ACCURACY AND MONOTONICITY OF SPECTRAL ELEMENT METHOD ON STRUCTURED MESHES

by

Hao Li

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Mathematics

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Xiangxiong Zhang, Chair

Department of Mathematics

Dr. Daniel Appelö

Department of Computational Mathematics, Science, and Engineering,
Michigan State University

Dr. Jingwei Hu

Department of Mathematics

Dr. Jie Shen

Department of Mathematics

Approved by:

Dr. Plamen D. Stefanov

To my parents

ACKNOWLEDGMENTS

My deepest gratitude is to my advisor Prof. Xiangxiong Zhang, who supported me for five years not only academically, but also mentally. Prof. Zhang provided inspired research advice when I was confused, taught me how to analyze questions, express ideas and think critically. I learned how valuable ideas come out by looking ways he works in. I am grateful to him for holding me to a high research standard and telling me how to embrace bigger challenges. His support and patience have been helping me overcome many frustrating situations. Beyond my research, he has been a role model and a tremendous influence on my life. I am very fortunate to have him as my Ph.D advisor.

I would like to thank Prof. Daniel Appelö, Prof. Jingwei Hu and Prof. Jie Shen who have graciously agreed to serve on my dissertation committee, spent time and provided invaluable feedback.

Parts of this dissertation are based on two publications, [1] and [2]. My coauthors Dr. Shusen Xie and Dr. Daniel Appelö have contributed considerably through their intelligent work.

I am grateful to the Department of Mathematics and graduate school of Purdue University for generous financial support in the past five years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

On rectangular meshes, the simplest spectral element method for elliptic equations is the classical Lagrangian Q^k finite element method with only $(k+1)$-point Gauss-Lobatto quadrature, which can also be regarded as a finite difference scheme on all Gauss-Lobatto points. We prove that this finite difference scheme is $(k+2)$-th order accurate for $k \geq 2$, whereas Q^k spectral element method is usually considered as a $(k+1)$-th order accurate scheme in L^2-norm. This result can be extended to linear wave, parabolic and linear Schrödinger equations.

Additionally, the Q^k finite element method for elliptic problems can also be viewed as a finite difference scheme on all Gauss-Lobatto points if the variable coefficients are replaced by their piecewise Q^k Lagrange interpolants at the Gauss Lobatto points in each rectangular cell, which is also proven to be $(k+2)$-th order accurate.

Moreover, the monotonicity and discrete maximum principle can be proven for the fourth order accurate Q^2 scheme for solving a variable coefficient Poisson equation, which is the first monotone and high order accurate scheme for a variable coefficient elliptic operator.

Last but not the least, we proved that certain high order accurate compact finite difference methods for convection diffusion problems satisfy weak monotonicity. Then a simple limiter can be designed to enforce the bound-preserving property when solving convection diffusion equations without losing conservation and high order accuracy.

1. INTRODUCTION

Accurate and efficient approximations of solutions to partial differential equations are important to numerous applications arising in engineering and the sciences. For the numerical methods solving the partial differential equations, we are interested in three practical perspectives: accuracy, efficiency and stability. The methods that are based on a variational formulation, such as spectral methods and finite element methods, are usually with accuracy guaranteed. High order numerical methods will help achieve the desired accuracy with low computation cost. For numerical stability, it is desired to have numerical solutions to preserve some discrete analogues of the key properties of the exact solution. The First three chapters are dedicated for accuracy analysis and the last two chapters will deal with stability issues.

1.1 Superconvergence Of Spectral Element Method And Its Finite Difference Type Implementation

Consider solving a two-dimensional elliptic equation with smooth coefficients on a rectangular domain (or some geometry that can be mapped to a rectangular smoothly) with homogeneous Dirichlet boundary condition by the classical spectral element method on a rectangular mesh. The variational problem from the elliptic equation is to find $u \in H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ satisfying

$$A(u, v) := \iint_\Omega (\nabla v^T \mathbf{a} \nabla u + \mathbf{b} \nabla u v + cuv)\, dxdy = (f, v), \quad \forall v \in H_0^1(\Omega), \tag{1.1}$$

where $\mathbf{a} = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$ is real symmetric positive definite and $\mathbf{b} = (b^1 \quad b^2)$.

Let h be the mesh size and $V_0^h \subseteq H_0^1(\Omega)$ be the piecewise polynomial space consisting of piecewise Q^k polynomials (i.e., tensor product of piecewise polynomials of degree k), then the continuous finite element solution is defined as $u_h \in V_0^h$ satisfying

$$A(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_0^h. \tag{1.2}$$

It is well-known that standard error estimates of (1.2) are $\|u-u_h\|_1 \leq Ch^k\|u\|_{k+1}$ and $\|u-u_h\|_0 \leq Ch^{k+1}\|u\|_{k+1}$ where $\|\cdot\|_k$ denotes $H^k(\Omega)$-norm, see [3]. For $k \geq 2$, $\mathcal{O}(h^{k+1})$ superconvergence for the gradient at Gauss quadrature points and $\mathcal{O}(h^{k+2})$ superconvergence for functions values at Gauss-Lobatto quadrature points were proven for one-dimensional case in [4]–[6] and for two-dimensional case in [7]–[10].

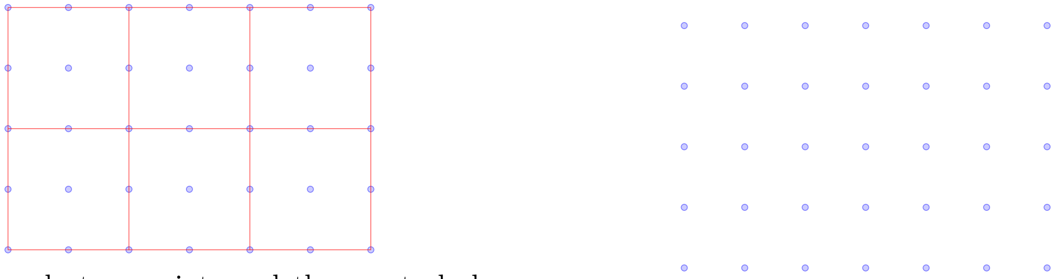
The spectral element method in the literature usually refers to implementing the scheme (1.2) with tensor product of m-point Gauss-Lobatto quadrature with $m \geq k+1$. For the Q^k spectral element method, the previous standard finite element error estimates still hold [11], i.e., the error in H^1-norm is k-th order and the error in L^2-norm is $(k+1)$-th order. It is also well known that the Lagrangian Q^k $(k \geq 2)$ continuous finite element method is $(k+2)$-th order accurate in the discrete 2-norm over all $(k+1)$-point Gauss-Lobatto quadrature points [8]–[10]. If using a very accurate quadrature in the finite element method for a variable coefficient operator $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u)$, then $(k+2)$-th order superconvergence at Gauss-Lobatto points holds trivially. In practice users might use over-integration $m > k+1$ for problems with variable coefficients, which will deteriorate the efficiency. In this dissertation, we prove that even the superconvergence of function values still hold for the simplest choice $m = k+1$ i.e. $(k+1)$-points Gauss-Lobatto quadrature, which is desired for the efficiency of having a diagonal mass matrix and for the convenience of implementation. In particular in the seismic community, where highly efficient simulation of the elastic wave equation is of important, the spectral method has become the method of choice, [12], [13].

It may not seem surprising that the $(k+2)$-th order superconvergence of (1.2) would be affected by the $(k+1)$-point Gauss-Lobatto quadrature, but it is actually quite difficult to prove by standard superconvergence techniques and never proved in the literature.

The spectral element scheme can be denoted as finding $u_h \in V_0^h$ satisfying

$$A_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h, \tag{1.3}$$

where $A_h(u_h, v_h)$ and $\langle f, v_h\rangle_h$ denote using tensor product of $(k+1)$-point Gauss-Lobatto quadrature for integrals $A(u_h, v_h)$ and (f, v_h) respectively. Such a scheme can be regarded as a finite difference type scheme on all Gauss-Lobatto points, see Figure 1.1.



(a) The quadrature points and the spectral element mesh.

(b) The corresponding finite difference grid

Figure 1.1. An illustration of Lagrangian Q^2 element and the 3×3 Gauss-Lobatto quadrature.

So the coincidence of the superconvergence points and degrees of freedom actually provides us a $(k+2)$-th order accurate finite difference type scheme. To be more specific, for homogeneous and non-homogeneous Dirichlet type boundary conditions, we can show that (1.3) with $k \geq 2$ is a $(k+2)$-th order accurate finite difference scheme in the discrete 2-norm under suitable smoothness assumptions on the exact solution and the coefficients.

We emphasize that such a superconvergence result cannot be proven by the standard quadrature estimate, i.e., the Bramble-Hilbert Lemma. In order to obtain desired estimate, we used a novel and very tight Gauss-Lobatto quadrature error estimate by counting all possible cancellations of quadrature errors across element boundaries. The $(k+2)$-th order accuracy over all Gauss-Lobatto points explains why people observe higher order accuracy in spectral element method than the L^2-estimate when the errors are only measured at the quadrature points.

The above superconvergence results to spectral element method can also be extended to the case for solving parabolic, hyperbolic equations and linear Schrödinger equation. Thus for the time-dependent problem, we can gain more from the superconvergence i.e. we get a $(k+2)$-th order accurate finite difference method, which is one more order accurate than the traditional spectral element method.

Based on the same idea as above, to have the coincidence of the superconvergence points and degrees of freedom and compute the bilinear form in the scheme (1.2), another convenient implementation is to replace the smooth coefficient $a(x,y)$ by a piecewise Q^2 polynomial $a_I(x,y)$ obtained by interpolating $a(x,y)$ at the quadrature points in each cell shown in

Figure 1.1. Then one can compute the integrals in the bilinear form exactly since the integrand is a polynomial. The same $(k+2)$-th order superconvergence of function values for such an approximated coefficient scheme will be proven in Chapter 4.

1.2 Monotonicity And Discrete Maximum Principle

Consider solving a two-dimensional Poisson equation with variable coefficient and Dirichlet boundary condition on a rectangular domain $\Omega = [0,1]^2$:

$$\begin{aligned} \mathcal{L}u \equiv -\nabla \cdot (a\nabla u) + cu = 0 \quad & \text{on} \quad \Omega, \\ u = g \quad & \text{on} \quad \partial\Omega, \end{aligned} \tag{1.4}$$

where $a(x,y) \in C^1(\bar{\Omega})$, $c(x,y) \in C^0(\bar{\Omega})$ with $0 < a_{\min} \leq a(x,y) \leq a_{\max}$ and $c(x,y) \geq 0$. For a smooth enough solution u, maximum principle holds [14]: $\mathcal{L}u \leq 0$ in $\Omega \Longrightarrow \max_{\bar{\Omega}} u \leq \max\left\{0, \max_{\partial\Omega} u\right\}$, and in particular,

$$\mathcal{L}u = 0 \text{ in } \Omega \Longrightarrow |u(x,y)| \leq \max_{\partial\Omega} |u|, \quad \forall (x,y) \in \Omega. \tag{1.5}$$

A linear approximation to $\mathcal{L}$ can be represented as a matrix L_h. The matrix L_h is called *monotone* if its inverse has nonnegative entries, i.e., $L_h^{-1} \geq 0$. All matrix inequalities in the following are entrywise inequalities. One sufficient condition for the discrete maximum principle is the *monotonicity* of the scheme [15], which was also used to prove convergence of numerical schemes, e.g., [16]–[19]. Monotonicity is a sufficient condition to achieve bound-preserving property. For various purposes, it is desired to have numerical schemes to satisfy (1.5) in the discrete sense or a monotone approximation of elliptic operators, e.g., constructing bound-preserving and positivity-preserving schemes for convection dominated convection-diffusion problems.

For discrete maximum principle to hold in P^2 FEM on a generic triangular mesh, it was proven in [20] that it is necessary and sufficient to require a very strong mesh constraint, which essentially gives either regular triangulation or equilateral triangulation. Thus usually discrete maximum principle is regarded as not true for high order accurate schemes on unstructured meshes. On structured meshes, there are a few fourth order accurate finite difference schemes that is monotone for discrete Laplacian. For instance, the classical fourth order accurate 9-point discrete Laplacian, which is a fourth order accurate compact finite difference scheme, forms an M-matrix thus is monotone. However, for a variable coefficient elliptic operator, even on structured meshes, no high order accurate schemes have been proven monotone before.

For proving monotonicity, the main viable tool in the literature is to use M-matrices which are inverse positive. All off-diagonal entries of M-matrices must be non-positive. Except the fourth order compact finite difference, all high order accurate schemes induce positive off-diagonal entries, destroying M-matrix structure, which is a major challenge of proving monotonicity. M-matrix factorization of the form $L_h = M_1 M_2$ were shown for special high order schemes for Laplacian but these M-matrix factorization seem ad hoc and do not apply to complicated variable coefficient problems. In [21], Lorenz proposed some matrix entry-wise inequality for ensuring a matrix to be a product of two M-matrices and applied it to Lagrangian P^2 finite element method on uniform regular triangular meshes for Laplacian. We were able to extend Lorenz's condition to the Q^2 spectral element method for a scalar variable coefficient problem $-\nabla \cdot (a\nabla u) + cu = f$ on uniform meshes. This is the first time a high order accurate is proven monotone for a variable coefficient problem. Following this approach, the fifth order accurate Q^3 scheme was proven monotone for Laplacian in [22].

For convection dominated convection-diffusion problems, we also proved that certain high order accurate compact finite difference methods with high order strong stability preserving time discretizations for convection diffusion problems satisfies weak monotonicity as in [23]. Then a simple limiter can be designed to enforce the bound-preserving property in compact finite difference schemes solving convection diffusion equations without losing conservation and high order accuracy.

1.3 Organization Of The Dissertation

In this dissertation, in Chapter 2, we analyze the accuracy of spectral element method for elliptic equation measured on the $(k+1)$ Gauss-Lobatto points, on which the method can be viewed as a $(k+2)$-th order finite difference method. In Chapter 3, we extend this result to second order linear parabolic, wave and Schrödinger equations. Then in Chapter 4, following the same idea, we describe how to construct high-order finite difference method for elliptic equations by replacing the coefficients with their piecewise Q^k interpolant and analyze its accuracy. In Chapter 5, we show that the discrete maximum principle can be proven for the method constructed in Chapter 2 in the case $k = 2$ under some mesh constraint when solving the variable coefficient Poisson equations. In Chapter 6, we present a class of high-order bound-preserving compact finite difference methods.

2. SUPERCONVERGENCE OF SPECTRAL ELEMENT METHOD FOR ELLIPTIC EQUATIONS

In this chapter, we analyze the accuracy of spectral element method for the elliptic equations with Dirichlet boundary conditions. The classical spectral element method with Lagrangian Q^k basis reduces to a finite difference scheme when all the integrals are approximated by the $(k+1)\times(k+1)$ Gauss-Lobatto quadrature. We prove that this finite difference scheme is $(k+2)$-th order accurate in the discrete 2-norm for the elliptic equations with Dirichlet boundary conditions, which is a superconvergence result of function values. We also give a convenient implementation for the case $k=2$, which is a simple fourth order accurate elliptic solver on a rectangular domain.

2.1 Introduction

2.1.1 Motivation

In this chapter we consider solving a two-dimensional elliptic equation with smooth coefficients on a rectangular domain by high order finite difference schemes, which are constructed via using suitable quadrature in the classical continuous finite element method on a rectangular mesh. Consider the following model problem as an example: a variable coefficient Poisson equation $-\nabla\cdot(a(\mathbf{x})\nabla u)=f, a(\mathbf{x})>0$ on a square domain $\Omega=(0,1)\times(0,1)$ with homogeneous Dirichlet boundary conditions. The variational form is to find $u\in H_0^1(\Omega)=\{v\in H^1(\Omega): v|_{\partial\Omega}=0\}$ satisfying

$$A(u,v)=(f,v),\quad \forall v\in H_0^1(\Omega),$$

where $A(u,v)=\iint_\Omega a\nabla u\cdot\nabla v dxdy$, $(f,v)=\iint_\Omega fv dxdy$. Let h be the mesh size of an uniform rectangular mesh and $V_0^h\subseteq H_0^1(\Omega)$ be the continuous finite element space consisting of piecewise Q^k polynomials (i.e., tensor product of piecewise polynomials of degree k), then the C^0-Q^k finite element solution is defined as $u_h\in V_0^h$ satisfying

$$A(u_h,v_h)=(f,v_h),\quad \forall v_h\in V_0^h. \tag{2.1}$$

Standard error estimates of (2.1) are $\|u-u_h\|_1 \le Ch^k\|u\|_{k+1}$ and $\|u-u_h\|_0 \le Ch^{k+1}\|u\|_{k+1}$ where $\|\cdot\|_k$ denotes $H^k(\Omega)$-norm, see [3]. For $k \ge 2$, $\mathcal{O}(h^{k+1})$ superconvergence for the gradient at Gauss quadrature points and $\mathcal{O}(h^{k+2})$ superconvergence for functions values at Gauss-Lobatto quadrature points were proven for one-dimensional case in [4]–[6] and for two-dimensional case in [7]–[10].

When implementing the scheme (2.1), integrals are usually approximated by quadrature. The most convenient implementation is to use $(k+1)\times(k+1)$ Gauss-Lobatto quadrature because they not only are superconvergence points but also can define all the degree of freedoms of Lagrangian Q^k basis. See Figure 1.1 for the case $k=2$. Such a quadrature scheme can be denoted as finding $u_h \in V_0^h$ satisfying

$$A_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h, \tag{2.2}$$

where $A_h(u_h, v_h)$ and $\langle f, v_h\rangle_h$ denote using tensor product of $(k+1)$-point Gauss-Lobatto quadrature for integrals $A(u_h, v_h)$ and (f, v_h) respectively.

It is well known that many classical finite difference schemes are exactly finite element methods with specific quadrature scheme, see [3]. We will write scheme (2.2) as an exact finite difference type scheme in Section 2.9 for $k=2$. Such a finite difference scheme not only provides an efficient and also convenient way for assembling the stiffness matrix especially for a variable coefficient problem, but also with has advantages inherited from the variational formulation, such as symmetry of stiffness matrix and easiness of handling boundary conditions in high order schemes. This is the variational approach to construct a high order accurate finite difference scheme .

2.1.2 Superconvergence Of C^0-Q^k Finite Element Method

Standard error estimates of (2.1) are $\|u-u_h\|_1 \le Ch^k\|u\|_{k+1}$ and $\|u-u_h\|_0 \le Ch^{k+1}\|u\|_{k+1}$ [3]. At certain quadrature or symmetry points the finite element solution or its derivatives have higher order accuracy, which is called superconvergence. Douglas and Dupont first proved that continuous finite element method using piecewise polynomial of degree k has $O(h^{2k})$ convergence at the knots in an one dimensional mesh [24], [25]. In [25], $O(h^{2k})$ was proven to be the best possible convergence rate. For $k \geq 2$, $\mathcal{O}(h^{k+1})$ for the derivatives at Gauss quadrature points and $\mathcal{O}(h^{k+2})$ for functions values at Gauss-Lobatto quadrature points were proven in [4]–[6].

For two dimensional cases, it was first showed in [26] that the $(k+2)$-th order superconvergence for $k \geq 2$ at vertices of all rectangular cells in a two dimensional rectangular mesh. Namely, the convergence rate at the knots is as least one order higher than the rate globally. Later on, the $2k$-th order (for $k \geq 2$) convergence rate at the knots was proven for Q^k elements solving $-\Delta u = f$, see [27], [28].

For the multi-dimensional variable coefficient case, when discussing the superconvergence of derivatives, it can be reduced to the Laplacian case. Superconvergence of tensor product elements for the Laplacian case can be established by extending one-dimensional results [8], [26]. See also [29] for the superconvergence of the gradient. The superconvergence of function values in rectangular elements for the variable coefficient case were studied in [9] by Chen with M-type projection polynomials and in [10] by Lin and Yan with the point-line-plane interpolation polynomials. In particular, let Z_0 denote the set of tensor product of $(k+1)$-point Gauss-Lobatto quadrature points for all rectangular cells, then the following superconvergence of function values for Q^k elements was shown in [9]:

$$\begin{aligned}\left(h^2 \sum_{(x,y)\in Z_0} |u(x,y)-u_h(x,y)|^2\right)^{1/2} &\leq Ch^{k+2}\|u\|_{k+2}, \quad k \geq 2, &(2.3)\\ \max_{(x,y)\in Z_0} |u(x,y)-u_h(x,y)| &\leq Ch^{k+2}|\ln h|\|u\|_{k+2,\infty,\Omega}, \quad k \geq 2. &(2.4)\end{aligned}$$

Classical quadrature error estimates imply that standard finite element error estimates still hold for (2.2), see [3], [30]. The focus of this chapter is to prove that the superconvergence of function values at Gauss-Lobatto points still holds with the Gauss-Lobatto quadrature. To be more specific, for Dirichlet type boundary conditions, we will show that (2.2) with $k \geq 2$ is a $(k+2)$-th order accurate finite difference scheme in the discrete 2-norm under suitable smoothness assumptions on the exact solution and the coefficients.

In this chapter, the main motivation to study superconvergence is to use it for constructing $(k+2)$-th order accurate finite difference schemes. For such a task, superconvergence points should define all degree of freedoms over the whole computational domain including boundary points. For high order finite element methods, this seems possible only on quite structured meshes such as rectangular meshes for a rectangular domain and equilateral triangles for a hexagonal domain, even though there are numerous superconvergence results for interior cells in unstructured meshes.

2.1.3 Related Work And Difficulty In Using Standard Tools

To illustrate our perspectives and difficulties, we focus on the case $k = 2$ in the following. For computing the bilinear form in the scheme (2.1), another convenient implementation is to replace the smooth coefficient $a(x, y)$ by a piecewise Q^2 polynomial $a_I(x, y)$ obtained by interpolating $a(x, y)$ at the quadrature points in each cell shown in Figure 1.1. Then one can compute the integrals in the bilinear form exactly since the integrand is a polynomial. Superconvergence of function values for such an approximated coefficient scheme was proven in Chapter 5 and the proof can be easily extended to higher order polynomials and three-dimensional cases. This result might seem surprising since interpolation error $a(x, y) - a_I(x, y)$ is of third order. On the other hand, all the tools used in Chapter 4 are standard in the literature.

From a practical point of view, (2.2) is interesting and practical since it gives a genuine finite difference scheme. It is straightforward to use standard tools in the literature for showing superconvergence still holds for accurate enough quadrature. Even though the 3×3 Gauss-Lobatto quadrature is fourth order accurate, the standard quadrature error estimates cannot be used directly to establish the fourth order accuracy of (2.2), as will be explained in detail in Remark 2.3.10 in Section 2.3.2.

We can also rewrite (2.2) for $k = 2$ as a finite difference scheme but its local truncation error is only second order as will be shown in Section 2.9.4. The phenomenon that truncation errors have lower orders was named *supraconvergence* in the literature. The second order truncation error makes it difficult to establish the fourth order accuracy following any traditional finite difference analysis approaches.

To construct high order finite difference schemes from variational formulation, we can also consider finite element method with P^2 basis on a regular triangular mesh in which two adjacent triangles form a rectangle [31]. Superconvergence of function values in C^0-P^2 finite element method at the three vertices and three edge centers can be proven [8], [9]. See also [32]. Even though the quadrature using only three edge centers is third order accurate, error cancellations happen on two adjacent triangles forming a rectangle, thus fourth order accuracy of the corresponding finite difference scheme is still possible. However, extensions to construct higher order finite difference schemes are much more difficult.

The main contribution is to give the proof of the $(k+2)$-th order accuracy of (2.2) with $k \geq 2$, which is an easy construction of high order finite difference schemes for variable coefficient problems. An important step is to obtain desired sharp quadrature estimate for the bilinear form, for which it is necessary to count in quadrature error cancellations between neighboring cells. Conventional quadrature estimating tools such as the Bramble-Hilbert Lemma only give the sharp estimate on each cell thus cannot be used directly. A key technique in this chapter is to apply the Bramble-Hilbert Lemma after integration by parts on proper interpolation polynomials to allow error cancellations.

In Section 2.2, we introduce our notations and assumptions. In Section 2.3, standard quadrature estimates are reviewed. Superconvergence of bilinear forms with quadrature is shown in Section 2.5. Then we prove the main result for homogeneous Dirichlet boundary conditions in Section 2.6 and for nonhomogeneous Dirichlet boundary conditions in Section 2.7. The Neumann boundary condition case is in Section 2.8. Section 2.9 provides a simple finite difference implementation of (2.2). Section 2.10 contains numerical tests. Concluding remarks are given in Section 2.11.

2.2 Notations And Assumptions

2.2.1 Notations and basic tools

Except the notations in the introduction, we have the following notations and common tools.

- We only consider a rectangular domain $\Omega = (0,1) \times (0,1)$ with its boundary denoted as $\partial\Omega$.

- Only for convenience, we assume Ω_h is an uniform rectangular mesh for $\bar{\Omega}$ and $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$ denotes any cell in Ω_h with cell center (x_e, y_e). The assumption of an uniform mesh is not essential to the discussion of superconvergence. All superconvergence results in this chapter can be easily extended to continuous finite element method with Q^k element on a quasi-uniform rectangular mesh, but not on a generic quadrilateral mesh or any curved mesh.

- $Q^k(e) = \left\{ p(x,y) = \sum\limits_{i=0}^{k} \sum\limits_{j=0}^{k} p_{ij} x^i y^j, (x,y) \in e \right\}$ is the set of tensor product of polynomials of degree k on a cell e.

- $V^h = \{p(x,y) \in C^0(\Omega_h) : p|_e \in Q^k(e), \quad \forall e \in \Omega_h\}$ denotes the continuous piecewise Q^k finite element space on Ω_h.

- $V_0^h = \{v_h \in V^h : v_h|_{\partial\Omega} = 0\}$.

- The norm and seminorms for $W^{k,p}(\Omega)$ and $1 \le p < +\infty$, with standard modification for $p = +\infty$:
$$\|u\|_{k,p,\Omega} = \left(\sum_{i+j \le k} \iint_\Omega |\partial_x^i \partial_y^j u(x,y)|^p dxdy \right)^{1/p},$$
$$|u|_{k,p,\Omega} = \left(\sum_{i+j = k} \iint_\Omega |\partial_x^i \partial_y^j u(x,y)|^p dxdy \right)^{1/p},$$
$$[u]_{k,p,\Omega} = \left(\iint_\Omega |\partial_x^k u(x,y)|^p dxdy + \iint_\Omega |\partial_y^k u(x,y)|^p dxdy \right)^{1/p}.$$
Notice that $[u]_{k+1,p,\Omega} = 0$ if u is a Q^k polynomial.

- For simplicity, sometimes we may use $\|u\|_{k,\Omega}$, $|u|_{k,\Omega}$ and $[u]_{k,\Omega}$ denote norm and seminorms for $H^k(\Omega) = W^{k,2}(\Omega)$.

- When there is no confusion, Ω may be dropped in the norm and seminorms, e.g., $\|u\|_k = \|u\|_{k,2,\Omega}$.

- For any $v_h \in V^h$, $1 \leq p < +\infty$ and $k \geq 1$, we will abuse the notation to denote the broken Sobolev norm and seminorms by the following symbols

$$\|v_h\|_{k,p,\Omega} := \left(\sum_e \|v_h\|_{k,p,e}^p\right)^{\frac{1}{p}}, \quad |v_h|_{k,p,\Omega} := \left(\sum_e |v_h|_{k,p,e}^p\right)^{\frac{1}{p}}, \quad [v_h]_{k,p,\Omega} := \left(\sum_e [v_h]_{k,p,e}^p\right)^{\frac{1}{p}}.$$

- Let $Z_{0,e}$ denote the set of $(k+1) \times (k+1)$ Gauss-Lobatto points on a cell e.
- $Z_0 = \bigcup_e Z_{0,e}$ denotes all Gauss-Lobatto points in the mesh Ω_h.
- Let $\|u\|_{l^2(\Omega)}$ and $\|u\|_{l^\infty(\Omega)}$ denote the discrete 2-norm and the maximum norm over Z_0 respectively:

$$\|u\|_{l^2(\Omega)} = \left[h^2 \sum_{(x,y)\in Z_0} |u(x,y)|^2\right]^{\frac{1}{2}}, \quad \|u\|_{l^\infty(\Omega)} = \max_{(x,y)\in Z_0} |u(x,y)|.$$

- When there is no confusion, for simplicity, sometimes we may use $\|u\|_{l^2}$ and $|u|_{l^\infty}$ to denote $\|u\|_{l^2(\Omega)}$ and $\|u\|_{l^\infty(\Omega)}$ respectively.
- For a continuous function $f(x,y)$, let $f_I(x,y)$ denote its piecewise Q^k Lagrange interpolant at $Z_{0,e}$ on each cell e, i.e., $f_I \in V^h$ satisfies:

$$f(x,y) = f_I(x,y), \quad \forall (x,y) \in Z_0.$$

- $P^k(t)$ denotes the set of polynomial of degree k of variable t.
- $(f,v)_e$ denotes the inner product in $L^2(e)$ and (f,v) denotes the inner product in $L^2(\Omega)$:

$$(f,v)_e = \iint_e fv\,dxdy, \quad (f,v) = \iint_\Omega fv\,dxdy = \sum_e (f,v)_e.$$

- $\langle f,v\rangle_{e,h}$ denotes the approximation to $(f,v)_e$ by using $(k+1)\times(k+1)$-point Gauss Lobatto quadrature with $k \geq 2$ for integration over cell e.
- $\langle f,v\rangle_h$ denotes the approximation to (f,v) by using $(k+1)\times(k+1)$-point Gauss Lobatto quadrature with $k \geq 2$ for integration over each cell e.

- $\hat{K} = [-1,1] \times [-1,1]$ denotes a reference cell.
- For $f(x,y)$ defined on e, consider $\hat{f}(s,t) = f(sh + x_e, th + y_e)$ defined on $\hat{K}$. Let $\hat{f}_I$ denote the Q^k Lagrange interpolation of $\hat{f}$ at the $(k+1) \times (k+1)$ Gauss Lobatto quadrature points on $\hat{K}$.
- $(\hat{f}, \hat{v})_{\hat{K}} = \iint_{\hat{K}} \hat{f}\hat{v}\, dsdt$.
- $\langle \hat{f}, \hat{v} \rangle_{\hat{K}}$ denotes the approximation to $(\hat{f}, \hat{v})_{\hat{K}}$ by using $(k+1) \times (k+1)$-point Gauss-Lobatto quadrature.
- On the reference cell $\hat{K}$, for convenience we use the superscript h over the ds or dt to denote we use $(k+1)$-point Gauss-Lobatto quadrature for the corresponding variable. For example,

  $$\iint_{\hat{K}} \hat{f} d^h s dt = \int_{-1}^{1} [w_1 \hat{f}(-1,t) + w_{k+1} \hat{f}(1,t) + \sum_{i=2}^{k} w_i \hat{f}(x_i, t)] dt.$$

  Since $(\hat{f}\hat{v})_I$ coincides with $\hat{f}\hat{v}$ at the quadrature points, we have

  $$\iint_{\hat{K}} (\hat{f}\hat{v})_I dx dy = \iint_{\hat{K}} (\hat{f}\hat{v})_I d^h x d^h y = \iint_{\hat{K}} \hat{f}\hat{v} d^h x d^h y = \langle \hat{f}, \hat{v} \rangle_{\hat{K}}.$$

- On the domain Ω, for convenience we use the superscript h over the dx or dy to denote we use $(k+1)$-point Gauss-Lobatto quadrature for the corresponding variable on each cell. For example, we have

  $$\iint_{\Omega} (fv)_I dx dy = \iint_{\Omega} (fv)_I d^h x d^h y = \iint_{\Omega} fv d^h x d^h y = \langle f, v \rangle_h.$$

The following are commonly used tools and facts:

- For n-dimensional problems, the following scaling argument will be used:

  $$h^{k-n/p} |v|_{k,p,e} = |\hat{v}|_{k,p,\hat{K}}, \quad h^{k-n/p} [v]_{k,p,e} = [\hat{v}]_{k,p,\hat{K}}, \quad 1 \leq p \leq \infty. \tag{2.5}$$

- There exist constants C_i $(i = 1, 2, 3, 4)$ independent of h such that l^2-norm and L^2-norm are equivalent for V^h:

$$\begin{aligned} C_1\|v_h\|_{l^2(\Omega)} \le \|v_h\|_0 \le C_2\|v_h\|_{l^2(\Omega)}, \quad \forall v \in V^h, \\ C_3\langle v_h, v_h\rangle_h \le \|v_h\|_0^2 \le C_4\langle v_h, v_h\rangle_h, \quad \forall v \in V^h. \end{aligned} \tag{2.6}$$

- Inverse estimates for polynomials:

$$\|v_h\|_{k+1,e} \le Ch^{-1}\|v_h\|_{k,e}, \quad \forall v_h \in V^h, k \ge 0. \tag{2.7}$$

- Sobolev's embedding in two and three dimensions: $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$.

- The embedding implies

$$\|\hat{f}\|_{0,\infty,\hat{K}} \le C\|\hat{f}\|_{k,2,\hat{K}}, \quad \forall \hat{f} \in H^k(\hat{K}), k \ge 2,$$

$$\|\hat{f}\|_{1,\infty,\hat{K}} \le C\|\hat{f}\|_{k+1,2,\hat{K}}, \quad \forall \hat{f} \in H^{k+1}(\hat{K}), k \ge 2.$$

- Cauchy-Schwarz inequalities in two dimensions:

$$\sum_e \|u\|_{k,e}\|v\|_{k,e} \le \left(\sum_e \|u\|_{k,e}^2\right)^{\frac{1}{2}} \left(\sum_e \|v\|_{k,e}^2\right)^{\frac{1}{2}}, \quad \|u\|_{k,1,e} = \mathcal{O}(h)\|u\|_{k,2,e}.$$

- Poincaré inequality: let $\bar{u}$ be the average of $u \in H^1(\Omega)$ on Ω, then

$$|u - \bar{u}|_{0,p,\Omega} \le C|\nabla u|_{0,p,\Omega}, \quad p \ge 1.$$

  If $\bar{u}$ is the average of $u \in H^1(e)$ on a cell e, we have

$$|u - \bar{u}|_{0,p,e} \le Ch|\nabla u|_{0,p,e}, \quad p \ge 1.$$

- For $k \ge 2$, the $(k+1) \times (k+1)$ Gauss-Lobatto quadrature is exact for integration of polynomials of degree $2k - 1 \ge k + 1$ on $\hat{K}$.

- Define the projection operator $\hat{\Pi}_1 : \hat{u} \in L^1(\hat{K}) \to \hat{\Pi}_1 \hat{u} \in Q^1(\hat{K})$ by

$$\iint_{\hat{K}} (\hat{\Pi}_1 \hat{u}) w ds dt = \iint_{\hat{K}} \hat{u} w ds dt, \quad \forall w \in Q^1(\hat{K}). \tag{2.8}$$

  Notice that all degree of freedoms of $\hat{\Pi}_1 \hat{u}$ can be represented as a linear combination of $\iint_{\hat{K}} \hat{u}(s,t) p(s,t) ds dt$ for $p(s,t) = 1, s, t, st$, thus the $H^1(\hat{K})$ (or $H^2(\hat{K})$) norm of $\hat{\Pi}_1 \hat{u}$ are determined by $\iint_{\hat{K}} \hat{u}(s,t) p(s,t) ds dt$. By Cauchy-Schwarz inequality $|\iint_{\hat{K}} \hat{u}(s,t) \hat{p}(s,t) ds dt| \leq \|\hat{u}\|_{0,2,\hat{K}} \|\hat{p}\|_{0,2,\hat{K}} \leq C \|\hat{u}\|_{0,2,\hat{K}}$, we have $\|\Pi_1 \hat{u}\|_{1,2,\hat{K}} \leq C \|\hat{u}\|_{0,2,\hat{K}}$, which means $\hat{\Pi}_1$ is a continuous linear mapping from $L^2(\hat{K})$ to $H^1(\hat{K})$. By a similar argument, one can show $\hat{\Pi}_1$ is a continuous linear mapping from $L^2(\hat{K})$ to $H^2(\hat{K})$.

2.2.2 Coercivity and elliptic regularity

We consider the elliptic variational problem of finding $u \in H_0^1(\Omega)$ to satisfy

$$A(u,v) := \iint_{\Omega} (\nabla v^T \mathbf{a} \nabla u + \mathbf{b} \nabla u v + c u v) \, dx dy = (f, v), \forall v \in H_0^1(\Omega), \tag{2.9}$$

where $\mathbf{a} = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$ is real symmetric positive definite and $\mathbf{b} = [b^1 \quad b^2]$. Assume the coefficients $\mathbf{a}$, $\mathbf{b}$ and c are smooth with uniform upper bounds, thus $A(u,v) \leq C\|u\|_1 \|v\|_1$ for any $u, v \in H_0^1(\Omega)$. We denote $\lambda_{\mathbf{a}}$ as the smallest eigenvalues of $\mathbf{a}$. Assume $\lambda_{\mathbf{a}}$ has a positive lower bound and $\nabla \cdot \mathbf{b} \leq 2c$, so that coercivity of the bilinear form can be easily achieved. Since

$$(\mathbf{b} \cdot \nabla u, v) = \int_{\partial\Omega} uv \mathbf{b} \cdot \mathbf{n} ds - (\nabla \cdot (v\mathbf{b}), u) = \int_{\partial\Omega} uv \mathbf{b} \cdot \mathbf{n} ds - (\mathbf{b} \cdot \nabla v, u) - (v \nabla \cdot \mathbf{b}, u),$$

we have

$$2(\mathbf{b} \cdot \nabla v, v) + 2(cv, v) = \int_{\partial\Omega} v^2 \mathbf{b} \cdot \mathbf{n} ds + ((2c - \nabla \cdot \mathbf{b}) v, v) \geq 0, \quad \forall v \in H_0^1(\Omega). \tag{2.10}$$

By the equivalence of two norms $|\cdot|_1$ and $\|\cdot\|_1$ for the space $H_0^1(\Omega)$ (see [3]), we conclude that the bilinear form $A(u,v) = (\mathbf{a}\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (cu, v)$ satisfies coercivity $A(v,v) \geq C\|v\|_1$ for any $v \in H_0^1(\Omega)$.

The coercivity can also be achieved if we assume $|\mathbf{b}| < 4\lambda_{\mathbf{a}}c$. By Young's inequality

$$|(\mathbf{b}\cdot\nabla v, v)| \leq \iint_\Omega \frac{|\mathbf{b}\cdot\nabla v|^2}{4c} + c|v|^2 dxdy \leq \left(\frac{|\mathbf{b}|^2}{4c}\nabla v, \nabla v\right) + (cv, v),$$

we have

$$A(v,v) \geq (\mathbf{a}\nabla v, \nabla v) + (cv, v) - |(\mathbf{b}\cdot\nabla v, v)| \geq \left((\lambda_{\mathbf{a}} - \frac{|\mathbf{b}|^2}{4c})\nabla v, \nabla v\right) > 0, \quad \forall v \in H_0^1(\Omega). \tag{2.11}$$

Let A^* be the dual operator of A, i.e., $A^*(u,v) = A(v,u)$. We need to assume the elliptic regularity holds for the dual problem of (2.9) :

$$w \in H_0^1(\Omega), A^*(w,v) = (f,v), \quad \forall v \in H_0^1(\Omega) \Longrightarrow \|w\|_2 \leq C\|f\|_0, \tag{2.12}$$

where C is independent of w and f. See [33], [34] for the elliptic regularity with Lipschitz continuous coefficients on a Lipschitz domain.

2.3 Quadrature Error Estimates

In the following, we will use $\hat{}$ for a function to emphasize the function is defined on or transformed to the reference cell $\hat{K} = [-1,1] \times [-1,1]$ from a mesh cell.

2.3.1 Standard estimates

By the abstract Bramble-Hilbert Lemma in [35], with the result $\|v\|_{m,p,\Omega} \leq C(|v|_{0,p,\Omega} + [v]_{m,p,\Omega})$ for any $v \in W^{m,p}(\Omega)$ [36], [37], the Bramble-Hilbert Lemma for Q^k polynomials can be stated as (see Exercise 3.1.1 and Theorem 4.1.3 in [38]):

Theorem 2.3.1. *If a continuous linear mapping $\hat{\Pi} : H^{k+1}(\hat{K}) \to H^{k+1}(\hat{K})$ satisfies $\hat{\Pi}\hat{v} = \hat{v}$ for any $\hat{v} \in Q^k(\hat{K})$, then*

$$\|\hat{u} - \hat{\Pi}\hat{u}\|_{k+1,\hat{K}} \leq C[\hat{u}]_{k+1,\hat{K}}, \quad \forall \hat{u} \in H^{k+1}(\hat{K}). \tag{2.13}$$

Thus if $l(\cdot)$ is a continuous linear form on the space $H^{k+1}(\hat{K})$ satisfying $l(\hat{v}) = 0, \forall \hat{v} \in Q^k(\hat{K})$, then

$$|l(\hat{u})| \leq C\|l\|_{k+1,\hat{K}}[\hat{u}]_{k+1,\hat{K}}, \quad \forall \hat{u} \in H^{k+1}(\hat{K}),$$

where $\|l\|_{k+1,\hat{K}}$ is the norm in the dual space of $H^{k+1}(\hat{K})$.

For Q^k element ($k \geq 2$), consider $(k+1) \times (k+1)$ Gauss-Lobatto quadrature, which is exact for integration of Q^{2k-1} polynomials.

It is straightforward to establish the interpolation error:

Theorem 2.3.2. *For a smooth function a, $|a - a_I|_{0,\infty,\Omega} = \mathcal{O}(h^{k+1})|a|_{k+1,\infty,\Omega}$.*

Let s_j, t_j and w_j ($j = 1, \cdots, k+1$) be the Gauss-Lobatto quadrature points and weight for the interval $[-1, 1]$. Notice $\hat{f}$ coincides with its Q^k interpolant $\hat{f}_I$ at the quadrature points and the quadrature is exact for integration of $\hat{f}_I$, the quadrature can be expressed on $\hat{K}$ as

$$\sum_{i=1}^{k+1}\sum_{j=1}^{k+1} \hat{f}(s_i, t_j)w_i w_j = \iint_{\hat{K}} \hat{f}_I(x, y)dxdy,$$

thus the quadrature error is related to interpolation error:

$$\iint_{\hat{K}} \hat{f}(x, y)dxdy - \sum_{i=1}^{k+1}\sum_{j=1}^{k+1} \hat{f}(s_i, t_j)w_i w_j = \iint_{\hat{K}} \hat{f}(x, y)dxdy - \iint_{\hat{K}} \hat{f}_I(x, y)dxdy.$$

We have the following estimates on the quadrature error:

Theorem 2.3.3. *For $n = 2$ and a sufficiently smooth function $a(x, y)$, if $k \geq 2$ and m is an integer satisfying $k \leq m \leq 2k$, we have*

$$\iint_e a(x, y)dxdy - \iint_e a_I(x, y)dxdy = \mathcal{O}(h^{m+\frac{n}{2}})[a]_{m,e} = \mathcal{O}(h^{m+n})[a]_{m,\infty,e}.$$

Proof. Let $E(a)$ denote the quadrature error for function $a(x,y)$ on e. Let $\hat{E}(\hat{a})$ denote the quadrature error for the function $\hat{a}(s,t) = a(sh + x_e, th + y_e)$ on the reference cell $\hat{K}$. Then for any $\hat{f} \in H^m(\hat{K})$ $(m \geq k \geq 2)$, since quadrature are represented by point values, with the Sobolev's embedding we have

$$|\hat{E}(\hat{f})| \leq C|\hat{f}|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{m,2,\hat{K}}.$$

Thus $\hat{E}(\cdot)$ is a continuous linear form on $H^m(\hat{K})$ and $\hat{E}(\hat{f}) = 0$ if $\hat{f} \in Q^{m-1}(\hat{K})$. With (2.5), the Bramble-Hilbert lemma implies

$$|E(a)| = h^n|\hat{E}(\hat{a})| \leq Ch^n[\hat{a}]_{m,2,\hat{K}} = \mathcal{O}(h^{m+\frac{n}{2}})[a]_{m,2,e} = \mathcal{O}(h^{m+n})[a]_{m,\infty,e}.$$

□

Theorem 2.3.4. *If $k \geq 2$, $(f, v_h) - \langle f, v_h \rangle_h = \mathcal{O}(h^{k+2})\|f\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V^h$.*

Proof. This result is a special case of Theorem 5 in [30]. For completeness, we include a proof. Let $\hat{E}(\cdot)$ denote the quadrature error term on the reference cell $\hat{K}$. Consider the projection (2.8). Let Π_1 denote the same projection on e. Since $\hat{\Pi}_1$ leaves $Q^0(\hat{K})$ invariant, by the Bramble-Hilbert lemma on $\hat{\Pi}_1$, we get $[\hat{v}_h - \hat{\Pi}_1\hat{v}_h]_{1,\hat{K}} \leq \|\hat{v}_h - \hat{\Pi}_1\hat{v}_h\|_{1,\hat{K}} \leq C[\hat{v}_h]_{1,\hat{K}}$ thus $[\hat{\Pi}_1\hat{v}_h]_{1,\hat{K}} \leq [\hat{v}_h]_{1,\hat{K}} + [\hat{v}_h - \hat{\Pi}_1\hat{v}_h]_{1,\hat{K}} \leq C[\hat{v}_h]_{1,\hat{K}}$. By setting $w = \hat{\Pi}_1\hat{v}_h$ in (2.8), we get $|\hat{\Pi}_1\hat{v}_h|_{0,\hat{K}} \leq |\hat{v}_h|_{0,\hat{K}}$. For $k \geq 2$, repeat the proof of Theorem 2.3.3, we can get

$$|\hat{E}(\hat{f}\hat{\Pi}_1\hat{v}_h)| \leq C[\hat{f}\hat{\Pi}_1\hat{v}_h]_{k+2,\hat{K}} \leq C([\hat{f}]_{k+2,\hat{K}}|\hat{\Pi}_1\hat{v}_h|_{0,\infty,\hat{K}} + [\hat{f}]_{k+1,\hat{K}}|\hat{\Pi}_1\hat{v}_h|_{1,\infty,\hat{K}}),$$

where the fact $[\hat{\Pi}_1\hat{v}_h]_{l,\infty,\hat{K}} = 0$ for $l \geq 2$ is used. The equivalence of norms over $Q^1(\hat{K})$ implies

$$\begin{aligned}|\hat{E}(\hat{f}\hat{\Pi}_1\hat{v}_h)| &\leq C([\hat{f}]_{k+2,\hat{K}}|\hat{\Pi}_1\hat{v}_h|_{0,\hat{K}} + [\hat{f}]_{k+1,\hat{K}}|\hat{\Pi}_1\hat{v}_h|_{1,\hat{K}}) \\ &\leq C([\hat{f}]_{k+2,\hat{K}}|\hat{v}_h|_{0,\hat{K}} + [\hat{f}]_{k+1,\hat{K}}|\hat{v}_h|_{1,\hat{K}}).\end{aligned}$$

Next consider the linear form $\hat{f} \in H^k(\hat{K}) \to \hat{E}(\hat{f}(\hat{v}_h - \hat{\Pi}_1\hat{v}_h))$. Due to the embedding $H^k(\hat{K}) \hookrightarrow C^0(\hat{K})$, it is continuous with operator norm $\leq C\|\hat{v}_h - \hat{\Pi}_1\hat{v}_h\|_{0,\hat{K}}$ since

$$\begin{aligned}|\hat{E}(\hat{f}(\hat{v}_h - \hat{\Pi}_1\hat{v}_h))| \leq C|\hat{f}(\hat{v}_h - \hat{\Pi}_1\hat{v}_h)|_{0,\infty,\hat{K}} &\leq C|\hat{f}|_{0,\infty,\hat{K}}|\hat{v}_h - \hat{\Pi}_1\hat{v}_h|_{0,\infty,\hat{K}} \\ &\leq C\|\hat{f}\|_{k,\hat{K}}\|\hat{v}_h - \hat{\Pi}_1\hat{v}_h\|_{0,\hat{K}}.\end{aligned}$$

For any $\hat{f} \in Q^{k-1}(\hat{K})$, $\hat{E}(\hat{f}\hat{v}_h) = 0$. By the Bramble-Hilbert lemma, we get

$$|\hat{E}(\hat{f}(\hat{v}_h - \hat{\Pi}_1\hat{v}_h))| \leq C[\hat{f}]_{k,\hat{K}}\|\hat{v}_h - \hat{\Pi}_1\hat{v}_h\|_{0,\hat{K}} \leq C[\hat{f}]_{k,\hat{K}}[\hat{v}_h]_{2,\hat{K}}.$$

So on a cell e, with (2.5), we get

$$E(fv_h) = h^n\hat{E}(\hat{f}\hat{v}_h) = Ch^{k+2}([f]_{k+2,e}|v_h|_{0,e} + [f]_{k+1,e}|v_h|_{1,e} + [f]_{k,e}[v_h]_{2,e}).$$

Summing over e and use Cauchy-Schwarz inequality, we get the desired result. □

Remark 2.3.5. By the Theorem 2.3.1, on the reference cell $\hat{K}$, for $a(x,y) \in H^{k+2}(e)$ and $k \geq 2$, we have

$$\iint_{\hat{K}} \hat{a}(s,t) - \hat{a}_I(s,t)dsdt \leq C[\hat{a}]_{k+2,\hat{K}} \leq C[\hat{a}]_{k+2,\infty,\hat{K}}, \tag{2.14}$$

and

$$\|\hat{a} - \hat{a}_I\|_{k+1,\hat{K}} \leq C[\hat{a}]_{k+1,\hat{K}}. \tag{2.15}$$

Lemma 2.3.6. *If $g \in H^{k+3}(\partial\Omega)$ and $v_h \in V^h$, then for $k \geq 3$, we have*

$$\int_{\partial\Omega}(g - g_I)v_h d\mu = \mathcal{O}(h^{k+2.5})\|g\|_{k+3,\partial\Omega}\|v_h\|_2.$$

Proof. Note

$$\int_{\partial\Omega}(g - g_I)v_h d\mu = \int_{\partial\Omega}(g - g_I)(v_h - \Pi_1 v_h)d\mu + \int_{\partial\Omega}(g - g_I)\Pi_1 v_h d\mu = I + II.$$

In the following, we will focus on the left boundary L_2 instead of $\partial\Omega$ since the same estimate can be applied to the top boundary L_1, bottom boundary L_3 and right boundary L_4 as well. Assume the left boundary of cell e is denoted as l_2^e.

For part I, by the Bramble-Hilbert Lemma, we have

$$\begin{aligned}
&\int_{L^2}(g-g_I)(v_h-\Pi_1 v_h)(-1,y)dy\\
=&h\sum_{e\cap L_2\neq\emptyset}\int_{-1}^{1}(\hat{g}-\hat{g}_I)(\hat{v}_h-\hat{\Pi}_1\hat{v}_h)(-1,t)dt\\
\leq&h\sum_{e\cap L_2\neq\emptyset}\left(\int_{-1}^{1}|\hat{g}-\hat{g}_I|^2(-1,t)dt\right)^{\frac{1}{2}}\left(\int_{-1}^{1}|\hat{v}_h-\hat{\Pi}_1\hat{v}_h|^2(-1,t)dt\right)^{\frac{1}{2}}\\
\leq&h\sum_{e\cap L_2\neq\emptyset}\left(\int_{-1}^{1}|\partial_t^{k+1}\hat{g}|^2(-1,t)dt\right)^{\frac{1}{2}}\left(\int_{-1}^{1}|\partial_t^{2}\hat{v}_h|^2(-1,t)dt\right)^{\frac{1}{2}}=\mathcal{O}(h^{k+3})\sum_{e\cap L_2\neq\emptyset}|g|_{k+1,l_2^e}|v_h|_{2,l_2^e}.
\end{aligned}$$

For part II, let $\hat{E}_1(\cdot)$ denote the quadrature error term on the reference cell $\hat{K}_1=[-1,1]$, then

$$\int_{L_2}(g-g_I)\Pi_1 v_h dy=\int_{L_2}(g-g_I)\Pi_1 v_h d\mu-\int_{L_2}(g-g_I)\Pi_1 v_h d^h y=h\sum_{e\cap L_2\neq\emptyset}\hat{E}_1\left((g-g_I)\Pi_1 v_h(-1,t)\right).$$

Following the proof of Theorem 2.3.4, we have

$$\begin{aligned}
&\hat{E}_1\left((\hat{g}-\hat{g}_I)\hat{\Pi}_1\hat{v}_h(-1,t)\right)\leq C[(g-g_I)\Pi_1 v_h(-1,t)]_{k+3,\hat{K}_1}\\
\leq&C\left[\hat{g}(-1,t)]_{k+3,\hat{K}_1}|\hat{v}_h|_{0,\hat{K}_1}+[\hat{g}(-1,t)]_{k+2,\hat{K}_1}|\hat{v}_h|_{1,\hat{K}_1}\right)=\mathcal{O}(h^{k+2})\|g\|_{k+3,l_2^e}\|v_h\|_{1,l_2^e}.
\end{aligned}$$

Thus

$$\int_{L_2}(g-g_I)\Pi_1 v_h dy=\mathcal{O}(h^{k+3})\sum_{e\cap L_2\neq\emptyset}\|g\|_{k+3,l_2^e}\|v_h\|_{1,l_2^e}.$$

For polynomial $\hat{q}(s,t)\in Q^{2k}(\hat{K})$, let s_α and ω_α ($\alpha=1,2,\cdots,k+2$) denote the quadrature points and weights in $(k+2)$-point Gauss-Lobatto quadrature rule for $s\in[-1,1]$. Since $\hat{q}^2(s,t)\in Q^{2k}(\hat{K})$, $(k+2)$-point Gauss-Lobatto quadrature is exact for s-integration thus

$$\int_{-1}^{1}\int_{-1}^{1}\hat{q}^2(s,t)dsdt=\sum_{\alpha=1}^{k+2}\omega_\alpha\int_{-1}^{1}\hat{q}^2(s_\alpha,t)dt,$$

which implies

$$\int_{-1}^{1} \hat{q}^2(\pm 1, t)dt \leq C \int_{-1}^{1}\int_{-1}^{1} \hat{q}^2(s,t)dsdt, \tag{2.16}$$

thus

$$h^{\frac{1}{2}}|q|_{0,l_2^e} \leq C|q|_{0,e}. \tag{2.17}$$

Above all, we have

$$\begin{aligned}&\int_{L^2}(g-g_I)v_h(-1,y)dy\\ =&\mathcal{O}(h^{k+3}) \sum_{e\cap L_2\neq\emptyset} \|g\|_{k+3,l_2^e}\|v_h\|_{2,l_2^e} = \mathcal{O}(h^{k+2.5}) \sum_{e\cap L_2\neq\emptyset} \|g\|_{k+3,l_2^e}\|v_h\|_{2,e} = \mathcal{O}(h^{k+2.5})\|g\|_{k+3,L_2}\|v_h\|_{2,\Omega},\end{aligned}$$

which implies the theorem. □

The following two results are also standard estimates obtained by applying the Bramble-Hilbert Lemma.

Lemma 2.3.7. *If $f \in H^2(\Omega)$ or $f \in V^h$, we have $(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}(h^2)|f|_2\|v_h\|_0, \quad \forall v_h \in V^h$.*

Proof. For simplicity, we ignore the subscript in v_h. Let $E(f)$ denote the quadrature error for integrating $f(x,y)$ on e. Let $\hat{E}(\hat{f})$ denote the quadrature error for integrating $\hat{f}(s,t) = f(x_e+sh, y_e+th)$ on the reference cell $\hat{K}$. Due to the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, we have

$$|\hat{E}(\hat{f}\hat{v})| \leq C|\hat{f}\hat{v}|_{0,\infty,\hat{K}} \leq C|\hat{f}|_{0,\infty,\hat{K}}|\hat{v}|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}}.$$

Thus the mapping $\hat{f} \to E(\hat{f}\hat{v})$ is a continuous linear form on $H^2(\hat{K})$ and its norm is bounded by $C\|\hat{v}\|_{0,\hat{K}}$. If $\hat{f} \in Q^1(\hat{K})$, then we have $\hat{E}(\hat{f}\hat{v}) = 0$. By the Bramble-Hilbert Lemma Theorem 2.3.1 on this continuous linear form, we get

$$|\hat{E}(\hat{f}\hat{v})| \leq C[\hat{f}]_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}}.$$

So on a cell e, we get

$$E(fv) = h^2 \hat{E}(\hat{f}\hat{v}) \le Ch^2[\hat{f}]_{2,\hat{K}} \|\hat{v}\|_{0,\hat{K}} \le Ch^2 |f|_{2,e} \|v\|_{0,e}. \tag{2.18}$$

Summing over all elements and use Cauchy-Schwarz inequality, we get the desired result. □

Theorem 2.3.8. *Assume all coefficients of* (2.9) *are in* $W^{2,\infty}(\Omega)$*. We have*

$$A(z_h, v_h) - A_h(z_h, v_h) = \mathcal{O}(h)\|v_h\|_2 \|z_h\|_1, \quad \forall v_h, z_h \in V^h.$$

Proof. Following the same arguments as in the proof of Lemma 2.18, we have

$$E(fv) \le Ch^2 |f|_{2,e} \|v\|_{0,e}, \forall f, v \in V^h.$$

Let $f = a^{11}(v_h)_x$ and $v = (z_h)_x$ in the estimate above, we get

$$\begin{aligned} &|(a^{11}(z_h)_x, (v_h)_x) - \langle a^{11}(z_h)_x, (v_h)_x \rangle_h| \le Ch^2 \|a^{11}(v_h)_x\|_2 \|(z_h)_x\|_0 \\ \le & Ch^2 \|a^{11}\|_{2,\infty} \|v_h\|_3 |z_h|_1 \le Ch \|a^{11}\|_{2,\infty} \|v_h\|_2 |z_h|_1, \end{aligned}$$

where the inverse estimate (2.7) is used in the last inequality. Similarly, we have

$$\begin{aligned} (a^{12}(z_h)_x, (v_h)_y) - \langle a^{12}(z_h)_x, (v_h)_y \rangle_h &= Ch \|a^{12}\|_{2,\infty} \|v_h\|_2 |z_h|_1, \\ (a^{22}(z_h)_y, (v_h)_y) - \langle a^{22}(z_h)_y, (v_h)_y \rangle_h &= Ch \|a^{22}\|_{2,\infty} \|v_h\|_2 |z_h|_1, \\ (b^1(z_h)_x, v_h) - \langle b^1(z_h)_x, v_h \rangle_h &= Ch \|b^1\|_{2,\infty} \|v_h\|_2 |z_h|_0, \\ (b^2(z_h)_y, v_h) - \langle b^2(z_h)_y, v_h \rangle_h &= Ch \|b^2\|_{2,\infty} \|v_h\|_2 |z_h|_0, \\ (cz_h, v_h) - \langle cz_h, v_h \rangle_h &= Ch \|c\|_{2,\infty} \|v_h\|_1 |z_h|_0, \end{aligned}$$

which implies

$$A(z_h, v_h) - A_h(z_h, v_h) = \mathcal{O}(h)\|v_h\|_2 \|z_h\|_1.$$

□

2.3.2 A refined consistency error

In this subsection, we will show how to establish the desired consistency error estimate for smooth enough coefficients:

$$A(u,v_h)-A_h(u,v_h)=\begin{cases}\mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h\\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h\end{cases}.$$

Theorem 2.3.9. *Assume $a(x,y)\in W^{k+2,\infty}(\Omega)$, $u\in H^{k+3}(\Omega)$, $k\geq 2$, then*

$$(a\partial_x u,\partial_x v_h)-\langle a\partial_x u,\partial_x v_h\rangle_h=\begin{cases}\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h, \quad (2.19a)\\ \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h, \quad (2.19b)\end{cases}$$

$$(a\partial_x u,\partial_y v_h)-\langle a\partial_x u,\partial_y v_h\rangle_h=\begin{cases}\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h, \quad (2.20a)\\ \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h, \quad (2.20b)\end{cases}$$

$$(a\partial_x u,v_h)-\langle a\partial_x u,v_h\rangle_h=\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h\in V_0^h, \tag{2.21}$$

$$(au,v_h)-\langle au,v_h\rangle_h=\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v_h\|_2, \quad \forall v_h\in V_0^h. \tag{2.22}$$

Remark 2.3.10. We emphasize that Theorem 2.3.9 cannot be proven by applying the Bramble-Hilbert Lemma directly. Consider the constant coefficient case $a(x,y)\equiv 1$ and $k=2$ as an example,

$$(\partial_x u,\partial_x v_h)-\langle\partial_x u,\partial_x v_h\rangle_h=\sum_e\left(\iint_e u_x(v_h)_x dxdy-\iint_e u_x(v_h)_x d^hxd^hy\right).$$

Since the 3×3 Gauss-Lobatto quadrature is exact for integrating Q^3 polynomials, by Theorem 2.3.1 we have

$$\left|\iint_e u_x(v_h)_x dxdy-\iint_e u_x(v_h)_x d^hxd^hy\right|=\left|\iint_{\hat K}\hat u_s(\hat v_h)_s dsdt-\iint_{\hat K}\hat u_s(\hat v_h)_s d^hsd^ht\right|\leq C[\hat u_s(\hat v_h)_s]_{4,\hat K}.$$

Notice that $\hat{v}_h$ is Q^2 thus $(\hat{v}_h)_{stt}$ does not vanish and $[(\hat{v}_h)_s]_{4,\hat{K}} \leq C|\hat{v}_h|_{3,\hat{K}}$. So by Bramble-Hilbert Lemma for Q^k polynomials, we can only get

$$\iint_e u_x(v_h)_x dxdy - \iint_e u_x(v_h)_x d^hxd^hy = \mathcal{O}(h^4)\|u\|_{5,e}\|v_h\|_{3,e}.$$

Thus by Cauchy-Schwarz inequality after summing over e, we only have

$$(\partial_x u, \partial_x v_h) - \langle \partial_x u, \partial_x v_h\rangle_h = \mathcal{O}(h^4)\|u\|_5\|v_h\|_3.$$

In order to get the desired estimate involving only the broken H^2-norm of v_h, we will take advantage of error cancellations between neighboring cells through integration by parts.

Proof. For simplicity, we ignore the subscript $_h$ of v_h in this proof and all the following v are in V^h which are Q^k polynomials in each cell. First, by Theorem 2.3.4, we easily obtain (2.21) and (2.22):

$$(au_x, v) - \langle au_x, v\rangle_h = \mathcal{O}(h^{k+2})\|au_x\|_{k+2}\|v\|_2 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2,$$

$$(au, v) - \langle au, v\rangle_h = \mathcal{O}(h^{k+2})\|au\|_{k+2}\|v\|_2 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v\|_2.$$

We will only discuss $(au_x, v_x) - \langle au_x, v_x\rangle_h$ and the same discussion also applies to derive (2.20a) and (2.20b).

Since we have

$$\begin{aligned}(au_x, v_x) - \langle au_x, v_x\rangle_h &= \sum_e \left(\iint_e au_xv_xdxdy - \iint_e au_xv_xd^hxd^hy\right)\\ =\sum_e \left(\iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_sdsdt - \iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_sd^hsd^ht\right) &= \sum_e \left(\iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_sdsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_sd^hsd^ht\right),\end{aligned}$$

where we use the fact $\hat{a}\hat{u}_s\hat{v}_s = (\hat{a}\hat{u}_s)_I\hat{v}_s$ on the Gauss-Lobatto quadrature points. For fixed t, $(\hat{a}\hat{u}_s)_I\hat{v}_s$ is a polynomial of degree $2k-1$ w.r.t. variable s, thus the $(k+1)$-point Gauss-Lobatto quadrature is exact for its s-integration, i.e.,

$$\iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_sd^hsd^ht = \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_sdsd^ht.$$

To estimate the quadrature error we introduce some intermediate values then do interpretation by parts,

$$\begin{aligned}
&\iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_s d^hsd^ht && (2.23)\\
=&\iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_s dsdt + \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I\hat{v}_s dsd^ht && (2.24)\\
=&\iint_{\hat{K}} \left[\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right]\hat{v}_s dsdt + \left(\iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v}dsd^ht - \iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v}dsdt\right) && (2.25)\\
&+\left(\int_{-1}^{1} (\hat{a}\hat{u}_s)_I\hat{v}dt\Big|_{s=-1}^{s=1} - \int_{-1}^{1} (\hat{a}\hat{u}_s)_I\hat{v}d^ht\Big|_{s=-1}^{s=1}\right) = I + II + III. && (2.26)
\end{aligned}$$

For the first term in (2.26), let $\overline{\hat{v}_s}$ be the cell average of $\hat{v}_s$ on $\hat{K}$, then

$$I = \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right)\overline{\hat{v}_s}dsdt + \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right)(\hat{v}_s - \overline{\hat{v}_s})dsdt.$$

By (2.14) we have

$$\left|\iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right)\overline{\hat{v}_s}dsdt\right| \leq C[\hat{a}\hat{u}_s]_{k+2,\hat{K}}\left|\overline{\hat{v}_s}\right| = \mathcal{O}(h^{k+2})\|\hat{a}\|_{k+2,\infty,e}\|\hat{u}\|_{k+3,e}\|\hat{v}\|_{1,e}.$$

By Cauchy-Schwarz inequality, the Bramble-Hilbert Lemma on interpolation error and Poincaré inequality, we have

$$\begin{aligned}
&\left|\iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right)(\hat{v}_s - \overline{\hat{v}_s})dsdt\right| \leq |\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I|_{0,\hat{K}}|\hat{v}_s - \overline{\hat{v}_s}|_{0,\hat{K}}\\
\leq& C[\hat{a}\hat{u}_s]_{k+1,\hat{K}}|\hat{v}|_{2,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,e}\|u\|_{k+2,e}\|v\|_{2,e}.
\end{aligned}$$

Thus we have

$$I = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,e}\|u\|_{k+3,e}\|v\|_{2,e}.$$

For the second term in (2.26), we can estimate it the same way as in the proof of Theorem 2.4. in [39]. For each $\hat{v} \in Q^k(\hat{K})$ we can define a linear form on $H^k(\hat{K})$ as

$$\hat{E}_{\hat{v}}(\hat{f}) = \iint_{\hat{K}} (\hat{F}_I)_s\hat{v}dsdt - \iint_{\hat{K}} (\hat{F}_I)_s\hat{v}dsd^ht,$$

where $\hat{F}$ is an antiderivative of $\hat{f}$ w.r.t. variable s. Due to the linearity of interpolation operator and differentiating operation, $\hat{E}_{\hat{v}}$ is well defined. By the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, we have

$$\hat{E}_{\hat{v}}(\hat{f}) \leq C\|\hat{F}\|_{0,\infty,\hat{K}}\|\hat{v}\|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{0,\infty,\hat{K}}\|\hat{v}\|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}} \leq C\|\hat{f}\|_{k,\hat{K}}\|\hat{v}\|_{0,\hat{K}},$$

where we use the fact that all the norms on $Q^k(\hat{K})$ are equivalent to derive the first inequality. The above inequalities imply that the mapping $\hat{E}_{\hat{v}}$ is a continuous linear form on $H^k(\hat{K})$. With projection Π_1 defined in (2.8), we have

$$\hat{E}_{\hat{v}}(\hat{f}) = \hat{E}_{\hat{v}-\Pi_1\hat{v}}(\hat{f}) + \hat{E}_{\Pi_1\hat{v}}(\hat{f}), \quad \forall \hat{v} \in Q^k(\hat{K}).$$

Notice that $\hat{F}$ by definition is an antiderivative of $\hat{f}$ w.r.t. only variable s. If $\hat{f} \in Q^{k-1}(\hat{K})$, then $\hat{F}_I$ is a polynomial of degree only $k-1$ w.r.t. to variable t thus $(\hat{F}_I)_s \in Q^{k-1}(\hat{K})$. The quadrature is exact for polynomials of degree $2k-1$, thus $Q^{k-1}(\hat{K}) \subset \ker \hat{E}_{\hat{v}-\Pi_1\hat{v}}$. So by the Bramble-Hilbert Lemma, we get

$$\hat{E}_{\hat{v}-\Pi_1\hat{v}}(\hat{f}) \leq C[f]_{k,\hat{K}}\|\hat{v}-\Pi_1\hat{v}\|_{0,\hat{K}} \leq C[f]_{k,\hat{K}}|\hat{v}|_{2,\hat{K}},$$

and we also have

$$\hat{E}_{\Pi_1\hat{v}}(\hat{f}) = \iint_{\hat{K}} (\hat{F}_I)_s \Pi_1 \hat{v} ds dt - \iint_{\hat{K}} (\hat{F}_I)_s \Pi_1 \hat{v} ds d^h t = 0.$$

Thus we have

$$\begin{aligned}&\iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v} ds d^h t - \iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v} ds dt = -\hat{E}_{\hat{v}}((\hat{a}\hat{u}_s)_s) = -\hat{E}_{\hat{v}-\Pi_1\hat{v}}((\hat{a}\hat{u}_s)_s)\\ \leq& C[(\hat{a}\hat{u}_s)_s]_{k,\hat{K}}|\hat{v}_h|_{2,\hat{K}} \leq C|\hat{a}\hat{u}_s|_{k+1,\hat{K}}|\hat{v}|_{2,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,e}\|u\|_{k+2,e}|v|_{2,e}\end{aligned}$$

Now we only need to discuss the line integral term. Let L_2 and L_4 denote the left and right boundary of Ω and let l_2^e and l_4^e denote the left and right edge of element e or $l_2^{\hat{K}}$ and $l_4^{\hat{K}}$ for $\hat{K}$. Since $(\hat{a}\hat{u}_s)_I\hat{v}$ mapped back to e will be $\frac{1}{h}(au_x)_I v$ which is continuous across l_2^e and

l_4^e, after summing over all elements e, the line integrals along the inner edges are canceled out and only the line integrals on L_2 and L_4 remain.

For a cell e adjacent to L_2, consider its reference cell $\hat{K}$, and define a linear form $\hat{E}(\hat{f}) = \int_{-1}^{1} \hat{f}(-1,t)dt - \int_{-1}^{1} \hat{f}(-1,t)d^h t$, then we have

$$\hat{E}(\hat{f}\hat{v}) \leq C|\hat{f}|_{0,\infty,l_2^{\hat{K}}} |\hat{v}|_{0,\infty,l_2^{\hat{K}}} \leq C\|\hat{f}\|_{2,l_2^{\hat{K}}} \|\hat{v}\|_{0,l_2^{\hat{K}}},$$

which means that the mapping $\hat{f} \rightarrow \hat{E}(\hat{f}\hat{v})$ is continuous with operator norm less than $C\|\hat{v}\|_{0,l_2^{\hat{K}}}$ for some C. Clearly we have

$$\hat{E}(\hat{f}\hat{v}) = \hat{E}(\hat{f}\Pi_1\hat{v}) + \hat{E}(\hat{f}(\hat{v} - \Pi_1\hat{v})).$$

By the Theorem 2.3.1 we get

$$\begin{aligned} &\hat{E}((\hat{a}\hat{u}_s)_I(\hat{v} - \Pi_1\hat{v})) \leq C[(\hat{a}\hat{u}_s)_I]_{k,l_2^{\hat{K}}} [\hat{v}]_{2,l_2^{\hat{K}}} \leq C(|\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I|_{k,l_2^{\hat{K}}} + |\hat{a}\hat{u}_s|_{k,l_2^{\hat{K}}})[\hat{v}]_{2,l_2^{\hat{K}}} \\ \leq &(|\hat{a}\hat{u}_s|_{k+1,l_2^{\hat{K}}} + |\hat{a}\hat{u}_s|_{k,l_2^{\hat{K}}})[\hat{v}]_{2,l_2^{\hat{K}}} = \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} [v]_{2,l_2^e}, \end{aligned}$$

where the first inequality comes from the accuracy of the $(k+1)$-point Gauss-Lobatto quadrature rule, i.e. $\hat{E}(\hat{f}) = 0$, $\forall \hat{f} \in P^{2k-1}(\hat{K})$. The $(k+1)$-point Gauss-Lobatto quadrature rule also gives

$$\hat{E}((\hat{a}\hat{u}_s)_I \Pi_1 \hat{v}) = 0.$$

For the third term in (2.26), we sum them up over all the elements. Then for the line integral along L_2

$$\begin{aligned} &\sum_{e \cap L_2 \neq \emptyset} \int_{-1}^{1} (\hat{a}\hat{u}_s)_I(-1,t)\hat{v}(-1,t)dt - \sum_{e \cap L_2 \neq \emptyset} \int_{-1}^{1} (\hat{a}\hat{u}_s)_I(-1,t)\hat{v}(-1,t)d^h t \\ = &\sum_{e \cap L_2 \neq \emptyset} \hat{E}((\hat{a}\hat{u}_s)_I \hat{v}) = \sum_{e \cap L_2 \neq \emptyset} \mathcal{O}(h^{k+2}) \|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} |v|_{2,l_2^e}. \end{aligned}$$

Let s_α and ω_α $(\alpha = 1, 2, \cdots, k+2)$ denote the quadrature points and weights in $(k+2)$-point Gauss-Lobatto quadrature rule for $s \in [-1, 1]$. Since $\hat{v}_{tt}^2(s,t) \in Q^{2k}(\hat{K})$, $(k+2)$-point Gauss-Lobatto quadrature is exact for s-integration thus

$$\int_{-1}^{1}\int_{-1}^{1} \hat{v}_{tt}^2(s,t)dsdt = \sum_{\alpha=1}^{k+2} \omega_\alpha \int_{-1}^{1} \hat{v}_{tt}^2(s_\alpha, t)dt,$$

which implies

$$\int_{-1}^{1} \hat{v}_{tt}^2(\pm 1, t)dt \leq C \int_{-1}^{1}\int_{-1}^{1} \hat{v}_{tt}^2(s,t)dsdt, \tag{2.27}$$

thus

$$h^{\frac{1}{2}}|v|_{2,l_2^e} \leq C[v]_{2,e}.$$

By Cauchy-Schwarz inequality and trace inequality, we have

$$\begin{aligned}
&\sum_{e\cap L_2 \neq \emptyset} \left(\int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} dt \Big|_{s=-1}^{s=1} - \int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} d^h t \Big|_{s=-1}^{s=1} \right) \\
= &\sum_{e\cap L_2 \neq \emptyset} \mathcal{O}(h^{k+2}) \|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} |v|_{2,l_2^e} \\
= &\sum_{e\cap L_2 \neq \emptyset} \mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} |v|_{2,e} = \mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k+1,\infty,\Omega} \|u\|_{k+2,L_2} |v|_{2,\Omega} \\
= &\mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k+1,\infty,\Omega} \|u\|_{k+3,\Omega} |v|_{2,\Omega}.
\end{aligned}$$

Combine all the estimates above, we get (2.19b). Since the $\frac{1}{2}$ order loss is only due to the line integral along the boundary $\partial\Omega$. If $v \in V_0^h$, $v_{yy} = 0$ on L_2 and L_4 so we have (2.19a). □

2.4 The M-type Projection

To establish the superconvergence of C^0-Q^k finite element method for multi-dimensional variable coefficient equations, it is necessary to use a special polynomial projection of the exact solution, which has two equivalent definitions. One is the M-type projection used in [9], [40]. The other one is the point-line-plane interpolation used in [10], [41].

For the sake of completeness, we review the relevant results regarding M-type projection, which is a more convenient tool. Most results in this section were considered and established for more general rectangular elements in [9]. For simplicity, we use some simplified proof and arguments for Q^k element in this section. We only discuss the two dimensional case and the extension to three dimensions is straightforward.

2.4.1 One dimensional case

The L^2-orthogonal Legendre polynomials on the reference interval $\hat{K} = [-1, 1]$ are given as

$$l_k(t) = \frac{1}{2^k k!}\frac{d^k}{dt^k}(t^2-1)^k : l_0(t) = 1, l_1(t) = t, l_2(t) = \frac{1}{2}(3t^2-1), \cdots$$

Define their antiderivatives as M-type polynomials:

$$M_{k+1}(t) = \frac{1}{2^k k!}\frac{d^{k-1}}{dt^{k-1}}(t^2-1)^k : M_0(t) = 1, M_1(t) = t, M_2(t) = \frac{1}{2}(t^2-1), M_3(t) = \frac{1}{2}(t^3-t), \cdots$$

which satisfy the following properties:

- $M_k(\pm 1) = 0, \forall k \geq 2.$
- If $j - i \neq 0, \pm 2$, then $M_i(t) \perp M_j(t)$, i.e., $\int_{-1}^{1} M_i(t)M_j(t)dt = 0.$
- Roots of $M_k(t)$ are the k-point Gauss-Lobatto quadrature points for $[-1, 1]$.

Since Legendre polynomials form a complete orthogonal basis for $L^2([-1, 1])$, for any $f(t) \in H^1([-1, 1])$, its derivative $f(t)$ can be expressed as Fourier-Legendre series

$$f'(t) = \sum_{j=0}^{\infty} b_{j+1} l_j(t), \quad b_{j+1} = (j + \frac{1}{2})\int_{-1}^{1} f(t)l_j(t)dt.$$

Define the M-type projection

$$f_k(t) = \sum_{j=0}^{k} b_j M_j(t),$$

where $b_0 = \frac{f(1)+f(-1)}{2}$ is determined by $b_1 = \frac{f(1)-f(-1)}{2}$ to make $f_k(\pm 1) = f(\pm 1)$. Since the Fourier-Legendre series converges in L^2, by Cauchy Schwarz inequality,

$$\lim_{k\to\infty} f_k(t) - f(t) = \lim_{k\to\infty} \int_{-1}^{t} [f_k(x) - f(x)]\, dx \le \lim_{k\to\infty} \sqrt{2}\|f_k(t) - f(t)\|_{L^2([-1,1])} = 0.$$

We get the M-type expansion of $f(t)$: $f(t) = \lim\limits_{k\to\infty} f_k(t) = \sum\limits_{j=0}^{\infty} b_j M_j(t)$. The remainder $R_k(t)$ of M-type projection is

$$R[f]_k(t) = f(t) - f_k(t) = \sum_{j=k+1}^{\infty} b_j M_j(t).$$

The following properties are straightforward to verify:

- $f_k(\pm 1) = f(\pm 1)$ thus $R_k(\pm 1) = 0$ for $k \ge 1$.
- $R[f]_k(t) \perp v(t)$ for any $v(t) \in P^{k-2}(t)$ on $[-1, 1]$, i.e., $\int_{-1}^{1} R[f]_k v dt = 0$.
- $R[f]_k(t) \perp v(t)$ for any $v(t) \in P^{k-1}(t)$ on $[-1, 1]$.
- For $j \ge 2$, $b_j = (j - \frac{1}{2})[f(t)l_{j-1}(t)|_{-1}^{1}] - \int_{-1}^{1} f(t)l(j-1)(t)dt$.
- For $j \le k$, $|b_j| \le C_k \|f\|_{0,\infty,\hat{K}}$.
- $\|R[f]_k(t)\|_{0,\infty,\hat{K}} \le C_k \|f\|_{0,\infty,\hat{K}}$.

2.4.2 Two dimensional case

Consider a function $\hat{f}(s,t) \in H^2(\hat{K})$ on the reference cell $\hat{K} = [-1, 1] \times [-1, 1]$, it has the expansion

$$\hat{f}(s,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \hat{b}_{i,j} M_i(s) M_j(t),$$

where

$$\begin{aligned}
\hat{b}_{0,0} &= \frac{1}{4}[\hat{f}(-1,-1) + \hat{f}(-1,1) + \hat{f}(1,-1) + \hat{f}(1,1)],\\
\hat{b}_{0,j}, \hat{b}_{1,j} &= \frac{2j-1}{4}\int_{-1}^{1}[\hat{f}_t(1,t) \pm \hat{f}_t(-1,t)]l_{j-1}(t)dt, \quad j \geq 1,\\
\hat{b}_{i,0}, \hat{b}_{i,1} &= \frac{2i-1}{4}\int_{-1}^{1}[\hat{f}_s(s,1) \pm \hat{f}_s(s,-1)]l_{i-1}(s)ds, \quad i \geq 1,\\
\hat{b}_{i,j} &= \frac{(2i-1)(2j-1)}{4}\iint_{\hat{K}} \hat{f}_{st}(s,t)l_{i-1}(s)l_{j-1}(t)dsdt, \quad i,j \geq 1.
\end{aligned}$$

Define the Q^k M-type projection of $\hat{f}$ on $\hat{K}$ and its remainder as

$$\hat{f}_{k,k}(s,t) = \sum_{i=0}^{k}\sum_{j=0}^{k}\hat{b}_{i,j}M_i(s)M_j(t), \quad \hat{R}[\hat{f}]_{k,k}(s,t) = \hat{f}(s,t) - \hat{f}_{k,k}(s,t).$$

For $f(x,y)$ on $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$, let $\hat{f}(s,t) = f(sh + x_e, th + y_e)$ then the Q^k M-type projection of f on e and its remainder are defined as

$$f_{k,k}(x,y) = \hat{f}_{k,k}(\frac{x - x_e}{h}, \frac{y - y_e}{h}), \quad R[f]_{k,k}(x,y) = f(x,y) - f_{k,k}(x,y).$$

Theorem 2.4.1. *The Q^k M-type projection is equivalent to the Q^k point-line-plane projection Π defined as follows:*

1. *$\Pi\hat{u} = \hat{u}$ at four corners of $\hat{K} = [-1,1] \times [-1,1]$.*

2. *$\Pi\hat{u} - \hat{u}$ is orthogonal to polynomials of degree $k-2$ on each edge of $\hat{K}$.*

3. *$\Pi\hat{u} - \hat{u}$ is orthogonal to any $v \in Q^{k-2}(\hat{K})$ on $\hat{K}$.*

Proof. We only need to show that M-type projection $\hat{f}_{k,k}(s,t)$ satisfies the same three properties. By $M_j(\pm 1) = 0$ for $j \geq 2$, we can derive that $\hat{f}_{k,k} = \hat{f}$ at $(\pm 1, \pm 1)$. For instance, $\hat{f}_{k,k}(1,1) = \hat{b}_{0,0} + \hat{b}_{1,0} + \hat{b}_{0,1} + \hat{b}_{1,1} = \hat{f}(1,1)$.

The second property is implied by $M_j(\pm 1) = 0$ for $j \geq 2$ and $M_j(t) \perp P^{k-2}(t)$ for $j \geq k+1$. For instance, at $s = 1$, $\hat{f}_{k,k}(1,t) - \hat{f}(1,t) = \sum\limits_{j=k+1}^{\infty}(\hat{b}_{0,j} + \hat{b}_{1,j})M_j(t) \perp P^{k-2}(t)$ on $[-1,1]$.

The third property is implied by $M_j(t) \perp P^{k-2}(t)$ for $j \geq k+1$. □

Lemma 2.4.2. *Assume $\hat{f} \in H^{k+1}(\hat{K})$ with $k \geq 2$, then*

1. $|\hat{b}_{i,j}| \leq C_k \|\hat{f}\|_{0,\infty,\hat{K}}, \quad \forall i, j \leq k.$
2. $|\hat{b}_{i,j}| \leq C_k |\hat{f}|_{i+j,2,\hat{K}}, \quad \forall i, j \geq 1, i+j \leq k+1.$
3. $|\hat{b}_{i,k+1}| \leq C_k |\hat{f}|_{k+1,2,\hat{K}}, \quad 0 \leq i \leq k+1.$
4. *If $\hat{f} \in H^{k+2}(\hat{K})$, then* $|\hat{b}_{i,k+1}| \leq C_k |\hat{f}|_{k+2,2,\hat{K}}, \quad 1 \leq i \leq k+1.$

Proof. First of all, similar to the one-dimensional case, through integration by parts, $\hat{b}_{i,j}$ can be represented as integrals of $\hat{f}$ thus $|\hat{b}_{i,j}| \leq C_k \|\hat{f}\|_{0,\infty,\hat{K}}$ for $i, j \leq k$.

By the fact that the antiderivatives (and higher order ones) of Legendre polynomials vanish at ± 1, after integration by parts for both variables, we have

$$|\hat{b}_{i,j}| \leq C_k \iint_{\hat{K}} |\partial_s^i \partial_t^j \hat{f}| ds dt \leq C_k |\hat{f}|_{i+j,2,\hat{K}}, \quad \forall i, j \geq 1, i+j \leq k+1.$$

For the third estimate, by integration by parts only for the variable t, we get

$$\forall i \geq 1, |\hat{b}_{i,k+1}| \leq C_k \iint_{\hat{K}} |\partial_s \partial_t^k \hat{f}| ds dt \leq C_k |\hat{f}|_{k+1,2,\hat{K}}.$$

For $\hat{b}_{0,k+1}$, from the first estimate, we have $|\hat{b}_{0,k+1}| \leq C_k \|\hat{f}\|_{0,\infty,\hat{K}} \leq C_k \|\hat{f}\|_{k+1,2,\hat{K}}$ thus $\hat{b}_{0,k+1}$ can be regarded as a continuous linear form on $H^{k+1}(\hat{K})$ and it vanishes if $\hat{f} \in Q^k(\hat{K})$. So by the Bramble-Hilbert Lemma, $|\hat{b}_{0,k+1}| \leq C_k [\hat{f}]_{k+1,2,\hat{K}}$.

Finally, by integration by parts only for the variable t, we get

$$|\hat{b}_{i,k+1}| \leq C_k \iint_{\hat{K}} |\partial_s \partial_t^{k+1} \hat{f}| ds dt \leq C_k |\hat{f}|_{k+2,2,\hat{K}}, \quad 1 \leq i \leq k+1.$$

□

Lemma 2.4.3. *For $k \geq 2$, we have*

1. $|\hat{R}[\hat{f}]_{k,k}|_{0,\infty,\hat{K}} \leq C_k [\hat{f}]_{k+1,\hat{K}}$, $|\hat{R}[\hat{f}]_{k,k}|_{0,2,\hat{K}} \leq C_k [\hat{f}]_{k+1,\hat{K}}$.
2. $|\partial_s \hat{R}[\hat{f}]_{k,k}|_{0,\infty,\hat{K}} \leq C_k [\hat{f}]_{k+1,\hat{K}}$, $|\partial_s \hat{R}[\hat{f}]_{k,k}|_{0,2,\hat{K}} \leq C_k [\hat{f}]_{k+1,\hat{K}}$.

3. $\iint_{\hat K} \partial_s \hat R[\hat f]_{k,k} dsdt = 0$

Proof. Lemma 2.4.2 implies $\|\hat f_{k,k}\|_{0,\infty,\hat K} \le C_k \|\hat f\|_{0,\infty,\hat K}$ and $\|\partial_s \hat f_{k,k}\|_{0,\infty,\hat K} \le C_k \|\hat f\|_{0,\infty,\hat K}$. Thus

$$\forall (s,t) \in \hat K, |\hat R[\hat f]_{k,k}(s,t)| \le |\hat f_{k,k}(s,t)| + |\hat f(s,t)| \le C_k \|\hat f\|_{0,\infty,\hat K} \le C_k \|\hat f\|_{k+1,\hat K}.$$

Notice that here C_k does not depend on (s,t). So $R[\hat f]_{k,k}(s,t)$ is a continuous linear form on $H^{k+1}(\hat K)$ and its operator norm is bounded by a constant independent of (s,t). Since it vanishes for any $\hat f \in Q^k(\hat K)$, by the Bramble-Hilbert Lemma, we get $|R[\hat f]_{k,k}(s,t)| \le C_k [\hat f]_{k+1,\hat K}$ where C_k does not depend on (s,t). So the L^∞ estimate holds and it implies the L^2 estimate.

The second estimate can be established similarly since we have

$$|\partial_s \hat R[\hat f]_{k,k}(s,t)| \le |\partial_s \hat f_{k,k}(s,t)| + |\partial_s \hat f(s,t)| \le C_k \|\hat f\|_{1,\infty,\hat K} \le C_k \|\hat f\|_{k+1,\hat K}.$$

The third equation is implied by the fact that $M_j(t) \perp 1$ for $j \ge 3$ and $M_j(t) \perp 1$ for $j \ge 2$. Another way to prove the third equation is to use integration by parts

$$\iint_{\hat K} \partial_s \hat R[\hat f]_{k+1,k+1} dsdt = \int_{-1}^{1} \left(\hat R[\hat f]_{k+1,k+1}(1,t) - \hat R[\hat f]_{k+1,k+1}(-1,t) \right) dt,$$

which is zero the second property in Theorem 2.4.1. □

For the discussion in the next few subsections, it is useful to consider the lower order part of the remainder of $\hat R[\hat f]_{k,k}$:

Lemma 2.4.4. *For $\hat f \in H^{k+2}(\hat K)$ with $k \ge 2$, define $\hat R[\hat f]_{k+1,k+1} - \hat R[\hat f]_{k,k} = \hat R_1 + \hat R_2$ with*

$$\begin{aligned} \hat R_1 &= \sum_{i=0}^{k} \hat b_{i,k+1} M_i(s) M_{k+1}(t), \\ \hat R_2 &= \sum_{j=0}^{k+1} \hat b_{k+1,j} M_{k+1}(s) M_j(t) = M_{k+1}(s) \hat b_{k+1}(t), \quad \hat b_{k+1}(t) = \sum_{j=0}^{k+1} \hat b_{k+1,j} M_j(t). \end{aligned} \tag{2.28}$$

They have the following properties:

1. $\iint_{\hat{K}} \partial_s \hat{R}_1 ds dt = 0.$

2. $|\partial_s \hat{R}_1|_{0,\infty,\hat{K}} \leq C_k |\hat{f}|_{k+2,2,\hat{K}}$, $|\partial_s \hat{R}_1|_{0,2,\hat{K}} \leq C_k |\hat{f}|_{k+2,2,\hat{K}}$.

3. $|\hat{b}_{k+1}(t)| \leq C_k |\hat{f}|_{k+1,\hat{K}}$, $|\hat{b}_{k+1}(t)| \leq C_k |\hat{f}|_{k+2,\hat{K}}$, $\forall t \in [-1,1]$.

Proof. The first equation is due to the fact that $M_{k+1}(t) \perp 1$ since $k \geq 2$.

Notice that $M_0(s) = 0$, by Lemma 2.4.2, we have

$$|\partial_s \hat{R}_1(s,t)| = \left| \sum_{i=1}^{k} \hat{b}_{i,k+1} M_i(s) M_{k+1}(t) \right| \leq C_k |\hat{f}|_{k+2,\hat{K}}.$$

So we get the L^∞ estimate for $|\partial_s \hat{R}_1(s,t)|$ thus the L^2 estimate.

Similar to the estimates in Lemma 2.4.2, we can show $|\hat{b}_{k+1,j}| \leq C_k |\hat{f}|_{k+1,\hat{K}}$ for $j \leq k+1$, thus $|b_{k+1}(t)| \leq C_k |\hat{f}|_{k+1,\hat{K}}$. Since $b_{k+1}(t) = \sum\limits_{j=1}^{k+1} \hat{b}_{k+1,j} M_j(t)$, by the last estimate in Lemma 2.4.2, we get $|\hat{b}_{k+1}(t)| \leq C_k |\hat{f}|_{k+2,\hat{K}}$. □

2.4.3 The C^0-Q^k projection

Now consider a function $u(x,y) \in H^{k+2}(\Omega)$, let $u_p(x,y)$ denote its piecewise Q^k M-type projection on each element e in the mesh Ω_h. The first two properties in Theorem 2.4.1 imply that $u_p(x,y)$ on each edge is uniquely determined by $u(x,y)$ along that edge. Thus $u_p(x,y)$ is continuous on Ω_h. The approximation error $u - u_p$ is one order higher at all Gauss-Lobatto points Z_0:

Theorem 2.4.5.

$$\|u - u_p\|_{l^2(\Omega)} = \mathcal{O}(h^{k+2}) \|u\|_{k+2}, \quad \forall u \in H^{k+2}(\Omega).$$

$$\|u - u_p\|_{l^\infty(\Omega)} = \mathcal{O}(h^{k+2}) \|u\|_{k+2,\infty}, \quad \forall u \in W^{k+2,\infty}(\Omega).$$

Proof. Consider any e with cell center (x_e, y_e), define $\hat{u}(s,t) = u(x_e + sh, y_e + th)$. Since the $(k+1)$ Gauss-Lobatto points are roots of $M_{k+1}(t)$, $\hat{R}_{k+1,k+1}[\hat{u}] - \hat{R}_{k,k}[\hat{u}]$ vanishes at $(k+1) \times (k+1)$ Gauss-Lobatto points on $\hat{K}$. By Lemma 2.4.3, we have $|\hat{R}_{k+1,k+1}[\hat{u}](s,t)| \leq C[\hat{u}]_{k+2,\hat{K}}$.

Mapping back to the cell e, with (2.5), at the $(k+1)\times(k+1)$ Gauss-Lobatto points on e, $|u-u_p|\leq Ch^{k+2-\frac{n}{2}}[u]_{k+2,e}$. Summing over all elements e, we get

$$\|u-u_p\|_{l^2(\Omega)}\leq C\left[h^n\sum_e h^{2k+4-n}[u]^2_{k+2,e}\right]^{\frac{1}{2}}=\mathcal{O}(h^{k+2})[u]_{k+2,\Omega}.$$

If further assuming $u\in W^{k+2,\infty}(\Omega)$, then at the $(k+1)\times(k+1)$ Gauss-Lobatto points on e, $|u-u_p|\leq Ch^{k+2-\frac{n}{2}}[u]_{k+2,e}\leq Ch^{k+2}[u]_{k+2,\infty,\Omega}$, which implies the second estimate. □

2.5 Superconvergence Of The Bilinear Form

The M-type projection in [9], [40] is a very convenient tool for discussing the superconvergence of function values. Let u_p be the M-type Q^k projection of the smooth exact solution u and its definition will be given in the following subsection. To establish the superconvergence of the original finite element method (2.1) for a generic elliptic problem (2.9) with smooth coefficients, one can show the following superconvergence of bilinear forms, see [9], [10] (see also 4 for a detailed proof):

$$A(u-u_p,v_h)=\begin{cases}\mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h,\\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h.\end{cases}$$

In this section we will show the superconvergence of the bilinear form A_h:

$$A_h(u-u_p,v_h)=\begin{cases}\mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h, \quad (2.29a)\\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h. \quad (2.29b)\end{cases}$$

Lemma 2.5.1. *Assume $\hat{f}(s,t)\in H^{k+3}(\hat{K})$, $k\geq 2$,*

$$\langle\hat{R}[\hat{f}]_{k+1,k+1}-\hat{R}[\hat{f}]_{k,k},1\rangle_{\hat{K}}=0,\quad |\langle\partial_s\hat{R}[\hat{f}]_{k+1,k+1},1\rangle_{\hat{K}}|\leq C|\hat{f}|_{k+3,\hat{K}}.$$

Proof. First, we have

$$\langle \hat{R}[\hat{f}]_{k+1,k+1} - \hat{R}[\hat{f}]_{k,k}, 1\rangle_{\hat{K}} = \langle M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s) + M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t), 1\rangle_{\hat{K}} = 0$$

due to the fact that roots of $M_{k+1}(t)$ are the $(k+1)$-point Gauss-Lobatto quadrature points for $[-1,1]$.

We have

$$\begin{aligned}
&\langle \partial_s\hat{R}[\hat{f}]_{k+1,k+1}, 1\rangle_{\hat{K}}\\
=&\langle \partial_s\hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle \partial_s(\hat{R}[\hat{f}]_{k+2,k+2} - \hat{R}[\hat{f}]_{k+1,k+1}), 1\rangle_{\hat{K}}\\
=&\langle \partial_s\hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle M_{k+2}(t)\sum_{i=0}^{k+1}\hat{b}_{i,k+2}M_i(s) + M_{k+2}(s)\sum_{j=0}^{k+2}\hat{b}_{k+2,j}M_j(t), 1\rangle_{\hat{K}}\\
=&\langle \partial_s\hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle M_{k+2}(t)\sum_{i=0}^{k}\hat{b}_{i+1,k+2}l_i(s), 1\rangle_{\hat{K}} + \langle l_{k+1}(s)\sum_{j=0}^{k+2}\hat{b}_{k+2,j}M_j(t), 1\rangle_{\hat{K}}.
\end{aligned}$$

Then by Lemma 2.4.3,

$$|\langle \partial_s\hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}}| \leq C|\hat{f}|_{k+3,\hat{K}}.$$

Notice that we have $\langle l_{k+1}(s)\sum_{j=0}^{k+2}\hat{b}_{k+2,j}M_j(t), 1\rangle_{\hat{K}} = 0$ since the $(k+1)$-point Gauss-Lobatto quadrature for s-integration is exact and $l_{k+1}(s)$ is orthogonal to 1. Lemma 2.4.2 implies $|\hat{b}_{i+1,k+2}| \leq C[\hat{f}]_{k+3,\hat{K}}$ for $i \geq 0$, thus we have

$$|\langle M_{k+2}(t)\sum_{i=0}^{k}\hat{b}_{i+1,k+2}l_i(s), 1\rangle_{\hat{K}}| \leq C[\hat{f}]_{k+3,\hat{K}}.$$

□

Lemma 2.5.2. *Assume $a(x,y) \in W^{k,\infty}(\Omega)$, $u(x,y) \in H^{k+3}(\Omega)$ and $k \geq 2$. Then*

$$\langle a(u-u_p)_x, (v_h)_x\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. As before, we ignore the subscript of v_h for simplicity. We have

$$\langle a(u-u_p)_x, v_x\rangle_h = \sum_e \langle a(u-u_p)_x, v_x\rangle_{e,h},$$

and on each cell e,

$$\begin{aligned}\langle a(u-u_p)_x, v_x\rangle_{e,h} &= \langle (R[u]_{k,k})_x, av_x\rangle_{e,h} = \langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} \\ =&\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}}.\end{aligned} \tag{2.30}$$

For the first term in (2.30), we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} = \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}(\hat{v}_s - \overline{\hat{v}_s})\rangle_{\hat{K}}.$$

By Lemma 2.5.1,

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}}\,\overline{\hat{v}_s}\rangle_{\hat{K}} \le C|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}}.$$

By Lemma 2.4.3,

$$|(\hat{R}[\hat{u}]_{k+1,k+1})_s|_{0,\infty,\hat{K}} \le C[\hat{u}]_{k+2,\hat{K}}.$$

By Bramble-Hilbert Lemma Theorem 2.3.1 we have

$$\begin{aligned}\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}} &= \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}}\,\overline{\hat{v}_s}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, (\hat{a}-\overline{\hat{a}})\overline{\hat{v}_s}\rangle_{\hat{K}} \\ \le& C(|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}} + |\hat{a}-\overline{\hat{a}}|_{0,\infty}|\hat{u}|_{k+2,\hat{K}}|\hat{v}|_{1,\hat{K}}) \\ \le& C(|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}} + |\hat{a}|_{1,\infty}|\hat{u}|_{k+2,\hat{K}}|\hat{v}|_{1,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,e}\|u\|_{k+3,e}\|v\|_{1,e},\end{aligned}$$

and

$$\begin{aligned}\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}(\hat{v}_s - \overline{\hat{v}_s})\rangle_{\hat{K}} &\le C[\hat{u}]_{k+2,2,\hat{K}}|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}_s - \overline{\hat{v}_s}|_{0,\infty,\hat{K}} \\ \le& C[\hat{u}]_{k+2,2,\hat{K}}|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}_s - \overline{\hat{v}_s}|_{0,2,\hat{K}} = \mathcal{O}(h^{k+2})[u]_{k+2,2,e}|a|_{0,\infty,e}|v|_{2,2,e}.\end{aligned}$$

Thus,

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,e}|u|_{k+3,2,e}\|v\|_{2,e}. \tag{2.31}$$

For the second term in (2.30), we have

$$\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle(M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s)+M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t))_s,\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s)+l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s),\hat{a}\hat{v}_s\rangle_{\hat{K}}-\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}.
\end{aligned}\tag{2.32}$$

Since $M_{k+1}(t)$ vanishes at $(k+1)$ Gauss-Lobatto points, we have

$$\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,3}l_i(s),\hat{a}\hat{v}_s\rangle_{\hat{K}}=0.$$

For the second term in (2.32),

$$\begin{aligned}
&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}(\hat{v}_s-\overline{\hat{v}_s})\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\hat{\Pi}_1\hat{a})\overline{\hat{v}_s}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{\Pi}_1\hat{a})\overline{\hat{v}_s}\rangle_{\hat{K}}\\
&+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\overline{\hat{a}})(\hat{v}_s-\overline{\hat{v}_s})\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\overline{\hat{a}}(\hat{v}_s-\overline{\hat{v}}_s)\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\hat{\Pi}_1\hat{a})\overline{\hat{v}}_s\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\overline{\hat{a}})(\hat{v}_s-\overline{\hat{v}}_s)\rangle_{\hat{K}},
\end{aligned}$$

where the last step is due to the facts that $(\hat{\Pi}_1\hat{a})\overline{\hat{v}_s}$ and $\overline{\hat{a}}(\hat{v}_s-\overline{\hat{v}}_s)$ are polynomials of degree at most $k-1$ with respect to variable s, the $(k+1)$-point Gauss-Lobatto quadrature on s-integration is exact for polynomial of degree $2k-1$, and $l_k(s)$ is orthogonal to polynomials of lower degree. With Lemma 2.4.2, we have

$$\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}\le C|\hat{u}|_{k+1,2,\hat{K}}(|\hat{a}|_{2,\infty}|\hat{v}|_{1,\hat{K}}+|\hat{a}|_{1,\infty}|\hat{v}|_{2,\hat{K}})=\mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+1,e}\|v\|_{2,e}.\tag{2.33}$$

Combined with (2.31), we have proved the estimate. □

Lemma 2.5.3. *Assume* $a(x,y) \in W^{2,\infty}(\Omega)$*,* $u(x,y) \in H^{k+2}(\Omega)$ *and* $k \geq 2$*. Then*

$$\langle a(u-u_p), v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. As before, we ignore the subscript of v_h for simplicity and

$$\langle a(u-u_p), v\rangle_h = \sum_e \langle a(u-u_p), v\rangle_{e,h}.$$

On each cell e we have

$$\langle a(u-u_p), v\rangle_{e,h} = \langle R[u]_{k,k}, av\rangle_{e,h} = h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v}\rangle_{\hat{K}} = h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} + h^2\langle \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}}. \tag{2.34}$$

For the first term in (2.34), due to the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, Bramble-Hilbert Lemma Theorem 2.3.1 and Lemma 2.4.3, we have

$$\begin{aligned}
&h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \leq Ch^2|R[\hat{u}]_{k,k}|_\infty|\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}|_\infty \leq Ch^2|\hat{u}|_{k+1,\hat{K}}\|\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\|_{2,\hat{K}} \\
&\leq Ch^2|\hat{u}|_{k+1,\hat{K}}(\|\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\|_{L^2(\hat{K})} + |\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) \\
&\leq Ch^2|\hat{u}|_{k+1,\hat{K}}(|\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{2,\infty,e}\|u\|_{k+1,e}\|v\|_{2,e}.
\end{aligned}$$

For the second term in (2.34), we have

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} = h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} - h^2\langle \hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}}.$$

By Lemma 2.4.3 and Lemma 2.5.1 we have

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \leq Ch^2|\hat{u}|_{k+2,\hat{K}}|\hat{a}\hat{v}|_{0,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{0,\infty,e}\|u\|_{k+2,e}\|v\|_{0,e},$$

and

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} = 0.$$

Thus, we have $\langle a(u-u_p), v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2}\|v_h\|_2$. □

Lemma 2.5.4. *Assume $a \in W^{2,\infty}(\Omega)$, $u \in H^{k+3}(\Omega)$ and $k \geq 2$. Then*

$$\langle a(u-u_p)_x, v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. As before, we ignore the subscript in v_h and we have

$$\langle a(u-u_p)_x, v\rangle_h = \sum_e \langle a(u-u_p)_x, v\rangle_{e,h}.$$

On each cell e, we have

$$\begin{aligned} \langle a(u-u_p)_x, v\rangle_{e,h} = \langle (R[u]_{k,k})_x, av\rangle_{e,h} = h\langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}\rangle_{\hat{K}} \\ = h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} - h\langle (\hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}\rangle_{\hat{K}}. \end{aligned} \tag{2.35}$$

For the first term in (2.35), we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} \leq \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\rangle_{\hat{K}}$$

Due to Lemma 2.5.1,

$$h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \leq Ch\|a\|_{0,\infty}|u|_{k+3,\hat{K}}\|v\|_{0,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{0,\infty}\|u\|_{k+3,e}\|v\|_{0,e},$$

and by the same arguments as in the proof of Lemma 2.5.3 we have

$$\begin{aligned} h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \leq Ch|(R[\hat{u}]_{k+1,k+1})_s|_\infty |\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}|_\infty \leq Ch|\hat{u}|_{k+2,\hat{K}}\|\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\|_{2,\hat{K}} \\ \leq Ch|\hat{u}|_{k+2,\hat{K}}(\|\hat{a}\hat{v} - \overline{\hat{a}\hat{v}}\|_{L^2(\hat{K})} + |\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) \leq Ch|\hat{u}|_{k+2,\hat{K}}(|\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2,e} \end{aligned}$$

Thus

$$h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3,e}\|v\|_{2,e}. \tag{2.36}$$

For the second term in (2.35), we have

$$\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k+1,k+1}-\hat{R}[\hat{u}]_{k,k})_s,\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle(M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s)+M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t))_s,\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s)+l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s),\hat{a}\hat{v}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}},
\end{aligned}$$

where the last step is due to that $M_{k+1}(t)$ vanishes at $(k+1)$ Gauss-Lobatto points. Then

$$\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}},
\end{aligned}$$

where the last step is due to the facts that $\hat{\Pi}_1(\hat{a}\hat{v})$ is a linear function in s thus the $(k+1)$-point Gauss-Lobatto quadrature on s-variable is exact, and $l_k(s)$ is orthogonal to linear functions.

By Lemma 2.4.4, Lemma 2.4.2, and Theorem 2.3.1, we have

$$\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}\\
\leq&C|u|_{k+1,\hat{K}}|\hat{a}\hat{v}|_{2,\hat{K}}\leq C|u|_{k+1,\hat{K}}(|\hat{a}|_{2,\infty,\hat{K}}|\hat{v}|_{0,\hat{K}}+|\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{1,\hat{K}}+|\hat{a}|_{0,\infty}|\hat{v}|_{2,\hat{K}})
\end{aligned}$$

Thus

$$h\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+1,e}\|v\|_{2,e}. \tag{2.37}$$

By (2.36) and (2.37) and sum up over all the cells, we get the desired estimate. □

Lemma 2.5.5. *Assume $a(x,y) \in W^{4,\infty}(\Omega)$, $u(x,y) \in H^{k+3}(\Omega)$ and $k \geq 2$. Then*

$$\langle a(u-u_p)_x, (v_h)_y \rangle_h = \begin{cases} \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h, \quad (2.38a) \\ \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h. \quad (2.38b) \end{cases}$$

Proof. We ignore the subscript in v_h and we have

$$\langle a(u-u_p)_x, v_y \rangle_h = \sum_e \langle a(u-u_p)_x, v_y \rangle_{e,h},$$

and on each cell e

$$\begin{aligned} &\langle a(u-u_p)_x, v_y \rangle_{e,h} = \langle (R[u]_{k,k})_x, a v_y \rangle_{e,h} = \langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}} \\ =&\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}}. \end{aligned} \quad (2.39)$$

By the same arguments as in the proof of Lemma 2.5.2, we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{1,\infty}|u|_{k+3,2,e}\|v\|_{2,e}, \quad (2.40)$$

and

$$\langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}} = -\langle l_k(s) \sum_{j=0}^{k+1} \hat{b}_{k+1,j} M_j(t), \hat{a}\hat{v}_t \rangle_{\hat{K}}.$$

For simplicity, we define

$$\hat{b}_{k+1}(t) := \sum_{j=0}^{k+1} \hat{b}_{k+1,j} M_j(t).$$

then by the third and fourth estimates in Lemma 2.4.4, we have

$$\begin{aligned} &|\hat{b}_{k+1}(t)| \leq C \sum_{j=0}^{k+1} |\hat{b}_{k+1,j}| \leq C|\hat{u}|_{k+1,\hat{K}}, \\ &|\hat{b}_{k+1}^{(m)}(t)| \leq C \sum_{j=m}^{k+1} |\hat{b}_{k+1,j}| \leq C|\hat{u}|_{k+2,\hat{K}}, \quad 1 \leq m, \end{aligned}$$

where $\hat{b}_{k+1}^{(m)}(t)$ is the m-th derivative of $\hat{b}_{k+1}(t)$. We use the same technique in the proof of Theorem 2.3.9 and we let $l_k = l_k(s)$, $b_{k+1} = b_{k+1}(t)$ in the following,

$$\begin{aligned}
&\langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t \rangle_{\hat{K}} = -\langle l_k(s)\hat{b}_{k+1}(t), \hat{a}\hat{v}_t \rangle_{\hat{K}} \\
= &-\iint_{\hat{K}} l_k(s)\hat{b}_{k+1}(t)\hat{a}\hat{v}_t d^h s d^h t = -\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v}_t d^h s d^h t \\
= &-\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v}_t d^h s d^h t + \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt - \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt,
\end{aligned}$$

and

$$\begin{aligned}
&-\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v}_t d^h s d^h t + \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt \\
= &\iint_{\hat{K}} \left[l_k\hat{b}_{k+1}\hat{a} - (l_k\hat{b}_{k+1}\hat{a})_I \right] \hat{v}_t dsdt + \iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v}_t dsdt - \iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v}_t d^h sdt \\
= &\iint_{\hat{K}} \left[l_k\hat{b}_{k+1}\hat{a} - (l_k\hat{b}_{k+1}\hat{a})_I \right] \hat{v}_t dsdt + \iint_{\hat{K}} \partial_t (l_k\hat{b}_{k+1}\hat{a})_I \hat{v} d^h sdt - \iint_{\hat{K}} \partial_t (l_k\hat{b}_{k+1}\hat{a})_I \hat{v} dsdt \\
&+ \left(\int_{-1}^{1} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v} ds \Big|_{t=-1}^{t=1} - \int_{-1}^{1} (l_k\hat{b}_{k+1}\hat{a})_I \hat{v} d^h s \Big|_{t=-1}^{t=1} \right) = I + II + III.
\end{aligned}$$

After integration by parts with respect to the variable s, we have

$$\iint_{\hat{K}} l_k(s)\hat{b}_{k+1}(t)\hat{a}\hat{v}_t dsdt = -\iint_{\hat{K}} M_{k+1}(s)\hat{b}_{k+1}(t)(\hat{a}_s\hat{v}_t + \hat{a}\hat{v}_{st}) dsdt. \tag{2.41}$$

Let $N(s)$ be the antiderivative of $M_{k+1}(s)$. After integration by parts, we have

$$-\iint_{\hat{K}} \hat{M}_{k+1}(s) b_{k+1}(t)(\hat{a}_s\hat{v}_t + \hat{a}\hat{v}_{st}) = \iint_{\hat{K}} \hat{b}_{k+1}(t)N(s)(\hat{a}_{ss}\hat{v}_t + 2\hat{a}_s\hat{v}_{st}) + \iint_{\hat{K}} \hat{b}_{k+1}(t)N(s)\hat{a}\hat{v}_{sst}.$$

After integration by parts on the t-variable,

$$-\iint_{\hat{K}} \hat{b}_{k+1}(t)N(s)\hat{a}\hat{v}_{sst} = \iint_{\hat{K}} \partial_t[\hat{b}_{k+1}(t)N(s)\hat{a}]\hat{v}_{ss} - \int_{-1}^{1} \hat{b}_{k+1}(t)N(s)\hat{a}\hat{v}_{ss} ds \Big|_{t=-1}^{t=1},$$

$$\iint_{\hat{K}} \partial_t[\hat{b}_{k+1}(t)N(s)\hat{a}]\hat{v}_{ss} = \iint_{\hat{K}} [\hat{b}_{k+1}(t)N(s)\hat{a} + \hat{b}_{k+1}(t)N(s)\hat{a}_t]\hat{v}_{ss}.$$

By Lemma 2.4.4, we have the estimate for the two double integral terms

$$\left|\iint_{\hat{K}} \hat{b}_{k+1}(t)N(s)(\hat{a}_{ss}\hat{v}_t + 2\hat{a}_s\hat{v}_{st})\right| \leq C|\hat{u}|_{k+1,2,\hat{K}}(|\hat{a}|_{2,\infty,\hat{K}}|\hat{v}|_{1,2,\hat{K}} + |\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{2,2,\hat{K}}),$$

$$\begin{aligned}&\left|\iint_{\hat{K}}[\hat{b}_{k+1}(t)N(s)\hat{a} + \hat{b}_{k+1}(t)N(s)\hat{a}_t]\hat{v}_{ss}\right| \\ \leq &C(|\hat{u}|_{k+2,2,\hat{K}}|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}|_{2,2,\hat{K}} + |\hat{u}|_{k+1,2,\hat{K}}|\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{2,2,\hat{K}}),\end{aligned}$$

which gives the estimate $Ch^{k+2}\|a\|_{2,\infty,\Omega}\|u\|_{k+2,e}\|v\|_{k+2,e}$ after mapping back to e.

After mapping back to e, we have

$$\int_{-1}^{1} \hat{b}_{k+1}(t)M_{k+1}(s)\hat{a}\hat{v}_{ss}ds\bigg|_{t=-1}^{t=1} = h\int_{x_e-h}^{x_e+h} b_{k+1}(y)M_{k+1}(\frac{x-x_e}{h})av_{xx}dx\bigg|_{y=y_e-h}^{y=y_e+h}.$$

Notice that we have

$$\begin{aligned}b_{k+1}(y_e+h) = \hat{b}_{k+1}(1) = \sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(1) = \hat{b}_{k+1,0} + \hat{b}_{k+1,1} \\ = (k+\frac{1}{2})\int_{-1}^{1}\partial_s\hat{u}(s,1)l_k(s)ds = (k+\frac{1}{2})\int_{x_e-h}^{x_e+h}\partial_x u(x,y_e+h)l_k(\frac{x-x_e}{h})dx,\end{aligned}$$

and similarly we get $b_{k+1}(y_e-h) = \hat{b}_{k+1}(-1) = (k+\frac{1}{2})\int_{x_e-h}^{x_e+h}\partial_x u(x,y_e-h)l_k(\frac{x-x_e}{h})dx$. Thus the term $b_{k+1}(y)M_{k+1}(\frac{x-x_e}{h})av_{xx}$ is continuous across the top/bottom edge of cells. Therefore, if summing over all elements e, the line integral on the inner edges are cancelled out. Let L_1 and L_3 denote the top and bottom boundary of Ω. Then the line integral after summing over e consists of two line integrals along L_1 and L_3. We only need to discuss one of them.

Let l_1 and l_3 denote the top and bottom edge of e. First, after integration by parts k times, we get

$$\hat{b}_{k+1}(1) = (k+\frac{1}{2})\int_{-1}^{1}\partial_s\hat{u}(s,1)l_k(s)ds = (-1)^k(k+\frac{1}{2})\int_{-1}^{1}\frac{\partial^{k+1}}{\partial s^{k+1}}\hat{u}(s,1)\frac{1}{2^k k!}(s^2-1)^k ds,$$

thus by Cauchy Schwarz inequality we get

$$|\hat{b}_{k+1}(1)| \leq C_k \sqrt{\int_{-1}^{1} \left[\frac{\partial^{k+1}}{\partial s^{k+1}} \hat{u}(s,1)\right]^2 ds} \leq C_k h^{k+\frac{1}{2}} |u|_{k+1,2,l_1}.$$

Second, since v_{xx}^2 is a polynomial of degree $2k$ w.r.t. y variable, by using $(k+2)$-point Gauss Lobatto quadrature for integration w.r.t. y-variable in $\iint_e v_{xx}^2 dxdy$, we get

$$\int_{x_e-h}^{x_e+h} v_{xx}^2(x, y_e+h)dx \leq Ch^{-1} \iint_e v_{xx}^2(x,y)dxdy.$$

So by Cauchy Schwarz inequality, we have

$$\int_{x_e-h}^{x_e+h} |v_{xx}(x, y_e+h)|dx \leq \sqrt{2h} \sqrt{\int_{x_e-h}^{x_e+h} v_{xx}^2(x, y_e+h)dx} \leq C|v|_{2,2,e}.$$

Thus the line integral along L_1 can be estimated by considering each e adjacent to L_1 in the reference cell:

$$\begin{aligned}
&\sum_{e \cap L_1 \neq \emptyset} \left| \int_{-1}^{1} \hat{b}_{k+1}(1) M_{k+1}(s) \hat{a}(s,1) \hat{v}_{ss}(s,1) ds \right| \\
\leq &\sum_{e \cap L_1 \neq \emptyset} C |\hat{a}|_{0,\infty,\hat{K}} |\hat{b}_{k+1}(1)| \int_{-1}^{1} |\hat{v}_{ss}(s,1)| ds \\
= &\mathcal{O}(h^{k+\frac{3}{2}}) \sum_{e \cap L_1 \neq \emptyset} |u|_{k+1,2,l_1} \int_{x_e-h}^{x_e+h} |v_{xx}(x, y_e+h)| dx \\
= &\mathcal{O}(h^{k+\frac{3}{2}}) \sum_{e \cap L_1 \neq \emptyset} |u|_{k+1,2,l_1} |v|_{2,2,e} \\
= &\mathcal{O}(h^{k+\frac{3}{2}}) \|u\|_{k+1,L_1} \|v\|_{2,\Omega} = \mathcal{O}(h^{k+\frac{3}{2}}) \|u\|_{k+2,\Omega} \|v\|_{2,\Omega},
\end{aligned}$$

where the trace inequality $\|u\|_{k+1,\partial\Omega} \leq C\|u\|_{k+2,\Omega}$ is used.

Combine all the estimates above and sum over all elements, we have the estimate for the term $\iint_{\hat{K}} l_k(s) \hat{b}_{k+1}(t) \hat{a} \hat{v}_t dsdt$:

$$\sum_e \iint_{\hat{K}} l_k(s) \hat{b}_{k+1}(t) \hat{a} \hat{v}_t dsdt = \begin{cases} \mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k+2,\infty} \|u\|_{k+3} \|v\|_2, & \forall v \in V^h, \quad (2.42a) \\ \mathcal{O}(h^{k+2}) \|a\|_{k+2,\infty} \|u\|_{k+3} \|v\|_2, & \forall v \in V_0^h. \quad (2.42b) \end{cases}$$

We get $(k+\frac{3}{2})$-th order for the above estimation. Since the $\frac{1}{2}$ order loss is only due to the line integral along L_1 and L_3, on which $v_{xx}=0$ if $v\in V_0^h$, we get $(k+2)$-th order.

Then we can do similar estimation as in Theorem 2.3.9 for I, II, III separately.

For term I, by Theorem 2.3.1 and the estimate (2.14), we have

$$\begin{aligned}
&\iint_{\hat{K}}\left[l_k\hat{b}_{k+1}\hat{a}-(l_k\hat{b}_{k+1}\hat{a})_I\right]\hat{v}_t dsdt\\
=&\iint_{\hat{K}}\left[l_k\hat{b}_{k+1}\hat{a}-(l_k\hat{b}_{k+1}\hat{a})_I\right]\overline{\hat{v}_t}dsdt+\iint_{\hat{K}}\left[l_k\hat{b}_{k+1}\hat{a}-(l_k\hat{b}_{k+1}\hat{a})_I\right](\hat{v}_t-\overline{\hat{v}_t})dsdt\\
\leq& C\left[l_k\hat{b}_{k+1}\hat{a}\right]_{k+2,\hat{K}}|\hat{v}|_{1,\hat{K}}+C\left[l_k\hat{b}_{k+1}\hat{a}\right]_{k+1,\hat{K}}|\hat{v}|_{2,\hat{K}}\\
\leq& C\left(\sum_{m=2}^{k+2}|\hat{a}|_{m,\infty,\hat{K}}\max_{t\in[-1,1]}|\hat{b}_{k+1}(t)|\right)|\hat{v}|_{1,\hat{K}}+C\left(\sum_{m=0}^{k+2}|\hat{a}|_{m,\infty,\hat{K}}\max_{t\in[-1,1]}|\hat{b}_{k+1}^{(k+2-m)}(t)|\right)|\hat{v}|_{1,\hat{K}}\\
+&C\left(\sum_{m=1}^{k+1}|\hat{a}|_{m,\infty,\hat{K}}\max_{t\in[-1,1]}|\hat{b}_{k+1}(t)|\right)|\hat{v}|_{2,\hat{K}}+C\left(\sum_{m=0}^{k+1}|\hat{a}|_{m,\infty,\hat{K}}\max_{t\in[-1,1]}|\hat{b}_{k+1}^{(k+1-m)}(t)|\right)|\hat{v}|_{2,\hat{K}}\\
=&\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2,e}\|v\|_{2,e}.
\end{aligned}$$

For term II, as in the proof of Theorem 2.3.9, we define the linear form as

$$\hat{E}_{\hat{v}}(\hat{f})=\iint_{\hat{K}}(\hat{F}_I)_t\hat{v}dsdt-\iint_{\hat{K}}(\hat{F}_I)_t\hat{v}d^hsdt,$$

for each $\hat{v}\in Q^k(\hat{K})$ and $\hat{F}$ is an antiderivative of $\hat{f}$ w.r.t. variable t. We can easily see that $\hat{E}_{\hat{v}}$ is well defined and $\hat{E}_{\hat{v}}$ is a continuous linear form on $H^k(\hat{K})$. With projection $\hat{\Pi}_1$ defined in (2.8), we have

$$\hat{E}_{\hat{v}}(\hat{f})=\hat{E}_{\hat{v}-\hat{\Pi}_1\hat{v}}(\hat{f})+\hat{E}_{\hat{\Pi}_1\hat{v}}(\hat{f}),\quad\forall\hat{v}\in Q^k(\hat{K}).$$

Since $Q^{k-1}(\hat{K})\subset\ker\hat{E}_{\hat{v}-\hat{\Pi}_1\hat{v}}$ thus

$$\hat{E}_{\hat{v}-\hat{\Pi}_1\hat{v}}(\hat{f})\leq C[f]_{k,\hat{K}}\|\hat{v}-\hat{\Pi}_1\hat{v}\|_{0,\hat{K}}\leq C[f]_{k,\hat{K}}|\hat{v}|_{2,\hat{K}}$$

and

$$\hat{E}_{\hat{\Pi}_1\hat{v}}(\hat{f})=\iint_{\hat{K}}(\hat{F}_I)_t\hat{\Pi}_1\hat{v}dsdt-\iint_{\hat{K}}(\hat{F}_I)_t\hat{\Pi}_1\hat{v}d^hsdt=0.$$

Thus we have

$$\iint_{\hat{K}} \partial_t(l_k\hat{b}_{k+1}\hat{a})_I \hat{v} d^h s dt - \iint_{\hat{K}} \partial_t(l_k\hat{b}_{k+1}\hat{a})_I \hat{v} ds dt = -\hat{E}_{\hat{v}}((l_k\hat{b}_{k+1}\hat{a})_t)$$
$$= - \hat{E}_{\hat{v}-\Pi_1\hat{v}}((l_k\hat{b}_{k+1}\hat{a})_t) \leq C[(l_k\hat{b}_{k+1}\hat{a})_t]_{k,\hat{K}} |\hat{v}_h|_{2,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,e}\|u\|_{k+2,e}|v|_{2,e}.$$

Now we only need to discuss term III. Let L_1 and L_3 denote the top and bottom boundaries of Ω and let l_1^e, l_3^e denote the top and bottom edges of element e (and $l_1^{\hat{K}}$ and $l_3^{\hat{K}}$ for $\hat{K}$). Notice that after mapping back to the cell e we have

$$b_{k+1}(y_e + h) = \hat{b}_{k+1}(1) = \sum_{j=0}^{k+1} \hat{b}_{k+1,j} M_j(1) = \hat{b}_{k+1,0} + \hat{b}_{k+1,1}$$
$$= (k + \frac{1}{2}) \int_{-1}^{1} \partial_s \hat{u}(s,1) l_k(s) ds = (k + \frac{1}{2}) \int_{x_e-h}^{x_e+h} \partial_x u(x, y_e + h) l_k(\frac{x - x_e}{h}) dx,$$

and similarly we get $b_{k+1}(y_e - h) = \hat{b}_{k+1}(-1) = (k + \frac{1}{2}) \int_{x_e-h}^{x_e+h} \partial_x u(x, y_e - h) l_k(\frac{x-x_e}{h}) dx$. Thus the term $l(\frac{x-x_e}{h}) b_{k+1}(y) a v$ is continuous across the top and bottom edges of cells. Therefore, if summing over all elements e, the line integral on the inner edges are cancelled out. So after summing over all elements, the line integral reduces to two line integrals along L_1 and L_3. We only need to discuss one of them. For a cell e adjacent to L_1, consider its reference cell $\hat{K}$ and define linear form $\hat{E}(\hat{f}) = \int_{-1}^{1} \hat{f}(s,1) ds - \int_{-1}^{1} \hat{f}(s,1) d^h s$, then we have

$$\hat{E}(\hat{f}\hat{v}) \leq C|\hat{f}|_{0,\infty,l_1^{\hat{K}}} |\hat{v}|_{0,\infty,l_1^{\hat{K}}} \leq C\|\hat{f}\|_{2,l_1^{\hat{K}}} \|\hat{v}\|_{0,l_1^{\hat{K}}},$$

thus the mapping $\hat{f} \to \hat{E}(\hat{f}\hat{v})$ is continuous with operator norm less than $C\|\hat{v}\|_{0,l_1^{\hat{K}}}$ for some C. Since $\hat{E}((\hat{a}\hat{u}_s)_I\hat{\Pi}_1\hat{v}) = 0$ we have

$$\begin{aligned}
&\sum_{e\cap L_1\neq\emptyset}\int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}ds - \int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}d^hs \\
=&\sum_{e\cap L_1\neq\emptyset}\hat{E}((l_k\hat{b}_{k+1}\hat{a})_I\hat{v}) = \sum_{e\cap L_1\neq\emptyset}\hat{E}((l_k\hat{b}_{k+1}\hat{a})_I(\hat{v}-\hat{\Pi}_1\hat{v})) \leq \sum_{e\cap L_1\neq\emptyset}C[(l_k\hat{b}_{k+1}\hat{a})_I]_{k,l_1^{\hat{K}}}[\hat{v}]_{2,l_1^{\hat{K}}} \\
\leq&\sum_{e\cap L_1\neq\emptyset}C(|l_k\hat{b}_{k+1}\hat{a}-(l_k\hat{b}_{k+1}\hat{a})_I|_{k,l_1^{\hat{K}}} + |l_k\hat{b}_{k+1}\hat{a}|_{k,l_1^{\hat{K}}})[\hat{v}]_{2,l_1^{\hat{K}}} \\
\leq&\sum_{e\cap L_1\neq\emptyset}(|l_k\hat{b}_{k+1}\hat{a}|_{k+1,l_1^{\hat{K}}} + |l_k\hat{b}_{k+1}\hat{a}|_{k,l_1^{\hat{K}}})[\hat{v}]_{2,l_1^{\hat{K}}} \leq \sum_{e\cap L_1\neq\emptyset}C\|\hat{a}\|_{k,\infty,\hat{K}}|\hat{b}_{k+1}(1)|[\hat{v}]_{2,l_1^{\hat{K}}},
\end{aligned}$$

where the first inequality is derived from $\hat{E}(\hat{f}(\hat{v}-\hat{\Pi}_1\hat{v})) = 0, \forall\hat{f}\in Q^{k-1}(\hat{K})$ and Theorem 2.3.1.

Since $l_k(t) = \frac{1}{2^k k!}\frac{d^k}{dt^k}(t^2-1)^k$, after integration by parts k times,

$$\hat{b}_{k+1}(1) = (k+\frac{1}{2})\int_{-1}^{1}\partial_s u(s,1)l_k(s)dx = (-1)^k(k+\frac{1}{2})\int_{-1}^{1}\partial_s^{k+1}u(s,1)L(s)ds,$$

where $L(s)$ is a polynomial of degree $2k$ by taking antiderivatives of $l_k(s)$ k times. Then by Cauchy-Schwarz inequality we have

$$\hat{b}_{k+1}(1) \leq C\left(\int_{-1}^{1}|\partial_s^{k+1}\hat{u}(s,1)|^2ds\right)^{\frac{1}{2}} \leq Ch^{k+\frac{1}{2}}|u|_{k+1,l_1^e}.$$

By (2.27), we get $|\hat{v}|_{2,l_1^{\hat{K}}} = h^{\frac{3}{2}}|\hat{v}|_{2,l_1^e} \leq Ch|v|_{2,e}$. Thus we have

$$\begin{aligned}
&\sum_{e\cap L_1\neq\emptyset}\int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}ds - \int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}d^hs \leq \sum_{e\cap L_1\neq\emptyset}C\|\hat{a}\|_{k,\infty,\hat{K}}|\hat{b}_{k+1}(1)||\hat{v}|_{2,l_1^{\hat{K}}} \\
=&\mathcal{O}(h^{k+\frac{3}{2}})\sum_{e\cap L_1\neq\emptyset}\|a\|_{k,\infty}|u|_{k+1,l_1^e}|v|_{2,e} = \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k,\infty}|u|_{k+1,L_1}\|v\|_{2,\Omega} = \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k,\infty}\|u\|_{k+2,\Omega}\|v\|_{2,\Omega},
\end{aligned}$$

where the trace inequality $\|u\|_{k+1,\partial\Omega} \leq C\|u\|_{k+2,\Omega}$ is used.

Combine all the estimates above, we get (2.38a). Since the $\frac{1}{2}$ order loss is only due to the line integral along L_1 and L_3, on which $v_{xx} = 0$ if $v \in V_0^h$, we get (2.38b). □

By all the discussions in this subsection, we have proven (2.29a) and (2.29b).

2.6 Homogeneous Dirichlet Boundary Conditions

2.6.1 V^h-ellipticity

In order to discuss the scheme (2.2), we need to show A_h satisfies V^h-ellipticity

$$\forall v_h \in V_0^h, \quad C\|v_h\|_1^2 \le A_h(v_h, v_h). \tag{2.43}$$

We first consider the V_h-ellipticity for the case $\mathbf{b} \equiv 0$.

Lemma 2.6.1. *Assume the coefficients in (2.9) satisfy that $\mathbf{b} \equiv 0$, both $c(x,y)$ and the eigenvalues of $\mathbf{a}(x,y)$ have a uniform upper bound and a uniform positive lower bound, then there exist two constants $C_1, C_2 > 0$ independent of mesh size h such that*

$$\forall v_h \in V_0^h, \quad C_1\|v_h\|_1^2 \le A_h(v_h, v_h) \le C_2\|v_h\|_1^2.$$

Proof. Let $Z_{0,\hat{K}}$ denote the set of $(k+1)\times(k+1)$ Gauss-Lobatto points on the reference cell $\hat{K}$. First we notice that the set $Z_{0,\hat{K}}$ is a $Q^k(\hat{K})$-unisolvent subset. Since the Gauss-Lobatto quadrature weights are strictly positive, we have

$$\forall \hat{p} \in Q^k(\hat{K}), \sum_{i=1}^{2}\langle \partial_i \hat{p}, \partial_i \hat{p}\rangle_{\hat{K}} = 0 \Longrightarrow \partial_i \hat{p} = 0 \text{ at quadrature points,}$$

where $i = 1, 2$ represents the spatial derivative on variable x_i respectively. Since $\partial_i \hat{p} \in Q^k(\hat{K})$ and it vanishes on a $Q^k(\hat{K})$-unisolvent subset, we have $\partial_i \hat{p} \equiv 0$. As a consequence, $\sqrt{\sum_{i=1}^n \langle \partial_i \hat{p}, \partial_i \hat{p}\rangle_h}$ defines a norm over the quotient space $Q^k(\hat{K})/Q^0(\hat{K})$. Since that $|\cdot|_{1,\hat{K}}$ is also a norm over the same quotient space, by the equivalence of norms over a finite dimensional space, we have

$$\forall \hat{p} \in Q^k(\hat{K}), \quad C_1|\hat{p}|_{1,\hat{K}}^2 \le \sum_{i=1}^{n}\langle \partial_i \hat{p}, \partial_i \hat{p}\rangle_{\hat{K}} \le C_2|\hat{p}|_{1,\hat{K}}^2.$$

On the reference cell $\hat{K}$, by the assumption on the coefficients, we have

$$C_1|\hat{v}_h|_{1,\hat{K}}^2 \le C_1\sum_{i}^{n}\langle \partial_i \hat{v}_h, \partial_i \hat{v}_h\rangle_{\hat{K}} \le \sum_{i,j=1}^{n}\left(\langle \hat{a}_{ij}\partial_i \hat{v}_h, \partial_j \hat{v}_h\rangle_{\hat{K}} + \langle \hat{c}\hat{v}_h, \hat{v}_h\rangle_{\hat{K}}\right) \le C_2\|\hat{v}_h\|_{1,\hat{K}}^2$$

Mapping these back to the original cell e and summing over all elements, by the equivalence of two norms $|\cdot|_1$ and $\|\cdot\|_1$ for the space $H_0^1(\Omega) \supset V_0^h$ [3], we get $C_1\|v_h\|_1^2 \leq A_h(v_h, v_h) \leq C_2\|v_h\|_1^2$. □

For discussing V_h-ellipticity when $\mathbf{b}$ is nonzero, by Young's inequality we have

$$|\langle \mathbf{b} \cdot \nabla v_h, v_h \rangle_h| \leq \sum_e \iint_e \frac{(\mathbf{b} \cdot \nabla v_h)^2}{4c} + c|v_h|^2 d^h x d^h y \leq \langle \frac{|\mathbf{b}|^2}{4c} \nabla v_h, \nabla v_h \rangle_h + \langle c v_h, v_h \rangle_h.$$

Thus we have

$$\langle \mathbf{a} \nabla v_h, \nabla v_h \rangle_h + \langle \mathbf{b} \cdot \nabla v_h, v_h \rangle_h + \langle c v_h, v_h \rangle_h \geq \langle \lambda_{\mathbf{a}} \nabla v_h, \nabla v_h \rangle_h - \langle \frac{|\mathbf{b}|^2}{4c} \nabla v_h, \nabla v_h \rangle_h,$$

where $\lambda_{\mathbf{a}}$ is smallest eigenvalue of $\mathbf{a}$. Then we have the following Lemma

Lemma 2.6.2. *Assume $4\lambda_{\mathbf{a}} c > |\mathbf{b}|^2$, then there exists a constant $C > 0$ independent of mesh size h such that*

$$\forall v_h \in V_0^h, \quad A_h(v_h, v_h) \geq C\|v_h\|_1^2.$$

2.6.2 Standard estimates for the dual problem

In order to apply the Aubin-Nitsche duality argument for establishing superconvergence of function values, we need certain estimates on a proper dual problem. Define $\theta_h := u_h - u_p$. Then we consider the dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega), \tag{2.44}$$

where $A^*(\cdot, \cdot)$ is the adjoint bilinear form of $A(\cdot, \cdot)$ such that

$$A^*(u, v) = A(v, u) = (\mathbf{a} \nabla v, \nabla u) + (\mathbf{b} \cdot \nabla v, u) + (cv, u).$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{2.45}$$

Notice that the right hand side of (2.45) is different from the right hand side of the scheme (2.2).

We need the following standard estimates on w_h for the dual problem.

Theorem 2.6.3. *Assume all coefficients in (2.9) are in $W^{2,\infty}(\Omega)$. Let w be defined in (2.44), w_h be defined in (2.45), and $\theta_h = u_h - u_p$. Assume elliptic regularity (2.12) and V^h ellipticity holds, we have*

$$\|w - w_h\|_1 \le Ch\|w\|_2,$$

$$\|w_h\|_2 \le C\|\theta_h\|_0.$$

Proof. By V^h ellipticity, we have $C_1\|w_h - v_h\|_1^2 \le A_h^*(w_h - v_h, w_h - v_h)$. By the definition of the dual problem, we have

$$A_h^*(w_h, w_h - v_h) = (\theta_h, w_h - v_h) = A^*(w, w_h - v_h), \quad \forall v_h \in V_0^h.$$

Thus for any $v_h \in V_0^h$, by Theorem 2.3.8, we have

$$\begin{aligned}
&C_1\|w_h - v_h\|_1^2 \le A_h^*(w_h - v_h, w_h - v_h)\\
=&A^*(w - v_h, w_h - v_h) + [A_h^*(w_h, w_h - v_h) - A^*(w, w_h - v_h)] + [A^*(v_h, w_h - v_h) - A_h^*(v_h, w_h - v_h)]\\
=&A^*(w - v_h, w_h - v_h) + [A(w_h - v_h, v_h) - A_h(w_h - v_h, v_h)]\\
\le&C\|w - v_h\|_1\|w_h - v_h\|_1 + Ch\|v_h\|_2\|w_h - v_h\|_1.
\end{aligned}$$

Thus

$$\|w - w_h\|_1 \le \|w - v_h\|_1 + \|w_h - v_h\|_1 \le C\|w - v_h\|_1 + Ch\|v_h\|_2. \tag{2.46}$$

Now consider $\Pi_1 w \in V_0^h$ where Π_1 is the piecewise Q^1 projection and its definition on each cell is defined through (2.8) on the reference cell. By the Bramble Hilbert Lemma Theorem 2.3.1 on the projection error, we have

$$\|w - \Pi_1 w\|_1 \le Ch\|w\|_2, \quad \|w - \Pi_1 w\|_2 \le C\|w\|_2, \tag{2.47}$$

thus $\|\Pi_1 w\|_2 \leq \|w\|_2 + \|w - \Pi_1 w\|_2 \leq C\|w\|_2$. By setting $v_h = \Pi_1 w$, from (2.46) we have

$$\|w - w_h\|_1 \leq C\|w - \Pi_1 w\|_1 + Ch\|\Pi_1 w\|_2 \leq Ch\|w\|_2. \tag{2.48}$$

By the inverse estimate on the piecewise polynomial $w_h - \Pi_1 w$, we get

$$\|w_h\|_2 \leq \|w_h - \Pi_1 w\|_2 + \|\Pi_1 w - w\|_2 + \|w\|_2 \leq Ch^{-1}\|w_h - \Pi_1 w\|_1 + C\|w\|_2. \tag{2.49}$$

By (2.47) and (2.48), we also have

$$\|w_h - \Pi_1 w\|_1 \leq \|w - \Pi_1 w\|_1 + \|w - w_h\|_1 \leq Ch\|w\|_2. \tag{2.50}$$

With (2.49), (2.50) and the elliptic regularity $\|w\|_2 \leq C\|\theta_h\|_0$, we get

$$\|w_h\|_2 \leq C\|w\|_2 \leq C\|\theta_h\|_0.$$

□

2.6.3 Superconvergence of function values

Theorem 2.6.4. *Assume $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ and $u(x,y) \in H^{k+3}(\Omega)$, $f(x,y) \in H^{k+2}(\Omega)$ with $k \geq 2$. Assume elliptic regularity (2.12) and V^h ellipticity holds. Then u_h, the numerical solution from scheme (2.2), is a $(k+2)$-th order accurate approximation to the exact solution u in the discrete 2-norm over all the $(k+1) \times (k+1)$ Gauss-Lobatto points:*

$$\|u_h - u\|_{l^2(\Omega)} = \mathcal{O}(h^{k+2})(\|u\|_{k+3,\Omega} + \|f\|_{k+2,\Omega}).$$

Proof. By Theorem 2.3.9 and Theorem 2.3.4, for any $v_h \in V_0^h$,

$$\begin{aligned}
& A_h(u - u_h, v_h) = [A(u, v_h) - A_h(u_h, v_h)] + [A_h(u, v_h) - A(u, v_h)] \\
= \quad & A(u, v_h) - A_h(u_h, v_h) + \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\
= \quad & [(f, v_h) - \langle f, v_h\rangle_h] + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|v_h\|_2.
\end{aligned}$$

Let $\theta_h = u_h - u_p$, then $\theta_h \in V_0^h$ due to the properties of the M-type projection. So by (2.29a) and Theorem 2.6.3, we get

$$\begin{aligned}
\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h - u, w_h) + A_h(u - u_p, w_h)\\
=& A_h(u - u_p, w_h) + \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2\\
=& \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|\theta_h\|_0,
\end{aligned}$$

thus

$$\|u_h - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}).$$

Finally, by the equivalence of the discrete 2-norm on Z_0 and the $L^2(\Omega)$ norm in finite-dimensional space V^h and Theorem 2.4.5, we obtain

$$\begin{aligned}
\|u_h - u\|_{l^2(\Omega)} \leq \|u_h - u_p\|_{l^2(\Omega)} + \|u_p - u\|_{l^2(\Omega)} \leq C\|u_h - u_p\|_0 + \|u_p - u\|_{l^2(\Omega)}\\
= \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}).
\end{aligned}$$

□

Remark 2.6.5. To extend the discussions to Neumann type boundary conditions, due to (2.29b) and Theorem 2.3.9, one can only prove $(k + \frac{3}{2})$-th order accuracy:

$$\|u_h - u\|_{l^2(\Omega)} = \mathcal{O}(h^{k+\frac{3}{2}})(\|u\|_{k+3} + \|f\|_{k+2}).$$

On the other hand, for solving a general elliptic equation, only $\mathcal{O}(h^{k+\frac{3}{2}})$ superconvergence at all Lobatto point can be proven for Neumann boundary conditions even for the full finite element scheme (2.1), see [9].

Remark 2.6.6. All key discussions can be extended to three-dimensional cases. For instance, M-type expansion has been used for discussing superconvergence for the three-dimensional case [9]. The most useful technique in Section 2.3.2 to obtain desired consistency error estimate is to derive error cancellations between neighboring cells through integration by parts on suitable interpolation polynomials, which still seems possible on rectangular meshes in three dimensions.

2.7 Nonhomogeneous Dirichlet Boundary Conditions

We consider a two-dimensional elliptic problem on $\Omega = (0,1)^2$ with nonhomogeneous Dirichlet boundary condition,

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b} \cdot \nabla u + cu &= f \text{ on } \Omega \\ u &= g \text{ on } \partial\Omega. \end{aligned} \tag{2.51}$$

Assume there is a function $\bar{g} \in H^1(\Omega)$ as a smooth extension of g so that $\bar{g}|_{\partial\Omega} = g$. The variational form is to find $\tilde{u} = u - \bar{g} \in H_0^1(\Omega)$ satisfying

$$A(\tilde{u}, v) = (f, v) - A(\bar{g}, v), \quad \forall v \in H_0^1(\Omega). \tag{2.52}$$

In practice, $\bar{g}$ is not used explicitly. By abusing notations, the most convenient implementation is to consider

$$g(x,y) = \begin{cases} 0, & \text{if} \quad (x,y) \in (0,1) \times (0,1), \\ g(x,y), & \text{if} \quad (x,y) \in \partial\Omega, \end{cases}$$

and $g_I \in V^h$ which is defined as the Q^k Lagrange interpolation at $(k+1) \times (k+1)$ Gauss-Lobatto points for each cell on Ω of $g(x,y)$. Namely, $g_I \in V^h$ is the piecewise P^k interpolation of g along the boundary grid points and $g_I = 0$ at the interior grid points. The numerical scheme is to find $\tilde{u}_h \in V_0^h$, s.t.

$$A_h(\tilde{u}_h, v_h) = \langle f, v_h \rangle_h - A_h(g_I, v_h), \quad \forall v_h \in V_0^h. \tag{2.53}$$

Then $u_h = \tilde{u}_h + g_I$ will be our numerical solution for (2.51). Notice that (2.53) is not a straightforward approximation to (2.52) since $\bar{g}$ is never used. Assuming elliptic regularity and V^h ellipticity hold, we will show that $u_h - u$ is of $(k+2)$-th order in the discrete 2-norm over all $(k+1) \times (k+1)$ Gauss-Lobatto points.

2.7.1 An auxiliary scheme

In order to discuss the superconvergence of (2.53), we need to prove the superconvergence of an auxiliary scheme. Notice that we discuss the auxiliary scheme only for proving the accuracy of (2.53). In practice one should not implement the auxiliary scheme since (2.53) is a much more convenient implementation with the same accuracy.

Let $\bar{g}_p \in V^h$ be the piecewise M-type Q^k projection of the smooth extension function $\bar{g}$, and define $g_p \in V^h$ as $g_p = \bar{g}_p$ on $\partial\Omega$ and $g_p = 0$ at all the inner grids. The auxiliary scheme is to find $\tilde{u}_h^* \in V_0^h$ satisfying

$$A_h(\tilde{u}_h^*, v_h) = \langle f, v_h \rangle_h - A_h(g_p, v_h), \quad \forall v_h \in V_0^h, \tag{2.54}$$

Then $u_h^* = \tilde{u}_h^* + g_p$ is the numerical solution for problem (2.52). Define $\theta_h = u_h^* - u_p$, then by Theorem 2.4.1 we have $\theta_h \in V_0^h$. Following Section 2.6.2, define the following dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega). \tag{2.55}$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{2.56}$$

Notice that the dual problem has homogeneous Dirichlet boundary conditions. By Theorem 2.3.9, Theorem 2.3.4, for any $v_h \in V_0^h$,

$$\begin{aligned} & A_h(u - u_h^*, v_h) = [A(u, v_h) - A_h(u_h^*, v_h)] + [A_h(u, v_h) - A(u, v_h)] \\ = \quad & A(u, v_h) - A_h(u_h^*, v_h) + \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\ = \quad & [(f, v_h) - \langle f, v_h \rangle_h] + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|v_h\|_2. \end{aligned}$$

By (2.29a) and Theorem 2.6.3, we get

$$
\begin{aligned}
&\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h^* - u, w_h) + A_h(u - u_p, w_h) \\
=&A_h(u - u_p, w_h) + \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 \\
=&\mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|\theta_h\|_0,
\end{aligned}
$$

thus $\|u_h^* - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})$. So Theorem 2.6.4 still holds for the auxiliary scheme (2.54):

$$
\|u_h^* - u\|_{l^2(\Omega)} = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}). \tag{2.57}
$$

2.7.2 The main result

In order to extend Theorem 2.6.4 to (2.53), we only need to prove

$$
\|u_h - u_h^*\|_0 = \mathcal{O}(h^{k+2}).
$$

The difference between (2.54) and (2.53) is

$$
A_h(\tilde{u}_h^* - \tilde{u}_h, v_h) = A_h(g_I - g_p, v_h), \quad \forall v_h \in V_0^h. \tag{2.58}
$$

We need the following Lemma.

Lemma 2.7.1. *Assuming $u \in H^{k+4}(\Omega)$ for $k \geq 2$, with g_I and g_p being defined as in this Section, then we have*

$$
A_h(g_I - g_p, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}\|v_h\|_{2,\Omega}, \quad \forall v_h \in V_0^h. \tag{2.59}
$$

Proof. For simplicity, we ignore the subscript $_h$ of v_h in this proof and all the following v are in V^h.

Notice that $g_I - g_p \equiv 0$ in interior cells. Thus we only consider cells adjacent to $\partial\Omega$. Let L_1, L_2, L_3 and L_4 denote the top, left, bottom and right boundary edges of $\bar{\Omega} = [0,1] \times [0,1]$

respectively. Without loss of generality, we consider cell $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$ adjacent to the left boundary L_2, i.e., $x_e - h = 0$. Let l_1^e, l_2^e, l_3^e and l_4^e denote the top, left, bottom and right boundary edges of e respectively.

On $l_2 \subset L_2$, Let $\phi_{ij}(x, y), i, j = 0, 1, \dots, k$, be Lagrange basis functions on edge l_2^e for the $(k+1) \times (k+1)$ Gauss-Lobatto points in cell e. Then $g_I - g_p = \sum_{i,j=0}^{k} \lambda_{ij}\phi_{ij}(x, y)$ and $|\lambda_{ij}| \leq \|g_I - g_p\|_{l^\infty(\Omega)}$. Due to Sobolev's embedding, we have $u \in W^{k+2,\infty}(\Omega)$. By Theorem 2.4.5, we have

$$\|g_I - g_p\|_{l^\infty(\Omega)} \leq \|u - u_p\|_{l^\infty(\Omega)} = \mathcal{O}(h^{k+2})\|u\|_{k+2,\infty,\Omega} = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}.$$

Thus we get $\forall v \in V_0^h$,

$$\langle a(g_I - g_p)_x, v_x\rangle_e = \langle a \sum_{i,j=0}^{k} \lambda_{ij}\phi_{ij}(x, y)_x, v_x\rangle_e \leq C\|a\|_{\infty,\Omega} \max_{i,j} |\lambda_{ij}||\langle \sum_{i,j=0}^{k} \phi_{ij}(x, y)_x, v_x\rangle_e|.$$

Since for polynomials on $\hat{K}$ all the norm are equivalent, we have

$$|\langle \sum_{i,j=0}^{k} \phi_{ij}(x, y)_x, v_x\rangle_e| = |\langle \sum_{i,j=0}^{k} \hat{\phi}_{ij}(s, t)_s, \hat{v}_s\rangle_{\hat{K}}| \leq C|\hat{v}_s|_{\infty,\hat{K}} \leq C|v|_{1,\hat{K}} = C|v|_{1,e},$$

which implies

$$\langle a(g_I - g_p)_x, v_x\rangle_h \leq C\|a\|_{\infty,\Omega} \sum_e \max_{i,j} |\lambda_{ij}||v|_{1,e} = \mathcal{O}(h^{k+2})\|a\|_{\infty,\Omega}\|u\|_{k+4,\Omega}\|v\|_{2,\Omega}$$

Similarly, for any $v \in V_0^h$, we have

$$\begin{aligned}
\langle a(g_I - g_p)_y, v_y\rangle_h = &\quad \mathcal{O}(h^{k+2})\|a\|_\infty\|u\|_{k+4}\|v\|_2, \\
\langle a(g_I - g_p)_x, v_y\rangle_h = &\quad \mathcal{O}(h^{k+2})\|a\|_\infty\|u\|_{k+4}\|v\|_2, \\
\langle \mathbf{b} \cdot \nabla(g_I - g_p), v\rangle_h = &\quad \mathcal{O}(h^{k+2})\|\mathbf{b}\|_\infty\|u\|_{k+4}\|v\|_2, \\
\langle c(g_I - g_p), v\rangle_h = &\quad \mathcal{O}(h^{k+2})\|c\|_\infty\|u\|_{k+4}\|v\|_2.
\end{aligned}$$

Thus we conclude that

$$A_h(g_I - g_p, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|v_h\|_2, \quad \forall v_h \in V_0^h.$$

□

By (2.58) and Lemma 2.7.1, we have

$$A_h(\tilde{u}_h^* - \tilde{u}_h, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|v_h\|_2, \quad \forall v_h \in V_0^h. \tag{2.60}$$

Let $\theta_h = \tilde{u}_h^* - \tilde{u}_h \in V_0^h$. Following Section 2.6.2, define the following dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega). \tag{2.61}$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{2.62}$$

By (2.60) and Theorem 2.6.3, we get

$$\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h^*(w_h, \theta_h) = A_h(\tilde{u}_h^* - \tilde{u}_h, w_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|w_h\|_2 = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|\theta_h\|_0,$$

thus $\|\tilde{u}_h^* - \tilde{u}_h\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})\|u\|_{k+4}$. By equivalence of norms for polynomials, we have

$$\|\tilde{u}_h^* - \tilde{u}_h\|_{l^2(\Omega)} \leq C\|\tilde{u}_h^* - \tilde{u}_h\|_0 = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}. \tag{2.63}$$

Notice that both $\tilde{u}_h$ and $\tilde{u}_h^*$ are constant zero along $\partial\Omega$, and $u_h|_{\partial\Omega} = g_I$ is the Lagrangian interpolation of g along $\partial\Omega$. With (2.57), we have proven the following main result.

Theorem 2.7.2. *Assume elliptic regularity (2.12) and V^h ellipticity holds. For a non-homogeneous Dirichlet boundary problem (2.51), with suitable smoothness assumptions for $k \geq 2$, $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$, the exact solution of (2.52) $u(x, y) = \tilde{u} + \bar{g} \in H^{k+4}(\Omega)$ and*

$f(x,y) \in H^{k+2}(\Omega)$, the numerical solution u_h by scheme (2.53) is a $(k+2)$-th order accurate approximation to u in the discrete 2-norm over all the $(k+1)\times(k+1)$ Gauss-Lobatto points:

$$\|u_h - u\|_{l^2(\Omega)} = \mathcal{O}(h^{k+2})(\|u\|_{k+4} + \|f\|_{k+2}).$$

2.8 Neumann Boundary Conditions

Consider the elliptic problem with Neumann boundary condition:

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b}\cdot\nabla u + cu =& f \text{ in } \Omega \\ (\mathbf{a}\nabla u)\cdot\mathbf{n} =& g \text{ on } \partial\Omega. \end{aligned} \tag{2.64}$$

The corresponding variational form is to find $u \in H^1(\Omega)$ to satisfy

$$A(u,v) = (f,v) + \int_{\partial\Omega} gv d\mu, \quad \forall v \in H^1(\Omega). \tag{2.65}$$

In this section we consider the problem:

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b}\cdot\nabla u + cu =& f \text{ in } \Omega \\ (\mathbf{a}\nabla u)\cdot\mathbf{n} =& g_I \text{ on } \partial\Omega. \end{aligned} \tag{2.66}$$

The corresponding variational form is to find $u \in H^1(\Omega)$ to satisfy

$$A(u,v) = (f,v) + \int_{\partial\Omega} g_I v d\mu, \quad \forall v \in H^1(\Omega). \tag{2.67}$$

The numerical scheme is to find $u_h \in V^h(\Omega)$ to satisfy

$$A_h(u_h,v_h) = \langle f, v_h\rangle_h + \int_{\partial\Omega} g v_h d^h\mu, \quad \forall v \in H^1(\Omega). \tag{2.68}$$

2.8.1 Quadrature error estimates

Theorem 2.8.1. *For the equation* (2.64), *assume the coefficients* $\mathbf{a}(x,y), \mathbf{b}(x,y), c \in W^{k+2,\infty}(\Omega)$ *and* $u \in H^{k+3}(\Omega)$ *is the solution of* (2.64), $k \geq 2$, *then we have*

$$A(u,v_h)-A_h(u,v_h) = \int_{\partial\Omega} gv d\mu - \int_{\partial\Omega} gv_h d^h\mu + \int_{\partial\Omega} (g_I - g)v_h d\mu + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. By the proof for (2.19a) and (2.19b), we have

$$\begin{aligned}
&(a^{11}\partial_x u, \partial_x v_h) - \langle a^{11}\partial_x u, \partial_x v_h\rangle_h \\
=&\sum_{e\cap L_4\neq\emptyset}\int_{-1}^{1}(\hat{a}^{11}\hat{u}_s)_I(1,t)\hat{v}_h(1,t)dt - \sum_{e\cap L_4\neq\emptyset}\int_{-1}^{1}(\hat{a}^{11}\hat{u}_s)_I(1,t)\hat{v}_h(1,t)d^h t \\
&- \sum_{e\cap L_2\neq\emptyset}\int_{-1}^{1}(\hat{a}^{11}\hat{u}_s)_I(-1,t)\hat{v}_h(-1,t)dt + \sum_{e\cap L_2\neq\emptyset}\int_{-1}^{1}(\hat{a}^{11}\hat{u}_s)_I(-1,t)\hat{v}_h(-1,t)d^h t \\
&+ \mathcal{O}(h^{k+2})\|a^{11}\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\
=&\int_{L_4}(a^{11}u_x)_I(1,y)v_h(1,y)dy - \int_{L_4}(a^{11}u_x)_I(1,y)v_h(1,y)d^h y \\
&- \int_{L_2}(a^{11}u_x)_I(-1,y)v_h(-1,y)dy + \int_{L_2}(a^{11}u_x)_I(-1,y)v_h(-1,y)d^h y \\
&+ \mathcal{O}(h^{k+2})\|a^{11}\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2.
\end{aligned} \tag{2.69}$$

Follow the same procedure we have

$$\begin{aligned}
&(a^{12}\partial_y u, \partial_x v_h) - \langle a^{12}\partial_y u, \partial_x v_h\rangle_h \\
=&\sum_{e\cap L_4\neq\emptyset}\int_{-1}^{1}(\hat{a}^{12}\hat{u}_t)_I(1,t)\hat{v}_h(1,t)dt - \sum_{e\cap L_4\neq\emptyset}\int_{-1}^{1}(\hat{a}^{12}\hat{u}_t)_I(1,t)\hat{v}_h(1,t)d^h t \\
&- \sum_{e\cap L_2\neq\emptyset}\int_{-1}^{1}(\hat{a}^{12}\hat{u}_t)_I(-1,t)\hat{v}_h(-1,t)dt + \sum_{e\cap L_2\neq\emptyset}\int_{-1}^{1}(\hat{a}^{12}\hat{u}_t)_I(-1,t)\hat{v}_h(-1,t)d^h t \\
&+ \mathcal{O}(h^{k+2})\|a^{12}\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2. \\
=&\int_{L_4}(a^{12}u_y)_I(1,y)v_h(1,y)dy - \int_{L_4}(a^{12}u_y)_I(1,y)v_h(1,y)d^h y \\
&- \int_{L_2}(a^{12}u_y)_I(-1,y)v_h(-1,y)dy + \int_{L_2}(a^{12}u_y)_I(-1,y)v_h(-1,y)d^h y \\
&+ \mathcal{O}(h^{k+2})\|a^{11}\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2.
\end{aligned} \tag{2.70}$$

Add (2.69) and (2.70) on both sides then we have

$$\begin{aligned}
&(a^{11}\partial_x u + a^{12}\partial_y u, \partial_x v_h) - \langle a^{11}\partial_x u + a^{12}\partial_y u, \partial_x v_h\rangle_h \\
=&\int_{L_4} (a^{11}u_x + a^{12}u_y)_I(1,y)v_h(1,y)dy - \int_{L_4} (a^{11}u_x + a^{12}u_y)_I(1,y)v_h(1,y)d^h y \\
&- \int_{L_2} (a^{11}u_x + a^{12}u_y)_I(-1,y)v_h(-1,y)dy + \int_{L_2} (a^{11}u_x + a^{12}u_y)_I(-1,y)v_h(-1,y)d^h y \\
&+ \mathcal{O}(h^{k+2})\left(\|a^{11}\|_{k+2,\infty} + \|a^{12}\|_{k+2,\infty}\right)\|u\|_{k+3}\|v_h\|_2.
\end{aligned} \tag{2.71}$$

Note on the boundary L^2 and L^4, by the Neumann boundary condition, we have

$$\begin{aligned}
a^{11}\partial_x u + a^{12}\partial_y u =& g, \text{ on } L^4 \\
a^{11}\partial_x u + a^{12}\partial_y u =& - g \text{ on } L^2,
\end{aligned}$$

thus with Lemma 2.3.4,

$$\begin{aligned}
&(a^{11}\partial_x u + a^{12}\partial_y u, \partial_x v_h) - \langle a^{11}\partial_x u + a^{12}\partial_y u, \partial_x v_h\rangle_h \\
=&\int_{L_4} g_I(1,y)v_h(1,y)dy - \int_{L_4} g(1,y)v_h(1,y)d^h y + \int_{L_2} g_I(-1,y)v_h(-1,y)dy - \int_{L_2} g(-1,y)v_h(-1,y)d^h y \\
&+ \mathcal{O}(h^{k+2})\left(\|a^{11}\|_{k+2,\infty} + \|a^{12}\|_{k+2,\infty}\right)\|u\|_{k+3}\|v_h\|_2
\end{aligned} \tag{2.72}$$

By the same argument and denoting the upper boundary and lower boundary with L_1 and L_3 respectively, we have

$$\begin{aligned}
&(a^{12}\partial_x u + a^{22}\partial_y u, \partial_y v_h) - \langle a^{12}\partial_x u + a^{22}\partial_y u, \partial_y v_h\rangle_h \\
=&\int_{L_1} g_I(x,1)v_h(x,1)dx - \int_{L_1} g(x,1)v_h(x,1)d^h x + \int_{L_3} g_I(x,-1)v_h(x,-1)dx - \int_{L_3} g(x,-1)v_h(x,-1)d^h x \\
&+ \mathcal{O}(h^{k+2})\left(\|a^{12}\|_{k+2,\infty} + \|a^{22}\|_{k+2,\infty}\right)\|u\|_{k+3}\|v_h\|_2.
\end{aligned} \tag{2.73}$$

By (2.21), (2.22), (2.72) and (2.73), we have

$$\begin{aligned}
&A(u, v_h) - \int_{\partial\Omega} gv d\mu - A_h(u, v_h) + \int_{\partial\Omega} gv_h d^h\mu \\
=&\int_{\partial\Omega} g_I v d\mu - \int_{\partial\Omega} gv_h d^h\mu - \int_{\partial\Omega} gv d\mu + \int_{\partial\Omega} gv_h d^h\mu + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 \qquad (2.74) \\
=&\int_{\partial\Omega} (g_I - g)v_h d\mu + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2.
\end{aligned}$$

□

By the same argument, we immediately have the following theorem.

Theorem 2.8.2. *For the equation (2.66), assume the coefficients* $\mathbf{a}(x, y), \mathbf{b}(x, y), c \in W^{k+2,\infty}(\Omega)$ *and* $u \in H^{k+3}(\Omega)$ *is the solution of (2.66),* $k \geq 2$*, then we have*

$$A(u, v_h) - A_h(u, v_h) = \int_{\partial\Omega} gv_h d\mu - \int_{\partial\Omega} gv_h d^h\mu + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Remark 2.8.3. For the homogeneous Neumann boundary case (2.64) and (2.66) are the same. To analyze the accuracy of scheme (2.68) to problem (2.64), one only need to estimate the extra error by $\int_{\partial\Omega}(g_I - g)v d\mu$. For simplicity, we only analyze the accuracy of scheme (2.68) to problem (2.66) for the case $k \geq 2$ and to problem (2.64) for the case $k \geq 3$.

2.8.2 Superconvergence of function values

The V^h-elliptic and standard estimates for the dual problem are stated as Lemma 2.6.1 Theorem 2.6.3, then we have the superconvergence of function values.

Theorem 2.8.4. *For problem (2.66), assume* $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ *and* $u(x, y) \in H^{k+3}(\Omega)$*,* $f(x, y) \in H^{k+2}(\Omega)$ *with* $k \geq 2$*. Assume elliptic regularity (2.12) and* V^h *ellipticity holds. Then the numerical solution* u_h *from scheme (2.68), in the discrete 2-norm over all the* $(k + 1) \times (k + 1)$ *Gauss-Lobatto points, is a* $(k + 2)$*-th order accurate approximation to the* *exact solution u if the coefficient matrix $\mathbf{a}$ is diagonal and is at least a $(k+\frac{3}{2})$-th order accurate approximation for general coefficient matrix $\mathbf{a}$:*

$$\|u_h - u\|_{l^2(\Omega)} = \begin{cases} \mathcal{O}(h^{k+2})\left(\|u\|_{k+3} + \|f\|_{k+2}\right)\|v_h\|_2, & \textit{if } \mathbf{a} \textit{ is diagonal} \\ \mathcal{O}(h^{k+\frac{3}{2}})\left(\|u\|_{k+3} + \|f\|_{k+2}\right)\|v_h\|_2, & \textit{otherwise.} \end{cases} .$$

Proof. By Theorem 2.8.2 and Theorem 2.3.4, for any $v_h \in V_0^h$,

$$\begin{aligned} & A_h(u-u_h, v_h) = [A(u,v_h) - A_h(u_h,v_h)] + [A_h(u,v_h) - A(u,v_h)] \\ = \;& A(u,v_h) - A_h(u_h,v_h) - \textstyle\int_{\partial\Omega} gv d\mu + \int_{\partial\Omega} gv_h d^h\mu + \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\ = \;& [(f,v_h) - \langle f, v_h\rangle_h] + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 \\ = \;& \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|v_h\|_2. \end{aligned}$$

Let $\theta_h = u_h - u_p$, then $\theta_h \in V_0^h$ due to the properties of the M-type projection. So by (2.29a) and Theorem 2.6.3, we get

$$\begin{aligned} & \|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h - u, w_h) + A_h(u - u_p, w_h) \\ =\;& A_h(u-u_p, w_h) + \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 \\ =\;& \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|\theta_h\|_0, \end{aligned}$$

thus

$$\|u_h - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}).$$

Finally, by the equivalence of the discrete 2-norm on Z_0 and the $L^2(\Omega)$ norm in finite-dimensional space V^h and Theorem 2.4.5, we obtain

$$\begin{aligned} \|u_h - u\|_{l^2(\Omega)} \leq \|u_h - u_p\|_{l^2(\Omega)} + \|u_p - u\|_{l^2(\Omega)} &\leq C\|u_h - u_p\|_0 + \|u_p - u\|_{l^2(\Omega)} \\ &= \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}). \end{aligned}$$

□

Theorem 2.8.5. *For problem* (2.64)*, assume* $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ *and* $u(x,y) \in H^{k+3}(\Omega)$*,* $f(x,y) \in H^{k+2}(\Omega)$ *with* $k \geq 3$*. Assume elliptic regularity* (2.12) *and* V^h *ellipticity holds. Then the numerical solution* u_h *from scheme* (2.68)*, in the discrete 2-norm over all the* $(k+1) \times (k+1)$ *Gauss-Lobatto points, is a* $(k+2)$*-th order accurate approximation to the exact solution* u *if the coefficient matrix* $\mathbf{a}$ *is diagonal and is at least a* $(k+\frac{3}{2})$*-th order accurate approximation for general coefficient matrix* $\mathbf{a}$*:*

$$\|u_h - u\|_{l^2(\Omega)} = \begin{cases} \mathcal{O}(h^{k+2})\left(\|u\|_{k+3} + \|f\|_{k+2} + \|g\|_{k+3,\partial\Omega}\right)\|v_h\|_2, & \textit{if } \mathbf{a} \textit{ is diagonal} \\ \mathcal{O}(h^{k+\frac{3}{2}})\left(\|u\|_{k+3} + \|f\|_{k+2} + \|g\|_{k+3,\partial\Omega}\right)\|v_h\|_2, & \textit{otherwise.} \end{cases}.$$

Proof. By Theorem 2.8.1 and Lemma 2.3.6, we have, $\forall v_h \in V^h$

$$A(u, v_h) - A_h(u, v_h) = \int_{\partial\Omega} g v d\mu - \int_{\partial\Omega} g v_h d^h\mu + \mathcal{O}(h^{k+2})\left(\|u\|_{k+3} + \|g\|_{k+3,\partial\Omega}\right)\|v_h\|_2.$$

Then the rest will be the same as the proof of Theorem 2.8.4. □

Remark 2.8.6. All key discussions can be extended to three-dimensional cases.

2.9 Finite Difference Implementation

In this section we present the finite difference implementation of the scheme (2.53) for the case $k = 2$ on a uniform mesh. The finite difference implementation of the nonhomogeneous Dirichlet boundary value problem is based on a homogeneous Neumann boundary value problem, which will be discussed first. We demonstrate how it is derived for the one-dimensional case then give the two-dimensional implementation. It provides efficient assembling of the stiffness matrix and one can easily implement it in MATLAB. Implementations for higher order elements or quasi-uniform meshes can be similarly derived, even though it will no longer be a conventional finite difference scheme on a uniform grid.

2.9.1 One-dimensional case

Consider a homogeneous Neumann boundary value problem $-(au) = f$ on $[0,1], u(0) = 0, u(1) = 0$, and its variational form is to seek $u \in H^1([0,1])$ satisfying

$$(au, v) = (f, v), \quad \forall v \in H^1([0,1]). \tag{2.75}$$

Consider a uniform mesh $x_i = ih$, $i = 0, 1, \ldots, n+1$, $h = \frac{1}{n+1}$. Assume n is odd and let $N = \frac{n+1}{2}$. Define intervals $I_k = [x_{2k}, x_{2k+2}]$ for $k = 0, \ldots, N-1$ as a finite element mesh for P^2 basis. Define

$$V^h = \{v \in C^0([0,1]) : v|_{I_k} \in P^2(I_k), k = 0, \ldots, N-1\}.$$

Let $\{v_i\}_{i=0}^{n+1} \subset V^h$ be a basis of V^h such that $v_i(x_j) = \delta_{ij}$, $i, j = 0, 1, \ldots, n+1$. With 3-point Gauss-Lobatto quadrature, the C^0-P^2 finite element method for (2.75) is to seek $u_h \in V^h$ satisfying

$$\langle au_h, v_i \rangle_h = \langle f, v_i \rangle_h, \quad i = 0, 1, \ldots, n+1. \tag{2.76}$$

Let $u_j = u_h(x_j)$, $a_j = a(x_j)$ and $f_j = f(x_j)$ then $u_h(x) = \sum\limits_{j=0}^{n+1} u_j v_j(x)$. We have

$$\sum_{j=0}^{n+1} u_j \langle av_j, v_i \rangle_h = \langle au_h, v_j \rangle_h = \langle f, v_i \rangle_h = \sum_{j=0}^{n+1} f_j \langle v_j, v_i \rangle_h, \quad i = 0, 1, \ldots, n+1.$$

The matrix form of this scheme is $\bar{S}\bar{\mathbf{u}} = \bar{M}\bar{\mathbf{f}}$, where

$$\bar{\mathbf{u}} = \Big[u_0, u_1, \ldots, u_n, u_{n+1}\Big]^T, \quad \bar{\mathbf{f}} = \Big[f_0, f_1, \ldots, f_n, f_{n+1}\Big]^T,$$

the stiffness matrix $\bar{S}$ is has size $(n+2) \times (n+2)$ with (i,j)-th entry as $\langle av_i, v_j \rangle_h$, and the lumped mass matrix M is a $(n+2) \times (n+2)$ diagonal matrix with diagonal entries $h\left(\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{2}{3}, \ldots, \frac{2}{3}, \frac{4}{3}, \frac{1}{3}\right)$.

Next we derive an explicit representation of the matrix $\bar{S}$. Since basis functions $v_i \in V^h$ and $u_h(x)$ are not C^1 at the knots x_{2k} $(k = 1, 2, \dots, N-1)$, their derivatives at the knots are double valued. We will use superscripts $+$ and $-$ to denote derivatives obtained from the right and from the left respectively, e.g., v_{2k}^+ and v_{2k+2}^- denote the derivatives of v_{2k} and v_{2k+2} respectively in the interval $I_k = [x_{2k}, x_{2k+2}]$. Then in the interval $I_k = [x_{2k}, x_{2k+2}]$ we have the following representation of derivatives

$$\begin{bmatrix} v_{2k}^+(x) \\ v_{2k+1}(x) \\ v_{2k+2}^-(x) \end{bmatrix} = \frac{1}{2h} \begin{bmatrix} -3 & 4 & -1 \\ -1 & 0 & 1 \\ 1 & -4 & 3 \end{bmatrix} \begin{bmatrix} v_{2k}(x) \\ v_{2k+1}(x) \\ v_{2k+2}(x) \end{bmatrix}. \tag{2.77}$$

By abusing notations, we use $(v_i)_{2k}$ to denote the average of two derivatives of v_i at the knots x_{2k}:

$$(v_i)_{2k} = \frac{1}{2}[(v_i)_{2k}^- + (v_i)_{2k}^+].$$

Let $[v_i]$ denote the difference between the right derivative and left derivative:

$$[v_i]_0 = [v_i]_{n+2} = 0, \quad [v_i]_{2k} := (v_i)_{2k}^+ - (v_i)_{2k}^-, \quad k = 1, 2, \dots, N-1.$$

Then at the knots, we have

$$(v_i)_{2k}^-(v_j)_{2k}^- + (v_i)_{2k}^+(v_j)_{2k}^+ = 2(v_i)_{2k}(v_j)_{2k} + \frac{1}{2}[v_i]_{2k}[v_j]_{2k}. \tag{2.78}$$

We also have

$$\langle av_j, v_i \rangle_{I_{2k}} = h\left[\frac{1}{3}a_{2k}(v_j)_{2k}^+(v_i)_{2k}^+ + \frac{4}{3}a_{2k+1}(v_j)_{2k+1}(v_i)_{2k+1} + \frac{1}{3}a_{2k+2}(v_j)_{2k+2}^-(v_i)_{2k+2}^-\right]. \tag{2.79}$$

Let $\mathbf{v}_i$ denote a column vector of size $n+2$ consisting of grid point values of $v_i(x)$. Plugging (2.78) into (2.79), with (2.77), we get

$$\langle av_j, v_i \rangle_h = \sum_{k=0}^{N-1} \langle av_j, v_i \rangle_{I_{2k}} = \frac{1}{h}\mathbf{v}_i^T(D^T WAD + E^T WAE)\mathbf{v}_j,$$

where A is a diagonal matrix with diagonal entries $a_0, a_1, \dots, a_n, a_{n+1}$, and

$$W = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \dots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n+2)\times(n+2)},$$

$$D = \frac{1}{2}\begin{pmatrix} -3 & 4 & -1 & & & & & & & & \\ -1 & 0 & 1 & & & & & & & & \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & & & \\ & & -1 & 0 & 1 & & & & & & \\ & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & \\ & & & & -1 & 0 & 1 & & & & \\ & & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & -1 & 0 & 1 & & \\ & & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & & & -1 & 0 & 1 \\ & & & & & & & & 1 & -4 & 3 \end{pmatrix}_{(n+2)\times(n+2)}, E = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & & & & & & & & \\ 0 & 0 & 0 & & & & & & & & \\ -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & & & \\ & & 0 & 0 & 0 & & & & & & \\ & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & \\ & & & & 0 & 0 & 0 & & & & \\ & & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & 0 & 0 & 0 & & \\ & & & & & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} \\ & & & & & & & & 0 & 0 & 0 \\ & & & & & & & & 0 & 0 & 0 \end{pmatrix}_{(n+2)\times(n+2)}.$$

Since $\{v_i\}_{i=0}^{n}$ are the Lagrangian basis for V^h, we have

$$\bar{S} = \frac{1}{h}(D^T WAD + E^T WAE). \tag{2.80}$$

Now consider the one-dimensional Dirichlet boundary value problem:

$$\begin{aligned} -(au) =& f \text{ on } [0,1], \\ u(0) = \sigma_1, &\quad u(1) = \sigma_2. \end{aligned}$$

Consider the same mesh as above and define

$$V_0^h = \{v \in C^0([0,1]) : v|_{I_k} \in P^2(I_k), k = 0, \dots, N-1; v(0) = v(1) = 0\}.$$

Then $\{v_i\}_{i=1}^{n} \subset V^h$ is a basis of V_0^h for $\{v_i\}_{i=0}^{n+1}$ defined above. The one-dimensional version of (2.53) is to seek $u_h \in V_0^h$ satisfying

$$\begin{aligned} \langle au_h, v_i\rangle_h &= \langle f, v_i\rangle_h - \langle ag_I, v_i\rangle_h, \quad i = 1, 2, \dots, n, \\ g_I(x) &= \sigma_0 v_0(x) + \sigma_1 v_{n+1}(x). \end{aligned} \tag{2.81}$$

Notice that we can obtain (2.81) by simply setting $u_h(0) = \sigma_0$ and $u_h(1) = \sigma_1$ in (2.76). So the finite difference implementation of (2.81) is given as follows:

1. Assemble the $(n+2) \times (n+2)$ stiffness matrix $\bar{S}$ for homogeneous Neumann problem as in (2.80).

2. Let S denote the $n \times n$ submatrix $\bar{S}(2:n+1, 2:n+1)$, i.e., $[\bar{S}_{ij}]$ for $i, j = 2, \cdots, n+1$.

3. Let $\mathbf{l}$ denote the $n \times 1$ submatrix $\bar{S}(2:n+1, 1)$ and $\mathbf{r}$ denote the $n \times 1$ submatrix $\bar{S}(2:n+1, n+2)$, which correspond to $v_0(x)$ and $v_{n+1}(x)$.

4. Let $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}^T$ and $\mathbf{f} = \begin{bmatrix} f_1 & f_2 & \cdots & f_n \end{bmatrix}^T$. Define $\mathbf{w} = \begin{bmatrix} \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{2}{3}, \ldots, \frac{2}{3}, \frac{4}{3} \end{bmatrix}$ as a column vector of size n. The scheme (2.81) can be implemented as

$$S\mathbf{u} = h\mathbf{w}^T\mathbf{f} - \sigma_0\mathbf{l} - \sigma_1\mathbf{r}.$$

2.9.2 Notations and tools for the two-dimensional case

We will need two operators:

- Kronecker product of two matrices: if A is $m \times n$ and B is $p \times q$, then $A \otimes B$ is $mp \times nq$ give by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \vdots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

- For a $m \times n$ matrix X, $vec(X)$ denotes the vectorization of the matrix X by rearranging X into a vector column by column.

The following properties will be used:

1. $(A \otimes B)(C \otimes D) = AC \otimes BD$.

2. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

3. $(B^T \otimes A)vec(X) = vec(AXB)$.

4. $(A \otimes B)^T = A^T \otimes B^T$.

Consider a uniform grid (x_i, y_j) for a rectangular domain $\bar{\Omega} = [0,1] \times [0,1]$ where $x_i = ih_x$, $i = 0, 1, \ldots, n_x + 1$, $h_x = \frac{1}{n_x+1}$ and $y_j = jh_y$, $j = 0, 1, \ldots, n_y + 1$, $h_y = \frac{1}{n_y+1}$.

Assume n_x and n_y are odd and let $N_x = \frac{n_x+1}{2}$ and $N_y = \frac{n_y+1}{2}$. We consider rectangular cells $e_{kl} = [x_{2k}, x_{2k+2}] \times [y_{2l}, y_{2l+2}]$ for $k = 0, \ldots, N_x - 1$ and $l = 0, \ldots, N_y - 1$ as a finite element mesh for Q^2 basis. Define

$$V^h = \{v \in C^0(\Omega) : v|_{e_{kl}} \in Q^2(e_{kl}), k = 0, \ldots, N_x - 1, l = 0, \ldots, N_y - 1\},$$

$$V_0^h = \{v \in C^0(\Omega) : v|_{e_{kl}} \in Q^2(e_{kl}), k = 0, \ldots, N_x - 1, l = 0, \ldots, N_y - 1; v|_{\partial\Omega} \equiv 0\}.$$

For the coefficients $\mathbf{a}(x,y) = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$, $\mathbf{b} = [b^1 \quad b^2]$ and c in the elliptic operator (2.9), consider their grid point values in the following form:

$$A^{kl} = \begin{pmatrix} a_{00} & a_{01} & \ldots & a_{0,n_x+1} \\ a_{10} & a_{11} & \ldots & a_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ a_{n_y+1,0} & a_{n_y+1,1} & \ldots & a_{n_y+1,,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad a_{ij} = a^{kl}(x_j, y_i), \quad k, l = 1, 2,$$

$$B^{m} = \begin{pmatrix} b_{00} & b_{01} & \ldots & b_{0,n_x+1} \\ b_{10} & b_{11} & \ldots & b_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ b_{n_y+1,0} & b_{n_y+1,1} & \ldots & b_{n_y+1,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad b_{ij} = b^{m}(x_j, y_i), \quad m = 1, 2,$$

$$C = \begin{pmatrix} c_{00} & c_{01} & \ldots & c_{0,n_x+1} \\ c_{10} & c_{11} & \ldots & c_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ c_{n_y+1,0} & c_{n_y+1,1} & \ldots & c_{n_y+1,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad c_{ij} = c(x_j, y_i).$$

Let $diag(\mathbf{x})$ denote a diagonal matrix with the vector $\mathbf{x}$ as diagonal entries and define

$$\bar{W}_x = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n_x+2)\times(n_x+2)},$$

$$\bar{W}_y = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n_y+2)\times(n_y+2)},$$

$$W_x = diag\left(\tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}\right)_{n_x\times n_x}, W_y = diag\left(\tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}\right)_{n_y\times n_y}.$$

Let $s = x$ or y, we define the D and E matrices with dimension $(n_s + 2) \times (n_s + 2)$ for each variable:

$$D_s = \frac{1}{2}\begin{pmatrix}
-3 & 4 & -1 & & & & & & & & \\
-1 & 0 & 1 & & & & & & & & \\
\frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & & & \\
 & & -1 & 0 & 1 & & & & & & \\
 & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & \\
 & & & & -1 & 0 & 1 & & & & \\
 & & & & & & \ddots & \ddots & \ddots & & \\
 & & & & & & -1 & 0 & 1 & & \\
 & & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\
 & & & & & & & & -1 & 0 & 1 \\
 & & & & & & & & 1 & -4 & 3
\end{pmatrix}, \quad E_s = \frac{1}{2}\begin{pmatrix}
0 & 0 & 0 & & & & & & & & \\
0 & 0 & 0 & & & & & & & & \\
-\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & & & \\
 & & 0 & 0 & 0 & & & & & & \\
 & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & \\
 & & & & 0 & 0 & 0 & & & & \\
 & & & & & & \ddots & \ddots & \ddots & & \\
 & & & & & & 0 & 0 & 0 & & \\
 & & & & & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} \\
 & & & & & & & & 0 & 0 & 0 \\
 & & & & & & & & 0 & 0 & 0
\end{pmatrix}.$$

Define an inflation operator $Infl : \mathbb{R}^{n_y\times n_x} \longrightarrow \mathbb{R}^{(n_y+2)\times(n_x+2)}$ by adding zeros:

$$Infl(U) = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & U & \vdots \\ 0 & \cdots & 0 \end{pmatrix}_{(n_y+2)\times(n_x+2)}$$

and its matrix representation is given as $\tilde{I}_x \otimes \tilde{I}_y$ where

$$\tilde{I}_x = \begin{pmatrix} \mathbf{0} \\ I_{n_x\times n_x} \\ \mathbf{0} \end{pmatrix}_{(n_x+2)\times n_x}, \tilde{I}_y = \begin{pmatrix} \mathbf{0} \\ I_{n_y\times n_y} \\ \mathbf{0} \end{pmatrix}_{(n_y+2)\times n_y}.$$

Its adjoint is a restriction operator $Res : \mathbb{R}^{(n_y+2)\times(n_x+2)} \longrightarrow \mathbb{R}^{n_y\times n_x}$ as

$$Res(X) = X(2 : n_y + 1, 2 : n_x + 1) \quad , \forall X \in \mathbb{R}^{(n_y+2)\times(n_x+2)},$$

and its matrix representation is $\tilde{I}_x^T \otimes \tilde{I}_y^T$.

2.9.3 Two-dimensional case

For $\bar{\Omega} = [0,1]^2$ we first consider an elliptic equation with homogeneous Neumann boundary condition:

$$-\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b}\nabla u + cu = f \text{ on } \Omega, \tag{2.82}$$

$$\mathbf{a}\nabla u \cdot \mathbf{n} = 0 \text{ on } \partial\Omega. \tag{2.83}$$

The variational form is to find $u \in H^1(\Omega)$ satisfying

$$A(u,v) = (f,v), \quad \forall v \in H^1(\Omega). \tag{2.84}$$

The C^0-Q^2 finite element method with 3×3 Gauss-Lobatto quadrature is to find $u_h \in V^h$ satisfying

$$\langle \mathbf{a}\nabla u_h, \nabla v_h \rangle_h + \langle \mathbf{b}\nabla u_h, v_h \rangle_h + \langle c u_h, v_h \rangle_h = \langle f, v_h \rangle_h, \quad \forall v_h \in V^h, \tag{2.85}$$

Let $\bar{U}$ be a $(n_y+2)\times(n_x+2)$ matrix such that its (j,i)-th entry is $\bar{U}(j,i) = u_h(x_{i-1}, y_{j-1})$, $i = 1,\ldots,n_x+2$, $j = 1,\ldots,n_y+2$. Let $\bar{F}$ be a $(n_y+2)\times(n_x+2)$ matrix such that its (j,i)-th entry is $\bar{F}(j,i) = f(x_{i-1}, y_{j-1})$. Then the matrix form of (2.85) is

$$\bar{S}vec(\bar{U}) = \bar{M}vec(\bar{F}), \quad \bar{M} = h_x h_y \bar{W}_x \otimes \bar{W}_y, \quad \bar{S} = \sum_{k,l=1}^{2} S_a^{kl} + \sum_{m=1}^{2} S_b^m + S_c, \tag{2.86}$$

where

$$\begin{aligned}
&S_a^{11} = \frac{h_y}{h_x}(D_x^T \otimes I_y) diag(vec(\bar{W}_y A^{11} \bar{W}_x))(D_x \otimes I_y) + \frac{h_y}{h_x}(E_x^T \otimes I_y) diag(vec(\bar{W}_y A^{11} \bar{W}_x))(E_x \otimes I_y),\\
&S_a^{12} = (D_x^T \otimes I_y) diag(vec(\bar{W}_y A^{12} \bar{W}_x))(I_x \otimes D_y) + (E_x^T \otimes I_y) diag(vec(\bar{W}_y A^{12} \bar{W}_x))(I_x \otimes E_y),\\
&S_a^{21} = (I_x \otimes D_y^T) diag(vec(\bar{W}_y A^{21} \bar{W}_x))(D_x \otimes I_y) + (I_x \otimes E_y^T) diag(vec(\bar{W}_y A^{21} \bar{W}_x))(E_x \otimes I_y),\\
&S_a^{22} = \frac{h_x}{h_y}(I_x \otimes D_y^T) diag(vec(\bar{W}_y A^{22} \bar{W}_x))(I_x \otimes D_y) + \frac{h_x}{h_y}(I_x \otimes E_y^T) diag(vec(\bar{W}_y A^{22} \bar{W}_x))(I_x \otimes E_y),\\
&S_b^1 = h_y diag(vec(\bar{W}_y B^1 \bar{W}_x))(D_x \otimes I_y), \quad S_b^2 = h_x diag(vec(\bar{W}_y B^2 \bar{W}_x))(I_x \otimes D_y),\\
&S_c = h_x h_y diag(vec(\bar{W}_y C \bar{W}_x).
\end{aligned}$$

Now consider the scheme (2.53) for nonhomogeneous Dirichlet boundary conditions. Its numerical solution can be represented as a matrix U of size $ny \times nx$ with (j,i)-entry $U(j,i) = u_h(x_i, y_j)$ for $i = 1, \cdots, nx; j = 1, \cdots, ny$. Similar to the one-dimensional case, its stiffness matrix can be obtained as the submatrix of $\bar{S}$ in (2.86). Let $\bar{G}$ be a $(n_y + 2)$ by $(n_x + 2)$ matrix with (j,i)-th entry as $\bar{G}(j,i) = g(x_{i-1}, y_{j-1})$, where

$$g(x,y) = \begin{cases} 0, & \text{if} \quad (x,y) \in (0,1) \times (0,1), \\ g(x,y), & \text{if} \quad (x,y) \in \partial\Omega. \end{cases}$$

In particular, $\bar{G}(j+1, i+1) = 0$ for $j = 1, \ldots, n_y$, $i = 1, \ldots, n_x$. Let F be a matrix of size $ny \times nx$ with (j,i)-entry as $F(j,i) = f(x_i, y_j)$ for $i = 1, \cdots, nx; j = 1, \cdots, ny$. Then the scheme (2.53) becomes

$$(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y) vec(U) = (W_x \otimes W_y) vec(F) - (\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S} vec(\bar{G}). \tag{2.87}$$

Even though the stiffness matrix is given as $S = (\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y)$, S should be implemented as a linear operator in iterative linear system solvers. For example, the matrix vector multiplication $(\tilde{I}_x^T \otimes \tilde{I}_y^T) S_a^{11} (\tilde{I}_x \otimes \tilde{I}_y) vec(U)$ is equivalent to the following linear operator from $\mathbb{R}^{ny \times nx}$ to $\mathbb{R}^{ny \times nx}$:

$$\frac{h_y}{h_x} \tilde{I}_y^T \left\{ I_y \left([\bar{W}_y A^{11} \bar{W}_x] \circ [I_y(\tilde{I}_y U \tilde{I}_x^T) D_x^T] \right) D_x + I_y \left([\bar{W}_y A^{11} \bar{W}_x] \circ [I_y(\tilde{I}_y U \tilde{I}_x^T) E_x^T] \right) E_x \right\} \tilde{I}_x,$$

where $\circ$ is the Hadamard product (i.e., entrywise multiplication).

2.9.4 The Laplacian case

For one-dimensional constant coefficient case with homogeneous Dirichlet boundary condition, the scheme can be written as a classical finite difference scheme $H\mathbf{u} = \mathbf{f}$ with

$$H = M^{-1} S = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & & & \\ -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & \\ & -1 & 2 & -1 & & & \\ & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & \\ & & & -1 & 2 & -1 & \\ & & & & \ddots & \ddots & \\ & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 \\ & & & & & -1 & 2 \end{pmatrix}$$

In other words, if x_i is a cell center, the scheme is

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i,$$

and if x_i is a knot away from the boundary, the scheme is

$$\frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i.$$

It is straightforward to verify that the local truncation error is only second order.

For the two-dimensional Laplacian case homogeneous Dirichlet boundary condition, the scheme can be rewritten as

$$(H_x \otimes I_y) + (I_x \otimes H_y) vec(U) = vec(F),$$

where H_x and H_y are the same H matrix above with size $n_x \times n_x$ and $n_y \times n_y$ respectively. The inverse of $(H_x \otimes I_y) + (I_x \otimes H_y)$ can be efficiently constructed via the eigen-decomposition of small matrices H_x and H_y:

1. Compute eigen-decomposition of $H_x = T_x \Lambda_x T_x^{-1}$ and $H_y = T_y \Lambda_y T_y^{-1}$.
2. The properties of Kronecker product imply that

   $$(H_x \otimes I_y) + (I_x \otimes H_y) = (T_x \otimes T_y)(\Lambda_x \otimes I_y + I_x \otimes \Lambda_y)(T_x^{-1} \otimes T_y^{-1}),$$

   thus

   $$[(H_x \otimes I_y) + (I_x \otimes H_y)]^{-1} = (T_x \otimes T_y)(\Lambda_x \otimes I_y + I_x \otimes \Lambda_y)^{-1}(T_x^{-1} \otimes T_y^{-1}).$$

3. It is nontrivial to determine whether H is diagonalizable. In all our numerical tests, H has no repeated eigenvalues. So if assuming Λ_x and Λ_y are diagonal matrices, the matrix vector multiplication $[(H_x \otimes I_y) + (I_x \otimes H_y)]^{-1} vec(F)$ can be implemented as a linear operator on F:

   $$T_y([T_y^{-1} F (T_x^{-1})^T]./\Lambda) T_x^T, \tag{2.88}$$

   where Λ is a $n_y \times n_x$ matrix with (i,j)-th entry as $\Lambda(i,j) = \Lambda_y(i,i) + \Lambda_x(j,j)$ and ./ denotes entry-wise division for two matrices of the same size.

For the 3D Laplacian, the matrix can be represented as $H_x \otimes I_y \otimes I_z + I_x \otimes H_y \otimes I_z + I_x \otimes I_y \otimes H_z$ thus can be efficiently inverted through eigen-decomposition of small matrices H_x, H_y and H_z as well.

Since the eigen-decomposition of small matrices H_x and H_y can be precomputed, and (2.88) costs only $\mathcal{O}(n^3)$ for a 2D problem on a mesh size $n \times n$, in practice (2.88) can be used as a simple preconditioner in conjugate gradient solvers for the following linear system equivalent to (2.87):

$$(W_x^{-1} \otimes W_y^{-1})(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y) vec(U) = vec(F) - (W_x^{-1} \otimes W_y^{-1})(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S} vec(G),$$

even though the multigrid method as reviewed in [42] is the optimal solver in terms of computational complexity.

2.10 Numerical Results

In this section we show a few numerical tests verifying the accuracy of the scheme (2.53) for $k = 2$ implemented as a finite difference scheme on a uniform grid. We first consider the following two dimensional elliptic equation:

$$-\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b} \cdot \nabla u + cu = f \quad \text{on } [0,1] \times [0,2] \tag{2.89}$$

where $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = 10 + 30y^5 + x\cos y + y$, $a_{12} = a_{21} = 2 + 0.5(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$, $a_{22} = 10 + x^5$, $\mathbf{b} = \mathbf{0}$, $c = 1 + x^4 y^3$, with an exact solution

$$u(x, y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3).$$

Table 2.1. A 2D elliptic equation with Dirichlet boundary conditions. The first column is the number of regular cells in a finite element mesh. The second column is the number of grid points in a finite difference implementation, i.e., number of degree of freedoms.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	3×7	3.94E-2	-	7.15E-2	-
4×8	7×15	1.23E-2	1.67	3.28E-2	1.12
8×16	15×31	1.46E-3	3.08	5.42E-3	2.60
16×32	31×63	1.14E-4	3.68	3.96E-4	3.78
32×64	63×127	7.75E-6	3.88	2.62E-5	3.92
64×128	127×255	5.02E-7	3.95	1.73E-6	3.92
128×256	255×511	3.23E-8	3.96	1.13E-7	3.94

The errors at grid points are listed in Table 2.1 for purely Dirichlet boundary condition and Table 2.2 for purely Neumann boundary condition. We observe fourth order accuracy in the discrete 2-norm for both tests, even though only $\mathcal{O}(h^{3.5})$ can be proven for Neumann boundary condition as discussed in Remark 2.6.5. Regarding the maximum norm of

Table 2.2. A 2D elliptic equation with Neumann boundary conditions.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	5×9	1.38E0	-	2.27E0	-
4×8	9×17	1.46E-1	3.24	2.52E-1	3.17
8×16	17×33	7.49E-3	4.28	1.64E-2	3.94
16×32	33×65	4.31E-4	4.12	1.02E-3	4.01
32×64	65×129	2.61E-5	4.04	7.47E-5	3.78

the superconvergence of the function values at Gauss-Lobatto points, one can only prove $\mathcal{O}(h^3 \log h)$ even for the full finite element scheme (2.1) since discrete Green's function is used, see [9].

Next we consider a three-dimensional problem $-\Delta u = f$ with homogeneous Dirichlet boundary conditions on a cube $[0,1]^3$ with the following exact solution

$$u(x,y,z) = \sin(\pi x)\sin(2\pi y)\sin(3\pi z) + (x - x^3)(y^2 - y^4)(z - z^2).$$

See Table 2.3 for the performance of the finite difference scheme. There is no essential difficulty to extend the proof to three dimensions, even though it is not very straightforward. Nonetheless we observe that the scheme is indeed fourth order accurate. The linear system is solved by the eigenvector method shown in Section 2.9.4. The discrete 2-norm over the set of all grid points Z_0 is defined as $\|u\|_{l^2(\Omega)} = \left[h^3 \sum_{(x,y,z)\in Z_0} |u(x,y,z)|^2\right]^{\frac{1}{2}}$.

Table 2.3. $-\Delta u = f$ in 3D with homogeneous Dirichlet boundary condition.

Finite Difference Grid	l^2 error	order	l^∞ error	order
$7 \times 7 \times 7$	1.51E-2	-	4.87E-2	-
$15 \times 15 \times 15$	9.23E-4	4.04	3.12E-3	3.96
$31 \times 31 \times 31$	5.68E-5	4.02	1.95E-4	4.00
$63 \times 63 \times 63$	3.54E-6	4.01	1.22E-5	4.00
$127 \times 127 \times 127$	2.21E-7	4.00	7.59E-7	4.00

Last we consider (2.89) with convection term and the coefficients **b** is incompressible $\nabla \cdot \mathbf{b} = 0$: $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = 100 + 30y^5 + x\cos y + y$, $a_{12} = a_{21} = 2 + 0.5(\sin(\pi x) +$

$x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$, $a_{22} = 100 + x^5$, $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, $b_1 = \psi_y$, $b_2 = -\psi_x$, $\psi = x \exp(x^2 + y)$, $c = 1 + x^4 y^3$, with an exact solution

$$u(x, y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3).$$

The errors at grid points are listed in Table 2.4 for Dirichlet boundary conditions.

Table 2.4. A 2D elliptic equation with convection term and Dirichlet boundary conditions.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	3×7	1.26E-1	-	2.71E-1	-
4×8	7×15	2.85E-2	2.15	9.70E-2	1.48
8×16	15×31	1.89E-3	3.92	7.25E-3	3.74
16×32	31×63	1.17E-4	4.01	4.01E-4	4.17
32×64	63×127	7.41E-6	3.98	2.54E-5	3.98

2.11 Concluding Remarks

In this chapter we have proven the superconvergence of function values in the simplest finite difference implementation of Q^k spectral element method for elliptic equations. In particular, for the case $k = 2$ the scheme (2.53) can be easily implemented as a fourth order accurate finite difference scheme as shown in Section 2.9. It provides only only an convenient approach for constructing fourth order accurate finite difference schemes but also the most efficient implementation of C^0-Q^k finite element method without losing superconvergence of function values. In the last section, we will show that discrete maximum principle can be proven for the scheme (2.53) in the case $k = 2$ when solving a variable coefficient Poisson equation.

3. ACCURACY OF SPECTRAL ELEMENT METHOD FOR WAVE, PARABOLIC AND SCHRÖDINGER EQUATIONS

3.1 Introduction

In this Chapter, we are actually extending the results in Chapter 2 to wave, parabolic and Schrödinger equations.

In recent years many such stable and high order accurate methods for wave equations have been developed. These include discontinuous Galerkin methods for first order hyperbolic systems [43]–[49] and wave equations in second order form [50]–[53], and finite differences with summation by parts operators [54]–[60], as well as spectral elements for wave equations [12], [13].

In this chapeter we study the rates of convergence of the error, as measured in norms over nodes for all degree of freedoms, for the spectral element method applied to linear wave and parabolic, and Schrödinger equations.

To be precise, we consider the Lagrangian Q^k $(k \geq 2)$ continuous finite element method for solving linear evolution PDEs with a second order operator $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u)$ on rectangular meshes implemented by $(k+1)$-point Gauss-Lobatto quadrature for all integrals. This is often referred to as the spectral element method in the literature and this is the notation we will use here.

For the Q^k spectral element method, it is well known that the standard finite element error estimates still hold [11], i.e., the error in H^1-norm is k-th order and the error in L^2-norm is $(k+1)$-th order. It is also well known that the Lagrangian Q^k $(k \geq 2)$ continuous finite element method is $(k+2)$-th order accurate in the discrete 2-norm over all $(k+1)$-point Gauss-Lobatto quadrature points [8]–[10]. If using a very accurate quadrature in the finite element method for a variable coefficient operator $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u)$, then $(k+2)$-th order superconvergence at Gauss-Lobatto points holds trivially. However, for the efficiency of having a diagonal mass matrix and for the convenience of implementation, the most popular method for wave equations is the simplest choice of quadrature, i.e. using $(k+1)$-point Gauss-Lobatto quadrature for Q^k elements in all integrals for both mass and stiffness matrices. In particular in the seismic community, where highly efficient simulation of the elastic wave equation is of important, the spectral method has become the method of choice, [12], [13].

When using this $(k+1)$-point Gauss-Lobatto quadrature for Lagrangian Q^k finite element method, the quadrature nodes coincide with the nodes defining the degrees of freedom, and the resulting method becomes the so-called spectral element method. Thus the spectral element method can also be regarded as a finite difference scheme at all Gauss-Lobatto points. For instance, consider solving $u_{tt} = u_{xx}$ on the interval $[0, 1]$ with homogeneous Dirichlet boundary conditions. Introduce the uniform grid $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ with spacing $h = 1/(N + 1)$ and N being odd. This grid gives a uniform partition of the interval $[0, 1]$ into uniform intervals $I_k = [x_{2k}, x_{2k+2}]$ $(k = 0, \cdots, \frac{N-1}{2})$. Then all 3-point Gauss-Lobatto quadrature points for intervals $I_k = [x_{2k}, x_{2k+2}]$ coincide with the grid points x_i. The Q^2 spectral element method on intervals $I_k = [x_{2k}, x_{2k+2}]$ $(k = 0, \cdots, \frac{N-1}{2})$ is equivalent to the following semi-discrete finite difference scheme [61], [62]:

$$\frac{d^2}{dt^2}u_i = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}, \quad \text{if } i \text{ is odd;} \tag{3.1a}$$

$$\frac{d^2}{dt^2}u_i = \frac{-u_{i-2} + 8u_{i-1} - 14u_i + 8u_{i+1} - u_{i+2}}{4h^2}, \quad \text{if } i \text{ is even.} \tag{3.1b}$$

The truncation error of (3.1) is only second order yet the dispersion error is fourth order, see Section 11 in [61]. Although the dispersion error results can in principle be extended to any order, the derivation and expressions become increasingly cumbersome. Further the dispersion error results are limited to unbounded or periodic domains and do not produce error estimates in the form of a norm of the error. In fact, as we have shown in [62], it is nontrivial and requires new analysis tools to establish the $(k + 2)$-th order superconvergence when $(k + 1)$-point Gauss-Lobatto quadrature is used. In [62], $(k + 2)$-th order accuracy at all Gauss-Lobatto points of Q^k spectral element method was proven for elliptic equations with Dirichlet boundary conditions. In this chapter, we extend those results and will prove that the Q^k spectral element method is a $(k + 2)$-th order accurate scheme for linear wave, parabolic and Schrödinger equations with Dirichlet boundary conditions. For Neumann boundary conditions, if $\mathbf{a}(\mathbf{x})$ is diagonal, i.e., there are no mixed second order derivatives in $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u)$, $(k+2)$-th order accuracy in discrete 2-norm can be proven. When mixed second order derivatives are involved, only $(k+\frac{3}{2})$-th order can be proven for Neumann boundary conditions, and we indeed observe some order loss in numerical tests.

This chapter explains the order of accuracy of Q^k spectral element method, when the errors are measured only at nodes of degree of freedoms. As mentioned above we consider the case of rectangular elements and a smooth coefficient $\mathbf{a}(\mathbf{x})$ in the term $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u)$. We note that this does include discretizations on regular meshes of curvilinear domains that can be smoothly mapped to rectangular meshes for the unit cube, e.g., the spectral element method for Δu on such a mesh for a curvilinear domain is equivalent to the spectral element method for $\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u$ on a reference uniform rectangular mesh where $\mathbf{a}(\mathbf{x})$ and $\mathbf{b}(\mathbf{x})$ emerge from the mapping between the curvilinear domain and the unit cube. It does however not include problems on unstructured quadrilateral meshes where the metric terms typically are non-smooth at element interfaces but we note that the numerical examples that we present indicate that such meshes may still exhibit larger rates than $k+1$.

This chapter is organized as follows. In Section 3.2, we introduce notation and assumptions. In Section 3.3, we review a few standard quadrature estimates. In Section 3.4, the superconvergence of elliptic projection is analyzed, which is parallel to the classic error estimation for hyperbolic and parabolic equations by involving elliptic projection of the corresponding elliptic operator, see [63]–[65]. We then prove the main result for homogeneous Dirichlet boundary conditions in Section 3.5, for the second-order wave equation in Section 3.5.1, parabolic equations in Section 3.5.2 and linear Schrödinger equation in Section 3.5.3. Neumann boundary conditions can be discussed similarly as summarized in Section 3.5.4. For problems with nonhomogeneous Dirichlet boundary conditions, a convenient implementation which maintains the $(k+2)$-th order of accuracy is given in Section 3.6. Numerical tests verifying the estimates are given in Section 3.7. Concluding remarks are given in Section 3.8

3.2 Equations, Notation, And Assumptions

3.2.1 Problem Setup

Let L be a linear second order differential operator with time dependent coefficients:

$$Lu = -\nabla \cdot (\mathbf{a}(\mathbf{x}, t)\nabla u) + \mathbf{b}(\mathbf{x}, t) \cdot \nabla u + c(\mathbf{x}, t)u,$$

where $\mathbf{a}(\mathbf{x}, t) = (a_{ij}(\mathbf{x}, t))$ is a positive symmetric definite operator for $t \in [0, T]$, i.e. there exists a constant $\alpha, \beta > 0$ such that $\alpha|\xi|^2 \leq \xi^T \mathbf{a}(\mathbf{x}, t)\xi \leq \beta|\xi|^2$, for all $(\mathbf{x}, t) \in \Omega \times [0, T], \xi \in \mathbb{R}^n$. Consider the following two initial-boundary value problems with smooth enough coefficients on a rectangular domain $\Omega = (0, 1) \times (0, 1)$ with its boundary $\partial\Omega$:

Given $0 < T < \infty$, find $u(\mathbf{x}, t)$ on $\bar{\Omega} \times [0, T]$ satisfying

$$\begin{aligned} u_t =& - Lu + f(\mathbf{x}, t) && \text{in } \Omega \times (0, T], \\ u(\mathbf{x}, t) =& 0 && \text{on } \partial\Omega \times [0, T], \\ u(\mathbf{x}, 0) =& u_0(\mathbf{x}) && \text{on } \Omega. \end{aligned} \tag{3.2}$$

Given $0 < T < \infty$, find $u(\mathbf{x}, t)$ on $\bar{\Omega} \times [0, T]$ satisfying

$$\begin{aligned} u_{tt} =& - Lu + f(\mathbf{x}, t) && \text{in } \Omega \times (0, T], \\ u(\mathbf{x}, t) =& 0 && \text{on } \partial\Omega \times [0, T], \\ u(\mathbf{x}, 0) =& u_0(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = u_1(\mathbf{x}) && \text{on } \Omega \times \{t = 0\}. \end{aligned} \tag{3.3}$$

We use $A(\cdot)$ to denote the bilinear form: for $u, v \in H^1(\Omega)$,

$$A(u, v) = \int_\Omega \nabla u^T \mathbf{a}(\mathbf{x}, t)\nabla v + \mathbf{b}(\mathbf{x}, t) \cdot \nabla u + c(\mathbf{x}, t)uv d\mathbf{x}. \tag{3.4}$$

For convenience, we assume Ω_h is an uniform rectangular mesh for $\bar{\Omega}$ and $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$ denotes any cell in Ω_h with cell center (x_e, y_e). Though we only discuss uniform meshes, the main result can be easily extended to nonuniform rectangular meshes with smoothly varying cells. Let $Q^k(e) = \left\{ p(x, y) = \sum\limits_{i=0}^{k} \sum\limits_{j=0}^{k} p_{ij} x^i y^j, (x, y) \in e \right\}$, denote the set of tensor product of polynomials of degree k on an element e. Then we use $V^h = \{p(x,y) \in C^0(\Omega_h) : p|_e \in Q^k(e), \quad \forall e \in \Omega_h\}$ to denote the continuous piecewise Q^k finite element space on Ω_h and $V_0^h = \{v_h \in V^h : v_h|_{\partial\Omega} = 0\}$. Further let $(u,v) = \int_\Omega uv d\mathbf{x}$ and let $\langle \cdot, \cdot \rangle_h$ and $A_h(\cdot,\cdot)$ denote approximation of the integrals by $(k+1)$-point Gauss-Lobatto quadrature for each spatial variable in each cell. Also, $u^{(i)}$ will denote the i-th time derivative of the function $u(\mathbf{x},t)$.

For the equations that we are interested in, assume the exact solution $u(\mathbf{x},t) \in H_0^1(\Omega) \cap H^2(\Omega)$ for any t, and define its discrete elliptic projection $R_h u \in V_0^h$ as

$$A_h(R_h u, v_h) = \langle -Lu, v_h \rangle_h, \quad \forall v_h \in V_0^h, \quad 0 \le t \le T. \tag{3.5}$$

Also, let $u_I \in V^h$ denote the piecewise Lagrangian Q^k interpolation polynomial of function u at $(k+1) \times (k+1)$ Gauss-Lobatto points in each rectangular cell.

We consider semi-discrete spectral element schemes whose initial conditions are defined by the elliptic projection and the Lagrange interpolant of the continuous initial data.

For problem (3.2) the scheme is to find $u_h(\mathbf{x},t) \in V_0^h$ satisfying

$$\begin{aligned} \langle u_h^{(1)}, v_h \rangle_h + A_h(u_h, v_h) =& \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h, \\ u_h(0) =& R_h u_0. \end{aligned} \tag{3.6}$$

We consider the semi-discrete spectral element scheme for problem (3.3) with special initial conditions: solve for $u_h(t) \in V_0^h$ satisfying

$$\begin{aligned} \langle u_h^{(2)}, v_h \rangle_h + A_h(u_h, v_h) =& \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h, \\ u_h(0) = R_h u_0, \quad u_h^{(1)}(0) =& (u_1)_I. \end{aligned} \tag{3.7}$$

3.2.2 Notation and basic tools

We will use the same notation as in Chapter 2, we may also need:

- Let superscript (i) denote i-th time derivatives for coefficients $\mathbf{a}, \mathbf{b}$, and c. For the time dependent operators L and A, the symbols $L^{(i)}$ and $A^{(i)}$ are defined as taking time derivatives only for coefficients:

$$L^{(i)}u = -\nabla \cdot (\mathbf{a}^{(i)}\nabla u) + \mathbf{b}^{(i)} \cdot \nabla u + c^{(i)}u,$$

and

$$A^{(i)}(u,v) = \int_\Omega \nabla u^T \mathbf{a}^{(i)} \nabla v + \mathbf{b}^{(i)} \cdot \nabla u + c^{(i)}uvd\mathbf{x}.$$

The symbol $A_h^{(i)}$ is similarly defined as taking time derivatives only for coefficients in A_h. With this notation, for $u(\mathbf{x},t)$ and time independent test function $v(\mathbf{x})$, we have Leibniz rule

$$(Lu)^{(m)} = \sum_{j=0}^{m} \binom{m}{j} L^{(m-j)}u^{(j)}, \quad [A(u,v)]^{(m)} = \sum_{j=0}^{m} \binom{m}{j} A^{(m-j)}(u^{(j)},v).$$

- By integration by parts, it is straightforward to verify

$$(L^{(m-j)}u^{(j)},v) = A^{(m-j)}(u^{(j)},v), \quad \forall v \in H_0^1(\Omega). \tag{3.8}$$

3.2.3 Assumption on the coercivity and the elliptic regularity

For the operator $A(u,v) := \int_\Omega [\nabla u^T \mathbf{a} \nabla v + (\mathbf{b}\cdot\nabla u)v + cuv]\, d\mathbf{x}$ where $\mathbf{a} = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$ is positive definite and $\mathbf{b} = (b^1 \quad b^2)$, assume the coefficients a_{ij}, b_j, $c \in C^{m_1}([0,T]; W^{m_2,\infty}(\Omega))$ for m_1, m_2 large enough. Thus for $t \in [0,T]$, $A(u,v) \le C\|u\|_1\|v\|_1$ for any $u, v \in H_0^1(\Omega)$. As discussed in Chapter 2, if we assume $\lambda_{\mathbf{a}}$ has a positive lower bound and $\nabla \cdot \mathbf{b} \le 2c$, where $\lambda_{\mathbf{a}}$ as the smallest eigenvalues of $\mathbf{a}$, the coercivity of the bilinear form can be easily achieved. For the V^h-ellipticity, as pointed out in Lemma 2.6.2 of Chapter 2, if $4\lambda_{\mathbf{a}}c > |\mathbf{b}|^2$, for $t \in [0,T]$,

$$C\|v_h\|_1^2 \le A_h(v_h,v_h), \quad \forall v_h \in V^h, \tag{3.9}$$

can be proven. In the rest of this chapter, we assume coercivity for the bilinear forms A, A^*, and A_h. We assume the elliptic regularity $\|w\|_2 \leq C\|f\|_0$ holds for the exact dual problem of finding $w \in H_0^1(\Omega)$ satisfying $A^*(w, v) = (f, v), \quad \forall v \in H_0^1(\Omega)$. See [33], [34] for the elliptic regularity with Lipschitz continuous coefficients on a Lipschitz domain.

We remark that in the case of the wave equation we also assume finite speed of propagation i.e. that there is an upper bound on the eigenvalues of $\mathbf{a}$.

3.3 Quadrature Error Estimates

For any continuous function $u(\mathbf{x}, t_0)$ with fixed time t_0, its M-type projection (defined in Section 2.4 of Chapter 2) on spatial variables is a continuous piecewise Q^k polynomial of $\mathbf{x}$, denoted as $u_p(\mathbf{x}, t_0) \in V^h$. The M-type projection was used to analyze superconvergence [9]. For $m \geq 0$, $(u_p)^{(m)} = \left(u^{(m)}\right)_p$, thus there is no ambiguity to use the notation $u_p^{(m)}$. The M-type projection has the following properties.

For convenience, we write (2.29a) and (2.29b) and Theorem 2.3.9 as the follows respectively.

Lemma 3.3.1. *For $i, j \geq 0$ and any fixed $t \in [0, T]$, assuming sufficiently smooth coefficients $\mathbf{a}, \mathbf{b}, c$ and function $u(\mathbf{x}, t) \in H^{k+3}(\Omega)$, we have*

$$A_h^{(i)}((u - u_p)^{(j)}, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u^{(j)}(t)\|_{k+3}\|v_h\|_2, & \text{if } v_h \in V_0^h \text{ or } \mathbf{a} \text{ is diagonal;} \\ \mathcal{O}(h^{k+\frac{3}{2}})\|u^{(j)}(t)\|_{k+3}\|v_h\|_2, & \text{otherwise.} \end{cases} \tag{3.10}$$

Lemma 3.3.2. *For the differential operator L and any fixed $t \in [0, T]$, assume $a_{ij}(\mathbf{x}, t)$, $b_i(\mathbf{x}, t)$, $c(\mathbf{x}, t) \in L^\infty\left([0, T]; W^{k+2,\infty}(\Omega)\right)$ and $u(\mathbf{x}, t) \in H^{k+3}(\Omega)$. For $k \geq 2$, we have*

$$A(u, v_h) - A_h(u, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u(t)\|_{k+3}\|v_h\|_2, & \text{if } v_h \in V_0^h \text{ or } \mathbf{a} \text{ is diagonal;} \\ \mathcal{O}(h^{k+\frac{3}{2}})\|u(t)\|_{k+3}\|v_h\|_2, & \text{otherwise.} \end{cases} \tag{3.11}$$

Remark 3.3.3. There is half order loss in (3.10) and (3.11), only when using $v \in V^h$ for non-diagonal $\mathbf{a}$, i.e., when solving second order equations containing mixed second order derivatives with Neumann boundary conditions.

We have the Gronwall's inequality in integral form as follows:

Lemma 3.3.4. *Let $\xi(t)$ be continuous on $[0, T]$ and*

$$\xi(t) \leq C_1 \int_0^t \xi(s) ds + \alpha(t)$$

for constant $C_1 \geq 0$ and $\alpha(t) \geq 0$ nondescreasing in t. Then $\xi(t) \leq \alpha(t)e^{C_1 t}$ thus $\xi(t) \leq \alpha(t)e^{C_1 T} = C\alpha(t)$ for all $0 \leq t \leq T$.

3.4 Error Estimates For The Elliptic Projection

Let $u_h(\mathbf{x}, t)$ denote the solution of the semi-discrete numerical scheme. Let $e(\mathbf{x}, t) = u_h(\mathbf{x}, t) - u_p(\mathbf{x}, t)$, then we can write

$$e = \theta_h + \rho_h,$$

where $\theta_h := u_h - R_h u \in V_0^h$ and $\rho_h := R_h u - u_p \in V_0^h$.

We have the following superconvergence result for $\|\rho_h^{(m)}(t)\|$, $m \geq 0$, $t \in [0, T]$.

Lemma 3.4.1. *If a_{ij}, b_j, $c \in C^m\left([0,T]; W^{k+2,\infty}(\Omega)\right)$, $u \in C^m\left([0,T]; H^{k+4}(\Omega)\right)$, then we have*

$$\|\rho_h^{(m)}(t)\|_1 \leq C h^{k+1} \sum_{j=0}^{m} (\|u^{(j)}(t)\|_{k+3} + \|(Lu)^{(j)}(t)\|_{k+2}), \tag{3.12}$$

$$\|\rho_h^{(m)}\|_{L^2([0,T];L^2(\Omega))} \leq C h^{k+2} \sum_{j=0}^{m} (\|u^{(j)}\|_{L^2([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^2([0,T];H^{k+2}(\Omega))}), \tag{3.13}$$

$$\|\rho_h^{(m)}\|_{L^\infty([0,T];L^2(\Omega))} \leq C h^{k+2} \sum_{j=0}^{m} (\|u^{(j)}\|_{L^\infty([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^\infty([0,T];H^{k+2}(\Omega))}), \tag{3.14}$$

where C is independent of h, u, f, and time t.

Proof. First we prove (3.12), with which we then prove (3.13) and (3.14) by the dual argument.

From the definition of the discrete elliptic projection (3.5) we have

$$A_h(\rho_h, v_h) = \epsilon(v_h), \quad \forall v_h \in V_0^h. \tag{3.15}$$

where

$$\epsilon(v_h) = \langle Lu, v_h \rangle_h - A_h(u_p, v_h).$$

Note that v_h is time independent. Taking m time derivatives of (3.15) yields

$$(A_h(\rho_h, v_h))^{(m)} = \sum_{j=0}^{m} \binom{m}{j} A_h^{(m-j)}(\rho_h^{(j)}, v_h) = \epsilon^{(m)}(v_h). \tag{3.16}$$

The term $\epsilon^{(m)}(v_h)$ can be rewritten as follows:

$$\begin{aligned}
\epsilon^{(m)}(v_h) &= \langle (Lu)^{(m)}, v_h \rangle_h - (A_h(u_p, v_h))^{(m)} \\
&= \left[((Lu)^{(m)}, v_h) - (A(u, v_h))^{(m)}\right] - \left[((Lu)^{(m)}, v_h) - \langle (Lu)^{(m)}, v_h \rangle_h\right] \\
&\quad + \left[(A(u, v_h))^{(m)} - (A_h(u, v_h))^{(m)}\right] + (A_h(u - u_p, v_h))^{(m)}.
\end{aligned}$$

By Leibniz rule and (3.8), we have

$$((Lu)^{(m)}, v_h) - (A(u, v_h))^{(m)} = \sum_{j=0}^{m} \binom{m}{j} \left[(L^{(m-j)} u^{(j)}, v_h) - A^{(m-j)}(u^{(j)}, v_h)\right] = 0.$$

By Lemma 2.3.4,

$$((Lu)^{(m)}, v_h) - \langle (Lu)^{(m)}, v_h \rangle_h = \mathcal{O}(h^{k+2}) \|(Lu)^{(m)}(t)\|_{k+2} \|v_h\|_2.$$

By Leibniz rule and Lemma 3.3.2,

$$\begin{aligned}(A(u,v_h))^{(m)}-(A_h(u,v_h))^{(m)} &= \sum_{j=0}^{m}\binom{m}{j}\left[A^{(m-j)}(u^{(j)},v_h)-A_h^{(m-j)}(u^{(j)},v_h)\right]\\ &= \mathcal{O}(h^{k+2})\sum_{j=0}^{m}\binom{m}{j}\|u^{(j)}(t)\|_{k+3}\|v_h\|_2.\end{aligned}$$

Now, Lemma 3.3.1 implies

$$(A_h(u-u_p,v_h))^{(m)}=\sum_{j=0}^{m}\binom{m}{j}A_h^{(m-j)}\left((u-u_p)^{(j)},v_h\right)=\mathcal{O}(h^{k+2})\sum_{j=0}^{m}\binom{m}{j}\|u^{(j)}(t)\|_{k+3}\|v_h\|_2.$$

Thus we have

$$\epsilon^{(m)}(v_h)=\mathcal{O}(h^{k+2})\left(\sum_{j=0}^{m}\|u^{(j)}(t)\|_{k+3}+\|(Lu)^{(m)}(t)\|_{k+2}\right)\|v_h\|_2. \tag{3.17}$$

For $i\geq 0$, by the V_h-ellipticity (3.9), (3.16), and (3.17) we have

$$\begin{aligned}&C\|\rho_h^{(i)}(t)\|_1^2\leq A_h(\rho_h^{(i)},\rho_h^{(i)})\\ =&\sum_{j=0}^{i}\binom{i}{j}A_h^{(i-j)}(\rho_h^{(j)},\rho_h^{(i)})-\sum_{j=0}^{i-1}\binom{i}{j}A_h^{(i-j)}(\rho_h^{(j)},\rho_h^{(i)})\\ =&\epsilon^{(i)}(\rho_h^{(i)})-\sum_{j=0}^{i-1}\binom{i}{j}A_h^{(i-j)}(\rho_h^{(j)},\rho_h^{(i)})\\ \leq&\mathcal{O}(h^{k+1})\left(\sum_{j=0}^{i}\|u^{(j)}\|_{k+3}+\|(Lu)^{(i)}\|_{k+2}\right)h\|\rho_h^{(i)}\|_2+C\sum_{j=0}^{i-1}\|\rho_h^{(j)}(t)\|_1\|\rho_h^{(i)}(t)\|_1\\ \leq&\left[\mathcal{O}(h^{k+1})\left(\sum_{j=0}^{i}\|u^{(j)}\|_{k+3}+\|(Lu)^{(i)}\|_{k+2}\right)+C\sum_{j=0}^{i-1}\|\rho_h^{(j)}(t)\|_1\right]\|\rho_h^{(i)}(t)\|_1,\end{aligned}$$

the last inequality follows from an application of an inverse estimate. Thus

$$\|\rho_h^{(i)}(t)\|_1\leq\mathcal{O}(h^{k+1})\left(\sum_{j=0}^{i}\|u^{(j)}\|_{k+3}+\|(Lu)^{(i)}\|_{k+2}\right)+C\sum_{j=0}^{i-1}\|\rho_h^{(j)}(t)\|_1. \tag{3.18}$$

Now (3.12) can be proven by induction as follows. First, set $i = 0$ in (3.18) to obtain (3.12) with $m = 0$. Second, assume (3.18) holds for $m = i - 1$, then (3.18) implies that (3.12) also holds for $m = i$.

For fixed $t \in [0, T]$, to estimate $\rho_h^{(m)}$ in L^2-norm, we consider the dual problem: find $\phi_h \in V_0^h$ satisfying: for $i \geq 0$,

$$A^*(\phi_h, v_h) = (\rho_h^{(i)}(t), v_h), \quad \forall v_h \in V_0^h. \tag{3.19}$$

Based on Theorem 5.3 in [62], by assuming the elliptic regularity and V^h ellipticity, problem (3.19) has a unique solution satisfying

$$\|\phi_h\|_2 \leq C\|\rho_h^{(i)}(t)\|_0. \tag{3.20}$$

Take $v_h = \rho_h^{(i)}$ in (3.19) then we have

$$\begin{aligned}
&\|\rho_h^{(i)}(t)\|_0^2 \\
=&A^*(\phi_h, \rho_h^{(i)}) = A(\rho_h^{(i)}, \phi_h) \\
=&\sum_{j=0}^{i} \binom{i}{j} A^{(i-j)}(\rho_h^{(j)}, \phi_h) - \sum_{j=0}^{i-1} \binom{i}{j} A^{(i-j)}(\rho_h^{(j)}, \phi_h) \\
=&\sum_{j=0}^{i} \binom{i}{j} \left(A_h^{(i-j)}(\rho_h^{(j)}, \phi_h) + E\left(A^{(i-j)}(\rho_h^{(j)}, \phi_h)\right)\right) - \sum_{j=0}^{i-1} \binom{i}{j} \left(\rho_h^{(j)}, (L^*)^{(i-j)}\phi_h\right).
\end{aligned}$$

Note that $\forall \chi \in V_0^h$, with (3.16) and (3.17),

$$\begin{aligned}
&\sum_{j=0}^{i} \binom{i}{j} A_h^{(i-j)}(\rho_h^{(j)}, \phi_h) \\
=&\sum_{j=0}^{i} \binom{i}{j} A_h^{(i-j)}(\rho_h^{(j)}, \phi_h - \chi) + \sum_{j=0}^{i} \binom{i}{j} A_h^{(i-j)}(\rho_h^{(j)}, \chi) \\
=&\sum_{j=0}^{i} \binom{i}{j} A_h^{(i-j)}(\rho_h^{(j)}, \phi_h - \chi) + \epsilon^{(i)}(\chi) \\
\leq& C\sum_{j=0}^{i} \|\rho_h^{(j)}(t)\|_1 \|\phi_h - \chi\|_1 + \mathcal{O}(h^{k+2}) \left(\sum_{j=0}^{i} \|u^{(j)}(t)\|_{k+3} + \|(Lu)^{(i)}(t)\|_{k+2}\right) \|\chi\|_2.
\end{aligned} \tag{3.21}$$

Let $\chi = \Pi_1 \phi_h$ where Π_1 is the L^2 projection to functions in the continuous piecewise Q^1 polynomial space, see [62]. Then we have $\|\phi_h - \chi\|_1 \leq Ch\|\phi_h\|_2$ and $\|\chi\|_2 \leq C\|\phi_h\|_2$. Inserting (3.12) and (3.20) into (3.21), we have

$$\sum_{j=0}^{i} \binom{i}{j} A_h^{(i-j)}(\rho_h^{(j)}, \phi_h) = \mathcal{O}(h^{k+2}) \left(\sum_{j=0}^{i} (\|u^{(j)}t)\|_{k+3} + \|(Lu)^{(i)}(t)\|_{k+2} \right) \|\phi_h\|_2. \tag{3.22}$$

Thus with (3.22), Lemma 3.3.2, and inverse inequality we have

$$\begin{aligned}
&\|\rho_h^{(i)}(t)\|_0^2 \\
\leq& \mathcal{O}(h^{k+2}) \left(\sum_{j=0}^{i} \|u^{(j)}(t)\|_{k+3} + \|(Lu)^{(i)}(t)\|_{k+2} \right) \|\phi_h\|_2 \\
&+ \mathcal{O}(h^{k+2}) \sum_{j=0}^{i} \|\rho_h^{(j)}(t)\|_{k+2} \|\phi_h\|_2 + C \sum_{j=0}^{i-1} \|\rho_h^{(j)}(t)\|_0 \|\phi_h\|_2 \\
=& \left[\mathcal{O}(h^{k+2}) \left(\sum_{j=0}^{i} \|u^{(j)}\|_{k+3} + \|(Lu)^{(i)}\|_{k+2} \right) + C \sum_{j=0}^{i-1} \|\rho_h^{(j)}(t)\|_0 \right] \|\phi_h\|_2 \\
\leq& \left(\mathcal{O}(h^{k+2}) \left(\sum_{j=0}^{i} \|u^{(j)}\|_{k+3} + \|(Lu)^{(i)}\|_{k+2} \right) + C \sum_{j=0}^{i-1} \|\rho_h^{(j)}(t)\|_0 \right) \|\rho_h^{(i)}(t)\|_0,
\end{aligned} \tag{3.23}$$

where (3.20) is applied in the last inequality.

With similar induction arguments as above, (3.23) implies

$$\|\rho_h^{(i)}(t)\|_0 \leq \mathcal{O}(h^{k+2}) \sum_{j=0}^{i} (\|u^{(j)}(t)\|_{k+3} + \|(Lu)^{(j)}(t)\|_{k+2}). \tag{3.24}$$

Take square for both sides of (3.24) then integrate from 0 to T and take square root for both sides, we can get (3.13). Take the maximum of the right hand side then the left hand side of (3.24) for $t \in [0, T]$, we can get (3.14). □

3.5 Accuracy Of The Semi-discrete Schemes

Throughout this section the generic constant C is independent of h. Although in principle it may depend on t though the coefficients $a_{ij}(t)$, $b_j(t)$, $c(t)$, we also treat it as independent of time since its time dependent version can always be replaced by a time independent constant after taking maximum over the ime interval $[0,T]$. In what follows we will state and prove the main theorems for wave, parabolic and the Schrödinger equations.

3.5.1 The hyperbolic problem

The main result for the wave equation can be stated as the following theorem.

Theorem 3.5.1. *If a_{ij}, b_j, $c \in C^2\left([0,T]; W^{k+2,\infty}(\Omega)\right)$, $u \in C^2\left([0,T]; H^{k+4}(\Omega)\right)$, then for the semi-discrete scheme (3.7) we have*

$$\begin{aligned}\|u_h - u\|_{L^2([0,T];l^2(\Omega))} \leq & Ch^{k+2}\left(\sum_{j=0}^{2}(\|u^{(j)}\|_{L^2([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^2([0,T];H^{k+2}(\Omega))})\right. \\ & \left. + \sum_{j=0}^{1}(\|u^{(j)}(0)\|_{k+3} + \|(Lu)^{(j)}(0)\|_{k+2})\right), \\ \|u_h - u\|_{L^\infty([0,T];l^2(\Omega))} \leq & Ch^{k+2}\sum_{j=0}^{2}(\|u^{(j)}\|_{L^\infty([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^\infty([0,T];H^{k+2}(\Omega))}),\end{aligned}$$

where C is independent of t, h, u, and f.

Proof. Note that for the numerical solution u_h we have

$$\langle u_h^{(2)}, v_h\rangle_h + A_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h. \tag{3.25}$$

The exact solution u satisfies $u_{tt} = -Lu + f$ thus the elliptic projection (3.5) satisfies

$$A_h(R_h u, v_h) = \langle u^{(2)} - f, v_h\rangle_h, \quad \forall v_h \in V_0^h.$$

Subtracting the two equations above, we get $\theta_h = u_h - R_h u$, which satisfies

$$\langle \theta_h^{(2)}, v_h\rangle_h + A_h(\theta_h, v_h) = -\langle \rho_h^{(2)}, v_h\rangle_h + \langle u^{(2)} - u_p^{(2)}, v_h\rangle, \quad \forall v_h \in V_0^h. \tag{3.26}$$

Note that

$$\frac{d}{dt}A_h(\theta_h, \theta_h) = A_h^{(1)}(\theta_h, \theta_h) + 2A_h(\theta_h, \theta_h^{(1)}) - \langle \mathbf{b}\cdot\nabla\theta_h, \theta_h^{(1)}\rangle_h + \langle \mathbf{b}\cdot\nabla\theta_h^{(1)}, \theta_h\rangle_h. \tag{3.27}$$

Thus by Lemma 2.3.7 and (2.6), we have

$$\begin{aligned}\langle \mathbf{b}\cdot\nabla\theta_h^{(1)},\theta_h\rangle_h =&(\mathbf{b}\cdot\nabla\theta_h^{(1)},\theta_h)+\mathcal{O}(h^2)|\mathbf{b}\theta_h|_2\|\nabla\theta_h^{(1)}\|_0\\ \le&(\mathbf{b}\cdot\nabla\theta_h^{(1)},\theta_h)+C\|\theta_h^{(1)}\|_0\|\theta_h\|_1\\ =&(\nabla\cdot(\mathbf{b}\theta_h),\theta_h^{(1)})+C\|\theta_h^{(1)}\|_0\|\theta_h\|_1\\ \le&C\|\theta_h^{(1)}\|_0\|\theta_h\|_1\le C\|\theta_h^{(1)}\|_{l^2}\|\theta_h\|_1,\end{aligned}\tag{3.28}$$

where an inverse inequality was applied to the first inequality and integration by parts in $\theta_h \in V_0^h$ yields the last equation.

Next we estimate $\|\theta_h^{(1)}(s)\|_0^2+\|\theta_h(s)\|_1^2$. Take $v_h=\theta_h^{(1)}$ in (3.26) and integrate with respect to t from 0 to s. With (3.27), we have

$$\begin{aligned}&\int_0^s \frac{d}{dt}\left(\frac{1}{2}\langle\theta_h^{(1)},\theta_h^{(1)}\rangle_h+\frac{1}{2}A_h(\theta_h,\theta_h)\right)dt\\ =&\frac{1}{2}\int_0^s A_h^{(1)}(\theta_h,\theta_h)-\langle\mathbf{b}\cdot\nabla\theta_h,\theta_h^{(1)}\rangle_h+\langle\mathbf{b}\cdot\nabla\theta_h^{(1)},\theta_h\rangle_h-2\langle\rho_h^{(2)},\theta_h^{(1)}\rangle_h+2\langle u^{(2)}-u_p^{(2)},\theta_h^{(1)}\rangle_h dt.\end{aligned}\tag{3.29}$$

With $\theta_h(0)=0$ and (3.28), this implies

$$\begin{aligned}&\frac{1}{2}(\|\theta_h^{(1)}(s)\|_{l^2}^2+A_h(\theta_h(s),\theta_h(s)))-\frac{1}{2}\|\theta_h^{(1)}(0)\|_{l^2}^2\\ \le&C\int_0^s(\|\theta_h\|_1^2+\|\theta_h^{(1)}\|_0\|\theta_h\|_1)dt+C\int_0^s\|\rho_h^{(2)}\|_0\|\theta_h^{(1)}\|_0dt\\ &+C\int_0^s\|u^{(2)}-u_p^{(2)}\|_{l^2}\|\theta_h^{(1)}\|_0dt\\ \le&C\int_0^s(\|\theta_h^{(1)}\|_0^2+\|\theta_h\|_1^2)dt+C\int_0^s(\|\rho_h^{(2)}\|_0^2+\|u^{(2)}-u_p^{(2)}\|_{l^2}^2)dt,\end{aligned}\tag{3.30}$$

where Cauchy-Schwarz inequality was used in the last inequality.

Thus with (2.6), (3.9), and (3.30) we have

$$\begin{aligned}&\|\theta_h^{(1)}(s)\|_0^2+\|\theta_h(s)\|_1^2\le C\|\theta_h^{(1)}(s)\|_{l^2}^2+CA_h(\theta_h(s),\theta_h(s))\\ \le&C\|\theta_h^{(1)}(0)\|_{l^2}^2+C\int_0^s(\|\theta_h^{(1)}\|_0^2+\|\theta_h\|_1^2)dt+C\int_0^s(\|\rho_h^{(2)}\|_0^2+\|u^{(2)}-u_p^{(2)}\|_{l^2}^2)dt.\end{aligned}\tag{3.31}$$

With the Gronwall inequality (3.3.4) we can eliminate the second term to find

$$\|\theta_h^{(1)}(s)\|_0^2 + \|\theta_h(s)\|_1^2 \leq C\|\theta_h^{(1)}(0)\|_{l^2}^2 + C\int_0^s \|\rho_h^{(2)}\|_0^2 + \|u^{(2)} - u_p^{(2)}\|_{l^2}^2 dt.$$

With (3.14) and Theorem 2.4.5 we have

$$\|\theta_h^{(1)}(s)\|_0^2 + \|\theta_h(s)\|_1^2 \leq C\|\theta_h^{(1)}(0)\|_{l^2}^2 + \mathcal{O}(h^{2k+4})\int_0^s \sum_{j=0}^{2}(\|u^{(j)}\|_{k+3} + \|(Lu)^{(j)}\|_{k+2})^2 dt,$$

i.e.

$$\|\theta_h^{(1)}(s)\|_0 + \|\theta_h(s)\|_1 \leq C\|\theta_h^{(1)}(0)\|_{l^2} + \mathcal{O}(h^{k+2})\int_0^s \sum_{j=0}^{2}(\|u^{(j)}\|_{k+3} + \|(Lu)^{(j)}\|_{k+2}) dt. \tag{3.32}$$

To estimate $\|\theta_h^{(1)}(0)\|_{l^2}$ we use Theorem 2.4.5, (3.14), and (2.6),

$$\begin{aligned}
\|\theta_h^{(1)}(0)\|_{l^2} =& \|(u_1)_I - (R_h u)^{(1)}(0)\|_{l^2} \\
=& \|(u_1)_I - (u_1)_p + (u_1)_p - (R_h u)^{(1)}(0)\|_{l^2} \\
\leq& \|(u_1)_I - (u_1)_p\|_{l^2} + \|(u_1)_p - (R_h u)^{(1)}(0)\|_{l^2} \\
=& \|u_1 - (u_1)_p\|_{l^2} + \|(u_1)_p - R_h(u^{(1)}(0))\|_{l^2} \\
=& \|u_1 - (u_1)_p\|_{l^2} + \|(u_1)_p - R_h(u_1)\|_{l^2} \\
=& \mathcal{O}(h^{k+2})(\|u_1\|_{k+3} + \|Lu_1\|_{k+2}).
\end{aligned}$$

Then we have

$$\begin{aligned}
&\|\theta_h^{(1)}\|_0 + \|\theta_h\|_1 \\
\leq& \mathcal{O}(h^{k+2})\left(\|u_1\|_{k+3} + \|Lu_1\|_{k+2} + \int_0^s \sum_{j=0}^{2}(\|u^{(j)}\|_{k+3} + \|(Lu)^{(j)}\|_{k+2}) dt\right).
\end{aligned} \tag{3.33}$$

Now with (3.13), (3.14), and Theorem 2.4.5, the proof is concluded. □

3.5.2 The parabolic problem

We now present the main result for the parabolic problem.

Theorem 3.5.2. *If a_{ij}, b_j, $c \in C^1([0,T]; W^{k+1,\infty}(\Omega))$, $u \in C^1([0,T]; H^{k+4}(\Omega))$, then for the semi-discrete scheme (3.6) we have*

$$\begin{aligned}\|u_h - u\|_{L^2([0,T];l^2(\Omega))} \leq & Ch^{k+2}\sum_{j=0}^{1}(\|u^{(j)}\|_{L^2([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^2([0,T];H^{k+2}(\Omega))}),\\ \|u_h - u\|_{L^\infty([0,T];l^2(\Omega))} \leq & Ch^{k+2}\sum_{j=0}^{1}(\|u^{(j)}\|_{L^\infty([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^\infty([0,T];H^{k+2}(\Omega))}),\end{aligned}$$

where C is independent of t, h, u, and f.

Proof. By our semi-discrete numerical scheme (3.6) and the definition of the elliptic projection (3.5), we have

$$\langle \theta_h^{(1)}, v_h\rangle_h + A_h(\theta_h, v_h) = -\langle \rho_h^{(1)}, v_h\rangle_h + \langle u^{(1)} - u_p^{(1)}, v_h\rangle, \quad \forall v_h \in V_0^h. \tag{3.34}$$

Take $v_h = \theta_h^{(1)}$ in (3.34) and integrate with respect to t from 0 to s,

$$\begin{aligned}&\int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h + \frac{1}{2}\frac{d}{dt}A_h(\theta_h, \theta_h)dt\\ =&\frac{1}{2}\int_0^s A_h^{(1)}(\theta_h, \theta_h) - \langle \mathbf{b}\cdot\nabla\theta_h, \theta_h^{(1)}\rangle_h + \langle \mathbf{b}\cdot\nabla\theta_h^{(1)}, \theta_h\rangle_h - 2\langle \rho_h^{(1)}, \theta_h^{(1)}\rangle_h + 2\langle u^{(1)} - u_p^{(1)}, \theta_h^{(1)}\rangle_h dt.\end{aligned} \tag{3.35}$$

Note that $\theta_h(0) = 0$, then with (2.6), (3.28), and (3.35) we have

$$\begin{aligned}&\int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h dt + \|\theta_h(s)\|_1^2 \leq \int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h dt + CA_h(\theta_h(s), \theta_h(s))\\ \leq& C\int_0^s \|\theta_h\|_1^2 dt + C\int_0^s \|\theta_h^{(1)}\|_{l^2}\|\theta_h\|_1 dt + C\int_0^s \|\rho_h^{(1)}\|_{l^2}\|\theta_h^{(1)}\|_{l^2} dt\\ &+ C\int_0^s \|u^{(1)} - u_p^{(1)}\|_{l^2}\|\theta_h^{(1)}\|_{l^2} dt\\ \leq& C\int_0^s \|\theta_h\|_1^2 dt + \int_0^s \epsilon\langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h + \frac{C}{4\epsilon}\|\theta_h\|_1^2 dt + \int_0^s \epsilon\langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h + \frac{C}{4\epsilon}\|\rho_h^{(1)}\|_0^2 dt\\ &+ \int_0^s \epsilon\langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h + \frac{C}{4\epsilon}\|u^{(1)} - u_p^{(1)}\|_{l^2}^2 dt,\end{aligned}$$

where Cauchy-Schwartz inequality was applied in the last inequality. Thus we have

$$(1-3\epsilon)\int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h dt + \|\theta_h(s)\|_1^2 \leq C(1+\frac{1}{4\epsilon})\int_0^s \|\theta_h\|_1^2 dt + \frac{C}{4\epsilon}\int_0^s \|\rho_h^{(1)}\|_0^2 dt + \frac{C}{4\epsilon}\int_0^s \|u^{(1)} - u_p^{(1)}\|_{l^2}^2 dt.$$

Now take ϵ small enough to make $1-3\epsilon \geq \frac{1}{2}$ then

$$\begin{aligned}\frac{1}{2}\int_0^s \langle \theta_h^{(1)}(s), \theta_h^{(1)}\rangle_h(s)dt + \|\theta_h(s)\|_1^2 \leq C\int_0^s \|\rho_h^{(1)}\|_0^2 dt + C\int_0^s \|u^{(1)} - u_p^{(1)}\|_{l^2}^2 dt \\ +C\int_0^s \left(\|\theta_h(t)\|_1^2 + \frac{1}{2}\int_0^t \langle \theta_h^{(1)}(\eta), \theta_h^{(1)}(\eta)\rangle_h d\eta\right) dt.\end{aligned} \tag{3.36}$$

Next, apply Gronwall's inequality to eliminate the last term of the right hand side of (3.36) to find

$$\frac{1}{2}\int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h dt + \|\theta_h\|_1^2 \leq C\int_0^s \|\rho_h^{(1)}\|_0^2 dt + C\int_0^s \|u^{(1)} - u_p^{(1)}\|_{l^2}^2 dt.$$

Using (3.13), (3.14), and Theorem 2.4.5 we have

$$\frac{1}{2}\int_0^s \langle \theta_h^{(1)}, \theta_h^{(1)}\rangle_h dt + \|\theta_h\|_1^2 \leq \mathcal{O}(h^{k+2})\int_0^s \sum_{j=0}^{1}(\|u^{(j)}\|_{k+3} + \|(Lu)^{(j)}\|_{k+2})dt,$$

concluding the proof. □

3.5.3 The linear Schrödinger equation

Consider the problem

$$\begin{cases} iu_t = -\Delta u + Vu + f, & \text{in } \Omega \times [0,T], \\ u(\mathbf{x},t) = 0, & \text{on } \partial\Omega \times [0,T], \\ u(\mathbf{x},0) = u_0(\mathbf{x}), & \text{in } \Omega, \end{cases} \tag{3.37}$$

where $\Omega \in R^2$ is a rectangular domain, the functions $u_0(\mathbf{x}), f(\mathbf{x},t)$, and the solution $u(\mathbf{x},t)$ are complex-valued while the potential function $V(\mathbf{x},t)$ is real-valued, non-negative, and bounded for all $(\mathbf{x},t) \in \Omega \times [0,T]$.

In this subsection we work with complex-valued functions and the definition of inner product and the induced norms are modified accordingly. For instance, for complex-valued v, $w \in L^2(\Omega)$, the inner product is defined as

$$(v,w) := \int_\Omega v\bar{w}d\mathbf{x}.$$

We assume all the functions of the function spaces defined previously are complex-valued for this subsection, such as $H^k(\Omega)$, $H_0^k(\Omega)$, V_0^h, etc.

The variational form of (3.37) is: for $t \in [0,T]$, find $u(t) \in H_0^1(\Omega)$ satisfying:

$$\begin{cases} i\,(u_t, v) - (\nabla u, \nabla v) - (Vu, v) = (f, v), & \forall v \in H_0^1(\Omega), \\ u(0) = u_0, & \forall v \in H_0^1(\Omega). \end{cases} \tag{3.38}$$

The semi-discrete numerical scheme discretizing (3.38) is to find $u_h \in V_0^h$ satisfying

$$\begin{cases} i\langle (u_h)_t, v_h\rangle_h - \langle \nabla u_h, \nabla v_h\rangle_h - \langle Vu_h, v_h\rangle_h = \langle f, v_h\rangle_h, & \forall v_h \in V_0^h, \\ u_h(0) = (u_0)_I, \end{cases} \tag{3.39}$$

and the elliptic projection $R_h u \in V_0^h$ is defined as

$$\langle \nabla R_h u, \nabla v_h\rangle_h + \langle V R_h u, v_h\rangle_h = \langle -\Delta u + Vu, v_h\rangle_h, \quad \forall v_h \in V_0^h. \tag{3.40}$$

As in Section 3.4, we split the error into two parts

$$e = \theta_h + \rho_h,$$

where $\theta_h = u_h - R_h u \in V_0^h$ and $\rho_h = R_h u - u_p \in V_0^h$. The estimates for $\rho_h^{(m)}$, $m \geq 0$ from Lemma 3.4.1 are still valid.

Theorem 3.5.3. *If $u \in C^1([0,T]; H^{k+4}(\Omega))$, then for the semi-discrete scheme (3.39) we have*

$$\begin{aligned} \|u_h - u\|_{L^2([0,T];l^2(\Omega))} \leq & Ch^{k+2} \sum_{j=0}^{1} (\|u^{(j)}\|_{L^2([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^2([0,T];H^{k+2}(\Omega))}), \\ \|u_h - u\|_{L^\infty([0,T];l^2(\Omega))} \leq & Ch^{k+2} \sum_{j=0}^{1} (\|u^{(j)}\|_{L^\infty([0,T];H^{k+3}(\Omega))} + \|(Lu)^{(j)}\|_{L^\infty([0,T];H^{k+2}(\Omega))}), \end{aligned}$$

where C is independent of t, h, u, and f.

Proof. As in the parabolic case we start by estimating θ_h.

$$\langle \theta_h^{(1)}, v_h \rangle_h + i\langle \nabla\theta_h, \nabla v_h\rangle_h + i\langle V\theta_h, v_h\rangle_h = -\langle \rho_h^{(1)}, v_h\rangle_h + \langle u^{(1)} - u_p^{(1)}, v_h\rangle_h, \quad \forall v_h \in V_0^h. \tag{3.41}$$

Taking $v_h = \theta_h$ in (3.41) and taking real part,

$$\begin{aligned}\frac{d}{dt}\|\theta_h\|_{l^2(\Omega)}^2 = \frac{d}{dt}\langle \theta_h, \theta_h\rangle_h =& 2Re\left(-\langle \rho_h^{(1)}, \theta_h\rangle_h + \langle u^{(1)} - u_p^{(1)}, \theta_h\rangle_h\right)\\ \leq& 2\left(\|\rho_h^{(1)}\|_{l^2(\Omega)} + \|u^{(1)} - u_p^{(1)}\|_{l^2(\Omega)}\right)\|\theta_h\|_{l^2(\Omega)}.\end{aligned}$$

Since $\frac{d}{dt}\|\theta_h\|_{l^2(\Omega)}^2 = 2\|\theta_h\|_{l^2(\Omega)}\frac{d}{dt}\|\theta_h\|_{l^2(\Omega)}$, it impilies

$$\frac{d}{dt}\|\theta_h\|_{l^2(\Omega)} \leq \|\rho_h^{(1)}\|_{l^2(\Omega)} + \|u^{(1)} - u_p^{(1)}\|_{l^2(\Omega)}.$$

Upon integrating this inequality with respect to t from 0 to s we have

$$\|\theta_h(s)\|_{l^2(\Omega)} \leq \|\theta_h(0)\|_{l^2(\Omega)} + \int_0^s (\|\rho_h^{(1)}\|_{l^2(\Omega)} + \|u^{(1)} - u_p^{(1)}\|_{l^2(\Omega)})dt.$$

Now, using Theorem 2.4.5, (3.14), and (2.6) we have

$$\begin{aligned}\|\theta_h(0)\|_{l^2} =& \|(u_0)_I - (R_h u)(0)\|_{l^2}\\ =& \|(u_0)_I - (u_0)_p + (u_0)_p - (R_h u)(0)\|_{l^2}\\ \leq& \|(u_0)_I - (u_0)_p\|_{l^2} + \|(u_0)_p - (R_h u)(0)\|_{l^2}\\ =& \|u_0 - (u_0)_p\|_{l^2} + \|(u_0)_p - R_h u_0\|_{l^2}\\ =& O(h^{k+2})(\|u_0\|_{k+3} + \|Lu_0\|_{k+2}).\end{aligned}$$

With this result in concert with (3.13), (3.14), and Theorem 2.4.5 we note

$$\|\theta_h(s)\|_{l^2(\Omega)} \leq \mathcal{O}(h^{k+2})\left(\|u_0\|_{k+3} + \|Lu_0\|_{k+2} + \int_0^s \sum_{j=0}^{1}(\|u^{(j)}\|_{k+3} + \|(Lu)^{(j)}\|_{k+2})dt\right).$$

Together with (3.13), (3.14), and Theorem 2.4.5, proof is concluded. □

3.5.4 Neumann boundary conditions and ℓ^∞-norm estimate

For Neumann type boundary conditions, due to Lemma 3.3.1 and Lemma 3.3.2, in general we can only prove $(k+\frac{3}{2})$-th order accuracy for the hyperbolic equation, parabolic equation, and linear Schrödinger equation. As explained in Remark 3.3.3, the half order loss happens for Neumann boundary condition only when the second order operator coefficient $\mathbf{a}$ is not diagonal, e.g., when the PDE contains second order mixed derivatives. If $\mathbf{a}$ is diagonal, then all results of $(k+2)$-th order in ℓ^2 norm in this Section can be easily extended to the Neumann boundary conditions.

For Lagrangian Q^k finite element method without any quadrature solving the elliptic equation with Dirichlet boundary conditions, the best superconvergence order in max norm of function values at Gauss-Lobatto that one can prove is $\mathcal{O}(|\log h|h^{k+2})$ in two dimensions, see [62] and references therein. Thus we do not expect better results can be proven in the Q^k spectral element method in ℓ^∞ norm over all nodes of degree of freedoms.

3.6 Implementation For Nonhomogeneous Dirichlet Boundary Conditions

Consider the hyperbolic problem on $\Omega=(0,1)^2$ with compatible nonhomogeneous Dirichlet boundary condition and initial value

$$\begin{aligned} u_{tt} =& -Lu+f(\mathbf{x},t) && \text{in } \Omega\times(0,T],\\ u(\mathbf{x},t) =& g && \text{on } \partial\Omega\times[0,T],\\ u(\mathbf{x},0) =& u_0(\mathbf{x}), \quad u_t(\mathbf{x},0)=u_1(\mathbf{x}) && \text{on } \Omega\times\{t=0\}. \end{aligned} \tag{3.42}$$

As in [62], [66], by abusing notation, we define

$$g(x,y,t)=\begin{cases} 0, & \text{if} \quad (x,y)\in(0,1)\times(0,1),\\ g(x,y,t), & \text{if} \quad (x,y)\in\partial\Omega, \end{cases}$$

and define $g_I\in V^h$ as the Q^k Lagrange interpolation at $(k+1)\times(k+1)$ Gauss-Lobatto points for each cell on Ω of $g(x,y,t)$. Namely, $g_I\in V^h$ is the piecewise Q^k interpolant of

g along $\partial\Omega$ at the boundary grid points and $g_I = 0$ at the interior grid points. Then the semi-discrete scheme for problem (3.42) is as follows: for $t \in [0, T]$, find $\tilde{u}_h \in V_0^h$ such that

$$\begin{aligned} \langle \tilde{u}_h^{(2)}, v_h \rangle_h + A_h(\tilde{u}_h, v_h) =& \langle f, v_h \rangle_h - A_h(g_I, v_h), \quad \forall v_h \in V_0^h, \\ \tilde{u}_h(0) = R_h u_0, \quad \tilde{u}_h^{(1)}(0) =& (u_1)_I. \end{aligned} \tag{3.43}$$

Then

$$u_h := \tilde{u}_h + g_I, \tag{3.44}$$

is the desired numerical solution. Notice that u_h and $\tilde{u}_h$ are the same at all interior grid points.

For the initial value of numerical solution, instead of using discrete elliptic projection, we can also use $\tilde{u}_h(0) = u(x, y, 0)_I$ in (3.43) where $u(x, y, 0)_I$ is the piecewise Lagrangian Q^k interpolation of $u(x, y, 0)$. In all numerical tests in Section 3.7, $(k+2)$-th order accuracy is still observed for the initial condition $\tilde{u}_h(0) = u(x, y, 0)_I$.

The treatment for nonhomogeneous Dirichlet boundary condition above can be extended naturally to the parabolic equation and linear Schrödinger equation,

Remark 3.6.1. For the $(k+2)$-th order accuracy of the scheme (3.43), it can be shown analogously as in [62], and in Section 3.4 and Section 3.5 by defining discrete elliptic projection as

$$R_h u := \tilde{R}_h u + g_I, \tag{3.45}$$

where $\tilde{R}_h u \in V_0^h$ satisfying

$$A_h(\tilde{R}_h u, v_h) = \langle -Lu, v_h \rangle_h - A_h(g_I, v_h), \quad \forall v_h \in V_0^h, \quad 0 \le t \le T.$$

3.7 Numerical Examples

In this section we present numerical examples for the wave equation, a parabolic equation and the Schrödinger equation.

3.7.1 Numerical examples for the wave equation

Timestepping

After semidiscretization the method (3.7) can be written as

$$\frac{d^2\mathbf{u}_h}{dt^2} = Q\mathbf{u}_h,$$

where $\mathbf{u}_h$ is a vector containing all the degrees of freedom and Q is a matrix. To evolve in time we expand the approximate solution around $t + \Delta$ and $t - \Delta t$

$$\mathbf{u}_h(t+\Delta t) + \mathbf{u}_h(t-\Delta t) = 2\mathbf{u}_h(t) + \Delta t^2 \frac{d^2\mathbf{u}_h(t)}{dt^2} + \frac{\Delta t^4}{12}\frac{d^4\mathbf{u}_h(t)}{dt^4} + \frac{\Delta t^6}{360}\frac{d^6\mathbf{u}_h(t)}{dt^4} + \mathcal{O}(\Delta t^8).$$

Replacing the even time derivatives with applications of the matrix Q we obtain, for example, a 6th order accurate explicit temporal approximation

$$\mathbf{u}_h(t+\Delta t) + \mathbf{u}_h(t-\Delta t) = 2\mathbf{u}_h(t) + \Delta t^2 Q^2\mathbf{u}_h(t) + \frac{\Delta t^4}{12}Q^4\mathbf{u}_h(t) + \frac{\Delta t^6}{360}Q^6\mathbf{u}_h(t).$$

Standing mode with Dirichlet conditions

In this experiment we solve the the wave equation $u_{tt} = u_{xx} + u_{yy}$ with homogenous Dirichlet boundary conditions in the square domain $(x, y) \in [-\pi, \pi]^2$. We take the initial data to be

$$u(x, y, 0) = \sin(x)\sin(y), \quad u_t(x, y, 0) = 0,$$

which results in the exact standing mode solution

$$u(x, y, 0) = \sin(x)\sin(y)\cos(\sqrt{2}t).$$

We consider the two cases $k = 2$ and $k = 4$ and discretize on three different sequences of grids. The first sequence contains only plain Cartesian of increasing refinement. The second sequence consists of the same grids as in the Cartesian sequence but with all the interior nodes perturbed by a two dimensional uniform random variable with each component drawn from $[-h/4, h/4]$. The nodes of the third sequence are

$$(x, y) = (\xi + 0.1 \sin(\xi) \sin(\eta), \eta + 0.1 \sin(\eta) \sin(\xi)), \qquad (\xi, \eta) = [-\pi, \pi]^2,$$

and this is refined in the same ways as the Cartesian sequence. Typical examples of the grids are displayed in Figure 3.1. Even though the equation contains no coefficients, variable coefficients are still involved for the second and the third sequences of grids. The variable coefficients are induced by the geometric transformations of the elements in the mesh to a reference rectangle element. However, on a randomly perturbed grid, the variable coefficients are not smooth across cell interfaces. The variable coefficients are smooth in a smoothly perturbed grid.

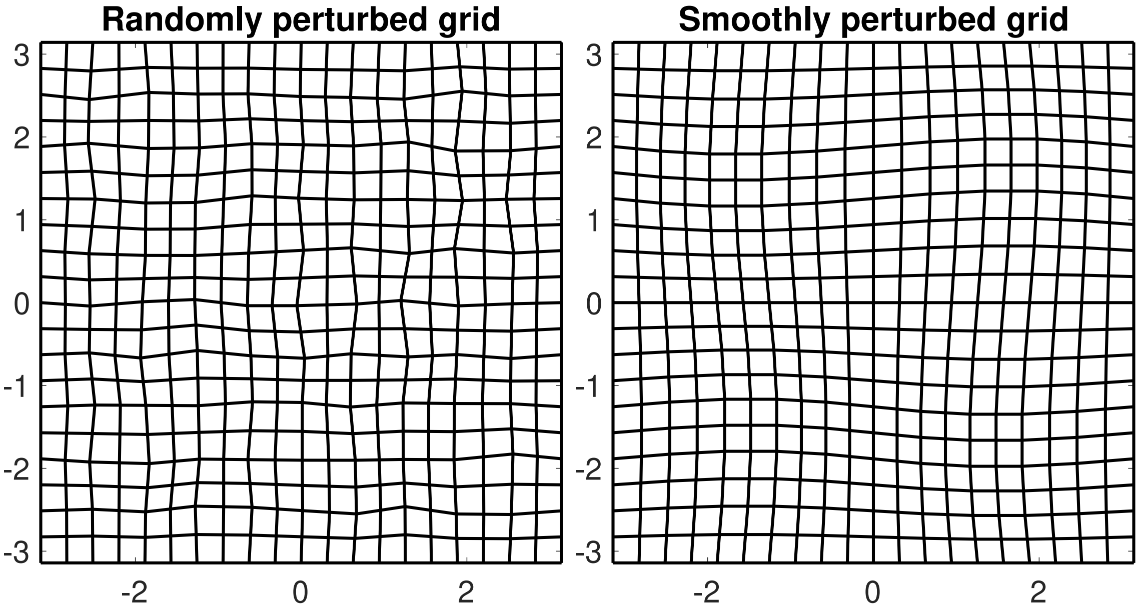


Figure 3.1. Two typical grids used in the numerical examples in Section 3.7.1 and 3.7.1.

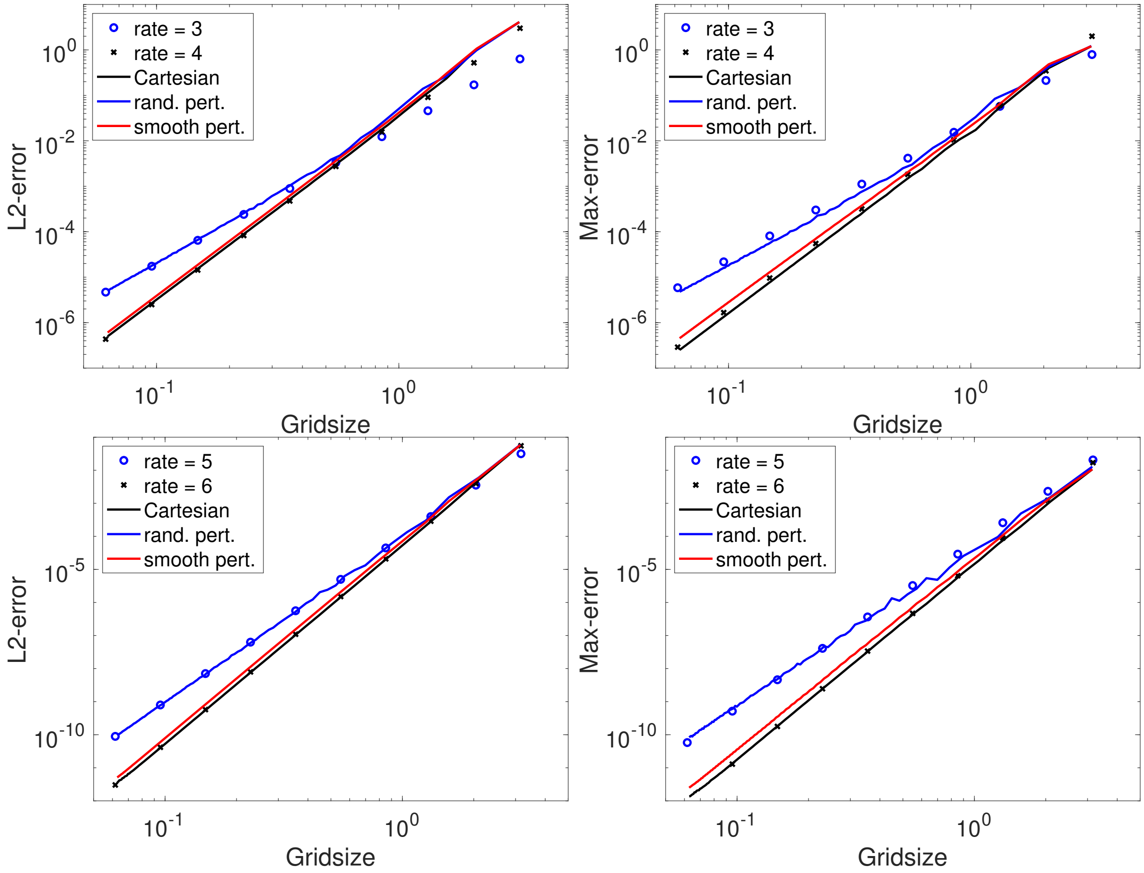


Figure 3.2. Dirichlet problem in a square. Errors measured in the l^2 and the l^∞ norms for the three different sequences of grids. The top row is for $k = 2$ and the bottom row is for $k = 4$.

We evolve the numerical solution until time 5 by the time stepping discussed in Section 3.7.1 of order of accuracy 4 when $k = 2$ and 6 when $k = 4$. To get clean measurements of the error we report the time integrated errors

$$\left(\int_0^5 \|u(\cdot,t) - u_h(\cdot,t)\|_{l^2}^2 \, dt\right)^{\frac{1}{2}}, \qquad \int_0^5 \|u(\cdot,t) - u_h(\cdot,t)\|_{l^\infty} \, dt,$$

for the spatial l^2 and l^∞ errors respectively.

The results are displayed in Figure 3.2. Note that here and in the rest of this section the solid lines in the figures are the computed errors, using many different grid sizes, and the symbols are indicating the slopes or rates of convergence of the curves. The Cartesian grids and smoothly perturbed grids satisfy the assumptions of the theory developed in this chapter while the second sequence of randomly perturbed grids does not. The results confirm the theoretical predictions for smooth variable coefficients as the rate of convergence is $k+2$ for the l^2-norm in the cases of the Cartesian meshes and the smoothly perturbed meshes. We also observe the rate $k+2$ in the l^∞-norm for these cases. For the non-smooth variable coefficients resulting from the randomly perturbed grid, which not covered by our theory, we see a rate of convergence of $k+1$ in the l^2-norm.

Standing mode in a sector of an annulus with Dirichlet conditions

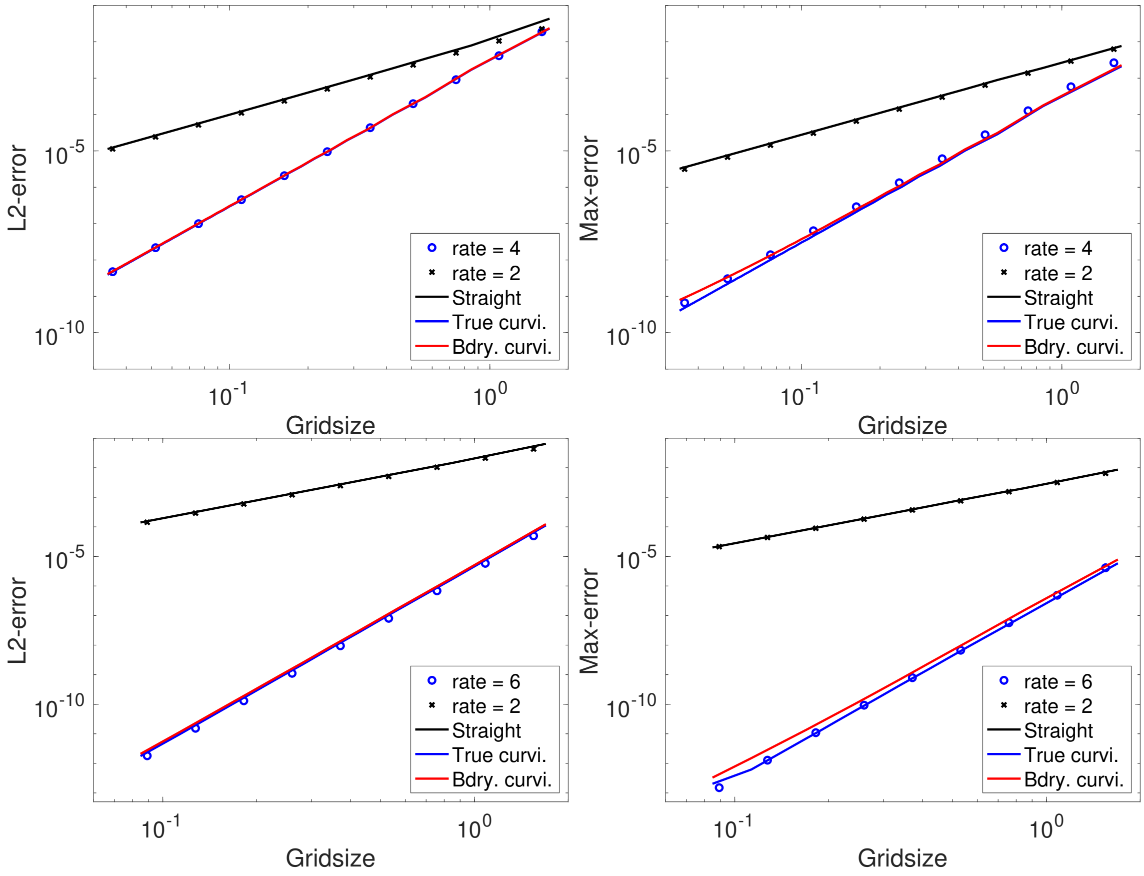


Figure 3.3. Dirichlet problem in an annular sector. Errors measured in the l^2 and the l^∞ norms for the three different sequences of grids. The top row is for $k=2$ and the bottom row is for $k=4$. These results are for the annular problem with homogenous Dirichlet boundary conditions.

In this experiment we solve the wave equation $u_{tt} = u_{xx} + u_{yy}$ and and with homogenous Dirichlet boundary conditions. The computational domain is the first quadrant of the annular region between two circles with radii $r_{\rm i} = 7.58834243450380438$ and $r_{\rm o} = 14.37253667161758967$, i.e. the domain is described by $(x, y) = (r\cos\theta, r\sin\theta)$ where

$$r_{\rm i} \leq r \leq r_{\rm o}, \quad 0 \leq \theta \leq \pi/2.$$

On this domain the standing mode

$$u(r, \theta, t) = J_4(r)\sin(4\theta)\cos(t),$$

is an exact solution and we use this solution to specify the initial conditions and to compute errors.

We consider the two cases $k = 2$ and $k = 4$ and discretize on three different sequences of grids. The first sequence uses a straight sided approximation of the annulus and all internal elements are quadrilaterals with straight sides. The second sequence uses curvilinear elements throughout the domain and all internal element boundaries conform with the polar coordinate transformation. After the smooth mapping to the unit square, smooth variable coefficients emerge due to the geometric terms. The metric terms are approximated with numerical differentiation using the values at the quadrature points. The third sequence is the same as the second sequence but all the internal element edges are straight. The meshes in the last sequence are likely close to those that would be provided by most grid generators.

We evolve the numerical solution until time 1 by the time stepping discussed in Section 3.7.1 of order of accuracy 4 when $k = 2$ and 6 when $k = 4$. Again, to get clean measurements of the error we report the time integrated errors

$$\left(\int_0^1 \|u(\cdot, t) - u_h(\cdot, t)\|_{l^2}^2 \, dt\right)^{\frac{1}{2}}, \qquad \int_0^1 \|u(\cdot, t) - u_h(\cdot, t)\|_{l^\infty} \, dt,$$

for the spatial l^2 and l^∞ errors respectively.

The results are displayed in Figure 3.3. Here, as expected, we only observe second order accuracy independent of k for the non-geometry-conforming meshes. We observe a convergence at the rate of $k+2$ in both the l^2-norm and l^∞-norm for the geometry-conforming meshes. The true curvilinear grids are covered by our theory since the variable coefficients due to the geometric transformation are smooth. For the third sequence of grids, since internal edges are straightsided, the variable coefficients from the geometric transformation are not smooth across edges thus this configuration is not covered by our theory. Nonetheless, its convergence rate is still $k + 2$.

Standing mode with Neumann conditions

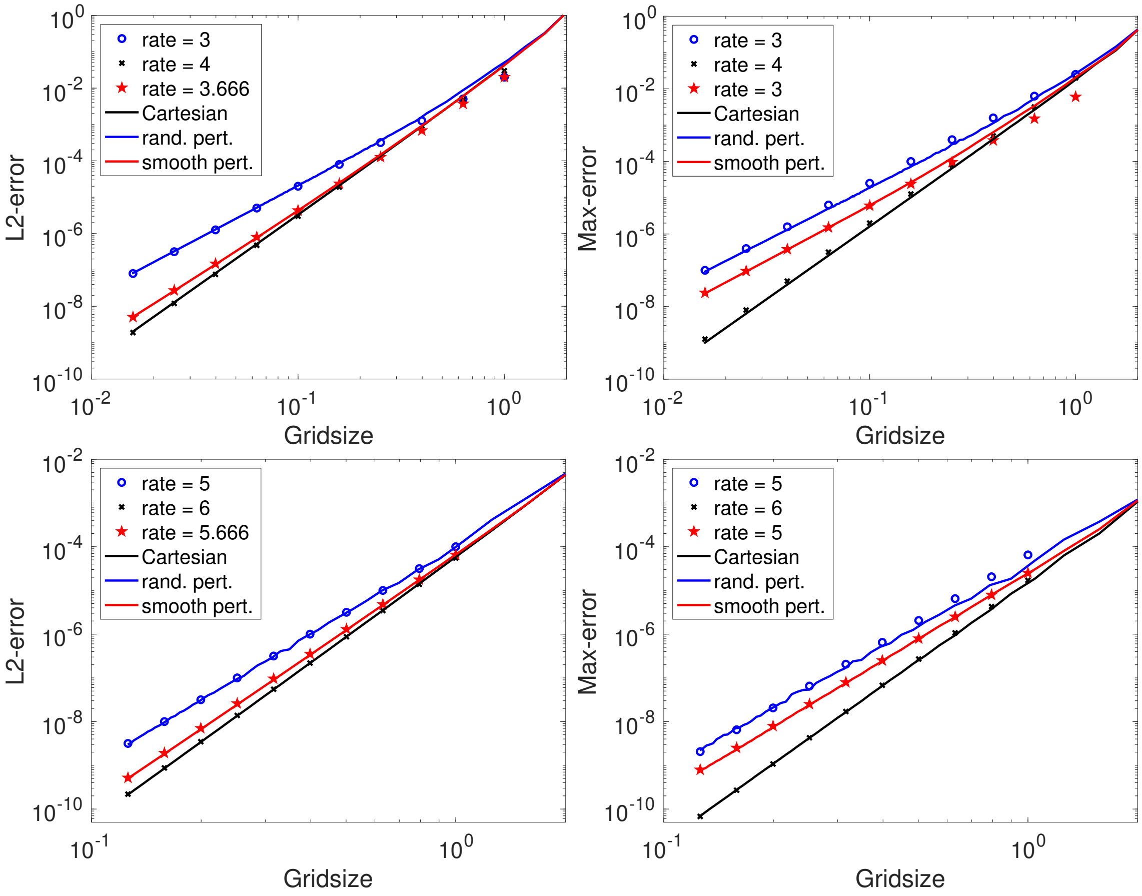


Figure 3.4. Neumann square problem. Errors measured in the l^2 and the l^∞ norms for the three different sequences of grids. The top row is for $k = 2$ and the bottom row is for $k = 4$.

In this experiment we approximate the solution to the wave equation $u_{tt} = u_{xx} + u_{yy}$ in the square domain $(x, y) \in [-\pi, \pi]^2$. Then with homogenous Neumann boundary conditions and initial data

$$u(x, y, 0) = \cos(x)\cos(y), \quad u_t(x, y, 0) = 0,$$

the exact standing mode solution is

$$u(x, y, 0) = \cos(x)\cos(y)\cos(\sqrt{2}t).$$

We consider the two cases $k = 2$ and $k = 4$ and discretize on the same three sequences of grids as those used in §3.7.1. We evolve the numerical solution until time 5 as above and we report the time integrated errors as above.

The results are displayed in Figure 3.4. For the Cartesian mesh we observe a rate of convergence $k + 2$ in the ℓ^2-norm, confirming our theory. For the smoothly perturbed grids, which corresponds to smooth variable coefficients resulting in mixed second order derivatives on the reference rectangular mesh, the rate in the l^2-norm appears to be $k + 5/3$. As explained in Section 3.5.4, only $(k + \frac{3}{2})$-th order can be proven when both mixed second order derivatives and Neumann boundary conditions are involved. As in the Dirichlet case, the randomly perturbed grid yields rates of convergence $k + 1$ in both norms.

Standing mode in a sector of an annulus with Neumann conditions

In this experiment we solve the the wave equation $u_{tt} = u_{xx} + u_{yy}$ with homogenous Neumann boundary conditions. The computational domain is again the first quadrant of the annular region between two circles, now with radii $r_{\rm i} = 5.31755312608399$ and $r_{\rm o} = 9.28239628524161$, to satisfy the boundary conditions. On this domain the standing mode

$$u(r, \theta, t) = J_4(r)\cos(4\theta)\cos(t),$$

is an exact solution and we use this solution to specify the initial conditions and to compute errors.

As in the previous examples we consider the two cases $k = 2$ and $k = 4$ and discretize on the same three different sequences of grids as was used in the Dirichlet example above. We evolve the numerical solution until time 1 in the same way as above and we report the time integrated errors.

The results are displayed in Figure 3.5. Here, the only grid satisfying our assumptions is the true curvilinear grid. For this case, the problem is equivalent to solving a variable coefficient problem $u_{tt} = u_{rr} + \frac{1}{r^2}u_{\theta\theta} + \frac{1}{r}u_r$ on rectangular meshes for polar coordinates $(r, \theta) \in [r_i, r_o] \times [0, \frac{\pi}{2}]$. Since there are no mixed second order derivatives, by our theory as explained in Section 3.5.4, $(k + 2)$-th order in the ℓ^2-norm can still be proven. We can see that the rate for the true curvilinear grid is indeed $k + 2$ in ℓ^2-norm, confirming our theory for Neumann boundary conditions.

3.7.2 Numerical tests for the parabolic equation

For problem (3.2) on the domain $\Omega = (0, \pi)^2$, we set $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ with $a_{11} = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)(1 + y + y^2 + x\cos y)$, $a_{12} = a_{21} = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)\left(1 + \frac{1}{2}(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4\right.$ $a_{22} = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)(1 + x^2)$, $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ with $b_1 = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)\left(\frac{1}{5} + x\right)$, $b_2 = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)\left(\frac{1}{5} - y\right)$, and $c = \left(\frac{3}{4} + \frac{1}{4}\sin(t)\right)(10 + x^4y^3)$. For time discretization in (3.6), we use the third order backward differentiation formula (BDF) method. Let $u(x, y, t) = (\frac{3}{4}+\frac{1}{4}\sin(t))(-\sin(y)\cos(y)\sin(x)^2)$ and we use a potential function f so that u is the exact solution. The time step is set as $\Delta t = \min(\frac{\Delta x}{10}, \frac{\Delta x}{10 b_M}, \frac{f_M}{10})$, where $b_M = \max_{\mathbf{x}\in\Omega, i=1,2} |b_i(0, \mathbf{x})|$ and $f_M = \max_{\mathbf{x}\in\Omega} |f(0, \mathbf{x})|$. The errors at time $T = 0.1$ are listed in Table 3.1, in which we observe order around $k + 2$ for the ℓ^2-norm.

3.7.3 Numerical tests for the linear Schrödinger equation

For problem (3.37) on the domain $(0, 2)^2$, a fourth-order explicit Adams-Bashforth as time discretization for (3.39). The solution and potential functions are as follows: $u(x, y, t) =$

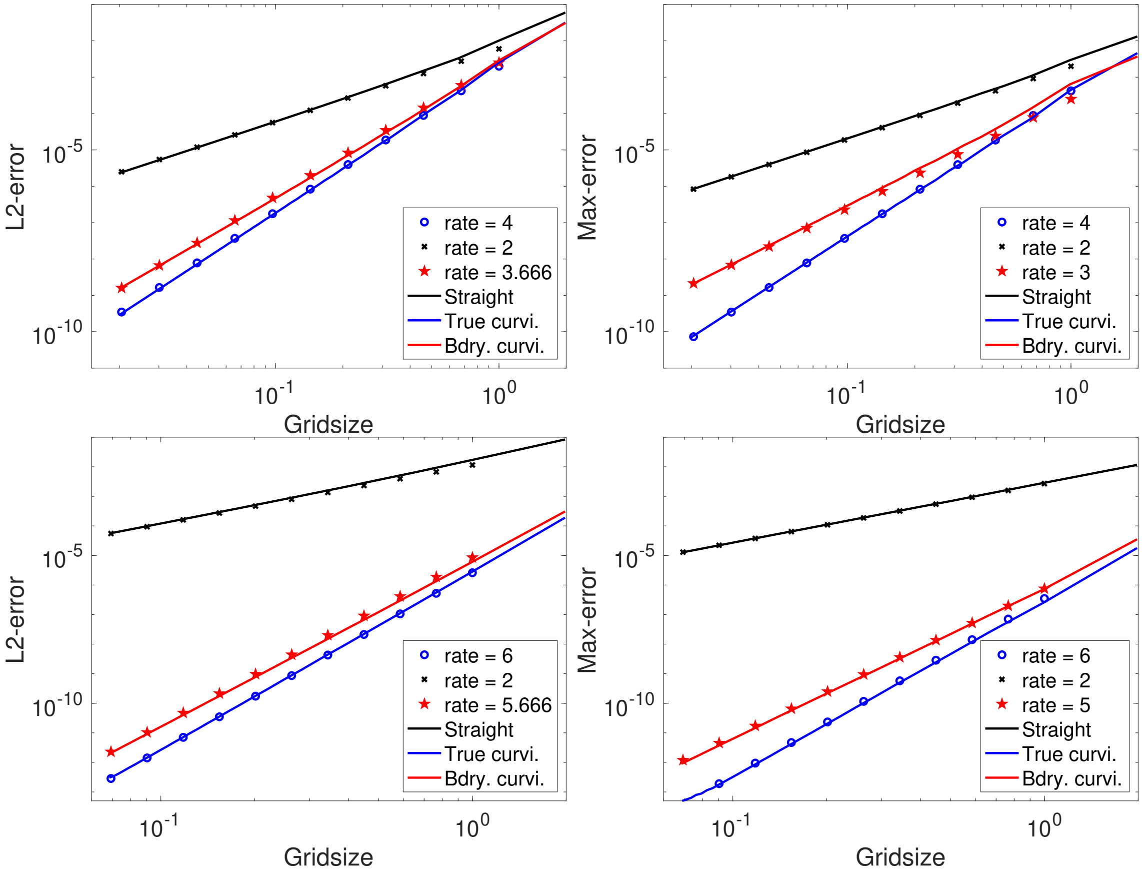


Figure 3.5. Neumann annular sector problem. Errors measured in the l^2 and the l^∞ norms for the three different sequences of grids. The top row is for $k = 2$ and the bottom row is for $k = 4$. These results are for the annular problem with homogenous Neumann conditions.

$e^{-it}e^{-\frac{x^2+y^2}{2}}$, $V(x,y) = \frac{x^2+y^2}{2}$, and $f(x,y,t) = 0$. The time step is set as $\Delta t = \frac{\Delta x^2}{500}$. Errors at time $T = 0.5$ are listed in Table 3.2, in which we observe order near $k+2$ for the ℓ^2-norm.

3.8 Concluding Remarks

We have proven that the Q^k ($k \geq 2$) spectral element method, when regarded as a finite difference scheme, is a $(k+2)$-th order accurate scheme in the discrete 2-norm for linear hyperbolic, parabolic and Schrödinger equations with Dirichlet boundary conditions, under smoothness assumptions of the exact solution and the differential operator coefficients. The same result holds for Neumann boundary conditions when there are no mixed second order

Table 3.1. A two-dimensional parabolic equation with Dirichlet boundary conditions.

Q^k polynomial	SEM Mesh	l^2 error	order	l^∞ error	order
k = 2	4 × 4	8.34E-3	-	4.57E-3	-
	8 × 8	6.59E-4	3.66	3.16E-4	3.85
	16 × 16	4.52E-5	3.86	2.36E-5	3.74
	32 × 32	2.91E-6	3.96	1.53E-6	3.94
k = 3	4 × 4	5.88E-4	-	1.71E-4	-
	8 × 8	2.24E-5	4.71	7.56E-6	4.50
	16 × 16	7.49E-7	4.90	2.52E-7	4.91
	32 × 32	2.38E-8	4.97	8.06E-9	4.96
k = 4	4 × 4	4.26E-5	-	1.16E-5	-
	8 × 8	7.62E-7	5.81	2.34E-7	5.63
	16 × 16	1.26E-8	5.92	4.12E-9	5.83
	32 × 32	2.00E-10	5.98	6.68E-11	5.95

Table 3.2. A two-dimensional linear Schrödinger equation with Dirichlet boundary conditions.

Q^k polynomial	SEM Mesh	l^2 error	order	l^∞ error	order
k = 2	4 × 4	9.98E-4	-	6.36E-4	-
	8 × 8	6.65E-5	3.91	4.01E-5	3.99
	16 × 16	4.10E-6	4.02	2.77E-6	3.85
	32 × 32	2.53E-7	4.02	1.79E-7	3.89
k = 3	4 × 4	4.06E-5	-	2.12E-5	-
	8 × 8	1.12E-6	5.18	5.56E-7	5.26
	16 × 16	3.22E-8	5.12	1.75E-8	4.99
	32 × 32	1.05E-9	4.94	5.33E-10	5.04
k = 4	4 × 4	1.61E-6	-	5.86E-7	-
	8 × 8	2.65E-8	5.92	9.93E-9	5.88
	16 × 16	3.95E-10	6.07	1.66E-10	5.90
	32 × 32	5.30E-12	6.22	2.66E-12	5.97

derivatives. This explains the observed order of accuracy when the errors of the spectral element method are only measured at nodes of degree of freedoms.

4. SUPERCONVERGENCE OF C^0-Q^k FINITE ELEMENT METHOD FOR ELLIPTIC EQUATIONS WITH APPROXIMATED COEFFICIENTS

4.1 Introduction

In this chapter, we prove that the superconvergence of C^0-Q^k finite element method at the Gauss Lobatto quadrature points still holds if variable coefficients in an elliptic problem are replaced by their piecewise Q^k Lagrange interpolants at the Gauss Lobatto points in each rectangular cell.

4.1.1 Motivations

Consider solving a variable coefficient Poisson equation

$$-\nabla \cdot (a\nabla u) = f, \quad a(x,y) > 0 \tag{4.1}$$

with homogeneous Dirichlet boundary conditions on a rectangular domain Ω as in Chapter 2. The variational form is to find $u \in H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ satisfying

$$A(u,v) = (f,v), \quad \forall v \in H_0^1(\Omega), \tag{4.2}$$

where $A(u,v) = \iint_\Omega a\nabla u \cdot \nabla v dxdy$, $(f,v) = \iint_\Omega f v dxdy$. Consider a rectangular mesh with mesh size h, the C^0-Q^k finite element solution of (4.2) is defined as $u_h \in V_0^h$ satisfying

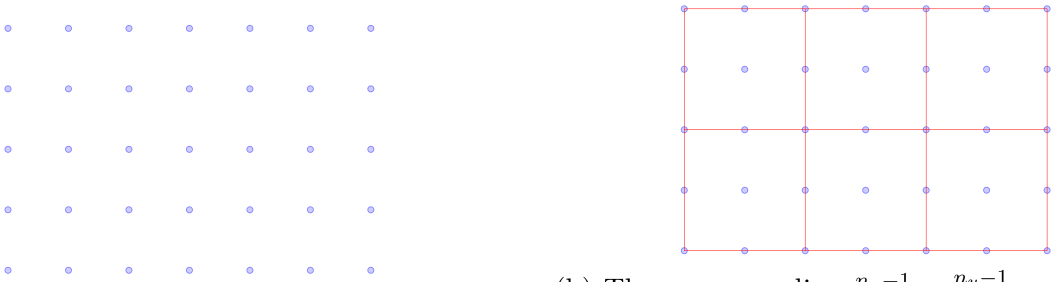
$$A(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_0^h. \tag{4.3}$$

For implementing finite element method (4.3), either some quadrature is used or the coefficient $a(x,y)$ is approximated by polynomials for computing $\iint_\Omega a u_h v_h \, dxdy$. In this chapter, we consider the implementation to approximate the smooth coefficient $a(x,y)$ by its Q^k Lagrangian interpolation polynomial in each cell. For instance, consider Q^2 element in two dimensions, tensor product of 3-point Lobatto quadrature form nine uniform points on each cell, see Figure 4.1. By point values of $a(x, y)$ at these nine points, we can obtain a Q^2 Lagrange interpolation polynomial on each cell. Let $a_I(x, y)$ and $f_I(x, y)$ denote the piecewise Q^k interpolation of $a(x, y)$ and $f(x, y)$ respectively. For a smooth functions $a \geq C > 0$, the interpolation error on each cell e is $\max_{\mathbf{x}\in e} |a_I(\mathbf{x}) - a(\mathbf{x})| = \mathcal{O}(h^{k+1})$ thus $a_I > 0$ if h is small enough. So if assuming the mesh is fine enough so that $a_I(x, y) \geq C > 0$, we consider the following scheme using the approximated coefficients $a_I(x, y)$: find $\tilde{u}_h \in V_0^h$ satisfying

$$A_I(\tilde{u}_h, v_h) := \iint_\Omega a_I \nabla \tilde{u} \cdot \nabla v dxdy = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h, \tag{4.4}$$

where $\langle f, v_h \rangle_h$ denotes using tensor product of $(k+1)$-point Gauss Lobatto quadrature for the integral (f, v_h). One can also simplify the computation of the right hand side by using $f_I(x, y)$, so we also consider the scheme to find $\tilde{u}_h$ satisfying

$$A_I(\tilde{u}_h, v_h) = (f_I, v_h), \quad \forall v_h \in V_0^h. \tag{4.5}$$



(a) A $n_x \times n_y$ finite difference grid

(b) The corresponding $\frac{n_x-1}{2} \times \frac{n_y-1}{2}$ mesh Ω_h for Q^2 element

Figure 4.1. An illustration of meshes.

The schemes (4.4) and (4.5) correspond to the equation

$$-\nabla \cdot (a_I(x, y) \nabla \tilde{u}(x, y)) = f(x, y). \tag{4.6}$$

At first glance, one might expect $(k+1)$-th order accuracy for a numerical method applying to (4.6) due to the interpolation error $a(x, y) - a_I(x, y) = \mathcal{O}(h^{k+1})$. But as we will show in Section 4.4.1, the difference between exact solutions u and $\tilde{u}$ to the two elliptic equations

(4.1) and (4.6) is $\mathcal{O}(h^{k+2})$ in $L^2(\Omega)$-norm under suitable assumptions. The main focus of this chapter is to show (4.4) and (4.5) are $(k+2)$-th order accurate finite difference type schemes via the superconvergence of finite element method. Such a result is very interesting from the perspective that a fourth order accurate scheme can be constructed even if the coefficients in the equation are approximated by quadratic polynomials, which does not seem to be considered before in the literature.

Since only grid point values of $a(x,y)$ and $f(x,y)$ are needed in scheme (4.4) or (4.5), they can be regarded as finite difference type schemes. Consider a uniform $n_x \times n_y$ grid for a rectangle Ω with grid points (x_i, y_j) and grid spacing h, where n_x and n_y are both odd numbers as shown in Figure 4.1(a). Then there is a mesh Ω_h of $(n_x-1)/2 \times (n_y-1)/2$ Q^2 elements so that Gauss-Lobatto points for all cells in Ω_h are exactly the finite difference grid points. By using the scheme (4.4) or (4.5) on the finite element mesh Ω_h shown in Figure 4.1(b), we obtain a fourth order finite difference scheme in the sense that $\tilde{u}_h$ is fourth order accurate in the discrete 2-norm at all grid points.

In practice the most convenient implementation is to use tensor product of $(k+1)$-point Gauss Lobatto quadrature for integrals in (4.2) i.e. scheme (2.2) as described in Chapter 2. Numerical tests suggest that the approximated coefficient scheme (4.5) is more accurate and robust than the quadrature scheme (2.2) in some cases.

4.2 Notations And Preliminaries

We will continue to use the notations in Section 3.2 of Chapter 2. Besides the error estimate in Section 2.3, we also need the following estimate.

Theorem 4.2.1. *For $k \geq 2$, $(f, v_h) - (f_I, v_h) = \mathcal{O}(h^{k+2})\|f\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V^h$.*

Proof. Repeat the proof of Theorem 2.3.4 for the function $f - f_I$ on a cell e, with the fact $[f_I]_{k+1,p,e} = [f_I]_{k+2,p,e} = 0$, we get

$$E[(f-f_I)v_h] = Ch^{k+2}([f]_{k+2,e}|v_h|_{0,e} + [f]_{k+1,e}|v_h|_{1,e} + [f-f_I]_{k,e}|v_h|_{2,e}).$$

By (2.13) on the Lagrange interpolation operator and the fact $[f-f_I]_{k,e} \leq \|f-f_I\|_{k+1,e}$, we get $[f-f_I]_{k,e} \leq Ch[f]_{k+1,e}$. Notice that $\langle f-f_I, v_h\rangle_h = 0$, with (2.5), we get

$$(f,v_h)-(f_I,v_h) = (f-f_I,v_h) - \langle f-f_I, v_h\rangle_h = \mathcal{O}(h^{k+2})\|f\|_{k+2}\|v_h\|_2, \forall v_h \in V^h.$$

□

4.3 Superconvergence Of The Bilinear Form

For convenience, in this subsection, we drop the subscript h in a test function $v_h \in V^h$. When there is no confusion, we may also drop $dxdy$ or $dsdt$ in a double integral.

Lemma 4.3.1. *Assume $a(x,y) \in W^{2,\infty}(\Omega)$. For $k \geq 2$,*

$$\iint_\Omega a(u-u_p)_x v_x \, dxdy = \mathcal{O}(h^{k+2})\|u\|_{k+2}\|v\|_2, \quad \forall v \in V^h.$$

Proof. For each cell e, we consider $\iint_e a(u-u_p)_x v_x \, dxdy$. Let $R[u]_{k,k} = u - u_p$ denote the M-type projection remainder on e. Then $R[u]_{k,k}$ can be splitted into lower order part $R[u]_{k,k} - R[u]_{k+1,k+1}$ and high order part $R[u]_{k+1,k+1}$.

$$\iint_e a(u-u_p)_x v_x \, dxdy = \iint_e a(R[u]_{k+1,k+1})_x v_x \, + \iint_e a(R[u]_{k,k} - R[u]_{k+1,k+1})_x v_x.$$

We first consider the high order part. Mapping everything to the reference cell $\hat{K}$ and let $\overline{\hat{a}\hat{v}_s}$ denote the average of $\hat{a}\hat{v}_s$ on $\hat{K}$. By the last property in Lemma 2.4.3, we get

$$\begin{aligned} &h^{2-n}\left|\iint_e a(R[u]_{k+1,k+1})_x v_x \, dxdy\right| = \left|\iint_{\hat{K}} \partial_s(\hat{R}[\hat{u}]_{k+1,k+1})\hat{a}\hat{v}_s dsdt\right| \\ &= \left|\iint_{\hat{K}} \partial_s(\hat{R}[\hat{u}]_{k+1,k+1})(\overline{\hat{a}\hat{v}_s} - \hat{a}\hat{v}_s) dsdt\right| \leq |\partial_s(\hat{R}[\hat{u}]_{k+1,k+1})|_{0,2,\hat{K}} |\overline{\hat{a}\hat{v}_s} - \hat{a}\hat{v}_s|_{0,2,\hat{K}}. \end{aligned}$$

By Poincaré inequality and Cauchy-Schwarz inequality, we have

$$|\overline{\hat{a}\hat{v}_s} - \hat{a}\hat{v}_s|_{0,2,\hat{K}} \leq C|\nabla(\hat{a}\hat{v}_s)|_{0,2,\hat{K}} \leq C|\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{1,2,\hat{K}} + C|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}|_{2,2,\hat{K}}.$$

Mapping back to the cell e, with (2.5), by Lemma 2.4.3, the higher order part is bounded by $Ch^{k+2}[u]_{k+2,2,e}(|a|_{1,\infty,e}|v|_{1,2,e} + |a|_{0,\infty,e}|v|_{2,2,e})$ thus

$$\begin{aligned}\sum_e \iint_e a(R[u]_{k+1,k+1})_x v_x \, dxdy &= \mathcal{O}(h^{k+2})\|a\|_{1,\infty,\Omega} \sum_e \|u\|_{k+2,e}\|v\|_{2,e} \\ &= \mathcal{O}(h^{k+2})\|a\|_{1,\infty,\Omega}\|u\|_{k+2,\Omega}\|v\|_{2,\Omega}.\end{aligned}$$

Now we only need to discuss the lower order part of the remainder. Let $R[u]_{k,k} - R[u]_{k+1,k+1} = R_1 + R_2$ which is defined similarly as in (2.28). For R_1, by the first two results in Lemma 2.4.4, we have

$$\begin{aligned}\iint_{\hat{K}} (\partial_s \hat{R}_1)\hat{a}\hat{v}_s = \iint_{\hat{K}} (\partial_s \hat{R}_1)(\hat{a}\hat{v}_s - \overline{\hat{a}\hat{v}_s}) &\leq |\partial_s \hat{R}_1|_{0,2,\hat{K}} |\overline{\hat{a}\hat{v}_s} - \hat{a}\hat{v}_s|_{0,2,\hat{K}} \\ &\leq C|\hat{u}|_{k+2,2,\hat{K}} |\overline{\hat{a}\hat{v}_s} - \hat{a}\hat{v}_s|_{0,2,\hat{K}}.\end{aligned}$$

By similar discussions above, we get

$$\sum_e \iint_e a(R_1)_x v_x \, dxdy = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,\Omega}\|u\|_{k+2,\Omega}\|v\|_{2,\Omega}.$$

For R_2, let $N(s)$ be the antiderivative of $M_{k+1}(s)$ then $N(\pm 1) = 0$. Let $\bar{\hat{a}}$ be the average of $\bar{\hat{a}}$ on $\hat{K}$ then $|\hat{a} - \bar{\hat{a}}|_{0,\infty,\hat{K}} \leq C|\hat{a}|_{1,\infty,\hat{K}}$. Since $M_{k+1}(s) \perp P^{k-2}(s)$, we have $\iint_{\hat{K}} \hat{b}_{k+1}(t)M_{k+1}(s)\hat{v}_{ss} = 0$. After integration by parts, by Lemma 2.4.4 we have

$$\begin{aligned}&\iint_{\hat{K}} (\partial_s \hat{R}_2)\hat{a}\hat{v}_s = -\iint_{\hat{K}} \hat{b}_{k+1}(t)M_{k+1}(s)(\hat{a}_s\hat{v}_s + \hat{a}\hat{v}_{ss}) \\ =&\iint_{\hat{K}} \hat{b}_{k+1}(t)N(s)(\hat{a}_{ss}\hat{v}_s + \hat{a}_s\hat{v}_{ss}) - \iint_{\hat{K}} \hat{b}_{k+1}(t)M_{k+1}(s)(\hat{a} - \bar{\hat{a}})\hat{v}_{ss} \\ \leq& C|\hat{u}|_{k+1,\hat{K}}(|\hat{a}|_{2,\infty,\hat{K}}|\hat{v}|_{1,2,\hat{K}} + |\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{2,2,\hat{K}}).\end{aligned}$$

Thus we can get

$$\sum_e \iint_e (\partial_x R_2)a\hat{v}_x dxdy = \mathcal{O}(h^{k+2})\|a\|_{2,\infty,\Omega}\|u\|_{k+1,\Omega}\|v\|_{2,\Omega}.$$

So we have $\iint_\Omega a(u - u_p)_x v_x \, dxdy = \mathcal{O}(h^{k+2})\|a\|_{2,\infty,\Omega}\|u\|_{k+2}\|v\|_2, \quad \forall v \in V^h.$ □

Lemma 4.3.2. *Assume* $c(x,y) \in W^{1,\infty}(\Omega)$*. For* $k \geq 2$*,*

$$\iint_\Omega c(u-u_p)v\,dxdy = \mathcal{O}(h^{k+2})\|u\|_{k+1}\|v\|_1, \quad \forall v \in V^h.$$

Proof. Let $\overline{\hat{c}\hat{v}}$ be the average of $\hat{c}\hat{v}$ on $\hat{K}$. Following similar arguments as in the proof Lemma 4.3.1,

$$\begin{aligned}\left|\iint_{\hat K} \hat{R}[\hat{u}]_{k,k}\hat{c}\hat{v}\right| &= \left|\iint_{\hat K} \hat{R}[\hat{u}]_{k,k}(\hat{c}\hat{v} - \overline{\hat{c}\hat{v}})\right| \leq |\hat{R}[\hat{u}]_{k,k}|_{0,2,\hat K}|\hat{c}\hat{v} - \overline{\hat{c}\hat{v}}|_{0,2,\hat K} \\ &\leq C[u]_{k+1,2,\hat K}[\hat{c}\hat{v}]_{1,2,\hat K} \leq C[u]_{k+1,2,\hat K}(|\hat{c}|_{0,\infty,\hat K}|\hat{v}|_{1,2,\hat K} + |\hat{c}|_{1,\infty,\hat K}|\hat{v}|_{0,2,\hat K}).\end{aligned}$$

So with (2.5) we have

$$\iint_e cR[u]_{k,k}vdxdy = h^n \iint_{\hat K}(R[\hat{u}]_{k,k})\hat{c}\hat{v}dsdt = \mathcal{O}(h^{k+2})\|c\|_{1,\infty,\Omega}\|u\|_{k+1,e}\|v\|_{1,e},$$

which implies the estimate. □

Lemma 4.3.3. *Assume* $b(x,y) \in W^{2,\infty}(\Omega)$*. For* $k \geq 2$*,*

$$\iint_\Omega b(u-u_p)_x v\,dxdy = \mathcal{O}(h^{k+2})\|u\|_{k+2}\|v\|_2, \quad \forall v \in V^h.$$

Proof. Let $\overline{\hat{b}\hat{v}}$ be the average of $\hat{b}\hat{v}$ on $\hat{K}$. Following similar arguments as in the proof Lemma 4.3.1, we have

$$\begin{aligned}&\left|\iint_{\hat K} \partial_s(\hat{R}[\hat{u}]_{k+1,k+1})\hat{b}\hat{v}\right| = \left|\iint_{\hat K} \partial_s(\hat{R}[\hat{u}]_{k+1,k+1})(\hat{b}\hat{v} - \overline{\hat{b}\hat{v}})\right| \\ &\leq |\partial_s(\hat{R}[\hat{u}]_{k+1,k+1})|_{0,2,\hat K}|\overline{\hat{b}\hat{v}} - \hat{b}\hat{v}|_{0,2,\hat K} \leq C[\hat{u}]_{k+2,2,\hat K}(|\hat{b}|_{1,\infty,\hat K}|\hat{v}|_{0,2,\hat K} + |\hat{b}|_{0,\infty,\hat K}|\hat{v}|_{1,2,\hat K}).\end{aligned}$$

$$\begin{aligned}\iint_{\hat K}(\partial_s\hat{R}_1)\hat{b}\hat{v} = \iint_{\hat K}(\partial_s\hat{R}_1)(\hat{b}\hat{v} - \overline{\hat{b}\hat{v}}) &\leq |\partial_s\hat{R}_1|_{0,2,\hat K}|\overline{\hat{b}\hat{v}} - \hat{b}\hat{v}|_{0,2,\hat K} \\ &\leq C|\hat{u}|_{k+2,2,\hat K}(|\hat{b}|_{1,\infty,\hat K}|\hat{v}|_{0,2,\hat K} + |\hat{b}|_{0,\infty,\hat K}|\hat{v}|_{1,2,\hat K}).\end{aligned}$$

Let $N(s)$ be the antiderivative of $M_{k+1}(s)$. After integration by parts, we have

$$
\begin{aligned}
&\iint_{\hat K} (\partial_s \hat R_2)\hat b \hat v = -\iint_{\hat K} \hat b_{k+1}(t) M_{k+1}(s)(\hat b_s \hat v + \hat b \hat v_s)\\
=&\iint_{\hat K} \hat b_{k+1}(t) N(s)(\hat b_{ss}\hat v + \hat b_s \hat v_s + \hat b \hat v_{ss})\\
\le& C|\hat u|_{k+1,2,\hat K}(|\hat b|_{2,\infty,\hat K}|\hat v|_{0,2,\hat K} + |\hat b|_{1,\infty,\hat K}|\hat v|_{1,2,\hat K} + |\hat b|_{0,\infty,\hat K}|\hat v|_{2,2,\hat K}).
\end{aligned}
$$

After combining all the estimates, with (2.5), we have

$$
\iint_e b(u-u_p)_x v \ = h^{n-1}\iint_{\hat K} \hat b (R[\hat u]_{k,k})_s \hat v \ = \mathcal{O}(h^{k+2})\|b\|_{2,\infty,\Omega}\|u\|_{k+2,e}\|v\|_{2,e}.
$$

□

Lemma 4.3.4. *Assume $a(x,y) \in W^{2,\infty}(\Omega)$. For $k \ge 2$,*

$$
\iint_\Omega a(u-u_p)_x (v_h)_y \, dxdy = \begin{cases} \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v_h\|_2, & \forall v_h \in V^h, \qquad (4.7a)\\ \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v_h\|_2, & \forall v_h \in V_0^h. \qquad (4.7b) \end{cases}
$$

Proof. Similar to the proof of Lemma 4.3.1, we have

$$
\begin{aligned}
&\left|\iint_e a(R[u]_{k+1,k+1})_x v_y \, dxdy\right| = h^{n-2}\left|\iint_{\hat K} \partial_s(\hat R[\hat u]_{k+1,k+1})\hat a \hat v_t ds dt\right|\\
=&h^{n-2}\left|\iint_{\hat K} \partial_s(\hat R[\hat u]_{k+1,k+1})(\overline{\hat a \hat v_t} - \hat a \hat v_t) ds dt\right| \le h^{n-2}|\partial_s(\hat R[\hat u]_{k+1,k+1})|_{0,2,\hat K}|\overline{\hat a \hat v_t} - \hat a \hat v_t|_{0,2,\hat K}\\
\le& C h^{k+2}\|a\|_{1,\infty,\Omega}\|u\|_{k+2,e}\|v\|_{2,e},
\end{aligned}
$$

and

$$
\iint_{\hat K} (\partial_s \hat R_1)\hat a \hat v_t = \iint_{\hat K} (\partial_s \hat R_1)(\hat a \hat v_t - \overline{\hat a \hat v_t}) \le |\partial_s \hat R_1|_{0,2,\hat K}|\overline{\hat a \hat v_t} - \hat a \hat v_t|_{0,2,\hat K}.
$$

Following the proof of Lemma 4.3.1, with (2.5), we get

$$
\sum_e \iint_e a(R_1)_x v_y \, dxdy = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,\Omega}\|u\|_{k+2,\Omega}\|v\|_{2,\Omega}.
$$

After integration by parts, we have

$$\iint_{\hat{K}} (\partial_s \hat{R}_2)\hat{a}\hat{v}_t = -\iint_{\hat{K}} \hat{b}_{k+1}(t)M_{k+1}(s)(\hat{a}_s\hat{v}_t + \hat{a}\hat{v}_{st}),$$

which is exactly the same integral estimates (2.41) in the proof of Lemma 2.5.5 in Chapter 2. By the same proof as (2.41), after combining all the estimates above and summing over all elements, we have the estimate for the term $\iint_{\hat{K}} l_k(s)\hat{b}_{k+1}(t)\hat{a}\hat{v}_t dsdt$:

$$\sum_e |\iint_{\hat{K}} (\partial_s \hat{R}_2)\hat{a}\hat{v}_t| = \begin{cases} \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2, & \forall v \in V^h, \\ \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2, & \forall v \in V_0^h. \end{cases}$$

Combine all the estimates above, we get (4.7a). Since the $\frac{1}{2}$ order loss is only due to the line integral along L_1 and L_3, on which $v_{xx} = 0$ if $v \in V_0^h$, we get (4.7b). □

4.4 The Main Result

4.4.1 Superconvergence of bilinear forms with approximated coefficients

Even though standard interpolation error is $a - a_I = \mathcal{O}(h^{k+1})$, as shown in the following discussion, the error in the bilinear forms is related to $\iint_e (a - a_I)\, dxdy$ on each cell e, which is the quadrature error thus the order is higher. We have the following estimate on the bilinear forms with approximated coefficients:

Lemma 4.4.1. *Assume $a(x,y) \in W^{k+2,\infty}(\Omega)$ and $u(x,y) \in H^2(\Omega)$, then $\forall v \in V^h$ or $\forall v \in H^2(\Omega)$,*

$$\begin{aligned}
\iint_\Omega a u_x v_x \, dxdy - \iint_\Omega a_I u_x v_x \, dxdy &= \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|v\|_2, \\
\iint_\Omega a u_x v_y \, dxdy - \iint_\Omega a_I u_x v_y \, dxdy &= \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|v\|_2, \\
\iint_\Omega a u_x v \, dxdy - \iint_\Omega a_I u_x v \, dxdy &= \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|v\|_1, \\
\iint_\Omega a u v \, dxdy - \iint_\Omega a_I u v \, dxdy &= \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_1\|v\|_1.
\end{aligned}$$

Proof. For every cell e in the mesh Ω_h, let $\overline{u_x v_x}$ be the cell average of $u_x v_x$. By Theorem 2.3.2 and Theorem 2.3.3 , we have

$$
\begin{aligned}
&\iint_e (a_I - a) u_x v_x \\
=&\iint_e (a_I - a)\overline{u_x v_x} + \iint_e (a_I - a)(u_x v_x - \overline{u_x v_x}) \\
=&\frac{1}{4h^2}\iint_e (a_I - a)\iint_e u_x v_x + \iint_e (a_I - a)(u_x v_x - \overline{u_x v_x}) \\
=&\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_{1,e}\|v\|_{1,e} + \mathcal{O}(h^{k+1})\|a\|_{k+1,\infty,\Omega}\iint_e |u_x v_x - \overline{u_x v_x}|.
\end{aligned}
$$

By Poincaré inequality and Cauchy-Schwarz inequality, we have

$$\iint_e |u_x v_x - \overline{u_x v_x}| = \mathcal{O}(h)\|\nabla(u_x v_x)\|_{0,1,e} = \mathcal{O}(h)\|u\|_{2,e}\|v\|_{2,e}$$

thus $\iint_e (a_I - a) u_x v_x = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_{2,e}\|v\|_{2,e}$. Summing over all elements e, we have $\iint_\Omega (a_I - a) u_x v_x = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|v\|_2$. Similarly we can establish the other three estimates. □

Lemma 4.4.1 implies that the difference in the solutions to (4.6) and (4.1) is $\mathcal{O}(h^{k+2})$ in the $L^2(\Omega)$-norm:

Theorem 4.4.2. *Assume $a(x,y) \in W^{k+2,\infty}(\Omega)$ and $a_I(x,y) \geq C > 0$. Let $u, \tilde{u} \in H_0^1(\Omega)$ be the solutions to*

$$A(u,v) := \iint a\nabla u \cdot \nabla v \, dxdy = (f,v), \quad \forall v \in H_0^1(\Omega)$$

and

$$A_I(\tilde{u},v) := \iint a_I\nabla \tilde{u} \cdot \nabla v \, dxdy = (f,v), \quad \forall v \in H_0^1(\Omega)$$

respectively, where $f \in L^2(\Omega)$. Then $\|u - \tilde{u}\|_0 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|f\|_0$.

Proof. By Lemma 4.4.1, for any $v \in H^2(\Omega)$ we have

$$\begin{aligned} A_I(u-\tilde{u},v) = A_I(u,v) - A_I(\tilde{u},v) &= [A_I(u,v) - A(u,v)] + [A(u,v) - A_I(\tilde{u},v)] \\ &= A_I(u,v) - A(u,v) = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|v\|_2. \end{aligned}$$

Let $w \in H_0^1(\Omega)$ be the solution to the dual problem

$$A_I(v,w) = (u-\tilde{u},v) \quad \forall v \in H_0^1(\Omega).$$

Since $a_I \geq C > 0$ and $|a_I(x,y)| \leq C|a(x,y)|$, the coercivity and boundedness of the bilinear form A_I hold [3]. Moreover, a_I is Lipschitz continuous because $a(x,y) \in W^{k+2,\infty}(\Omega)$. Thus the solution w exists and the elliptic regularity $\|w\|_2 \leq C\|u-\tilde{u}\|_0$ holds on a convex domain, e.g., a rectangular domain Ω, see [34]. Thus,

$$\|u-\tilde{u}\|_0^2 = (u-\tilde{u},u-\tilde{u}) = A_I(u-\tilde{u},w) = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|u\|_2\|w\|_2.$$

With elliptic regularity $\|w\|_2 \leq C\|u-\tilde{u}\|_0$ and $\|u\|_2 \leq C\|f\|_0$, we get

$$\|u-\tilde{u}\|_0 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty,\Omega}\|f\|_0.$$

□

Remark 4.4.3. For even number $k \geq 4$, $(k+1)$-point Newton-Cotes quadrature rule has the same error order as the $(k+1)$-point Gauss-Lobatto quadrature rule. Thus Theorem 4.4.2 still holds if we redefine $a_I(x,y)$ as the Q^k interpolant of $a(x,y)$ at the uniform $(k+1)\times(k+1)$ Newton-Cotes points in each cell if $k \geq 4$ is even.

4.4.2 The variable coefficient Poisson equation

Let $u(x,y) \in H_0^1(\Omega)$ be the exact solution to

$$A(u,v) := \iint_\Omega a\nabla u \cdot \nabla v \, dxdy = (f,v), \quad \forall v \in H_0^1(\Omega).$$

Let $\tilde{u}_h \in V_0^h(\Omega)$ be the solution to

$$A_I(\tilde{u}_h, v_h) := \iint_\Omega a_I \nabla \tilde{u}_h \cdot \nabla v_h \, dxdy = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h(\Omega).$$

Theorem 4.4.4. *For $k \geq 2$, let u_p be the piecewise Q^k M-type projection of $u(x,y)$ on each cell e in the mesh Ω_h. Assume $a \in W^{k+2,\infty}(\Omega)$ and $u, f \in H^{k+2}(\Omega)$, then*

$$A_I(\tilde{u}_h - u_p, v_h) = \mathcal{O}(h^{k+2})(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2})\|v_h\|_2, \quad \forall v_h \in V_0^h.$$

Proof. For any $v_h \in V^h$, we have

$$\begin{aligned}
&A_I(\tilde{u}_h, v_h) - A_I(u_p, v_h)\\
=&(f, v_h) - A_I(u_p, v_h) + \langle f, v_h \rangle_h - (f, v_h)\\
=&A(u, v_h) - A_I(u_p, v_h) + \langle f, v_h \rangle_h - (f, v_h)\\
=&[A(u, v_h) - A_I(u, v_h)] + [A_I(u - u_p, v_h) - A(u - u_p, v_h)] + A(u - u_p, v_h) + \langle f, v_h \rangle_h - (f, v_h).
\end{aligned}$$

Lemma 4.4.1 implies $A(u, v_h) - A_I(u, v_h) = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_2\|v_h\|_2$. Theorem 2.3.4 gives $\langle f, v_h \rangle_h - (f, v_h) = \mathcal{O}(h^{k+2})\|f\|_{k+2}\|v_h\|_2$. By Lemma 4.3.1, $A(u - u_p, v_h) = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2}\|v_h\|_2$.

For the second term $A_I(u - u_p, v_h) - A(u - u_p, v_h) = \iint_\Omega (a - a_I)\nabla(u - u_p)\nabla v_h$, by Theorem 2.3.2 and Lemma 2.4.3, we have

$$\begin{aligned}
\left| \iint_\Omega (a - a_I)(u - u_p)_x \partial_x v_h \right| &\leq |a - a_I|_{0,\infty,\Omega} \sum_e \iint_e |(u - u_p)_x \partial_x v_h|\\
&\leq |a - a_I|_{0,\infty,\Omega} \sum_e |(u - u_p)_x|_{0,2,e} |v_h|_{1,2,e}\\
&= \mathcal{O}(h^{2k+1})\|a\|_{k+1,\infty,\Omega} \sum_e \|u\|_{k+1,e}\|v_h\|_{1,e}\\
&= \mathcal{O}(h^{2k+1})\|a\|_{k+1,\infty,\Omega}\|u\|_{k+1}\|v_h\|_1.
\end{aligned}$$

□

Theorem 4.4.5. *Assume $a(x,y) \in W^{k+2,\infty}(\Omega)$ is positive and $u(x,y), f(x,y) \in H^{k+2}(\Omega)$. Assume the mesh is fine enough so that the piecewise Q^k interpolant satisfies $a_I(x,y) \geq C >$*

0. *Then $\tilde{u}_h$ is a $(k+2)$-th order accurate approximation to u in the discrete 2-norm over all the $(k+1)\times(k+1)$ Gauss-Lobatto points:*

$$\|\tilde{u}_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2}).$$

Proof. Let $\theta_h = \tilde{u}_h - u_p$. By the definition of u_p and Theorem 2.4.1, it is straightforward to show $\theta_h = 0$ on $\partial\Omega$. By the Aubin-Nitsche duality method, let $w \in H_0^1(\Omega)$ be the solution to the dual problem

$$A_I(v, w) = (\theta_h, v) \quad \forall v \in H_0^1(\Omega).$$

By the same discussion as in the proof of Theorem 4.4.2, the solution w exists and the regularity $\|w\|_2 \leq C\|\theta_h\|_0$ holds.

Let w_h be the finite element projection of w, i.e., $w_h \in V_0^h$ satisfies

$$A_I(v_h, w_h) = (\theta_h, v_h) \quad \forall v_h \in V_0^h.$$

Since $w_h \in V_0^h$, by Theorem 4.4.4, we have

$$\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_I(\theta_h, w_h) = \mathcal{O}(h^4)(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2})\|w_h\|_2. \tag{4.8}$$

Let $w_I = \Pi_1 w$ be the piecewise Q^1 projection of w on Ω_h as defined in (2.8). By the Bramble-Hilbert Lemma, we get $\|w - w_I\|_{2,e} \leq C[w]_{2,e} \leq C\|w\|_{2,e}$ thus

$$\|w - w_I\|_2 \leq C\|w\|_2.$$

By the inverse estimate on the piecewise polynomial $w_h - w_I$, we have

$$\|w_h\|_2 \leq \|w_h - w_I\|_2 + \|w_I - w\|_2 + \|w\|_2 \leq Ch^{-1}\|w_h - w_I\|_1 + C\|w\|_2. \tag{4.9}$$

With coercivity, Galerkin orthogonality and Cauchy Schwarz inequality, we get

$$C\|w_h - w_I\|_1^2 \leq A_I(w_h - w_I, w_h - w_I) = A_I(w_h - w_I, w - w_I) \leq C\|w - w_I\|_1 \|w_h - w_I\|_1,$$

which implies

$$\|w_h - w_I\|_1 \leq C\|w - w_I\|_1 \leq Ch\|w\|_2. \tag{4.10}$$

With (4.9), (4.10) and the elliptic regularity $\|w\|_2 \leq C\|\theta_h\|_0$, we get

$$\|w_h\|_2 \leq C\|w\|_2 \leq C\|\theta_h\|_0. \tag{4.11}$$

By (4.8) and (4.11) we have

$$\|\theta_h\|_0^2 \leq \mathcal{O}(h^{k+2})(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2})\|\theta_h\|_0,$$

i.e.,

$$\|\tilde{u}_h - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2}).$$

Finally, by the equivalency between the discrete 2-norm on Z_0 and the $L^2(\Omega)$ norm in the space V^h, with Theorem 2.4.5, we obtain

$$\|\tilde{u}_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|a\|_{k+2,\infty}\|u\|_{k+2} + \|f\|_{k+2}).$$

□

Remark 4.4.6. To extend Theorem 4.4.5 to homogeneous Neumann boundary conditions or mixed homogeneous Dirichlet and Neumann boundary conditions, dual problems with the same homogeneous boundary conditions as in primal problems should be used. Then all the estimates such as Theorem 4.4.4 hold not only for $v \in V_0^h$ but also for any v in V^h.

Remark 4.4.7. With Theorem 4.2.1, all the results hold for the scheme (4.5).

Remark 4.4.8. It is straightforward to verify that all results hold in three dimensions. Notice that the in three dimensions the discrete 2-norm is

$$\|u\|_{2,Z_0} = \left[h^3 \sum_{\mathbf{x} \in Z_0} |u(\mathbf{x})|^2 \right]^{\frac{1}{2}}.$$

Remark 4.4.9. For discussing superconvergence of the scheme (2.2), we have to consider the dual problem of the bilinear form A instead and the exact Galerkin orthogonality in (2.2) no longer holds. In order for the proof above holds, we need to show the Galerkin orthogonality in (2.2) holds up to $\mathcal{O}(h^{k+2})\|v_h\|_2$ for a test function $v_h \in V_h$, which is very difficult to establish. This is the main difficulty to extend the proof of Theorem 4.4.5 to the Gauss Lobatto quadrature scheme (2.2), which will be analyzed in next section by different techniques.

4.4.3 General elliptic problems

In this section, we discuss extensions to more general elliptic problems. Consider an elliptic variational problem of finding $u \in H_0^1(\Omega)$ to satisfy

$$A(u,v) := \iint_\Omega (\nabla v^T \mathbf{a} \nabla u + \mathbf{b} \nabla u v + cuv)\, dxdy = (f,v), \forall v \in H_0^1(\Omega),$$

where $\mathbf{a}(x,y) = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is positive definite and $\mathbf{b} = [b_1 \quad b_2]$. Assume the coefficients $\mathbf{a}$, $\mathbf{b}$ and c are smooth, and $A(u,v)$ satisfies coercivity $A(v,v) \geq C\|v\|_1$ and boundedness $|A(u,v)| \leq C\|u\|_1\|v\|_1$ for any $u, v \in H_0^1(\Omega)$.

By the estimates in Section 4.3, we first have the following estimate on the Q^k M-type projection u_p:

Lemma 4.4.10. *Assume $a_{ij}(x,y), b_i(x,y) \in W^{2,\infty}(\Omega)$ and $b_i(x,y) \in W^{2,\infty}(\Omega)$, then*

$$A(u-u_p, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u\|_{k+2}\|v_h\|_2, & \forall v_h \in V_0^h, \\ \mathcal{O}(h^{k+1.5})\|u\|_{k+2}\|v_h\|_2, & \forall v_h \in V^h. \end{cases}$$

If $a_{12} = a_{21} \equiv 0$, *then*

$$A(u - u_p, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Let $\mathbf{a}_I$, b_I and c_I denote the corresponding piecewise Q^k Lagrange interpolation at Gauss-Lobatto points. We are interested in the solution $\tilde{u}_h \in V_0^h$ to

$$A_I(\tilde{u}_h, v_h) := \iint_\Omega (\nabla v_h^T \mathbf{a}_I \nabla \tilde{u}_h + \mathbf{b}_I \nabla \tilde{u}_h v_h + c_I \tilde{u}_h v_h)\, dxdy = \langle f, v_h \rangle_h, \forall v_h \in V_0^h.$$

We need to assume that A_I still satisfies coercivity $A_I(v, v) \geq C\|v\|_1$ and boundedness $|A_I(u, v)| \leq C\|u\|_1\|v\|_1$ for any $u, v \in H_0^1(\Omega)$, so that the solution $u \in H_0^1(\Omega)$ of the following problem exists and is unique:

$$A_I(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega).$$

We also need the elliptic regularity to hold for the dual problem:

$$A_I(v, w) = (f, v), \quad \forall v \in H_0^1(\Omega).$$

For instance, if $\mathbf{b} \equiv 0$, it suffices to require that eigenvalues of $\mathbf{a}_I + c_I \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ has a uniform positive lower bound on Ω, which is achievable on fine enough meshes if $\mathbf{a} + c \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ are positive definite. This implies the coercivity of A_I. The boundedness of A_I follows from the smoothness of coefficients. Since $\mathbf{a}_I$ and c_I are Lipschitz continuous, the elliptic regularity for A_I holds on a convex domain [34].

By Lemma 4.4.1 and Lemma 4.4.10, it is straightforward to extend Theorem 4.4.4 to the general elliptic case:

Theorem 4.4.11. *For $k \geq 2$, assume $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ and $u, f \in H^{k+2}(\Omega)$, then*

$$A_I(\tilde{u}_h - u_p, v_h) = \begin{cases} \mathcal{O}(h^{k+2})(\|u\|_{k+2} + \|f\|_{k+2})\|v_h\|_2, & \forall v_h \in V_0^h, \\ \mathcal{O}(h^{k+1.5})(\|u\|_{k+2} + \|f\|_{k+2})\|v_h\|_2, & \forall v_h \in V^h, \end{cases}.$$

And if $a_{12} = a_{21} \equiv 0$, then

$$A_I(\tilde{u}_h - u_p, v_h) = \mathcal{O}(h^{k+2})(\|u\|_{k+2} + \|f\|_{k+2})\|v_h\|_2, \quad \forall v_h \in V^h.$$

With suitable assumptions, it is straightforward to extend the proof of Theorem 4.4.5 to the general case:

Theorem 4.4.12. *For $k \geq 2$, assume $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ and $u, f \in H^{k+2}(\Omega)$, Assume the approximated bilinear form A_I satisfies coercivity and boundedness and the elliptic regularity still holds for the dual problem of A_I. Then $\tilde{u}_h$ is a $(k+2)$-th order accurate approximation to u in the discrete 2-norm over all the $(k+1) \times (k+1)$ Gauss-Lobatto points:*

$$\|\tilde{u}_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|u\|_{k+2} + \|f\|_{k+2}).$$

Remark 4.4.13. With Neumann type boundary conditions, due to Lemma 4.3.4, we can only prove $(k+1.5)$-th order accuracy

$$\|\tilde{u}_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+1.5})(\|u\|_{k+2} + \|f\|_{k+2}),$$

unless there are no mixed second order derivatives in the elliptic equation, i.e., $a_{12} = a_{21} \equiv 0$. We emphasize that even for the full finite element scheme (4.3), only $(k+1.5)$-th order accuracy at all Lobatto points can be proven for a general elliptic equation with Neumann type boundary conditions.

4.5 Numerical Results

In this section we show some numerical tests of C^0-Q^2 finite element method on an uniform rectangular mesh and verify the order of accuracy at Z_0, i.e., all Gauss-Lobatto points. The following four schemes will be considered:

1. Full Q^2 finite element scheme (4.3) where integrals in the bilinear form are approximated by 5×5 Gauss quadrature rule, which is exact for Q^9 polynomials thus exact for $A(u_h, v_h)$ if the variable coefficient is a Q^5 polynomial.

2. The Gauss Lobatto quadrature scheme (2.2): all integrals are approximated by 3×3 Gauss Lobatto quadrature.

3. The schemes (4.4) and (4.5).

The last three schemes are finite difference type since only grid point values of the coefficients are needed. In (4.4) and (4.5), $A_I(u_h, v_h)$ can be exactly computed by 4×4 Gauss quadrature rule since coefficients are Q^2 polynomials. An alternative finite difference type implementation of (4.4) and (4.5) is to precompute integrals of Lagrange basis functions and their derivatives to form a sparse tensor, then multiply the tensor to the vector consisting of point values of the coefficient to form the stiffness matrix. With either implementation, computational cost to assemble stiffness matrices in schemes (4.4) and (4.5) is higher than the stiffness matrix assembling in the simpler scheme (2.2) since the Lagrangian Q^k basis are delta functions at Gauss-Lobatto points.

4.5.1 Accuracy

We consider the following example with either purely Dirichlet or purely Neumann boundary conditions:

$$\nabla \cdot (a \nabla u) = f \quad \text{on } [0,1] \times [0,2]$$

where $a(x,y) = 1 + 0.1x^3y^5 + \cos(x^3y^2 + 1)$ and $u(x,y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$. The nonhomogeneous boundary condition should be computed in a way consistent with the computation of integrals in the bilinear form.

Table 4.1. The errors of C^0-Q^2 for a Poisson equation with Dirichlet boundary conditions at Lobatto points.

	FEM with Approximated Coefficients (4.4)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.22E-1	-	3.96E-1	-
4×8	4.83E-2	2.20	1.51E-1	1.39
8×16	2.54E-3	4.25	1.16E-2	3.71
16×32	1.49E-4	4.09	7.52E-4	3.95
32×64	9.22E-6	4.01	5.14E-5	3.87
	FEM using Gauss Lobatto Quadrature (2.2)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.24E-1	-	4.30E-1	-
4×8	4.43E-2	2.34	1.37E-1	1.65
8×16	2.27E-3	4.29	8.61E-3	4.00
16×32	1.32E-4	4.11	4.87E-4	4.14
32×64	8.13E-6	4.02	3.09E-5	3.97
	FEM with Approximated Coefficients (4.5)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.78E-1	-	6.31E-1	-
4×8	2.76E-2	3.33	8.69E-2	2.86
8×16	1.28E-3	4.43	3.77E-3	4.53
16×32	8.96E-5	3.83	3.36E-4	3.49
32×64	5.79E-6	3.95	2.41E-5	3.80
	Full FEM Scheme			
Mesh	l^2 error	order	l^∞ error	order
2×4	1.48E-2	-	3.79E-2	-
4×8	1.05E-2	0.50	3.76E-2	0.01
8×16	7.32E-4	3.84	4.04E-3	3.22
16×32	4.54E-5	4.01	2.83E-4	3.83
32×64	2.85E-6	3.99	1.75E-5	4.02

Table 4.2. The errors of C^0-Q^2 for a Poisson equation with Neumann boundary conditions at Lobatto points.

	FEM with Approximated Coefficients (4.4)			
Mesh	l^2 error	order	l^∞ error	order
2×4	3.44E0	-	5.39E0	-
4×8	1.83E-1	4.23	3.51E-1	3.93
8×16	1.38E-2	3.73	3.43E-2	3.36
16×32	8.37E-4	4.04	2.21E-3	3.96
32×64	5.13E-5	4.03	1.41E-4	3.96
	FEM using Gauss Lobatto Quadrature (2.2)			
Mesh	l^2 error	order	l^∞ error	order
2×4	3.43E0	-	4.95E0	-
4×8	1.81E-1	4.25	3.11E-1	3.99
8×16	1.37E-2	3.72	2.81E-2	3.47
16×32	8.33E-4	4.04	1.76E-3	4.00
32×64	5.11E-5	4.03	1.12E-4	3.97
	FEM with Approximated Coefficients (4.5)			
Mesh	l^2 error	order	l^∞ error	order
2×4	3.64E0	-	5.06E0	-
4×8	1.60E-1	4.51	2.54E-1	4.32
8×16	1.26E-2	3.67	2.39E-2	3.41
16×32	7.67E-4	4.03	1.67E-3	3.84
32×64	4.71E-5	4.03	1.09E-4	3.94
	Full FEM Scheme			
Mesh	l^2 error	order	l^∞ error	order
2×4	8.45E-2	-	2.13E-1	-
4×8	1.56E-2	2.43	5.66E-2	1.91
8×16	9.12E-4	4.10	5.14E-3	3.46
16×32	5.47E-5	4.06	3.24E-4	3.99
32×64	3.37E-6	4.02	2.22E-5	3.87

Table 4.3. An elliptic equation with mixed second order derivatives and Neumann boundary conditions.

	FEM with Approximated Coefficients (4.4)			
Mesh	l^2 error	order	l^∞ error	order
2×4	1.92E0	-	3.47E0	-
4×8	2.16E-1	3.15	6.05E-1	2.52
8×16	1.45E-2	3.90	6.12E-2	3.30
16×32	9.08E-4	4.00	4.05E-3	3.92
32×64	5.66E-5	4.00	2.76E-4	3.88
	FEM using Gauss Lobatto Quadrature (2.2)			
Mesh	l^2 error	order	l^∞ error	order
2×4	1.38E0	-	2.27E0	-
4×8	1.46E-1	3.24	2.52E-1	3.17
8×16	7.49E-3	4.28	1.64E-2	3.94
16×32	4.31E-4	4.12	1.02E-3	4.01
32×64	2.61E-5	4.04	7.47E-5	3.78
	FEM with Approximated Coefficients (4.5)			
Mesh	l^2 error	order	l^∞ error	order
2×4	1.89E0	-	2.84E0	-
4×8	1.04E-1	4.18	1.45E-1	4.30
8×16	5.62E-3	4.21	1.86E-2	2.96
16×32	3.24E-4	4.12	1.67E-3	3.48
32×64	1.95E-5	4.05	1.32E-4	3.66
	Full FEM Scheme			
Mesh	l^2 error	order	l^∞ error	order
2×4	1.46E-1	-	4.31E-1	-
4×8	1.64E-2	3.16	6.55E-2	2.71
8×16	7.08E-4	4.53	3.42E-3	4.26
16×32	4.44E-5	4.06	4.84E-4	2.82
32×64	2.95E-6	3.85	7.96E-5	2.60

Table 4.4. An elliptic equation with mixed second order derivatives and Dirichlet boundary conditions.

	FEM with Approximated Coefficients (4.4)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.64E-2	-	7.01E-2	-
4×8	4.68E-3	2.50	1.92E-2	1.87
8×16	4.78E-4	3.29	2.70E-3	2.83
16×32	3.69E-5	3.69	2.43E-4	3.47
32×64	2.53E-6	3.87	1.82E-5	3.74
64×128	1.65E-7	3.94	1.25E-6	3.87
	FEM using Gauss Lobatto Quadrature (2.2)			
Mesh	l^2 error	order	l^∞ error	order
2×4	3.94E-2	-	7.15E-2	-
4×8	1.23E-2	1.67	3.28E-2	1.12
8×16	1.46E-3	3.08	5.42E-3	2.60
16×32	1.14E-4	3.68	3.96E-4	3.78
32×64	7.75E-6	3.88	2.62E-5	3.92
	FEM with Approximated Coefficients (4.5)			
Mesh	l^2 error	order	l^∞ error	order
2×4	4.08E-2	-	7.67E-2	-
4×8	1.01E-2	2.02	3.39E-2	1.18
8×16	5.22E-4	4.27	1.72E-3	4.30
16×32	3.14E-5	4.05	9.57E-5	4.17
32×64	1.99E-6	3.98	5.71E-6	4.07
	Full FEM Scheme			
Mesh	l^2 error	order	l^∞ error	order
2×4	7.35E-2	-	1.99E-1	-
4×8	5.94E-3	3.63	2.43E-2	3.03
8×16	4.31E-4	3.79	2.01E-3	3.60
16×32	2.83E-5	3.93	1.76E-4	3.93
32×64	1.68E-6	4.07	8.41E-6	4.07

Table 4.5. A Poisson equation with coefficient $\min_{(x,y)} a(x,y) \approx 0.001$.

	FEM with Approximated Coefficients (4.4)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.78E-1	-	4.52E-1	-
4×8	6.22E-2	2.16	2.08E-1	1.12
8×16	1.09E-2	2.51	8.44E-2	1.30
16×32	1.31E-3	3.05	1.81E-2	2.22
32×64	1.08E-4	3.60	1.75E-3	3.38
64×128	7.24E-6	3.90	1.52E-4	3.53
	FEM using Gauss Lobatto Quadrature (2.2)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.81E-1	-	4.59E-1	-
4×8	4.69E-2	2.58	1.37E-1	1.74
8×16	5.06E-3	3.21	3.75E-2	1.87
16×32	7.04E-4	2.85	7.86E-3	2.25
32×64	6.74E-5	3.39	1.21E-3	2.70
64×128	4.94E-6	3.77	1.17E-4	3.37
	FEM with Approximated Coefficients (4.5)			
Mesh	l^2 error	order	l^∞ error	order
2×4	2.68E-1	-	5.48E-1	-
4×8	2.91E-1	3.21	1.59E-1	1.78
8×16	3.51E-3	3.05	4.02E-2	1.98
16×32	2.86E-4	3.62	3.60E-3	3.48
32×64	1.86E-5	3.94	2.31E-4	3.96
64×128	1.17E-6	4.00	1.53E-5	3.91

The errors at Z_0 are shown in Table 4.1 and Table 4.2. We can see that the four schemes are all fourth order in the discrete 2-norm on Z_0. Even though we did not discuss the max norm error on Z_0, we should expect a $|\ln h|$ factor in the order of l^∞ error over Z_0 due to (2.4), which was proven upon the discrete Green's function.

Next we consider an elliptic equation with purely Dirichlet or purely Neumann boundary conditions:

$$\nabla \cdot (\mathbf{a}\nabla u) + cu = f \quad \text{on } [0,1]\times[0,2]$$

where $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = 10 + 30y^5 + x\cos y + y$, $a_{12} = a_{21} = 2 + 0.5(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$, $a_{22} = 10 + x^5$, $c = 1 + x^4y^3$ and $u(x,y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$. The errors at Z_0 are listed in Table 4.3 and Table 4.4. Recall that only $\mathcal{O}(h^{3.5})$ can be proven due to the mixed second order derivatives for the Neumann boundary conditions as discussed in Remark 4.4.13, we observe around fourth order accuracy for (4.4) and (4.5) for Neumann boundary conditions in this particular example.

4.5.2 Robustness

In Table 4.1 and Table 4.2, the errors of approximated coefficient schemes (4.4), (4.5) and the Gauss Lobatto quadrature scheme (2.2) are close to one another. We observe that the scheme (4.5) tends to be more accurate than (4.4) and (2.2) when the coefficient $a(x,y)$ is closer to zero in the Poisson equation. See Table 4.5 for errors of solving $\nabla\cdot(a\nabla u) = f \quad \text{on } [0,1]\times[0,2]$ with Dirichlet boundary conditions, $a(x,y) = 1 + \varepsilon x^3y^5 + \cos(x^3y^2 + 1)$ and $u(x,y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$ where $\varepsilon = 0.001$. Here the smallest value of $a(x,y)$ is around $\varepsilon = 0.001$. We remark that the difference among three schemes is much smaller for larger ε such as $\varepsilon = 0.1$ as in Table 4.1.

4.6 Concluding Remarks

We have shown that the classical superconvergence of functions values at Gauss Lobatto points in C^0-Q^k finite element method for an elliptic problem still holds if replacing the coefficients by their piecewise Q^k Lagrange interpolants at the Gauss Lobatto points. Such a superconvergence result can be used for constructing a fourth order accurate finite difference type scheme by using Q^2 approximated variable coefficients. Numerical tests suggest that this is an efficient and robust implementation of C^0-Q^2 finite element method without affecting the superconvergence of function values.

5. ON THE MONOTONICITY AND DISCRETE MAXIMUM PRINCIPLE OF THE FINITE DIFFERENCE IMPLEMENTATION OF C^0-Q^2 FINITE ELEMENT METHOD

5.1 Introduction

5.1.1 Monotonicity and discrete maximum principle

Consider a Poisson equation with variable coefficients and Dirichlet boundary conditions on a two dimensional rectangular domain $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} \mathcal{L}u \equiv -\nabla \cdot (a\nabla u) + cu = 0 \quad &\text{on} \quad \Omega, \\ u = g \quad &\text{on} \quad \partial\Omega, \end{aligned} \tag{5.1}$$

where $a(x,y) \in C^1(\bar{\Omega})$, $c(x,y) \in C^0(\bar{\Omega})$ with $0 < a_{\min} \leq a(x,y) \leq a_{\max}$ and $c(x,y) \geq 0$. For a smooth function $u \in C^2(\Omega) \cap C(\bar{\Omega})$, maximum principle holds [14]: $\mathcal{L}u \leq 0$ in $\Omega \Longrightarrow \max_{\bar{\Omega}} u \leq \max\left\{0, \max_{\partial\Omega} u\right\}$, and in particular,

$$\mathcal{L}u = 0 \text{ in } \Omega \Longrightarrow |u(x,y)| \leq \max_{\partial\Omega} |u|, \quad \forall (x,y) \in \Omega. \tag{5.2}$$

For various purposes, it is desired to have numerical schemes to satisfy (5.2) in the discrete sense. A linear approximation to $\mathcal{L}$ can be represented as a matrix L_h. The matrix L_h is called *monotone* if its inverse has nonnegative entries, i.e., $L_h^{-1} \geq 0$. All matrix inequalities in this chapter are entrywise inequalities. One sufficient condition for the discrete maximum principle is the *monotonicity* of the scheme, which was also used to prove convergence of numerical schemes, e.g., [16]–[19].

In this chapter, we will discuss the monotonicity and discrete maximum principle of the simplest finite difference implementation of the continuous finite element method with Q^2 basis (i.e., tensor product of quadratic polynomial) for (5.1), which is a fourth order accurate scheme.

5.1.2 Second order schemes and M-Matrices

The second order centered difference $u \approx \frac{u_{i-1}-2u_i+u_{i+1}}{\Delta x^2}$ for solving $-u(x) = f(x), u(0) = u(1) = 0$ results in a tridiagonal $(-1, 2, -1)$ matrix, which is an M-Matrix. Nonsingular M-Matrices are inverse-positive matrices and it is the most convenient tool for constructing inverse-positive matrices. There are many equivalent definitions or characterizations of M-Matrices, see [67]. One convenient characterization of nonsingular M-Matrices are nonsingular matrices with nonpositive off-diagonal entries and positive diagonal entries, and all row sums are non-negative with at least one row sum is positive.

The continuous finite element method with piecewise linear basis forms an M-Matrix for the variable coefficient problem (5.1) on triangular meshes under reasonable mesh constraints [68]. The M-Matrix structure in linear finite element method also holds for a nonlinear elliptic equation [69]. For solving $-\Delta u = f$ on regular triangular meshes, linear finite element method reduces to the 5-point discrete Laplacian. Linear finite element method or the 5-point discrete Laplacian is the most popular method in the literature for constructing schemes satisfying a discrete maximum principle and bound-preserving properties.

Almost all high order accurate schemes result in positive off-diagonal entries in L_h for solving $-\Delta u = f$ thus L_h is no longer an M-Matrix. The only known exceptions are the fourth order accurate 9-point discrete Laplacian and the fourth order accurate compact finite difference scheme.

5.1.3 Existing high order accurate monotone methods for two-dimensional Laplacian

There are at least three kinds of high order accurate schemes which have been proven to satisfy $L_h^{-1} \geq 0$ for the Laplacian operator $\mathcal{L}u = -\Delta u$:

1. Both the fourth order accurate 9-point discrete Laplacian scheme [16], [70] and the fourth order accurate compact finite difference scheme [1], [71] for $-\Delta u = f$ can be written as $S\mathbf{u} = W\mathbf{f}$ with S being an M-Matrix and $W \geq 0$, thus $L_h^{-1} = S^{-1}M \geq 0$.

2. In [72], [73], Bramble and Hubbard constructed a fourth order accurate finite difference discrete Laplacian operator for which L_h is not an M-Matrix but monotonicity $L_h^{-1} \geq 0$ is ensured through an M-Matrix factorization $L_h = M_1 M_2$, i.e., L_h is a product of two M-Matrices.

3. Finite element method with quadratic polynomial (P2 FEM) basis on a regular triangular mesh can be implemented as a finite difference scheme defined at vertices and edge centers of triangles [31]. The error estimate of P2 FEM is third order in L^2-norm. The error at at vertices and edge centers are fourth order accurate in l^2-norm due to superconvergence. The stiffness matrix is not an M-Matrix but its monotonicity was proven in [21].

For discrete maximum principle to hold in P2 FEM on a generic triangular mesh, it was proven in [20] that it is necessary and sufficient to require a very strong mesh constraint, which essentially gives either regular triangulation or equilateral triangulation. Thus, the discrete maximum principle holds in P2 FEM on a regular triangulation or an equilateral triangulation. For finite element method with cubic and higher order polynomials on regular triangular meshes, it was shown that the discrete maximum principle fails in [74].

5.1.4 Other known results regarding discrete maximum principle

For one-dimensional Laplacian, discrete maximum principle was proven for arbitrarily high order finite element method using discrete Green's function in [75]. The discrete Green's function was also used to analyze P1 FEM in two dimensions [76]. Discontinuous coefficients were considered and a nonlinear scheme was constructed in [77]. Piecewise constant coefficient in one dimension was considered in [78]. A numerical study for high order FEM with very accurate Gauss quadrature in two dimensions showed that DMP was violated on non-uniform unstructured meshes for variable coefficients in [79]. A more general operator $\nabla(\mathbf{a}\nabla u)$ with matrix coefficients $\mathbf{a}$ was considered for linear FEM in [80]. See [81] for an anisotropic computational example.

5.1.5 Existing inverse-positive approaches when L_h is not an M-Matrix

In this chapter, we will focus on the finite difference implementation of continuous finite element method with Q^2 basis (Q2 FEM), which will be reviewed in Section 5.2. The matrix L_h in such a scheme is not an M-Matrix due to its off-diagonal positive entries. There are at least three methods to study whether $L_h^{-1} \geq 0$ holds when M-Matrix structure is lost:

1. An M-Matrix factorization of the form $L_h = M_1 M_2$ was shown in [73] and [82]. In Appendix 5.6, we will demonstrate an M-Matrix factorization for the finite difference implementation of Q^2 FEM solving $-\Delta u = f$.
2. Perturbation of M-Matrices by positive off-diagonal entries without losing monotonicity was discussed in [83].
3. In [21], Lorenz proposed a sufficient condition for ensuring $L_h = M_1 M_2$. Lorenz's condition will be reviewed in Section 5.3.3.

The main result of this chapter is to prove that $L_h^{-1} \geq 0$ and a discrete maximum principle holds under some mesh constraint in the fourth order accurate finite difference implementation of Q^2 FEM solving (5.1) by verifying the Lorenz's condition.

5.1.6 Extensions to the discrete maximum principle for parabolic equations

Classical solutions to the parabolic equation $u_t = \nabla \cdot (a \nabla u)$ satisfy a maximum principle [14]. With suitable boundary conditions and initial value $u(x, y, 0)$ such as periodic or homogeneous Dirichlet boundary conditions and initial minimum $\min_{\Omega} u(x, y, 0) = 0$, the solution to the initial value problem satisfies the following maximum principle:

$$\min_{(x,y)} u(x, y, 0) \leq u(x, y, t) \leq \max_{(x,y)} u(x, y, 0). \tag{5.3}$$

Now consider solving $u_t = \nabla \cdot (a \nabla u)$ with backward Euler time discretization, then U^{n+1} satisfies an elliptic equation of the form (5.1):

$$-\nabla \cdot (a \nabla U^{n+1}) + \frac{1}{\Delta t} U^{n+1} = \frac{1}{\Delta t} U^n. \tag{5.4}$$

If S_h denotes spatial discretization for $-\nabla\cdot(a\nabla u)$, then the numerical scheme can be written as $U^{n+1}=(I+\Delta t S_h)^{-1}U^n$. Let $\mathbf{1}=\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^T$. Then for suitable boundary conditions usually we have $S_h\mathbf{1}=\mathbf{0}$ since S_h approximates a differential operator. So we have $(I+\Delta t S_h)\mathbf{1}=\mathbf{1}$ thus $(I+\Delta t S_h)^{-1}\mathbf{1}=\mathbf{1}$. If we further have the monotonicity $(I+\Delta t S_h)^{-1}\geq 0$, then each row of the $(I+\Delta t S_h)^{-1}$ has nonnegative entries and sums to one, thus the discrete maximum principle holds $\min_j U_j^n \leq U_j^{n+1}\leq \max_j U_j^n$, which is a desired and useful property in many applications. For instance, second order centered difference or P1 finite element method has been used to construct schemes satisfying the discrete maximum principle in solving phase field equations [84]–[86]. In the rest of the chapter, we will only focus on discussing the equation (5.1), even though all discussions can be extended to solving the parabolic equation with backward Euler time discretization.

To the best of our knowledge, this is the first time that a high order accurate scheme under suitable mesh constraints is proven to be monotone in the sense $L_h^{-1}\geq 0$ for solving a variable coefficient $a(\mathbf{x})$ in (5.1) in two dimensions. For simplicity, we only discuss an uniform mesh in this chapter, even though the main results can be extended to non-uniform meshes. However, an additional mesh constraint is expected for the discrete maximum principle to hold. See such a mesh constraint of non-uniform meshes for Q1 FEM in [87] and P2 FEM for one-dimensional problem in [75].

This chapter is organized as follows. In Section 5.2, we describe the fourth order accurate finite difference implementation of C^0-Q^2 finite element method. In Section 5.3, we review the sufficient conditions to ensure monotonicity and discrete maximum principle. In Section 5.4, we prove that the fourth order accurate finite difference implementation of C^0-Q^2 finite element method is monotone under some mesh constraints. Numerical tests are given in Section 5.5. Concluding remarks are given in Section 5.6.

5.2 Finite Difference Implementation Of C^0-Q^2 Finite Element Method

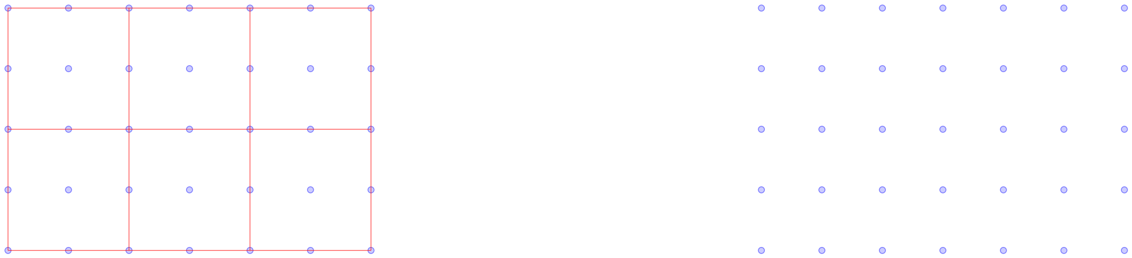
Consider solving the following elliptic equation on $\Omega = (0,1) \times (0,1)$ with Dirichlet boundary conditions:

$$\begin{aligned} \mathcal{L}u \equiv -\nabla \cdot (a\nabla u) + cu = f \quad \text{on} \quad \Omega, \\ u = g \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{5.5}$$

Assume there is a function $\bar{g} \in H^1(\Omega)$ as an extension of g so that $\bar{g}|_{\partial\Omega} = g$. The variational form of (5.1) is to find $\tilde{u} = u - \bar{g} \in H_0^1(\Omega)$ satisfying

$$\mathcal{A}(\tilde{u}, v) = (f, v) - \mathcal{A}(\bar{g}, v), \quad \forall v \in H_0^1(\Omega), \tag{5.6}$$

where $\mathcal{A}(u,v) = \iint_\Omega a\nabla u \cdot \nabla v dxdy + \iint_\Omega cuvdxdy$, $(f,v) = \iint_\Omega fvdxdy$.



(a) The quadrature points and a FEM mesh.

(b) The corresponding finite difference grid.

Figure 5.1. An illustration of Q^2 element and the 3×3 Gauss-Lobatto quadrature.

Let h be the mesh size of the rectangular mesh and $V_0^h \subseteq H_0^1(\Omega)$ be the continuous finite element space consisting of piecewise Q^2 polynomials (i.e., tensor product of piecewise quadratic polynomials), then the most convenient implementation of C^0-Q^2 finite element method is to use 3×3 Gauss-Lobatto quadrature rule for all the integrals, see Figure 5.1. Such a numerical scheme can be defined as: find $u_h \in V_0^h$ satisfying

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h - \mathcal{A}_h(g_I, v_h), \quad \forall v_h \in V_0^h, \tag{5.7}$$

where $\mathcal{A}_h(u_h, v_h)$ and $\langle f, v_h \rangle_h$ denote using tensor product of 3-point Gauss-Lobatto quadrature for integrals $\mathcal{A}(u_h, v_h)$ and (f, v_h) respectively, and g_I is the piecewise Q^2 Lagrangian interpolation polynomial at the 3 × 3 quadrature points shown in Figure 5.1 of the following function:

$$g(x,y) = \begin{cases} 0, & \text{if} \quad (x,y) \in (0,1) \times (0,1), \\ g(x,y), & \text{if} \quad (x,y) \in \partial\Omega. \end{cases}$$

Then $\bar{u}_h = u_h + g_I$ is the numerical solution for the problem (5.5). We emphasize that (5.7) is not a straightforward approximation to (5.6) since $\bar{g}$ is never used. It was proven in chapter 2 that the scheme (5.7) is fourth order accurate if coefficients and exact solutions are smooth. Notice that $\bar{u}_h$ satisfies:

$$\mathcal{A}_h(\bar{u}_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h. \tag{5.8}$$

See chapter 2 for the detailed finite difference implementation and proof of fourth order accuracy for the scheme (5.7).

5.2.1 One-dimensional case

Now consider the one-dimensional Dirichlet boundary value problem:

$$\begin{aligned} -(au) + cu =& f \text{ on } (0,1), \\ u(0) = \sigma_0, & \quad u(1) = \sigma_1. \end{aligned}$$

Consider a uniform mesh $x_i = ih$, $i = 0, 1, \ldots, n+1$, $h = \frac{1}{n+1}$. Assume n is odd and let $M = \frac{n+1}{2}$. Define intervals $I_k = [x_{2k}, x_{2k+2}]$ for $k = 0, \ldots, M-1$ as a finite element mesh for P^2 basis. Define

$$V^h = \{v \in C^0([0,1]) : v \in P^2(I_k), k = 0, \ldots, M-1\}.$$

Let $\{\phi_i\}_{i=0}^{n+1} \subset V^h$ be a basis for V^h so that $\phi_i(x_j) = \delta_{ij}$, $i, j = 0, 1, \ldots, n+1$. Let $u_0 = \sigma_0$, $u_i = u_h(x_i)$ and $u_{n+1} = \sigma_1$, then $u_h, \bar{u}_h \in V^h$ can be represented as

$$u_h(x) = \sum_{i=1}^{n} u_i \phi_i(x), \quad \bar{u}_h(x) = \sum_{i=0}^{n+1} u_i \phi_i(x).$$

Let $f_j = f(x_j)$, then (5.8) becomes

$$\langle au_h, \phi_i \rangle_h + \langle cu_h, \phi_i \rangle_h = \langle f, \phi_i \rangle_h, \quad i = 1, \ldots, n; u_0 = \sigma_0,\, u_{n+1} = \sigma_1,$$

which are

$$\sum_{j=0}^{n+1} u_j \left(\langle a\phi_j, \phi_i \rangle_h + \langle c\phi_j, \phi_i \rangle_h \right) = \sum_{j=0}^{n+1} f_j \langle \phi_j, \phi_i \rangle_h, \quad i = 1, \ldots, n;$$

$$u_0 = \sigma_0,\, u_{n+1} = \sigma_1.$$

The matrix form is $S\bar{\mathbf{u}} = M\bar{\mathbf{f}}$ where

$$\bar{\mathbf{u}} = \begin{bmatrix} u_0 & u_1 & u_2 \cdots & u_n & u_{n+1} \end{bmatrix}^T, \quad \bar{\mathbf{f}} = \begin{bmatrix} \sigma_0 & f_1 & f_2 \cdots & f_n & \sigma_1 \end{bmatrix}^T.$$

The scheme can be written as $\mathcal{L}_h(\bar{\mathbf{u}}) = \bar{\mathbf{f}}$. The linear operator $\mathcal{L}_h$ has the matrix representation $L_h = M^{-1}S$.

For the Laplacian $\mathcal{L}u = -u$, we have

$$\mathcal{L}_h(\bar{\mathbf{u}})_0 = u_0 = \sigma_0, \quad \mathcal{L}_h(\bar{\mathbf{u}})_{n+1} = u_{n+1} = \sigma_1, \tag{5.9a}$$

$$\text{if } i \text{ is odd, i.e., } x_i \text{ is a cell center,} \tag{5.9b}$$

$$\mathcal{L}_h(\bar{\mathbf{u}})_i = \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i, \tag{5.9c}$$

$$\text{if } i \text{ is even, i.e., } x_i \text{ is a cell end,} \tag{5.9d}$$

$$\mathcal{L}_h(\bar{\mathbf{u}})_i = \frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i. \tag{5.9e}$$

For the variable coefficient operator $\mathcal{L}u = -(au) + cu$, we have

$$\mathcal{L}_h(\bar{\mathbf{u}})_0 = u_0 = \sigma_0, \quad \mathcal{L}_h(\bar{\mathbf{u}})_{n+1} = u_{n+1} = \sigma_1, \tag{5.10a}$$

and if x_i is a cell center, we have

$$\mathcal{L}_h(\bar{\mathbf{u}})_i = \frac{-(3a_{i-1} + a_{i+1})u_{i-1} + 4(a_{i-1} + a_{i+1})u_i - (a_{i-1} + 3a_{i+1})u_{i+1}}{4h^2} + c_i u_i = f_i; \tag{5.10b}$$

and if x_i is a cell end, then

$$\begin{aligned}\mathcal{L}_h(\bar{\mathbf{u}})_i = &\frac{(3a_{i-2} - 4a_{i-1} + 3a_i)u_{i-2} - (4a_{i-2} + 12a_i)u_{i-1}}{8h^2} \\ &+\frac{(a_{i-2} + 4a_{i-1} + 18a_i + 4a_{i+1} + a_{i+2})u_i}{8h^2} \\ &+\frac{-(12a_i + 4a_{i+2})u_{i+1} + (3a_{i+2} - 4a_{i+1} + 3a_i)u_{i+2}}{8h^2} + c_i u_i = f_i. \qquad (5.10c)\end{aligned}$$

5.2.2 Two-dimensional case

Consider a uniform grid (x_i, y_j) for a rectangular domain $[0, 1] \times [0, 1]$ where $x_i = ih$, $i = 0, 1, \ldots, n + 1$ and $y_j = jh$, $j = 0, 1, \ldots, n + 1$, $h = \frac{1}{n+1}$, where n must be odd. Let u_{ij} denote the numerical solution at (x_i, y_j). Let $\mathbf{u}$ denote an abstract vector consisting of u_{ij} for $i, j = 1, 2, \cdots, n$. Let $\bar{\mathbf{u}}$ denote an abstract vector consisting of u_{ij} for $i, j = 0, 1, 2, \cdots, n, n + 1$. Let $\bar{\mathbf{f}}$ denote an abstract vector consisting of f_{ij} for $i, j = 1, 2, \cdots, n$ and the boundary condition g at the boundary grid points.

The scheme (5.8) for solving (5.5) can still be written as $\mathcal{L}_h(\bar{\mathbf{u}}) = \bar{\mathbf{f}}$.

Two-dimensional Laplacian

For the Laplacian $\mathcal{L}u = -\Delta u$, $\mathcal{L}_h(\bar{\mathbf{u}})$ can be expressed as the following. If $(x_i, y_j) \in \partial\Omega$, then

$$\mathcal{L}_h(\bar{\mathbf{u}})_{i,j} = u_{i,j} = g_{i,j}.$$

If (x_i, y_j) is an interior grid point and a cell center , $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\frac{-u_{i-1,j} - u_{i+1,j} + 4u_{i,j} - u_{i,j+1} - u_{i+1,j}}{h^2} = f_{i,j}. \qquad (5.11a)$$

For interior grid points, there are three types: cell center, edge center and knots. See Figure 5.2. If (x_i, y_j) is an interior grid point and an edge center for an edge parallel to x-axis, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\frac{-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}}{h^2} + \frac{u_{i,j-2} - 8u_{i,j-1} + 14u_{i,j} - 8u_{i,j+1} + u_{i,j+2}}{4h^2} = f_{i,j}. \qquad (5.11b)$$

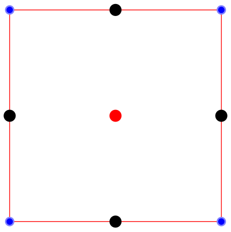


Figure 5.2. Three types of interior grid points: red cell center, blue knots and black edge centers for a finite element cell.

If (x_i, y_j) is an interior grid point and an edge center for an edge parallel to y-axis, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is similarly defined as above. If (x_i, y_j) is an interior grid point and a knot (x_i, y_j), $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\begin{aligned} &\frac{u_{i-2,j} - 8u_{i-1,j} + 14u_{i,j} - 8u_{i+1,j} + u_{i+2,j}}{4h^2} \\ +&\frac{u_{i,j-2} - 8u_{i,j-1} + 14u_{i,j} - 8u_{i,j+1} + u_{i,j+2}}{4h^2} = f_{i,j}. \end{aligned} \tag{5.11c}$$

If ignoring the denominator h^2, then the stencil of the operator $\mathcal{L}_h$ at interior grid points can be represented as:

$$\text{cell center} \quad \begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array} \qquad \text{knots} \quad \begin{array}{ccccc} & & \frac{1}{4} & & \\ & & -2 & & \\ \frac{1}{4} & -2 & 7 & -2 & \frac{1}{4} \\ & & -2 & & \\ & & \frac{1}{4} & & \end{array}$$

$$\text{edge center (edge parallel to } y\text{-axis)} \quad \begin{array}{ccccc} & & -1 & & \\ \frac{1}{4} & -2 & \frac{11}{2} & -2 & \frac{1}{4} \\ & & -1 & & \end{array}$$

edge center (edge parallel to x-axis) $\begin{array}{ccc} & \frac{1}{4} & \\ & -2 & \\ -1 & \frac{11}{2} & -1 \\ & -2 & \\ & \frac{1}{4} & \end{array}$

5.2.3 Two-dimensional variable coefficient case

For $\mathcal{L}u = -\nabla \cdot (a\nabla u) + cu$, $\mathcal{L}_h(\bar{\mathbf{u}})$ will have exactly the same stencil size as the Laplacian case. At boundary points $(x_i, y_j) \in \partial\Omega$, $\mathcal{L}_h(\bar{\mathbf{u}}) = \bar{\mathbf{f}}$ becomes

$$\mathcal{L}_h(\bar{\mathbf{u}})_{i,j} = u_{i,j} = g_{i,j}. \tag{5.12a}$$

If (x_i, y_j) is an interior grid point and a cell center, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\begin{aligned} &\frac{-(3a_{i-1,j} + a_{i+1,j})u_{i-1,j} + 4(a_{i-1,j} + a_{i+1,j})u_{i,j} - (a_{i-1,j} + 3a_{i+1,j})u_{i+1,j}}{4h^2} \\ &+ \frac{-(3a_{i,j-1} + a_{i,j+1})u_{i,j-1} + 4(a_{i,j-1} + a_{i,j+1})u_{i,j} - (a_{i,j-1} + 3a_{i,j+1})u_{i,j+1}}{4h^2} + c_{ij}u_{ij}. \end{aligned} \tag{5.12b}$$

If (x_i, y_j) is an interior grid point and a knot, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\begin{aligned} &\frac{(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})u_{i-2,j} - (4a_{i-2,j} + 12a_{i,j})u_{i-1,j}}{8h^2} \\ &+ \frac{(a_{i-2,j} + 4a_{i-1,j} + 18a_{i,j} + 4a_{i+1,j} + a_{i+2,j})u_{i,j}}{8h^2} \\ &+ \frac{-(12a_{i,j} + 4a_{i+2,j})u_{i+1,j} + (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})u_{i+2,j}}{8h^2} \\ &+ \frac{(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})u_{i,j-2} - (4a_{i,j-2} + 12a_{i,j})u_{i,j-1}}{8h^2} \\ &+ \frac{(a_{i,j-2} + 4a_{i,j-1} + 18a_{i,j} + 4a_{i,j+1} + a_{i,j+2})u_{i,j}}{8h^2} \\ &+ \frac{-(12a_{i,j} + 4a_{i,j+2})u_{i,j+1} + (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})u_{i,j+2}}{8h^2} + c_{ij}u_{ij}. \end{aligned} \tag{5.12c}$$

If (x_i, y_j) is an interior grid point and an edge center for an edge parallel to y-axis, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\begin{aligned}
&\frac{(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})u_{i-2,j} - (4a_{i-2,j} + 12a_{i,j})u_{i-1,j}}{8h^2} \\
+&\frac{(a_{i-2,j} + 4a_{i-1,j} + 18a_{i,j} + 4a_{i+1,j} + a_{i+2,j})u_{i,j}}{8h^2} \\
+&\frac{-(12a_{i,j} + 4a_{i+2,j})u_{i+1,j} + (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})u_{i+2,j}}{8h^2} \\
+&\frac{-(3a_{i,j-1} + a_{i,j+1})u_{i,j-1} + 4(a_{i,j-1} + a_{i,j+1})u_{i,j} - (a_{i,j-1} + 3a_{i,j+1})u_{i,j+1}}{4h^2} + c_{ij}u_{ij}.
\end{aligned} \tag{5.12d}$$

If (x_i, y_j) is an interior grid point and an edge center for an edge parallel to x-axis, $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}$ is equal to

$$\begin{aligned}
&\frac{(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})u_{i,j-2} - (4a_{i,j-2} + 12a_{i,j})u_{i,j-1}}{8h^2} \\
+&\frac{(a_{i,j-2} + 4a_{i,j-1} + 18a_{i,j} + 4a_{i,j+1} + a_{i,j+2})u_{i,j}}{8h^2} \\
+&\frac{-(12a_{i,j} + 4a_{i,j+2})u_{i,j+1} + (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})u_{i,j+2}}{8h^2} \\
+&\frac{-(3a_{i-1,j} + a_{i+1,j})u_{i-1,j} + 4(a_{i-1,j} + a_{i+1,j})u_{i,j} - (a_{i-1,j} + 3a_{i+1,j})u_{i+1,j}}{4h^2} + c_{ij}u_{ij}.
\end{aligned} \tag{5.12e}$$

5.3 Sufficient Conditions For Monotonicity And Discrete Maximum Principle

5.3.1 Discrete maximum principle

Assume there are N grid points in the domain Ω and N^∂ grid points on $\partial\Omega$. Define

$$\mathbf{u} = \begin{pmatrix} u_1 & u_2 & \cdots & u_N \end{pmatrix}^T, \quad \mathbf{u}^\partial = \begin{pmatrix} u_1^\partial & u_2^\partial & \cdots & u_{N^\partial}^\partial \end{pmatrix}^T,$$

$$\tilde{\mathbf{u}} = \begin{pmatrix} u_1 & u_2 & \cdots & u_N & u_1^\partial & u_2^\partial & \cdots & u_{N^\partial}^\partial \end{pmatrix}^T.$$

A finite difference scheme can be written as

$$\begin{aligned}\mathcal{L}_h(\tilde{\mathbf{u}})_i = \sum_{j=1}^{N} b_{ij}u_j + \sum_{j=1}^{N^\partial} b_{ij}^\partial u_j^\partial =& f_i, \quad 1 \le i \le N,\\ u_i^\partial =& g_i, \quad 1 \le i \le N^\partial.\end{aligned}$$

The matrix form is

$$\tilde{L}_h\tilde{\mathbf{u}} = \tilde{\mathbf{f}}, \tilde{L}_h = \begin{pmatrix} L_h & B^\partial \\ 0 & I \end{pmatrix}, \tilde{\mathbf{u}} = \begin{pmatrix} \mathbf{u} \\ \mathbf{u}^\partial \end{pmatrix}, \tilde{\mathbf{f}} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}.$$

The discrete maximum principle is

$$\mathcal{L}_h(\tilde{\mathbf{u}})_i \le 0, 1 \le i \le N \Longrightarrow \max_i u_i \le \max\{0, \max_i u_i^\partial\} \tag{5.13}$$

which implies

$$\mathcal{L}_h(\tilde{\mathbf{u}})_i = 0, 1 \le i \le N \Longrightarrow |u_i| \le \max_i |u_i^\partial|.$$

The following result was proven in [15]:

Theorem 5.3.1. *A finite difference operator $\mathcal{L}_h$ satisfies the discrete maximum principle (5.13) if $\tilde{L}_h^{-1} \ge 0$ and all row sums of $\tilde{L}_h$ are non-negative.*

Let $\bar{\mathbf{u}}$ and $\bar{\mathbf{f}}$ be the same vectors as defined in Section 5.2. For the same finite difference scheme, the matrix form can also be written as

$$\bar{L}_h\bar{\mathbf{u}} = \bar{\mathbf{f}}.$$

Notice that there exist two permutation matrices P_1 and P_2 such that $\bar{\mathbf{u}} = P_1\tilde{\mathbf{u}}$ and $\bar{\mathbf{f}} = P_2\tilde{\mathbf{f}}$. Since the matrix vector form of the same scheme is also $\tilde{L}_h\tilde{\mathbf{u}} = \tilde{\mathbf{f}}$, we obtain $P_2^{-1}\bar{L}_hP_1 = \tilde{L}_h$. Notice that a permutation matrix P is inverse-positive and the signs of row sums will not be altered after multiplying P to $\tilde{L}_h$. Thus we have

Theorem 5.3.2. *If $\bar{L}_h$ is inverse-positive and row sums of $\bar{L}_h$ are non-negative, then $\mathcal{L}_h$ satisfies the discrete maximum principle (5.13).*

Notice that $\tilde{L}_h^{-1} = \begin{pmatrix} L_h^{-1} & -L_h^{-1}B^\partial \\ 0 & I \end{pmatrix}$, thus we have

Theorem 5.3.3. *If $\bar{L}_h^{-1} \geq 0$, then $\tilde{L}_h^{-1} \geq 0$ and thus $L_h^{-1} \geq 0$.*

Let $\mathbf{1}$ denote a vector of suitable size with 1 as entries, then for all schemes in Section 5.2, $\mathcal{L}_h(\mathbf{1}) \geq 0$, which implies the row sums of $\bar{L}_h$ are non-negative. Thus from now on, we only need to discuss the monotonicity of the matrix $\bar{L}_h$.

5.3.2 Characterizations of nonsingular M-Matrices

M-Matrices belong to the set of Z-matrices which are matrices with nonpositive off-diagonal entries. Nonsingular M-Matrices are always inverse-positive. See [67] for the definition and various characterization of nonsingular M-Matrices. The following is a convenient sufficient condition to characterize nonsingular M-Matrices:

Theorem 5.3.4. *For a real square matrix A with positive diagonal entries and non-positive off-diagonal entries, A is a nonsingular M-Matrix if all the row sums of A are non-negative and at least one row sum is positive.*

Proof. By condition C_{10} in [67], A is a nonsingular M-Matrix if and only if $A + aI$ is nonsingular for any $a \geq 0$. Since all the row sums of A are non-negative and at least one row sum is positive, the matrix A is irreducibly diagonally dominant thus nonsingular, and $A + aI$ is strictly diagonally dominant thus nonsingular for any $a > 0$. □

Definition 1. Let $\mathcal{N} = \{1, 2, \ldots, n\}$. For $\mathcal{N}_1, \mathcal{N}_2 \subset \mathcal{N}$, we say a matrix A of size $n \times n$ connects $\mathcal{N}_1$ with $\mathcal{N}_2$ if

$$\forall i_0 \in \mathcal{N}_1, \exists i_r \in \mathcal{N}_2, \exists i_1, \ldots, i_{r-1} \in \mathcal{N} \quad \text{s.t.} \quad a_{i_{k-1}i_k} \neq 0, \quad k = 1, \cdots, r. \tag{5.14}$$

If perceiving A as a directed graph adjacency matrix of vertices labeled by $\mathcal{N}$, then (5.14) simply means that there exists a directed path from any vertex in $\mathcal{N}_1$ to at least one vertex in $\mathcal{N}_2$. In particular, if $\mathcal{N}_1 = \emptyset$, then any matrix A connects $\mathcal{N}_1$ with $\mathcal{N}_2$.

Given a square matrix A and a column vector $\mathbf{x}$, we define

$$\mathcal{N}^0(A\mathbf{x}) = \{i : (A\mathbf{x})_i = 0\}, \quad \mathcal{N}^+(A\mathbf{x}) = \{i : (A\mathbf{x})_i > 0\}.$$

By condition L_{36} in [67], we have the following characterization of nonsingular M-Matrices:

Theorem 5.3.5. *For a real square matrix A with non-positive off-diagonal entries, if there is a vector $\mathbf{x} > 0$ with $A\mathbf{x} \geq 0$ s.t. A connects $\mathcal{N}^0(A\mathbf{x})$ with $\mathcal{N}^+(A\mathbf{x})$, then A is a nonsingular M-Matrix thus $A^{-1} \geq 0$.*

5.3.3 Lorenz's sufficient condition for monotonicity

All results in this subsection were first shown in [21]. For completeness, we include a detailed proof.

Given a matrix $A = [a_{ij}] \in \mathbb{R}^{n\times n}$, define its diagonal, positive and negative off-diagonal parts as $n \times n$ matrices A_d, A_a, A_a^+, A_a^-:

$$(A_d)_{ij} = \begin{cases} a_{ii}, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \neq j \end{cases}, \quad A_a = A - A_d,$$

$$(A_a^+)_{ij} = \begin{cases} a_{ij}, & \text{if} \quad a_{ij} > 0, \quad i \neq j \\ 0, & \text{otherwise.} \end{cases}, \quad A_a^- = A_a - A_a^+.$$

Lemma 5.3.6. *If A is monotone, then for any two matrices $B \geq C$, $A^{-1}B \geq A^{-1}C$.*

Proof. For any two column vectors $\mathbf{b} \geq \mathbf{c}$, we have

$$\mathbf{b} - \mathbf{c} \geq 0 \Rightarrow A^{-1}(\mathbf{b} - \mathbf{c}) \geq 0 \Rightarrow A^{-1}\mathbf{b} \geq A\mathbf{c}.$$

By considering $\mathbf{b}$ and $\mathbf{c}$ as column vectors of B and C, we get $A^{-1}B \geq A^{-1}C$. □

Lemma 5.3.7. *If A is an M-Matrix, then $A_d \geq A$ and $A^{-1} \geq A_d^{-1}$.*

Proof. $A_d \geq A$ is trivial. A is monotone, thus

$$A_d \geq A \Rightarrow A^{-1}A_d \geq A^{-1}A = I.$$

And $A_d^{-1} \geq 0$ implies

$$A^{-1}A_d \geq I \Rightarrow A^{-1}A_dA_d^{-1} \geq IA_d^{-1} \Rightarrow A^{-1} \geq A_d^{-1}.$$

□

Theorem 5.3.8. *If $A_a \leq 0$ and there exists a nonzero vector $\mathbf{e} \in \mathbb{R}^n$ such that $\mathbf{e} \geq 0$ and $A\mathbf{e} \geq 0$. Moreover, A connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$. Then the following hold:*

- $\mathbf{e} > 0$.
- $a_{ii} > 0$, $\forall i \in N$.
- *A is a M-Matrix and $A^{-1} \geq 0$.*

Proof. Assume there is one index i such that $e_i = 0$, then

$$0 \leq (A\mathbf{e})_i = \sum_{j \neq i} a_{ij}e_j \leq 0 \Rightarrow (A\mathbf{e})_i = 0 \Rightarrow \sum_{j \neq i} a_{ij}e_j = 0 \Rightarrow a_{ij}e_j = 0, \forall j.$$

Thus if $a_{ij} < 0$, then $e_j = 0$, which implies $(A\mathbf{e})_j = 0$ by the same argument as above. Therefore, A has no off-diagonal nonzero entry a_{kl} such that $k \in \mathcal{N}^0(A\mathbf{e})$ and $l \in \mathcal{N}^+(A\mathbf{e})$. In other words, if A represents the graph adjacency matrix for a directed graph of vertices indexed by $1, 2, \cdots, n$, then any edge starting from a vertex $i \in \mathcal{N}^0(A\mathbf{e})$ points to vertices in $\mathcal{N}^0(A\mathbf{e})$, thus there is no directed path from $i \in \mathcal{N}^0(A\mathbf{e})$ to any vertex in $\mathcal{N}^+(A\mathbf{e})$, which contradicts to the assumption that A connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$. With $\mathbf{e} > 0$, the rest is proven by following Theorem 5.3.5. □

corollary 5.3.9. *If A is a nonsingular M-Matrix, $\mathbf{f} \in \mathbb{R}^n$ is a nonzero vector with $\mathbf{f} \geq 0$ and A connects $\mathcal{N}^0(\mathbf{f})$ with $\mathcal{N}^+(\mathbf{f})$, then $A^{-1}\mathbf{f} > 0$.*

Proof. By using $\mathbf{e} = A^{-1}\mathbf{f} \geq 0$ in Theorem 5.3.8, we get $A^{-1}\mathbf{f} > 0$. □

Theorem 5.3.10. *If $A \leq M_1 M_2 \cdots M_k L$ where $M_1, \cdots, M_k$ are nonsingular M-Matrices and $L_a \leq 0$, and there exists a nonzero vector $\mathbf{e} \geq 0$ such that one of the matrices $M_1, \cdots, M_k, L$ connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$. Then A is a product of $k+1$ nonsingular M-Matrices thus $A^{-1} \geq 0$.*

Proof. Let $M = M_1 M_2 \cdots M_k$, then M is monotone. By Lemma 5.3.6, we get

$$M^{-1}A \leq L, \tag{5.15}$$

thus

$$(M^{-1}A)_a \leq 0. \tag{5.16}$$

For each M_i, $i = 1, \ldots, k$, by Lemma 5.3.7, we have

$$(M_i)^{-1} \geq ((M_i)_d)^{-1} \Rightarrow M^{-1} \geq (M_k)_d^{-1} \cdots (M_1)_d^{-1}, \tag{5.17}$$

which implies

$$M^{-1}A\mathbf{e} \geq cA\mathbf{e}, \tag{5.18}$$

for some positive number c.

If L connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$, then $M^{-1}A$ also connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$ because (5.15) implies that $(M^{-1}A)_{ij} \neq 0$ whenever $L_{ij} \neq 0$ for any $i \neq j$. By (5.18), $\mathcal{N}^+(A\mathbf{e}) \subset \mathcal{N}^+(M^{-1}A\mathbf{e})$ and $\mathcal{N}^0(M^{-1}A\mathbf{e}) \subset \mathcal{N}^0(A\mathbf{e})$, thus $M^{-1}A$ also connects $\mathcal{N}^0(M^{-1}A\mathbf{e})$ with $\mathcal{N}^+(M^{-1}A\mathbf{e})$. With (5.16), by Theorem 5.3.8, $M^{-1}A$ is a nonsingular M-Matrix thus A is a product of $k+1$ M-Matrices which implies A is monotone.

If M_i connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$ for some $1 \leq i \leq k$. Let $M = M_1 \ldots M_{i-1}$. Similar to (5.17) and (5.18), we get

$$(M)^{-1}A\mathbf{e} \geq c_2 A\mathbf{e}, \quad c_2 > 0, \tag{5.19}$$

which implies that M_i connects $\mathcal{N}^0((M)^{-1}Ae)$ with $\mathcal{N}^+((M)^{-1}A\mathbf{e})$. By Corollary 5.3.9, we know $M_i^{-1}(M)^{-1}A\mathbf{e} > 0$, thus $M^{-1}A\mathbf{e} > 0$. With (5.16), through Theorem 5.3.8 we find $M^{-1}A$ is a M-Matrix thus A is a product of $k+1$ M-Matrices which implies A is monotone. □

Theorem 5.3.11. *If A_a^- has a decomposition: $A_a^- = A^z + A^s = (a_{ij}^z) + (a_{ij}^s)$ with $A^s \leq 0$ and $A^z \leq 0$, such that*

$$A_d + A^z \text{ is a nonsingular } M\text{-Matrix,} \tag{5.20a}$$

$$A_a^+ \leq A^z A_d^{-1} A^s \text{ or equivalently } \forall a_{ij} > 0 \text{ with } i \neq j, a_{ij} \leq \sum_{k=1}^{n} a_{ik}^z a_{kk}^{-1} a_{kj}^s, \tag{5.20b}$$

$$\exists \mathbf{e} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \mathbf{e} \geq 0 \text{ with } A\mathbf{e} \geq 0 \text{ s.t. } A^z \text{ or } A^s \text{ connects } \mathcal{N}^0(A\mathbf{e}) \text{ with } \mathcal{N}^+(A\mathbf{e}). \tag{5.20c}$$

Then A is a product of two nonsingular M-Matrices thus $A^{-1} \geq 0$.

Proof. By (5.20b), we have

$$A = A_d + A^z + A^s + A_a^+ \leq (A_d + A^z)(I + A_d^{-1} A^s). \tag{5.21}$$

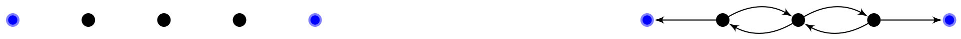
By (5.20c), either $A_d + A^z$ or $I + A_d^{-1} A^s$ connects $\mathcal{N}^0(A\mathbf{e})$ with $\mathcal{N}^+(A\mathbf{e})$. By applying Theorem 5.3.10 for the case $k = 1$, $M_1 = A_d + A^z$ and $L = I + A_d^{-1} A^s$, we get $A^{-1} \geq 0$. □

5.4 The Main Result

For a general matrix, conditions (5.20) in Theorem 5.3.11 can be difficult to verify. We will first derive a simplified version of Theorem 5.3.11 then verify it for the schemes in Section 5.2.

5.4.1 A simplified sufficient condition for monotonicity

We will take advantage of the directed graph described by the 5-point discrete Laplacian, i.e., the second order centered difference scheme, which has similar off-diagonal negative entry patterns as the schemes in Section 5.2.



(a) Grid points.

(b) The directed graph.

Figure 5.3. An illustration of the directed graph described by off-diagonal entries of the matrix in (5.22): the domain $[0, 1]$ is discretized by a uniform 5-point grid; the black points are interior grid points and the blue ones are the boundary grid points. There is a directed path from any interior grid point to at least one of the boundary points.

For the one-dimensional problem $-u = f, x \in (0, 1)$ with $u(0) = u(1)$, the scheme can be written as $u_0 = \sigma_0, u_{n+1} = \sigma_1, \frac{-u_{i-1}+2u_i-u_{i+1}}{h^2} = f_i, i = 1, \cdots, n$. The matrix vector form is $K\bar{\mathbf{u}} = \bar{\mathbf{f}}$ where

$$K = \frac{1}{h^2}\begin{pmatrix} h^2 & & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & & h^2 \end{pmatrix}, \tag{5.22}$$

which described the directed graph illustrated in Figure 5.3. Let $\mathbf{1}$ denote a vector of suitable size with each entry as 1, then $(K\mathbf{1})_i = \begin{cases} 0, & i = 1, \cdots, n \\ 1, & i = 0, n+1 \end{cases}$. By Figure 5.3, it is easy to see that K connects $\mathcal{N}^0(K\mathbf{1})$ with $\mathcal{N}^+(K\mathbf{1})$.

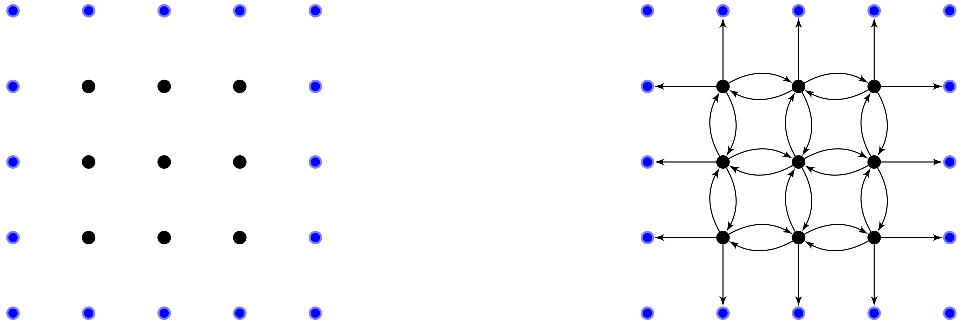
Next we consider the second order accurate 5-point discrete Laplacian scheme for solving $-\Delta u = f$ on $\Omega = (0, 1) \times (0, 1)$ with homogeneous Dirichlet boundary conditions:

$$\begin{aligned} u_{i,j} &= 0, (x_i, y_j) \in \partial\Omega; \\ \frac{-u_{i-1,j} - u_{i+1,j} + 4u_{i,j} - u_{i,j+1} - u_{i+1,j}}{h^2} &= f_{ij}, (x_i, y_j) \in \Omega. \end{aligned}$$

See Figure 5.4 for the directed graph described by its matrix representation. Let K be the matrix representation of the 5-point discrete Laplacian scheme, then

$$(K\mathbf{1})_{i,j} = \begin{cases} 1, & \text{if } (x_i, y_j) \in \partial\Omega, \\ 0, & \text{if } (x_i, y_j) \in \Omega. \end{cases}$$

By Figure 5.4, it is easy to see that K connects $\mathcal{N}^0(K\mathbf{1})$ with $\mathcal{N}^+(K\mathbf{1})$.



(a) Grid points.

(b) The directed graph.

Figure 5.4. An illustration of the directed graph described by off-diagonal entries in the 5-point discrete Laplacian matrix: the domain $[0,1] \times [0,1]$ is discretized by a uniform 5×5 grid; the black points are interior grid points and the blue ones are the boundary grid points. There is a directed path from any interior grid point to at least one of the boundary grid points.

Let $A := \bar{L}_h$ denote the matrix representation of any scheme in Section 5.2. Then

$$(A\mathbf{1})_{i,j} = \begin{cases} 1, & \text{if } (x_i, y_j) \in \partial\Omega, \\ c_{ij} \geq 0, & \text{if } (x_i, y_j) \in \Omega. \end{cases}$$

Therefore, $\mathcal{N}^+(K\mathbf{1}) \subset \mathcal{N}^+(A\mathbf{1})$ implies $\mathcal{N}^0(A\mathbf{1}) \subset \mathcal{N}^0(K\mathbf{1})$, thus K also connects $\mathcal{N}^0(A\mathbf{1})$ with $\mathcal{N}^+(A\mathbf{1})$. Notice that indices of nonzero off-diagonal entries in K is a subset of indices of nonzero entries in A_a^-, thus A_a^- also connects $\mathcal{N}^0(A\mathbf{1})$ with $\mathcal{N}^+(A\mathbf{1})$. So the vector $\mathbf{e}$ can be set as $\mathbf{1}$ in (5.20c). If assuming $c(x,y) > 0$, then $A\mathbf{1} > 0$ thus the condition (5.20c) is trivially satisfied.

By Theorem 5.3.4, for any decomposition of off-diagonal negative entries $A_a^- = A^z + A^s$, $A_d + A^z$ is an M-Matrix if $(A_d + A^z)\mathbf{1} \neq \mathbf{0}$ and $(A_d + A^z)\mathbf{1} \geq 0$. So Theorem 5.3.11 for the schemes (5.10) and (5.12) can be simplified as

Theorem 5.4.1. *Let A denote the matrix representation of the schemes solving $-\nabla\cdot(a\nabla)u + cu = f$ in Section 5.2. Assume A_a^- has a decomposition $A_a^- = A^z + A^s$ with $A^s \leq 0$ and $A^z \leq 0$. Then $A^{-1} \geq 0$ if the following are satisfied:*

1. *$(A_d + A^z)\mathbf{1} \neq \mathbf{0}$ and $(A_d + A^z)\mathbf{1} \geq 0$;*
2. *$A_a^+ \leq A^z A_d^{-1} A^s$;*
3. *For $c(x,y) \geq 0$, either A^z or A^s has the same sparsity pattern as A_a^-. If $c(x,y) > 0$, then this condition can be removed.*

5.4.2 One-dimensional Laplacian case

As a demonstration of how to apply Theorem 5.4.1, we first consider the scheme (5.9). Let A be the matrix representation of the linear operator $\mathcal{L}_h$ in the scheme (5.9). Let $\mathcal{A}_d$ and $\mathcal{A}_a^\pm$ be linear operators corresponding to the matrices A_d and $A_a^\pm$ respectively.

Consider the following decomposition of $\mathcal{A}_a^- = \mathcal{A}^z + \mathcal{A}^s$ with $\mathcal{A}^z = \mathcal{A}^s = \frac{1}{2}\mathcal{A}_a^-$:

$$\begin{aligned}
&\mathcal{A}^z(\bar{\mathbf{u}})_0 = \mathcal{A}^s(\bar{\mathbf{u}})_0 = 0, \quad \mathcal{A}^z(\bar{\mathbf{u}})_{n+1} = \mathcal{A}^s(\bar{\mathbf{u}})_{n+1} = 0,\\
&\mathcal{A}^z(\bar{\mathbf{u}})_i = \mathcal{A}^s(\bar{\mathbf{u}})_i = \frac{-u_{i-1} - u_{i+1}}{2h^2}, \quad \text{if } x_i \text{ is a cell center,}\\
&\mathcal{A}^z(\bar{\mathbf{u}})_i = \mathcal{A}^s(\bar{\mathbf{u}})_i = \frac{-8u_{i-1} - 8u_{i+1}}{8h^2}, \quad \text{if } x_i \text{ is an interior cell end.}
\end{aligned}$$

The operator $\mathcal{A}_d$ and $\mathcal{A}_a^+$ are given as:

$$\begin{aligned}
&\mathcal{A}_d(\bar{\mathbf{u}})_0 = u_0, \quad \mathcal{A}_d(\bar{\mathbf{u}})_{n+1} = u_{n+1},\\
&\mathcal{A}_d(\bar{\mathbf{u}})_i = \frac{2u_i}{h^2}, \quad \text{if } x_i \text{ is a cell center,}\\
&\mathcal{A}_d(\bar{\mathbf{u}})_i = \frac{14u_i}{4h^2}, \quad \text{if } x_i \text{ is an interior cell end.}
\end{aligned}$$

$$\begin{aligned}
&\mathcal{A}_a^+(\bar{\mathbf{u}})_0 = 0, \quad \mathcal{A}_a^+(\bar{\mathbf{u}})_{n+1} = 0,\\
&\mathcal{A}_a^+(\bar{\mathbf{u}})_i = 0, \quad \text{if } x_i \text{ is a cell center,}\\
&\mathcal{A}_a^+(\bar{\mathbf{u}})_i = \frac{u_{i-2} + u_{i+2}}{4h^2}, \quad \text{if } x_i \text{ is an interior cell end.}
\end{aligned}$$

Obviously, A^z and A^s both have have the same sparsity pattern as A_a^-. It is straightforward to verify $[\mathcal{A}^d + \mathcal{A}^z](\mathbf{1})$ is a non-negative nonzero vector. So we only need to verify $A_a^+ \leq A^z A_d^{-1} A^s$ to apply Theorem 5.4.1. Since $A^z A_d^{-1} A^s \geq 0$, we only need to compare nonzero coefficients in $\mathcal{A}_a^+(\bar{\mathbf{u}})_i$ and $\mathcal{A}^z\left(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]\right)_i$.

When x_i is an interior cell end, $x_{i\pm1}$ are cell centers, and we have

$$\mathcal{A}^s(\bar{\mathbf{u}})_{i-1} = \frac{-u_{i-2} - u_i}{2h^2}, \quad \mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i-1} = \frac{h^2 \mathcal{A}^s(\bar{\mathbf{u}})_{i-2}}{2},$$

$$\mathcal{A}^z(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})])_i = \frac{-\mathcal{A}_d^{-1}[-\mathcal{A}^s(\bar{\mathbf{u}})]_{i-1} - \mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i+1}}{h^2} = \frac{u_{i-2} + 2u_i + u_{i+2}}{4h^2}.$$

We can verify $A_a^+ \leq A^z A_d^{-1} A^s$ by comparing only the coefficients of $u_{i\pm2}$ in $\mathcal{A}_a^+(\bar{\mathbf{u}})_i$ and $\mathcal{A}^z\left(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]\right)_i$ because $A^z A_d^{-1} A^s \geq 0$. By Theorem 5.4.1, we get $A^{-1} \geq 0$.

5.4.3 One-dimensional variable coefficient case

As we have seen in the previous discussion, all the operators are either zero or identity at the boundary points thus do not affect the discussion verifying the condition (5.20b). For the sake of simplicity, we only consider the interior grid points for the linear operators. With the positive and negative parts for a number f defined as:

$$f^+ = \frac{|f| + f}{2}, \quad f^- = \frac{|f| - f}{2},$$

the linear operators $\mathcal{A}_d$, $\mathcal{A}_a^\pm$ are

$$\text{if } x_i \text{ is a cell center,} \quad \mathcal{A}_d(\bar{\mathbf{u}})_i = \left(\frac{a_{i-1} + a_{i+1}}{h^2} + c_i\right) u_i;$$

$$\text{if } x_i \text{ is an interior cell end,}$$

$$\mathcal{A}_d(\bar{\mathbf{u}})_i = \left(\frac{a_{i-2} + 4a_{i-1} + 18a_i + 4a_{i+1} + a_{i+2}}{8h^2} + c_i\right) u_i.$$

$$\text{if } x_i \text{ is a cell center,} \quad \mathcal{A}_a^+(\bar{\mathbf{u}})_i = 0;$$

$$\text{if } x_i \text{ is an interior cell end,}$$

$$\mathcal{A}_a^+(\bar{\mathbf{u}})_i = \frac{(3a_{i-2} - 4a_{i-1} + 3a_i)^+ u_{i-2} + (3a_{i+2} - 4a_{i+1} + 3a_i)^+ u_{i+2}}{8h^2}.$$

$$\text{If } x_i \text{ is a cell center,} \quad \mathcal{A}_a^-(\bar{\mathbf{u}})_i = \frac{-(3a_{i-1} + a_{i+1})u_{i-1} - (a_{i-1} + 3a_{i+1})u_{i+1}}{4h^2};$$

$$\text{If } x_i \text{ is an interior cell end,} \quad \mathcal{A}_a^-(\bar{\mathbf{u}})_i = \frac{-(3a_{i-2} - 4a_{i-1} + 3a_i)^- u_{i-2}}{8h^2}$$

$$+ \frac{-(4a_{i-2} + 12a_i)u_{i-1} - (12a_i + 4a_{i+2})u_{i+1} - (3a_i - 4a_{i+1} + 3a_{i+2})^- u_{i+2}}{8h^2}.$$

We can easily verify that $(A_d + A^z)\mathbf{1} \geq 0$ for the following $\mathcal{A}^z$:

$$\text{if } x_i \text{ is a cell center,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_i = \epsilon \frac{-(3a_{i-1} + a_{i+1})u_{i-1} - (a_{i-1} + 3a_{i+1})u_{i+1}}{4h^2},$$

$$\text{if } x_i \text{ is an interior cell end,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_i =$$

$$\frac{-(3a_{i-2} - 4a_{i-1} + 3a_i)^- u_{i-2} - [4a_{i-2} + 12a_i - (3a_{i-2} - 4a_{i-1} + 3a_i)^+]u_{i-1}}{8h^2}$$

$$+ \frac{-[12a_i + 4a_{i+2} - (3a_i - 4a_{i+1} + 3a_{i+2})^+]u_{i+1} - (3a_i - 4a_{i+1} + 3a_{i+2})^- u_{i+2}}{8h^2},$$

where $\epsilon > 0$ is a small number. Moreover, A^z has the same sparsity pattern as A_a^- for any $\epsilon > 0$. For $\epsilon < 1$ we can verify that $A^s = A_a^- - A^z \leq 0$:

If x_i is a cell center, $\quad \mathcal{A}^s(\bar{\mathbf{u}})_i = (1-\epsilon)\dfrac{-(3a_{i-1}+a_{i+1})u_{i-1} - (a_{i-1}+3a_{i+1})u_{i+1}}{4h^2}$,

If x_i is an interior cell end,

$$\mathcal{A}^s(\bar{\mathbf{u}})_i = \frac{-(3a_{i-2}-4a_{i-1}+3a_i)^+u_{i-1} - (3a_i - 4a_{i+1}+3a_{i+2})^+u_{i+1}}{8h^2}.$$

Now we only need to compare nonzero coefficients in $\mathcal{A}_a^+(\bar{\mathbf{u}})_i$ and $\mathcal{A}^z\left(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]\right)_i$ for x_i being an interior cell end. When x_i is an interior cell end, $x_{i\pm1}$ are cell centers, and we have

$$\mathcal{A}^s(\bar{\mathbf{u}})_{i-1} = (1-\epsilon)\frac{-(3a_{i-2}+a_i)u_{i-2} - (a_{i-2}+3a_i)u_i}{4h^2},$$

$$A^s(\bar{\mathbf{u}})_{i-2} = \frac{-(3a_{i-4}-4a_{i-3}+3a_{i-2})^+u_{i-3} - (3a_{i-2}-4a_{i-1}+3a_i)^+u_{i-1}}{8h^2},$$

$$\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i-1} = \frac{h^2\mathcal{A}^s(\bar{\mathbf{u}})_{i-1}}{(a_{i-2}+a_i+h^2c_{i-1})} = (1-\epsilon)\frac{-(3a_{i-2}+a_i)u_{i-2} - (a_{i-2}+3a_i)u_i}{4(a_{i-2}+a_i+h^2c_{i-1})}.$$

It suffices to focus on the coefficient of u_{i-2} in $\mathcal{A}^z(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})])_i$ and the discussion for the coefficient of u_{i+2} is similar. Notice that $\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i-2}$ will contribute nothing to the coefficient of u_{i-2}. So the coefficient of u_{i-2} in $\mathcal{A}^z(\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})])_i$ is

$$(1-\epsilon)\frac{(3a_{i-2}+a_i)(4a_{i-2}+12a_i-(3a_{i-2}-4a_{i-1}+3a_i)^+)}{32h^2(a_{i-2}+a_i+h^2c_{i-1})}.$$

Thus to ensure $A_a^+ \leq A^zA_d^-A^s$, it suffices to have the following holds for any interior cell end x_i:

$$(1-\epsilon)\frac{(3a_{i-2}+a_i)(4a_{i-2}+12a_i-(3a_{i-2}-4a_{i-1}+3a_i)^+)}{32h^2(a_{i-2}+a_i+h^2c_{i-1})} \geq \frac{(3a_{i-2}-4a_{i-1}+3a_i)^+}{8h^2}.$$

Equivalently, we need the following inequality holds for any cell center x_i:

$$(1-\epsilon)\frac{(3a_{i-1}+a_{i+1})(4a_{i-1}+12a_{i+1}-(3a_{i-1}-4a_i+3a_{i+1})^+)}{32h^2(a_{i-1}+a_{i+1}+h^2c_i)} \geq \frac{(3a_{i-1}-4a_i+3a_{i+1})^+}{8h^2}. \tag{5.23}$$

Notice that ϵ can be any fixed number in $[0, 1)$ so that $A_d + A^z$ is an M-Matrix and $A^s \leq 0$. And ϵ must be strictly positive so that A^z has the same sparsity pattern as A_a^-. Thus if there is one fixed $\epsilon \in (0, 1)$ so that (5.23) holds for any cell center x_i, then by Theorem 5.4.1, $A^{-1} \geq 0$. A sufficient condition for (5.23) to hold for any cell center x_i with some fixed $\epsilon \in (0, 1)$ is to have the following inequality for any cell center x_i:

$$\frac{(3a_{i-1} + a_{i+1})(4a_{i-1} + 12a_{i+1} - (3a_{i-1} - 4a_i + 3a_{i+1})^+)}{32h^2(a_{i-1} + a_{i+1} + h^2 c_i)} > \frac{(3a_{i-1} - 4a_i + 3a_{i+1})^+}{8h^2}. \tag{5.24}$$

If $3a_{i-1} - 4a_i + 3a_{i+1} \leq 0$, then (5.24) holds trivially. We only need to discuss the case $3a_{i-1} - 4a_i + 3a_{i+1} > 0$, for which (5.24) becomes

$$(3a_{i-1} + a_{i+1})(a_{i-1} + 4a_i + 9a_{i+1}) > 4(a_{i-1} + a_{i+1} + h^2 c_i)(3a_{i-1} - 4a_i + 3a_{i+1}). \tag{5.25}$$

So we have proven the first result for the variable coefficient case:

Theorem 5.4.2. *For the scheme (5.10) solving $-(au) + cu = f$ with $a(x) > 0$ and $c(x) \geq 0$, its matrix representation $A = \bar{L}_h$ satisfies $A^{-1} \geq 0$ if (5.25) holds for any cell center x_i.*

The constraint (5.25) will be satisfied for small enough h. The proof of the following two theorems are included in the Appendix 5.6.

Theorem 5.4.3. *For the scheme (5.10) solving $-(au) + cu = f$ with $a(x) > 0$ and $c(x) \geq 0$ on a uniform mesh, its matrix representation $A = \bar{L}_h$ satisfies $A^{-1} \geq 0$ if any of the following constraints is satisfied for each finite element cell $I_i = [x_{i-1}, x_{i+1}]$:*

- *There exists some $\lambda \in (\frac{3}{13}, 1)$ such that*

$$h^2 c_i < \frac{13(1-\lambda)\min\limits_{I_i} a^2(x)}{6\max\limits_{I_i} a(x) - 4\min\limits_{I_i} a(x)}, \qquad h\frac{\max\limits_{x\in I_i} |a(x)|}{\min\limits_{x\in I_i} a(x)} < \frac{\sqrt{39\lambda} - 3}{6}.$$

- $2h\max\limits_{I_i} |a(x)| + h^2 c_i \left(1 - \frac{2}{3}\frac{\min_{I_i} a(x)}{\max_{I_i} a(x)}\right) < \frac{5}{3}\frac{\min_{I_i} a^2(x)}{\max_{I_i} a(x)}$.

- *If $c(x) \equiv 0$, then we only need $h\frac{\max_{x\in I_i}|a(x)|}{\min_{x\in I_i} a(x)} < \frac{\sqrt{39}-3}{6}$.*

- *If $a(x) \equiv a > 0$, then we only need $h^2 c_i < 5a$.*

Theorem 5.4.4. *For the scheme (5.10) solving $-(au)+cu = f$ with $a(x) > 0$ and $c(x) \geq 0$, its matrix representation $A = \bar{L}_h$ satisfies $A^{-1} \geq 0$ if the following mesh constraint is achieved for all cell centers x_i:*

$$h^2 \left(\frac{3}{2} c_i + \max_{x \in (x_{i-1}, x_{i+1})} a(x) \right) < \frac{74}{45} \min\{a_{i-1}, a_i, a_{i+1}\}. \tag{5.26a}$$

If $a(x)$ is a concave function, then (5.26a) can be replaced by

$$h^2 c_i < 3 \min\{a_{i-1}, a_i, a_{i+1}\}. \tag{5.26b}$$

Remark 5.4.5. For solving heat equation with backward Euler time discretization (5.4), the mesh constraints in Theorem 5.4.3 and Theorem 5.4.4 imply that a lower bound for $\frac{\Delta t}{h^2}$ is a sufficient condition for ensuring monotonicity. Numerical tests suggest that a lower bound on $\frac{\Delta t}{h^2}$ is also a necessary condition, see Section 5.5. A lower bound constraint on the time step is common for high order accurate spatial discretizations with backward Euler to satisfy monotonicity, e.g., [88].

5.4.4 Two-dimensional variable coefficient case

Next we apply Theorem 5.4.1 to the scheme (5.12). The splitting $A_a^- = A^z + A^s$ is quite similar to one-dimensional case due to its stencil pattern.

Let $A := \bar{L}_h$ be the matrix representation of the linear operator $\mathcal{L}_h$ in the scheme (5.12). We only consider interior grid points since $\mathcal{L}_h$ is identity operator on boundary points which do not affect applying Theorem 5.4.1. We first have

$$\text{if } x_{ij} \text{ is a cell center,} \quad \mathcal{A}_d(\bar{\mathbf{u}})_{ij} = \left(\frac{a_{i-1,j} + a_{i+1,j} + a_{i,j-1} + a_{i,j+1}}{h^2} + c_{ij}\right) u_{ij};$$

if x_{ij} is an edge center for an edge parallel to y-axis,

$$\mathcal{A}_d(\bar{\mathbf{u}})_{ij} = \left(\frac{(a_{i-2,j} + 4a_{i-1,j} + 18a_{ij} + 4a_{i+1,j} + a_{i+2,j}) + 8(a_{i,j-1} + a_{i,j+1})}{8h^2} + c_{ij}\right) u_{ij};$$

if x_{ij} is an edge center for an edge parallel to x-axis,

$$\mathcal{A}_d(\bar{\mathbf{u}})_{ij} = \left(\frac{(a_{i,j-2} + 4a_{i,j-1} + 18a_{ij} + 4a_{i,j+1} + a_{i,j+2}) + 8(a_{i-1,j} + a_{i+1,j})}{8h^2} + c_{ij}\right) u_{ij};$$

if x_{ij} is a knot,

$$\mathcal{A}_d(\bar{\mathbf{u}})_{ij} = \left(\frac{a_{i-2,j} + 4a_{i-1,j} + 18a_{ij} + 4a_{i+1,j} + a_{i+2,j}}{8h^2}\right.$$
$$\left. + \frac{(a_{i,j-2} + 4a_{i,j-1} + 18a_{ij} + 4a_{i,j+1} + a_{i,j+2})}{8h^2} + c_{ij}\right) u_{ij}.$$

For the operator $\mathcal{A}_a^+$, it is given as

$$\text{if } x_{ij} \text{ is a cell center,} \quad \mathcal{A}_a^+(\bar{\mathbf{u}})_{ij} = 0;$$

if x_{ij} is an edge center for an edge parallel to y-axis,

$$\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij} = \frac{(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^+ u_{i-2,j} + (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^+ u_{i+2,j}}{8h^2};$$

if x_{ij} is an edge center for an edge parallel to x-axis,

$$\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij} = \frac{(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^+ u_{i,j-2} + (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^+ u_{i,j+2}}{8h^2};$$

if x_{ij} is a knot, $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij} =$

$$\frac{(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^+ u_{i-2,j} + (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^+ u_{i+2,j}}{8h^2}$$
$$+ \frac{(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^+ u_{i,j-2} + (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^+ u_{i,j+2}}{8h^2}.$$

Let $\epsilon \in (0,1)$ be a fixed number. We consider the following $A^z \leq 0$ so that $(A_d + A^z)\mathbf{1} \geq 0$:

$$\text{if } x_{ij} \text{ is a cell center,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_{ij} = -\epsilon\frac{(3a_{i-1,j} + a_{i+1,j})u_{i-1,j}}{4h^2}$$
$$-\,\epsilon\frac{(a_{i-1,j} + 3a_{i+1,j})u_{i+1,j} + (3a_{i,j-1} + a_{i,j+1})u_{i,j-1} + (a_{i,j-1} + 3a_{i,j+1})u_{i,j+1}}{4h^2};$$

$$\text{if } x_{ij} \text{ is an edge center for an edge parallel to } y\text{-axis,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_{ij} =$$
$$\frac{-(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^- u_{i-2,j} - [4a_{i-2,j} + 12a_{i,j} - (3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^+]u_{i-1,j}}{8h^2}$$
$$+\,\frac{-[12a_{i,j} + 4a_{i+2,j} - (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^+]u_{i+1,j} - (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^- u_{i+2,j}}{8h^2}$$
$$+\,\epsilon\frac{-(3a_{i,j-1} + a_{i,j+1})u_{i,j-1} - (a_{i,j-1} + 3a_{i,j+1})u_{i,j+1}}{4h^2};$$

$$\text{if } x_{ij} \text{ is an edge center for an edge parallel to } x\text{-axis,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_{ij} =$$
$$\frac{-(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^- u_{i,j-2} - [4a_{i,j-2} + 12a_{i,j} - (3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^+]u_{i,j-1}}{8h^2}$$
$$+\,\frac{-[12a_{i,j} + 4a_{i,j+2} - (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^+]u_{i,j+1} - (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^- u_{i,j+2}}{8h^2}$$
$$+\,\epsilon\frac{-(3a_{i-1,j} + a_{i+1,j})u_{i-1,j} - (a_{i-1,j} + 3a_{i+1,j})u_{i+1,j}}{4h^2};$$

$$\text{if } x_{ij} \text{ is a knot,} \quad \mathcal{A}^z(\bar{\mathbf{u}})_{ij} =$$
$$frac{-}(3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^- u_{i-2,j} - [4a_{i-2,j} + 12a_{i,j} - (3a_{i-2,j} - 4a_{i-1,j} + 3a_{i,j})^+]u_{i-1,j}8h^2$$
$$+\,\frac{-[12a_{i,j} + 4a_{i+2,j} - (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^+]u_{i+1,j} - (3a_{i+2,j} - 4a_{i+1,j} + 3a_{i,j})^- u_{i+2,j}}{8h^2}$$
$$+\,\frac{-(3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^- u_{i,j-2} - [4a_{i,j-2} + 12a_{i,j} - (3a_{i,j-2} - 4a_{i,j-1} + 3a_{i,j})^+]u_{i,j-1}}{8h^2}$$
$$+\,\frac{-[12a_{i,j} + 4a_{i,j+2} - (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^+]u_{i,j+1} - (3a_{i,j+2} - 4a_{i,j+1} + 3a_{i,j})^- u_{i,j+2}}{8h^2};$$

Then $A^s = A_a^- - A^z$ is given as:

$$\begin{aligned}
&\text{if } x_i \text{ is a cell center,} \quad \mathcal{A}^s(\bar{\mathbf{u}})_{ij} = \\
&-(1-\epsilon)\frac{(3a_{i-1,j}+a_{i+1,j})u_{i-1,j}+(a_{i-1,j}+3a_{i+1,j})u_{i+1,j}}{4h^2} \\
&-(1-\epsilon)\frac{(3a_{i,j-1}+a_{i,j+1})u_{i,j-1}+(a_{i,j-1}+3a_{i,j+1})u_{i,j+1}}{4h^2}; \\
&\text{if } x_{ij} \text{ is an edge center for an edge parallel to } y\text{-axis,} \quad \mathcal{A}^s(\bar{\mathbf{u}})_{ij} = \\
&\frac{-(3a_{i-2,j}-4a_{i-1,j}+3a_{i,j})^+u_{i-1,j}-(3a_{i+2,j}-4a_{i+1,j}+3a_{i,j})^+u_{i+1,j}}{8h^2} \\
&+(1-\epsilon)\frac{-(3a_{i,j-1}+a_{i,j+1})u_{i,j-1}-(a_{i,j-1}+3a_{i,j+1})u_{i,j+1}}{4h^2};
\end{aligned}$$

$$\begin{aligned}
&\text{if } x_{ij} \text{ is an edge center for an edge parallel to } x\text{-axis,} \quad \mathcal{A}^s(\bar{\mathbf{u}})_{ij} = \\
&\frac{-(3a_{i,j-2}-4a_{i,j-1}+3a_{i,j})^+u_{i,j-1}-(3a_{i,j+2}-4a_{i,j+1}+3a_{i,j})^+u_{i,j+1}}{8h^2} \\
&+(1-\epsilon)\frac{-(3a_{i-1,j}+a_{i+1,j})u_{i-1,j}-(a_{i-1,j}+3a_{i+1,j})u_{i+1,j}}{4h^2}; \\
&\text{if } x_{ij} \text{ is a knot,} \quad \mathcal{A}^s(\bar{\mathbf{u}})_{ij} = \\
&\frac{-(3a_{i-2,j}-4a_{i-1,j}+3a_{i,j})^+u_{i-1,j}-(3a_{i+2,j}-4a_{i+1,j}+3a_{i,j})^+u_{i+1,j}}{8h^2} \\
&+\frac{-(3a_{i,j-2}-4a_{i,j-1}+3a_{i,j})^+u_{i,j-1}-(3a_{i,j+2}-4a_{i,j+1}+3a_{i,j})^+u_{i,j+1}}{8h^2};
\end{aligned}$$

For the positive off-diagonal entries, $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$ is nonzero only for x_{ij} being an edge center or a cell center. Thus to verify $A_a^+ \leq A^z A_d^{-1} A^s$, it suffices to compare $\mathcal{A}^z \left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ with $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$ for x_{ij} being an edge center or a cell center.

If x_{ij} is an edge center for an edge parallel to y-axis, then $x_{i\pm1,j}$ are cell centers. Since everything here has a symmetric structure, we only need to compare the coefficients of $u_{i-2,j}$

in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ and $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$, and the comparison for the coefficients of $u_{i+2,j}$ will be similar.

$$\begin{aligned}\mathcal{A}^s(\bar{\mathbf{u}})_{i-1,j} &= -(1-\epsilon)\frac{(3a_{i-2,j}+a_{ij})u_{i-2,j}+(a_{i-2,j}+3a_{i,j})u_{i,j}}{4h^2}\\ -(1-\epsilon)&\frac{(3a_{i-1,j-1}+a_{i-1,j+1})u_{i-1,j-1}+(a_{i-1,j-1}+3a_{i-1,j+1})u_{i-1,j+1}}{4h^2},\\ \mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i-1,j} &= -(1-\epsilon)\frac{(3a_{i-2,j}+a_{ij})u_{i-2,j}+(a_{i-2,j}+3a_{ij})u_{i,j}}{4(a_{i-2,j}+a_{ij}+a_{i-1,j+1}+a_{i-1,j-1}+h^2c_{i-1,j})}\\ -(1-\epsilon)&\frac{(3a_{i-1,j-1}+a_{i-1,j+1})u_{i-1,j-1}+(a_{i-1,j-1}+3a_{i-1,j+1})u_{i-1,j+1}}{4(a_{i-2,j}+a_{ij}+a_{i-1,j+1}+a_{i-1,j-1}+h^2c_{i-1,j})}.\end{aligned}$$

Since the coefficient of $u_{i-2,j}$ in $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$ is $(3a_{i-2,j}-4a_{i-1,j}+3a_{ij})^+/(8h^2)$, we only need to discuss the case $3a_{i-2,j}-4a_{i-1,j}+3a_{ij}>0$, for which the coefficient of $u_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ becomes

$$\frac{a_{i-2,j}+4a_{i-1,j}+9a_{ij}}{8h^2}\frac{(1-\epsilon)(3a_{i-2,j}+a_{ij})}{4(a_{i-2,j}+a_{ij}+a_{i-1,j+1}+a_{i-1,j-1}+h^2c_{i-1,j})}.$$

To ensure the coefficient of $u_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ is no less than the coefficient of $u_{i-2,j}$ in $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$, we need

$$\frac{(1-\epsilon)(a_{i-2,j}+4a_{i-1,j}+9a_{ij})(3a_{i-2,j}+a_{ij})}{32h^2(a_{i-2,j}+a_{ij}+a_{i-1,j+1}+a_{i-1,j-1}+h^2c_{i-1,j})}\geq\frac{3a_{i-2,j}-4a_{i-1,j}+3a_{ij}}{8h^2}.$$

Similar to the one-dimensional case, it suffices to require

$$\frac{(a_{i-2,j}+4a_{i-1,j}+9a_{ij})(3a_{i-2,j}+a_{ij})}{4(a_{i-2,j}+a_{ij}+a_{i-1,j+1}+a_{i-1,j-1}+h^2c_{i-1,j})}>3a_{i-2,j}-4a_{i-1,j}+3a_{ij}.$$

Equivalently, we need the following inequality holds for any cell center x_{ij}:

$$\frac{(a_{i-1,j}+4a_{i,j}+9a_{i+1,j})(3a_{i-1,j}+a_{i+1,j})}{4(a_{i-1,j}+a_{i+1,j}+a_{i,j+1}+a_{i,j-1}+h^2c_{i,j})}>3a_{i-1,j}-4a_{i,j}+3a_{i+1,j}. \tag{5.27a}$$

Notice that (5.27a) was derived for comparing $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ and $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$ for x_{ij} being an edge center of an edge parallel to y-axis. If x_{ij} is an edge center of an edge parallel to x-axis, then we can derive a similar constraint:

$$\frac{(a_{i,j-1}+4a_{i,j}+9a_{i,j+1})(3a_{i,j-1}+a_{i,j+1})}{4(a_{i,j-1}+a_{i,j+1}+a_{i+1,j}+a_{i-1,j}+h^2c_{i,j})} > 3a_{i,j-1}-4a_{i,j}+3a_{i,j+1}. \tag{5.27b}$$

If x_{ij} is a knot, then $x_{i\pm1,j}$ are edge centers for an edge parallel to x-axis. Since everything here has a symmetric structure, we only need to compare the coefficients of $u_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ and $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$, and the comparison for the coefficients of $u_{i+2,j}$, $u_{i,j-2}$ and $u_{i,j+2}$ will be similar.

$$\begin{aligned}
&\mathcal{A}^s(\bar{\mathbf{u}})_{i-1,j} = (1-\epsilon)\frac{-(3a_{i-2,j}+a_{i,j})u_{i-2,j}-(a_{i-2,j}+3a_{i,j})u_{i,j}}{4h^2}\\
&+\frac{-(3a_{i-1,j-2}-4a_{i-1,j-1}+3a_{i-1,j})^+u_{i-1,j-1}-(3a_{i-1,j+2}-4a_{i-1,j+1}+3a_{i-1,j})^+u_{i-1,j+1}}{8h^2}\\
&\mathcal{A}_d^{-1}[\mathcal{A}^s(\bar{\mathbf{u}})]_{i-1,j} =\\
&(1-\epsilon)\frac{-(3a_{i-2,j}+a_{i,j})u_{i-2,j}-(a_{i-2,j}+3a_{i,j})u_{i,j}}{\frac{1}{2}(a_{i-1,j-2}+4a_{i-1,j-1}+18a_{i-1,j}+4a_{i-1,j+1}+a_{i-1,j+2})+4(a_{i-2,j}+a_{i,j})+4h^2c_{i-1,j}}\\
&+\frac{-(3a_{i-1,j-2}-4a_{i-1,j-1}+3a_{i-1,j})^+u_{i-1,j-1}-(3a_{i-1,j+2}-4a_{i-1,j+1}+3a_{i-1,j})^+u_{i-1,j+1}}{(a_{i-1,j-2}+4a_{i-1,j-1}+18a_{i-1,j}+4a_{i-1,j+1}+a_{i-1,j+2})+8(a_{i-2,j}+a_{i,j})+8h^2c_{i-1,j}}.
\end{aligned}$$

For the same reason as above we still only consider the case where $3a_{i-2,j}-4a_{i-1,j}+3a_{ij}>0$. So the coefficient of $u_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ is

$$\frac{1}{4h^2}\frac{(1-\epsilon)(a_{i-2,j}+4a_{i-1,j}+9a_{ij})(3a_{i-2,j}+a_{i,j})}{(a_{i-1,j-2}+4a_{i-1,j-1}+18a_{i-1,j}+4a_{i-1,j+1}+a_{i-1,j+2})+8(a_{i-2,j}+a_{i,j})+8c_{i-1,j}h^2}.$$

To ensure the coefficient of $u_{i-2,j}$ in $\mathcal{A}^z\left[\mathcal{A}_d^{-1}\left(\mathcal{A}^s(\bar{\mathbf{u}})\right)\right]_{ij}$ is no less than the coefficient of $u_{i-2,j}$ in $\mathcal{A}_a^+(\bar{\mathbf{u}})_{ij}$, we only need

$$\frac{2(a_{i-2,j}+4a_{i-1,j}+9a_{ij})(3a_{i-2,j}+a_{i,j})}{(a_{i-1,j-2}+4a_{i-1,j-1}+18a_{i-1,j}+4a_{i-1,j+1}+a_{i-1,j+2})+8(a_{i-2,j}+a_{i,j})+8c_{i-1,j}h^2}$$

$$> 3a_{i-2,j}-4a_{i-1,j}+3a_{ij}.$$

Equivalently, we need the following inequality holds for any edge center x_{ij} for an edge parallel to x-axis:

$$\frac{2(a_{i-1,j}+4a_{i,j}+9a_{i+1,j})(3a_{i-1,j}+a_{i+1,j})}{(a_{i,j-2}+4a_{i,j-1}+18a_{i,j}+4a_{i,j+1}+a_{i,j+2})+8(a_{i-1,j}+a_{i+1,j})+8c_{i,j}h^2} > 3a_{i-1,j}-4a_{i,j}+3a_{i+1,j}. \tag{5.28a}$$

We also need the following inequality holds for any edge center x_{ij} for an edge parallel to y-axis:

$$\frac{2(a_{i,j-1}+4a_{i,j}+9a_{i,j+1})(3a_{i,j-1}+a_{i,j-1})}{(a_{i-2,j}+4a_{i-1,j}+18a_{i,j}+4a_{i+1,j}+a_{i+2,j})+8(a_{i,j-1}+a_{i,j+1})+8c_{i,j}h^2} > 3a_{i,j-1}-4a_{i,j}+3a_{i,j+1}. \tag{5.28b}$$

We have similar result to the one-dimensional case as following:

Theorem 5.4.6. *For the scheme (5.12) solving $-\nabla\cdot(a\nabla u)+cu=f$ with $a(x)>0$ and $c(x)\geq 0$, its matrix representation $A=\bar{L}_h$ satisfies $A^{-1}\geq 0$ if (5.27) holds for any cell center x_{ij}, (5.28a) holds for x_{ij} being any edge center of an edge parallel to x-axis and (5.28b) holds for x_{ij} being any edge center of an edge parallel to y-axis.*

The constraints (5.27), (5.28a) and (5.28b) can be satisfied for small h.

Theorem 5.4.7. *For the scheme (5.12) solving $-\nabla(a(x)\nabla u)+cu=f$ with $a(x)>0$ and $c(x)\geq 0$, its matrix representation $A=\bar{L}_h$ satisfies $A^{-1}\geq 0$ if the following mesh constraint is achieved for all edge centers x_{ij}:*

$$\min_{J_{ij}} a(x)^2 > \frac{49}{61}\max_{J_{ij}} a(x)^2 + \frac{8}{61}\left(3\max_{J_{ij}} a(x) - 2\min_{J_{ij}} a(x)\right)h^2 c_{ij},$$

where J_{ij} is the union of two finite element cells: if x_{ij} is an edge center of an edge parallel to x-axis, then $J_{ij}=[x_{i-1},x_{i+1}]\times[y_{j-2},y_{j+2}]$; if x_{ij} is an edge center of an edge parallel to y-axis, then $J_{ij}=[x_{i-2},x_{i+2}]\times[y_{j-1},y_{j+1}]$.

Theorem 5.4.8. *For the scheme (5.12) solving $-\nabla\cdot(a\nabla u)+cu=f$ with $a(x)>0$ and $c(x)\geq 0$ on a uniform mesh, its matrix representation $A=\bar{L}_h$ satisfies $A^{-1}\geq 0$ if any of the following mesh constraints is satisfied for any edge center x_{ij}:*

- *There exists some $\lambda\in(\frac{49}{61},1)$ such that*

$$h^2c_{ij}<\frac{61(1-\lambda)\min\limits_{J_{ij}}a^2(x)}{8\left(3\max\limits_{J_{ij}}a(x)-2\min\limits_{J_{ij}}a(x)\right)},\qquad h\frac{\max\limits_{x\in Jij}|\nabla a(x)|}{\min\limits_{x\in J_{ij}}a(x)}<\frac{\sqrt{122\lambda}-7\sqrt{2}}{28}.$$

- $\frac{49\sqrt{2}}{3}h\max\limits_{J_{ij}}|\nabla a(x)|+2h^2c_{ij}\left(1-\frac{2}{3}\frac{\min_{J_{ij}}a(x)}{\max_{J_{ij}}a(x)}\right)<\frac{\min_{J_{ij}}a^2(x)}{\max_{J_{ij}}a(x)}$.

- *If $c(x)\equiv 0$, then we only need $h\frac{\max_{x\in J_{ij}}|\nabla a(x)|}{\min_{x\in J_{ij}}a(x)}<\frac{\sqrt{122}-7\sqrt{2}}{28}$.*

- *If $a(x)\equiv a>0$, then we only need $h^2c_{ij}<\frac{3}{2}a$.*

Here the definition of J_{ij} is the same as in Theorem 5.4.7.

The proof of Theorem 5.4.7 is included in the Appendix 5.6. The proof of Theorem 5.4.8 is very similar to the proof of Theorem 5.4.3 thus omitted. Since the two-dimensional case is more complicated, it does not seem possible to derive a similar mesh constraint involving second order derivatives of $a(x,y)$ as in Theorem 5.4.4. For instance, by Theorem 5.4.4, if $a(x)>0$ is concave and $c(x)\equiv 0$, then the one-dimensional scheme (5.10) satisfies $\bar{L}_h^{-1}\geq 0$ without any mesh constraint. For the two-dimensional scheme (5.12), even if assuming $a(x,y)>0$ is concave and $c(x,y)\equiv 0$, constraints (5.27), (5.28a) and (5.28b) are not all satisfied for any h.

5.5 Numerical Tests

In this section we show some numerical tests of scheme (5.12) on an uniform rectangular mesh and verify the inverse non-negativity of $\mathcal{L}_h$. See chapter 2 for numerical tests on the fourth order accuracy of this scheme. In order to minimize round-off errors, we redefine (5.12a) to its equivalent expression $\mathcal{L}_h(\bar{\mathbf{u}})_{i,j}=\frac{1}{h^2}u_{i,j}=\frac{1}{h^2}g_{i,j}$ so that all nonzero entries in

$\bar{L}_h$ have similar magnitudes. By Theorem 5.3.3, we have $L_h^{-1} \geq 0$ whenever $\bar{L}_h^{-1} \geq 0$. Even though $L_h^{-1} \geq 0$ is not sufficient to ensure the discrete maximum principle, in practice only L_h^{-1} is used directly thus its positivity is also important.

We first consider the following equation with purely Dirichlet conditions:

$$-\nabla \cdot (a\nabla u) + cu = f \quad \text{on } [0,1] \times [0,2] \tag{5.29}$$

where $c(x) \equiv 10$ and $a(x,y) = 1 + d\cos(\pi x)\cos(\pi y)$ with $d = 0.5, 0.9$, and 0.99. The smallest entries in L_h^{-1} and $\bar{L}_h^{-1}$ are listed in Table 5.1, in which -10^{-18} should be regarded as the numerical zero. As we can see, $L_h^{-1} \geq 0$ and $\bar{L}_h^{-1} \geq 0$ are achieved when h is small enough.

Table 5.1. Minimum of entries in $\bar{L}_h^{-1}$ and L_h^{-1} for Poisson equation (5.29) with smooth coefficients.

Finite Element Mesh	$d = 0.5$		$d = 0.9$		$d = 0.99$	
	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}
2×4	$-7.32E-18$	$7.48E-06$	$-3.90E-04$	$6.37E-06$	$-7.41E-04$	$6.14E-06$
4×8	$-1.31E-18$	$1.23E-07$	$-4.02E-19$	$9.95E-08$	$-1.65E-04$	$9.44E-08$
8×16	$-3.96E-19$	$1.91E-09$	$-4.91E-19$	$1.52E-09$	$-1.77E-05$	$1.44E-09$
16×32	$-1.92E-19$	$2.98E-11$	$-7.60E-19$	$2.35E-11$	$-1.06E-18$	$2.22E-11$

Next we consider (5.12) solving (5.29) with $c(x,y) \equiv 0$ and a_{ij} being random uniformly distributed random numbers in the interval $(d, d+1)$. Notice that the larger d is, the smaller $\frac{\max\limits_{ij}\{a_{ij}\}}{\min\limits_{ij}\{a_{ij}\}}$ is. When $d = 10$, we have $\frac{\max\limits_{ij}\{a_{ij}\}}{\min\limits_{ij}\{a_{ij}\}} < \sqrt{\frac{61}{49}}$, thus $L_h^{-1} \geq 0$ and $\bar{L}_h^{-1} \geq 0$ are guaranteed by Theorem 5.4.7. In Table 5.2 we can see that the upper bound on $\frac{\max\limits_{ij}\{a_{ij}\}}{\min\limits_{ij}\{a_{ij}\}}$ is indeed a necessary condition to have $\bar{L}_h^{-1} \geq 0$, even though constraints in Theorem 5.4.7 may not be sharp since we still have the positivity when $d = 1$. We have tested $d = 0.3$ many times and never observed negative entries in $\bar{L}_h^{-1}$ and L_h^{-1}.

Last we consider solving the heat equation $u_t = \Delta u$ on $[0,1] \times [0,2]$ with backward Euler time discretization $-\Delta u^{n+1} + \frac{1}{\Delta t} u^{n+1} = \frac{u^n}{\Delta t}$, corresponding to (5.29) with $a(x,y) \equiv 1$ and $c = \frac{1}{\Delta t}$. By Theorem 5.4.8, $\frac{\Delta t}{h^2} > \frac{2}{3}$, is a sufficient condition to ensure $\bar{L}_h^{-1} \geq 0$ and $L_h^{-1} \geq 0$. In Table 5.3, we can see that it is necessary to have a lower bound constraint on $\frac{\Delta t}{h^2}$ but $\frac{\Delta t}{h^2} > \frac{2}{3}$ is not sharp at all. In Figure 5.5, we can see the minimum of entries in $\bar{L}_h^{-1}$ and L_h^{-1}

Table 5.2. Minimum of all entries of $\bar{L}_h^{-1}$ and L_h^{-1} for $a(x, y)$ being random coefficients.

Finite Element Mesh	$d = 0.1$		$d = 1$		$d = 10$	
	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}
2×4	$-1.00E-03$	$6.60E-05$	$-8.15E-18$	$4.73E-05$	$-1.98E-16$	$6.74E-06$
4×8	$-2.14E-04$	$3.22E-06$	$-3.46E-18$	$9.95E-07$	$-5.10E-17$	$1.35E-07$
8×16	$-6.73E-05$	$2.88E-08$	$-5.24E-19$	$1.65E-08$	$-1.81E-17$	$2.21E-09$
16×32	$-2.34E-05$	$3.61E-10$	$-9.01E-19$	$2.02E-10$	$-8.37E-18$	$3.56E-11$

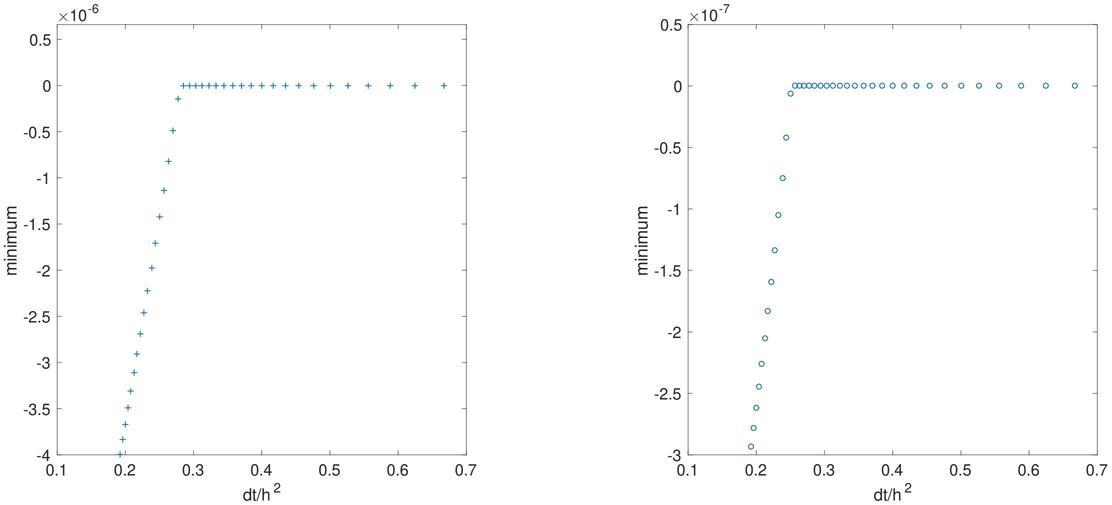
decreases for smaller $\frac{\Delta t}{h^2}$. The lower bound to ensure the inverse non-negativity of $\bar{L}_h^{-1}$ and L_h^{-1} seems to be near $\frac{\Delta t}{h^2} = \frac{1}{3.6}$.

Table 5.3. Minimum of all entries of $\bar{L}_h^{-1}$ and L_h^{-1} for solving heat equation with backward Euler.

Finite Element Mesh	$\Delta t = \frac{3h^2}{2}$		$\Delta t = \frac{h^2}{2}$		$\Delta t = \frac{h^2}{4}$	
	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}	$\bar{L}_h^{-1}$	L_h^{-1}
2×4	0	$7.95E-06$	0	$3.21E-07$	$-9.14E-05$	$-5.34E-07$
4×8	0	$1.01E-09$	0	$1.93E-13$	$-2.28E-05$	$-1.00E-07$
8×16	0	$7.74E-17$	0	$2.58E-25$	$-5.71E-06$	$-2.51E-08$
16×32	0	$2.63E-30$	0	$2.73E-48$	$-1.43E-06$	$-6.27E-09$

5.6 Concluding Remarks

In this paper we have proven that the simplest fourth order accurate finite difference implementation of C^0-Q^2 finite element method is monotone thus satisfies a discrete maximum principle for solving a variable coefficient problem $-\nabla\cdot(a(x, y)\nabla u)+c(x, y)u = f$ under some suitable mesh constraints. The main results in this paper can be used to construct high order spatial discretization preserving positivity or maximum principle for solving time-dependent diffusion problems implicitly by backward Euler time discretization.



(a) Minimum of entries in $\bar{L}_h^{-1}$

(b) Minimum of entries in L_h^{-1}

Figure 5.5. Minimum of all entries of $\bar{L}_h^{-1}$ and L_h^{-1} on 16×32 mesh with different time steps.

Appendix A: M-Matrix factorization for discrete Laplacian

The matrix form of (5.9) can be written as $\frac{1}{h^2}\bar{L}_h\bar{\mathbf{u}} = \bar{\mathbf{f}}$. As an example, if there are seven interior grid points in the mesh for $(0,1)$, then the matrix $\bar{L}_h$ is given by

$$\bar{L}_h = \begin{pmatrix} 1 &&&&&&&& \\ -1 & 2 & -1 &&&&&& \\ \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} &&&& \\ && -1 & 2 & -1 &&&& \\ && \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} && \\ &&&& -1 & 2 & -1 && \\ &&&& \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} \\ &&&&&& -1 & 2 & -1 \\ &&&&&&&& 1 \end{pmatrix}$$

The matrix $\bar{L}_h$ can be written as a product of two nonsingular M-Matrices $\bar{L}_h = M_1M_2$ where

$$M_1 = \begin{pmatrix} 1 &&&&&&&& \\ & 1 &&&&&&& \\ & -\frac{1}{4} & 1 & -\frac{1}{4} &&&&& \\ &&& 1 &&&&& \\ &&& -\frac{1}{4} & 1 & -\frac{1}{4} &&& \\ &&&&& 1 &&& \\ &&&&& -\frac{1}{4} & 1 & -\frac{1}{4} & \\ &&&&&&& 1 & \\ &&&&&&&& 1 \end{pmatrix}, M_2 = \begin{pmatrix} 1 &&&&&&&& \\ -1 & 2 & -1 &&&&&& \\ & -\frac{3}{2} & 3 & -\frac{3}{2} &&&&& \\ && -1 & 2 & -1 &&&& \\ &&& -\frac{3}{2} & 3 & -\frac{3}{2} &&& \\ &&&& -1 & 2 & -1 && \\ &&&&& -\frac{3}{2} & 3 & -\frac{3}{2} & \\ &&&&&& -1 & 2 & -1 \\ &&&&&&&& 1 \end{pmatrix}.$$

Such a factorization is not unique and it does not seem to have further physical or geometrical meanings.

For the scheme (5.11), we can find two linear operators $\mathcal{A}_1$ and $\mathcal{A}_2$ are with their matrix representations A_1 and A_2 being nonsingular M-Matrices, such that $\mathcal{L}_h(\bar{\mathbf{u}}) = \mathcal{A}_2(\mathcal{A}_1(\bar{\mathbf{u}}))$.

Definition of $\mathcal{A}_1$ is given as

- At boundary points:

$$v_{i,j} = \mathcal{A}_1(\bar{\mathbf{u}})_{i,j} = u_{i,j} := g_{ij}.$$

- At interior knots:

$$v_{i,j} = \mathcal{A}_1(\bar{\mathbf{u}})_{i,j} = u_{i,j}.$$

- At interior cell center:

$$v_{i,j} = \mathcal{A}_1(\bar{\mathbf{u}})_{i,j} = 2u_{i,j} - \frac{1}{4}u_{i-1,j} - \frac{1}{4}u_{i+1,j} - \frac{1}{4}u_{i,j-1} - \frac{1}{4}u_{i,j+1}.$$

- At interior edge center (an edge parallel to x-axis):

$$v_{i,j} = \mathcal{A}_1(\bar{\mathbf{u}})_{i,j} = -\frac{1}{6}u_{i-1,j} + \frac{4}{3}u_{i,j} - \frac{1}{6}u_{i+1,j}.$$

- At interior edge center (an edge parallel to y-axis):

$$v_{i,j} = \mathcal{A}_1(\bar{\mathbf{u}})_{i,j} = -\frac{1}{6}u_{i,j-1} + \frac{4}{3}u_{i,j} - \frac{1}{6}u_{i,j+1}.$$

Definition of $\mathcal{A}_2$ is given as:

- At boundary points:

$$\mathcal{A}_2(\bar{\mathbf{v}})_{i,j} = v_{i,j}.$$

- At an interior knot:

$$\mathcal{A}_2(\bar{\mathbf{v}})_{i,j} = -\frac{3}{2}v_{i-1,j} + 3v_{i,j} - \frac{3}{2}v_{i+1,j} - \frac{3}{2}v_{i,j-1} + 3v_{i,j} - \frac{3}{2}v_{i,j+1}$$

- At an interior cell center:

$$\begin{aligned}\mathcal{A}_2(\bar{\mathbf{v}})_{i,j} =& 2v_{i,j} - \frac{3}{8}v_{i-1,j} - \frac{3}{8}v_{i+1,j} - \frac{3}{8}v_{i,j-1} - \frac{3}{8}v_{i,j+1}\\ &-\frac{1}{8}v_{i-1,j+1} - \frac{1}{8}v_{i+1,j+1} - \frac{1}{8}v_{i-1,j-1} - \frac{1}{8}v_{i+1,j+1}.\end{aligned}$$

- At an interior edge center (an edge parallel to x-axis):

$$\begin{aligned}\mathcal{A}_2(\bar{\mathbf{v}})_{i,j} = -\frac{7}{16}v_{i-1,j} + \frac{15}{4}v_{i,j} - \frac{7}{16}v_{i+1,j} - v_{i,j+1} - v_{i,j-1} - \frac{3}{16}v_{i-1,j-1} - \frac{3}{16}v_{i+1,j-1}\\ -\frac{3}{16}v_{i-1,j+1} - \frac{3}{16}v_{i+1,j+1} - \frac{1}{32}v_{i-1,j+2} - \frac{1}{32}v_{i+1,j+2} - \frac{1}{32}v_{i-1,j-2} - \frac{1}{32}v_{i+1,j-2}.\end{aligned}$$

- At an interior edge center (an edge parallel to y-axis):

$$\begin{aligned}\mathcal{A}_2(\bar{\mathbf{v}})_{i,j} = &-\frac{7}{16}v_{i,j-1} + \frac{15}{4}v_{i,j} - \frac{7}{16}v_{i,j+1} - v_{i+1,j} - v_{i-1,j} - \frac{3}{16}v_{i-1,j-1} - \frac{3}{16}v_{i-1,j+1} \\ &-\frac{3}{16}v_{i+1,j-1} - \frac{3}{16}v_{i+1,j+1} - \frac{1}{32}v_{i+2,j-1} - \frac{1}{32}v_{i+2,j+1} - \frac{1}{32}v_{i-2,j-1} - \frac{1}{32}v_{i-2,j+1}.\end{aligned}$$

It is straightforward to verify that $\mathcal{L}_h(\bar{\mathbf{u}}) = \mathcal{A}_2(\bar{\mathbf{v}})$ where $\bar{\mathbf{v}} = \mathcal{A}_1(\bar{\mathbf{u}})$. Obviously, matrices of $\mathcal{A}_1$ and $\mathcal{A}_2$ have positive diagonal entries and nonpositive off-diagonal entries. Moreover, $\mathcal{A}_1(\mathbf{1}) \geq 0$ and $\mathcal{A}_2(\mathbf{1}) \geq 0$ thus A_1 and A_2 satisfy the row sum conditions in Theorem 5.3.4. So A_1 and A_2 are both nonsingular M-matrices and the matrix representation of $\mathcal{L}_h$ is A_2A_1. However, this kind of M-Matrix factorization cannot be extended to the variable coefficient case.

Appendix B

Proof of Theorem 5.4.3. If $c(x) \equiv 0$, then (5.25) reduces to

$$(28a_{i-1} + 20a_{i+1})a_i + 4a_{i+1}a_{i-1} > 9a_{i-1}^2 + 3a_{i+1}^2.$$

A convenient sufficient condition is to require

$$52\min\{a_{i-1}^2, a_i^2, a_{i+1}^2\} > 12\max\{a_{i-1}^2, a_i^2, a_{i+1}^2\},$$

which is equivalent to

$$\frac{\max\{a_{i-1}, a_i, a_{i+1}\}}{\min\{a_{i-1}, a_i, a_{i+1}\}} < \sqrt{\frac{13}{3}}.$$

Let $a(x^1) = \max\{a_{i-1}, a_i, a_{i+1}\}$ and $a(x^2) = \min\{a_{i-1}, a_i, a_{i+1}\}$. Then the inequality above is equivalent to

$$\frac{a(x^1) - a(x^2)}{a(x^2)} < \frac{\sqrt{39} - 3}{3}.$$

By the Mean Value Theorem, there is some $\xi \in (x_{i-1}, x_{i+1})$ such that $a(x^1) - a(x^2) = a(\xi)(x^2 - x^1)$. Since $|x^2 - x^1| \leq 2h$, we have

$$|a(x^1) - a(x^2)| \leq \max_{x \in (x_{i-1}, x_{i+1})} |a(x)| \, 2h.$$

Thus a sufficient condition is to require

$$h \frac{\max\limits_{x \in (x_{i-1}, x_{i+1})} |a(x)|}{\min\limits_{x \in (x_{i-1}, x_{i+1})} a(x)} < \frac{\sqrt{39} - 3}{6}.$$

For $c(x) \geq 0$, (5.25) reduces to

$$(28a_{i-1} + 20a_{i+1})a_i + 4a_{i+1}a_{i-1} > 9a_{i-1}^2 + 3a_{i+1}^2 + 4h^2c_i(3a_{i-1} - 4a_i + 3a_{i+1}),$$

for which a sufficient condition is

$$13 \min_{I_i} a^2(x) > 3 \max_{I_i} a^2(x) + h^2 c_i (6 \max_{I_i} a(x) - 4 \min_{I_i} a(x)). \tag{5.30}$$

One sufficient condition for (5.30) is to have

$$\begin{aligned} \exists \lambda \in (0,1), \quad h^2 c_i (6 \max_{I_i} a(x) - 4 \min_{I_i} a(x)) &< 13(1-\lambda) \min_{I_i} a^2(x), \\ 3 \max_{I_i} a^2(x) &< 13\lambda \min_{I_i} a^2(x). \end{aligned}$$

By similar discussions above, a sufficient condition for $3 \max\limits_{I_i} a^2(x) < 13\lambda \min\limits_{I_i} a^2(x)$ is to have $\lambda > \frac{3}{13}$ and

$$h \frac{\max\limits_{x \in I_i} |a(x)|}{\min\limits_{x \in I_i} a(x)} < \frac{\sqrt{39\lambda} - 3}{6}.$$

The inequality (5.30) is also equivalent to

$$10 \min_{I_i} a^2(x) > 3(\max_{I_i} a^2(x) - \min_{I_i} a^2(x)) + h^2 c_i (6 \max_{I_i} a(x) - 4 \min_{I_i} a(x)).$$

Let $a^2(x^1) = \max_{I_i} a^2(x)$ and $a^2(x^2) = \min_{I_i} a^2(x)$, then by the Mean Value Theorem on the function $a^2(x)$, there is some $\xi \in (x_{i-1}, x_{i+1})$ such that

$$a^2(x^1) - a^2(x^1) = 2a(\xi)a(\xi)(x^1 - x^2) \leq 4h \max_{I_i} a(x) \max_{I_i} |a(x)|.$$

So it suffices to have

$$10 \min_{I_i} a^2(x) > 12h \max_{I_i} a(x) \max_{I_i} |a(x)| + h^2 c_i (6 \max_{I_i} a(x) - 4 \min_{I_i} a(x)),$$

which can be simplified to

$$2h \max_{I_i} |a(x)| + h^2 c_i (1 - \frac{2}{3} \frac{\min_{I_i} a(x)}{\max_{I_i} a(x)}) < \frac{5}{3} \frac{\min_{I_i} a^2(x)}{\max_{I_i} a(x)}.$$

If $a(x) \equiv a > 0$, it is straightforward to verify that (5.25) is equivalent to $hc_i < 5a$. □

of Theorem 5.4.4. For a smooth coefficient $a(x)$, by Taylor's Theorem,

$$a(x+h) = a(x) + ha(x) + \frac{1}{2}h^2 a(\xi_1), \xi_1 \in [x, x+h],$$

$$a(x-h) = a(x) - ha(x) + \frac{1}{2}h^2 a(\xi_2), \xi_2 \in [x-h, x].$$

With the Intermediate Value Theorem for $a(x)$, we get

$$a(x) = \frac{1}{2}[a(x+h) + a(x-h) - h^2 a(\xi)], \quad \xi \in (\xi_2, \xi_1) \subset [x-h, x+h].$$

Thus we can rewrite a_i as $a_i = \frac{1}{2}(a_{i-1} + a_{i+1} - d_i h^2)$ where

$$d_i := \frac{a_{i-1} + a_{i+1} - 2a_i}{h^2} = a(\xi), \text{ for some } \xi \in (x_{i-1}, x_{i+1}).$$

If $c(x) \equiv 0$, then (5.25) reduces to $(28a_{i-1}+20a_{i+1})a_i+4a_{i+1}a_{i-1} > 9a_{i-1}^2+3a_{i+1}^2$. Introducing an arbitrary number $\lambda \in (0,2]$, it is equivalent to

$$\begin{aligned}
4a_{i+1}a_{i-1} + (4-2\lambda)a_i(7a_{i-1}+5a_{i+1}) + 2\lambda a_i(7a_{i-1}+5a_{i+1}) &> 9a_{i-1}^2 + 3a_{i+1,}^2 \\
(12\lambda+4)a_{i+1}a_{i-1} + (4-2\lambda)a_i(7a_{i-1}+5a_{i+1}) + (7\lambda-9)a_{i-1}^2 + (5\lambda-3)a_{i+1}^2 & \\
> \lambda h^2 d_i(7a_{i-1}+5a_{i+1}), & \\
(\frac{4}{\lambda}-2)a_i + a_{i-1}\frac{(5\lambda-3)\theta^2+(12\lambda+4)\theta+(7\lambda-9)}{\lambda(5\theta+7)} > h^2 d_i, \quad \theta = \frac{a_{i+1}}{a_{i-1}}, & \\
\left(\frac{4}{\lambda}-2\right)a_i + \left(\frac{\frac{41}{5}\theta-9}{\lambda(5\theta+7)}+1\right)a_{i-1} + \left(1-\frac{3}{5\lambda}\right)a_{i+1} > h^2 d_i. &
\end{aligned}$$

Notice that $\frac{\frac{41}{5}\theta-9}{5\theta+7} > -\frac{9}{7}$. By taking $\frac{9}{7} \le \lambda \le 2$, it suffices to require

$$(1-\frac{9}{7\lambda})a_{i-1} + (\frac{4}{\lambda}-2)a_i + (1-\frac{3}{5\lambda})a_{i+1} > h^2 d_i, \tag{5.31}$$

as a sufficient condition of the above inequalities. If $a(x)$ is a concave function, then it satisfies $a(x_i) = a(\frac{x_{i-1}+x_{i-1}}{2}) \ge \frac{1}{2}a(x_{i-1}) + \frac{1}{2}a(x_{i+1})$, which implies $a_{i-1}+a_{i+1}-2a_i \le 0$, thus (5.31) holds trivially. Otherwise, (5.31) holds for $\lambda = \frac{9}{7}$ if the following mesh constraint is satisfied:

$$h^2 \max_{x\in(x_{i-1},x_{i+1})} a(x) < \frac{74}{45}\min\{a_{i-1}, a_i, a_{i+1}\}.$$

If $c(x) \ge 0$, for any $\lambda \in (0,2]$, (5.25) is equivalent to

$$\begin{aligned}
(12\lambda+4)a_{i+1}a_{i-1} + (4-2\lambda)a_i(7a_{i-1}+5a_{i+1}) + (7\lambda-9)a_{i-1}^2 + (5\lambda-3)a_{i+1}^2 & \\
> \lambda h^2 d_i(7a_{i-1}+5a_{i+1}) + 4h^2 c_i(a_{i-1}+a_{i+1}+2d_i h^2). &
\end{aligned} \tag{5.32}$$

If assuming $d_i h^2 \le \frac{74}{45}\min\{a_{i-1}, a_i, a_{i+1}\}$, then $d_i h^2 \le \lambda_1 a_{i-1} + \lambda_2 a_{i+1}$ for any two positive numbers λ_1, λ_2 satisfying $\lambda_1+\lambda_2 = \frac{74}{45}$. In particular, for $\lambda_1 = \frac{563}{540}$, we get $d_i h^2 \le \frac{563}{540}a_{i-1} + \frac{65}{108}a_{i+1}$, which implies

$$a_{i-1} + a_{i+1} + 2d_i h^2 \le \frac{119}{270}(7a_{i-1}+5a_{i+1}).$$

By replacing $a_{i-1}+a_{i+1}+2d_ih^2$ by the inequality above in (5.32), we get a sufficient condition for (5.32) as following:

$$\begin{aligned}(12\lambda+4)a_{i+1}a_{i-1}+(4-2\lambda)a_i(7a_{i-1}+5a_{i+1})+(7\lambda-9)a_{i-1}^2+(5\lambda-3)a_{i+1}^2\\ >\lambda h^2d_i(7a_{i-1}+5a_{i+1})+4h^2c_i\frac{119}{270}(7a_{i-1}+5a_{i+1}).\end{aligned}\tag{5.33}$$

Similar to the derivation of (5.31), we can derive a sufficient condition of (5.33) as

$$h^2\left(1.5c_i+\max_{x\in(x_{i-1},x_{i+1})}a(x)\right)<\frac{74}{45}\min\{a_{i-1},a_i,a_{i+1}\}.$$

If $d_i\leq 0$, then a sufficient condition for (5.32) is

$$\frac{(12\lambda+4)a_{i+1}a_{i-1}+(4-2\lambda)a_i(7a_{i-1}+5a_{i+1})+(7\lambda-9)a_{i-1}^2+(5\lambda-3)a_{i+1}^2}{a_{i-1}+a_{i+1}}>4h^2c_i,$$

from which we can derive a sufficient condition as

$$4h^2c_i<(7\lambda-9)a_{i-1}+(5-\frac{5}{2}\lambda)a_i+(5\lambda-3)a_{i+1},$$

for which a sufficient condition by setting $\lambda=2$ is $h^2c_i<3\min\{a_{i-1},a_i,a_{i+1}\}$. □

of Theorem 5.4.7. Since (5.27a) and (5.28a) are equivalent to

$$\begin{aligned}4(7a_{i-1,j}+5a_{i+1,j})a_{ij}+4a_{i-1,j}a_{i+1,j}+16a_{ij}(a_{i,j-1}+a_{i,j+1})\\ >9a_{i-1,j}^2+3a_{i+1,j}^2+12(a_{i-1,j}+a_{i+1,j})(a_{i,j-1}+a_{i,j+1})+4(3a_{i-1,j}-4a_{ij}+3a_{i+1,j})h^2c_{ij}\end{aligned}$$

and

$$\begin{aligned}a_{i-1,j}a_{i+1,j}+2a_{ij}a_{i-1,j}+4a_{ij}(a_{i,j-2}+4a_{i,j-1}+18a_{i,j}+4a_{i,j+1}+a_{i,j+2})>18a_{i-1,j}^2+6a_{i+1,j}^2\\ +14a_{ij}a_{i+1,j}+3(a_{i-1,j}+a_{i+1,j})(a_{i,j-2}+4a_{i,j-1}+4a_{i,j+1}+a_{i,j+2})+8(3a_{i-1,j}-4a_{ij}+3a_{i+1,j})h^2c_{ij}.\end{aligned}$$

A sufficient condition is to require

$$7\min_{I_{ij}} a(x)^2 > 5\max_{I_{ij}} a(x)^2 + \frac{2}{3}(3\max_{I_{ij}} a(x) - 2\min_{I_{ij}} a(x))h^2 c_{ij} \tag{5.34}$$

for all cell centers x_{ij} of cell $I_{ij} = [x_{i-1}, x_{i+1}] \times [y_{i-1}, y_{i+1}]$, and the following mesh constraints for all edge centers x_{ij}:

$$61\min_{J_{ij}} a(x)^2 > 49\max_{J_{ij}} a(x)^2 + 8(3\max_{J_{ij}} a(x) - 2\min_{J_{ij}} a(x))h^2 c_{ij}, \tag{5.35}$$

where we J_{ij} is the union of two cells: if x_{ij} is an edge center of an edge parallel to x-axis, then $J_{ij} = I_{i,j-1} \cup I_{i,j+1}$; if x_{ij} is an edge center of an edge parallel to y-axis, then $J_{ij} = I_{i-1,j} \cup I_{i+1,j}$. Notice that (5.35) implies (5.34), thus it suffices to have (5.35) only. □

6. A HIGH ORDER ACCURATE BOUND-PRESERVING COMPACT FINITE DIFFERENCE SCHEME FOR SCALAR CONVECTION DIFFUSION EQUATIONS

In this chapter, we show that the classical fourth order accurate compact finite difference scheme with high order strong stability preserving time discretizations for convection diffusion problems satisfies a weak monotonicity property, which implies that a simple limiter can enforce the bound-preserving property without losing conservation and high order accuracy. Higher order accurate compact finite difference schemes satisfying the weak monotonicity will also be discussed.

6.1 Introduction

6.1.1 The bound-preserving property

Consider the initial value problem for a scalar convection diffusion equation $u_t + f(u)_x = a(u)_{xx}, \quad u(x,0) = u_0(x)$, where $a(u) \geq 0$. Assume $f(u)$ and $a(u)$ are well-defined smooth functions for any $u \in [m, M]$ where $m = \min_x u_0(x)$ and $M = \max_x u_0(x)$. Its exact solution satisfies:

$$\min_x u_0(x) = m \leq u(x,t) \leq M = \max_x u_0(x), \quad \forall t \geq 0. \tag{6.1}$$

In this chapter, we are interested in constructing a high order accurate finite difference scheme satisfying the bound-preserving property (6.1).

For a scalar problem, it is desired to achieve (6.1) in numerical solutions mainly for the physical meaning. For instance, if u denotes density and $m = 0$, then negative numerical solutions are meaningless. In practice, in addition to enforcing (6.1), it is also critical to strictly enforce the global conservation of numerical solutions for a time-dependent convection dominated problem. Moreover, the computational cost for enforcing (6.1) should not be significant if it is needed for each time step.

6.1.2 Popular methods for convection problems

For the convection problems, i.e., $a(u) \equiv 0$, a straightforward way to achieve the above goals is to require a scheme to be monotone, total-variational-diminishing (TVD), or satisfying a discrete maximum principle, which all imply the bound-preserving property. But most schemes satisfying these stronger properties are at most second order accurate. For instance, a monotone scheme and traditional TVD finite difference and finite volume schemes are at most first order accurate [89]. Even though it is possible to have high order TVD finite volume schemes in the sense of measuring the total variation of reconstruction polynomials [90], [91], such schemes can be constructed only for the one-dimensional problems. The second order central scheme satisfies a discrete maximum principle $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$ where u_j^n denotes the numerical solution at n-th time step and j-th grid point [92]. Any finite difference scheme satisfying such a maximum principle can be at most second order accurate, see Harten's example in [93]. By measuring the extrema of reconstruction polynomials, third order maximum-principle-satisfying schemes can be constructed [94] but extensions to multi-dimensional nonlinear problems are very difficult.

For constructing high order accurate schemes, one can enforce only the bound-preserving property for fixed known bounds, e.g., $m = 0$ and $M = 1$ if u denotes the density ratio. Even though high order linear schemes cannot be monotone, high order finite volume type spatial discretizations including the discontinuous Galerkin (DG) method satisfy a weak monotonicity property [23], [93], [95]. Namely, in a scheme consisting of any high order finite volume spatial discretization and forward Euler time discretization, the cell average is a monotone function of the point values of the reconstruction or approximation polynomial at Gauss-Lobatto quadrature points. Thus if these point values are in the desired range $[m, M]$, so are the cell averages in the next time step. A simple and efficient local bound-preserving limiter can be designed to control these point values without destroying conservation. Moreover, this simple limiter is high order accurate, see [23] and the appendix in [96]. With strong stability preserving (SSP) Runge-Kutta or multistep methods [97], which are convex combinations of several formal forward Euler steps, a high order accurate finite volume or DG scheme can be rendered bound-preserving with this limiter. These results can be easily extended to multiple dimensions on cells of general shapes. However, for a general finite difference scheme, the weak monotonicity does not hold.

For enforcing only the bound-preserving property in high order schemes, efficient alternatives include a flux limiter [98], [99] and a sweeping limiter in [100]. These methods are designed to directly enforce the bounds without destroying conservation thus can be used on any conservative schemes. Even though they work well in practice, it is nontrivial to analyze and rigorously justify the accuracy of these methods especially for multi-dimensional nonlinear problems.

6.1.3 The weak monotonicity in compact finite difference schemes

Even though the weak monotonicity does not hold for a general finite difference scheme, in this paper we will show that some high order compact finite difference schemes satisfy such a property, which implies a simple limiting procedure can be used to enforce bounds without destroying accuracy and conservation.

To demonstrate the main idea, we first consider a fourth order accurate compact finite difference approximation to the first derivative on the interval $[0, 1]$:

$$\frac{1}{6}(f_{i+1} + 4f_i + f_{i-1}) = \frac{f_{i+1} - f_{i-1}}{2\Delta x} + \mathcal{O}(\Delta x^4),$$

where f_i and f_i are point values of a function $f(x)$ and its derivative $f(x)$ at uniform grid points x_i $(i = 1, \cdots, N)$ respectively. For periodic boundary conditions, the following tridiagonal linear system needs to be solved to obtain the implicitly defined approximation to the first order derivative:

$$\frac{1}{6}\begin{pmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix} = \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 & & & -1 \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ 1 & & & -1 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}. \tag{6.2}$$

We refer to the tridiagonal $\frac{1}{6}(1,4,1)$ matrix as a weighting matrix. For the one-dimensional scalar conservation laws with periodic boundary conditions on $[0,1]$:

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x), \tag{6.3}$$

the semi-discrete fourth order compact finite difference scheme can be written as

$$\frac{d\bar{u}_i}{dt} = -\frac{1}{2\Delta x}[f(u_{i+1}) - f(u_{i-1})], \tag{6.4}$$

where $\bar{u}_i$ is defined as $\bar{u}_i = \frac{1}{6}(u_{i-1} + 4u_i + u_{i+1})$. Let $\lambda = \frac{\Delta t}{\Delta x}$, then (6.4) with the forward Euler time discretization becomes

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{1}{2}\lambda[f(u_{i+1}^n) - f(u_{i-1}^n)]. \tag{6.5}$$

The following weak monotonicity holds under the CFL $\lambda \max_u |f(u)| \leq \frac{1}{3}$:

$$\begin{aligned}
\bar{u}_i^{n+1} &= \frac{1}{6}(u_{i-1}^n + 4u_i^n + u_{i+1}^n) - \frac{1}{2}\lambda[f(u_{i+1}^n) - f(u_{i-1}^n)] \\
&= \frac{1}{6}[u_{i-1} + 3\lambda f(u_{i-1}^n)] + \frac{1}{6}[u_{i+1}^n - 3\lambda f(u_{i+1}^n)] + \frac{4}{6}u_i^n \\
&= H(u_{i-1}^n, u_i^n, u_{i+1}^n) = H(\uparrow, \uparrow, \uparrow),
\end{aligned}$$

where ↑ denotes that the partial derivative with respect to the corresponding argument is non-negative. Therefore $m \leq u_i^n \leq M$ implies $m = H(m,m,m) \leq \bar{u}_i^{n+1} \leq H(M,M,M) = M$, thus

$$m \leq \frac{1}{6}(u_{i-1}^{n+1} + 4u_i^{n+1} + u_{i+1}^{n+1}) \leq M. \tag{6.6}$$

If there is any overshoot or undershoot, i.e., $u_i^{n+1} > M$ or $u_i^{n+1} < m$ for some i, then (6.6) implies that a local limiting process can eliminate the overshoot or undershoot. Here we consider the special case $m = 0$ to demonstrate the basic idea of this limiter, and for simplicity we ignore the time step index $n+1$. In Section 6.2 we will show that $\frac{1}{6}(u_{i-1}+4u_i+u_{i+1}) \geq 0, \forall i$ implies the following two facts:

1. $\max\{u_{i-1}, u_i, u_{i+1}\} \geq 0$;

2. If $u_i < 0$, then $\frac{1}{2}(u_{i-1})_+ + \frac{1}{2}(u_{i+1})_+ \geq -u_i > 0$, where $(u)_+ = \max\{u, 0\}$.

By the two facts above, when $u_i < 0$, then the following three-point stencil limiting process can enforce positivity without changing $\sum_i u_i$:

$$v_{i-1} = u_{i-1} + \frac{(u_{i-1})_+}{(u_{i-1})_+ + (u_{i+1})_+} u_i; \quad v_{i+1} = u_{i+1} + \frac{(u_{i+1})_+}{(u_{i-1})_+ + (u_{i+1})_+} u_i,$$
$$\text{replace} \quad u_{i-1},\, u_i,\, u_{i+1} \quad \text{by} \quad v_{i-1},\, 0,\, v_{i+1} \quad \text{respectively.}$$

In Section 6.2.2, we will show that such a simple limiter can enforce the bounds of u_i without destroying accuracy and conservation. Thus with SSP high order time discretizations, the fourth order compact finite difference scheme solving (6.3) can be rendered bound-preserving by this limiter. Moreover, in this paper we will show that such a weak monotonicity and the limiter can be easily extended to more general and practical cases including two-dimensional problems, convection diffusion problems, inflow-outflow boundary conditions, higher order accurate compact finite difference approximations, compact finite difference schemes with a total-variation-bounded (TVB) limiter [101]. However, the extension to non-uniform grids is highly nontrivial thus will not be discussed. In this paper, we only focus on uniform grids.

6.1.4 The weak monotonicity for diffusion problems

Although the weak monotonicity holds for arbitrarily high order finite volume type schemes solving the convection equation (6.3), it no longer holds for a conventional high order linear finite volume scheme or DG scheme even for the simplest heat equation, see the appendix in [96]. Toward satisfying the weak monotonicity for the diffusion operator, an unconventional high order finite volume scheme was constructed in [102]. Second order accurate DG schemes usually satisfies the weak monotonicity for the diffusion operator on general meshes [103]. The only previously known high order linear scheme in the literature satisfying the weak monotonicity for scalar diffusion problems is the third order direct DG (DDG) method with special parameters [104], which is a generalized version of interior penalty DG method. On the other hand, arbitrarily high order nonlinear positivity-preserving DG schemes for diffusion problems were constructed in [96], [105], [106].

In this paper we will show that the fourth order accurate compact finite difference and a few higher order accurate ones are also weakly monotone, which is another class of linear high order schemes satisfying the weak monotonicity for diffusion problems.

It is straightforward to verify that the backward Euler or Crank-Nicolson method with the fourth order compact finite difference methods satisfies a maximum principle for the heat equation but it can be used be as a bound-preserving scheme only for linear problems. The method is this chapter is explicit thus can be easily applied to nonlinear problems. It is difficult to generalize the maximum principle to an implicit scheme. Regarding positivity-preserving implicit schemes, see [107] for a study on weak monotonicity in implicit schemes solving convection equations. See also [108] for a second order accurate implicit and explicit time discretization for the BGK equation.

Although high order compact finite difference methods have been extensively studied in the literature, e.g., [71], [101], [109]–[112], this is the first time that the weak monotonicity in compact finite difference approximations is discussed. This is also the first time a weak monotonicity property is established for a high order accurate finite difference type scheme. The weak monotonicity property suggests it is possible to post process the numerical solution without losing conservation by a simple limiter to enforce global bounds. Moreover, this approach allows an easy justification of high order accuracy of the constructed bound-preserving scheme.

For extensions to two-dimensional problems, convection diffusion problems, and sixth order and eighth order accurate schemes, the discussion about the weak monotonicity in general becomes more complicated since the weighting matrix may become a five-diagonal matrix instead of the tridiagonal $\frac{1}{6}(1,4,1)$ matrix in (6.2). Nonetheless, we demonstrate that the same simple three-point stencil limiter can still be used to enforce bounds because we can factor the more complicated weighting matrix as a product of a few of tridiagonal $\frac{1}{c+2}(1,c,1)$ matrices with $c \geq 2$.

The chapter is organized as follows: in Section 6.2 we demonstrate the main idea for the fourth order accurate scheme solving one-dimensional problems with periodic boundary conditions. Two-dimensional extensions are discussed in in Section 6.3. Section 6.5 is the extension to higher order accurate schemes. Inflow-outflow boundary conditions and Dirichlet boundary conditions are considered in Section 6.6. Numerical tests are given in Section 6.7. Section 6.8 consists of concluding remarks.

6.2 A Fourth Order Accurate Scheme For One-dimensional Problems

In this section we first show the fourth order compact finite difference with forward Euler time discretization satisfies the weak monotonicity. Then we discuss how to design a simple limiter to enforce the bounds of point values. To eliminate the oscillations, a total variation bounded (TVB) limiter can be used. We also show that the TVB limiter does not affect the bound-preserving property of $\bar{u}_i$, thus it can be combined with the bound-preserving limiter to ensure the bound-preserving and non-oscillatory solutions for shocks. High order time discretizations will be discussed in Section 6.2.5.

6.2.1 One-dimensional convection problems

Consider a periodic function $f(x)$ on the interval $[0,1]$. Let $x_i = \frac{i}{N}$ $(i = 1, \cdots, N)$ be the uniform grid points on the interval $[0,1]$. Let $\mathbf{f}$ be a column vector with numbers $f_1, f_2, \cdots, f_N$ as entries, where $f_i = f(x_i)$. Let W_1, W_2, D_x and D_{xx} denote four linear operators as follows:

$$W_1\mathbf{f} = \frac{1}{6}\begin{pmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}, D_x\mathbf{f} = \frac{1}{2}\begin{pmatrix} 0 & 1 & & & -1 \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ 1 & & & -1 & 0 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix},$$

$$W_2\mathbf{f} = \frac{1}{12}\begin{pmatrix} 10 & 1 & & & 1 \\ 1 & 10 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 10 & 1 \\ 1 & & & 1 & 10 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}, D_{xx}\mathbf{f} = \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix}.$$

The fourth order compact finite difference approximation to the first order derivative (6.2) with periodic assumption for $f(x)$ can be denoted as $W_1\mathbf{f} = \frac{1}{\Delta x}D_x\mathbf{f}$. The fourth order compact finite difference approximation to $f(x)$ is $W_2\mathbf{f} = \frac{1}{\Delta x^2}D_{xx}\mathbf{f}$. The fourth compact finite difference approximations can be explicitly written as

$$\mathbf{f} = \frac{1}{\Delta x}W_1^{-1}D_x\mathbf{f}, \quad \mathbf{f} = \frac{1}{\Delta x^2}W_2^{-1}D_{xx}\mathbf{f},$$

where W_1^{-1} and W_2^{-1} are the inverse operators. For convenience, by abusing notations we let $W_1^{-1}f_i$ denote the i-th entry of the vector $W_1^{-1}\mathbf{f}$.

Then the scheme (6.4) solving the scalar conservation laws (6.3) with periodic boundary conditions on the interval $[0,1]$ can be written as $W_1\frac{d}{dt}u_i = -\frac{1}{2\Delta x}[f(u_{i+1}) - f(u_{i-1})]$, and the scheme (6.5) is equivalent to $W_1u_i^{n+1} = W_1u_i^n - \frac{1}{2}\lambda[f(u_{i+1}^n) - f(u_{i-1}^n)]$. As shown in Section 6.1.3, the scheme (6.5) satisfies the weak monotonicity.

Theorem 6.2.1. *Under the CFL constraint $\frac{\Delta t}{\Delta x}\max_u|f(u)| \leq \frac{1}{3}$,if $u_i^n \in [m, M]$, then u^{n+1} computed by the scheme* (6.5) *satisfies* (6.6)*.*

6.2.2 A three-point stencil bound-preserving limiter

In this subsection, we consider a more general constraint than (6.6) and we will design a simple limiter to enforce bounds of point values based on it. Assume we are given a sequence of periodic point values u_i $(i = 1, \cdots, N)$ satisfying

$$m \leq \frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) \leq M, \quad i = 1, \cdots, N, \quad c \geq 2, \tag{6.7}$$

where $u_0 := u_N$, $u_{N+1} := u_1$ and $c \geq 2$ is a constant. We have the following results:

Lemma 6.2.2. *The constraint* (6.7) *implies the following for stencil* $\{i-1, i, i+1\}$*:*

(1) $\min\{u_{i-1}, u_i, u_{i+1}\} \leq M, \quad \max\{u_{i-1}, u_i, u_{i+1}\} \geq m.$

(2) If $u_i > M$*, then* $\frac{(u_i-M)_+}{(M-u_{i-1})_+ + (M-u_{i+1})_+} \leq \frac{1}{c}$.
If $u_i < m$*, then* $\frac{(m-u_i)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+} \leq \frac{1}{c}$.
Here the subscript $+$ *denotes the positive part, i.e.,* $(a)_+ = \max\{a, 0\}$.

Remark 6.2.3. The first statement in Lemma 6.2.2 states that there do not exist three consecutive overshoot points or three consecutive undershoot points. But it does not necessarily imply that at least one of three consecutive point values is in the bounds $[m, M]$. For instance, consider the case for $c = 4$ and N is even, define $u_i \equiv 1.1$ for all odd i and $u_i \equiv -0.1$ for all even i, then $\frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) \in [0, 1]$ for all i but none of the point values u_i is in $[0, 1]$.

Remark 6.2.4. Lemma 6.2.2 implies that if u_i is out of the range $[m, M]$, then we can set $u_i \leftarrow m$ for undershoot (or $u_i \leftarrow M$ for overshoot) without changing the local sum $u_{i-1} + u_i + u_{i+1}$ by decreasing (or increasing) its neighbors $u_{i\pm1}$.

Proof. We only discuss the upper bound. The inequalities for the lower bound can be similarly proved. First, if $u_{i-1}, u_i, u_{i+1} > M$ then $\frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) > M$ which is a contradiction to (6.7). Second, (6.7) implies $u_{i-1} + cu_i + u_{i+1} \leq (c+2)M$, thus $c(u_i - M) \leq (M - u_{i-1}) + (M - u_{i+1}) \leq (M - u_{i-1})_+ + (M - u_{i+1})_+$. If $u_i > M$, we get $(M - u_{i-1})_+ + (M - u_{i+1})_+ > 0$. Moreover, $\frac{(u_i-M)_+}{(M-u_{i-1})_+ + (M-u_{i+1})_+} = \frac{u_i-M}{(M-u_{i-1})_+ + (M-u_{i+1})_+} \leq \frac{1}{c}$. □

For simplicity, we first consider a limiter to enforce only the lower bound without destroying global conservation. For $m = 0$, this is a positivity-preserving limiter.

Remark 6.2.5. Even though a **for** loop is used, Algorithm 1 is a local operation to an undershoot point since only information of two immediate neighboring points of the undershoot point are needed. Thus it is not a sweeping limiter.

Theorem 6.2.6. *The output of Algorithm 1 satisfies* $\sum\limits_{i=1}^{N} v_i = \sum\limits_{i=1}^{N} u_i$ *and* $v_i \geq m$.

Algorithm 1 A limiter for periodic data u_i to enforce the lower bound.

```
Require: The input $u_i$ satisfies $\bar{u}_i = \frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) \geq m, i = 1, \cdots, N$, with $c \geq 2$.
  Let $u_0$, $u_{N+1}$ denote $u_N$, $u_1$ respectively.
Ensure: The output satisfies $v_i \geq m, i = 1, \cdots, n$ and $\sum_{i=1}^{N} v_i = \sum_{i=1}^{N} u_i$.
  First set $v_i = u_i$, $i = 1, \cdots, N$. Let $v_0$, $v_{N+1}$ denote $v_N$, $v_1$ respectively.
  for $i = 1, \cdots, N$ do
    if $u_i < m$ then
      $v_{i-1} \leftarrow v_{i-1} - \frac{(u_{i-1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+$
      $v_{i+1} \leftarrow v_{i+1} - \frac{(u_{i+1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+$
      $v_i \leftarrow m$
    end if
  end for
```

Proof. First of all, notice that the algorithm only modifies the undershoot points and their immediate neighbors.

Next we will show the output satisfies $v_i \geq m$ case by case:

- If $u_i < m$, the i-th step in **for** loops sets $v_i = m$. After the $(i+1)$-th step in **for** loops, we still have $v_i = m$ because $(u_i - m)_+ = 0$.

- If $u_i = m$, then $v_i = m$ in the final output because $(u_i - m)_+ = 0$.

- If $u_i > m$, then limiter may decrease it if at least one of its neighbors u_{i-1} and u_{i+1} is below m:

$$\begin{aligned} v_i &= u_i - \frac{(u_i - m)_+(m - u_{i-1})_+}{(u_{i-2} - m)_+ + (u_i - m)_+} - \frac{(u_i - m)_+(m - u_{i+1})_+}{(u_i - m)_+ + (u_{i+2} - m)_+} \\ &\geq u_i - \frac{1}{c}(u_i - m)_+ - \frac{1}{c}(u_i - m)_+ > m, \end{aligned}$$

  where the inequalities are implied by Lemma 6.2.2 and the fact $c \geq 2$.

Finally, we need to show the local sum $v_{i-1} + v_i + v_{i+1}$ is not changed during the i-th step if $u_i < m$. If $u_i < m$, then after $(i-1)$-th step we still have $v_i = u_i$ because $(u_i - m)_+ = 0$. Thus in the i-th step of **for** loops, the point value at x_i is increased by the amount $m - u_i$, and the point values at x_{i-1} and x_{i+1} are decreased by $\frac{(u_{i-1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+ + \frac{(u_{i+1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+ = m - u_i$. So $v_{i-1} + v_i + v_{i+1}$ is not changed during the i-th step. Therefore the limiter ensures the output $v_i \geq m$ without changing the global sum. □

The limiter described by Algorithm 1 is a local three-point stencil limiter in the sense that only undershoots and their neighbors will be modified, which means the limiter has no influence on point values that are neither undershoots nor neighbors to undershoots. Obviously a similar procedure can be used to enforce only the upper bound. However, to enforce both the lower bound and the upper bound, the discussion for this three-point stencil limiter is complicated for a saw-tooth profile in which both neighbors of an overshoot point are undershoot points. Instead, we will use a different limiter for the saw-tooth profile. To this end, we need to separate the point values $\{u_i, i = 1, \cdots, N\}$ into two classes of subsets consisting of consecutive point values.

In the following discussion, a *set* refers to a set of consecutive point values $u_l, u_{l+1}, u_{l+2}, \cdots, u_{m-1}, u_m$. For any set $S = \{u_l, u_{l+1}, \cdots, u_{m-1}, u_m\}$, we call the first point value u_l and the last point value u_m as *boundary points*, and call the other point values $u_{l+1}, \cdots, u_{m-1}$ as *interior points*. A set of class I is defined as a set satisfying the following:

1. It contains at least four point values.

2. Both *boundary points* are in $[m, M]$ and all *interior points* are out of range.

3. It contains both undershoot and overshoot points.

Notice that in a set of class I, at least one undershoot point is next to an overshoot point. For given point values $u_i, i = 1, \cdots, N$, suppose all the sets of class I are $S_1 = \{u_{m_1}, u_{m_1+1}, \cdots, u_{n_1}\}$, $S_2 = \{u_{m_2}, \cdots, u_{n_2}\}$, $\cdots$, $S_K = \{u_{m_K}, \cdots, u_{n_K}\}$, where $m_1 < m_2 < \cdots < u_{m_K}$.

A set of class II consists of point values between S_i and S_{i+1} and two boundary points u_{n_i} and $u_{m_{i+1}}$. Namely they are $T_0 = \{u_1, u_2, \cdots, u_{m_1}\}$, $T_1 = \{u_{n_1}, \cdots, u_{m_2}\}$, $T_2 = \{u_{n_2}, \cdots, u_{m_3}\}$, $\cdots$, $T_K = \{u_{n_K}, \cdots, u_N\}$. For periodic data u_i, we can combine T_K and T_0 to define $T_K = \{u_{n_K}, \cdots, u_N, u_1, \cdots, u_{m_1}\}$.

In the sets of class I, the undershoot and the overshoot are neighbors. In the sets of class II, the undershoot and the overshoot are separated, i.e., an overshoot is not next to any undershoot. We remark that the sets of class I are hardly encountered in the numerical tests but we include them in the discussion for the sake of completeness. When there are no sets of class I, all point values form a single set of class II. We will use the same procedure as in Algorithm 1 for T_i and a different limiter for S_i to enforce both the lower bound and the upper bound.

Theorem 6.2.7. *Assume periodic data* $u_i (i = 1, \cdots, N)$ *satisfies* $\bar{u}_i = \frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) \in [m, M]$, $c \geq 2$ *for all* $i = 1, \cdots, N$ *with* $u_0 := u_N$ *and* $u_{N+1} := u_1$, *then the output of Algorithm 2 satisfies* $\sum_{i=1}^{N} v_i = \sum_{i=1}^{N} u_i$ *and* $v_i \in [m, M]$, $\forall i$.

Proof. First we show the output $v_i \in [m, M]$. Consider **Step II**, which only modifies the undershoot and overshoot points and their immediate neighbors. Notice that the operation

Algorithm 2 A bound-preserving limiter for periodic data u_i satisfying $\bar{u}_i \in [m, M]$

```
Require: the input $u_i$ satisfies $\bar{u}_i = \frac{1}{c+2}(u_{i-1} + cu_i + u_{i+1}) \in [m, M]$, $c \geq 2$. Let $u_0$, $u_{N+1}$
    denote $u_N$, $u_1$ respectively.
Ensure: the output satisfies $v_i \in [m, M], i = 1, \cdots, N$ and $\sum_{i=1}^N v_i = \sum_{i=1}^N u_i$.
Step 0: First set $v_i = u_i$, $i = 1, \cdots, N$. Let $v_0$, $v_{N+1}$ denote $v_N$, $v_1$ respectively.
Step I: Find all the sets of class I $S_1, \cdots, S_K$ (all local saw-tooth profiles) and all the
    sets of class II $T_1, \cdots, T_K$.
Step II: For each $T_j$ $(j = 1, \cdots, K)$, the same limiter as in Algorithm 1 (but for both
    upper bound and lower bound) is used:
for all index $i$ in $T_j$ do
    if $u_i < m$ then
        $v_{i-1} \leftarrow v_{i-1} - \frac{(u_{i-1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+$
        $v_{i+1} \leftarrow v_{i+1} - \frac{(u_{i+1}-m)_+}{(u_{i-1}-m)_+ + (u_{i+1}-m)_+}(m - u_i)_+$
        $v_i \leftarrow m$
    end if
    if $u_i > M$ then
        $v_{i-1} \leftarrow v_{i-1} + \frac{(M-u_{i-1})_+}{(M-u_{i-1})_+ + (M-u_{i+1})_+}(u_i - M)_+$
        $v_{i+1} \leftarrow v_{i+1} + \frac{(M-u_{i+1})_+}{(M-u_{i-1})_+ + (M-u_{i+1})_+}(u_i - M)_+$
        $v_i \leftarrow M$
    end if
end for
Step III: for each saw-tooth profile $S_j = \{u_{m_j}, \cdots, u_{n_j}\}$ $(j = 1, \cdots, K)$, let $N_0$ and $N_1$
    be the numbers of undershoot and overshoot points in $S_j$ respectively.
Set $U_j = \sum_{i=m_j}^{n_j} v_i$.
for $i = m_j + 1, \cdots, n_j - 1$ do
    if $u_i > M$ then
        $v_i \leftarrow M$.
    end if
    if $u_i < m$ then
        $v_i \leftarrow m$.
    end if
end for
Set $V_j = N_1 M + N_0 m + v_{m_j} + v_{n_j}$.
Set $A_j = v_{m_j} + v_{n_j} + N_1 M - (N_1 + 2)m$, $B_j = (N_0 + 2)M - v_{m_j} - v_{n_j} - N_0 m$.
if $V_j - U_j > 0$ then
    for $i = m_j, \cdots, n_j$ do
        $v_i \leftarrow v_i - \frac{v_i - m}{A_j}(V_j - U_j)$
    end for
else
    for $i = m_j, \cdots, n_j$ do
        $v_i \leftarrow v_i + \frac{M - v_i}{B_j}(U_j - V_j)$
    end for
end if
```

described by lines 6-8 will not increase the point value of neighbors to an undershoot point thus it will not create new overshoots. Similarly, the operation described by lines 11-13 will not create new undershoots. In other words, no new undershoots (or overshoots) will be created when eliminating overshoots (or undershoots) in **Step II**.

Each interior point u_i in any T_j belongs to one of the following four cases:

1. $u_i \leq m$ or $u_i \geq M$.

2. $m < u_i < M$ and $u_{i-1}, u_{i+1} \leq M$.

3. $m < u_i < M$ and $u_{i-1}, u_{i+1} \geq m$.

4. $m < u_i < M$ and $u_{i-1} > M, u_{i+1} < m$ (or $u_{i+1} > M, u_{i-1} < m$).

We want to show $v_i \in [m, M]$ after **Step II**. For the first three cases, by the same arguments as in the proof of Theorem 6.2.6, we can easily show that the output point values are in the range $[m, M]$. For case (1), after **Step II**, if $u_i \leq m$ then $v_i = m$; if $u_i \geq M$ then $v_i = M$. For case (2), $v_i \neq u_i$ only if at least one of u_{i-1} and u_{i+1} is an undershoot. If so, then

$$\begin{aligned} v_i &= u_i - \frac{(u_i - m)_+ (m - u_{i-1})_+}{(u_{i-2} - m)_+ + (u_i - m)_+} - \frac{(u_i - m)_+ (m - u_{i+1})_+}{(u_i - m)_+ + (u_{i+2} - m)_+} \\ &\geq u_i - \frac{1}{c}(u_i - m)_+ - \frac{1}{c}(u_i - m)_+ > m. \end{aligned}$$

Similarly, for case (3), $v_i \neq u_i$ only if at least one of u_{i-1} and u_{i+1} is an overshoot, and we can show $v_i < M$.

Notice that case (2) and case (3) are not exclusive to each other, which however does not affect the discussion here. When case (2) and case (3) overlap, we have $u_i, u_{i-1}, u_{i+1} \in [m, M]$ thus $v_i = u_i \in [m, M]$ after **Step II**.

For case (4), without loss of generality, we consider the case when $u_{i+1} > M, u_i \in [m, M], u_{i-1} < m$, and we need to show that the output $v_i \in [m, M]$. By Lemma 6.2.2, we know that Algorithm 2 will decrease the value at x_i by at most $\frac{1}{c}(u_i - m)$ to eliminate the undershoot at x_{i-1} then increase the point value at x_i by at most $\frac{1}{c}(M-u_i)$ to eliminate the overshoot at x_{i+1}. So after **Step II**,

$$\begin{aligned} v_i &\leq u_i + \frac{1}{c}(M-u_i) \leq M \quad (\text{because} \quad c \geq 2, u_i < M); \\ v_i &\geq u_i - \frac{1}{c}(u_i - m) \geq m \quad (\text{because} \quad c \geq 2, u_i > m). \end{aligned}$$

Thus we have $v_i \in [m, M]$ after **Step II**. By the same arguments as in the proof of Theorem 6.2.6, we can also easily show the boundary points are in the range $[m, M]$ after **Step II**. It is straightforward to verify that $\sum_{i=1}^{N} v_i = \sum_{i=1}^{N} u_i$ after **Step II** because the operations described by lines 6-8 and lines 11-13 do not change the local sum $v_{i-1} + v_i + v_{i+1}$.

Next we discuss **Step III** in Algorithm 2. Let $\bar{N} = 2 + N_0 + N_1 = n_j - m_j + 1$ be the cardinality of $S_j = \{u_{m_j}, \cdots, u_{n_j}\}$.

We need to show that the average value in each saw-tooth profile S_j is in the range $[m, M]$ after **Step II** before **Step III**. Otherwise it is impossible to enforce the bounds in S_j without changing the sum in S_j. In other words, we need to show $\bar{N}m \leq U_j = \sum_{v_i \in S_j} v_i \leq \bar{N}M$. We will prove the claim by conceptually applying the upper or lower bound limiter Algorithm 1 to S_j. Consider a boundary point of S_j, e.g., $u_{m_j} \in [m, M]$, then during **Step II** the point value at x_{m_j} can be unchanged, moved down at most $\frac{1}{c}(u_{m_j} - m)$ or moved up at most $\frac{1}{c}(M - u_{m_j})$. We first show the average value in S_j after **Step II** is not below m:

(a) Assume both boundary point values of S_j are unchanged during **Step II**. If applying Algorithm 1 to S_j after **Step II**, by the proof of Theorem 6.2.6, we know that the output values would be greater than or equal to m with the same sum, which implies that $\sum_{v_i \in S_j} v_i \geq \bar{N}m$.

(b) If a boundary point value of S_j is increased during **Step II**, the same discussion as in (a) still holds because an increased boundary value does not affect the discussion for the lower bound.

(c) If a boundary point value v_{m_j} of S_j is decreased during **Step II**, then with the fact that it is decreased by at most the amount $\frac{1}{c}(u_{m_j} - m)$, the same discussion as in (a) still holds.

Similarly if applying the upper bound limiter similar to Algorithm 1 to S_j after **Step II**, then by the similar arguments as above, the output values would be less than or equal to M with the same sum, which implies $\sum_{v_i \in S_j} v_i \le \bar{N} M$.

Now we can show the output $v_i \in [m, M]$ for each S_j after **Step III**:

1. Assume $V_j = N_1 M + N_0 m + v_{m_j} + v_{n_j} > U_j$ before the **for** loops in **Step III**. Then after **Step III**: if $u_i < m$ we get $v_i = m$; if $u_i \ge m$ we have

$$\begin{aligned} M \ge \quad v_i \quad & - \frac{v_i - m}{A_j}(V_j - U_j) \\ = \quad v_i \quad & - \frac{v_i - m}{v_{m_j} + v_{n_j} + N_1 M - (N_1 + 2)m}(v_{m_j} + v_{n_j} + N_1 M + N_0 m - U_j) \\ \ge \quad v_i \quad & - \frac{v_i - m}{v_{m_j} + v_{n_j} + N_1 M - (N_1 + 2)m}(v_{m_j} + v_{n_j} + N_1 M + N_0 m - \bar{N} m) \\ = \quad v_i \quad & -(v_i - m) = m. \end{aligned}$$

2. Assume $V_j = N_1 M + N_0 m + v_{m_j} + v_{n_j} \le U_j$ before the **for** loops in **Step III**. Then after **Step III**: if $u_i > M$ we get $v_i = M$; if $u_i \ge M$ we have

$$\begin{aligned} m \le \quad v_i \quad & + \frac{M - v_i}{B_j}(U_j - V_j) \\ = \quad v_i \quad & + \frac{M - v_i}{(N_0 + 2)M - v_{m_j} - v_{n_j} - N_0 m}(U_j - v_{m_j} - v_{n_j} - N_1 M - N_0 m) \\ \le \quad v_i \quad & + \frac{M - v_i}{(N_0 + 2)M - v_{m_j} - v_{n_j} - N_0 m}(\bar{N} M - v_{m_j} - v_{n_j} - N_1 M - N_0 m) \\ = \quad v_i \quad & +(M - v_i) = M. \end{aligned}$$

Thus we have shown all the final output values are in the range $[m, M]$.

Finally it is straightforward to verify that $\sum_{i=1}^{N} v_i = \sum_{i=1}^{N} u_i$. □

The limiters described in Algorithm 1 and Algorithm 2 are high order accurate limiters in the following sense. Assume $u_i (i = 1, \cdots, N)$ are high order accurate approximations to point values of a very smooth function $u(x) \in [m, M]$, i.e., $u_i - u(x_i) = \mathcal{O}(\Delta x^k)$. For fine enough uniform mesh, the global maximum points are well separated from the global minimum points in $\{u_i, i = 1, \cdots, N\}$. In other words, there is no saw-tooth profile in $\{u_i, i =$

$1, \cdots, N\}$. Thus Algorithm 2 reduces to the three-point stencil limiter for smooth profiles on fine resolved meshes. Under these assumptions, the amount which limiter increases/decreases each point value is at most $(u_i - M)_+$ and $(m - u_i)_+$. If $(u_i - M)_+ > 0$, which means $u_i > M \geq u(x_i)$, we have $(u_i - M)_+ = O(\Delta x^k)$ because $(u_i - M)_+ < u_i - u(x_i) = O(\Delta x^k)$. Similarly, we get $(m - u_i)_+ = O(\Delta x^k)$. Therefore, for point values u_i approximating a smooth function, the limiter changes u_i by $O(\Delta x^k)$.

6.2.3 A TVB limiter

The scheme (6.5) can be written into a conservation form:

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x}(\hat{f}_{i+\frac{1}{2}} - \hat{f}_{i-\frac{1}{2}}), \tag{6.8}$$

which is suitable for shock calculations and involves a numerical flux

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2}(f(u_{i+1}^n) + f(u_i^n)). \tag{6.9}$$

To achieve nonlinear stability and eliminate oscillations for shocks, a TVB (total variation bounded in the means) limiter was introduced for the scheme (6.8) in [101]. In this subsection we will show that the bound-preserving property of $\bar{u}_i$ (6.6) still holds for the scheme (6.8) with the TVB limiter in [101]. Thus we can use both the TVB limiter and the bound-preserving limiter in Algorithm (2) at the same time.

The compact finite difference scheme with the limiter in [101] is

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x}(\hat{f}^{(m)}_{i+\frac{1}{2}} - \hat{f}^{(m)}_{i-\frac{1}{2}}), \tag{6.10}$$

where the numerical flux $\hat{f}^{(m)}_{i+\frac{1}{2}}$ is the modified flux approximating (6.9).

First we write $f(u) = f^+(u) + f^-(u)$ with the requirement that $\frac{\partial f^+(u)}{\partial u} \geq 0$, and $\frac{\partial f^-(u)}{\partial u} \leq 0$. The simplest such splitting is the Lax-Friedrichs splitting $f^\pm(u) = \frac{1}{2}(f(u) \pm \alpha u), \alpha =$

$\max\limits_{u\in[m,M]} |f(u)|$. Then we write the flux $\hat{f}_{i+\frac{1}{2}}$ as $\hat{f}_{i+\frac{1}{2}} = \hat{f}^+_{i+\frac{1}{2}} + \hat{f}^-_{i+\frac{1}{2}}$, where $\hat{f}^\pm_{i+\frac{1}{2}}$ are obtained by adding superscripts ± in (6.9). Next we define

$$d\hat{f}^+_{i+\frac{1}{2}} = \hat{f}^+_{i+\frac{1}{2}} - f^+(\bar{u}_i), \quad d\hat{f}^-_{i+\frac{1}{2}} = f^-(\bar{u}_{i+1}) - \hat{f}^-_{i+\frac{1}{2}}.$$

Here $d\hat{f}^\pm_{i+\frac{1}{2}}$ are the differences between the numerical fluxes $\hat{f}^\pm_{i+\frac{1}{2}}$ and the first-order, upwind fluxes $f^+(\bar{u}_i)$ and $f^-(\bar{u}_{i+1})$. The limiting is defined by

$$d\hat{f}^{+(m)}_{i+\frac{1}{2}} = \tilde{m}(d\hat{f}^+_{i+\frac{1}{2}}, \Delta^+ f^+(\bar{u}_i), \Delta^+ f^+(\bar{u}_{i-1})), \quad d\hat{f}^{-(m)}_{i+\frac{1}{2}} = \tilde{m}(d\hat{f}^-_{i+\frac{1}{2}}, \Delta^+ f^-(\bar{u}_i), \Delta^+ f^-(\bar{u}_{i+1})),$$

where $\Delta^+ v_i \equiv v_{i+1} - v_i$ is the usual forward difference operator, and the modified *minmod* function $\tilde{m}$ is defined by

$$\tilde{m}(a_1, \dots, a_k) = \begin{cases} a_1, & \text{if } |a_1| \le p\Delta x^2, \\ m(a_1, \dots, a_k), & \text{otherwise}, \end{cases} \tag{6.11}$$

where p is a positive constant independent of Δx and m is the *minmod* function

$$m(a_1, \dots, a_k) = \begin{cases} s \min_{1\le i\le k} |a_i|, & \text{if } sign(a_1) = \cdots = sign(a_k) = s, \\ 0, & \text{otherwise}. \end{cases}$$

The limited numerical flux is then defined by $\hat{f}^{+(m)}_{i+\frac{1}{2}} = f^+(\bar{u}_i) + d\hat{f}^{+(m)}_{i+\frac{1}{2}}$, $\quad \hat{f}^{-(m)}_{i+\frac{1}{2}} = f^-(\bar{u}_{i+1}) - d\hat{f}^{-(m)}_{i+\frac{1}{2}}$,and $\hat{f}^{(m)}_{i+\frac{1}{2}} = \hat{f}^{+(m)}_{i+\frac{1}{2}} + \hat{f}^{-(m)}_{i+\frac{1}{2}}$. The following result was proved in [101]:

Lemma 6.2.8. *For any n and Δt such that $0 \le n\Delta t \le T$, scheme (6.10) is TVBM (total variation bounded in the means): $TV(\bar{u}^n) = \sum_i |\bar{u}^n_{i+1} - \bar{u}^n_i| \le C$, where C is independent of Δt, under the CFL condition $\max_u(\frac{\partial}{\partial u} f^+(u) - \frac{\partial}{\partial u} f^-(u))\frac{\Delta t}{\Delta x} \le \frac{1}{2}$.*

Next we show that the TVB scheme still satisfies (6.6).

Theorem 6.2.9. *If $u^n_i \in [m, M]$, then under a suitable CFL condition, the TVB scheme (6.10) satisfies $m \le \frac{1}{6}(u^{n+1}_{i-1} + 4u^{n+1}_i + u^{n+1}_{i+1}) \le M$.*

Proof. Let $\lambda = \frac{\Delta t}{\Delta x}$, then we have

$$\begin{aligned}\bar{u}_i^{n+1} &= \bar{u}_i^n - \lambda(\hat{f}^{(m)}_{i+\frac{1}{2}} - \hat{f}^{(m)}_{i-\frac{1}{2}}) \\ &= \frac{1}{4}(\bar{u}_i^n - 4\lambda\hat{f}^{+(m)}_{i+\frac{1}{2}}) + \frac{1}{4}(\bar{u}_i^n - 4\lambda\hat{f}^{-(m)}_{i+\frac{1}{2}}) + \frac{1}{4}(\bar{u}_i^n + 4\lambda\hat{f}^{+(m)}_{i-\frac{1}{2}}) + \frac{1}{4}(\bar{u}_i^n + 4\lambda\hat{f}^{-(m)}_{i-\frac{1}{2}}).\end{aligned}$$

We will show $\bar{u}_i^{n+1} \in [m, M]$ by proving that the four terms satisfy

$$\begin{aligned}\bar{u}_i^n - 4\lambda\hat{f}^{+(m)}_{i+\frac{1}{2}} &\in [m - 4\lambda f^+(m), M - 4\lambda f^+(M)], \\ \bar{u}_i - 4\lambda\hat{f}^{-(m)}_{i+\frac{1}{2}} &\in [m - 4\lambda f^-(m), M - 4\lambda f^-(M)], \\ \bar{u}_i^n + 4\lambda\hat{f}^{+(m)}_{i-\frac{1}{2}} &\in [m + 4\lambda f^+(m), M + 4\lambda f^+(M)], \\ \bar{u}_i + 4\lambda\hat{f}^{-(m)}_{i-\frac{1}{2}} &\in [m + 4\lambda f^-(m), M + 4\lambda f^-(M)],\end{aligned}$$

under the CFL condition

$$\lambda \max_u |f^{(\pm)}(u)| \leq \frac{1}{12}. \tag{6.12}$$

We only discuss the first term since the proof for the rest is similar. We notice that $u - 4\lambda f^+(u)$ and $u - 12\lambda f^+(u)$ are monotonically increasing functions of u under the CFL constraint (6.12), thus $u \in [m, M]$ implies $u - 4\lambda f^+(u) \in [m - 4\lambda f^+(m), M - 4\lambda f^+(M)]$ and $u - 12\lambda f^+(u) \in [m - 12\lambda f^+(m), M - 12\lambda f^+(M)]$. For convenience, we drop the time step n, then we have

$$\bar{u}_i - 4\lambda\hat{f}^{+(m)}_{i+\frac{1}{2}} = \bar{u}_i - 4\lambda(f^+(\bar{u}_i) + d\hat{f}^{+(m)}_{i+\frac{1}{2}}),$$

where the value of $d\hat{f}^{+(m)}_{i+\frac{1}{2}}$ has four possibilities:

1. If $d\hat{f}^{+(m)}_{i+\frac{1}{2}} = 0$, then

$$\bar{u}_i - 4\lambda\hat{f}^{+(m)}_{i+\frac{1}{2}} = \bar{u}_i - 4\lambda f^+(\bar{u}_i) \in [m - 4\lambda f^+(m), M - 4\lambda f^+(M)].$$

2. If $d\hat{f}^{+(m)}_{i+\frac{1}{2}} = d\hat{f}^{+}_{i+\frac{1}{2}}$, then we get

$$\begin{aligned}\bar{u}_i - 4\lambda \hat{f}^{+(m)}_{i+\frac{1}{2}} &= \frac{1}{6}(u_{i-1} + 4u_i + u_{i+1}) - 4\lambda \frac{f^+(u_i) + f^+(u_{i+1})}{2} \\ &= \frac{1}{6}u_{i-1} + \frac{2}{3}(u_i - 3\lambda f^+(u_i)) + \frac{1}{6}(u_{i+1} - 12\lambda f^+(u_{i+1})).\end{aligned}$$

By the monotonicity of the function $u - 12\lambda f^+(u)$ and $u - 3\lambda f^+(u)$, we have

$$u_i - 3\lambda f^+(u_i) \in [m - 3\lambda f^+(m), M - 3\lambda f^+(M)],$$

$$u_{i+1} - 12\lambda f^+(u_{i+1}) \in [m - 12\lambda f^+(m), M - 12\lambda f^+(M)],$$

which imply $\bar{u}_i - 4\lambda \hat{f}^{+(m)}_{i+\frac{1}{2}} \in [m - 4\lambda f^+(m), M - 4\lambda f^+(M)]$.

3. If $d\hat{f}^{+(m)}_{i+\frac{1}{2}} = \Delta^+ f^+(\bar{u}_i)$, $\bar{u}_i - 4\lambda \hat{f}^{+(m)}_{i+\frac{1}{2}} = \bar{u}_i - 4\lambda f^+(\bar{u}_{i+1})$. If $\Delta^+ f^+(\bar{u}_i) > 0$, $\bar{u}_i - 4\lambda f^+(\bar{u}_{i+1}) < \bar{u}_i - 4\lambda f^+(\bar{u}_i) \leq M - 4\lambda f^+(M)$, which implies the upper bound holds. Due to the definition of the *minmod* function, we can get $0 < \Delta^+ f^+(\bar{u}_i) < d\hat{f}^{+}_{i+\frac{1}{2}}$. Thus, $\hat{f}^{+}_{i+\frac{1}{2}} = \frac{f^+(u_i)+f^+(u_{i+1})}{2} = f^+(\bar{u}_i) + d\hat{f}^{+}_{i+\frac{1}{2}} > f^+(\bar{u}_i) + \Delta^+ f^+(\bar{u}_i) = f^+(\bar{u}_{i+1})$. Then, $\bar{u}_i - 4\lambda f^+(\bar{u}_{i+1}) > \bar{u}_i - 4\lambda \frac{f^+(u_i)+f^+(u_{i+1})}{2} \geq m - 4\lambda f^+(m)$, which gives the lower bound. For the case $\Delta^+ f^+(\bar{u}_i) < 0$, the proof is similar.

4. If $d\hat{f}^{+(m)}_{i+\frac{1}{2}} = \Delta^+ f^+(\bar{u}_{i-1})$, the proof is the same as the previous case.

□

6.2.4 One-dimensional convection diffusion problems

We consider the one-dimensional convection diffusion problems with periodic boundary conditions: $u_t + f(u)_x = a(u)_{xx}$, $u(x, 0) = u_0(x)$, where $a(u) \geq 0$. Let $\mathbf{f}^n$ denote the column vector with entries $f(u^n_1), \cdots, f(u^n_N)$. By notations introduced in Section 6.2.1, the fourth-order compact finite difference with forward Euler can be denoted as:

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \frac{\Delta t}{\Delta x} W_1^{-1} D_x \mathbf{f}^n + \frac{\Delta t}{\Delta x^2} W_2^{-1} D_{xx} \mathbf{a}^n. \tag{6.13}$$

Recall that we have abused the notation by using $W_1 f_i^n$ to denote the i-th entry of the vector $W_1\mathbf{f}^n$ and we have defined $\bar{u}_i = W_1 u_i$. We now define

$$\tilde{u}_i = W_2 u_i.$$

Notice that W_1 and W_2 are both circulant thus they both can be diagonalized by the discrete Fourier matrix, so W_1 and W_2 commute. Thus we have

$$\tilde{\bar{u}}_i = (W_2 W_1 \mathbf{u})_i = (W_1 W_2 \mathbf{u})_i = \bar{\tilde{u}}_i.$$

Let $f_i^n = f(u_i^n)$ and $a_i^n = a(u_i^n)$, then the scheme (6.13) can be written as

$$\bar{\tilde{u}}_i^{n+1} = \bar{\tilde{u}}_i^n - \frac{\Delta t}{\Delta x} W_2 D_x f_i^n + \frac{\Delta t}{\Delta x^2} W_1 D_{xx} a_i^n.$$

Theorem 6.2.10. *Under the CFL constraint* $\frac{\Delta t}{\Delta x}\max_u |f(u)| \leq \frac{1}{6}$, $\frac{\Delta t}{\Delta x^2}\max_u a(u) \leq \frac{5}{24}$, *if* $u_i^n \in [m, M]$, *then the scheme* (6.13) *satisfies that* $m \leq \bar{\tilde{u}}_i^{n+1} \leq M$.

Proof. Let $\lambda = \frac{\Delta t}{\Delta x}$ and $\mu = \frac{\Delta t}{\Delta x^2}$. We can rewrite the scheme (6.13) as

$$\mathbf{u}^{n+1} = \frac{1}{2}(\mathbf{u}^n - 2\lambda W_1^{-1} D_x \mathbf{f}^n) + \frac{1}{2}(\mathbf{u}^n + 2\mu W_2^{-1} D_{xx}\mathbf{a}^n),$$

$$W_2 W_1 \mathbf{u}^{n+1} = \frac{1}{2} W_2 (W_1 \mathbf{u}^n - 2\lambda D_x \mathbf{f}^n) + \frac{1}{2} W_1 (W_2 \mathbf{u}^n + 2\mu D_{xx}\mathbf{a}^n),$$
$$\bar{\tilde{u}}_i^{n+1} = \frac{1}{2} W_2 (\bar{u}_i^n - 2\lambda D_x f_i^n) + \frac{1}{2} W_1 (\tilde{u}_i^n + 2\mu D_{xx} a_i^n).$$

By Theorem 6.2.1, we have $\bar{u}_i^n - 2\lambda D_x f_i^n \in [m, M]$. We also have

$$\tilde{u}_i^n + 2\mu D_{xx} a_i^n = \frac{1}{12}(u_{-1}^n + 10u_i^n + u_{i+1}^n) + 2\mu(a_{i-1}^n - 2a_i^n + a_{i+1}^n)$$
$$= \left(\frac{5}{6}u_i^n - 4\mu a_i^n\right) + \left(\frac{1}{12}u_{i-1}^n + 2\mu a_{i-1}^n\right) + \left(\frac{1}{12}u_{i+1}^n + 2\mu a_{i+1}^n\right).$$

Due to monotonicity under the CFL constraint and the assumption $a(u) \geq 0$, we get $\tilde{u}_i^n + 2\mu D_{xx} a_i^n \in [m, M]$. Thus we get $\bar{\tilde{u}}_i^{n+1} \in [m, M]$ since it is a convex combination of $\bar{u}_i^n - 2\lambda D_x f_i^n$ and $\tilde{u}_i^n + 2\mu D_{xx} a_i^n$. □

Given point values u_i satisfying $\bar{\tilde{u}}_i \in [m, M]$ for any i, Lemma 6.2.2 no longer holds since $\bar{\tilde{u}}_i$ has a five-point stencil. However, the same three-point stencil limiter in Algorithm 2 can still be used to enforce the lower and upper bounds. Given $\bar{\tilde{u}}_i = W_2 W_1 u_i$ $i = 1, \cdots, N$, conceptually we can obtain the point values u_i by first computing $\bar{u}_i = W_2^{-1}\bar{\tilde{u}}_i$ then computing $u_i = W_1^{-1}\bar{u}_i$. Thus we can apply the limiter in Algorithm 2 twice to enforce $u_i \in [m, M]$:

1. Given $\bar{\tilde{u}}_i \in [m, M]$, compute $\bar{u}_i = W_2^{-1}\bar{\tilde{u}}_i$ which are not necessarily in the range $[m, M]$. Then apply the limiter in Algorithm 2 to $\bar{u}_i, i = 1, \cdots, N$. Let $\bar{v}_i$ denote the output of the limiter. Since we have

$$\bar{\tilde{u}}_i = \tilde{\bar{u}}_i = \frac{1}{c+2}(\bar{u}_{i-1} + c\bar{u}_i + \bar{u}_{i+1}), \quad c = 10,$$

   all discussions in Section 6.2.2 are still valid, thus we have $\bar{v}_i \in [m, M]$.

2. Compute $u_i = W_1^{-1}\bar{v}_i$. Apply the limiter in Algorithm 2 to $u_i, i = 1, \cdots, N$. Let v_i denote the output of the limiter. Then we have $v_i \in [m, M]$.

6.2.5 High order time discretizations

For high order time discretizations, we can use strong stability preserving (SSP) Runge-Kutta and multistep methods, which are convex combinations of formal forward Euler steps. Thus if using the limiter in Algorithm 2 for fourth order compact finite difference schemes considered in this section on each stage in a SSP Runge-Kutta method or each time step in a SSP multistep method, the bound-preserving property still holds.

In the numerical tests, we will use a fourth order SSP multistep method and a fourth order SSP Runge-Kutta method [97]. Now consider solving $u_t = F(u)$. The SSP coefficient C for a SSP time discretization is a constant so that the high order SSP time discretization is stable in a norm or a semi-norm under the time step restriction $\Delta t \leq C\Delta t_0$, if under the time step restriction $\Delta t \leq \Delta t_0$ the forward Euler is stable in the same norm or semi-norm. The fourth order SSP Multistep method (with SSP coefficient $C_{ms} = 0.1648$) and the fourth order SSP Runge-Kutta method (with SSP coefficient $C_{rk} = 1.508$) will be used in the numerical tests. See [97] for their definitions.

In Section 6.2.2 we have shown that the limiters in Algorithm 1 and Algorithm 2 are high order accurate provided u_i are high order accurate approximations to a smooth function $u(x) \in [m, M]$. This assumption holds for the numerical solution in a multistep method in each time step, but it is no longer true for inner stages in the Runge-Kutta method. So only SSP multistep methods with the limiter Algorithm 2 are genuinely high order accurate schemes. For SSP Runge-Kutta methods, using the bound-preserving limiter for compact finite difference schemes might result in an order reduction. The order reduction for bound-preserving limiters for finite volume and DG schemes with Runge-Kutta methods was pointed out in [23] due to the same reason. However, such an order reduction in compact finite difference schemes is more prominent, as we will see in the numerical tests.

6.3 Extensions To Two-dimensional Problems

In this section we consider initial value problems on a square $[0,1] \times [0,1]$ with periodic boundary conditions. Let $(x_i, y_j) = (\frac{i}{N_x}, \frac{j}{N_y})$ $(i = 1, \cdots, N_x, j = 1, \cdots, N_y)$ be the uniform grid points on the domain $[0,1] \times [0,1]$. For a periodic function $f(x, y)$ on $[0,1] \times [0,1]$, let $\mathbf{f}$ be a matrix of size $N_x \times N_y$ with entries f_{ij} representing point values $f(u_{ij})$. We first define two linear operators W_{1x} and W_{1y} from $\mathbb{R}^{N_x \times N_y}$ to $\mathbb{R}^{N_x \times N_y}$:

$$W_{1x}\mathbf{f} = \frac{1}{6}\begin{pmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{pmatrix}_{N_x \times N_x} \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1,N_y} \\ f_{21} & f_{22} & \cdots & f_{2,N_y} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N_x-1,1} & f_{N_x-1,2} & \cdots & f_{N_x-1,N_y} \\ f_{N_x,1} & f_{N_x,2} & \cdots & f_{N_x,N_y} \end{pmatrix},$$

$$W_{1y}\mathbf{f} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1,N_y} \\ f_{21} & f_{22} & \cdots & f_{2,N_y} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N_x-1,1} & f_{N_x-1,2} & \cdots & f_{N_x-1,N_y} \\ f_{N_x,1} & f_{N_x,2} & \cdots & f_{N_x,N_y} \end{pmatrix} \frac{1}{6}\begin{pmatrix} 4 & 1 & & & 1 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ 1 & & & 1 & 4 \end{pmatrix}_{N_y \times N_y}.$$

We can define W_{2x}, W_{2y}, D_x, D_y, W_{2x} and W_{2y} similarly such that the subscript x denotes the multiplication of the corresponding matrix from the left for the x-index and the subscript y denotes the multiplication of the corresponding matrix from the right for the y-index. We abuse the notations by using $W_{1x}f_{ij}$ to denote the (i,j) entry of $W_{1x}\mathbf{f}$. We only discuss the forward Euler from now on since the discussion for high order SSP time discretizations are the same as in Section 6.2.5.

6.3.1 Two-dimensional convection equations

Consider solving the two-dimensional convection equation: $u_t+f(u)_x+g(u)_y=0, \quad u(x,y,0)=u_0(x,y)$. By the our notations, the fourth order compact scheme with the forward Euler time discretization can be denoted as:

$$u_{ij}^{n+1}=u_{ij}^{n}-\frac{\Delta t}{\Delta x}W_{1x}^{-1}D_x f_{ij}^{n}-\frac{\Delta t}{\Delta y}W_{1y}^{-1}D_y g_{ij}^{n}. \tag{6.14}$$

We define $\bar{\mathbf{u}}^n=W_{1x}W_{1y}\mathbf{u}^n$, then by applying $W_{1y}W_{1x}$ to both sides, (6.14) becomes

$$\bar{u}_{ij}^{n+1}=\bar{u}_{ij}^{n}-\frac{\Delta t}{\Delta x}W_{1y}D_x f_{ij}^{n}-\frac{\Delta t}{\Delta y}W_{1x}D_y g_{ij}^{n}. \tag{6.15}$$

Theorem 6.3.1. *Under the CFL constraint*

$$\frac{\Delta t}{\Delta x}\max_u|f(u)|+\frac{\Delta t}{\Delta y}\max_u|g(u)|\leq\frac{1}{3}, \tag{6.16}$$

if $u_{ij}^n\in[m,M]$, then the scheme (6.15) satisfies $\bar{u}_{ij}^{n+1}\in[m,M]$.

Proof. For convenience, we drop the time step n in u_{ij}^n, f_{ij}^n, and introduce:

$$U=\begin{pmatrix}u_{i-1,j+1} & u_{i,j+1} & u_{i+1,j+1}\\ u_{i-1,j} & u_{i,j} & u_{i+1,j}\\ u_{i-1,j-1} & u_{i,j-1} & u_{i+1,j-1}\end{pmatrix}, \quad F=\begin{pmatrix}f_{i-1,j+1} & f_{i,j+1} & f_{i+1,j+1}\\ f_{i-1,j} & f_{i,j} & f_{i+1,j}\\ f_{i-1,j-1} & f_{i,j-1} & f_{i+1,j-1}\end{pmatrix}.$$

Let $\lambda_1 = \frac{\Delta t}{\Delta x}$ and $\lambda_2 = \frac{\Delta t}{\Delta y}$, then the scheme (6.15) can be written as

$$\begin{aligned}\bar{u}_{ij}^{n+1} &= W_{1y}W_{1x}u_{ij}^n - \lambda_1 W_{1y}D_x f_{ij}^n - \lambda_2 W_{1x}D_y g_{ij}^n,\\ &= \frac{1}{36}\begin{pmatrix}1 & 4 & 1\\ 4 & 16 & 4\\ 1 & 4 & 1\end{pmatrix} : U - \frac{\lambda_1}{12}\begin{pmatrix}-1 & 0 & 1\\ -4 & 0 & 4\\ -1 & 0 & 1\end{pmatrix} : F - \frac{\lambda_2}{12}\begin{pmatrix}1 & 4 & 1\\ 0 & 0 & 0\\ -1 & -4 & -1\end{pmatrix} : G,\end{aligned}$$

where : denotes the sum of all entrywise products in two matrices of the same size. Obviously the right hand side above is a monotonically increasing function with respect to u_{lm} for $i-1 \le l \le i+1$, $j-1 \le m \le j+1$ under the CFL constraint (6.16). The monotonicity implies the bound-preserving result of $\bar{u}_{ij}^{n+1}$. □

Given $\bar{u}_{ij}$, we can recover point values u_{ij} by obtaining first $v_{ij} = W_{1x}^{-1}\bar{u}_{ij}$ then $u_{ij} = W_{1y}^{-1}v_{ij}$. Thus similar to the discussions in Section 6.2.4, given point values u_{ij} satisfying $\bar{u}_{ij} \in [m, M]$ for any i and j, we can use the limiter in Algorithm 2 in a dimension by dimension fashion to enforce $u_{ij} \in [m, M]$:

1. Given $\bar{u}_{ij} \in [m, M]$, compute $v_{ij} = W_{1x}^{-1}\bar{u}_{ij}$ which are not necessarily in the range $[m, M]$. Then apply the limiter in Algorithm 2 to v_{ij} $(i = 1, \cdots, N_x)$ for each fixed j. Since we have
$$\bar{u}_{ij} = \frac{1}{c+2}(v_{i-1,j} + cv_{i,j} + v_{i+1,j}), \quad c = 4,$$
all discussions in Section 6.2.2 are still valid. Let $\bar{v}_{ij}$ denote the output of the limiter, thus we have $\bar{v}_{ij} \in [m, M]$.

2. Compute $u_{ij} = W_{1y}^{-1}\bar{v}_{ij}$. Then we have
$$\bar{v}_{ij} = \frac{1}{c+2}(u_{i,j-1} + cu_{i,j} + u_{i,j+1}), \quad c = 4.$$
Apply the limiter in Algorithm 2 to u_{ij} $(j = 1, \cdots, N_y)$ for each fixed i. Then the output values are in the range $[m, M]$.

6.3.2 Two-dimensional convection diffusion equations

Consider the two-dimensional convection diffusion problem:

$$u_t + f(u)_x + g(u)_y = a(u)_{xx} + b(u)_{xx}, \quad u(x,y,0) = u_0(x,y),$$

where $a(u) \geq 0$ and $b(u) \geq 0$. A fourth-order accurate compact finite difference scheme can be written as

$$\frac{d\mathbf{u}}{dt} = -\frac{1}{\Delta x} W_{1x}^{-1} D_x \mathbf{f} - \frac{1}{\Delta y} W_{1y}^{-1} D_y \mathbf{g} + \frac{1}{\Delta x^2} W_{2x}^{-1} D_{xx} \mathbf{a} + \frac{1}{\Delta y^2} W_{2y}^{-1} D_{yy} \mathbf{b}.$$

Let $\lambda_1 = \frac{\Delta t}{\Delta x}$, $\lambda_2 = \frac{\Delta t}{\Delta y}$, $\mu_1 = \frac{\Delta t}{\Delta x^2}$ and $\mu_2 = \frac{\Delta t}{\Delta y^2}$. With the forward Euler time discretization, the scheme becomes

$$u_{ij}^{n+1} = u_{ij}^n - \lambda_1 W_{1x}^{-1} D_x f_{ij}^n - \lambda_2 W_{1y}^{-1} D_y g_{ij}^n + \mu_1 W_{2x}^{-1} D_{xx} a_{ij}^n + \mu_2 W_{2y}^{-1} D_{yy} b_{ij}^n. \tag{6.17}$$

We first define $\bar{\mathbf{u}} = W_{1x} W_{1y} \mathbf{u}$ and $\tilde{\mathbf{u}} = W_{2x} W_{2y} \mathbf{u}$, where $W_1 = W_{1x} W_{1y}$ and $W_2 = W_{2x} W_{2y}$. Due to the fact $W_1 W_2 = W_2 W_1$, we have

$$\tilde{\bar{\mathbf{u}}} = W_{2x} W_{2y} (W_{1x} W_{1y} \mathbf{u}) = W_{1x} W_{1y} (W_{2x} W_{2y} \mathbf{u}) = \bar{\tilde{\mathbf{u}}}.$$

The scheme (6.17) is equivalent to the following form:

$$\begin{aligned} \tilde{\bar{u}}_{ij}^{n+1} = \tilde{\bar{u}}_{ij}^n &- \lambda_1 W_{1y} W_{2x} W_{2y} D_x f_{ij}^n - \lambda_2 W_{1x} W_{2x} W_{2y} D_y g_{ij}^n \\ &+ \mu_1 W_{1x} W_{1y} W_{2y} D_{xx} a_{ij}^n + \mu_2 W_{1x} W_{1y} W_{2x} D_{yy} b_{ij}^n. \end{aligned}$$

Theorem 6.3.2. *Under the CFL constraint*

$$\frac{\Delta t}{\Delta x} \max_u |f(u)| + \frac{\Delta t}{\Delta y} \max_u |g(u)| \leq \frac{1}{6}, \frac{\Delta t}{\Delta x^2} \max_u a(u) + \frac{\Delta t}{\Delta y^2} \max_u b(u) \leq \frac{5}{24}, \tag{6.18}$$

if $u_{ij}^n \in [m, M]$, then the scheme (6.17) satisfies $\tilde{\bar{u}}_{ij}^{n+1} \in [m, M]$.

Proof. By using $\tilde{\bar{u}}^n_{ij} = \frac{1}{2}\tilde{\bar{u}}^n_{ij} + \frac{1}{2}\tilde{\bar{u}}^n_{ij}$, we obtain

$$\begin{aligned}\tilde{\bar{u}}^{n+1}_{ij} &= \frac{1}{2}W_{2x}W_{2y}[\bar{u}^n_{ij} - 2\lambda_1 W_{1y}D_x f^n_{ij} - 2\lambda_2 W_{1x}D_y g^n_{ij}] \\ &+\frac{1}{2}W_{1x}W_{1y}[\tilde{u}^n_{ij} + 2\mu_1 W_{2y}D_{xx}a^n_{ij} + 2\mu_2 W_{2x}D_{yy}b^n_{ij}].\end{aligned}$$

Let $\bar{v}_{ij} = \bar{u}^n_{ij} - 2\lambda_1 W_{1y}D_x f^n_{ij} - 2\lambda_2 W_{1x}D_y g^n_{ij}$, $\tilde{w}_{ij} = \tilde{u}^n_{ij} + 2\mu_1 W_{2y}D_{xx}a^n_{ij} + 2\mu_2 W_{2x}D_{yy}b^n_{ij}$. Then by the same discussion as in the proof of Theorem 6.3.1, we can show $\bar{v}_{ij} \in [m, M]$. For $\tilde{w}_{ij}$, it can be written as

$$\tilde{w}_{ij} = \frac{1}{144}\begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : U + \frac{\mu_1}{6}\begin{pmatrix} 1 & -2 & 1 \\ 10 & -20 & 10 \\ 1 & -2 & 1 \end{pmatrix} : A + \frac{\mu_2}{6}\begin{pmatrix} 1 & 10 & 1 \\ -2 & -20 & -2 \\ 1 & 10 & 1 \end{pmatrix} : B,$$

$$A = \begin{pmatrix} a_{i-1,j+1} & a_{i,j+1} & a_{i+1,j+1} \\ a_{i-1,j} & a_{i,j} & a_{i+1,j} \\ a_{i-1,j-1} & a_{i,j-1} & a_{i+1,j-1} \end{pmatrix}, \quad B = \begin{pmatrix} b_{i-1,j+1} & b_{i,j+1} & b_{i+1,j+1} \\ b_{i-1,j} & b_{i,j} & b_{i+1,j} \\ b_{i-1,j-1} & b_{i,j-1} & b_{i+1,j-1} \end{pmatrix}.$$

Under the CFL constraint (6.18), $\tilde{w}_{ij}$ is a monotonically increasing function of u^n_{ij} involved thus $\tilde{w}_{ij} \in [m, M]$. Therefore, $\tilde{\bar{u}}^{n+1}_{ij} \in [m, M]$. □

Given $\tilde{\bar{u}}_{ij}$, we can recover point values u_{ij} by obtaining first $\tilde{u}_{ij} = W_{1x}^{-1}W_{1y}^{-1}\tilde{\bar{u}}_{ij}$ then $u_{ij} = W_{2x}^{-1}W_{2y}^{-1}\tilde{u}_{ij}$. Thus similar to the discussions in the previous subsection, given point values u_{ij} satisfying $\tilde{\bar{u}}_{ij} \in [m, M]$ for any i and j, we can use the limiter in Algorithm 2 dimension by dimension several times to enforce $u_{ij} \in [m, M]$:

1. Given $\tilde{\bar{u}}_{ij} \in [m, M]$, compute $\tilde{u}_{ij} = W_{1x}^{-1}W_{1y}^{-1}\tilde{\bar{u}}_{ij}$ and apply the limiting algorithm in the previous subsection to ensure $\tilde{u}_{ij} \in [m, M]$.

2. Compute $v_{ij} = W_{2x}^{-1}\tilde{u}_{ij}$ which are not necessarily in the range $[m, M]$. Then apply the limiter in Algorithm 2 to v_{ij} for each fixed j. Since we have

$$\tilde{u}_{ij} = \frac{1}{c+2}(v_{i-1,j} + cv_{i,j} + v_{i+1,j}), c = 10,$$

all discussions in Section 6.2.2 are still valid. Let $\tilde{v}_{ij}$ denote the output of the limiter, thus we have $\tilde{v}_{ij} \in [m, M]$.

3. Compute $u_{ij} = W_{2y}^{-1}\tilde{v}_{ij}$. Then we have $\tilde{v}_{ij} = \frac{1}{c+2}(u_{i,j-1} + cu_{i,j} + u_{i,j+1}), \quad c = 10$. Apply the limiter in Algorithm 2 to u_{ij} for each fixed i. Then the output values are in the range $[m, M]$.

6.4 Two-dimensional Incompressible Navier-Stokes Equation

In this section we consider the two-dimensional incompressible Navier-Stokes equation in the vorticity stream-function form:

$$\omega_t + (u\omega)_x + (v\omega)_y = \frac{1}{Re}\Delta\omega, \tag{6.19a}$$

$$\psi = \Delta\omega, \tag{6.19b}$$

$$\langle u, v\rangle = \langle -\psi_y, \psi_x\rangle, \tag{6.19c}$$

where ω is the vorticity, ψ is the stream function, $\langle u, v\rangle$ is the velocity and Re is the Reynolds number. The equation (6.19c) implies the incompressiblility condition

$$u_x + v_y = 0. \tag{6.20}$$

Due to (6.20), (6.19a) is equivalent to

$$\omega_t + u\omega_x + v\omega_y = \frac{1}{Re}\Delta\omega, \tag{6.21}$$

for which the initial value problem satisfies the same bound-preserving property as discussed before:

$$\min_{x,y} \omega(x, y, 0) = m \leq \omega(x, y, t) \leq M = \max_{x,y} \omega(x, y, 0).$$

If solving (6.21) directly, it is usually easier to achieve a bound-preserving scheme. But for the sake of conservation, it is desired to solve the conservative form equation (6.19a). In order to enforce the bound-preserving property for (6.19a) without losing accuracy, the divergence free constraint (6.20) must be properly used since the bound-preserving property may not hold for (6.19a) without (6.20), see [23], [102], [103].

For simplicity, we only consider a periodic boundary condition on a square $[0,1]\times[0,1]$. Let $(x_i, y_j) = (\frac{i}{N_x}, \frac{j}{N_y})$ $(i = 1,\cdots,N_x, j = 1,\cdots,N_y)$ be the uniform grid points on the domain $[0,1]\times[0,1]$. All notations are the same as in the previous section.

Standard fourth order compact finite difference schemes can be used to solve the Poisson equation (6.19b). Efficient Fourier-based Poisson solvers can be constructed.

6.4.1 Incompressible Euler equations

For simplicity, we first consider how to achieve the weak monotonicity for the incompressible Euler equations

$$\omega_t + (u\omega)_x + (v\omega)_y = 0. \tag{6.22}$$

A fourth order compact finite difference scheme with the forward Euler method for (6.22) can be given as

$$\omega_{ij}^{n+1} = \omega_{ij}^n - \lambda_1[W_{1x}^{-1}D_x(\mathbf{u}^n\circ\boldsymbol{\omega}^n)]_{ij} - \lambda_2[W_{1y}^{-1}D_y(\mathbf{u}^n\circ\boldsymbol{\omega}^n)]_{ij}, \tag{6.23}$$

and it is equivalent to

$$\begin{aligned}\bar{\omega}_{ij}^{n+1} &= \bar{\omega}_{ij}^n - \lambda_1[W_{1y}D_x(\mathbf{u}^n\circ\boldsymbol{\omega}^n)]_{ij} - \lambda_2[W_{1x}D_y(\mathbf{v}^n\circ\boldsymbol{\omega}^n)]_{ij}\\ &= \frac{1}{36}\begin{pmatrix}1 & 4 & 1\\ 4 & 16 & 4\\ 1 & 4 & 1\end{pmatrix} : \Omega^n - \frac{\lambda_1}{12}\begin{pmatrix}-1 & 0 & 1\\ -4 & 0 & 4\\ -1 & 0 & 1\end{pmatrix} : (U^n\circ\Omega^n) - \frac{\lambda_2}{12}\begin{pmatrix}1 & 4 & 1\\ 0 & 0 & 0\\ -1 & -4 & -1\end{pmatrix} : (V^n\circ\Omega^n),\end{aligned} \tag{6.24}$$

where $\circ$ denotes the matrix Hadamard product, and

$$U = \begin{pmatrix}u_{i-1,j+1} & u_{i,j+1} & u_{i+1,j+1}\\ u_{i-1,j} & u_{i,j} & u_{i+1,j}\\ u_{i-1,j-1} & u_{i,j-1} & u_{i+1,j-1}\end{pmatrix}, V = \begin{pmatrix}v_{i-1,j+1} & v_{i,j+1} & v_{i+1,j+1}\\ v_{i-1,j} & v_{i,j} & v_{i+1,j}\\ v_{i-1,j-1} & v_{i,j-1} & v_{i+1,j-1}\end{pmatrix}, \Omega = \begin{pmatrix}\omega_{i-1,j+1} & \omega_{i,j+1} & \omega_{i+1,j+1}\\ \omega_{i-1,j} & \omega_{i,j} & \omega_{i+1,j}\\ \omega_{i-1,j-1} & \omega_{i,j-1} & \omega_{i+1,j-1}\end{pmatrix}.$$

By the discussions in the Section 6.3.1, we can easily conclude that $\bar{\omega}_{ij}^{n+1}$ is a monotonically increasing function with respect to all ω_{ij}^n involved in (6.24) under the CFL condition

$$\frac{\Delta t}{\Delta x}\max_{ij}|u_{ij}^n| + \frac{\Delta t}{\Delta y}\max_{ij}|v_{ij}^n| \leq \frac{1}{3}.$$

However, to obtain $\bar{\omega}_{ij}^{n+1} \in [m, M]$, the monotonicity is sufficient only if the following consistency condition holds:

$$\omega_{ij}^n \equiv m \Rightarrow \bar{\omega}_{ij}^{n+1} = m, \quad \omega_{ij}^n \equiv M \Rightarrow \bar{\omega}_{ij}^{n+1} = M. \tag{6.25}$$

Plugging $\omega_{ij}^n \equiv m$ in (6.24), we get

$$\begin{aligned}\omega_{ij}^{n+1} =& m + \frac{\Delta t}{6\Delta x}\left(\frac{u_{i+1,j-1}^n - u_{i-1,j-1}^n}{2} + \frac{4(u_{i+1,j}^n - u_{i-1,j}^n)}{2} + \frac{u_{i+1,j+1}^n - u_{i-1,j+1}^n}{2}\right)\\ &+ \frac{\Delta t}{6\Delta y}\left(\frac{v_{i-1,j+1}^n - v_{i-1,j-1}^n}{2} + \frac{4(v_{i,j+1}^n - v_{i,j-1}^n)}{2} + \frac{v_{i+1,j+1}^n - v_{i+1,j-1}^n}{2}\right).\end{aligned}$$

Thus the consistency (6.25) holds only if the velocity $\langle u^n, v^n\rangle$ satisfies:

$$\begin{aligned}&\frac{1}{6\Delta x}\left(\frac{u_{i+1,j-1}^n - u_{i-1,j-1}^n}{2} + \frac{4(u_{i+1,j}^n - u_{i-1,j}^n)}{2} + \frac{u_{i+1,j+1}^n - u_{i-1,j+1}^n}{2}\right)\\ +&\frac{1}{6\Delta y}\left(\frac{v_{i-1,j+1}^n - v_{i-1,j-1}^n}{2} + \frac{4(v_{i,j+1}^n - v_{i,j-1}^n)}{2} + \frac{v_{i+1,j+1}^n - v_{i+1,j-1}^n}{2}\right) = 0.\end{aligned} \tag{6.26}$$

Thus we get the following bound-preserving result:

Theorem 6.4.1. *If the velocity $\langle u^n, v^n\rangle$ satisfies the constraint (6.26) and $\omega_{ij}^n \in [m, M]$, then under the CFL constraint*

$$\frac{\Delta t}{\Delta x}\max_{ij}|u_{ij}^n| + \frac{\Delta t}{\Delta y}\max_{ij}|v_{ij}^n| \leq \frac{1}{3},$$

the scheme (6.24) satisfies $\bar{\omega}_{ij}^{n+1} \in [m, M]$.

6.4.2 A discrete divergence free velocity field

Note that (6.26) is a discrete divergence free constraint and we can reconstruct a fourth order accurate velocity field satisfying (6.26) by direct difference.

In the following discussion, we may discard the superscript n sometimes for convenience since everything discussed is at time step n. Given ω_{ij}, we first compute ψ_{ij} by a fourth order compact finite difference scheme (i.e., the nine-point discrete Laplacian) for the Poisson equation (6.19b). Then by the fourth order compact finite difference we have

$$-D_y \mathbf{\Psi} = W_{1y}\mathbf{u}_{re}, \quad D_x \mathbf{\Psi} = W_{1x}\mathbf{v}_{re}, \tag{6.27}$$

where

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1,N_y} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2,N_y} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{N_x-1,1} & \psi_{N_x-1,2} & \cdots & \psi_{N_x-1,N_y} \\ \psi_{N_x,1} & \psi_{N_x,2} & \cdots & \psi_{N_x,N_y} \end{pmatrix}_{N_x \times N_y} .$$

Write (6.26) in matrix form, then we have

$$D_x W_{1y}\mathbf{u} + D_y W_{1x}\mathbf{v} = 0.$$

We can clearly see that $\mathbf{u}_{re}$ and $\mathbf{v}_{re}$ constructed in (6.27) satisfy (6.26) immediately.

Now we let

$$\mathbf{u} = \mathbf{u}_{re}, \quad \mathbf{v} = \mathbf{v}_{re}.$$

6.4.3 A fourth order accurate bound-preserving scheme

For the Euler equations (6.22), the following implementation of the fourth order compact finite difference with forward Euler time discretization scheme can preserve the bounds:

1. Given $\omega_{ij}^n \in [m, M]$, solve the Poisson equation by the fourth order accurate nine-point discrete Laplacian scheme to obtain point values of the stream function ψ_{ij}.

2. Reconstruct $\mathbf{u}$ and $\mathbf{v}$ by (6.27).

3. Obtain $\bar{\omega}_{ij}^{n+1} \in [m, M]$ by scheme (6.24).

4. Apply the limiting procedure in Section 6.3.1 to obtain $\omega_{ij}^{n+1} \in [m, M]$.

For high order SSP time discretizations, we should use the same implementation above for each time stage or time step.

For the Navier-Stokes equations (6.19a), the scheme can be written as

$$\begin{aligned}\omega_{ij}^{n+1} =& \omega_{ij}^{n} - \lambda_1 [W_{1x}^{-1} D_x(\mathbf{u}^n \circ \boldsymbol{\omega}^n)]_{ij} - \lambda_2 [W_{1y}^{-1} D_y(\mathbf{v}^n \circ \boldsymbol{\omega}^n)]_{ij} \\ &+ \frac{\mu_1}{Re} W_{2x}^{-1} D_{xx} \omega_{ij}^n + \frac{\mu_2}{Re} W_{2y}^{-1} D_{yy} \omega_{ij}^n,\end{aligned}$$

which is equivalent to

$$\begin{aligned}\tilde{\bar{\omega}}_{ij}^{n+1} = \tilde{\bar{\omega}}_{ij}^{n} - W_{2x} W_{2y} \Bigg[& \frac{1}{36} \begin{pmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{pmatrix} : \Omega^n - \frac{\lambda_1}{12} \begin{pmatrix} -1 & 0 & 1 \\ -4 & 0 & 4 \\ -1 & 0 & 1 \end{pmatrix} : (U^n \circ \Omega^n) \\ & - \frac{\lambda_2}{12} \begin{pmatrix} 1 & 4 & 1 \\ 0 & 0 & 0 \\ -1 & -4 & -1 \end{pmatrix} : (V^n \circ \Omega^n) \Bigg]_{ij} \\ & + \frac{\mu_1}{Re} W_{1x} W_{1y} W_{2y} D_{xx} \omega_{ij}^n + \frac{\mu_2}{Re} W_{1x} W_{1y} W_{2x} D_{yy} \omega_{ij}^n.\end{aligned} \tag{6.28}$$

Following the discussions in Section 6.4.1 and Section 6.3.2, we obtain the following result:

Theorem 6.4.2. *If the velocity $\langle u^n, v^n \rangle$ satisfies the constraint (6.26) and $\omega_{ij}^n \in [m, M]$, then under the CFL constraint*

$$\frac{\Delta t}{\Delta x} \max_{ij} |u_{ij}^n| + \frac{\Delta t}{\Delta y} \max_{ij} |v_{ij}^n| \leq \frac{1}{6}, \quad \frac{\Delta t}{Re \Delta x^2} + \frac{\Delta t}{Re \Delta y^2} \leq \frac{5}{24},$$

the scheme (6.28) satisfies $\tilde{\bar{\omega}}_{ij}^{n+1} \in [m, M]$.

We can use the same implementation in this subsection to first compute $\tilde{\bar{\omega}}_{ij}^{n+1} \in [m, M]$ by (6.28) then apply the limiting procedure in Section 6.3.2 to recover point values $\omega_{ij}^{n+1} \in [m, M]$.

6.5 Higher Order Extensions

The weak monotonicity may not hold for a generic compact finite difference operator. See [71] for a general discussion of compact finite difference schemes. In this section we demonstrate how to construct a higher order accurate compact finite difference scheme satisfying the weak monotonicity. Following Section 6.2 and Section 6.3, we can use these compact finite difference operators to construct higher order accurate bound-preserving schemes.

6.5.1 Higher order compact finite difference operators

Consider a compact finite difference approximation to the first order derivative in the following form:

$$\beta_1 f_{i-2} + \alpha_1 f_{i-1} + f_i + \alpha_1 f_{i+1} + \beta_1 f_{i+2} = b_1 \frac{f_{i+2} - f_{i-2}}{4\Delta x} + a_1 \frac{f_{i+1} - f_{i-1}}{2\Delta x}, \tag{6.29}$$

where $\alpha_1, \beta_1, a_1, b_1$ are constants to be determined. To obtain a sixth order accurate approximation, there are many choices for $\alpha_1, \beta_1, a_1, b_1$. To ensure the approximation in (6.29) satisfies the weak monotonicity for solving scalar conservation laws under some CFL condition, we need $\alpha_1 > 0, \beta_1 > 0$. By requirements above, we obtain

$$\beta_1 = \frac{1}{12}(-1 + 3\alpha_1), \quad a_1 = \frac{2}{9}(8 - 3\alpha_1), \quad b_1 = \frac{1}{18}(-17 + 57\alpha_1), \quad \alpha_1 > \frac{1}{3}. \tag{6.30}$$

With (6.30), the approximation (6.29) is sixth order accurate and satisfies the weak monotonicity as discussed in Section 6.2.1. The truncation error of the approximation (6.29) and (6.30) is $\frac{4}{7!}(9\alpha_1 - 4)\Delta x^6 f^{(7)} + \mathcal{O}(\Delta x^8)$, so if setting

$$\alpha_1 = \frac{4}{9}, \quad \beta_1 = \frac{1}{36}, \quad a_1 = \frac{40}{27}, \quad b_1 = \frac{25}{54}, \tag{6.31}$$

we have an eighth order accurate approximation satisfying the weak monotonicity.

Now consider the fourth order compact finite difference approximations to the second derivative in the following form:

$$\beta_2 f_{i-2} + \alpha_2 f_{i-1} + f_i + \alpha_2 f_{i+1} + \beta_2 f_{i+2} = b_2 \frac{f_{i+2} - 2f_i + f_{i-2}}{4\Delta x^2} + a_2 \frac{f_{i+1} - 2f_i + f_{i-1}}{\Delta x^2},$$

$$a_2 = \frac{1}{3}(4 - 4\alpha_2 - 40\beta_2), \quad b_2 = \frac{1}{3}(-1 + 10\alpha_2 + 46\beta_2).$$

with the truncation error $\frac{-4}{6!}(-2+11\alpha_2-124\beta_2)\Delta x^4 f^{(6)}$. The fourth order scheme discussed in Section 6.2 is the special case with $\alpha_2 = \frac{1}{10}, \quad \beta_2 = 0, \quad a_2 = \frac{6}{5}, \quad b_2 = 0$. If $\beta_2 = \frac{11\alpha_2 - 2}{124}$, we get a family of sixth-order schemes satisfying the weak monotonicity:

$$a_2 = \frac{-78\alpha_2 + 48}{31}, \quad b_2 = \frac{291\alpha_2 - 36}{62}, \quad \alpha_2 > 0. \tag{6.32}$$

The truncation error of the sixth order approximation is $\frac{4}{31\cdot 8!}(1179\alpha_2 - 344)\Delta x^6 f^{(8)}$. Thus we obtain an eighth order approximation satisfying the weak monotonicity if

$$\alpha_2 = \frac{344}{1179}, \beta_2 = \frac{23}{2358}, a_2 = \frac{320}{393}, b_2 = \frac{310}{393}, \tag{6.33}$$

with truncation error $\frac{-172}{5676885}\Delta x^8 f^{(10)}$.

6.5.2 Convection problems

For the rest of this section, we will mostly focus on the family of sixth order schemes since the eighth order accurate scheme is a special case of this family. For $u_t + f(u)_x = 0$ with periodic boundary conditions on the interval $[0, 1]$, we get the following semi-discrete scheme:

$$\frac{d}{dt}\mathbf{u} = -\frac{1}{\Delta x}\widetilde{W}_1^{-1}\widetilde{D}_x \mathbf{f},$$

$$\widetilde{W}_1\mathbf{u} = \frac{\beta_1}{1+2\alpha_1+2\beta_1}\begin{pmatrix} \frac{1}{\beta_1} & \frac{\alpha_1}{\beta_1} & 1 & & & 1 & \frac{\alpha_1}{\beta_1} \\ \frac{\alpha_1}{\beta_1} & \frac{1}{\beta_1} & \frac{\alpha_1}{\beta_1} & 1 & & & 1 \\ 1 & \frac{\alpha_1}{\beta_1} & \frac{1}{\beta_1} & \frac{\alpha_1}{\beta_1} & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & \frac{\alpha_1}{\beta_1} & \frac{1}{\beta_1} & \frac{\alpha_1}{\beta_1} & 1 \\ 1 & & & 1 & \frac{\alpha_1}{\beta_1} & \frac{1}{\beta_1} & \frac{\alpha_1}{\beta_1} \\ \frac{\alpha_1}{\beta_1} & 1 & & & 1 & \frac{\alpha_1}{\beta_1} & \frac{1}{\beta_1} \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-2} \\ u_{N-1} \\ u_N \end{pmatrix},$$

$$\widetilde{D}_x\mathbf{f} = \frac{1}{4(1+2\alpha_1+2\beta_1)}\begin{pmatrix} 0 & 2a_1 & b_1 & & & -b_1 & -2a_1 \\ -2a_1 & 0 & 2a_1 & b_1 & & & -b_1 \\ -b_1 & -2a_1 & 0 & 2a_1 & b_1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -b_1 & -2a_1 & 0 & 2a_1 & b_1 \\ b_1 & & & -b_1 & -2a_1 & 0 & 2a_1 \\ 2a_1 & b_1 & & & -b_1 & -2a_1 & 0 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{N-2} \\ f_{N-1} \\ f_N \end{pmatrix},$$

where f_i and u_i are point values of functions $f(u(x))$ and $u(x)$ at uniform grid points x_i $(i = 1, \cdots, N)$ respectively. We have a family of sixth-order compact schemes with forward Euler time discretization:

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \frac{\Delta t}{\Delta x}\widetilde{W}_1^{-1}\widetilde{D}_x\mathbf{f}. \tag{6.34}$$

Define $\bar{\mathbf{u}} = \widetilde{W}_1\mathbf{u}$ and $\lambda = \frac{\Delta t}{\Delta x}$, then scheme (6.34) can be written as

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\lambda}{4(1+2\alpha_1+2\beta_1)}(b_1 f_{i+2}^n + 2a_1 f_{i+1}^n - 2a_1 f_{i-1}^n - b_1 f_{i-2}^n).$$

Following the lines in Section 6.2.1, we can easily conclude that the scheme (6.34) satisfies $\bar{u}_i^{n+1} \in [m, M]$ if $u_i^n \in [m, M]$, under the CFL constraint

$$\frac{\Delta t}{\Delta x}|f(u)| \leq \min\{\frac{9}{8-3\alpha_1}, \frac{6(3\alpha_1-1)}{57\alpha_1-17}\}.$$

Given $\bar{u}_i \in [m, M]$, we also need a limiter to enforce $u_i \in [m, M]$. Notice that $\bar{u}_i$ has a five-point stencil instead of a three-point stencil in Section 6.2.2. Thus in general the extensions of Section 6.2.2 for sixth order schemes are more complicated. However, we can still use the same limiter as in Section 6.2.2 because the five-diagonal matrix $\widetilde{W}_1$ can be represented as a product of two tridiagonal matrices.

Plugging in $\beta_1 = \frac{1}{12}(-1+3\alpha_1)$, we have $\widetilde{W}_1 = \widetilde{W}_1^{(1)}\widetilde{W}_1^{(2)}$, where

$$\widetilde{W}_1^{(1)} = \frac{1}{c_1^{(1)}+2}\begin{pmatrix} c_1^{(1)} & 1 & & & 1 \\ 1 & c_1^{(1)} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & c_1^{(1)} & 1 \\ 1 & & & 1 & c_1^{(1)} \end{pmatrix}, c_1^{(1)} = \frac{6\alpha_1}{3\alpha_1-1} - \frac{\sqrt{2}\sqrt{7-24\alpha_1+27\alpha_1^2}}{\sqrt{1-6\alpha_1+9\alpha_1^2}},$$

$$\widetilde{W}_1^{(2)} = \frac{1}{c_1^{(2)}+2}\begin{pmatrix} c_1^{(2)} & 1 & & & 1 \\ 1 & c_1^{(2)} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & c_1^{(2)} & 1 \\ 1 & & & 1 & c_1^{(2)} \end{pmatrix}, c_1^{(2)} = \frac{6\alpha_1}{3\alpha_1-1} + \frac{\sqrt{2}\sqrt{7-24\alpha_1+27\alpha_1^2}}{\sqrt{1-6\alpha_1+9\alpha_1^2}}.$$

In other words, $\bar{\mathbf{u}} = \widetilde{W}_1\mathbf{u} = \widetilde{W}_1^{(1)}\widetilde{W}_1^{(2)}\mathbf{u}$. Thus following the limiting procedure in Section 6.2.4, we can still use the same limiter in Section 6.2.2 twice to enforce the bounds of point values if $c_1^{(1)}, c_1^{(2)} \geq 2$, which implies $\frac{1}{3} < \alpha_1 \leq \frac{5}{9}$. In this case we have $\min\{\frac{9}{8-3\alpha_1}, \frac{6(3\alpha_1-1)}{57\alpha_1-17}\} = \frac{6(3\alpha_1-1)}{57\alpha_1-17}$, thus the CFL for the weak monotonicity becomes $\lambda|f(u)| \leq \frac{6(3\alpha_1-1)}{57\alpha_1-17}$. We summarize the results in the following theorem.

Theorem 6.5.1. *Consider a family of sixth order accurate schemes (6.34) with*

$$\beta_1 = \frac{1}{12}(-1+3\alpha_1), \quad a_1 = \frac{2}{9}(8-3\alpha_1), \quad b_1 = \frac{1}{18}(-17+57\alpha_1), \quad \frac{1}{3} < \alpha_1 \leq \frac{5}{9},$$

which includes the eighth order scheme (6.31) as a special case. If $u_i^n \in [m, M]$ for all i, under the CFL constraint $\frac{\Delta t}{\Delta x}\max_u |f(u)| \leq \frac{6(3\alpha_1-1)}{57\alpha_1-17}$, we have $\bar{u}_i^{n+1} \in [m, M]$.

Given point values u_i satisfying $\widetilde{W}_1^{(1)}\widetilde{W}_1^{(2)}u_i = \widetilde{W}_1 u_i = \bar{u}_i \in [m, M]$ for any i, we can apply the limiter in Algorithm 2 twice to enforce $u_i \in [m, M]$:

1. Given $\bar{u}_i \in [m, M]$, compute $v_i = [\widetilde{W}_1^{(1)}]^{-1}\bar{u}_i$ which are not necessarily in the range $[m, M]$. Then apply the limiter in Algorithm 2 to $v_i, i = 1, \cdots, N$. Let $\bar{v}_i$ denote the output of the limiter. Since we have $\bar{u}_i = \frac{1}{c_1^{(1)}+2}(v_{i-1} + c_1^{(1)}v_i + v_{i+1}), c_1^{(1)} > 2$, all discussions in Section 6.2.2 are still valid, thus we have $\bar{v}_i \in [m, M]$.

2. Compute $u_i = [\widetilde{W}_1^{(2)}]^{-1}\bar{v}_i$. Apply the limiter in Algorithm 2 to $u_i, i = 1, \cdots, N$. Since we have $\bar{v}_i = \frac{1}{c_1^{(2)}+2}(u_{i-1} + c_1^{(2)}u_i + u_{i+1}), c_1^{(2)} > 2$, all discussions in Section 6.2.2 are still valid, thus the output are in $[m, M]$.

6.5.3 Diffusion problems

For simplicity we only consider the diffusion problems and the extension to convection diffusion problems can be easily discussed following Section 6.2.4. For the one-dimensional scalar diffusion equation $u_t = g(u)_{xx}$ with $g(u) \geq 0$ and periodic boundary conditions on an interval $[0, 1]$, we get the sixth order semi-discrete scheme: $\frac{d}{dt}\mathbf{u} = \frac{1}{\Delta x^2}\widetilde{W}_2^{-1}\widetilde{D}_{xx}\mathbf{g}$, where

$$\widetilde{W}_2\mathbf{u} = \frac{\beta_2}{1+2\alpha_2+2\beta_2}\begin{pmatrix} \frac{1}{\beta_2} & \frac{\alpha_2}{\beta_2} & 1 & & & 1 & \frac{\alpha_2}{\beta_2} \\ \frac{\alpha_2}{\beta_2} & \frac{1}{\beta_2} & \frac{\alpha_2}{\beta_2} & 1 & & & 1 \\ 1 & \frac{\alpha_2}{\beta_2} & \frac{1}{\beta_2} & \frac{\alpha_2}{\beta_2} & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & \frac{\alpha_2}{\beta_2} & \frac{1}{\beta_2} & \frac{\alpha_2}{\beta_2} & 1 \\ 1 & & & 1 & \frac{\alpha_2}{\beta_2} & \frac{1}{\beta_2} & \frac{\alpha_2}{\beta_2} \\ \frac{\alpha_2}{\beta_2} & 1 & & & 1 & \frac{\alpha_2}{\beta_2} & \frac{1}{\beta_2} \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-2} \\ u_{N-1} \\ u_N \end{pmatrix},$$

$$\widetilde{D}_{xx}\mathbf{g} = \tfrac{1}{4(1+2\alpha_2+2\beta_2)} \begin{pmatrix} -8a_2-2b_2 & 4a_2 & 2b_2 & & & 2b_2 & 4a_2 \\ 4a_2 & -8a_2-2b_2 & 4a_2 & 2b_2 & & & 2b_2 \\ 2b_2 & 4a_2 & -8a_2-2b_2 & 4a_2 & 2b_2 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 2b_2 & 4a_2 & -8a_2-2b_2 & 4a_2 & 2b_2 \\ 2b_2 & & & 2b_2 & 4a_2 & -8a_2-2b_2 & 4a_2 \\ 4a_2 & 2b_2 & & & 2b_2 & 4a_2 & -8a_2-2b_2 \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_{N-2} \\ g_{N-1} \\ g_N \end{pmatrix},$$

where g_i and u_i are values of functions $g(u(x))$ and $u(x)$ atx_i respectively.

As in the previous subsection, we prefer to factor $\widetilde{W}_2$ as a product of two tridiagonal matrices. Plugging in $\beta_2 = \frac{11\alpha_2-2}{124}$, we have: $\widetilde{W}_2 = \widetilde{W}_2^{(1)}\widetilde{W}_2^{(2)}$, where

$$\widetilde{W}_2^{(1)} = \frac{1}{c_2^{(1)}+2} \begin{pmatrix} c_2^{(1)} & 1 & & & 1 \\ 1 & c_2^{(1)} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & c_2^{(1)} & 1 \\ 1 & & & 1 & c_2^{(1)} \end{pmatrix}, c_2^{(1)} = \frac{62\alpha_2}{11\alpha_2-2} - \frac{\sqrt{2}\sqrt{128-726\alpha_2+2043\alpha_2^2}}{\sqrt{4-44\alpha_2+121\alpha_2^2}},$$

$$\widetilde{W}_2^{(2)} = \frac{1}{c_2^{(2)}+2} \begin{pmatrix} c_2^{(2)} & 1 & & & 1 \\ 1 & c_2^{(2)} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & c_2^{(2)} & 1 \\ 1 & & & 1 & c_2^{(2)} \end{pmatrix}, c_2^{(2)} = \frac{62\alpha_2}{11\alpha_2-2} + \frac{\sqrt{2}\sqrt{128-726\alpha_2+2043\alpha_2^2}}{\sqrt{4-44\alpha_2+121\alpha_2^2}}.$$

To have $c_2^{(1)}, c_2^{(2)} \geq 2$, we need $\frac{2}{11} < \alpha_2 \leq \frac{60}{113}$. The forward Euler gives

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \frac{\Delta t}{\Delta x^2}\widetilde{W}_2^{-1}\widetilde{D}_{xx}\mathbf{g}. \tag{6.35}$$

Define $\tilde{u}_i = \widetilde{W}_2 u_i$ and $\mu = \frac{\Delta t}{\Delta x^2}$, then the scheme (6.35) can be written as

$$\tilde{u}_i^{n+1} = \tilde{u}_i^n + \frac{\mu}{4(1+2\alpha_2+2\beta_2)}\left[2b_2 g_{i-2}^n + 4a_2 g_{i-1}^n + (-8a_2-2b_2)g_i^n + 4a_2 g_{i+1}^n + 2b_2 g_{i+2}^n\right].$$

Theorem 6.5.2. *Consider a family of sixth order accurate schemes* (6.35) *with*

$$\beta_2 = \frac{11\alpha_2 - 2}{124}, a_2 = \frac{-78\alpha_2 + 48}{31}, \quad b_2 = \frac{291\alpha_2 - 36}{62}, \quad \frac{2}{11} < \alpha_2 \leq \frac{60}{113},$$

which includes the eighth order scheme (6.33) *as a special case. If $u_i^n \in [m, M]$ for all i, under the CFL $\frac{\Delta t}{\Delta x^2} g(u) < \frac{124}{3(116-111\alpha_2)}$, the scheme satisfies $\tilde{u}^{n+1} \in [m, M]$.*

As in the previous subsection, given point values u_i satisfying $\widetilde{W}_2^{(1)}\widetilde{W}_2^{(2)}u_i = \widetilde{W}_2 u_i = \tilde{u}_i \in [m, M]$ for any i, we can apply the limiter in Algorithm 2 twice to enforce $u_i \in [m, M]$. The matrices $\widetilde{W}_1$ and $\widetilde{W}_2$ commute because they are both circulant matrices thus diagonalizable by the discrete Fourier matrix. The discussion for the sixth order scheme solving convection diffusion problems is also straightforward.

6.6 Extensions To General Boundary Conditions

Since the compact finite difference operator is implicitly defined thus any extension to other type boundary conditions is not straightforward. In order to maintain the weak monotonicity, the boundary conditions must be properly treated. In this section we demonstrate a high order accurate boundary treatment preserving the weak monotonicity for inflow and outflow boundary conditions. For convection problems, we can easily construct a fourth order accurate boundary scheme. For convection diffusion problems, it is much more complicated to achieve weak monotonicity near the boundary thus a straightforward discussion gives us a third order accurate boundary scheme.

6.6.1 Inflow-outflow boundary conditions for convection problems

For simplicity, we consider the following initial boundary value problem on the interval $[0, 1]$ as an example: $u_t + f(u)_x = 0, \quad u(x, 0) = u_0(x), \quad u(0, t) = L(t)$, where we assume $f(u) > 0$ so that the inflow boundary condition at the left cell end is a well-posed boundary condition. The boundary condition at $x = 1$ is not specified thus understood as an outflow boundary condition. We further assume $u_0(x) \in [m, M]$ and $L(t) \in [m, M]$ so that the exact solution is in $[m, M]$.

Consider a uniform grid with $x_i = i\Delta x$ for $i = 0, 1, \cdots, N, N+1$ and $\Delta x = \frac{1}{N+1}$. Then a fourth order semi-discrete compact finite difference scheme is given by

$$\frac{d}{dt}\frac{1}{6}\begin{pmatrix} 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \end{pmatrix}\begin{pmatrix} u_0 \\ \vdots \\ u_{N+1} \end{pmatrix} = \frac{1}{2\Delta x}\begin{pmatrix} -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} f_0 \\ \vdots \\ f_{N+1} \end{pmatrix}.$$

With forward Euler time discretization, the scheme is equivalent to

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{1}{2}\lambda(f_{i+1}^n - f_{i-1}^n), \quad i = 1, \cdots, N. \tag{6.36}$$

Here $u_0^n = L(t^n)$ is given as boundary condition for any n. Given u_i^n for $i = 0, 1, \cdots, N+1$, the scheme (6.36) gives $\bar{u}_i^{n+1}$ for $i = 1, \cdots, N$, from which we still need u_{N+1}^{n+1} to recover interior point values u_i^{n+1} for $i = 1, \cdots, N$.

Since the boundary condition at $x_{N+1} = 1$ can be implemented as outflow, we can use $\bar{u}_i^{n+1}$ for $i = 1, \cdots, N$ to obtain a reconstructed u_{N+1}^{n+1}. If there is a cubic polynomial $p_i(x)$ so that u_{i-1}, u_i, u_{i+1} are its point values at x_{i-1}, x_i, x_{i+1}, then $\frac{1}{2\Delta x}\int_{x_{i-1}}^{x_{i+1}} p_i(x)\,dx = \frac{1}{6}u_{i-1} + \frac{4}{6}u_i + \frac{1}{6}u_{i+1} = \bar{u}_i$, due to the exactness of the Simpson's quadrature rule for cubic polynomials. To this end, we can consider a unique cubic polynomial $p(x)$ satisfying four equations: $\frac{1}{2\Delta x}\int_{x_{j-1}}^{x_{j+1}} p(x)\,dx = \bar{u}_j^{n+1}, \quad j = N-3, N-2, N-1, N$. If $\bar{u}_j^{n+1}$ are fourth order accurate approximations to $\frac{1}{6}u(x_{j-1}, t^{n+1}) + \frac{4}{6}u(x_j, t^{n+1}) + \frac{1}{6}u(x_{j+1}, t^{n+1})$, then $p(x)$ is a fourth order accurate approximation to $u(x, t^{n+1})$ on the interval $[x_{N-4}, x_{N+1}]$. So we get a fourth order accurate u_{N+1}^{n+1} by

$$p(x_{N+1}) = -\frac{2}{3}\bar{u}_{N-3} + \frac{17}{6}\bar{u}_{N-2} - \frac{14}{3}\bar{u}_{N-1} + \frac{7}{2}\bar{u}_N. \tag{6.37}$$

Since (6.37) is not a convex linear combination, $p(x_{N+1})$ may not lie in the bound $[m, M]$. Thus to ensure $u_{N+1}^{n+1} \in [m, M]$ we can define

$$u_{N+1}^{n+1} := \max\{\min\{p(x_{N+1}), M\}, m\}. \tag{6.38}$$

Obviously Theorem 6.2.1 still holds for the scheme (6.36). For the forward Euler time discretization, we can implement the bound-preserving scheme as follows:

1. Given u_i^n for all i, compute $\bar{u}_i^{n+1}$ for $i = 1, \cdots, N$ by (6.36).
2. Obtain boundary values $u_0^{n+1} = L(t^{n+1})$ and u_{N+1}^{n+1} by (6.37) and (6.38).
3. Given $\bar{u}_i^{n+1}$ for $i = 1, \cdots, N$ and two boundary values u_0^{n+1} and u_{N+1}^{n+1}, recover point values u_i^{n+1} for $i = 1, \cdots, N$ by solving the tridiagonal linear system (the superscript $n+1$ is omitted):
$$\frac{1}{6}\begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} \bar{u}_1 - \frac{1}{6}u_0 \\ \bar{u}_2 \\ \vdots \\ \bar{u}_{N-1} \\ \bar{u}_N - \frac{1}{6}u_{N+1} \end{pmatrix}.$$
4. Apply the limiter in Algorithm 2 to the point values u_i^{n+1} for $i = 1, \cdots, N$.

6.6.2 Dirichlet boundary conditions for one-dimensional convection diffusion equations

Consider the initial boundary value problem for a one-dimensional scalar convection diffusion equation on the interval $[0, 1]$:

$$u_t + f(u)_x = g(u)_{xx}, \quad u(x,t) = u_0(x), \quad u(0,t) = L(t), \quad u(1,t) = R(t), \tag{6.39}$$

where $g(u) \geq 0$. We further assume $u_0(x) \in [m, M]$ and $L(t), R(t) \in [m, M]$ so that the exact solution is in $[m, M]$.

We demonstrate how to treat the boundary approximations so that the scheme still satisfies some weak monotonicity such that a certain convex combination of point values is in the range $[m, M]$ at the next time step. Consider a uniform grid with $x_i = i\Delta x$

for $i = 0, 1, \cdots, N, N+1$ where $\Delta x = \frac{1}{N+1}$. The fourth order compact finite difference approximations at the interior points can be written as:

$$W_1 \begin{pmatrix} f_{x,1} \\ f_{x,2} \\ \vdots \\ f_{x,N-1} \\ f_{x,N} \end{pmatrix} = \frac{1}{\Delta x} D_x \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \end{pmatrix} + \begin{pmatrix} -\frac{f_{x,0}}{6} - \frac{f_0}{2\Delta x} \\ 0 \\ \vdots \\ 0 \\ -\frac{f_{x,N+1}}{6} + \frac{f_{N+1}}{2\Delta x} \end{pmatrix},$$

$$W_1 = \frac{1}{6} \begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix}, \quad D_x = \frac{1}{2} \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{pmatrix},$$

$$W_2 \begin{pmatrix} g_{xx,1} \\ g_{xx,2} \\ \vdots \\ g_{xx,N-1} \\ g_{xx,N} \end{pmatrix} = \frac{1}{\Delta x^2} D_{xx} \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N \end{pmatrix} + \begin{pmatrix} -\frac{g_{xx,0}}{12} + \frac{g_0}{\Delta x^2} \\ 0 \\ \vdots \\ 0 \\ -\frac{g_{xx,N+1}}{12} + \frac{g_{N+1}}{\Delta x^2} \end{pmatrix},$$

$$W_2 = \frac{1}{12} \begin{pmatrix} 10 & 1 & & & \\ 1 & 10 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 10 & 1 \\ & & & 1 & 10 \end{pmatrix}, \quad D_{xx} = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix},$$

where $f_{x,i}$ and $g_{xx,i}$ denotes the values of $f(u)_x$ and $g(u)_{xx}$ at x_i respectively. Let

$$F = \begin{pmatrix} -\frac{f_{x,0}}{6} - \frac{f_0}{2\Delta x} \\ 0 \\ \vdots \\ 0 \\ -\frac{f_{x,N+1}}{6} + \frac{f_{N+1}}{2\Delta x} \end{pmatrix}, \quad G = \begin{pmatrix} -\frac{g_{xx,0}}{12} + \frac{g_0}{\Delta x^2} \\ 0 \\ \vdots \\ 0 \\ -\frac{g_{xx,N+1}}{12} + \frac{g_{N+1}}{\Delta x^2} \end{pmatrix}.$$

Define $W := W_1 W_2 = W_2 W_1$. Here W_2 and W_1 commute because they have the same eigenvectors, which is due to the fact that $2W_2 - W_1$ is the identity matrix. Let $\mathbf{u} = \Big(u_1\ u_2\ \cdots\ u_N\Big)^T$, $\mathbf{f} = \Big(f(u_1)\ f(u_2)\ \cdots\ f(u_N)\Big)^T$ and $\mathbf{g} = \Big(g(u_1)\ g(u_2)\ \cdots\ g(u_N)\Big)^T$. Then a fourth order compact finite difference approximation to (6.39) at the interior grid points is $\frac{d}{dt}\mathbf{u} + W_1^{-1}(\frac{1}{\Delta x} D_x \mathbf{f} + F) = W_2^{-1}(\frac{1}{\Delta x^2} D_{xx}\mathbf{g} + G)$ which is equivalent to

$$\frac{d}{dt}(W\mathbf{u}) + \frac{1}{\Delta x} W_2 D_x \mathbf{f} - \frac{1}{\Delta x^2} W_1 D_{xx} \mathbf{g} = -W_2 F + W_1 G.$$

If $u_i(t) = u(x_i, t)$ where $u(x,t)$ is the exact solution to the problem, then it satisfies

$$u_{t,i} + f_{x,i} = g_{xx,i}, \tag{6.40}$$

where $u_{t,i} = \frac{d}{dt} u_i(t)$, $f_{x,i} = f(u_i)_x$ and $g_{xx,i} = g(u_i)_{xx}$. If we use (6.40) to simplify $-W_2 F + W_1 G$, then the scheme is still fourth order accurate. In other words, setting $-f_{x,i} + g_{xx,i} = u_{t,i}$

does not affect the accuracy. Plugging (6.40) in the original $-W_2F+W_1G$, we can redefine $-W_2F+W_1G$ as

$$-W_2F+W_1G := \begin{pmatrix} -\frac{1}{18}u_{t,0}+\frac{1}{12}f_{x,0}+\frac{5}{12\Delta x}f_0+\frac{2}{3\Delta x^2}g_0 \\ -\frac{1}{72}u_{t,0}+\frac{1}{24}f_0+\frac{1}{6\Delta x^2}g_0 \\ 0 \\ \vdots \\ 0 \\ -\frac{1}{72}u_{t,N+1}-\frac{1}{24}f_{N+1}+\frac{1}{6\Delta x^2}g_{N+1} \\ -\frac{1}{18}u_{t,N+1}+\frac{1}{12}f_{x,N+1}-\frac{5}{12\Delta x}f_{N+1}+\frac{2}{3\Delta x^2}g_{N+1} \end{pmatrix}.$$

So we now consider the following fourth order accurate scheme:

$$\frac{d}{dt}(W\mathbf{u})+\frac{1}{\Delta x}W_2D_x\mathbf{f}-\frac{1}{\Delta x^2}W_1D_{xx}\mathbf{g} = \begin{pmatrix} -\frac{1}{18}u_{t,0}+\frac{1}{12}f_{x,0}+\frac{5}{12\Delta x}f_0+\frac{2}{3\Delta x^2}g_0 \\ -\frac{1}{72}u_{t,0}+\frac{1}{24}f_0+\frac{1}{6\Delta x^2}g_0 \\ 0 \\ \vdots \\ 0 \\ -\frac{1}{72}u_{t,N+1}-\frac{1}{24}f_{N+1}+\frac{1}{6\Delta x^2}g_{N+1} \\ -\frac{1}{18}u_{t,N+1}+\frac{1}{12}f_{x,N+1}-\frac{5}{12\Delta x}f_{N+1}+\frac{2}{3\Delta x^2}g_{N+1} \end{pmatrix}. \tag{6.41}$$

The first equation in (6.41) is

$$\frac{d}{dt}(\frac{4u_0+41u_1+14u_2+u_3}{72}) = \frac{1}{24\Delta x}(10f_0+f_1-10f_2-f_3)+\frac{1}{6\Delta x^2}(4g_0-7g_1+2g_2+g_3)+\frac{1}{12}f_{x,0}.$$

After multiplying $\frac{72}{60}=\frac{6}{5}$ to both sides, it becomes

$$\begin{aligned}\frac{d}{dt}(\frac{4u_0+41u_1+14u_2+u_3}{60}) &= \frac{1}{20\Delta x}(10f_0+f_1-10f_2-f_3) \\ &\quad +\frac{1}{5\Delta x^2}(4g_0-7g_1+2g_2+g_3)+\frac{1}{10}f_{x,0}.\end{aligned} \tag{6.42}$$

In order for the scheme (6.42) to satisfy a weak monotonicity in the sense that $\frac{4u_0^{n+1}+41u_1^{n+1}+14u_2^{n+1}+u_3^{n+1}}{60}$ in (6.42) with forward Euler can be written as a monotonically increasing function of u_i^n under some CFL constraint, we still need to find an approximation to $f(u)_{x,0}$ using only u_0, u_1, u_2, u_3, with which we have a straightforward third order approximation to $f(u)_{x,0}$:

$$f_{x,0} = \frac{1}{\Delta x}(-\frac{11}{6}f_0 + 3f_1 - \frac{3}{2}f_2 + \frac{1}{3}f_3) + \mathcal{O}(\Delta x^3). \tag{6.43}$$

Then (6.42) becomes

$$\begin{aligned} \frac{d}{dt}(\frac{4u_0 + 41u_1 + 14u_2 + u_3}{60}) &= \frac{1}{60\Delta x}(19f_0 + 21f_1 - 39f_2 - f_3) \\ &\quad + \frac{1}{5\Delta x^2}(4g_0 - 7g_1 + 2g_2 + g_3). \end{aligned} \tag{6.44}$$

The second to second last equations of (6.41) can be written as

$$\begin{aligned} \frac{d}{dt}(\frac{u_{i-2} + 14u_{i-1} + 42u_i + 14u_{i+1} + u_{i+2}}{72}) &= \frac{1}{24\Delta x}(f_{i-2} + 10f_{i-1} \\ -10f_{i+1} - f_{i+2}) + \frac{1}{6\Delta x^2}(g_{i-2} + 2g_{i-1} - 6g_i + 2g_{i+1} + g_{i+2}), &\quad 2 \le i \le N-1, \end{aligned} \tag{6.45}$$

which satisfies a straightforward weak monotonicity under some CFL constraint.

The last equation in (6.41) is

$$\begin{aligned} \frac{d}{dt}(\frac{4u_{N+1} + 41u_N + 14u_{N-1} + u_{N-2}}{72}) &= \frac{1}{24\Delta x}(f_{N-2} + 10f_{N-1} - f_N \\ -10f_{N+1}) + \frac{1}{6\Delta x^2}&(g_{N-2} + 2g_{N-1} - 7g_N + 4g_{N+1}) + \frac{1}{12}f_{x,N+1}. \end{aligned}$$

After multiplying $\frac{72}{60} = \frac{6}{5}$ to both sides, it becomes

$$\begin{aligned} \frac{d}{dt}(\frac{u_{N-2} + 14u_{N-1} + 41u_N + 4u_{N+1}}{60}) &= \frac{1}{20\Delta x}(f_{N-2} + 10f_{N-1} - f_N \\ -10f_{N+1}) + \frac{1}{5\Delta x^2}&(g_{N-2} + 2g_{N-1} - 7g_N + 4g_{N+1}) + \frac{1}{10}f_{x,N+1}. \end{aligned}$$

Similar to the boundary scheme at x_0, we should use a third-order approximation:

$$f_{x,N+1} = \frac{1}{\Delta x}(-\frac{1}{3}f_{N-2} + \frac{3}{2}f_{N-1} - 3f_N + \frac{11}{6}f_{N+1}) + \mathcal{O}(\Delta x^3). \tag{6.46}$$

Then the boundary scheme at x_{N+1} becomes

$$\begin{aligned}\frac{d}{dt}(\frac{u_{N-2} + 14u_{N-1} + 41u_N + 4u_{N+1}}{60}) = \frac{1}{60\Delta x}(f_{N-2} + 39f_{N-1} - 21f_N \\ -19f_{N+1}) + \frac{1}{5\Delta x^2}(g_{N-2} + 2g_{N-1} - 7g_N + 4g_{N+1}).\end{aligned} \tag{6.47}$$

To summarize the full semi-discrete scheme, we can represent the third order scheme (6.44), (6.45) and (6.47), for the Dirichlet boundary conditions as:

$$\frac{d}{dt}\widetilde{W}\tilde{\mathbf{u}} = -\frac{1}{\Delta x}\widetilde{D}_x f(\tilde{\mathbf{u}}) + \frac{1}{\Delta x^2}\widetilde{D}_{xx} g(\tilde{\mathbf{u}}),$$

where

$$\widetilde{W} = \frac{1}{72}\begin{pmatrix} \frac{24}{5} & \frac{246}{5} & \frac{84}{5} & \frac{6}{5} & & & & & \\ 1 & 14 & 42 & 14 & 1 & & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & 1 & 14 & 42 & 14 & 1 & \\ & & & & & \frac{6}{5} & \frac{84}{5} & \frac{246}{5} & \frac{24}{5} \end{pmatrix}_{N\times(N+2)}, \tilde{\mathbf{u}} = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_N \\ u_{N+1} \end{pmatrix}_{(N+2)\times 1},$$

$$\widetilde{D}_x = \frac{1}{24}\begin{pmatrix} -\frac{38}{5} & -\frac{42}{5} & \frac{78}{5} & \frac{2}{5} & & & & & \\ -1 & -10 & 0 & 10 & 1 & & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & -1 & -10 & 0 & 10 & 1 & \\ & & & & & -\frac{2}{5} & -\frac{78}{5} & \frac{42}{5} & \frac{38}{5} \end{pmatrix}_{N\times(N+2)}, \widetilde{D}_{xx} = \frac{1}{6}\begin{pmatrix} \frac{24}{5} & -\frac{42}{5} & \frac{12}{5} & \frac{6}{5} & & & & & \\ 1 & 2 & -6 & 2 & 1 & & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & 1 & 2 & -6 & 2 & 1 & \\ & & & & & \frac{6}{5} & \frac{12}{5} & -\frac{42}{5} & \frac{24}{5} \end{pmatrix}_{N\times(N+2)}.$$

Let $\bar{\mathbf{u}} = \widetilde{W}\tilde{\mathbf{u}}$, $\lambda = \frac{\Delta t}{\Delta x}$ and $\mu = \frac{\Delta t}{\Delta x^2}$. With forward Euler, it becomes

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{1}{2}\lambda \widetilde{D}_x \tilde{f}_i + \mu \widetilde{D}_{xx} \tilde{g}_i, \quad i = 1, \cdots, N. \tag{6.48}$$

We state the weak monotonicity without proof.

Theorem 6.6.1. *Under the CFL constraint* $\frac{\Delta t}{\Delta x}\max_u |f(u)| \leq \frac{4}{19}$, $\frac{\Delta t}{\Delta x^2}\max_u g(u) \leq \frac{695}{1596}$, *if* $u_i^n \in [m, M]$, *then the scheme* (6.48) *satisfies* $\bar{u}_i^{n+1} \in [m, M]$.

We notice that

$$\begin{aligned}
\bar{u}_1^{n+1} = & \quad \tfrac{1}{60}(4u_0^{n+1} + 41u_1^{n+1} + 14u_2^{n+1} + u_3^{n+1}) \\
= & \quad \tfrac{u_0^{n+1}+4u_1^{n+1}+u_2^{n+1}}{6} + \tfrac{1}{10}\tfrac{u_1^{n+1}+4u_2^{n+1}+u_3^{n+1}}{6} - \tfrac{1}{10}u_0^{n+1}, \\
\bar{u}_N^{n+1} = & \quad \tfrac{1}{60}(u_{N-2}^{n+1} + 14u_{N-1}^{n+1} + 41u_N^{n+1} + 4u_{N+1}^{n+1}) \\
= & \quad \tfrac{1}{10}\tfrac{u_{N-2}^{n+1}+4u_{N-1}^{n+1}+u_N^{n+1}}{6} + \tfrac{u_{N-1}^{n+1}+4u_N^{n+1}+u_{N+1}^{n+1}}{6} - \tfrac{1}{10}u_{N+1}^{n+1}.
\end{aligned}$$

Recall that the boundary values are given: $u_0^{n+1} = L(t^{n+1}) \in [m, M]$ and $u_{N+1}^{n+1} = R(t^{n+1}) \in [m, M]$, so we have

$$\begin{aligned}
\frac{10}{11}\frac{u_0^{n+1} + 4u_1^{n+1} + u_2^{n+1}}{6} + \frac{1}{11}\frac{u_1^{n+1} + 4u_2^{n+1} + u_3^{n+1}}{6} &\leq \frac{10}{11}M + \frac{1}{11}M = M, \\
\frac{10}{11}\frac{u_0^{n+1} + 4u_1^{n+1} + u_2^{n+1}}{6} + \frac{1}{11}\frac{u_1^{n+1} + 4u_2^{n+1} + u_3^{n+1}}{6} &\geq \frac{10}{11}m + \frac{1}{11}m = m, \\
\frac{1}{11}\frac{u_{N-2}^{n+1} + 4u_{N-1}^{n+1} + u_N^{n+1}}{6} + \frac{10}{11}\frac{u_{N-1}^{n+1} + 4u_N^{n+1} + u_{N+1}^{n+1}}{6} &\leq \frac{10}{11}M + \frac{1}{11}M = M, \\
\frac{1}{11}\frac{u_{N-2}^{n+1} + 4u_{N-1}^{n+1} + u_N^{n+1}}{6} + \frac{10}{11}\frac{u_{N-1}^{n+1} + 4u_N^{n+1} + u_{N+1}^{n+1}}{6} &\geq \frac{10}{11}m + \frac{1}{11}m = m.
\end{aligned}$$

Thus define $\mathbf{w}^{n+1} = \left(w_1^{n+1}, w_2^{n+1}, w_3^{n+1}, \ldots, w_{N-1}^{n+1}, w_N^{n+1}\right)^T$ as follows and we have:

$$\begin{aligned}
& m \leq w_i^{n+1} := \bar{u}_i^{n+1} \leq M, \quad i = 2, \cdots, N-1, \\
& m \leq w_1^{n+1} := \frac{10}{11}\frac{u_0^{n+1} + 4u_1^{n+1} + u_2^{n+1}}{6} + \frac{1}{11}\frac{u_1^{n+1} + 4u_2^{n+1} + u_3^{n+1}}{6} \leq M, \\
& m \leq w_N^{n+1} := \frac{1}{11}\frac{u_{N-3}^{n+1} + 4u_{N-2}^{n+1} + u_{N-1}^{n+1}}{6} + \frac{10}{11}\frac{u_{N-2}^{n+1} + 4u_{N-1}^{n+1} + u_N^{n+1}}{6} \leq M.
\end{aligned}$$

By the notations above, we get

$$\mathbf{w}^{n+1} = K\bar{\mathbf{u}}^{n+1} + \mathbf{u}_{bc}^{n+1} = \widetilde{\widetilde{\widetilde{W}}}\tilde{\mathbf{u}}, \tag{6.49}$$

$$K = \begin{pmatrix} \frac{10}{11} & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & \frac{10}{11} \end{pmatrix}_{N\times N}, \mathbf{u}_{bc} = \frac{1}{11}\begin{pmatrix} u_0 \\ 0 \\ \vdots \\ 0 \\ u_{N+1} \end{pmatrix}_{N\times 1}, \widetilde{\widetilde{\widetilde{W}}} = \frac{1}{72}\begin{pmatrix} \frac{120}{11} & \frac{492}{11} & \frac{168}{11} & \frac{12}{11} & & & & \\ 1 & 14 & 42 & 14 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & 14 & 42 & 14 & 1 \\ & & & & \frac{12}{11} & \frac{168}{11} & \frac{492}{11} & \frac{120}{11} \end{pmatrix}_{N\times(N+2)}.$$

We notice that $\widetilde{\widetilde{\widetilde{W}}}$ can be factored as a product of two tridiagonal matrices:

$$\frac{1}{72}\begin{pmatrix} \frac{120}{11} & \frac{492}{11} & \frac{168}{11} & \frac{12}{11} & & & & \\ 1 & 14 & 42 & 14 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & 14 & 42 & 14 & 1 \\ & & & & \frac{12}{11} & \frac{168}{11} & \frac{492}{11} & \frac{120}{11} \end{pmatrix} = \frac{1}{12}\begin{pmatrix} \frac{120}{11} & \frac{12}{11} & & & \\ 1 & 10 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 10 & 1 \\ & & & \frac{12}{11} & \frac{120}{11} \end{pmatrix}_{N\times N} \frac{1}{6}\begin{pmatrix} 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 & \\ & & & & 1 & 4 & 1 \end{pmatrix}_{N\times(N+2)},$$

which can be denoted as $\widetilde{\widetilde{\widetilde{W}}} = \widetilde{W}_2\widetilde{W}_1$. Fortunately, all the diagonal entries of $\widetilde{W}_1$ and $\widetilde{W}_2$ are in the form of $\frac{c}{c+2}, c > 2$. So given $\bar{u}_i = \widetilde{W}u_i \in [m, M]$, we construct $w_i^{n+1} \in [m, M]$. We can apply the limiter in Algorithm 2 twice to enforce $u_i \in [m, M]$:

1. Given u_i^n for all i, use the scheme (6.48) to obtain $\bar{u}_i^{n+1} \in [m, M]$ for $i = 1, \cdots, N$. Then construct $w_i^{n+1} \in [m, M]$ for $i = 1, \cdots, N$ by (6.49).

2. Notice that $\widetilde{W}_2$ is a matrix of size $N \times N$. Compute $\mathbf{v} = \widetilde{W}_2^{-1}\mathbf{w}^{n+1}$. Apply the limiter in Algorithm 2 to v_i and let $\bar{v}_i$ denote the output values. Since we have $\widetilde{W}_2 v_i \in [m, M]$, i.e.,

$$\begin{aligned}
m \leq {}& \tfrac{10}{11}v_1 + \tfrac{1}{11}v_2 && \leq M,\\
m \leq {}& \tfrac{1}{12}v_1 + \tfrac{10}{12}v_2 + \tfrac{1}{12}v_3 && \leq M,\\
& \vdots \\
m \leq {}& \tfrac{1}{12}v_{N-2} + \tfrac{10}{12}v_{N-1} + \tfrac{1}{12}v_N && \leq M,\\
m \leq {}& \tfrac{1}{11}v_{N-1} + \tfrac{10}{11}v_N && \leq M.
\end{aligned}$$

Following the discussions in Section 6.2.2, it implies $\bar{v}_i \in [m, M]$.

3. Obtain values of u_i^{n+1}, $i = 1, \cdots, N$ by solving a $N \times N$ system:

$$\frac{1}{6}\begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix}\begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_{N-1}^{n+1} \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \vdots \\ \bar{v}_{N-1} \\ \bar{v}_N \end{pmatrix} - \frac{1}{6}\mathbf{u}_{bc}^{n+1}.$$

4. Apply the limiter in Algorithm 2 to u_i^{n+1} to ensure $u_i^{n+1} \in [m, M]$.

6.7 Numerical Tests

6.7.1 One-dimensional problems with periodic boundary conditions

In this subsection, we test the fourth order and eighth order accurate compact finite difference schemes with the bound-preserving limiter. The time step is taken to satisfy both the CFL condition required for weak monotonicity in Theorem 6.2.1 and Theorem 6.2.10 and the SSP coefficient for high order SSP time discretizations.

Example 1. One-dimensional linear convection equation. Consider $u_t + u_x = 0$ with and initial condition $u_0(x) = \frac{1}{2} + \sin^4(x)$ and periodic boundary conditions on the interval $[0, 2\pi]$

with a uniform N-point grid. The L^1 and L^∞ errors for the fourth order scheme with a smooth initial condition at time $T = 10$ are listed in Table 6.1 where $\Delta x = \frac{2\pi}{N}$, the time step is taken as $\Delta t = C_{ms}\frac{1}{3}\Delta x$ for the multistep method, and $\Delta t = 5C_{ms}\frac{1}{3}\Delta x$ for the Runge-Kutta method so that the number of spatial discretization operators computed is the same as in the one for the multistep method. We can observe the fourth order accuracy for the multistep method and obvious order reductions for the Runge-Kutta method.

The errors for smooth initial conditions $u_0(x) = \frac{1}{2} + \frac{1}{2}\sin^4(x)$ at time $T = 10$ for the eighth order accurate scheme are listed in Table 6.2. For the eighth order accurate scheme, the time step to achieve the weak monotonicity is $\Delta t = C_{ms}\frac{6}{25}\Delta x$ for the fourth-order SSP multistep method. On the other hand, we need to set $\Delta t = \Delta x^2$ in fourth order accurate time discretizations to verify the eighth order spatial accuracy. To this end, the time step is taken as $\Delta t = C_{ms}\frac{6}{25}\Delta x^2$ for the multistep method, and $\Delta t = 5C_{ms}\frac{6}{25}\Delta x^2$ for the Runge-Kutta method. We can observe the eighth order accuracy for the multistep method and the order reduction for $N = 160$ is due to the roundoff errors. We can also see an obvious order reduction for the Runge-Kutta method.

Table 6.1. Fourth order scheme accuracy test.

	Fourth order SSP multistep				Fourth order SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	3.44E-2	-	6.49E-2	-	3.41E-2	-	6.26E-2	-
40	3.12E-3	3.47	6.19E-3	3.39	3.14E-3	3.44	6.62E-3	3.24
80	1.82E-4	4.10	2.95E-4	4.39	1.86E-4	4.08	3.82E-4	4.11
160	1.10E-5	4.05	1.85E-5	4.00	1.29E-5	3.85	4.48E-5	3.09
320	6.81E-7	4.02	1.15E-6	4.01	1.42E-6	3.18	1.03E-5	2.13

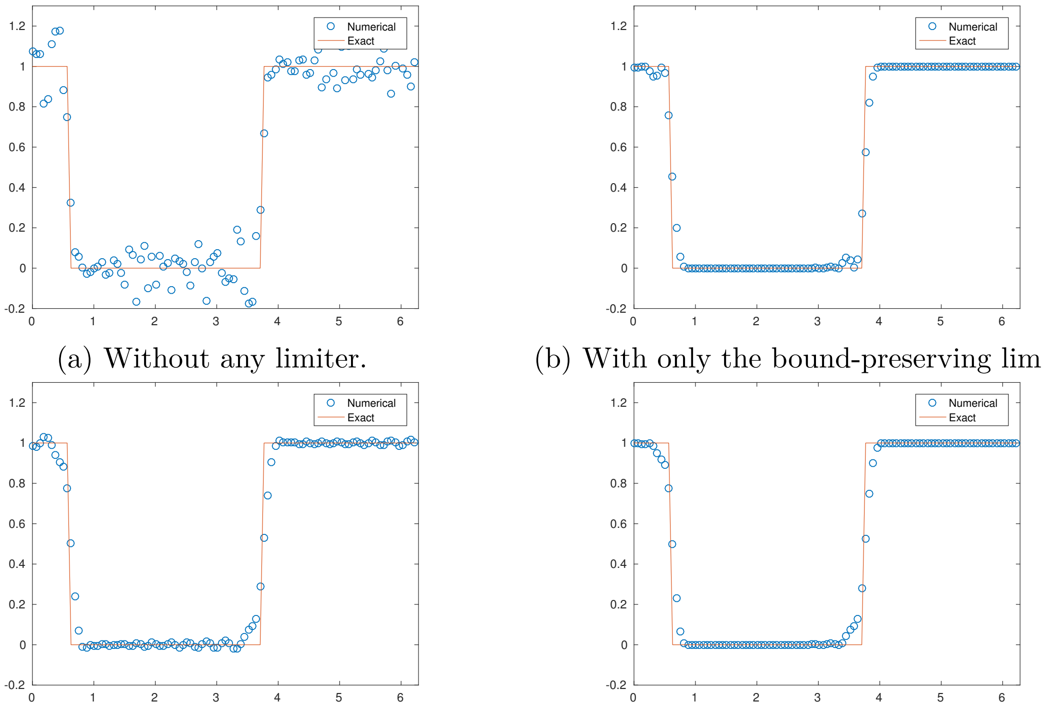
Table 6.2. Eighth order scheme accuracy test.

	Fourth order SSP multistep				Fourth order SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10	6.31E-2	-	1.01E-1	-	6.44E-2	-	9.58E-2	-
20	3.35E-5	7.55	5.59E-4	7.49	3.39E-4	7.57	5.79E-4	7.37
40	9.58E-7	8.45	1.49E-6	8.55	1.52E-6	7.80	4.32E-6	7.06
80	3.50E-9	8.10	5.51E-9	8.08	5.34E-8	4.83	2.31E-7	4.23
160	6.57E-11	5.74	1.01E-10	5.77	2.40E-9	4.48	1.45E-8	3.99

Next, we consider the following discontinuous initial data:

$$u_0(x) = \begin{cases} 1, & \text{if} \quad 0 < x \leq \pi, \\ 0, & \text{if} \quad \pi < x \leq 2\pi. \end{cases} \tag{6.50}$$

See Figure 6.1 for the performance of the bound-preserving limiter and the TVB limiter on the fourth order scheme at $T = 10$. Fourth order compact finite difference and fourth order SSP multistep with $\Delta t = \frac{1}{3} C_{ms} \Delta x$ and 100 grid points are used. The TVB parameter in (6.11) is $p = 5$. We observe that the TVB limiter can reduce oscillations but cannot remove the overshoot/undershoot. When both limiters are used, we can obtain a non-oscillatory bound-preserving numerical solution. See Figure 6.2 for the performance of the bound-preserving limiter on the eighth order scheme at $T = 10$ with $\Delta t = C_{ms} \frac{6}{25} \Delta x$ and 100 grid points.



(a) Without any limiter.

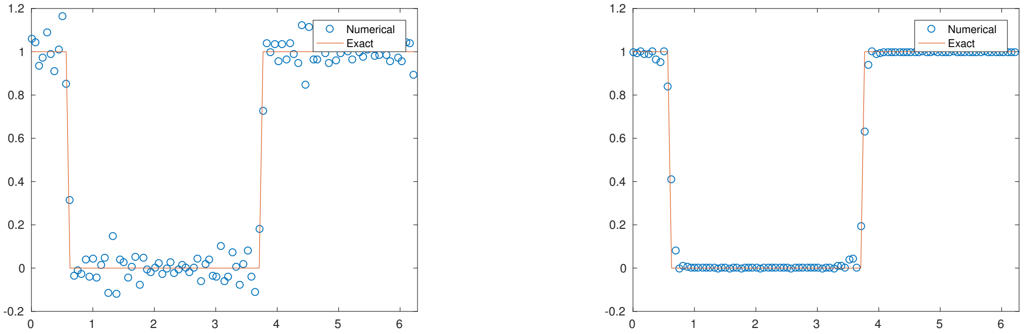
(b) With only the bound-preserving limiter.

(c) With only the TVB limiter.

(d) With both limiters.

Figure 6.1. Fourth order scheme for linear convection with discontinuous initial data.

Example 2. One dimensional Burgers' equation. Consider the Burgers' equation $u_t + (\frac{u^2}{2})_x = 0$ with a periodic boundary condition on $[-\pi, \pi]$. For the initial data $u_0(x) = \sin(x) + 0.5$,



(a) Without any limiter.

(b) With the bound-preserving limiter.

Figure 6.2. Eighth order scheme for linear convection with discontinuous initial data.

the exact solution is smooth up to $T = 1$, then it develops a moving shock. We list the errors of the fourth order scheme at $T = 0.5$ in Table 6.3 where the time step is $\Delta t = \frac{1}{3}C_{ms}\Delta x$ for SSP multistep and $\Delta t = \frac{5}{3}C_{ms}\Delta x$ for SSP Runge-Kutta with $\Delta x = \frac{2\pi}{N}$. We observe the expected fourth order accuracy for the multistep time discretization. At $T = 1.2$, the exact solution contains a shock near $x = -2.5$. The errors on the smooth region $[-2, \pi]$ at $T = 1.2$ are listed in Table 6.4 where high order accuracy is lost. Some high order schemes can still be high order accurate on a smooth region away from the shock in this test, see [91]. We emphasize that in all our numerical tests, Step III in Algorithm 2 was never triggered. In other words, set of Class I is rarely encountered in practice. So the limiter Algorithm 2 is a local three-point stencil limiter for this particular example rather than a global one. The loss of accuracy in smooth regions is possibly due to the fact that compact finite difference operator is defined globally thus the error near discontinuities will pollute the whole domain.

The solutions of the fourth order compact finite difference and the fourth order SSP multistep with the bound-preserving limiter and the TVB limiter at time $T = 2$ are shown in Figure 6.3, for which the exact solution is in the range $[-0.5, 1.5]$. With 100 grid points and time step $\Delta t = \frac{1}{3\max_x |u_0(x)|}C_{ms}\Delta x$, the TVB parameter in (6.11) is set as $p = 5$. The TVB limiter alone does not eliminate the overshoot or undershoot. When both the bound-preserving and the TVB limiters are used, we can obtain a non-oscillatory bound-preserving numerical solution.

Table 6.3. The fourth order scheme with limiter for the Burgers' equation. Smooth solutions.

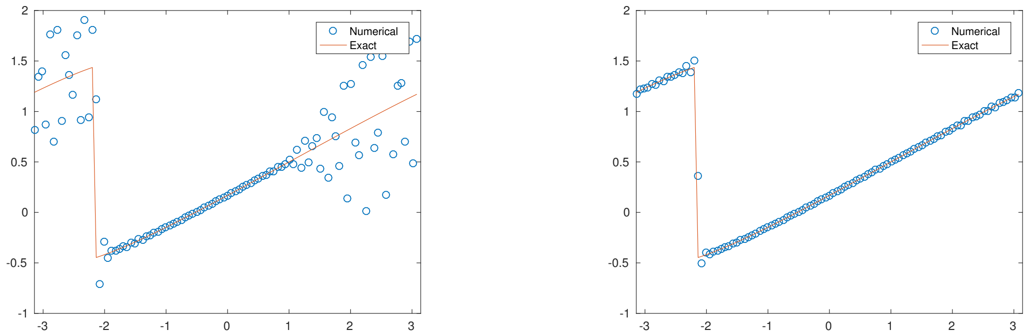
	Fourth order SSP multistep				Fourth SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	6.92E-4	-	5.24E-3	-	7.79E-4	-	5.61E-3	-
40	3.28E-5	4.40	3.62E-4	3.85	4.45E-5	4.13	4.77E-4	3.56
80	1.90E-6	4.11	2.00E-5	4.18	3.53E-6	3.66	2.09E-5	4.51
160	1.15E-6	4.04	1.24E-6	4.01	4.93E-7	2.84	5.47E-6	1.93
320	7.18E-9	4.00	7.67E-8	4.01	8.78E-8	2.49	1.73E-6	1.66

Table 6.4. Burgers' equation. The errors are measured in the smooth region away from the shock.

	Fourth order SSP multistep				Fourth SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	1.59E-2	-	5.26E-2	-	1.62E-2	-	5.39E-2	-
40	2.10E-3	2.92	1.38E-2	1.93	2.11E-3	2.94	1.39E-2	1.95
80	6.35E-4	1.73	6.56E-3	1.07	6.48E-4	1.70	7.01E-3	0.99
160	1.48E-4	2.10	1.65E-3	1.99	1.51E-4	2.10	1.66E-3	2.08
320	3.12E-5	2.25	6.10E-4	1.43	3.14E-5	2.26	6.13E-4	1.44

Example 3. One dimensional convection diffusion equation.

Consider the linear convection diffusion equation $u_t + cu_x = du_{xx}$ with a periodic boundary condition on $[0, 2\pi]$. For the initial $u_0(x) = \sin(x)$, the exact solution is $u(x,t) = exp(-dt)sin(x - ct)$ which is in the range $[-1, 1]$. We set $c = 1$ and $d = 0.001$. The errors of the fourth order scheme at $T = 1$ are listed in the Table 6.5 in which $\Delta t = C_{ms}min\{\frac{1}{6}\frac{\Delta x}{c}, \frac{5}{24}\frac{\Delta x^2}{d}\}$ for SSP multistep and $\Delta t = 5C_{ms}min\{\frac{1}{6}\frac{\Delta x}{c}, \frac{5}{24}\frac{\Delta x^2}{d}\}$ for SSP Runge-Kutta with $\Delta x = \frac{2\pi}{N}$. We observe the expected fourth order accuracy for the SSP multistep method. Even though the bound-preserving limiter is triggered, the order reduction for the Runge-Kutta method is not observed for the convection diffusion equation. One possible explanation is that the source of such an order reduction is due to the lower order accuracy of inner stages in the Runge-Kutta method, which is proportional to the time step. Compared to $\Delta t = \mathcal{O}(\Delta x)$ for a pure convection, the time step is $\Delta t = \mathcal{O}(\Delta x^2)$ in a convection diffusion problem thus the order reduction is much less prominent. See the Table 6.6 for



(a) Without any limiter.

(b) With both limiters.

Figure 6.3. Burgers' equation at $T = 2$.

the errors at $T = 1$ of the eighth order scheme with $\Delta t = C_{ms} \min\{\frac{3}{25}\frac{\Delta x^2}{c}, \frac{131}{530}\frac{\Delta x^2}{d}\}$ for SSP multistep and $\Delta t = 5C_{ms} \min\{\frac{3}{25}\frac{\Delta x^2}{c}, \frac{131}{530}\frac{\Delta x^2}{d}\}$ for SSP Runge-Kutta where $\Delta x = \frac{2\pi}{N}$.

Table 6.5. The fourth order compact finite difference with limiter for linear convection diffusion.

	Fourth order SSP multistep				Fourth order SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
20	3.30E-5	-	5.19E-5	-	3.60E-5	-	6.09E-5	-
40	2.11E-6	3.97	3.30E-6	3.97	2.44E-6	4.00	3.52E-6	4.12
80	1.33E-7	3.99	2.09E-7	3.98	1.37E-7	4.04	2.15E-7	4.03
160	8.36E-9	3.99	1.31E-8	3.99	8.46E-9	4.02	1.33E-8	4.02
320	5.24E-10	4.00	8.23E-10	4.00	5.29E-10	4.00	8.31E-10	4.00

Example 4. Nonlinear degenerate diffusion equations.

A representative test for validating the positivity-preserving property of a scheme solving nonlinear diffusion equations is the porous medium equation, $u_t = (u^m)_{xx}, m > 1$. We consider the Barenblatt analytical solution given by

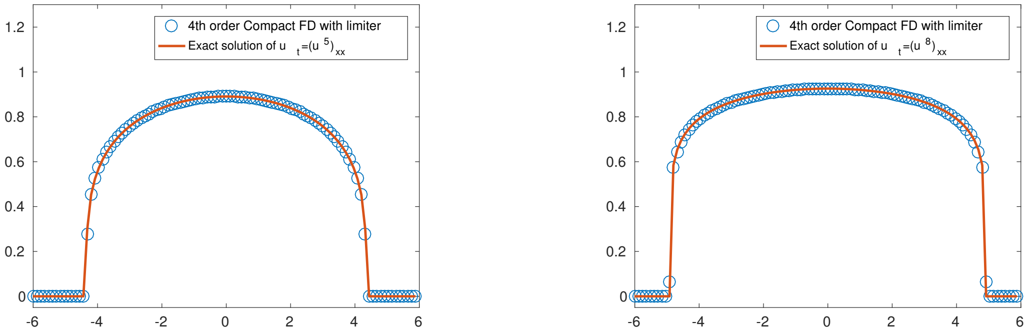
$$B_m(x,t) = t^{-k}[(1 - \frac{k(m-1)}{2m}\frac{|x|^2}{t^{2k}})_+]^{1/(m-1)},$$

where $u_+ = \max\{u, 0\}$ and $k = (m+1)^{-1}$. The initial data is the Barenblatt solution at $T = 1$ with periodic boundary conditions on $[6, 6]$. The solution is computed till time $T = 2$. High order schemes without any particular positivity treatment will generate negative solutions [102], [103], [106]. See Figure 6.4 for solutions of the fourth order scheme and the

Table 6.6. The eighth order compact finite difference with limiter for linear convection diffusion.

	SSP multistep				SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10	3.85E-7	-	5.96E-7	-	3.85E-7	-	5.95E-7	-
20	1.40E-9	8.10	2.20E-9	8.08	1.42E-9	8.08	2.23E-9	8.06
40	5.46E-12	8.01	8.60E-12	8.00	5.48E-12	8.02	8.69E-12	8.01
80	3.53E-12	0.63	6.46E-12	0.41	1.06E-12	2.37	3.29E-12	1.40

SSP multistep method with $\Delta t = \frac{1}{3m} C_{ms} \Delta x$ and 100 grid points. Numerical solutions are strictly nonnegative. Without the bound-preserving limiter, negative values emerge near the sharp gradients.



(a) $m = 5$.

(b) $m = 8$.

Figure 6.4. The fourth order compact finite difference with limiter for the porous medium equation.

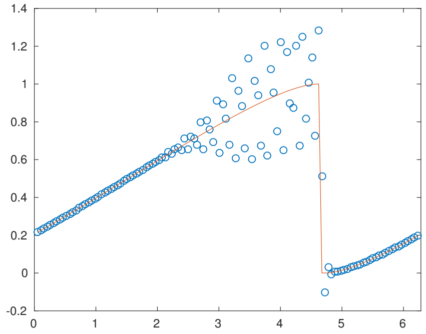
6.7.2 One-dimensional problems with non-periodic boundary conditions

Example 5. One-dimensional Burgers' equation with inflow-outflow boundary condition. Consider $u_t + (\frac{u^2}{2})_x = 0$ on interval $[0, 2\pi]$ with inflow-outflow boundary condition and smooth initial condition $u(x,0) = u_0(x)$. Let $u_0(x) = \frac{1}{2}\sin(x) + \frac{1}{2} \geq 0$, we can set the left boundary condition as inflow $u(0,t) = L(t)$ and right boundary as outflow, where $L(t)$ is obtained from the exact solution of initial-boundary value problem for the same initial data and a periodic boundary condition. We test the fourth order compact finite difference and fourth order SSP multistep method with the bound-preserving limiter. The errors at $T = 0.5$

are listed in Table 6.7 where $\Delta t = C_{ms}\Delta x$ and $\Delta x = \frac{2\pi}{N}$. See Figure 6.5 for the shock at $T = 3$ on a 120-point grid with $\Delta t = C_{ms}\Delta x$.

Table 6.7. Burgers' equation. The fourth order scheme. Inflow and outflow boundary conditions.

N	L^∞ error	order	L^1 error	order
20	1.15E-4	-	7.80E-4	-
40	4.10E-6	4.81	2.00E-5	5.29
80	2.17E-7	4.24	9.43E-7	4.40
160	1.22E-8	4.15	4.87E-8	4.28
320	7.41E-10	4.05	2.87E-9	4.09



(a) Without any limiter.

(b) With the bound-preserving limiter.

Figure 6.5. Burgers' equation. The fourth order scheme. Inflow and outflow boundary conditions.

Example 6. One-dimensional convection diffusion equation with Dirichlet boundary conditions. We consider equation $u_t + cu_x = du_{xx}$ on $[0, 2\pi]$ with boundary conditions $u(0,t) = \cos(-ct)e^{-dt}$ and $u(2\pi, t) = \cos(2\pi - ct)e^{-dt}$. The exact solution is $u(x, y, t) = \cos(x - ct)e^{-dt}$. We set $c = 1$ and $d = 0.01$. We test the third order boundary scheme proposed in Section 6.6.2 and the fourth order interior compact finite difference with the fourth order SSP multistep time discretization. The errors at $T = 1$ are listed in Table 6.8 where $\Delta t = C_{ms} \min\{\frac{4}{19}\frac{\Delta x}{c}, \frac{695}{1596}\frac{\Delta x^2}{d}\}$, $\Delta x = \frac{2\pi}{N}$.

Table 6.8. A linear convection diffusion equation with Dirichlet boundary conditions.

N	L^∞ error	order	L^1 error	order
10	1.68E-3	-	8.76E-3	-
20	1.47E-4	3.51	7.12E-4	3.62
40	8.35E-6	4.14	4.27E-5	4.06
80	4.44E-7	4.23	2.28E-6	4.23
160	2.30E-8	4.27	1.10E-7	4.37

6.7.3 Two-dimensional problems with periodic boundary conditions

In this subsection we test the fourth order compact finite difference scheme solving two-dimensional problems with periodic boundary conditions.

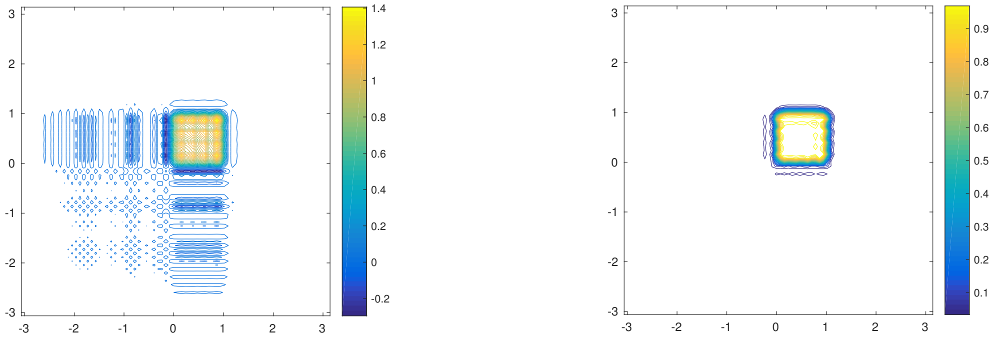
Example 7. Two-dimensional linear convection equation. Consider $u_t + u_x + u_y = 0$ on the domain $[0, 2\pi] \times [0, 2\pi]$ with a periodic boundary condition. The scheme is tested with a smooth initial condition $u_0(x, y) = \frac{1}{2} + \frac{1}{2}\sin^4(x + y)$ to verify the accuracy. The errors at time $T = 1$ are listed in Table 6.9 where $\Delta t = C_{ms}\frac{1}{6}\Delta x$ for the SSP multistep method and $\Delta t = 5C_{ms}\frac{1}{6}\Delta x$ for the SSP Runge-Kutta method with $\Delta x = \Delta y = \frac{2\pi}{N}$. We can observe the fourth order accuracy for the multistep method on resolved meshes and obvious order reductions for the Runge-Kutta method.

Table 6.9. Fourth order accurate compact finite difference with limiter for the 2D linear equation.

	Fourth order SSP multistep				Fourth order SSP Runge-Kutta			
$N \times N$ Mesh	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	4.70E-2	-	1.17E-1	-	8.45E-2	-	1.07E-1	-
20×20	5.47E-3	3.10	8.97E-3	3.71	5.56E-3	3.93	9.09E-3	3.56
40×40	3.04E-4	4.17	5.09E-4	4.13	2.88E-4	4.27	6.13E-4	3.89
80×80	1.78E-5	4.09	2.99E-5	4.09	1.95E-5	3.89	6.77E-5	3.18
160×160	1.09E-6	4.03	1.85E-6	4.01	2.65E-6	2.88	1.26E-5	2.43

We also test the following discontinuous initial data:

$$u_0(x, y) = \begin{cases} 1, \text{ if } (x, y) \in [-0.2, 0.2] \times [-0.2, 0.2], \\ 0, \text{ otherwise.} \end{cases}$$



(a) Without any limiter.

(b) With the bound-preserving limiter.

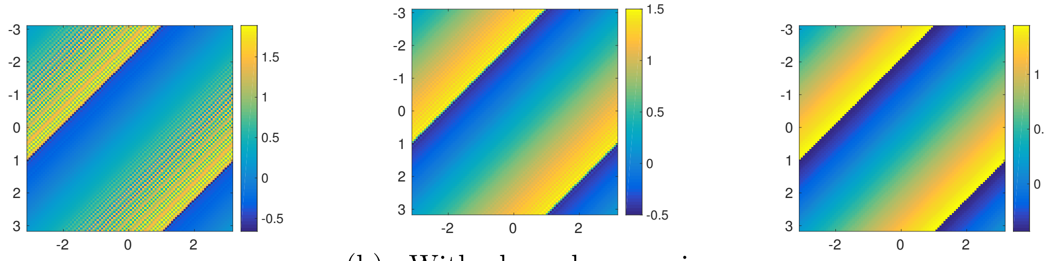
Figure 6.6. Fourth order compact finite difference for the 2D linear convection.

The numerical solutions on a 80×80 mesh at $T = 0.5$ are shown in Figure 6.6 with $\Delta t = \frac{1}{6} C_{ms} \Delta x$ and $\Delta x = \Delta y = \frac{2\pi}{N}$. Fourth order SSP multistep method is used.

Example 8. Two-dimensional Burgers' equation. Consider $u_t + (\frac{u^2}{2})_x + (\frac{u^2}{2})_y = 0$ with $u_0(x, y) = 0.5 + \sin(x + y)$ and periodic boundary conditions on $[-\pi, \pi] \times [-\pi, \pi]$. At time $T = 0.2$, the solution is smooth and the errors at $T = 0.2$ on a $N \times N$ mesh are shown in the Table 6.10 in which $\Delta t = C_{ms} \frac{\Delta x}{6 \max_x |u_0(x)|}$ for multistep and $\Delta t = 5C_{ms} \frac{\Delta x}{6 \max_x |u_0(x)|}$ for Runge-Kutta with $\Delta x = \Delta y = \frac{2\pi}{N}$. At time $T = 1$, the exact solution contains a shock. The numerical solutions of the fourth order SSP multistep method on a 100×100 mesh are shown in Figure 6.7 where $\Delta t = \frac{1}{6 \max_x |u_0(x)|} C_{ms} \Delta x$. The bound-preserving limiter ensures the solution to be in the range $[-0.5, 1.5]$.

Table 6.10. Fourth order compact finite difference scheme with the bound-preserving limiter for the 2D Burgers' equation.

	SSP multistep				SSP Runge-Kutta			
$N \times N$ Mesh	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	1.08E-2	-	4.48E-3	-	9.16E-3	-	3.73E-2	-
20×20	4.73E-4	4.52	3.76E-3	3.58	2.90E-4	4.98	2.14E-3	4.12
40×40	1.90E-5	4.64	1.45E-4	4.69	2.03E-5	3.83	1.12E-4	4.25
80×80	9.99E-7	4.25	7.43E-6	4.29	2.35E-6	3.12	1.54E-5	2.86
160×160	5.87E-8	4.09	4.26E-7	4.13	3.62E-7	2.70	5.13E-6	1.59



(a) Without any limiter.

(b) With bound-preserving limiter.

(c) The exact solution.

Figure 6.7. The fourth order scheme for 2D Burgers' equation.

Example 9. Two-dimensional convection diffusion equation.

Consider the equation $u_t + c(u_x + u_y) = d(u_{xx} + u_{yy})$ with $u_0(x, y) = \sin(x + y)$ and a periodic boundary condition on $[0, 2\pi] \times [0, 2\pi]$. The errors at time $T = 0.5$ for $c = 1$ and

$d = 0.001$ are listed in Table 6.11, in which $\Delta t = C_{ms}\min\{\frac{\Delta x}{6c}, \frac{5\Delta x^2}{48d}\}$ for the fourth-order SSP multistep method, and $\Delta t = 5C_{ms}\min\{\frac{\Delta x}{6c}, \frac{5\Delta x^2}{48d}\}$ for the fourth-order SSP Runge-Kutta method, where $\Delta x = \Delta y = \frac{2\pi}{N}$.

Table 6.11. Fourth order compact finite difference with limiter for the 2D convection diffusion equation.

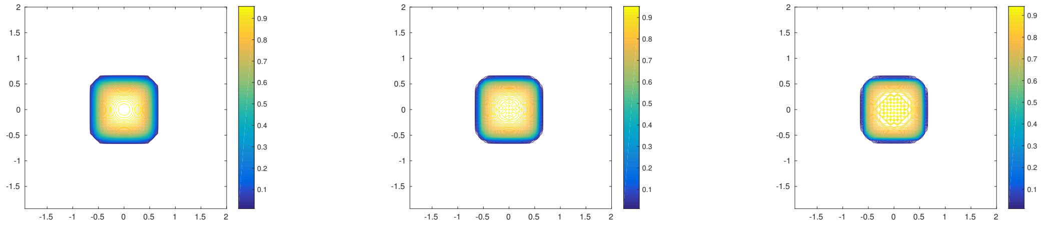
	Fourth order SSP multistep				Fourth order SSP Runge-Kutta			
N	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	6.26E-4	-	9.67E-4	-	6.68E-4	-	9.59E-4	-
20×20	3.62E-5	4.11	5.61E-5	4.11	3.60E-5	4.21	6.09E-5	3.98
40×40	2.20E-6	4.04	3.45E-6	4.02	2.24E-6	4.00	3.52E-6	4.12
80×80	1.35E-7	4.02	2.13E-7	4.01	1.37E-7	4.04	2.15E-7	4.03
160×160	8.45E-9	4.01	1.33E-8	4.01	8.46E-9	4.02	1.33E-8	4.02

Example 10. Two-dimensional porous medium equation.

We consider the equation $u_t = \Delta(u^m)$ with the following initial data

$$u_0(x,y) = \begin{cases} 1, \text{ if } (x,y) \in [-0.5, 0.5] \times [-0.5, 0.5], \\ 0, \text{ if } (x,y) \in [-2,2] \times [-2,2]/[-1,1] \times [-1,1], \end{cases}$$

and a periodic boundary condition on domain $[-2,2] \times [-2,2]$. See Figure 6.8 for the solutions at time $T = 0.01$ for SSP multistep method with $\Delta t = \frac{5}{48\max_x |u_0(x)|} C_{ms}\Delta x$ and $\Delta x = \Delta y = \frac{1}{15}$. The numerical solutions are strictly non-negative, which is nontrivial for high order accurate schemes. High order schemes without any positivity treatment will generate negative solutions in this test, see [102], [103], [106].



(a) $m = 3$.

(b) $m = 4$.

(c) $m = 5$.

Figure 6.8. The fourth order scheme with limiter for 2D porous medium equations $u_t = \Delta(u^m)$.

6.7.4 Two-dimensional incompressible Navier-Stokes equation

In this subsection, we test the bound-preserving fourth order compact finite difference scheme for the two-dimensional incompressible flow.

Example 11. Consider solving (6.19) on the domain $[0,2\pi]\times[0,2\pi]$ with a periodic boundary condition. We use a smooth solution $\omega(x,y,t)=\sin(2x)\sin(2y)\exp(\frac{-8}{Re}t)$ to test the accuracy of the proposed scheme. The errors for $Re=1000$ at $T=1$ are listed in Table 6.12. The time step is taken as $\Delta t=C_{ms}\min\{\frac{1}{12\max_x|u_0|}\Delta x, \frac{5Re}{48}\Delta x^2\}$ for the SSP multistep method and $\Delta t=5C_{ms}\min\{\frac{1}{12\max_x|u_0|}\Delta x, \frac{5Re}{48}\Delta x^2\}$ for the SSP Runge-Kutta method.

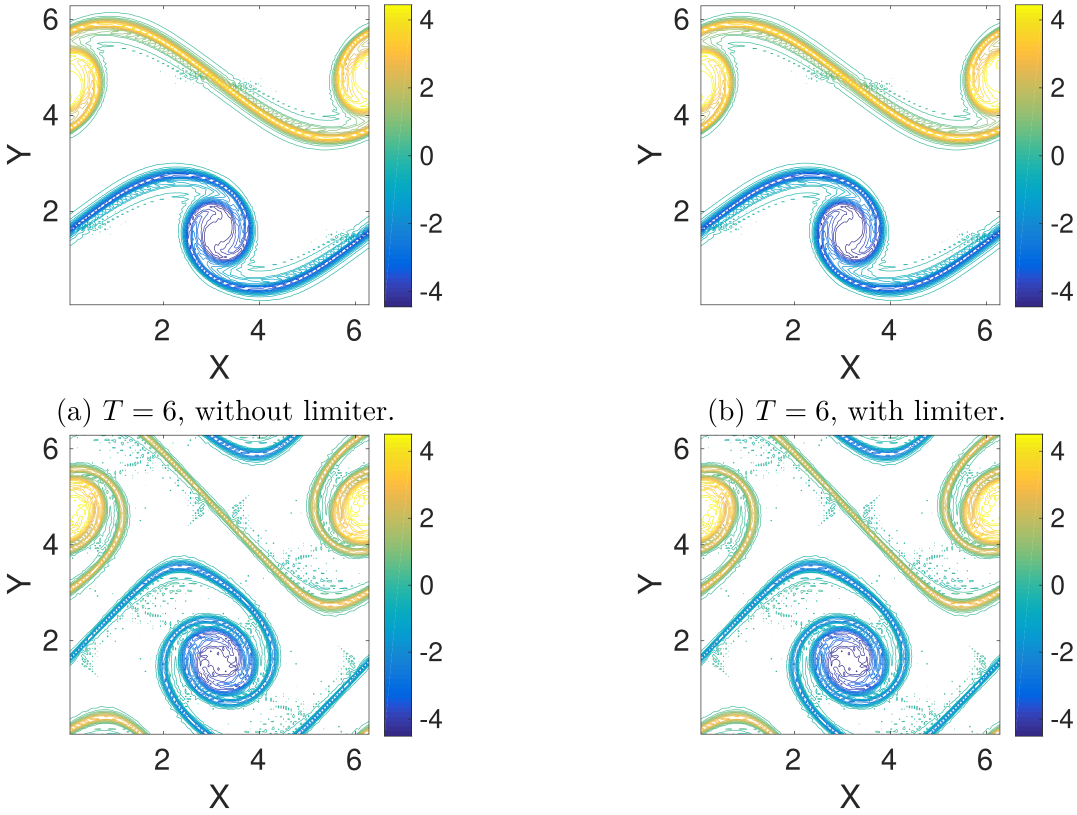
Table 6.12. Fourth order compact finite difference scheme with the bound-preserving limiter for the incompressible Navier-Stokes equation.

	SSP multistep				SSP Runge-Kutta			
$N\times N$	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	2.76E-5	-	6.58E-5	-	3.67E-5	-	8.76E-5	-
20×20	1.80E-6	3.94	4.29E-6	3.94	2.16E-6	4.09	5.16E-6	4.09
40×40	1.22E-7	3.88	3.07E-7	3.80	1.33E-7	4.02	3.35E-7	3.94
80×80	7.93E-9	3.95	1.97E-8	3.97	8.42E-9	3.99	2.09E-8	4.00
160×160	5.03E-10	3.98	1.24E-9	3.98	5.17E-10	4.03	1.28E-9	4.03

Example 12. Double Shear Layer Problem. We test the double shear layer problem on the domain $[0,2\pi]\times[0,2\pi]$ with a periodic boundary condition. The initial condition is

$$\omega(x,y,0)=\begin{cases}\delta cos(x)-\frac{1}{\rho}sech^2((y-\frac{\pi}{2})/\rho),\ y\le\pi\\ \delta cos(x)+\frac{1}{\rho}sech^2((\frac{3\pi}{2}-y)/\rho), y>\pi\end{cases}$$

with $\delta=0.05$ and $\rho=\pi/15$. The vorticity ω for $Re=5000$ at time $T=6$ and $T=8$ are shown in Figure 6.9. We use the fourth order compact finite difference with SSP multistep method on a 120×120 mesh solving the Navier-Stokes equation (6.19) with $Re=5000$. The time step is $\Delta t=C_{ms}\min\{\frac{1}{12\max_x|u_0|}\Delta x, \frac{5Re}{48}\Delta x^2\}$. Although one can barely see any difference between the results with the limiter and without the limiter from the contour, we point out that the numerical solutions of the scheme with the bound-preserving limiter are ensured to be in the range $[-\delta-\frac{1}{\rho},\delta+\frac{1}{\rho}]$. In the numerical solutions, we observe some obvious oscillations, which would be reduced if the TVB limiter is also used. On the other hand, if the physical diffusion in the Navier-Stokes equation is resolved on a fine enough mesh, the physical diffusion can also smooth out the oscillations, as we will see in the next example.



(a) $T = 6$, without limiter.

(b) $T = 6$, with limiter.

(c) $T = 8$, with limiter.

(d) $T = 8$, with limiter.

Figure 6.9. Double shear layer problem.

Example 13. Vortex Patch Problem. We test the vortex patch problem in the domain $[0, 2\pi] \times [0, 2\pi]$ with a periodic boundary condition. The initial condition is

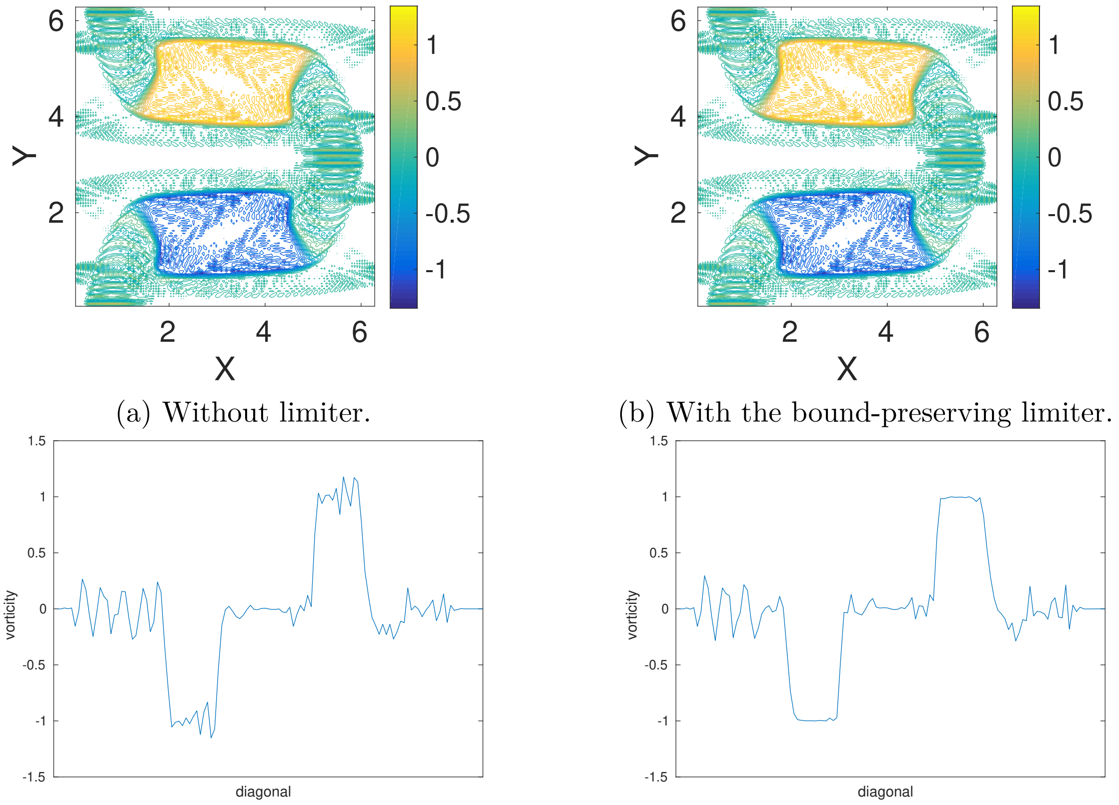
$$\omega(x, y, 0) = \begin{cases} -1, & (x, y) \in [\frac{\pi}{2}, \frac{3\pi}{2}] \times [\frac{\pi}{4}, \frac{3\pi}{4}]; \\ 1, & (x, y) \in [\frac{\pi}{2}, \frac{3\pi}{2}] \times [\frac{5\pi}{4}, \frac{7\pi}{4}]; \\ 0, & \text{otherwise.} \end{cases}$$

Numerical solutions for incompressible Euler and Navier-Stokes equations are shown in Figure 6.10, Figure 6.11 and Figure 6.12. In Figure 6.10, we use fourth order accurate compact finite difference scheme to solve the incompressible Euler equation at $T = 5$ on a 120×120 mesh. The time step is $\Delta t = C_{ms}\frac{1}{6\max|u_0|}\Delta x$. The second row is the cut along the diagonal of the two-dimensional array. We can observe that the solutions generated by the compact finite difference scheme with only the bound-preserving limiter are still highly oscillatory for the Euler equation, thus in principle other limiters should be used to eliminate the oscillations, e.g., the TVB limiter. On the other hand, we solve the incompressible Navier-Stokes equation with $Re = 1000$, at time $T = 5$ on a 60×60 mesh with the time step $\Delta t = C_{ms}\min\{\frac{1}{12\max|u_0|}\Delta x, \frac{5Re}{48}\Delta x^2\}$. The second row is the cut along the diagonal of the two-dimensional array. We observe that the numerical solution is non-oscillatory on a fine enough mesh. Notice that the oscillations in Figure 6.10 and Figure 6.11 suggest that the artificial viscosity induced by the bound-preserving limiter is quite low, thus it is the physical diffusion in the Navier-Stokes equation that starts to smooth out the numerical oscillations in Figure 6.12. For Figure 6.12, on a 120×120 mesh we use time step is $\Delta t = C_{ms}\min\{\frac{1}{12\max|u_0|}\Delta x, \frac{5Re}{48}\Delta x^2\}$.

The same phenomenon was also observed for the high order positivity-preserving discontinuous Galerkin scheme for the compressible Navier-Stokes system in [96]. In other words, for solving convection diffusion problems on resolved meshes, the bound-preserving limiter is enough for high order schemes producing satisfying results and there is no need to use the TVB limiter.

6.8 Concluding Remarks

In this chapter we have demonstrated that fourth order accurate compact finite difference schemes for convection diffusion problems with periodic boundary conditions satisfy a weak monotonicity property, and a simple three-point stencil limiter can enforce bounds without destroying the global conservation. Since the limiter is designed based on an intrinsic property in the high order finite difference schemes, the accuracy of the limiter can be easily justified. This is the first time that the weak monotonicity is established for a high order accurate finite difference scheme, complementary to results regarding the weak monotonicity property of high order finite volume and discontinuous Galerkin schemes in [23], [93], [95].



(a) Without limiter.

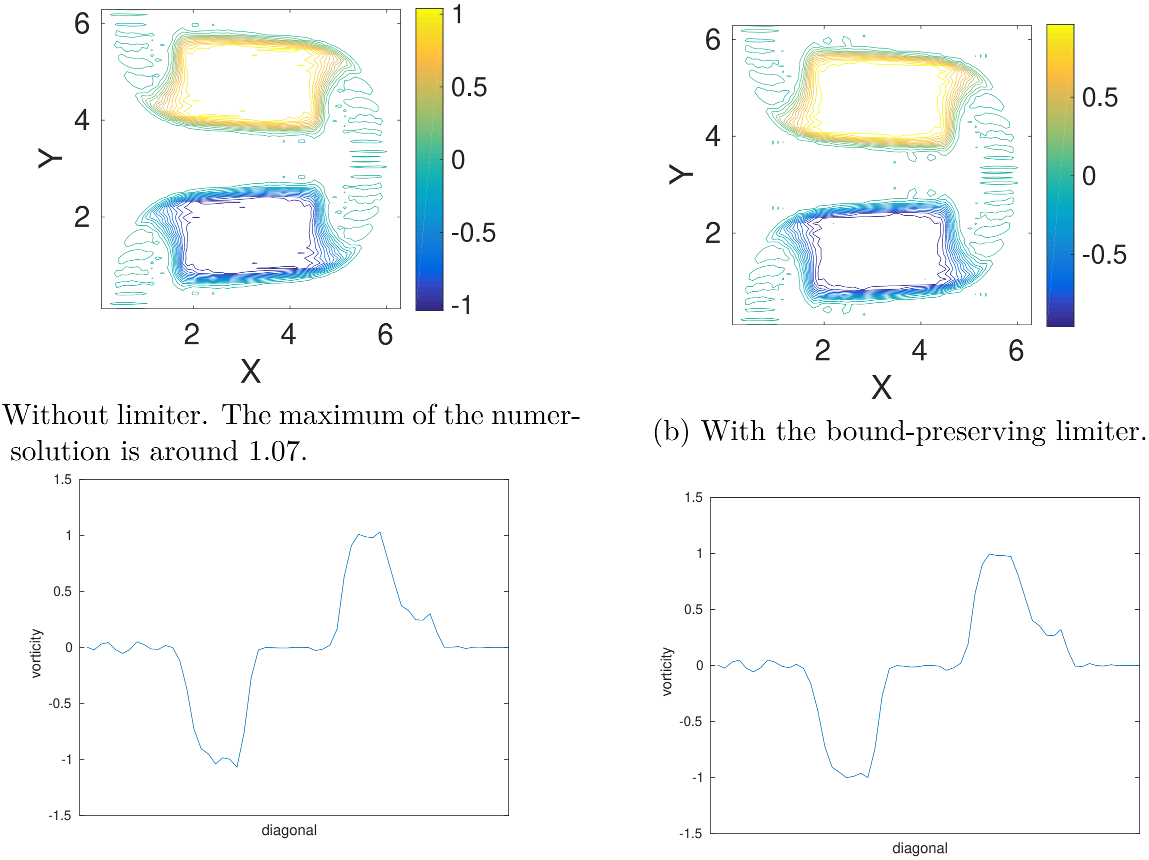
(b) With the bound-preserving limiter.

(c) Without limiter.

(d) With the bound-preserving limiter.

Figure 6.10. Vortex patch for incompressible Naiver-Stokes equation on the 120×120 mesh.

We have discussed extensions to two dimensions, higher order accurate schemes and general boundary conditions, for which the five-diagonal weighting matrices can be factored as a product of tridiagonal matrices so that the same simple three-point stencil bound-preserving limiter can still be used. We have also proved that the TVB limiter in [101] does not affect the bound-preserving property. Thus with both the TVB and the bound-preserving limiters, the numerical solutions of high order compact finite difference scheme can be rendered non-oscillatory and strictly bound-preserving without losing accuracy and global conservation. To generalize the bound-preserving scheme to incompressible flows, we have proposed a vector field which satisfies a discrete divergence free constraint. Extensive numerical results suggest the good performance of the high order bound-preserving compact finite difference schemes.



(a) Without limiter. The maximum of the numerical solution is around 1.07.

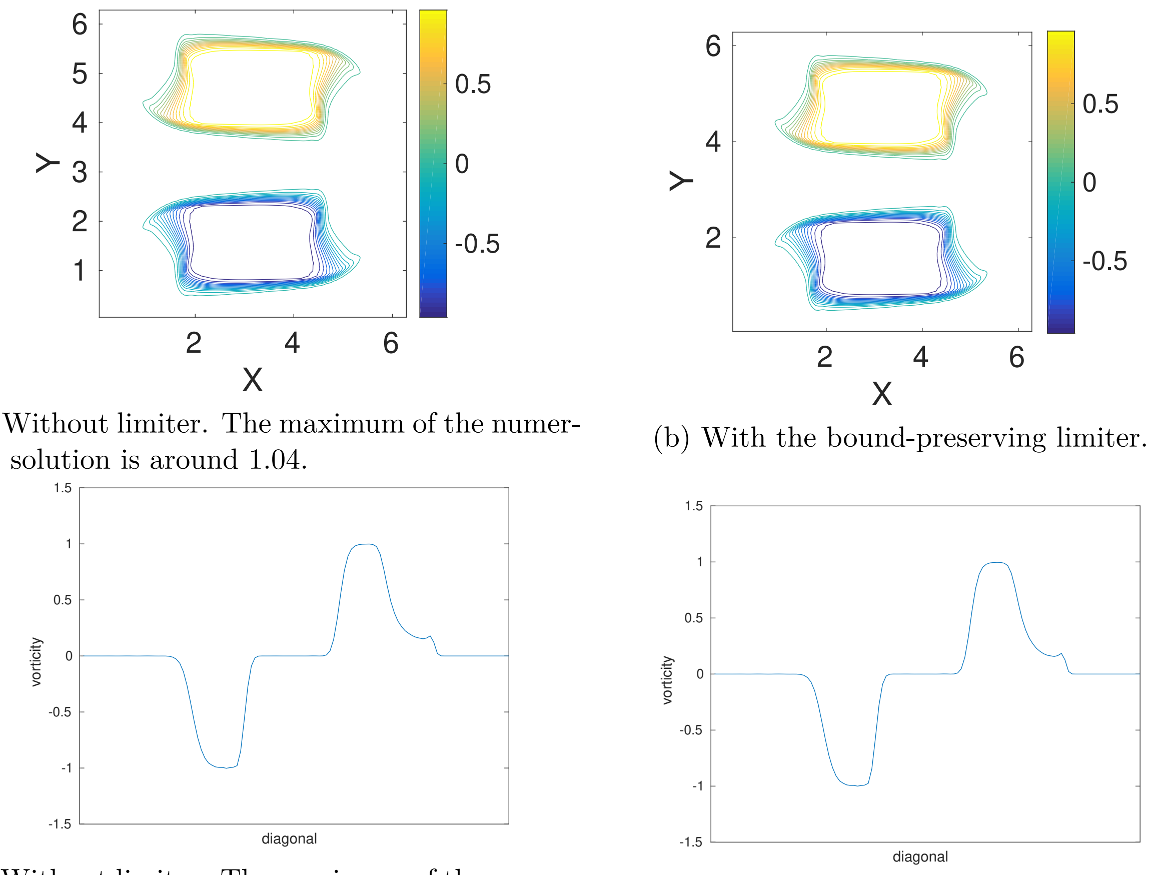
(b) With the bound-preserving limiter.

(c) Without limiter. The maximum of the numerical solution is around 1.07.

(d) With the bound-preserving limiter.

Figure 6.11. Vortex patch for incompressible Naiver-Stokes equation on the 60×60 mesh.

For more generalizations and applications, there are certain complications. For using compact finite difference schemes on non-uniform meshes, one popular approach is to introduce a mapping to a uniform grid but such a mapping results in an extra variable coefficient which may affect the weak monotonicity. Thus any extension to non-uniform grids is much less straightforward. For applications to systems, e.g., preserving positivity of density and pressure in compressible Euler equations, the weak monotonicity can be easily extended to a weak positivity property. However, the same three-point stencil limiter cannot enforce the positivity for pressure. One has to construct a new limiter for systems.



(a) Without limiter. The maximum of the numerical solution is around 1.04.

(b) With the bound-preserving limiter.

(c) Without limiter. The maximum of the numerical solution is around 1.04.

(d) With the bound-preserving limiter.

Figure 6.12. Vortex patch for Naiver-Stokes equation on the 120 × 120 mesh.

6.9 Appendix A: Comparison With The Nine-point Discrete Laplacian

Consider solving the two-dimensional Poisson equations $u_{xx} + u_{yy} = f$ with either homogeneous Dirichlet boundary conditions or periodic boundary conditions on a rectangular domain. Let $\mathbf{u}$ be a $N_x \times N_y$ matrix with entries $u_{i,j}$ denoting the numerical solutions at a uniform grid $(x_i, y_j) = (\frac{i}{Nx}, \frac{j}{Ny})$. Let $\mathbf{f}$ be a $N_x \times N_y$ matrix with entries $f_{i,j} = f(x_i, y_j)$. The fourth order compact finite difference method in Section 6.3.2 for $u_{xx} + u_{yy} = f$ can be written as:

$$\frac{1}{\Delta x^2} W_{2x}^{-1} D_{xx}\mathbf{u} + \frac{1}{\Delta y^2} W_{2y}^{-1} D_{yy}\mathbf{u} = f(\mathbf{u}). \tag{6.1}$$

For convenience, we introduce two matrices,

$$U = \begin{pmatrix} u_{i-1,j+1} & u_{i,j+1} & u_{i+1,j+1} \\ u_{i-1,j} & u_{i,j} & u_{i+1,j} \\ u_{i-1,j-1} & u_{i,j-1} & u_{i+1,j-1} \end{pmatrix}, \quad F = \begin{pmatrix} f_{i-1,j+1} & f_{i,j+1} & f_{i+1,j+1} \\ f_{i-1,j} & f_{i,j} & f_{i+1,j} \\ f_{i-1,j-1} & f_{i,j-1} & f_{i+1,j-1} \end{pmatrix}.$$

Notice that the scheme (6.1) is equivalent to

$$\frac{1}{\Delta x^2} W_{2y} D_{xx} \mathbf{u} + \frac{1}{\Delta y^2} W_{2x} D_{yy} \mathbf{u} = W_{2x} W_{2y} f(\mathbf{u}),$$

which can be written as

$$\frac{1}{12\Delta x^2} \begin{pmatrix} 1 & -2 & 1 \\ 10 & -20 & 10 \\ 1 & -2 & 1 \end{pmatrix} : U + \frac{1}{12\Delta y^2} \begin{pmatrix} 1 & 10 & 1 \\ -2 & -20 & -2 \\ 1 & 10 & 1 \end{pmatrix} : U = \frac{1}{144} \begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : F, \tag{6.2}$$

where : denotes the sum of all entrywise products in two matrices of the same size.

In particular, if $\Delta x = \Delta y = h$, the scheme above reduces to

$$\frac{1}{6h^2} \begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix} : U = \frac{1}{144} \begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : F.$$

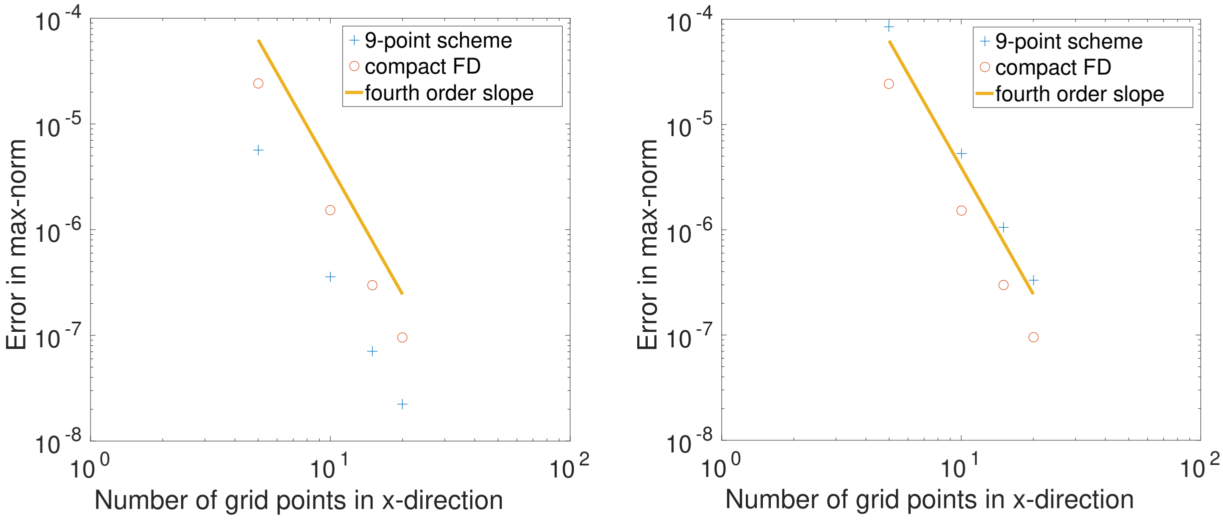
Recall that the classical nine-point discrete Laplacian [113] for the Poisson equation can be written as

$$\frac{1}{12\Delta x^2} \begin{pmatrix} 1 & -2 & 1 \\ 10 & -20 & 10 \\ 1 & -2 & 1 \end{pmatrix} : U + \frac{1}{12\Delta y^2} \begin{pmatrix} 1 & 10 & 1 \\ -2 & -20 & -2 \\ 1 & 10 & 1 \end{pmatrix} : U = \frac{1}{12} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 0 \end{pmatrix} : F, \tag{6.3}$$

which reduces to the following under the assumption $\Delta x = \Delta y = h$,

$$\frac{1}{6h^2}\begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix} : U = \frac{1}{12}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 0 \end{pmatrix} : F.$$

Both schemes (6.2) and (6.3) are fourth order accurate and they have the same stencil in the left hand side. As to which scheme produces smaller errors, it seems to be problem dependent, see Figure 6.13. Figure 6.13 shows the errors of two schemes (6.2) and (6.3) using uniform grids with $\Delta x = \frac{2}{3}\Delta y$ for solving the Poisson equation $u_{xx}+u_{yy} = f$ on a rectangle $[0, 1]\times[0, 2]$ with Dirichlet boundary conditions. For solution 1, we have $u(x, y) = \sin(\pi x)\sin(\pi y) + 2x$, for solution 2, we have $u(x, y) = \sin(\pi x)\sin(\pi y) + 4x^4y^4$.



(a) Solution 1.

(b) Solution 2.

Figure 6.13. Error comparison.

6.10 Appendix B: M-matrices And A Discrete Maximum Principle

Consider solving the heat equation $u_t = u_{xx} + u_{yy}$ with a periodic boundary condition. It is well known that a discrete maximum principle is satisfied under certain time step constraints if the spatial discretization is the nine-point discrete Laplacian or the compact scheme (6.1) with backward Euler and Crank-Nicolson time discretizations. For simplicity, we only consider the compact scheme (6.1) and the discussion for the nine-point discrete Laplacian is similar. Assume $\Delta x = \Delta y = h$. For backward Euler, the scheme can be written as

$$\frac{1}{144}\begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : (U^{n+1} - U^n) = \frac{\Delta t}{6h^2}\begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix} : U^{n+1},$$

thus

$$\frac{1}{144}\begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : U^{n+1} - \frac{\Delta t}{6h^2}\begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix} : U^{n+1} = \frac{1}{144}\begin{pmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{pmatrix} : U^n.$$

Let A and B denote the matrices corresponding to the operator in the left hand side and right hand side above respectively, then the scheme can be written as

$$A\mathbf{u}^{n+1} = B\mathbf{u}^n,$$

and A is a M-Matrix (diagonally dominant, positive diagonal entries and non-positive off diagonal entries) under the following constraint which allows very large time steps:

$$\frac{\Delta t}{h^2} \geq \frac{5}{48}.$$

The inverses of M-Matrices have non-negative entries, e.g., see [107]. Thus A^{-1} has non-negative entries. Moreover, it is straightforward to check that $A\mathbf{e} = \mathbf{e}$ where $\mathbf{e} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$. Thus $A^{-1}\mathbf{e} = \mathbf{e}$, which implies the sum of each row of A^{-1} is 1 thus each row of A^{-1} multiplying any vector V is a convex combination of entries of V. It is also obvious that each entry of B is non-negative and the sum of each row of B is 1. Therefore, $\mathbf{u}^{n+1} = A^{-1}B\mathbf{u}^n$ satisfies a discrete maximum principle:

$$\min_{i,j} u^n_{i,j} \leq u^{n+1}_{i,j} \leq \max_{i,j} u^n_{i,j}.$$

For the second order accurate Crank-Nicolson time discretization, the scheme can be written as

$$\frac{1}{144}\begin{pmatrix}1 & 10 & 1\\ 10 & 100 & 10\\ 1 & 10 & 1\end{pmatrix} : (U^{n+1} - U^n) = \frac{\Delta t}{6h^2}\begin{pmatrix}1 & 4 & 1\\ 4 & -20 & 4\\ 1 & 4 & 1\end{pmatrix} : \frac{U^{n+1}+U^n}{2},$$

thus

$$\left[\frac{1}{144}\begin{pmatrix}1 & 10 & 1\\ 10 & 100 & 10\\ 1 & 10 & 1\end{pmatrix} - \frac{\Delta t}{12h^2}\begin{pmatrix}1 & 4 & 1\\ 4 & -20 & 4\\ 1 & 4 & 1\end{pmatrix}\right] : U^{n+1} =$$

$$\left[\frac{1}{144}\begin{pmatrix}1 & 10 & 1\\ 10 & 100 & 10\\ 1 & 10 & 1\end{pmatrix} + \frac{\Delta t}{12h^2}\begin{pmatrix}1 & 4 & 1\\ 4 & -20 & 4\\ 1 & 4 & 1\end{pmatrix}\right] : U^{n}.$$

Let the matrix-vector form of the scheme above be $A\mathbf{u}^{n+1} = B\mathbf{u}^n$. Then for A to be an M-Matrix, we only need $\frac{\Delta t}{h^2} \geq \frac{5}{24}$. However, for B to have non-negative entries, we need $\frac{\Delta t}{h^2} \leq \frac{5}{12}$. Thus the Crank-Nicolson method can ensure a discrete maximum principle if the time step satisfies,

$$\frac{5}{24}h^2 \leq \Delta t \leq \frac{5}{12}h^2.$$

6.11 Appendix C: Fast Poisson Solvers

6.11.1 Dirichlet boundary conditions

Consider solving the Poisson equation $u_{xx} + u_{yy} = f(x, y)$ on a rectangular domain $[0, L_x] \times [0, L_y]$ with homogeneous Dirichlet boundary conditions. Assume we use the grid $x_i = i\Delta x$, $i = 0, \cdots, N_x + 1$ with uniform spacing $\Delta x = \frac{L_x}{N_x+1}$ for the x-variable and $y_j = j\Delta y$, $j = 0, \cdots, N_y + 1$ with uniform spacing $\Delta y = \frac{L_y}{N_y+1}$ for y-variable. Let $\mathbf{u}$ be a $N_x \times N_y$ matrix such that its (i, j) entry $u_{i,j}$ is the numerical solution at interior grid points (x_i, y_j). Let $\mathbf{F}$ be a $(N_x + 2) \times (N_y + 2)$ matrix with entries $f(x_i, y_j)$ for $i = 0, \cdots, N_x + 1$ and $j = 0, \cdots, N_y + 1$.

To obtain the matrix representation of the operator in (6.2) and (6.3), we consider two operators:

- Kronecker product of two matrices: if A is $m \times n$ and B is $p \times q$, then $A \otimes B$ is $mp \times nq$ give by
$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \vdots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$
- For a $m \times n$ matrix X, $\mathrm{vec}(X)$ denotes a column vector of size mn made of the columns of X stacked atop one another from left to right.

The following properties will be used:

1. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
2. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
3. $(B^T \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB)$.

We define two tridiagonal square matrices of size $N_x \times N_x$:

$$D_{xx} = \begin{pmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix}, W_{2x} = \frac{1}{12}\begin{pmatrix} 10 & 1 & & & & \\ 1 & 10 & 1 & & & \\ & 1 & 10 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 10 & 1 \\ & & & & 1 & 10 \end{pmatrix}.$$

Let $\overline{W}_{2x}$ denote a $N_x \times (N_x + 2)$ tridiagonal matrix of the following form:

$$\overline{W}_{2x} = \frac{1}{12}\begin{pmatrix} 1 & 10 & 1 & & & \\ & 1 & 10 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 10 & 1 \end{pmatrix}. \tag{6.4}$$

The matrices D_{yy}, W_{2y} and $\overline{W}_{2y}$ are similarly defined.

Then the scheme (6.2) can be written in a matrix-vector form:

$$\frac{1}{\Delta x^2} D_{xx}\mathbf{u}W_{2y}^T + \frac{1}{\Delta y^2} W_{2x}\mathbf{u}D_{yy}^T = \overline{W}_{2x}\mathbf{F}\overline{W}_{2y}^T,$$

or equivalently,

$$\left(W_{2y} \otimes \frac{1}{\Delta x^2} D_{xx} + \frac{1}{\Delta y^2} D_{yy} \otimes W_{2x}\right) \operatorname{vec}(\mathbf{u}) = (\overline{W}_{2x} \otimes \overline{W}_{2y}) \operatorname{vec}(\mathbf{F}). \tag{6.5}$$

Let $\mathbf{h}_x = [h_1, h_2, \cdots, h_{N_x}]^T$ with $h_i = \frac{i}{N_x+1}$, and $\sin(m\pi\mathbf{h}_x)$ denote a column vector of size N_x with its i-th entry being $\sin(m\pi h_i)$. Then $\sin(m\pi\mathbf{h}_x)$ are the eigenvectors of D_{xx} and W_{2x} with the associated eigenvalues being $2\cos(\frac{m\pi}{N_x+1}) - 2$ and $\frac{5}{6} + \frac{1}{6}\cos(\frac{m\pi}{N_x+1})$ respectively for $m = 1, \cdots, N_x$. Let

$$S_x = [\sin(\pi\mathbf{h}_x), \sin(2\pi\mathbf{h}_x), \cdots, \sin(N_x\pi\mathbf{h}_x)]$$

be the $N_x \times N_x$ eigenvector matrix, then S_x is a symmetric matrix. Let Λ_{1x} denote a diagonal matrix with diagonal entries $2\cos(\frac{m\pi}{N_x+1}) - 2$ and Λ_{2x} denote a diagonal matrix with diagonal entries $\frac{5}{6} + \frac{1}{6}\cos(\frac{m\pi}{N_x+1})$, then we have $D_{xx} = S_x\Lambda_{1x}S_x^{-1}$ and $W_{2x} = S_x\Lambda_{2x}S_x^{-1}$, thus

$$W_{2y} \otimes D_{xx} = (S_y\Lambda_{2y}S_y^{-1}) \otimes (S_x\Lambda_{1x}S_x^{-1}) = (S_y \otimes S_x)(\Lambda_{2y} \otimes \Lambda_{1x})(S_y^{-1} \otimes S_x^{-1}).$$

The scheme can be written as

$$(S_y \otimes S_x)(\frac{1}{\Delta x^2}\Lambda_{2y} \otimes \Lambda_{1x} + \frac{1}{\Delta y^2}\Lambda_{1y} \otimes \Lambda_{2x})(S_y^{-1} \otimes S_x^{-1}) \operatorname{vec}(\mathbf{u}) = (\overline{W}_{2y} \otimes \overline{W}_{2x}) \operatorname{vec}(\mathbf{F}).$$

Let Λ be a $N_x \times N_y$ matrix with $\Lambda_{i,j}$ being equal to

$$\frac{1}{3\Delta x^2}\left(\cos(\frac{i\pi}{N_x+1}) - 1\right)\left(\cos(\frac{m\pi}{N_y+1}) + 5\right) + \frac{1}{3\Delta y^2}\left(\cos(\frac{m\pi}{N_x+1}) + 5\right)\left(\cos(\frac{j\pi}{N_y+1}) - 1\right),$$

then vec(Λ) are precisely the diagonal entries of the diagonal matrix $\frac{1}{\Delta x^2}\Lambda_{2y}\otimes\Lambda_{1x}+\frac{1}{\Delta y^2}\Lambda_{1y}\otimes\Lambda_{2x}$, then the scheme above is equivalent to

$$S_x(\Lambda\circ(S_x^{-1}\mathbf{u}S_y^{-1}))S_y=\overline{W}_{2x}\mathbf{F}\overline{W}_{2y}^T,$$

where the symmetry of S matrices is used. The solution is given by

$$\mathbf{u}=S_x\{[S_x^{-1}(\overline{W}_{2x}\mathbf{F}\overline{W}_{2y}^T)S_y^{-1}]./\Lambda\}S_y, \tag{6.6}$$

where ./ denotes the entrywise division for two matrices of the same size.

Since the multiplication of the matrices S and S^{-1} can be implemented by the *Discrete Sine Transform*, (6.6) gives a fast Poisson solver.

For nonhomogeneous Dirichlet boundary conditions, the fourth order accurate compact finite difference scheme can also be written in the form of (6.5):

$$\left(W_{2y}\otimes\frac{1}{\Delta x^2}D_{xx}+\frac{1}{\Delta y^2}D_{yy}\otimes W_{2x}\right)\text{vec}(\mathbf{u})=\text{vec}(\tilde{\mathbf{F}}), \tag{6.7}$$

where $\tilde{\mathbf{F}}$ consists of both $\mathbf{F}$ and the Dirichlet boundary conditions. Thus the scheme can still be efficiently implemented by the *Discrete Sine Transform*.

6.11.2 Periodic boundary conditions

For periodic boundary conditions on a rectangular domain, we should consider the uniform grid $x_i=i\Delta x$, $i=1,\cdots,N_x$ with $\Delta x=\frac{L_x}{N_x}$ and $y_j=j\Delta y$, $j=1,\cdots,N_y$ with uniform spacing $\Delta y=\frac{L_y}{N_y}$, then the fourth order accurate compact finite difference scheme can still be written in the form of (6.5) with the D_{xx}, D_{yy}, W_{2x} and W_{2y} matrices being redefined as circulant matrices:

$$D_{xx} = \begin{pmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ 1 & & & & 1 & -2 \end{pmatrix}, W_{2x} = \frac{1}{12}\begin{pmatrix} 10 & 1 & & & & 1 \\ 1 & 10 & 1 & & & \\ & 1 & 10 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 10 & 1 \\ 1 & & & & 1 & 10 \end{pmatrix}.$$

The Discrete Fourier Matrix is the eigenvector matrix for any circulant matrices, and the corresponding eigenvalues are for D_{xx} and W_{2x} are $2\cos(\frac{m2\pi}{N_x}) - 2$ and $\frac{1}{6}\cos(\frac{m2\pi}{N_x}) + \frac{5}{6}$ for $m = 0, 1, 2, \cdots, Nx - 1$. The matrix $W_{2y} \otimes \frac{1}{\Delta x^2} D_{xx} + \frac{1}{\Delta y^2} D_{yy} \otimes W_{2x}$ is singular because its first eigenvalue $\Lambda_{1,1}$ is zero. Nonetheless, the scheme can still be implemented by solving (6.6) with Fast Fourier Transform. For the zero eigenvalue, we can simply reset the division by eigenvalue zero to zero. Since the eigenvector for eigenvalue zero is $\mathbf{e} = \begin{pmatrix} 1 \; 1 \cdots 1 \end{pmatrix}^T$, and the columns of the Discrete Fourier Matrix are orthogonal to one another, resetting the division by eigenvalue zero to zero simply means that we obtain a numerical solution satisfying $\sum_i \sum_j u_{i,j} = 0$. And this is also the least square solution to the singular linear system.

6.11.3 Neumann boundary conditions

For Dirichlet and periodic boundary conditions, we can invert the matrix coefficient matrix in (6.5) using eigenvectors of much smaller matrices W_{2x} and D_{xx} due to the fact that $W_{2x} - \frac{1}{12} D_{xx}$ is the identity matrix Id. Here we discuss how to achieve a fourth order accurate boundary approximation for Neumann boundary conditions by keeping $W_{2x} - \frac{1}{12} D_{xx} =$

Id. We first consider a one-dimensional problem with homogeneous Neumann boundary conditions:

$$u(x) = f(x), x \in [0, L_x],$$
$$u(0) = u(L_x) = 0.$$

Assume we use the uniform grid $x_i = i\Delta x$, $i = 0, \cdots, N_x + 1$ with $\Delta x = \frac{L_x}{N_x+1}$. The two boundary point values u_0 and u_{N_x+1} can be expressed in terms of interior point values through boundary conditions. For approximating the boundary conditions, we can apply the fourth order one-sided difference at $x = 0$:

$$u(0) \approx \frac{-25u(0) + 48u(\Delta x) - 36u(2\Delta x) + 16u(3\Delta x) - 3u(4\Delta x)}{12\Delta x}$$

which implies the finite difference approximation:

$$u_0 = \frac{48u_1 - 36u_2 + 16u_3 - 3u_4}{25}.$$

Define two column vectors:

$$\mathbf{u} = [u_1, u_2, \cdots, u_{N_x}]^T, \quad \mathbf{f} = [f(x_0), f(x_1), \cdots, f(x_{N_x}), f(x_{N_x+1})]^T,$$

then a fourth order accurate compact finite difference scheme can be written as

$$\frac{1}{\Delta x^2}\overline{D}_{xx} I_x \mathbf{u} = \overline{W}_{2x}\mathbf{f},$$

where $\overline{W}_{2x}$ is the same as in (6.4), and $\overline{D}_{xx}$ is a matrix of size $N_x \times (N_x+2)$ and I_x is a matrix of size $(N_x+2) \times N_x$:

$$\overline{D}_{xx} = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}, I_x = \begin{pmatrix} \frac{48}{25} & -\frac{36}{25} & \frac{16}{25} & -\frac{3}{25} & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & -\frac{3}{25} & \frac{16}{25} & -\frac{36}{25} & \frac{48}{25} \end{pmatrix}.$$

Now consider solving the Poisson equation $u_{xx} + u_{yy} = f(x,y)$ on a rectangular domain $[0, L_x] \times [0, L_y]$ with homogeneous Neumann boundary conditions. Assume we use the grid $x_i = i\Delta x$, $i = 0, \cdots, N_x + 1$ with $\Delta x = \frac{L_x}{N_x+1}$ and $y_j = j\Delta y$, $j = 0, \cdots, N_y + 1$ with uniform spacing $\Delta y = \frac{L_y}{N_y+1}$. Let $\mathbf{u}$ be a $N_x \times N_y$ matrix such that $u_{i,j}$ is the numerical solution at (x_i, y_j) and $\mathbf{F}$ be a $(N_x+2) \times (N_y+2)$ matrix with entries $f(x_i, y_j)$ ($i = 0, \cdots, N_x + 1$, $j = 0, \cdots, N_y + 1$). Then a fourth order accurate compact finite difference scheme can be written as

$$\frac{1}{\Delta x^2}\overline{D}_{xx} I_x \mathbf{u} I_y^T \overline{W}_{2y}^T + \frac{1}{\Delta y^2}\overline{W}_{2x} I_x \mathbf{u} I_y^T \overline{D}_{yy}^T = \overline{W}_{2x}\mathbf{F}\overline{W}_{2y}^T.$$

Let $D_{xx} = \overline{D}_{xx} I_x$ and $W_{2x} = \overline{W}_{2x} I_x$, then the scheme can be written as (6.5).

Notice that $W_{2x} - \frac{1}{12}D_{xx} = (\overline{W}_{2x} - \frac{1}{12}\overline{D}_{xx})I_x$ is still the identity matrix thus W_{2x} and D_{xx} still have the same eigenvectors. Let S be the eigenvector matrix and Λ_1 and Λ_2 be diagonal matrices with eigenvalues, then the scheme can still be implemented as (6.6). The eigenvectors S and the eigenvalues can be obtained by computing eigenvalue problems for two small matrices D_{xx} of size $N_x \times N_x$ and D_{yy} of size $N_y \times N_y$. If such a Poisson problem needs to be solved in each time step in a time-dependent problem such as the incompressible flow equations, then this is an efficient Poisson solver because S and Λ can be computed before time evolution without considering eigenvalue problems for any matrix of size $N_x N_y \times N_x N_y$.

For nonhomogeneous Neumann boundary conditions, the point values of u along the boundary should be expressed in terms of interior ones as follows:

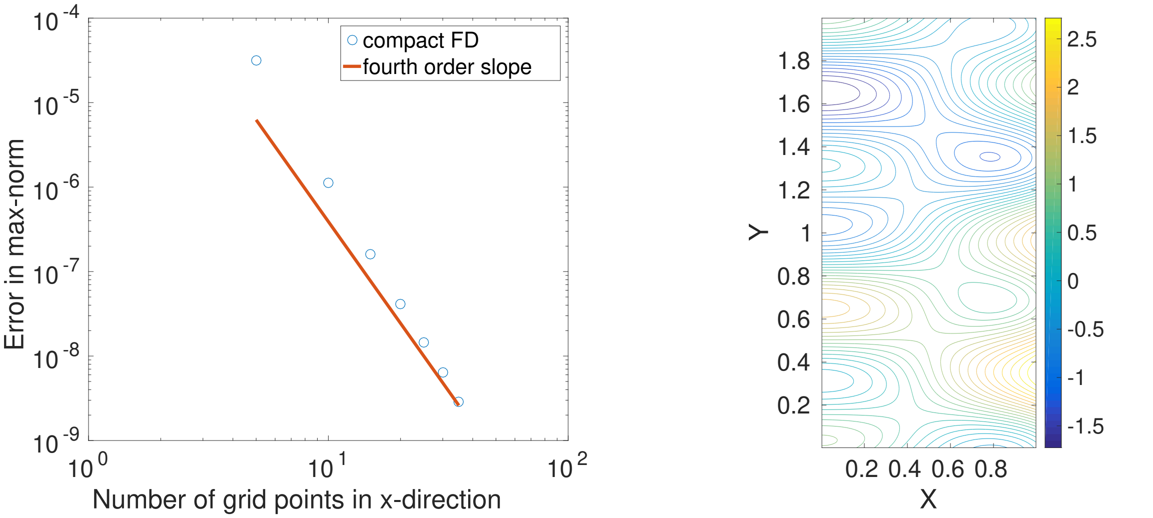
1. First obtain the point values except the two cell ends (i.e., corner points of the rectangular domain) for each of the four boundary line segments. For instance, if the left boundary condition is $\frac{\partial u}{\partial x}(0, y) = g(y)$, then we obtain

$$u_{0,j} = \frac{48u_{1,j} - 36u_{2,j} + 16u_{3,j} - 3u_{4,j} + 12\Delta x g(y_j)}{25}, \quad j = 1, \cdots, N_y.$$

2. Second, obtain the approximation at four corners using the point values along the boundary. For instance, if the bottom boundary condition is $\frac{\partial u}{\partial y}(x, 0) = h(x)$, then

$$u_{0,0} = \frac{48u_{1,0} - 36u_{2,0} + 16u_{3,0} - 3u_{4,0} + 12\Delta y h(0)}{25}$$

The scheme can still be written as (6.7) with $\tilde{\mathbf{F}}$ consisting of $\mathbf{F}$ and the nonhomogeneous boundary conditions. Notice that the matrix in (6.7) is singular thus we need to reset the division by eigenvalue zero to zero, which however no longer means that the obtained solution satisfies $\sum_i \sum_j u_{i,j} = 0$ since the eigenvectors are not necessarily orthogonal to one another. See Figure 6.14 for the accuracy test of the fourth order compact finite difference scheme using uniform grids with $\Delta x = \frac{3}{2}\Delta y$ for solving the Poisson equation $u_{xx} + u_{yy} = f$ on a rectangle $[0, 1] \times [0, 2]$ with nonhomogeneous Neumann boundary conditions. The exact solution is $u(x, y) = \cos(\pi x)\cos(3\pi y) + \sin(\pi y) + x^4$. Since the solutions to Neumann boundary conditions are unique up to any constant, when computing errors, we need to add a constant $\frac{1}{N_x}\frac{1}{N_y}\sum_{i,j}[u(x_i, y_j) - u_{i,j}]$ to each entry of $\mathbf{u}$.



(a) Convergence rate.

(b) The contour of the solution.

Figure 6.14. Accuracy test for Neumann boundary condition.

REFERENCES

[1] H. Li, S. Xie, and X. Zhang, “A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations,” *SIAM Journal on Numerical Analysis*, vol. 56, no. 6, pp. 3308–3345, 2018.

[2] H. Li, D. Appelö, and X. Zhang, *Accuracy of spectral element method for wave, parabolic and schrödinger equations*, 2021. arXiv: 2103.00400 `[math.NA]`.

[3] P. G. Ciarlet, “Basic error estimates for elliptic problems,” *Handbook of Numerical Analysis*, vol. 2, pp. 17–351, 1991.

[4] P. Lesaint and M. Zlamal, “Superconvergence of the gradient of finite element solutions,” *RAIRO. Analyse numérique*, vol. 13, no. 2, pp. 139–166, 1979.

[5] C. Chen, “Superconvergent points of Galerkin’s method for two point boundary value problems,” *Numerical Mathematics A Journal of Chinese Universities*, vol. 1, pp. 73–79, 1979.

[6] M. Bakker, “A note on C^0 Galerkin methods for two-point boundary problems,” *Numer. Math.*, vol. 38, no. 3, pp. 447–453, 1982.

[7] J. Douglas Jr, T. Dupont, and M. F. Wheeler, “An L^∞ estimate and a superconvergence result for a Galerkin method for elliptic equations based on tensor products of piecewise polynomials,” 1974.

[8] L. Wahlbin, *Superconvergence in Galerkin finite element methods*. Springer, 2006.

[9] C. Chen, *Structure theory of superconvergence of finite elements (In Chinese)*. Hunan Science and Technology Press, Changsha, 2001.

[10] Q. Lin and N. Yan, *Construction and Analysis for Efficient Finite Element Method (In Chinese)*. Hebei University Press, 1996.

[11] Y. Maday and E. M. Rønquist, “Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries,” *Computer Methods in Applied Mechanics and Engineering*, vol. 80, no. 1-3, pp. 91–115, 1990.

[12] D. Komatitsch, J.-P. Vilotte, R. Vai, J. M. Castillo-Covarrubias, and F. J. Sánchez-Sesma, “The spectral element method for elastic wave equations: Application to 2-D and 3-D seismic problems,” *International Journal for numerical methods in engineering*, vol. 45, no. 9, pp. 1139–1164, 1999.

[13] D. Komatitsch and J. Tromp, “Introduction to the spectral element method for three-dimensional seismic wave propagation,” *Geophysical journal international*, vol. 139, no. 3, pp. 806–822, 1999.

[14] L. C. Evans, *Partial Differential Equations*, 2nd ed., ser. Graduate Studies in Mathematics. American Mathematical Society, 2010, vol. 019.

[15] P. G. Ciarlet, “Discrete maximum principle for finite-difference operators,” *Aequationes Math.*, vol. 4, no. 3, pp. 338–352, 1970.

[16] J. Bramble and B. Hubbard, “On the formulation of finite difference analogues of the Dirichlet problem for Poisson’s equation,” *Numer. Math.*, vol. 4, no. 1, pp. 313–327, 1962.

[17] P. G. Ciarlet and P.-A. Raviart, “Maximum principle and uniform convergence for the finite element method,” *Comput. Methods Appl. Mech. Engrg.*, vol. 2, no. 1, pp. 17–31, 1973.

[18] O. Axelsson and L. Kolotilina, “Monotonicity and discretization error estimates,” *SIAM J. Numer. Anal.*, vol. 27, no. 6, pp. 1591–1611, 1990.

[19] P. J. Ferket and A. A. Reusken, “A finite difference discretization method for elliptic problems on composite grids,” *Computing*, vol. 56, no. 4, pp. 343–369, 1996.

[20] W. Höhn and H. D. Mittelmann, “Some remarks on the discrete maximum-principle for finite elements of higher order,” *Computing*, vol. 27, no. 2, pp. 145–154, 1981.

[21] J. Lorenz, “Zur inversmonotonie diskreter probleme,” *Numer. Math.*, vol. 27, no. 2, pp. 227–238, 1977.

[22] L. J. Cross and X. Zhang, “On the monotonicity of high order discrete laplacian,” *arXiv preprint arXiv:2010.07282*, 2020.

[23] X. Zhang and C.-W. Shu, “On maximum-principle-satisfying high order schemes for scalar conservation laws,” *Journal of Computational Physics*, vol. 229, no. 9, pp. 3091–3120, 2010.

[24] J. Douglas, “Some superconvergence results for Galerkin methods for the approximate solution of two-point boundary problems,” *Topics in numerical analysis*, pp. 89–92, 1973.

[25] J. Douglas and T. Dupont, “Galerkin approximations for the two point boundary problem using continuous, piecewise polynomial spaces,” *Numer. Math.*, vol. 22, no. 2, pp. 99–109, 1974.

[26] J. Douglas Jr, T. Dupont, and M. F. Wheeler, "An l^∞ estimate and a superconvergence result for a galerkin method for elliptic equations based on tensor products of piecewise polynomials," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 8, no. R2, pp. 61–66, 1974.

[27] C. Chen and S. Hu, "The highest order superconvergence for bi-k degree rectangular elements at nodes: A proof of $2k$-conjecture," *Math. Comp.*, vol. 82, no. 283, pp. 1337–1355, 2013.

[28] W. He and Z. Zhang, "$2k$ superconvergence of Q_k finite elements by anisotropic mesh approximation in weighted Sobolev spaces," *Mathematics of Computation*, vol. 86, no. 306, pp. 1693–1718, 2017.

[29] W. He, Z. Zhang, and Q. Zou, "Ultraconvergence of high order fems for elliptic problems with variable coefficients," *Numerische Mathematik*, vol. 136, no. 1, pp. 215–248, May 2017.

[30] P. G. Ciarlet and P.-A. Raviart, "The combined effect of curved boundaries and numerical integration in isoparametric finite element methods," in *The mathematical foundations of the finite element method with applications to partial differential equations*, Elsevier, 1972, pp. 409–474.

[31] J. Whiteman, "Lagrangian finite element and finite difference methods for poisson problems," in *Numerische Behandlung von Differentialgleichungen*, Springer, 1975, pp. 331–355.

[32] Y. Huang and J. Xu, "Superconvergence of quadratic finite elements on mildly structured grids," *Math. Comp.*, vol. 77, no. 263, pp. 1253–1268, 2008.

[33] G. Savaré, "Regularity results for elliptic equations in Lipschitz domains," *Journal of Functional Analysis*, vol. 152, no. 1, pp. 176–201, 1998.

[34] P. Grisvard, *Elliptic problems in nonsmooth domains.* SIAM, 2011, vol. 69.

[35] F. Brezzi and L. Marini, "On the numerical solution of plate bending problems by hybrid methods," *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, vol. 9, no. R3, pp. 5–50, 1975.

[36] K. Smith, "Inequalities for formally positive integro-differential forms," *Bulletin of the American Mathematical Society*, vol. 67, no. 4, pp. 368–370, 1961.

[37] S. Agmon, *Lectures on elliptic boundary value problems.* American Mathematical Soc., 2010, vol. 369.

[38] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems.* Society for Industrial and Applied Mathematics, 2002.

[39] H. Li and X. Zhang, "Superconvergence of c^0-q^k finite element method for elliptic equations with approximated coefficients," *Journal of Scientific Computing*, vol. 82, no. 1, pp. 1–28, 2020.

[40] C. Chen, "Superconvergence of finite element solutions and their derivatives," *Numerical Mathematics A Journal of Chinese Universities*, vol. 3, no. 2, pp. 118–125, 1981.

[41] Q. Lin, N. Yan, and A. Zhou, "A rectangle test for interpolated finite elements," in *Proc. Sys. Sci. and Sys. Eng.(Hong Kong), Great Wall Culture Publ. Co*, 1991, pp. 217–229.

[42] J. Xu and L. Zikatanov, "Algebraic multigrid methods," *Acta Numerica*, vol. 26, pp. 591–721, 2017.

[43] J. Hesthaven and T. Warburton, "Nodal high-order methods on unstructured grids: I. time-domain solution of Maxwell's equations," *J. Comput. Phys.*, vol. 181, pp. 186–221, 2002.

[44] P. Monk and G. Richter, "A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media," *Journal of Scientific Computing*, vol. 22-23, no. 1-3, pp. 443–477, 2005, ISSN: 0885-7474.

[45] E. T. Chung and B. Engquist, "Optimal discontinuous Galerkin methods for wave propagation," *SIAM Journal on Numerical Analysis*, vol. 44, no. 5, pp. 2131–2158, 2006.

[46] E. T. Chung and B. Engquist, "Optimal discontinuous Galerkin methods for the acoustic wave equation in higher dimensions," *SIAM Journal on Numerical Analysis*, vol. 47, no. 5, pp. 3820–3848, 2009.

[47] J. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods*, ser. Texts in Applied Mathematics 54. New York: Springer-Verlag, 2008.

[48] T. Warburton, "A low-storage curvilinear discontinuous Galerkin method for wave problems," *SIAM Journal on Scientific Computing*, vol. 35, no. 4, A1987–A2012, 2013.

[49] X. Meng, C.-W. Shu, and B. Wu, "Optimal error estimates for discontinuous Galerkin methods based on upwind-biased fluxes for linear hyperbolic equations," *Mathematics of Computation*, vol. 85, no. 299, pp. 1225–1261, 2016.

[50] B. Riviere and M. Wheeler, "Discontinuous finite element methods for acoustic and elastic wave problems. part i: Semidiscrete error estimates," *Contemporary Mathematics*, vol. 329, pp. 271–282, 2003.

[51] M. J. Grote, A. Schneebeli, and D. Schötzau, "Discontinuous Galerkin finite element method for the wave equation," *SIAM Journal on Numerical Analysis*, vol. 44, no. 6, pp. 2408–2431, 2006, ISSN: 00361429.

[52] C.-S. Chou, C.-W. Shu, and Y. Xing, "Optimal energy conserving local discontinuous Galerkin methods for second-order wave equation in heterogeneous media," *Journal of Computational Physics*, vol. 272, pp. 88–107, 2014, ISSN: 0021-9991.

[53] D. Appelö and T. Hagstrom, "A new discontinuous Galerkin formulation for wave equations in second order form," *SIAM Journal On Numerical Analysis*, vol. 53, no. 6, pp. 2705–2726, 2015.

[54] K. Mattsson, "Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients," *Journal of Scientific Computing*, vol. 51, no. 3, pp. 650–682, Jun. 2012.

[55] K. Mattsson and J. Nordström, "Summation by parts operators for finite difference approximations of second derivatives," *J. Comput. Phys.*, vol. 199, pp. 503–540, 2004.

[56] K. Mattsson, F. Ham, and G. Iaccarino, "Stable and accurate wave–propagation in discontinuous media," *J. Comput. Phys.*, vol. 227, pp. 8753–8767, 2008.

[57] K. Virta and K. Mattsson, "Acoustic wave propagation in complicated geometries and heterogeneous media," *Journal of Scientific Computing*, vol. 61, no. 1, pp. 90–118, 2014.

[58] M. Almquist, S. Wang, and J. Werpers, "Order-preserving interpolation for summation-by-parts operators at nonconforming grid interfaces," *SIAM J. Sci. Comput.*, vol. 41, A1201–A1227, 2019.

[59] S. Wang, "An improved high order finite difference method for non–conforming grid interfaces for the wave equation," *J. Sci. Comput.*, vol. 77, pp. 775–792, 2018.

[60] S. Wang, K. Virta, and G. Kreiss, "High order finite difference methods for the wave equation with non-conforming grid interfaces," *Journal of Scientific Computing*, vol. 68, no. 3, pp. 1002–1028, 2016.

[61] G. Cohen, *Higher-order numerical methods for transient wave equations.* Springer Science & Business Media, 2001.

[62] H. Li and X. Zhang, “Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations,” *Journal of Scientific Computing*, vol. 82, no. 2, pp. 1–39, 2020.

[63] M. F. Wheeler, “A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations,” *SIAM Journal on Numerical Analysis*, vol. 10, no. 4, pp. 723–759, 1973.

[64] P. H. Sammon, “Convergence estimates for semidiscrete parabolic equation approximations,” *SIAM Journal on Numerical Analysis*, vol. 19, no. 1, pp. 68–92, 1982.

[65] T. Dupont, “L2-estimates for Galerkin methods for second order hyperbolic equations,” *SIAM journal on numerical analysis*, vol. 10, no. 5, pp. 880–889, 1973.

[66] M. S. Gockenbach, *Understanding and implementing the finite element method*. Siam, 2006, vol. 97.

[67] R. J. Plemmons, “M-matrix characterizations. I-nonsingular M-matrices,” *Numer. Anal. Appl.*, vol. 18, no. 2, pp. 175–188, 1977.

[68] J. Xu and L. Zikatanov, “A monotone finite element scheme for convection-diffusion equations,” *Math. Comp.*, vol. 68, no. 228, pp. 1429–1446, 1999.

[69] J. Karátson and S. Korotov, “Discrete maximum principles for fem solutions of some nonlinear elliptic interface problems,” *Int. J. Numer. Anal. Model*, vol. 6, no. 1, pp. 1–16, 2009.

[70] J. H. Bramble, “Fourth-order finite difference analogues of the Dirichlet problem for Poisson’s equation in three and four dimensions,” *Math. Comp.*, vol. 17, no. 83, pp. 217–222, 1963.

[71] S. K. Lele, “Compact finite difference schemes with spectral-like resolution,” *Journal of Computational Physics*, vol. 103, no. 1, pp. 16–42, 1992.

[72] J. Bramble and B. Hubbard, “On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type,” *J. Math. and Phys.*, vol. 43, no. 1-4, pp. 117–132, 1964.

[73] J. H. Bramble and B. E. Hubbard, “New monotone type approximations for elliptic problems,” *Math. Comp.*, vol. 18, no. 87, pp. 349–367, 1964.

[74] T. Vejchodskỳ, “Angle conditions for discrete maximum principles in higher-order FEM,” in *Numerical Mathematics and Advanced Applications 2009*, Springer, 2010, pp. 901–909.

[75] T. Vejchodskỳ and P. Šolín, "Discrete maximum principle for higher-order finite elements in 1D," *Math. Comp.*, vol. 76, no. 260, pp. 1833–1846, 2007.

[76] A. Drăgănescu, T. Dupont, and L. Scott, "Failure of the discrete maximum principle for an elliptic finite element problem," *Math. Comp.*, vol. 74, no. 249, pp. 1–23, 2005.

[77] Z. Li and K. Ito, "Maximum principle preserving schemes for interface problems with discontinuous coefficients," *SIAM J. Sci. Comput.*, vol. 23, no. 1, pp. 339–361, 2001.

[78] T. Vejchodskỳ and P. Šolín, "Discrete maximum principle for a 1D problem with piecewise-constant coefficients solved by hp-FEM," *J. Numer. Math.*, vol. 15, no. 3, pp. 233–243, 2007.

[79] G. Payette, K. Nakshatrala, and J. Reddy, "On the performance of high-order finite elements with respect to maximum principles and the nonnegative constraint for diffusion-type equations," *Internat. J. Numer. Methods Engrg.*, vol. 91, no. 7, pp. 742–771, 2012.

[80] S. Korotov, M. Křížek, and J. Šolc, "On a discrete maximum principle for linear fe solutions of elliptic problems with a nondiagonal coefficient matrix," in *International Conference on Numerical Analysis and Its Applications*, Springer, 2008, pp. 384–391.

[81] D. Kuzmin, M. J. Shashkov, and D. Svyatskiy, "A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems," *J. Comput. Phys.*, vol. 228, no. 9, pp. 3448–3463, 2009.

[82] E. Bohl and J. Lorenz, "Inverse monotonicity and difference schemes of higher order. a summary for two-point boundary value problems," *Aequationes Math.*, vol. 19, no. 1, pp. 1–36, 1979.

[83] F. Bouchon, "Monotonicity of some perturbations of irreducibly diagonally dominant m-matrices," *Numer. Math.*, vol. 105, no. 4, pp. 591–601, 2007.

[84] T. Tang and J. Yang, "Implicit-explicit scheme for the Allen-Cahn equation preserves the maximum principle," *J. Comput. Math*, vol. 34, no. 5, pp. 471–481, 2016.

[85] J. Shen, T. Tang, and J. Yang, "On the maximum principle preserving schemes for the generalized Allen–Cahn equation," *Commun. Math. Sci*, vol. 14, no. 6, pp. 1517–1534, 2016.

[86] J. Xu, Y. Li, S. Wu, and A. Bousquet, "On the stability and accuracy of partially and fully implicit schemes for phase field modeling," *Comput. Methods Appl. Mech. Engrg.*, vol. 345, pp. 826–853, 2019.

[87] I. Christie and C. Hall, “The maximum principle for bilinear elements,” *Internat. J. Numer. Methods Engrg.*, vol. 20, no. 3, pp. 549–553, 1984.

[88] T. Qin and C.-W. Shu, “Implicit positivity-preserving high-order discontinuous galerkin methods for conservation laws,” *SIAM Journal on Scientific Computing*, vol. 40, no. 1, A81–A107, 2018.

[89] R. J. LeVeque, *Numerical Methods for Conservation Laws*. Birkhauser Basel, 1992.

[90] R. Sanders, “A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws,” *Mathematics of Computation*, vol. 51, no. 184, pp. 535–558, 1988.

[91] X. Zhang and C.-W. Shu, “A genuinely high order total variation diminishing scheme for one-dimensional scalar conservation laws,” *SIAM Journal on Numerical Analysis*, vol. 48, no. 2, pp. 772–795, 2010.

[92] D. Levy and E. Tadmor, “Non-oscillatory central schemes for the incompressible 2-D Euler equations,” *Mathematical Research Letters*, vol. 4, pp. 321–340, 1997.

[93] X. Zhang and C.-W. Shu, “Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 467, no. 2134, pp. 2752–2776, 2011.

[94] X.-D. Liu and S. Osher, “Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I,” *SIAM Journal on Numerical Analysis*, vol. 33, no. 2, pp. 760–779, 1996.

[95] X. Zhang, Y. Xia, and C.-W. Shu, “Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes,” *Journal of Scientific Computing*, vol. 50, no. 1, pp. 29–62, 2012.

[96] X. Zhang, “On positivity-preserving high order discontinuous galerkin schemes for compressible navier–stokes equations,” *Journal of Computational Physics*, vol. 328, pp. 301–343, 2017.

[97] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.

[98] Z. Xu, “Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: One-dimensional scalar problem,” *Mathematics of Computation*, vol. 83, no. 289, pp. 2213–2238, 2014.

[99] T. Xiong, J.-M. Qiu, and Z. Xu, "High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, A583–A608, 2015.

[100] Y. Liu, Y. Cheng, and C.-W. Shu, "A simple bound-preserving sweeping technique for conservative numerical approximations," *Journal of Scientific Computing*, vol. 73, no. 2-3, pp. 1028–1071, 2017.

[101] B. Cockburn and C.-W. Shu, "Nonlinearly stable compact schemes for shock calculations," *SIAM Journal on Numerical Analysis*, vol. 31, no. 3, pp. 607–627, 1994.

[102] X. Zhang, Y. Liu, and C.-W. Shu, "Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations," *SIAM Journal on Scientific Computing*, vol. 34, no. 2, A627–A658, 2012.

[103] Y. Zhang, X. Zhang, and C.-W. Shu, "Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection–diffusion equations on triangular meshes," *Journal of Computational Physics*, vol. 234, pp. 295–316, 2013.

[104] Z. Chen, H. Huang, and J. Yan, "Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes," *Journal of Computational Physics*, vol. 308, pp. 198–217, 2016.

[105] Z. Sun, J. A. Carrillo, and C.-W. Shu, "A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials," *Journal of Computational Physics*, vol. 352, pp. 76–104, 2018.

[106] S. Srinivasan, J. Poggie, and X. Zhang, "A positivity-preserving high order discontinuous galerkin scheme for convection–diffusion equations," *Journal of Computational Physics*, vol. 366, pp. 120–143, 2018.

[107] T. Qin and C.-W. Shu, "Implicit positivity-preserving high order discontinuous Galerkin methods for conservation laws," *to appear in SIAM Journal on Scientific Computing*,

[108] J. Hu, R. Shu, and X. Zhang, "Asymptotic-preserving and positivity-preserving implicit-explicit schemes for the stiff BGK equation," *to appear in SIAM Journal on Numerical Analysis*, 2017.

[109] M. H. Carpenter, D. Gottlieb, and S. Abarbanel, "The stability of numerical boundary treatments for compact high-order finite-difference schemes," *Journal of Computational Physics*, vol. 108, no. 2, pp. 272–295, 1993.

[110] A. I. Tolstykh, *High accuracy non-centered compact difference schemes for fluid dynamics applications*. World Scientific, 1994, vol. 21.

[111] W. Spotz and G. Carey, “High-order compact finite difference methods,” in *Preliminary Proceedings International Conference on Spectral and High Order Methods, Houston, TX*, 1995, pp. 397–408.

[112] A. I. Tolstykh and M. V. Lipavskii, “On performance of methods with third-and fifth-order compact upwind differencing,” *Journal of Computational Physics*, vol. 140, no. 2, pp. 205–232, 1998.

[113] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

PUBLICATIONS

- **Hao Li**, Shusen Xie and Xiangxiong Zhang. A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations, *SIAM Journal on Numerical Analysis*, 56(6), 3308-3345.

- **Hao Li** and Xiangxiong Zhang. Superconvergence of C^0-Q^k finite element method for elliptic equations with approximated coefficients, *Journal of Scientific Computing* 82, 1 (2020).

- **Hao Li** and Xiangxiong Zhang. Superconvergence of high order finite difference schemes based on variation formulation for elliptic equations, *Journal of Scientific Computing* 82, 36 (2020).

- **Hao Li** and Xiangxiong Zhang. On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method, *Numerische Mathematik* 145, 437-472 (2020).

- **Hao Li**, Daniel Appelö, and Xiangxiong Zhang. Accuracy of spectral element method for wave, parabolic and Schrödinger equations, arXiv:2103.00400.