H^1}$ |
|----------|----------------------|-------------------|----------------------|
| $\epsilon$ | 0 | 0.5 | 1 | 0 | 0.5 | 1 | 0 | 0.5 | 1 |
| 1 | 2.01 | 1.97 | 1.98 | 0.99 | 1.00 | 1.00 | 1.96 | 2.01 | 2.01 |
| 2 | 2.16 | 2.75 | 3.01 | 1.99 | 1.99 | 1.99 | 2.07 | 2.65 | 3.13 |
| 3 | 3.90 | 3.93 | 3.92 | 2.95 | 2.94 | 2.97 | 3.98 | 4.01 | 4.01 |

Figure 2.4: The errors of numerical solutions on perturbed grids with $\alpha = 0.5$ for $\epsilon = 1$.

Table 2.3: The convergence rates of various errors on perturbed grids with $\alpha = 0, 0.5, 1$ for fixed $\epsilon = 1$ and different $p$.

| $p$/rate | $\|v - v_h\|_{L^2}$ | $|u - u_h|_{H^1}$ | $|I_hu - u_h|_{H^1}$ |
|----------|----------------------|-------------------|----------------------|
| $\alpha$ | 0 | 0.5 | 1 | 0 | 0.5 | 1 | 0 | 0.5 | 1 |
| 1 | 1.96 | 1.99 | 1.98 | 0.99 | 1.00 | 1.00 | 1.61 | 1.95 | 2.01 |
| 2 | 2.62 | 2.99 | 3.02 | 2.00 | 2.01 | 2.00 | 2.19 | 2.76 | 3.15 |
| 3 | 3.19 | 3.68 | 3.81 | 2.95 | 2.96 | 2.95 | 3.31 | 3.93 | 3.96 |
Chapter 3
APPLICATION TO CONVECTIVE WAVE EQUATIONS

The content in this chapter has been published in SIAM J. Numer. Anal., 57(5), 2019, pp. 2469-2492 under the name "An energy-based discontinuous Galerkin method for the wave equation with advection" [92]

In this chapter, we extend the energy-based DG method [3] to the wave equations with advection. The energy form of the wave equations with advection is not the simple sum of kinetic and potential energy as in [3]. In our work, both subsonic and supersonic advection is allowed and error estimates in the energy-norm are established. We prove a suboptimal convergence rate by 1 for central fluxes and by 1/2 for upwind fluxes. For problems in one space dimension we prove optimal estimates in the upwind case, and observe optimal convergence in $L^2$ for upwind fluxes in experiments on regular meshes.

3.1. Introduction

Regularly hyperbolic partial differential equations [23, Ch. 5] arising as the Euler-Lagrange equations associated to a Lagrangian, $L(x, t, u, \frac{\partial u}{\partial x_j}, \frac{\partial u}{\partial t})$ are a good target for a general formulation of DG methods. In [3], the energy-based DG formulation is restricted to a class of Lagrangians of the form $L = \frac{1}{2} |\frac{\partial u}{\partial t}|^2 - U(\nabla u, u)$. In this work, we focus on the scalar wave equation with constant advection and $L$ is given by

$$\frac{1}{2} \left( \frac{\partial u}{\partial t} + w \cdot \nabla u \right)^2 - \frac{c^2}{2} |\nabla u|^2,$$

which leads to the equation

$$\left( \frac{\partial}{\partial t} + w \cdot \nabla \right)^2 u = c^2 \Delta u,$$
and an associated energy density

\[ \mathcal{E} = \frac{1}{2} \left( \frac{\partial u}{\partial t} + \mathbf{w} \cdot \nabla u \right)^2 + \frac{c^2}{2} |\nabla u|^2. \]

Besides being a simple example of a second order regularly hyperbolic partial differential equation which cannot be directly treated by the method proposed in [3], the advective wave equation is a physically interesting model of sound propagation in a uniform flow. Moreover, we believe our methods could be generalized to treat more general models used in aeroacoustics.

The DG method developed in this chapter for (3.1) guarantees energy stability based on simply defined upwind or central fluxes without mesh-dependent parameters and only introduces one extra field. Lastly, our scheme allows both subsonic and supersonic background flows by developing different forms of upwind fluxes.

Here is the outline of this chapter. In Section 3.2, the DG formulation is developed. We focus on the analysis of the spatial semi-discretization; energy and error estimates are presented in Section 3.3. Section 3.4 displays numerical examples in both one and two space dimensions.

**3.2. DG formulation**

We consider the following second-order wave equation with a constant advection \( \mathbf{w} \),

\[ \left( \frac{\partial}{\partial t} + \mathbf{w} \cdot \nabla \right)^2 u = c^2 \Delta u + f(x, t), \quad x \in \Omega, \quad t \geq 0, \quad (3.1) \]

with suitable initial and boundary conditions, which are described later. By introducing a second scalar variable

\[ v = \frac{\partial u}{\partial t} + \mathbf{w} \cdot \nabla u, \]

we produce a system which is first order in time

\[ \begin{cases} \frac{\partial u}{\partial t} + \mathbf{w} \cdot \nabla u - v = 0, \\ \frac{\partial v}{\partial t} + \mathbf{w} \cdot \nabla v - c^2 \Delta u = f. \end{cases} \quad (3.2) \]
The energy density for system (3.2) is given by

$$E(u, v) = \frac{1}{2}v^2 + \frac{c^2}{2} |\nabla u|^2,$$

Dividing the domain $\Omega$ into non-overlapping subdomains $\Omega_j$ with $\Omega = \cup_j \Omega_j$, let $n$ denote the outward-pointing unit normal. Then we find the change of energy on an element $\Omega_j$ is given by boundary and source contributions,

$$\frac{d}{dt} \int_{\Omega_j} E(u, v) = \int_{\Omega_j} vf + \int_{\partial \Omega_j} c^2 v \nabla u \cdot n - \frac{1}{2} c^2 |\nabla u|^2 w \cdot n - \frac{1}{2} v^2 w \cdot n.$$ \hspace{1cm} (3.3)

Let $(u^h, v^h)$ restricted to $\Omega_j$ be the DG approximation of $(u, v)$. To discretize we require that the components of $(u^h, v^h)$ are polynomials of degree $q$ and $s$, respectively, that is, elements of $\mathcal{P}^{(q,s)} = \Pi^q \times \Pi^s$.\hspace{1cm} (1.1)

Now we seek approximations satisfying a discrete energy identity analogous to (3.3) to the system (3.2). Consider the discrete energy in $\Omega_j$,

$$E^h_j(t) = \int_{\Omega_j} \frac{1}{2}(v^h)^2 + \frac{1}{2} c^2 |\nabla u^h|^2,$$ \hspace{1cm} (3.4)

and its time derivative,

$$\frac{dE^h_j}{dt} = \int_{\Omega_j} v^h \frac{\partial u^h}{\partial t} + c^2 \nabla u^h \cdot \nabla \frac{\partial u^h}{\partial t}.$$ \hspace{1cm} (3.5)

To develop a weak form compatible with the discrete energy (3.4), choosing $(\phi_u, \phi_v) \in (\Pi^q, \Pi^s)$, we multiply the first equation of (3.2) by $-c^2 \Delta \phi_u$, the second equation of (3.2) by $\phi_v$, then integrate them over $\Omega_j$ and add flux terms which vanish for the continuous problem. This yields

$$\int_{\Omega_j} -c^2 \Delta \phi_u \left( \frac{\partial u^h}{\partial t} + w \cdot \nabla u^h - v^h \right) =$$

$$\int_{\Omega_j} -c^2 \nabla \phi_u \cdot n \left( \frac{\partial u^h}{\partial t} + w \cdot \nabla u^h - v^* \right) - c^2 \nabla \phi_u \cdot (\nabla u^* - \nabla u^h) w \cdot n,$$ \hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

\hspace{1cm} (3.5)

In the numerical experiments we also use the tensor product spaces $Q^q \times Q^s$. Our analysis can be extended to that case with no essential changes.
\[
\int_{\Omega_j} \phi_v \left( \frac{\partial v^h}{\partial t} + w \cdot \nabla v^h - c^2 \Delta u^h \right) = \int_{\Omega_j} \phi_v f + \\
\int_{\partial \Omega_j} c^2 \phi_v (\nabla u^* - \nabla u^h) \cdot n - (v^* - v^h) \phi_v w \cdot n. \quad (3.6)
\]

In what follows, an integration by parts in (3.5) and (3.6) results in the following alternative form,

\[
\int_{\Omega_j} c^2 \nabla \phi_u \cdot \nabla \left( \frac{\partial u^h}{\partial t} + w \cdot \nabla u^h - v^h \right) = \\
\int_{\partial \Omega_j} c^2 (v^* - v^h) \nabla \phi_u \cdot n - c^2 \nabla \phi_u \cdot (\nabla u^* - \nabla u^h) w \cdot n, \quad (3.7)
\]

\[
\int_{\Omega_j} \phi_v \frac{\partial v^h}{\partial t} + \phi_v w \cdot \nabla v^h + c^2 \nabla u^h \cdot \nabla \phi_v = \int_{\Omega_j} \phi_v f + \\
\int_{\partial \Omega_j} c^2 \phi_v \nabla u^* \cdot n - (v^* - v^h) \phi_v w \cdot n. \quad (3.8)
\]

To solve a system generated by (3.7) and (3.8), an equation which determines the mean value of \(\frac{\partial u^h}{\partial t}\) must be supplemented to (3.7). Precisely, we have

\[
\int_{\Omega_j} \tilde{\phi}_u \left( \frac{\partial u^h}{\partial t} + w \cdot \nabla u^h - v^h \right) = 0, \quad \forall \tilde{\phi}_u \in \Pi^0. \quad (3.9)
\]

Note that this equation does not change the discrete energy (3.4).

Let \(\Phi = (\phi_u, \phi_v, \tilde{\phi}_u)\) and \(U = (u^h, v^h)\), we obtain the final form,

\[
\mathcal{B}(\Phi, U) = \sum_j \int_{\Omega_j} \left[ (c^2 \nabla \phi_u \cdot \nabla + \tilde{\phi}_u) \left( \frac{\partial u^h}{\partial t} + w \cdot \nabla u^h - v^h \right) + \phi_v \frac{\partial v^h}{\partial t} + \phi_v w \cdot \nabla v^h \\
+ c^2 \nabla \phi_v \cdot \nabla u^h \right] - \sum_j \int_{\partial \Omega_j} \left[ c^2 (v^* - v^h) \nabla \phi_u \cdot n + c^2 \phi_v \nabla u^* \cdot n \\
- c^2 \nabla \phi_u \cdot (\nabla u^* - \nabla u^h) w \cdot n - (v^* - v^h) \phi_v w \cdot n \right].
\]

Let \(\mathcal{N}\) be the space of arbitrary constants on an element \(\Omega_j\), we then have the following semidiscrete problem,
Problem 1. Find $U = (u^h, v^h) \in \mathcal{P}^{q,s}$ such that

$$
\mathcal{B}(\Phi, U) = \sum_j \int_{\Omega_j} \phi v f, \quad \forall \Phi \in \mathcal{P}^{q,s} \times \mathcal{N}.
$$

We then state the result as follows.

**Theorem 3.1.** Let $U(t)$ and the fluxes $v^*, \nabla u^*$ be given. Then $\frac{dU}{dt}$ is uniquely determined, and the energy identity

$$
\frac{dE^h_j}{dt} = \int_{\Omega_j} v^h f + \int_{\partial\Omega_j} \left[ -\frac{1}{2}c^2|\nabla u^h|^2 w \cdot n - \frac{1}{2}(v^h)^2 w \cdot n + c^2(v^* - v^h)\nabla u^h \cdot n \\
- c^2\nabla u^h \cdot (\nabla u^* - \nabla u^h)w \cdot n + c^2 v^h \nabla u^* \cdot n - v^h(v^* - v^h)w \cdot n \right], \quad (3.10)
$$

holds.

**Proof.** The system on each element $\Omega_j$ is linear with respect to the time derivatives, and the mass matrix of $\frac{dV^h}{dt}$ is nonsingular. The number of linear equations for $\frac{dV^h}{dt}$, which equals the number of independent equations in (3.7) plus the equation in (3.9), matches the dimensionality of $\Pi^q$. If the data $v^h, v^*, \nabla u^h, \nabla u^*$ vanish in (3.7), we must have $\frac{dV^h}{dt} = 0$, and so the linear system is invertible. By setting $\Phi = (U, 0)$ in Problem 1, then (3.10) follows directly. \qed

3.2.1. Fluxes

To complete the DG formulation (3.7)–(3.8), we need to specify the numerical fluxes $\nabla u^*, v^*$ at both inter-element and physical boundaries. Denote the traces of data from outside of an element by “+”, while from inside of an element by “−”. Further, we adopt the common notations

$$
\{\{v\}\} = \frac{1}{2}(v^+ + v^-), \quad [[v]] = v^+ n^+ + v^- n^-,
$$

$$
\{\{\nabla u\}\} = \frac{1}{2}(\nabla u^+ + \nabla u^-), \quad [[\nabla u]] = \nabla u^+ \cdot n^+ + \nabla u^- \cdot n^-.
$$

We first consider the inter-element boundaries. Label two elements sharing a common boundary by 1 and 2. Then their net contribution to the energy derivative is the boundary integral
The scheme is energy conservation if $J^h = 0$, and a typical example is the central flux,

$$v^* = \{\{v\}\}, \; \nabla v^* = \{\{\nabla u\}\}.$$  \hfill (3.11)

To obtain an energy dissipating scheme, $J^h < 0$, we define upwind fluxes containing jumps of DG solutions at the inter-element boundaries. We first consider the case with $|\mathbf{w} \cdot \mathbf{n}| \leq c$.

Let $\xi > 0$, which has a same units with $c$, be the flux splitting parameter,

$$v \nabla u \cdot \mathbf{n} = \frac{1}{4\xi} (v + \xi \nabla u \cdot \mathbf{n})^2 - \frac{1}{4\xi} (v - \xi \nabla u \cdot \mathbf{n})^2 =: F^+ - F^-.$$  

Then we define the numerical fluxes by forcing $(F^+, F^-)$ to be computed by the data from the outside and inside of an element, respectively. Specifically, we enforce the following equation for $l = 1, 2$,

$$v^* - \xi \nabla u^* \cdot \mathbf{n}_l = v^h_l - \xi \nabla u^h_l \cdot \mathbf{n}_l, \; \xi > 0.$$  

Solving the above equations and additionally setting the tangential components of $\nabla u^*$ to be the average of the values from each side, we derive what we call the Sommerfeld flux:

$$v^* = \{\{v^h\}\} - \frac{\xi}{2} [[\nabla v^h]], \; \nabla v^* = -\frac{1}{2\xi} [[v^h]] + \{\{\nabla u^h\}\}.$$  \hfill (3.12)

By this choice we obtain

$$J^h = -\left( \frac{\xi c^2}{2} [[\nabla u^h]]^2 + \frac{c^2}{2\xi} \left|[[v^h]]\right|^2 - \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right)|[\nabla u^h][[v^h]] \cdot \mathbf{w} \right).$$
Let subscript $\tau$ be the orthogonal projection of any vector onto the tangent space of the element boundary, we may rewrite the formula above for $J^h$ as

$$
J^h = -\left(\frac{\xi^2}{2}||[\nabla u^h]|^2 + \frac{c^2}{2\xi}||[v^h]|^2\right) - \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)[[\nabla u^h]](([[v^h]] \cdot n)(w \cdot n) + ([v^h]_{\tau} \cdot w_{\tau}))
$$

$$
= -\left(\frac{\xi^2}{2}||[\nabla u^h]|^2 + \frac{c^2}{2\xi}||[v^h]|^2\right) - \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)[[\nabla u^h]]([v^h] \cdot n)(w \cdot n).
$$

Since

$$
[[\nabla u]]([v] \cdot n) \leq \frac{\alpha}{2}||[\nabla u]|^2 + \frac{1}{2\alpha}||[v]|^2, \quad \alpha > 0,
$$

we conclude that

$$
J^h \leq -\frac{\xi^2}{2}||[\nabla u^h]|^2 + \frac{c^2}{2\xi}||[v^h]|^2 + \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)|w \cdot n|\left(\frac{\alpha}{2}||[\nabla u^h]|^2 + \frac{1}{2\alpha}||[v^h]|^2\right).
$$

Then the numerical energy will not grow, $J^h \leq 0$, if

$$
-\frac{\xi^2}{2} + \frac{c^2}{2\xi} + \frac{\xi}{2}|w \cdot n|\frac{\alpha}{2} \leq 0 \quad \text{and} \quad -\frac{c^2}{2\xi} + \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)|w \cdot n|\frac{1}{2\alpha} \leq 0,
$$

equivalently,

$$
\frac{(c^2 + \xi^2)|w \cdot n|}{2c^2} \leq \frac{2\xi^2c^2}{(c^2 + \xi^2)|w \cdot n|}.
$$

(3.13)

In what follows, we claim the existence of $\alpha$ which satisfies (3.13). Since

$$
\frac{2\xi^2c^2}{(c^2 + \xi^2)|w \cdot n|} - \frac{(c^2 + \xi^2)|w \cdot n|}{2c^2} = \frac{4\xi^2c^4 - (c^2 + \xi^2)^2|w \cdot n|^2}{2c^2|w \cdot n|(c^2 + \xi^2)};
$$

then if $4\xi^2c^4 - (c^2 + \xi^2)^2|w \cdot n|^2 \geq 0$, that is,

$$
|w \cdot n| \leq \frac{2\xi c^2}{c^2 + \xi^2},
$$

we conclude that (3.13) can be satisfied. Thus we have the following results:

a. the numerical energy is dissipated if $|w \cdot n| < \frac{2\xi c^2}{c^2 + \xi^2}$;

b. the numerical energy is conserved if $|w \cdot n| = \frac{2\xi c^2}{c^2 + \xi^2}$.

Particularly, if $\xi = c$, we have a dissipated energy when $|w \cdot n| < c$ and a conserved energy when $|w \cdot n| = c$. 

42
To simplify the analysis for the rest of chapter, we introduce a general parametrization for numerical fluxes,

\[ v^* = (\sigma v_1^h + (1 - \sigma)v_2^h) - \eta[[\nabla u^h]], \quad \nabla u^* = -\beta[[v^h]] + ((1 - \sigma)\nabla u_1^h + \sigma \nabla u_2^h), \]

where \(0 \leq \sigma \leq 1, \beta, \eta \geq 0\). For this general flux form, we have

\[ J^h = - \left( c^2 \eta [[\nabla u^h]]^2 + c^2 \beta \left| [[v^h]] \right|^2 - (c^2 \beta + \eta) [[\nabla u^h]] ([|[[v^h]]| \cdot n](\mathbf{w} \cdot n)) \right), \]

The previous cases for different numerical fluxes are:

*Central flux*: \(\sigma = \frac{1}{2}, \beta = \eta = 0\).

*Sommerfeld flux*: \(\sigma = \frac{1}{2}, \beta = \frac{1}{2} \xi, \eta = \frac{\xi}{2}\).

Next, we consider the case with \(c < |\mathbf{w} \cdot n|\). For this, we define the *upwind fluxes* only come from one element. Specifically,

\[ v^* = v_1^h, \quad \nabla u^* = \nabla u_1^h, \quad (3.14) \]

or

\[ v^* = v_2^h, \quad \nabla u^* = \nabla u_2^h. \quad (3.15) \]

By (3.14), we have

\[ J^h = \frac{1}{2} \left( c^2 |\nabla u_1^h - \nabla u_2^h|^2 + (v_1^h - v_2^h)^2 \right) \mathbf{w} \cdot \mathbf{n}_2 + c^2 (\nabla u_1^h - \nabla u_2^h) \cdot \mathbf{n}_1 (v_1^h - v_2^h) \]
\[ \leq \frac{(\mathbf{w} \cdot \mathbf{n}_2 + c)}{2} \left( c^2 |\nabla u_1^h - \nabla u_2^h|^2 + (v_1^h - v_2^h)^2 \right), \quad (3.16) \]

then \(J^h \leq 0\) if \(\mathbf{w} \cdot \mathbf{n}_2 \leq -c\). By (3.15), we get

\[ J^h = \frac{1}{2} \left( c^2 |\nabla u_1^h - \nabla u_2^h|^2 + (v_1^h - v_2^h)^2 \right) \mathbf{w} \cdot \mathbf{n}_1 + c^2 (\nabla u_1^h - \nabla u_2^h) \cdot \mathbf{n}_2 (v_1^h - v_2^h) \]
\[ \leq \frac{(\mathbf{w} \cdot \mathbf{n}_1 + c)}{2} \left( c^2 |\nabla u_1^h - \nabla u_2^h|^2 + (v_1^h - v_2^h)^2 \right). \quad (3.17) \]

then \(J^h \leq 0\) if \(\mathbf{w} \cdot \mathbf{n}_1 \leq -c\).
3.2.2. Boundary conditions

In this section, we consider the physical boundaries. Specifically, we discuss the inflow boundaries with \( w \cdot n < 0 \) and outflow boundaries \( w \cdot n > 0 \).

3.2.2.1. Inflow boundary conditions

A homogeneous Dirichlet boundary condition, \( u(x,t) = 0 \), i.e., \( \frac{\partial u(x,t)}{\partial t} = 0 \), is given on an inflow boundary, \( w \cdot n < 0 \). This implies

\[
v(x,t) = \frac{\partial u(x,t)}{\partial t} + w \cdot \nabla u(x,t) = w \cdot \nabla u(x,t).
\]

To define the numerical fluxes, we enforce the following conditions

\[
\begin{cases}
v^* - w \cdot \nabla u^* = 0, \\
v^* - \xi \nabla u^* \cdot n = v^h - \xi \nabla u^h \cdot n,
\end{cases}
\]

\[ (\nabla u^*)_\tau = 0, \]

with \( |w \cdot n| \leq c \). Exploiting the fact that \( w \cdot \nabla u^* = (w \cdot n)(\nabla u^* \cdot n) + w_\tau \cdot (\nabla u^*)_\tau \), we solve the above system to obtain

\[
\nabla u^* \cdot n = \frac{\xi \nabla u^h \cdot n - v^h}{\xi - w \cdot n}, \quad v^* = \frac{w \cdot n}{\xi - w \cdot n}(\xi \nabla u^h \cdot n - v^h).
\]

By a direct calculation, we have

\[
-c^2 \nabla u^h \cdot \nabla u^* - v^h v^* =
\]

\[
- \frac{c^2 \xi}{\xi - w \cdot n}(\nabla u^h \cdot n)^2 + \frac{c^2 - \xi w \cdot n}{\xi - w \cdot n}(\nabla u^h \cdot n) v^h + \frac{w \cdot n}{\xi - w \cdot n}(v^h)^2,
\]

and

\[
c^2(v^* - v^h) \nabla u^h \cdot n + c^2 v^h \nabla u^* \cdot n = \frac{c^2 \xi w \cdot n}{\xi - w \cdot n}(\nabla u^h \cdot n)^2 - \frac{c^2}{\xi - w \cdot n}(v^h)^2.
\]
Let the subscript $I$ denote faces with inflow boundaries. Inserting (3.18) into (3.10) and using (3.19) and (3.20) to simplify the resulting equation, we get

\[
\frac{dE^h_{Ij}}{dt} = \int_{\partial\Omega_{Ij}} \left( \frac{1}{2} c^2 |\nabla u^h|^2 + \frac{1}{2} (v^h)^2 - c^2 \nabla u^h \cdot \nabla u^* - v^h v^* \right) w \cdot n \\
+ c^2 (v^* - v^h) \nabla u^h \cdot n + c^2 v^h \nabla u^* \cdot n \\
= \int_{\partial\Omega_{Ij}} \frac{c^2 w \cdot n}{2} |\nabla u^h|^2 + \frac{c^2 w \cdot n}{2} (\nabla u^h \cdot n)^2 + \left( \frac{w \cdot n}{2} + \frac{(w \cdot n)^2 - c^2}{\xi - w \cdot n} \right) \cdot (v^h)^2 \\
+ \frac{(c^2 - \xi w \cdot n)w \cdot n}{\xi - w \cdot n} (\nabla u^h \cdot n)v^h.
\]

Let $a = \frac{c^2 w \cdot n}{2}$, $b = \frac{(c^2 - \xi (w \cdot n)w \cdot n)}{2(\xi - w \cdot n)}$, and $d = \frac{w \cdot n}{2} + \frac{(w \cdot n)^2 - c^2}{\xi - w \cdot n}$. Since $w \cdot n < 0$ on the inflow boundaries, we will have an energy dissipating scheme if $ad > b^2$. Now, let us claim the fact $ad > b^2$. Since

\[
ad = \frac{c^2 \xi^2 (w \cdot n)^2 + 2c^4 (w \cdot n)^2 - 2c^4 \xi (w \cdot n) - c^2 (w \cdot n)^4}{4(\xi - w \cdot n)^2},
\]

and

\[
b^2 = \frac{c^4 (w \cdot n)^2 + \xi^2 (w \cdot n)^4 - 2c^2 \xi (w \cdot n)^3}{4(\xi - w \cdot n)^2},
\]

we find that the numerator of $ad - b^2$ is

\[
(c^2 \xi^2 + c^4)(w \cdot n)^2 - (c^2 + \xi^2)(w \cdot n)^4 + 2c^2 \xi (w \cdot n)^3 - 2c^4 \xi (w \cdot n)
\]

\[
= (c^2 - (w \cdot n)^2)(w \cdot n)((c^2 + \xi^2)w \cdot n - 2c^2 \xi),
\]

then we conclude that

\[
ad > b^2 \quad \text{when} \quad -c < w \cdot n < 0.
\]

Thus we have the following results:

a. the numerical energy is dissipated when $-c < w \cdot n < 0$;

b. the numerical energy is conserved when $w \cdot n = -c$ and $(\nabla u^h)_r = 0$ on inflow boundaries.

For the case $w \cdot n < -c$, we must impose two boundary conditions,

\[
u = 0, \quad \nabla u \cdot n = 0.
\]

(3.21)
Then \( v^* = 0 \) and \( \nabla u^* = 0 \) follow directly. From (3.17) we conclude that the numerical energy is decreasing.

### 3.2.2.2. Outflow boundary conditions

A radiation boundary condition is assumed on the outflow boundaries, \( w \cdot n > 0 \). Specifically, we enforce the following conditions

\[
\begin{align*}
  v^* + \xi \nabla u^* \cdot n &= 0, \\
  v^* - \xi \nabla u^* \cdot n &= v^h - \xi \nabla u^h \cdot n, \\
  (\nabla u^*)_{\tau} &= (\nabla u^h)_{\tau},
\end{align*}
\]

with \( 0 < w \cdot n \leq c \). Solving the system, we obtain

\[
\nabla u^* \cdot n = \frac{\xi \nabla u^h \cdot n - v^h}{2\xi}, \quad v^* = \frac{v^h - \xi \nabla u^h \cdot n}{2}.
\]

(3.22)

By a direct calculation we find that

\[
-c^2 \nabla u^h \cdot \nabla u^* - v^h v^*
\]

\[
= -\frac{c^2}{2}(\nabla u^h \cdot n)^2 - \frac{1}{2}(v^h)^2 + \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)(\nabla u^h \cdot n)v^h - c^2|\nabla u^h|_{\tau}^2,
\]

(3.23)

and

\[
c^2(v^* - v^h)\nabla u^h \cdot n + c^2v^h \nabla u^* \cdot n = -\frac{c^2\xi}{2}(\nabla u^h \cdot n)^2 - \frac{c^2}{2\xi}(v^h)^2.
\]

(3.24)

Let the subscript \( O \) denote faces with outflow boundaries. Inserting (3.22) in (3.10) and applying (3.23) and (3.24) to the resulting equation, we get

\[
\frac{dE_{\mathcal{O}}^h}{dt} = \int_{\partial\Omega_{\mathcal{O}}} \left[ \left(\frac{1}{2}c^2|\nabla u^h|_{\tau}^2 + \frac{1}{2}(v^h)^2 - c^2 \nabla u^h \cdot \nabla u^* - v^h v^*\right)w \cdot n \\
+ c^2(v^* - v^h)\nabla u^h \cdot n + c^2v^h \nabla u^* \cdot n \right]
\]

\[
= \int_{\partial\Omega_{\mathcal{O}}} \left[ -c^2|\nabla u^h|_{\tau}^2 w \cdot n + \left(\frac{c^2}{2\xi} + \frac{\xi}{2}\right)(\nabla u^h \cdot n)v^h w \cdot n - \frac{c^2\xi}{2}(\nabla u^h \cdot n)^2 - \frac{c^2}{2\xi}(v^h)^2 \right].
\]
Since \( \mathbf{w} \cdot \mathbf{n} > 0 \), for positive \( \delta \) we have that
\[
\left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right) (\nabla u^h \cdot \mathbf{n}) v^h \mathbf{w} \cdot \mathbf{n} \leq \frac{\delta}{2} \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right) \mathbf{w} \cdot \mathbf{n} (\nabla u^h \cdot \mathbf{n})^2 + \frac{1}{2\delta} \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right) \mathbf{w} \cdot \mathbf{n} (v^h)^2,
\]
thus the numerical energy decreases if
\[
\frac{\delta}{2} \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right) \mathbf{w} \cdot \mathbf{n} - \frac{c^2}{2\xi} \leq 0, \quad \frac{1}{2\delta} \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right) \mathbf{w} \cdot \mathbf{n} - \frac{c^2}{2\xi} \leq 0.
\]
are satisfied. Alternatively, we only need to show that
\[
\frac{(c^2 + \xi^2) \mathbf{w} \cdot \mathbf{n}}{2c^2} \leq \delta \leq \frac{2c^2 \xi^2}{(c^2 + \xi^2) \mathbf{w} \cdot \mathbf{n}}.
\]
Since
\[
\frac{2c^2 \xi^2}{(c^2 + \xi^2) \mathbf{w} \cdot \mathbf{n}} - \frac{(c^2 + \xi^2) \mathbf{w} \cdot \mathbf{n}}{2c^2} = \frac{4c^4 \xi^2 - (c^2 + \xi^2)^2 (\mathbf{w} \cdot \mathbf{n})^2}{2c^2 (c^2 + \xi^2) \mathbf{w} \cdot \mathbf{n}},
\]
\( \delta \) exists if \( 4c^4 \xi^2 - (c^2 + \xi^2)^2 (\mathbf{w} \cdot \mathbf{n})^2 \geq 0 \), i.e. \( (\mathbf{w} \cdot \mathbf{n})^2 \leq \frac{4c^4 \xi^2}{(c^2 + \xi^2)^2} \). Then we have following results:

a. the numerical energy is dissipated when \( 0 < \mathbf{w} \cdot \mathbf{n} < \frac{2\xi^2}{c^2 + \xi^2} \). Specially, if \( \xi = c \), the condition reduces to \( 0 < \mathbf{w} \cdot \mathbf{n} < c \);

b. the numerical energy is conserved when \( \mathbf{w} \cdot \mathbf{n} = \frac{2\xi^2}{c^2 + \xi^2} \) and \( (\nabla u^h)_{\tau} = 0 \) on the outflow boundaries. Specifically, if \( \xi = c \), the condition reduces to \( \mathbf{w} \cdot \mathbf{n} = c \) and \( (\nabla u^h)_{\tau} = 0 \).

Lastly, for the case with \( \mathbf{w} \cdot \mathbf{n} > c \) we impose no boundary conditions. That means the numerical fluxes are given by
\[
\mathbf{v}^* = \mathbf{v}^h, \quad \nabla \mathbf{u}^* = \nabla \mathbf{u}^h.
\]
(3.25)

By combining (3.16) and (3.25), we conclude that the numerical energy is dissipated.

**Theorem 3.2.** Suppose the following fluxes are imposed:

i. The Sommerfeld flux (3.18) at subsonic inflow boundaries, \( -c < \mathbf{w} \cdot \mathbf{n} < 0 \),

ii. (3.21) at supersonic inflow boundaries, \( \mathbf{w} \cdot \mathbf{n} < -c \),

iii. The Sommerfeld flux (3.22) at subsonic outflow boundaries, \( 0 < \mathbf{w} \cdot \mathbf{n} < \frac{2\xi^2}{c^2 + \xi^2} \),

iv. (3.25) at supersonic outflow boundaries \( \mathbf{w} \cdot \mathbf{n} > c \),
v. At inter-element boundaries either the central flux (3.11), or

a. the Sommerfeld flux (3.12) if \(|\mathbf{w} \cdot \mathbf{n}| < \frac{2c^2}{\xi + c^2}\),

b. (3.14) if \(\mathbf{w} \cdot \mathbf{n}_1 > c\),

c. (3.15) if \(\mathbf{w} \cdot \mathbf{n}_2 > c\).

Also suppose that the parameter \(\xi\) defining the Sommerfeld flux at any boundary satisfies \(|\mathbf{w} \cdot \mathbf{n}| \leq \frac{2c^2}{\xi + c^2}\). Then the discrete energy \(E^h(t) = \sum_j E_j^h(t)\) with \(E_j^h(t)\) defined in (3.4) satisfies

\[
\frac{dE^h}{dt} = \sum_j \int_{\Omega_j} v^h f - \sum_j \int_{F_j} \left[ c^2 \eta[[\nabla u^h]]^2 + c^2 \beta[[v^h]]^2 \right.
- c^2 \beta[[v^h]] \cdot (\nabla u_1^h(\mathbf{w} \cdot \mathbf{n}_1) + \nabla u_2^h(\mathbf{w} \cdot \mathbf{n}_2)) - \eta[[\nabla u^h]][[v^h]] \cdot \mathbf{w}
\left. + \sum_j \int_{B_{ij}} \left[ \frac{c^2 \mathbf{w} \cdot \mathbf{n}}{2} |(\nabla u^h)_r|^2 + \frac{c^2 \mathbf{w} \cdot \mathbf{n}}{2} (\nabla u^h \cdot \mathbf{n})^2 + \left( \frac{\mathbf{w} \cdot \mathbf{n}}{2} + \frac{(\mathbf{w} \cdot \mathbf{n})^2 - c^2}{\xi - \mathbf{w} \cdot \mathbf{n}} \right) (v^h)^2 \right.ight.
+ \left( \frac{c^2 - \xi \mathbf{w} \cdot \mathbf{n}}{\xi - \mathbf{w} \cdot \mathbf{n}} (\nabla u^h \cdot \mathbf{n}) v^h \right)
+ \sum_j \int_{B_{i\Omega}} \left[ - \frac{c^2}{2} |(\nabla u^h)_r|^2 \mathbf{w} \cdot \mathbf{n} \right.ight.
+ \left( \frac{c^2}{2 \xi} + \frac{\xi}{2} (\nabla u^h \cdot \mathbf{n}) v^h \mathbf{w} \cdot \mathbf{n} - \frac{c^2 \xi}{2} (\nabla u^h \cdot \mathbf{n})^2 - \frac{c^2}{2 \xi} (v^h)^2 \right]
\]
\[
\leq \sum_j \int_{\Omega_j} v^h f \leq \sqrt{2E^h} \|f\|_{L^2}. \quad (3.26)
\]

3.3. Error estimates in the energy norm

To derive the error estimates, we define the errors by

\[ e_u = u - u^h, \quad e_v = v - v^h, \]

and let \(\mathbf{D}^h = (e_u, e_v)\). Then the fundamental Galerkin orthogonality is given by

\[ \mathcal{B}(\Phi, \mathbf{D}^h) = 0. \]

For the rest of analysis, we follow the standard approach of comparing the DG solution \((u^h, v^h)\) to an arbitrary polynomial approximation \((\tilde{u}^h, \tilde{v}^h) \in \mathcal{P}^q\) \(Q\). Define by the differences

\[ \tilde{e}_u = \tilde{u}^h - u^h, \quad \tilde{e}_v = \tilde{v}^h - v^h, \quad \delta_u = \tilde{u}^h - u, \quad \delta_v = \tilde{v}^h - v. \]
and let
\[ \tilde{D}^h = (\tilde{e}_u, \tilde{e}_v) \in P^{q,s}, \quad \tilde{D}_0^h = (\tilde{e}_u, \tilde{e}_v, 0) \in P^{q,s} \times N, \quad \Delta^h = (\delta_u, \delta_v). \]
Then from the relation \( D^h = \tilde{D}^h - \Delta^h \), we find the following error equality
\[ B(\tilde{D}_0^h, \tilde{D}^h) = B(\tilde{D}_0^h, \Delta^h). \]
Define the energy of \( \tilde{D}^h \) by
\[ E^h = \frac{1}{2} \sum_j \int_{\Omega_j} \tilde{e}_v^2 + c^2|\nabla \tilde{e}_u|^2. \]
Then by the same analysis that lead to (3.26), we find that
\[ \frac{dE^h}{dt} = B(\tilde{D}_0^h, \Delta^h) \]
\[ + \sum_j \int_{B_{ij}} \left[ c^2\eta[[\nabla \tilde{e}_u]]^2 + c^2\beta[[\tilde{e}_v]]^2 - c^2\beta[[\tilde{e}_v]] \cdot (\nabla \tilde{e}_u)(w \cdot n_1) \right. \]
\[ + \nabla \tilde{e}_u(w \cdot n_2) - \eta[[\nabla \tilde{e}_u]][[\tilde{e}_v]] \cdot w \]
\[ + \sum_j \int_{B_{ij}} \left[ c^2\eta[[\nabla \tilde{e}_u]]^2 + c^2\beta[[\tilde{e}_v]]^2 - c^2\beta[[\tilde{e}_v]] \cdot (\nabla \tilde{e}_u)(w \cdot n_1) \right. \]
\[ + \left( \frac{c^2 - \xi w \cdot n}{\xi - w \cdot n} (\nabla \tilde{e}_u)(w \cdot n_1) \tilde{e}_v \right) \]
\[ + \sum_j \int_{B_{ij}} \left[ - \frac{c^2}{2}(\nabla \tilde{e}_u)(w \cdot n_1)^2 w \cdot n \right. \]
\[ + \frac{c^2}{2}(\nabla \tilde{e}_u)(w \cdot n_1) \tilde{e}_v w \cdot n - \frac{c^2}{2}(\nabla \tilde{e}_u)(w \cdot n_1)^2 - \frac{c^2}{2}(\tilde{e}_v)^2. \quad (3.27) \]

The strategy of the error analysis in this section is to choose suitable polynomial approximations \((\tilde{u}^h, \tilde{v}^h)\). The polynomials can both approximate \((u, v)\) and eliminate some of the (potentially) larger terms in \( B(\tilde{D}_0^h, \Delta^h) \). In what follows, for simplicity, we will assume that \((u^h, v^h) = (\tilde{u}^h, \tilde{v}^h)\) at \( t = 0 \). We note that in the numerical experiments it is beneficial to subtract off a function which satisfies the initial conditions, thus solving a forced equation with zero initial data.
3.3.1. General case

For the following analysis, we use the $L^2$ projection of $v$ and a (broken) $H^1$ seminorm projection of $u$. Precisely, on each element $\Omega_j$, we impose for all time $t$

$$
\int_{\Omega_j} \nabla \phi_u \cdot \nabla \delta_u = \int_{\Omega_j} \phi_u \delta_u = \int_{\Omega_j} \delta_u = 0, \quad \forall (\phi_u, \phi_v) \in \mathcal{P}^{q,s}.
$$

(3.28)

Then, integrating by parts, we obtain the following expression for $\mathcal{B}(\tilde{D}_0^h, \Delta^h)$:

$$
\mathcal{B}(\tilde{D}_0^h, \Delta^h) = \sum_j \int_{\Omega_j} \left[ c^2 \nabla \tilde{e}_u \cdot \nabla \left( \frac{\partial \delta_u}{\partial t} \right) - c^2 \Delta \tilde{e}_u \nu \cdot \nabla \delta_u + c^2 \nabla \tilde{e}_u \nu \cdot \Delta \delta_u + \tilde{e}_v \nu \cdot \frac{\partial \delta_u}{\partial t} \right]
$$

$$
- \nu \cdot \nabla \tilde{e}_u \nu \cdot \nabla \delta_u + c^2 \nabla \tilde{e}_u \nu \cdot \nabla \delta_u \right] - \sum_j \int_{\partial \Omega_j} \left[ - c^2 \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu + c^2 \delta_v \nu \cdot \nu \nabla \tilde{e}_u \nu \cdot \nu
$$

$$
+ c^2 \nu \cdot \nabla \delta_u \cdot \nu - c^2 \nabla \tilde{e}_u \nu \cdot \left( \nabla \delta_u - \nabla \delta_u \right) \nu \cdot \nu - \delta_v \nu \cdot \nu \cdot \nu.
$$

Next, we rewrite the volume integral $\int_{\Omega_j} \Delta \tilde{e}_u \nu \cdot \nabla \delta_u$. For example in $\mathbb{R}^3$, we have

$$
\int_{\Omega_j} \Delta \tilde{e}_u \nu \cdot \nabla \delta_u = \int_{\Omega_j} \left( \nabla \nu \cdot \nabla \tilde{e}_u \nu \cdot \nu + \nabla \nu \times \nu \times \nu \right) \cdot \nabla \delta_u,
$$

(3.29)

and the formulas in other dimensions are analogous. Further, by a direct calculation, (3.29) gives

$$
\int_{\Omega_j} \nabla^2 \tilde{e}_u \nu \cdot \nabla \delta_u = \int_{\Omega_j} \nabla \nu \cdot \nabla \tilde{e}_u \nu \cdot \nu \cdot \nabla \delta_u + \nabla \nu \cdot \nu \cdot \nu \cdot \nabla \delta_u.
$$

Then invoking (3.28) the volume integrals in $\mathcal{B}(\tilde{D}_0^h, \Delta^h)$ will vanish and $\mathcal{B}(\tilde{D}_0^h, \Delta^h)$ is simplified to

$$
\mathcal{B}(\tilde{D}_0^h, \Delta^h) = - \sum_j \int_{\partial \Omega_j} c^2 \left( \nu \times \nu \nabla \tilde{e}_u \nu \cdot \nu \times \nu \nu \cdot \nu - c^2 \nu \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu
$$

$$
+ c^2 \delta_v \nabla \tilde{e}_u \nu \cdot \nu + c^2 \delta_v \nabla \delta_u \cdot \nu - c^2 \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu
$$

$$
= - \sum_j \int_{\partial \Omega_j} c^2 \left( - \nabla \delta_u \nu \cdot \nu \tilde{e}_u \nu \cdot \nu + \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu \right)
$$

$$
- c^2 \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu
$$

$$
+ c^2 \delta_v \nabla \tilde{e}_u \nu \cdot \nu + c^2 \delta_v \nabla \delta_u \cdot \nu - c^2 \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu
$$

$$
= - \sum_j \int_{\partial \Omega_j} c^2 \delta_v \nabla \tilde{e}_u \nu \cdot \nu - \delta_v \nabla \tilde{e}_u \nu \cdot \nu + c^2 \delta_v \nabla \delta_u \cdot \nu - c^2 \nabla \tilde{e}_u \nu \cdot \nu \nabla \delta_u \cdot \nu.
$$

50
Combining contributions from neighboring elements, we derive

$$B(\tilde{D}_0^h, \Delta^h) = -\sum_j \int_{F_j} \left[ c^2[\nabla \tilde{e}_u] \delta_v - \left[ (\tilde{e}_v) \delta_v + c^2[[\tilde{e}_v]] \right] \cdot w + c^2[\tilde{e}_v] \cdot \nabla \delta_u^* 
- c^2 \nabla \tilde{e}_u \cdot \nabla \delta_u w \cdot n_1 - c^2 \nabla \tilde{e}_{u_2} \cdot \nabla \delta_u^* w \cdot n_2 \right]$$

$$- \sum_j \int_{B_j} \left[ c^2 \delta_v \nabla \tilde{e}_v \cdot n - \delta_v \tilde{e}_v w \cdot n + c^2 \tilde{e}_v \nabla \delta_u^* \cdot n - c^2 \nabla \tilde{e}_v \cdot \nabla \delta_u^* w \cdot n \right].$$

Here, we have introduced the fluxes $\delta_v^*$, $\nabla \delta_u^*$ built from $\delta_v$, $\nabla \delta_u$ according to the specification in Section 3.2.1. For the rest of the analysis, $C, C_0, C_1$ will be constants independent of the solution and the element diameter $h$ for a shape-regular mesh. In addition, denote by $||\cdot||$ a Sobolev norm and $|\cdot|$ the associated seminorm. We then have the following error estimate.

**Theorem 3.3.** Let $\bar{q} = \min(q - 1, s)$. Then there exist numbers $C_0, C_1$ depending only on $s, q, \eta, \beta, \xi$, and the shape-regularity of the mesh, such that for smooth solutions $u \in L^\infty(0, T; H^{q+2}(\Omega)), v \in L^\infty(0, T; H^{q+1}(\Omega))$ and time $T$

$$||\nabla e_u(\cdot, T)||^2_{L^2(\Omega)} + ||e_v(\cdot, T)||^2_{L^2(\Omega)} \leq (C_0 T + C_1 T^2) \max_{t \leq T} \left[ h^{2q} (||u(\cdot, t)||^2_{H^{q+2}(\Omega)} + ||v(\cdot, t)||^2_{H^{q+1}(\Omega)}) \right], \quad (3.30)$$

where

$$\theta = \begin{cases} 
\bar{q}, & \beta, \eta \geq 0, \quad |w \cdot n| \leq \frac{2\xi}{c_2 + 1}, \\
\bar{q} + \frac{1}{2}, & \beta, \eta > 0, \quad |w \cdot n| \leq \frac{2\xi}{c_2 + 1}.
\end{cases}$$

**Proof.** By the Bramble-Hilbert lemma (e.g. [24]), for $\bar{q} = \min(q - 1, s)$, we have

$$||\delta_v||^2_{L^2(\Omega)} + ||\nabla \delta_u||^2_{L^2(\Omega)} \leq Ch^{2q+2} \left( ||u(\cdot, t)||^2_{H^{q+2}(\Omega)} + ||v(\cdot, t)||^2_{H^{q+1}(\Omega)} \right), \quad (3.31)$$

$$||\frac{\partial \delta_v}{\partial t}||^2_{L^2(\Omega)} \leq Ch^{2s+2} \left( \frac{\partial ||v(\cdot, t)||^2_{H^{q+2}(\Omega)}}{\partial t} \right)_{H^{q+1}(\Omega)}, \quad (3.32)$$

$$||\delta_v^*||^2_{L^2(\Omega)} + ||\nabla \delta_u^* \cdot n||^2_{L^2(\Omega)} \leq Ch^{2q+1} \left( ||u(\cdot, t)||^2_{H^{q+2}(\Omega)} + ||v(\cdot, t)||^2_{H^{q+1}(\Omega)} \right), \quad (3.33)$$

$$||\tilde{e}_v||^2_{L^2(\Omega)} + ||\nabla \tilde{e}_u \cdot n||^2_{L^2(\Omega)} \leq C h^{-1} \left( ||\tilde{e}_v||^2_{L^2(\Omega)} + ||\nabla \tilde{e}_u||^2_{L^2(\Omega)} \right). \quad (3.34)$$

First, consider the case with either central flux and Sommerfeld flux. By the same estimates as those leading to Theorem 3.2 in (3.27), we conclude that the time derivative of the error energy satisfies

$$\frac{dE^h}{dt} \leq B(\tilde{D}_0^h, \Delta^h).$$
Combining the Cauchy-Schwarz inequality with (3.33) and (3.34), we obtain that

\[
\mathcal{B}(\tilde{D}_0^h, \Delta^h) \leq C \sum_j \| \nabla \tilde{e}_u \cdot n \|_{L^2(\partial \Omega_j)} \| \delta^*_u \|_{L^2(\partial \Omega_j)} + \| \tilde{e}_v \|_{L^2(\partial \Omega_j)} \| \nabla \delta^*_u \cdot n \|_{L^2(\partial \Omega_j)} \\
+ \| \nabla \tilde{e}_u \|_{L^2(\partial \Omega_j)} \| \nabla \delta^*_u \|_{L^2(\partial \Omega_j)} + \| \delta^*_u \|_{L^2(\partial \Omega_j)} \| \tilde{e}_v \|_{L^2(\partial \Omega_j)} \\
\leq C \sqrt{h^q} h^q \| u(\cdot, t) \|_{H^{q+2}(\Omega)} + \| v(\cdot, t) \|_{H^{q+1}(\Omega)}.)
\]

Then a direct integration in time with \((\tilde{e}_u, \tilde{e}_v) = 0\) at \(t = 0\) gives us

\[
\mathcal{E}^h(T) \leq CT^2 \max_{t \leq T} h^{2q} \left( \| u(\cdot, t) \|_{H^{q+2}(\Omega)}^2 + \| v(\cdot, t) \|_{H^{q+1}(\Omega)}^2 \right).
\]

For dissipative fluxes, \(\beta, \eta > 0\), the estimates can be improved. The contribution from the inter-element boundaries is given by

\[
\Theta_1 = - \sum_j \int_{F_j} \left[ c^2 \left( \| \nabla \tilde{e}_u \|_{\Omega_j} \| \delta^*_u \left( \nabla \tilde{e}_u \cdot n \right) \|_{\Omega_j} + \| \tilde{e}_v \|_{\Omega_j} \| \nabla \delta^*_u \|_{\Omega_j} \right) - c^2 \nabla \tilde{e}_u \cdot \nabla \delta^*_u \cdot n \right] \\
- c^2 \nabla \tilde{e}_u \cdot \nabla \delta^*_u \cdot n - \sum_j \int_{F_j} \left[ c^2 \eta \| \nabla \tilde{e}_u \|_{\Omega_j}^2 + c^2 \beta \| \tilde{e}_u \|_{\Omega_j}^2 \right] \\
- c^2 \beta \| \tilde{e}_v \| \left( \nabla \tilde{e}_u \cdot (w \cdot n_1) + \nabla \tilde{e}_u \cdot (w \cdot n_2) \right) - \eta \| \nabla \tilde{e}_u \| \| \tilde{e}_v \| \cdot w,
\]

which leads to

\[
\Theta_1 \leq C \sum_j \left( \| \delta^*_u \|_{L^2(F_j)}^2 + \| \nabla \delta^*_u \|_{L^2(F_j)}^2 \right) \\
\leq C h^{2q+1} \left( \| u(\cdot, t) \|_{H^{q+2}(\Omega)}^2 + \| v(\cdot, t) \|_{H^{q+1}(\Omega)}^2 \right).
\]

At the inflow physical boundaries, we have

\[
\Theta_2 = - \sum_j \int_{B_{ij}} \left[ c^2 \delta^*_u \nabla \tilde{e}_u \cdot n - \delta^*_u \tilde{e}_v \cdot w \cdot n + c^2 \tilde{e}_v \nabla \delta^*_u \cdot n - c^2 (\nabla \tilde{e}_u \cdot n) (\nabla \delta^*_u \cdot n) w \cdot n \right] \\
- c^2 (\nabla \tilde{e}_u) \cdot (\nabla \delta^*_u) \cdot w \cdot n + \sum_j \int_{B_{ij}} \left[ \frac{c^2 w \cdot n}{2} (\nabla \tilde{e}_u) \cdot n \right] + \sum_j \int_{B_{ij}} \left[ \frac{c^2 w \cdot n}{2} (\nabla \tilde{e}_u) \cdot n \right] \\
+ \frac{(w \cdot n)^2 - c^2}{\xi - w \cdot n} \| \tilde{e}_v \| + \frac{(c^2 - \xi w \cdot n) w \cdot n}{\xi - w \cdot n} \| \nabla \tilde{e}_u \cdot n \| \tilde{e}_v \|,
\]

where we have used the fact

\[
\nabla \tilde{e}_u \cdot \nabla \delta^*_u = (\nabla \tilde{e}_u \cdot n) (\nabla \delta^*_u \cdot n) + (\nabla \tilde{e}_u) \cdot (\nabla \delta^*_u).
\]

(3.36)
Further, we obtain that

\[ \Theta_2 \leq C \sum_j \left( \| \delta_v^* \|^2_{L^2(B_{jI})} + \| \nabla \delta_u^* \|^2_{L^2(B_{jI})} \right) \]

\[ \leq C h^{2q+1} \left( |u(\cdot, t)|^2_{H^{q+2}(\Omega)} + |v(\cdot, t)|^2_{H^{q+2}(\Omega)} \right). \]  

(3.37)

Similarly, by (3.36), at the outflow boundaries we have

\[ \Theta_3 = - \sum_j \int_{B_{jO}} \left[ c^2 \delta_v^* \nabla \tilde{e}_v \cdot n - \delta_v^* \tilde{e}_v w \cdot n + c^2 \tilde{e}_v \nabla \delta_u^* \cdot n - c^2 (\nabla \tilde{e}_u \cdot n)(\nabla \delta_u^* \cdot n)w \cdot n \right. \]

\[ \left. - c^2 (\nabla \tilde{e}_u)_\tau \cdot (\nabla \delta_u^*)_\tau w \cdot n \right] + \sum_j \int_{B_{jO}} \left[ - \frac{c^2}{2} |(\nabla \tilde{e}_u)_\tau|^2 w \cdot n \right. \]

\[ \left. + \left( \frac{c^2}{2\xi} + \frac{\xi}{2} \right)(\nabla \tilde{e}_u \cdot n)\tilde{e}_v w \cdot n - \frac{c^2}{2} (\nabla \tilde{e}_u \cdot n)\tilde{e}_v w \cdot n \right. \]

resulting in

\[ \Theta_3 \leq C \sum_j \left( \| \delta_v^* \|^2_{L^2(B_{jO})} + \| \nabla \delta_u^* \|^2_{L^2(B_{jO})} \right) \]

\[ \leq C h^{2q+1} \left( |u(\cdot, t)|^2_{H^{q+2}(\Omega)} + |v(\cdot, t)|^2_{H^{q+2}(\Omega)} \right). \]  

(3.38)

Combining (3.27) with (3.35)–(3.38) yields

\[ E^h(T) \leq CT \max_{t \leq T} h^{2q+1} \left( |u(\cdot, t)|^2_{H^{q+2}(\Omega)} + |v(\cdot, t)|^2_{H^{q+1}(\Omega)} \right). \]

Since \( e_v = \tilde{e}_v - \delta_v, e_u = \tilde{e}_u - \delta_u \), (3.30) follows from the triangle inequality and an invocation of (3.31).

Remark. A similar analysis yields the same results for the supersonic boundaries with \( |w \cdot n| > c \).

3.3.2. Improved estimates for one dimension

We can further prove an optimal estimate if the problem is in one dimension. In what follows, we assume \( s = q - 1 \) and seek polynomial approximation \((\tilde{u}^h, \tilde{v}^h)\) such that the boundary terms in \( B(\bar{D}_0^h, \Delta^h) \) vanish:

\[ \delta_v^* = \frac{\partial \delta_u^*}{\partial x} = 0. \]  

(3.39)
This can be achieved if we impose the boundary conditions at the end points of the element
\(\Omega_j = (x_{j-1}, x_j)\) as follows

\[
(1 + \beta - \alpha)\delta_v + (\eta + \alpha) \frac{\partial \delta_u}{\partial x} = 0, \quad x = x_{j-1},
\]

\[
(\beta + \alpha)\delta_v - (1 + \eta - \alpha) \frac{\partial \delta_u}{\partial x} = 0, \quad x = x_j.
\]

(3.40) (3.41)

As in [3], we find that we need to assume \(\alpha(1 - \alpha) = \beta \eta\). It is clear that this condition is satisfied by the Sommerfeld flux but it is not true for the central flux. Given (3.40)–(3.41), \(\delta_u\) and \(\delta_v\) are constructed by requiring

\[
\int_{x_{j-1}}^{x_j} \phi \frac{\partial \delta_u}{\partial x} = \int_{x_{j-1}}^{x_j} \phi \delta_v = \int_{x_{j-1}}^{x_j} \delta_u = 0,
\]

(3.42)

where \(\phi\) is an arbitrary polynomial of degree \(q - 2\). By the Bramble-Hilbert lemma, we have the following inequality

\[
\left\| \frac{\partial \delta_u}{\partial t} \right\|_{H^1(\Omega)} + \left\| \frac{\partial \delta_v}{\partial t} \right\|_{L^2(\Omega)} \leq C h^q \left( \left\| \frac{\partial u}{\partial t} \right\|_{H^{q+1}(\Omega)} + \left\| \frac{\partial v}{\partial t} \right\|_{H^q(\Omega)} \right),
\]

(3.43)

for \((u, v) \in L^\infty(0, T; H^{q+2}(\Omega)) \times L^\infty(0, T; H^{q+1}(\Omega))\). Then, by the same computations as those in Section 3.3.1 and with (3.39) and (3.42), we have that

\[
\mathcal{B}(\mathbf{\hat{D}}_h^b, \Delta_h) = \sum_j \int_{x_{j-1}}^{x_j} c^2 \frac{\partial \hat{\varepsilon}_u}{\partial x} \frac{\partial \delta_u}{\partial t} + \hat{\varepsilon}_v \frac{\partial \delta_v}{\partial t}.
\]

Combining with (3.43) gives an optimal estimate

\[
\frac{d\mathcal{E}_h}{dt} \leq C h^q \sqrt{\mathcal{E}_h} \left( |u(\cdot, t)|^2_{H^{q+2}(\Omega)} + |v(\cdot, t)|^2_{H^{q+1}(\Omega)} \right)^{1/2}.
\]
3.4. Numerical experiments

In this section, numerical simulations are presented to show the convergence in the $L^2$ norm for the method we proposed. Precisely, we use a standard modal formulation with a tensor-product Legendre basis for all cases, and the 4-stage fourth order Runge-Kutta scheme (RK4) method is used to march in time. In addition, the flux splitting parameter introduced in Section 3.2.1 is set to be $\xi = c$ in the Sommerfeld flux for all cases. Finally, we choose a sufficiently small time step size to make the temporal errors dominated by the spatial errors. We note that a study of the spectrum of the spatial discretization establishes that its spectral radius scales with $(c + |w|)\frac{h^2}{q^2}$, with some variability depending on whether $q$ is even or odd. This is comparable to the results in [3].

3.4.1. Periodic boundary conditions in one space dimension

We consider the following problem

$$u_{tt} + 2wu_{tx} + w^2u_{xx} = c^2u_{xx}, \quad x \in (0, 1), \quad t \geq 0,$$

with the initial condition $u(x, 0) = \sin(2\pi x)$ and periodic boundary condition $u(0, t) = u(1, t)$. Then the problem admits an exact traveling wave solution

$$u(x, t) = \cos(2c\pi t)\sin(2\pi(x - wt)), \quad t \geq 0.$$

The discretization is performed on a uniform mesh with element vertices $x_i = ih, \ i = 0, \ldots, n, \ h = 1/n$. The problem is evolved until final time $T = 0.4$ with time step $\Delta t = CFL \times h$ for the degree of approximation polynomials $q = (1, 2, 3, 4, 5, 6)$. In addition, both the central flux and the upwind flux are used in the numerical simulations. Precisely, we test three different cases: $|w| = c$, $|w| < c$, and $|w| > c$. These choices are consistent with our theory. Note that if $|w| > c$ the upwind flux is taken from a single element. We also consider two different choices for the degrees of the approximation spaces: either the approximation degree of $v^h$ is one less than the approximation degree of $u^h$, or $u^h$ and $v^h$ are in the same space. Finally, we present the $L^2$-error for both $u^h$ and $v^h$. 

55
In this work, we did not attempt to precisely characterize the largest time steps we could take based on accuracy/stability requirements. We simply chose convenient sufficiently small values. Precisely, to investigate the convergence due to the spatial discretization, we set the following CFL conditions:

a. central flux: \( \text{CFL} = \frac{0.075}{2\pi} \) for \( q = (1, 2, 3, 4, 5) \) and \( \text{CFL} = \frac{0.00375}{2\pi} \) for \( q = 6 \);

b. upwind flux with \( |w| < c \): \( \text{CFL} = \frac{0.1125}{2\pi} \) for \( q = (1, 2, 3, 4, 5) \) and \( \text{CFL} = \frac{0.01125}{2\pi} \) for \( q = 6 \);

c. upwind flux with \( |w| > c \): \( \text{CFL} = \frac{0.075}{2\pi} \) for \( q = (1, 2, 3, 4, 5) \) and \( \text{CFL} = \frac{0.0075}{2\pi} \) for \( q = 6 \).

The aggressive reductions in time step for \( q = 6 \), though not necessary, were convenient. First, since the spatial accuracy exceeds the temporal accuracy by a fairly significant factor for \( q = 6 \) compared with other cases, we found that we could not observe convergence at the design order for \( q = 6 \) without reducing the time step; second, as the spectral radius of the spatial discretization matrix is proportional to \( q^2 \), the stability requirements are stricter for higher order approximation, although here we found that the scheme was stable for \( q = 6 \) using the same steps as in the other cases.

In our initial numerical experiments, we observed that the convergence was somewhat irregular for all cases when we used \( L^2 \)-projection to determine the non-zero initial conditions. This could be possibly remedied for the upwind flux with a special projection required by the analysis; see [22] for an example which discusses a special projection for the LDG method with alternating fluxes. Here, we propose a simpler method which is to transform the problem to one with zero initial data by

\[ u(x, t) = \tilde{u}(x, t) + u_0(x)e^{-t^2}, \]

where \( u_0(x) \) is the initial condition for \( u(x) \), and then numerically solve for \( \tilde{u} \) which is zero initially.

The \( L^2 \) errors for \( u \) are plotted against the spatial grid size \( h \) in Figure 3.1 when the upwind flux is used. Linear regression estimates of the rate of convergence can be found in Table 3.2 when the approximation degrees for \( u^h \) and \( v^h \) are same; and in Table 3.1 when the degree of \( v^h \) is one less than that of \( u^h \). Here, we only use the ten finest grids to obtain
Figure 3.1: Plots of the errors in $u$ as a function of $h$ in 1d with upwind flux for periodic boundary condition. In the legend, $q$ is the maximum degree of the approximation of $u$, solid lines represent the case of $u^h$ and $v^h$ in the same approximation space, dotted lines represent the case of $v^h$ one degree lower than $u^h$. 

(a) $w = 0.5, c = 1$

(b) $w = 1, c = 0.5$

(c) $w = c = 0.5$
the rates of convergence.

<table>
<thead>
<tr>
<th>Degree ((q)) of approx. for (u)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit (u (w = 0.5, c = 1))</td>
<td>0.90</td>
<td>3.00</td>
<td>4.05</td>
<td>5.03</td>
<td>5.92</td>
<td>6.91</td>
</tr>
<tr>
<td>Rate fit (u (w = 0.5, c = 0.5))</td>
<td>0.92</td>
<td>3.00</td>
<td>4.01</td>
<td>5.00</td>
<td>6.14</td>
<td>7.00</td>
</tr>
<tr>
<td>Rate fit (u (w = 1, c = 0.5))</td>
<td>0.88</td>
<td>2.99</td>
<td>4.01</td>
<td>5.03</td>
<td>6.04</td>
<td>6.93</td>
</tr>
</tbody>
</table>

Table 3.1: Linear regression estimates of the convergence rate of \(u\) in 1d with upwind flux for periodic boundary condition, approximation for \(v\) is one degree lower than \(u\).

<table>
<thead>
<tr>
<th>Degree ((q)) of approx. of (u)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit (u (w = 0.5, c = 1))</td>
<td>0.97</td>
<td>3.00</td>
<td>4.01</td>
<td>5.00</td>
<td>5.98</td>
<td>6.95</td>
</tr>
<tr>
<td>Rate fit (u (w = 0.5, c = 0.5))</td>
<td>1.91</td>
<td>3.01</td>
<td>4.00</td>
<td>5.00</td>
<td>6.00</td>
<td>6.89</td>
</tr>
<tr>
<td>Rate fit (u (w = 1, c = 0.5))</td>
<td>0.97</td>
<td>2.99</td>
<td>4.00</td>
<td>5.01</td>
<td>5.99</td>
<td>6.90</td>
</tr>
</tbody>
</table>

Table 3.2: Linear regression estimates of the convergence rate of \(u\) in 1d with upwind flux for periodic boundary condition, the approximation for \(v\) is one degree lower than \(u\).

For \(q \geq 2\) we observe the optimal convergence, \(q + 1\), for \(u\) with the two choices of approximation space for \(v\). However, from the graphs, we see that there are sometimes noticeable differences in accuracy. Generally speaking, errors are smaller when \(v^h\) is taken from the same space as \(u^h\).

The \(L^2\) errors in \(u\) are plotted against the grid-spacing \(h\) for the central flux in Figure 3.2. Linear regression estimates of the convergence rate can be found in Table 3.4 when \(u^h\) and \(v^h\) are in the same approximation space and in Table 3.3 when \(u^h\) and \(v^h\) are in different spaces. Excluding the special case \(|w| = c\), we observe optimal convergence, \(q + 1\), for \(u\) when \(q\) is odd. For even \(q\) the rate of convergence is only \(q\) for \(u\).
Figure 3.2: Plots of the errors in $u$ as a function of grid spacing $h$ in 1d with central flux for periodic boundary condition. In the legend, $q$ is the maximum degree of the approximation space of $u$, solid lines represent the case of $u^h$ and $v^h$ in the same space, dotted lines represent the case of $v^h$ one degree lower than $u^h$. 
Table 3.3: Linear regression estimates of the convergence rate for $u$ and $v$ in 1d with central flux for periodic boundary condition, the approximation for $v$ is one degree lower than $u$.

<table>
<thead>
<tr>
<th>Degree ($q$) of approx. of $u$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ ($w = 0.5, c = 0.5$)</td>
<td>2.00</td>
<td>1.99</td>
<td>4.05</td>
<td>3.71</td>
<td>6.01</td>
<td>6.27</td>
</tr>
<tr>
<td>Rate fit $u$ ($w = 0.5, c = 1$)</td>
<td>2.00</td>
<td>2.00</td>
<td>4.03</td>
<td>4.03</td>
<td>5.99</td>
<td>5.91</td>
</tr>
<tr>
<td>Rate fit $u$ ($w = 1, c = 0.5$)</td>
<td>2.00</td>
<td>1.99</td>
<td>4.13</td>
<td>4.11</td>
<td>5.81</td>
<td>5.60</td>
</tr>
</tbody>
</table>

Table 3.4: Linear regression estimates of the convergence rate of $u$ and $v$ in 1d with central flux for periodic boundary condition, $u$ and $v$ are in the same approximation space.

<table>
<thead>
<tr>
<th>Degree ($q$) of approx. of $u$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ ($w = 0.5, c = 0.5$)</td>
<td>2.00</td>
<td>3.01</td>
<td>3.99</td>
<td>4.99</td>
<td>6.00</td>
<td>6.73</td>
</tr>
<tr>
<td>Rate fit $u$ ($w = 0.5, c = 1$)</td>
<td>1.99</td>
<td>2.00</td>
<td>4.03</td>
<td>4.01</td>
<td>6.02</td>
<td>6.01</td>
</tr>
<tr>
<td>Rate fit $u$ ($w = 1, c = 0.5$)</td>
<td>2.00</td>
<td>2.00</td>
<td>4.01</td>
<td>4.03</td>
<td>6.04</td>
<td>6.01</td>
</tr>
</tbody>
</table>

3.4.2. Periodic boundary conditions in two space dimensions

In this section, we test our method on the problem

$$\left(\frac{\partial}{\partial t} + w \cdot \nabla\right)^2 u = c^2 \Delta u, \quad (x, y) \in (0,1) \times (0,1), \quad t \geq 0,$$

with periodic boundary conditions $u(0, y, t) = u(1, y, t)$, $u(x, 0, t) = u(x, 1, t)$. Let $w = (w_x, w_y)$, we approximate the exact solution

$$u(x, y, t) = \sin(2c\pi t) \left( \sin(2\pi(x - w_xt)) + \sin(2\pi(y - w_yt)) \right).$$

The discretization is performed with elements whose vertices are on the Cartesian grid defined by $x_i = ih$, $y_j = jh$, $i, j = 0, 1, \ldots, n$ with $h = 1/n$. Here, we only focus on the case where $u^h$ and $v^h$ are in the same space. The problem is evolved until final time $T = 0.2$ by the classic fourth order Runge-Kutta method. We also test two different fluxes: central flux and upwind flux. In addition, we have $CFL = 0.075/(2\pi)$ for the central flux and
CFL = 0.0375/(2\pi) for the upwind flux. Note that at an interface with supersonic normal flow the upwind flux is from one side of an element.

(a) \( w_x = w_y = c = 1 \)  
(b) \( w_x = 0.5, w_y = 1.5, c = 1 \)  
(c) \( w_x = w_y = 0.5, c = 1 \)

Figure 3.3: Plots of the error in \( u \) as a function of spacing-grid \( h \) in 2d with the upwind flux and periodic boundary conditions. In the legend, \( q \) is the degree of the approximation of \( u \) for both \( x \) and \( y \) directions.

Figure 3.3 shows the errors for \( u \) obtained with the upwind flux plotted against the grid-spacing \( h \). Linear regression estimates of the rate of convergence can be found in Table 3.5. We observe an optimal convergence, \( q + 1 \), for \( u \) when \( q \geq 2 \).
Figure 3.4: Plots of the error in $u$ as a function of $h$ in 2d with the central flux and periodic boundary conditions. In the legend, $q$ is the degree of approximation to $u$ and $v$ for both $x$ and $y$ directions.
The $L^2$ error for $u$ with the central flux is plotted against the grid-spacing $h$ in Figure 3.4. Linear regression estimates of the rate of convergence can be found in Table 3.6 for $u$. Similar to the one-dimensional case, convergence is optimal for $u$ when $q$ is odd and suboptimal by one when $q$ is even except in the special case of sonic boundaries, $|w \cdot n| = c$.

![Graphs showing the error in $u$ as a function of $h$ for $q = 1, 2, 3, 4, 5, 6$.](image)

(a) $w_x = w_y = 0.5, c = 1$

(b) $w_x = w_y = 0.5, c = 1$

Figure 3.5: Plots of the error in $u$ as a function of $h$ in 2d with upwind (left) and central (right) fluxes for Dirichlet boundary condition on inflow boundaries and a radiation boundary condition on outflow boundaries. In the legend, $q$ is the degree of the approximation to $u$ for both $x$ and $y$ directions.

3.4.3. Dirichlet and radiation boundary conditions in two space dimensions

<table>
<thead>
<tr>
<th>Degree ($q$) of approx. of $u$ and $v$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ ($w_x = 1, w_y = 1, c = 1$)</td>
<td>1.77</td>
<td>3.04</td>
<td>3.99</td>
<td>5.00</td>
<td>6.00</td>
<td>6.97</td>
</tr>
<tr>
<td>Rate fit $u$ ($w_x = 0.5, w_y = 1.5, c = 1$)</td>
<td>1.05</td>
<td>2.93</td>
<td>4.00</td>
<td>4.99</td>
<td>5.99</td>
<td>6.96</td>
</tr>
<tr>
<td>Rate fit $u$ ($w_x = 0.5, w_y = 0.5, c = 1$)</td>
<td>1.07</td>
<td>2.95</td>
<td>4.02</td>
<td>4.98</td>
<td>5.99</td>
<td>7.00</td>
</tr>
</tbody>
</table>

Table 3.5: Linear regression estimates of the convergence rate of $u$ and $v$ in 2d with upwind flux for periodic boundary condition and $q_x = q_y = q$. 

63
Lastly, we test a problem with a Dirichlet boundary condition on inflow boundaries (left and bottom) and radiation boundary conditions on outflow boundaries (right and top). Further, we consider the following manufactured solution

\[ u(x, y, t) = x(1 - x)^2 y(1 - y)^2 \exp(x + y) \sin(t), \]

and solve

\[ \left( \frac{\partial}{\partial t} + \mathbf{w} \cdot \nabla \right)^2 u = c^2 \Delta u + f, \quad (x, y) \in (0, 1) \times (0, 1), \quad t > 0, \]

with \( f \) determined by \( u \). Note that for this specific choice we have that \( u(x, y, t) = 0 \) on the inflow boundaries and \( u(x, y, t) = u_x(x, y, t) = u_y(x, y, t) = 0 \) on the outflow boundaries. For the following numerical experiments, the degrees for the space of \( u^h \) and \( v^h \) are chosen to be same with \( q_x = q_y = q \). The problem is evolved until final time \( T = 0.2 \). For the purpose of stability, we only consider the subsonic case, \( w_x = w_y = 0.5 \) with \( c = 1 \), and compare both upwind and central fluxes.

In Figure 3.5, we present the errors for \( u \) against the grid-spacing \( h \) for both fluxes. The linear regression estimates of the rate of convergence can be found in Table 3.7. The rates of convergence are very similar to those for the periodic problem.
Table 3.7: Linear regression estimates of the convergence rate for $u$ in $2d$ with Dirichlet boundary condition on inflow boundaries, radiation boundary condition on outflow boundaries and $q_x = q_y = q$. Here the first two rows correspond to the upwind flux and the last two to the central flux.
Chapter 4

APPLICATION TO SEMI-LINEAR WAVE EQUATIONS

The content in this chapter has been submitted under the name "An energy-based discontinuous Galerkin method for semilinear wave equations" [6]

In this chapter, we generalize the energy-based discontinuous Galerkin method proposed in [3] to second-order semilinear wave equations. A stability and convergence analysis is presented along with numerical experiments demonstrating optimal convergence for certain choices of the inter-element fluxes. Applications to the sine-Gordon equation include simulations of breathers, kink, and anti-kink solitons.

4.1. Introduction

We have successfully used energy-based discontinuous Galerkin methods to solve many second-order linear hyperbolic problems. However, although it seems clear that the method can be adapted to any second-order linear hyperbolic system, the formulation for nonlinear problems presented in [3] is both incomplete and inconvenient. In particular, the analogue of (4.7) proposed in [3] involves a nonlinear function of $\phi_u$. Thus the equation would typically be overdetermined. Moreover, to guarantee the energy estimate the system must be satisfied for $\phi_u = u$, which directly leads to a nonlinear problem to calculate $\frac{\partial u}{\partial t}$. Our main result in this work is to show how all these potential issues can be avoided for semilinear problems.

The remainder of the paper is organized as follows. In Section 4.2 we introduce the semi-discretization, proposing a number of inter-element fluxes and proving the basic energy estimate. In Section 4.3 we prove a suboptimal error estimate and present several numerical experiments in Section 4.4. The latter demonstrates optimal convergence for certain choices of flux: specifically an energy-conserving alternating flux as well as two energy-dissipating fluxes. We also present simulations of soliton solutions of the sine-Gordon equation.
4.2. Problem formulation

We consider semilinear wave equations of the form

\[ \frac{\partial^2 u}{\partial t^2} + \theta \frac{\partial u}{\partial t} = c^2 \Delta u + f(u), \quad x \in \Omega \subset \mathbb{R}^d, \quad t \geq 0, \quad (4.1) \]

where \( c > 0 \) is the sound wave speed, \( \theta \geq 0 \) is the dissipation coefficient, and \( f(u) \) is a smooth function with \( \lim_{u \to 0} \frac{f(u)}{u} \) bounded. The initial conditions are given by

\[ u(x, 0) = g_1(x), \quad \frac{\partial u(x, 0)}{\partial t} = g_2(x), \quad x \in \Omega \subset \mathbb{R}^d. \]

Note that when \( \theta = 0 \), (4.1) is the Euler-Lagrange equation derived from the Lagrangian

\[ L = \frac{1}{2} \left( \frac{\partial u}{\partial t} \right)^2 - \frac{c^2}{2} |\nabla u|^2 - F(u), \]

where \( F'(u) = -f(u) \). To derive an energy-based DG formulation for problem (4.1), we introduce a second scalar variable to produce a system which is first order in time,

\[ \begin{cases} 
\frac{\partial u}{\partial t} - v = 0, \\
\frac{\partial v}{\partial t} + \theta v - c^2 \Delta u - f(u) = 0.
\end{cases} \quad (4.2) \]

The energy takes the form

\[ E = \frac{1}{2} \int_{\Omega} v^2 + c^2 |\nabla u|^2 + 2F(u). \quad (4.3) \]

Note that \( F(u) > 0 \) corresponds to a defocusing equation and \( F(u) < 0 \) gives a focusing equation. In the rest of the analysis in this paper, we investigate the defocusing equation with \( F(u) > 0 \), although the method formulation applies in either case. The change of the energy is given by boundary contributions and a volume integral related to the dissipation:

\[ \frac{dE}{dt} = -\theta \int_{\Omega} \left( \frac{\partial u}{\partial t} \right)^2 \, dx + \int_{\partial \Omega} c^2 v \nabla u \cdot n \, dS, \quad (4.4) \]

where \( n \) is the outward-pointing unit normal.
We note that in our error analysis we will make the stronger assumption $f(u) < 0$, which can be enforced after a transformation of variables if we only assume the ratio is bounded. Then the defocusing assumption holds since

$$F(u) = -\int_0^u f(z) dz = \int_0^u \left( -\frac{f(z)}{z} \right) z \, dz > 0.$$  

4.2.1. Semi-discrete DG formulation

We develop an energy-based DG scheme for problem (4.1) through the reformulation (4.2). Let the domain $\Omega$ be discretized by non-overlapping elements $\Omega_j; \Omega = \bigcup_j \Omega_j$. Choose the components of the approximations, $(u^h, v^h)$ to $(u, v)$, restricted to $\Omega_j$, to be polynomials or tensor-product polynomials of degree $q$ and $s$ respectively\(^1\),

$$U^q_h = \left\{ u^h(x, t) : u^h(x, t) \in \Pi^q(\Omega_j), \ x \in \Omega_j, \ t \geq 0 \right\},$$

$$V^s_h = \left\{ v^h(x, t) : v^h(x, t) \in \Pi^s(\Omega_j), \ x \in \Omega_j, \ t \geq 0 \right\}.$$

We seek an approximation to the system (4.2) which satisfies a discrete energy estimate analogous to (4.3). Consider discrete energy in $\Omega_j$,

$$E^h_j(t) = \frac{1}{2} \int_{\Omega_j} (v^h)^2 + c^2 |\nabla u^h|^2 \, dx + \sum_k \omega_{k,j} F(u^h(x_{k,j})), \quad (4.5)$$

and its time derivative

$$\frac{dE^h_j(t)}{dt} = \int_{\Omega_j} v^h \frac{\partial v^h}{\partial t} + c^2 \nabla u^h \cdot \nabla \frac{\partial u^h}{\partial t} \, dx - \sum_k \omega_{k,j} f(u^h(x_{k,j})) \frac{\partial u^h}{\partial t}(x_{k,j}), \quad (4.6)$$

where we have omitted $t$ in $u^h(x_{k,j})$ for simplicity. Note here we use a quadrature rule with nodes $x_{k,j}$ in $\Omega_j$ and positive weights $\omega_{k,j}$ to approximate the integration of the nonlinear terms; in our experiments we use 16-point Gauss rules. To obtain a weak form which is compatible with the discrete energy (4.5) and (4.6), we choose $\phi_u \in U^q_h, \phi_v \in V^s_h$ and test the first equation of (4.2) with $-c^2 \Delta \phi_u$, the second equation of (4.2) with $\phi_v$ and add terms\(^1\) for simplicity we abuse notation and let $\Pi^r$ denote either the polynomials of degree $r$ or the tensor-product polynomials of degree $r$ in each coordinate on a reference element.

68
which vanish for the continuous problem. This yields the following equations,

\[
\int_{\Omega_j} -c^2 \Delta \phi_u \left( \frac{\partial u^h}{\partial t} - v^h \right) \, dx = \int_{\partial \Omega_j} c^2 \nabla \phi_u \cdot n \left( v^* - \frac{\partial u^h}{\partial t} \right) \, dS \\
+ \sum_k \omega_{k,j} \phi_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial u^h}{\partial t}(x_{k,j}) - v^h(x_{k,j}) \right),
\]

\[
\int_{\Omega_j} \phi_v \frac{\partial v^h}{\partial t} - c^2 \phi_v \Delta v^h + \theta \phi_v v^h \, dx - \sum_k \omega_{k,j} \phi_v(x_{k,j}) f(u^h(x_{k,j})) \\
= \int_{\partial \Omega_j} c^2 \phi_v ((\nabla u)^* \cdot n - \nabla u^h \cdot n) \, dS,
\]

where \( v^* \) and \( (\nabla u)^* \) are numerical fluxes on both inter-element and physical boundaries. In what follows, we apply integration by parts to obtain an alternative form,

\[
\int_{\Omega_j} c^2 \nabla \phi_u \cdot \nabla \left( \frac{\partial u^h}{\partial t} - v^h \right) \, dx - \sum_k \omega_{k,j} \phi_u(x_{k,j}) f(u^h(x_{k,j})) \\
= \int_{\partial \Omega_j} c^2 \phi_u \cdot n (v^* - v^h) \, dS, \quad (4.7)
\]

and

\[
\int_{\Omega_j} \phi_u \frac{\partial u^h}{\partial t} + c^2 \nabla \phi_v \cdot \nabla u^h + \theta \phi_v v^h \, dx - \sum_k \omega_{k,j} \phi_v(x_{k,j}) f(u^h(x_{k,j})) \\
= \int_{\partial \Omega_j} c^2 \phi_v (\nabla u)^* \cdot n \, dS. \quad (4.8)
\]

Now by setting \( \phi_u = u^h \) and \( \phi_v = v^h \) and recalling (4.6) we arrive at

\[
\frac{dE^h}{dt} = \sum_j \frac{dE^h_j}{dt} = - \sum_j \int_{\Omega_j} \theta (v^h)^2 \, dx + \sum_j \int_{\partial \Omega_j} c^2 \phi_u \cdot n (v^* - v^h) + c^2 \phi_v (\nabla u)^* \cdot n dS.
\]

Note that if \( \frac{f(u)}{u} = 0 \), then we need to impose an extra equation which determines the mean value of \( \frac{\partial u^h}{\partial t} \),

\[
\int_{\Omega_j} \bar{\phi}_u \left( \frac{\partial u^h}{\partial t} - v^h \right) \, dx = 0.
\]

Here, \( \bar{\phi}_u \) is an arbitrary constant function and this equation does not affect the energy.
The innovation here in comparison with the weak form proposed in [3] is the appearance
of $\phi_u \frac{f(u^h)}{u^h}$ instead of $f(\phi_u)$ in (4.7). This exchange obviously yields an invertible linear
system for computing $\frac{\partial u^h}{\partial t}$. The energy estimate still holds as the two terms are identical for
the special choice $\phi_u = u^h$.

4.2.2. Fluxes

To complete the formulation of energy-based DG scheme proposed in Section 4.2.1, we
must specify the numerical fluxes $v^*, (\nabla u)^*$ both at inter-element and physical boundaries. Denote $'+'$ to be the trace of data from the outside of the element, $'-'$ to be the trace of
data from the inside of the element. Here, we introduce the common notation for averages and jumps,

$$\{v^h\} \equiv \frac{1}{2}(v^h)^+ + \frac{1}{2}(v^h)^-,$$

and

$$\{\nabla u^h\} \equiv \frac{1}{2}(\nabla u^h)^+ + \frac{1}{2}(\nabla u^h)^-.$$ 

4.2.2.1. Inter-element boundaries

To analyze the problem, we label two elements sharing one inter-element boundary by
1 and 2. Then, besides the volume dissipation, if any, their net contribution to the discrete
ergy $E^h(t)$ is the boundary integral of

$$J = c^2\nabla u^h_1 \cdot \mathbf{n}_1 (v^* - v^h_1) + c^2 v^h_1 (\nabla u)^* \cdot \mathbf{n}_1 + c^2\nabla u^h_2 \cdot \mathbf{n}_2 (v^* - v^h_2) + c^2 v^h_2 (\nabla u)^* \cdot \mathbf{n}_2. \quad (4.9)$$

We first introduce the so-called central flux,

$$v^* \equiv \frac{1}{2} (v^h_1 + v^h_2), \quad (\nabla u)^* \equiv \frac{1}{2} (\nabla u^h_1 + \nabla u^h_2). \quad (4.10)$$
Plug this back into (4.9) and use \( n_1 = -n_2 \). Then we have

\[
J = \frac{1}{2} \left( c^2 \nabla u_1^h \cdot n_1 (v_2^h - v_1^h) + c^2 v_1^h (\nabla u_1^h + \nabla u_2^h) \cdot n_1 \right) \\
+ \frac{1}{2} \left( c^2 \nabla u_2^h \cdot n_2 (v_1^h - v_2^h) + c^2 v_2^h (\nabla u_1^h + \nabla u_2^h) \cdot n_2 \right) = 0.
\]

Second, we propose an alternating flux,

\[
v^* \equiv v_1^h, \quad (\nabla u)^* \equiv \nabla u_2^h,
\]

or

\[
v^* \equiv v_2^h, \quad (\nabla u)^* \equiv \nabla u_1^h.
\]

Using (4.11) as an example, we have

\[
J = c^2 \nabla u_1^h \cdot n_1 (v_1^h - v_1^h) + c^2 v_1^h \nabla u_2^h \cdot n_1 + c^2 \nabla u_2^h \cdot n_2 (v_1^h - v_2^h) + c^2 v_2^h \nabla u_2^h \cdot n_2 = 0.
\]

If \( \theta = 0 \), then it is clear that both the central flux (4.10) and the alternating flux (4.11) or (4.12) lead to an energy-conserving energy-based DG scheme since \( J = 0 \). To develop an energy-dissipating scheme for \( \theta = 0 \), we introduce a Sommerfeld flux which yields \( J < 0 \) in the presence of jumps. Let us denote a flux splitting parameter by \( \xi > 0 \) which has the same units as the wave speed \( c \) and note that,

\[
v \nabla u \cdot n = \frac{1}{4\xi} (v + \xi \nabla u \cdot n)^2 - \frac{1}{4\xi} (v - \xi \nabla u \cdot n)^2.
\]

Then we enforce

\[
\begin{aligned}
v^* - \xi (\nabla u)^* \cdot n_1 &= v_1^h - \xi (\nabla u_1^h) \cdot n_1, \\
v^* - \xi (\nabla u)^* \cdot n_2 &= v_2^h - \xi (\nabla u_2^h) \cdot n_2.
\end{aligned}
\]

Solving system (4.13) yields

\[
v^* = \{\{v^h]\} - \frac{\xi}{2}[\nabla u^h], \quad (\nabla u)^* = \{\{\nabla u^h]\} - \frac{1}{2\xi}[\nabla u^h].
\]

71
Plugging (4.14) into (4.9) we obtain

\[ J = - \left( \frac{\xi c^2}{2} \left[ [\nabla u^h] \right]^2 + \frac{c^2}{2\xi} \left[ [u^h] \right]^2 \right) < 0. \]

Thus we have an energy-dissipating scheme even when \( \theta = 0 \) if the Sommerfeld flux is used.

4.2.2.2. Physical boundaries

In this section, we focus on the boundary condition,

\[ \gamma \frac{\partial u(x, t)}{\partial t} + \eta \nabla u(x, t) \cdot n = 0, \quad x \in \partial \Omega, \quad (4.15) \]

where \( \gamma^2 + \eta^2 = 1 \) and \( \gamma, \eta \geq 0 \). Note that we have a homogeneous Dirichlet boundary condition if \( \eta = 0 \) and a homogeneous Neumann boundary condition when \( \gamma = 0 \). On the one hand, multiplying (4.15) by \( \gamma \nabla u \cdot n \) gives

\[ \gamma^2 \frac{\partial u}{\partial t} \nabla u \cdot n + \gamma \eta (\nabla u \cdot n)^2 = 0, \quad (4.16) \]

on the other hand, multiplying (4.15) by \( \eta \frac{\partial u}{\partial t} \) yields,

\[ \eta \gamma (\frac{\partial u}{\partial t})^2 + \eta^2 \frac{\partial u}{\partial t} \nabla u \cdot n = 0. \quad (4.17) \]

Combining (4.16) and (4.17), from (4.4) we have

\[ \frac{dE}{dt} = - \int_{\Omega} \theta \left( \frac{\partial u}{\partial t} \right)^2 \, dx - \int_{\partial \Omega} \gamma \eta \left( \frac{\partial u}{\partial t} \right)^2 + (\nabla u \cdot n)^2 \, dS \leq 0. \]

The numerical fluxes \( v^* \) and \( (\nabla u)^* \) are chosen to be consistent with the physical boundary condition (4.15),

\[ \gamma v^* + \eta (\nabla u)^* \cdot n = 0. \]

By a similar analysis as in [3], we find that if we choose

\[ v^* = v^h - (\gamma - a\eta) \rho, \quad (\nabla u)^* = \nabla u^h - (\eta + a\gamma) \rho n, \]

with

\[ \rho = \gamma v^h + \eta (\nabla u^h) \cdot n, \]

72
then the contribution to the discrete energy from the physical boundaries is given by

\[
\left. \frac{dE_h}{dt} \right|_{\partial\Omega} = -\int_{\partial\Omega} \gamma\eta \left( (v^*)^2 + ((\nabla u)^* \cdot \mathbf{n})^2 \right) + \rho^2 ((1-a^2)\gamma\eta + a(\gamma - \eta)),
\]

which yields a nonincreasing contribution to the energy if

\[
b = (1-a^2)\gamma\eta + a(\gamma - \eta^2) \geq 0.
\]

4.2.3. Stability of the scheme

We are now ready to state the stability of the proposed energy-based DG scheme. To make the statement concise, we introduce a general formulation for the fluxes on the inter-element boundaries,

\[
v^* ≡ \alpha v_1^h + (1-\alpha)v_2^h - \tau \left[ \nabla u^h \right], \quad (\nabla u)^* ≡ (1-\alpha)\nabla u_1^h + \alpha \nabla u_2^h - \beta \left[ v^h \right],
\]

(4.18)

with \(0 \leq \alpha \leq 1\) and \(\beta, \tau \geq 0\). Note that the previous cases correspond to:

Central flux : \(\alpha = 0.5, \tau = \beta = 0\).

Alternating flux : \(\alpha = 0, \tau = \beta = 0\) or \(\alpha = 1, \tau = \beta = 0\).

Sommerfeld flux : \(\alpha = 0.5, \tau = \frac{\xi}{2}, \beta = \frac{1}{2\xi}\).

For the general flux formulation (4.18), we find that the contribution to the discrete energy from the inter-element boundaries is the boundary integral of

\[
J = -\left( \beta ||[v^h]||^2 + \tau ||[\nabla u^h]||^2 \right) \leq 0.
\]

**Theorem 4.1.** The discrete energy \(E_h(t) = \sum_j E_j^h(t)\) with \(E_j^h(t)\) defined in (4.5) satisfies

\[
\frac{dE_h}{dt} = -\sum_j \int_{\Omega_j} \theta (v^h)^2 \, dx - \sum_j \int_{F_j} \left( \beta ||[v^h]||^2 + \tau ||[\nabla u^h]||^2 \right) \, dS
\]

\[
- \int_{\partial\Omega} \gamma\eta \left( (v^*)^2 + ((\nabla u)^* \cdot \mathbf{n})^2 \right) + b\rho^2 \, dS.
\]

If the flux parameters \(\tau, \beta\) and \(b\) are non-negative, then \(E_h(t) \leq E_h(0)\).
4.3. Error estimates

To analyze the numerical error of the scheme, we define the errors by

$$e_u = u - u^h, \quad e_v = v - v^h,$$

and compare $\langle u^h, v^h \rangle$ with an arbitrary polynomial $\langle \tilde{u}^h, \tilde{v}^h \rangle$, $\tilde{u}^h \in U_h^q$, $\tilde{v}^h \in V_h^s$ with $q - 2 \leq s \leq q$. To proceed, we denote the difference

$$\tilde{e}_u = \tilde{u}^h - u^h, \quad \tilde{e}_v = \tilde{v}^h - v^h, \quad \delta_u = \tilde{u}^h - u, \quad \delta_v = \tilde{v}^h - v,$$

and the numerical error energy

$$E = \sum_j \int_{\Omega_j} \frac{1}{2} |\nabla \tilde{e}_u|^2 + \frac{1}{2} e_v^2 \, dx - \frac{1}{2} \sum_{k,j} \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j})$$

$$- \sum_{k,j} \int_0^{\tilde{e}_u(x_{k,j})} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz.$$  \hspace{1cm} (4.20)

Here we assume $\frac{f(u)}{u} < 0$. However, this restriction can be relaxed as we show in the Remark below. First, we claim that $E$ is non-negative for small errors. Note that for small $z$

$$\frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} = -z \left. \frac{d}{dw} \left( \frac{f(w)}{w} \right) \right|_{w=\tilde{u}^h(x_{k,j})-\vartheta z}, \quad \vartheta \in [0, 1].$$

Then

$$\int_0^{\tilde{e}_u(x_{k,j})} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz = O(\tilde{e}_u^2(x_{k,j})).$$

Thus

$$\left| \int_0^{\tilde{e}_u(x_{k,j})} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz \right| < -\frac{1}{2} \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}),$$

74
which guarantees the positivity of $E$. Since both the continuous solution $(u, v)$ and the numerical solution $(u^h, v^h)$ satisfy (4.7) and (4.8), we have

$$\int_{\Omega_j} c^2 \nabla \phi_u \cdot \nabla \left( \frac{\partial e_u}{\partial t} - e_v \right) \, dx + \sum_k \omega_{k,j} \phi_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial u^h}{\partial t}(x_{k,j}) - v^h(x_{k,j}) \right)$$

$$= \int_{\partial \Omega_j} c^2 \nabla \phi_u \cdot n (e_v^* - e_v) \, dS, \quad (4.21)$$

and, using (4.19),

$$\int_{\Omega_j} \phi_v \frac{\partial e_v}{\partial t} + c^2 \nabla \phi_v \cdot \nabla e_u + \theta \phi_v e_v \, dx - \sum_k \omega_{k,j} \phi_v(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right)$$

$$= \int_{\partial \Omega_j} c^2 \phi_v(\nabla e_u)^* \cdot n \, dS. \quad (4.22)$$

Now, by using the relations $e_u = \tilde{e}_u - \delta_u$, $e_v = \tilde{e}_v - \delta_v$, choosing $\phi_u = \tilde{e}_u$, $\phi_v = \tilde{e}_v$ and then summing (4.21) and (4.22), we obtain

$$\int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla (\tilde{e}_u - \delta_u) - \theta \tilde{e}_v(\tilde{e}_v - \delta_v) \, dx - \sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial u^h}{\partial t}(x_{k,j}) - v^h(x_{k,j}) \right)$$

$$- \omega_{k,j} \tilde{e}_v(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right)$$

$$+ \int_{\partial \Omega_j} c^2 \nabla \tilde{e}_u \cdot n (\tilde{e}_v^* - \delta_v^* - (\tilde{e}_v - \delta_v)) + c^2 \tilde{e}_v ((\nabla \tilde{e}_u)^* \cdot n - (\nabla \delta_u)^* \cdot n) \, dS. \quad (4.23)$$

An integration by parts in the volume integral $\int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla \delta_v \, dx$ simplifies (4.23) to

$$\int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla (\tilde{e}_u - \delta_u) + c^2 \Delta \tilde{e}_u \delta_v + \tilde{e}_u \frac{\partial \delta_v}{\partial t} + c^2 \nabla \tilde{e}_v \cdot \nabla \delta_v - \theta \tilde{e}_v(\tilde{e}_v - \delta_v) \, dx$$

$$- \sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial u^h}{\partial t}(x_{k,j}) - v^h(x_{k,j}) \right) - \omega_{k,j} \tilde{e}_v(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right)$$

$$+ \int_{\partial \Omega_j} c^2 \nabla \tilde{e}_u \cdot n (\tilde{e}_v^* - \tilde{e}_v) + c^2 \tilde{e}_v ((\nabla \tilde{e}_u)^* \cdot n - c^2 \tilde{e}_u (\nabla \delta_u)^* \cdot n) \, dS. \quad (4.24)$$
We now must choose \((\tilde{u}^h, \tilde{v}^h)\) to achieve an acceptable error. On \(\Omega_j\), we impose for all time
\(t\) and \(\forall \phi_u \in U^0_h, \forall \phi_v \in U^s_h\),
\[
\int_{\Omega_j} \nabla \phi_u \cdot \nabla \delta_u \, dx = 0, \quad \int_{\Omega_j} \phi_v \delta_v \, dx = 0, \quad \int_{\Omega_j} \delta_u \, dx = 0,
\]
then equation (4.24) yields
\[
\int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla \frac{\partial \tilde{e}_u}{\partial t} + \frac{\partial \tilde{e}_v}{\partial t} \, dx =
\]
\[-\sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial u^h(x_{k,j})}{\partial t} - v^h(x_{k,j}) \right) - \omega_{k,j} \tilde{e}_v(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right)\]
\[-\int_{\partial \Omega_j} \theta \tilde{e}_v^2 \, dx + \int_{\partial \Omega_j} c^2 \nabla \tilde{e}_u \cdot \mathbf{n} (\tilde{e}_v^* - \tilde{e}_v) + c^2 \tilde{e}_v (\nabla \tilde{e}_u)^* \cdot \mathbf{n} - c^2 \nabla \tilde{e}_u \cdot \mathbf{n} \delta_v^* - c^2 \tilde{e}_v (\nabla \delta_u)^* \cdot \mathbf{n} \, dS \]
\[= -\int_{\Omega_j} \theta \tilde{e}_v^2 \, dx + \sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right)\]
\[+ \sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial \tilde{e}_u(x_{k,j})}{\partial t} - \tilde{e}_v(x_{k,j}) - \left( \frac{\partial \delta_u(x_{k,j})}{\partial t} - \delta_v(x_{k,j}) \right) \right)\]
\[+ \int_{\partial \Omega_j} c^2 \nabla \tilde{e}_u \cdot \mathbf{n} (\tilde{e}_v^* - \tilde{e}_v) + c^2 \tilde{e}_v (\nabla \tilde{e}_u)^* \cdot \mathbf{n} - c^2 \nabla \tilde{e}_u \cdot \mathbf{n} \delta_v^* - c^2 \tilde{e}_v (\nabla \delta_u)^* \cdot \mathbf{n} \, dS. \quad (4.25)\]

First, we estimate the third term on the right-hand side of (4.25),
\[
\sum_k \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial \tilde{e}_u(x_{k,j})}{\partial t} - \tilde{e}_v(x_{k,j}) \right)\]
\[= \sum_k \omega_{k,j} \left( \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) \tilde{e}_u(x_{k,j}) \frac{\partial \tilde{e}_u(x_{k,j})}{\partial t}\]
\[+ \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u(x_{k,j}) \frac{\partial \tilde{e}_u(x_{k,j})}{\partial t} - \omega_{k,j} \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \tilde{e}_u(x_{k,j}) \tilde{e}_v(x_{k,j})\]
\[= \frac{d}{dt} \sum_k \int_0^{\tilde{u}_u(x_{k,j})} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz\]
\[-\sum_k \omega_{k,j} \frac{1}{2} \frac{d}{dt} \left( \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) \tilde{e}_u(x_{k,j}) + \omega_{k,j} \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \tilde{e}_u(x_{k,j}) \tilde{e}_v(x_{k,j})\]
\[-\sum_k \int_0^{\tilde{u}_u(x_{k,j})} \omega_{k,j} \frac{d}{dt} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j}) - z} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz\]
\[+ \frac{1}{2} \frac{d}{dt} \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}),\]
then substituting this estimate into (4.25) and recalling (4.20) we conclude:

\[
\frac{d\mathcal{E}}{dt} = \sum_j \int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla \frac{\partial \tilde{e}_u}{\partial t} + \tilde{e}_v \frac{\partial \tilde{e}_v}{\partial t} \, dx - \sum_j \frac{1}{2} \frac{d}{dt} \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}) \\
- \sum_j \frac{d}{dt} \sum_k \int_0 r_{u,k,j} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j})} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz \\
- \sum_k \omega_{k,j} \int_0 r_{u,k,j} \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial \delta_u(x_{k,j})}{\partial t} - \delta_v(x_{k,j}) \right) - \sum_j \frac{1}{2} \frac{d}{dt} \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}) \\
+ \int_0 r_{u,k,j} \omega_{k,j} \frac{d}{dt} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j})} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z - \sum_j \int_{\Omega_j} \theta \tilde{e}_v^2 \, dx \, dz \\
+ \sum_j \int_{\partial \Omega_j} c^2 \nabla \tilde{e}_u \cdot \mathbf{n}(\tilde{e}_v - \tilde{e}_v) + c^2 \tilde{e}_v(\nabla \tilde{e}_u)^* \cdot \mathbf{n} - c^2 \nabla \tilde{e}_u \cdot \mathbf{n} \delta^*_v - c^2 \tilde{e}_v(\nabla \delta_u)^* \cdot \mathbf{n} \, dS.
\]

Combining the contributions from neighboring elements in (4.26) we obtain:

\[
\frac{d\mathcal{E}}{dt} = \sum_j \int_{\Omega_j} c^2 \nabla \tilde{e}_u \cdot \nabla \frac{\partial \tilde{e}_u}{\partial t} + \tilde{e}_v \frac{\partial \tilde{e}_v}{\partial t} \, dx - \sum_j \frac{1}{2} \frac{d}{dt} \sum_k \omega_{k,j} \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}) \\
- \sum_j \frac{d}{dt} \sum_k \int_0 r_{u,k,j} \omega_{k,j} \left( \frac{f(\tilde{u}^h(x_{k,j}) - z)}{\tilde{u}^h(x_{k,j})} - \frac{f(\tilde{u}^h(x_{k,j}))}{\tilde{u}^h(x_{k,j})} \right) z \, dz \\
- \sum_k \omega_{k,j} \int_0 r_{u,k,j} \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \tilde{e}_u^2(x_{k,j}) + \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial \delta_u(x_{k,j})}{\partial t} - \delta_v(x_{k,j}) \right) \\
- \sum_k \omega_{k,j} \int_0 r_{u,k,j} \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \tilde{e}_u(x_{k,j}) \tilde{e}_v(x_{k,j}) - \omega_{k,j} \tilde{e}_v(x_{k,j}) \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \left( \frac{\partial \delta_u(x_{k,j})}{\partial t} - \delta_v(x_{k,j}) \right) \\
- \sum_k \omega_{k,j} \int_0 r_{u,k,j} \frac{d}{dt} \left( \frac{f(u^h(x_{k,j}) - z)}{u^h(x_{k,j})} - \frac{f(u^h(x_{k,j}))}{u^h(x_{k,j})} \right) z \, dz - \sum_j \int_{\Omega_j} \theta \tilde{e}_v^2 \, dx \, dz \\
- \sum_j \int_{B_j} \gamma \left( (\tilde{e}_v^*)^2 + ((\nabla \tilde{e}_u)^* \cdot \mathbf{n})^2 \right) + \beta (\tilde{e}_v + \eta \nabla \tilde{e}_u \cdot \mathbf{n})^2 + c^2 \nabla \tilde{e}_u \cdot \mathbf{n} \delta^*_v + c^2 \tilde{e}_v(\nabla \delta_u)^* \cdot \mathbf{n} \, dS \\
- \sum_j \int_{F_j} \left( \beta [||\tilde{e}_v||^2 + \tau [||\nabla \tilde{e}_u||^2] - c^2 [||\nabla \tilde{e}_u||^2 + c^2 ||\tilde{e}_v||^2] \cdot (\nabla \delta_u)^* \right) \, dS.
\]

Here, $F_j$ represents inter-element boundaries and $B_j$ represents physical boundaries. Besides, we introduce the fluxes $\delta^*_v, \nabla \delta^*_u$ built from $\delta_v, \nabla \delta_u$ according to the specification in Section 4.2.2. In what follows, $C$ is a constant independent of the solution and element diameter $h$.
for a shape-regular mesh. Denote Sobolev norms by \( ||\cdot|| \) and the associated seminorms by \( |\cdot| \). We then have the following error estimate.

**Theorem 4.2.** Let \( \bar{q} = \min(q-1, s) \), \( q-2 \leq s \leq q \), \( \frac{f(u)}{u} \leq -L \), \( L > 0 \) be smooth. Then there exist numbers \( C_0, C_1 \) depending only on \( s, q, \xi, \beta, \tau, b \), the bounds of \( \frac{df(u)}{du}, \frac{f(u)}{u}, \frac{d}{dt} \left( \frac{f(u)}{u} \right) \) and the shape regularity of the mesh, such that for smooth solutions \( u \in L^{\infty}(0, T; H^{q+2}(\Omega)), v \in L^{\infty}(0, T; H^{q+1}(\Omega)) \) time \( T \), and \( h \) sufficiently small

\[
||\nabla u(\cdot, T)||^2_{L^2(\Omega)} + ||e_v(\cdot, T)||^2_{L^2(\Omega)} \leq C_0 e^{C_1 T} \max_{t \leq T} \left[ h^{2 \bar{q}} \left( |u(\cdot, t)|_{H^{q+2}(\Omega)} + |v(\cdot, t)|_{H^{q+1}(\Omega)} \right) \right]
\]

\[
+ h^{2(s+1)} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)}^2 + |v(\cdot, t)|_{H^{s+1}(\Omega)}^2 + |u(\cdot, t)|_{H^{s+1}(\Omega)}^2 \right), \tag{4.28}
\]

where

\[
\zeta = \begin{cases} 
\bar{q}, & \beta, \tau, b \geq 0, \\
\bar{q} + \frac{1}{2}, & \beta, \tau, b > 0.
\end{cases}
\]

**Proof.** From the Bramble-Hilbert lemma (e.g., [24]), we have for \( \bar{q} = \min(q, s - 1) \)

\[
||\delta u||^2_{L^2(\Omega)} \leq Ch^{2s+2} |u(\cdot, t)|_{H^{s+1}(\Omega)}^2, \\
||\delta v||^2_{L^2(\Omega)} \leq Ch^{2s+2} |v(\cdot, t)|_{H^{s+1}(\Omega)}^2, \tag{4.29}
\]

\[
\left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)}^2 \leq Ch^{2s+2} \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)}^2.
\]

Now we estimate the nonlinear volume integrals in (4.27). By the Cauchy-Schwartz inequality, the Cauchy inequality and (4.29) we obtain:

\[
- \sum_{k,j} \omega_{k,j} \frac{1}{2} \frac{d}{dt} \left( \frac{f(\tilde{u}(x_{k,j}))}{\tilde{u}(x_{k,j})} \right) \tilde{e}_u(x_{k,j})^2 + \omega_{k,j} \frac{f(u(x_{k,j}))}{u(x_{k,j})} \tilde{e}_u(x_{k,j}) \tilde{e}_v(x_{k,j})
\]

\[
- \sum_{k,j} \omega_{k,j} \tilde{e}_u(x_{k,j}) \frac{f(u(x_{k,j}))}{u(x_{k,j})} \left( \frac{\partial \delta u(x_{k,j})}{\partial t} - \delta_v(x_{k,j}) \right)
\]

\[
- \sum_{k,j} \int_0^{\tilde{e}_u(x_{k,j})} \omega_{k,j} \frac{d}{dt} \left( \frac{f(\tilde{u}(x_{k,j}) - z)}{\tilde{u}(x_{k,j}) - z} - \frac{f(\tilde{u}(x_{k,j}))}{\tilde{u}(x_{k,j})} \right) z \, dz
\]

\[
\leq C \mathcal{E} + Ch^{s+1} \sqrt{\mathcal{E}} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)} + |v(\cdot, t)|_{H^{s+1}(\Omega)} \right), \tag{4.30}
\]

78
and
\[
\sum_{k,j} \omega_{k,j} \varepsilon_v(x_{k,j}) \left( f(u(x_{k,j})) - f(u^h(x_{k,j})) \right) \leq \sum_{k,j} C \omega_{k,j} |\varepsilon_v(x_{k,j})| |u(x_{k,j}) - u^h(x_{k,j})| \\
\leq \sum_{k,j} C \omega_{k,j} |\varepsilon_v(x_{k,j})| (|\varepsilon_u(x_{k,j})| + |\delta_u(x_{k,j})|) \leq C\mathcal{E} + Ch^{s+1} \sqrt{\mathcal{E}} |u(\cdot, t)|_{H^{s+1}(\Omega)}.
\]

Then, using (4.30)-(4.31), (4.27) is simplified to
\[
\frac{d\mathcal{E}}{dt} \leq C\mathcal{E} + Ch^{s+1} \sqrt{\mathcal{E}} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)} + |v(\cdot, t)|_{H^{s+1}(\Omega)} + |u(\cdot, t)|_{H^{s+1}(\Omega)} \right) \\
- \sum_j \int_{B_j} e^2 \nabla e_u \cdot n \delta_u^e + e^2 \varepsilon_v (\nabla \delta_u)^e \cdot n + \gamma \eta \left( (\varepsilon_v)^2 + ((\nabla e_u)^e \cdot n)^2 \right) + b(\gamma e_v + \eta \nabla e_u \cdot n)^2 dS \\
- \sum_j \int_{F_j} (\beta[[\varepsilon_u]]^2 + \tau[[\nabla e_u]]^2) - c^2[[\nabla e_u]]^2 \delta_u^e + c^2[[\varepsilon_u]](\nabla \delta_u)^e dS.
\]

Now, we only need to consider the boundary integrals. We use the same analysis as in [3] and complete the estimates for the following cases:

**Case I:** $\beta = 0$ or $\tau = 0$,
\[
\frac{d\mathcal{E}}{dt} \leq C\mathcal{E} + Ch^{s+1} \sqrt{\mathcal{E}} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)} + |v(\cdot, t)|_{H^{s+1}(\Omega)} + |u(\cdot, t)|_{H^{s+1}(\Omega)} \right) \\
+ Ch^{\frac{s}{2}} \sqrt{\mathcal{E}} \left( |u(\cdot, t)|_{H^{\frac{s+2}{2}}(\Omega)} + |v(\cdot, t)|_{H^{\frac{s+1}{2}}(\Omega)} \right).
\]

Then, combining a direct integration of (4.32) in time with the assumption $\varepsilon_u = \varepsilon_v = 0$ at $t = 0$, we obtain
\[
\sqrt{\mathcal{E}}(T) \leq C \left( e^{CT} - 1 \right) \max_{t \leq T} \left[ h^q \left( |u(\cdot, t)|_{H^{q+2}(\Omega)} + |v(\cdot, t)|_{H^{q+1}(\Omega)} \right) \\
+ h^{s+1} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)} + |v(\cdot, t)|_{H^{s+1}(\Omega)} + |u(\cdot, t)|_{H^{s+1}(\Omega)} \right) \right],
\]
\[
\text{since } \varepsilon_u = \varepsilon_u + \delta_u, \varepsilon_v = \varepsilon_v + \delta_v, \text{ then (4.28) follows from the triangle inequality and (4.33).}
\]

**Case II:** $\beta, \tau, b > 0$,
\[
\frac{d\mathcal{E}}{dt} \leq C\mathcal{E} + Ch^{s+1} \sqrt{\mathcal{E}} \left( \left| \frac{\partial u(\cdot, t)}{\partial t} \right|_{H^{s+1}(\Omega)} + |v(\cdot, t)|_{H^{s+1}(\Omega)} + |u(\cdot, t)|_{H^{s+1}(\Omega)} \right) \\
+ Ch^{\frac{s}{2}+1/2} \sqrt{\mathcal{E}} \left( |u(\cdot, t)|_{H^{\frac{s+2}{2}}(\Omega)} + |v(\cdot, t)|_{H^{\frac{s+1}{2}}(\Omega)} \right).
\]

Then again (4.28) with $\zeta = \bar{q} + \frac{1}{2}$ follows directly from an integration in time of (4.34) combined with the triangle inequality.

\[79\]
Remark: If \( \frac{f(u)}{u} \geq 0 \) for some \( u \) we may introduce a new variable \( u = e^{\alpha t}w \), \( \alpha > 0 \) and use the energy-based DG scheme to solve for \( w \). Then so long as \( \alpha^2 + \alpha \theta - \frac{f(u)}{u} \) is positive the hypotheses above are satisfied and so the energy and error estimates hold. This applies, for example, to the sine-Gordon equation. However, in our numerical experiments we solve for \( u \) rather than \( w \).

Remark: For 1-dimensional problems, we can improve the error estimate to \( h^{s+1} \) by constructing \((\bar{u}^h, \bar{v}^h)\) to make the boundary term in (4.27) vanish as in [3].

Remark: We note that the error estimate appears to be overly conservative for the problems that we consider in the numerical experiments section. There we do not observe worse than linear growth of the error in time.

4.4. Numerical experiments

In this section we present numerical experiments to evaluate the performance of our scheme. In all cases we use a standard modal formulation and use the \( L_2 \) norm in space to evaluate the error. We present the numerical experiments in both one and two dimensions. For two-dimensional problems, we consider a simple square domain and use the tensor product of the Legendre polynomials to be the basis functions. All the numerical experiments are marched in time by a fourth-order Runge-Kutta (RK4) scheme and the flux splitting parameter is chosen to be \( \xi = c = 1 \). In all experiments we choose the time step size sufficiently small to guarantee that the temporal error is dominated by the spatial error.

4.4.1. Convergence in 1D

In this section we consider the sine-Gordon equation with a dissipating term, i.e, \( \theta = 1 \). Particularly, to investigate the order of convergence of our method, we solve the problem

\[
\frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} - \sin(u) + f(x, t), \quad x \in (-20, 20), \quad t \geq 0,
\]
with a standing breather solution

solution 1: \[ u(x, t) = 4 \arctan \frac{\sqrt{0.75} \cos(0.5t)}{0.5 \cosh(\sqrt{0.75}x)}, \quad x \in (-20, 20), \quad t \geq 0. \] (4.35)

The initial conditions, boundary conditions and the external forcing \( f(x, t) \) are chosen so that (4.35) is the exact solution. We note that our theoretical results establish convergence in the energy norm, but here we investigate the convergence of the solution itself.

As seen below, in our simulations we find the convergence rate for low degrees \( q \leq 3 \) is not regular for some cases, so for comparison we also give the results for the manufactured solution

solution 2: \[ u(x, t) = e^{\sin(x-t)}, \quad x \in (-20, 20), \quad t \geq 0. \] (4.36)

The corresponding initial conditions, boundary conditions and external forcing are determined by the manufactured solution (4.36). For these two examples, we use the same space and time discretization, the only difference is the solution itself.

The discretization is performed on the computational domain \((-20, 20)\) with the element vertices \( x_j = -20 + (j - 1)h, \ j = 1, 2, \ldots, N + 1, \ h = \frac{40}{N}. \) We evolve the discretized problems until the final time \( T = 2 \) with the time step \( \Delta t = 0.075h/(2\pi). \) We present the \( L^2 \) error for \( u. \) The degrees of the approximation space for \( u^h \) are set to be \( q = (1, 2, 3, 4, 5, 6). \)

We test four different fluxes: the central flux denoted by \( C.-\text{flux}, \) the alternating flux with \( \alpha = 0, \) denoted by \( A.-\text{flux}, \) the Sommerfeld flux denoted by \( S.-\text{flux}, \) and the upwind flux in (4.18) with \( \alpha = 0, \tau = \frac{\xi}{2}, \beta = \frac{1}{2\xi} \) denoted by \( A.S.-\text{flux}. \) Note that both the \( C.-\text{flux} \) and \( A.-\text{flux} \) are energy-conserving methods; both the \( S.-\text{flux} \) and \( A.S.-\text{flux} \) are energy-dissipating methods even when \( \theta = 0. \) We want to point out that \( \alpha = 1 \) has a similar performance to \( \alpha = 0 \) in the cases \( A.-\text{flux} \) and \( A.S.-\text{flux} \); thus we only show the results for \( \alpha = 0 \) in the rest of the paper. We also consider two different approximation spaces: either \( u^h \) and \( v^h \) in the same space, i.e, \( s = q, \) or the degree of the approximation space of \( v^h \) one less than \( u^h, \) i.e, \( s = q - 1. \)

In experiments not shown, we observed that the convergence rate was somewhat irregular for all cases when \( L^2 \) projection was used to compute the initial conditions. One may use
a special projection for the initial conditions to solve this problem; see for example the approach in [22] which discusses a projection for the local DG method with the alternating flux. But here, we adopt a simpler idea as in [92]: transform the problem into one with zero initial conditions,

\[ u(x, t) = u_0(x) + \tilde{u}(x, t), \]

where \( u_0(x) = u(x, 0) \). Then we get \( u \) by numerically solving for \( \tilde{u}(x, t) \).

The \( L^2 \) errors for \( u \) and both problems one and two are presented in Tables 4.5 through 4.4. For the energy-dissipating schemes we observe that the convergence rate for \( u^h \) is predictable for both problems, optimal convergence independent of the degree of \( v^h \). For the energy-conserving schemes we see that the behavior of the convergence rate for \( u^h \) is predictable when \( q \geq 4 \) for both problems with the A.-flux, optimal convergence for both \( s = q \) and \( s = q - 1 \). For the central flux we note suboptimal convergence for \( u^h \) when \( s = q - 1, q \geq 3 \) and optimal convergence in \( u^h \) when \( s = q, q \geq 3 \). For the lower order schemes (\( q \leq 3 \)) the behavior of the convergence rate for problem 1 is unpredictable for all fluxes and approximation spaces. Generally speaking, the error levels for both problems with all fluxes are comparable, but the rates of convergence are more predictable for dissipating fluxes with high order approximation (\( q \geq 3 \)).

4.4.2. Soliton solutions of the sine-Gordon equation in 1D

In this section, we consider the sine-Gordon equation without the dissipating term, i.e.,

\[ \theta = 0, \quad \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} - \sin(u), \quad x \in (-20, 20), \quad t \geq 0. \]  

(4.37)

This equation appears in a number of physical applications and is famous for its soliton and multi-soliton solutions. Here, we focus on investigating these soliton solutions: breather soliton, kink soliton, anti-kink soliton and multi-soliton solutions: kink-kink collision, kink-antikink collision. In the numerical simulations, the number of elements is chosen to be
Table 4.1: $L^2$ errors in $u$ for problem 1 (4.35) when the A.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$, and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$s$</th>
<th>$N$</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td>1.78e-03(-)</td>
<td>4.87e-04(1.87)</td>
<td>1.98e-04(2.22)</td>
<td>1.12e-04(1.99)</td>
<td>7.12e-05(2.02)</td>
<td>4.92e-05(2.03)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>6.74e-01(-)</td>
<td>6.73e-01(0.00)</td>
<td>7.03e-01(-0.11)</td>
<td>6.96e-01(0.04)</td>
<td>6.91e-01(0.03)</td>
<td>6.87e-01(0.03)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td>3.62e-04(-)</td>
<td>2.16e-05(4.06)</td>
<td>6.18e-06(3.99)</td>
<td>8.19e-06(-0.98)</td>
<td>1.12e-06(8.99)</td>
<td>6.88e-07(2.65)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>3.23e-02(-)</td>
<td>2.58e-03(3.64)</td>
<td>2.86e-04(5.43)</td>
<td>2.98e-03(-8.14)</td>
<td>5.69e-04(7.41)</td>
<td>3.13e-04(3.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>800</td>
<td>1200</td>
<td>1600</td>
<td>2000</td>
<td>2400</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td>3.99e-03(-)</td>
<td>5.67e-05(6.14)</td>
<td>3.60e-06(3.98)</td>
<td>2.21e-07(4.03)</td>
<td>1.37e-08(4.00)</td>
<td>1.34e-09(3.36)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
<td>1.95e-03(-)</td>
<td>9.72e-05(4.33)</td>
<td>9.94e-06(3.29)</td>
<td>2.02e-06(2.30)</td>
<td>4.31e-08(5.55)</td>
<td>3.18e-09(3.76)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>100</td>
<td>120</td>
<td>140</td>
<td>160</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td></td>
<td>7.12e-06(-)</td>
<td>2.35e-06(4.96)</td>
<td>9.50e-07(4.98)</td>
<td>4.40e-07(4.99)</td>
<td>2.26e-07(4.99)</td>
<td>1.26e-07(4.99)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td>8.64e-06(-)</td>
<td>2.88e-06(4.93)</td>
<td>1.16e-06(4.97)</td>
<td>5.39e-07(4.98)</td>
<td>2.77e-07(4.98)</td>
<td>1.54e-07(4.99)</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td></td>
<td>3.88e-07(-)</td>
<td>1.01e-07(6.02)</td>
<td>3.41e-08(5.97)</td>
<td>1.36e-08(5.98)</td>
<td>6.10e-09(5.99)</td>
<td>3.01e-09(5.99)</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
<td>4.41e-07(-)</td>
<td>1.15e-07(6.02)</td>
<td>3.88e-08(5.97)</td>
<td>1.54e-08(5.98)</td>
<td>6.94e-09(5.98)</td>
<td>3.43e-09(5.99)</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td></td>
<td>2.06e-08(-)</td>
<td>4.57e-09(6.74)</td>
<td>1.28e-09(6.97)</td>
<td>4.37e-10(6.98)</td>
<td>1.72e-10(6.97)</td>
<td>7.63e-11(6.92)</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td></td>
<td>2.25e-08(-)</td>
<td>5.02e-09(6.73)</td>
<td>1.41e-09(6.97)</td>
<td>4.80e-10(6.97)</td>
<td>1.90e-10(6.96)</td>
<td>8.41e-11(6.90)</td>
</tr>
</tbody>
</table>

Table 4.2: $L^2$ errors in $u$ for problem 2 (4.36) when the A.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$. 

<table>
<thead>
<tr>
<th>$q$</th>
<th>$s$</th>
<th>$N$</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td>5.84e-01(-)</td>
<td>4.25e-01(0.46)</td>
<td>3.50e-01(0.48)</td>
<td>3.05e-01(0.48)</td>
<td>2.73e-01(0.50)</td>
<td>2.50e-01(0.49)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>3.84e+00(-)</td>
<td>3.94e+00(-0.01)</td>
<td>4.35e+00(-0.01)</td>
<td>4.36e+00(-0.01)</td>
<td>4.36e+00(0.00)</td>
<td>4.37e+00(0.00)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td>7.73e-04(-)</td>
<td>9.79e-05(2.98)</td>
<td>2.92e-05(2.98)</td>
<td>1.24e-05(2.98)</td>
<td>6.36e-06(2.99)</td>
<td>3.06e-06(2.98)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>1.17e-02(-)</td>
<td>2.92e-03(2.00)</td>
<td>1.30e-03(2.00)</td>
<td>7.30e-04(2.01)</td>
<td>4.67e-04(2.00)</td>
<td>3.24e-04(2.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td>1.46e-02(-)</td>
<td>6.81e-04(4.42)</td>
<td>4.20e-05(4.02)</td>
<td>2.29e-06(4.20)</td>
<td>1.43e-07(4.00)</td>
<td>9.39e-09(3.93)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
<td>3.71e-02(-)</td>
<td>2.74e-03(3.76)</td>
<td>1.78e-04(3.94)</td>
<td>1.12e-05(3.99)</td>
<td>7.03e-07(4.00)</td>
<td>4.39e-08(4.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>100</td>
<td>120</td>
<td>140</td>
<td>160</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td></td>
<td>8.23e-05(-)</td>
<td>2.59e-05(5.18)</td>
<td>1.06e-05(4.92)</td>
<td>4.96e-06(4.90)</td>
<td>2.41e-06(5.41)</td>
<td>1.32e-06(5.08)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td>9.51e-05(-)</td>
<td>3.05e-05(5.10)</td>
<td>1.21e-05(5.07)</td>
<td>5.55e-06(5.06)</td>
<td>2.83e-06(5.04)</td>
<td>1.57e-06(5.03)</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td></td>
<td>3.74e-06(-)</td>
<td>1.01e-06(5.88)</td>
<td>3.47e-07(5.84)</td>
<td>1.41e-07(5.86)</td>
<td>6.39e-08(5.91)</td>
<td>3.14e-08(6.03)</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
<td>4.18e-06(-)</td>
<td>1.10e-06(5.98)</td>
<td>3.69e-07(6.00)</td>
<td>1.47e-07(5.99)</td>
<td>6.57e-08(6.01)</td>
<td>3.24e-08(5.99)</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td></td>
<td>2.18e-07(-)</td>
<td>4.60e-08(6.98)</td>
<td>1.29e-08(6.99)</td>
<td>4.35e-09(7.03)</td>
<td>1.74e-09(6.88)</td>
<td>7.89e-10(6.70)</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td></td>
<td>2.31e-07(-)</td>
<td>4.91e-08(6.93)</td>
<td>1.38e-08(6.97)</td>
<td>4.71e-09(6.97)</td>
<td>1.86e-09(6.97)</td>
<td>8.12e-10(7.03)</td>
</tr>
<tr>
<td>$q$</td>
<td>$s$</td>
<td>$error$</td>
<td>$N$</td>
<td>400</td>
<td>800</td>
<td>1200</td>
<td>1600</td>
<td>2000</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1.18e-03 (–)</td>
<td>3.98e-04 (3.36)</td>
<td>7.82e-04 (–3.35)</td>
<td>2.52e-04</td>
<td>6.50e-04 (5.21)</td>
<td>2.85e-01 (–)</td>
<td>4.11e-04 (9.44)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.99e-03 (–)</td>
<td>1.06e-03 (2.23)</td>
<td>7.32e-05 (6.60)</td>
<td>4.53e-05 (1.67)</td>
<td>1.52e-05 (4.89)</td>
<td>8.98e-06 (2.89)</td>
<td>3.39e-03</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.18e-03 (–)</td>
<td>3.98e-04 (3.36)</td>
<td>7.82e-04 (–3.35)</td>
<td>2.52e-04 (5.08)</td>
<td>6.50e-04 (–5.21)</td>
<td>2.85e-01 (–)</td>
<td>4.11e-04 (9.44)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.99e-03 (–)</td>
<td>1.06e-03 (2.23)</td>
<td>7.32e-05 (6.60)</td>
<td>4.53e-05 (1.67)</td>
<td>1.52e-05 (4.89)</td>
<td>8.98e-06 (2.89)</td>
<td>3.39e-03</td>
</tr>
</tbody>
</table>

Table 4.3: $L^2$ errors in $u$ for problem 1 (4.35) when the C.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$s$</th>
<th>$error$</th>
<th>$N$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1600</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1.18e-03 (–)</td>
<td>3.98e-04 (3.36)</td>
<td>7.82e-04 (–3.35)</td>
<td>2.52e-04 (5.08)</td>
<td>6.50e-04 (–5.21)</td>
<td>2.85e-01 (–)</td>
<td>4.11e-04 (9.44)</td>
<td>3.85e-04 (0.16)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.99e-03 (–)</td>
<td>1.06e-03 (2.23)</td>
<td>7.32e-05 (6.60)</td>
<td>4.53e-05 (1.67)</td>
<td>1.52e-05 (4.89)</td>
<td>8.98e-06 (2.89)</td>
<td>3.39e-03</td>
<td>1.17e-03 (1.50)</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1.18e-03 (–)</td>
<td>3.98e-04 (3.36)</td>
<td>7.82e-04 (–3.35)</td>
<td>2.52e-04 (5.08)</td>
<td>6.50e-04 (–5.21)</td>
<td>2.85e-01 (–)</td>
<td>4.11e-04 (9.44)</td>
<td>3.85e-04 (0.16)</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.99e-03 (–)</td>
<td>1.06e-03 (2.23)</td>
<td>7.32e-05 (6.60)</td>
<td>4.53e-05 (1.67)</td>
<td>1.52e-05 (4.89)</td>
<td>8.98e-06 (2.89)</td>
<td>3.39e-03</td>
<td>1.17e-03 (1.50)</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1.18e-03 (–)</td>
<td>3.98e-04 (3.36)</td>
<td>7.82e-04 (–3.35)</td>
<td>2.52e-04 (5.08)</td>
<td>6.50e-04 (–5.21)</td>
<td>2.85e-01 (–)</td>
<td>4.11e-04 (9.44)</td>
<td>3.85e-04 (0.16)</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.99e-03 (–)</td>
<td>1.06e-03 (2.23)</td>
<td>7.32e-05 (6.60)</td>
<td>4.53e-05 (1.67)</td>
<td>1.52e-05 (4.89)</td>
<td>8.98e-06 (2.89)</td>
<td>3.39e-03</td>
<td>1.17e-03 (1.50)</td>
</tr>
</tbody>
</table>

Table 4.4: $L^2$ errors in $u$ for problem 2 (4.36) when the C.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$. 84
<table>
<thead>
<tr>
<th>q</th>
<th>s</th>
<th>error</th>
<th>N</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>8.80e-03(-)</td>
<td>6.70e-03(0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.09e-02(-)</td>
<td>5.35e-03(1.03)</td>
<td>3.63e-03(0.96)</td>
<td>2.74e-03(0.97)</td>
<td>2.20e-03(0.98)</td>
<td>1.84e-03(0.98)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>4.75e-05(-)</td>
<td>1.03e-03(-4.44)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>4.85e-05(-)</td>
<td>6.20e-05(-0.35)</td>
<td>1.85e-05(2.98)</td>
<td>2.40e-05(-0.89)</td>
<td>1.39e-06(12.76)</td>
<td>2.06e-05(-2.17)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>1.91e-02(-)</td>
<td>4.23e-05(8.81)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>9.81e-03(-)</td>
<td>3.52e-05(8.12)</td>
<td>2.06e-06(4.09)</td>
<td>1.48e-07(3.80)</td>
<td>7.12e-09(4.38)</td>
<td>4.40e-10(4.02)</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>5.14e-06(-)</td>
<td>1.66e-06(5.07)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>4.12e-06(-)</td>
<td>1.34e-06(5.05)</td>
<td>5.32e-07(5.05)</td>
<td>2.44e-07(5.05)</td>
<td>1.25e-07(5.04)</td>
<td>6.89e-08(5.04)</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>2.76e-07(-)</td>
<td>7.11e-08(6.08)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td>2.22e-07(-)</td>
<td>5.72e-08(6.07)</td>
<td>1.91e-08(6.02)</td>
<td>7.54e-09(6.02)</td>
<td>3.37e-09(6.02)</td>
<td>1.66e-09(6.02)</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>1.45e-08(-)</td>
<td>3.14e-09(6.86)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>1.18e-08(-)</td>
<td>2.50e-09(6.80)</td>
<td>7.22e-10(7.01)</td>
<td>2.45e-10(7.01)</td>
<td>9.62e-11(7.01)</td>
<td>4.22e-11(7.01)</td>
</tr>
</tbody>
</table>

Table 4.5: $L^2$ errors in $u$ for problem 1 (4.35) when the S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.

<table>
<thead>
<tr>
<th>q</th>
<th>s</th>
<th>error</th>
<th>N</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>5.22e-01(-)</td>
<td>2.76e-01(0.92)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>6.65e-01(-)</td>
<td>4.18e-01(0.67)</td>
<td>3.23e-01(0.64)</td>
<td>2.71e-01(0.61)</td>
<td>2.38e-01(0.58)</td>
<td>2.14e-01(0.58)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>8.25e-04(-)</td>
<td>1.04e-04(2.99)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>7.08e-04(-)</td>
<td>8.97e-05(2.98)</td>
<td>2.67e-05(2.99)</td>
<td>1.13e-05(2.99)</td>
<td>5.82e-06(2.97)</td>
<td>3.38e-06(2.98)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>1.42e-02(-)</td>
<td>6.62e-04(4.42)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>1.09e-02(-)</td>
<td>4.81e-04(5.45)</td>
<td>2.24e-05(4.43)</td>
<td>1.21e-06(4.21)</td>
<td>7.19e-08(4.08)</td>
<td>4.41e-09(4.03)</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>6.40e-05(-)</td>
<td>2.07e-05(5.06)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>4.35e-05(-)</td>
<td>1.39e-05(5.11)</td>
<td>5.50e-06(5.09)</td>
<td>2.52e-06(5.08)</td>
<td>1.28e-06(5.07)</td>
<td>7.05e-07(5.06)</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>3.41e-06(-)</td>
<td>8.90e-07(6.02)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td>2.23e-06(-)</td>
<td>5.77e-07(6.06)</td>
<td>1.92e-07(6.04)</td>
<td>7.56e-08(6.03)</td>
<td>3.38e-08(6.02)</td>
<td>1.67e-08(6.02)</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td>u - u^h</td>
<td></td>
<td>_{L^2}</td>
<td></td>
<td>1.95e-07(-)</td>
<td>4.09e-08(6.99)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td>1.31e-07(-)</td>
<td>2.76e-08(6.98)</td>
<td>7.71e-09(7.00)</td>
<td>2.62e-09(7.01)</td>
<td>1.03e-09(7.02)</td>
<td>4.49e-10(7.02)</td>
</tr>
</tbody>
</table>

Table 4.6: $L^2$ errors in $u$ for problem 2 (4.36) when the S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$. 
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
\textbf{q} & \textbf{s} & \textbf{N} & 400 & 800 & 1200 & 1600 & 2000 & 2400 \\
\hline
1 & 0 & \text{error} & 8.84e-03 (-) & 3.42e-03(1.03) & 2.92e-03(0.97) & 2.20e-03(0.98) & 1.77e-03(0.98) & 1.48e-03(0.98) \\
1 & 1 & & 9.14e-02(1) & 1.63e-02(2.49) & 6.90e-03(2.12) & 1.02e-02(-1.37) & 4.34e-03(3.85) & 3.64e-03(0.97) \\
2 & 0 & \text{error} & 4.32e-05(-) & 3.64e-05(0.25) & 6.45e-06(4.27) & 1.07e-05(-1.75) & 1.07e-06(10.33) & 9.55e-07(0.60) \\
2 & 1 & & 4.69e-05(-) & 5.12e-05(-0.13) & 8.11e-05(2.56) & 2.41e-05(-0.99) & 1.39e-06(12.80) & 2.03e-06(-2.09) \\
\hline
\textbf{q} & \textbf{s} & \textbf{N} & 50 & 100 & 200 & 400 & 800 & 1600 \\
\hline
3 & 0 & \text{error} & 7.43e-02(-) & 5.89e-05(10.30) & 3.94e-06(3.90) & 2.54e-07(3.95) & 1.48e-08(4.10) & 9.18e-10(4.01) \\
3 & 1 & & 1.34e-02(-) & 4.22e-05(8.31) & 2.54e-06(4.05) & 2.16e-07(3.56) & 7.42e-09(4.86) & 4.48e-10(4.05) \\
4 & 0 & \text{error} & 5.49e-06(-) & 1.75e-06(5.13) & 6.90e-07(5.10) & 3.16e-07(5.07) & 1.61e-07(5.05) & 8.88e-08(5.04) \\
4 & 1 & & 4.26e-06(-) & 1.36e-06(5.11) & 5.38e-07(5.09) & 2.46e-07(5.07) & 1.25e-07(5.06) & 6.91e-08(5.05) \\
5 & 0 & \text{error} & 3.70e-07(-) & 1.01e-07(5.79) & 3.51e-08(5.82) & 1.42e-08(5.87) & 6.46e-09(5.90) & 3.21e-09(5.93) \\
5 & 1 & & 2.31e-07(-) & 5.95e-08(6.07) & 1.98e-08(6.04) & 7.79e-09(6.04) & 3.47e-09(6.04) & 1.70e-09(6.04) \\
6 & 0 & \text{error} & 1.77e-08(-) & 3.75e-09(6.95) & 1.01e-09(7.20) & 3.35e-10(7.15) & 1.29e-10(7.12) & 5.61e-11(7.09) \\
6 & 1 & & 1.22e-08(-) & 2.63e-09(8.66) & 7.29e-10(7.04) & 2.47e-10(7.03) & 9.67e-11(7.02) & 4.23e-11(7.02) \\
\hline
\end{tabular}
\caption{\textbf{L$^2$ errors in $u$ for problem 1 (4.35) when the A.S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.}}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
\textbf{q} & \textbf{s} & \textbf{N} & 400 & 800 & 1200 & 1600 & 2000 & 2400 \\
\hline
1 & 0 & \text{error} & 5.50e-01(-) & 2.49e-01(0.92) & 1.97e-01(0.96) & 1.48e-01(0.99) & 1.18e-01(1.02) & 9.88e-02(1.00) \\
1 & 1 & & 9.92e-01(-) & 5.85e-01(0.76) & 4.29e-01(0.76) & 3.47e-01(0.74) & 2.95e-01(0.73) & 2.59e-01(0.71) \\
2 & 0 & \text{error} & 6.88e-04(-) & 8.64e-05(2.99) & 2.56e-05(3.00) & 1.08e-05(3.00) & 5.55e-06(2.98) & 3.21e-06(3.00) \\
2 & 1 & & 7.07e-04(-) & 8.95e-05(2.98) & 2.67e-05(2.98) & 1.13e-05(2.99) & 5.82e-06(2.97) & 3.38e-06(2.98) \\
3 & 0 & \text{error} & 1.37e-02(-) & 7.26e-04(4.24) & 3.58e-05(4.34) & 1.96e-06(4.19) & 1.18e-07(4.06) & 7.31e-09(4.01) \\
3 & 1 & & 1.37e-02(-) & 7.12e-04(4.26) & 3.03e-05(4.55) & 1.40e-06(4.44) & 7.54e-08(4.21) & 4.47e-09(4.07) \\
4 & 0 & \text{error} & 5.66e-05(-) & 1.73e-05(5.32) & 6.70e-06(5.19) & 3.04e-06(5.14) & 1.54e-06(5.11) & 8.43e-07(5.09) \\
4 & 1 & & 4.58e-05(-) & 1.43e-05(5.23) & 5.58e-06(5.15) & 2.54e-06(5.11) & 1.29e-06(5.09) & 7.08e-07(5.08) \\
5 & 0 & \text{error} & 2.95e-06(-) & 8.00e-07(5.80) & 2.80e-07(5.83) & 1.13e-07(5.86) & 5.16e-08(5.89) & 2.57e-08(5.91) \\
5 & 1 & & 2.36e-06(-) & 5.95e-07(6.06) & 1.97e-07(6.05) & 7.76e-08(6.04) & 3.46e-08(6.04) & 1.70e-08(6.04) \\
6 & 0 & \text{error} & 1.65e-07(-) & 3.35e-08(7.13) & 9.16e-09(7.11) & 3.07e-09(7.09) & 1.20e-09(7.07) & 5.20e-10(7.06) \\
6 & 1 & & 1.36e-07(-) & 2.82e-08(7.06) & 7.81e-09(7.05) & 2.64e-09(7.04) & 1.03e-09(7.04) & 4.50e-10(7.03) \\
\hline
\end{tabular}
\caption{\textbf{L$^2$ errors in $u$ for problem 2 (4.36) when the A.S.-flux is used. $q$ is the degree of $u^h$, $s$ is the degree of $v^h$ and $N$ is the number of the cells with uniform mesh size $h = 40/N$.}}
\end{table}
$N = 120$. We impose no-flux conditions at the computational domain boundaries,

$$\frac{\partial u}{\partial x}(-20, t) = \frac{\partial u}{\partial x}(20, t) = 0, \quad t \geq 0.$$ 

4.4.2.1. Standing breather soliton

To numerically simulate the breather soliton solution of the sine-Gordon equation (4.37), we consider the initial conditions,

$$u(x, 0) = 4 \arctan \frac{\sqrt{0.75}}{0.5 \cosh(\sqrt{0.75} x)}, \quad \partial u \partial t(x, 0) = 0, \quad x \in (-20, 20).$$

These conditions correspond to an exact standing breather soliton solution

$$u(x, t) = 4 \arctan \frac{\sqrt{0.75} \cos(0.5t)}{0.5 \cosh(\sqrt{0.75} x)}.$$  

**Time history of the numerical energy**: we first study the numerical energy of the DG approximations to the standing breather solution. As above, we consider the four different fluxes: A.-flux, C.-flux, A.S.-flux and S.-flux; we also consider the cases where both $u^h, v^h$ are in the same approximation space ($s = q$) and when the degree of the approximation space for $v^h$ is one less than $u^h$ ($s = q - 1$). The degree of the approximation space for $u$ is fixed to be $q = 4$. We evolve the numerical solution until $T = 120$ with $h = 1/3$ and use the 4-stage Runge Kutta method with $\Delta t = 0.195h/(2\pi)$.

In Figure 4.1, we present the numerical energy for the schemes with S.-flux, A.S-flux, A.-flux and C.-flux. From the left to the right are the cases where $u^h, v^h$ are in different approximation spaces, $s = q - 1$, and the same approximation space, $s = q$, respectively. Overall, we observe that the change of the numerical energy is not significant compared with the initial energy even for dissipating schemes. The A.S-flux and S.-flux produce energy dissipating schemes and they have somewhat different performance depending on $s$, but even then the energy is conserved to around 7 digits.

**The numerical standing breather soliton**: the numerical standing breather solutions are shown in Figure 4.2 and Figure 4.3. In the simulation, $u^h, v^h$ are chosen to be in
the same approximation space with $q = s = 4$ and the S.-flux is used. Figure 4.2 shows both exact and numerical breather solutions at several times, $t = 0, 45, 90, 120$ respectively. Figure 4.3 presents the space-time plot of the breather solution from $t = 0$ to $t = 120$. We find that the numerical results match well with the analytic solution.

**Time history of the $L^2$ error:** the time history of the $L^2$ errors for the standing breather soliton solution with A.-flux, C.-flux, A.S.-flux and S.-flux are plotted in Figure 4.4 for both $s = q$ and $s = q - 1$. Particularly, $q$ is set to be 4 in the numerical simulation. The top panel is for energy-conserving schemes with the A.-flux and C.-flux from the left to right. The bottom panel is for energy-dissipating schemes with the A.S.-flux and S.-flux from the left to right. The error dynamics for all schemes except for the C.-flux are quite similar to each other and for the two values of $s$ tested. For the C-flux., however, the errors display noticeably different patterns. Nonetheless, the peak errors for all eight experiments are comparable. Finally, considering that the standing breather solution is periodic in time, we note that the $L^2$ error grows linearly in time.
Figure 4.2: Plots of the standing breather with the degree of approximation space \( q = s = 4 \). The S.-flux is used in the simulation. From the top to the bottom, the left to the right, the numerical and exact breather solutions at \( t = 0, 45, 90, 120 \) are plotted.
4.4.2.2. Kink soliton and antikink soliton

For the kink soliton solution, the sine-Gordon equation (4.37) is solved with the initial condition, \( x \in (-20, 20) \),

\[
 u(x, 0) = 4 \arctan \left( \exp \left( \frac{x}{\sqrt{1 - \mu^2}} \right) \right), \quad \frac{\partial u}{\partial t}(x, 0) = -\frac{2\mu}{\sqrt{1 - \mu^2}} \text{sech} \left( \frac{x}{\sqrt{1 - \mu^2}} \right).
\]

The analytic kink solution

\[
 u(x, t) = 4 \arctan \left( \exp \left( \frac{x - \mu t}{\sqrt{1 - \mu^2}} \right) \right)
\]

is a traveling wave increasing monotonically from 0 to \( 2\pi \) as \( x \) varies from \(-\infty\) to \( \infty \). In contrast with the kink soliton (4.38), for the antikink soliton solution we solve the sine-Gordon equation (4.37) with the initial conditions, \( x \in (-20, 20) \),

\[
 u(x, 0) = 4 \arctan \left( \exp \left( -\frac{x}{\sqrt{1 - \mu^2}} \right) \right), \quad \frac{\partial u}{\partial t}(x, 0) = \frac{2\mu}{\sqrt{1 - \mu^2}} \text{sech} \left( \frac{x}{\sqrt{1 - \mu^2}} \right),
\]
Figure 4.4: Plots of the history of the $L^2$ errors for $u$, standing breather. The first row is for energy-conserving schemes, from the left to right the A.-flux and the C.-flux respectively. The second row is for energy-dissipating schemes, from the left to right the A.S.-flux and the S.-flux respectively. The degree of the approximation space for $u$ is $q = 4$ and for $v$ is $s$. 
which leads to an analytic antikink solution

\[ u(x, t) = 4 \arctan \left( \exp \left( -\frac{x - \mu t}{\sqrt{1 - \mu^2}} \right) \right). \]

(4.39)

Compared with the kink solution (4.38), the antikink soliton (4.39) is also a traveling wave solution, but the solution varies monotonically from \( 2\pi \) to \( 0 \) as \( x \) varies from \( -\infty \) to \( \infty \).

In the numerical simulation, the velocity for both the kink soliton and the antikink soliton is chosen to be \( \mu = 0.2 \). For the kink soliton, an energy-conserving scheme with the A.-flux is used, while the C.-flux is used in the simulation of the antikink soliton. We take \( u^h \) and \( v^h \) to be in different approximation spaces, i.e., \( s = q - 1 \), with \( q = 4 \). Finally, the problem is evolved with a 4-stage Runge Kutta time integrator until \( T = 80 \) with time step size \( \Delta t = 0.01 \).

The space-time plots of the kink and antikink solitons are shown in Figure 4.5. From the left to the right are the kink soliton and the antikink soliton respectively. From the left graph, we see that the kink soliton increases monotonically from \( 0 \) to \( 2\pi \) and the antikink soliton decreases from \( 2\pi \) to \( 0 \) monotonically in the right graph. Both kink and antikink solitons move from the left to the right and keep their original shape.

4.4.2.3. Kink-kink collision and kink-antikink collision

To numerically simulate the kink-kink collision, we use the superposition of two kink solitons as the initial condition for (4.37), one moves from the left to the right and the other moves from the left to the right as follows,

\[ u(x, 0) = 4 \arctan \left( \exp \left( \frac{x + 10}{\sqrt{1 - \mu^2}} \right) \right) + 4 \arctan \left( \exp \left( \frac{x - 10}{\sqrt{1 - \mu^2}} \right) \right), \quad x \in (-20, 20), \]

\[ \frac{\partial u}{\partial t}(x, 0) = -\frac{2\mu}{\sqrt{1 - \mu^2}} \text{sech} \left( \frac{x + 10}{\sqrt{1 - \mu^2}} \right) + \frac{2\mu}{\sqrt{1 - \mu^2}} \text{sech} \left( \frac{x - 10}{\sqrt{1 - \mu^2}} \right), \quad x \in (-20, 20). \]

Similarly, for the kink-antikink soliton collision we choose the superposition of a kink soliton and an antikink soliton as the initial conditions; the kink soliton moves from the left to the
Figure 4.5: From the left to the right, plots for the kink and antikink solitons respectively. The degree of approximation space for $u$ is $q = 4$ and for $v$ is $s = 3$. An energy-based DG scheme with the A.-flux is used in the simulation of kink soliton and the C.-flux is used for the simulation of antikink soliton.

right and the antikink soliton moves from the right to the left as follows,

$$u(x,0) = 4 \arctan \left( \exp \left( \frac{x + 10}{\sqrt{1 - \mu^2}} \right) \right) + 4 \arctan \left( \exp \left( - \frac{x - 10}{\sqrt{1 - \mu^2}} \right) \right), \quad x \in (-20, 20),$$

$$\frac{\partial u}{\partial t}(x,0) = -\frac{2 \mu}{\sqrt{1 - \mu^2}} \sech \left( \frac{x + 10}{\sqrt{1 - \mu^2}} \right) - \frac{2 \mu}{\sqrt{1 - \mu^2}} \sech \left( - \frac{x - 10}{\sqrt{1 - \mu^2}} \right), \quad x \in (-20, 20).$$

Note that we simply use the superposition of two kink solitons (kink and antikink solitons) to be the initial conditions rather than the analytic solution of the corresponding collisions.

The parameter $\mu$ is chosen to be 0.2 in the numerical simulation. For the kink-kink collision soliton, an energy-dissipating scheme with the A.S.-flux is used, and the S.-flux is used in the simulation of kink-antikink collision soliton. Besides, $u^h, v^h$ are assumed to be in different approximation spaces, i.e., $s = q - 1$, with $q = 4$. Finally, the problem is evolved with a 4-stage Runge-Kutta time integrator until $T = 80$ with time step size $\Delta t = 0.01$. 
The plots of the kink-kink and the kink-antikink soliton collisions are shown in Figure 4.6. In the left graph we observe that initially the two kinks move towards each other at the same speed. The kink with the profile from 0 to $2\pi$ moves from left to right and the kink with profile from $2\pi$ to $4\pi$ moves from right to left. After a certain time, they collide with each other and are immediately reflected, keeping their original shape while moving in the opposite direction. The space-time plot of the kink-antikink collision is shown in the right graph. We see that the kink and antikink solitons move towards each other at the same speed. Here the kink with profile from $2\pi$ to $4\pi$ moves left to right and the antikink with profile from $4\pi$ to $2\pi$ moves right to left. After the collision, they move away from each other with their original velocity and direction but changed profiles.

Figure 4.6: From the left to the right plots of the kink-kink and kink-antikink collisions respectively: the degree of approximation space for $u$ is $q = 4$ and for $v$ is $s = 3$. An energy-based DG scheme with A.S.-flux is used in the simulation of kink-kink collision and the S.-flux is used for the simulation of the kink-antikink collision.
4.4.3. Convergence in 2D

In this section we investigate the convergence rate of the proposed energy-based DG scheme in 2D. Specifically, we set $\theta = 0$ and $f(u) = -4u^3$, i.e,

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - 4u^3 + f_1(x, y, t), \quad (x, y) \in (0, 1) \times (0, 1), \quad t \geq 0. \quad (4.40)$$

We construct a manufactured solution

$$u(x, y, t) = \cos(2\pi x) \cos(2\pi y) \sin(2\pi t), \quad (x, y) \in (0, 1) \times (0, 1), \quad t \geq 0, \quad (4.41)$$

to solve (4.40). The initial conditions, boundary conditions and external forcing $f_1(x, y, t)$ are determined by $u$ in (4.41).

The discretization is performed with elements whose vertices are on the Cartesian grids defined by $x_i = ih, y_j = jh, i, j = 0, 1, \cdots, n$ with $h = 1/n$. We evolve the solution with the RK4 time integrator until the final time $T = 0.2$ with a time step size of $\Delta t = 0.075h/(2\pi)$. As in the 1D test, we use four different fluxes: C.-flux, A.-flux, A.S.-flux and S.-flux, but only consider the case where $u^h$ and $v^h$ are in the same approximation space, i.e, $q_x = s_x = q$ and $q_y = s_y = q$. In Figure 4.7, the $L^2$ errors for $u$ are plotted against the mesh size $h_x = h_y = h$. Table 4.9 presents the linear regression estimates of the convergence rate for $u$ based on the data in Figure 4.7. Note that we only use the ten finest grids to compute the convergence rate here. From Table 4.9, we observe the optimal convergence rate of $q + 1$

<table>
<thead>
<tr>
<th>Degree ($q$) of approx. of $u$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit with A.-flux</td>
<td>0.19</td>
<td>2.00</td>
<td>3.89</td>
<td>4.98</td>
<td>5.73</td>
<td>6.77</td>
</tr>
<tr>
<td>Rate fit with C.-flux</td>
<td>1.40</td>
<td>1.99</td>
<td>4.31</td>
<td>4.93</td>
<td>6.18</td>
<td>6.64</td>
</tr>
<tr>
<td>Rate fit with A.S.-flux</td>
<td>0.89</td>
<td>2.86</td>
<td>4.14</td>
<td>5.03</td>
<td>6.03</td>
<td>7.02</td>
</tr>
<tr>
<td>Rate fit with S.-flux</td>
<td>1.05</td>
<td>2.87</td>
<td>4.07</td>
<td>5.02</td>
<td>6.02</td>
<td>7.01</td>
</tr>
</tbody>
</table>

Table 4.9: Linear regression estimates of the convergence rate for $u$ with C.-flux, A.-flux, S.-flux and A.S.-flux for the 2D test problem. The approximation degrees for $u^h, v^h$ are $q_x = q_y = s_x = s_y = q$. 

95
Figure 4.7: The plot of $L^2$ errors for $u$ for the 2D convergence test: from left to right, top to bottom A.-flux, C.-flux, A.S.-flux and S.-flux respectively. The approximation degrees for $u^h, v^h$ are $q_x = q_y = s_x = s_y = q$.
for the A.S.-flux and the S.-flux when \( q \geq 2 \) and an order reduction by 1 compared with the optimal convergence rate for \( q = 1 \). For the A.-flux and C.-flux, we observe optimal convergence for \( q \geq 3 \), and an order reduction by 1 for \( q = 2 \). When \( q = 1 \), the A.-flux has an order reduction by 2 compared with the optimal rate and for the C.-flux an order reduction by \( \frac{1}{2} \) compared with optimal. These observations are consistent with the results in 1D.

4.4.4. Time history of the numerical energy in 2D

We compare the numerical energy for both cases with \( \theta = 0 \) and \( \theta = 1 \) in this section. Precisely, we consider,

\[
\frac{\partial^2 u}{\partial t^2} + \theta \frac{\partial u}{\partial t} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - 4u^3, \quad (x, y) \in (0, 1) \times (0, 1), \quad t > 0, \tag{4.42}
\]

with initial conditions

\[
u(x, y, 0) = -\cos(2\pi x) \cos(2\pi y), \quad \frac{\partial u}{\partial t}(x, y, 0) = \cos(2\pi x) \cos(2\pi y), \quad (x, y) \in (0, 1) \times (0, 1),
\]

and flux free physical boundary conditions, \( t > 0 \),

\[
\frac{\partial u}{\partial x}(0, y, t) = \frac{\partial u}{\partial x}(1, y, t) = 0, \quad y \in (0, 1); \quad \frac{\partial u}{\partial y}(x, 0, t) = \frac{\partial u}{\partial y}(x, 1, t) = 0, \quad x \in (0, 1).
\]

The space discretization is same as in Section 4.4.3 with \( n = 5 \). The degree of the approximation space is set to be \( q_x = q_y = s_x = s_y = 4 \). Finally, the problems are evolved with the RK4 time integrator until the final time \( T = 10 \) with time step size chosen to be \( \Delta t = 0.075 h/(2\pi) \). In Figure 4.8, the left graph shows the numerical energy evolution with four different fluxes for the problem with dissipating term, \( \theta = 1 \); the right graph presents the numerical energy evolution with four different fluxes for the problem without dissipating term, \( \theta = 0 \). We observe that for the case without dissipating term both the A.-flux and C.-flux conserve the numerical energy; both S.-flux and A.S.-flux are energy dissipating but the total dissipation is small. For the case with dissipating term, \( \theta = 1 \), the numerical energy dissipates for all fluxes, the numerical energy evolution for schemes with A.S.-flux, S.-flux,
Figure 4.8: The plots of energy for DG solutions of (4.42) with four different fluxes. For the left graph, the dissipating term is considered, i.e., $\theta = 1$; while the right graph does not contain the dissipating term, i.e., $\theta = 0$.

C.-flux and A.-flux are on top of each other and the numerical energy dissipates very fast.

4.4.5. Focusing equation

Finally, we consider a focusing problem whose energy is indefinite. Specifically, we test the problem

$$\frac{\partial^2 u}{\partial t^2} + \theta \frac{\partial u}{\partial t} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + 4u^3, \quad (x, y) \in (0, 1) \times (0, 1), \quad t > 0, \quad (4.43)$$

for both $\theta = 0$ and $\theta = 1$. We use the same initial data as in Section 4.4.4

$$u(x, y, 0) = -\cos(2\pi x) \cos(2\pi y), \quad \frac{\partial u}{\partial t}(x, y, 0) = \cos(2\pi x) \cos(2\pi y), \quad (x, y) \in (0, 1) \times (0, 1),$$

and periodic boundary conditions are imposed in both $x$ and $y$ directions with $u(0, y, t) = u(1, y, t)$, $y \in (0, 1)$ and $u(x, 0, t) = u(x, 1, t)$, $x \in (0, 1)$.

The space discretization is the same with the one in Section 4.4.3 and we set $n = 5$. The degree of the approximation space is set to be $q_x = q_y = s_x = s_y = 4$. Finally, the problems are evolved with the RK4 time integrator and the S.-flux with time step size $\Delta t = 0.075h/(2\pi)$. Figure 4.9 shows the time evolution of $u$. From the left column to the right column are for the problem (4.43) without ($\theta = 0$) and with ($\theta = 1$) the dissipating term.
Figure 4.9: Plots of $u$ for the focusing equation (4.43) at times $t = 0, 0.69, 6, 10$ with the S-flux and $q_x = q_y = s_x = s_y = q = 4$. For the left column, $\theta = 0$; for the right column, $\theta = 1$. 
respectively. On the left column, we observe that the solution $u$ seems to be approximately periodic in time; it recovers its original shape around $t = 0.69$ at first. The right column is for $\theta = 1$, we note that the solution loses its energy as time goes by; at $t = 0.69$, it has a similar shape to the case $\theta = 0$, but the amplitude of the solution is smaller.
In this chapter, we combine the energy-based discontinuous Galerkin (DG) methods with staggered grids to overcome the stiffness coming from the high order piecewise polynomial approximations. In a single dimension with periodic boundary conditions we prove bounds on the spatial operators of our method which admits a tame CFL number $\frac{c^2}{h} < 0.15$ independent of the order of the method. For problems on bounded domains and in higher dimensions we demonstrate numerically that the method is explicit and can march with large timesteps while being high order accurate in time and space.

5.1. Introduction

The discontinuous Galerkin method is spectrally convergent with the order $q$ of the approximation and is used for many practical problems. But very high order methods, say $q > 10$, are seldom used in practice. A reason for this is that the spectral radius of the spatial differential operator discretized with DG grows as $q^2/h$. This rapidly growing numerical stiffness forces the use of excessively small time steps, effectively prohibiting the use of very high order methods.

The source of this numerical stiffness is the approximation by polynomials. Heuristically this can be understood by comparing a wave $w = e^{iqx}$ and its $q$ times larger derivative $w_x = iqw$ with a polynomial $T_q(x)$ and its derivative $T_q'(x)$. To make things concrete we assume $T_q$ is a Chebyshev polynomial $T_q = \cos(q \cos^{-1}(x))$. Then $|T_q(x)| \leq 1$, $-1 \leq x \leq 1, \forall q$, as for the wave, but the derivative $|T_q'(\pm1)| = q^2$, is $q$ times larger at the edges. As this behavior at the edges is not unique to Chebyshev polynomials, but rather a fundamental property of polynomial approximations on a fixed interval, it may appear that the possibility for DG methods with large time steps is bleak. Fortunately, polynomials typically behave better at
the element center, for example for Chebyshev polynomials we have $|T_q'(0)| \leq q$, a property that can be exploited to tame the CFL condition [83].

This chapter is organized as follows. In Section 5.2, we recall the DG method in [3] and proposed a DG method with staggered grids. We investigate the operator bounds of semi-discretization DG schemes for both non-staggered and staggered grids in one dimension with periodic boundary conditions in Section 5.3. Section 5.4 presents the algorithm of local time stepping. In Section 5.5, we show some numerical experiments to verify the convergence and spectral radius of our method.

5.2. Energy Based Discontinuous Galerkin Method for the Scalar Wave Equation

We consider the scalar wave equation written as a first order system in time

\[
\frac{\partial u(x,t)}{\partial t} = v(x,t), \quad \frac{\partial v(x,t)}{\partial t} = \nabla \cdot (c^2(x)\nabla u(x,t)) + f(x,t),
\]

on the domain

\[x = (x_1, \ldots, x_d) \in S \subset \mathbb{R}^d, \ t > 0,\]

with initial conditions

\[u(x,0) = g(x), \ \frac{\partial u}{\partial t}(x,0) = v(x,0) = h(x),\]

and boundary conditions

\[\gamma \frac{\partial u}{\partial t} + \kappa (n \cdot \nabla u) = 0, \ (x_1, x_2) \in \partial S.\]

Here $c$ is the speed of sound, $n$ is the outward pointing normal and we normalize $\gamma^2 + \kappa^2 = 1$. The choice $\kappa = 0$ corresponds to a homogenous Dirichlet boundary condition on $u$; $\gamma = 0$ corresponds to a homogenous Neumann boundary and the choice $\gamma \kappa > 0$, will dissipate the energy of the system and can be thought of as a low order non-reflecting boundary condition.
5.2.1. Staggered Formulation

Let the finite element meshes
\[ \Omega = \bigcup_j \Omega_j \quad \text{and} \quad \Omega^* = \bigcup_k \Omega_k^* \]
be staggered. Each of them is a discretization of \( S \) consisting of geometry-conforming and non-overlapping (possibly curved) quadrilaterals with piecewise smooth boundaries. Here, the staggered meshes \( \Omega \) and \( \Omega^* \) satisfy:

i. both \( \Omega \) and \( \Omega^* \) are meshes for the same computational domain \( S \);

ii. vertices, edges and faces of \( \Omega \) and \( \Omega^* \) do not coincide, but they may and will intersect.

On each element \( \Omega_j \) the approximation to the displacement, \( u^h \), is the tensor product polynomials in the spaces \( (\mathbb{Q}^{q_u}(\Omega_j))^d \). And on each element \( \Omega_k^* \) the approximation to the velocity, \( v^h \), is the tensor product polynomials in the spaces \( (\mathbb{Q}^{q_v}(\Omega_k^*))^d \).

Then, the energy based discontinuous Galerkin method can be stated as follows: on each element \( \Omega_j \) and \( \Omega_k^* \), require that for all test functions \( \phi \in (\mathbb{Q}^{q_u}(\Omega_j))^d \), \( \psi \in (\mathbb{Q}^{q_v}(\Omega_k^*))^d \),

the following variational formulation holds:

\[
\int_{\Omega_j} \nabla \phi \cdot \left( \frac{\partial \nabla u^h}{\partial t} - \nabla v^h \right) = \int_{\partial \Omega_j} (\nabla \phi \cdot \vec{n}) \left( v^* - v^h \right), \tag{5.3}
\]

\[
\int_{\Omega_k^*} \psi \frac{\partial v^h}{\partial t} + \nabla \psi \cdot (c^2(x) \nabla u^h) = \int_{\partial \Omega_k^*} \psi (c^2 \nabla u) \cdot \vec{n}^* \tag{5.4}
\]

where \( v^* \) and \( (\nabla u)^* \) are numerical fluxes at the element boundaries and will be specified later. Additionally we also require

\[
\int_{\Omega_j} \frac{\partial u^h}{\partial t} - v^h = 0. \tag{5.5}
\]
5.3. Operator Bounds in One Dimension on Periodic Domains

In this section, we focus on establishing the bounds for the energy-based DG and the staggered energy-based DG spatial operator for the second-order wave equation (5.1) and (5.2). We restrict the analysis to uniform grids and periodic boundary conditions and constant coefficients. As the key ingredient to taming the CFL condition is to evaluate certain terms with derivatives only \( \text{near} \) the element centers we expect that the analysis can be extended to smoothly changing grids and to variable coefficients (we demonstrate the applicability of the analysis in the variable coefficient case in the experiments section below.) It may also be possible to extend the analysis to problems with Dirichlet or Neumann boundary conditions by using the image principle, as done for Hermite methods in [5], however, we don’t pursue this here.

5.3.0.1. Operator Bounds for the Non-staggered Scheme

Denote the broken finite element spaces

\[
U_h^q_u = \{ w : w|_{I_j} \in \mathbb{Q}^q_u(I_j) \}, \quad V_h^q_v = \{ w : w|_{I_j} \in \mathbb{Q}^q_v(I_j) \}.
\]

Consider the problem (2.3)–(2.4) in one space dimension with \( A \) constant, \( \phi_u = \phi, \phi_v = \psi \) and the external forcing \( f = 0 \) in (2.4). Then the energy-based DG scheme reads as: find \( u^h(\cdot, t) \in U_h^q_u \) and \( v^h(\cdot, t) \in V_h^q_v \) such that for any \( \phi \in U_h^q_u \) and \( \psi \in V_h^q_v \) and for all \( j \) (here the sum over \( j \) is to be interpreted based on the assumption of periodic boundary conditions)

\[
\sum_j \int_{x_j}^{x_{j+1}} \phi_x \left( \frac{\partial u_x^h}{\partial t} - v_x^h \right) dx = \sum_j \phi_x (u_x^s - v_x^h)|_{x_j}^{x_{j+1}},
\]

\[
\sum_j \int_{x_j}^{x_{j+1}} \psi \frac{\partial v_x^h}{\partial t} + \psi_x u_x^h dx = \sum_j \psi_x u_x^s|_{x_j}^{x_{j+1}},
\]

where the numerical fluxes are defined in (2.9)–(2.10) in one dimension. Particularly the coefficients \( \theta, \beta, \tau \) in (2.9)–(2.10) are independent of mesh size \( h \) in this chapter. To study the constraints on the allowable time step size as a function of the orders of approximation
and define the operator

\[
\int_\Omega \mathcal{L}(u^h, v^h)(\phi, \psi) dx = \sum_j \int_{x_j}^{x_{j+1}} \phi_x v_x^h - \psi_x u_x^h dx + \sum_j \phi_x (v^* - v^h) |_{x_j}^{x_{j+1}} + \psi u_x^* |_{x_j}^{x_{j+1}},
\]

and the operator norm

\[
\|\mathcal{L}\| \equiv \sup_{u^h, \psi \in U^q_u, v^h, \psi \in V^q_v} \frac{\int_\Omega \mathcal{L}(u^h, v^h)(\phi, \psi) dx}{\|u_x^h\|_{L^2(\Omega)} + \|v^h\|_{L^2(\Omega)}^{1/2}(\|\phi_x\|_{L^2(\Omega)} + \|\psi\|_{L^2(\Omega)}^{1/2})^{1/2}}.
\]

Assuming that a bound \(\|\mathcal{L}\| \leq R_\Delta\) can be established, time-stability for method of lines discretization combined with one-step methods follows from Kreiss-Wu theory [54].

Here \(R\) is the radius of the stability domain of the one-step method under consideration.

It is well known, [44], that one-step methods based on Taylor expansion with \(q_{\text{Taylor}} = 3, 4, 7, 8, 11, 12, 15, 16, \ldots\) terms have stability domains that include the imaginary axis and that \(R \sim q_{\text{Taylor}}\). Thus, if we can establish bounds on \(\|\mathcal{L}\|\) that grow linearly in \(q_u\) and \(q_v\) we should expect that the fully discrete method can time-march at a CFL condition of \(O(1)\) when the spatial and temporal orders are matched. In what follows we will see that such a bound can be established for the staggered method but not for the non-staggered method.

For the non-staggered method we obtain the following bound (note that we have suppressed the superscript \(h\)).

**Theorem 5.1.** Let the energy-based DG spatial operator \(\mathcal{L}\) be defined as in (5.6) and let the numerical fluxes be defined as in (2.9) and (2.10). Further, let \(C = \max\{C_1, C_2\}\), with \(C_1 \leq \sqrt{3}\) and \(C_2 \leq \sqrt{2}/2\), be a positive constant independent of \(q_u, q_v\) and \(h\), then the following estimates holds:

\[
\|\mathcal{L}\| \leq \frac{2C}{R_\Delta} \left( q_u^2 + \max\{4\pi q_u^2, 4\beta(q_v + 1)^2\} + 2(q_v + 1)q_u \right).
\]

**Proof.** By the triangle inequality we have that, for \(u, \phi \in U^q_u, v, \psi \in V^q_v\), the following inequality holds

\[
\int_\Omega \mathcal{L}(u, v)(\phi, \psi) dx \leq \sum_j \left| \int_{x_j}^{x_{j+1}} \phi_{xx} v dx \right| + \int_{x_j}^{x_{j+1}} \psi_x u_x dx + |\phi_x| |v^*| |_{x_j}^{x_{j+1}} + |\phi_x| |v^*| \big|_{x_j}^{x_{j+1}}
\]

\[
+ |\psi| |u_x^*| \big|_{x_j}^{x_{j+1}} + |\psi| |u_x^*| \big|_{x_j}^{x_{j+1}} := \Lambda_1 + \Lambda_2 + \Lambda_3
\]

105
We now bound each of the terms, starting with the volume term. By applying the Cauchy-Schwarz inequality to $\Lambda_1$, we have

$$\Lambda_1 \leq \sum_j |\phi_{xx}|_{L^2(I_j)} v_{L^2(I_j)} + |\psi_x|_{L^2(I_j)} u_{L^2(I_j)}.$$

To proceed we rewrite the numerical fluxes (2.9) and (2.10) as

$$v^*_{x_j} = \left( \theta v^+ + (1 - \theta) v^- - \tau (u_x^+ - u_x^-) \right) \bigg|_{x_j},$$

$$u^*_{x_j} = \left( (1 - \theta) u_x^+ + \theta u_x^- - \beta (v^+ - v^-) \right) \bigg|_{x_j},$$

and

$$v^*_{x_{j+1}} = \left( \theta v^+ + (1 - \theta) v^- - \tau (u_x^- - u_x^+) \right) \bigg|_{x_{j+1}},$$

$$u^*_{x_{j+1}} = \left( (1 - \theta) u_x^+ + \theta u_x^- - \beta (v^- - v^+) \right) \bigg|_{x_{j+1}},$$

so that the second term becomes

$$\Lambda_2 = \sum_j |\phi_{xx}| \left( \theta |v^+| + (1 - \theta) |v^-| + \tau |u_x^-| + \tau |u_x^+| \right) \bigg|_{x_{j+1}} + |\phi_x| \left( \theta |v^+| + (1 - \theta) |v^-| + \tau |u_x^-| + \tau |u_x^+| \right) \bigg|_{x_j}.$$

By the triangle inequality we arrive at

$$\Lambda_2 \leq \sum_j |\phi_{xx}| \left( |\theta v^+| + (1 - \theta) |v^-| + \tau |u_x^-| + \tau |u_x^+| \right) \bigg|_{x_{j+1}}$$

$$+ |\phi_x| \left( |\theta v^+| + (1 - \theta) |v^-| + \tau |u_x^-| + \tau |u_x^+| \right) \bigg|_{x_j}.$$

Similarly we find

$$\Lambda_3 = \sum_j |\psi| \left( (1 - \theta) u_x^+ + \theta u_x^- - \beta (v^- - v^+) \right) \bigg|_{x_{j+1}} + |\psi| \left( (1 - \theta) u_x^- + \theta u_x^+ - \beta (v^+ - v^-) \right) \bigg|_{x_j},$$
and

\[ \Lambda_3 \leq \sum_j |\psi| \left( (1 - \theta) |u_x^+| + \theta |u_x^-| + \beta |v^-| + \beta |v^+| \right) \bigg|_{x_{j+1}} + |\psi| \left( (1 - \theta) |u_x^-| + \theta |u_x^+| + \beta |v^-| + \beta |v^+| \right) \bigg|_{x_j}. \]

Now, by the polynomial inverse inequalities encapsulated in Lemma 1 and 2 in [70] (scaled from \([-1,1]\) to \(I_j\)), we find that

\[ \Lambda_1 \leq \sum_j 2C_1 \frac{(q_u - 1)^2}{h} \|\phi_x\|_{L^2(I_j)} \|v\|_{L^2(I_j)} + 2C_1 \frac{q_v^2}{h} \|\psi\|_{L^2(I_j)} \|u_x\|_{L^2(I_j)}, \]

\[ \Lambda_2 \leq \sum_j 2C_2^2 \left( \frac{2q_u(q_u + 1)}{h} \|\phi_x\|_{L^2(I_j)} \|v\|_{L^2(I_j)} + \frac{4\tau q_u^2}{h} \|\phi_x\|_{L^2(I_j)} \|u_x\|_{L^2(I_j)} \right), \]

and

\[ \Lambda_3 \leq \sum_j 2C_2^2 \left( \frac{2(q_v + 1)q_u}{h} \|\psi\|_{L^2(I_j)} \|u_x\|_{L^2(I_j)} + \frac{4\beta(q_v + 1)^2}{h} \|\phi_x\|_{L^2(I_j)} \|v\|_{L^2(I_j)} \right). \]

Adding up \(\Lambda_i, \ i = 1, 2, 3\), in combination with Cauchy-Schwarz, we have that (with \(C = \max\{C_1, C_2^2\}\))

\[
\int_{\Omega} \mathcal{L}(u,v)(\phi,\psi) \, dx \leq \frac{2}{h} C \left( \max\{(q_u - 1)^2, q_u^2\} + \max\{4\tau q_u^2, 4\beta(q_v + 1)^2\} + 2(q_v + 1)q_u \right) \cdot \left( \|u_x\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)} \right)^{1/2} \left( \|\phi_x\|_{L^2(\Omega)}^2 + \|\psi\|_{L^2(\Omega)}^2 \right)^{1/2},
\]

from which the result follows. \(\square\)

5.3.1. Staggered Scheme

For the staggered version of the method we introduce element centers \(\rho_j = (x_j + x_{j+1})/2\) as well as the overlapping grid composed of the elements \(I_{j+\frac{1}{2}} = [\rho_j, \rho_{j+1}]\). For the staggered method we again approximate the displacement \(u(x,t)\) by the local polynomial

\[ u_j^h(x,t) = \sum_{k=0}^{q_u} \hat{u}_k(t) \phi_k(x), \quad x \in I_j, \]

while the velocity is approximated on the staggered grid by the local polynomial

\[ \tilde{v}_j^h(x,t) = \sum_{k=0}^{q_v} \hat{v}_k(t) \psi_k(x), \quad x \in I_{j+\frac{1}{2}}. \]
Then the global approximations, \( u^h(x, t) \) and \( v^h(x, t) \) are elements of the broken finite element spaces

\[
U_{h}^{q_u} = \{ w : w|_{I_j} \in Q_{\Omega}(I_j) \}, \quad \tilde{V}_{h}^{q_v} = \{ w : w|_{I_{j+\frac{1}{2}}} \in Q_{\Omega}(I_{j+\frac{1}{2}}) \},
\]

respectively.

The staggered energy-based DG scheme then consists of finding \( u^h(\cdot, t) \in U_{h}^{q_u} \) and \( \tilde{v}^h(\cdot, t) \in \tilde{V}_{h}^{q_v} \) such that for any \( \phi \in U_{h}^{q_u} \) and \( \psi \in \tilde{V}_{h}^{q_v} \) and for all \( j \) (again we assume periodic boundary conditions)

\[
\sum_j \int_{x_j}^{x_{j+1}} \phi_x \frac{\partial u^h}{\partial t} \, dx + \int_{x_j}^{x_{j+1}} \phi_{xx} \tilde{v}^h \, dx + \int_{x_j}^{x_{j+1}} \phi_{xx} v^h \, dx = \sum_j -\phi_x v^*|_{x_j} + \phi_x v^*|^{x_{j+1}}_x, \quad (5.7)
\]

\[
\sum_j \int_{\rho_j}^{\rho_{j+1}} \psi \frac{\partial \tilde{v}^h}{\partial t} \, dx + \int_{\rho_j}^{\rho_{j+1}} \psi_{xx} u^h \, dx + \int_{\rho_j}^{\rho_{j+1}} \psi_{xx} u^h \, dx = \sum_j -\psi u^*|_{\rho_j} + \psi u^*|^{\rho_{j+1}}_{\rho_j}. \quad (5.8)
\]

Note that the second and third integrals in (5.7) are against \( v^h \) on two elements and similarly the second and third integrals in (5.8) are against \( u^h \) on two elements as well.

To connect the element solutions in a stable fashion we now use the numerical fluxes

\[
v^* = \tilde{v}^h - \tau_c(u^h_x^- - u^h_x^+), \quad (5.9)
\]

\[
u^*_x = u^h_x - \beta_c(\tilde{v}^h_x^- - \tilde{v}^h_x^+). \quad (5.10)
\]

Note here that there is no ambiguity in defining the \( v \) and the \( u_x \) part of the respective fluxes above since they are evaluated at the element center of \( \tilde{v}^h \) and \( u^h \).

Providing periodic boundary conditions, we add the equations (5.7) and (5.8) up and do integration by parts to obtain

\[
\int_\Omega \phi_x \frac{\partial u^h}{\partial t} + \psi \frac{\partial \tilde{v}^h}{\partial t} \, dx = \sum_j \int_{x_j}^{x_{j+1}} \phi_x \tilde{v}^h \, dx + \int_{\rho_j}^{\rho_{j+1}} \phi_x v^h \, dx + \sum_j \phi_x (v^* - \tilde{v}^h)|_{x_j}^{x_{j+1}} + \phi_x v^h|_{\rho_j}^{\rho_{j+1}} - \sum_j \int_{\rho_j}^{\rho_{j+1}} \psi_x u^h \, dx + \int_{x_j+1}^{x_{j+1}} \psi_x u^h \, dx + \sum_j \psi u^*_x|^{\rho_{j+1}}_{\rho_j}. \quad (5.11)
\]

Here, \( \rho_j^- \) and \( \rho_j^+ \) represent the data from \( I_{j-\frac{1}{2}} \) and \( I_{j+\frac{1}{2}} \), respectively. With the choice of
fluxes in (5.9)–(5.10) and substituting \((u^h, \tilde{v}^h)\) for \((\phi, \psi)\) in (5.11)

\[
\int_{\Omega} u^h \frac{\partial u^h}{\partial t} + \tilde{v}^h \frac{\partial \tilde{v}^h}{\partial t} \, dx = \sum_j -\tau_c(u^h_+ - u^h_-)^2|_{x_j} - \beta_c(\tilde{v}^h_+ - \tilde{v}^h_-)^2|_{\rho_j}. \tag{5.12}
\]

The semi-discrete stability of the method follows directly from (5.12) provided \(\beta_c, \tau_c \geq 0\).

5.3.1.1. Operator Bounds for the Staggered Scheme

Similarly to the non-staggered scheme. To study the constraints on the allowable time steps size as a function of the orders of approximation \(q_u\) and \(q_v\), we follow [70] and define the operator

\[
\int_{\Omega} L_c(u^h, \tilde{v}^h) (\phi, \psi) \, dx = \sum_j \int_{x_j}^{x_{j+1}} -\phi x \tilde{v}^h \, dx - \int_{\rho_j}^{x_{j+1}} \phi x \tilde{v}^h \, dx + \phi x v^*|_{x_j}^{x_{j+1}} + \sum_j \int_{\rho_j}^{x_{j+1}} -\psi x u^h \, dx - \int_{x_{j+1}}^{x_{j+1} + 1} \psi x u^h \, dx + \psi u^*|_{\rho_j}^{x_{j+1}}, \tag{5.13}
\]

and the operator norm

\[
||L_c|| := \sup_{u^h, \phi \in U_h^{q_u}, \tilde{v}^h, \psi \in V_h^{q_v}} \frac{\int_{\Omega} L_c(u^h, \tilde{v}^h) (\phi, \psi) \, dx}{(\|u^h\|_{L^2(\Omega)}^2 + \|\tilde{v}^h\|_{L^2(\Omega)}^2)^{1/2} \left(\|\phi \|_{L^2(\Omega)}^2 + \|\psi \|_{L^2(\Omega)}^2\right)^{1/2}}.
\]

**Theorem 5.2.** Let the staggered energy-based DG spatial operator \(L_c\) be defined as in (5.13) and let the numerical fluxes be defined as in (5.9) and (5.10). Further, let \(C_2 \leq \frac{\sqrt{2}}{2}, C_{3,\frac{1}{2}} \leq \frac{4\sqrt{3} + 2}{3} \) and \(C_{4,\frac{1}{2}} \leq 4\sqrt{\frac{3}{\pi}}\left(\frac{3}{4}\right)^{-\frac{1}{4}}\) be positive constants independent of \(q_u, q_v\) and \(h\), then the following estimates holds:

\[
||L_c|| \leq \frac{4\sqrt{2}}{h} C_{3,\frac{1}{2}} (q_u - 1) + \frac{2\sqrt{2}}{h} C_{3,\frac{1}{2}} q_v + \frac{4}{h} C_{4,\frac{1}{2}} \sqrt{q_u(q_u + 1)} + \frac{8}{h} C_{4,\frac{1}{2}} \max\{\tau_c q_u^2, \beta_c (q_v + 1)^2\},
\]

with \(\tau_c = O\left(\frac{1}{q_u}\right), \beta_c = O\left(\frac{1}{q_v}\right)\).

**Proof.** In the following, we suppress the superscript \(h\). Consider the small intervals \((x_j, x_{j+\frac{1}{4}}), (x_{j+\frac{1}{4}}, x_{j+\frac{1}{2}}), (x_{j+\frac{1}{2}}, x_{j+\frac{3}{4}}), (x_{j+\frac{3}{4}}, x_{j+1}), (x_{j+1}, x_{j+\frac{5}{4}}), (x_{j+\frac{5}{4}}, x_{j+\frac{3}{2}})\) and the numerical fluxes.
(5.9)–(5.10) with integration by parts, for \(u, \phi \in U_h^{q_e}, \bar{v}, \psi \in \tilde{V}_h^{q_e}\), we have that

\[
\int_{\Omega} \mathcal{L}_c (u, \bar{v})(\phi, \psi) \, dx = \sum_j \int_{x_j}^{x_{j+1}} \phi_x \bar{v}_x \, dx - \int_{x_j}^{x_{j+1}} \phi_{xx} \bar{v} \, dx - \int_{x_j}^{x_{j+1}} \phi_{xx} \bar{v} \, dx + \int_{x_j}^{x_{j+1}} \phi_{xx} \bar{v} \, dx
\]

\[
+ \sum_j \int_{x_j}^{x_{j+1}} \psi u_{xx} \, dx - \int_{x_j}^{x_{j+1}} \psi_x u_{xx} \, dx - \int_{x_j}^{x_{j+1}} \psi_{xx} u_{x} \, dx + \int_{x_j}^{x_{j+1}} \psi_{xx} u_{x} \, dx
\]

\[
+ \sum_j -\tau_c \phi_x^+ (u_x^+ - u_x^-) |_{x_j} + \tau_c \phi_x^- (u_x^+ - u_x^-) |_{x_{j+1}} + \beta_c \psi^+ (\bar{v}^+ - \bar{v}^-) |_{x_{j+\frac{3}{4}}}
\]

\[
+ \sum_j \beta_c \psi^- (\bar{v}^+ - \bar{v}^-) |_{x_{j+\frac{3}{4}}} = \psi u_{x} |_{x_{j+\frac{3}{4}}} + \psi u_{x} |_{x_{j+\frac{3}{4}}} - \phi_x \bar{v} |_{x_{j+\frac{3}{4}}} + \phi_x \bar{v} |_{x_{j+\frac{3}{4}}}.
\]

Then by triangle inequality, we have the following inequality holds

\[
\int_{\Omega} \mathcal{L}_c (u, v)(\phi, \psi) \, dx \leq \Theta_{u,1} + \Theta_{u,2} + \Theta_{u,3} + \Theta_{u,4} + \Theta_{v,1} + \Theta_{v,2} + \Theta_{v,3} + \Theta_{v,4},
\]

where

\[
\Theta_{u,1} = \sum_j \int_{x_j}^{x_{j+1}} |\phi_x \bar{v}_x| \, dx + \int_{x_{j+\frac{1}{4}}}^{x_{j+\frac{3}{4}}} |\phi_{xx} \bar{v}| \, dx,
\]

\[
\Theta_{u,2} = \sum_j \int_{x_{j+\frac{1}{4}}}^{x_{j+\frac{3}{4}}} |\phi_{xx} \bar{v}| \, dx + \int_{x_{j+\frac{1}{4}}}^{x_{j+\frac{3}{4}}} |\phi_{xx} \bar{v}| \, dx,
\]

\[
\Theta_{u,3} = \sum_j |\phi_x| |\bar{v}| |_{x_{j+\frac{3}{4}}} + |\phi_x| |\bar{v}| |_{x_{j+\frac{3}{4}}},
\]

\[
\Theta_{u,4} = \tau_c \sum_j |\phi_x^+ |u_x^+| |_{x_j} + |\phi_x^- |u_x^-| |_{x_j} + |\phi_x^- |u_x^+| |_{x_{j+1}} + |\phi_x^- |u_x^-| |_{x_{j+1}},
\]

and

\[
\Theta_{v,1} = \sum_j \int_{x_{j+\frac{1}{4}}}^{x_{j+\frac{3}{4}}} |\psi u_{xx}| \, dx + \int_{x_{j+\frac{1}{4}}}^{x_{j+\frac{3}{4}}} |\psi_x u_{x}| \, dx,
\]

\[
\Theta_{v,2} = \sum_j \int_{x_{j+1}}^{x_{j+\frac{5}{4}}} |\psi_x u_{xx}| \, dx + \int_{x_{j+1}}^{x_{j+\frac{5}{4}}} |\psi_{xx} u_{x}| \, dx,
\]

\[
\Theta_{v,3} = \sum_j |\psi| |u_x| |_{x_{j+\frac{3}{4}}} + |\psi| |u_x| |_{x_{j+\frac{3}{4}}},
\]

\[
\Theta_{v,4} = \beta_c \sum_j |\psi^+ |\bar{v}^+| |_{x_{j+\frac{3}{4}}} + |\psi^+ |\bar{v}^-| |_{x_{j+\frac{3}{4}}} + |\psi^- |\bar{v}^+| |_{x_{j+\frac{3}{4}}} + |\psi^- |\bar{v}^-| |_{x_{j+\frac{3}{4}}}.
\]

We first bound the volume integral terms in (5.14)–(5.15). By using the Cauchy-Schwarz
inequality and Lemma 3 in [70] to $\Theta_{u,1}$, we have
\[ \Theta_{u,1} \leq \sum_j \left( 2 \frac{\sqrt{2}}{h} C_{3,1} q_v \| \phi_x \|_{L^2([x_j, x_{j+\frac{1}{4}}])} \right) \| \tilde{v} \|_{L^2([x_j, x_{j+\frac{1}{4}}])} + \| \phi_x \|_{L^2([x_j, x_{j+\frac{1}{4}}])} \| \tilde{v} \|_{L^2([x_j, x_{j+\frac{1}{4}}])}. \]

Then adopting inverse inequality and Lemma 3 in [70], scaling from $[-1, 1]$ to $I_{j \frac{1}{2}} = [x_{j - \frac{1}{2}}, x_{j + \frac{1}{2}}]$ and from $[-1, 1]$ to $I_{j} = [x_{j}, x_{j+1}]$, we find that
\[ \Theta_{u,1} \leq \sum_j \left( 2 \frac{\sqrt{2}}{h} C_{3,1} q_v \| \phi_x \|_{L^2(I_{j})} \| \tilde{v} \|_{L^2(I_{j})} + \| \phi_x \|_{L^2(I_{j})} \| \tilde{v} \|_{L^2(I_{j})} \right). \]

Similarly, we obtain the bounds for $\Theta_{u,2}$, $\Theta_{v,1}$ and $\Theta_{v,2}$ as follows
\[ \Theta_{u,2} \leq \sum_j \left( 4 \frac{\sqrt{2}}{h} C_{3,1} q_v \| u_x \|_{L^2(I_{j})} \| \psi \|_{L^2(I_{j})} \right), \]

\[ \Theta_{v,1} \leq \sum_j \left( 4 \frac{\sqrt{2}}{h} C_{3,1} q_v \| u_x \|_{L^2(I_{j})} \| \psi \|_{L^2(I_{j})} \right), \]

and
\[ \Theta_{v,2} \leq \sum_j \left( 4 \frac{\sqrt{2}}{h} C_{3,1} q_v \| u_x \|_{L^2(I_{j})} \| \psi \|_{L^2(I_{j})} \right), \]

Next, we bound the boundary terms in (5.14)–(5.15). For $\Theta_{u,3}$, use inverse inequality and Lemma 4 in [70], scale from $[-1, 1]$ to $I_{j \frac{1}{2}} = [x_{j - \frac{1}{2}}, x_{j + \frac{1}{2}}]$ and $[-1, 1]$ to $I_{j} = [x_{j}, x_{j+1}]$, we arrive at
\[ \Theta_{u,3} \leq \sum_j \left( \frac{2}{h} C_{4,1} \sqrt{q_v(q_v + 1)} \right) \left( \| \phi_x \|_{L^2(I_{j})} \| \tilde{v} \|_{L^2(I_{j})} + \| \phi_x \|_{L^2(I_{j})} \| \tilde{v} \|_{L^2(I_{j})} \right), \]

similarly, we have
\[ \Theta_{v,3} \leq \sum_j \left( \frac{2}{h} C_{4,1} \sqrt{q_v(q_v + 1)} \right) \left( \| u_x \|_{L^2(I_{j})} \| \psi \|_{L^2(I_{j})} + \| u_x \|_{L^2(I_{j})} \| \psi \|_{L^2(I_{j})} \right). \]
For $\Theta_{u,4}$, using Lemma 2 in [70], scaling from $[-1, 1]$ to $I_j = [x_j, x_{j+1}]$ and $[-1, 1]$ to $I_{j+\frac{1}{2}} = [x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$, then

$$\Theta_{u,4} \leq \sum_j \frac{2}{h} C_2^2 \beta_c q_u^2 \left( 2 \| \phi_x \|_{L^2(I_j)} \| u_x \|_{L^2(I_j)} + \| \phi_x \|_{L^2(I_j)} \| u_x \|_{L^2(I_{j-1})} + \| \phi_x \|_{L^2(I_j)} \| u_x \|_{L^2(I_{j+1})} \right).$$

By a similar analysis, we have

$$\Theta_{v,4} \leq \sum_j \frac{2}{h} C_2^2 \beta_c (q_v + 1)^2 \left( 2 \| \psi \|_{L^2(I_{j+\frac{1}{2}})} \| \tilde{v} \|_{L^2(I_{j+\frac{1}{2}})} + \| \psi \|_{L^2(I_{j+\frac{1}{2}})} \| \tilde{v} \|_{L^2(I_{j-\frac{1}{2}})} + \| \psi \|_{L^2(I_{j+\frac{1}{2}})} \| \tilde{v} \|_{L^2(I_{j+\frac{3}{2}})} \right).$$

Adding all $\Theta_{u,i}$ and $\Theta_{v,i}$, $i = 1, 2, 3, 4$ together, we conclude that

$$\int_{\Omega} L_c(u, \tilde{v})(\phi, \psi)dx \leq \frac{2\sqrt{2}}{h} C_{3,\frac{1}{2}} q_v \| \phi_x \|_{L^2(\Omega)} \| \tilde{v} \|_{L^2(\Omega)} + \frac{4\sqrt{2}}{h} C_{3,\frac{1}{2}} (q_u - 1) \| \phi_x \|_{L^2(\Omega)} \| \tilde{v} \|_{L^2(\Omega)}$$

$$+ \frac{4}{h} C_{4,\frac{1}{2}} \sqrt{(q_v + 1)q_u} \| \phi_x \|_{L^2(\Omega)} \| \tilde{v} \|_{L^2(\Omega)} + \frac{8}{h} C_2^2 \tau_c q_u^2 \| \phi_x \|_{L^2(\Omega)} \| u_x \|_{L^2(\Omega)}$$

$$+ \frac{4\sqrt{2}}{h} C_{3,\frac{1}{2}} (q_u - 1) \| u_x \|_{L^2(\Omega)} \| \psi \|_{L^2(\Omega)} + \frac{2\sqrt{2}}{h} C_{3,\frac{1}{2}} q_u \| u_x \|_{L^2(\Omega)} \| \psi \|_{L^2(\Omega)}$$

$$+ \frac{4}{h} C_{4,\frac{1}{2}} \sqrt{q_u(q_v + 1)} \| \psi \|_{L^2(\Omega)} \| u_x \|_{L^2(\Omega)} + \frac{8}{h} C_2^2 \beta_c (q_v + 1)^2 \| \psi \|_{L^2(\Omega)} \| \tilde{v} \|_{L^2(\Omega)}$$

$$\leq \left( \frac{4\sqrt{2}}{h} C_{3,\frac{1}{2}} (q_u - 1) + \frac{2\sqrt{2}}{h} C_{3,\frac{1}{2}} q_u + \frac{4}{h} C_{4,\frac{1}{2}} \sqrt{q_u(q_v + 1)} + \frac{8}{h} C_2^2 \max \{ \tau_c q_u, \beta_c (q_v + 1)^2 \} \right) \left( \| u_x \|_{L^2(\Omega)} + \| \tilde{v} \|_{L^2(\Omega)} \right)^{1/2} \left( \| \phi_x \|_{L^2(\Omega)} + \| \psi \|_{L^2(\Omega)} \right)^{1/2}.$$

Finally, by Cauchy-Schwarz inequality, we get

$$\| L_c \| \leq \frac{4\sqrt{2}}{h} C_{3,\frac{1}{2}} (q_u - 1) + \frac{2\sqrt{2}}{h} C_{3,\frac{1}{2}} q_u + \frac{4}{h} C_{4,\frac{1}{2}} \sqrt{q_u(q_v + 1)} + \frac{8}{h} C_2^2 \max \{ \tau_c q_u, \beta_c (q_v + 1)^2 \},$$

from which the result follows.

**Remark** To handle a bounded domain $x \in [x_L, x_R]$ we simply move the leftmost and rightmost vertices on the dual grid to coincide with the boundary, i.e. $\rho_{-1} = x_0 = x_L$ and $\rho_N = x_N = x_R$. 

112
Then, for example, at the left most (half) elements we have the variational form
\[
\int_{x_0}^{x_1} \phi_x \frac{\partial u^h}{\partial t} - \phi_x \tilde{v}^h \, dx = \left[ \phi_x (v^* - \tilde{v}^h) \right]_{x_0}^{x_1},
\]
\[
\int_{x_0}^{x_1} \psi_x \frac{\partial \tilde{v}^h}{\partial t} + \psi_x u^h \, dx = \left[ \psi w^* \right]_{x_0}^{x_1}.
\]
Focusing on the contribution at the boundary we find after replacing \( \phi = u^h \) and \( \psi = \tilde{v}^h \):
\[-u^h_x (v^* - \tilde{v}^h) - \tilde{v}^h w^*.
\]
For homogenous Dirichlet conditions we can choose \( v^* = 0, w^* = u^h_x \) and for homogenous Neumann conditions we can choose \( v^* = \tilde{v}^h, w^* = 0 \), both yield zero contribution to the change in the energy.

5.4. Local Timestepping

In this section, we present the time integrator which is used in this chapter. To implement the schemes we proposed, we need to solve ordinary systems which have the following form
\[
\frac{dw}{dt} = Aw, \tag{5.16}
\]
where \( A \) is the matrix from spatial discretization and \( w \) is a vector containing the degrees of freedom on the corresponding DG element. First, let’s recall the \( q_t \)-th order Taylor time integrator which is used to solve (5.16): Given the value of \( w \) at time \( t = t_n \), the value of \( w \) at \( t = t_n + h_t \) is obtained by
\[
w(t_n + h_t) \approx w(t_n) + \frac{dw(t_n)}{dt} h_t + \frac{d^2 w(t_n)}{dt^2} \frac{h_t^2}{2!} + \cdots + \frac{d^q w(t_n)}{dt^q} \frac{h_t^q}{q!} \tag{5.17}
\]
Here, the values of the time derivatives of \( w \) at \( t = t_n \) are obtained by (5.16). For example, \( \frac{d^2 w(t_n)}{dt^2} = A \frac{dw(t_n)}{dt} \). To state the local time stepping, we need to introduce the concept of number of gaps (NOG):

Suppose we have DG elements \( K_{i_1, i_2, \ldots, i_d}, i_k \in [1, M_k], k = 1, 2, \ldots, d \) in a \( d \)-dimensional problem. Then \( \text{NOG} = m \) means that the DG elements \( K_{i_1, i_2, \ldots, i_d} \) with \( i_1 \in [1, M_1], \ldots, i_{j-1} \in [1, M_{j-1}] \),
Now we are ready to state the local time stepping we are using:

**Algorithm 1** Local time stepping (one step evolve from $t = t_n$ to $t = t + \Delta t$)

- Give the value of $w$ at $t = t_n$, say $w(t_n)$

- Divide the time step $\Delta t$ into $p$ sub-steps with $\Delta t = p\delta t$

- use (5.16)–(5.17) to evolve solution from $w(t_n)$ to $w(t_n + \delta t)$ and store all derivatives $\frac{d^jw(t_n)}{dt^j}, j = 1, 2, \cdots, p$

- Loop $i = 2, \cdots, p$
  - for the DG elements which are close to physical boundaries (NOG = $m$), use (5.16)–(5.17) to evolve solution from $w(t_n + (i - 1)\delta t)$ to $w(t_n + i\delta t)$
  - for the DG elements which are left, use the formula
    
    \[
    w(t_n + i\delta t) = w(t_n + (i - 1)\delta t) + \sum_{k=1}^{p} \left( \sum_{j=k}^{p} \frac{d^jw(t_n)}{dt^j} \frac{((i - 1)\delta t)^{j-k}}{(j-k)!} \right) \frac{((i - 1)\delta t)^k}{k!}.
    \]

The values of $q_t$, $p$ and $m$ will be specified later based on different problems.

5.5. Numerical Experiments

In this section, we present some numerical results to investigate the convergence of our method with staggered grids in $L^2$ norm. In all cases we use a modal formulation associated with the tensor product of Legendre polynomial basis. For all tests, we use central fluxes with $\tau_e, \beta_e = 0$. The spectral radius for both semi-discretization and full-discretization are also investigated.
5.5.1. Convergence in one dimension with periodic boundary conditions

Here we evolve the exact solution

\[
\begin{align*}
    u(x,t) &= \sin(\omega(x + t)), \\
    v(x,t) &= \omega \cos(\omega(x + t)),
\end{align*}
\]

on the periodic domain \( x \in [-1,1] \) until \( T = 2.2 \). The sound wave speed is set to be \( c = 1 \). We discretize using the staggered scheme with \( q_u = 2, 3, 6, 7, 10, 11, 14, 16, 18, 19, 22, 23, 26, 27 \) and \( q_v = q_u - 1 \). In order to make it possible to observe the rates of convergence we set \( \omega = 2q_u\pi \) for and \( \omega = 4q_u\pi \) for \( q_u = 2,6,10,14,16,18,22,26 \) and \( q_u = 3,7,11,19,23,27 \), respectively.

In time we use Taylor series time stepping with \( q_u \) terms (the stability domain of all of these Taylor series methods contain the imaginary axis) and throughout we keep the ratio \( \Delta t/h = 0.1 \). The \( L^2 \)-errors in the solution \( u^h \) as a function of the element size \( h \) are displayed in Figure 5.1. As can be seen from the figure the rates of convergence (as indicated by the dashed lines) appear to be optimal, i.e. \( q_u + 1 \), when \( q_u = 3,7,\ldots \) and suboptimal by one, i.e. \( q_u \) when \( q_u = 2,6,\ldots \). This is consistent with the analysis and numerical experiments for the non-staggered scheme; see [3].

![Figure 5.1](image_url)

Figure 5.1: To the left are the \( L^2 \)-errors in \( u^h \) for \( q_u = 2,6,\ldots \) and to the right for \( q_u = 3,7,\ldots \). The dashed lines are indicates rates of convergence 2,6,10,\ldots to the left and 4,8,12,\ldots to the right.
5.5.2. Spectral radii in one dimension

In this section, we explore the spectral radii of the proposed staggered scheme in Section 5.3.1. Specifically, we verify that the largest spectral radius is proportional to $q_u$ with periodic boundary conditions. And we also investigate the stability of the full discretization with Dirichlet and Neumann boundary conditions and local time stepping proposed in Section 5.4 by investigating the spectral radii of the full discretization matrix. Finally, the sound wave speed $c$ is set to be 1 for this section.

5.5.2.1. Spectral radii of semi-discretization operator

Here, the computational domain is chosen to be $[-1, 1]$. Assume we have a periodic boundary condition. The discretization is performed on a staggered uniform mesh with mesh size $h$. We first write the semi-discretization as a system of ordinary differential equations

$$\frac{dW}{dt} = AW,$$

where $W$ is a vector containing the modes describing the element-wise expansions of the displacement and the velocity. Practically speaking we extract $A$ one column at a time by computing the time derivative resulting from setting a single mode in the projection of the initial data to be 1 and all other to be zero. With purely central fluxes, the eigenvalues of $A$ will be imaginary and based on the estimates on the operator norm of $L_c$ in Section 5.3.1 we expect them to grow linearly with $q_u$ (here we set $q_v = q_a - 1$).

In Figure 5.2 we plot the spectral radii of the matrix $A$, i.e. the eigenvalue of $A$ with the largest magnitude (denoted by $\lambda_\infty$), scaled by $(h/q_u)$ for three different element sizes $h = 2/5, 2/10, 2/20$. As can be seen the growth of the spectral radii appears to be asymptotically linear in $q_u$ (i.e. constant when scaled by $q_u^{-1}$), which verifies the theoretic finding in Section 5.3.1.

5.5.2.2. Spectral radii of full discretization operator

In this section, the computational domain is chosen to be $[-1, 1.5]$. And we have a
Figure 5.2: The figure displays the spectral radii of the time stepping matrix $A$ scaled by the element size $h$ and the reciprocal of $q_u$ as a function of $q_u$.

homogeneous Neumann boundary condition at the left boundary $x = -1$ and a homogeneous Dirichlet boundary condition at the right boundary $x = 1.5$. The discretization is performed on a staggered uniform mesh with mesh size $h$. We then can update the solution by a full discretization matrix $B$ as

$$W(t_{n+1}) = BW(t_n).$$

Here, $W$ is a vector containing the modes describing the element-wise expansions of the displacement and the velocity as in Section 5.5.2.1. The full discretization matrix $B$ is obtained by a similar way as we obtain semi-discretization matrix $A$ in Section 5.5.2.1, that is, we extract $B$ one column at a time by computing the solution at the next step resulting from setting a single mode in the projection of the initial data to be 1 and all others to be zero with the local time stepping proposed in Section 5.4. To have a stable scheme, the modulus of eigenvalues ($\lambda$) of $B$ should be less or equal to 1 or at least not too larger than 1, say $\max |\lambda| \leq 1 + C1e^{-7}$, $C \in [0, 10)$.

For the local time stepping proposed in Section 5.4, there are three factors $(q_t, p, m)$ we need to specify when it is used. For all the simulations in this chapter, when the local time stepping is used, the order of Taylor series $q_t$ and the number of subcells $p$ are chosen based on the degree of the approximation space for $u, q_u$; we always set $q_t = p = q_u + 1$. We are
next going to investigate the suitable value for NOG, i.e., \( m \) to guarantee a stable scheme.

We first fix the number of DG elements for \( u \) to be 10, i.e., \( h = 2.5/10 \), then the number of DG elements for \( v \) is 11 under the framework of staggered grids, see the Remark in Section 5.3.1.1. The degrees of the approximation spaces for \( u \) and \( v \) are chosen to be \( q_u = 14 \) and \( q_v = 13 \), respectively. The ratio between time step size \( \Delta t \) and the mesh size \( h \) are fixed as \( \frac{\Delta t}{h} = 0.1 \).

In Figure 5.3, on the left, we show spectral radius for the full discretization with NOG = 2, we observe that the largest eigenvalue’s modulus is larger than 1 up to \( 6e - 4 \). On the right, we present the spectral radius for the full discretization with NOG = 3, and we find that the largest eigenvalue’s modulus is larger than 1 up to \( C1e - 7 \). Thus, we conclude that NOG = 3 gives a stable scheme when the number of DG elements for \( u \) equals 10.

![Figure 5.3](image_url)

Figure 5.3: To the left are the spectrum for \( q_u = 14 \) and \( q_v = 13 \) with NOG = 2 and to the right are the spectrum for \( q_u = 14 \) with NOG = 3.

Next, we show that we actually can fix the the value of NOG for different number of DG elements for \( u \). In Figure 5.4, we fix the NOG = 3 and increase the number of DG elements for \( u \) from 20 to 80. We observe that the largest eigenvalue’s modulus is larger than 1 up to \( C1e - 7 \) which is same with the case in Figure 5.3 when the number of DG elements for \( u \) equals 10. Thus, our scheme is stable with a small value of NOG.

5.5.3. Convergence in two dimensions with Dirichlet boundary condition

In this section, we investigate the convergence of the staggered energy based DG scheme
with the local time stepping in Section 5.4 and variable sound wave speed $c(x,y)$ in two dimensions, solving

$$\frac{\partial^2 u}{\partial t^2} = \nabla \cdot (c^2(x,y)\nabla u) + f(x,y,t), \quad (x,y) \in [-1,1] \times [-1,1], \quad t > 0,$$

where $c(x,y) = 1 + x^2 + y^2$. Further, we construct manufactured solutions

$$u(x,y,t) = \sin(\sqrt{k_1^2 + k_2^2}\pi t) \sin(k_1\pi x) \sin(k_2\pi y),$$

$$v(x,y,t) = \sqrt{k_1^2 + k_2^2}\pi \cos(\sqrt{k_1^2 + k_2^2}\pi t) \sin(k_1\pi x) \sin(k_2\pi y).$$

Then the initial condition, boundary condition and the external forcing function $f(x,y,t)$ are determined by this manufactured solution. To observe the desired convergence order, in the numerical simulation, we set $k_1 = k_2 = q = 2$ for $q_u = q_v = 2, 3$ and $k_1 = k_2 = 2q$ for $q_u = q_v = q = 6, 7$ with $q$ is the degree of the approximation space for both $u$ and $v$ in both direction $x$ and direction $y$. Clearly, we have a homogeneous Dirichlet boundary condition for these choices.

The discretization is performed with staggered elements. The vertices for the DG elements of $u$ are on the Cartesian grids defined by $x_i = ih, y_j = jh, i, j = 0, 1, \cdots, n$ with $h = 2/n$. The elements for $v$ are staggered with the DG elements for $u$ followed the rule of the Remark in Section 5.3.1.1. Then we have $n^2$ DG elements for $u$ and $(n + 1)^2$ DG elements for $v$. Figure 5.5 gives an illustration of the staggered grids with $n = 3$. Finally,
we evolve the solution by the local time stepping proposed in Section 5.4 with \( p = q_t = q + 1 \) and NOG = 3 until final time \( T = 0.5 \), and the ratio of time step size \( \Delta t \) and mesh size \( h \) are set to be \( \frac{\Delta t}{h} = 0.1 \).

The \( L^2 \) errors for \( u \) are plotted against the mesh size \( h \) in Figure 5.6. Table 5.1 presents the linear regression estimates of the convergence rate for \( u \) based on the data in Figure 5.6. From Table 5.1, we observe an optimal convergence rate of \( q + 1 \) when \( q = 3, 6, 7 \) and a suboptimal convergence by one for \( q = 2 \).

<table>
<thead>
<tr>
<th>Degree (( q )) of approx. of ( u )</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit with C.-flux</td>
<td>2.00</td>
<td>4.27</td>
<td>7.21</td>
<td>8.13</td>
</tr>
</tbody>
</table>

Table 5.1: Linear regression estimates of the convergence rate for \( u \) with central flux in two dimensions. The degree of the approximation space for \( u \) and \( v \) are \( q \) for both \( x \) and \( y \) directions.
5.5.4. Spectral radii in two dimensions

In this section, we investigate the stability of the full discretization with local time stepping proposed in Section 5.4 in two dimensions with homogeneous Dirichlet boundary conditions. The computational domain is chosen to be the same as in Section 5.5.3. And the spatial discretization is also the same as in 5.5.3; here, we set $n = 10$. Then, the number of DG elements for $u$ is 100 and for $v$ is 121. The degree of the approximation space for $u$ and $v$ are set to be $q$ for both $x$ and $y$ directions, and $p, q_t$ in the local time stepping are set to be $p = q_t = q + 1$ and NOG = 3.

In Figure 5.7, we present the spectral radius of the full discretization with different values of $p$. The top panel is for $q = 2, 3$ from the left to the right, respectively. And the bottom panel if for $q = 6, 7$ from the left to the right, respectively. We observe that the spectral radius is strictly less than 1 for all $q = 2, 3, 6, 7$. In other words, our scheme is stable.
Figure 5.7: The top panel, from the left to the right are the spectral radii when $q = 2, 3$, respectively. The bottom panel, from the left to the right are the spectral radii when $q = 6, 7$, respectively.
Chapter 6

GALERKIN DIFFERENCE BASIS

In this chapter, we investigate a new class of difference methods based on the Galerkin formulation for the scalar second-order wave equation. The method combines the energy-based discontinuous Galerkin (DG) method with Galerkin difference basis functions. The method reduces the computational cost of the inversion of stiffness matrix from $O(N^2)$ to $O(N)$ with $N$ to be the degree of freedom (DOF) in one DG element by introducing a new class of basis functions which are generalized eigenvectors of the stiffness matrix. The resulting method also overcomes the typical numerical stiffness associated with high order piecewise polynomial approximations and possesses nice dispersion properties.

6.1. Introduction

As shown in Chapter 5, the spectral radii of the resulting semi-discretization for the second-order wave equation with energy-based discontinuous Galerkin methods are proportional to $p^2$ with $p$ to be the approximation order of the piecewise basis functions. This forces us to use a very small time step size to guarantee the stability of the method for high order spatial approximations, which is not computationally efficient in practice. But it is well known that the high order numerical schemes ($q$ is higher than fourth order) can be orders of magnitude faster, more efficient and have much smaller dispersion error than lower order schemes [45,53]. Thus, it is critical to overcome the numerical stiffness of the proposed scheme in [3]. The reason for this numerical stiffness is the behavior of the polynomial basis functions at the boundaries of a fixed interval. One can refer to Chapter 5 for a detailed explanation. In Chapter 5, we combine staggered grids with an energy-based DG scheme to overcome the stiffness from the boundaries. Here, we modify the polynomial basis functions.
to improve the numerical efficiency, allowing larger time step sizes. Particularly, we use the
galerkin difference basis functions proposed in [12] which admit the following properties:

i. obtain high order accuracy by including neighboring DOFs rather than introducing extra
   interior DOFs;

ii. impose only low order continuity at element boundaries [52];

iii. resemble a compact finite difference scheme away from boundaries [57].

This chapter is organized as follows. In Section 6.2, we recall the Galerkin difference basis
functions proposed in [12] and the energy-based DG method developed in [3]. We introduce
a new class of basis functions which are simply the generalized eigenvalues of the stiffness
matrix in Section 6.3. We also analyze and verify that the corresponding computational cost
is $O(N)$ with $N$ being the number of degrees of freedom on one DG element. Section 6.4
presents the dispersion analysis of the energy-based DG method with Galerkin difference
basis functions. We show the spectral radius of the semi-discretization formulation and
observe it is proportional to the approximation order $p$ in Section 6.5. Finally, in Section
6.6, we show numerical experiments that verify the convergence of our method.

6.2. Preliminaries

We consider the wave equation in the first order form in time and second order form in
space

$$u_t = v,$$

$$v_t = \nabla \cdot (c^2 \nabla u) + f, \quad t > 0, \quad (x, y, z) \in \Omega$$

with suitable initial and boundary conditions described later.

In what follows we will consider energy based DG methods implemented on $d$-dimensional
tensor product elements. Each element will be mapped to the $d$-dimensional unit cube which
will be discretized by an equidistant Cartesian grid. Precisely, in each dimension we consider
a grid

\[ x_i = ih, \ i = 0, \ldots, N, \ h = 1/N. \]

To this end we will use \( N \) to denote the number of intervals, \( n \) to denote the number of elements in one dimension, \( m \) to denote the number of elements in multiple dimensions and \( p \) to be the polynomial degree of the approximation. Also, we will only consider \( p \) odd and will often use the integer \( q = (p + 1)/2 \).

6.2.1. The Galerkin Difference Basis

We now describe the Galerkin difference (GD) basis we will use. Note that the description here is slightly different compared to the original description in [12] but the basis is identical.

We first consider the case far away from a boundary. The goal is to construct a polynomial basis of odd degree \( p \) on an equidistant grid with grid spacing \( h \). The generating basis function \( \Phi_p(x) \) is centered around \( x = 0 \) and the basis itself is simply a translation of the generating basis function. Thus an element in the basis centered around \( x_i = ih \) becomes, \( \phi_{p,i} = \Phi_p(x - ih) \). The generating basis function \( \Phi_p(x) \) is symmetric, \( \Phi_p(x) = \Phi_p(-x) \), and has compact support on \( x \in [-qh, qh] \), where \( p = 2q - 1 \) (recall that \( p \) is odd). In Figure

![Figure 6.1](image-url)

Figure 6.1: To the left, the generating basis function \( \Phi_p(x) \) for \( p = 1, 3, 5, 7, 9 \). Note that \( \Phi_p(x) = \Phi_p(-x) \). Note that each function is vertically offset by \( q - 1 \). In the middle, the “right” four degree 7 Lagrange polynomials that make up \( \Phi_7(x) \), the part of \( L_j \) that is used is in bold. To the right, a three dimensional representation of the generating basis function \( \Phi_p(x) \) for \( p \) up to 39.
6.1 we display the non-zero part of \( \Phi_p(x) \) for \( p = 1, 3, 5, 7, 9 \), and \( x > 0 \). In the lower left corner, where \( p = 1 \), we recognize the classic finite element hat function and as \( p \) increases we see that \( \Phi_p(x) \) becomes increasingly similar to the Cardinal Sinc function.

Now, an explicit formula for \( \Phi_p(x) \) inside each of the \( q \) positive intervals \([jh, (j+1)h)\) is obtained as follows. Let \( L_j(x) \) be the Lagrange interpolating polynomial on the grid \( G_j = \{-jh, \ldots, (p-j)h\} \) with the property that \( L_j(0) = 1 \), then

\[
\Phi_p(x) = L_j(x), \quad x \in [(q-j-1)h, (q-j)h).
\]

A continuous function \( u(x,t) \) can then be approximated by a linear combination of nodal values

\[
u(x,t) \approx \sum_{i=k-(q-1)}^{k+q} u_i \phi_{p,i}(x), \quad x \in [kh, (k+1)h),
\]

where \( u_i = u(x_i,t) \).

6.2.1.1. Modification Near Boundaries

Near boundaries the basis must be modified. In [12], three approaches for handling boundaries are described: ghost basis, extrapolation basis and use of modified equations. Here we will exclusively use the extrapolation basis, which we describe next.

The extrapolation procedure draws from the standard practice to use ghost points in finite difference methods. First, note that the \( q - 1 \) additional ghost basis functions associated with the \( q - 1 \) first grid points outside the computational domain are the only ghost basis with support inside the computational domain. In the ghost basis approach the degrees of freedom at the ghost points are retained as unknowns but in the extrapolation approach they are eliminated in favor of modifying the basis itself near the boundary.

As the name suggests, the elimination is done by extrapolating the nodal values inside the computational domain to the ghost points. For example consider \( p = 3 \). Then \( q - 1 = 1 \) and one ghost point value, \( u_{-1} \), must be determined. As the basis is fourth order accurate,
the ghost point value is determined by fourth order accurate extrapolation,

\[ u_{-1} = 4u_0 - 6u_1 + 4u_2 - u_3. \]

To understand how the modified basis is constructed, consider evaluating the approximation \( u(x) \) inside the computational domain where the ghost basis has support. In this case this means \( x \in (x_0, x_1) \) and the approximation is

\[ u(x) = u_{-1}\phi_{-1}(x) + u_0\phi_0(x) + u_1\phi_1(x) + u_2\phi_2(x). \]

To obtain a value for \( u_{-1} \) we use the extrapolation condition

\[
\begin{align*}
  u(x) &= (4u_0 - 6u_1 + 4u_2 - u_3)\phi_{-1}(x) + u_0\phi_0(x) + u_1\phi_1(x) + u_2\phi_2(x) \\
  &= u_0[\phi_0(x) + 4\phi_{-1}(x)] + u_1[\phi_1(x) - 6\phi_{-1}(x)] + u_2[\phi_2(x) + 4\phi_{-1}(x)] + u_3[-\phi_{-1}(x)] \\
  &= u_0[\phi_0(x) + 4\phi_{-1}(x)] + u_1[\phi_1(x) - 6\phi_{-1}(x)] + u_2[\phi_2(x) + 4\phi_{-1}(x)] \\
  &+ u_3[\phi_3(x) - \phi_{-1}(x)].
\end{align*}
\]

In the last step, we used the fact the support of \( \phi_3 \) vanish in \((x_0, x_1)\). Thus the modified basis functions are

\[ \tilde{\phi}_0 = \phi_0 + 4\phi_{-1}, \quad \tilde{\phi}_1 = \phi_1 - 6\phi_{-1}, \quad \tilde{\phi}_2 = \phi_2 + 4\phi_{-1}, \quad \tilde{\phi}_3 = \phi_3 - \phi_{-1}. \]

The extension to larger \( p \) requires the basis to be modified in a wider band near the boundaries and the extrapolation is done at the order matching the interior scheme. The handling of the right boundary is analogous.

6.2.1.2. Extension to Higher Dimensions

The extension to higher dimensions simply amounts to the tensor product approximation built off the one dimensional basis. For example, in two dimensions we have

\[
 u(x, y, t) \approx \sum_{i=k_x-(q-1)}^{k_x+q} \sum_{j=k_y-(q-1)}^{k_y+q} u_{i,j}\tilde{\phi}_{p,i,j}(x, y),
\]

127
for \((x, y) \in [k_x h, (k_x + 1) h) \times [k_y h, (k_y + 1) h]\) with \(u_{i,j} = u(x_i, y_j, t)\) and
\[
\phi_{p,i,j}(x, y) = \begin{cases} 
\phi_{p,i}(x)\phi_{p,j}(y), & p < i < N - p, \quad p < j < N - p, \\
\phi_{p,i}(x)\tilde{\phi}_{p,j}(y), & p < i < N - p, \quad 0 \leq j \leq N, \\
\tilde{\phi}_{p,i}(x)\phi_{p,j}(y), & 0 \leq i \leq N, \quad p < j < N - p, \\
\tilde{\phi}_{p,i}(x)\tilde{\phi}_{p,j}(y), & i, j \leq p, \quad i, j \geq N - p.
\end{cases}
\]

Below, for notational convenience, we will not explicitly distinguish between the modified basis functions and the interior basis functions and simply drop the tilde notation. Also, we will use the notation \(\mathbb{Q}_{p,N}\) to denote the one dimensional space spanned by the \((N + 1)\) Galerkin difference basis functions associated with the nodal degrees of freedom.

6.2.2. Energy Based Discontinuous Galerkin Method for the Wave Equation

We consider a mesh that discretizes \(\Omega\) into non-overlapping box shaped elements \(\Omega^k\) with \(\Omega = \bigcup_{k=1}^m \Omega^k\). On each element \(\Omega^k = \bigotimes_{i \in S} [L_i^k, R_i^k]\), where \(S = \{x, y\}\) or \(S = \{x, y, z\}\) depending if we are working in two or three dimensions. Let \((\mathbb{Q}_{p,N})^d\) be the space of functions spanned by the tensor product of the one dimensional Galerkin difference basis on an element, i.e. a test function \(\varphi\) in \((\mathbb{Q}_{p,N})^d\) can be expressed (in three dimensions) as
\[
\varphi(x, y, z) = \prod_{z \in S} \phi_{p,i,z}(z).
\]
Here we assume the same degree of approximation and the same number of degrees of freedom in each dimension but remark that these can also be chosen independently.

Let \(A\) in (2.3)–(2.4) be a diagonal matrix with \(c\) to be the diagonal value; let \(\psi\) substitute \(\psi_u\) and \(\psi_v\) in (2.3)–(2.4) and \(\tilde{\phi}_u = 1\) in (2.5); set external forcing \(f\) to be 0. Then on each DG element \(\Omega^k\), we have our discretization
\[
\begin{align*}
\int_{\Omega^k} \nabla \varphi \cdot \left( \frac{\partial \nabla u}{\partial t} - \nabla v \right) d\Omega^k &= \int_{\partial \Omega^k} (\vec{n} \cdot \nabla \varphi)(v^* - v) d\Omega^k, \\
\int_{\Omega^k} \varphi \frac{\partial v}{\partial t} + c^2 \nabla \varphi_v \cdot \nabla u d\Omega^k &= \int_{\partial \Omega^k} c^2 \varphi (\vec{n} \cdot (\nabla u)^*) d\Omega^k,
\end{align*}
\]
for all $\varphi \in (Q_{p,N})^d$. As equation (6.1) vanishes for constants we augment it by the independent equation
\[
\int_{\Omega^k} \left( \frac{\partial u}{\partial t} - v \right) \, d\Omega^k = 0. \tag{6.3}
\]
The numerical fluxes $v^*$ and $(\nabla u)^*$ are defined in (2.9)–(2.10). Particularly, we choose the coefficients in (2.9)–(2.10) as in (2.11)–(2.13) for the numerical simulations in this chapter.

### 6.3. Efficient Formulation on Cartesian Grids

In this section we restrict our attention to the case of constant speed of sound and Cartesian grids and show how a simple simultaneous diagonalization of the mass and stiffness matrix can be used to construct practical implementations of (6.1), (6.2) and (6.3) whose cost scales linearly with the total number of degrees of freedom.

On each element $\Omega^k$, e.g, in three dimensions, we approximate the solution by tensor product expansions
\[
\begin{align*}
 u(x, y, z, t) &= \sum_{l_x=0}^{N} \sum_{l_y=0}^{N} \sum_{l_z=0}^{N} u_{l_x,l_y,l_z} \phi_{p,l_x}(x) \phi_{p,l_y}(y) \phi_{p,l_z}(z), \\
 v(x, y, z, t) &= \sum_{l_x=0}^{N} \sum_{l_y=0}^{N} \sum_{l_z=0}^{N} v_{l_x,l_y,l_z} \phi_{p,l_x}(x) \phi_{p,l_y}(y) \phi_{p,l_z}(z).
\end{align*}
\]

On element $\Omega^k$, let $U^k$, $V^k$ be column vectors containing the nodal values of $u$ and $v$, that is $u_{l_x,l_y,l_z}$ and $v_{l_x,l_y,l_z}$, respectively. Then, in $d$-dimensions we may write the nodal based version of the method as the system of ordinary differential equations
\[
\dot{S} \frac{dU^k}{dt} = \hat{S} V^k \\
+ \sum_{j=1}^{d} (-\theta) \left[ (D^{R,R} - D^{L,L}) V^k - D^{R,L} V^{k+1} + D^{L,R} V^{k-1} \right] \\
- \sum_{j=1}^{d} \tau \left[ (C^{R,R} - C^{L,L}) U^k + C^{R,L} U^{k+1} - C^{L,R} U^{k-1} \right], \tag{6.4}
\]

129
\[
M \frac{dV^k}{dt} = -c^2 S U^k \\
+ c^2 \sum_{j=1}^{d} \theta \left( E^{R,R} - E^{L,L} \right) U^k + (1 - \theta) \left( E^{R,L} U^{k+1} - E^{L,R} U^{k-1} \right) \\
- c^2 \sum_{j=1}^{d} \tau \left[ \left( B^{R,R} - B^{L,L} \right) V^k + B^{R,L} V^{k+1} - B^{L,R} V^{k-1} \right].
\] (6.5)

Here we abuse the notation in that for each coordinate direction in the sums we use the superscript \(k \pm 1\) to denote the element “right” and “left” of an element \(k\) in the \(j\)th direction.

The definitions of the mass matrix \(M\), the stiffness matrices \(S\), \(\hat{S}\) and the lift matrices \(B, C, D, E\) are given later.

6.3.1. Complexity with Galerkin Difference Basis

Now, to compute the time derivatives \(\frac{dU^k}{dt}\) and \(\frac{dV^k}{dt}\) we must evaluate the matrix vector products on the right hand side and the action of the lift matrices on \(U^k\) and \(V^k\). As the matrices are sparse it is possible to do this at a cost that scales as \(\sim f(p)N^\kappa\), with \(f(p)\) being a low degree polynomial in \(p\) and \(\kappa = d\) for the volume terms and \(\kappa = d - 1\) for the surface terms. In other words the cost scales linearly with the number of degrees of freedom.

Further, due to the tensor product structure of the mass matrix we have that the element mass matrix \(M\) can be composed as a Kronecker product of the one dimensional matrix, which we denote \(M_j\),

\[
M = \bigotimes_{j=1}^{d} M_j,
\]

with \(M_{j,kl} = \int_{L_j}^{R_j} \phi_{j,k} \phi_{j,l} dx_j\). Now, as the one dimensional mass matrices have bandwidth \(p\) so will its \(LU\)-factors. Let \(L_j U_j \equiv M_j\) then by the Hadamard product property we have

\[
M = \bigotimes_{j=1}^{d} M_j = \bigotimes_{j=1}^{d} L_j U_j = (\bigotimes_{j=1}^{d} L_j)(\bigotimes_{j=1}^{d} U_j) \equiv LU.
\]

Thus, as the cost of each substitution is \(O(p(N+1))\) the cost of solving \(Mx = b\) is \(O((p(N+1))^d)\), which again is linear in the degrees of freedom. Unfortunately, the stiffness matrix \(\hat{S}\)
is a sum of Kronecker products

\[ S = S_1 \otimes M_2 \otimes M_3 + M_1 \otimes S_2 \otimes M_3 + M_1 \otimes M_2 \otimes S_3, \]

with \( S_{j,kl} = \int_{L_j}^{R_j} \frac{d\phi_{j,k}}{dx_j} \frac{d\phi_{j,l}}{dx_j}, \) supplied by an extra equation from (6.3), so that a direct solve of \( \hat{S}x = b \) on all elements will scale as \( O((N + 1)^{2d}) \). As \( N \) may be much larger than \( p \) this scaling implies that the resulting method would not be competitive with an implementation using a standard polynomial basis. We must thus seek an improved method. Here, we want to mention that it is acceptable to iteratively solve the scheme as long as it is only done for a few elements.

6.3.2. Optimal Computational Complexity by Simultaneous Diagonalization

The above mentioned complexity for evolving \( U \) is not competitive for practical computations. In this section we show that it is possible to make a simple (computational) change of basis that results in a method with linear complexity. Precisely the new basis is found by solving the generalized eigenvalue problem for each of the one dimensional matrices, \( S_j, \) \( j = 1, \ldots, d. \) That is, the new basis vectors are solutions to,

\[ S_j \psi_{j,k_j} = \lambda_{j,k_j} M_j \psi_{j,k_j}, \quad k_j = 0, \ldots, N_j, \quad j = 1, \ldots, d. \]

We normalize the eigenvectors according to

\[ \psi_{j,k_j} \leftarrow \frac{\psi_{j,k_j}}{((\psi_{j,k_j})^T M_j \psi_{j,k_j})^{1/2}}. \]

Let \( \Psi_j \) be the matrix containing the new one dimensional basis

\[ \Psi_j = \begin{pmatrix} \psi_{j,0} & \psi_{j,1} & \cdots & \psi_{j,N_j} \end{pmatrix}, \]

then the \( d \)-dimensional basis is

\[ \Psi = \otimes_{j=1}^d \Psi_j. \]

Now, we define \( \bar{U} \) and \( \bar{V} \) by

\[ U = \Psi \bar{U}, \quad V = \Psi \bar{V}, \]

131
then equation (6.1) and (6.2) become

\[ \Psi^T S \Psi \left( \frac{d \bar{U}}{dt} - \bar{V} \right) = f_u, \quad (6.6) \]

\[ \Psi^T M \Psi \frac{d \bar{V}}{dt} + c^2 \Psi^T S \Psi \bar{U} = f_v, \quad (6.7) \]

where,

\[ f_u = \sum_{j=1}^{d} \sum_{j=1}^{d} \psi^T \left[ (D_{R,R} - D_{L,L}) \psi^{k} \bar{U}^{k \pm 1} + D_{L,R} \psi^{k \pm 1} \right] \]

\[ - \sum_{j=1}^{d} \psi^T \left[ (C_{R,R} - C_{L,L}) \psi^{k} + C_{R,L} \psi^{k \pm 1} - C_{L,R} \psi^{k \pm 1} \right] \]

\[ f_v = c^2 \sum_{j=1}^{d} \psi^T \left[ (E_{R,R} - E_{L,L}) \psi^{k} + (1 - \theta) \psi^{k \pm 1} \right] \]

\[ - c^2 \sum_{j=1}^{d} \beta \psi^T \left[ (B_{R,R} - B_{L,L}) \psi^{k} + B_{R,L} \psi^{k \pm 1} - B_{L,R} \psi^{k \pm 1} \right] \]

In the new basis we have that the mass matrix diagonalizes

\[ \Psi^T M \Psi = \otimes_{j=1}^{d} (\Psi_j^T M_j \Psi_j) = I, \]

as does the differentiation matrix

\[ \Psi^T S \Psi = \sum_{j=1}^{d} (\Psi_j^T M_j \Psi_j) \otimes \cdots \otimes (\Psi_{j-1}^T M_{j-1} \Psi_{j-1}) \otimes (\Psi_j^T S_j \Psi_j) \otimes (\Psi_{j+1}^T M_{j+1} \Psi_{j+1}) \otimes \cdots \]

\[ \cdots \otimes (\Psi_d^T M_d \Psi_d) = \sum_j \Lambda_j, \]

where,

\[ \Lambda_j = I \otimes \cdots \otimes I \otimes \text{diag}(\lambda_{j,0}, \ldots, \lambda_{j,N_j}) \otimes I \otimes \cdots \otimes I. \]

In the above equations we use \( I \) to denote the identity matrix of size inferred by the context.

Note that as one of the eigenvalues of \( S \) is zero one of the equations in (6.6) vanishes. Suppose we have ordered the unknowns so that this corresponds to the first entry in \( \bar{U} \), then
we simply enforce the additional independent equation
\[ \frac{d\bar{U}_1}{dt} = \bar{V}_1. \]

We thus conclude that the cost of all the volume terms scales linearly with the number of degrees of freedom. We now turn to the evaluation of the surface terms in the new basis.

Consider first the surface terms in the Galerkin difference basis. In a single dimension the elements in the four different surface terms are of the form
\[ \tilde{B}_{X,Y}^{kl} = \phi_k(X)\phi_l(Y), \quad \tilde{D}_{X,Y}^{kl} = \frac{d\phi_k}{dx}(X)\phi_l(Y), \quad \tilde{E}_{X,Y}^{kl} = \phi_k(X)\frac{d\phi_l}{dx}(Y), \quad \tilde{C}_{X,Y}^{kl} = \frac{d\phi_k}{dx}(X)\frac{d\phi_l}{dx}(Y), \]
where \( \{X, Y\} \in \{\{L, L\}, \{L, R\}, \{R, L\}, \{R, R\}\} \).

Now due to the local support properties of the Galerkin difference basis the number of nonzero elements in the above matrices are 1 for \( \tilde{B}_{X,Y}^{kl} \), \( (p+1) \) for \( \tilde{D}_{X,Y}^{kl} \) and \( \tilde{E}_{X,Y}^{kl} \), and \( (p+1)^2 \) for \( \tilde{C}_{X,Y}^{kl} \).

The \( d \)-dimensional version of the surface matrices can again be constructed by Kronecker products. For example we have that
\[ D_j^{X,Y} = M_1 \otimes \cdots \otimes M_{j-1} \otimes \tilde{B}_j^{X,Y} \otimes M_{j+1} \otimes \cdots \otimes M_d. \]

Applying the change of basis we have that
\[ \Psi^T D_j^{X,Y} \Psi = (\Psi^T M_1 \Psi_1) \otimes \cdots \otimes (\Psi^T_{j-1} M_{j-1} \Psi_{j-1}) \otimes (\Psi^T_j \tilde{B}_j^{X,Y} \Psi_j) \otimes (\Psi^T_{j+1} M_{j+1} \Psi_{j+1}) \otimes \cdots \]
\[ \cdots \otimes (\Psi^T_M M_d \Psi_d) = I \otimes \cdots \otimes I \otimes (\Psi_j \tilde{B}_j^{X,Y} \Psi_j) \otimes I \otimes \cdots \otimes I. \]

Thus applying \( \Psi^T D_j^{X,Y} \Psi \) to the \( (N+1)^d \) dimensional vector \( \bar{V} \) can be done at a cost that scales with \( (p+1)(N+1)^d \). Similarly, the cost of applying \( \Psi^T B_j^{X,Y} \Psi, \Psi^T C_j^{X,Y} \Psi, \) and \( \Psi^T E_j^{X,Y} \Psi \) can be done at a cost of \( (N+1)^d, (p+1)^2(N+1)^d, \) and \( (p+1)(N+1)^d \), respectively.

6.3.3. Numerical Verification of the Computational Complexity

We now present timing results that illustrate the theoretical findings of Section 6.3. We consider a two dimensional problem defined in the domain \([0, 1] \times [0, 1]\) and use the upwind
flux with $\xi = c$ (the other fluxes give similar timing results). Here, we use the manufactured solution

$$u(x, t) = \sin(ck\pi t) \sin(k\pi x) + \cos(ck\pi t) \cos(k\pi y)$$

with $c = 1$ and $k = 16$ to construct the corresponding initial and boundary conditions.

In addition, we consider both one single DG element ($m = 1$ i.e., $n = 1$) and thirty-six DG elements ($m = 36$, i.e., $n = 6$). Specifically, we let $N = 3, 4, 5 \cdots, 42$ when $m = 1$ for both standard basis and improved basis; $N = 3, 4, 5 \cdots, 29$ when $m = 36$ for both standard basis and improved basis. Further, the degree of the approximation space is chosen to be $p = 3$. Finally, we use the inline function CPU\_TIME() in FORTRAN to record the elapsed CPU time which reflects the total time of evolving 10 steps by a 4-stage fourth order Runge-Kutta time integrator.

In Figure 6.2, we observe that the CPU time is proportional to $(n^2(N + 1))^2$ for the standard Galerkin basis and proportional to $n^2(N + 1)^2$ for the Galerkin difference basis obtained by the simultaneous diagonalization. Also, the CPU time for the Galerkin difference basis obtained by the simultaneous diagonalization is much smaller than the standard Galerkin basis.

### 6.4. Dispersion Analysis

To investigate how well the proposed scheme in Section 6.2.2 preserves the wave propagation properties, we use the standard analysis as in [2, 46, 48, 66]. Here, we consider

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad a < x < b,$$

with initial condition $u(x, 0) = e^{ikx}$ and periodic boundary condition $u(a, t) = u(b, t)$. Now, seeking spatially periodic solutions of the form

$$u(x, t) = e^{i(kx - \omega t)},$$

where

which corresponds to an exact dispersion $\omega = \pm ck$ for (6.8).

Partitioning the computational domain into non-overlapping uniform elements $I^k =$
\[ [x^k, x^{k+1}], k = 0, \ldots, n \] with \( H = x^{k+1} - x^k = (b - a)/n \). For each element \( I^k \), there are \((N + 1)\) equidistant nodal DOFs with spacing \( h = \frac{H}{N} \). Then for the problem (6.8), we have semi-discretization formulations as in (6.4) and (6.5) with \( d = 1 \) in \( I^k \). Let the vectors \( U^k = (U_0^k, U_1^k, \ldots, U_N^k) \) and \( V^k = (V_0^k, V_1^k, \ldots, V_N^k) \) hold the nodal approximations of \( u(x, t) \) and \( v(x, t) \) in \( I^k \), respectively. Next, we seek solutions which have the following forms
\[
U_l^k = \hat{U}_l^k e^{i(\alpha(x^k + lh) - \omega t)}, \quad V_l^k = \hat{V}_l^k e^{i(\alpha(x^k + lh) - \omega t)}, \quad l = 0, \ldots, N. \tag{6.9}
\]
Further, we assume the periodicity of the solution satisfies
\[
W_{k+1}^0 = e^{i\kappa H} W_0^k, \quad W_{k-1}^N = e^{-i\kappa H} W_N^k \tag{6.10}
\]
with \( W \) represents \( U \) or \( V \). To condense the notations, we have omitted the superscript \( k \) for the rest of this section. Let \( Z = (U, V)^T \) and combine with (6.4)-(6.5), (6.9)-(6.10), we then obtain the following eigenvalue problem
\[
\hat{A} Z = -i\Omega Z, \quad \hat{A} = \hat{A}_1 \hat{A}_2,
\]
where
\[
\Omega = \frac{\omega H}{c}, \quad \hat{A}_1 = \frac{H}{c} \begin{pmatrix} \hat{S}^{-1} \\ M^{-1} \end{pmatrix}, \quad \hat{A}_2 = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},
\]
with
\[
A_{11} = \tau(C^{L,L} - C^{R,R}) - \tau(e^{iK}C^{R,L} - e^{-iK}C^{L,R}),
\]
\[
A_{12} = \hat{S} + (-\theta)(D^{R,R} - D^{L,L}) + \theta(e^{iK}D^{R,L} - e^{-iK}D^{L,R}),
\]
\[
A_{21} = -c^2S + \theta c^2(E^{R,R} - E^{L,L}) + (1 - \theta)c^2(e^{iK}E^{R,L} - e^{-iK}E^{L,R}),
\]
\[
A_{22} = -\beta c^2(B^{R,R} - B^{L,L}) - \beta c^2(e^{iK}B^{R,L} - e^{-iK}B^{L,R}).
\]
and \( K = \kappa H \). Note that the values of \( \Omega \) are complex in general, \( \Omega = \Omega_r + i\Omega_i \), with a non-positive \( \Omega_i \) which represents the numerical damping of the corresponding scheme, the real part \( \Omega_r \) recovers an approximation of frequency \( \frac{\omega H}{c} \).
For the simulations in this section, we set the computational domain to be \( x \in [0, 1] \), i.e., \( a = 0, b = 1 \); the order of the approximation space \( p = 3 \); the number of DG elements to be \( n = 5 \) and the DOF in each DG element to be \( N + 1 = 4 \). Figure 6.5 shows the dispersion relation of the central flux and the alternating flux. As expected, we see that for \( K \) small, the numerical phase velocity is very close to the physical wave speed. As we already know, both the central flux and the alternating flux yield conservative schemes which are consistent with the results in Figure 6.6. Figure 6.3 presents the dispersion relation of the upwind flux. When \( K \) is small, the numerical phase velocity also reflects the physical wave speed. Compare the results in Figure 6.3 and Figure 6.5–6.6, we find that both conserving schemes (central flux or alternating flux) and the dissipating scheme can recover the physical mode when \( K \) is small. The conserving schemes admit more complicated phenomena: the spurious modes do not damp for small value of \( K \); for the energy-dissipating scheme, however, the unphysical modes are severely damped. In Figure 6.4, we show the dispersion relation of physical modes of dissipating schemes with a range of orders of approximation. We see that the numerical phase velocity is very close to the physical wave speed when \( K \) is small and improves for a broader range of \( K \) as the order of the approximation increases.

6.5. Spectral Radius

In this section, we study the spectral radius of the semi-discretization of our scheme. Consider the problem

\[
\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in [a, b]
\]

with periodic boundary condition \( u(a, t) = u(b, t) \). The computational domain \([a, b]\) is divided into non-overlapping uniform intervals \( I^k = [x^k, x^{k+1}], k = 0, 1, \cdots, n \) with \( H = (b - a)/n \). For each DG element \( I^k \), it has \( N + 1 \) equidistant DOFs with \( h = H/N \). We approximate \( u(x, t) \) and \( v(x, t) \) in \( I^k \) with the nodal values \( U^k_l \) and \( V^k_l \) by

\[
u(x, t) = \sum_{l=0}^{N} V^k_l \phi_{p,l}(x), \quad v(x, t) = \sum_{l=0}^{N} V^k_l \phi_{p,l}(x),
\]
respectively. Let $W^k = [W^k_0, W^k_1, \ldots, W^k_N]$, $W = [W^0, W^1, \ldots, W^n]$ with $W$ represents $U$ or $V$ and $Z = [U, V]^T$. Then we can write the semi-discretization (6.4)-(6.5) as a system of ordinary differential equations

$$\frac{dZ}{dt} = \mathcal{L}Z.$$

Practically speaking we extract $\mathcal{L}$ one column at a time by computing the time derivative in the semi-discretization (6.4)-(6.5) with one element in $Z$ to be 1 and all other to be zero.

For the simulation in this section, we choose the computational domain to be $x \in [0, 1]$ with $a = 0, b = 1$, the number of DG elements to be $n = 1$. Further, we test different degrees of approximation space $p = (1, 3, 5, 7, 9, 11)$ with different DOFs $N + 1 = (13, 21, 29)$ in the DG element. Finally, the central flux is denoted by $C$, the alternating flux is denoted by $A$ and the upwind flux is denoted by $U$.

Figure 6.7 shows the magnitude of the maximum eigenvalue as a function of $p$. For the conserving scheme (central flux or alternating flux), the magnitude of the maximum eigenvalue is proportional to $p$ for DOFs up to 29, and the slope for the alternating flux is larger than the central flux. For the dissipating scheme (upwind flux), the magnitude of the maximum eigenvalue is proportional to $p$ when DOF is slightly less than 29.

6.6. Numerical experiments

In this section, we present some numerical results to investigate the convergence of our method in the $L^2$ norm. In all cases we used a nodal formulation associated with the proposed basis functions in Section 6.3.2 and marched in time by 4-stage fourth order Runge-Kutta scheme (RK4). We also set sound wave speed $c = 1$ and the flux splitting parameter $\xi$ in the upwind scheme equals $c$. 

137
6.6.1. Periodic boundary conditions in one dimension

To investigate the order of accuracy of our method, we solve

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t \geq 0,$$

with periodic boundary condition \( u(0, t) = u(1, t) \). It admits an exact solution

$$u(x, t) = \sin(16c\pi t) \sin(16\pi x), \quad t \geq 0. \quad (6.11)$$

The discretization is performed on a uniform mesh with element vertices \( x_i = iH, \ i = 0, \ldots, n, \ H = 1/n \) and subcells vertices on element \((i + 1)\) is \( x_{ij} = x_i + jh, \ i = 0, \ldots, n - 1, \ j = 0, \ldots, N, \ h = H/N \). We evolve the solution until \( T = 0.4 \) with time step \( \Delta t = CFL \times h \), \( CFL = 0.0075/(2\pi) \) to guarantee the error is dominated by space error. We present the \( L^2 \) error for \( u^h \) with the degree of approximation polynomials \( p = (1, 3, 5, 7, 9) \). We refine the grid spacing \( h \) by increasing the number of DG elements and fixing the the degrees of freedom in each element to be \( N + 1 = 11 \). Further, we test three different fluxes: the central flux, the alternating flux and the upwind flux. We denote C.-flux to be the central flux, A.-flux to be the alternating flux and U.-flux to be the upwind flux.

The \( L^2 \) errors for \( u \) and \( v \) are plotted against the grid spacing \( h \) in Figure 6.8 for \( u^h \) when the upwind flux, the central flux and the alternating flux are used, respectively. Linear regression estimates of the rate of convergence can be found in Table 6.1. Note that we only use the ten finest grids to obtain the rate of convergence. When \( p = (3, 5, 7) \), we observe an optimal convergence, \( p + 1 \), for \( u \) with all three different fluxes. When \( p = 9 \), we observe a super convergence rate, around \( p + 3/2 \), for \( u \) with three different fluxes. Finally, when \( p = 1 \), the central flux performs best, an optimal convergence; the alternating flux does not converge and the upwind flux has a sub-optimal convergence \( p \).
Table 6.1: Linear regression estimates of the convergence rate of $u$ in one dimensional with periodic boundary condition.

<table>
<thead>
<tr>
<th>Degree ($p$) of approx. for $u$</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ U.-flux</td>
<td>0.93</td>
<td>4.16</td>
<td>6.01</td>
<td>7.95</td>
<td>10.73</td>
</tr>
<tr>
<td>Rate fit $u$ C.-flux</td>
<td>1.81</td>
<td>4.38</td>
<td>6.14</td>
<td>8.08</td>
<td>10.64</td>
</tr>
<tr>
<td>Rate fit $u$ A.-flux</td>
<td>0.04</td>
<td>3.98</td>
<td>6.01</td>
<td>7.97</td>
<td>10.60</td>
</tr>
</tbody>
</table>

6.6.2. Dirichlet and Neumann boundary conditions in one dimension

We now test our method on the problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 31/32), \quad t \geq 0,$$

with a homogeneous Dirichlet boundary condition $u(0, t) = 0$ at $x = 0$ and a homogeneous Neumann boundary condition $u_x(31/32, t) = 0$ at $x = 31/32$. It has a same solution (6.11) as in Section 6.6.1.

The discretization is performed on a uniform mesh with element vertices $x_i = iH$, $i = 0, \cdots, n$, $H = 31/32/n$ and subcells vertices on element $(i+1)$ is $x_{ij} = x_i + jh$, $i = 0, \cdots, n-1$, $j = 0, \cdots, N$, $h = H/N$. We evolve the solution until $T = 0.4$ with time step $\Delta t = \text{CFL} \times h$, $\text{CFL} = 0.0075/(2\pi)$ to remove the effect of the temporal error we set. As in Section 6.6.1, we refine the grid spacing $h$ by increasing the number of elements $n$ and set $N = 10$.

The $L^2$ error of $u$ is plotted against the grid spacing $h$ in Figure 6.9 with three different fluxes. The linear regression estimate of the convergence rate is shown in Table 6.2, again we only use the ten finest grids. We observe a similar results as in Section 6.6.1 with periodic boundary condition when $p = (1, 3, 5, 7)$. And an optimal convergence for $p = 9$.

6.6.3. Periodic boundary conditions in two dimensions

Lastly, we examine our method on

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \quad (x, y) \in (0, 1) \times (0, 1), \quad t \geq 0,$$
Table 6.2: Linear regression estimates of the convergence rate of $u$ in one dimension with Dirichlet boundary condition on the left and Neumann boundary condition on the right. first 14 points for $q = 9$

<table>
<thead>
<tr>
<th>Degree ($p$) of approx. for $u$</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ U.-flux</td>
<td>0.88</td>
<td>4.20</td>
<td>6.01</td>
<td>7.96</td>
<td>9.92</td>
</tr>
<tr>
<td>Rate fit $u$ C.-flux</td>
<td>1.82</td>
<td>4.17</td>
<td>6.07</td>
<td>8.05</td>
<td>10.16</td>
</tr>
<tr>
<td>Rate fit $u$ A.-flux</td>
<td>0.03</td>
<td>4.04</td>
<td>5.97</td>
<td>7.96</td>
<td>9.99</td>
</tr>
</tbody>
</table>

with periodic boundary conditions and initial condition chosen so that

$$u(x, y, t) = \sin(16\pi t)(\sin(16\pi x) + \cos(16\pi y)).$$

We consider regular Cartesian grids with the elements whose vertices are defined by $x_i = iH$, $y_j = jH$, $i, j = 0, 1, \ldots, n$, $H = 1/n$ and subcells vertices on element $(i + 1, j + 1)$ is $x_{ik} = x_i + kh$, $i = 0, \ldots, n - 1$, $k = 0, \ldots, N$, and $y_{jl} = y_j + lh$, $j = 0, \ldots, n - 1$, $l = 0, \ldots, N$, $h = H/N$. We evolve the solution until $T = 0.2$ and the time step size is chosen to be $\Delta t = \text{CFL}h$, with $\text{CFL} = 0.0075/(2\pi)$ to guarantee the temporal error is controlled by the space error. We also fixed $N = 10$ for the experiment in this section.

The $L^2$ error for $u$ is presented in Figure 6.10 with the upwind flux, the central flux and the alternating flux, respectively. The corresponding convergence rate from linear regression is shown in Table 6.3, note that we still only use the ten finest grids here. The results in two dimensions are very similar to the one dimensional results.
Table 6.3: Linear regression estimates of the convergence rate of $u$ in two dimensions with periodic boundary conditions.

<table>
<thead>
<tr>
<th>Degree ($p$) of approx. for $u$</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate fit $u$ U.-flux</td>
<td>0.40</td>
<td>3.67</td>
<td>6.10</td>
<td>7.97</td>
<td>11.33</td>
</tr>
<tr>
<td>Rate fit $u$ C.-flux</td>
<td>1.83</td>
<td>4.11</td>
<td>6.16</td>
<td>7.92</td>
<td>11.24</td>
</tr>
<tr>
<td>Rate fit $u$ A.-flux</td>
<td>0.21</td>
<td>3.97</td>
<td>6.03</td>
<td>7.97</td>
<td>11.41</td>
</tr>
</tbody>
</table>

Figure 6.2: Plots of CPU time for the standard Galerkin difference basis (left column) and the Galerkin difference basis by simultaneous diagonalization (right column) as a function of degrees of freedom (DOF) in two dimensions. The first row is for $n = 1$ and the second row is for $n = 6$. The sound wave speed $c = 1$, and the splitting parameter for upwind flux is $\xi = c$. 
Figure 6.3: On the top, we show the numerical dispersion relation for the linear acoustic operator with the Sommerfeld flux. The black dashed lines represent the exact case, ‘p.-mode’ represents the physical mode and ‘s.-mode’ represent the spurious mode. On the bottom, we illustrate the dissipation associated with the eight modes. The sound wave speed is $c = 1$ and the degree of the approximation spaces for both $u$ and $v$ is $p = 3$. The flux splitting parameter $\xi = c = 1$. 
Figure 6.4: On the top, we show the numerical dispersion relation for the linear acoustic operator with the Sommerfeld flux. The black dashed line represents the exact case. We present the numerical dispersion relation for the physical mode at different orders $p = (1, 3, 5, 7, 9)$. On the bottom, we illustrate the dissipation associated with the different orders $p = (1, 3, 5, 7, 9)$. The sound wave speed $c$ and the flux splitting parameter $\xi$ are chosen to be same $c = \xi = 1$. 
Figure 6.5: On the top, we show the numerical dispersion relation for the linear acoustic operator with the central flux. On the bottom, we show the numerical dispersion relation for the linear acoustic operator with the alternating flux. The black dashed lines represent the exact case. 'p.-mode' represents the physical mode and 's.-mode' represent the spurious mode. The sound wave speed is $c = 1$ and the degree of the approximation spaces for both $u$ and $v$ is $p = 3$. 
Figure 6.6: On the left, we show the numerical dissipation for the linear acoustic operator with the central flux for all modes. On the right, we show the numerical dissipation for the linear acoustic operator with the alternating flux for all modes. The sound wave speed is \( c = 1 \) and the degree of the approximation spaces for both \( u \) and \( v \) is \( p = 3 \). ‘p.-mode’ represents physical mode and ‘s.-mode’ represents the spurious mode.

Figure 6.7: Spectral radii (scaled by \( h = 1/N \)) as a function of the degree of approximation \( (p) \) for the three different fluxes. The sound wave speed is \( c = 1 \), and the splitting parameter \( \xi \) for upwind flux is \( \xi = c \).
Figure 6.8: Plots of the error in $u$ as a function of $h$ in one dimension with upwind flux, central flux and alternating flux for periodic boundary conditions. In the legend, $p$ is the degree of the approximation space of $u$. 
Figure 6.9: Plots of the error in \( u \) as a function of \( h \) in one dimension with upwind flux, central flux and alternating flux for Dirichlet boundary condition on the left and Neumann boundary condition on the right. In the legend, \( p \) is the degree of the approximation space of \( u \).
Figure 6.10: Plots of the error in $u$ as a function of $h$ in two dimensions with upwind flux, central flux and alternating flux for periodic boundary conditions. In the legend, $p$ is the degree of the approximation space of $u$. 
In conclusion, the main results of the thesis are as follows.

In Chapter 2, we have investigated the convergence property of the energy-based discontinuous Galerkin method proposed in [3] for second-order scalar wave equations. By choosing spacial flux parameters, and introducing a special elliptic projection operator, we improved the suboptimal convergent results in [3] to optimal convergence with Cartesian meshes. For quadrilateral meshes, an optimal convergence has been obtained by choosing special flux parameters, and introducing special elliptic projection operator and adding a penalty term to penalize the term containing the jumps of solutions at mesh interfaces. Numerical experiments were presented to verify the theoretical findings.

In Chapter 3, we have generalized the energy-based DG method of [3] to the wave equation with advection, a problem for which the energy density takes a more complicated form than a simple sum of a term involving the time derivative and a term involving space derivatives. We have shown that the new form can be handled by introducing a second variable which, unlike what was done in [3, 4], involves both space and time derivatives. We prove error estimates completely analogous with those shown in [3] for the isotropic wave equation, including cases with both subsonic and supersonic background flows. Numerical experiments also demonstrate optimal convergence on regular grids when an upwind flux is used.

In Chapter 4, we have demonstrated that the energy-based DG method for second-order wave equations can be generalized to semilinear problems. In particular we: 1) modified the weak form proposed in [3] so that the time derivatives of the approximate solution can be computed via the solution of a linear system of equations in each element; 2) established the stability of the method by proving energy estimates for a wide choice of fluxes with mesh-independent parametrizations, including energy-conserving central or alternating fluxes as
well as dissipative upwind fluxes; 3) derived suboptimal estimates of convergence in the energy norm; 4) observed, for polynomial degrees above 3, optimal convergence in the $L^2$ norm for the energy-conserving alternating flux as well as for dissipative methods based on Sommerfeld flux splitting.

In Chapter 5, we have proposed a new scheme to overcome the typical numerical stiffness associated with the high-order piecewise polynomial approximations in the discontinuous Galerkin framework. Particularly, we established the energy-based DG method for the scalar second-order wave equation with spatial staggered elements in the bulk of the computational domain. In the one dimensional problem with periodic boundary condition, we proved the the Tame CFL condition $c^2 \Delta t \frac{\epsilon}{\Delta x} < 0.15$. Numerical experiments demonstrated the optimal convergence in $L^2$ norm for both one and two dimensional problems. For the two dimensional problem and the problems with a bounded domain, we numerically verified that our methods admitted larger time-step sizes by combining local time-stepping with staggered grids.

In Chapter 6, we have combined the energy-based DG method with Galerkin difference basis functions to solve the second-order scalar wave equations. The scheme helps to overcome the numerical stiffness coming from the high-order piecewise polynomial approximations. In addition, we generated new basis functions from Galerkin difference basis functions by simultaneous diagonalization. The new basis functions help reduce the computational cost of the inversion of the stiffness matrix from $O(N^2)$ to $O(N)$ with $N$ the number of degrees of the freedom. The numerical experiments demonstrated the optimal convergence in the $L^2$ norm with the new basis functions on regular meshes. Based on the experiments and on computations of the spectral radius, we demonstrated that larger times step sizes could be used compared with the energy-based DG method in [3].


[52] K. Kormann, M. Kronbichler, and B. Müller. Derivation of strictly stable high order
difference approximations for variable-coefficient PDE. *Journal of Scientific


[54] H. O. Kreiss and L. Wu. On the stability definition of difference approximations for the
1993.

method for elliptic problems: stability and energy error estimates. *SIAM Journal on

transport equation. In *Mathematical Aspects of Finite Elements in Partial

[57] S. K. Lele. Compact finite difference schemes with spectral-like resolution. *Journal of

[58] D. Levy, C.-W. Shu, and J. Yan. Local discontinuous Galerkin methods for nonlinear


[60] Q. Lin and N.-n. Yan. The construction and analysis of high efficiency finite element

penalty discontinuous Galerkin method of Cahn–Hilliard equation. *Numerical

equation with boundary effect. *Journal of Computational Physics*, 215(1):197–218,
2006.

[63] J. Melenk and S. Sauter. Convergence analysis for finite element discretizations of the
Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Mathematics


