Recent Progress on Q^k Spectral Element Method: Accuracy, Monotonicity and Applications

Xiangxiong Zhang[0000−0002−1090−7189]

Abstract We first briefly review some recently proven new results about Q^k spectral element method for second order linear PDEs, including its accuracy as a finite difference method in ℓ^2-norm and monotonicity. Then we discuss some miscellaneous extensions including the accuracy for Helmholtz equation and applications of monotone discrete Laplacian. In particular, the Q^2 spectral element method gives a fourth order accurate monotone discrete Laplacian, with which one can obtain explicit convergence rates of Picard and Newton iterations for solving a special semilinear PDE.

Key words: Spectral element method, monotone discrete Laplacian, Gauss-Lobatto quadrature, superconvergence, Helmholtz equation, Picard iteration, Newton's method

1 Introduction

In the vast computational science and engineering literature, spectral element methods usually refer to various finite element methods using high order polynomial basis. In this paper, Q^k spectral element method specifically refers to the classical continuous finite element method for second order partial differential equations (PDEs) with Q^k Lagrangian basis on rectangular meshes, implemented by $(k+1)$-point Gauss-Lobatto quadrature for all integrals.

Such a scheme can be regarded as a finite difference scheme defined at all the Gauss-Lobatto quadrature points. For example, consider solving $-u''(x) = f(x)$ on $x \in (0,1)$, and consider a uniform grid $x_i = ih, i = 0, 1, \cdots, n+1$ with mesh size $h = \frac{1}{n+1}$, the explicit equivalent finite difference form of Q^2 spectral element method on the uniform mesh with intervals $I_j = [x_{2j}, x_{2j+2}]$ can be written as

Xiangxiong Zhang
Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA. e-mail: zhan1966@purdue.edu

$$\frac{-u_{i-1}+2u_i-2u_{i+1}}{h^2}=f_i, \quad i \text{ is odd}, \tag{1a}$$

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

Since 1960s, it has been very well known that finite element methods with suitable quadrature are finite difference methods. What makes the scheme (1) in particular much more interesting is the fact that it is a high order accurate monotone scheme [25]. Let $-\Delta_h$ be the discrete Laplacian matrix, e.g., $-\Delta_h$ is the tridiagonal $\frac{1}{h^2}(-1,2,-1)$ matrix for the simplest centered finite difference approximation $-u''(x_i) \approx \frac{-u_{i-1}+2u_i-2u_{i+1}}{h^2}$. Then a discrete Laplacian scheme or the matrix $-\Delta_h$ is called *monotone*, if the inverse matrix has non-negative entries $(-\Delta_h)^{-1} \geq 0$, where the inequality is for each entry of the matrix.

The second order finite difference is monotone due to the well known fact that the tridiagonal $\frac{1}{h^2}(-1,2,-1)$ matrix is an M-matrix [41, 11]. The most useful generalization of this property is that linear finite element method forms an M-matrix for scalar variable coefficient operator $-\nabla \cdot (a(\boldsymbol{x})\nabla u)$ on triangular meshes under a mild mesh angle constraints [46].

In the past decade, there have been some efforts pursuing implicit high order accurate bound-preserving and positivity-preserving schemes for convection diffusion problems, which are naturally related to monotone schemes. Though monotonicity is in general lost for high order finite element method on unstructured meshes [18], it is still possible to have monotone high order finite element method on structured meshes [36].

In particular, it has been proven in [25] that the Q^2 spectral element method (1) is unconditionally monotone for the Laplacian in two dimensions, and the same scheme is monotone under mesh constraints for a variable coefficient operator $-\nabla \cdot (a(\boldsymbol{x})\nabla u) + c(\boldsymbol{x})u$ with a scalar coefficient $a(\boldsymbol{x})$. Thus a convenient high order monotone scheme to replace the second order finite difference would be (1).

The rest of the paper is organized as follows. We first briefly review some recent results for this scheme including accuracy, monotonicity and GPU implementation in Section 2, then we discuss some miscellaneous extensions and applications of these results including ℓ^2-norm *a priori* error estimate of Q^k spectral element method for the Helmholtz equation in Section 3 and applications of monotonicity including ℓ^∞ estimate of the discrete Laplacian of Q^2 spectral element method and explicit convergence rates of Picard and Newton iterations for solving a semi-linear PDE in Section 4. Concluding remarks are given in Section 5.

2 Review of recent progress: accuracy, monotonicity and implementation

2.1 Accuracy as a finite difference scheme

Since the Q^2 spectral element method (1) is a monotone scheme, it is interesting to understand its order of accuracy as a finite difference scheme. In one dimension, using Green's function [13], it is straightforward to show (1) is a fourth order finite difference scheme in ℓ^2-norm for very smooth solutions. In multiple dimensions, the Q^k $(k \geq 2)$ spectral element method is $(k+2)$-th order accurate as a finite difference scheme in ℓ^2-norm, but it is quite technical to establish a rigorous *a priori* error estimate to the desired $(k+2)$-th order.

The main difficulty is related to the fact that the $(k+2)$-th order convergence for function values at all Gauss-Lobatto quadrature points is a superconvergence result for Q^k element $(k \geq 2)$. The standard error estimate only states that it is $(k+1)$-th order accurate for function values in L^2-norm [38]. For Q^k finite element method without using any quadrature, such a superconvergence result has been proven since 1980s by two different approaches: one is via M-type projection in [6, 7, 8] and the other one is by analyzing superconvergence of bilinear forms [32, 31].

It may seem straightforward to obtain a rigorous *a priori* error estimate for the $(k+2)$-th order convergence by combining existing superconvergence theory and quadrature error estimates, since $(k+1)$-point Gauss-Lobatto quadrature is $(k+2)$-th order accurate. However, as explained in [27], a straightforward quadrature error estimate by standard Bramble-Hilbert Lemma is not enough, and it is necessary to derive a sharp quadrature error estimate via counting error cancellations, which should be combined with M-type projection to show the desired results.

It is easy to check that the finite difference scheme (1) is only a second order approximation to the second order derivative, but it is a fourth order accurate scheme for solving second order PDEs. The rigorous *a priori* error estimate for the $(k+2)$-th order convergence of Q^k $(k \geq 2)$ spectral elememt method as a finite difference scheme in ℓ^2-norm has been established for elliptic equations with Dirichlet boundary conditions in [27], and for wave, parabolic and Schrödinger equations in [23]. See [22, Section 2.8] for the discussion for Neumann boundary. In Section 3, we will discuss the *a priori* error estimate for solving the Helmholtz equation.

2.2 Provable monotonicity results

Even though arbitrarily high order finite element method can be proven monotone for Laplacian in one dimension using the Green's function [45], it is simply not true that Q^k spectral element method is monotone for Laplacian in multiple dimensions for any k. In two dimensions, Q^k spectral element method with $k \geq 9$ is no longer unconditionally monotone for Laplacian, as indicated by numerical results in [10].

To be more specific, for solving a one-dimensional Poisson equation $-u'' = f$ on the domain $(0, 1)$ with homogeneous Dirichlet boundary condition, consider the classical continuous finite element method using P^k polynomial basis on a uniform mesh with all the integrals approximated by (k + 1)-point Gauss-Lobatto quadrature, which is equivalent to a finite difference scheme at all Gauss-Lobatto points excluding two domain boundary points. The finite difference scheme can be written as $S\mathbf{u} = M\mathbf{f}$ or $H\mathbf{u} = \mathbf{f}$, where S is the stiffness matrix, $M \geq 0$ is the diagonal mass matrix and $H = M^{-1}S$. The result in [45] simply implies that $S^{-1} \geq 0$ and $H^{-1} \geq 0$ for any polynomial degree k.

In two dimensions, the Q^k spectral element method on a uniform mesh has a stiffness matrix $S \otimes M + M \otimes S = (M \otimes M)(H \otimes I + \otimes H)$. The monotonicity in two dimensions is equivalent to $(H \otimes I + I \otimes H)^{-1} \geq 0$. It may seem possible that $H^{-1} \geq 0$ could imply $(H \otimes I + I \otimes H)^{-1} \geq 0$, but this is simply not true for $k \geq 9$ as shown in [10], i.e., Q^k spectral element method is not monotone for $k \geq 9$ in two dimensions.

For $k = 2$, it is proven that $H \otimes I + I \otimes H$ is a product of two M-matrices thus still monotone [25]. For $k = 3$, it is proven that $H \otimes I + I \otimes H$ is a product of four M-matrices thus still monotone [10]. Both the proof in [25] and [10] can be extended to three dimensions. For proving that a matrix is a product of two M-matrices, a convenient sufficient condition is due to Lorenz [36]. See [25, 11, 10] for details of how to use Lorenz's conditions.

For three dimensional Laplacian, the monotonicity of Q^k spectral element method is equivalent to the monotonicity of the matrix $(H \otimes I \otimes I + I \otimes H \otimes I + I \otimes I \otimes H)^{-1}$. As shown by the numerical tests in Figure 1, Q^k spectral element method is no longer monotone for $k \geq 4$ in three dimensions.

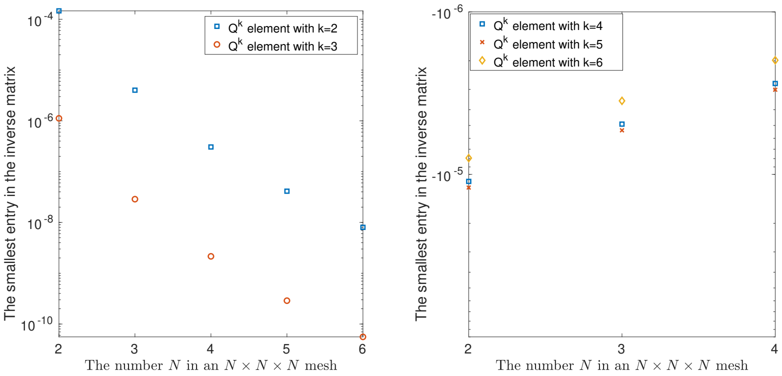

Fig. 1 The smallest entry in the matrix $(H \otimes I \otimes I + I \otimes H \otimes I + I \otimes I \otimes H)^{-1}$ for Q^k spectral element method on a uniform $N \times N \times N$ mesh for solving $-\Delta u = f$ in three dimensions with homogeneous Dirichlet boundary on a cube. The figure on the right suggests that Q^k spectral element method is not monotone for $k \geq 4$ in three dimensions.

One of the motivating applications for studying monotonicity is to construct implicit bound-preserving schemes for solving the heat equation $u_t = u_{xx}$. If using the scheme (1), the semi-discrete ODE is given as

$$u_i'(t) = -\frac{-u_{i-1} + 2u_i - 2u_{i+1}}{h^2}, \quad i \text{ is odd}, \tag{2a}$$

$$u_i'(t) = -\frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2}, \quad i \text{ is even}. \tag{2b}$$

Due to the negative sign in front of $u_{i\pm 2}$, the exact solution to the initial value problem of this ODE is simply not bound-preserving. This is a counter-intuitive result since the solution to the ODE (2) is a more accurate approximation to the heat equation than the semi-discrete ODE from centered difference

$$u_i'(t) = -\frac{-u_{i-1} + 2u_i - 2u_{i+1}}{h^2}, \quad \forall i,$$

which has a bound-preserving solution due to the ODE theory and the fact that the right hand side matrix has negative diagonal entries, non-negative off-diagonal entries and it is diagonally dominant. For instance, if one solves (2) by matrix exponential method, then it is fourth order accurate in space and exact in time, but it is not strictly bound-preserving especially if time step is very small. For backward Euler time discretization of (2), a sufficient condition to ensure monotonicity thus bound-preserving property is to require $\Delta t > \frac{2}{3}h^2$, which is practical since usually one would not use a small time step like $\Delta t = \mathcal{O}(h^2)$ for implicit time stepping.

We summarize a few useful conclusions and observations as follows:

1. For the discrete Laplacian, Q^2 and Q^3 spectral element methods are unconditionally monotone on uniform meshes. The proof of the two-dimensional case is given in [25] for Q^2, and in [10] for Q^3. The monotonicity proof can be easily extended to three dimensions and Neumann boundary conditions.
2. For the discrete Laplacian, the Lagrangian P^2 finite element method on a regular structured triangular mesh is unconditionally monotone [36]. This scheme is also a fourth order finite difference scheme [10]. On unstructured triangular meshes, the monotonicity does not hold [18].
3. For Laplacian, Q^k element method are not unconditionally monotone for $k \geq 9$ in two dimensions, and not unconditionally monotone for $k \geq 4$ in three dimensions. On the other hand, it is possible that very high order schemes can be still monotone if using different implementations, e.g., different quadrature rules.
4. It is possible to construct traditional finite difference schemes that are fourth order accurate and monotone [3, 4, 24, 29], however only for Dirichlet boundary conditions. It is difficult to extend these schemes to Neumann boundary conditions.
5. All the known high order accurate monotone discrete Laplacian schemes include:
    - The classical fourth order compact finite difference schemes (see Appendix in [29]), e.g., 9-point discrete Laplacian: $-\Delta_h$ is an M-matrix in two-dimensions.

- A fourth order accurate traditional finite difference scheme [3, 4]: $-\Delta_h$ is a product of two M-matrices in two dimensions.
- A fourth order accurate finite difference by P^2 spectral element method: $-\Delta_h$ is a product of two M-matrices in two dimensions.
- A fourth order accurate finite difference by Q^2 finite element method: $-\Delta_h$ is a product of two M-matrices in two dimensions.
- A fifth order accurate finite difference by Q^3 finite element method: $-\Delta_h$ is a product of fourth M-matrices in two dimensions.

6. The monotonicity of Q^2 spectral element method on quasi-uniform meshes is given in [11].
7. For scalar variable coefficient elliptic operators $-\nabla \cdot (a(\boldsymbol{x})\nabla u) + c(\boldsymbol{x})u$, monotonicity of Q^2 spectral element method still holds under mesh constraints [25]. See extensions and applications to Allen-Cahn equation in [43], Keller-Segel equation in [19], Fokker-Planck equation in [33], and compressible Navier-Stokes equations in [34].
8. For a matrix coefficient elliptic operator, even the Q^1 finite element method needs a very stringent mesh constraint for monotonicity [28]. Thus we do not expect a convenient monotonicity result for the Q^2 spectral element method.
9. For solving $u_t = \Delta u$, the semi-discrete ODE by Q^2 spectral element method in space is not bound-preserving. With backward Euler time discretization, Q^2 spectral element method is monotone thus bound-preserving if $\Delta t > \frac{2}{3}h^2$ in two dimensions. It is an open problem to prove monotonicity of higher order implicit time stepping methods.

2.3 Simple and efficient implementation on GPU

For three dimensional Laplacian, the Q^k spectral element method on rectangular meshes has the structure $S \otimes M \otimes M + M \otimes S \otimes M + M \otimes M \otimes S$. Since 1980s, it is well known that such a tensor structure can be used to invert the stiffness matrix directly via its eigenvalue decomposition, yet with a complexity $O(N^{\frac{4}{3}})$ where N is the total degree of freedoms (DoFs). Modern graphic processing units (GPU) are designed to compute tensor multiplications efficiently. As of 2023, softwares such as MATLAB and JAX in Python allow one to implement tensor matrix multiplications on GPU without any low level coding, e.g., the GPU code is the same as CPU code. One only needs to write a few lines in these softwares to implement such a direct solver. It is demonstrated in [35] that the MATLAB double precision implementation achieves 0.8 second for inverting the 3D discrete Laplacian for Q^k spectral element method with one billion DoFs (e.g., a 1000^3 grid) on one Nividia A100 80G GPU card. At the time of writing this paper, JAX in Python for double precision computation can only handle a smaller problem like DoFs being 800^3 on the same A100 GPU card.

3 Accuracy for the Helmholtz equation

For simplicity, we only consider the Helmholtz equation on a square domain $\Omega = (0,1)\times(0,1)$ with a homogeneous Robin boundary condition,

$$-\Delta u(\boldsymbol{x}) - \omega^2 u(\boldsymbol{x}) = f(\boldsymbol{x}), \qquad \boldsymbol{x}\in\Omega, \tag{3a}$$

$$\frac{\partial u}{\partial n} - \mathbb{i}\omega u = 0, \qquad \boldsymbol{x}\in\partial\Omega. \tag{3b}$$

There is no essential difficulty to extend the main arguments in this section to three dimensions. The main results in this section can also be easily extended to a more general equation like $-\nabla\cdot(a(\boldsymbol{x})\nabla u) - \omega^2(\boldsymbol{x})u = f$ following the arguments in [27] and [22, Section 2.8].

The variational form of (3) is to seek $u \in H^1(\Omega)$ satisfying

$$A(u,v) := \iint_\Omega [\nabla u\cdot\nabla\bar{v} - \omega^2 u\bar{v}]d\boldsymbol{x} - \mathbb{i}\int_{\partial\Omega}\omega u\bar{v}ds = \iint_\Omega f\bar{v}d\boldsymbol{x}, \quad \forall v\in H^1(\Omega).$$

3.1 Notation

We will use the same notation as in [27]:

- 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 paper 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.
- $V_0^h = \{v_h\in V^h : v_h = 0 \quad \text{on} \quad \partial\Omega\}$.
- The norm and seminorms for $W^{k,p}(\Omega)$ and $1\le p < +\infty$:

$$\|u\|_{k,p,\Omega} = \left(\sum_{i+j\le k}\iint_\Omega |\partial_x^i\partial_y^j u(x,y)|^p dxdy\right)^{1/p},$$

$$|u|_{k,p,\Omega} = \left(\sum_{i+j=k}\iint_\Omega |\partial_x^i\partial_y^j u(x,y)|^p dxdy\right)^{1/p},$$

$$[u]_{k,p,\Omega} = \left(\iint_\Omega |\partial_x^k u(x,y)|^p dxdy + \iint_\Omega |\partial_y^k u(x,y)|^p dxdy \right)^{1/p}.$$

- 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**

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

- 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 $\|u\|_{2,Z_0}$ and $\|u\|_{\infty,Z_0}$ denote the discrete 2-norm and the maximum norm over Z_0 respectively:

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

- $(f,v)_e = \iint_e f\bar{v}\, dxdy, \quad (f,v) = \iint_\Omega f\bar{v}\, dxdy = \sum_e (f,v)_e.$
- $\langle f,v\rangle_h = \iint_\Omega f\bar{v}\, dx^h dy^h$ 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. **Notice that we use dx^h to denote that quadrature is used.**

3.2 Q^k spectral element method

We consider a uniform rectangular mesh and a continous piecewise Q^k polynomial space V^h. With notation above, we define a bilinear form with quadrature as

$$A_h(u,v) := \langle \nabla u, \nabla v\rangle_h - \omega^2 \langle u,v\rangle_h - \mathbb{i}\omega\langle u,v\rangle_{\partial\Omega,h} = \iint_\Omega [\nabla u \cdot \nabla \bar{v} - \omega^2 u\bar{v}] d\boldsymbol{x}^h - \mathbb{i} \int_{\partial\Omega} \omega u\bar{v} ds^h,$$

where $d\boldsymbol{x}^h$ and ds^h denote that $(k+1)$-point Gauss-Lobatto quadrature is used for each finite element cell or interval. The exact solution $u \in H^1(\Omega)$ satisfies

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

The Q^k spectral element method solution $u_h \in V^h$ satisfies

$$A_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V^h \subset H^1(\Omega).$$

3.3 Quadrature error estimates

We first need a sharp quadrature estimate. Following the proof in [22, Theorem 2.8.1], we have

Theorem 1 *For $k \geq 2$, assume $u \in H^{k+3}(\Omega)$ and $\frac{\partial u}{\partial n} - \mathbb{i}\omega u = 0, \forall \boldsymbol{x} \in \partial\Omega$, then*

$$A(u, v_h) - A_h(u, v_h) = O(h^{k+2})\|u\|_{k+3}\|v_h\|_2 + O(h^{k+2})\omega^2\|u\|_{k+2}\|v_h\|_2.$$

Remark 1 Without the assumption on the boundary condition $\frac{\partial u}{\partial n} - \mathbb{i}\omega u = 0, \forall \boldsymbol{x} \in \partial\Omega$, then following Theorem 3.7 in [27], one can only prove $O(h^{k+1.5})$. The half order loss is due to an extra boundary term, which can be cancelled if using boundary conditions of u.

By Theorem 3.3 in [27], a standard quadrature estimate is given by

Theorem 2 *If $f \in H^{k+2}(\Omega)$ with $k \geq 2$, then $(f, v_h) - \langle f, v_h\rangle_h = O(h^{k+2})\|f\|_{k+2}\|v_h\|_2$.*

3.4 M-type projection

The M-type projection in [7, 8] is a convenient tool for discussing the superconvergence of function values at Gauss-Lobatto points. We refer to [26, 27] for its detailed definition. Given a smooth function $u(x, y)$, let $u_p(x, y)$ be the M-type Q^k projection of $u(x, y)$, then $u_p \in V^h$ and $u - u_p$ has smaller errors at Gauss-Lobatto points, e.g., Theorem 4.2 in [27] is given as

Theorem 3 *For $k \geq 2$,*

$$\|u - u_p\|_{2,Z_0} = O(h^{k+2})\|u\|_{k+2}, \quad \forall u \in H^{k+2}(\Omega).$$

$$\|u - u_p\|_{\infty,Z_0} = O(h^{k+2})\|u\|_{k+2,\infty}, \quad \forall u \in W^{k+2,\infty}(\Omega).$$

Thus, in order to prove Q^k spectral element method is a $(k+2)$-th order accurate finite difference scheme for smooth solutions in ℓ^2-norm, it suffices to prove $\|u_h - u_p\|_{2,Z_0} = O(h^{k+2})$. Next, we need the superconvergence of the bilinear form:

Theorem 4 *For $k \geq 2$, assume $u \in H^{k+4}(\Omega)$ and $\frac{\partial u}{\partial n} - \mathbb{i}\omega u = 0, \forall \boldsymbol{x} \in \partial\Omega$, then*

$$A_h(u - u_p, v_h) = O(h^{k+2})(\|u\|_{k+3} + \omega\|u\|_{k+4} + \omega^2\|u\|_{k+2})\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. By Lemma 4.5-4.8 in [27], we first have

$$\langle\nabla(u - u_p), \nabla v_h\rangle_h - \omega^2\langle(u - u_p), v_h\rangle_h = O(h^{k+2})(\|u\|_{k+3} + \omega^2\|u\|_{k+2})\|v_h\|_2.$$

For the line integral, with the properties of M-type projection (see [26, 27]), following the proof of Lemma 2.3.6 in [22], we can get

$$|\mathbb{i}\omega\langle\nabla(u-u_p),\nabla v_h\rangle_{\partial\Omega,h}| = O(h^{k+3})\omega\|u\|_{k+3,\partial\Omega}\|v_h\|_{2,\partial\Omega} = O(h^{k+2.5})\omega\|u\|_{k+4,\Omega}\|v_h\|_{2,\Omega}.$$

Notice that the arguments in [22, Lemma 2.3.6] do not work for the case $k = 2$ due to the fact that 3-point Gauss-Lobatto quadrature is exact only for polynomials of degree 3 but not polynomials of degree $k + 2 = 4$. In order to establish the desired result for $k = 2$, one possibility is to apply the discrete integration by part in the proof of [22, Theorem 2.8.1] to use $\langle\nabla(u - u_p), \nabla v_h\rangle_h$ to generate a boundary line integral term, which can be used to reduce the error term $\langle\nabla(u-u_p), \nabla v_h\rangle_{\partial\Omega,h}$. □

3.5 Dual problem and regularity for the Helmholtz equation

Define $\theta_h = u_h - u_p$, then $\theta_h \in V^h \subset H^1(\Omega)$. One critical step in this section is to consider the following dual problem: seek $w \in H^1(\Omega)$ satisfying

$$A^*(w, v) := A(v, w) = (v, \theta_h), \quad \forall v \in H^1(\Omega).$$

In other words, w satisfies the Helmholtz equation $-\Delta w - \omega^2 w = \theta_h$ with the homogeneous Robin boundary condition.

Consider a finite element solution for the dual problem: seek $w_h \in V^h$ satisfying

$$A_h^*(w_h, v_h) := A_h(v_h, w_h) = (v_h, \theta_h), \quad \forall v_h \in V^h.$$

We first need to establish a convergence result for finite element scheme solving Helmholtz equation: $\|w-w_h\| \leq Ch\|w\|_2$. Since the bilinear form for the Helmholtz equation is indefinite, its variational problem satisfies a Garding inequality [42], for which Galerkin methods are asymptotically quasi-optimal, e.g., convergence can be proven if assuming h is small enough. In particular, assuming $\omega^2 h \ll 1$, it satisfies standard estimate

$$\|w - w_h\|_1 \leq C(\omega) \min_{v_h\in V^h} \|w - v_h\|_1$$

which can be found in [14, 15]. For simplicity, we do not discuss the specific dependency of $C(\omega)$ on ω and polynomial degree k, which is of interest for high frequency problems in the pre-asymptotic region. We refer to [20, 21, 39, 40, 5, 2, 17] and references therenin for pre-asymptotic analysis.

We also need the regularity result for solving Helmholtz equation $-\Delta w - \omega^2 w = \theta_h$. In particular, the regularity result in [16] states that

$$\|w\|_2 \leq C(\omega^3 + \omega^{-1})\|\theta_h\|_0.$$

See also [39, 40] for a regularity result, and [12] for a sharp estimate of the regularity coefficients for high frequencies.

Then by the proof of Theorem 5.3 in [27], we obtain a standard estimate for w_h:

Theorem 5 *Assume h is small enough, then*

$$\|w_h\|_2 \leq C(\omega)\|w\|_2 \leq C(\omega)(\omega^3 + \omega^{-1})\|\theta_h\|_0.$$

3.6 The main result: superconvergence of function values

Now we can put all the results together in this section to obtain

Theorem 6 *Assume the exact solution $u(x, y) \in H^{k+4}(\Omega)$ and satisfies the homogeneoug Robin boundary condition, $f(x, y) \in H^{k+2}(\Omega)$. Assume h is sufficiently small. Then 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:*

$$\|u_h - u\|_{2,Z_0} = O(h^{k+2})(\|u\|_{k+3} + \omega^2\|u\|_{k+2} + \omega\|u\|_{k+4} + \|f\|_{k+2})C(\omega)(\omega^3 + \omega^{-1}).$$

Proof. Recall $\theta_h = u_h - u_p \in V^h$, with Theorem 1, Theorem 2, Theorem 4 and Theorem 5, we have

$$\begin{aligned}
&\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h - u_p, w_h) \\
&= [A_h(u_h, w_h) - A(u, w_h)] + [A(u, w_h) - A_h(u, w_h)] + A_h(u - u_p, w_h) \\
&= \langle f, w_h\rangle_h - (f, w_h) + [A(u, w_h) - A_h(u, w_h)] + A_h(u - u_p, w_h) \\
&= O(h^{k+2})(\|u\|_{k+3} + \omega^2\|u\|_{k+2} + \omega\|u\|_{k+4} + \|f\|_{k+2})\|w_h\|_2 \\
&= O(h^{k+2})(\|u\|_{k+3} + \omega^2\|u\|_{k+2} + \omega\|u\|_{k+4} + \|f\|_{k+2})C(\omega)(\omega^3 + \omega^{-1})\|\theta_h\|_0.
\end{aligned}$$

So we obtain

$$\|\theta_h\|_0 = O(h^{k+2})(\|u\|_{k+3} + \omega^2\|u\|_{k+2} + \omega\|u\|_{k+4} + \|f\|_{k+2})C(\omega)(\omega^3 + \omega^{-1}).$$

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 3, we obtain

$$\|u_h - u\|_{2,Z_0} \leq \|u_h - u_p\|_{2,Z_0} + \|u_p - u\|_{2,Z_0} \leq C\|u_h - u_p\|_0 + \|u_p - u\|_{2,Z_0} = O(h^{k+2}).$$

□

3.7 Numerical tests

For simplicity we test the accuracy Q^k spectral element method for solving $-\Delta u - \omega^2 u = f$ on $[0, 1]^2$ with a homogeneous Neumann boundary condition. We consider a simple exact solution $u(x, y) = \cos(\omega\pi x)\cos(\omega\pi y)$, and the ℓ^2-norm error for $\omega = 100$ is given in Figure 2, in which we observe $(k + 2)$-th order convergence for small enough h.

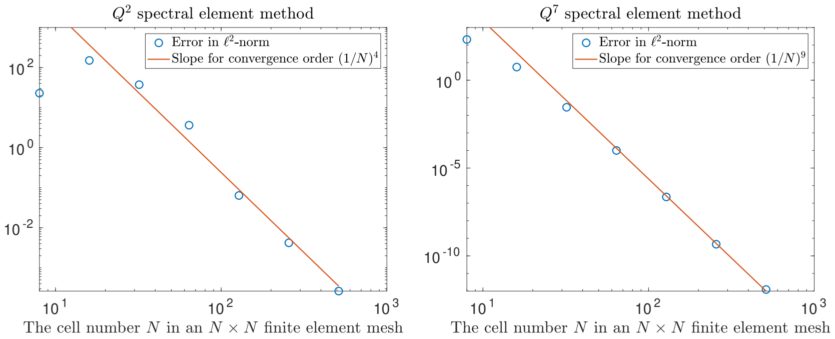

Fig. 2 The Q^k spectral element method on a $N \times N$ mesh is also a finite difference scheme at all $(Nk+1) \times (Nk+1)$ Gauss-Lobatto quadrature points. The error in ℓ^2-norm at all $(Nk+1) \times (Nk+1)$ grid points is listed.

4 Applications of Monotone Discrete Laplacian

In this section, we first demonstrate how to find the estimate $\| - \Delta_h\|_\infty$ for two monotone schemes: the classical second order centered difference (or equivalently Q^1 finite element method with quadrature) and the fourth order finite difference scheme (1) (or equivalently Q^2 spectral element method) in multiple dimensions. Then we discuss the applications.

For simplicity we start with the Poisson equation $-\Delta u + v(\boldsymbol{x})u = f$ with a given potential function $v(\boldsymbol{x}) \geq 0$ on $\Omega = (0,1)^d$ with $d = 1, 2, 3$ and homogeneous Dirichlet boundary conditions. Extensions to Neumann and periodic boundary conditions are straightforward.

4.1 Second order finite difference

The second order finite difference

$$(K+V)\mathbf{u} = \mathbf{f}$$

where V is diagonal matrix with entries $v(x_i) \geq 0$ and

$$K = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}, \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \ddots \\ u_{N-1} \\ u_N \end{pmatrix}, \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \ddots \\ f_{N-1} \\ f_N \end{pmatrix}.$$

It is easy to verify that the matrix $K+V$ satisfies Theorem 11 thus is monotone. The 2D scheme can be written as

$$(K \otimes I + I \otimes K + V)vec(U) = vec(F),$$

where U and F denote 2D arrays of grid point values and $vec(U)$ is the vector by arranging U column by column, and V is still a diagonal matrix.

For 3D scheme, following notation in [35], the matrix can be written as

$$K \otimes I \otimes I + I \otimes K \otimes I + I \otimes I \otimes K + V, \tag{4}$$

where V is still a diagonal matrix. The matrix $K \otimes I + I \otimes K + V$ and the matrix (4) still satisfy Theorem 11 thus they are monotone. The scheme can also be regarded as $Q1$ finite element method on uniform rectangular meshes with 2-point Gauss-Lobatto quadrature for all integrals. The monotonicity can be extended to P1 finite element method on 2D unstructured triangular meshes [46]. Next, we focus on the discrete Laplacian $-\Delta_h = \begin{cases} K, & d=1 \\ (K \otimes I + I \otimes K), & d=2 \\ (K \otimes I \otimes I + I \otimes K \otimes I + I \otimes I \otimes K), & d=3. \end{cases}$

It can be easily shown the corresponding graph of discrete Laplacian is strongly connected thus the matrix $-\Delta_h$ is irreducible, see [25]. By Corollary 1 in the Appendix, we only need to find a vector $\mathbf{z}$ such that $(-\Delta_h)\mathbf{z} \geq \mathbf{1}$ to obtain estimates for $\|(-\Delta_h)^{-1}\|_2$ and $\|(-\Delta_h)^{-1}\|_\infty$.

For the K matrix, in order to find $\mathbf{z}$ such that $K\mathbf{z} = \mathbf{1}$, first think about the exact solution to the problem $-u'' = 1, u(0) = u(1) = 0$, which is $z(x) = \frac{1}{2}x(1-x)$.

Let $\mathbf{z} = z(\mathbf{x})$ where $\mathbf{x}$ is the grid points for the corresponding scheme, i.e., $\mathbf{x} = \begin{bmatrix} h & 2h & \cdots & nh \end{bmatrix}^T$ with $h = \frac{1}{n+1}$. It is straightforward to verify that $K\mathbf{z} = \mathbf{1}$. On the other hand, since $0 \leq z(x) \leq \frac{1}{8}$ for $x \in (0,1)$, we have $\|\mathbf{z}\|_\infty \leq \frac{1}{8}$, thus $\|K^{-1}\|_\infty = \|\mathbf{z}\|_\infty \leq \frac{1}{8}$.

For $K \otimes I + I \otimes K$, we only need to consider an array $Z = \frac{1}{2}\mathbf{z}\mathbf{1}^T + \frac{1}{2}\mathbf{1}\mathbf{z}^T$, then

$$(K \otimes I + I \otimes K)vec(Z) = vec(ZK^T + KZ) = \frac{1}{2}vec(\mathbf{z}\mathbf{1}^T K^T + \mathbf{1}\mathbf{z}^T K^T + K\mathbf{z}\mathbf{1}^T + K\mathbf{1}\mathbf{z}^T)$$

$$> \frac{1}{2}vec(\mathbf{1}\mathbf{z}^T K^T + K\mathbf{z}\mathbf{1}^T) = vec(\mathbf{1}\mathbf{1}^T),$$

where $K\mathbf{1} = h^2 \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \end{bmatrix}^T \geq 0$ and $\mathbf{z} \geq 0$ is used. Thus, $\|(K \otimes I + I \otimes K)^{-1}\|_\infty \leq \|vec(Z)\|_\infty < \frac{1}{8}$.

Obviously the discussion can be easily extended to three dimensions by considering a 3D array

$$Z = \frac{1}{3}(\mathbf{z}\mathbf{1}^T) \otimes_{outer} \mathbf{1} + \frac{1}{3}(\mathbf{1}\mathbf{z}^T) \otimes_{outer} \mathbf{1} + \frac{1}{3}(\mathbf{1}\mathbf{1}^T) \otimes_{outer} \mathbf{z},$$

where $A \otimes_{outer} \mathbf{v}$ denotes the outer product between a matrix A and a column vector $\mathbf{v}$, and their outer product is a 3D array. By using this 3D array Z, we can easily find

$$\|(K \otimes I \otimes I + I \otimes K \otimes I + I \otimes I \otimes K)^{-1}\|_\infty \leq \|vec(Z)\|_\infty < \frac{1}{8}.$$

We emphasize that this estimate is sharp in one dimension but not in multiple dimensions.

Remark 2 If the domain is $(-L, L)^d$ instead of $(0, 1)^d$, then similiar discussion gives

$$\|(-\Delta_h)^{-1}\|_\infty \leq \|vec(Z)\|_\infty \leq \frac{L^2}{2}.$$

Here we also mention the eigenvectors of K. The eigen-decomposition of $-\Delta_h$ can be used as a simple preconditioner used in conjugate gradient method for inverting $-\Delta_h + \mathbb{V}$ in multiple dimensions [35]. Let $\mathbf{x} = \begin{bmatrix}\frac{1}{n+1} & \frac{2}{n+1} & \cdots & \frac{n}{n+1}\end{bmatrix}^T$ and $\sin(m\pi\mathbf{x}) = \begin{bmatrix}\sin(m\pi\frac{1}{n+1}) & \sin(m\pi\frac{2}{n+1}) & \cdots & \sin(m\pi\frac{n}{n+1})\end{bmatrix}^T$. Then K has eigenvectors $\sin(m\pi\mathbf{x})$ with eigenvalues $\lambda_m = \frac{(2-2\cos\frac{m\pi}{n+1})}{h^2} = \frac{(n+1)^2}{(2L)^2}(2 - 2\cos\frac{m\pi}{n+1})$ for $m = 1, \cdots, n$ for the domain $\Omega = [-L, L]^d$.

Remark 3 The eigen-value decomposition of K can be written as $K = T\Lambda T$ where T is the matrix of orthonormal eigenvectors with the property $T^{-1} = T$.

4.2 Q^2 spectral element method: fourth order finite difference

If using Q^2 spectral element method with 3-point Gauss-Lobatto quadrature for all integrals, we get a finite difference scheme [27, 43]

$$(H + V)\mathbf{u} = \mathbf{f}$$

or

$$(H \otimes I + I \otimes H + V)vec(U) = vec(F),$$

where

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

The matrices $H + V$ and $H \otimes I + I \otimes H + V$ are no longer M-matrices. It is proven in [25] that they are products of M-matrices thus still monotone under the following mesh size constraints:

1. $h^2 \max_i v(x_i) < 5$ for 1D.

2. $h^2 \max_i v(x_i, y_j) < \frac{3}{2}$ for 2D.

It can also be extended to quasi-uniform rectangular grids [11]. The monotonicity proof in [25] can be extended to the 3D case, for which the discrete Laplacian matrix is given as $-\Delta_h = \begin{cases} H, & d=1 \\ (H\otimes I + I\otimes H), & d=2 \\ (H\otimes I\otimes I + I\otimes H\otimes I + I\otimes I\otimes H), & d=3. \end{cases}$

Notice that $-\Delta_h$ is monotone unconditionally on a uniform grid without any mesh size constraint [25, 10].

It can be easily verified that we also have $H\mathbf{z} = \mathbf{1}$ where $\mathbf{z}$ is the same vector in previous subsection. Thus in the one dimension case, we have $\|(-\Delta_h)\|_\infty \leq \frac{1}{8}$.

Unfortunately we do not have $H\mathbf{1} \geq 0$, thus the 2D case is different from the second order finite difference. Define a vector $\mathbf{v} = \begin{bmatrix}1 & 2 & 2 & \cdots & 2 & 1\end{bmatrix}^T$ then we have $H\mathbf{v} \geq 0$.

For $-\Delta_h = H\otimes I + I\otimes H$, we need to consider an array $Z = \frac{1}{2}\mathbf{z}\mathbf{v}^T + \frac{1}{2}\mathbf{1}\mathbf{v}^T$, then

$$(H\otimes I + I\otimes H)vec(Z) = vec(ZH^T + HZ) = \frac{1}{2}vec(\mathbf{z}\mathbf{v}^T H^T + \mathbf{v}\mathbf{z}^T H^T + H\mathbf{z}\mathbf{v}^T + H\mathbf{v}\mathbf{z}^T)$$

$$> \frac{1}{2}vec(\mathbf{v}\mathbf{z}^T H^T + H\mathbf{z}\mathbf{v}^T) > vec(\mathbf{1}\mathbf{1}^T),$$

where $H\mathbf{z} \geq \mathbf{1}$, $H\mathbf{v} \geq 0$, $\mathbf{v} \geq \mathbf{1}$ and $\mathbf{z} \geq 0$ are used. Thus, we can only obtain

$$\|(H\otimes I + I\otimes H)^{-1}\|_\infty < \|vec(Z)\|_\infty \leq \frac{1}{4}.$$

Remark 4 If the domain is $(-L, L)^d$ instead of $(0,1)^d$, then similiar discussion gives

$$\|(-\Delta_h)^{-1}\|_\infty \leq \|vec(Z)\|_\infty \leq L^2.$$

For three dimensions, if considering a 3D array

$$Z = \frac{1}{3}(\mathbf{z}\mathbf{v}^T) \otimes_{outer} \mathbf{v} + \frac{1}{3}(\mathbf{v}\mathbf{z}^T) \otimes_{outer} \mathbf{v} + \frac{1}{3}(\mathbf{v}\mathbf{v}^T) \otimes_{outer} \mathbf{z},$$

we can easily find

$$\|(H\otimes I\otimes I + I\otimes H\otimes I + I\otimes I\otimes H)^{-1}\|_\infty < \|vec(Z)\|_\infty \leq \frac{1}{4}.$$

The eigen-decomposition of H can be computed either numerically or expressed explicitly as follows. Let $\mathbf{x}$ and $\sin(m\pi\mathbf{x})$ be the same as in previous subsection. Let x_j^1 and x_j^2 be the larger and smaller root of the following quadratic equation of x:

$$2\cos\frac{j\pi}{n+1}x^2 + (1+\cos^2\frac{j\pi}{n+1})x - 4\cos\frac{j\pi}{n+1} = 0. \tag{5}$$

$$b_j = \begin{cases} x_j^1 & j \le \frac{n+1}{2} \\ x_j^2 & j > \frac{n+1}{2} \end{cases}.$$

The m-th eigenvector of H is given as $\sin(m\pi\mathbf{x}) \circ \begin{bmatrix} 1 & b_m & 1 & b_m & 1 & b_m \cdots & 1 & b_m & 1 \end{bmatrix}^T$ with eigenvalue $\frac{(n+1)^2}{(2L)^2}(2 - 2b_m \cos\frac{m\pi}{n+1})$ for the domain $\Omega = [-L, L]^d$.

4.3 Picard iteration and Newton's iteration for nonlinear equations

Consider solving a semilinear equation $-\Delta u + vu + \beta|u|^2 u = f$ with cubic nonlinear term and a non-negative potential function $v \ge 0$ on homogeneous Dirichlet boundary conditions on $\Omega = [0,1]^d$. For simplicity, we denote the numerical approximation by

$$-\Delta_h u + Vu + \beta \operatorname{diag}(|u|^2) u = f, \tag{6}$$

where $-\Delta_h$ is a monotone discrete Laplacian, V is diagonal matrix, u is a vector of numerical solutions and $\operatorname{diag}(u^2)$ denotes a diagonal matrix with diagonal entries $|u_i|^2$. The system (6) is closely related to the Gross-Pitaevskii eigenvalue problem [37]. See [9] and references for more details on the background. We assume both right hand side f and the exact solution to (6) are positive. Note that the ground state of Gross-Pitaevskii eigenvalue problem can be proven positive [30].

The Picard iteration for a semi-linear equation $A(u)u = f$ takes the form $A(u^n)u^{n+1} = f$, and we have

$$-\Delta_h u^{n+1} + Vu^{n+1} + \beta \operatorname{diag}(|u^n|^2) u^{n+1} = f. \tag{7}$$

The Newton's method for $N(u) = 0$ takes the form $u^{n+1} = u^n - [\frac{\delta N}{\delta u}(u^n)]^{-1} N(u^n)$, equivalent to $[\frac{\delta N}{\delta u}(u^n)]u^{n+1} = [\frac{\delta N}{\delta u}(u^n)]u^n - N(u^n)$, which can be written as:

$$-\Delta_h u^{n+1} + Vu^{n+1} + \beta 3 \operatorname{diag}(|u^n|^2) u^{n+1} = f + 2\beta(u^n)^3. \tag{8}$$

To see why the cubic polynomial system (6) has a unique solution, we note that $-\Delta_h u + Vu + \beta u^3 = f$ is the first order optimality condition for the minimizing the energy function:

$$F(u) = -\frac{1}{2}u^T(\Delta_h + V)u + \frac{1}{4}\beta \sum_i u_i^4 - \langle f, u\rangle. \tag{9}$$

Since $F''(u) = -\Delta_h + V + 3\beta \operatorname{diag}(u^2)$ and $-\Delta_h$ has positive eigenvalues as discussed in previous subsections, $F(u)$ is strongly convex for the case $\beta \ge 0$ and non-negative potential $V(\boldsymbol{x}_i) \ge 0$, it thus has a unique minimizer. For minimizing the function, there are other straightforward algorithms, e.g.,

- Gradient descent (forward Euler):

$$u^{n+1} = u^n - \Delta t F'(u^n) = u^n - \Delta t_k [-\Delta_h u^n + V u^n + \beta \operatorname{diag}(|u^n|^2) u^n - f]$$

where the step size Δt_k can be implemented by exact line search (steepest descent).

- Backward Euler for the linearized equation [1]:

$$u^{n+1} = u^n - \Delta t[-\Delta_h u^{n+1} + V u^{n+1} + \beta \operatorname{diag}(|u^n|^2) u^{n+1} - f]. \tag{10}$$

- Newton's method: with step size being constant 1 for minimizing $F(u)$ will give exactly the same Newton iteration above for solving the equation directly.

In both Picard iteration and Newton's method, we need to invert a matrix $A_h = -\Delta_h + V + \beta \operatorname{diag}(|u^n|^2)$ or $A_h = -\Delta_h + V + 3\beta \operatorname{diag}(|u^n|^2)$. Recall that V is diagonal with non-negative diagonal entries. If $\beta > 0$ and $(-\Delta_h)\mathbf{z} = \mathbf{1}$, then $A_h \mathbf{z} \geq (-\Delta_h)\mathbf{z} = \mathbf{1}$ because $\mathbf{z} \geq 0$. If A_h is monotone, then we have $\|A_h^{-1}\|_2 \leq \|A_h^{-1}\|_\infty \leq \|\mathbf{z}\|_\infty$.

4.4 Convergence rate of Picard iteration for $\beta = 1$

From now on, we only focus on the Q^2 spectral element scheme (1) for $-\Delta_h$ on $\Omega = [0,1]^2$. Recall V is a diagonal matrix with non-negative diagonal entries being point values of the potential function $V(\boldsymbol{x}_i)$, and $|u^n|^2$ in (6) can be regarded as a diagonal matrix with diagonal entries being $|u^n|_i^2$.

Theorem 7 *On the domain $\Omega = [0,1]^2$, for the discrete Laplacian being the fourth order finite difference obtained from Q^2 spectral element method, for $\beta = 1$ and non-negative potential function $v(\boldsymbol{x}_i) \geq 0$, assume the exact solution to* (6) *satisfies $0 \leq u_i \leq 1$ and the right hand side function is bounded $0 \leq f_i \leq 4, \forall i$. If the mesh size is small enough $h \leq 1$ and $h^2 v(\boldsymbol{x}_i) \leq \frac{1}{2}$ for all i, then*

1. *The Picard iteration* (7) *with random initial guess $u_i^0 \in [0,1]$ satisfies $0 \leq u_i^n < 1$.*
2. *The Picard iteration converges with a rate $\|u^{n+1} - u\|_\infty \leq \frac{3}{4}\|u^n - u\|_\infty$.*

Remark 5 The results above are not sharp since the estimate on $\|(-\Delta_h)^{-1}\|_\infty$ in this paper is not sharp.

Proof. By the discussion in Section 4.2, the matrix $A_h = -\Delta_h + V + \operatorname{diag}(|u^n|^2)$ is monotone if $h^2(v(\boldsymbol{x}_i) + |u_i^n|^2) < \frac{3}{2}$.

We first prove $0 \leq u_i^n < 1$ by induction. With $u_i^0 \in [0,1]$, we first have $u^1 = (-\Delta_h + V + \operatorname{diag}|u^0|^2)^{-1} f$. Under the assumption $h^2(v(\boldsymbol{x}_i) \leq \frac{1}{2}$ and $h \leq 1$, by discussion in Section 4.2, we have $(-\Delta_h + V + \operatorname{diag}|u^0|^2)^{-1} \geq 0$ and $\|(-\Delta_h + V + \operatorname{diag}|u^0|^2)^{-1}\|_\infty < \frac{1}{4}$, thus $u^1 = (-\Delta_h + V + \operatorname{diag}|u^0|^2)^{-1} f \geq 0$ and $\|u^1\|_\infty \leq \|(-\Delta_h + V + \operatorname{diag}|u^0|^2)^{-1}\|_\infty \|f\|_\infty < 1$ gives $0 \leq u_i^1 < 1$. With induction assumption $0 \leq u_i^n < 1$, the same discussion easily gives $0 \leq u_i^{n+1} < 1$.

Let u be the exact solution to $-\Delta_h u + V(x) u + \operatorname{diag}|u|^2 u = f$. Let $e^{n+1} = u^{n+1} - u$, then

$$
\begin{aligned}
&-\Delta_h e^{n+1} + V e^{n+1} + \operatorname{diag}|u^n|^2 u^{n+1} - \operatorname{diag}|u|^2 u = 0\\
&-\Delta_h e^{n+1} + V e^{n+1} + \operatorname{diag}|u^n|^2 u^{n+1} - \operatorname{diag}|u^n|^2 u + \operatorname{diag}|u^n|^2 u - \operatorname{diag}|u|^2 u = 0\\
&-\Delta_h e^{n+1} + V e^{n+1} + \operatorname{diag}|u^n|^2 e^{n+1} = -\operatorname{diag}(|u^n|^2 - u^2)u\\
&A_h e^{n+1} = -(u^n + u)(e^n)u,
\end{aligned}
$$

where $u(u^n+u)e^n$ are entryways product for vectors. So $e^{n+1} = -A_h^{-1}u(u^n+u)e^n$, which is standard for Picard iteration. Here we have already proven $|u_i^n| < 1$, so the mesh size satisfies $h^2(v(\boldsymbol{x}_i)+|u_i^n|^2) < \frac{3}{2}$. Thus A_h is monotone and we have estimate $\|A_h^{-1}\|_\infty \le \frac{1}{4}$. With both $0 \le u \le 1$ and $0 \le u_i^n < 1$, we have $\|e^n\|_\infty \le 1$, so we have

$$
\|e^{n+1}\|_\infty \le \|A_h^{-1}\|_\infty \|u\|_\infty \|u^n+u\|_\infty \|e^n\|_\infty \le \frac{1}{4}(\|e^n\|_\infty + 2\|u\|_\infty)\|e^n\|_\infty \le \frac{3}{4}\|e^n\|_\infty.
$$

□

4.5 Convergence rate of Newton's method for $\beta = 1$

For Newton's method, we need to make stronger assumptions on h and f:

Theorem 8 *On the domain $\Omega = [0,1]^2$, for the discrete Laplacian being the fourth order finite difference obtained from Q^2 spectral element method, for $\beta = 1$ and non-negative potential function $v(\boldsymbol{x}_i) \ge 0$, assume the exact solution to* (6) *satisfies $0 \le u_i \le 1$ and the right hand side function is bounded $0 \le f_i \le 4 - a, \forall i$ where $a > 0$ is small. If the mesh size is small enough $h^2 \le \frac{1}{3}$ and $h^2 v(\boldsymbol{x}_i) \le \frac{1}{2}$ for all i, then*

1. *Newton's method* (8) *with random initial guess $u_i^0 \in [0,1]$ satisfies $0 \le u_i^n < 1 - \frac{a}{4}$.*
2. *Newton's method* (8) *converges with a rate $\|u^{n+1} - u\|_\infty \le (1 - \frac{3a}{16})\|u^n - u\|_\infty^2$.*

Remark 6 The results above are not sharp since the estimate on $\|(-\Delta_h)^{-1}\|_\infty$ in this paper is not sharp.

Proof. By the discussion in Section 4.2, the matrix $A_h = -\Delta_h + V + 3\operatorname{diag}(|u^n|^2)$ is monotone if $h^2(v(\boldsymbol{x}_i) + 3|u_i^n|^2) < \frac{3}{2}$.

We first prove $0 \le u_i^n < 1 - \frac{a}{4}$ by induction. With $0 \le u_i \le 1$, we first have $u^1 = (-\Delta_h + V + 3\operatorname{diag}|u^0|^2)^{-1} f$. Under the assumption $h^2(v(\boldsymbol{x}_i) \le \frac{1}{2}$ and $h^3 \le \frac{1}{3}$, by discussion in Section 4.2, we have $(-\Delta_h + V + 3\operatorname{diag}|u^0|^2)^{-1} \ge 0$ and $\|(-\Delta_h + V + 3\operatorname{diag}|u^0|^2)^{-1}\|_\infty < \frac{1}{4}$, thus $u^1 = (-\Delta_h + V + 3\operatorname{diag}|u^0|^2)^{-1} f \ge 0$ and $\|u^1\|_\infty \le \|(-\Delta_h + V + 3\operatorname{diag}|u^0|^2)^{-1}\|_\infty \|f\|_\infty < 1 - \frac{a}{4}$ gives $0 \le u_i^1 < 1 - \frac{a}{4}$. With induction assumption $0 \le u_i^n < 1 - \frac{a}{4}$, the same discussion gives $u^{n+1} = A_h^{-1}(f + 2\operatorname{diag}|u^n|^2 u^n) \ge 0$ and $\|u^{n+1}\|_\infty < 1 - \frac{a}{4}$.

Let $A_h = -\Delta_h + V + 3\operatorname{diag}(|u^n|^2)$ and $e^n = u^n - u$. Notice that $\operatorname{diag}|u^2|u$ can also be written as a vector u^3 with entries u_i^3, we have

$$
\begin{aligned}
&-\Delta_h e^{n+1} + V e^{n+1} + 3(u^n)^2 u^{n+1} - u^3 = 0 + 2(u^n)^3\\
&-\Delta_h e^{n+1} + V e^{n+1} + 3(u^n)^2 u^{n+1} - 3(u^n)^2 u + 3(u^n)^2 u - u^3 - 2(u^n)^3 = 0\\
&A_h e^{n+1} + 3(u^n)^2 u - u^3 - 2(u^n)^3 = 0\\
&A_h e^{n+1} = -3(u^n)^2 u + u^3 + 2(u^n)^3\\
&A_h e^{n+1} = -3(u^n)^2 u + 3u^3 - 2u^3 + 2(u^n)^3\\
&A_h e^{n+1} = -3u[(u^n)^2 - u^2] + 2(u^n - u)(u^2 + uu^n + |u^n|^2)\\
&A_h e^{n+1} = e^n[-3u(u^n + u) + 2(u^2 + uu^n + |u^n|^2)\\
&A_h e^{n+1} = e^n[2|u^n|^2 - uu^n - u^2]\\
&A_h e^{n+1} = |e^n|^2(2u^n + u),
\end{aligned}
$$

which is all standard for Newton's method. Here we have estimate $\|A_h^{-1}\|_\infty \le \frac{1}{4}$. Also, $0 \le u_i^n \le 1 - \frac{a}{4}$ implies $\|e^n\|_\infty \le 1$. With $\|u^n\|_\infty \le 1 - \frac{a}{4}$, we have

$$
\|e^{n+1}\|_\infty \le \frac{1}{4}\|2u^n + u\|_\infty \|e^n\|_\infty^2 \le \frac{1}{4}(3\|u^n\|_\infty + \|e^n\|_\infty)\|e^n\|_\infty^2 \le (1 - \frac{3a}{16})\|e^n\|_\infty^2.
$$

□

4.6 Convergence of a new Picard iteration for $\beta = -1$

Consider a constant c and an equivalent system $-\Delta_h u + Vu + \beta|u|^2 u + cu = f + cu$ with $c + V(x) + \beta u^2 \ge 0$. The Picard iteration for $A(u)u = B(u)$ is $A(u^n)u^{n+1} = B(u^n)$:

$$
-\Delta_h u^{n+1} + (V + c)u^{n+1} + \beta|u^n|^2 u^{n+1} = f + cu^n.
$$

In particular, let $c = \frac{1}{\Delta t}$, then this is exactly the backward Euler scheme (10).

Let $e^{n+1} = u^{n+1} - u$, then

$$
\begin{aligned}
-\Delta_h e^{n+1} + (V + c)e^{n+1} + \beta|u^n|^2 u^{n+1} - \beta u^3 &= ce^n\\
-\Delta_h e^{n+1} + (V + c)e^{n+1} + \beta|u^n|^2 u^{n+1} - \beta|u^n|^2 u + \beta|u^n|^2 u - \beta u^3 &= ce^n\\
-\Delta_h e^{n+1} + (V + c)e^{n+1} + \beta|u^n|^2 e^{n+1} &= -\beta u(|u^n|^2 - u^2) + ce^n\\
A_h^c e^{n+1} &= [c - \beta u(u^n + u)]e^n.
\end{aligned}
$$

If assuming $V + c + \beta|u^n|^2 \ge 0$, with suitable mesh size assumptions, we can still have $[A_h^c]^{-1} \ge 0$ and estimate $\|[A_h^c]^{-1}\|_\infty \le \frac{1}{4}$, then similar convergence discussion follows.

4.7 Positivity of SCF for a nonlinear eigenvalue problem

We conclude this section by considering another application of monotonicity for the nonlinear eigenvalue problem of finding the smallest eigenvalue for

$$-\Delta u + V(\boldsymbol{x})u + \beta|u|^2 u = \lambda u.$$

It is proven in [30] that the ground state eigenfunction $u(\boldsymbol{x}) > 0$. In general, it is not easy to solve such a nonlinear eigenvalue problem numerically, especially with a convergent proof. The self consistent field (SCF) iteration is a simple popular method for nonlinear eigenvalue problems given as

$$-\Delta_h u^{n+1} + V u^{n+1} + \beta|u^n|^2 u^{n+1} = \lambda_n u^{n+1},$$

where λ_n is the lowest eigenvalue for the matrix $-\Delta_h + V + \beta \operatorname{diag}|u^n|^2$. In practice, SCF may diverge. If there is a pair (u, v) such that v is the lowest eigenvector of $A(u) = -\Delta_h + V + \beta \operatorname{diag}|u|^2$ and u is the lowest eigenvector of $A(v) = -\Delta_h + V + \beta \operatorname{diag}|v|^2$, then SCF starting with u will stuck with this pair. On the other hand, its local convergence of SCF can usually be proven [47].

Let $A_h(u^n) = -\Delta_h + V + \beta \operatorname{diag}(|u^n|^2)$ where $\operatorname{diag}(|u_n|^2)$ is a diagonal matrix with entries $|u^n_i|^2$. Notice that $V + \beta \operatorname{diag}(|u^n|^2)$ is a diagonal matrix thus does not change connectivity of a graph. Since $-\Delta_h$ is irreducible, A_h is also irreducible.

For $\beta > 0$, it is the easier *defocusing case*. Assume $h^2 \max_i(v(x_i) + |u^n_i|^2) \leq \frac{3}{2}$, then monotonicity of A_h holds for fourth order scheme (no mesh constraints for second order scheme). By Perron-Frobenius Theorem, i.e., Theorem 9 in the Appendix, the smallest eigenvalue of A_h is positive and simple, and has a positive eigenvector.

For the *focusing case* $\beta < 0$, let c be a fixed positive constant such that $c + \min_i(v(x_i) - |u^n_i|^2) \geq 0$. Consider $A^c_h = -\Delta_h + V - \operatorname{diag}(|u^n|^2) + cI$. For $h^2 \max_i(v(x_i) - |u^n_i|^2 + c) \leq \frac{3}{2}$, then monotonicity of A_h holds for fourth order scheme (no mesh constraints for second order scheme). By Perron-Frobenius Theorem, i.e., Theorem 9 in the Appendix, the smallest eigenvalue of A^c_h is positive and simple, and has a positive eigenvector. This implies that the smallest eigenvalue A_h is simple and has a positive eigenvector.

So monotonicity of the discrete Laplacian ensures SCF returns a positive iterate $u^{n+1} > 0$:

$$-\Delta u_{n+1} + V u_{n+1} + \beta \operatorname{diag}|u_n|^2 u_{n+1} = \lambda_n u_{n+1},$$

where λ_n is the smallest eigenvalue of $A_h(u^n)$.

5 Concluding Remarks

We have reviewed some recent progress of Q^k spectral element method including its accuracy as a finite difference scheme and provable monotonicity results. We have also discussed its accuracy for the Helmholtz equation and the applications of monotonicity for solving certain nonlinear problems.

Novelty statement

To the best of the author's knowledge, the results in Section 3 and Section 4 have not appeared in any paper in the literature. They are not included in any other paper, preprint or report written by the author.

Acknowledgements Research was supported by NSF DMS-1913120. The author is deeply grateful to Dr. Théophile Chaumont-Frelet at Inria for discussions on the Helmholtz equation, and to Prof. Jianfeng Lu at Duke University for discussions on nonlinear eigenvalue problems.

Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

Appendix

We first list a few concepts:

- A matrix $A \in \mathbb{R}^{n\times n}$ is called *monotone* if its inverse is nonnegative $A^{-1} \geq 0$ (entrywise inequality).
- A matrix $A \in \mathbb{C}^{n\times n}$ is called *reducible* if there exists a permutation matrix P such that PAP^T is block upper triangular.
- A square matrix A is *irreducible* if it is not reducible. A matrix is irreducible if and only the graph it represents is strongly connected.

Lemma 1 *For a nonsingluar irreducible matrix A, A^{-1} is also irreducible.*

Proof. Assume A^{-1} is reducible, then $PA^{-1}P^T = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$. Since A^{-1} is nonsingular, so are B_{11} and B_{22}. Thus $A = P^T \begin{bmatrix} B_{11}^{-1} & -B_{11}^{-1}B_{12}B_{22}^{-1} \\ 0 & B_{22}^{-1} \end{bmatrix} P$, which is a contradiction with A being irreducible. □

The following results can be found in [44]:

Lemma 2 *If $A \geq 0$ is irreducible, then either $\rho(A) = \sum_{j=1}^n a_{ij}, \quad \forall i$, or*

$$\min_i \left(\sum_{j=1}^n a_{ij} \right) \leq \rho(A) \leq \max_i \left(\sum_{j=1}^n a_{ij} \right).$$

Theorem 9 (Perron-Frobenius) *If $A \geq 0$ is irreducible, then*

1. *The spectral radius $\rho(A)$ is a simple eigenvalue of A with an eigenvector $x > 0$.*
2. *$\rho(A)$ increases when any entry of A increases.*

Theorem 10 *The positive eigenvector (Perron-Frobenius eigenvector) for an irreducible nonnegative matrix is unique.*

Proof. Let $x > 0$ be the left Perron-Frobenius eigenvector then $x^T A = \rho(A)x^T$. If there exists another eigenvector $y > 0$ for an eigenvalue λ, then $Ay = \lambda y \Rightarrow x^T Ay = \lambda x^T y$. Since $x^T A = \rho(A)x^T \Rightarrow x^T Ay = \rho(A)x^T y$, we get $(\rho(A) - \lambda)x^T y = 0$ and $x^T y > 0 \Rightarrow \rho(A) = \lambda$. Thus there is only one eigenvalue with positive eigenvectors. □

By the results above, we get

Corollary 1 *For a real monotone and irreducible matrix A, its inverse $A^{-1} \geq 0$ is irreducible. Let a^{ij} be entries of A^{-1}, then*

1. *$\rho(A^{-1}) \leq \max_i \left(\sum\limits_{j=1}^{n} a^{ij}\right) = \|A^{-1}\|_\infty$.*
2. *If a vector $\mathbf{z}$ satisifies $A\mathbf{z} \geq \mathbf{1}$, then $A\mathbf{1} \leq \mathbf{z}$, thus $\|A^{-1}\|_2 \leq \|A^{-1}\|_\infty \leq \|\mathbf{z}\|_\infty$.*
3. *A has a positive eigenvalue with a positive eigenvector. If assume A is also symmetric, then the smallest eigenvalue of A is positive and simple with a positive eigenvector, and*
$$\|A^{-1}\|_2 = \rho(A^{-1}) \leq \|A^{-1}\|_\infty.$$

Nonsingular M-matrices are monotone matrices. There are many equivalent definitions or characterizations of M-matrices, see [41]. The following is a convenient sufficient but not necessary characterization of nonsingular M-matrices [25]:

Theorem 11 *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.*

By condition K_{35} in [41], a sufficient and necessary characterization is,

Theorem 12 *For a real square matrix A with positive diagonal entries and non-positive off-diagonal entries, A is a nonsingular M-matrix if and only if that there exists a positive diagonal matrix D such that AD has all positive row sums.*

Non-negative row sum is not a necessary condition for M-matrices. For instance, the following matrix A is an M-matrix by Theorem 12:

$$A = \begin{bmatrix} 10 & 0 & 0 \\ -10 & 2 & -10 \\ 0 & 0 & 10 \end{bmatrix}, D = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, AD = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 4 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

References

1. Bao, W., Du, Q.: Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow. SIAM Journal on Scientific Computing **25**(5), 1674–1697 (2004)
2. Bernkopf, M., Chaumont-Frelet, T., Melenk, J.M.: Wavenumber-explicit stability and convergence analysis of hp finite element discretizations of Helmholtz problems in piecewise smooth media. arXiv:2209.03601 (2022)
3. Bramble, J.H.: Fourth-order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions. Math. Comp. **17**(83), 217–222 (1963)
4. Bramble, J.H., Hubbard, B.E.: New monotone type approximations for elliptic problems. Math. Comp. **18**(87), 349–367 (1964)
5. Chaumont-Frelet, T., Nicaise, S.: Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problems. IMA Journal of Numerical Analysis **40**(2), 1503–1543 (2020)
6. Chen, C.: Superconvergent points of Galerkin's method for two point boundary value problems. Numerical Mathematics A Journal of Chinese Universities **1**, 73–79 (1979)
7. Chen, C.: Superconvergence of finite element solutions and its derivatives [j]. Numerical Mathematics A Journal of Chinese Universities **2**, 002 (1981)
8. Chen, C.: Structure theory of superconvergence of finite elements (In Chinese). Hunan Science and Technology Press, Changsha (2001)
9. Chen, Z., Lu, J., Lu, Y., Zhang, X.: On the convergence of Sobolev gradient flow for the Gross-Pitaevskii eigenvalue problem. to appear in SIAM Journal on Numerical Analysis (2023)
10. Cross, L.J., Zhang, X.: Monotonicity of Q^3 spectral element method for discrete Laplacian. arXiv:2010.07282 (2023)
11. Cross, L.J., Zhang, X.: On the monotonicity of Q^2 spectral element method for Laplacian on quasi-uniform rectangular meshes. to appear in Communications in Computational Physics (2023)
12. Cummings, P., Feng, X.: Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. Mathematical Models and Methods in Applied Sciences **16**(01), 139–160 (2006)
13. Douglas, J., Dupont, T.: Galerkin approximations for the two point boundary problem using continuous, piecewise polynomial spaces. Numer. Math. **22**(2), 99–109 (1974)
14. Douglas Jr, J., Santos, J.E., Sheen, D., Bennethum, L.S.: Frequency domain treatment of one-dimensional scalar waves. Mathematical models and methods in applied sciences **3**(02), 171–194 (1993)
15. Douglas Jr, J., Sheen, D., Santos, J.E.: Approximation of scalar waves in the space-frequency domain. Mathematical Models and Methods in Applied Sciences **4**(04), 509–531 (1994)
16. Feng, X., Sheen, D.: An elliptic regularity coefficient estimate for a problem arising from a frequency domain treatment of waves. Transactions of the American Mathematical Society **346**(2), 475–487 (1994)
17. Galkowski, J., Spence, E.A.: Sharp preasymptotic error bounds for the Helmholtz h-FEM. arXiv preprint arXiv:2301.03574 (2023)
18. Höhn, W., Mittelmann, H.D.: Some remarks on the discrete maximum-principle for finite elements of higher order. Computing **27**(2), 145–154 (1981)
19. Hu, J., Zhang, X.: Positivity-preserving and energy-dissipative finite difference schemes for the Fokker–Planck and Keller–Segel equations. IMA Journal of Numerical Analysis **43**(3), 1450–1484 (2023)
20. Ihlenburg, F., Babuška, I.: Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. Computers & Mathematics with Applications **30**(9), 9–37 (1995)
21. Ihlenburg, F., Babuska, I.: Finite element solution of the Helmholtz equation with high wave number part II: the hp version of the FEM. SIAM Journal on Numerical Analysis **34**(1), 315–358 (1997)

22. Li, H.: Accuracy and monotonicity of spectral element method on structured meshes. Ph.D. thesis, Purdue University (2021)
23. Li, H., Appelö, D., Zhang, X.: Accuracy of Spectral Element Method for Wave, Parabolic, and Schrödinger Equations. SIAM Journal on Numerical Analysis **60**(1), 339–363 (2022)
24. Li, H., Xie, S., Zhang, X.: A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations. SIAM Journal on Numerical Analysis **56**(6), 3308–3345 (2018)
25. Li, H., Zhang, X.: On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method. Numerische Mathematik pp. 1–36 (2020)
26. Li, H., Zhang, X.: Superconvergence of $C^0 - Q^k$ finite element method for elliptic equations with approximated coefficients. Journal of Scientific Computing **82**(1), 1 (2020)
27. Li, H., Zhang, X.: Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations. Journal of Scientific Computing **82**(2), 36 (2020)
28. Li, H., Zhang, X.: A monotone Q^1 finite element method for anisotropic elliptic equations. arXiv:2310.16274 (2023)
29. Li, H., Zhang, X.: A high order accurate bound-preserving compact finite difference scheme for two-dimensional incompressible flow. Communications on Applied Mathematics and Computation pp. 1–29 (2023)
30. Lieb, E.H., Seiringer, R., Yngvason, J.: Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional. The Stability of Matter: From Atoms to Stars: Selecta of Elliott H. Lieb pp. 685–697 (2001)
31. Lin, Q., Yan, N.: Construction and Analysis for Efficient Finite Element Method (In Chinese). Hebei University Press (1996)
32. Lin, Q., Yan, N., Zhou, A.: A rectangle test for interpolated finite elements. In: Proc. Sys. Sci. and Sys. Eng.(Hong Kong), Great Wall Culture Publ. Co, pp. 217–229 (1991)
33. Liu, C., Gao, Y., Zhang, X.: Structure Preserving Schemes for Fokker–Planck Equations of Irreversible Processes. Journal of Scientific Computing **98**(1), 4 (2024)
34. Liu, C., Zhang, X.: A positivity-preserving implicit-explicit scheme with high order polynomial basis for compressible Navier-Stokes equations. Journal of Computational Physics **493**, 112496 (2023)
35. Liu, X., Shen, J., Zhang, X.: A simple GPU implementation of spectral-element methods for solving 3D Poisson type equations on cartesian meshes. arXiv:2310.00226 (2023)
36. Lorenz, J.: Zur inversmonotonie diskreter probleme. Numer. Math. **27**(2), 227–238 (1977)
37. Lu, J., Ying, L.: Sparsifying preconditioner for soliton calculations. Journal of Computational Physics **315**, 458–466 (2016)
38. Maday, Y., Rønquist, E.M.: Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries. Computer Methods in Applied Mechanics and Engineering **80**(1-3), 91–115 (1990)
39. Melenk, J., Sauter, S.: Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. Mathematics of Computation **79**(272), 1871–1914 (2010)
40. Melenk, J.M., Sauter, S.: Wavenumber Explicit Convergence Analysis for Galerkin Discretizations of the Helmholtz Equation. SIAM Journal on Numerical Analysis **49**(3), 1210–1243 (2011). DOI 10.1137/090776202
41. Plemmons, R.J.: M-matrix characterizations. I—-nonsingular M-matrices. Linear Algebra and its Applications **18**(2), 175–188 (1977)
42. Schatz, A.H.: An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Mathematics of Computation **28**(128), 959–962 (1974)
43. Shen, J., Zhang, X.: Discrete maximum principle of a high order finite difference scheme for a generalized Allen–Cahn equation. Communications in Mathematical Sciences **20**(5), 1409–1436 (2022)
44. Varga, R.S.: Matrix iterative analysis, vol. 27. Springer Science & Business Media (1999)
45. Vejchodský, T., Šolín, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comp. **76**(260), 1833–1846 (2007)

46. Xu, J., Zikatanov, L.: A monotone finite element scheme for convection-diffusion equations. Mathematics of Computation of the American Mathematical Society **68**(228), 1429–1446 (1999)
47. Yang, C., Gao, W., Meza, J.C.: On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. SIAM journal on matrix analysis and applications **30**(4), 1773–1788 (2009)