Mathematical Models and Methods in Applied Sciences
Vol. 17, No. 8 (2007) 1279–1305

World Scientific
www.worldscientific.com

IMPROVED ACCURACY FOR ALTERNATING-DIRECTION METHODS FOR PARABOLIC EQUATIONS BASED ON REGULAR AND MIXED FINITE ELEMENTS

TODD ARBOGAST

Department of Mathematics, The University of Texas at Austin
1 University Station C1200, Austin, TX 78712, USA
arbogast@ices.utexas.edu

CHIEH-SEN HUANG*

Department of Applied Mathematics and National Center for Theoretical Sciences
National Sun Yat-sen University, Kaohsiung 804, Taiwan, R.O.C.
huangcs@math.nsysu.edu.tw

SONG-MING YANG

Department of Applied Mathematics, National Sun Yat-sen University
Kaohsiung 804, Taiwan, R.O.C.
minjey@pchome.com.tw

Received 7 December 2005
Revised 17 November 2006
Communicated by F. Brezzi

An efficient modification by Douglas and Kim of the usual alternating directions method reduces the splitting error from $\mathcal{O}(k^2)$ to $\mathcal{O}(k^3)$ in time step k. We prove convergence of this modified alternating directions procedure, for the usual non-mixed Galerkin finite element and finite difference cases, under the restriction that k/h^2 is sufficiently small, where h is the grid spacing. This improves the results of Douglas and Gunn, who require k/h^4 to be sufficiently small, and Douglas and Kim, who require that the locally one-dimensional operators commute. We propose a similar and efficient modification of alternating directions for mixed finite element methods that reduces the splitting error to $\mathcal{O}(k^3)$, and we prove convergence in the noncommuting case, provided that k/h^2 is sufficiently small. Numerical computations illustrating the mixed finite element results are also presented. They show that our proposed modification can lead to a significant reduction in the alternating direction splitting error.

Keywords: Alternating directions method; splitting error.

AMS Subject Classification: 65M12, 65M60, 65N12, 80M10

*Corresponding author

1. Introduction

We consider in this paper the approximation of a parabolic problem on a bounded domain $\Omega \subset \mathbb{R}^d$ of the form

$$cu_t - \nabla\cdot(a\nabla u) = f, \quad x \in \Omega, \quad 0 < t \le T, \tag{1.1}$$

$$u = g, \quad x \in \partial\Omega, \quad 0 < t \le T, \tag{1.2}$$

$$u = u_0, \quad x \in \Omega, \quad t = 0, \tag{1.3}$$

where u_t is the time partial derivative of $u(x,t)$, $c(x)$ and $a(x)$ are uniformly positive on $\bar{\Omega}$, and $c(x)$, $a(x)$, $f(x,t)$, $g(x,t)$ and $u_0(x)$ are sufficiently smooth for our purposes. Since the 50s, scientists have formulated time-stepping procedures to numerically approximate the solutions of such problems. The Alternating Direction (AD) methods were first introduced in 1955 by Douglas, Peaceman and Rachford.[4,8,13] They noted that

$$-\nabla\cdot(a\nabla) = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(a\frac{\partial}{\partial x_i}\right)$$

is a sum of d one-dimensional operators and, thereby, treated the spatial variables of (1.1) individually in a cyclic fashion. This locally one-dimensional approach produces a *splitting error* over an approach that treats the full d-dimensional problem at once.

An AD method can be interpreted as being a perturbation of some underlying implicit numerical time-stepping method, such as Crank–Nicolson or backward Euler. The spatial variable splitting error terms form a perturbation of the same order in the time step k as the truncation error terms associated with the Crank–Nicolson method, $\mathcal{O}(k^2)$, and of higher order with the backward Euler method, $\mathcal{O}(k)$. Thus, the asymptotic rate of convergence for the AD method should be of the same order in the spatial and temporal discretization parameters as that for its associated underlying method. However, at practical levels of discretization, the actual errors associated with an AD method can be much larger than that of the underlying method. To rectify this, Douglas and Kim[7] (cf. Ref. 5) proposed the Modified Alternating Direction iteration algorithms (AD-M), sometimes referred to as AD-II, AD with Improved Initialization (see (2.11)–(2.13) below). They modify the right-hand side of an AD algorithm to reduce the order of the splitting error from $\mathcal{O}(k^2)$ to $\mathcal{O}(k^3)$.

This paper has two main results. In Ref. 7, Douglas and Kim give a convergence proof for AD-M under the assumption that the order in which the individual spatial variables are treated is immaterial. More precisely, if A_m is the discrete approximation to the one-dimensional operator $-\partial/\partial x_m(a\partial/\partial x_m)$, they require that A_{m_1} and A_{m_2} commute for all m_1 and m_2. This condition generally does not hold in practice, for example, when a is not constant. Some four decades ago, Douglas and Gunn[6] provided a general formulation and proof of convergence of the AD method in the non-commutative case. However, they required the restriction that k/h^4 must be

sufficiently small, where h is the grid spacing. Our first main result is a convergence proof for the AD-M under the constraint that merely k/h^2 is sufficiently small.

Our second main result is an improved treatment of mixed finite element methods.[3,14] These methods approximate the flux variable $\mathbf{q} = -a\nabla u$ simultaneously with the scalar variable u, and they are often employed to solve for flow fields in physics and engineering problems. Douglas and Pietra[9] formulated an AD iterative technique for solving the algebraic systems associated with mixed finite element methods for second order elliptic problems. We propose a modification similar to the AD-M method for the non-mixed formulation. Our modification reduces the splitting error from $\mathcal{O}(k^2)$ to $\mathcal{O}(k^3)$. A similar convergence theory to that our first main result is developed for this mixed AD-M method. That is, we obtain convergence for the noncommuting case with the restriction that k/h^2 is sufficiently small.

The rest of the paper is organized as follows. In Sec. 2, we define the non-mixed AD and AD-M methods for the usual Galerkin finite element or finite difference discretizations of parabolic equations. We present our convergence proof for AD-M in Sec. 3. After a review of the standard mixed finite element method, given in Sec. 4, the AD-M for mixed finite element methods is formulated in Sec. 5. Our convergence results for mixed methods are given in Sec. 6. Finally, in Sec. 7, we present some numerical experiments illustrating the utility of our mixed methods, and confirming our theoretical results.

2. The Basic Methods in Non-Mixed Form

Although our results do not require the following constraint, for simplicity we tacitly suppose that $\Omega \subset \mathbb{R}^d$ admits a rectangular spatial grid of maximal spacing h. This is the usual situation considered since AD can be implemented efficiently in this case. Also for simplicity, we take a uniform time step $k = T/N > 0$. We define discrete times $t^n = nk$ and use the notation φ^n in place of $\varphi(t^n)$ and $\varphi^{n+1/2}$ in place of $\varphi((t^n + t^{n+1})/2)$.

Loosely speaking, we let A be the d-dimensional linear operator obtained from finite difference or finite element approximation of $-\nabla\cdot(a\nabla)$ over the grid on Ω with order of accuracy $\mathcal{O}(h^s)$. We assume that

$$A = \sum_{m=1}^{d} A_m,$$

where each A_m can be inverted relatively easily (cf. Ref. 5). Normally, A_m is a one-dimensional linear operator obtained from approximation of $-\partial_m(a\partial_m)$, where $\partial_m = \partial/\partial x_m$, on an x_m-line of Omega over the grid. However, we need to be more precise about exactly what we mean by the operator A_m, especially for the next section where we need to apply it multiple times to the true solution. It is confusing to explain the finite difference and finite element cases together, so we present only the more difficult finite element case and leave it to the reader to translate things to the finite difference case.

We begin by rewriting our differential system (1.1)–(1.2) in variational form. Let $(\cdot,\cdot)$ denote the inner-product in $L^2(\Omega)$. We find $u \in H_0^1(\Omega) + g$, where g is extended to all of Ω, such that

$$(cu_t, v) + (a\nabla u, \nabla v) = (f, v), \quad v \in H_0^1(\Omega). \tag{2.1}$$

Let $V_h \subset H_0^1(\Omega)$ denote our finite element space, with nodal basis

$$V_h = \operatorname*{span}_i \{v_i\}.$$

We will approximate u in $V_h + g$. Now A and A_m are symmetric, positive semidefinite matrices, with ij-entries

$$A_{ij} = (a\nabla v_i, \nabla v_j) \quad \text{and} \quad A_{m,ij} = (a\partial_m v_i, \partial_m v_j).$$

We also need the positive definite matrix

$$C_{ij} = (cv_i, v_j)$$

(which is diagonal if mass lumping is used, and so trivially inverted) and the vector $F^{n+\theta}$ given by

$$F_i^{n+\theta} = (f, v_i) - \left(c\frac{g^{n+1} - g^n}{k}, v_i\right) - \big(a\nabla[\theta g^{n+1} + (1-\theta)g^n], \nabla v_i\big),$$

for θ chosen immediately below.

2.1. *The backward Euler and Crank–Nicolson methods*

The basic time-stepping algorithms for (1.1) can be written together using $\theta = 1$ for backward Euler and $\theta = 1/2$ for Crank–Nicolson.

We approximate u^n by $u_h^n \in V_h + g^n$, which has the expansion

$$u_h^n = \sum_i U_i^n v_i + g^n, \tag{2.2}$$

in terms of the vector U^n which satisfies

$$C\frac{U^{n+1} - U^n}{k} + A\big(\theta U^{n+1} + (1-\theta)U^n\big) = F^{n+\theta}, \quad n = 0, 1, \ldots, N-1, \tag{2.3}$$

where U^0 is the vector of nodal values for some given and accurate approximation to $u_0 - g^0$ in V_h. It is well known that the local time truncation error is $\mathcal{O}(k)$ for backward Euler and $\mathcal{O}(k^2)$ for Crank–Nicolson, which we can write in the compact form $\mathcal{O}(k^{3-2\theta})$. Unfortunately, we must solve an implicit multi-dimensional problem for each time step.

2.2. *The AD method*

The Douglas–Gunn algorithm[6] for AD time discretization of (2.3) asks that, for each $n = 0, 1, \ldots, N-1$, we find $w^{n,m}$, for $m = 1, \ldots, d$, such that

$$C\frac{w^{n,m} - w^n}{k} + \sum_{i=1}^{m} A_i\big(\theta w^{n,i} + (1-\theta)w^n\big) + \sum_{i=m+1}^{d} A_i w^n = F^{n+\theta}, \tag{2.4}$$

and then set $w^{n+1} = w^{n,d}$, where w^n approximates u^n in the usual sense:

$$w^n_{\mathrm{AD},h} = \sum_i w^n_i \, v_i + g^n. \tag{2.5}$$

Note that we need to solve implicitly only a one-dimensional problem at each stage. This algorithm is written in the efficiently computable, equivalent form

$$(C + \theta k A_1) w^{n,1} = \left(C - (1-\theta) k A_1 - k \sum_{i=2}^{d} A_i \right) w^n + k F^{n+\theta}, \tag{2.6}$$

$$(C + \theta k A_m) w^{n,m} = C w^{n,m-1} + \theta k A_m w^n, \quad m = 2, \ldots, d, \tag{2.7}$$

$$w^{n+1} = w^{n,d}. \tag{2.8}$$

Multiply (2.7) by $(1 + \theta k C^{-1} A_1) \cdots (1 + \theta k C^{-1} A_{m-1})$, sum on m, and add (2.6) to see that

$$C \frac{w^{n+1} - w^n}{k} + A\big(\theta w^{n+1} + (1-\theta) w^n\big) + B(w^{n+1} - w^n) = F^{n+\theta}, \tag{2.9}$$

where

$$\begin{aligned} B &= k^{-1} C \big[(1 + \theta k C^{-1} A_1) \cdots (1 + \theta k C^{-1} A_d) - 1 - \theta k C^{-1} A \big] \\ &= \theta^2 k \sum_{1 \le m_1 < m_2 \le d} A_{m_1} C^{-1} A_{m_2} + \theta^3 k^2 \sum_{1 \le m_1 < m_2 < m_3 \le d} A_{m_1} C^{-1} A_{m_2} C^{-1} A_{m_3} \\ &\quad + \cdots + \theta^d k^{d-1} A_1 C^{-1} A_2 \cdots C^{-1} A_d. \end{aligned} \tag{2.10}$$

This equation is similar to (2.3), and from it we conclude, without being very precise, that the splitting error is

$$B(w^{n+1} - w^n) = kB\left(\frac{w^{n+1} - w^n}{k} \right) = \mathcal{O}\left(k^2 \sum_{i,j} \left| \frac{\partial}{\partial t} \frac{\partial^2}{\partial x_i^2} \frac{\partial^2}{\partial x_j^2} (u - g) \right| \right) = \mathcal{O}(k^2),$$

provided that w (i.e. $w_{\mathrm{AD},h} - g$) is a good approximation of $u - g$, and a, c, g and u are sufficiently smooth.

2.3. *The AD-M method*

The Douglas–Kim AD-M algorithm[7] adds a term to the right-hand side of (2.6); that is, for each $n = 1, 2, \ldots, N-1$, we find $W^{n,m}$ for $m = 1, \ldots, d$ such that

$$\begin{aligned} (C + \theta k A_1) W^{n,1} &= \left(C - (1-\theta) k A_1 - k \sum_{i=2}^{d} A_i \right) W^n \\ &\quad + k F^{n+\theta} + B(W^n - W^{n-1}), \end{aligned} \tag{2.11}$$

$$(C + \theta k A_m) W^{n,m} = C W^{n,m-1} + \theta k A_m W^n, \quad m = 2, \ldots, d, \tag{2.12}$$

$$W^{n+1} = W^{n,d}, \tag{2.13}$$

where W^n approximates u^n via

$$w^n_h = \sum_i W^n_i \, v_i + g^n. \tag{2.14}$$

We retain the property that only one-dimensional problems need to be solved implicitly at each stage. These equations imply

$$C\frac{W^{n+1}-W^n}{k}+A\big(\theta W^{n+1}+(1-\theta)W^n\big)+B(W^{n+1}-2W^n+W^{n-1})=F^{n+\theta}, \tag{2.15}$$

from which we see that the splitting error

$$\begin{aligned} B(W^{n+1}-2W^n+W^{n-1}) &= k^2B\bigg(\frac{W^{n+1}-2W^n+W^{n-1}}{k^2}\bigg) \\ &= \mathcal{O}\bigg(k^3\sum_{i,j}\bigg|\frac{\partial^2}{\partial t^2}\frac{\partial^2}{\partial x_i^2}\frac{\partial^2}{\partial x_j^2}(u-g)\bigg|\bigg)=\mathcal{O}(k^3) \end{aligned}$$

is improved in convergence order from the AD method. The price we pay is that we need some good approximation to W^1, which we might obtain, for example, by solving AD several times with a smaller time step.

3. A New Proof of Convergence for AD-M

As noted in the Introduction, Douglas and Kim[7] gave a convergence proof for AD-M under the assumption that the A_m commute. Douglas and Gunn[6] handled non-commutative problems, but required the restriction that kh^{-4} be sufficiently small. We significantly relax this constraint in this section.

Recall that $(\cdot,\cdot)$ is the inner-product in $L^2(\Omega)$, and let $\|\cdot\|$ denote the induced $L^2(\Omega)$-norm. We will also use some of the $H^\ell(\Omega)$-norms, denoted by $\|\cdot\|_\ell$. We make the following reasonable assumptions regarding the underlying discretizations.

Assumption 3.1. *For some constant $\mathcal{C}$ independent of h and k, $kh^{-2}\le\mathcal{C}$. Moreover, for any vector $\Psi\in\mathbb{R}^{\dim V_h}$,*

$$|\Psi|\le\mathcal{C}\|\psi\|,$$

where $\psi=\sum_i\Psi_iv_i$.

The latter inequality above holds if the mesh is quasi-uniform.

Assumption 3.2. *For some constant $\mathcal{C}$ depending on the smoothness of a, c and g, but not on h or k, the discrete spatial operators A_m and B satisfy:*

1. *Each A_m is bounded in the sense that $\|A_m\|\le\mathcal{C}h^{-2}$;*
2. *B satisfies the bound $|Bv|\le\mathcal{C}kh^{-4}|v|$.*

Moreover, $0<a_*\le a\le a^*<\infty$ and $0<c_*\le c\le c^*<\infty$ on $\bar\Omega$.

Note that Part 1 holds for the usual finite element spaces, and it (with Assumption 3.1) implies Part 2.

Assumption 3.3. *For some constant $\mathcal{C}$ depending on the smoothness of u, a, c and g, but not on h or k, the following hold.*

1. *The initial approximation is the elliptic projection, meaning that* $u_h^0 = w_h^0 \in V_h + g^0$ *satisfies*

$$(a\nabla(u - u_h^0), \nabla v) = 0, \quad v \in V_h. \tag{3.1}$$

Moreover,

$$\|u^0 - u_h^0\| + h\|u^0 - u_h^0\|_1 \le \mathcal{C}h^s.$$

2. *The solution* u_h *to the full scheme* (2.3) *approximates the true solution* u *of* (1.1)–(1.3) *in the sense that the error* $\mathcal{E}^n = u^n - u_h^n$ *satisfies*

$$\left(\sum_{n=0}^{N-1} \left\| \frac{\mathcal{E}^{n+1} - \mathcal{E}^n}{k} \right\|^2 k\right)^{1/2} + \max_{0\le m\le N} \|\mathcal{E}^m\| \le \mathcal{C}\{h^s + k^{3-2\theta}\},$$

$$\max_{0\le m\le N} \|\mathcal{E}^m\|_1 \le \mathcal{C}\{h^{s-1} + k^{3-2\theta}\}.$$

It is well known that this assumption holds for the usual s-order accurate finite element spaces, when we carefully choose the approximation to the initial condition.[1,15] In fact, Part 2 can be proven to hold using techniques similar to what we use in the remainder of this section.

To prove convergence of AD-M, let $E^n = U^n - W^n$ and subtract (2.15) from (2.3) to obtain

$$C\frac{E^{n+1} - E^n}{k} + A\big(\theta E^{n+1} + (1-\theta)E^n\big) = B(W^{n+1} - 2W^n + W^{n-1}). \tag{3.2}$$

In integral form, with $e^n = u_h^n - w_h^n$, the ith component is

$$\left(c\frac{e^{n+1} - e^n}{k}, v_i\right) + \big(a\nabla[\theta e^{n+1} + (1-\theta)e^n], \nabla v_i\big) = \big(B(W^{n+1} - 2W^n + W^{n-1})\big)_i. \tag{3.3}$$

Take the dot product of this with the test vector $E^{n+1} - E^n$, i.e. the test function $e^{n+1} - e^n$, to obtain for $n = 1, 2, \ldots, N-1$,

$$\begin{aligned}\left\| c^{1/2}\frac{e^{n+1} - e^n}{k} \right\|^2 k &+ \theta(a\nabla e^{n+1}, \nabla e^{n+1}) \\ &+ (1-2\theta)(a\nabla e^{n+1}, \nabla e^n) - (1-\theta)(a\nabla e^n, \nabla e^n) \\ &= B(W^{n+1} - 2W^n + W^{n-1}) \cdot (E^{n+1} - E^n).\end{aligned}$$

Thus, for either $\theta = 1$ or $\theta = 1/2$,

$$\begin{aligned} c_* \left\| \frac{e^{n+1} - e^n}{k} \right\|^2 k &+ \frac{1}{2}\big((a\nabla e^{n+1}, \nabla e^{n+1}) - (a\nabla e^n, \nabla e^n)\big) \\ &\le \mathcal{C}|B(W^{n+1} - 2W^n + W^{n-1})|\,\|e^{n+1} - e^n\| \\ &\le \mathcal{C}|B(W^{n+1} - 2W^n + W^{n-1})|^2 k + \frac{c_*}{2}\left\| \frac{e^{n+1} - e^n}{k} \right\|^2 k. \end{aligned} \tag{3.4}$$

Now for $\Psi \in \mathbb{R}^{\dim V_h}$, with $\psi = \sum_i \Psi_i v_i$, note that

$$\begin{aligned}
B\Psi &= \theta^2 k \sum_{1\le m_1<m_2\le d} A_{m_1} C^{-1} A_{m_2} \Psi + \cdots + \theta^d k^{d-1} A_1 C^{-1} A_2 \cdots C^{-1} A_d \Psi \\
&= \theta^2 k \sum_{1\le m_1<m_2\le d} A_{m_1} C^{-1} \big((a\partial_{m_2} v_i, \partial_{m_2}\psi) \big)_i \\
&\quad + \cdots + \theta^d k^{d-1} A_1 C^{-1} A_2 \cdots C^{-1} \big((a\partial_d v_i, \partial_d \psi) \big)_i \\
&= \theta^2 k \sum_{1\le m_1<m_2\le d} A_{m_1} C^{-1} \big((v_i, -\partial_{m_2} a \partial_{m_2}\psi) \big)_i \\
&\quad + \cdots + \theta^d k^{d-1} A_1 C^{-1} A_2 \cdots C^{-1} \big((v_i, -\partial_d a \partial_d \psi) \big)_i .
\end{aligned}$$

In this way, we see that B can be applied naturally to a function. Using Assumptions 3.1 and 3.2, we determine that

$$\begin{aligned}
&|B(W^{n+1} - 2W^n + W^{n-1})| \\
&\quad \le |B(e^{n+1} - 2e^n + e^{n-1})| + |B(\mathcal{E}^{n+1} - 2\mathcal{E}^n + \mathcal{E}^{n-1})| \\
&\qquad + |B((u-g)^{n+1} - 2(u-g)^n + (u-g)^{n-1})| \\
&\quad \le \mathcal{C}\bigg\{ \frac{k^2}{h^4} \bigg[\bigg\| \frac{e^{n+1} - e^n}{k} \bigg\| + \bigg\| \frac{e^n - e^{n-1}}{k} \bigg\| \bigg] \\
&\qquad + \bigg\| \frac{\mathcal{E}^{n+1} - \mathcal{E}^n}{k} \bigg\| + \bigg\| \frac{\mathcal{E}^n - \mathcal{E}^{n-1}}{k} \bigg\| + \bigg(\int_{t^{n-1}}^{t^{n+1}} \|(u-g)_{tt}\|_2^2 \, dt \bigg)^{1/2} k^{3/2} \bigg\}. \qquad (3.5)
\end{aligned}$$

We remark that the last term above was easily estimated to be $\mathcal{O}(k^2)$, which is all we need for our error estimate. However, it should be $\mathcal{O}(k^3)$, as noted above, but this is not so easy to prove rigorously in the finite element case.

Combining (3.4) and (3.5), we obtain

$$\begin{aligned}
&c_* \bigg\| \frac{e^{n+1} - e^n}{k} \bigg\|^2 k + (a\nabla e^{n+1}, \nabla e^{n+1}) - (a\nabla e^n, \nabla e^n) \\
&\qquad \le \mathcal{C} |B(W^{n+1} - 2W^n + W^{n-1})|^2 k \\
&\qquad \le \mathcal{C}\bigg\{ \bigg(\frac{k}{h^2} \bigg)^4 \bigg[\bigg\| \frac{e^{n+1} - e^n}{k} \bigg\|^2 k + \bigg\| \frac{e^n - e^{n-1}}{k} \bigg\|^2 k \bigg] \\
&\qquad\quad + \bigg\| \frac{\mathcal{E}^{n+1} - \mathcal{E}^n}{k} \bigg\|^2 k + \bigg\| \frac{\mathcal{E}^n - \mathcal{E}^{n-1}}{k} \bigg\|^2 k + \int_{t^{n-1}}^{t^{n+1}} \|(u-g)_{tt}\|_2^2 \, dt \, k^4 \bigg\}.
\end{aligned}$$

Summing on n from 1 to $m-1$, we see that

$$c_* \sum_{n=1}^{m-1} \left\| \frac{e^{n+1}-e^n}{k} \right\|^2 k + (a\nabla e^m, \nabla e^m)$$
$$\le (a\nabla e^1, \nabla e^1) + \mathcal{C}\left\{ \left(\frac{k}{h^2}\right)^4 \sum_{n=1}^{m-1} \left\| \frac{e^{n+1}-e^n}{k} \right\|^2 k + \left\| \frac{e^1-e^0}{k} \right\|^2 k \right.$$
$$\left. + \sum_{n=0}^{m-1} \left\| \frac{\mathcal{E}^{n+1}-\mathcal{E}^n}{k} \right\|^2 k + \int_0^T \|(u-g)_{tt}\|_2^2 \, dt \, k^4 \right\}.$$

For k/h^2 sufficiently small, we can remove the second term on the right-hand side, and the estimation of the error generated by the AD-M method is reduced to bounding the errors in w_h^0 and w_h^1. With Assumption 3.3, this gives us two of the convergence results stated below.

Theorem 3.1. *Under Assumptions* 3.1–3.3, *the solution* w_h^n *of the AD-M* (2.11)–(2.13) *converges to the solution* u *of* (1.1)–(1.3) *in the sense that the error* $\delta^n = u^n - w_h^n$ *satisfies*

$$\left(\sum_{n=0}^{N-1} \left\| \frac{\delta^{n+1}-\delta^n}{k} \right\|^2 k \right)^{1/2} + \max_{1\le m\le N} \|\delta^m\|$$
$$\le \mathcal{C}\left\{ h^s + k^{3-2\theta} + \|u_h^1 - w_h^1\|_1 + \left\| \frac{\delta^1-\delta^0}{k} \right\| \sqrt{k} \right\},$$
$$\max_{1\le m\le N} \|\delta^m\|_1 \le \mathcal{C}\left\{ h^{s-1} + k^{3-2\theta} + \|u_h^1 - w_h^1\|_1 + \left\| \frac{\delta^1-\delta^0}{k} \right\| \sqrt{k} \right\},$$

provided that k *and* k/h^2 *are sufficiently small, wherein* $\mathcal{C}$ *depends on the smoothness of* u, a, c *and* g, *but not on* h *or* k, *and* u_h *is the solution to the unmodified method* (2.2)–(2.3).

It remains to prove the estimate on $\max \|\delta^m\|$. Returning to (3.3), we take test vector $\theta E^{n+1} + (1-\theta)E^n$ to obtain, for $n = 1, 2, \ldots, N-1$,

$$\left(c\frac{e^{n+1}-e^n}{k}, \theta e^{n+1} + (1-\theta)e^n \right)$$
$$+ \left(a\nabla[\theta e^{n+1} + (1-\theta)e^n], \nabla[\theta e^{n+1} + (1-\theta)e^n] \right)$$
$$= B(W^{n+1} - 2W^n + W^{n-1}) \cdot \left(\theta E^{n+1} + (1-\theta)E^n \right).$$

Again, for either $\theta = 1$ or $\theta = 1/2$, we can estimate

$$\frac{1}{2}\left(\|c^{1/2}e^{n+1}\|^2 - \|c^{1/2}e^n\|^2 \right)$$
$$\le \mathcal{C}|B(W^{n+1} - 2W^n + W^{n-1})| \, \|\theta e^{n+1} + (1-\theta)e^n\| k$$
$$\le \mathcal{C}\left\{ |B(W^{n+1} - 2W^n + W^{n-1})|^2 k + \left(\|e^{n+1}\|^2 + \|e^n\|^2 \right) k \right\}$$

$$\le \mathcal{C}\Big\{ \big(\|e^{n+1}\|^2 + \|e^n\|^2\big)k + \left\|\frac{e^{n+1}-e^n}{k}\right\|^2 k + \left\|\frac{e^n - e^{n-1}}{k}\right\|^2 k$$
$$+ \left\|\frac{\mathcal{E}^{n+1}-\mathcal{E}^n}{k}\right\|^2 k + \left\|\frac{\mathcal{E}^n - \mathcal{E}^{n-1}}{k}\right\|^2 k + \int_{t^{n-1}}^{t^{n+1}} \|(u-g)_{tt}\|_2^2\, dt\, k^4 \Big\},$$

using (3.5) again. Now sum on n from 1 to $m-1$ to obtain that

$$\|c^{1/2}e^m\|^2 \le \|c^{1/2}e^1\|^2 + \mathcal{C}\Big\{ \sum_{n=1}^{m} \|e^n\|^2 k + \sum_{n=0}^{m-1} \left\|\frac{e^{n+1}-e^n}{k}\right\|^2 k$$
$$+ \sum_{n=0}^{m-1} \left\|\frac{\mathcal{E}^{n+1}-\mathcal{E}^n}{k}\right\|^2 k + \int_0^T \|(u-g)_{tt}\|_2^2\, dt\, k^4 \Big\}.$$

We can remove the second term on the right-hand side using the Gronwall inequality, provided that k is sufficiently small. The third term on the right-hand side is estimated using the first result of the theorem. Assumption 3.3 completes the proof in the finite element case. The finite difference case is similar to the above, only simpler, since it is trivial to apply B to a set of finite difference points of the true solution.

4. The Mixed Finite Element Formulation

We rewrite (1.1)–(1.3) in mixed form by defining

$$\mathbf{q} = -a\nabla u, \tag{4.1}$$

and then, with $\alpha(x) = 1/a(x)$, we have

$$\alpha\mathbf{q} + \nabla u = 0, \quad x \in \Omega, \quad 0 < t \le T, \tag{4.2}$$
$$cu_t + \nabla\cdot\mathbf{q} = f, \quad x \in \Omega, \quad 0 < t \le T, \tag{4.3}$$
$$u = g, \quad x \in \partial\Omega, \quad 0 < t \le T, \tag{4.4}$$
$$u = u_0, \quad x \in \Omega, \quad t = 0. \tag{4.5}$$

Define the function spaces

$$\mathbf{V} = H(\mathrm{div};\Omega) = \{\mathbf{q} \in (L^2(\Omega))^d : \nabla\cdot\mathbf{q} \in L^2(\Omega)\},$$
$$W = L^2(\Omega).$$

If (4.2) is tested by a function in $\mathbf{V}$ and (4.3) is tested by a function in W, the weak form of (4.2)–(4.4) of interest for the mixed method results; that is, we find for each time $(\mathbf{q}, u) \in V \times W$ such that

$$(\alpha\mathbf{q}, \mathbf{v}) - (\nabla\cdot\mathbf{v}, u) = -\langle g, \mathbf{v}\cdot\nu\rangle, \quad \mathbf{v} \in \mathbf{V}, \tag{4.6}$$
$$(cu_t, w) + (\nabla\cdot\mathbf{q}, w) = (f, w), \qquad w \in W, \tag{4.7}$$

where the inner-product $(\cdot,\cdot)$ is taken in W or W^d, as appropriate, $\langle\cdot,\cdot\rangle$ is the inner-product in $L^2(\partial\Omega)$, and ν is the outer unit normal to $\partial\Omega$.

The mixed finite element method approximates the solution in a properly chosen subspace $\mathbf{V}_h \times W_h$ of $\mathbf{V} \times W$ (see, e.g., Ref. 14). In semi-discrete form (i.e. discretizing space only), we seek $(\mathbf{q}_h(t), u_h(t)) \in \mathbf{V}_h \times W_h$ such that

$$(\alpha \mathbf{q}_h, \mathbf{v}) - (\nabla \cdot \mathbf{v}, u_h) = -\langle g, \mathbf{v} \cdot \nu \rangle, \quad \mathbf{v} \in \mathbf{V}_h, \tag{4.8}$$

$$(cu_{h,t}, w) + (\nabla \cdot \mathbf{q}_h, w) = (f, w), \quad w \in W_h. \tag{4.9}$$

5. AD-M for Mixed Finite Elements

We derive our AD-M method only for the case that $d = 2$; the extension to $d = 3$ is straightforward. We assume that $\mathbf{V}_h \times W_h$ is the Raviart–Thomas space[14] RT_{s-1} of index $s - 1 \geq 0$, though spaces with similar properties could be used instead, such as the Brezzi–Douglas–Fortin–Marini spaces.[2] The usual basis for $\mathbf{V}_h$,

$$\mathbf{V}_h = \operatorname*{span}_{i,j} \{\mathbf{v}_{x,i}, \mathbf{v}_{y,j}\},$$

has the properties that the vector function $\mathbf{v}_{x,i}$ has a vanishing y-component, is supported in at most two grid elements sharing an edge with normal in the x-direction, and is discontinuous in the y-direction; similarly, $\mathbf{v}_{y,j}$ has a vanishing x-component, is supported in at most two grid elements sharing an edge with normal in the y-direction, and is discontinuous in the x-direction. The usual basis for W_h,

$$W_h = \operatorname*{span}_{\ell} \{w_\ell\},$$

is piecewise discontinuous over the grid.

We now reduce (4.8)–(4.9) to a system of linear equations. First consider the matrix A arising from the form $(\alpha \mathbf{v}_1, \mathbf{v}_2)$, for two basis functions of $\mathbf{V}_h$. Any mixture of x and y basis functions results in

$$(\alpha \mathbf{v}_{x,i}, \mathbf{v}_{y,j}) = 0,$$

so A is block diagonal with blocks

$$A_{x,i_1,i_2} = (\alpha \mathbf{v}_{x,i_1}, \mathbf{v}_{x,i_2}),$$
$$A_{y,j_1,j_2} = (\alpha \mathbf{v}_{y,j_1}, \mathbf{v}_{y,j_2}).$$

Note that both A_x and A_y are invertible. Moreover, with the appropriate ordering, these matrices are banded with bands densely concentrated around the diagonal. For the x-basis functions, we use standard (i, j) ordering of the elements, with i advancing fastest. Starting from $(1, 1)$, we progress through the grid, numbering all $\mathbf{v}_{x,\ell}$ with support in (i, j). However, for the y-basis functions, we use (i, j) ordering of the elements with j advancing fastest. Thus these matrices have just a few bands near the diagonal, and so it is easy to solve linear subsystems involving A_x and A_y.

With any ordering of the elements, we also have the nonsingular, block diagonal matrix C defined by

$$C_{\ell_1,\ell_2} = (cw_{\ell_1}, w_{\ell_2}).$$

Finally, let

$$B_{x,i,\ell} = (w_\ell, \nabla\cdot\mathbf{v}_{x,i}),$$
$$B_{y,j,\ell} = (w_\ell, \nabla\cdot\mathbf{v}_{y,j}),$$

which are sparse, but not particularly well structured.

Let us expand the solution in the basis as

$$u_h(x,y,t) = \sum_\ell z_\ell(t) w_\ell(x,y), \tag{5.1}$$

$$\mathbf{q}_h(x,y,t) = \sum_i \mu_i(t)\mathbf{v}_{x,i}(x,y) + \sum_j \lambda_j(t)\mathbf{v}_{y,j}(x,y). \tag{5.2}$$

Then Eqs. (4.8)–(4.9) reduce to the system of linear equations

$$A_x\mu - B_x z = G_x, \tag{5.3}$$
$$A_y\lambda - B_y z = G_y, \tag{5.4}$$
$$Cz_t + B_x^T\mu + B_y^T\lambda = F, \tag{5.5}$$

where

$$G_{x,i} = -\langle g, \mathbf{v}_{x,i}\cdot\nu\rangle,$$
$$G_{y,j} = -\langle g, \mathbf{v}_{y,j}\cdot\nu\rangle,$$
$$F_\ell = (f, w_\ell).$$

5.1. *Backward Euler and Crank–Nicolson time discretization*

When we employ backward Euler ($\theta = 1$) or Crank–Nicolson ($\theta = 1/2$) time discretization, we begin with some approximation of z^0 and define μ^0 and λ^0 from

$$A_x\mu^0 - B_x z^0 = G_x^0, \tag{5.6}$$
$$A_y\lambda^0 - B_y z^0 = G_y^0, \tag{5.7}$$

and then, for $n = 0, 1, \ldots, N-1$, our system becomes

$$A_x\mu^{n+1} - B_x z^{n+1} = G_x^{n+1}, \tag{5.8}$$
$$A_y\lambda^{n+1} - B_y z^{n+1} = G_y^{n+1}, \tag{5.9}$$
$$C\frac{z^{n+1}-z^n}{k} + B_x^T\big(\theta\mu^{n+1} + (1-\theta)\mu^n\big) + B_y^T\big(\theta\lambda^{n+1} + (1-\theta)\lambda^n\big) = F^{n+\theta}. \tag{5.10}$$

This is an indefinite saddle-point problem, and, therefore, generally difficult to solve.

Now let

$$M_x = B_x^T A_x^{-1} B_x \quad\text{and}\quad M_y = B_y^T A_y^{-1} B_y.$$

By solving (5.6) and (5.7) or (5.8) and (5.9) for μ^n and λ^n, we reduce (5.10) to

$$C\frac{z^{n+1}-z^n}{k}+(M_x+M_y)\big(\theta z^{n+1}+(1-\theta)z^n\big)=\mathcal{F}^{n+1}, \tag{5.11}$$

where

$$\mathcal{F}^{n+1}=F^{n+\theta}-B_x^TA_x^{-1}\big(\theta G_x^{n+1}+(1-\theta)G_x^n\big)-B_y^TA_y^{-1}\big(\theta G_y^{n+1}+(1-\theta)G_y^n\big). \tag{5.12}$$

Now (5.11) is positive definite, but unfortunately, M_x+M_y is a full matrix, and so it is still difficult to solve.

5.2. *An Uzawa mixed AD method*

An efficiently computable Uzawa AD algorithm (see Ref. 9) iterates on $n=0$, $1,\dots,N-1$ as follows:

x-sweep:

$$A_x\mu^{n,1}-B_xz^{n,1}=G_x^{n+1}, \tag{5.13}$$

$$C\frac{z^{n,1}-z^n}{k}+B_x^T\big(\theta\mu^{n,1}+(1-\theta)\mu^n\big)+B_y^T\lambda^n=F^{n+\theta}, \tag{5.14}$$

y-sweep:

$$A_y\lambda^{n+1}-B_yz^{n+1}=G_y^{n+1}, \tag{5.15}$$

$$C\frac{z^{n+1}-z^n}{k}+B_x^T\big(\theta\mu^{n,1}+(1-\theta)\mu^n\big)+B_y^T\big(\theta\lambda^{n+1}+(1-\theta)B_y^T\lambda^n\big)=F^{n+\theta}, \tag{5.16}$$

corrector step:

$$A_x\mu^{n+1}-B_xz^{n+1}=G_x^{n+1}. \tag{5.17}$$

Eliminating μ^n and λ^n, the Uzawa AD algorithm becomes

$$C\frac{z^{n,1}-z^n}{k}+M_x\big(\theta z^{n,1}+(1-\theta)z^n\big)+M_yz^n=\mathcal{F}^{n,1}, \tag{5.18}$$

$$C\frac{z^{n+1}-z^n}{k}+M_x\big(\theta z^{n,1}+(1-\theta)z^n\big)+M_y\big(\theta z^{n+1}+(1-\theta)z^n\big)=\mathcal{F}^{n+1}, \tag{5.19}$$

where

$$\mathcal{F}^{n,1}=F^{n+\theta}-B_x^{\mathrm{T}}A_x^{-1}\big(\theta G_x^{n+1}+(1-\theta)G_x^n\big)-B_y^{\mathrm{T}}A_y^{-1}G_y^n. \tag{5.20}$$

Subtract (5.18) from (5.19), multiply the result by θkM_xC^{-1}, and combine with (5.19) to obtain the single equation

$$\begin{aligned}C\frac{z^{n+1}-z^n}{k}&+(M_x+M_y)\big(\theta z^{n+1}+(1-\theta)z^n\big)+\theta^2kM_xC^{-1}M_y(z^{n+1}-z^n)\\&=\mathcal{F}^{n+1}-\theta^2kM_xC^{-1}B_y^TA_y^{-1}(G_y^{n+1}-G_y^n).\end{aligned} \tag{5.21}$$

Comparing this with (5.11) shows that the splitting error is

$$\theta^2kM_xC^{-1}\big[M_y(z^{n+1}-z^n)+B_y^{\mathrm{T}}A_y^{-1}(G_y^{n+1}-G_y^n)\big]=\theta^2kM_xC^{-1}B_y^{\mathrm{T}}(\lambda^{n+1}-\lambda^n),$$

which is $\mathcal{O}(k^2)$ for a sufficiently smooth solution u and boundary condition g.

5.3. *A new Uzawa mixed AD-M method*

Equations (5.21) are not efficiently computable, but they do illuminate the splitting error and suggest, similar to (2.9) in Ref. 7, that we can reduce it to $\mathcal{O}(k^3)$ by adding terms to make the splitting error equal to

$$\begin{aligned}\theta^2 k M_x C^{-1}\big[M_y(z^{n+1}-2z^n+z^{n-1})+B_y^T A_y^{-1}(G_y^{n+1}-2G_y^n+G_y^{n-1})\big]\\ =\theta^2 k M_x C^{-1}B_y^T(\lambda^{n+1}-2\lambda^n+\lambda^{n-1})=\mathcal{O}(k^3).\end{aligned}$$

Then the local splitting error would be higher order in k than the local error for the backward Euler or Crank–Nicolson approximation, and also for the local splitting error of the AD method itself.

Although A_x and A_y have bands concentrated near the diagonal, their inverses may be full, so to avoid computing the inverse of A_x and A_y, we propose the efficiently computable algorithm, for $n=1,2,\dots,N-1$,
x-sweep:

$$A_x\mu^{n,1}-B_x z^{n,1}+\theta k B_x C^{-1}B_y^{\mathrm{T}}(\lambda^n-\lambda^{n-1})=G_x^{n+1}, \tag{5.22}$$

$$C\frac{z^{n,1}-z^n}{k}+B_x^{\mathrm{T}}\big(\theta\mu^{n,1}+(1-\theta)\mu^n\big)+B_y^{\mathrm{T}}\lambda^n=F^{n+\theta}, \tag{5.23}$$

y-sweep:

$$A_y\lambda^{n+1}-B_y z^{n+1}=G_y^{n+1}, \tag{5.24}$$

$$C\frac{z^{n+1}-z^n}{k}+B_x^{\mathrm{T}}\big(\theta\mu^{n,1}+(1-\theta)\mu^n\big)+B_y^{\mathrm{T}}\big(\theta\lambda^{n+1}+(1-\theta)\lambda^n\big)=F^{n+\theta}, \tag{5.25}$$

corrector step:

$$A_x\mu^{n+1}-B_x z^{n+1}=G_x^{n+1}. \tag{5.26}$$

After some manipulation, we have

$$\begin{aligned}C\frac{z^{n+1}-z^n}{k}+(M_x+M_y)\big(\theta z^{n+1}+(1-\theta)z^n\big)\\ +\theta^2 k M_x C^{-1}M_y(z^{n+1}-2z^n+z^{n-1})\\ =\mathcal{F}^{n+1}-\theta^2 k M_x C^{-1}B_y^T A_y^{-1}(G_y^{n+1}-2G_y^n+G_y^{n-1}),\end{aligned} \tag{5.27}$$

which has the promised $\mathcal{O}(k^3)$ splitting error.

6. Convergence of Mixed AD-M

We suppose that $\mathbf{V}_h\times W_h$ approximates $\mathbf{V}\times W$ as in the case of RT_{s-1}, i.e.

$$\min_{\mathbf{v}\in\mathbf{V}_h}\|\mathbf{q}-\mathbf{v}\|\le\mathcal{C}\|\mathbf{q}\|_s h^s \quad\text{and}\quad \min_{w\in W_h}\|u-w\|\le\mathcal{C}\|u\|_s h^s. \tag{6.1}$$

All the usual mixed spaces satisfy the property that $\nabla\cdot\mathbf{V}_h = W_h$, and they each have a linear projection operator $\pi : \mathbf{V} \cap (L^p(\Omega))^d \to \mathbf{V}_h$, where $p > 2$, such that

$$\|\mathbf{v} - \pi\mathbf{v}\| \le \mathcal{C}\|\mathbf{v}\|_s h^s, \tag{6.2}$$

$$(\nabla\cdot(\mathbf{v} - \pi\mathbf{v}), w) = 0, \quad w \in W_h, \tag{6.3}$$

$$\|\nabla\cdot(\mathbf{v} - \pi\mathbf{v})\| \le \mathcal{C}\|\nabla\cdot\mathbf{v}\|_s h^s. \tag{6.4}$$

We also let $\mathcal{P}$ be the linear orthogonal projection operator of $L^2(\Omega)$ onto W_h. We will also need below $\mathcal{P}_0$, which is the linear orthogonal projection operator of $L^2(\Omega)$ onto the space of discontinuous constants (i.e. the scalar space of RT_0). Trivially,

$$\|c - \mathcal{P}_0 c\|_{L^\infty} \le \mathcal{C}\|\nabla c\|_{L^\infty} h,$$

where $\|\cdot\|_{L^\infty}$ is the L^∞-norm.

Mixed methods on rectangles have many interesting and important superconvergence properties. For example, it is known[12,10,11] that the weighted L^2-projection $\mathcal{P}^\alpha_{\mathbf{V}}$, defined by

$$\big(\alpha(\mathbf{v} - \mathcal{P}^\alpha_{\mathbf{V}}\mathbf{v}), \tilde{\mathbf{v}}\big) = 0, \quad \tilde{\mathbf{v}} \in \mathbf{V}_h,$$

has the property that

$$\|\pi\mathbf{v} - \mathcal{P}^\alpha_{\mathbf{v}}\mathbf{v}\| \le \mathcal{C}\|\mathbf{v}\|_{s+1} h^{s+1}. \tag{6.5}$$

Recalling (5.1)–(5.2), we have from (4.6) and either (5.6)–(5.7) or (5.26), (5.24) that the errors $e^n = u^n - u^n_h$ and $\sigma^n = \mathbf{q}^n - \mathbf{q}^n_h$ satisfy, for each $n = 0, 1, \ldots, N$,

$$(\alpha\pi\sigma^n, \mathbf{v}) - (\nabla\cdot\mathbf{v}, \mathcal{P}e^n) = \big(\alpha(\pi\mathbf{q}^n - \mathbf{q}^n, \mathbf{v}\big), \quad \mathbf{v} \in \mathbf{V}_h. \tag{6.6}$$

Solving either (5.6)–(5.7) or (5.26), (5.24) for μ^n and λ^n, we can rewrite (5.27) as

$$\begin{aligned} C\frac{z^{n+1} - z^n}{k} &+ B_x^{\mathrm{T}}\big(\theta\mu^{n+1} + (1-\theta)\mu^n\big) + B_y^{\mathrm{T}}\big(\theta\lambda^{n+1} + (1-\theta)\lambda^n\big) \\ &= \theta F^{n+1} + (1-\theta)F^n - \theta^2 k M_x C^{-1} B_y^{\mathrm{T}}(\lambda^{n+1} - 2\lambda^n + \lambda^{n-1}). \end{aligned} \tag{6.7}$$

From (4.7) at times t^n and t^{n+1}, then, we obtain for $w_i \in W_h$ that

$$\begin{aligned} \left(c\mathcal{P}\frac{e^{n+1} - e^n}{k}, w_i\right) &+ \big(\nabla\cdot[\theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n], w_i\big) \\ &= \theta^2 k\big(M_x C^{-1} B_y^T(\lambda^{n+1} - 2\lambda^n + \lambda^{n-1})\big)_i \\ &\quad + \left(c\left[\mathcal{P}\frac{u^{n+1} - u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right], w_i\right), \end{aligned} \tag{6.8}$$

wherein we introduced π trivially using (6.3).

We bound the splitting error as follows. First note that

$$\begin{aligned} B_y^{\mathrm{T}}\lambda &= M_y z + B_y^{\mathrm{T}} A_y^{-1} G_y \\ &= -M_y E + M_y U + B_y^{\mathrm{T}} A_y^{-1} G_y \\ &= -M_y E + B_y^{\mathrm{T}} A_y^{-1}(B_y U + G_y), \end{aligned}$$

where E and U are the vectors of finite element coefficients of $\mathcal{P}e$ and $\mathcal{P}u$, respectively. Now, if we take test function $\mathbf{v}_{y,j}$ in (4.6) and let $\mathcal{Q}$ be the α-weighted linear orthogonal projection of $\big(L^2(\Omega)\big)^2$ onto $\mathbf{V}_h$, we see that

$$(\alpha\mathcal{Q}\mathbf{q}, \mathbf{v}_{y,j}) - (\nabla\cdot\mathbf{v}_{y,j}, \mathcal{P}u) = -\langle g, \mathbf{v}_{y,j}\cdot\nu\rangle,$$

and, with Q being the vector of finite element coefficients of $\mathcal{Q}\mathbf{q}$,

$$A_y Q - B_y U = G_y.$$

Combining, we have that

$$B_y^{\mathrm{T}}\lambda = -M_y E + B_y^{\mathrm{T}} Q.$$

The second step in bounding the splitting error is to note that $\|A_x\| + \|A_y\| + \|C\| \le \mathcal{C}$ and $\|A_x^{-1}\| + \|A_y^{-1}\| + \|C^{-1}\| \le \mathcal{C}$, and that we have only $\|B_x\| + \|B_y\| \le \mathcal{C}h^{-1}$. Thus,

$$\begin{aligned} &|\theta^2 k M_x C^{-1} B_y^{\mathrm{T}}(\lambda^{n+1} - 2\lambda^n + \lambda^{n-1})| \\ &\quad\le |k M_x C^{-1} M_y (E^{n+1} - 2E^n + E^{n-1})| + |k M_x C^{-1} B_y^{\mathrm{T}}(Q^{n+1} - 2Q^n + Q^{n-1})| \\ &\quad\le \mathcal{C}kh^{-2}\big\{h^{-2}\|\mathcal{P}(e^{n+1} - 2e^n + e^{n-1})\| + \|\partial_2(\mathcal{Q}q_2^{n+1} - 2\mathcal{Q}q_2^n + \mathcal{Q}q_2^{n-1})\|\big\} \\ &\quad\le \mathcal{C}\frac{k}{h^2}\left\{\frac{k}{h^2}\left(\left\|\mathcal{P}\frac{e^{n+1} - e^n}{k}\right\| + \left\|\mathcal{P}\frac{e^n - e^{n-1}}{k}\right\|\right) + k\int_{t^{n-1}}^{t^{n+1}} \|\partial_2(\mathcal{Q}q_2)_{tt}\|\,dt\right\} \\ &\quad\le \mathcal{C}\frac{k}{h^2}\left\{\frac{k}{h^2}\left(\left\|\mathcal{P}\frac{e^{n+1} - e^n}{k}\right\| + \left\|\mathcal{P}\frac{e^n - e^{n-1}}{k}\right\|\right) + k^{3/2}\left(\int_{t^{n-1}}^{t^{n+1}} \|q_{2,tt}\|_1^2\,dt\right)^{1/2}\right\}, \end{aligned} \tag{6.9}$$

using the stability of the L^2-projection $\mathcal{Q}$ in H^1.

The overall analysis of (6.8) proceeds much as in the non-mixed case. First take the difference of (6.6) at times t^{n+1} and t^n, and then choose the test function $\mathbf{v} = \theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n$. Combine the result with $w = \mathcal{P}(e^{n+1} - e^n)$ in (6.8), and obtain that

$$\begin{aligned} &\left(c\mathcal{P}\frac{e^{n+1} - e^n}{k}, \mathcal{P}(e^{n+1} - e^n)\right) + \big(\alpha(\pi\sigma^{n+1} - \pi\sigma^n), \theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n\big) \\ &\qquad = \theta^2 k M_x C^{-1} B_y^{\mathrm{T}}(\lambda^{n+1} - 2\lambda^n + \lambda^{n-1})\cdot(E^{n+1} - E^n) \\ &\qquad\quad + \left(\mathcal{P}\left\{c\left[\mathcal{P}\frac{u^{n+1} - u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right]\right\}, \mathcal{P}(e^{n+1} - e^n)\right) \\ &\qquad\quad + \big(\alpha(\pi\mathbf{q}^{n+1} - \mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^{n+1} - \pi\mathbf{q}^n + \mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^n), \theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n\big), \end{aligned}$$

wherein we introduced the operators $\mathcal{P}$ and $\mathcal{P}^{\alpha}_{\mathbf{v}}$. Again assuming that k/h^2 is sufficiently small, after some manipulation similar to that in Sec. 3, we obtain

$$\begin{aligned}
&\frac{1}{2}\left\|\sqrt{c}\mathcal{P}\frac{e^{n+1}-e^n}{k}\right\|^2 k + \frac{1}{2}\big(\|\sqrt{\alpha}\pi\sigma^{n+1}\|^2 - \|\sqrt{\alpha}\pi\sigma^n\|^2\big)\\
&\quad\le \mathcal{C}\left\{\left(\frac{k}{h^2}\right)^4\left[\left\|\mathcal{P}\frac{e^{n+1}-e^n}{k}\right\|^2 k + \left\|\mathcal{P}\frac{e^n-e^{n-1}}{k}\right\|^2 k\right]\right.\\
&\qquad + k^4\int_{t^{n-1}}^{t^{n+1}} \|q_{2,tt}\|_1^2\,dt + \left\|\mathcal{P}\left(c\left[\mathcal{P}\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right]\right)\right\|^2 k\\
&\qquad \left. + \big\|(\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}^{n+1} - \pi\mathbf{q}^{n+1}) - (\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}^n - \pi\mathbf{q}^n)\big\|^2 k^{-1} + \|\sigma^{n+1}\|^2 k + \|\sigma^n\|^2 k\right\}.
\end{aligned}$$

For the time truncation error, we expand

$$c = (c - \mathcal{P}_0 c) + \mathcal{P}_0 c$$

and bound

$$\begin{aligned}
&\left\|\mathcal{P}\left((c-\mathcal{P}_0 c)\left[\mathcal{P}\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right]\right)\right\|^2 k\\
&\quad\le \|c-\mathcal{P}_0 c\|_{L^\infty}\left\{\left\|\mathcal{P}\left(\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right)\right\|\right.\\
&\qquad \left. + \theta\|u_t^{n+1} - \mathcal{P}u_t^{n+1}\| + (1-\theta)\|u_t^n - \mathcal{P}u_t^n\|\right\}^2 k\\
&\quad\le \mathcal{C}\left\{k^{6-4\theta}h^2\int_{t^n}^{t^{n+1}} \|\partial_t^{4-2\theta}u\|^2\,dt + \big(\|u_t^{n+1}\|_s^2 + \|u_t^n\|_s^2\big)h^{2(s+1)}k\right\},
\end{aligned}$$

since $\mathcal{P}$ is bounded in the L^2-norm. The remaining term is then

$$\left\|\mathcal{P}_0 c\,\mathcal{P}\left[\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n\right]\right\|^2 k \le \mathcal{C}k^{6-4\theta}\int_{t^n}^{t^{n+1}} \|\partial_t^{4-2\theta}u\|^2\,dt.$$

Moreover, we have that

$$\begin{aligned}
\big\|(\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}^{n+1} - \pi\mathbf{q}^{n+1}) - (\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}^n - \pi\mathbf{q}^n)\big\|^2 k^{-1} &= \left\|\int_{t^n}^{t^{n+1}} (\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q} - \pi\mathbf{q})_t\,dt\right\|^2 k^{-1}\\
&\le \left(\int_{t^n}^{t^{n+1}} \|\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}_t - \pi\mathbf{q}_t\|\,dt\right)^2 k^{-1} \le \int_{t^n}^{t^{n+1}} \|\mathcal{P}^{\alpha}_{\mathbf{v}}\mathbf{q}_t - \pi\mathbf{q}_t\|^2\,dt\\
&\le h^{2(s+1)}\int_{t^n}^{t^{n+1}} \|\mathbf{q}_t\|_{s+1}^2\,dt.
\end{aligned}$$

Finally, Gronwall's lemma implies that, for k and k/h^2 sufficiently small,

$$\begin{aligned}
&\sum_{n=1}^{N-1} \left\| \mathcal{P}\frac{e^{n+1}-e^n}{k} \right\|^2 k + \max_{1\le n\le N} \|\pi\sigma^n\|^2 \\
&\quad \le \mathcal{C}\biggl\{ \|\pi\sigma^1\|^2 + \left\| \mathcal{P}\frac{e^1-e^0}{k} \right\|^2 k + k^4 \int_0^T \|q_{2,tt}\|_1^2 \, dt + k^{6-4\theta} \int_0^T \|\partial_t^{4-2\theta} u\|^2 \, dt \\
&\qquad + h^{2(s+1)} \biggl(\sum_{n=1}^{N} \|u_t^n\|_s^2 k + \int_0^T \|\mathbf{q}_t\|_{s+1}^2 \, dt \biggr) \biggr\}. \qquad (6.10)
\end{aligned}$$

For our second estimate of (6.8), take the θ-weighted average of (6.6) at times t^{n+1} and t^n, and choose the test function $\mathbf{v} = \theta\pi\sigma^n + (1-\theta)\pi\sigma^n$. Combined with (6.8) using $w = \theta\mathcal{P}e^{n+1} + (1-\theta)\mathcal{P}e^n$, it follows that

$$\begin{aligned}
&\biggl(c\mathcal{P}\frac{e^{n+1}-e^n}{k}, \theta\mathcal{P}e^{n+1} + (1-\theta)\mathcal{P}e^n \biggr) \\
&\qquad + \bigl(\alpha(\theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n), \theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n \bigr) \\
&\qquad = \theta^2 k M_x C^{-1} B_y^T (\lambda^{n+1} - 2\lambda^n + \lambda^{n-1}) \cdot [\theta\mathcal{P}E^{n+1} + (1-\theta)\mathcal{P}E^n] \\
&\qquad\quad + \biggl(\mathcal{P}\biggl\{ c\biggl[\mathcal{P}\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n \biggr] \biggr\}, \theta\mathcal{P}e^{n+1} + (1-\theta)\mathcal{P}e^n \biggr) \\
&\qquad\quad + \bigl(\alpha[\theta(\pi\mathbf{q}^{n+1} - \mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^{n+1}) + (1-\theta)(\pi\mathbf{q}^n - \mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^n)], \theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n \bigr).
\end{aligned}$$

Again after some manipulation, we have that

$$\begin{aligned}
&\frac{1}{2}\bigl(\|\sqrt{c}\mathcal{P}e^{n+1}\|^2 - \|\sqrt{c}\mathcal{P}e^n\|^2 \bigr) + \frac{1}{2}\|\sqrt{\alpha}(\theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n)\|^2 k \\
&\quad \le \mathcal{C}\biggl\{ \left\| \mathcal{P}\frac{e^{n+1}-e^n}{k} \right\|^2 k + \left\| \mathcal{P}\frac{e^n-e^{n-1}}{k} \right\|^2 k + k^4 \int_{t^{n-1}}^{t^{n+1}} \|q_{2,tt}\|_1^2 \, dt \\
&\qquad + \left\| \mathcal{P}\biggl(c\biggl[\mathcal{P}\frac{u^{n+1}-u^n}{k} - \theta u_t^{n+1} - (1-\theta)u_t^n \biggr] \biggr) \right\|^2 k + \|\mathcal{P}e^{n+1}\|^2 k + \|\mathcal{P}e^n\|^2 k \\
&\qquad + \|\mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^{n+1} - \pi\mathbf{q}^{n+1}\|^2 k + \|\mathcal{P}_{\mathbf{v}}^{\alpha}\mathbf{q}^n - \pi\mathbf{q}^n\|^2 k \biggr\}.
\end{aligned}$$

Gronwall's lemma and the previous estimate (6.10) implies that

$$\begin{aligned}
&\max_{1\le n\le N} \|\mathcal{P}e^n\|^2 + \sum_{n=1}^{N-1} \|\theta\pi\sigma^{n+1} + (1-\theta)\pi\sigma^n)\|^2 k \\
&\quad \le \mathcal{C}\biggl\{ \|\mathcal{P}e^1\|^2 + \|\pi\sigma^1\|^2 + \left\| \mathcal{P}\frac{e^1-e^0}{k} \right\|^2 k \\
&\qquad + k^4 \int_0^T \|q_{2,tt}\|_1^2 \, dt + k^{6-4\theta} \int_0^T \|\partial_t^{4-2\theta} u\|^2 \, dt \\
&\qquad + h^{2(s+1)} \biggl(\sum_{n=1}^{N} \bigl(\|u_t^n\|_s^2 + \|\mathbf{q}^n\|_{s+1}^2 \bigr) k + \int_0^T \|\mathbf{q}_t\|_{s+1}^2 \, dt \biggr) \biggr\}. \qquad (6.11)
\end{aligned}$$

Our results (6.10) and (6.11) lead to the following theorem.

Theorem 6.1. *Assuming* (6.1), *the solution* $(u_h^n, \mathbf{q}_h)$ *of the mixed AD-M* (5.22)–(5.26) *converges to the solution* $(u, \mathbf{q})$ *of* (4.6)–(4.7) *in the sense that the errors* $e^n = u^n - u_h^n$ *and* $\sigma^n = \mathbf{q}^n - \mathbf{q}_h^n$ *satisfy*

$$\begin{aligned}
&\left(\sum_{n=1}^{N-1} \left\| \mathcal{P} \frac{e^{n+1} - e^n}{k} \right\|^2 k \right)^{1/2} + \max_{1 \le n \le N} \|\mathcal{P} e^n\| + \max_{1 \le n \le N} \|\pi \sigma^n\| \\
&\quad \le \mathcal{C} \Bigg\{ \|\mathcal{P} e^1\| + \|\pi \sigma^1\| + \left\| \mathcal{P} \frac{e^1 - e^0}{k} \right\| \sqrt{k} \\
&\qquad + k^2 \left(\int_0^T \|q_{2,tt}\|_1^2 \, dt \right)^{1/2} + k^{3-2\theta} \left(\int_0^T \|\partial_t^{4-2\theta} u\|^2 \, dt \right)^{1/2} \\
&\qquad + h^{s+1} \left[\left(\sum_{n=1}^{N} \left(\|u_t^n\|_s^2 + \|\mathbf{q}^n\|_{s+1}^2 \right) k \right)^{1/2} + \left(\int_0^T \|\mathbf{q}_t\|_{s+1}^2 \, dt \right)^{1/2} \right] \Bigg\},
\end{aligned}$$

provided that k *and* k/h^2 *are sufficiently small, wherein* $\mathcal{C}$ *depends on the smoothness of* a, c *and* g, *but not on* h *or* k. *Moreover,*

$$\begin{aligned}
&\left(\sum_{n=1}^{N-1} \left\| \frac{e^{n+1} - e^n}{k} \right\|^2 k \right)^{1/2} + \max_{1 \le n \le N} \|e^n\| + \max_{1 \le n \le N} \|\sigma^n\| \\
&\quad \le \mathcal{C} \Bigg\{ \|\mathcal{P} e^1\| + \|\pi \sigma^1\| + \left\| \mathcal{P} \frac{e^1 - e^0}{k} \right\| \sqrt{k} \\
&\qquad + k^2 \left(\int_0^T \|q_{2,tt}\|_1^2 \, dt \right)^{1/2} + k^{3-2\theta} \left(\int_0^T \|\partial_t^{4-2\theta} u\|^2 \, dt \right)^{1/2} \\
&\qquad + h^s \left[\max_{1 \le n \le N} \|u^n\|_s + \max_{1 \le n \le N} \|\mathbf{q}^n\|_s \right. \\
&\qquad \left. + \left(\sum_{n=1}^{N} \|u_t^n\|_s^2 \, k \right)^{1/2} + \left(\int_0^T \left(\|u_t\|_s^2 + \|\mathbf{q}_t\|_s^2 \right) dt \right)^{1/2} \right] \Bigg\}.
\end{aligned}$$

The last estimate follows from the above argument using only H^s-smoothness (i.e. not invoking superconvergence) and adding the projection errors to the right-hand side.

7. Numerical Results

In this section, we present some numerical experiments illustrating the utility of our mixed method for RT_0 and confirming our theoretical results. We test only the Crank–Nicolson procedures. The errors reported are measured in discrete L^2-norms. For the scalar solution u, this is

$$|||e|||_{L^\infty(L^2)} = \max_n \left\{ \sum_\ell (e_\ell^n)^2 h^2 \right\}^{1/2},$$

where $e_\ell^n = u^n - u_h^n$ is the error at the center of grid cell ℓ, wherein u is the exact solution of (4.2)–(4.5) and u_h is the approximation from either the full Crank-Nicholson (C-N) system (5.8)–(5.10), the AD method (5.13)–(5.17), or our AD-M method (5.22)–(5.26). This norm is $\mathcal{O}(h^2)$ close to $\|\mathcal{P}e\|$, and so it should exhibit superconvergence of order $\mathcal{O}(h^2 + k^2)$.

We also report the errors

$$|||e_t|||_{L^2(L^2)} = \left\{ \sum_n \sum_\ell \left(\frac{e_\ell^{n+1} - e_\ell^n}{k} \right)^2 h^2 \, k \right\}^{1/2},$$

and, for the vector solution $\mathbf{q}$,

$$|||\sigma|||_{L^\infty(L^2)} = \max_n \left\{ \sum_i (\sigma_{x,i}^n)^2 h^2 + \sum_j (\sigma_{y,j}^n)^2 h^2 \right\}^{1/2},$$

where $\sigma^n = \mathbf{q}^n - \mathbf{q}_h^n$ are the errors at the center of the cell edges in the x and y directions, respectively, which are $\mathcal{O}(h^2)$ close to $\pi\sigma$. Again, these norms should exhibit superconvergence of order $\mathcal{O}(h^2 + k^2)$.

Note that we should use a scaling of $h \sim k$, since the overall error is $\mathcal{O}(h^2 + k^2)$. However, the condition $k/h^2 \to 0$ is required for the theoretical results. We have been unable to find an example that requires this condition, however. Thus, we use a single discretization parameter n so that $h = k = 1/n$. We note in passing that the condition $k/h^2 \to 0$ is natural for the Backward Euler methods combined with RT_0, since then the superconvergent errors are $\mathcal{O}(h^2 + k)$.

In our AD-M method, for practical purposes, λ^1 was obtained by running 10 micro-time steps of AD using one-tenth of the time step. We use the unit square as Ω and $T = 1.0$. In all our test cases, we choose a specific solution $u(x, y, t)$ and coefficient $a(x, y)$, and then we determine f, g and u_0 so that (4.2)–(4.5) are satisfied.

7.1. *Smooth examples*

In this set of examples, based on those of Ref. 7. In Table 1, we show the results for the exact solution

$$u_+(x, y, t) = \sin(2\pi t) + \sin(2\pi x) + \sin(2\pi y)$$

and $a(x, y) = 1$, for which $f(x, y) = 2\cos(2\pi t)\pi + 4\sin(2\pi x)\pi^2 + 4\sin(2\pi y)\pi^2$. In this example, the AD method does not introduce a larger splitting error, so all three methods are comparable in their errors, at least for large values of n (i.e. small values of $h = k$). We see second-order convergence for the full Crank–Nicolson system for all three norms. Moreover, we see nearly second-order convergence for the two alternating direction methods, but the rate is somewhat degraded to about 1.6 to 1.8 for some of the norms. It appears that in this simple example, the splitting errors actually cancel some of the approximation error, giving less overall error for AD and AD-M than for C-N in some norms for small values of n.

Table 1. Discrete errors with exact solution u_+, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	6.51e–3	1.51e–3	3.79e–4	9.47e–5	2.03
	AD	6.05e–3	1.51e–3	3.79e–4	9.47e–5	2.00
	AD-M	6.09e–3	1.51e–3	3.79e–4	9.47e–5	2.00
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	2.21e–2	6.18e–3	1.68e–3	4.46e–4	1.88
	AD	1.90e–2	5.95e–3	1.68e–3	4.49e–4	1.80
	AD-M	1.10e–2	3.28e–3	1.15e–3	3.66e–4	1.62
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	6.83e–2	1.77e–2	4.58e–3	1.17e–3	1.95
	AD	4.89e–2	1.43e–2	4.27e–3	1.17e–3	1.79
	AD-M	8.44e–2	2.08e–2	4.91e–3	1.18e–3	2.06

Table 2. Discrete errors with exact solution $u_\times$, and the observed convergence rate.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	1.09e–2	2.73e–3	6.81e–4	1.70e–4	2.00
	AD	6.28e–2	1.58e–2	3.94e–3	9.86e–4	2.00
	AD-M	2.70e–2	4.50e–3	8.84e–4	1.94e–4	2.37
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	2.86e–2	7.84e–3	2.12e–3	5.63e–4	1.89
	AD	2.59e–1	6.92e–2	2.27e–2	8.63e–3	1.63
	AD-M	9.75e–2	1.33e–2	2.19e–3	5.10e–4	2.53
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	1.67e–1	4.15e–2	1.03e–2	2.59e–3	2.00
	AD	1.80e–0	6.45e–1	2.70e–1	1.17e–1	1.31
	AD-M	6.40e–1	1.14e–1	2.14e–2	4.20e–3	2.42

In Table 2, we show the results for the exact solution

$$u_\times(x, y, t) = \big(\sin(2\pi t) + 1\big)\big(\sin(2\pi x) + 1\big)\big(\sin(2\pi y) + 1\big)$$

and diffusion coefficient $a(x, y) = 1$. We see second-order convergence for C-N. In this example, the AD method has much more error than C-N. AD produces a large splitting error that degrades the effectiveness of the algorithm. It does not even appear that we have entered the region of asymptotic convergence for this method, since the convergence rates are less than expected for e_t and σ.

On the other hand, the AD-M method produces an error larger but comparable to C-N. The AD-M splitting error is much smaller than that for AD. We observe somewhat better rates of convergence (greater than 2), because the splitting error is being removed at the rate of $\mathcal{O}(k^3)$.

We show the reduction in splitting error in Fig. 1, where we plot the base 10 log of the error $\|\|\|e\|\|\|_{L^\infty(L^2)}$ for the base 10 log of n, where $n = 5$, 10, 20, 40, 80 and 160. The data are for exact solution $u(x, y, t) = \big(\sin(\pi t)+1\big)\big(\sin(\pi x)+1\big)\big(\sin(\pi y)+1\big)$. The graph clearly shows a slope of about 3, i.e. $\mathcal{O}(k^3)$ convergence, for small n. The slope quickly reduces to about 2, i.e. $\mathcal{O}(k^2)$ convergence.

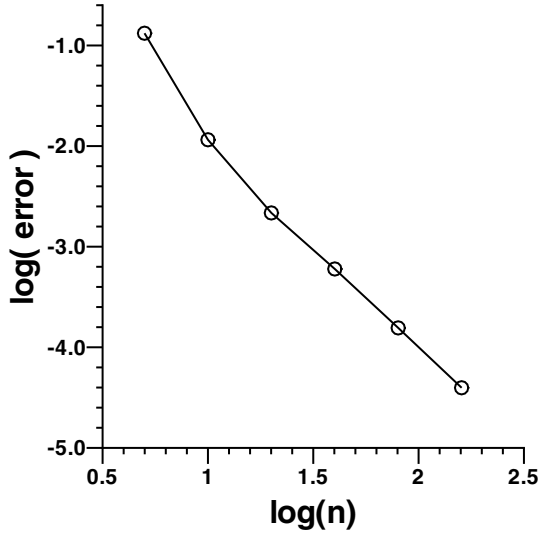

Fig. 1. The log of the error $|||e|||_{L^\infty(L^2)}$ versus the log of n.

Table 3. Discrete L^2-errors for $n = 80$ with exact solution $u_\times$.

Error	Method	$a = a_1$	$a = a_2$	$a = a_3$
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	6.81e–4	6.79e–4	6.85e–4
	AD	3.94e–3	2.43e–3	6.08e–3
	AD-M	8.84e–4	7.77e–4	9.64e–4
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	2.12e–3	1.87e–3	3.59e–3
	AD	2.27e–2	1.49e–2	3.00e–2
	AD-M	2.19e–3	1.91e–3	3.24e–3
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	1.03e–2	6.40e–3	1.74e–2
	AD	2.70e–1	1.37e–1	5.29e–1
	AD-M	2.14e–2	1.17e–2	4.49e–2

In our final smooth experiment, we test our AD-M with a variable coefficient. We again take the exact solution $u_\times$, but set a to one of the three choices

$$a_1(x,y) = 1,$$
$$a_2(x,y) = \frac{1}{2+\cos(3\pi x)\cos(2\pi y)},$$
$$a_3(x,y) = \begin{cases} 1 + 0.5\sin(5\pi x) + y^3, & \text{if } x \le 0.5, \\ \dfrac{1.5}{1+(x-0.5)^2} + y^3, & \text{otherwise.} \end{cases}$$

In Table 3, we present the results, which show again that AD is inferior to AD-M, which is comparable to the full solution.

7.2. *Nonsmooth examples*

In this set of examples, we consider the true solution

$$u_\alpha(x, y, t) = |xyt - 0.25|^\alpha,$$

for some parameter α. Note that the solution has bounded partial derivatives (of any type) only up to order α. According to Theorem 6.1, we should take $\alpha > 3$ to have the regularity demanded of the solution for superconvergence, and $\alpha > 2$ to obtain convergence of order $\mathcal{O}(k + h)$ (since the time error will degenerate to first order).

In the first set of tests, we take $a = 1$, which means that the 1-D operators commute. It is reasonable to expect that there is no requirement that k/h^2 be sufficiently small in this case, assuming the results of Douglas and Kim[7] extend to mixed methods. Indeed we see in Table 4 good convergence for this case when $\alpha = 3.1$, meaning that $\mathbf{q}$ has 2 derivatives.

When $\alpha = 2.1$, C-N behaves as expected in Table 5, which shows the convergence rate of e_t and σ degrading to about $\mathcal{O}(k + h)$. The fact that the error norm of e retains its superconvergence is unexpected. Both AD and AD-M follow the general

Table 4. Discrete errors with exact solution $u_{3.1}$ and $a = 1$, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	1.14e–4	2.85e–5	7.13e–6	1.78e–6	2.00
	AD	1.32e–3	3.41e–4	8.62e–5	2.16e–5	2.00
	AD-M	2.52e–4	4.51e–5	9.10e–6	2.02e–6	2.32
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	1.84e–4	4.63e–5	1.18e–5	2.97e–6	1.98
	AD	1.78e–3	4.63e–4	1.18e–4	3.00e–5	1.97
	AD-M	5.28e–4	1.23e–4	3.12e–5	8.23e–6	2.00
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	8.71e–4	2.17e–4	5.49e–5	1.38e–5	1.99
	AD	5.29e–2	2.02e–2	7.42e–3	2.67e–3	1.44
	AD-M	6.96e–3	1.36e–3	2.49e–4	4.79e–5	2.40

Table 5. Discrete errors with exact solution $u_{2.1}$ and $a = 1$, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	1.04e–4	2.65e–5	9.06e–6	1.89e–6	1.89
	AD	8.78e–4	2.22e–4	5.75e–5	1.38e–5	1.99
	AD-M	1.39e–4	3.02e–5	9.43e–6	2.09e–6	1.99
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	6.90e–4	4.41e–4	2.37e–4	1.08e–4	0.89
	AD	1.02e–3	3.78e–4	1.66e–4	7.10e–5	1.27
	AD-M	7.51e–4	3.70e–4	1.85e–4	8.47e–5	1.04
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	2.04e–3	9.24e–4	4.90e–4	2.07e–4	1.08
	AD	2.94e–2	1.07e–2	3.87e–3	1.39e–3	1.47
	AD-M	2.93e–3	9.24e–4	3.75e–4	1.62e–4	1.38

Table 6. Discrete errors with exact solution $u_{1.9}$ and $a = 1$, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	1.74e–4	5.07e–5	1.73e–5	5.30e–6	1.67
	AD	8.49e–4	2.15e–4	5.17e–5	1.49e–5	1.96
	AD-M	1.84e–4	5.33e–5	1.48e–5	4.98e–6	1.75
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	1.64e–3	1.47e–3	9.17e–4	4.43e–4	0.66
	AD	1.37e–3	9.13e–4	5.50e–4	2.56e–4	0.80
	AD-M	1.36e–3	9.48e–4	5.80e–4	2.77e–4	0.76
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	4.55e–3	2.66e–3	1.68e–3	7.71e–4	0.83
	AD	2.67e–2	9.79e–3	3.55e–3	1.29e–3	1.46
	AD-M	3.73e–3	1.63e–3	8.19e–4	3.84e–4	1.08

Table 7. Discrete errors with exact solution $u_{3.1}$ and nonsmooth a, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	7.04e–5	1.77e–5	4.44e–6	1.11e–6	2.00
	AD	5.11e–3	1.47e–3	3.88e–4	9.88e–5	1.90
	AD-M	8.32e–4	1.11e–4	1.56e–5	2.41e–6	2.81
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	1.91e–4	4.40e–5	1.12e–5	2.93e–6	2.01
	AD	6.97e–3	1.97e–3	5.17e–4	1.33e–4	1.91
	AD-M	1.57e–3	3.27e–4	7.77e–5	1.99e–5	2.10
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	5.03e–3	1.26e–3	3.16e–4	7.92e–5	2.00
	AD	7.89e–1	3.70e–1	1.51e–1	5.76e–2	1.26
	AD-M	1.46e–1	2.87e–2	5.23e–3	9.46e–4	2.43

Table 8. Discrete errors with exact solution $u_{2.1}$ and nonsmooth a, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	7.14e–5	2.34e–5	8.24e–6	1.91e–6	1.72
	AD	3.46e–3	9.15e–4	2.34e–4	5.85e–5	1.96
	AD-M	3.37e–4	5.61e–5	1.18e–5	3.09e–6	2.26
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	9.14e–4	6.44e–4	3.49e–4	1.60e–4	0.84
	AD	3.90e–3	1.12e–3	3.67e–4	1.37e–4	1.61
	AD-M	1.59e–3	6.51e–4	3.09e–4	1.42e–4	1.15
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	5.58e–3	2.41e–3	1.24e–3	5.28e–4	1.12
	AD	4.61e–1	1.92e–1	7.37e–2	2.71e–2	1.36
	AD-M	4.64e–2	1.32e–2	4.47e–3	1.71e–3	1.59

results of C-N, but with somewhat greater error (and AD is worse than AD-M). When $\alpha = 1.9$ (Table 6), we lose sufficient regularity to have full good convergence, but we nevertheless retain a fractional rate of convergence (as we should expect) of order about 0.9 for e_t and σ, and a bit better partial superconvergence rate for e.

Table 9. Discrete errors with exact solution $u_{1.9}$ and nonsmooth a, and the observed convergence rates.

Error	Method	$n = 20$	$n = 40$	$n = 80$	$n = 160$	Rate
$\|\|\|e\|\|\|_{L^\infty(L^2)}$	C-N	1.85e–4	6.77e–5	2.50e–5	6.64e–6	1.58
	AD	3.08e–3	8.02e–4	1.97e–4	5.21e–5	1.97
	AD-M	3.68e–4	7.70e–5	2.35e–5	6.84e–6	1.90
$\|\|\|e_t\|\|\|_{L^2(L^2)}$	C-N	2.15e–3	2.20e–3	1.38e–3	6.66e–4	0.86
	AD	3.77e–3	1.55e–3	8.30e–4	3.89e–4	1.07
	AD-M	2.41e–3	1.42e–3	8.64e–4	4.24e–4	0.82
$\|\|\|\sigma\|\|\|_{L^\infty(L^2)}$	C-N	1.15e–2	6.72e–3	4.17e–3	1.93e–3	0.84
	AD	3.86e–1	1.57e–1	5.97e–2	2.19e–2	1.38
	AD-M	5.41e–2	1.49e–2	7.24e–3	3.60e–3	1.03

In the second set of nonsmooth tests, we also take a nonsmooth a given by

$$a(x,y) = \begin{cases} 2 + \sin(xy^2) + 32(x-0.5)(y-0.5), & x \le 0.5, y \le 0.5, \\ 2 + \sin(xy^2) + 8(x-0.5)(y-0.5), & x > 0.5, y > 0.5, \\ 2 + \sin(xy^2), & \text{otherwise.} \end{cases}$$

In this case, the 1-D operators do *not* commute. Nevertheless, we do not seem to require that k/h^2 be sufficiently small (contrary to what Theorem 6.1 suggests). The results for $\alpha = 3.1$, 2.1 and 1.9 are given in Tables 7–9, and they agree qualitatively with the previous test cases.

8. Conclusions

We have shown that the AD and AD-M algorithms for finite difference and Galerkin approximations to second order parabolic equations converge optimally if only $k/h^2 \to 0$ (not $k/h^4 \to 0$).

We have shown that the AD-M modification in Ref. 7 can be applied to mixed finite element procedures. Moreover, we formulated an efficient Uzawa AD-M implementation. The resulting method has splitting error of size $\mathcal{O}(k^3)$. For RT_s, the Uzawa AD and AD-M converge optimally at the rate $\mathcal{O}(k^r + h^{s+1})$ provided that $k/h^2 \to 0$, where $r = 1$ for backward Euler and $r = 2$ for Crank–Nicolson time discretization. Moreover, both methods exhibit superconvergence. In discrete norms, the scalar and vector variables converge with order $\mathcal{O}(k^r + h^{s+2})$.

Numerical results using Crank–Nicolson and RT_0 show that $k \sim h$ works well for AD and AD-M, suggesting that the condition $k/h^2 \to 0$ is not actually needed (though we cannot prove this now). The numerical results also clearly show that the splitting error is higher order, and was seen to be $\mathcal{O}(k^3)$. Generally, we saw that the AD-M error was comparable to C-N, but AD had more error.

It is clear from the algorithms that the AD-M modification requires little extra computation compared to AD, but it can lead to a significant reduction in the splitting perturbation error associated with the AD method for mixed finite elements. Moreover, AD-M is much easier to solve than C-N alone, but often produces comparable error.

Acknowledgments

This work was partially supported by the National Science Council of Taiwan and National Center for High-Performance Computing, Taiwan. The work of the first author was supported in part by U.S. National Science Foundation grant DMS-0417431.

References

1. S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods* (Springer-Verlag, 1994).
2. F. Brezzi, J. Douglas, Jr., M. Fortin and L. D. Marini, Efficient rectangular mixed finite elements in two and three space variables, *RAIRO Mod. Math. Anal. Numér.* **21** (1987) 581–604.
3. F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer-Verlag, 1991).
4. J. Douglas, Jr., On the numerical integration of $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial t}$ by implicit methods, *J. Soc. Indust. Appl. Math.* **3** (1955) 42–65.
5. J. Douglas, Jr. and T. Dupont, Alternating-direction Galerkin methods on rectangles, in *Numerical Solution of Partial Differential Equations*, *II* SYNSPADE 1970, (Academic Press, 1971), pp. 133–214.
6. J. Douglas, Jr. and J. Gunn, A general formulation of alternating direction methods: Part I. Parabolic and hyperbolic problems, *Numer. Math.* **6** (1964) 428–453.
7. J. Douglas, Jr. and S. Kim, Improved accuracy for locally one-dimensional methods for parabolic equations, *Math. Mod. Meth. Appl. Sci.* **11** (2001) 1563–1579.
8. J. Douglas, Jr. and D. W. Peaceman, Numerical solution of two-dimensional heat flow problems, *Amer. Inst. Chem. Eng. J.* **1** (1955) 505–512.
9. J. Douglas, Jr. and P. Pietra, A description of some alternating–direction iterative techniques for mixed finite element methods, in *Mathematical and Computational Methods in Seismic Exploration and Reservoir Modeling*, ed. W. E. Fitzgibbon, (SIAM, 1986), pp. 37–53.
10. J. Douglas, Jr. and J. Wang, Superconvergence for mixed finite element methods on rectangular domains, *Calcolo* **26** (1989) 121–134.
11. R. Durán, Superconvergence for rectangular mixed finite elements, *Numer. Math.* **58** (1990) 287–298.
12. M. Nakata, A. Weiser and M. F. Wheeler, Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains, in *The Mathematics of Finite Elements and Applications V*, ed. J. R. Whiteman (Academic Press, 1985).
13. D. W. Peaceman and H. Rachford, The numerical solution of parabolic and elliptic equations, *J. Soc. Indust. Appl. Math.* **3** (1955) 28–41.

14. R. A. Raviart and J. M. Thomas, A mixed finite element method for 2nd order elliptic problems, in *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Math. Vol. 606, eds. I. Galligani and E. Magenes (Springer-Verlag, 1977), pp. 292–315.
15. 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.