Journal of Computational and Applied Mathematics 223 (2009) 938–952

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Split least-squares finite element methods for linear and nonlinear parabolic problems ☆

Hongxing Rui [a,*], Sang Dong Kim [b], Seokchan Kim [c]

[a] School of Mathematics, Shandong University, Jinan, Shandong, 250100, China

[b] Department of Mathematics, Kyungpook National University, Daegu 702-701, South Korea

[c] Department of Applied Mathematics, Changwon National University, Changwon 641-773, South Korea

ARTICLE INFO

Article history:
Received 9 June 2007
Received in revised form 5 March 2008

MSC:
65M12
65M15
65M60

Keywords:
Split
Least squares
Finite element
Error estimates
Parabolic problem

ABSTRACT

In this paper, we propose some least-squares finite element procedures for linear and nonlinear parabolic equations based on first-order systems. By selecting the least-squares functional properly each proposed procedure can be split into two independent symmetric positive definite sub-procedures, one of which is for the primary unknown variable u and the other is for the expanded flux unknown variable $\boldsymbol{\sigma}$. Optimal order error estimates are developed. Finally we give some numerical examples which are in good agreement with the theoretical analysis.

1. Introduction

The purpose of this paper is to consider the least-squares finite element procedures for linear and nonlinear parabolic problems written as first-order systems. It is well known that, compared to mixed element methods, the least-squares finite element method has two typical advantages as follows: it is not subject to the Ladyzhenkaya–Babuska–Brezzi [13,1,4] consistency condition, so the choice of approximation spaces becomes flexible, and it results in a symmetric positive definite system.

Least-squares finite element methods for elliptic problems, based on first-order systems, were introduced by [12] where a least-squares residual minimization is introduced for the mixed system in primary unknown variable u and expanded unknown flux $\boldsymbol{\sigma}$. Then an elegant theory for least-squares finite element approximation for general elliptic boundary value problems was established, see, for example, [12,10,11,16,5,6] and the references therein. Concerning the parabolic problems, [14] and [15] introduced the least-squares finite element procedure with semi-discretization in time and fully discrete scheme. They also established the *a posterior* error estimates and constructed adaptive algorithms.

In this paper we consider the least-squares finite element procedures for linear and nonlinear parabolic problems. Like [14,15] we define the least-squares functionals using weight-factors. By selecting different weight-factors we get different

☆ The work was supported by the Korea Research Foundation under contract number KRF-2005-070-C00017, the National Natural Science Foundation of China Grant No. 10771124 and the Research Fund for Doctoral Program of High Education by the State Education Ministry of China No. 20060422006.

* Corresponding author.
E-mail addresses: hxrui@sdu.edu.cn (H. Rui), skim@knu.ac.kr (S.D. Kim), sckim@changwon.ac.kr (S. Kim).

0377-0427/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2008.03.030

procedures. We show that all the procedures presented in this paper can be divided into two independent sub-procedures, one of which is for the primary unknown variable u and the other is for the expanded flux $\boldsymbol{\sigma}$. The key point used to explain the split of the procedure is Lemma 2.1 which was obtained by integration by parts. Similar results have been found and used by [8] to prove the coercivity of least-squares bilinear formats and by [2,3] to establish connections between least-squares and mixed methods. The last two papers also show that not only is the pressure the same as in the Galerkin method, but also the flux is the same as in the mixed method under some conditions on the finite element spaces.

In this paper three procedures were presented for linear parabolic problems. In the first procedure the sub-procedure for the primary unknown u is the same as the standard Galerkin finite element procedure. In the second procedure one of the sub-procedures is for the expanded flux $\boldsymbol{\sigma}$ only. The third one is a procedure with second-order approximation in time increment. We give one procedure to deal with the nonlinear problem. For these schemes we give the optimal order error estimates. Finally we give some numerical examples.

The remainder of this paper is organized as follows. In Section 2 we introduce the split least-squares schemes for linear problems. In Section 3, we establish the optimal order error estimates. In Section 4, we give a least-squares finite element procedure for nonlinear problems. Finally in Section 5 we give some numerical examples.

Throughout this paper, the notations of standard Sobolev spaces $\mathrm{L}^2(\Omega)$, $\mathrm{H}^k(\Omega)$ and associated norms $\|\cdot\| = \|\cdot\|_{\mathrm{L}^2(\Omega)}$, $\|\cdot\|_k = \|\cdot\|_{\mathrm{H}^k(\Omega)}$ are adopted as those in [7]. For simplicity we use $\|\cdot\|_{L^\infty(H^{m+1})}$ and $\|\cdot\|_{L^2(H^{m+1})}$ to represent $\|\cdot\|_{L^\infty(J;(H^{m+1}(\Omega))^d)}$ and $\|\cdot\|_{L^2(J;(H^{m+1}(\Omega))^d)}$ respectively for $J = (0, T)$ and $d \leq 3$. A constant C (with or without subscript) stands for a generic positive constant independent of the mesh parameter h_u, h_{σ} and Δt, it may be different at different occurrence.

2. Least-squares procedure for linear problems

In this section we present three least-squares finite element procedures for linear problems. For simplicity we just consider the homogeneous boundary condition. The same idea can be used to deal with problems with non-homogeneous boundary condition.

Consider the following parabolic problem on a bounded domain $\Omega \subset \mathcal{R}^d$, $d = 2, 3$:

$$\begin{cases} \phi u_t - \operatorname{div}(\mathcal{A}\nabla u) = f, & \text{in } \Omega \times J, \\ u = 0, & \text{on } \Gamma_D \times J, \\ \mathcal{A}\nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \times J, \end{cases} \tag{2.1}$$

subject to the initial condition

$$u(x, 0) = u_0(x) \quad \text{on } \Omega \times J, \tag{2.2}$$

where $\partial\Omega = \Gamma_D \cap \Gamma_N$, $\mathbf{n}$ is the outward unit normal vector, $J = (0, T]$ is the time interval and ϕ is a continuous function satisfying $\phi_1 \leq \phi \leq \phi_2$ with two positive constants ϕ_1 and ϕ_2. We further assume that $\mathcal{A} = (a_{ij}(x))_{i,j=1}^d$ is a bounded, symmetric and positive definite matrix in Ω, i.e., there exist positive constants α and β such that,

$$\alpha\|\boldsymbol{\xi}\|^2 \leq (\mathcal{A}\boldsymbol{\xi}, \boldsymbol{\xi}) \leq \beta\|\boldsymbol{\xi}\|^2, \quad \forall \boldsymbol{\xi} \in \mathcal{R}^d. \tag{2.3}$$

In some applications, the problem (2.1) appears as a first-order system for both u and $\boldsymbol{\sigma} = -\mathcal{A}\nabla u$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_d)$, as follows:

$$\begin{cases} \phi u_t + \operatorname{div}\boldsymbol{\sigma} - f = 0, & \text{in } \Omega \times J, \\ \boldsymbol{\sigma} + \mathcal{A}\nabla u = 0, & \text{in } \Omega \times J, \\ u = 0, & \text{on } \Gamma_D \times J, \\ \boldsymbol{\sigma} \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \times J. \end{cases} \tag{2.4}$$

For example, in the compressible miscible displacement problem [9], u represents the pressure and $\boldsymbol{\sigma}$ represents the Darcy velocity or flux. In this case the approximations to both u and $\boldsymbol{\sigma}$ are necessary. We consider the least-squares mixed element approximations for (2.4).

First we consider the first-order approximation in time increment. Let Δt be a time increment. With $t^n = n\Delta t$, $u^n = u(t^n, \cdot)$, put

$$\delta_t u^n := \frac{u^n - u^{n-1}}{\Delta t}, \qquad \rho_1^n := \phi(\delta_t u^n - u_t^n). \tag{2.5}$$

It is clear that

$$\rho_1^n = O\left(\int_{t^{n-1}}^{t^n} \|u_{tt}\|\mathrm{d}t\right) = O\left(\left(\Delta t \int_{t^{n-1}}^{t^n} \|u_{tt}\|^2\mathrm{d}t\right)^{\frac{1}{2}}\right). \tag{2.6}$$

Define two function spaces

$$\mathrm{V} = \{v \in \mathrm{H}^1(\Omega) : v = 0 \text{ on } \Gamma_D\}, \tag{2.7}$$

$$\mathbf{W} = \{\boldsymbol{\tau} \in (\mathrm{L}^2(\Omega))^d : \operatorname{div}\boldsymbol{\tau} \in \mathrm{L}^2(\Omega), \boldsymbol{\tau} \cdot \mathbf{n} = 0 \text{ on } \Gamma_N\}. \tag{2.8}$$

From (2.4) we know that for $n \ge 1$, $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ satisfy that

$$\begin{cases} \phi^{-\frac{1}{2}}(\phi u^n + \Delta t \,\mathrm{div}\, \boldsymbol{\sigma}^n - F_1^n) = 0, & \text{in } \Omega \times J, \\ \mathcal{A}^{-\frac{1}{2}}(\boldsymbol{\sigma}^n + \mathcal{A}\nabla u^n) = 0, & \text{in } \Omega \times J, \end{cases} \tag{2.9}$$

where $F_1^n = \phi u^{n-1} + \Delta t f^n + \Delta t \rho_1^n$.

For $(v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}$, define the first kind of least-squares functional $J_1^n(v, \boldsymbol{\tau})$ as follows.

$$J_1^n(v, \boldsymbol{\tau}) = \|\phi^{-\frac{1}{2}}(\phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau} - F_1^n)\|^2 + \Delta t \|\mathcal{A}^{-\frac{1}{2}}(\boldsymbol{\tau} + \mathbf{A}\nabla v)\|^2. \tag{2.10}$$

The least-squares minimization problem corresponding to (2.9) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \boldsymbol{\sigma}^n \in \mathbf{W}$ such that

$$J_1^n(u^n, \boldsymbol{\sigma}^n) = \inf_{(v,\boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}} J_1^n(v, \boldsymbol{\tau}). \tag{2.11}$$

Define the bilinear form $a(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau})$ corresponding to the least-squares functional J_1^n as

$$a(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = (\phi^{-1}(\phi u + \Delta t \,\mathrm{div}\, \boldsymbol{\sigma}), \phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau}) + \Delta t (\mathcal{A}^{-1}(\boldsymbol{\sigma} + \mathcal{A}\nabla u), \boldsymbol{\tau} + \mathcal{A}\nabla v). \tag{2.12}$$

The weak statement of the minimization problem (2.11) becomes: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$a(u^n, \boldsymbol{\sigma}^n; v, \boldsymbol{\tau}) = (\phi^{-1} F_1^n, \phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau}), \quad \forall (v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}. \tag{2.13}$$

Noticing the definition of F_1^n, (2.13) becomes

$$a(u^n, \boldsymbol{\sigma}^n; v, \boldsymbol{\tau}) = (u^{n-1} + \Delta t \phi^{-1}(f^n + \rho_1^n), \phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau}), \quad \forall (v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}. \tag{2.14}$$

Now we consider the second weak formulation different from (2.13). From (2.4) we have that for $n \ge 1$, $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ satisfy that

$$\begin{cases} \phi^{-\frac{1}{2}}(\phi u^n + \Delta t \,\mathrm{div}\, \boldsymbol{\sigma}^n - F_1^n) = 0, & \text{in } \Omega \times J, \\ \mathcal{A}^{-\frac{1}{2}}(\boldsymbol{\sigma}^n + \mathcal{A}\nabla u^n - G^n) = 0, & \text{in } \Omega \times J, \end{cases} \tag{2.15}$$

where $G^n = \boldsymbol{\sigma}^{n-1} + \mathcal{A}\nabla u^{n-1}$. For $(v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}$, define the second kind of least-squares functional $J_2^n(v, \boldsymbol{\tau})$ as follows.

$$J_2^n(v, \boldsymbol{\tau}) = \|\phi^{-\frac{1}{2}}(\phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau} - F^n)\|^2 + \Delta t \|\mathcal{A}^{-\frac{1}{2}}(\boldsymbol{\tau} + \mathcal{A}\nabla v - G^n)\|^2. \tag{2.16}$$

The least-squares minimization problem corresponding to (2.15) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$J_2^n(u^n, \boldsymbol{\sigma}^n) = \inf_{(v,\boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}} J_2^n(v, \boldsymbol{\tau}). \tag{2.17}$$

Similarly to (2.14), the weak statement of (2.17) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$\begin{aligned} a(u^n, \boldsymbol{\sigma}^n; v, \boldsymbol{\tau}) = {} & (u^{n-1} + \Delta t \phi^{-1}(f^n + \rho_1^n), \phi v + \Delta t \,\mathrm{div}\, \boldsymbol{\tau}), + \Delta t (\mathcal{A}^{-1}\boldsymbol{\sigma}^{n-1} + \nabla u^{n-1}, \boldsymbol{\tau} + \mathcal{A}\nabla v) \\ & \forall (v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}. \end{aligned} \tag{2.18}$$

In order to approximate the formulations (2.14) and (2.18), we need to construct the finite element spaces. Let $\mathcal{T}_{h_u}$ and $\mathcal{T}_{h_\sigma}$ be two families of regular finite element partitions of the domain Ω, which are either identical or not. Let h_u and h_σ denote the largest of the diameters of the element in $\mathcal{T}_{h_u}$ and $\mathcal{T}_{h_\sigma}$ respectively. Based on $\mathcal{T}_{h_u}$ and $\mathcal{T}_{h_\sigma}$, respectively, we construct the finite element spaces $\mathrm{V}_h \subset V$ and $\mathbf{W}_h \subset \mathbf{W}$ with the following approximation properties:

$$\inf_{v_h \in \mathrm{V}_h} \{\|v - v_h\| + h_u \|\nabla(v - v_h)\|\} \le C h_u^{m+1} \|v\|_{m+1}, \tag{2.19}$$

$$\inf_{\boldsymbol{\tau}_h \in \mathbf{W}_h} \|\boldsymbol{\tau} - \boldsymbol{\tau}_h\| \le C h_{\boldsymbol{\sigma}}^{k+1} \|\boldsymbol{\tau}\|_{k+1}, \tag{2.20}$$

$$\inf_{\boldsymbol{\tau}_h \in \mathbf{W}_h} \|\,\mathrm{div}(\boldsymbol{\tau} - \boldsymbol{\tau}_h)\| \le C h_{\boldsymbol{\sigma}}^{k_1} \|\boldsymbol{\tau}\|_{k_1+1}, \tag{2.21}$$

for $v \in \mathrm{V} \cap H^{m+1}(\Omega)$ and $\boldsymbol{\tau} \in \mathbf{W} \cap (H^{k_1+1}(\Omega))^d$. It is clear that when assumption (2.20) holds we can deduce $k_1 = k$, and when $\mathbf{W}_h$ is selected as any of the Raviart–Thomas mixed element space [17] we can choose $k_1 = k + 1$. In this paper we always suppose $k_1 = k + 1$ when $\mathbf{W}_h$ is any of the Raviart–Thomas mixed element space [17] and $k_1 = k$ otherwise.

We select the initial approximation $u_h^0 \in \mathrm{V}_h$, $\boldsymbol{\sigma}_h^0 \in \mathbf{W}_h$ such that

$$\begin{cases} \|u_0 - u_h^0\|_j \le C h_u^{m+1-j} \|u_0\|_{m+1}, & j = 0, 1, \\ \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_h^0\| \le C h_{\boldsymbol{\sigma}}^{k+1} \|\boldsymbol{\sigma}_0\|_{k+1}, \end{cases} \tag{2.22}$$

where $\boldsymbol{\sigma}_0 = A\nabla u_0$. The first least-squares finite element procedure based on (2.14) reads as follows.

Scheme (I). For $n \ge 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that

$$a(u_h^n, \boldsymbol{\sigma}_h^n; v_h, \boldsymbol{\tau}_h) = (u_h^{n-1} + \Delta t \phi^{-1} f^n, \phi v_h + \Delta t \,\mathrm{div}\, \boldsymbol{\tau}_h), \quad \forall\, (v_h, \boldsymbol{\tau}_h) \in \mathrm{V}_h \times \mathbf{W}_h. \tag{2.23}$$

Based on (2.18) the second least-squares finite element procedure reads as follows.

Scheme (II). For $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that

$$a(u_h^n, \boldsymbol{\sigma}_h^n; v_h, \boldsymbol{\tau}_h) = (u_h^{n-1} + \Delta t \phi^{-1} f^n, \phi v_h + \Delta t \operatorname{div} \boldsymbol{\tau}_h) + \Delta t(\mathcal{A}^{-1}\boldsymbol{\sigma}_h^{n-1} + \nabla u_h^{n-1}, \boldsymbol{\tau}_h + \mathcal{A}\nabla v_h), \quad \forall (v_h, \boldsymbol{\tau}_h) \in \mathrm{V}_h \times \mathbf{W}_h. \tag{2.24}$$

Now let us mention about the bilinear form $a(\cdot, \cdot; \cdot, \cdot)$ in the following lemma, which leads to decoupled systems.

Lemma 2.1. *For any $u, v \in \mathrm{V}$ and $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{W}$ we have that,*

$$a(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = (\phi u, v) + \Delta t(\mathcal{A}\nabla u, \nabla v) + \Delta t(\mathcal{A}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + \Delta t^2(\phi^{-1} \operatorname{div} \boldsymbol{\sigma}, \operatorname{div} \boldsymbol{\tau}). \tag{2.25}$$

Proof. A direct calculation shows that

$$a(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = (\phi u, v) + \Delta t(\mathcal{A}\nabla u, \nabla v) + \Delta t(\mathcal{A}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + \Delta t^2(\phi^{-1} \operatorname{div} \boldsymbol{\sigma}, \operatorname{div} \boldsymbol{\tau}) + \Delta t((u, \operatorname{div} \boldsymbol{\tau}) + (v, \operatorname{div} \boldsymbol{\sigma}) + (\nabla u, \boldsymbol{\tau}) + (\nabla v, \boldsymbol{\sigma})),$$

Integrating by parts shows that

$$(u, \operatorname{div} \boldsymbol{\tau}) + (v, \operatorname{div} \boldsymbol{\sigma}) + (\nabla u, \boldsymbol{\tau}) + (\nabla v, \boldsymbol{\sigma}) = 0, \tag{2.26}$$

which completes the proof. □

Using Lemma 2.1, we have the decoupling equivalent form of each scheme (I) or (II) alternatively by putting $\boldsymbol{\tau}_h = 0$ and $v_h = 0$ in (2.23) or (2.24).

Equivalent Form of Scheme (I). With the initial guess $(u_h^0, \boldsymbol{\sigma}_h^0) \in \mathrm{V}_h \times \mathbf{W}_h$, for $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that for all $v_h \in \mathrm{V}_h$ and $\boldsymbol{\tau}_h \in \mathbf{W}_h$

$$(\phi u_h^n, v_h) + \Delta t(\mathcal{A}\nabla u_h^n, \nabla v_h) = (\phi u_h^{n-1} + \Delta t f^n, v_h), \tag{2.27}$$

$$(\mathcal{A}^{-1}\boldsymbol{\sigma}_h^n, \boldsymbol{\tau}_h) + \Delta t(\phi^{-1} \operatorname{div} \boldsymbol{\sigma}_h^n, \operatorname{div} \boldsymbol{\tau}_h) = (u_h^{n-1} + \Delta t \phi^{-1} f^n, \operatorname{div} \boldsymbol{\tau}_h). \tag{2.28}$$

Equivalent Form of Scheme (II). With the initial guess $(u_h^0, \boldsymbol{\sigma}_h^0) \in \mathrm{V}_h \times \mathbf{W}_h$, for $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that for all $v_h \in \mathrm{V}_h$ and $\boldsymbol{\tau}_h \in \mathbf{W}_h$

$$(\phi u_h^n, v_h) + \Delta t(\mathcal{A}\nabla u_h^n, \nabla v_h) = (\phi u_h^{n-1} + \Delta t f^n, v_h) + \Delta t(\boldsymbol{\sigma}_h^{n-1} + \mathcal{A}\nabla u_h^{n-1}, \nabla v_h) \tag{2.29}$$

$$(\mathcal{A}^{-1}\boldsymbol{\sigma}_h^n, \boldsymbol{\tau}_h) + \Delta t(\phi^{-1} \operatorname{div} \boldsymbol{\sigma}_h^n, \operatorname{div} \boldsymbol{\tau}_h) = (\mathcal{A}^{-1}\boldsymbol{\sigma}_h^{n-1}, \boldsymbol{\tau}_h) + \Delta t(\phi^{-1} f^n, \operatorname{div} \boldsymbol{\tau}_h). \tag{2.30}$$

Note that each Scheme (I) or (II) is split into two independent symmetric positive definite systems. Sub-procedure (2.27) is the same as the standard Galerkin finite element procedure for parabolic problems. Sub-procedure (2.30) is a procedure for the unknown flux $\boldsymbol{\sigma}_h^n$ with first-order approximation in time increment.

It clear that both problems (2.23) and (2.24) have a unique solution.

Now we consider the second-order approximation in time increment. Let

$$\rho_2^n := \phi\left(\delta_t u^n - u_t^{n-\frac{1}{2}}\right) + \frac{1}{2} \operatorname{div}(\boldsymbol{\sigma}^n + \boldsymbol{\sigma}^{n-1}) - \operatorname{div} \boldsymbol{\sigma}^{n-\frac{1}{2}}, \tag{2.31}$$

which can be estimated as

$$\rho_2^n = O\left(\Delta t^{\frac{3}{2}} \left(\int_{t^{n-1}}^{t^n} (|u_{ttt}|^2 + |\operatorname{div} \boldsymbol{\sigma}_{tt}|^2)\mathrm{d}t\right)^{\frac{1}{2}}\right). \tag{2.32}$$

From (2.4) we know that for $n \geq 1$, $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ satisfy that

$$\begin{cases} \phi^{-\frac{1}{2}}\left(\phi u^n + \dfrac{\Delta t}{2} \operatorname{div} \boldsymbol{\sigma}^n - F_2^n\right) = 0, & \text{in } \Omega \times J, \\ \mathcal{A}^{-\frac{1}{2}}(\boldsymbol{\sigma}^n + \mathcal{A}\nabla u^n - G^n) = 0, & \text{in } \Omega \times J, \end{cases} \tag{2.33}$$

where G^n is the same as in (2.15),

$$F_2^n = \phi u^{n-1} + \Delta t f^{n-\frac{1}{2}} - \frac{\Delta t}{2} \operatorname{div} \boldsymbol{\sigma}^{n-1} + \Delta t \rho_2^n. \tag{2.34}$$

For $(v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}$, define the least-squares functional $J_3^n(v, \boldsymbol{\tau})$ as follows.

$$J_3^n(v, \boldsymbol{\tau}) = \left\|\phi^{-\frac{1}{2}}\left(\phi v + \frac{\Delta t}{2} \operatorname{div} \boldsymbol{\tau} - F_2^n\right)\right\|^2 + \frac{\Delta t}{2}\|\mathbf{A}^{-\frac{1}{2}}(\boldsymbol{\tau} + \mathcal{A}\nabla v - G^n)\|^2. \tag{2.35}$$

The least-squares minimization problem corresponding to (2.33) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$J_3^n(u^n, \boldsymbol{\sigma}^n) = \inf_{v \in \mathrm{V}, \boldsymbol{\tau} \in \mathbf{W}} J_3^n(v, \boldsymbol{\tau}). \tag{2.36}$$

Define the bilinear form $b(\cdot, \cdot; \cdot, \cdot)$ as

$$b(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = \left(u + \frac{\Delta t}{2}\phi^{-1}\operatorname{div}\boldsymbol{\sigma}, \phi v + \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\tau}\right) + \frac{\Delta t}{2}(\mathcal{A}^{-1}\boldsymbol{\sigma} + \nabla u, \boldsymbol{\tau} + \mathcal{A}\nabla v). \tag{2.37}$$

Noticing the definition of F_2^n in (2.34), the weak statement of the minimization problem (2.36) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$\begin{aligned} b(u^n, \boldsymbol{\sigma}^n; v, \boldsymbol{\tau}) \;=\; & \left(u^{n-1} + \Delta t\phi^{-1}\left(f^{n-\frac{1}{2}} - \frac{1}{2}\operatorname{div}\boldsymbol{\sigma}^{n-1} + \rho_2^n\right), \phi v + \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\tau}\right), \\ & + \frac{\Delta t}{2}(\mathcal{A}^{-1}\boldsymbol{\sigma}^{n-1} + \nabla u^{n-1}, \boldsymbol{\tau} + \mathcal{A}\nabla v) \quad \forall (v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}. \end{aligned} \tag{2.38}$$

Then the corresponding least-squares finite element procedure reads as follows.

Scheme (III). With the initial guess $(u_h^0, \boldsymbol{\sigma}_h^0) \in \mathrm{V}_h \times \mathbf{W}_h$, for $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that

$$\begin{aligned} b(u_h^n, \boldsymbol{\sigma}_h^n; v_h, \boldsymbol{\tau}_h) \;=\; & \left(u_h^{n-1} + \Delta t\phi^{-1}\left(f^{n-\frac{1}{2}} - \frac{1}{2}\operatorname{div}\boldsymbol{\sigma}_h^{n-1}\right), \phi v_h + \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\tau}_h\right) \\ & + \frac{\Delta t}{2}(\mathcal{A}^{-1}\boldsymbol{\sigma}_h^{n-1} + \nabla u_h^{n-1}, \boldsymbol{\tau}_h + \mathcal{A}\nabla v_h), \quad \forall\, (v_h, \boldsymbol{\tau}_h) \in \mathrm{V}_h \times \mathbf{W}_h. \end{aligned} \tag{2.39}$$

Similarly to Lemma 2.1 we know that the following lemma holds.

Lemma 2.2. *For any $u, v \in \mathrm{V}$ and $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{W}$ we have that,*

$$b(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = (\phi u, v) + \frac{\Delta t}{2}(\mathcal{A}\nabla u, \nabla v) + \frac{\Delta t}{2}(\mathcal{A}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + \left(\frac{\Delta t}{2}\right)^2 (\phi^{-1}\operatorname{div}\boldsymbol{\sigma}, \operatorname{div}\boldsymbol{\tau}). \tag{2.40}$$

Using Lemma 2.2 we have a decoupling equivalent form of Scheme (III).

Equivalent Form of Scheme (III). With the initial guess $(u_h^0, \boldsymbol{\sigma}_h^0) \in \mathrm{V}_h \times \mathbf{W}_h$, for $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that

$$\begin{aligned} (\phi u_h^n, v_h) + \frac{\Delta t}{2}(\mathcal{A}\nabla u_h^n, \nabla v_h) \;=\; & \left(\phi u_h^{n-1} + \Delta t f^{n-\frac{1}{2}} - \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\sigma}_h^{n-1}, v_h\right) + \frac{\Delta t}{2}(\boldsymbol{\sigma}_h^{n-1} + \mathcal{A}\nabla u_h^{n-1}, \nabla v_h), \\ & \forall\, v_h \in \mathrm{V}_h. \end{aligned} \tag{2.41}$$

$$\begin{aligned} (\mathcal{A}^{-1}\boldsymbol{\sigma}_h^n, \boldsymbol{\tau}_h) + \frac{\Delta t}{2}(\phi^{-1}\operatorname{div}\boldsymbol{\sigma}_h^n, \operatorname{div}\boldsymbol{\tau}_h) \;=\; & (\mathcal{A}^{-1}\boldsymbol{\sigma}_h^{n-1}, \boldsymbol{\tau}_h) + \Delta t\left(\phi^{-1}\left(f^{n-\frac{1}{2}} - \frac{1}{2}\operatorname{div}\boldsymbol{\sigma}_h^{n-1}\right), \operatorname{div}\boldsymbol{\tau}_h\right), \\ & \forall\, \boldsymbol{\tau}_h \in \mathbf{W}_h. \end{aligned} \tag{2.42}$$

Then this scheme also can be split into two independent sub-procedures. Sub-procedure (2.42) is a procedure for the unknown flux $\boldsymbol{\sigma}_h^n$ with second-order approximation in time increment.

Remark 2.3. Results similar to Lemma 2.1 or Lemma 2.2 have been found and used by [8] to prove the coercivity of least-squares bilinear formats and by [2,3] to establish connections between least-squares and mixed methods.

3. Error estimates

In this section we give the error estimates for the schemes described in Section 2.

We first discuss the error estimate for Scheme (I) in the following Theorem 3.1.

Theorem 3.1. *Suppose $(u_h^n, \boldsymbol{\sigma}_h^n) \in V_h \times \mathbf{W}_h$ is the solution of Scheme (I). Under the assumption $\|u_h^0 - u^0\| = O(h_u^{m+1-j}\|u^0\|_{H^{m+1-j}})$, $j = 0, 1$, there exists a positive constant C independent of h_u, h_σ and Δt such that*

$$\|u_h^n - u^n\|_s \leq Ch_u^{m+1-s}(\|u\|_{L^\infty(H^{m+1})} + \|u_t\|_{L^2(H^{m+1})}) + C\Delta t\|u_{tt}\|_{L^2(L^2)}, \quad s = 0, 1, \tag{3.1}$$

$$\begin{aligned} \|\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}\| + \Delta t^{\frac{1}{2}}\|\operatorname{div}(\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}^n)\| \;\leq\; & C(h_{\boldsymbol{\sigma}}^{k+1}\|\boldsymbol{\sigma}^n\|_{k+1} + \Delta t^{\frac{1}{2}} h_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}^n\|_{k_1+1} + \Delta t\|u_{tt}\|_{L^2(L^2)}) \\ & + C\min\{h_u^m, \Delta t^{-\frac{1}{2}} h_u^{m+1}\}(\|u\|_{L^\infty(H^{m+1})} + \|u_t\|_{L^2(H^{m+1})}). \end{aligned} \tag{3.2}$$

Proof. Since Scheme (I) is equivalent to (2.27) and (2.28), from the error estimates of the finite element method for parabolic problems (see [18] and [19] for example), we know that (3.1) holds.

We next consider $\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i$, $1 \le i \le n \le \frac{T}{\Delta t}$. Subtracting (2.14) from (2.23) and setting $v_h = 0$, using Lemma 2.1, we have

$$\begin{aligned}(\mathcal{A}^{-1}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i), \boldsymbol{\tau}_h) + \Delta t(\phi^{-1}\operatorname{div}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i), \operatorname{div}\boldsymbol{\tau}_h) &= (u_h^{i-1} - u^{i-1}, \operatorname{div}\boldsymbol{\tau}_h) - \Delta t(\phi^{-1}\rho_1^i, \operatorname{div}\boldsymbol{\tau}_h) \quad \forall \boldsymbol{\tau}_h \in \mathbf{W}_h \\ &= -(\nabla(u_h^{i-1} - u^{i-1}), \boldsymbol{\tau}_h) - \Delta t(\phi^{-1}\rho_1^i, \operatorname{div}\boldsymbol{\tau}_h). \end{aligned} \tag{3.3}$$

Let $\boldsymbol{\sigma}_I^i \in \mathbf{W}_h$ be an interpolant of $\boldsymbol{\sigma}^i$ such that

$$\begin{cases} \|\boldsymbol{\sigma}_I^i - \boldsymbol{\sigma}^i\| \le C h_\sigma^{k+1}\|\boldsymbol{\sigma}^i\|_{k+1}, \\ \|\operatorname{div}(\boldsymbol{\sigma}_I^i - \boldsymbol{\sigma}^i)\| \le C h_\sigma^{k_1}\|\boldsymbol{\sigma}^i\|_{k_1+1}. \end{cases} \tag{3.4}$$

Denote by

$$\xi_\sigma^i = \boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}_I^i. \tag{3.5}$$

Setting $\boldsymbol{\tau}_h = \xi_\sigma^i = \boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}_I^i$ in (3.3), and using the ϵ-inequality, (2.6), we have

$$\begin{aligned} &\|\mathcal{A}^{-\frac12}\xi_\sigma^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_\sigma^i\|^2 \\ &\quad = (\mathcal{A}^{-1}(\boldsymbol{\sigma}^i - \boldsymbol{\sigma}_I^i); \xi_\sigma^i) + \Delta t(\phi^{-1}\operatorname{div}(\boldsymbol{\sigma}^i - \boldsymbol{\sigma}_I^i); \operatorname{div}\xi_\sigma^i) - (\nabla(u_h^{i-1} - u^{i-1}), \xi_\sigma^i) + \Delta t(\phi^{-1}\rho_1^i, \operatorname{div}\xi_\sigma^i) \\ &\quad \le \frac12\left(\|\mathcal{A}^{-\frac12}\xi_\sigma^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_\sigma^i\|^2\right) + C\left[\|\boldsymbol{\sigma}^i - \boldsymbol{\sigma}_I^i\|^2 + \Delta t\|\operatorname{div}(\boldsymbol{\sigma}^i - \boldsymbol{\sigma}_I^i)\|^2 + \|\nabla(u_h^{i-1} - u^{i-1})\|^2 + \Delta t\|\rho_1^i\|^2\right] \\ &\quad \le \frac12\left(\|\mathcal{A}^{-\frac12}\xi_\sigma^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_\sigma^i\|^2\right) + C\Delta t^2\|u_{tt}\|^2_{L^2(L^2)} \\ &\qquad + C\left(h_\sigma^{2(k+1)}\|\boldsymbol{\sigma}^i\|_{k+1}^2 + \Delta t h_\sigma^{2k_1}\|\boldsymbol{\sigma}^i\|_{k_1+1}^2 + \|\nabla(u_h^{i-1} - u^{i-1})\|^2\right). \end{aligned} \tag{3.6}$$

By using

$$\begin{aligned} -(\nabla(u_h^{i-1} - u^{i-1}), \xi_\sigma^i) &= (u_h^{i-1} - u^{i-1}, \operatorname{div}\xi_\sigma^i) \\ &\le C\Delta t^{-\frac12}\|u_h^{i-1} - u^{i-1}\|\Delta t^{\frac12}\|\phi^{\frac12}\operatorname{div}\xi_\sigma^i\|, \end{aligned}$$

we have the following estimate instead of (3.6),

$$\begin{aligned} \|\mathcal{A}^{-\frac12}\xi_\sigma^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_\sigma^i\|^2 &\le \frac12\left(\|\mathcal{A}^{-\frac12}\xi_\sigma^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_\sigma^i\|^2\right) + C\Delta t^2\|u_{tt}\|^2_{L^2(L^2)} \\ &\quad + C\left(h_\sigma^{2(k+1)}\|\boldsymbol{\sigma}^i\|_{k+1}^2 + \Delta t h_\sigma^{2k_1}\|\boldsymbol{\sigma}^i\|_{k_1+1}^2 + \Delta t^{-1}\|u_h^{i-1} - u^{i-1}\|^2\right). \end{aligned} \tag{3.7}$$

Then, using (3.1) and (3.7), and the positive definiteness of $\mathcal{A}$, we have that

$$\begin{aligned} \|\xi_\sigma^i\| + \Delta t^{\frac12}\|\operatorname{div}\xi_\sigma^i\| &\le C(h_\sigma^{k+1}\|\boldsymbol{\sigma}^i\|_{k+1} + \Delta t^{\frac12}h_\sigma^{k_1}\|\boldsymbol{\sigma}^i\|_{k_1+1} + \Delta t\|u_{tt}\|_{L^2(L^2)}) \\ &\quad + C\min\{h_u^m, \Delta t^{-\frac12}h_u^{m+1}\}(\|u\|_{L^\infty(H^{m+1})} + \|u_t\|_{L^2(H^{m+1})}). \end{aligned} \tag{3.8}$$

Combining (3.8) with (3.4) completes the proof. □

For the error estimates for Scheme (II), for any $i \le n \le \frac{T}{\Delta t}$ we define the auxiliary projection $\tilde{u}_h^i \in \mathrm{V}_h$ satisfying

$$(\mathcal{A}\nabla(\tilde{u}_h^i - u^i), \nabla v_h) = 0, \quad \forall v_h \in \mathrm{V}_h. \tag{3.9}$$

From this definition we have that

$$(\mathcal{A}\nabla\delta_t(\tilde{u}_h^i - u^i), \nabla v_h) = 0, \quad \forall v_h \in \mathrm{V}_h. \tag{3.10}$$

From [7] it is easy to see that that

$$\begin{cases} \|\tilde{u}_h^i - u^i\|_j \le C h_u^{m+1-j}\|u^i\|_{m+1}, \quad j = 0, 1, \\ \|\delta_t(\tilde{u}_h^i - u^i)\|_j \le C h_u^{m+1-j}\left(\dfrac{1}{\Delta t}\displaystyle\int_{t^{i-1}}^{t^i}\|u_t\|_{m+1}^2 \mathrm{d}t\right)^{\frac12}, \quad j = 0, 1. \end{cases} \tag{3.11}$$

Theorem 3.2. *Suppose $(u_h^n, \boldsymbol{\sigma}_h^n) \in V_h \times \mathbf{W}_h$ is the solution of Scheme* (II)*. The initial guess satisfies $\|\boldsymbol{\sigma}_h^0 - \boldsymbol{\sigma}^0\| = O(h_u^{k_1}\|\boldsymbol{\sigma}^0\|_{H^{k_1}})$. When h_u, h_σ and Δt are sufficiently small, there exists a positive constant C independent of h_u, h_σ and Δt such that*

$$\|\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}^n\| + \left(\sum_{i=1}^n \Delta t\|\operatorname{div}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i)\|^2\right)^{\frac12} \le C(h_\sigma^{k_1}\|\boldsymbol{\sigma}\|_{L^\infty(H^{k_1+1})} + h_\sigma^{k+1}\|\boldsymbol{\sigma}_t\|_{L^2(H^{k+1})} + \Delta t\|u_{tt}\|_{L^2(L^2)}). \tag{3.12}$$

Further, if $u_h^0 = \tilde{u}_h^0$ holds there, we have that

$$\|u_h^n - u^n\| \le C h_u^{m+1}(\|u\|_{L^\infty(H^{m+1})} + \|u_t\|_{L^2(H^{m+1})}) + C h_{\boldsymbol\sigma}^{k_1}\|\boldsymbol\sigma\|_{L^\infty(H^{k_1+1})} + C h_{\boldsymbol\sigma}^{k+1}\|\boldsymbol\sigma_t\|_{L^2(H^{k+1})} + C\Delta t\|u_{tt}\|_{L^2(L^2)}. \tag{3.13}$$

Proof. Subtracting (2.18) from (2.24) we have that

$$\begin{aligned} a(u_h^i - u^i, \boldsymbol\sigma_h^i - \boldsymbol\sigma^i; v_h, \boldsymbol\tau_h) &= (u_h^{i-1} - u^{i-1}, \phi v_h + \Delta t \operatorname{div}\boldsymbol\tau_h) + \Delta t(\mathcal{A}^{-1}(\boldsymbol\sigma_h^{i-1} - \boldsymbol\sigma^{i-1}) + \nabla(u_h^{i-1} - u^{i-1}), \boldsymbol\tau_h \\ &\quad + \mathcal{A}\nabla v_h), -\Delta t(\phi^{-1}\rho_1^i, \phi v_h + \Delta t \operatorname{div}\boldsymbol\tau_h), \quad \forall (v_h, \boldsymbol\tau_h) \in \mathrm{V}_h \times \mathbf{W}_h. \end{aligned} \tag{3.14}$$

Setting $v_h = 0$, using Lemma 2.1 and the divergence theorem, we have for $\boldsymbol\tau_h \in \mathbf{W}_h$ that

$$(\mathcal{A}^{-1}(\boldsymbol\sigma_h^i - \boldsymbol\sigma^i), \boldsymbol\tau_h) + \Delta t(\phi^{-1}\operatorname{div}(\boldsymbol\sigma_h^i - \boldsymbol\sigma^i), \operatorname{div}\boldsymbol\tau_h) = (\mathcal{A}^{-1}(\boldsymbol\sigma_h^{i-1} - \boldsymbol\sigma^{i-1}), \boldsymbol\tau_h) - \Delta t(\phi^{-1}\rho_1^i, \operatorname{div}\boldsymbol\tau_h), \tag{3.15}$$

which can be written as

$$\begin{aligned} &(\mathcal{A}^{-1}\xi_{\boldsymbol\sigma}^i, \boldsymbol\tau_h) + \Delta t(\phi^{-1}\operatorname{div}\xi_{\boldsymbol\sigma}^i, \operatorname{div}\boldsymbol\tau_h) \\ &\quad = (\mathcal{A}^{-1}\xi_{\boldsymbol\sigma}^{i-1}, \boldsymbol\tau_h) + \Delta t\left(\mathcal{A}^{-1}\delta_t(\boldsymbol\sigma^i - \boldsymbol\sigma_I^i), \boldsymbol\tau_h\right) + \Delta t\left(\phi^{-1}\operatorname{div}(\boldsymbol\sigma^i - \boldsymbol\sigma_I^i), \operatorname{div}\boldsymbol\tau_h\right) - \Delta t(\phi^{-1}\rho_1^i, \operatorname{div}\boldsymbol\tau_h) \\ &\quad \le \frac{1}{2}\|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^{i-1}\|^2 + \frac12\|\mathcal{A}^{-\frac12}\boldsymbol\tau_h\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{-\frac12}\delta_t(\boldsymbol\sigma^i - \boldsymbol\sigma_I^i)\|^2 + \frac{\Delta t}{2}\|\boldsymbol\tau_h\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}(\boldsymbol\sigma^i - \boldsymbol\sigma_I^i)\|^2 \\ &\qquad + \frac{\Delta t}{4}\|\phi^{-\frac12}\operatorname{div}\boldsymbol\tau_h\|^2 + \Delta t\|\phi^{-\frac12}\rho_1^i\|^2 + \frac{\Delta t}{4}\|\phi^{-\frac12}\operatorname{div}\boldsymbol\tau_h\|^2, \end{aligned} \tag{3.16}$$

where the notation δ_t is defined in (2.5).

Note that ϕ is bounded below and above, $0 < \phi_1 \le \phi \le \phi_2$. Then putting $\boldsymbol\tau_h = \xi_{\boldsymbol\sigma}^i$ in (3.16) and using the ϵ-inequality we have

$$\begin{aligned} \|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_{\boldsymbol\sigma}^i\|^2 &\le \frac{1+\Delta t}{2}\|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^i\|^2 + \frac{\Delta t}{2}\|\phi^{-\frac12}\operatorname{div}\xi_{\boldsymbol\sigma}^i\|^2 + \frac12\|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^{i-1}\|^2 \\ &\quad + C\Delta t\left[\|\delta_t(\boldsymbol\sigma - \boldsymbol\sigma_I)^i\|^2 + \|\operatorname{div}(\boldsymbol\sigma - \boldsymbol\sigma_I)^i\|^2 + \|\rho_1^i\|^2\right]. \end{aligned} \tag{3.17}$$

Since

$$\|\delta_t(\boldsymbol\sigma - \boldsymbol\sigma_I)^i\| = \left\|\frac{1}{\Delta t}\int_{t^{i-1}}^{t^i}(\boldsymbol\sigma - \boldsymbol\sigma_I)_t\,\mathrm{d}t\right\| \le C h_{\boldsymbol\sigma}^{k+1}\left(\frac{1}{\Delta t}\int_{t^{i-1}}^{t^i}\|\boldsymbol\sigma_t\|_{k+1}^2\,\mathrm{d}t\right)^{\frac12},$$

applying (3.4) and (2.6) to (3.17) we have

$$\begin{aligned} \|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^i\|^2 + \Delta t\|\phi^{-\frac12}\operatorname{div}\xi_{\boldsymbol\sigma}^i\|^2 &\le \|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^{i-1}\|^2 + \Delta t\|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^i\|^2 + C\Delta t^2\int_{t^{i-1}}^{t^i}\|u_{tt}\|^2\mathrm{d}t \\ &\quad + C\left[h_{\boldsymbol\sigma}^{2(k+1)}\int_{t^{i-1}}^{t^i}\|\boldsymbol\sigma_t\|_{k+1}^2\mathrm{d}t + \Delta t\, h_{\boldsymbol\sigma}^{2k_1}\|\boldsymbol\sigma\|_{L^\infty(H^{k_1+1})}^2\right]. \end{aligned} \tag{3.18}$$

Carrying out summation for $i = 1, 2, \ldots, n$ we have that

$$\begin{aligned} \|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^n\|^2 + \sum_{i=1}^n\Delta t\|\phi^{-\frac12}\operatorname{div}\xi_{\boldsymbol\sigma}^n\|^2 &\le \sum_{i=1}^n\Delta t\|\mathcal{A}^{-\frac12}\xi_{\boldsymbol\sigma}^i\|^2 + \|\mathcal{A}^{-\frac12}(\boldsymbol\sigma_h^0 - \boldsymbol\sigma_I^0)\|^2 + C\Delta t^2\int_0^T\|u_{tt}\|^2\mathrm{d}t \\ &\quad + C\left[h_{\boldsymbol\sigma}^{2(k+1)}\int_0^T\|\boldsymbol\sigma_t\|_{k+1}^2\mathrm{d}t + h_{\boldsymbol\sigma}^{2k_1}\|\boldsymbol\sigma\|_{L^\infty(H^{k_1+1})}^2\right]. \end{aligned} \tag{3.19}$$

Noticing $\boldsymbol\sigma_h^0 - \boldsymbol\sigma_I^0 = (\boldsymbol\sigma_h^0 - \boldsymbol\sigma^0) + (\boldsymbol\sigma^0 - \boldsymbol\sigma_I^0)$, using Gronwall's inequality shows that

$$\|\xi_{\boldsymbol\sigma}^n\|^2 + \sum_{i=1}^n\Delta t\|\operatorname{div}\xi_{\boldsymbol\sigma}^i\|^2 \le C\left[h_{\boldsymbol\sigma}^{2k_1}\|\boldsymbol\sigma\|_{L^\infty(H^{k_1+1})}^2 + h_{\boldsymbol\sigma}^{2(k+1)}\|\boldsymbol\sigma_t\|_{L^2(H^{k+1})}^2 + \Delta t^2\|u_{tt}\|_{L^2(L^2)}^2\right]. \tag{3.20}$$

Combining with (3.4) completes the proof of (3.12).

Now we consider the estimate of $u_h^i - u^i$, for $i \le n \le \frac{T}{\Delta t}$. Letting $\boldsymbol\tau_h = 0$ in (3.14), using Lemma 2.1 and the divergence theorem lead to

$$\begin{aligned} (\phi(u_h^i - u^i), v_h) + \Delta t(\mathcal{A}\nabla(u_h^i - u^i), \nabla v_h) &= (\phi(u_i^{i-1} - u^{i-1}), v_h) - \Delta t(\rho_1^i, v_h) + \Delta t(\boldsymbol\sigma_h^{i-1} - \boldsymbol\sigma^{i-1}, \nabla v_h) \\ &\quad + \Delta t(\mathcal{A}\nabla(u_h^{i-1} - u^{i-1}), \nabla v_h), \quad \forall v_h \in \mathrm{V}_h. \end{aligned} \tag{3.21}$$

With the use of the definition of $\tilde{u}_h^i$, we have

$$\begin{aligned}&(\phi(u_h^i-\tilde{u}_h^i),v_h)+\Delta t(\mathcal{A}\nabla(u_h^i-\tilde{u}_h^i),\nabla v_h)\\&\quad=(\phi(u_h^{i-1}-\tilde{u}_h^{i-1}),v_h)+\Delta t(\phi\delta_t(u^i-\tilde{u}_h^i),v_h)-\Delta t(\operatorname{div}(\boldsymbol{\sigma}_h^{i-1}-\boldsymbol{\sigma}^{i-1}),v_h)-\Delta t(\rho_1^i,v_h)\\&\qquad+\Delta t(\mathcal{A}\nabla(u_h^{i-1}-\tilde{u}_h^{i-1}),\nabla v_h),\quad\forall v_h\in \mathrm{V}_h.\end{aligned}\tag{3.22}$$

Let

$$\xi_u^i=u_h^i-\tilde{u}_h^i,\qquad \eta_u^i=u^i-\tilde{u}_h^i.\tag{3.23}$$

With the choice $v_h=\xi_u^i=u_h^i-\tilde{u}_h^i$ in (3.22), it follows that

$$\begin{aligned}\|\phi^{\frac12}\xi_u^i\|^2+\Delta t\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2\le\ &\frac{1+\Delta t}{2}\|\phi^{\frac12}\xi_u^i\|^2+\frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2+\frac12\|\phi^{\frac12}\xi_u^{i-1}\|^2+\frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^{i-1}\|^2\\&+C\Delta t\left[\|\delta_t(u^i-\tilde{u}_h^i)\|^2+\|\operatorname{div}(\boldsymbol{\sigma}_h^{i-1}-\boldsymbol{\sigma}^{i-1})\|^2+\|\rho_1^i\|^2\right],\end{aligned}\tag{3.24}$$

which can be reduced to

$$\begin{aligned}\|\phi^{\frac12}\xi_u^i\|^2+\Delta t\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2\le\ &\|\phi^{\frac12}\xi_u^{i-1}\|^2+\Delta t\|\phi^{\frac12}\xi_u^i\|^2+\Delta t\|\mathcal{A}^{\frac12}\nabla\xi_u^{i-1}\|^2\\&+C\Delta t\left[\|\delta_t(u^i-\tilde{u}_h^i)\|^2+\|\operatorname{div}(\boldsymbol{\sigma}_h^{i-1}-\boldsymbol{\sigma}^{i-1})\|^2+\|\rho_1^i\|^2\right].\end{aligned}\tag{3.25}$$

Summing (3.25) from $i=1$ to n and noticing (3.12) we have

$$\begin{aligned}\|\phi^{\frac12}\xi_u^n\|^2+\Delta t\|\mathbf{A}^{\frac12}\nabla\xi_u^n\|^2\le\ &\sum_{i=1}^n\Delta t\|\phi^{\frac12}\xi_u^i\|^2+\|\phi^{\frac12}\xi_u^0\|^2+\Delta t\|\mathcal{A}^{\frac12}\nabla\xi_u^0\|^2+C(h_u^{2(m+1)}\|u_t\|^2_{L^2(H^{m+1})}\\&+\Delta t^2\|u_{tt}\|^2_{L^2(L^2)})+C\sum_{i=1}^n\Delta t\|\operatorname{div}(\boldsymbol{\sigma}_h^{i-1}-\boldsymbol{\sigma}^{i-1})\|^2\\\le\ &\sum_{i=1}^n\Delta t\|\phi^{\frac12}\xi_u^n\|^2+C\left[h_u^{2(m+1)}\|u_t\|^2_{L^2(H^{m+1})}+\Delta t^2\|u_{tt}\|^2_{L^2(L)}\right]\\&+C\left[h_{\boldsymbol{\sigma}}^{2k_1}\|\boldsymbol{\sigma}\|^2_{L^\infty(H^{k_1+1})}+h_{\boldsymbol{\sigma}}^{2(k+1)}\|\boldsymbol{\sigma}_t\|^2_{L^2(H^{k+1})}\right].\end{aligned}\tag{3.26}$$

Therefore we can apply Gronwall's inequality to (3.26). Hence it follows that

$$\|\xi_u^n\|+\Delta t^{\frac12}\|\nabla\xi_u^n\|\le C\left[h_u^{m+1}\|u_t\|_{L^2(H^{m+1})}+\Delta t\|u_{tt}\|_{L^2(L^2)}\right]+C\left[h_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}\|_{L^\infty(H^{k_1+1})}+h_{\boldsymbol{\sigma}}^{k+1}\|\boldsymbol{\sigma}_t\|_{L^2(H^{k+1})}\right].$$

Finally, combining (3.27) with (3.11) completes the proof. □

Remark 3.3. Instead of $u_h^0=\tilde{u}_h^0$ if we suppose $\|u_h^0-u^0\|_j=O(h_u^{m+1-j})$, from the proof we know that replacing (3.13) we have a estimate

$$\|u_h^n-u^n\|\le C(h_u^{m+1}+\Delta t^{\frac12}h_u^m+h_{\boldsymbol{\sigma}}^{k_1}+\Delta t).$$

Now we give the error estimate for Scheme (III). For this purpose, define $\tilde{\boldsymbol{\sigma}}_h^i\in\mathbf{W}_h$ such that

$$(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}^i,\boldsymbol{\tau}_h)+(\phi^{-1}\operatorname{div}(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}^i),\ \operatorname{div}\boldsymbol{\tau}_h)=0,\quad\forall\boldsymbol{\tau}_h\in\mathbf{W}_h.\tag{3.27}$$

It is clear that $\tilde{\boldsymbol{\sigma}}_h^i$ exist uniquely. By splitting $\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}^i$ as $\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}^i=(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}_I^i)+(\boldsymbol{\sigma}_I^i-\boldsymbol{\sigma}^i)$ and using (3.27), we have

$$\begin{aligned}&(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}_I^i,\boldsymbol{\tau}_h)+(\phi^{-1}\operatorname{div}(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}_I^i),\operatorname{div}\boldsymbol{\tau}_h)=(\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i,\boldsymbol{\tau}_h)+(\phi^{-1}\operatorname{div}(\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i),\ \operatorname{div}\boldsymbol{\tau}_h)\\&\quad\le\frac12\|\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i\|^2+\frac12\|\boldsymbol{\tau}_h\|^2+\frac12\|\phi^{-\frac12}\operatorname{div}(\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i)\|+\frac12\|\phi^{-\frac12}\operatorname{div}\boldsymbol{\tau}_h\|^2,\quad\forall\boldsymbol{\tau}_h\in\mathbf{W}_h.\end{aligned}\tag{3.28}$$

Let $\boldsymbol{\tau}_h=\boldsymbol{\sigma}_h^i-\boldsymbol{\sigma}_I^i$. We have the following error estimate,

$$\|\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}^i\|+\|\operatorname{div}(\tilde{\boldsymbol{\sigma}}^i-\boldsymbol{\sigma}^i)\|\le Ch_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}^i\|_{k_1+1}.\tag{3.29}$$

From (3.28) we also get that

$$\begin{aligned}(\delta_t(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}_I^i),\boldsymbol{\tau}_h)+(\phi^{-1}\operatorname{div}\delta_t(\tilde{\boldsymbol{\sigma}}_h^i-\boldsymbol{\sigma}_I^i),\ \operatorname{div}\boldsymbol{\tau}_h)\ =\ &(\delta_t(\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i),\boldsymbol{\tau}_h)+(\phi^{-1}\operatorname{div}\delta_t(\boldsymbol{\sigma}^i-\boldsymbol{\sigma}_I^i),\ \operatorname{div}\boldsymbol{\tau}_h),\\&\forall\boldsymbol{\tau}_h\in\mathbf{W}_h.\end{aligned}\tag{3.30}$$

Let $\boldsymbol{\tau}_h=\delta_t(\boldsymbol{\sigma}_h^i-\boldsymbol{\sigma}_I^i)$. We have the following error estimate similarly,

$$\|\delta_t(\tilde{\boldsymbol{\sigma}}^i-\boldsymbol{\sigma}^i)\|+\|\operatorname{div}\delta_t(\tilde{\boldsymbol{\sigma}}^i-\boldsymbol{\sigma}^i)\|\le Ch_{\boldsymbol{\sigma}}^{k_1}\left(\frac{1}{\Delta t}\int_{t^{i-1}}^{t^i}\|\boldsymbol{\sigma}_t\|_{k_1+1}\mathrm{d}t\right)^{\frac12}.\tag{3.31}$$

Theorem 3.4. *Suppose* $(u_h^n, \boldsymbol{\sigma}_h^n) \in V_h \times \mathbf{W}_h$ *is the solution of Scheme (III). Under the assumption* $\|\boldsymbol{\sigma}_h^0 - \boldsymbol{\sigma}^0\| = O(h_{\boldsymbol{\sigma}}^{k+1}\|\boldsymbol{\sigma}^0\|_{H^{k+1}})$, *then*

$$\|\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}\| + \left[\sum_{i=1}^n \Delta t \|\operatorname{div}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i + \boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})\|^2\right]^{\frac{1}{2}} \le C\left[h_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}\|_{L^\infty(H^{k_1+1})} + h_{\boldsymbol{\sigma}}^{k+1}\|\boldsymbol{\sigma}_t\|_{L^\infty(H^{k+1})}\right] + \Delta t^2\left[\|u_t^{(3)}\|_{L^2(L^2)} + \|\boldsymbol{\sigma}_{tt}\|_{L^2(H^1)}\right]. \tag{3.32}$$

Moreover, if $\|\operatorname{div}(\boldsymbol{\sigma}_h^0 - \boldsymbol{\sigma}^0)\| = O(h_{\boldsymbol{\sigma}}^{k_1}\|\operatorname{div}\boldsymbol{\sigma}^0\|_{H^{k_1}})$ *and* $u_h^0 = \tilde{u}_h^0$ *we have*

$$\|u_h^n - u^n\| \le Ch_u^{m+1}\left[\|u\|_{L^\infty(H^{m+1})} + \|u_t\|_{L^2(H^{m+1})}\right] + \Delta t^2\left[\|u_t^{(4)}\|_{L^2(L^2)} + \|\boldsymbol{\sigma}_t^{(3)}\|_{L^2(H^1)} + \|u_t^{(3)}\|_{L^\infty(L^2)} + \|\boldsymbol{\sigma}_{tt}\|_{L^\infty(H^1)}\right] + h_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}\|_{L^\infty(H^{k_1+1})} + h_{\boldsymbol{\sigma}}^{k+1}\|\boldsymbol{\sigma}_t\|_{L^2(H^{k_1+1})} + \|\boldsymbol{\sigma}\|_{L^\infty(H^{k+1})}. \tag{3.33}$$

Here C denotes a positive constant C independent of h_u, h_σ *and* Δt.

Proof. First note that subtracting (2.38) from (2.39) leads to

$$\begin{aligned} b(u_h^i - u^i, \boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i; v_h, \boldsymbol{\tau}_h) = &\left(u_h^{i-1} - u^{i-1} - \frac{\Delta t}{2}\phi^{-1}(\operatorname{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})), \phi v_h + \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\tau}_h\right) \\ &+ \frac{\Delta t}{2}\left(\mathcal{A}^{-1}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}) + \nabla(u_h^{i-1} - u^{i-1}), \boldsymbol{\tau}_h + \mathcal{A}\nabla v_h\right), \\ &- \Delta t\left(\phi^{-1}\rho_2^i, \phi v_h + \frac{\Delta t}{2}\operatorname{div}\boldsymbol{\tau}_h\right), \quad \forall (v_h, \boldsymbol{\tau}_h) \in V_h \times \mathbf{W}_h. \end{aligned} \tag{3.34}$$

Using Lemma 2.2 and (3.34) with a chosen $v_h = 0$, it follows that

$$\begin{aligned} &\frac{\Delta t}{2}(\mathcal{A}^{-1}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i), \boldsymbol{\tau}_h) + \left(\frac{\Delta t}{2}\right)^2(\phi^{-1}\operatorname{div}(\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i), \operatorname{div}\boldsymbol{\tau}_h) \\ &= \frac{\Delta t}{2}(u_h^{i-1} - u^{i-1}, \operatorname{div}\boldsymbol{\tau}_h) - \left(\frac{\Delta t}{2}\right)^2(\phi^{-1}\operatorname{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}), \operatorname{div}\boldsymbol{\tau}_h) \\ &\quad + \frac{\Delta t}{2}(\mathcal{A}^{-1}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}), \boldsymbol{\tau}_h) + \frac{\Delta t}{2}(\nabla(u_h^{i-1} - u^{i-1}), \boldsymbol{\tau}_h), \\ &\quad - \frac{(\Delta t)^2}{2}(\phi^{-1}\rho_2^i, \operatorname{div}\boldsymbol{\tau}_h), \quad \forall (v_h, \boldsymbol{\tau}_h) \in V_h \times \mathbf{W}_h. \end{aligned} \tag{3.35}$$

Let us split into $\boldsymbol{\sigma}_h^i - \boldsymbol{\sigma}^i = \xi_{\sigma}^i - \eta_{\sigma}^i$ where

$$\xi_{\sigma}^i = \boldsymbol{\sigma}_h^i - \tilde{\boldsymbol{\sigma}}_h^i \in \mathbf{W}_h, \qquad \eta_{\sigma}^i = \boldsymbol{\sigma}^i - \tilde{\boldsymbol{\sigma}}_h^i. \tag{3.36}$$

By the definition of $\tilde{\boldsymbol{\sigma}}_h^i$ in (3.27), we have

$$(\phi^{-1}\operatorname{div}(\eta_{\sigma}^i + \eta_{\sigma}^{i-1}), \operatorname{div}\boldsymbol{\tau}_h) = -(\eta_{\sigma}^i + \eta_{\sigma}^{i-1}, \boldsymbol{\tau}_h).$$

It is clear that

$$(u_h^{i-1} - u^{i-1}, \operatorname{div}\boldsymbol{\tau}_h) + (\nabla(u_h^{i-1} - u^{i-1}), \boldsymbol{\tau}_h) = 0.$$

Hence (3.35) reduces to: for all $\boldsymbol{\tau}_h \in \mathbf{W}_h$

$$\begin{aligned} &(\mathcal{A}^{-1}(\xi_{\sigma}^i - \xi_{\sigma}^{i-1}), \boldsymbol{\tau}_h) + \frac{\Delta t}{2}(\phi^{-1}\operatorname{div}(\xi_{\sigma}^i + \xi_{\sigma}^{i-1}), \operatorname{div}\boldsymbol{\tau}_h) \\ &= \Delta t(\mathcal{A}^{-1}\,\delta_t\eta_{\sigma}^i, \boldsymbol{\tau}_h) - \frac{\Delta t}{2}(\eta_{\sigma}^i + \eta_{\sigma}^{i-1}, \boldsymbol{\tau}_h) - \Delta t(\phi^{-1}\rho_2^i, \operatorname{div}\boldsymbol{\tau}_h). \end{aligned} \tag{3.37}$$

Letting $\boldsymbol{\tau}_h = \xi_{\sigma}^i + \xi_{\sigma}^{i-1} \in \mathbf{W}_h$ in (3.37) and using the Cauchy inequality, we have

$$\begin{aligned} &\|\mathcal{A}^{-\frac{1}{2}}\xi_{\sigma}^i\|^2 - \|\mathcal{A}^{-\frac{1}{2}}\xi_{\sigma}^{i-1}\|^2 + \frac{\Delta t}{2}\|\phi^{-\frac{1}{2}}\operatorname{div}(\xi_{\sigma}^i + \xi_{\sigma}^{i-1})\|^2 \le \Delta t\|\xi_{\sigma}^i + \xi_{\sigma}^{i-1}\|^2 + \frac{\Delta t}{4}\|\phi^{-\frac{1}{2}}\operatorname{div}(\xi_{\sigma}^i + \xi_{\sigma}^{i-1})\|^2 \\ &\quad + \Delta t\left[\frac{1}{2}\|\mathcal{A}^{-1}\,\delta_t\eta_{\sigma}^i\|^2 + \frac{1}{8}\|\eta_{\sigma}^i + \eta_{\sigma}^{i-1}\|^2 + \|\phi^{-\frac{1}{2}}\rho_2^i\|^2\right]. \end{aligned} \tag{3.38}$$

Summing (3.38) for $i = 1, 2, \ldots, n$ we can deduce that

$$\|\mathcal{A}^{-\frac{1}{2}}\xi_\sigma^n\|^2 + \frac{1}{2}\sum_{i=1}^n \Delta t\|\phi^{-\frac{1}{2}}\,\mathrm{div}(\xi_\sigma^i + \xi_\sigma^{i-1})\|^2$$
$$\leq C\sum_{i=1}^n \Delta t\|\xi_\sigma^i\|^2 + C\|\xi_\sigma^0\|^2 + C\sum_{i=1}^n \Delta t[\|\delta_t\eta_\sigma^i\|^2 + \|\eta_\sigma^i\|^2 + \|\rho_2^i\|^2] + C\Delta t\|\eta_\sigma^0\|^2. \quad (3.39)$$

Using Gronwall's Lemma we can get that,

$$\|\xi_\sigma^n\|^2 + \sum_{i=1}^n \Delta t\|\,\mathrm{div}(\xi_\sigma^i + \xi_\sigma^{i-1})\|^2 \leq Ch_\sigma^{2(k+1)}(\|\boldsymbol{\sigma}\|^2_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_t\|^2_{L^2(H^{k+1})}) + C\Delta t^4(\|u_t^{(3)}\|^2_{L^2(L^2)} + \|\mathrm{div}\,\boldsymbol{\sigma}_{tt}\|^2_{L^2(L^2)}). \quad (3.40)$$

Combining with (3.4) completes the proof of (3.32).

Choosing $\boldsymbol{\tau}_h = \frac{1}{\Delta t}(\xi_\sigma^i - \xi_\sigma^{i-1}) = \delta_t\xi_\sigma^i$ in (3.37) we have that

$$\Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|^2 + \frac{1}{2}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^i\|^2 - \frac{1}{2}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^{i-1}\|^2$$
$$= \Delta t(\mathcal{A}^{-1}\delta_t\eta_\sigma^i, \delta_t\xi_\sigma^i) - \frac{\Delta t}{2}(\eta_\sigma^i + \eta_\sigma^{i-1}, \delta_t\xi_\sigma^i) - \Delta t(\phi^{-1}\rho_2^i, \mathrm{div}\,\delta_t\xi_\sigma^i),$$
$$\leq C\Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|(\|\delta_t\eta_\sigma^i\| + \|\eta_\sigma^i + \eta_\sigma^{i-1}\|) - (\phi^{-1}\rho_2^i, \mathrm{div}\,\xi_\sigma^i) + (\phi^{-1}\rho_2^{i-1}, \mathrm{div}\,\xi_\sigma^{i-1}) - \Delta t(\phi^{-1}\delta_t\rho_2^i,\ \mathrm{div}\,\xi_\sigma^{i-1}), \quad (3.41)$$

where we have used the equivalence

$$\Delta t(\phi^{-1}\rho_2^i,\ \mathrm{div}\,\delta_t\xi_\sigma^i) = (\phi^{-1}\rho_2^i, \mathrm{div}\,\xi_\sigma^i) - (\phi^{-1}\rho_2^{i-1}, \mathrm{div}\,\xi_\sigma^{i-1}) - \Delta t(\phi^{-1}\delta_t\rho_2^i, \mathrm{div}\,\xi_\sigma^{i-1}).$$

For convenience we introduce a notation ρ_2^0 and $\delta_t\rho_2^1 = \frac{\rho_2^1 - \rho_2^0}{\Delta t}$. Making summation over $i = 1, 2, \ldots, n$ and using the Cauchy inequality result in

$$\sum_{i=1}^n \Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|^2 + \frac{1}{2}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^n\|^2 \leq C\sum_{i=1}^n \Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|^2(\|\delta_t\eta_\sigma^i\| + \|\eta_\sigma^i + \eta_\sigma^{i-1}\|) + \frac{1}{2}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^0\|^2$$
$$-(\phi^{-1}\rho_2^n, \mathrm{div}\,\xi_\sigma^n) + (\phi^{-1}\rho_2^0,\ \mathrm{div}\,\xi_\sigma^0) + \sum_{i=2}^n \Delta t(\phi^{-1}\delta_t\rho_2^i, \mathrm{div}\,\xi_\sigma^{i-1}) + \Delta t(\phi^{-1}\delta_t\rho_2^1, \mathrm{div}\,\xi_\sigma^0). \quad (3.42)$$

Since

$$(\phi^{-1}\rho_2^0,\ \mathrm{div}\,\xi_\sigma^0) + \Delta t(\phi^{-1}\delta_t\rho_2^1,\ \mathrm{div}\,\xi_\sigma^0) = (\phi^{-1}\rho_2^1,\ \mathrm{div}\,\xi_\sigma^0),$$

using the ϵ-inequality we have that

$$\sum_{i=1}^n \Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|^2 + \frac{1}{2}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^n\|^2 \leq \frac{1}{2}\sum_{i=1}^n \Delta t\|\mathcal{A}^{-\frac{1}{2}}\delta_t\xi_\sigma^i\|^2 + \frac{1}{4}\|\phi^{-\frac{1}{2}}\,\mathrm{div}\,\xi_\sigma^n\|^2 + C\|\mathrm{div}\,\xi_\sigma^0\|^2$$
$$+C\sum_{i=1}^n \Delta t(\|\delta_t\eta_\sigma^i\|^2 + \|\eta_\sigma^i\|^2) + C(\|\eta_\sigma^0\|^2 + \|\rho_2^n\|^2 + \|\rho_2^1\|^2) + C\sum_{i=2}^n \Delta t\|\delta_t\rho_2^i\|^2 + C\sum_{i=1}^n \Delta t\|\phi^{-\frac{1}{2}}\mathrm{div}\,\xi_\sigma^i\|^2. \quad (3.43)$$

Moving the first two terms of the right-hand side to the left side, then Gronwall's inequality results in

$$\|\,\mathrm{div}\,\xi_\sigma^n\|^2 + \sum_{i=1}^n \Delta t\|\delta_t\xi_\sigma^i\|^2 \leq Ch_\sigma^{2(k+1)}(\|\boldsymbol{\sigma}\|^2_{L^\infty(H^{k+1})} + \|\boldsymbol{\sigma}_t\|^2_{L^2(H^{k+1})}) + Ch_\sigma^{2k_1}\|\,\mathrm{div}\,\boldsymbol{\sigma}^0\|^2_{H^{k_1}}$$
$$+C\Delta t^4(\|u_t^{(4)}\|^2_{L^2(L^2)} + \|\boldsymbol{\sigma}_t^{(3)}\|^2_{L^2(L^2)} + \|u_t^{(3)}\|^2_{L^\infty(L^2)} + \|\boldsymbol{\sigma}_{tt}\|^2_{L^\infty(L^2)}). \quad (3.44)$$

Now we consider the estimate of $u_h^n - u^n$. Choosing $\boldsymbol{\tau}_h = 0$ in (3.34) we have that,

$$(\phi(u_h^i - u^i), v_h) + \frac{\Delta t}{2}(\mathcal{A}\nabla(u_h^i - u^i), \nabla v_h) = (\phi(u_h^{i-1} - u^{i-1}), v) - \Delta t(\rho_2^i, v_h) - \Delta t(\mathrm{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}), v_h)$$
$$+\frac{\Delta t}{2}(\mathcal{A}\nabla(u_h^{i-1} - u^{i-1}), \nabla v_h), \quad \forall v_h \in \mathrm{V}_h. \quad (3.45)$$

Denote by

$$\xi_u^n = u_h^n - \tilde{u}_h^n, \qquad \eta_u^n = u^n - \tilde{u}_h^n. \quad (3.46)$$

From (3.45) we have that

$$(\phi\xi_u^i, v_h) + \frac{\Delta t}{2}(\mathcal{A}\nabla\xi_u^i, \nabla v_h) = (\phi\xi_u^{i-1}, v_h) + \Delta t(\phi\delta_t\eta_u^i, v_h) - \Delta t(\mathrm{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}), v_h)$$
$$- \Delta t(\rho_2^i, v_h) + \frac{\Delta t}{2}(\mathcal{A}\nabla\xi_u^{i-1}, \nabla v_h), \quad \forall v_h \in \mathrm{V}_h. \tag{3.47}$$

Setting $v_h = \xi_u^i$ and using the Cauchy inequality we have that

$$\begin{aligned}\|\phi^{\frac12}\xi_u^i\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2 &\le \frac12\|\phi^{\frac12}\xi_u^i\|^2 + \frac12\|\phi^{\frac12}\xi_u^{i-1}\|^2 + \frac{\Delta t}{6}\|\phi^{\frac12}\xi_u^i\|^2 + C\Delta t\|\delta_t\eta_u^i\|^2\\ &\quad + \frac{\Delta t}{6}\|\phi^{\frac12}\xi_u^i\|^2 + C\Delta t\|\,\mathrm{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})\|^2\\ &\quad + \frac{\Delta t}{6}\|\phi^{\frac12}\xi_u^i\|^2 + C\Delta t\|\rho_2^i\|^2 + \frac{\Delta t}{4}\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2 + \frac{\Delta t}{4}\|\mathcal{A}^{\frac12}\nabla\xi_u^{i-1}\|^2\\ &= \frac{1+\Delta t}{2}\|\phi^{\frac12}\xi_u^i\|^2 + \frac{\Delta t}{4}\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2 + \frac12\|\phi^{\frac12}\xi_u^{i-1}\|^2 + \frac{\Delta t}{4}\|\mathcal{A}^{\frac12}\nabla\xi_u^{i-1}\|^2\\ &\quad + C\Delta t[\|\delta_t\eta_u^i\|^2 + \|\mathrm{div}\,(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})\|^2 + \|\rho_2^i\|^2],\end{aligned} \tag{3.48}$$

then

$$\begin{aligned}\|\phi^{\frac12}\xi_u^i\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^i\|^2 &\le \|\phi^{\frac12}\xi_u^{i-1}\|^2 + \Delta t\|\phi^{\frac12}\xi_u^i\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^{i-1}\|^2\\ &\quad + C\Delta t[\|\delta_t\eta_u^i\|^2 + \|\mathrm{div}\,(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})\|^2 + \|\rho_2^i\|^2].\end{aligned} \tag{3.49}$$

Summing (3.49) over $i = 1, 2, \ldots, n$, we have that

$$\|\phi^{\frac12}\xi_u^n\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^n\|^2 \le \sum_{i=1}^n \Delta t\|\phi^{\frac12}\xi_u^n\|^2 + \|\phi^{\frac12}\xi_u^0\|^2 + \frac{\Delta t}{2}\|\mathcal{A}^{\frac12}\nabla\xi_u^0\|^2$$
$$+ Ch_u^{2(m+1)}\|u_t\|^2_{L^2(H^{m+1})} + \Delta t^2(\|u_t^{(3)}\|^2_{L^2(L^2)} + \|\,\mathrm{div}\,\boldsymbol{\sigma}_{tt}\|^2_{L^2(L^2)}) + C\sum_{i=1}^n \Delta t\|\,\mathrm{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1})\|^2. \tag{3.50}$$

Since

$$\mathrm{div}(\boldsymbol{\sigma}_h^{i-1} - \boldsymbol{\sigma}^{i-1}) = \mathrm{div}(\xi_{\sigma}^{i-1}) + \mathrm{div}(\tilde{\boldsymbol{\sigma}}_h^{i-1} - \boldsymbol{\sigma}^{i-1}),$$

noticing (3.44) and (3.31), by Gronwall's inequality shows that

$$\begin{aligned}\|u_h^n - \tilde{u}_h^n\| + \Delta t^{\frac12}\|\nabla\xi_u^n\| &\le Ch_u^{m+1}\|u_t\|_{L^2(H^{m+1})} + C\Delta t^2(\|u_t^{(4)}\|_{L^2(L^2)} + \|\boldsymbol{\sigma}_t^{(3)}\|_{L^2(H^1)})\\ &\quad + C\Delta t^2(\|u_t^{(3)}\|_{L^\infty(L^2)} + \|\boldsymbol{\sigma}_{tt}\|_{L^\infty(H^1)}) + Ch_{\boldsymbol{\sigma}}^{k_1}\|\boldsymbol{\sigma}\|_{L^\infty(H^{k_1+1})} + Ch_{\boldsymbol{\sigma}}^{k+1}(\|\boldsymbol{\sigma}_t\|_{L^2(H^{k+1})} + \|\boldsymbol{\sigma}\|_{L^\infty(H^{k+1})}).\end{aligned}$$

Combining with (3.11) completes the proof. □

4. Least-squares procedure for nonlinear problems

In this section we give a least-squares finite element procedure for nonlinear parabolic problems. We consider the following problem on a bounded domain $\Omega \subset \mathcal{R}^d$:

$$\begin{cases}\phi(u)u_t - \mathrm{div}(\mathcal{A}(u)\nabla u) = f(u), & \text{in } \Omega \times J,\\ u = 0, & \text{on } \Gamma_D \times J,\\ \mathcal{A}(u)\nabla u \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \times J,\end{cases} \tag{4.1}$$

subject to the initial condition

$$u(x, 0) = u_0(x) \quad \text{on } \Omega \times J. \tag{4.2}$$

The coefficient $\phi(u)$ is a strictly positive function and the coefficient matrix $\mathcal{A}(u) = (a_{ij}(u))_{i,j=1}^d$ is a bounded, symmetric and positive definite matrix, i.e., there exist two positive constants ϕ_1 and ϕ_2 and two positive constants α and β such that, for $u \in \mathcal{R}^1$

$$\phi_1 \le \phi(u) \le \phi_2, \qquad \alpha\|\boldsymbol{\xi}\|^2 \le (\mathcal{A}(u)\boldsymbol{\xi}, \boldsymbol{\xi}) \le \beta\|\boldsymbol{\xi}\|^2, \quad \forall \boldsymbol{\xi} \in \mathcal{R}^d. \tag{4.3}$$

In general the coefficients $\phi(u)$, $\mathcal{A}(u)$ and $f(u)$ are also dependent on time variable t and space variable x. Since our main purpose is to consider the nonlinearity, for convenience we just consider the dependence of the coefficients on u.

Introducing $\boldsymbol{\sigma} = -\mathcal{A}(u)\nabla u$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_d)$, the nonlinear problem (4.1) appears as a first-order system for both u and as follows:

$$\begin{cases} \phi(u)u_t + \operatorname{div}\boldsymbol{\sigma} - f = 0, & \text{in } \Omega \times J, \\ \boldsymbol{\sigma} + \mathcal{A}(u)\nabla u = 0, & \text{in } \Omega \times J, \\ u = 0, & \text{on } \Gamma_D \times J, \\ \boldsymbol{\sigma} \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \times J. \end{cases} \tag{4.4}$$

We approximate the above equation by

$$\begin{cases} \phi(u^{n-1})\delta_t u^n + \operatorname{div}\boldsymbol{\sigma}^n - f(u^{n-1}) - \rho_3^n = 0, & \text{in } \Omega \times J, \\ \boldsymbol{\sigma}^n + \mathcal{A}(u^{n-1})\nabla u^n - \rho_4^n = 0, & \text{in } \Omega \times J, \\ u = 0, & \text{on } \Gamma_D \times J, \\ \boldsymbol{\sigma} \cdot \mathbf{n} = 0 & \text{on } \Gamma_N \times J, \end{cases} \tag{4.5}$$

where the truncation errors ρ_3^n and ρ_4^n are defined as follows

$$\rho_3^n = \phi(u^{n-1})\delta_t u^n - \phi(u^n)u_t^n, \qquad \rho_4^n = (\mathcal{A}(u^{n-1}) - \mathcal{A}(u^n))\nabla u^n.$$

When the solution and the coefficients are sufficiently smooth we have that

$$\rho_3^n = O(\Delta t), \qquad \rho_4^n = O(\Delta t). \tag{4.6}$$

From (4.4) we know that for $n \geq 1$, $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ satisfy that

$$\begin{cases} \phi(u^{n-1})^{-\frac{1}{2}}(\phi(u^{n-1})u^n + \Delta t \operatorname{div}\boldsymbol{\sigma}^n - F_3^n) = 0, & \text{in } \Omega \times J, \\ \mathcal{A}(u^{n-1})^{-\frac{1}{2}}(\boldsymbol{\sigma}^n + \mathcal{A}(u^{n-1})\nabla u^n - \rho_4^n) = 0, & \text{in } \Omega \times J, \end{cases} \tag{4.7}$$

where

$$F_3^n = \phi(u^{n-1})u^{n-1} + \Delta t f(u^{n-1}) + \Delta t \rho_3^n.$$

For $(v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}$, define the least-squares functional $J_4^n(v, \boldsymbol{\tau})$ as follows.

$$J_4^n(v, \boldsymbol{\tau}) = \|\phi(u^{n-1})^{-\frac{1}{2}}(\phi(u^{n-1})v + \Delta t \operatorname{div}\boldsymbol{\tau} - F_3^n)\|^2 + \Delta t \|\tilde{\mathcal{A}}(u^{n-1})^{\frac{1}{2}}(\boldsymbol{\tau} + \mathcal{A}(u^{n-1})\nabla v - \rho_4^n)\|^2. \tag{4.8}$$

The least-squares minimization problem corresponding to (4.7) is: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$J_4^n(u^n, \boldsymbol{\sigma}^n) = \inf_{v \in \mathrm{V}, \boldsymbol{\tau} \in \mathbf{W}} J_4^n(v, \boldsymbol{\tau}). \tag{4.9}$$

Define the bilinear form $a(w; u, \boldsymbol{\sigma}; v, \boldsymbol{\tau})$ as

$$\begin{aligned} a(w; u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) &= \left(\frac{1}{\phi(w)}(\phi(w)u + \Delta t \operatorname{div}\boldsymbol{\sigma}), \phi(w)v + \Delta t \operatorname{div}\boldsymbol{\tau}\right) + \Delta t\left(\mathcal{A}(w)^{-1}(\boldsymbol{\sigma} + \mathcal{A}(w)\nabla u), \boldsymbol{\tau} + \mathcal{A}(w)\nabla v\right) \\ &= \left(u + \frac{\Delta t}{\phi(w)}\operatorname{div}\boldsymbol{\sigma}, \phi(w)v + \Delta t \operatorname{div}\boldsymbol{\tau}\right) + \Delta t\left(\mathcal{A}(w)^{-1}\boldsymbol{\sigma} + \nabla u, \boldsymbol{\tau} + \mathcal{A}(w)\nabla v\right). \end{aligned} \tag{4.10}$$

Noticing the definition of F_3^n, the weak statement of the minimization problem (4.9) becomes: find $(u^n, \boldsymbol{\sigma}^n) \in \mathrm{V} \times \mathbf{W}$ such that

$$\begin{aligned} a(u^{n-1}; u^n, \boldsymbol{\sigma}^n; v, \boldsymbol{\tau}) &= \left(\phi(u^{n-1})u^{n-1} + \Delta t(f(u^{n-1}) + \rho_3^n), v + \frac{\Delta t}{\phi(u^{n-1})}\operatorname{div}\boldsymbol{\tau}\right) \\ &\quad + \Delta t\left(\rho_4^n, \mathcal{A}(u^{n-1})^{-1}\boldsymbol{\tau} + \nabla v\right), \forall (v, \boldsymbol{\tau}) \in \mathrm{V} \times \mathbf{W}. \end{aligned} \tag{4.11}$$

Selecting the initial approximation $u_h^0 \in \mathrm{V}_h$, $\boldsymbol{\sigma}_h^0 \in \mathbf{W}_h$ similarly as before, the least-squares mixed finite element procedure based on (4.11) reads as follows, which was obtained by diminishing the truncation error terms from (4.11).

Scheme (IV). For $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in \mathrm{V}_h \times \mathbf{W}_h$ such that

$$a(u_h^{n-1}; u_h^n, \boldsymbol{\sigma}_h^n; v, \boldsymbol{\tau}) = \left(\phi(u_h^{n-1})u_h^{n-1} + \Delta t f(u_h^{n-1}), v_h + \frac{\Delta t}{\phi(u_h^{n-1})}\operatorname{div}\boldsymbol{\tau}_h\right), \quad \forall\, (v_h, \boldsymbol{\tau}_h) \in \mathrm{V}_h \times \mathbf{W}_h. \tag{4.12}$$

Similarly to Lemma 2.1 we can prove the following lemma.

Lemma 4.1. *For any $u, v \in \mathrm{V}$ and $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{W}$ we have that,*

$$a(w; u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = (\phi(w)u, v) + \Delta t(\mathcal{A}(w)\nabla u, \nabla v) + \Delta t(\mathcal{A}(w)^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + \Delta t^2\left(\frac{1}{\phi(w)}\operatorname{div}\boldsymbol{\sigma}, \operatorname{div}\boldsymbol{\tau}\right). \tag{4.13}$$

Using Lemma 4.1, we have the decoupling equivalent form of Scheme (IV).

Equivalent Form of Scheme (IV). With the initial guess $(u_h^0, \boldsymbol{\sigma}_h^0) \in V_h \times \mathbf{W}_h$, for $n \geq 1$ find $(u_h^n, \boldsymbol{\sigma}_h^n) \in V_h \times \mathbf{W}_h$ such that for all $v_h \in V_h$ and $\boldsymbol{\tau}_h \in \mathbf{W}_h$

$$(\phi u_h^n, v_h) + \Delta t(\mathcal{A}(u_h^{n-1})\nabla u_h^n, \nabla v_h) = (\phi(u_h^{n-1})u_h^{n-1} + \Delta t f(u_h^{n-1}), v_h), \tag{4.14}$$

$$(\tilde{\mathcal{A}}(u_h^{n-1})\boldsymbol{\sigma}_h^n, \boldsymbol{\tau}_h) + \Delta t\left(\frac{1}{\phi(u_h^{n-1})}\operatorname{div}\boldsymbol{\sigma}_h^n, \operatorname{div}\boldsymbol{\tau}_h\right) = (u_h^{n-1}, \operatorname{div}\boldsymbol{\tau}_h) + \Delta t\left(\frac{1}{\phi(u_h^{n-1})}f(u_h^{n-1}), \operatorname{div}\boldsymbol{\tau}_h\right). \tag{4.15}$$

Now we discuss the error estimate for Scheme (IV).

Theorem 4.2. *Suppose the analytical solution $(u, \boldsymbol{\sigma})$ is sufficiently smooth. Suppose also that the coefficients ϕ, $\mathcal{A}$ and f are Lipschitz continuous bounded functions and satisfy (4.3). $(u_h^n, \boldsymbol{\sigma}_h^n) \in V_h \times \mathbf{W}_h$ is the solution of Scheme* (I). *Under the assumption $\|u_h^0 - u^0\| = O(h_u^{m+1-j}\|u^0\|_{H^{m+1-j}})$, $j = 0, 1$, there exists a positive constant C independent of h_u, h_σ and Δt such that*

$$\|u_h^n - u^n\|_s \leq C(h_u^{m+1-s} + C\Delta t), \quad s = 0, 1, \tag{4.16}$$

$$\|\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}\| + \Delta t^{\frac{1}{2}}\|\operatorname{div}(\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}^n)\| \leq C(h_\sigma^{k+1} + \Delta t^{\frac{1}{2}}h_\sigma^{k_1} + \Delta t + \min\{h_u^m, h_u^{m+1}\Delta t^{-\frac{1}{2}}\}). \tag{4.17}$$

Proof. Since Scheme (IV) is equivalent to (4.14) and (4.15), we use the equivalent form of Scheme (IV) in error estimates. Setting $\boldsymbol{\tau} = 0$ in (4.11) and noticing Lemma 4.1 we have that

$$(\phi(u^{n-1})u^n, v) + \Delta t(\mathcal{A}(u^{n-1})\nabla u^n, \nabla v) = \left(\phi(u^{n-1})u^{n-1} + \Delta t(f(u^{n-1}) + \rho_3^n), v\right) + \Delta t\left(\rho_4^n, \nabla v\right), \quad \forall v \in V. \tag{4.18}$$

Comparing (4.14) and (4.18), from the error estimates of the finite element method for nonlinear parabolic problems (see [18] or [19], for example), we know that (4.16) holds.

Now we consider the error estimates for $\boldsymbol{\sigma}_h^n$. Subtracting (4.11) from (4.12) and letting $v_h = 0$ we have that,

$$\begin{aligned} & a(u_h^{n-1}; u_h^n, \boldsymbol{\sigma}_h{}^n; 0, \boldsymbol{\tau}_h) - a(u^{n-1}; u^n, \boldsymbol{\sigma}^n; 0, \boldsymbol{\tau}_h) \\ &= \left(u_h^{n-1} - u^{n-1} + \frac{\Delta t}{\phi(u_h^{n-1})}f(u_h^{n-1}) - \frac{\Delta t}{\phi(u^{n-1})}f(u^{n-1}) - \frac{\Delta t}{\phi(u^{n-1})}\rho_3^n, \Delta t\operatorname{div}\boldsymbol{\tau}_h\right), \quad \forall\, \boldsymbol{\tau}_h \in \mathbf{W}_h. \end{aligned} \tag{4.19}$$

Using Lemma 2.1, we have

$$\begin{aligned} & (\tilde{\mathcal{A}}(u_h^{n-1})\boldsymbol{\sigma}_h^n - \mathcal{A}(u^{n-1})^{-}\boldsymbol{\sigma}^n, \boldsymbol{\tau}_h) + \Delta t\left(\frac{1}{\phi(u_h^{n-1})}\operatorname{div}\boldsymbol{\sigma}_h^n - \frac{1}{\phi(u^{n-1})}\operatorname{div}\boldsymbol{\sigma}^n, \operatorname{div}\boldsymbol{\tau}_h\right) \\ &= (u_h^{n-1} - u^{n-1}, \operatorname{div}\boldsymbol{\tau}_h) - \Delta t\left(\frac{1}{\phi(u_h^{n-1})}f(u_h^{n-1}) - \frac{1}{\phi(u^{n-1})}f(u^{n-1}) - \frac{1}{\phi(u^{n-1})}\rho_3^n, \operatorname{div}\boldsymbol{\tau}_h\right) \\ &= -(\nabla(u_h^{n-1} - u^{n-1}), \boldsymbol{\tau}_h) - \Delta t\left(\frac{f(u_h^{n-1})}{\phi(u_h^{n-1})} - \frac{f(u^{n-1})}{\phi(u^{n-1})} - \frac{1}{\phi(u^{n-1})}\rho_3^n, \operatorname{div}\boldsymbol{\tau}_h\right), \quad \forall \boldsymbol{\tau}_h \in \mathbf{W}_h, \end{aligned} \tag{4.20}$$

or equivalently,

$$\begin{aligned} & (\tilde{\mathcal{A}}(u_h^{n-1})(\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}^n), \boldsymbol{\tau}_h) + \Delta t\left(\frac{1}{\phi(u_h^{n-1})}\operatorname{div}(\boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}^n), \operatorname{div}\boldsymbol{\tau}_h\right) \\ &= ((\tilde{\mathcal{A}}(u_h^{n-1}) - \mathcal{A}(u^{n-1})^{-})\boldsymbol{\sigma}^n, \boldsymbol{\tau}_h) + \Delta t\left(\left(\frac{1}{\phi(u_h^{n-1})} - \frac{1}{\phi(u^{n-1})}\right)\operatorname{div}\boldsymbol{\sigma}^n, \operatorname{div}\boldsymbol{\tau}_h\right) \\ &\quad - (\nabla(u_h^{n-1} - u^{n-1}), \boldsymbol{\tau}_h) - \Delta t\left(\frac{f(u_h^{n-1})}{\phi(u_h^{n-1})} - \frac{f(u^{n-1})}{\phi(u^{n-1})} - \frac{1}{\phi(u^{n-1})}\rho_3^n, \operatorname{div}\boldsymbol{\tau}_h\right), \quad \forall \boldsymbol{\tau}_h \in \mathbf{W}_h. \end{aligned} \tag{4.21}$$

Let $\boldsymbol{\sigma}_I^n \in \mathbf{W}_h$ be the same interpolant of $\boldsymbol{\sigma}^n$ satisfying (3.4). With a chosen $\boldsymbol{\tau}_h = \boldsymbol{\sigma}_h^n - \boldsymbol{\sigma}_I^n$ in (4.21), and using the notation ξ_σ^i defined in (3.5) we have that

$$\begin{aligned} C_0(\|\xi_\sigma^n\|^2 + \Delta t\|\operatorname{div}\xi_\sigma^n\|^2) &\leq (\tilde{\mathcal{A}}(u_h^{n-1})\xi_\sigma^n, \xi_\sigma^n) + \Delta t\left(\frac{1}{\phi(u_h^{n-1})}\operatorname{div}\xi_\sigma^n, \operatorname{div}\xi_\sigma^n\right) \\ &= \left((\tilde{\mathcal{A}}(u_h^{n-1}) - \mathcal{A}(u^{n-1})^{-})\boldsymbol{\sigma}^n, \xi_\sigma^n\right) + \Delta t\left(\left(\frac{1}{\phi(u_h^{n-1})} - \frac{1}{\phi(u^{n-1})}\right)\operatorname{div}\boldsymbol{\sigma}^n, \operatorname{div}\xi_\sigma^n\right) \\ &\quad - (\nabla(u_h^{n-1} - u^{n-1}), \xi_\sigma^n) - \Delta t\left(\frac{f(u_h^{n-1})}{\phi(u_h^{n-1})} - \frac{f(u^{n-1})}{\phi(u^{n-1})} - \frac{1}{\phi(u^{n-1})}\rho_3^n, \operatorname{div}\xi_\sigma^n\right) \\ &\quad - (\tilde{\mathcal{A}}(u_h^{n-1})(\boldsymbol{\sigma}_I^n - \boldsymbol{\sigma}^n), \xi_\sigma^n) - \Delta t\left(\frac{1}{\phi(u_h^{n-1})}\operatorname{div}(\boldsymbol{\sigma}_I^n - \boldsymbol{\sigma}^n), \operatorname{div}\xi_\sigma^n\right). \end{aligned} \tag{4.22}$$

Table 4.1
Result of the first example

N, N_t	$D=10$		$D=1$		$D=0.1$	
	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}
5	0.319	0.267	3.33E−2	2.79E−2	2.36E−3	1.97E−3
10	7.93E−2	6.22E−2	9.47E−3	7.42E−3	1.50E−3	1.17E−3
20	2.06E−2	1.53E−2	3.02E−3	2.25E−3	8.46E−4	6.29E−4
40	5.45E−3	3.95E−3	1.07E−3	7.79E−4	4.47E−4	3.24E−4

Since ϕ, $\frac{1}{\phi}$, $\mathcal{A}$ and f are uniformly bounded and Lipschitz continuous, using the ϵ-inequality, (4.6) and (4.22), we have

$$\begin{aligned} C_0(\|\xi_\sigma^n\|^2+\Delta t\|\operatorname{div}\xi_\sigma^n\|^2) &\le C\|u_h^{n-1}-u^{n-1}\|\|\xi_\sigma^n\|+C\Delta t\|u_h^{n-1}-u^{n-1}\|\|\operatorname{div}\xi_\sigma^n\| \\ &\quad +C\|\nabla(u_h^{n-1}-u^{n-1})\|\|\xi_\sigma^n\|+C\Delta t(\|u_h^{n-1}-u^{n-1}\|+\|\rho_3^n\|)\|\operatorname{div}\xi_\sigma^n\| \\ &\quad +C\|\boldsymbol{\sigma}_I^n-\boldsymbol{\sigma}^n\|\|\xi_\sigma^n\|+C\Delta t\|\operatorname{div}(\boldsymbol{\sigma}_I^n-\boldsymbol{\sigma}^n)\|\|\operatorname{div}\xi_\sigma^n\|. \\ &\le \frac{C_0}{2}(\|\boldsymbol{\sigma}_h^n-\boldsymbol{\sigma}_I^n\|^2+\Delta t\|\operatorname{div}(\boldsymbol{\sigma}_h^n-\boldsymbol{\sigma}_I^n)\|^2) \\ &\quad +C(1+\Delta t)\|u_h^{n-1}-u^{n-1}\|^2+C\|\nabla(u_h^{n-1}-u^{n-1})\|^2+C\Delta t\|\rho_3^n\|^2 \\ &\quad +C\|\boldsymbol{\sigma}_I^n-\boldsymbol{\sigma}^n\|^2+C\Delta t\|\operatorname{div}(\boldsymbol{\sigma}_I^n-\boldsymbol{\sigma}^n)\|^2. \end{aligned} \tag{4.23}$$

Then, using (4.16) with $s=0$ and $s=1$, we have

$$\begin{aligned} &\|\xi_\sigma{}^n\|+\Delta t^{\frac{1}{2}}\|\operatorname{div}\xi_\sigma^n\| \\ &\quad \le C(h_\sigma^{k+1}\|\boldsymbol{\sigma}^n\|_{k+1}+\Delta t^{\frac{1}{2}}h_\sigma^{k_1}\|\boldsymbol{\sigma}^n\|_{k_1+1}+\Delta t\|u_{tt}\|_{L^2(L^2)})+Ch_u^m(\|u\|_{L^\infty(H^{m+1})}+\|u_t\|_{L^2(H^{m+1})}). \end{aligned} \tag{4.24}$$

Finally, combining (4.24) with (3.4) completes the proof. □

5. Numerical examples

In real implementation we can select the sub-procedure (2.27) to solve u_h and the sub-procedure (2.30) to solve $\boldsymbol{\sigma}_h$. (2.27) is the usual Galerkin finite element procedure, so it is sufficient to give some numerical examples to examine the sub-procedure (2.30) for $\boldsymbol{\sigma}_h$.

We consider the following problem

$$u_t-\operatorname{div}(\mathbf{D}\nabla u)=f, \quad \text{in } \Omega\times J,$$

with proper Dirichlet boundary condition and initial condition. For simplicity we let D be a constant, Ω be an unit square, $\Omega=(a,b)\times(a,b)$, and the time interval be $(0,T)=(0,0.5)$. The boundary and initial conditions are selected according to the analytical solution.

We divide $(0,T)$ into N_t equal time intervals, $\Delta t=\frac{T}{N_t}$, and divide Ω into $N\times N$ uniform square elements, $h=\frac{1}{N}$. Based on this triangulation we select the lowest order Raviart–Thomas mixed element as the test function space. For $\sigma_h=(\sigma_{h,1},\sigma_{h,2})$, a finite element approximation to $\sigma=(\sigma_1,\sigma_2)$, the set of nodal points for $\sigma_{h,1}$ is denoted by V_1 and the set of nodal points for $\sigma_{h,2}$ is denoted by V_2.

In the first example, the analytical solution is

$$u=\sin(\pi x)\sin(\pi y)\exp(-t), \qquad \boldsymbol{\sigma}=-D\nabla u.$$

For a set of simulations, different mesh sizes and different values of the diffusion coefficient D are taken and their corresponding errors are listed in Table 4.1. Here e_{σ,l^∞} and e_{σ,l^2} are defined as

$$e_{\sigma,l^\infty}=\max\{\max_{p\in V_1}|(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_{h,1})(p,T)|,\max_{p\in V_2}|(\boldsymbol{\sigma}_2-\boldsymbol{\sigma}_{h,2})(p,T)|\},$$

$$e_{\sigma,l^2}=\left(\sum_{p\in V_1}|(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_{h,1})(p,T)|^2h^2+\sum_{p\in V_2}|(\boldsymbol{\sigma}_2-\boldsymbol{\sigma}_{h,2})(p,T)|^2h^2\right)^{\frac{1}{2}}.$$

In the second example, the analytical solution is

$$u=x(1-x)y(1-y)\exp(x-y-t), \qquad \boldsymbol{\sigma}=-D\nabla u.$$

For different mesh sizes and different values of the diffusion coefficient D the errors are listed in Table 4.2.

In the third example, the analytical solution is

$$u=x(1-x)y(1-y)\exp(x+y+t), \qquad \boldsymbol{\sigma}=-D\nabla u.$$

For different mesh sizes and different values of the diffusion coefficient D the errors are listed in Table 4.3.

The numerical examples given above are in good agreement with the theoretical analysis, which shows that the scheme is stable and convergent.

Table 4.2
Result of the second example

N, N_t	$D = 10$		$D = 1$		$D = 0.1$	
	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}
5	3.40E−2	3.06E−2	3.11E−3	2.96E−3	4.43E−4	3.02E−4
10	9.62E−3	7.25E−3	9.86E−4	6.88E−4	1.46E−4	9.81E−5
20	3.05E−3	1.75E−3	3.12E−4	1.65E−4	7.51E−5	4.68E−5
40	8.90E−4	4.28E−4	9.10E−5	4.37E−5	3.87E−5	2.32E−5

Table 4.3
Result of the third example

N, N_t	$D = 10$		$D = 1$		$D = 0.1$	
	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}	e_{σ,l^∞}	e_{σ,l^2}
5	0.222	0.157	2.05E−2	1.44E−2	1.49E−3	1.43E−3
10	5.73E−2	3.64E−2	5.37E−3	3.22E−3	7.10E−4	5.06E−4
20	1.56E−2	8.65E−3	1.48E−3	7.43E−4	3.89E−4	2.29E−4
40	4.23E−3	2.07E−3	4.06E−4	2.10E−4	2.03E−4	1.13E−4

Acknowledgements

The authors would like to express their sincere thanks to the anonymous referees for their helpful comments and suggestions, which greatly improved the quality of this paper.

References

[1] I. Babuska, The finite element with Lagrange multipliers, Numer. Math. 20 (1973) 179–192.
[2] P. Bochev, M. Gunzburger, On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles, SIAM J. Numer. Anal. 43 (2005) 340–362.
[3] P. Bochev, M. Gunzburger, A locally conservative least-squares methods for Darcy flows, Comm. Numer. Methods Eng. 24 (2008) 97–110.
[4] I. Brezzi, On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers, RAIRO Anal. Numer. 8 (1974) 129–151.
[5] Z. Cai, R. Lazarov, T.A. Manteuffel, S.F. McCormick, First-order system least squares for second-order partial differential equations I, SIAM J. Numer. Anal. 31 (6) (1994) 1785–1799.
[6] Z. Cai, T.A. Manteuffel, S.F. McCormick, First-order system least squares for second-order partial differential equations. II, SIAM J. Numer. Anal. 34 (2) (1997) 425–454.
[7] P. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.
[8] Y. Cao, M. Gunzberger, Least-squares finite element approximations to solutions of interface problems 35 (1998) 393–405.
[9] Jim Douglas Jr., J.E. Roberts, Numerical methods for a model for compressible miscible displacement in porous media, Math. Comp. 40 (1983) 441–459.
[10] G. Fix, M. Gunzburger, On least squares approximation to indefinite problems of the mixed type, Internat. J. Numer. Meth. Engrg. (1978) 453–469.
[11] G. Fix, M. Gunzburger, R. Nicolaides, On mixed finite element methods for first-order elliptic systems, Numer. Math. 37 (1981) 29–48.
[12] D. Jesperson, A least-squares decomposition methods for solving elliptic equations, Math. Comp. 31 (1977) 873–880.
[13] O.A. Ladyzhenskaya, The Mathematical Theory of Viscous Incompressible Flows, Gordon and Brench, London, 1969.
[14] M. Majidi, G. Starke, Least-square Galerkin methods for parabolic problems: Semidiscretization in time, SIAM J. Numer. Anal. 39 (2001) 1302–1323.
[15] M. Majidi, G. Starke, Least-square Galerkin methods for parabolic problems: The fully discretization case and adaptive algorithms, SIAM J. Numer. Anal. 39 (2002) 1648–1666.
[16] A. Pehlivanov, G.F. Carey, D. Lazarov, Least-squares mixed finite elements for second-order elliptic problems, SIAM J. Numer. Anal. 31 (5) (1994) 1368–1377.
[17] P.A. Raviart, J.M. Thomas, A mixed finite element method for second order elliptic problems, in: Lecture Notes in Mathematics, vol. 606, Springer-Verlag, 1977.
[18] V. Thomee, Galerkin Finite Element Methods for Parabolic Problems, Springer-Verlag, Berlin, 1984.
[19] M.F. Wheeler, A priori L^2 error estimates for Galerkin approximations to parabolic partial differential equations, SIAM J. Numer. Anal. 10 (1973) 723–759.