Chapter 4

Error Analysis

4.1 Mixed finite element method for the flow equation

First let us consider the weak solution (3.10) and (3.11) of the flow equations. It is known that (3.10)-(3.11) has a unique solution (see, e.g., [20]). Assume that, for the permeability function, we have

$$0 < K_{\text{min}} = \kappa(\phi_0) \leq \kappa(\phi) \leq \kappa(\phi_f) = K_{\text{max}} < \infty. \tag{4.1}$$

We also assume that $a(\phi)$, f_1, and f_2 are Lipschitz continuous functions in c and ϕ, i.e. for $i = 1, 2$ we have

$$\|a(\phi_1) - a(\phi_2)\| \leq K\|\phi_1 - \phi_2\|, \tag{4.2}$$

$$\|f_i(c_1, \phi_1) - f_i(c_2, \phi_2)\| \leq K(\|c_1 - c_2\| + \|\phi_1 - \phi_2\|). \tag{4.3}$$

Note that (4.2) holds for $\kappa(\phi)$ defined by (2.1) and (4.3) is satisfied by (2.6). If we are given a concentration approximation C and a porosity approximation Φ at a time

$t \in J$, then the mixed method for pressure and velocity consists of $\mathbf{U} \in \mathbf{V}_h^g$, and $P \in W_h$ such that

$$\begin{aligned} (a(\Phi)\mathbf{U}, \mathbf{v}) &= (P, \nabla \cdot \mathbf{v}) - < g_2, \mathbf{v} \cdot \nu >_{\Gamma_2}, \qquad \mathbf{v} \in \mathbf{V}_h^0, && (4.4) \\ (\nabla \cdot \mathbf{U}, w) &= (-f_1(C, \Phi), w), \qquad w \in W_h. && (4.5) \end{aligned}$$

Existence and uniqueness of $\mathbf{U}$ and P is proved in [20]. Using techniques from [21] and [20], we define a projection of the exact solution into the finite element space. Define the map $(\tilde{\mathbf{U}}, \tilde{P}) : J \to \mathbf{V}_h \times W_h$ by:

$$\begin{aligned} (a(\phi)\tilde{\mathbf{U}}(t), \mathbf{v}) &= (\tilde{P}(t), \nabla \cdot \mathbf{v}) - < g_2, \mathbf{v} \cdot \nu >_{\Gamma_2}, \qquad \mathbf{v} \in \mathbf{V}_h, && (4.6) \\ (\nabla \cdot \tilde{\mathbf{U}}(t), w) &= (-f_1(c(t), \phi), w), \qquad w \in W_h, && (4.7) \end{aligned}$$

where $c(t)$ is the exact solution. In [9] and [20] this map is shown to exist. By Theorem 2.1 of Brezzi [9] and (A_p) we get

$$\begin{aligned} \|\mathbf{u} - \tilde{\mathbf{U}}\|_{H(\mathrm{div})} &+ \|p - \tilde{P}\| \\ &\leq K\left(\inf_{\mathbf{v} \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}\|_{H(\mathrm{div})} + \inf_{w \in W_h} \|p - w\|\right) \\ &\leq K(\|\mathbf{u}\|_1 + \|\nabla \cdot \mathbf{u}\|_1 + \|p\|_1)h && (4.8) \end{aligned}$$

where the constant K depends on the constants in (4.1), but is independent of $h, \mathbf{u}, p$ and c. We next estimate the difference between the discrete solution $(\mathbf{U}, P)$ and the projection $(\tilde{\mathbf{U}}, \tilde{P})$.

Lemma 1 *Given a concentration approximation C and a porosity approximation Φ at a time $t \in J$, the mixed method solution $(\mathbf{U}, P)$ of (4.4)-(4.5) satisfies:*

$$\|P - \tilde{P}\| + \|\mathbf{U} - \tilde{\mathbf{U}}\|_{H(\mathrm{div})} \leq K(\|\Phi - \phi\| + \|C - c\|) \tag{4.9}$$

for some constant K dependent on $\|\tilde{\mathbf{U}}\|_\infty$, where $\tilde{\mathbf{U}}$, and $\tilde{P}$ defined in (4.6) and (4.7) above.

Proof: After subtracting (4.6),(4.7) from (4.4),(4.5) we get

$$\begin{aligned} (a(\Phi)\mathbf{U} - a(\phi)\tilde{\mathbf{U}}, \mathbf{v}) &= (P - \tilde{P}, \nabla \cdot \mathbf{v}), \qquad \mathbf{v} \in \mathbf{V}_h^0, & (4.10) \\ (\nabla \cdot (\mathbf{U} - \tilde{\mathbf{U}}), w) &= (-f_1(C, \Phi) + f_1(c, \phi), w), \quad w \in W_h. & (4.11) \end{aligned}$$

Set $\mathbf{v} = \mathbf{U} - \tilde{\mathbf{U}}$ and $w = P - \tilde{P}$ and add (4.10) and (4.11) to obtain

$$(a(\Phi)\mathbf{U} - a(\phi)\tilde{\mathbf{U}}, \mathbf{U} - \tilde{\mathbf{U}}) = (-f_1(C, \Phi) + f_1(c, \phi), P - \tilde{P}). \tag{4.12}$$

Adding and subtracting $a(\Phi)\tilde{\mathbf{U}}$ to the first term in the LHS of (4.12) gives

$$\begin{aligned} (a(\Phi)\mathbf{U} - a(\phi)\tilde{\mathbf{U}}, \mathbf{U} - \tilde{\mathbf{U}}) &= (a(\Phi)(\mathbf{U} - \tilde{\mathbf{U}}), \mathbf{U} - \tilde{\mathbf{U}}) \\ &+ ((a(\Phi) - a(\phi))\tilde{\mathbf{U}}, \mathbf{U} - \tilde{\mathbf{U}}). & (4.13) \end{aligned}$$

Then (4.12) becomes

$$\begin{aligned} (a(\Phi)(\mathbf{U} - \tilde{\mathbf{U}}), \mathbf{U} - \tilde{\mathbf{U}}) &= (-f_1(C, \Phi) + f_1(c, \phi), P - \tilde{P}) \\ &- ((a(\Phi) - a(\phi))\tilde{\mathbf{U}}, \mathbf{U} - \tilde{\mathbf{U}}). & (4.14) \end{aligned}$$

By (4.1) we have

$$(a(\Phi)(\mathbf{U}-\tilde{\mathbf{U}}), \mathbf{U}-\tilde{\mathbf{U}}) \geq K\|\mathbf{U}-\tilde{\mathbf{U}}\|^2. \tag{4.15}$$

Moreover, using (4.2)

$$\begin{aligned}((a(\Phi)-a(\phi))\tilde{\mathbf{U}}, \mathbf{U}-\tilde{\mathbf{U}}) &\leq K\|\Phi-\phi\|\ \|\tilde{\mathbf{U}}\|_\infty\|\mathbf{U}-\tilde{\mathbf{U}}\| \\ &\leq K(\epsilon\|\mathbf{U}-\tilde{\mathbf{U}}\|^2+\frac{1}{\epsilon}\|\Phi-\phi\|^2\|\tilde{\mathbf{U}}\|_\infty^2). \end{aligned} \tag{4.16}$$

Using (4.3) and Schwartz inequality we get

$$\begin{aligned}(-f_1(C,\Phi)+f_1(c,\phi), P-\tilde{P}) &\leq \|-f_1(C,\Phi)+f_1(c,\phi)\|\|P-\tilde{P}\| \\ &\leq K(\|C-c\|+\|\Phi-\phi\|)\|P-\tilde{P}\| \\ &\leq K(\frac{1}{\epsilon}(\|C-c\|^2+\|\Phi-\phi\|^2) \\ &+ \epsilon\|P-\tilde{P}\|^2). \end{aligned} \tag{4.17}$$

To estimate the last term in (4.17), consider the auxiliary problem:

$$\begin{aligned}\nabla\cdot(\nabla\varphi) &= P-\tilde{P} && \text{in } \Omega \\ \varphi &= 0 && \text{on } \Gamma_2 \\ \nabla\varphi\cdot\nu &= 0 && \text{on } \Gamma_1\cup\Gamma_3. \end{aligned}$$

Clearly by the elliptic regularity [10] we have

$$\|\varphi\|_2 \leq K\|P-\tilde{P}\|. \tag{4.18}$$

Set $\psi=\nabla\varphi$ and let $\Pi\psi\in\mathbf{V}_h$ be the interpolant of ψ in RT0 [10] satisfying

$$(\nabla\cdot(\Pi\psi-\psi), w)=0 \qquad \forall w\in W_h. \tag{4.19}$$

The interpolation Π satisfies the following approximation property [10]:

$$\|\Pi\psi - \psi\| \leq Kh\|\psi\|_1. \tag{4.20}$$

Then by (4.10) and (4.19) with $\mathbf{v} = \Pi\psi \ \in \mathbf{V}_h^0$ we have

$$\begin{aligned} \|P - \tilde{P}\|^2 &= (P - \tilde{P}, \nabla \cdot \psi) \\ &= (P - \tilde{P}, \nabla \cdot \Pi\psi) \\ &= (a(\Phi)\mathbf{U} - a(\phi)\tilde{\mathbf{U}}, \Pi\psi). \end{aligned}$$

Similarly to (4.12) - (4.17) we get

$$\|P - \tilde{P}\|^2 \leq K(\|\Phi - \phi\|\|\tilde{\mathbf{U}}\|_\infty + \|\mathbf{U} - \tilde{\mathbf{U}}\|)\|\Pi\psi\|.$$

Now,

$$\|\Pi\psi\| = \|\Pi\psi - \psi + \psi\| \leq \|\Pi\psi - \psi\| + \|\psi\|.$$

Since, by (4.18) and (4.20),

$$\|\Pi\psi - \psi\| \leq Kh\|\psi\|_1 \leq K\|P - \tilde{P}\|,$$

and

$$\|\psi\| = \|\nabla\varphi\| \leq K\|\varphi\|_1 \leq K\|P - \tilde{P}\|,$$

then

$$\|P - \tilde{P}\| \leq K(\|\Phi - \phi\| + \|\mathbf{U} - \tilde{\mathbf{U}}\|). \tag{4.23}$$

Now (4.17) implies

$$(-f_1(C,\Phi)+f_1(c,\phi),P-\tilde{P}) \;\leq\; K[\frac{1}{\epsilon}(\|C-c\|^2+\|\Phi-\phi\|^2) + \epsilon(\|\Phi-\phi\|^2+\|\mathbf{U}-\tilde{\mathbf{U}}\|^2)]. \tag{4.24}$$

Substituting (4.15), (4.16) and (4.24) into (4.14) we get

$$\|\mathbf{U}-\tilde{\mathbf{U}}\|^2 \leq K[\frac{1}{\epsilon}(\|C-c\|^2 + \|\Phi-\phi\|^2)+\epsilon(\|\Phi-\phi\|^2+\|\mathbf{U}-\tilde{\mathbf{U}}\|^2)]. \tag{4.25}$$

Taking ϵ small enough we get

$$\|\mathbf{U}-\tilde{\mathbf{U}}\| \leq K(\|\Phi-\phi\|+\|C-c\|). \tag{4.26}$$

Furthermore, set now $w=\nabla\cdot(\mathbf{U}-\tilde{\mathbf{U}}) \in W_h$ in (4.11) to get

$$(\nabla\cdot(\mathbf{U}-\tilde{\mathbf{U}}),\nabla\cdot(\mathbf{U}-\tilde{\mathbf{U}})) \;=\; (-f_1(C,\Phi)+f_1(c,\phi),\nabla\cdot(\mathbf{U}-\tilde{\mathbf{U}})),$$

which implies, using (4.3) and Schwartz inequality,

$$\|\nabla\cdot(\mathbf{U}-\tilde{\mathbf{U}})\| \leq K(\|C-c\|+\|\Phi-\phi\|). \tag{4.27}$$

A combination of (4.23), (4.26), and (4.27) proves the lemma. ■

The following result follows from (4.8) and Lemma 1

Theorem 1 *For a given approximation C and Φ at some $t\in J$, the solution $(\mathbf{U},P)$ of (4.4)-(4.5) satisfies*

$$\|p-P\|+\|\mathbf{u}-\mathbf{U}\|_{H(\mathrm{div})} \leq K(h+\|\Phi-\phi\|+\|C-c\|).$$

We now proceed with bounding $\|C-c\|$ and $\|\Phi-\phi\|$.

4.2 Transport equation

Consider the discretized transport equations (3.20)-(3.21). For $n = 0$, set $C^0(x) = \pi_2 c^0(x)$ where $\pi_2 c^0(x)$ is the L^2-projection of c^0; that is

$$(C^0, w) = (c^0, w), \qquad\qquad w \in W_h. \tag{4.28}$$

Using techniques from [17] we will compare the approximation solution to an elliptic projection. Let $\overline{C\Phi}(\cdot, t) \in W_h, \bar{\mathbf{Q}}(\cdot, t) \in \bar{\mathbf{V}}_h$ denote the mixed method solutions (unsplit solutions) to the elliptic problem associated with (2.4); that is, for each $t \in [0, T]$,

$$\begin{aligned}
(D^{-1}\bar{\mathbf{Q}}(.,t), \mathbf{v}) \;&-\; (\overline{C\Phi}(.,t), \nabla\cdot\mathbf{v}) =< c_B, \mathbf{v}\cdot\nu >_{\Gamma_1} \qquad \mathbf{v} \in \bar{\mathbf{V}}_h && (4.29)\\
(\nabla\cdot\bar{\mathbf{Q}}(.,t), w) \;&=\; (\nabla\cdot\mathbf{q}(.,t), w)\\
&=\; -((c\phi)_t, w) - (\nabla\cdot\mathbf{F}(.,t), w) + (f_2(.,t), w), \quad w \in W_h. && (4.30)
\end{aligned}$$

From the standard mixed method theory [34, 10],

$$\|c\phi - \overline{C\Phi}\| \le K\|c\phi\|_1 h. \tag{4.31}$$

Let the data and the solution pair $(c, \mathbf{q})$ be sufficiently smooth and let $e = \phi - \Phi$, $\xi = C\Phi - \overline{C\Phi}$, $\theta = C - \bar{C}$, and $\eta = \mathbf{Q} - \bar{\mathbf{Q}}$. Let $\tilde{\mathbf{F}} = \Pi\,\mathbf{F} \in V_h$. By (4.19) $\tilde{\mathbf{F}}$ satisfies

$$(\nabla\cdot(\tilde{\mathbf{F}} - \mathbf{F})(.,t), w) = 0, \qquad \forall\, w \in W_h. \tag{4.32}$$

Moreover, (see [34, 10])

$$\int_e (\tilde{\mathbf{F}} - \mathbf{F})\cdot\nu d\sigma = 0$$

for any edge (face) e of an element $\mathcal{T}$. Assume that $\mathbf{G}^{n-1}$ satisfies the following approximation result

$$\|\mathbf{G}^{n-1} - \tilde{\mathbf{F}}^{n-1}\| \leq K[\|\xi^{n-1}\| + \|\mathbf{u}^{n-1} - U^{n-1}\| + h^2 + \Delta t].$$

A similar bound without the term $\|\mathbf{u}^{n-1} - U^{n-1}\|$ is shown in [17]. This additional term appears here due to dependence of $\mathbf{F}$ on $\mathbf{u}$. Using Theorem 1 we get

$$\|\mathbf{G}^{n-1} - \tilde{\mathbf{F}}^{n-1}\| \leq K[\|\xi^{n-1}\| + \|e^{n-1}\| + \|\theta^{n-1}\| + h + \Delta t]. \tag{4.33}$$

Assume also that $0 < D^{-1} < d^*$, where d^* is a positive constant. Using the above assumptions and notation, we now state an error estimate for the method described above.

Theorem 2 *For Δt sufficiently small, we have*

$$\max_n \|\xi^n\| \quad + \quad \left(\sum_{n=0}^{N} \|(D^n)^{-\frac{1}{2}} \eta^n\|^2 \Delta t_n \right)^{\frac{1}{2}} \leq K(d^*)(h + \Delta t)$$

where $K(d^*)$ is a constant which depends on d^* but not h or Δt.

Proof: Subtracting (4.29) from (3.20) we get

$$((D^{-1}\eta)^n, \mathbf{v}) - (\xi^n, \nabla \cdot \mathbf{v}) = 0, \tag{4.34}$$

and subtracting (4.30) from (3.21) and using (4.32) we get

$$\begin{aligned} (\partial_t (C\Phi)^n, w) + (\nabla \cdot \eta^n, w) \quad &= \quad ((c\phi)_t^n, w) - (\nabla \cdot \mathbf{G}^{n-1}, w) \\ &+ \quad (\nabla \cdot \tilde{\mathbf{F}}^n, w) + (\tilde{f}_2^n - f_2^n, w). \end{aligned}$$

Subtract $(\partial_t((\overline{C\Phi})^n), w)$ from both sides to get

$$\begin{aligned} (\partial_t \xi^n, w) &+ (\nabla \cdot \eta^n, w) = -(\partial_t (\overline{C\Phi})^n, w) + ((c\phi)_t^n, w) \\ &+ (\nabla \cdot (\tilde{\mathbf{F}}^n - \mathbf{G}^{n-1}), w) + (\tilde{f}_2^n - f_2^n, w). \end{aligned} \tag{4.35}$$

We set $v = \eta^n$,$w = \xi^n$ and add (4.34) and (4.35) to obtain

$$\begin{aligned} (\partial_t \xi^n, \xi^n) &+ ((D^{-1}\eta)^n, \eta^n) = -(\partial_t (\overline{C\Phi})^n, \xi^n) + ((c\phi)_t^n, \xi^n) \\ &+ (\nabla \cdot (\tilde{\mathbf{F}}^n - \mathbf{G}^{n-1}), \xi^n) + (\tilde{f}_2^n - f_2^n, \xi^n). \end{aligned} \tag{4.36}$$

We now estimate the right-hand side of (4.36) term-by-term. First, by (4.20), (4.34), and (4.33) we have

$$\begin{aligned} (\nabla \cdot (\tilde{\mathbf{F}}^n - \mathbf{G}^{n-1}), \xi^n) &= ((D^n)^{-1}\eta^n, \tilde{\mathbf{F}}^n - \mathbf{G}^{n-1}) \\ &\leq \frac{1}{2}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2 + 2\|(D^n)^{-\frac{1}{2}}\tilde{\mathbf{F}}^n - \mathbf{G}^{n-1}\|^2 \\ &\leq \frac{1}{2}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2 + K(d^*)\|\tilde{\mathbf{F}}^n - \tilde{\mathbf{F}}^{n-1}\|^2 \\ &+ K(d^*)\|\tilde{\mathbf{F}}^{n-1} - \mathbf{G}^{n-1}\|^2 \\ &\leq \frac{1}{2}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2 + K(d^*)(\|\tilde{\mathbf{F}}^n - \mathbf{F}^n\|^2 \\ &+ \|\mathbf{F}^n - \mathbf{F}^{n-1}\|^2 + \|\mathbf{F}^{n-1} - \tilde{\mathbf{F}}^{n-1}\|^2) \\ &+ K(d^*)\|\tilde{\mathbf{F}}^{n-1} - \mathbf{G}^{n-1}\|^2 \\ &\leq \frac{1}{2}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2 + K(d^*)(\|\xi^{n-1}\|^2 \\ &+ \|e^{n-1}\| + \|\theta^{n-1}\|^2 + h^2 + \Delta t^2), \end{aligned} \tag{4.37}$$

where $K(d^*)$ is a constant which is dependent on d^* but independent of h and Δt. Next,

$$(-\partial_t(\overline{C\Phi})^n + (c\phi)_t^n, \xi^n) \leq \| - \partial_t(\overline{C\Phi})^n + (c\phi)_t^n\| \, \|\xi^n\|.$$

We have

$$\| - \partial_t(\overline{C\Phi})^n + (c\phi)_t^n\| \leq \| - (\overline{C\Phi})^n)_t + (c\phi)_t^n\| + K\Delta t.$$

Now, from (4.31) and

$$\|(c\phi)_t - (\overline{C\Phi})_t\| \quad \leq \quad Kh,$$

which can be obtained by differencing (4.29)-(4.30) in time, see [35], we get

$$(-\partial_t(\overline{C\Phi})^n + (c\phi)_t^n, \xi^n) \leq K(h^2 + \Delta t^2 + \|\xi^n\|^2). \tag{4.38}$$

Finally, by (4.3) we have

$$(\tilde{f}_2^n - f_2^n, \xi^n) \leq K(\|\theta^n\|^2 + \|e^n\|^2 + \|\xi^n\|^2). \tag{4.39}$$

Substituting (4.37),(4.38), and (4.39) into (4.36) and collecting terms we get

$$\begin{aligned}
(\partial_t(\xi^n, \xi^n) \quad &+ \quad \frac{1}{2}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2 \leq \\
&\leq \quad K(d^*)(\|\xi^{n-1}\|^2 + \|\theta^{n-1}\|^2 + h^2 \\
&+ \quad \Delta t^2 + \|e^n\|^2 + \|\theta^n\|^2 + \|\xi^n\|^2).
\end{aligned} \tag{4.40}$$

Using

$$\begin{aligned}
(\partial_t(\xi^n, \xi^n) \quad &= \quad \frac{1}{2\Delta t_n}(\|\xi^n\|^2 - \|\xi^{n-1}\|^2 + \|\xi^n - \xi^{n-1}\|^2) \\
&\geq \quad \frac{1}{2\Delta t_n}(\|\xi^n\|^2 - \|\xi^{n-1}\|^2)
\end{aligned}$$

and multiplying by $2\Delta t_n$ we get

$$\begin{aligned}\|\xi^n\|^2 - \|\xi^{n-1}\|^2 &+ \|(D^n)^{-\frac{1}{2}}\eta^n\|^2\Delta t_n \leq \\ &\leq K^*(d^*)\Delta t_n(\|\xi^{n-1}\|^2 + \|\theta^{n-1}\|^2 + h^2 \\ &+ \Delta t^2 + \|e^n\|^2 + \|\theta^n\|^2 + \|\xi^n\|^2).\end{aligned}$$

Summing on n

$$\begin{aligned}\|\xi^N\|^2 + \|\xi^0\| &+ \sum_{n=1}^{N}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2\Delta t_n \leq \\ &\leq K(d^*)\sum_{n=1}^{N}\Delta t_n(\|\xi^{n-1}\|^2 + \|\xi^n\|^2 + h^2 \\ &+ \Delta t^2 + \|e^n\|^2 + \|\theta^n\|^2).\end{aligned}$$

Using that

$$\|\theta^n\| \leq K(\|e^n\| + \|\xi^n\|), \tag{4.41}$$

and (4.49) we obtain

$$\begin{aligned}\|\xi^N\|^2 &+ \sum_{n=1}^{N}\|(D^n)^{-\frac{1}{2}}\eta^n\|^2\Delta t_n \leq & (4.42)\\ &\leq K(d^*)\sum_{n=1}^{N}\Delta t_n(\|\xi^{n-1}\|^2 + \|\xi^n\|^2 + h^2 & (4.43)\\ &+ \Delta t^2 + \Delta t^2\sum_{m=0}^{N-1}\|\xi^m\|^2). & (4.44)\end{aligned}$$

We need to bound

$$A = \sum_{n=1}^{N}\Delta t_n\sum_{m=0}^{n-1}\Delta t^2\|\xi^m\|.$$

let $\mathcal{F}(n) = \sum_{m=0}^{n-1} \Delta t^2 \|\xi^m\|^2$ which is an increasing function. Now

$$\begin{aligned} A &= \sum_{n=1}^{N} \Delta t_n \mathcal{F}(n) \\ &\leq \mathcal{F}(N) \sum_{n=1}^{N} \Delta t_n \\ &= T \sum_{m=0}^{N-1} \Delta t^2 \|\xi^m\|^2. \end{aligned} \tag{4.45}$$

Substituting (4.45) into (4.42) and applying the discrete analogue of Gronwall's Lemma [22] with Δt sufficiently small gives

$$\max_n \|\xi^n\| \quad + \quad \left(\sum_{n=0}^{N} \|((D)^n)^{-\frac{1}{2}} \eta^n\|^2 \Delta t \right)^{\frac{1}{2}} \leq K(d^*) \left(h + \Delta t \right). \qquad \blacksquare$$

4.3 Porosity update

Writing the porosity update equation as

$$\phi_t = f_1(\phi, c) \tag{4.46}$$

and now from the Taylor series expansion we have:

$$\begin{aligned} \phi^{m+1} &= \phi^m + \Delta t \phi_t^m + O(\Delta t^2) \\ &= \phi^m + \Delta t f_1(\phi^m, c^m) + O(\Delta t^2). \end{aligned} \tag{4.47}$$

Suppose we use Euler's method to solve (4.46), then we have

$$\Phi^{m+1} = \Phi^m + \Delta t f_1(\Phi^m, C^m). \tag{4.48}$$

Subtracting (4.48) from (4.47), and letting $e^m = \phi^m - \Phi^m$ we get

$$e^{m+1} = e^m + \Delta t(f_1(\phi^m, c^m) - f_1(\Phi^m, C^m)) + O(\Delta t^2)$$

taking norms

$$\|e^{m+1}\| - \|e^m\| \leq \Delta t \|f_1(\phi^m, c^m) - f_1(\Phi^m, C^m)\|$$

and since by (4.3)

$$\|f_1(\phi^m, c^m) - f_1(\Phi^m, C^m)\| \quad \leq \quad K(\|\phi^m - \Phi^m\| + \|c^m - C^m\|),$$

then we have

$$\|e^{m+1}\| - \|e^m\| \leq K\Delta t(\|e^m\| + \|\theta^m\|).$$

using (4.41) and summing on m

$$\|e^N\| - \|e^0\| \leq K\Delta t \sum_{m=0}^{N-1}(\|e^m\| + \|\xi^m\|),$$

and by the discrete form of Gronwall's Lemma we get

$$\|e^N\| = \|\phi^N - \Phi^n\| \leq K\Delta t \sum_{m=0}^{N-1} \|\xi^m\| + Kh\Delta t \tag{4.49}$$

A combination of (4.49), (4.41), Theorem 1, and Theorem 2 gives

Theorem 3

$$\max_n(\|e^n\| + \|\theta^n\| + \|p^n - P^n\| + \|\mathbf{u}^n - U^n\|_{H(\mathrm{div})}) \leq K(h + \Delta t).$$

4.4 Operator splitting

In this section we bound the error due to operator splitting in the transport equation. Operator splitting, or fractional steps, methods were developed by N.N. Yanenko [38] and his collaborators for solving problems in theoretical mechanics numerically. At the present time these methods are widely used in problems of various kinds.

Theoretically the success of splitting is primarily determined by the splitting error. Therefore, we investigate the accuracy of splitting into three nonlinear operators in an abstract Cauchy problem and show that it is of first order in time for general evolutionary equations. We also obtain a general formula for the leading term. This is an extension of the work of Bobylev and Ohwada [8], where only two nonlinear operators were considered

Consider the initial value problem of the form

$$\begin{aligned} \frac{\partial}{\partial t}\mathcal{U}(\xi,t) &= L(\mathcal{U}), \quad \xi \in R^d,\ t>0 \\ \mathcal{U}(\xi,0) &= \mathcal{U}_0(\xi), \end{aligned} \tag{4.50}$$

where $L = A + B + C$ and where A, B, and C are nonlinear operators acting from a Banach space V to V, $\mathcal{U} \in V$. The above equation is split into three equations:

$$\frac{dx}{dt} = A(x), \qquad x(0) = \mathcal{U}_0, \qquad 0 \le t \le \Delta t$$

$$\frac{dy}{dt} = B(y), \qquad y(0) = x(\Delta t), \qquad 0 \le t \le \Delta t$$

$$\frac{dz}{dt} = C(z), \qquad z(0) = y(\Delta t), \qquad 0 \le t \le \Delta t$$

and the approximate solution is

$$\mathcal{U}(., \Delta t) \approx z(\Delta t).$$

For brevity, we express the solution of an abstract Cauchy problem

$$\frac{dU}{dt} = L(U), \quad U|_{t=0} = U_0 \in V,$$

as

$$U(t) = S^t_{A+B+C}(U_0).$$

Then the above splitting is an approximation of the operator $S^{\Delta t}_L$:

$$S^{\Delta t}_{A+B+C} \approx S^{\Delta t}_C S^{\Delta t}_B S^{\Delta t}_A. \tag{4.51}$$

Theorem 4 *Assume that $A, B,$ and C have continuous second derivatives. Then the splitting error is*

$$\begin{aligned} S^{\Delta t}_{A+B+C} - S^{\Delta t}_C S^{\Delta t}_B S^{\Delta t}_A &= \frac{\Delta t^2}{2}[A'B + A'C + B'C \\ &- B'A - C'A - C'B] + O(\Delta t^3) \end{aligned} \tag{4.52}$$

and if $A, B,$ and C are linear then the error is

$$\frac{\Delta t^2}{2}[AB + AC + BC - BA - CA - CB]. \tag{4.53}$$

Proof: by the assumption of the theorem $S^{\Delta t}_{A+B+C}$ can be expressed as:

$$\begin{aligned} S^{\Delta t}_{A+B+C}(U_0) &= U_0 + \Delta t[A(U_0) + B(U_0) + C(U_0)] \\ &+ \frac{\Delta t^2}{2}[A'_{U_0} + B'_{U_0} + C'_{U_0}][A(U_0) + B(U_0) + C(U_0)] + O(\Delta t^3), \end{aligned}$$

also,

$$\begin{aligned} x(\Delta t) &= S^{\Delta t}_A(U_0) = U_0 + \Delta t A(U_0) + \frac{\Delta t^2}{2} A'_{U_0} A(U_0) + O(\Delta t^3), \\ y(\Delta t) &= S^{\Delta t}_B(x(\Delta t)) = x(\Delta t) + \Delta t B(x(\Delta t)) + \frac{\Delta t^2}{2} B'_{x(\Delta t)} B(x(\Delta t)) + O(\Delta t^3), \\ z(\Delta t) &= S^{\Delta t}_C(y(\Delta t)) = y(\Delta t) + \Delta t C(y(\Delta t)) + \frac{\Delta t^2}{2} C'_{y(\Delta t)} C(y(\Delta t)) + O(\Delta t^3). \end{aligned}$$

Noting that

$$\begin{aligned} B[x(\Delta t)] &= B[U_0 + \Delta t A(U_0) + ...] = B(U_0) + \Delta t B'_{U_0} A(U_0) + O(\Delta t^2), \\ B'_{x(\Delta t)} B[x(\Delta t)] &= B'_{U_0} B(U_0) + \Delta t[B''_{U_0} + B'_{U_0} B'_{U_0} B(U_0)] + O(\Delta t), \end{aligned}$$

we have

$$\begin{aligned} S^{\Delta t}_B(S^{\Delta t}_A(U_0)) &= U_0 + \Delta t[A(U_0) + B(U_0)] \\ &+ \frac{\Delta t^2}{2}[A'_{U_0} A(U_0) + B'_{U_0} B(U_0) + 2B'_{U_0} A(U_0)] + O(\Delta t^3). \end{aligned}$$

Also note

$$\begin{aligned} C[y(\Delta t)] &= C[U_0 + \Delta t[A(U_0) + B(U_0)] + O(\Delta t^2)] \\ &= C(U_0) + \Delta t C'_{U_0}[A(U_0) + B(U_0)] + O(\Delta t), \end{aligned}$$

and

$$C'_{y(\Delta t)}C(y(\Delta t)) = C'_{U_0}C(U_0) + O(\Delta t).$$

Then we get

$$\begin{aligned} S_C^{\Delta t}(S_B^{\Delta t}(S_A^{\Delta t}(U_0))) &= U_0 + \Delta t[A(U_0) + B(U_0) + C(U_0)] \\ &+ \frac{\Delta t^2}{2}[A'_{U_0}A(U_0) + B'_{U_0}B(U_0) + C'_{U_0}C(U_0) \\ &+ 2B'_{U_0}A(U_0) + 2C'_{U_0}[A(U_0) + B(U_0)]] + O(\Delta t^3). \end{aligned}$$

Hence, we get

$$\begin{aligned} S_{A+B+C}^{\Delta t}(U_0) - S_C^{\Delta t}(S_B^{\Delta t}(S_A^{\Delta t}(U_0))) &= \frac{\Delta t^2}{2}[A'_{U_0}B(U_0) + A'_{U_0}C(U_0) + B'_{U_0}C(U_0) \\ &- B'_{U_0}A(U_0) - C'_{U_0}A(U_0) - C'_{U_0}B(U_0)] \\ &+ O(\Delta t^3). \end{aligned}$$

If A, B, and C are linear then

$$A' = A, B' = B, C' = C,$$

and the error is

$$\frac{\Delta t^2}{2}[AB + AC + BC - BA - CA - CB]. \qquad \blacksquare$$

Remark 4.4.1 *Note that in the linear case if the operators commute then the splitting error is second-order.*