SIAM J. NUMER. ANAL.
Vol. 60, No. 4, pp. 2150–2181

CONVERGENCE ANALYSIS OF A FULLY DISCRETE ENERGY-STABLE NUMERICAL SCHEME FOR THE Q-TENSOR FLOW OF LIQUID CRYSTALS*

VARUN M. GUDIBANDA†, FRANZISKA WEBER‡, AND YUKUN YUE‡

Abstract. We present a fully discrete convergent finite difference scheme for the Q-tensor flow of liquid crystals based on the energy-stable semidiscrete scheme by Zhao et al. [*Comput. Methods Appl. Mech. Engrg.*, 2017, pp. 803–825]. We prove stability properties of the scheme and show convergence to weak solutions of the Q-tensor flow equations. We demonstrate the performance of the scheme in numerical simulations.

Key words. convergence, finite difference method, Q-tensor flow, energy-stable

MSC codes. 65M06, 65M12

DOI. 10.1137/20M1383550

1. Introduction.

Liquid crystals constitute a state of matter that is intermediate between solids and liquids. On one hand, they have properties that are typical for fluids—in particular they have the ability to flow—and on the other hand, they exhibit properties of solids—as an example, their molecules are oriented in a crystal-like manner. A common characteristic of materials exhibiting a liquid crystal phase is that they consist of elongated molecules of identical size. They may be pictured as "rods" or "ribbons" and are subject to molecular interactions that make them align alike [11].

Liquid crystals play an important role in nature: As an example, phospholipids, which constitute the main component of cell membranes, are a form of liquid crystal. They also appear in many daily applications, such as soaps, shampoos, and detergents. Further applications include displays of electronic devices (LCD), where one makes use of the optical properties of liquid crystals in the presence or absence of an electric field, thermometers, optical switches [6, 16], and biotechnological applications. One generally distinguishes three types of liquid crystals: nematics, cholesterics, and smectics. We focus here on the numerical discretization of a liquid crystal model for nematic liquid crystals, the so-called Q-tensor model.

1.1. Q-tensor model.

In the Q-tensor model by Landau and de Gennes [5], the main orientation of the liquid crystal molecules is represented by the Q-tensor, a symmetric, trace-free matrix that is assumed to minimize the Landau–de Gennes free energy

$$E_{LG}(Q) = \int_\Omega \mathcal{F}_B(Q) + \mathcal{F}_E(Q)$$

in equilibrium situations. Here $\Omega \in \mathbb{R}^d$, $d = 2, 3$, is the spatial domain occupied by the liquid crystal molecules, $\mathcal{F}_B$ is a bulk potential, and $\mathcal{F}_E$ is the elastic energy given

*Received by the editors December 1, 2020; accepted for publication (in revised form) May 9, 2022; published electronically August 15, 2022.
https://doi.org/10.1137/20M1383550

†Department of Mathematics, University of Wisconsin–Madison, Madison, WI 53706-1325 USA (gudibanda@wisc.edu).
‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA (franzisw@andrew.cmu.edu, yukuny@andrew.cmu.edu).

by

$$\mathcal{F}_B(Q) = \frac{a}{2}\operatorname{tr}(Q^2) - \frac{b}{3}\operatorname{tr}(Q^3) + \frac{c}{4}(\operatorname{tr}(Q^2))^2,$$
$$\mathcal{F}_E(Q) = \frac{L_1}{2}|\nabla Q|^2 + \frac{L_2}{2}|\operatorname{div} Q|^2 + \frac{L_3}{2}\sum_{i,j,k=1}^{d} \partial_i Q_{jk}\partial_k Q_{ji},$$

where a, b, c, L_1, L_2, L_3 are constants with $c, L_1, L_2, L_3 > 0$.

Nonequilibrium situations can be described by the gradient flow [1, 9],

$$\begin{aligned}\frac{\partial Q_{ij}}{\partial t} = M\Bigg(& L_1\Delta Q_{ij} + \frac{L_2+L_3}{2}\left(\sum_{k=1}^{d}(\partial_{ik}Q_{jk} + \partial_{jk}Q_{ik}) - \frac{2}{d}\sum_{k,\ell=1}^{d}\partial_{k\ell}Q_{k\ell}\delta_{ij}\right)\\ &- \left(aQ_{ij} - b\left((Q^2)_{ij} - \frac{1}{d}\operatorname{tr}(Q^2)\delta_{ij}\right) + c\operatorname{tr}(Q^2)Q_{ij}\right)\Bigg),\end{aligned} \tag{1.1}$$

where $M > 0$ is a constant, and one approach to obtaining equilibrium states is to follow this gradient flow. Adding the dynamics of the mean flow of the liquid crystal fluid to this, one obtains the Beris–Edwards system [2].

Analysis of the Q-tensor flow has been done (e.g., [7, 4, 15]), and numerical methods for the Q-tensor flow have been constructed in [17, 10, 8, 13]. To the best of our knowledge, none of these methods has been shown to be convergent to a weak solution of (1.1). An exception is the work by Cai, Shen, and Xu [3], where under the assumption of smallness of the initial data, convergence of a time discretization in two dimensions (2D) is proved. Our goal is to show convergence to weak solutions of a fully discrete method for (1.1) in 2D and 3D under only the natural assumption that the initial energy is bounded. Our numerical method is based on the invariant energy quadratization idea by Zhao et al. [17], which we combine with a finite difference discretization in space.

This method takes as a basis the reformulation of the Q-tensor flow using the auxiliary variable r:

$$r(Q) = \sqrt{2\left(\frac{a}{2}\operatorname{tr}(Q^2) - \frac{b}{3}\operatorname{tr}(Q^3) + \frac{c}{4}\operatorname{tr}^2(Q^2) + A_0\right)}, \tag{1.2}$$

where $A_0 > 0$ is a constant ensuring that r is positive. Defining

$$S(Q) = aQ - b\left[Q^2 - \frac{1}{d}\operatorname{tr}(Q^2)I\right] + c\operatorname{tr}(Q^2)Q, \tag{1.3}$$

it follows that

$$\frac{\delta r(Q)}{\delta Q} = \frac{S(Q)}{r(Q)} := P(Q) \tag{1.4}$$

for symmetric, trace free tensors Q. Then one can formally write the gradient flow (1.1) as a system for (Q, r):

$$Q_t = M\left(L_1\Delta Q + \frac{L_2+L_3}{2}\alpha(Q) - rP(Q)\right) := MH, \tag{1.5a}$$

$$r_t = P(Q) : Q_t, \tag{1.5b}$$

where

$$\alpha(Q)_{ij} = \sum_{k=1}^{d} (\partial_{ik} Q_{jk} + \partial_{jk} Q_{ik}) - \frac{2}{d} \sum_{k,\ell=1}^{d} \partial_{k\ell} Q_{k\ell} \delta_{ij}.$$

It is easy to see that this reformulation comes with a formal energy law: Multiplying the first equation (1.5a) with $-H$ and (1.5b) with r, adding and integrating, and integrating by parts, we obtain

$$\frac{d}{dt} \frac{1}{2} \int_\Omega \left(L_1 |\nabla Q|^2 + (L_2 + L_3) |\operatorname{div} Q|^2 + r^2 \right) dx = -M \int_\Omega |H|^2 dx. \tag{1.6}$$

In [17], a time discretization of the system (1.5) is proposed that retains a discrete version of the energy law (1.6). Based on this prior work, we propose a fully discrete finite difference method for (1.5) and prove its convergence to weak solutions of (1.5) as defined in Definition 2.3.

We then proceed to showing that weak solutions of (1.5) are in fact weak solutions of (1.1) and so achieve convergence to the original system (1.1). To the best of our knowledge, this is the first convergence proof for a fully discrete numerical scheme discretizing (1.1). The proof is based on the derivation of discrete energy stability of the fully discrete scheme, then using this to derive the existence of a precompact sequence that allows us to pass to the limit in the approximations. We proceed to showing Lipschitz continuity of the function P and use a Lax–Wendroff type argument to show that the limit of the approximating sequence is a weak solution of (1.5). The last step is to show that weak solutions of (1.5) are in fact weak solutions of (1.1). We achieve this through showing that a weak form of the chain rule holds in this case. We conclude with numerical experiments in 2D. Our scheme and analysis is for the three-dimensional case but adaptations to 2D can be made easily.

2. Preliminaries.

NOTATION 2.1. *We introduce the following general notation for matrix-valued functions* $A, B : \mathbb{R}^d \to \mathbb{R}^{d\times d}$*:*

- $A : B = \sum_{i,j=1}^{d} A_{ij} B_{ij}$,
- $\langle A, B \rangle = \int_\Omega A : B \, dx$,
- $|A| := |A|_F = \sqrt{A : A}$,
- $\|A\|_{L^2}^2 = \int_\Omega |A|_F^2 \, dx$,
- $\partial_i A = (\partial_i A_{jk})_{jk}$, $\partial_i = \partial_{x_i}$,
- $\nabla A = (\partial_1 A, \ldots, \partial_d A)$,
- $|\nabla A|^2 = \sum_{i=1}^{d} |\partial_i A|_F^2$,
- $\|\nabla A\|_{L^2}^2 = \int_\Omega |\nabla A|^2 \, dx$.

We assume $\Omega \subset \mathbb{R}^d$ is a bounded, connected domain with Lipschitz boundary and $Q_0 : \Omega \to \mathbb{R}^{d\times d} \in (H^1(\Omega))^{d\times d}$ takes values in the symmetric trace-free $d \times d$ matrices and satisfies $Q_0|_{\partial\Omega} = 0$. Fix $T > 0$ an arbitrary time horizon. We then define weak solutions of (1.1) as follows.

DEFINITION 2.2. *By a weak solution of* (1.1)*, we mean a function* $Q : [0, T] \times \Omega \to \mathbb{R}^{d\times d}$ *that is trace-free and symmetric for every* (t, x) *and satisfies*

$$Q \in L^\infty(0, T; H^1(\Omega)), \quad Q_t \in L^2([0, T] \times \Omega),$$

and

$$\begin{aligned}
&\int_0^T \int_\Omega Q : \partial_t \varphi dx dt - \int_\Omega Q(T,x) : \varphi(T,x) dx + \int_\Omega Q_0(x) : \varphi(0,x) dx \\
&= M \int_0^T \int_\Omega \left(L_1 \sum_{i,j=1}^d \nabla Q_{ij} \cdot \nabla \varphi_{ij} \right. \\
&\quad \left. + \frac{L_2 + L_3}{2} \sum_{i,j,k=1}^d \left(\partial_k Q_{jk} \partial_i \varphi_{ij} + \partial_k Q_{ik} \partial_j \varphi_{ij} - \frac{2}{d} \partial_i Q_{ki} \partial_k \varphi_{jj} \right) \right) dx dt \\
&\quad + M \int_0^T \int_\Omega \left(aQ - b \left((Q^2) - \frac{1}{d} tr(Q^2) I \right) + c\, tr(Q^2) Q \right) : \varphi \, dx dt
\end{aligned} \tag{2.1}$$

for all smooth $\varphi = (\varphi_{ij})_{i,j=1}^d : [0,T] \times \Omega \to \mathbb{R}^{d \times d}$ *that are compactly supported within* Ω *for almost every* $t \in [0,T]$. *Furthermore,* Q *satisfies the energy inequality*

$$\begin{aligned}
&\frac{1}{2} \int_\Omega L_1 |\nabla Q(t,x)|^2 + (L_2 + L_3) |\operatorname{div} Q(t,x)|^2 + 2\mathcal{F}_B\left(Q(t,x)\right) dx \\
&\leq \frac{1}{2} \int_\Omega L_1 |\nabla Q_0|^2 + (L_2 + L_3) |\operatorname{div} Q_0|^2 + 2\mathcal{F}_B\left(Q_0\right) dx - M \int_0^t \int_\Omega |H(s,x)|^2 \, dx \, ds
\end{aligned} \tag{2.2}$$

for every $t \in [0,T]$.

Similarly, we define weak solutions of the reformulation (1.5).

Definition 2.3. *By a weak solution of* (1.5), *we mean a pair of functions* $Q : [0,T] \times \Omega \to \mathbb{R}^{d \times d}$ *and* $r : [0,T] \times \Omega \to \mathbb{R}$, *with* $Q(t,x)$ *trace-free and symmetric for every* (t,x), *and satisfying*

$$Q \in L^\infty(0,T; H^1(\Omega)), \quad Q_t \in L^2([0,T] \times \Omega), \quad r \in L^\infty(0,T; L^2(\Omega))$$

and

$$\begin{aligned}
&\int_0^T \int_\Omega Q : \partial_t \varphi dx dt - \int_\Omega Q(T,x) : \varphi(T,x) dx + \int_\Omega Q_0(x) : \varphi(0,x) dx \\
&= M \int_0^T \int_\Omega \left(L_1 \sum_{i,j=1}^d \nabla Q_{ij} \cdot \nabla \varphi_{ij} \right. \\
&\quad \left. + \frac{L_2 + L_3}{2} \sum_{i,j,k=1}^d \left(\partial_k Q_{jk} \partial_i \varphi_{ij} + \partial_k Q_{ik} \partial_j \varphi_{ij} - \frac{2}{d} \partial_i Q_{ki} \partial_k \varphi_{jj} \right) \right) dx dt \\
&\quad + M \int_0^T \int_\Omega r P(Q) : \varphi \, dx dt,
\end{aligned} \tag{2.3}$$

and

$$\int_0^T \int_\Omega r\, \phi_t dx dt - \int_\Omega r(T,x) \phi(T,x) dx + \int_\Omega r_0(x) \phi(0,x) dx = - \int_0^T \int_\Omega P(Q) : Q_t \, \phi \, dx dt, \tag{2.4}$$

where $r_0 = r(Q_0)$ *and* $P(Q)$ *is defined in* (1.4) *for all smooth* $\varphi = (\varphi_{ij})_{i,j=1}^d : [0,T] \times \Omega \to \mathbb{R}^{d \times d}$ *and* $\phi : [0,T] \times \Omega \to \mathbb{R}$ *that are compactly supported within* Ω *for every* $t \in [0,T]$. *Furthermore,* (Q,r) *satisfies for a.e.* $t \in [0,T]$ *the energy inequality*

$$\frac{1}{2}\int_\Omega L_1|\nabla Q(t,x)|^2 + (L_2+L_3)|\operatorname{div} Q(t,x)|^2 + |r(t,x)|^2\,dx \tag{2.5}$$
$$\leq \frac{1}{2}\int_\Omega L_1|\nabla Q_0|^2 + (L_2+L_3)|\operatorname{div} Q_0|^2 + |r_0|^2\,dx - M\int_0^t\int_\Omega |H(s,x)|^2\,dx\,ds.$$

3. The numerical scheme. We start by introducing notation to define our numerical scheme. We let $\Delta t > 0$ be a time step size and $t^n := n\Delta t$ time levels at which we intend to compute approximations. For the ease of notation, we present the scheme for the case $\Omega = [0,1]^3$ and $h > 0$ is a uniform grid size in each spatial dimension. Extensions to square prisms of different side lengths and nonuniform grid sizes are not hard but notationally cumbersome, and therefore we restrict our analysis to the cube in $\mathbb{R}^3$ and uniform mesh sizes. The two-dimensional case can easily be derived from the three-dimensional scheme presented here. We let $x_{ijk} = (x_i, y_j, z_k) = (ih, jh, kh)$ be grid points, $i,j,k = 0,\dots,N+1$, with $N+1 = 1/h \in \mathbb{N}$. For approximations $(f_{ijk})_{ijk=0}^{N+1}$ on this grid, we define the averages

$$f_{ijk}^{n+\frac{1}{2}} = \frac{f_{ijk}^{n+1} + f_{ijk}^n}{2}, \quad \overline{f_{ijk}}^{n+\frac{1}{2}} = \frac{3}{2}f_{ijk}^n - \frac{1}{2}f_{ijk}^{n-1}$$

and difference operators

(3.1)
$$D_t^\pm f_{ijk}^n = \pm\frac{f_{ijk}^{n\pm1} - f_{ijk}^n}{\Delta t},$$
$$D_1^\pm f_{ijk}^n = \pm\frac{f_{i\pm1,j,k}^n - f_{ijk}^n}{h}, \quad D_2^\pm f_{ijk}^n = \pm\frac{f_{i,j\pm1,k}^n - f_{ijk}^n}{h}, \quad D_3^\pm f_{ijk}^n = \pm\frac{f_{i,j,k\pm1}^n - f_{ijk}^n}{h},$$
$$D_1^c f_{ijk}^n = \frac{f_{i+1,j,k}^n - f_{i-1,j,k}^n}{2h}, \quad D_2^c f_{ijk}^n = \frac{f_{i,j+1,k}^n - f_{i,j-1,k}^n}{2h}, \quad D_3^c f_{ijk}^n = \frac{f_{i,j,k+1}^n - f_{i,j,k-1}^n}{2h}$$

for $i,j,k = 1,\dots,N$. We will also need the discrete gradient, Laplacian, and divergence operators:

$$\nabla_h^\pm f_{ijk} = (D_1^\pm f_{ijk}, D_2^\pm f_{ijk}, D_3^\pm f_{ijk})^\top,$$
$$\Delta_h f_{ijk} = \sum_{\alpha=1}^3 D_\alpha^- D_\alpha^+ f_{ijk}, \quad (\operatorname{div}_h f_{ijk})_\beta = \sum_{\alpha=1}^3 D_\alpha^c (f_{ijk})_{\alpha\beta},$$

where $(f_{ijk})_{\alpha\beta}$ is the (α,β)-entry of the 3×3-matrix f_{ijk}. We approximate the initial data using cell averages,

$$Q_{ijk}^0 = \frac{1}{h^3}\int_{\mathcal{C}_{ijk}} Q_0(x)dx, \quad r_{ijk}^0 = \frac{1}{h^3}\int_{\mathcal{C}_{ijk}} r(Q_0(x))dx, \quad i,j,k = 1,\dots,N,$$

where $\mathcal{C}_{ijk} = [x_i - 0.5h, x_i + 0.5h) \times [y_j - 0.5h, y_j + 0.5h) \times [z_k - 0.5h, z_k + 0.5h)$, for $x_{ijk} = (x_i, y_j, z_k)$, and use Dirichlet boundary conditions

(3.2)
$$Q_{0,j,k} = Q_{N+1,j,k} = Q_{i,0,k} = Q_{i,N+1,k} = Q_{i,j,0} = Q_{i,j,N+1} = 0, \;\; i,j,k = 0,1,\dots,N,N+1.$$

For ease of notation throughout, we will also impose boundary conditions on ghost nodes

(3.3)
$$Q_{-1,j,k} = Q_{N+2,j,k} = Q_{i,-1,k} = Q_{i,N+2,k} = Q_{i,j,-1} = Q_{i,j,N+2} = 0,$$
$$i,j,k = 0,1,\dots,N,N+1.$$

We then propose the following method:

$$\begin{cases} D_t^+ Q_{ijk}^n = M\left(L_1 \Delta_h Q_{ijk}^{n+\frac{1}{2}} - r_{ijk}^{n+\frac{1}{2}} \overline{P}_{ijk}^{n+\frac{1}{2}} + \frac{L_2+L_3}{2} \alpha_{ijk}^{n+\frac{1}{2}}\right) := M H_{ijk}^{n+\frac{1}{2}}, \\ r_{ijk}^{n+1} - r_{ijk}^n = \overline{P}_{ijk}^{n+\frac{1}{2}} : (Q_{ijk}^{n+1} - Q_{ijk}^n). \end{cases} \tag{3.4}$$

Here, Q_{ijk}^n is an approximation for Q and r_{ijk}^n is an approximation of r at spatial point (x_i, y_j, z_k) and time step n. We defined $\alpha_{ijk}^{n+\frac{1}{2}} = \frac{\alpha^{n+1}+\alpha^n}{2}$, where $\alpha^n = \alpha_h(Q_{ijk}^n)$ and $\alpha^{n+1} = \alpha_h(Q_{ijk}^{n+1})$ and α_h is a discretization of $\alpha(Q)$:

$$\left(\alpha_h(Q_{ijk}^n)\right)_{ws} = \sum_{\beta=1}^{3} \left[D_w^c D_\beta^c \left(Q_{ijk}^n\right)_{s\beta} + D_s^c D_\beta^c \left(Q_{ijk}^n\right)_{w\beta} \right] - \frac{2}{3} \sum_{\beta,\gamma=1}^{3} D_\beta^c D_\gamma^c \left(Q_{ijk}^n\right)_{\beta\gamma} \delta_{ws}, \tag{3.5}$$

where the notation $(Q_{ijk}^n)_{ws}$ indicates the element in row w and column s of the matrix Q_{ijk}^n.

4. Analysis of the numerical scheme. For the proof of the energy stability of this scheme, we will need the following useful lemma, which is proved in the appendix (as Lemma A.1).

LEMMA 4.1. *Let A_{ijk} and B_{ijk} be scalar quantities at grid point (x_i, y_j, z_k) such that $A_{ijk} = 0$ at boundary values, i.e., boundary conditions* (3.2), (3.3). *Then*

$$\sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^+ B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^- A_{ijk}, \qquad \sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^- B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^+ A_{ijk},$$

and

$$\sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^c B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^c A_{ijk}$$

for $\beta = 1, 2$ or 3.

4.1. Energy stability. We start by defining the following norms and seminorms for difference approximations. For sequences of approximations $\{f_{ijk}\}$, $\{g_{ijk}\}$, $\{A_{ijk}\}$, and $\{B_{ijk}\}$ of scalar- or vector-valued functions $f, g : \Omega \to \mathbb{R}^d$ and matrix-valued functions $A, B : \Omega \to \mathbb{R}^{d\times d}$, defined on our grid, we let

$$\langle A, B\rangle_h = h^3 \sum_{i,j,k=0}^{N+1} A_{ijk} : B_{ijk}, \qquad \langle f, g\rangle_h = h^3 \sum_{i,j,k=0}^{N+1} f_{ijk} \cdot g_{ijk},$$

$$\|A\|_h^2 = h^3 \sum_{i,j,k=0}^{N+1} |A_{ijk}|_F^2, \qquad \|f\|_h^2 = h^3 \sum_{i,j,k=0}^{N+1} |f_{ijk}|^2,$$

$$\|\nabla_h A\|_h^2 = \sum_{m=1}^{d} \left\|D_m^- A\right\|_h^2.$$

We start by using Lemma 4.1 to show some simple summation by parts identities that will be useful later in the proofs of the energy stability of the scheme.

LEMMA 4.2. *Let* $\{A_{ijk}\}$ *and* $\{B_{ijk}\}$ *be grid functions satisfying homogeneous Dirichlet boundary conditions* (3.2). *Then*

$$\langle A, \Delta_h B\rangle_h = -\langle \nabla_h A, \nabla_h B\rangle_h.$$

Proof. We write

$$\begin{aligned}\langle A, \Delta_h B\rangle_h &= h^3 \sum_{i,j,k=0}^{N+1} \sum_{\alpha=1}^{3} A_{ijk} : (D_\alpha^+ D_\alpha^- B_{ijk}) \\ &= -h^3 \sum_{i,j,k=0}^{N+1} \sum_{\alpha=1}^{3} (D_\alpha^- A_{ijk}) : (D_\alpha^- B_{ijk}) = -\langle \nabla_h A, \nabla_h B\rangle_h,\end{aligned}$$

where we used Lemma 4.1 and the boundary conditions for the second equality. ∎

For the α-term, we have the following.

LEMMA 4.3. *Let* $\{A_{ijk}\}$ *and* $\{B_{ijk}\}$ *be symmetric and trace-free grid functions satisfying homogeneous Dirichlet boundary conditions,* (3.2). *Then*

$$\langle A, \alpha_h(B)\rangle_h = -2\langle \operatorname{div}_h A, \operatorname{div}_h B\rangle_h.$$

Proof. We compute (denoting $\alpha_{ijk} := \alpha_h(B_{ijk})$)

$$\begin{aligned}\langle A, \alpha_h(B)\rangle_h &= h^3 \sum_{i,j,k=0}^{N+1} \sum_{w,s=1}^{3} (A_{ijk})_{ws} (\alpha_{ijk})_{ws} \\ &= h^3 \sum_{i,j,k=0}^{N+1} \Big(\sum_{w,s=1}^{3} \sum_{\beta=1}^{3} \Big[(A_{ijk})_{ws} D_w^c D_\beta^c (B_{ijk})_{s\beta} + (A_{ijk})_{ws} D_s^c D_\beta^c (B_{ijk})_{w\beta} \Big] \\ &\qquad - \frac{2}{3} \sum_{w,s=1}^{3} \sum_{\beta,\gamma=1}^{3} (A_{ijk})_{ws} D_\beta^c D_\gamma^c (B_{ijk})_{\beta\gamma} \delta_{ws} \Big).\end{aligned}$$

Focusing on the last term of the inner sum and using that $\delta_{ws} = 1 \Leftrightarrow w = s$

$$\begin{aligned}\sum_{w,s=1}^{3} \sum_{\beta,\gamma=1}^{3} (A_{ijk})_{ws} D_\beta^c D_\gamma^c (B_{ijk})_{\beta\gamma} \delta_{ws} &= \sum_{\beta,\gamma=1}^{3} \sum_{w=1}^{3} (A_{ijk})_{ww} D_\beta^c D_\gamma^c (B_{ijk})_{\beta\gamma} \\ &= \sum_{\beta,\gamma=1}^{3} D_\beta^c D_\gamma^c (B_{ijk})_{\beta\gamma} \sum_{w=1}^{3} (A_{ijk})_{ww} \\ &= 0,\end{aligned} \tag{4.1}$$

where in the last equality we used that A_{ijk} is trace-free. For the first two terms, we use Lemma 4.1 and the symmetry assumption to obtain

$$\begin{aligned}\langle A, \alpha_h(B)\rangle_h &= -h^3 \sum_{i,j,k=0}^{N+1} \sum_{w,s,\beta=1}^{3} \Big[D_w^c (A_{ijk})_{ws} D_\beta^c (B_{ijk})_{s\beta} + D_s^c (A_{ijk})_{ws} D_\beta^c (B_{ijk})_{w\beta} \Big] \\ &= -h^3 \sum_{i,j,k=0}^{N+1} \sum_{w,s,\beta=1}^{3} \Big[D_w^c (A_{ijk})_{sw} D_\beta^c (B_{ijk})_{s\beta} + D_s^c (A_{ijk})_{ws} D_\beta^c (B_{ijk})_{w\beta} \Big] \\ &= -2 \langle \operatorname{div}_h A, \operatorname{div}_h B\rangle_h,\end{aligned}$$

which completes the proof. ∎

Next, we show that the scheme preserves the trace-free and symmetry property of Q. To this end, we rewrite the scheme (3.4) as $\mathbb{A}(Q^{n+1}) = \mathbb{F}(Q^n)$, where

(4.2)
$$\begin{cases} \mathbb{A}(Q_{ijk}^{n+1}) & = \frac{Q_{ijk}^{n+1}}{\Delta t} - \frac{ML_1}{2}\Delta_h Q_{ijk}^{n+1} + \frac{M}{2}\left(\overline{P}_{ijk}^{n+\frac{1}{2}} : Q_{ijk}^{n+1}\right)\overline{P}_{ijk}^{n+\frac{1}{2}} - M\frac{L_2+L_3}{4}\alpha_h(Q_{ijk}^{n+1}), \\ \mathbb{F}(Q_{ijk}^{n}) & = \frac{Q_{ijk}^{n}}{\Delta t} + \frac{ML_1}{2}\Delta_h Q_{ijk}^{n} + \frac{M}{2}\left(\overline{P}_{ijk}^{n+\frac{1}{2}} : Q_{ijk}^{n}\right)\overline{P}_{ijk}^{n+\frac{1}{2}} - Mr_{ijk}^{n}\overline{P}_{ijk}^{n+\frac{1}{2}} \\ & \quad + M\frac{L_2+L_3}{4}\alpha_h(Q_{ijk}^{n}), \end{cases}$$

where we have used that

$$r_{ijk}^{n+\frac{1}{2}} = \frac{1}{2}\overline{P}_{ijk}^{n+\frac{1}{2}} : Q_{ijk}^{n+1} - \frac{1}{2}\overline{P}_{ijk}^{n+\frac{1}{2}} : Q_{ijk}^{n} + r_{ijk}^{n}.$$

PROPOSITION 4.4. *If Q^n and Q^{n-1} are trace-free and symmetric, then Q^{n+1} computed by the scheme* (3.4) *is also trace-free and symmetric.*

Proof. Since we assume that Q_{ijk}^n, Q_{ijk}^{n-1} are trace-free, it follows that also $\overline{P}_{ijk}^{n+\frac{1}{2}}$ is trace-free. Moreover, $\alpha_h(Q)$ is trace-free without any assumptions on Q. Hence, we find that $\operatorname{tr}(\mathbb{F}(Q_{ijk}^n)) = 0$. But since $\mathbb{A}(Q_{ijk}^{n+1}) = \mathbb{F}(Q_{ijk}^n)$, then we must have $\operatorname{tr}(\mathbb{A}(Q_{ijk}^{n+1})) = 0$. Hence,

$$\operatorname{tr}(\mathbb{A}(Q_{ijk}^{n+1})) = \frac{\operatorname{tr}(Q_{ijk}^{n+1})}{\Delta t} - \frac{ML_1}{4}\Delta_h \operatorname{tr}(Q_{ijk}^{n+1}) = 0.$$

Taking the inner product of this with $\operatorname{tr}(Q_{ijk}^{n+1})$ we then find

$$\frac{\left\|\operatorname{tr}(Q^{n+1})\right\|_h^2}{\Delta t} - \frac{ML_1}{4}\langle\Delta_h \operatorname{tr}(Q^{n+1}), \operatorname{tr}(Q^{n+1})\rangle_h = 0.$$

We use Lemma 4.1 for the second term:

$$\begin{aligned} \langle\Delta_h \operatorname{tr}(Q^{n+1}), \operatorname{tr}(Q^{n+1})\rangle_h &= h^3 \sum_{i,j,k=0}^{N+1}\sum_{\ell=1}^{3}\left(D_\ell^+ D_\ell^- \sum_{w=1}^{3}(Q_{ijk}^{n+1})_{ww}\right)\sum_{s=1}^{3}(Q_{ijk}^{n+1})_{ss} \\ &= -h^3 \sum_{i,j,k=0}^{N+1}\sum_{\ell=1}^{3}\left(D_\ell^- \sum_{w=1}^{3}(Q_{ijk}^{n+1})_{ww}\right)\left(D_\ell^- \sum_{s=1}^{3}(Q_{ijk}^{n+1})_{ss}\right) \\ &= -\left\|\nabla_h \operatorname{tr}(Q^{n+1})\right\|_h^2. \end{aligned}$$

Thus we must have that $\operatorname{tr}(Q_{ijk}^{n+1}) = 0$ for all $i, j, k = 1, \ldots, N$ and we see that the trace-free condition is preserved.

For the symmetry, we notice that if Q_{ijk}^n and Q_{ijk}^{n-1} are symmetric, then also $\overline{P}_{ijk}^{n+\frac{1}{2}}$ and $\alpha_h(Q_{ijk}^n)$ are symmetric. Hence $\mathbb{F}(Q_{ijk}^n) = (\mathbb{F}(Q_{ijk}^n))^\top$ and therefore $\mathbb{A}(Q_{ijk}^{n+1}) = (\mathbb{A}(Q_{ijk}^{n+1}))^\top$. Denoting $V_{ijk}^{n+1} := Q_{ijk}^{n+1} - (Q_{ijk}^{n+1})^\top$, this implies

$$\frac{V_{ijk}^{n+1}}{\Delta t} - \frac{ML_1}{2}\Delta_h V_{ijk}^{n+1} - M\frac{L_2+L_3}{4}\alpha_h(V_{ijk}^{n+1}) = 0.$$

Note that V_{ijk}^{n+1} is skew-symmetric and trace-free. We take the inner product with V_{ijk}^{n+1} and obtain

$$0 = \frac{\left\|V^{n+1}\right\|_h^2}{\Delta t} - \frac{ML_1}{2}\langle\Delta_h V^{n+1}, V^{n+1}\rangle_h - M\frac{L_2+L_3}{4}\langle\alpha_h(V^{n+1}), V^{n+1}\rangle_h.$$

Using Lemma 4.2, this can be rewritten as

$$0 = \frac{\left\|V^{n+1}\right\|_h^2}{\Delta t} + \frac{ML_1}{2}\left\|\nabla_h V^{n+1}\right\|_h^2 - M\frac{L_2+L_3}{2}\langle \alpha_h(V^{n+1}), V^{n+1}\rangle_h. \tag{4.3}$$

The term involving α on the right hand side (RHS) is

$$\begin{aligned}\langle V^{n+1}, \alpha_h(V^{n+1})\rangle_h = h^3 \sum_{i,j,k=0}^{N+1} \Bigg(\sum_{w,s=1}^{3}\sum_{\beta=1}^{3} \Bigg[&\left(V_{ijk}^{n+1}\right)_{ws} D_w^c D_\beta^c \left(V_{ijk}^{n+1}\right)_{s\beta} \\ &+ \left(V_{ijk}^{n+1}\right)_{ws} D_s^c D_\beta^c \left(V_{ijk}^{n+1}\right)_{w\beta}\Bigg] \\ &- \frac{2}{3}\sum_{w,s=1}^{3}\sum_{\beta,\gamma=1}^{3} \left(V_{ijk}^{n+1}\right)_{ws} D_\beta^c D_\gamma^c \left(V_{ijk}^{n+1}\right)_{\beta\gamma} \delta_{ws}\Bigg) \\ = h^3 \sum_{i,j,k=0}^{N+1}\sum_{w,s=1}^{3}\sum_{\beta=1}^{3} \Bigg[&\left(V_{ijk}^{n+1}\right)_{ws} D_w^c D_\beta^c \left(V_{ijk}^{n+1}\right)_{s\beta} \\ &+ \left(V_{ijk}^{n+1}\right)_{ws} D_s^c D_\beta^c \left(V_{ijk}^{n+1}\right)_{w\beta}\Bigg],\end{aligned}$$

using that V^{n+1} is trace-free, as in (4.1) (replacing A and B by V^{n+1}). Using Lemma 4.1 and the skew-symmetry of V^{n+1}, the remaining terms are

$$\begin{aligned}\langle V^{n+1}, \alpha_h(V^{n+1})\rangle_h &= -h^3 \sum_{i,j,k=0}^{N+1}\sum_{w,s=1}^{3}\sum_{\beta=1}^{3} \Bigg[D_w^c \left(V_{ijk}^{n+1}\right)_{ws} D_\beta^c \left(V_{ijk}^{n+1}\right)_{s\beta} \\ &\qquad + D_s^c \left(V_{ijk}^{n+1}\right)_{ws} D_\beta^c \left(V_{ijk}^{n+1}\right)_{w\beta}\Bigg] \\ &= h^3 \sum_{i,j,k=0}^{N+1}\sum_{w,s=1}^{3}\sum_{\beta=1}^{3} \Bigg[D_w^c \left(V_{ijk}^{n+1}\right)_{sw} D_\beta^c \left(V_{ijk}^{n+1}\right)_{s\beta} \\ &\qquad - D_s^c \left(V_{ijk}^{n+1}\right)_{ws} D_\beta^c \left(V_{ijk}^{n+1}\right)_{w\beta}\Bigg] = 0.\end{aligned}$$

Plugging this into (4.3), we see that $V_{ijk}^{n+1} = 0$ for all i, j, k. ∎

The next theorem guarantees the existence of a unique solution of the system of equations (3.4) (or (4.2)).

THEOREM 4.5. *The operator $\mathbb{A}$ is symmetric and positive definite for grid functions that are symmetric and trace-free.*

Proof. Let $Q^1 = (Q_{ijk}^1)_{ijk}$ and $Q^2 = (Q_{ijk}^2)_{ijk}$, and then

$$\begin{aligned}\langle \mathbb{A}(Q^1), Q^2\rangle_h = &\frac{1}{\Delta t}\langle Q^1, Q^2\rangle_h - \frac{ML_1}{2}\langle \Delta_h Q^1, Q^2\rangle_h + \frac{M}{2}\langle \overline{P}^{n+\frac{1}{2}} : Q^1, \overline{P}^{n+\frac{1}{2}} : Q^2\rangle_h \\ &- M\frac{L_2+L_3}{4}\langle \alpha_h(Q^1), Q^2\rangle_h.\end{aligned}$$

By Lemmas 4.2 and 4.3, we have

$$\langle \Delta_h Q^1, Q^2\rangle_h = -\langle \nabla_h Q^1, \nabla_h Q^2\rangle_h, \quad \langle \alpha_h(Q^1), Q^2\rangle_h = -2\langle \operatorname{div}_h Q^1, \operatorname{div}_h Q^2\rangle_h.$$

Therefore,

$$\begin{aligned}\langle \mathbb{A}(Q^1), Q^2\rangle_h =& \frac{1}{\Delta t}\langle Q^1, Q^2\rangle_h + \frac{ML_1}{2}\langle \nabla_h Q^1, \nabla_h Q^2\rangle_h + \frac{M}{2}\langle \overline{P}^{n+\frac{1}{2}} : Q^1, \overline{P}^{n+\frac{1}{2}} : Q^2\rangle_h \\ &+ \frac{M\left(L_2+L_3\right)}{2}\left\langle \mathrm{div}_h^- Q^1, \mathrm{div}_h^- Q^2\right\rangle_h,\end{aligned}$$

which we see is symmetric in Q^1 and Q^2. Moreover,

$$\begin{aligned}\langle \mathbb{A}(Q), Q\rangle_h = \frac{1}{\Delta t}\left\|Q\right\|_h^2 + \frac{ML_1}{2}\left\|\nabla_h Q\right\|_h^2 + \frac{M}{2}\left\|\overline{P}^{n+\frac{1}{2}} : Q\right\|_h^2 \\ + \frac{M\left(L_2+L_3\right)}{2}\left\|\mathrm{div}_h\, Q\right\|_h^2 \geq 0,\end{aligned}$$

with equality if and only if $Q \equiv 0$. ∎

By the previous two results, we conclude there is a unique solution Q^{n+1} to $\mathbb{A}(Q^{n+1}) = \mathbb{F}(Q^n)$ that is trace-free and symmetric.

Next, we will prove an energy estimate for the scheme.

THEOREM 4.6. *Define the energy*

$$E^n = \frac{L_1}{2}\left\|\nabla_h Q^n\right\|_h^2 + \frac{L_2+L_3}{2}\left\|\mathrm{div}_h\, Q^n\right\|_h^2 + \frac{1}{2}\left\|r^n\right\|_h^2, \tag{4.4}$$

where Q^n, r^n *solve* (3.4). *Then*

$$E^{n+1} - E^n = -\Delta t M\left\|H^{n+\frac{1}{2}}\right\|_h^2.$$

Proof. We take the inner product of the first equation in (3.4) with $-\Delta t H_{ijk}^{n+\frac{1}{2}}$, multiply by h^d, and sum over all grid points,

$$-\left\langle Q^{n+1} - Q^n, H^{n+\frac{1}{2}}\right\rangle_h = -\Delta t\, M\left\|H^{n+\frac{1}{2}}\right\|_h^2. \tag{4.5}$$

Taking the inner product of the second equation with $r_{ijk}^{n+\frac{1}{2}}$, multiplying by h^d, and summing over all grid points gives

$$\begin{aligned}&\left\langle r^{n+1} - r^n, r^{n+\frac{1}{2}}\right\rangle_h = \left\langle r^{n+\frac{1}{2}}\overline{P}^{n+\frac{1}{2}}, Q^{n+1} - Q^n\right\rangle_h \\ \iff\quad &\frac{1}{2}\left\|r^{n+1}\right\|_h^2 - \frac{1}{2}\left\|r^n\right\|_h^2 = \left\langle r^{n+\frac{1}{2}}\overline{P}^{n+\frac{1}{2}}, Q^{n+1} - Q^n\right\rangle_h.\end{aligned} \tag{4.6}$$

Next we shall work with the term $\langle Q^{n+1} - Q^n, H^{n+\frac{1}{2}}\rangle_h$.

$$\begin{aligned}\left\langle Q^{n+1} - Q^n, H^{n+\frac{1}{2}}\right\rangle_h =& \left\langle Q^{n+1} - Q^n, L_1\Delta_h Q^{n+\frac{1}{2}} - q^{n+\frac{1}{2}}\overline{P}^{n+\frac{1}{2}} + \frac{L_2+L_3}{2}\alpha^{n+\frac{1}{2}}\right\rangle_h \\ =& \frac{L_1}{2}\left\langle Q^{n+1} - Q^n, \Delta_h Q^{n+1} + \Delta_h Q^n\right\rangle_h - \left\langle Q^{n+1} - Q^n, q^{n+\frac{1}{2}}\overline{P}^{n+\frac{1}{2}}\right\rangle_h \\ &+ \frac{L_2+L_3}{4}\left\langle Q^{n+1} - Q^n, \alpha^{n+1} + \alpha^n\right\rangle_h.\end{aligned} \tag{4.7}$$

We shall deal with these terms individually. Using bilinearity, we have

$$\begin{aligned}\left\langle Q^{n+1} - Q^n, \Delta_h Q^{n+1} + \Delta_h Q^n\right\rangle_h =& \left\langle Q^{n+1}, \Delta_h Q^{n+1}\right\rangle_h + \left\langle Q^{n+1}, \Delta_h Q^n\right\rangle_h \\ &- \left\langle Q^n, \Delta_h Q^{n+1}\right\rangle_h - \left\langle Q^n, \Delta_h Q^n\right\rangle_h.\end{aligned} \tag{4.8}$$

We shall focus on the first term of the RHS. Using Lemma 4.2 and the boundary conditions, we have

$$\langle Q^{n+1}, \Delta_h Q^{n+1}\rangle_h = -\|\nabla_h Q^{n+1}\|_h^2.$$

Similarly, we see that $\langle Q^n, \Delta_h Q^n\rangle_h = -\|\nabla_h Q^n\|_h^2$, $\langle Q^n, \Delta_h Q^{n+1}\rangle_h = -\langle \nabla_h Q^n, \rangle \nabla_h Q^{n+1}{}_h$ and $\langle Q^{n+1}, \Delta_h Q^n\rangle_h = -\langle \nabla_h Q^{n+1}, \nabla_h Q^n\rangle_h$. Hence putting all these results into (4.8) we find

$$\begin{aligned}\langle Q^{n+1} - Q^n, \Delta_h Q^{n+1} + \Delta_h Q^n\rangle_h &\\ = -\left\|\nabla_h Q^{n+1}\right\|_h^2 - \langle \nabla_h Q^{n+1}, \nabla_h Q^n\rangle_h + \langle \nabla_h Q^n, \nabla_h Q^{n+1}\rangle_h + \|\nabla_h Q^n\|_h^2 &\\ = -\left\|\nabla_h Q^{n+1}\right\|_h^2 + \|\nabla_h Q^n\|_h^2 .&\end{aligned}$$

Next we consider the term $\langle Q^{n+1} - Q^n, \alpha^{n+1} + \alpha^n\rangle_h$ in (4.7):

$$\langle Q^{n+1} - Q^n, \alpha^{n+1} + \alpha^n\rangle_h = \langle Q^{n+1}, \alpha^{n+1}\rangle_h + \langle Q^{n+1}, \alpha^n\rangle_h - \langle Q^n, \alpha^{n+1}\rangle_h - \langle Q^n, \alpha^n\rangle_h.$$

Using Lemma 4.3 for each of these terms, we obtain

$$\langle Q^{n+1} - Q^n, \alpha^{n+1} + \alpha^n\rangle_h = -2\left\|\mathrm{div}_h\, Q^{n+1}\right\|_h^2 + 2\left\|\mathrm{div}_h\, Q^n\right\|_h^2 .$$

Overall we have shown that

$$\begin{aligned}\langle Q^{n+1} - Q^n, H^{n+\frac{1}{2}}\rangle_h = \frac{L_1}{2}\left(-\left\|\nabla_h Q^{n+1}\right\|_h^2 + \|\nabla_h Q^n\|_h^2\right)&\\ + \frac{L_2 + L_3}{2}\left(-\left\|\mathrm{div}_h\, Q^{n+1}\right\|_h^2 + \left\|\mathrm{div}_h\, Q^n\right\|_h^2\right)&\\ - \langle Q^{n+1} - Q^n, r^{n+\frac{1}{2}}\overline{P}^{n+\frac{1}{2}}\rangle_h.&\end{aligned}$$

Combining this with (4.5) and (4.6), we obtain

$$\begin{aligned}\frac{L_1}{2}\left(\left\|\nabla_h Q^{n+1}\right\|_h^2 - \|\nabla_h Q^n\|_h^2\right) + \frac{L_2 + L_3}{2}\left(\left\|\mathrm{div}_h\, Q^{n+1}\right\|_h^2 - \left\|\mathrm{div}_h\, Q^n\right\|_h^2\right)&\\ + \frac{1}{2}\left(\left\|r^{n+1}\right\|_h^2 - \|r^n\|_h^2\right) = -\Delta t\, M\left\|H^{n+\frac{1}{2}}\right\|_h^2 .&\quad \square\end{aligned}$$

Based on the energy estimate, Theorem 4.6, we can derive further stability bounds on the approximations $\{Q_{ijk}^n\}$ and $\{r_{ijk}^n\}$. Specifically, it follows from the bound on $\{H^{n+\frac{1}{2}}\}$ that $D_t^+ Q_{ijk}^n = M H_{ijk}^{n+\frac{1}{2}}$ is bounded:

COROLLARY 4.7. *We have*

$$\Delta t \sum_{n=0}^{N_T - 1} \left\|D_t^+ Q^n\right\|_h^2 \le E^0,$$

where N_T is such that $T = N_T \Delta t$.

Using this corollary and the energy estimate, we can also derive a uniform (in Δt and h) bound on $\|Q^n\|_h$.

LEMMA 4.8. *The following estimate holds for any $\Delta t, h > 0$:*

$$\|Q^m\|_h \le T^{\frac{1}{2}} \left(\Delta t \sum_{n=0}^{m-1} \left\| D_t^+ Q^n \right\|_h^2 \right)^{\frac{1}{2}} + \|Q^0\|_h < \infty$$

for any $0 \le m \le N_T$, where N_T is such that $N_T \Delta t = T$.

Proof. Note that

$$\begin{aligned}
\left\|Q^{n+1}\right\|_h^2 - \|Q^n\|_h^2 &= h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^{n+1}|_F^2 - h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^n|_F^2 \\
&= h^3 \sum_{i,j,k=0}^{N+1} (Q_{ijk}^{n+1} + Q_{ijk}^n) : (Q_{ijk}^{n+1} - Q_{ijk}^n) \\
&\le \left[h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^{n+1} + Q_{ijk}^n|_F^2 \right]^{\frac{1}{2}} \left[h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^{n+1} - Q_{ijk}^n|_F^2 \right]^{\frac{1}{2}} \\
&\le \left[\left(h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^{n+1}|_F^2 \right)^{\frac{1}{2}} + \left(h^3 \sum_{i,j,k=0}^{N+1} |Q_{ijk}^n|_F^2 \right)^{\frac{1}{2}} \right] \left[h^3 \sum_{i,j,k=0}^{N+1} \left| \frac{Q_{ijk}^{n+1} - Q_{ijk}^n}{\Delta t} \right|_F^2 \right]^{\frac{1}{2}} \Delta t \\
&= \left(\left\|Q^{n+1}\right\|_h + \|Q^n\|_h \right) \|D_t^+ Q^n\|_h \, \Delta t,
\end{aligned}$$

and so

$$\left\|Q^{n+1}\right\|_h - \|Q^n\|_h \le \|D_t^+ Q^n\|_h \, \Delta t$$

for any $0 \le n \le N_T$. Summing over n on both sides, we obtain that for any $0 \le m \le N_T$,

$$\begin{aligned}
\|Q^m\|_h \le \sum_{n=0}^{m-1} \|D_t^+ Q^n\|_h \, \Delta t + \|Q^0\|_h &\le \left(\sum_{n=0}^{m-1} \Delta t \right)^{\frac{1}{2}} \left(\sum_{n=0}^{m-1} \|D_t^+ Q^n\|_h^2 \, \Delta t \right)^{\frac{1}{2}} + \|Q^0\|_h \\
&\le T^{\frac{1}{2}} \left(\sum_{n=0}^{m-1} \|D_t^+ Q^n\|_h^2 \, \Delta t \right)^{\frac{1}{2}} + \|Q^0\|_h.
\end{aligned}$$

∎

4.2. Lipschitz continuity of $P(Q)$. In order to derive a stability bound on $\{D_t^+ r_{ijk}^n\}$, we need an auxiliary result, which is the Lipschitz continuity of $P(Q)$. Recall that we can write $P(Q)$ as $P(Q) = \frac{S(Q)}{r(Q)}$, where S and r have been defined in (1.3) and (1.2). Note that we can express the Frobenius norm as

$$|Q|_F = \sqrt{\operatorname{tr}(Q^2)} = \sqrt{\sum_{i=1}^{d} \lambda_i^2},$$

where λ_i is the ith eigenvalue of matrix Q.

We start with a few preliminary lemmas. First, note that $r(Q)$ is bounded from below by some constant $A > 0$ (see also [17, Theorem 2.1]). We will also need an

upper bound. Since $c > 0$, there exists constant $K_1 > 0$ such that for any Q for which $|Q|_F \geq K_1$,

$$\left| a \operatorname{tr}(Q^2) - \frac{2b}{3} \operatorname{tr}(Q^3) + 2A_0 \right| \leq \frac{c}{4} \operatorname{tr}^2(Q^2).$$

Then

$$\begin{aligned} r(Q) \geq \sqrt{\frac{c}{2}\operatorname{tr}^2(Q^2) - \left| a\operatorname{tr}(Q^2) - \frac{2b}{3}\operatorname{tr}(Q^3) + 2A_0 \right|} &\geq \sqrt{\frac{c}{2}\operatorname{tr}^2(Q^2) - \frac{c}{4}\operatorname{tr}^2(Q^2)} \\ &= \frac{\sqrt{c}}{2} \operatorname{tr}(Q^2), \end{aligned}$$

and

$$\begin{aligned} r(Q) \leq \sqrt{\frac{c}{2}\operatorname{tr}^2(Q^2) + \left| a \operatorname{tr}(Q^2) - \frac{2b}{3}\operatorname{tr}(Q^3) + 2A_0 \right|} &\leq \sqrt{\frac{c}{2}\operatorname{tr}^2(Q^2) + \frac{c}{4}\operatorname{tr}^2(Q^2)} \\ &\leq \sqrt{c}\operatorname{tr}(Q^2). \end{aligned}$$

So whenever $|Q|_F \geq K_1$, we can bound $r(Q)$ by

$$\frac{\sqrt{c}}{2} |Q|_F^2 = \frac{\sqrt{c}}{2} \operatorname{tr}(Q^2) \leq r(Q) \leq \sqrt{c} \operatorname{tr}(Q^2) = \sqrt{c}\, |Q|_F^2. \tag{4.9}$$

On the other hand, when Q is bounded by constant K_1, we have

$$\begin{aligned} r(Q) &\leq \sqrt{2\left[\frac{|a|}{2}\operatorname{tr}(Q^2) + \frac{|b|}{3}|\operatorname{tr}(Q^3)| + \frac{c}{4}\operatorname{tr}^2(Q^2) + A_0\right]} \\ &\leq \sqrt{2\left(\frac{|a|}{2}K_1^2 + \frac{|b|}{3}K_1^3 + \frac{c}{4}K_1^4 + A_0\right)} \triangleq K_2, \end{aligned}$$

where we have used the fact that

$$\operatorname{tr}(Q^4) = \sum_{i=1}^{d} \lambda_i^4 \leq \left(\sum_{i=1}^{d} \lambda_i^2 \right)^2 \leq \operatorname{tr}^2(Q^2),$$

and then

$$|\operatorname{tr}(Q^3)| \leq \operatorname{tr}^{\frac{1}{2}}(Q^4) \operatorname{tr}^{\frac{1}{2}}(Q^2) \leq \operatorname{tr}^{\frac{3}{2}}(Q^2) = K_1^3.$$

Combining the two results, we obtain that

$$r(Q) \leq K_2 + \sqrt{c}|Q|_F^2 \tag{4.10}$$

for some constant $K_2 > 0$. This bound will be used subsequently. The following lemmas are important steps toward our Lipschitz estimate for $P(Q)$.

LEMMA 4.9. *For any Q, there exist constants C_1, C_2, and C_3 such that*

$$\frac{1}{r(Q)} \leq C_1, \quad \frac{|Q|_F}{r(Q)} \leq C_2, \quad \frac{|Q|_F^2}{r(Q)} \leq C_3.$$

Proof. The first estimate follows from the fact that $r(Q)$ is bounded from below. For the third estimate, we split it into two cases. When $|Q|_F \le K_1$, we have

$$\frac{|Q|_F^2}{r(Q)} \le \frac{K_1^2}{A}.$$

When $|Q|_F \ge K_1$, by (4.9), we know that

$$\frac{|Q|_F^2}{r(Q)} \le \frac{|Q|_F^2}{\frac{\sqrt{c}}{2}|Q|_F^2} \le \frac{2}{\sqrt{c}}.$$

Define $C_3 \triangleq \max\left\{\frac{K_1^2}{A}, \frac{2}{\sqrt{c}}\right\}$, and then $\frac{|Q|_F^2}{r(Q)} \le C_3$. To prove the second estimate, we note that if $|Q|_F \le 1$, then

$$\frac{|Q|_F}{r(Q)} \le \frac{1}{r(Q)} \le C_1.$$

Else, we have that

$$\frac{|Q|_F}{r(Q)} \le \frac{|Q|_F^2}{r(Q)} \le C_3.$$

Defining $C_2 \triangleq \max\{C_1, C_3\}$, we obtain $\frac{|Q|_F}{r(Q)} \le C_2$, which completes the proof of the lemma. ∎

LEMMA 4.10. *For any matrix Q, $|\frac{S(Q)}{r(Q)^{\frac{3}{2}}}|_F$ is uniformly bounded.*

Proof. Note that

$$\begin{aligned}|S(Q)|_F &= \left|aQ - b\left[Q^2 - \frac{1}{d}\mathrm{tr}(Q^2)I\right] + c\mathrm{tr}(Q^2)Q\right|_F \\ &\le |a||Q|_F + |b||Q^2|_F + \frac{|b|}{d}|Q|_F^2 + c|Q|_F^2|Q|_F.\end{aligned}$$

Since

$$|Q^2|_F = \sqrt{\mathrm{tr}(Q^4)} \le \sqrt{\mathrm{tr}^2(Q^2)} = \mathrm{tr}(Q^2) = |Q|_F^2,$$

we obtain

$$|S(Q)|_F \le |a||Q|_F + \frac{(d+1)|b|}{d}|Q|_F^2 + c\,|Q|_F^3.$$

Then from Lemma 4.9, we obtain that

$$\begin{aligned}\left|\frac{S(Q)}{r(Q)^{\frac{3}{2}}}\right|_F &\le \frac{|a|\,|Q|_F + \frac{(d+1)|b|}{d}|Q|_F^2 + c\,|Q|_F^3}{r(Q)^{\frac{3}{2}}} \\ &\le \frac{|a|}{r(Q)^{\frac{1}{2}}}\frac{|Q|_F}{r(Q)} + \frac{\frac{(d+1)|b|}{d}}{r(Q)^{\frac{1}{2}}}\frac{|Q|_F^2}{r(Q)} + c\left(\frac{|Q|_F^2}{r(Q)}\right)^{\frac{3}{2}} \\ &\le \frac{|a|}{A^{\frac{1}{2}}}C_2 + \frac{(d+1)|b|}{dA^{\frac{1}{2}}}C_3 + c\,C_3^{\frac{3}{2}} \triangleq K_3,\end{aligned}$$

which proves the lemma. ∎

Now we are in a position to prove that $P(Q)$ is Lipschitz continuous with respect to the Frobenius norm.

THEOREM 4.11. *There exists a constant $L > 0$ such that for any matrices $Q, \delta Q \in \mathbb{R}^{3\times 3}$,*

$$|P(Q+\delta Q) - P(Q)|_F \le L|\delta Q|_F.$$

Proof. We will split the proof into two cases.

Case 1. δQ is so large such that $|\delta Q|_F \ge 2|Q|_F$ and $|\delta Q|_F \ge \max\{2K_1, K_3\sqrt{K_2}\} \triangleq G$. In this case, we can see that

$$|\delta Q + Q|_F \ge |\delta Q|_F - |Q|_F \ge \frac{1}{2}|\delta Q|_F \ge K_1,$$

and therefore, by (4.9), we have

$$r(Q+\delta Q) \le \sqrt{c}|Q+\delta Q|_F \le \sqrt{c}(|Q|_F + |\delta Q|_F) \le \frac{3\sqrt{c}}{2}|\delta Q|_F. \tag{4.11}$$

We use this to compute the difference between $P(Q+\delta Q)$ and $P(Q)$,

$$\begin{aligned}
|P(Q+\delta Q) - P(Q)|_F &\le |P(Q+\delta Q)|_F + |P(Q)|_F \\
&= \left|\frac{S(Q+\delta Q)}{r(Q+\delta Q)}\right|_F + \left|\frac{S(Q)}{r(Q)}\right|_F \\
&= \left|\frac{S(Q+\delta Q)}{r(Q+\delta Q)^{\frac{3}{2}}}\right|_F \sqrt{r(Q+\delta Q)} + \left|\frac{S(Q)}{r(Q)^{\frac{3}{2}}}\right|_F \sqrt{r(Q)} \\
&\overset{\text{Lem 4.9}}{\le} K_3\sqrt{r(Q+\delta Q)} + K_3\sqrt{r(Q)} \\
&\overset{(4.10),(4.11)}{\le} \frac{3K_3\sqrt{c}}{2}|\delta Q|_F + K_3(\sqrt{K_2} + c^{\frac{1}{4}}|Q|_F) \\
&\le \frac{3K_3\sqrt{c}}{2}|\delta Q|_F + |\delta Q|_F + \frac{K_3 c^{\frac{1}{4}}}{2}|\delta Q|_F \\
&= \left(\frac{3K_3\sqrt{c}}{2} + 1 + \frac{K_3 c^{\frac{1}{4}}}{2}\right)|\delta Q|_F,
\end{aligned}$$

which proves the result in this case.

Case 2. $|\delta Q|_F \le 2|Q|_F$ or $|\delta Q|_F \le G$.

In this case, we write the difference of $P(Q+\delta Q)$ and $P(Q)$ as

$$\begin{aligned}
|P(Q+\delta Q) - P(Q)|_F &= \left|\frac{S(Q+\delta Q)}{r(Q+\delta Q)} - \frac{S(Q)}{r(Q)}\right|_F \\
&= \left|\frac{S(Q+\delta Q) - S(Q)}{r(Q)} + S(Q+\delta Q)\left(\frac{1}{r(Q+\delta Q)} - \frac{1}{r(Q)}\right)\right|_F \\
&\le \underbrace{\left|\frac{S(Q+\delta Q) - S(Q)}{r(Q)}\right|_F}_{\text{I}} + \underbrace{\left|\frac{S(Q+\delta Q)}{r(Q+\delta Q)^{\frac{3}{2}}}\,\frac{\sqrt{r(Q+\delta Q)}\,[r(Q+\delta Q) - r(Q)]}{r(Q)}\right|_F}_{\text{II}}.
\end{aligned}$$

To compute I, we expand $S(Q+\delta Q)$ by plugging $(Q+\delta Q)$ into (1.3):

$$\begin{aligned}S(Q+\delta Q) =& S(Q) + a\,\delta Q - b\,(Q\,\delta Q + \delta Q\,Q) - b\,(\delta Q)^2 + \frac{2b}{d}\,\mathrm{tr}\,(Q\delta Q)\,I \\ &+ \frac{b}{d}\,\mathrm{tr}\left((\delta Q)^2\right) I + c\,\mathrm{tr}((\delta Q)^2)\delta Q + 2c\,\mathrm{tr}(Q\,\delta Q)Q \\ &+ 2c\,\mathrm{tr}(Q\,\delta Q)\delta Q + c\,\mathrm{tr}(Q^2)\delta Q + c\,\mathrm{tr}\left((\delta Q)^2\right) Q.\end{aligned}$$

Then by Lemma 4.9, we have

$$\begin{aligned}\mathrm{I} \le& |a|\frac{|\delta Q|_F}{r(Q)} + \frac{2(d+1)|b|}{d}\frac{|Q|_F|\delta Q|_F}{r(Q)} + \frac{(d+1)|b|}{d}\frac{|\delta Q|_F^2}{r(Q)} \\ &+ c\frac{|\delta Q|_F^3 + 3|Q|_F^2|\delta Q|_F + 3|\delta Q|_F^2|Q|_F}{r(Q)} \\ \le& |a|\,C_1\,|\delta Q|_F + \frac{2(d+1)C_2|b|}{d}|\delta Q|_F + \frac{(d+1)|b|}{d}\frac{|\delta Q|_F^2}{r(Q)} \\ &+ 3c\,C_3|\delta Q|_F + c\,\frac{|\delta Q|_F^3}{r(Q)} + 3c\frac{|Q|_F|\delta Q|_F^2}{r(Q)}.\end{aligned} \tag{4.12}$$

We still need to bound $\frac{|\delta Q|_F^2}{r(Q)}$, $\frac{|\delta Q|_F^3}{r(Q)}$, and $\frac{|Q|_F|\delta Q|_F^2}{r(Q)}$ in terms of δQ. Based on our assumption in this case, if $|\delta Q|_F \le 2|Q|_F$, then

$$\begin{aligned}\frac{|\delta Q|_F^2}{r(Q)} \le& \frac{2|Q|_F\,|\delta Q|_F}{r(Q)} \le 2C_2\,|\delta Q|_F, \\ \frac{|\delta Q|_F^3}{r(Q)} \le& \frac{4|Q|_F^2|\delta Q|_F}{r(Q)} \le 4C_3\,|\delta Q|_F, \\ \text{and}\quad \frac{|Q|_F|\delta Q|_F^2}{r(Q)} \le& \frac{2|Q|_F^2|\delta Q|_F}{r(Q)} \le 2C_3|\delta Q|_F.\end{aligned}$$

Hence, plugging this into (9), we obtain

$$\mathrm{I} \le \left(|a|\,C_1 + \frac{4(d+1)}{d}|b|\,C_2 + 13c\,C_3\right)|\delta Q|_F.$$

On the other hand, if $|\delta Q|_F \le G$, then we can bound $\frac{|\delta Q|_F^2}{r(Q)}$, $\frac{|\delta Q|_F^3}{r(Q)}$, and $\frac{|Q|_F|\delta Q|_F^2}{r(Q)}$ by

$$\frac{|\delta Q|_F^2}{r(Q)} \le GC_1|\delta Q|_F, \quad \frac{|\delta Q|_F^3}{r(Q)} \le G^2C_1|\delta Q|_F, \quad \frac{|Q|_F|\delta Q|_F^2}{r(Q)} \le G\,C_2|\delta Q|_F.$$

Plugging these into (9), we arrive at

$$\mathrm{I} \le \left(|a|\,C_1 + \frac{2(d+1)}{d}|b|\,C_2 + \frac{d+1}{d}|b|\,G\,C_1 + 3c\,C_3 + c\,G^2\,C_1 + 3c\,G\,C_2\right)|\delta Q|_F.$$

Therefore, $\mathrm{I} \le Z_1|\delta Q|_F$ for some constant Z_1 depending on C_i, $i = 1, 2, 3$, G, a, b, and c. To bound term II, note that

$$\begin{aligned}\mathrm{II} \le& \left|\frac{S(Q+\delta Q)}{r(Q+\delta Q)^{\frac{3}{2}}}\right|_F \frac{\sqrt{r(Q+\delta Q)}|r(Q+\delta Q) - r(Q)|}{r(Q)} \\ \overset{\text{Lem 4.10}}{\le}& K_3\,\frac{\sqrt{r(Q+\delta Q)}\,|r(Q+\delta Q) - r(Q)|}{r(Q)}\end{aligned}$$

$$= K_3 \frac{\sqrt{r(Q+\delta Q)}\ |r(Q+\delta Q)^2 - r(Q)^2|}{r(Q)\ [r(Q+\delta Q) + r(Q)]}$$

$$\leq K_3 \frac{|r(Q+\delta Q)^2 - r(Q)^2|}{r(Q)^{\frac{3}{2}}}, \tag{4.13}$$

where we have used the fact

$$r(Q+\delta Q) + r(Q) \geq 2\sqrt{r(Q+\delta Q)\, r(Q)} \geq \sqrt{r(Q+\delta Q)\, r(Q)}.$$

Expanding $r(Q+\delta Q)$ by plugging $(Q+\delta Q)$ into (1.2), we have

$$\begin{aligned} r(Q+\delta Q)^2 =& r(Q)^2 + 2a\,\mathrm{tr}(Q\,\delta Q) + a\,\mathrm{tr}((\delta Q)^2) - \frac{2b}{3}\,\mathrm{tr}((\delta Q)^3) - 2b\,\mathrm{tr}(Q^2\,\delta Q) \\ & - 2b\,\mathrm{tr}(Q\,(\delta Q)^2) + \frac{c}{2}\mathrm{tr}^2((\delta Q)^2) + 2c\,\mathrm{tr}^2(Q\,\delta Q) \\ & + 2c\,\mathrm{tr}(Q^2)\,\mathrm{tr}(Q\,\delta Q) + c\,\mathrm{tr}(Q^2)\,\mathrm{tr}((\delta Q)^2) + 2c\,\mathrm{tr}((\delta Q)^2)\,\mathrm{tr}(Q\,\delta Q). \end{aligned}$$

We plug this into (4.13),

$$\begin{aligned} \frac{1}{K_3}\,\mathrm{II} \leq & \frac{2|a|\,|Q|_F|\delta Q|_F + |a|\,|\delta Q|_F^2 + \frac{2|b|}{3}|\delta Q|_F^3 + 2|b||Q|_F^2|\delta Q|_F + 2|b||Q|_F|\delta Q|_F^2}{r(Q)^{\frac{3}{2}}} \\ & + \frac{\frac{c}{2}|\delta Q|_F^4 + 3c|Q|_F^2|\delta Q|_F^2 + 2c\,|Q|_F^3|\delta Q|_F + 2c|\delta Q|_F^3|Q|_F}{r(Q)^{\frac{3}{2}}}. \end{aligned}$$

In a similar way as for term I, we can find constant Z_2 such that $\mathrm{II} \leq Z_2|\delta Q|_F$. To sum up, if we choose $L = \max\{\frac{3K_3\sqrt{c}}{2} + 1 + \frac{K_3 c^{\frac{1}{4}}}{2}, Z_1, Z_2\}$, then

$$|P(Q+\delta Q) - P(Q)|_F \leq L|\delta Q|_F$$

for any Q and δQ. ∎

Using the Lipschitz continuity of $P(Q)$, it is now easy to prove the following bound on $\{D_t^+ r_{ijk}^n\}$.

LEMMA 4.12. *We have*

$$\Delta t \sum_{n=0}^{m} \left(h^3 \sum_{i,j,k=1}^{N} |D_t^+ r_{ijk}^n| \right)^2 \leq C < \infty$$

for $0 \leq m \leq N_T$, *where* N_T *is such that* $N_T \Delta t = T$ *and* $C > 0$ *is a constant independent of* h *and* Δt.

Proof. We take absolute values of the scheme for r_{ijk}^n, the second equation in (3.4) divided by Δt,

$$\left| D_t^+ r_{ijk}^n \right| = \left| \overline{P}_{ijk}^{n+\frac{1}{2}} : D_t^+ Q_{ijk}^n \right|,$$

and sum over $i, j, k = 1, \ldots N$, then multiply by h^3, square and sum over $n = 0, \ldots, m$, and use Hölder's inequality:

$$\sum_{n=0}^{m}\left(h^3\sum_{i,j,k=1}^{N}\left|D_t^+ r_{ijk}^n\right|\right)^2 = \sum_{n=0}^{m}\left(h^3\sum_{i,j,k=1}^{N}\left|\overline{P}_{ijk}^{n+\frac{1}{2}} : D_t^+ Q_{ijk}^n\right|\right)^2$$
$$\leq \sum_{n=0}^{m}\left(h^3\sum_{i,j,k=1}^{N}\left|\overline{P}_{ijk}^{n+\frac{1}{2}}\right|^2\right)\left(h^3\sum_{i,j,k=1}^{N}\left|D_t^+ Q_{ijk}^n\right|^2\right).$$

Next, we use the Lipschitz continuity of $P(Q)$, and then Lemma 4.8,

$$\sum_{n=0}^{m}\left(h^3\sum_{i,j,k=1}^{N}\left|D_t^+ r_{ijk}^n\right|\right)^2 \leq Ch^6\sum_{n=0}^{m}\sum_{i,j,k=1}^{N}\left(\left|Q_{ijk}^n\right|^2+\left|Q_{ijk}^{n-1}\right|^2\right)\sum_{i,j,k=1}^{N}\left|D_t^+ Q_{ijk}^n\right|^2$$
$$\leq C\sum_{n=0}^{m}\left\|D_t^+ Q^n\right\|_h^2 \max_{0\leq\ell\leq m}\left\|Q^\ell\right\|_h^2$$
$$\overset{\text{Lem. 4.8}}{\leq} C\sum_{n=0}^{m}\left\|D_t^+ Q^n\right\|_h^2.$$

Multiplying by Δt and using Corollary 4.7, we obtain the result. ∎

5. Convergence of the scheme. Using the estimates established in the previous section, we proceed to proving convergence of the scheme (3.4) to a weak solution of (1.5). To do so, we define piecewise constant interpolations of the grid functions $\{Q_{ijk}^n\}$, $\{r_{ijk}^n\}$, and $\{\overline{P}_{ijk}^{n+\frac{1}{2}}\}$,

(5.1)
$$Q_{h,\Delta t}^n(x)=\sum_{i,j,k=0}^{N+1} Q_{ijk}^n\,\chi_{C_{ijk}},\quad r_{h,\Delta t}^n(x)=\sum_{i,j,k=0}^{N+1} r_{ijk}^n\,\chi_{C_{ijk}},\quad P_{h,\Delta t}^n(x)=\sum_{i,j,k=0}^{N+1}\overline{P}_{ijk}^{n+\frac{1}{2}}\,\chi_{C_{ijk}}.$$

where $C_{ijk}=[(i-1/2)h,(i+1/2)h]\times[(j-1/2)h,(j+1/2)h]\times[(k-1/2)h,(k+1/2)h]$ and χ_A is the characteristic function of the set A. Then, we define piecewise constant interpolations in time,

$$Q_{h,\Delta t}(t,x)=\sum_{n=0}^{N_T-1} Q_{h,\Delta t}^n(x)\chi_{S_n}(t), \tag{5.2}$$

$$r_{h,\Delta t}(t,x)=\sum_{n=0}^{N_T-1} r_{h,\Delta t}^n(x)\chi_{S_n}(t), \tag{5.3}$$

$$P_{h,\Delta t}(t,x)=\sum_{n=0}^{N_T-1} P_{h,\Delta t}^n(x)\chi_{S_n}(t), \tag{5.4}$$

where $T=N_T\Delta t$ and $S_n=[n\Delta t,(n+1)\Delta t)$. We will show that a subsequence of these converges to a weak solution of (1.5).

THEOREM 5.1. *The piecewise constant interpolations* (5.2)–(5.4) *computed using scheme* (3.4) *converge up to a subsequence to a weak solution of* (1.5) *(as in Definition* 2.3*) as* $h,\Delta t\to 0$.

Proof. **Step 1: Compactness.** We apply the first order finite difference operator D_t^+ on $Q_{h,\Delta t}$ and $r_{h,\Delta t}$,

$$
\begin{aligned}
D_t^+ Q_{h,\Delta t}(t,x) &= \sum_{n=0}^{N_T-1} \frac{Q_{h,\Delta t}^{n+1}(x) - Q_{h,\Delta t}^{n}(x)}{\Delta t} \chi_{S_n}, \quad \text{and} \\
D_t^+ r_{h,\Delta t}(t,x) &= \sum_{n=0}^{N_T-1} \frac{r_{h,\Delta t}^{n+1}(x) - r_{h,\Delta t}^{n}(x)}{\Delta t} \chi_{S_n}.
\end{aligned} \tag{5.5}
$$

From the energy stability of the scheme, Theorem 4.6, it follows that $\{\nabla_h Q_{h,\Delta t}\} \subset L^\infty(0,T;L^2(\Omega))$ and $\{r_{h,\Delta t}\} \subset L^\infty(0,T;L^2(\Omega))$ uniformly in $\Delta t, h > 0$. Corollary 4.7 yields $\{D_t^+ Q_{h,\Delta t}\} \subset L^2([0,T]\times\Omega)$ uniformly in $h, \Delta t > 0$. Moreover, from Lemma 4.8, we get

$$
\|Q_{h,\Delta t}(t)\|_{L^2(\Omega)} \le T^{\frac{1}{2}} \|D_t^+ Q_{h,\Delta t}\|_{L^2([0,T]\times\Omega)} + \|Q_{h,\Delta t}(0)\|_{L^2(\Omega)} < \infty,
$$

and hence $\{Q_{h,\Delta t}\} \subset L^\infty([0,T];L^2(\Omega))$. Therefore, we can apply a discretized version of the Aubin–Lions lemma [14, Lemma A.1] to conclude that there exists $Q \in L^2([0,T],H^1(\Omega))$ and a subsequence $\{Q_{h_m,\Delta t_m}\}_m$ such that $Q_{h_m,\Delta t_m} \to Q$ in $L^2([0,T]\times\Omega)$ as $m \to \infty$. Due to the uniform bounds, we also obtain $\nabla_{h_m} Q_{h_m,\Delta t_m} \rightharpoonup \nabla Q$ in $L^2([0,T]\times\Omega)$ and we can extract a weakly convergent subsequence of $\{D_t^+ Q_{h_m}, \Delta t_m\}_m$ and $\{r_{h_m,\Delta t_m}\}_m$, for simplicity still indexed by m. In summary, we have the following:

$$
\begin{aligned}
&Q_{h_m,\Delta t_m} \to Q \qquad \text{in } L^2([0,T]\times\Omega), \qquad D_t^+ Q_{h_m,\Delta t_m} \rightharpoonup Q_t \qquad \text{in } L^2([0,T]\times\Omega), \\
&\nabla_h Q_{h_m,\Delta t_m} \rightharpoonup \nabla Q \qquad \text{in } L^2([0,T]\times\Omega),
\end{aligned} \tag{5.6}
$$

and

$$
r_{h_m,\Delta t_m} \overset{*}{\rightharpoonup} g \quad \text{in } L^\infty([0,T];L^2(\Omega)). \tag{5.7}
$$

Since we have shown that $P(Q)$ is Lipschitz continuous with respect to Q in Theorem 4.11, we obtain from the strong convergence of $\{Q_{h_m,\Delta t_m}\}_m$ that

$$
P(\,Q_{h_m,\Delta t_m}\,) \to P(Q) \qquad \text{in } L^2([0,T]\times\Omega). \tag{5.8}
$$

Step 2: Passing to the limit $m \to \infty$**.** Next, we show that the sequences $\{Q_{h_m,\Delta t_m}\}_m$, $\{r_{h_m,\Delta t_m}\}_m$ converge to a weak solution of (1.5), that is, that the limit (Q,r) is a weak solution in the sense of Definition 2.3. We start with the equation for the variable r. From the numerical scheme (3.4), it follows that

$$
D_t^+ r_{h,\Delta t} = P_{h,\Delta t}(t,x) : D_t^+ Q_{h,\Delta t}. \tag{5.9}
$$

For any smooth test function ϕ with compact support in $[0,T]\times\Omega$, we have $P(Q)\,\phi \in L^2([0,T]\times\Omega)$. Therefore, using the weak convergence $D_t^+ Q_{h_m,\Delta t_m} \rightharpoonup Q_t$ in $L^2([0,T]\times\Omega)$ we find that

$$
\int_0^T \int_\Omega P(Q) : D_t^+ Q_{h_m,\Delta t_m} \phi \, dx dt \overset{m\to\infty}{\longrightarrow} \int_0^T \int_\Omega P(Q) : Q_t \phi \, dx dt. \tag{5.10}
$$

Moreover, by the strong convergence of $P_{h_m,\Delta t_m}$ in $L^2([0,T]\times\Omega)$, we have

$$\left|\int_0^T \int_\Omega (\, P_{h_m,\Delta t_m} - P(Q)\,) : D_t^+ Q_{h_m,\Delta t_m} \phi \, dxdt\right|$$
$$\leq \|\phi\|_{L^\infty(\Omega\times[0,T])} \|P_{h_m,\Delta t_m} - P(Q)\|_{L^2([0,T]\times\Omega)} \|D_t^+ Q_{h_m,\Delta t_m}\|_{L^2([0,T]\times\Omega)} \overset{m\to\infty}{\longrightarrow} 0.$$

Therefore, we can multiply ϕ on both sides of (5.9), integrate over both space Ω and time interval $[0,T]$, and apply (5.10) to obtain

$$\begin{aligned}\int_0^T \int_\Omega D_t^+ r_{h_m,\Delta t_m} \phi \, dxdt &= \int_0^T \int_\Omega P_{h_m,\Delta t_m} : D_t^+ Q_{h_m,\Delta t_m} \phi \, dxdt \\ &= \int_0^T \int_\Omega (\, P_{h_m,\Delta t_m} - P(Q)\,) : D_t^+ Q_{h_m,\Delta t_m} \phi \, dxdt \\ &\quad + \int_0^T \int_\Omega P(Q) : D_t^+ Q_{h_m,\Delta t_m} \phi \, dxdt \\ &\overset{m\to\infty}{\longrightarrow} \int_0^T \int_\Omega P(Q) : Q_t \, \phi \, dxdt.\end{aligned}$$

For the LHS of (5.9), we combine the definition of the piecewise constant functions, (5.1) and (5.5), and rename the integration variables so that the difference operator acts on the smooth test function:

(5.11)
$$\begin{aligned}\text{LHS} &= \sum_{n=0}^{N_T-1} \int_{S_n} \int_\Omega \frac{r^{n+1}_{h_m,\Delta t_m}(x) - r^n_{h_m,\Delta t_m}(x)}{\Delta t_m} \phi(t,x)\, dxdt \\ &= \sum_{n=1}^{N_T-1} \int_{S_n} \int_\Omega r^n_{h_m,\Delta t_m} \frac{\phi(x,t-\Delta t_m) - \phi(t,x)}{\Delta t_m}\, dxdt \\ &\quad + \frac{1}{\Delta t_m} \int_{S_{N_T-1}} \int_\Omega r^{N_T}_{h_m,\Delta t_m} \phi(t,x)\, dxdt - \frac{1}{\Delta t_m} \int_{S_0} \int_\Omega r^0_{h_m,\Delta t_m} \phi(t,x)\, dxdt \\ &= \int_0^T \int_\Omega r_{h_m,\Delta t_m} \frac{\phi(x,t-\Delta t_m) - \phi(t,x)}{\Delta t_m}\, dxdt \\ &\quad + \frac{1}{\Delta t_m} \int_{S_{N_T-1}} \int_\Omega r^{N_T}_{h_m,\Delta t_m} \phi(t,x)\, dxdt - \frac{1}{\Delta t_m} \int_{S_0} \int_\Omega r^0_{h_m,\Delta t_m} \phi(t,x)\, dxdt.\end{aligned}$$

When ϕ has compact support in $[0,T)\times\Omega$, the second term on the RHS vanishes, and we can use the weak* convergence of $\{r_{h,\Delta t}\}$, (5.7), to pass the limit $h, \Delta t \to 0$,

$$\text{LHS} \overset{m\to\infty}{\longrightarrow} -\int_0^T \int_\Omega g\phi_t dxdt - \int_\Omega r^0(x)\phi(0,x)dx.$$

Using [12, Lemma 1.1, p. 250], this implies that r is weakly continuous in time on $L^1(\Omega)$, since $P(Q) : Q_t \in L^2([0,T];L^1(\Omega))$ by the Lipschitz continuity of P. Lemma A.2 then implies that also $\int r_{h,\Delta t}(t,x)\phi(t,x)dx \to \int g(t,x)\phi(t,x)dx$ for every $t \in [0,T]$ up to a subsequence as $h, \Delta t \to 0$, and hence we can pass to the limit in the LHS (5.11) when ϕ is compactly supported in $[0,T]\times\Omega$. Thus the limit g satisfies (2.4).

Next we show that the limit Q satisfies (2.3). We take the inner product of the first equation in (3.4) with a smooth matrix-valued function $\varphi = (\varphi_{\alpha\beta})_{\alpha,\beta=1}^d$:

$[0,T]\times\Omega\to\mathbb{R}^{d\times d}$ integrated over $S_n\times C_{ijk}$, i.e., $\iint_{S_n\times C_{ijk}}\varphi\,dxdt$, and then sum over n and i,j,k. We obtain

$$\sum_{n=0}^{N_T-1}\sum_{i,j,k=1}^{N}\int_{S_n}\int_{C_{ijk}}\frac{Q_{ijk}^{n+1}-Q_{ijk}^{n}}{\Delta t}:\varphi\,dxdt$$
$$=\sum_{n=0}^{N_T-1}\sum_{i,j,k=1}^{N}\int_{S_n}\int_{C_{ijk}}M\left(L_1\Delta_h Q_{ijk}^{n+\frac12}-r_{ijk}^{n+\frac12}\overline{P}_{ijk}^{n+\frac12}+\frac{L_2+L_3}{2}\alpha_{ijk}^{n+\frac12}\right):\varphi\,dxdt.$$

We rewrite this in terms of the piecewise constant functions (5.1):

(5.12)
$$\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}D_t^+Q_{h_m,\Delta t_m}^{n}:\varphi\,dxdt=\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}M\left(L_1\Delta_h Q_{h_m,\Delta t_m}^{n+\frac12}-r_{h_m,\Delta t_m}^{n+\frac12}\overline{P}_{h_m,\Delta t_m}^{n+\frac12}\right.$$
$$\left.+\frac{L_2+L_3}{2}\alpha_h(Q_{h_m,\Delta t_m})^{n+\frac12}\right):\varphi\,dxdt.$$

(Here $\alpha_h(Q_{h,\Delta t})^{n+\frac12}=\frac12(\alpha_h(Q_{h,\Delta t}^n)+\alpha_h(Q_{h,\Delta t}^{n+1}))$.) Since $\{D_t^+Q_{h,\Delta t}\}$ is weakly convergent in L^2 (c.f. (5.6)), we can pass to the limit $m\to\infty$ in the LHS and obtain

$$\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}D_t^+Q_{h_m,\Delta t_m}^{n}:\varphi\,dxdt\longrightarrow\int_0^T\int_{\Omega}Q_t:\varphi\,dxdt.$$

Integrating by parts, we obtain the LHS of (2.3). To deal with the RHS of (5.12), we introduce the discrete forward and difference operators D_k^+ and D_k^- for matrix functions $\varphi=(\varphi_{\alpha\beta})_{\alpha\beta}$, $1\le\alpha,\beta\le d$. Similar to (3.1), D_k^+ denotes the forward difference in the coordinate direction k. For example, for $x=(x_1,x_2,x_3)$ and $k=1$, we define

$$\left(D_1^{\pm}\varphi(x)\right)_{\alpha\beta}=\pm\frac{\varphi_{\alpha\beta}(t,x_1\pm h,x_2,x_3)-\varphi_{\alpha\beta}(t,x_1,x_2,x_3)}{h}.$$

In addition, we introduce the discrete gradient and divergence operators for smooth φ:

$$(\nabla_h^{\pm}\varphi)_{\alpha\beta}=\left((D_1^{\pm}\varphi)_{\alpha\beta},(D_2^{\pm}\varphi)_{\alpha\beta},(D_3^{\pm}\varphi)_{\alpha\beta}\right)^{\top},\qquad(\mathrm{div}_h\,\varphi)_{\beta}=\sum_{\alpha=1}^{d}(D_i^c\varphi)_{\alpha\beta},$$

where $\varphi_{\alpha\beta}$ is the (α,β)-entry of the matrix φ. Renaming the integration variables such that the difference operators act on the test functions in the RHS of (5.12) and then using (5.6) and (5.7), the RHS of (5.12) satisfies

$$\mathrm{RHS}=-ML_1\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}\nabla_{h_m}^{-}Q_{h_m,\Delta t_m}^{n+\frac12}\cdot\nabla_{h_m}^{-}\varphi$$
$$-M\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}r_{h_m,\Delta t_m}^{n+\frac12}\overline{P}_{h_m,\Delta t_m}^{n+\frac12}:\varphi\,dxdt$$
$$-M\frac{(L_2+L_3)}{2}\sum_{n=0}^{N_T-1}\int_{S_n}\int_{\Omega}\sum_{\alpha,\beta,\gamma=1}^{d}\left(D_\gamma^cQ_{h_m,\Delta t_m}^{n}\right)_{\beta\gamma}(D_\alpha^c\varphi)_{\alpha\beta}$$

$$
\begin{aligned}
&- M\frac{(L_2+L_3)}{2}\sum_{n=0}^{N_T-1}\int_{S_n}\int_\Omega \sum_{\alpha,\beta,\gamma=1}^{d}\left(D_\gamma^c Q^n_{h_m,\Delta t_m}\right)_{\alpha\gamma}(D_\beta^c\varphi)_{\alpha\beta}\\
&+\frac{M(L_2+L_3)}{d}\sum_{n=0}^{N_T-1}\int_{S_n}\int_\Omega \sum_{\alpha\beta,\gamma=1}^{d}\left(D_\alpha^c Q^n_{h_m,\Delta t_m}\right)_{\gamma\alpha}(D_\gamma^c\varphi)_{\beta\beta}\\
&\overset{m\to\infty}{\longrightarrow} -ML_1\int_0^T\int_\Omega\sum_{\alpha,\beta=1}^{d}\nabla Q_{\alpha\beta}\cdot\nabla\varphi_{\alpha\beta}\,dx\,dt\\
&- M\lim_{m\to\infty}\int_0^T\int_\Omega r_{h_m,\Delta t_m}P_{h_m,\Delta t_m}:\varphi\,dx\,dt\\
&- M\,\frac{L_2+L_3}{2}\int_0^T\int_\Omega\sum_{\alpha,\beta,\gamma=1}^{d}\Big(\partial_\gamma Q_{\beta\gamma}\partial_\alpha\varphi_{\alpha\beta}+\partial_\gamma Q_{\alpha\gamma}\partial_\beta\varphi_{\alpha\beta}\\
&-\frac{2}{d}\partial_\alpha Q_{\gamma\alpha}\partial_\gamma\varphi_{\beta\beta}\Big)\,dxdt,
\end{aligned}
$$

where $\nabla_h^- Q^{n+\frac12}_{h,\Delta t}\cdot\nabla_h^-\varphi=\sum_{\alpha,\beta=1}^d\nabla_h^-(Q^{n+\frac12}_{h,\Delta t})_{\alpha\beta}\cdot(\nabla_h^-\varphi)_{\alpha\beta}$. It remains to show

$$
\lim_{m\to\infty}\int_0^T\int_\Omega r_{h_m,\Delta,t_m}P_{h_m,\Delta t_m}:\varphi\,dx\,dt=\int_0^T\int_\Omega g\,P(Q):\varphi\,dx\,dt. \tag{5.13}
$$

To prove (5.13), we take the difference of the two terms, that is,

$$
\begin{aligned}
&\left|\int_0^T\int_\Omega r_{h_m,\Delta t_m}P_{h_m,\Delta t_m}:\varphi\,dx\,dt-\int_0^T\int_\Omega g\,P(Q):\varphi\,dx\,dt\right|\\
&=\left|\int_0^T\int_\Omega r_{h_m,\Delta t_m}\left(P_{h_m,\Delta t_m}-P(Q)\right):\varphi\,dx\,dt-\int_0^T\int_\Omega (g-r_{h_m,\Delta t_m})\,P(Q):\varphi\,dx\,dt\right|\\
&\le\underbrace{\left|\int_0^T\int_\Omega r_{h_m,\Delta t_m}\left(P_{h_m,\Delta t_m}-P(Q)\right):\varphi\,dx\,dt\right|}_{\mathrm{I}}+\underbrace{\left|\int_0^T\int_\Omega (g-r_{h_m,\Delta t_m})\,P(Q):\varphi\,dx\,dt\right|}_{\mathrm{II}}.
\end{aligned}
$$

By Cauchy–Schwarz inequality, (5.8), and the energy estimate, Theorem 4.6,

$$
\mathrm{I}\le\|\varphi\|_{L^\infty(\Omega\times[0,T])}\,\|P_{h_m,\Delta t_m}-P(Q)\|_{L^2([0,T]\times\Omega)}\,\|r_{h_m,\Delta t_m}\|_{L^2([0,T]\times\Omega)}\to 0.
$$

Note that $P(Q)\,\varphi\in L^2([0,T]\times\Omega)$ and $r_{h_m,\Delta t_m}\rightharpoonup g$ in L^2, and therefore $\mathrm{II}\to 0$. This proves (5.13). Combining the estimates for the left and the RHS, we see that Q satisfies (2.3). The trace-free condition and the symmetry are linear constraints and therefore conserved under the L^2-convergence of $Q_{h,\Delta t}$. The energy inequality is a direct result by passing the limits in Theorem 4.6 and using Fatou's lemma. Hence the limit (Q,r) is a weak solution in the sense of Definition 2.3. ∎

5.1. Equivalence of weak formulations ($r = r(Q)$). Now that we have established that the scheme converges to a weak solution of (1.5), it remains to show that such a weak solution is in fact a weak solution of (1.1). To do so, we show that the limit g established above in (5.7) satisfies $g = r(Q)$ weakly, where Q is the limit of $Q_{h_m,\Delta t_m}$ and $r(Q)$ is defined in (1.2). Plugging this into the weak formulation (2.1), we see that Q is in fact a weak solution in the sense of Definition 2.2. We thus need to prove the following lemma.

LEMMA 5.2. *Assume that (Q, g) is a weak solution in the sense of Definition* 2.3. *Then for any smooth ψ with compact support in $(0,T)\times\Omega$ (compactly supported in both time and space), we have*

$$\int_0^T \int_\Omega g\,\psi dxdt = \int_0^T \int_\Omega r(Q)\,\psi\,dxdt,$$

where $r(Q)$ is defined in (1.2).

Proof. Since (Q, g) is a weak solution of (1.5), we have that

$$-\int_0^T \int_\Omega g\psi_t\,dxdt = \int_0^T \int_\Omega P(Q) : Q_t\,\psi\,dxdt \tag{5.14}$$

for ψ smooth and compactly supported in $(0,T)\times\Omega$. For the RHS, if Q is a smooth function, we can use chain rule and integration by parts to get

$$\int_0^T \int_\Omega P(Q) : Q_t\,\psi = \int_0^T \int_\Omega r(Q)_t\,\psi = -\int_0^T \int_\Omega r(Q)\psi_t.$$

Since $Q \in L^2([0,T], H^1(\Omega))$ and $Q_t \in L^2([0,T]\times\Omega)$, we can find a sequence of smooth function $\{Q_n\}_n$ with $Q_n \to Q$ in $L^2([0,T], H^1(\Omega))$ and $(Q_n)_t \to Q_t$ in $L^2([0,T]\times\Omega)$. We note that by mean value theorem,

$$r(Q) - r(Q_n) = P(\tilde{Q}) : (Q - Q_n)$$

for some $\tilde{Q} = \lambda_1 Q + \lambda_2 Q_n$ where $\lambda_1, \lambda_2 \in [0,1]$ and $\lambda_1 + \lambda_2 = 1$. Noting that $P(Q)$ is Lipschitz continuous with respect to Q, so $|P(\tilde{Q})|_F \le \tilde{L}|\tilde{Q}|_F$ for some constant $\tilde{L} > 0$. Therefore,

$$|r(Q) - r(Q_n)|_F = \left|P(\tilde{Q}) : (Q - Q_n)\right|_F \le \tilde{L}\,(|Q|_F + |Q_n|_F)\;|Q - Q_n|_F\,.$$

Integrating it over time and space we obtain

$$\begin{aligned}\|r(Q) - r(Q_n)\|_{L^1([0,T]\times\Omega)} \le \tilde{L}\;&\big(\|Q\|_{L^2([0,T]\times\Omega)}\\ &+\|Q_n\|_{L^2([0,T]\times\Omega)}\big)\;||Q - Q_n||_{L^2([0,T]\times\Omega)} \to 0,\end{aligned}$$

since Q and Q_n are both bounded in $L^2([0,T]\times\Omega)$. So if we use smooth functions to approximate Q, we obtain

$$\left|\int_0^T \int_\Omega (\,r(Q) - r(Q_n)\,)\,\psi_t\right| \le \|\psi_t\|_{L^\infty([0,T]\times\Omega)}\;\|r(Q) - r(Q_n)\|_{L^1([0,T]\times\Omega)} \to 0.$$

On the other hand, using the Lipschitz continuity of $P(Q)$, we arrive at

$$\begin{aligned}&\left|\int_0^T \int_\Omega (\,P(Q) : Q_t\,\psi - P(Q_n) : (Q_n)_t\,\psi\,)\right|\\ &\le \left|\int_0^T \int_\Omega (\,P(Q) : Q_t - P(Q) : (Q_n)_t)\psi\right| + \left|\int_0^T \int_\Omega (\,P(Q) : (Q_n)_t - P(Q_n) : (Q_n)_t)\psi\right|\\ &\le \|P(Q)\|_{L^2}\,\|\psi\|_{L^\infty}\,\|Q_t - (Q_n)_t\|_{L^2} + \|P(Q) - P(Q_n)\|_{L^2}\,\|\psi\|_{L^\infty}\,\|(Q_n)_t\|_{L^2} \overset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

Therefore, we have for any $Q \in L^2([0,T], H^1(\Omega))$ with $Q_t \in L^2([0,T]\times\Omega)$

$$\int_0^T \int_\Omega P(Q) : Q_t \, \psi = \lim_{n\to\infty} \int_0^T \int_\Omega P(Q_n) : (Q_n)_t \, \psi$$
$$= -\lim_{n\to\infty} \int_0^T \int_\Omega r(Q_n)\psi_t = -\int_0^T \int_\Omega r(Q)\psi_t.$$

We use this in equation (5.14) to obtain

$$\int_0^T \int_\Omega g\psi_t = -\int_0^T \int_\Omega P(Q) : Q_t \, \psi = \int_0^T \int_\Omega r(Q)\psi_t.$$

From [12, Lemma 1.1, p. 250], we obtain that g as well as $r(Q)$ are absolutely continuous and satisfy for every test function $\psi \in L^\infty(\Omega)$ and almost every $t \in [0,T]$

$$\int_\Omega g(t,x)\psi(x)\,dx = \int_\Omega r(Q(t,x))\psi(x)\,dx + \int_\Omega f(x)\psi(x)\,dx$$

for some $f \in L^2(\Omega)$. However, since g satisfies (2.4), by letting T be 0 in (2.4), we find

$$\int_\Omega g(0,x)\psi(x)\,dx = \int_\Omega r^0(x)\psi(x)\,dx,$$

and so $f = 0$ in $L^2(\Omega)$. This proves the lemma. ∎

This lemma shows that

$$\int_0^T \int_\Omega g\, P(Q) : \varphi \, dx\, dt = \int_0^T \int_\Omega r(Q)\, P(Q) : \varphi \, dx\, dt$$

for any smooth and compactly supported $\varphi : [0,T]\times\Omega \to \mathbb{R}^{d\times d}$. Plugging this into (2.3), we see that the identity becomes (2.1) and hence any weak solution in the sense of Definition 2.3 is in fact a weak solution in the sense of Definition 2.2. Hence we have shown the following.

THEOREM 5.3. *Approximations computed by the numerical scheme* (3.4) *converge as* $\Delta t, h \to 0$*, up to a subsequence, to weak solutions of* (1.1) *as in Definition* 2.2.

6. Numerical results in 2D. We shall now present some numerical experiments in 2D. In this case, the term $\alpha(Q)$ in (1.5a) simplifies to $\alpha(Q) = \Delta Q$. We therefore denote $L := L_1 + \frac{1}{2}(L_2 + L_3)$. We will use the parameters

$$a = -0.3, \quad b = -4, \quad c = 4, \quad A_0 = 500, \quad M = 1, \tag{6.1}$$

unless specified otherwise. The scheme has been implemented in MATLAB and the code used to run the following numerical examples can be found at github.com/VarunMG/Liquid-Crystal-Energy-Stable.

6.1. Numerical example 1: Convergence test. First we check whether the formal second order of accuracy of the scheme manifests in practice when simulating a numerical example with smooth solution. We consider the domain $\Omega = [0,2]^2$, $L = 0.001$ and the initial condition

$$Q_0 = \mathbf{n}_0\mathbf{n}_0^\top - \frac{|\mathbf{n}_0|^2}{2} I_2, \tag{6.2}$$

TABLE 1
Errors and rates for spatial refinement in example (6.2), (6.3).

h	Error for Q_{11}	Order for Q_{11}	Error for Q_{12}	Order for Q_{12}	Error for r	Order for r
0.2	1.3509×10^{-2}	NaN	2.3646×10^{-2}	NaN	6.7561×10^{-3}	NaN
0.1	3.7509×10^{-3}	1.8486	6.4006×10^{-3}	1.8854	1.2878×10^{-3}	2.3912
0.05	9.9049×10^{-4}	1.9210	1.6690×10^{-3}	1.9392	3.8885×10^{-4}	1.7277
0.025	2.6162×10^{-4}	1.9206	4.4341×10^{-4}	1.9123	1.5189×10^{-4}	1.3562

where

$$\mathbf{n_0}(x,y) = \begin{pmatrix} x(2-x)y(2-y) \\ \sin(\pi x)\sin(0.5\pi y) \end{pmatrix}. \tag{6.3}$$

6.1.1. Refinement in space. We compute up to time $T = 0.4$ using 400 time steps and we will use a reference solution $(Q^{\text{ref}}, r^{\text{ref}})$ to show the spatial accuracy of our scheme. The reference solution is computed with 400 grid points in each spatial direction and 4000 time steps. All the errors are measured in L^2-norm

$$\mathcal{E}^Q_{\alpha\beta} = \left\| Q^{\text{ref}}_{\alpha\beta}(T,\cdot) - (Q_h)_{\alpha\beta}(T,\cdot) \right\|_{L^2(\Omega)}, \quad \mathcal{E}^r = \left\| r^{\text{ref}}(T,\cdot) - r_h(T,\cdot) \right\|_{L^2(\Omega)},$$

where $\alpha, \beta \in \{1,2\}$. We compute the numerical solutions with $n = 10, 20, 40, 80$ grid points in each spatial direction. The L^2-errors and convergence rates for Q_{11}, Q_{12}, and r are reported in Table 1. (Note that due to the symmetry and the trace-free property, $Q_{11} = -Q_{22}$ and $Q_{12} = Q_{21}$.) We note that for the components of Q the expected second order convergence rate is almost achieved whereas the convergence rate for the variable r is lower. We suspect that more mesh refinement may be needed to see the optimal order for the variable r.

Figure 1 shows the decay of the discrete energy for $n = 80$ and $N_T = 400$ up to time $T = 0.4$. As predicted by the theory, the energy decays monotonically.

6.1.2. Refinement in time. We use the same setting (initial value and parameters) as for the spatial accuracy test and compute up to time $T = 0.4$ with 100 grid

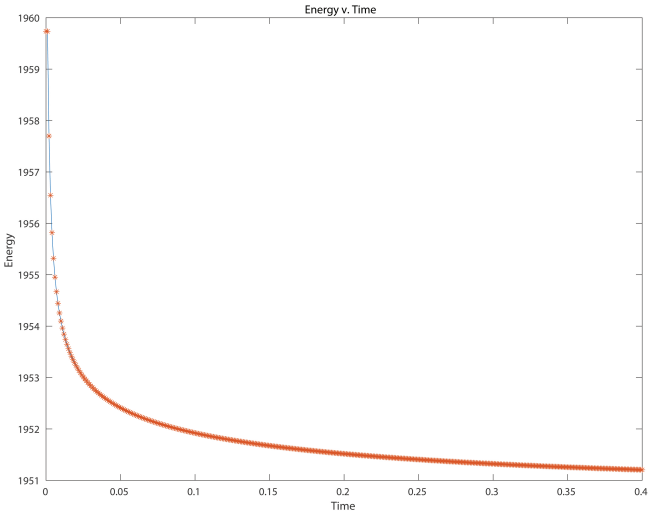


FIG. 1. *Energy decay when* $T = 0.4$ *with* 400 *time steps and* 80 *grid points in each spatial direction.*

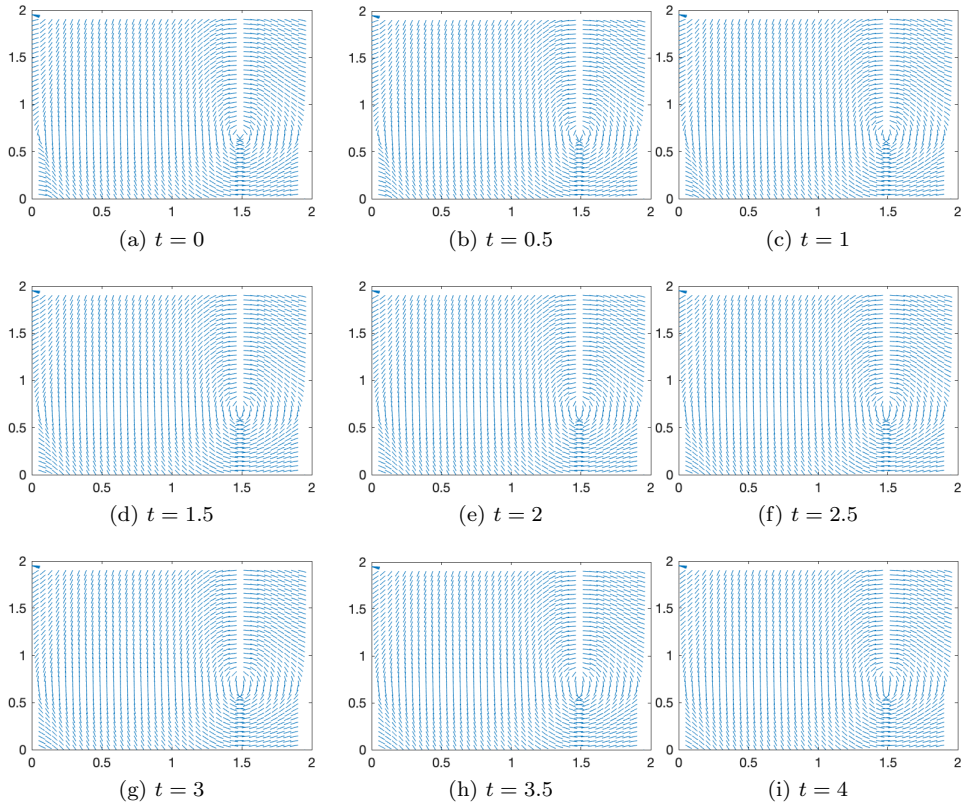
TABLE 2
Errors and rates for time refinement in example (6.2), (6.3).

Δt	Error for Q_{11}	Order for Q_{11}	Error for Q_{12}	Order for Q_{12}	Error for r	Order for r
0.01	7.87395×10^{-4}	NaN	1.49178×10^{-3}	NaN	9.15588×10^{-4}	NaN
5×10^{-3}	1.94110×10^{-4}	2.02022	3.67711×10^{-4}	2.02039	2.19902×10^{-4}	2.05784
2.5×10^{-3}	4.81199×10^{-5}	2.01217	9.11505×10^{-5}	2.01225	5.38631×10^{-5}	2.02949
1.25×10^{-3}	1.19280×10^{-5}	2.01223	2.25937×10^{-5}	2.01233	1.32752×10^{-5}	2.02056
6.25×10^{-4}	2.91895×10^{-6}	2.03083	5.52893×10^{-6}	2.03085	3.23961×10^{-6}	2.03485
3.125×10^{-4}	6.71610×10^{-7}	2.11975	1.27212×10^{-6}	2.11976	7.44387×10^{-7}	2.12170

points in each spatial direction. Similar as above, we will compare the approximations with different time step sizes with a reference solution which is computed with the same number of spatial points and 8000 time steps. The errors and convergence orders for the approximations with 40, 80, 160, 320, 640 time steps are shown in Table 2. We observe second order accuracy as expected.

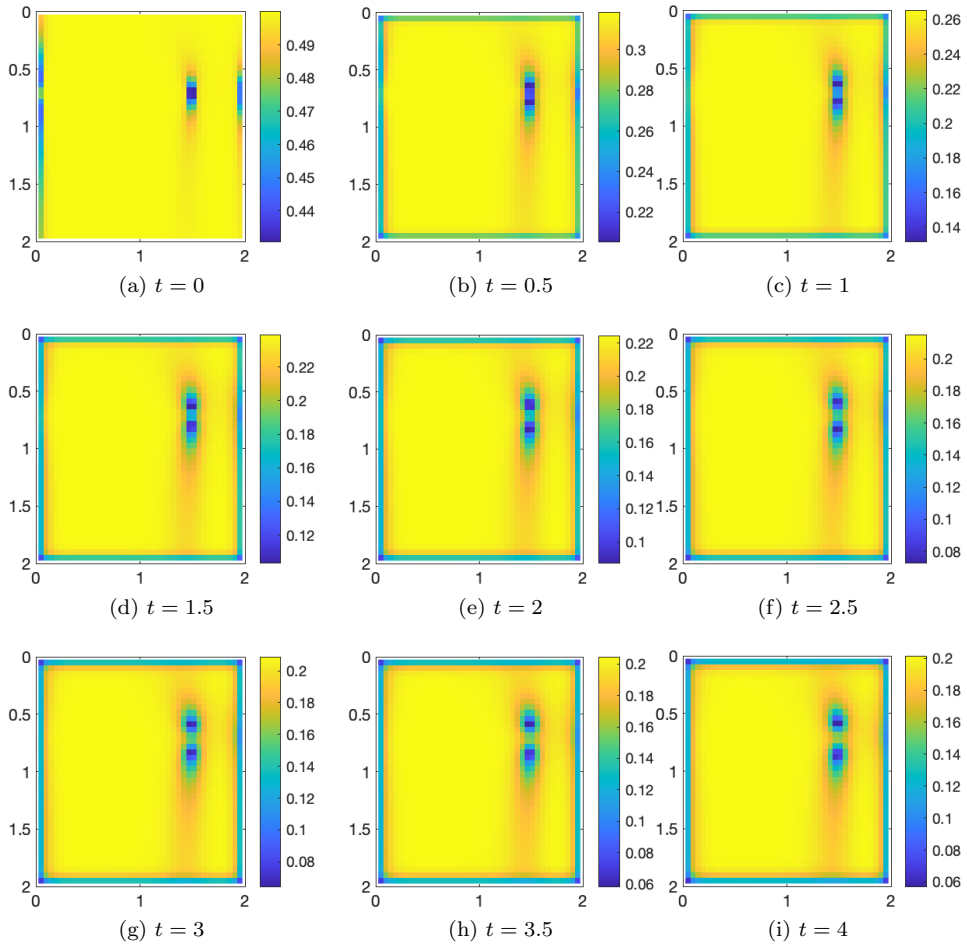
6.2. Numerical example 2: Defects in liquid crystals. We consider the domain $\Omega = [0, 2] \times [0, 2]$ and $L = 0.001$. In this example, we will study the dynamics of defects in liquid crystals. For the initial condition, we take

$$Q_0 = \mathbf{n}_0 \mathbf{n}_0^\top - \frac{|\mathbf{n}_0|^2}{2} I_2, \tag{6.4}$$



(a) $t = 0$ (b) $t = 0.5$ (c) $t = 1$
(d) $t = 1.5$ (e) $t = 2$ (f) $t = 2.5$
(g) $t = 3$ (h) $t = 3.5$ (i) $t = 4$

FIG. 2. *Simulation for initial data* (6.4), (6.5).



(a) $t = 0$ (b) $t = 0.5$ (c) $t = 1$

(d) $t = 1.5$ (e) $t = 2$ (f) $t = 2.5$

(g) $t = 3$ (h) $t = 3.5$ (i) $t = 4$

FIG. 3. *The largest eigenvalue of Q in the simulation for initial data* (6.4), (6.5).

where

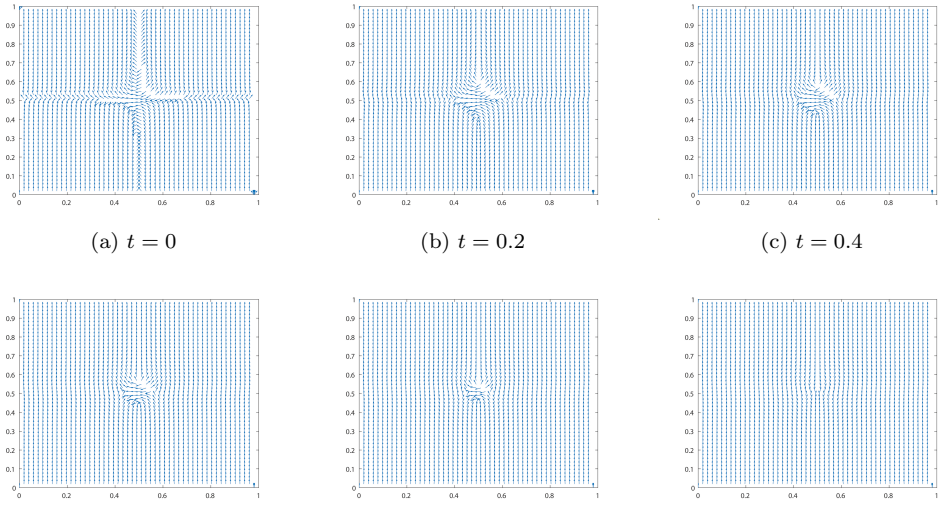
$$\mathbf{n_0}(x,y) = \begin{pmatrix} \log(x^2+1)\,(x-2)^2\,\sin(\frac{\pi y}{2})\,(e^{1.5}-e^x) \\ (y-2)\,(y-3)\,\sin(\frac{\pi y}{10})\,\sin(\frac{\pi x}{2})\,(0.7-y) \end{pmatrix}. \tag{6.5}$$

We use 40 grid points in space in each dimension and 4000 time steps up to $T = 4$. As we can see from Figure 2, initially, there is only one defect, which is located at $(1.5, 0.7)$. This configuration is not stable and generally splits into two different defects. They move away from each other and toward the boundary. Figure 3 depicts the largest eigenvalue of matrix Q at different times. We observe that as the two defects move, the largest eigenvalue decays in a neighborhood of the defects rapidly to 0. The eigenvalue is generally decreasing and tends to 0 everywhere as time evolves. This behavior is a consequence of the boundary condition and the energy dissipation property.

6.3. Numerical example 3: "Disappearing hole". We consider $\Omega = [0,1] \times [0,1]$ and use the parameters $a = -0.2, b = 1, c = 1, L = 0.0025$. As an initial condition, we use (6.4) with

$$\widetilde{\mathbf{n}}_0(x,y) = \begin{pmatrix} x(1-x)y(1-y) \\ \sin(2\pi x)\sin(2\pi y); \end{pmatrix}, \quad \mathbf{n}_0 = \frac{\widetilde{\mathbf{n}}_0}{|\widetilde{\mathbf{n}}_0|}, \tag{6.6}$$

and 50 grid points in space in each dimension and 100 time steps. The simulation is displayed in Figure 4. We observe that the initial misalignment disappears first along the axes and then propagates in a shrinking circle toward the center of the domain and eventually disappears. This behavior was stable with respect to mesh refinement. The discrete energy (4.4) decays at first rapidly and then approaches a constant state corresponding to the alignment of the director field along the y-axis as seen in Figure 5.



(a) $t = 0$ (b) $t = 0.2$ (c) $t = 0.4$

(d) $t = 0.6$ (e) $t = 0.8$ (f) $t = 1$

FIG. 4. *Simulation for initial data* (6.4), (6.6).

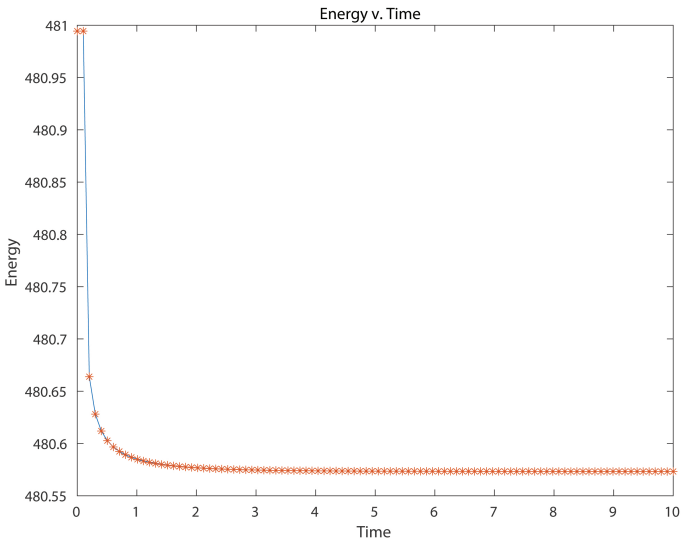


FIG. 5. *Energy when* $T = 10$ *with* 100 *time steps and* 50 *grid points in each spatial direction.*

Appendix A. Some lemmas.

LEMMA A.1. *Let A_{ijk} and B_{ijk} be scalar quantities at grid point (x_i, y_j, z_k) such that $A_{ijk} = 0$ at boundary values, i.e., boundary conditions* (3.2), (3.3). *Then*

$$\sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^+ B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^- A_{ijk}, \quad \sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^- B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^+ A_{ijk},$$

and

$$\sum_{i,j,k=0}^{N+1} A_{ijk} D_\beta^c B_{ijk} = -\sum_{i,j,k=0}^{N+1} B_{ijk} D_\beta^c A_{ijk}$$

for $\beta = 1, 2$ or 3.

Proof. We shall prove this for the case where $\beta = 1$; the other cases follow similarly. Note that

$$\begin{aligned}
\sum_{i,j,k=0}^{N+1} A_{ijk} D_1^+ B_{ijk} &= \frac{1}{h}\left(\sum_{i,j,k=1}^{N} A_{ijk} B_{(i+1)jk} - \sum_{i,j,k=1}^{N} A_{ijk} B_{ijk}\right) \\
&= \frac{1}{h}\left(\sum_{j,k=0}^{N+1}\sum_{i=2}^{N+1} A_{(i-1)jk} B_{ijk} - \sum_{j,k=0}^{N+1}\sum_{i=1}^{N} A_{ijk} B_{ijk}\right) \\
&= \frac{1}{h}\left(\sum_{i,j,k=0}^{N+1} A_{(i-1)jk} B_{ijk} - \sum_{i,j,k=0}^{N+1} A_{ijk} B_{ijk}\right) \\
&= -\sum_{i,j,k=0}^{N+1} B_{ijk} D_1^- A_{ijk},
\end{aligned}$$

where we used the boundary conditions (3.2) and (3.3) for A_{ijk}. For the second identity, using the same trick, we obtain

$$\begin{aligned}
\sum_{i,j,k=0}^{N+1} A_{ijk} D_1^- B_{ijk} &= \frac{1}{h}\left(\sum_{i,j,k=1}^{N} A_{ijk} B_{ijk} - \sum_{i,j,k=1}^{N} A_{ijk} B_{(i-1)jk}\right) \\
&= \frac{1}{h}\left(\sum_{j,k=0}^{N+1}\sum_{i=1}^{N} A_{ijk} B_{ijk} - \sum_{j,k=0}^{N+1}\sum_{i=0}^{N-1} A_{(i+1)jk} B_{ijk}\right) \\
&= \frac{1}{h}\left(\sum_{i,j,k=0}^{N+1} A_{ijk} B_{ijk} - \sum_{i,j,k=0}^{N+1} A_{(i+1)jk} B_{ijk}\right) \\
&= -\sum_{i,j,k=0}^{N+1} B_{ijk} D_1^+ A_{ijk}.
\end{aligned}$$

For the third identity,

$$\sum_{i,j,k=0}^{N+1} A_{ijk} D_1^c B_{ijk} = \frac{1}{2h}\left(\sum_{j,k=0}^{N+1}\sum_{i=1}^{N} A_{ijk} B_{(i+1)jk} - \sum_{j,k=0}^{N+1}\sum_{i=1}^{N} A_{ijk} B_{(i-1)jk}\right)$$

$$= \frac{1}{2h}\left(\sum_{j,k=0}^{N+1}\sum_{i=2}^{N+1} A_{(i-1)jk}B_{ijk} - \sum_{j,k=0}^{N+1}\sum_{i=0}^{N-1} A_{(i+1)jk}B_{ijk}\right)$$
$$= \frac{1}{2h}\left(\sum_{i,j,k=0}^{N+1} A_{(i-1)jk}B_{ijk} - \sum_{i,j,k=0}^{N+1} A_{(i+1)jk}B_{ijk}\right)$$
$$= -\sum_{i,j,k=0}^{N+1} B_{ijk}D_1^c A_{ijk},$$

where we have used boundary values of A_{ijk} and B_{ijk}. ∎

We believe the following lemma is a standard result from real analysis but we did not find a suitable reference to refer to and therefore provide the proof here for completeness.

LEMMA A.2. *Assume that $\{g_{h,\Delta t}\}_{h,\Delta t}$ is a sequence of piecewise constant functions converging weak*, as $h, \Delta t \to 0$, in $L^\infty([0,T];L^2(\Omega))$ to some limit $g \in L^\infty([0,T];L^2(\Omega))$ that is weakly continuous in time in $L^1(\Omega)$, i.e., $\int g(s,x)\phi(x)dx \to \int g(t,x)\phi(x)dx$ when $s \to t$ for $\phi \in L^\infty(\Omega)$. In addition, assume that*

$$\left\|D_t^+ g_{h,\Delta t}\right\|_{L^2([0,T];L^1(\Omega))} \le C,$$

where C is a constant independent of h and Δt. Then, up to a subsequence,

$$\int_\Omega g_{h,\Delta t}\phi(x)dx \overset{h,\Delta t\to 0}{\longrightarrow} \int_\Omega g(t,x)\phi(x)dx$$

for all $t \in [0,T]$ and $\phi \in L^\infty(\Omega)$.

Proof. Let $\phi \in L^\infty(\Omega)$. As $\{g_{h,\Delta t}\}$ is weak* convergent in $L^\infty([0,T];L^2(\Omega))$ we can find a dense set $\mathcal{T} := \{t_i\}_{i=1}^\infty \subset [0,T]$ such that for a (diagonal) subsequence $\{h_m, \Delta t_m\}_{m=1}^\infty$

$$\int_\Omega g_{h_m,\Delta t_m}(t_i,x)\phi(x)dx \overset{m\to\infty}{\longrightarrow} \int_\Omega g(t_i,x)\phi(x)dx \quad \text{for all } t_i \in \mathcal{T}.$$

Fix $\epsilon > 0$ arbitrary and $t \in [0,T]$. Then since g is weakly continuous, we can find an interval $I \subset [0,T]$ such that $t \in I$ and for all $s \in I$,

$$\left|\int_\Omega g(t,x)\phi(x)dx - \int_\Omega g(s,x)\phi(x)dx\right| < \frac{\epsilon}{3}.$$

Next, we pick $M_1 \in \mathbb{N}$ large enough, such that for all $m \ge M_1$ and all $t_j \in \mathcal{T} \cap I$,

$$\left|\int_\Omega g_{h_m,\Delta t_m}(t_j,x)\phi(x)dx - \int_\Omega g(t_j,x)\phi(x)dx\right| < \frac{\epsilon}{3}.$$

We observe that we can write for $s \ge t \in [0,T]$,

$$\int_\Omega \left(g_{h,\Delta t}(s,x) - g_{h,\Delta t}(t,x)\right)\phi(x)dx = \Delta t \int_\Omega \left(\sum_{\ell=\lfloor \frac{t}{\Delta t}\rfloor}^{\lfloor \frac{s}{\Delta t}\rfloor - 1} D_t^+ g_h^\ell(x)\right)\phi(x)dx.$$

Thus,

$$\begin{aligned}
&\left|\int_\Omega \left(g_{h,\Delta t}(s,x) - g_{h,\Delta t}(t,x)\right)\phi(x)dx\right| \\
&\quad\le \Delta t \int_\Omega \left(\sum_{\ell=\lfloor \frac{t}{\Delta t}\rfloor}^{\lfloor \frac{s}{\Delta t}\rfloor -1} \left|D_t^+ g_h^\ell(x)\right|\right)|\phi(x)|dx \\
&\quad\le \Delta t \sum_{\ell=\lfloor \frac{t}{\Delta t}\rfloor}^{\lceil \frac{s}{\Delta t}\rceil -1} \left\|D_t^+ g_h^\ell\right\|_{L^1(\Omega)} \|\phi\|_{L^\infty} \\
&\quad\le \Delta t \left(\sum_{\ell=\lfloor \frac{t}{\Delta t}\rfloor}^{\lceil \frac{s}{\Delta t}\rceil -1} \left\|D_t^+ g_h^\ell\right\|_{L^1(\Omega)}^2\right)^{1/2} \left(\left\lceil \frac{s}{\Delta t}\right\rceil - \left\lfloor \frac{t}{\Delta t}\right\rfloor\right)^{1/2} \|\phi\|_{L^\infty} \\
&\quad\le \left(\Delta t \sum_{\ell=\lfloor \frac{t}{\Delta t}\rfloor}^{\lceil \frac{s}{\Delta t}\rceil -1} \left\|D_t^+ g_h^\ell\right\|_{L^1(\Omega)}^2\right)^{1/2} (s-t+\Delta t)^{1/2} \|\phi\|_{L^\infty} \\
&\quad\le \left\|D_t^+ g_{h,\Delta t}\right\|_{L^2([0,T];L^1(\Omega))} (s-t+\Delta t)^{1/2} \|\phi\|_{L^\infty} \\
&\quad\le C\,(s-t+\Delta t)^{1/2}.
\end{aligned}$$

So we pick $M_2 \ge M_1$ large enough and $J \subset I$ such that for $m \ge M_2$ and $t_j \in J$,

$$\left|\int_\Omega \left(g_{h_m,\Delta t_m}(t_j,x) - g_{h,\Delta t}(t,x)\right)\phi(x)dx\right| \le C\,(t_j - t + 2\Delta t_m)^{1/2} < \frac{\epsilon}{3}.$$

Then we have for $m \ge M_2$ (and $t_j \in J$),

$$\begin{aligned}
\left|\int_\Omega (g(t,x) - g_{h_m,\Delta t_m}(t,x))\phi(x)dx\right| &\le \left|\int_\Omega (g(t,x) - g(t_j,x))\phi(x)dx\right| \\
&\quad + \left|\int_\Omega (g(t_j,x) - g_{h_m,\Delta t_m}(t_j,x))\phi(x)dx\right| \\
&\quad + \left|\int_\Omega (g_{h_m,\Delta t_m}(t_j,x) - g_{h_m,\Delta t_m}(t,x))\phi(x)dx\right| \\
&\le \epsilon,
\end{aligned}$$

which proves the result. ∎

Acknowledgment. We thank Max Hirsch for the careful reading of our manuscript and pointing out several mistakes and typos.

REFERENCES

[1] J. M. Ball, *Mathematics and liquid crystals*, Mol. Cryst. Liquid Cryst., 647 (2017), pp. 1–27, https://doi.org/10.1080/15421406.2017.1289425.

[2] A. Beris, B. Edwards, B. Edwards, and C. Edwards, *Thermodynamics of Flowing Systems: With Internal Microstructure*, Oxford Eng. Sci. Ser., Oxford University Press, Oxford, 1994.

[3] Y. Cai, J. Shen, and X. Xu, *A stable scheme and its convergence analysis for a 2D dynamic Q-tensor model of nematic liquid crystals*, Math. Models Methods Appl. Sci., 27 (2017), pp. 1459–1488, https://doi.org/10.1142/S0218202517500245.

[4] A. Contreras, X. Xu, and W. Zhang, *An elementary proof of eigenvalue preservation for the co-rotational Beris-Edwards system*, J. Nonlinear Sci., 29 (2019), pp. 789–801.

[5] P. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Internat. Ser. Monogr Phys., Clarendon Press, Oxford, 1995.

[6] J. W. Goodby, E. Chin, and J. S. Patel, *Eutectic mixtures of ferroelectric liquid crystals*, J. Phys. Chem., 93 (1989), pp. 8067–8072, https://doi.org/10.1021/j100361a020.

[7] G. Iyer, X. Xu, and A. D. Zarnescu, *Dynamic cubic instability in a 2D Q-tensor model for liquid crystals*, Math. Models Methods Appl. Sci., 25 (2015), pp. 1477–1517.

[8] H. Mori, E. C. Gartland, J. R. Kelly, and P. J. Bos, *Multidimensional director modeling using the Q tensor representation in a liquid crystal cell and its application to the pi-cell with patterned electrodes*, Jpn. J. Appl. Phys., 38 (1999), pp. 135–146, https://doi.org/10.1143/jjap.38.135.

[9] N. J. Mottram and C. J. P. Newton, *Introduction to Q-Tensor Theory*, https://arxiv.org/abs/1409.3542, 2014.

[10] J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev., 61 (2019), pp. 474–506.

[11] A. M. Sonnet and E. Virga, *Dissipative Ordered Fluids, Theories for Liquid Crystals*, Springer, New York, 2012.

[12] R. Temam, *Navier—Stokes Equations: Theory and Numerical Analysis*, AMS Chelsea Publishing, New York, 1985.

[13] O. M. Tovkach, C. Conklin, M. C. Calderer, D. Golovaty, O. D. Lavrentovich, J. Viñals, and N. J. Walkington, *Q-tensor model for electrokinetics in nematic liquid crystals*, Phys. Rev. Fluids, 2 (2017), 053302, https://doi.org/10.1103/PhysRevFluids.2.053302.

[14] K. Trivisa and F. Weber, *A convergent explicit finite difference scheme for a mechanical model for tumor growth*, ESAIM Math. Model. Numer. Anal., 51 (2017), pp. 35–62, https://doi.org/10.1051/m2an/2016014.

[15] M. Wang, W. Wang, and Z. Zhang, *From the Q-tensor flow for the liquid crystal to the harmonic map flow*, Arch. Ration. Mech. Anal., 225 (2017), pp. 663–683, https://doi.org/10.1007/s00205-017-1111-6.

[16] K. F. Wissbrun, *Orientation development in liquid crystal polymers*, in Orienting Polymers, J. L. Ericksen, ed., Springer, New York, 1984, pp. 1–26, https://doi.org/10.1007/BFb0072149.

[17] J. Zhao, X. Yang, Y. Gong, and Q. Wang, *A novel linear second order unconditionally energy stable scheme for a hydrodynamic Q-tensor model of liquid crystals*, Comput. Methods Appl. Mech. Engrg., 318 (2017), pp. 803–825, https://doi.org/10.1016/j.cma.2017.01.031.