DOI: 10.14464/gammas.v4i1.500 GAMMAS 2022, Vol. 4, No. 1, p. 15–28

International Association
of Applied Mathematics and Mechanics
– Archive for Students –

Monolithic, non-iterative and iterative time discretization methods for linear coupled elliptic–parabolic systems

Abdullah Mujahid[a,⋆]

[a] Universität Augsburg, Augsburg, Germany and Universität Stuttgart, Stuttgart, Germany

received 14.01.2022, accepted 30.05.2022, published 04.07.2022

⋆ corresponding author: abdullah.mujahid@math.uni-augsburg.de
supervisor: Robert Altmann, Universität Augsburg, Augsburg, Germany and Benjamin Unger, Universität Stuttgart, Stuttgart, Germany

Abstract: *We compare four numerical methods for the time discretization of linear coupled elliptic–parabolic systems. The monolithic method arising from an implicit Euler discretization is the primary method for solving the coupled system. An accelerated solution via non-iterative decoupling is possible by the semi-explicit Euler discretization, using a novel methodology from related delay differential equations. For poroelasticity, the fixed-stress splitting and the undrained splitting methods enable iterative decoupled solves. We present formulations for the iterative methods in an abstract form and derive the a priori convergence result. Finally, through numerical experiments, the a priori convergence results for the four methods are compared.*

Keywords: coupled elliptic–parabolic PDEs, implicit Euler method, semi-explicit Euler method, iterative decoupling methods, poroelasticity

1 Introduction

The quasi-static Biot poroelasticity model [4] results in a coupled elliptic–parabolic partial differential equation (PDE) system. This model finds application in geomechanics, biomedicine, petroleum engineering, and many more areas. Linear thermoelasticity [5] is another example of a coupled elliptic–parabolic PDE system with the same structure as poroelasticity. The primary numerical time integration method for solving the coupled elliptic–parabolic PDE system is via the *implicit Euler discretization* [10]. This is combined with either the finite element method (FEM) or the finite volume method (FVM) [12] for spatial discretization. Although the implicit method is unconditionally stable, it does not allow for specialized preconditioners for the block solve of the elliptic and the parabolic parts separately. For poroelasticity applications, specialized iterative methods exist, which decouple the elliptic and the parabolic equations iteratively. This includes the *fixed-stress* and the *undrained* splitting methods which are analysed in [14] for the convergence of the iterative schemes. For the optimal convergence rate of the iterative schemes, tuning of an additional stabilization parameter is required. An accelerated method for solving the coupled elliptic–parabolic PDE system is the *semi-explicit Euler discretization* for time integration. A priori convergence of the semi-explicit time discretization together with FEM for spatial discretization is analysed using a novel idea through the theory of delay differential equations [2]. This convergence result is established for a restrictive class of problems which satisfy a certain weak coupling condition. This is also established for certain nonlinear problems [1, 3]. In practice, the restrictive class actually encompasses a broad range of applications.

This article is devoted to a comparison of the mentioned time discretization methods for an abstract linear coupled elliptic–parabolic PDE system. The abstract setting and the problem are introduced in Section 2. The result on the existence of a weak solution for the coupled elliptic–parabolic PDE is stated in Section 3. The main contribution of this article is the formulation and proof of the existence of fixed-points for the iterative methods to decouple the abstract problem. The iterative decoupling is possible via two ways, either by first solving the decoupled parabolic equation and then solving the elliptic equation (Section 4.1), or first solving the decoupled elliptic equation and then solving the parabolic equation (Section 4.2). The convergence for these iterative methods for the abstract problem is shown for a restrictive class of problems satisfying some weak coupling conditions. We emphasize here that in contrast to the fixed-points of the above iterative methods for the abstract problem, the fixed-points for the iterative methods for the poroelasticity system are derived in [14] without any restrictive weak coupling conditions. In Section 5, the a priori convergence results for the implicit and the semi-explicit method are collected from [10] and [2]. Furthermore, the a priori convergence estimates for the iterative methods are derived using the contraction condition for fixed-points. Finally, in Section 6, for two different test problems, the convergence of the four methods

1. implicit Euler discretization,
2. semi-explicit Euler discretization,
3. elliptic–parabolic iterative decoupling with implicit Euler discretization, and
4. parabolic–elliptic iterative decoupling with implicit Euler discretization

is compared.

2 Abstract setting

To introduce the abstract problem of interest in this paper, we consider the following setting. Let $\Omega \subset \mathbb{R}^m$, $m \in \{2,3\}$ be a bounded Lipschitz domain and for $T > 0$ let $[0,T]$ be the time interval over which the problem is defined. Let $\mathcal{V} := [\mathcal{H}_0^1(\Omega)]^m$ and $\mathcal{Q} := \mathcal{H}_0^1(\Omega)$ be the Hilbert spaces with the associated Gelfand triples $\mathcal{V} \hookrightarrow \mathcal{H}_{\mathcal{V}} \simeq \mathcal{H}_{\mathcal{V}}^* \hookrightarrow \mathcal{V}^*$ and $\mathcal{Q} \hookrightarrow \mathcal{H}_{\mathcal{Q}} \simeq \mathcal{H}_{\mathcal{Q}}^* \hookrightarrow \mathcal{Q}^*$, respectively, where $\mathcal{H}_{\mathcal{V}} := [\mathcal{L}^2(\Omega)]^m$ and $\mathcal{H}_{\mathcal{Q}} := \mathcal{L}^2(\Omega)$ are the pivot spaces cf. [7, Sec. 5.2] . Let $a\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $b\colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ be symmetric, continuous and elliptic bilinear forms. Let $c\colon \mathcal{H}_{\mathcal{Q}} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ be a symmetric bilinear form that is continuous and elliptic in the pivot space $\mathcal{H}_{\mathcal{Q}}$. Let $d\colon \mathcal{V} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ be a bounded bilinear form. The constants defining the properties of the bilinear forms are collected in the following assumption.

Assumption 1 (Bilinear forms)**.** For the bilinear form $a\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, there exist some positive constants c_{a} and C_{a}, such that

$$a(u,u) \ge c_{\mathrm{a}} \|u\|_{\mathcal{V}}^2, \qquad a(u,v) \le C_{\mathrm{a}} \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}$$

for all $u, v \in \mathcal{V}$. For the bilinear form $b\colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$ there exist some positive constants c_{b} and C_{b} such that

$$b(p,p) \ge c_{\mathrm{b}} \|p\|_{\mathcal{Q}}^2, \qquad b(p,q) \le C_{\mathrm{b}} \|p\|_{\mathcal{Q}} \|q\|_{\mathcal{Q}}$$

for all $p, q \in \mathcal{Q}$. For the bilinear form $c\colon \mathcal{H}_{\mathcal{Q}} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ there exist some positive constants c_{c} and C_{c} such that

$$c(p,p) \ge c_{\mathrm{c}} \|p\|_{\mathcal{H}_{\mathcal{Q}}}^2, \qquad c(p,q) \le C_{\mathrm{c}} \|p\|_{\mathcal{H}_{\mathcal{Q}}} \|q\|_{\mathcal{H}_{\mathcal{Q}}}$$

for all $p, q \in \mathcal{H}_{\mathcal{Q}}$. The continuity of the bilinear form $d\colon \mathcal{V} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ is defined in two ways, i.e., there exist some positive constants C_{d} and $\tilde{C}_{\mathrm{d}}$ such that

$$d(u,p) \le C_{\mathrm{d}} \|u\|_{\mathcal{V}} \|p\|_{\mathcal{H}_{\mathcal{Q}}}, \qquad d(u,p) \le \tilde{C}_{\mathrm{d}} \|u\|_{\mathcal{H}_{\mathcal{V}}} \|p\|_{\mathcal{Q}}$$

for all $u \in \mathcal{V}$ and $p \in \mathcal{Q}$. The bilinear form $\mathfrak{a} \in \{a, b, c\}$ defines a natural norm $\|y\|_{\mathfrak{a}} := \sqrt{\mathfrak{a}(y,y)}$ where y belongs to an appropriate space $\mathcal{Y}$ defined by the bilinear form. Furthermore,

$$c_{\mathfrak{a}} \|y\|_{\mathcal{Y}}^2 \le \|y\|_{\mathfrak{a}}^2 \le C_{\mathfrak{a}} \|y\|_{\mathcal{Y}}^2.$$

Given the bilinear forms, an abstract linear elliptic–parabolic PDE in the weak form with Dirichlet boundary conditions reads as follows.

Definition 2 (Abstract problem)**.** Given source terms $f\colon [0,T] \to \mathcal{V}^*$, $g\colon [0,T] \to \mathcal{Q}^*$ which are sufficiently smooth, find abstract functions $(u,p) : [0,T] \to \mathcal{V} \times \mathcal{Q}$ such that for a. e. $t \in [0,T]$ it holds that

$$a(u,v) - d(v,p) = \langle f, v\rangle, \qquad (1a)$$

$$d(\dot{u},q) + c(\dot{p},q) + b(p,q) = \langle g, q\rangle \qquad (1b)$$

for all test functions $v \in \mathcal{V}$, $q \in \mathcal{Q}$. Furthermore, the initial conditions

$$u(0,\cdot) = u^0 \in \mathcal{V}, \qquad p(0,\cdot) = p^0 \in \mathcal{H}_{\mathcal{Q}}, \qquad (1c)$$

are consistently defined through (1a), *i.e.*, we assume $a(u^0, v) - d(v, p^0) = \langle f^0, v\rangle$ for all $v \in \mathcal{V}$.

The bilinear form d couples the two equations (1a) and (1b). Without the coupling term, (1a) is elliptic in u and (1b) is parabolic in p. To measure the coupling between the two equations, we can introduce

$$\omega := \frac{C_\mathrm{d}^2}{c_\mathrm{a} c_\mathrm{c}}.$$

The quasi-static Biot poroelasticity model is one example for the abstract problem defined in Definition 2. Linear thermoelasticity is another example.

Example 3 (Poroelasticity [4])**.** The quasi-static Biot poroelasticity model is given by

$$-\nabla\cdot\sigma(u)+\alpha\nabla p = f, \tag{2a}$$

$$\frac{\partial}{\partial t}\Big(\frac{1}{M}p+\alpha\nabla\cdot u\Big)-\nabla\cdot\Big(\frac{\kappa}{\nu}\nabla p\Big)=g. \tag{2b}$$

The quantities of interest are the displacement of the solid porous medium $u:[0,T]\times\Omega\to\mathbb{R}^m$ and the pressure of the fluid in the pores $p:[0,T]\times\Omega\to\mathbb{R}$. The point-wise Cauchy stress tensor $\sigma(t,x):\mathbb{R}^m\to\mathbb{R}^m$ for $(t,x)\in[0,T]\times\Omega$ encodes the internal forces in the solid porous medium and relates linearly to symmetric gradient of displacement $\varepsilon(u)$. This is given by Hooke's law characterized by the positive Lamé constants λ and μ [6]. The right-hand sides $f:[0,T]\times\Omega\to\mathbb{R}^m$ and $g:[0,T]\times\Omega\to\mathbb{R}$ represent appropriate source terms. The quantities α, M, and $\frac{\kappa}{\nu}$ are positive constant poroelasticity parameters. The definitions of the bilinear forms in Definition 2 for the case of poroelasticity (2) take the form

$$a(u,v) := \int_\Omega \sigma(u):\varepsilon(v)\,\mathrm{d}x, \qquad b(p,q) := \int_\Omega \frac{\kappa}{\nu}\nabla p\cdot\nabla q\,\mathrm{d}x,$$

$$c(p,q) := \int_\Omega \frac{1}{M}p\,q\,\mathrm{d}x, \qquad d(u,q) := \int_\Omega \alpha\,(\nabla\cdot u)\,q\,\mathrm{d}x.$$

These definitions satisfy the assumed properties for the bilinear forms in Assumption 1 cf. [2].

Example 4 (Thermoelasticity [5])**.** Thermoelasticity describes the displacement $u:[0,T]\times\Omega\to\mathbb{R}^m$ of a solid body deforming under the influence of a temperature field $p:[0,T]\times\Omega\to\mathbb{R}$. The structure of the bilinear forms in thermoelasticity are similar to those of the poroelasticity, except for the reinterpretation of some constants. Here, the constant α in the bilinear form d (cf. Example 3), which couples the two equations is interpreted as the thermal expansion coefficient.

It is convenient to adopt the operator notation for (1). Hence, for the bilinear forms, we associate operators and consider (1) in operator notation.

Definition 5 (Operators associated with bilinear forms)**.** Let $\mathscr{A}:\mathcal{V}\to\mathcal{V}^*$, $\mathscr{B}:\mathcal{Q}\to\mathcal{Q}^*$, $\mathscr{C}:\mathcal{H}_\mathcal{Q}\to\mathcal{H}_\mathcal{Q}^*$ and $\mathscr{D}:\mathcal{H}_\mathcal{Q}\to\mathcal{V}^*$ be the operators associated with the the bilinear forms a, b, c and d, respectively. The definitions are given as

$$\langle\mathscr{A}u,v\rangle := a(u,v), \qquad \langle\mathscr{B}p,q\rangle := b(p,q),$$

$$\langle\mathscr{C}p,q\rangle := c(p,q), \qquad \langle\mathscr{D}p,v\rangle := d(v,p).$$

We consider $\mathscr{D}^*:\mathcal{V}\to\mathcal{H}_\mathcal{Q}^*$ for the adjoint of $\mathscr{D}$.

The operator notation of (1) reads for a. e. $t\in[0,T]$,

$$\mathscr{A}u-\mathscr{D}p=f \quad \text{in } \mathcal{V}^*, \tag{3a}$$

$$\mathscr{D}^*\dot{u}+\mathscr{C}\dot{p}+\mathscr{B}p=g \quad \text{in } \mathcal{Q}^*. \tag{3b}$$

3 Existence theorems

We collect results on the existence of unique solutions for elliptic and parabolic PDEs in the weak form. This will then be used to prove the existence of a unique weak solution of the coupled system (1). For elliptic PDEs, the Lax–Milgram theorem provides the unique existence of a solution.

Theorem 6 (Lax–Milgram [11, Ch. 6, Th. 1])**.** *Consider a continuous elliptic bilinear form $a:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ defined on a Hilbert space $\mathcal{V}$ and let $f\in\mathcal{V}^*$. Then there exists a unique solution $u\in\mathcal{V}$ such that*

$$a(u,v)=\langle f,v\rangle \quad \textit{for all } v\in\mathcal{V}.$$

For parabolic PDEs, an analogue of the Lax–Milgram theorem is the Lions theorem which we state below in a slightly modified form.

Theorem 7 (J. L. Lions [13, Ch. 3, Th. 4.2 with Th. 1.1])**.** *Consider the Hilbert spaces $\mathcal{Q}\subset\mathcal{H}_\mathcal{Q}\subset\mathcal{Q}^*$ with a Gelfand triple structure, a continuous and elliptic bilinear form $b:\mathcal{Q}\times\mathcal{Q}\to\mathbb{R}$ and a monotone operator $\mathscr{T}:\mathcal{H}_\mathcal{Q}\to\mathcal{H}_\mathcal{Q}$ with an appropriate extension $\widehat{\mathscr{T}}:\mathcal{Q}^*\to\mathcal{Q}^*$ defined as*

$$\Big\langle\widehat{\mathscr{T}}\frac{\mathrm{d}}{\mathrm{d}t}p,q\Big\rangle := \Big\langle\frac{\mathrm{d}}{\mathrm{d}t}(\mathscr{T}p),q\Big\rangle := \frac{\mathrm{d}}{\mathrm{d}t}(\mathscr{T}p,q)_{\mathcal{H}_\mathcal{Q}}$$

for all $p,q\in\mathcal{H}_\mathcal{Q}$. Then, given $g\in\mathscr{L}^2(0,T;\mathcal{Q}^*)$ and $p^0\in\mathcal{H}_\mathcal{Q}$, there exists a unique

$$p\in\mathscr{L}^2(0,T;\mathcal{Q})\cap\mathscr{H}^1(0,T;\mathcal{Q}^*)\hookrightarrow\mathscr{C}([0,T];\mathcal{H}_\mathcal{Q}),$$

satisfying the evolution problem for a. e. $t\in[0,T]$,

$$\Big\langle\widehat{\mathscr{T}}\frac{\mathrm{d}}{\mathrm{d}t}p(t),q\Big\rangle+b(p,q)=\langle g(t),q\rangle \ \textit{for all } q\in\mathcal{Q},$$

$$p(0)=p^0.$$

Theorem 7 follows from [13, Ch. 3,Th. 1.1], where we check the monotone property of the first infinitesimal generator operator.

Remark 8 (Extension of $\mathscr{T}$ to $\widehat{\mathscr{T}}$)**.** Since the operator $\mathscr{T}: \mathcal{H}_{\mathcal{Q}} \to \mathcal{H}_{\mathcal{Q}}$ preserves the regularity (*i.e.*, no spatial differentiation operation is involved) and $g \in \mathcal{L}^2(0,T;\mathcal{Q}^*)$, it is customary to seek $\frac{\mathrm{d}}{\mathrm{d}t}p \in \mathcal{L}^2(0,T;\mathcal{Q}^*)$ (cf. [15, Ch. 23]). Furthermore, we identify $\widehat{\mathscr{T}}$ with $\mathscr{T}$ while the mapping is understood implicitly. ♣

Theorem 9 (Existence of a unique weak solution)**.** *Consider given consistent initial data $u^0 \in \mathcal{V}$, $p^0 \in \mathcal{H}_{\mathcal{Q}}$ and right-hand sides $f \in \mathcal{H}^1(0,T;\mathcal{V}^*)$, $g \in \mathcal{L}^2(0,T;\mathcal{Q}^*)$. Under Assumption 1 and Definition 5, system (3) has a unique weak solution*

$$p \in \mathcal{L}^2(0,T;\mathcal{Q}) \cap \mathcal{H}^1(0,T;\mathcal{Q}^*) \hookrightarrow \mathcal{C}([0,T];\mathcal{H}_{\mathcal{Q}}) \text{ and}$$
$$u \in \mathcal{C}([0,T];\mathcal{V}).$$

Proof. The assumptions imply the existence of $\mathcal{A}^{-1}$. Hence, solving for u from (3a) yields

$$u(t) = \mathcal{A}^{-1}\big(f(t) + \mathcal{D}p(t)\big)$$

which on substituting in (3b) after differentiating w.r.t t gives

$$\big(\mathcal{C} + \mathcal{D}^*\mathcal{A}^{-1}\mathcal{D}\big)\dot{p} + \mathcal{B}p = g - \mathcal{D}^*\mathcal{A}^{-1}\dot{f} \tag{4}$$

which is a parabolic PDE in p. The right-hand side satisfies $g - \mathcal{D}^*\mathcal{A}^{-1}\dot{f} \in \mathcal{L}^2(0,T;\mathcal{Q}^*)$ which is easily verifiable from the definition of the operators. The bilinear form associated with the operator $\mathcal{B}$ is elliptic by assumption. To check the monotonicity of the operator $\mathscr{T} := \mathcal{C} + \mathcal{D}^*\mathcal{A}^{-1}\mathcal{D}$, we see that the bilinear form associated with $\mathcal{C}$ is elliptic in $\mathcal{H}_{\mathcal{Q}}$ and $\langle \mathcal{D}^*\mathcal{A}^{-1}\mathcal{D}p, p\rangle = \langle \mathcal{A}^{-1}\mathcal{D}p, \mathcal{D}p\rangle \geq 0$ for all $p \in \mathcal{H}_{\mathcal{Q}}$. Hence, all the assumptions of Theorem 7 are satisfied and we are guaranteed a unique solution

$$p \in \mathcal{L}^2(0,T;\mathcal{Q}) \cap \mathcal{H}^1(0,T;\mathcal{Q}^*) \hookrightarrow \mathcal{C}([0,T];\mathcal{H}_{\mathcal{Q}}).$$

Next, we observe that $f \in \mathcal{H}^1(0,T;\mathcal{V}^*) \hookrightarrow \mathcal{C}([0,T];\mathcal{V}^*)$ and $\mathcal{D}p \in \mathcal{C}([0,T];\mathcal{V}^*)$. Hence, by Theorem 6, we get a unique $u \in \mathcal{C}([0,T];\mathcal{V})$. □

4 Fixed-points of iterative decoupling methods

For poroelasticity, there exist iterative decoupling methods which allow for specialized solves of the mechanics (2a) and flow (2b) equations, separately. The convergence of the fixed-stress splitting and the undrained splitting iterative methods for poroelasticity are analysed in [14]. We present the formulation of the iterative decoupling in the general case for the abstract problem (1).

For defining the iterative methods, let the index i denote the iteration number for the field values at a given time $t \in [0,T]$. Define the error between the field values at two successive iterates as

$$e_u^{i+1} := u^{i+1} - u^i, \qquad e_p^{i+1} := p^{i+1} - p^i,$$

and denote their time derivatives as

$$\dot{e}_u^{i+1} = \dot{u}^{i+1} - \dot{u}^i, \qquad \dot{e}_p^{i+1} = \dot{p}^{i+1} - \dot{p}^i.$$

Note that $e_u^{i+1}(0) = 0$ and $e_p^{i+1}(0) = 0$. For the product space $\mathcal{S} := \mathcal{C}([0,T];\mathcal{V}) \times \mathcal{C}([0,T];\mathcal{H}_{\mathcal{Q}})$, define a distance function $d_t : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ using the error terms as

$$\begin{aligned} d_t^2\big((u^{i+1},p^{i+1}),(u^i,p^i)\big) &:= \\ c_a\int_0^t \|\dot{e}_u^{i+1}(s)\|_{\mathcal{V}}^2\,\mathrm{d}s &+ c_c\int_0^t \|\dot{e}_p^{i+1}(s)\|_{\mathcal{H}_{\mathcal{Q}}}^2\,\mathrm{d}s \\ &+ \max_{0\leq s\leq t} c_b\|e_p^{i+1}(s)\|_{\mathcal{Q}}^2. \end{aligned} \tag{5}$$

In the forthcoming analysis, we will use the following result.

Lemma 10. *Suppose $\xi, \delta > 0$. Then*

$$2(1+\delta)(\delta+\xi) < 1$$

if and only if

$$\xi < \frac{1}{2} \quad \text{and} \quad \delta < -\frac{1}{2}(1+\xi) + \frac{1}{2}\sqrt{(1-\xi)^2+2}.$$

Proof. It is easy to see the *if* part of the statement. To see that the converse is true, we first observe that $\xi, \delta > 0$ together with

$$2(1+\delta)(\xi+\delta) = 2\xi + 2\delta\xi + 2\delta(1+\delta) < 1,$$

implies $\xi < \frac{1}{2}$. Next, we observe that the discriminant of the quadratic equation

$$\delta^2 + (\xi+1)\delta + \xi - \frac{1}{2} = 0 \tag{6}$$

is $(1-\xi)^2 + 2 \geq 9/4$. Hence

$$2(1+\delta)(\delta+\xi) < 1 \quad \text{or} \quad (\delta-\delta_1)(\delta-\delta_2) < 0$$

if $\delta_1 < \delta < \delta_2$, where δ_1, δ_2 are roots of (6). But $\delta_1 < 0$ and since $\xi < \frac{1}{2}$ we get

$$\delta_2 = -\frac{1}{2}(1+\xi) + \frac{1}{2}\sqrt{(1-\xi)^2+2} > 0.$$ □

4.1 Parabolic–elliptic iterative decoupling

To decouple system (3), we formally set

$$\mathscr{D}^* u^{i+1} := \mathscr{D}^* u^i + L_{\mathscr{F}} \mathscr{F}(p^{i+1} - p^i) \tag{7}$$

for a given operator $\mathscr{F}\colon \mathscr{H}_{\mathscr{Q}} \to \mathscr{H}_{\mathscr{Q}}$ and a constant $L_{\mathscr{F}} \in \mathbb{R}_+$. When the derivative of (7) is substituted in (3b), (3) yields the decoupled system

$$\mathscr{A} u^{i+1} - \mathscr{D} p^{i+1} = f, \tag{8a}$$

$$\mathscr{D}^* \dot{u}^i + L_{\mathscr{F}} \mathscr{F}(\dot{p}^{i+1} - \dot{p}^i) + \mathscr{C} \dot{p}^{i+1} + \mathscr{B} p^{i+1} = g. \tag{8b}$$

We immediately see from Theorem 7 and the proof of Theorem 9, that it is sufficient to assume that the operator $L_{\mathscr{F}}\mathscr{F}$ is monotone for unique solvability of p^{i+1} from (8b) given (u^i, p^i). This leads to Lemma 11. We assume the following properties of $\mathscr{F}$:

$$0 \le L_{\mathscr{F}} \langle \mathscr{F} p, p\rangle, \tag{9a}$$

$$\langle \mathscr{F} p, q\rangle \le C_{\mathscr{F}} \|p\|_{\mathscr{H}_{\mathscr{Q}}} \|q\|_{\mathscr{H}_{\mathscr{Q}}} \tag{9b}$$

for all $p, q \in \mathscr{H}_{\mathscr{Q}}$.

Lemma 11. *Let $\zeta_{\mathscr{F}}\colon \mathscr{S} \to \mathscr{S}$ be defined for* a. e. t *in* $[0, T]$ *as $\zeta_{\mathscr{F}}(u^i, p^i) := (u^{i+1}, p^{i+1})$ with (u^{i+1}, p^{i+1}) being the solution to* (8)*. If $L_{\mathscr{F}}\mathscr{F}$ is monotone, then $\zeta_{\mathscr{F}}$ is well-defined.*

Example 12 (Fixed-stress splitting in poroelasticity [14])**.** The decoupling condition for the fixed-stress splitting in poroelasticity is given by

$$\nabla\cdot u^{i+1} = \nabla\cdot u^i + L\frac{\alpha}{K_{\mathrm{dr}}}(p^{i+1} - p^i)$$

where K_{dr} is another poroelasticity parameter, and $L \in \mathbb{R}_+$ is an additional stabilization parameter. Here $L_{\mathscr{F}}\mathscr{F}$ is monotone with $L_{\mathscr{F}} = L$, and $\mathscr{F} = \frac{\alpha^2}{K_{\mathrm{dr}}}\mathscr{I}$ where $\mathscr{I}$ is the identity operator.

Theorem 13. *Consider $\zeta_{\mathscr{F}}\colon \mathscr{S} \to \mathscr{S}$ from Lemma 11. Let $\xi_{\mathscr{F}} := (L_{\mathscr{F}}^2 C_{\mathscr{F}}^2)/c_{\mathrm{c}}^2$. Suppose the conditions of Lemma 11 hold and*

$$\xi_{\mathscr{F}} < \frac{1}{2}, \quad \omega < -\frac{1}{2}\big(1 + \xi_{\mathscr{F}}\big) + \frac{1}{2}\sqrt{\big(1 - \xi_{\mathscr{F}}\big)^2 + 2}. \tag{10}$$

Then $\zeta_{\mathscr{F}}$ is a contraction map and has a unique fixed point.

Proof. Taking the difference between two successive iterates of the system obtained after differentiating (8a) w.r.t. t gives

$$a(\dot{e}_u^{i+1}, v) = d(v, \dot{e}_p^{i+1}), \tag{11a}$$

$$\begin{aligned} L_{\mathscr{F}}\langle \mathscr{F}\dot{e}_p^{i+1}, q\rangle &+ c(\dot{e}_p^{i+1}, q) \\ &+ b(e_p^{i+1}, q) = -d(\dot{e}_u^i, q) + L_{\mathscr{F}}\langle \mathscr{F}\dot{e}_p^i, q\rangle \end{aligned} \tag{11b}$$

for all test functions $v \in \mathcal{V}$, $q \in \mathscr{Q}$. Testing (11a) with $v = \dot{e}_u^{i+1}$ and using Assumption 1 together with an application of the weighted Young's inequality yields

$$\begin{aligned} c_{\mathrm{a}}\|\dot{e}_u^{i+1}\|_{\mathcal{V}}^2 \le a(\dot{e}_u^{i+1}, \dot{e}_u^{i+1}) &= d(\dot{e}_u^{i+1}, \dot{e}_p^{i+1}) \\ &\le C_{\mathrm{d}}\|\dot{e}_u^{i+1}\|_{\mathcal{V}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}} \\ &\le \frac{\delta}{2}\|\dot{e}_u^{i+1}\|_{\mathcal{V}}^2 + \frac{1}{2\delta}C_{\mathrm{d}}^2\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2, \end{aligned}$$

which on choosing $\delta = c_{\mathrm{a}}$ simplifies to

$$c_{\mathrm{a}}\|\dot{e}_u^{i+1}\|_{\mathcal{V}}^2 \le \omega c_{\mathrm{c}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2. \tag{12}$$

Next, testing (11b) with $q = \dot{e}_p^{i+1}$, together with Assumption 1 we have

$$\begin{aligned} L_{\mathscr{F}}\langle \mathscr{F}\dot{e}_p^{i+1}, \dot{e}_p^{i+1}\rangle + c_{\mathrm{c}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_p^{i+1}\|_b^2 & \\ \le C_{\mathrm{d}}\|\dot{e}_u^i\|_{\mathcal{V}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}& \\ + L_{\mathscr{F}}\langle \mathscr{F}\dot{e}_p^i, \dot{e}_p^{i+1}\rangle&, \end{aligned}$$

which on using (9) and the weighted Young's inequality simplifies to

$$\begin{aligned} c_{\mathrm{c}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_p^{i+1}\|_b^2 \le \frac{\delta}{2}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2 &+ \frac{1}{2\delta}C_{\mathrm{d}}^2\|\dot{e}_u^i\|_{\mathcal{V}}^2 \\ + \frac{\delta}{2}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2& \\ + \frac{1}{2\delta}L_{\mathscr{F}}^2 C_{\mathscr{F}}^2\|\dot{e}_p^i\|_{\mathscr{H}_{\mathscr{Q}}}^2&. \end{aligned}$$

Choosing $\delta = \frac{1}{2}c_{\mathrm{c}}$ and simplifying the right-hand side gives

$$\begin{aligned} c_{\mathrm{c}}\|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}}^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|e_p^{i+1}\|_b^2 \le 2\Big(\frac{C_{\mathrm{d}}^2}{c_{\mathrm{a}}\,c_{\mathrm{c}}} + \frac{L_{\mathscr{F}}^2 C_{\mathscr{F}}^2}{c_{\mathrm{c}}^2}\Big)& \\ \big\{c_{\mathrm{a}}\|\dot{e}_u^i\|_{\mathcal{V}}^2 + c_{\mathrm{c}}\|\dot{e}_p^i\|_{\mathscr{H}_{\mathscr{Q}}}^2\big\}&. \end{aligned}$$

Integrating and taking the supremum yields

$$\begin{aligned} &\Big\{c_{\mathrm{c}}\int_0^t \|\dot{e}_p^{i+1}(s)\|_{\mathscr{H}_{\mathscr{Q}}}^2\,\mathrm{d}s + \max_{0\le s\le t} c_{\mathrm{b}}\|e_p^{i+1}(s)\|_{\mathscr{Q}}^2\Big\} \\ &\quad\le 2\Big(\frac{C_{\mathrm{d}}^2}{c_{\mathrm{a}}\,c_{\mathrm{c}}} + \frac{L_{\mathscr{F}}^2 C_{\mathscr{F}}^2}{c_{\mathrm{c}}^2}\Big) \\ &\qquad\Big\{c_{\mathrm{a}}\int_0^t \|\dot{e}_u^i(s)\|_{\mathcal{V}}^2\,\mathrm{d}s + c_{\mathrm{c}}\int_0^t \|\dot{e}_p^i(s)\|_{\mathscr{H}_{\mathscr{Q}}}^2\,\mathrm{d}s\Big\}. \end{aligned} \tag{13}$$

Integrating (12), combining with (13), and adding appropriate positive terms on the right-hand side, we finally obtain

$$\begin{aligned}&\Big\{c_{\mathrm a}\int_0^t \|\dot e_u^{i+1}(s)\|_{\mathcal V}^2\,\mathrm ds + c_{\mathrm c}\int_0^t \|\dot e_p^{i+1}(s)\|_{\mathscr H_{\mathscr Q}}^2\,\mathrm ds \\ &\quad + \max_{0\le s\le t} c_{\mathrm b}\|e_p^{i+1}(s)\|_{\mathscr Q}^2\Big\} \\ &\le 2\big(1+\omega\big)\Big(\omega + \frac{L_{\mathscr F}^2 C_{\mathscr F}^2}{c_{\mathrm c}^2}\Big) \\ &\qquad \Big\{c_{\mathrm a}\int_0^t \|\dot e_u^{i}(s)\|_{\mathcal V}^2\,\mathrm ds + c_{\mathrm c}\int_0^t \|\dot e_p^{i}(s)\|_{\mathscr H_{\mathscr Q}}^2\,\mathrm ds \\ &\qquad + \max_{0\le s\le t} c_{\mathrm b}\|e_p^{i}(s)\|_{\mathscr Q}^2\Big\}. \end{aligned} \tag{14}$$

With the definition of the distance function (5), (14) reads

$$d_t^2\big((u^{i+1},p^{i+1}),(u^i,p^i)\big) \le \gamma_{\mathscr F}^2\, d_t^2\big((u^i,p^i),(u^{i-1},p^{i-1})\big)$$

with

$$\gamma_{\mathscr F}^2 := 2\big(1+\omega\big)\Big(\omega + \frac{L_{\mathscr F}^2 C_{\mathscr F}^2}{c_{\mathrm c}^2}\Big).$$

For this to be a contraction, *i.e.*, $\gamma_{\mathscr F} < 1$, using Lemma 10 we get the conditions (10). From contraction mapping principles [8, Th. 3.1], the iterative system (8) admits a fixed-point under the conditions (10). □

4.2 Elliptic–parabolic iterative decoupling

An alternate way to decouple the equations (3) is to formally set

$$p^{i+1} := p^i - L_{\mathscr U}\mathscr U(u^{i+1}-u^i) \tag{15}$$

for a given operator $\mathscr U\colon \mathcal V \to \mathscr H_{\mathscr Q}$ and a constant $L_{\mathscr U} \in \mathbb R_+$. When (15) is substituted in (3a), (3) yields the decoupled system

$$\mathscr A u^{i+1} - \mathscr D p^i + L_{\mathscr U}\mathscr D\mathscr U(u^{i+1}-u^i) = f, \tag{16a}$$

$$\mathscr D^*\dot u^{i+1} + \mathscr C\dot p^{i+1} + \mathscr B p^{i+1} = g. \tag{16b}$$

We immediately see from Theorem 6 and the proof of Theorem 9 that it is sufficient for the operator $\mathscr D\mathscr U$ to be continuous and monotone for unique solvability of u^{i+1} from (16a) given (u^i, p^i). Hence, this leads to Lemma 14. Defining $C_{\mathscr D\mathscr U}$ to be the positive continuity constant of $\mathscr D\mathscr U$, we formally have

$$0 \le \langle \mathscr D\mathscr U u, u\rangle, \tag{17a}$$

$$\langle \mathscr D\mathscr U u, v\rangle \le C_{\mathscr D\mathscr U}\|u\|_{\mathcal V}\|v\|_{\mathcal V} \tag{17b}$$

for all test functions $u, v \in \mathcal V$.

Lemma 14. *Let $\zeta_{\mathscr U}\colon \mathscr S \to \mathscr S$ be defined for* a. e. t *in* $[0,T]$ *as $\zeta_{\mathscr U}(u^i,p^i) := (u^{i+1},p^{i+1})$ where (u^{i+1},p^{i+1}) being the solution to* (16). *If $\mathscr D\mathscr U$ is continuous and monotone, then $\zeta_{\mathscr U}$ is well-defined.*

Example 15 (Undrained splitting in poroelasticity [14])**.** The decoupling condition for the undrained splitting in poroelasticity is given by

$$p^{i+1} = p^i - LM\alpha\nabla\cdot\big(u^{i+1}-u^i\big)$$

where $L \in \mathbb R_+$ is an additional stabilization parameter. This equation fits (15) for $L_{\mathscr U} = L$ and $\mathscr U = M\alpha\nabla\cdot$. But,

$$\langle \mathscr D\mathscr U u, u\rangle = \langle M\alpha^2\nabla\nabla\cdot u, u\rangle = -M\alpha^2\langle\nabla\cdot u, \nabla\cdot u\rangle \le 0$$

is not monotone. However, with the free parameter $L_{\mathscr U}$, we can ensure the necessary condition that $\mathscr A + L_{\mathscr U}\mathscr D\mathscr U$ is elliptic.

Theorem 16. *Consider $\zeta_{\mathscr U}\colon \mathscr S \to \mathscr S$ from Lemma 14. Let $\xi_{\mathscr D\mathscr U} := (L_{\mathscr U}^2 C_{\mathscr D\mathscr U}^2)/c_{\mathrm a}^2$. Suppose the conditions of the Lemma 14 hold and*

$$\xi_{\mathscr D\mathscr U} < \frac12, \quad \omega < -\frac12\big(1+\xi_{\mathscr D\mathscr U}\big) + \frac12\sqrt{\big(1-\xi_{\mathscr D\mathscr U}\big)^2 + 2}. \tag{18}$$

Then $\zeta_{\mathscr U}$ is a contraction map and has a unique fixed point.

Proof. Taking the difference between two successive iterates of the system obtained after differentiating (16a) w.r.t. t gives

$$\begin{aligned} a(\dot e_u^{i+1}, v) + L_{\mathscr U}\langle \mathscr D\mathscr U\dot e_u^{i+1}, v\rangle &= \\ d(v,\dot e_p^i) &+ L_{\mathscr U}\langle \mathscr D\mathscr U\dot e_u^i, v\rangle, \end{aligned} \tag{19a}$$

$$d(\dot e_u^{i+1}, q) + c(\dot e_p^{i+1}, q) + b(e_p^{i+1}, q) = 0 \tag{19b}$$

for all test functions $v \in \mathcal V$, $q \in \mathscr Q$. Testing (19a) with $v = \dot e_u^{i+1}$, using Assumption 1, and (17), we obtain

$$\begin{aligned} c_{\mathrm a}\|\dot e_u^{i+1}\|_{\mathcal V}^2 &\le C_{\mathrm d}\|\dot e_u^{i+1}\|_{\mathcal V}\|\dot e_p^i\|_{\mathscr H_{\mathscr Q}} \\ &\quad + L_{\mathscr U}C_{\mathscr D\mathscr U}\|\dot e_u^i\|_{\mathcal V}\|\dot e_u^{i+1}\|_{\mathcal V}. \end{aligned}$$

On application of the weighted Young's inequality for the terms on the right-hand side, the above estimate becomes

$$\begin{aligned} c_{\mathrm a}\|\dot e_u^{i+1}\|_{\mathcal V}^2 &\le \frac{\delta}{2}\|\dot e_u^{i+1}\|_{\mathcal V}^2 + \frac{1}{2\delta}C_{\mathrm d}^2\|\dot e_p^i\|_{\mathscr H_{\mathscr Q}}^2 \\ &\quad + \frac{\delta}{2}\|\dot e_u^{i+1}\|_{\mathcal V}^2 + \frac{1}{2\delta}L_{\mathscr U}^2 C_{\mathscr D\mathscr U}^2\|\dot e_u^i\|_{\mathcal V}^2. \end{aligned}$$

Choosing $\delta = \frac12 c_{\mathrm a}$ leads to

$$\begin{aligned} c_{\mathrm a}\|\dot e_u^{i+1}\|_{\mathcal V}^2 &\le 2\Big(\frac{C_{\mathrm d}^2}{c_{\mathrm a}c_{\mathrm c}} + \frac{L_{\mathscr U}^2 C_{\mathscr D\mathscr U}^2}{c_{\mathrm a}^2}\Big) \\ &\qquad \Big\{c_{\mathrm a}\|\dot e_u^i\|_{\mathcal V}^2 + c_{\mathrm c}\|\dot e_p^i\|_{\mathscr H_{\mathscr Q}}^2\Big\}. \end{aligned} \tag{20}$$

Next, testing (19b) with $q = \dot{e}_p^{i+1}$ and using the weighted Young's inequality, we obtain the estimate

$$\begin{aligned} c_{\mathrm{c}} \|\dot{e}_p^{i+1}\|^2_{\mathscr{H}_{\mathscr{Q}}} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_p^{i+1}\|_b^2 \le C_{\mathrm{d}} \|\dot{e}_u^{i+1}\|_{\mathscr{V}} \|\dot{e}_p^{i+1}\|_{\mathscr{H}_{\mathscr{Q}}} &\le \frac{\delta}{2}\|\dot{e}_p^{i+1}\|^2_{\mathscr{H}_{\mathscr{Q}}} \\ &\quad + \frac{1}{2\delta} C_{\mathrm{d}}^2 \|\dot{e}_u^{i+1}\|^2_{\mathscr{V}}. \end{aligned}$$

Choosing $\delta = c_{\mathrm{c}}$, integrating, and taking the supremum, the previous inequality simplifies to

$$\begin{aligned} c_{\mathrm{c}} \int_0^t \|\dot{e}_p^{i+1}(s)\|^2_{\mathscr{H}_{\mathscr{Q}}} \,\mathrm{d}s + \max_{0\le s\le t} c_{\mathrm{b}} \|e_p^{i+1}(s)\|^2_{\mathscr{Q}} \\ \le \omega c_{\mathrm{a}} \int_0^t \|\dot{e}_u^{i+1}(s)\|^2_{\mathscr{V}} \,\mathrm{d}s. \end{aligned} \tag{21}$$

Integrating (20) and combining with (21), we get

$$\begin{aligned} &\Big\{ c_{\mathrm{a}} \int_0^t \|\dot{e}_u^{i+1}(s)\|^2_{\mathscr{V}} \,\mathrm{d}s + c_{\mathrm{c}} \int_0^t \|\dot{e}_p^{i+1}(s)\|^2_{\mathscr{H}_{\mathscr{Q}}} \,\mathrm{d}s \\ &\quad + \max_{0\le s\le t} c_{\mathrm{b}} \|e_p^{i+1}(s)\|^2_{\mathscr{Q}} \Big\} \\ &\le 2\big(1+\omega\big)\Big(\omega + \frac{L_{\mathscr{U}}^2 C_{\mathscr{D}\mathscr{U}}^2}{c_{\mathrm{a}}^2}\Big) \\ &\quad \Big\{ c_{\mathrm{a}} \int_0^t \|\dot{e}_u^{i}(s)\|^2_{\mathscr{V}} \,\mathrm{d}s + c_{\mathrm{c}} \int_0^t \|\dot{e}_p^{i}(s)\|^2_{\mathscr{H}_{\mathscr{Q}}} \,\mathrm{d}s \\ &\quad + \max_{0\le s\le t} c_{\mathrm{b}} \|e_p^{i}(s)\|^2_{\mathscr{Q}} \Big\}. \end{aligned} \tag{22}$$

Using the distance function (5), (22) can be written

$$d_t^2\big((u^{i+1}, p^{i+1}), (u^i, p^i)\big) \le \gamma_{\mathscr{U}}^2 d_t^2\big((u^i, p^i), (u^{i-1}, p^{i-1})\big)$$

with

$$\gamma_{\mathscr{U}}^2 := 2\big(1+\omega\big)\Big(\omega + \frac{L_{\mathscr{U}}^2 C_{\mathscr{D}\mathscr{U}}^2}{c_{\mathrm{a}}^2}\Big).$$

For this to be a contraction, *i.e.*, $\gamma_{\mathscr{U}} < 1$, using Lemma 10 we get the conditions (18). From contraction mapping principles [8, Th. 3.1], the iterative system (16) admits a fixed-point under the conditions (18). □

Remark 17. In poroelasticity (Example 3), the bilinear form a takes the special form

$$\begin{aligned} a(u,v) &= \int_\Omega \sigma(u) : \varepsilon(v) \,\mathrm{d}x \\ &= \int_\Omega \Big\{ \lambda(\nabla\cdot u)(\nabla\cdot v) + 2\mu\varepsilon(u) : \varepsilon(v) \Big\} \,\mathrm{d}x. \end{aligned}$$

Hence the operator $\mathscr{A}$ splits into two parts

$$\langle \mathscr{A} u, v\rangle = \frac{\lambda}{\alpha^2}\langle \mathscr{D}^* u, \mathscr{D}^* v\rangle + 2\mu \int_\Omega \varepsilon(u) : \varepsilon(v) \,\mathrm{d}x$$

over symmetric gradients (cf. [14, Hyp. H2]). This allows the fixed-stress and the undrained iterative methods to have contraction maps with stronger norms. Hence, there are no weak coupling conditions like (10) or (18). ♣

5 Temporal discretization

Consider a uniform partition of $[0, T]$ with time step-size $\tau := \frac{T}{N}$ for $N \in \mathbb{N}$, and partition points $t_n := n\tau$. Let the index n in the superscripts of the quantities denote the time instance and for the iterative methods, let the index i denote the iteration number. The quantities u^n and p^n denote the approximation of the functions u and p, respectively, at the time point t_n. For the iterative methods, the quantities $u^{n,i}$ and $p^{n,i}$ denote the approximation of the functions u and p, respectively, at the time point t_n after i iterations. To measure the progress of the iterations, we define the quantity

$$\begin{aligned} \mathrm{ERROR} = \|u^{n+1,i+1} - u^{n+1,i}\|_{\mathscr{V}} \\ + \|p^{n+1,i+1} - p^{n+1,i}\|_{\mathscr{H}_{\mathscr{Q}}}. \end{aligned} \tag{23}$$

A termination criterion ERROR < TOL can be defined for some prespecified tolerance TOL. Let J_n be the iteration number when iterative methods are terminated at time point t_n. We define the discrete time derivative $D_\tau u^{n+1} := (u^{n+1} - u^n)/\tau$ to approximate the continuous time derivatives $\dot{u}(t_{n+1})$. For the iterative methods, the continuous time derivative $\dot{u}^i(t_{n+1})$ is approximated by $D_\tau u^{n+1,i} = (u^{n+1,i} - u^{n,J_n})/\tau$. Furthermore, define $f^n := f(t_n)$ and $g^n := g(t_n)$. We assume that the initial data $u^0 \in \mathscr{V}$ and $p^0 \in \mathscr{H}_{\mathscr{Q}}$ is consistent in the sense of $a(u^0, v) - d(v, p^0) = \langle f^0, v\rangle$ for all $v \in \mathscr{V}$.

System (3) can be rewritten with operator matrices as

$$\begin{bmatrix} 0 & 0 \\ \mathscr{D}^* & \mathscr{C} \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} -\mathscr{A} & \mathscr{D} \\ 0 & -\mathscr{B} \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} + \begin{bmatrix} f \\ g \end{bmatrix}. \tag{24}$$

In the four time-integration methods to be introduced below, for every n in $\{0, \dots, N-1\}$, we seek (u^{n+1}, p^{n+1}) given the initial values (u^0, p^0) and the right-hand sides $\{f^n\}_{n=0}^N, \{g^n\}_{n=0}^N$.

In deriving the a priori convergence estimates, the following easily verifiable lemma is used.

Lemma 18. *On a Hilbert space $\mathscr{H}$, a symmetric and positive bilinear form $\mathfrak{a} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ defines a natural norm $\|\cdot\|_{\mathfrak{a}}^2 := \mathfrak{a}(\cdot,\cdot)$, and it holds that*

$$2\mathfrak{a}(u, u-v) = \|u\|_{\mathfrak{a}}^2 - \|v\|_{\mathfrak{a}}^2 + \|u-v\|_{\mathfrak{a}}^2$$

for all $u, v \in \mathscr{H}$.

Notation

We use the short notation $\mathscr{L}^2(\mathscr{H})$ for $\mathscr{L}^2(0,T;\mathscr{H})$. The symbol $\lesssim$ is used to appropriately absorb the constants independent of the temporal discretization parameter. It also absorbs norms of the regularized solutions of (1) (cf. [10, Prop. 2.1]).

5.1 Implicit Euler method

The implicit Euler discretization of (24) reads: for every n in $\{0,\dots,N-1\}$, given (u^n,p^n) compute (u^{n+1},p^{n+1}) as the solution of

$$\begin{bmatrix} \mathscr{A} & -\mathscr{D} \\ \mathscr{D}^* & \mathscr{C}+\tau\mathscr{B} \end{bmatrix} \begin{bmatrix} u^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} f^{n+1} \\ \tau g^{n+1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathscr{D}^* & \mathscr{C} \end{bmatrix} \begin{bmatrix} u^{n} \\ p^{n} \end{bmatrix}. \tag{25}$$

The following lemma states the solvability of system (25).

Lemma 19. *For* $(u^n,p^n) \in \mathscr{V}\times\mathscr{Q}$ *and* $(f^n,g^n)\in\mathscr{V}^*\times\mathscr{Q}^*$, *system* (25) *has a unique solution* $(u^{n+1},p^{n+1})\in\mathscr{V}\times\mathscr{Q}$.

Proof. Consider the operator matrix

$$\mathscr{M} := \begin{bmatrix} \mathscr{A} & -\mathscr{D} \\ \mathscr{D}^* & \mathscr{C}+\tau\mathscr{B} \end{bmatrix}.$$

We observe for all $(u,p)\in\mathscr{V}\times\mathscr{Q}$ that

$$\begin{aligned} \left\langle \mathscr{M}\begin{bmatrix} u \\ p \end{bmatrix}, \begin{bmatrix} u \\ p \end{bmatrix} \right\rangle &= \langle \mathscr{A}u,u\rangle - \langle \mathscr{D}p,u\rangle \\ &\quad + \langle \mathscr{D}^*u,p\rangle + \langle(\mathscr{C}+\tau\mathscr{B})p,p\rangle \\ &\geq c_{\mathrm{a}}\|u\|^2_{\mathscr{V}} + c_{\mathrm{c}}\|p\|^2_{\mathscr{H}_{\mathscr{Q}}} + \tau c_{\mathrm{b}}\|p\|^2_{\mathscr{Q}}. \end{aligned}$$

Hence, $\mathscr{M}$ is invertible for all $\tau > 0$ and system (25) is uniquely solvable for (u^{n+1},p^{n+1}). □

The implicit Euler discretization is unconditionally stable and the following a priori convergence estimate holds.

Theorem 20 ([10]). *Let* (u,p) *be the solution of the original system* (1) *given by Theorem 9 for sufficiently smooth right-hand sides* $f\colon [0,T]\to\mathscr{V}^*$, $g\colon[0,T]\to\mathscr{Q}^*$ *and consistent initial data* $u^0\in\mathscr{V}$, $p^0\in\mathscr{H}_{\mathscr{Q}}$. *Then for all* $n\leq N$, *the solution* $(u^n,p^n)\in\mathscr{V}\times\mathscr{Q}$ *of the discretized implicit system* (25) *satisfies*

$$\begin{aligned} &\|u(t_n)-u^n\|^2_{\mathscr{V}} + \|p(t_n)-p^n\|^2_{\mathscr{H}_{\mathscr{Q}}} \\ &\qquad \sum_{m=1}^{n}\tau\|p(t_m)-p^m\|^2_{\mathscr{Q}} \lesssim t_n\tau^2. \end{aligned}$$

5.2 Semi-explicit Euler method

The semi-explicit Euler discretization of (24) reads: for every n in $\{0,\dots,N-1\}$, given (u^n,p^n) solve

$$\begin{bmatrix} \mathscr{A} & 0 \\ \mathscr{D}^* & \mathscr{C}+\tau\mathscr{B} \end{bmatrix} \begin{bmatrix} u^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} f^{n+1} \\ \tau g^{n+1} \end{bmatrix} + \begin{bmatrix} 0 & \mathscr{D} \\ \mathscr{D}^* & \mathscr{C} \end{bmatrix} \begin{bmatrix} u^{n} \\ p^{n} \end{bmatrix} \tag{26}$$

for (u^{n+1},p^{n+1}). The following lemma states the solvability of (26).

Lemma 21. *For* $(u^n,p^n)\in\mathscr{V}\times\mathscr{Q}$ *and* $(f^n,g^n)\in\mathscr{V}^*\times\mathscr{Q}^*$, *system* (26) *has a unique solution* $(u^{n+1},p^{n+1})\in\mathscr{V}\times\mathscr{Q}$.

Proof. From Assumption 1, the operators $\mathscr{C}+\tau\mathscr{B}$ and $\mathscr{A}$ are invertible. Due to the lower triangular structure of the operator matrix on the left-hand side of (26), block solves are possible and hence (u^{n+1},p^{n+1}) is uniquely determined. □

The a priori convergence estimate of the semi-explicit method is derived in [2] by showing that system (3) can be approximated by a related delay differential equation system with sufficient regularity conditions imposed on the history function for well-posedness of the delay system. The implicit Euler discretization of the delay system, with delay chosen to be the step-size of the time discretization, yields the semi-explicit Euler method. Whenever the sufficient weak coupling condition

$$\omega \leq 1$$

is satisfied, the following a priori convergence estimate holds. After omitting the details of the delay system, a simplified result from [2] is presented in Theorem 22, cf. [2, Th. 3.9] for details.

Theorem 22 ([2]). *Consider system* (1) *satisfying a weak coupling condition* $\omega\leq 1$. *Let* (u,p) *be the solution of* (1) *for sufficiently smooth right-hand sides* $f\colon[0,T]\to\mathscr{V}^*$, $g\colon[0,T]\to\mathscr{Q}^*$ *and consistent initial data* $u^0\in\mathscr{V}$, $p^0\in\mathscr{H}_{\mathscr{Q}}$. *Then for all* $n\leq N$, *the solution* $(u^n,p^n)\in\mathscr{V}\times\mathscr{Q}$ *of the discretized semi-explicit system* (26) *satisfies*

$$\begin{aligned} &\|u(t_n)-u^n\|^2_{\mathscr{V}} + \|p(t_n)-p^n\|^2_{\mathscr{H}_{\mathscr{Q}}} \\ &\qquad + \sum_{m=1}^{n}\tau\|p(t_m)-p^m\|^2_{\mathscr{Q}} \lesssim e^{t_n}(1+t_n)\tau^2. \end{aligned}$$

5.3 Parabolic–elliptic iterative method

The implicit discretization of the decoupled system (8) yields

$$\begin{bmatrix}\mathscr{A} & -\mathscr{D}\\ 0 & \mathscr{C}+\tau\mathscr{B}+L_{\mathscr{F}}\mathscr{F}\end{bmatrix}\begin{bmatrix}u^{n+1,i+1}\\ p^{n+1,i+1}\end{bmatrix} = \begin{bmatrix} f^{n+1}\\ \tau g^{n+1}\end{bmatrix} + \begin{bmatrix}0 & 0\\ \mathscr{D}^* & \mathscr{C}\end{bmatrix}\begin{bmatrix}u^{n,J_n}\\ p^{n,J_n}\end{bmatrix} + \begin{bmatrix}0 & 0\\ -\mathscr{D}^* & L_{\mathscr{F}}\mathscr{F}\end{bmatrix}\begin{bmatrix}u^{n+1,i}\\ p^{n+1,i}\end{bmatrix}. \tag{27}$$

The discrete problem reads: for every n in $\{0,\dots,N-1\}$, given (u^{n,J_n},p^{n,J_n}), set $(u^{n+1,0},p^{n+1,0}) := (u^{n,J_n},p^{n,J_n})$ as initial iterates, and solve for $(u^{n+1,i+1},p^{n+1,i+1})$ from (27) with $i \geq 0$ until the termination criterion is satisfied. The solvability of (27) is given in the following lemma.

Lemma 23. *Given $(u^{n,J_n},p^{n,J_n}),(u^{n+1,i},p^{n+1,i}) \in \mathscr{V}\times\mathscr{Q}$ and $(f^n,g^n)\in\mathscr{V}^*\times\mathscr{Q}^*$, the parabolic–elliptic discrete system (27) has a unique solution*

$$(u^{n+1,i+1},p^{n+1,i+1}) \in \mathscr{V}\times\mathscr{Q}.$$

Proof. Since $L_{\mathscr{F}}\mathscr{F}$ is monotone and $\mathscr{C}+\tau\mathscr{B}$ is elliptic, $\mathscr{C}+\tau\mathscr{B}+L_{\mathscr{F}}\mathscr{F}$ is invertible. Moreover $\mathscr{A}$ is also invertible, hence a unique solution exists which can be obtained by block solves. □

The iterative block solves of (27) is presented in Algorithm 1, which is called at all time steps.

Algorithm 1 Parabolic–elliptic iterative method

```
Set the initial iterates $u^{n+1,0} = u^{n,J_n}$, $p^{n+1,0} = p^{n,J_n}$, TOL and $i = 0$.
while ERROR > TOL do
    Given $u^{n+1,i}$, $p^{n+1,i}$, $u^{n,J_n}$, $p^{n,J_n}$ compute $p^{n+1,i+1}$ from the second row of (27).
    Compute $u^{n+1,i+1}$ using $u^{n+1,i}$, $p^{n+1,i+1}$, $u^{n,J_n}, p^{n,J_n}$ from first row of (27).
    Compute ERROR from (23).
    $i \leftarrow i+1$
end while
$u^{n+1,J_{n+1}} \leftarrow u^{n+1,i}$ and $p^{n+1,J_{n+1}} \leftarrow p^{n+1,i}$
```

Theorem 24. *Let (u,p) be the solution of the original system (1) given by Theorem 9 for sufficiently smooth right-hand sides $f\colon [0,T]\to\mathscr{V}^*$, $g\colon[0,T]\to\mathscr{Q}^*$ and consistent initial data $u^0\in\mathscr{V}$, $p^0\in\mathscr{H}_{\mathscr{Q}}$. Furthermore, let the conditions of Theorem 13 hold. Then for all $n\leq N$, the solution $(u^{n,J_n},p^{n,J_n})\in\mathscr{V}\times\mathscr{Q}$ of the discretized system (27) satisfies*

$$\left\|u(t_n)-u^{n,J_n}\right\|^2_{\mathscr{V}} + \left\|p(t_n)-p^{n,J_n}\right\|^2_{\mathscr{H}_{\mathscr{Q}}} + \sum_{m=1}^{n}\tau\left\|p(t_m)-p^{m,J_m}\right\|^2_{\mathscr{Q}} \lesssim e^{\frac{1}{1+\delta}t_n}\Big(\Big(\frac{1}{\tau}+\frac{1}{\tau^2}\Big)\mathrm{TOL}^2 + t_n\tau^2\Big)$$

for some arbitrary constant $\delta > 0$.

Proof. Taking the difference between the original system (3) and the discrete system obtained after dividing the second row of (27) by τ, we get

$$a(u(t_{n+1})-u^{n+1,i+1},v) - d(v,p(t_{n+1})-p^{n+1,i+1}) = 0, \tag{28a}$$

$$\begin{aligned} &d\big(\dot u(t_{n+1}) - \frac{1}{\tau}(u^{n+1,i}-u^{n,J_n}),q\big)\\ &+c\big(\dot p(t_{n+1}) - \frac{1}{\tau}(p^{n+1,i+1}-p^{n,J_n}),q\big)\\ &+b\big(p(t_{n+1})-p^{n+1,i+1},q\big)\\ &-L_{\mathscr{F}}\langle\mathscr{F}\frac{1}{\tau}\big(p^{n+1,i+1}-p^{n+1,i}\big),q\rangle = 0. \end{aligned} \tag{28b}$$

Using the ideas of [10] and [1] for a priori convergence, we introduce the error between the field values as

$$\begin{aligned}\eta_u^{n+1,i+1} &:= u(t_{n+1}) - u^{n+1,i+1},\\ \eta_p^{n+1,i+1} &:= p(t_{n+1}) - p^{n+1,i+1},\end{aligned}$$

and the error of approximating the derivative as

$$\begin{aligned}\theta_u^{n+1} &:= u(t_{n+1}) - u(t_n) - \tau\dot u(t_{n+1}),\\ \theta_p^{n+1} &:= p(t_{n+1}) - p(t_n) - \tau\dot p(t_{n+1}).\end{aligned}$$

Observe that

$$\begin{aligned}\tau\dot u(t_{n+1}) - (u^{n+1,i}-u^{n,J_n}) &= \tau\dot u(t_{n+1}) - (u(t_{n+1})-u(t_n))\\ &\quad + (u(t_{n+1})-u(t_n))\\ &\quad - (u^{n+1,i}-u^{n,J_n})\\ &= -\theta_u^{n+1} + \eta_u^{n+1,i} - \eta_u^{n,J_n}\end{aligned}$$

and similarly

$$\tau\dot p(t_{n+1}) - (p^{n+1,i+1}-p^{n,J_n}) = -\theta_p^{n+1} + \eta_p^{n+1,i+1} - \eta_p^{n,J_n}.$$

Substituting the above in the system resulting from multiplying (28b) by τ we get

$$a(\eta_u^{n+1,i+1},v) - d(v,\eta_p^{n+1,i+1}) = 0, \tag{29a}$$

$$\begin{aligned} &d(\eta_u^{n+1,i}-\eta_u^{n,J_n},q)\\ &+c(\eta_p^{n+1,i+1}-\eta_p^{n,J_n},q) + \tau b(\eta_p^{n+1,i+1},q)\\ &+L_{\mathscr{F}}\langle\mathscr{F}\eta_p^{n+1,i+1} - \mathscr{F}\eta_p^{n+1,i},q\rangle =\\ &\qquad d(\theta_u^{n+1},q) + c(\theta_p^{n+1},q). \end{aligned} \tag{29b}$$

Adding the resulting equations from testing of (29a) with $v = \eta_u^{n+1,i} - \eta_u^{n,J_n}$ and (29b) with $q = \eta_p^{n+1,i+1}$ gives

$$\begin{aligned} a(\eta_u^{n+1,i}, \eta_u^{n+1,i} - \eta_u^{n,J_n}) + \tau b(\eta_p^{n+1,i+1}, \eta_p^{n+1,i+1}) \\ + c(\eta_p^{n+1,i+1} - \eta_p^{n,J_n}, \eta_p^{n+1,i+1}) \\ = d(\theta_u^{n+1}, \eta_p^{n+1,i+1}) + c(\theta_p^{n+1}, \eta_p^{n+1,i+1}) \\ - a(\eta_u^{n+1,i+1} - \eta_u^{n+1,i}, \eta_u^{n+1,i} - \eta_u^{n,J_n}) \\ - L_{\mathscr{F}} \langle \mathscr{F}\eta_p^{n+1,i+1} - \mathscr{F}\eta_p^{n+1,i}, \eta_p^{n+1,i+1} \rangle. \end{aligned} \tag{30}$$

Note that $\eta_u^{n+1,i+1} = \eta_u^{n+1,i} + \eta_u^{n+1,i+1} - \eta_u^{n+1,i}$ is used for the terms with the bilinear form a. For the terms on the right-hand sides of (30), using the Cauchy-Schwarz and the weighted Young's inequalities, gives

$$\begin{aligned} d(\theta_u^{n+1}, \eta_p^{n+1,i+1}) &\le \tilde{C}_{\mathrm{d}} \|\theta_u^{n+1}\|_{\mathscr{H}_{\mathscr{V}}} \|\eta_p^{n+1,i+1}\|_{\mathscr{Q}} \\ &\le \frac{\tilde{C}_{\mathrm{d}}}{\sqrt{c_{\mathrm{b}}}} \|\theta_u^{n+1}\|_{\mathscr{H}_{\mathscr{V}}} \|\eta_p^{n+1,i+1}\|_b \\ &\le \frac{1}{2\delta} \frac{\tilde{C}_{\mathrm{d}}{}^2}{c_{\mathrm{b}}} \|\theta_u^{n+1}\|^2_{\mathscr{H}_{\mathscr{V}}} + \frac{\delta}{2} \|\eta_p^{n+1,i+1}\|_b^2, \end{aligned}$$

$$\begin{aligned} c(\theta_p^{n+1}, \eta_p^{n+1,i+1}) &\le C_{\mathrm{c}} \|\theta_p^{n+1}\|_{\mathscr{H}_{\mathscr{Q}}} \|\eta_p^{n+1,i+1}\|_{\mathscr{H}_{\mathscr{Q}}} \\ &\le \frac{C_{\mathrm{c}} C_{\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}}}{\sqrt{c_{\mathrm{b}}}} \|\theta_p^{n+1}\|_{\mathscr{H}_{\mathscr{Q}}} \|\eta_p^{n+1,i+1}\|_b \\ &\le \frac{1}{2\delta} \frac{C_{\mathrm{c}}^2 C^2_{\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}}}{c_{\mathrm{b}}} \|\theta_p^{n+1}\|^2_{\mathscr{H}_{\mathscr{Q}}} \\ &\quad + \frac{\delta}{2} \|\eta_p^{n+1,i+1}\|_b^2 \end{aligned}$$

where $C_{\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}}$ is the embedding constant of the embedding $\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}$,

$$\begin{aligned} & - a(\eta_u^{n+1,i+1} - \eta_u^{n+1,i}, \eta_u^{n+1,i} - \eta_u^{n,J_n}) \\ & \quad \le \frac{1}{2} \frac{C_{\mathrm{a}}^2}{c_{\mathrm{a}}} \|\eta_u^{n+1,i+1} - \eta_u^{n+1,i}\|^2_{\mathscr{V}} + \frac{1}{2} \|\eta_u^{n+1,i} - \eta_u^{n,J_n}\|_a^2 \end{aligned}$$

and

$$\begin{aligned} & - L_{\mathscr{F}} \langle \mathscr{F}\eta_p^{n+1,i+1} - \mathscr{F}\eta_p^{n+1,i}, \eta_p^{n+1,i+1} \rangle \\ & \quad \le \frac{1}{2\delta} \frac{L_{\mathscr{F}}^2 C_{\mathscr{F}}^2}{c_{\mathrm{b}}} \|\eta_p^{n+1,i+1} - \eta_p^{n+1,i}\|^2_{\mathscr{H}_{\mathscr{Q}}} + \frac{\delta}{2} \|\eta_p^{n+1,i+1}\|_b^2. \end{aligned}$$

On using Lemma 18 for terms on the left-hand side of (30) together with the above estimates yields

$$\begin{aligned} & \|\eta_u^{n+1,i}\|_a^2 + \|\eta_u^{n+1,i} - \eta_u^{n,J_n}\|_a^2 - \|\eta_u^{n,J_n}\|_a^2 \\ & \quad + \|\eta_p^{n+1,i+1}\|_c^2 + \|\eta_p^{n+1,i+1} - \eta_p^{n,J_n}\|_c^2 - \|\eta_p^{n,J_n}\|_c^2 \\ & \quad + 2\tau \|\eta_p^{n+1,i+1}\|_b^2 \\ & \le \frac{1}{\delta} \frac{\tilde{C}_{\mathrm{d}}^2}{c_{\mathrm{b}}} \|\theta_u^{n+1}\|^2_{\mathscr{H}_{\mathscr{V}}} + \delta \|\eta_p^{n+1,i+1}\|_b^2 \\ & \quad + \frac{1}{\delta} \frac{C_{\mathrm{c}}^2 C^2_{\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}}}{c_{\mathrm{b}}} \|\theta_p^{n+1}\|^2_{\mathscr{H}_{\mathscr{Q}}} + \delta \|\eta_p^{n+1,i+1}\|_b^2 \\ & \quad + \frac{C_{\mathrm{a}}^2}{c_{\mathrm{a}}} \|\eta_u^{n+1,i+1} - \eta_u^{n+1,i}\|^2_{\mathscr{V}} + \|\eta_u^{n+1,i} - \eta_u^{n,J_n}\|_a^2 \\ & \quad + \frac{1}{\delta} \frac{L_{\mathscr{F}}^2 C_{\mathscr{F}}^2}{c_{\mathrm{b}}} \|\eta_p^{n+1,i+1} - \eta_p^{n+1,i}\|^2_{\mathscr{H}_{\mathscr{Q}}} + \delta \|\eta_p^{n+1,i+1}\|_b^2. \end{aligned} \tag{31}$$

Using the estimates

$$\begin{aligned} & \|\theta_u^{n+1}\|_{\mathscr{H}_{\mathscr{V}}} \le \tau^2 \|\ddot{u}\|_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{V}})} \text{ and} \\ & \|\theta_p^{n+1}\|_{\mathscr{H}_{\mathscr{Q}}} \le \tau^2 \|\ddot{p}\|_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{Q}})} \end{aligned}$$

taken from [1], choosing $\delta = \frac{\tau}{4}$, and droppping a positive term on the left-hand side, (31) simplifies to

$$\begin{aligned} & \|\eta_u^{n+1,i}\|_a^2 - \|\eta_u^{n,J_n}\|_a^2 + \|\eta_p^{n+1,i+1}\|_c^2 - \|\eta_p^{n,J_n}\|_c^2 \\ & \quad + \frac{5\tau}{4} \|\eta_p^{n+1,i+1}\|_b^2 \\ & \le 4 \frac{\tilde{C}_{\mathrm{d}}{}^2}{c_{\mathrm{b}}} \tau^3 \|\ddot{u}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{V}})} + 4 \frac{C_{\mathrm{c}}^2 C^2_{\mathscr{Q} \hookrightarrow \mathscr{H}_{\mathscr{Q}}}}{c_{\mathrm{b}}} \tau^3 \|\ddot{p}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{Q}})} \\ & \quad + \frac{C_{\mathrm{a}}^2}{c_{\mathrm{a}}} \|\eta_u^{n+1,i+1} - \eta_u^{n+1,i}\|^2_{\mathscr{V}} \\ & \quad + \frac{4}{\tau} \frac{L_{\mathscr{F}}^2 C_{\mathscr{F}}^2}{c_{\mathrm{b}}} \|\eta_p^{n+1,i+1} - \eta_p^{n+1,i}\|^2_{\mathscr{H}_{\mathscr{Q}}}. \end{aligned} \tag{32}$$

Observe that

$$\begin{aligned} \|\eta_u^{n+1,i+1}\|_a^2 &= \|\eta_u^{n+1,i} + (\eta_u^{n+1,i+1} - \eta_u^{n+1,i})\|_a^2 \\ &\le \frac{1+\delta}{\delta} \|\eta_u^{n+1,i}\|_a^2 \\ &\quad + (1+\delta) C_{\mathrm{a}} \|\eta_u^{n+1,i+1} - \eta_u^{n+1,i}\|^2_{\mathscr{V}} \end{aligned}$$

where the inequality follows from the triangle inequality and subsequently the weighted Young's inequality. With this observation, (32) results in

$$
\begin{aligned}
&\frac{\delta}{1+\delta}\Big(\|\eta_u^{n+1,i+1}\|_a^2-\|\eta_u^{n,J_n}\|_a^2\Big)-\frac{1}{1+\delta}\|\eta_u^{n,J_n}\|_a^2\\
&\qquad+\|\eta_p^{n+1,i+1}\|_c^2-\|\eta_p^{n,J_n}\|_c^2+\frac{5\tau}{4}\|\eta_p^{n+1,i+1}\|_b^2\\
&\le 4\frac{\tilde{C}_{\mathrm{d}}^2}{c_{\mathrm{b}}}\tau^3\|\ddot{u}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{V}})}+4\frac{C_{\mathrm{c}}^2C^2_{\mathscr{Q}\hookrightarrow\mathscr{H}_{\mathscr{Q}}}}{c_{\mathrm{b}}}\tau^3\|\ddot{p}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{Q}})}\\
&\quad+\Big(\frac{C_{\mathrm{a}}^2}{c_{\mathrm{a}}}+\delta C_{\mathrm{a}}\Big)\|\eta_u^{n+1,i+1}-\eta_u^{n+1,i}\|_{\mathscr{V}}^2\\
&\qquad+\frac{4}{\tau}\frac{L_{\mathscr{F}}^2C_{\mathscr{F}}^2}{c_{\mathrm{b}}}\|\eta_p^{n+1,i+1}-\eta_p^{n+1,i}\|^2_{\mathscr{H}_{\mathscr{Q}}}.
\end{aligned}
$$

After J_{n+1} iterations, the error given by (23) reaches below a prescribed tolerance of TOL, and we finally obtain after a renaming of the index n by l

$$
\begin{aligned}
&\frac{\delta}{1+\delta}\Big(\|\eta_u^{l+1,J_{l+1}}\|_a^2-\|\eta_u^{l,J_l}\|_a^2\Big)-\frac{1}{1+\delta}\|\eta_u^{l,J_l}\|_a^2\\
&\qquad+\|\eta_p^{l+1,J_{l+1}}\|_c^2-\|\eta_p^{l,J_l}\|_c^2+\frac{5\tau}{4}\|\eta_p^{l+1,J_{l+1}}\|_b^2\\
&\le 4\frac{\tilde{C}_{\mathrm{d}}^2}{c_{\mathrm{b}}}\tau^3\|\ddot{u}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{V}})}+4\frac{C_{\mathrm{c}}^2C^2_{\mathscr{Q}\hookrightarrow\mathscr{H}_{\mathscr{Q}}}}{c_{\mathrm{b}}}\tau^3\|\ddot{p}\|^2_{\mathscr{L}^\infty(\mathscr{H}_{\mathscr{Q}})}\\
&\quad+\Big(\frac{C_{\mathrm{a}}^2}{c_{\mathrm{a}}}+\delta C_{\mathrm{a}}+\frac{4}{\tau}\frac{L_{\mathscr{F}}^2C_{\mathscr{F}}^2}{c_{\mathrm{b}}}\Big)\mathrm{TOL}^2. \qquad (33)
\end{aligned}
$$

Summing (33) over index l from 0 to $n-1$ while applying the discrete Grönwall's inequality [9] yields

$$
\begin{aligned}
&\Big\|u(t_n)-u^{n,J_n}\Big\|_{\mathscr{V}}^2+\Big\|p(t_n)-p^{n,J_n}\Big\|^2_{\mathscr{H}_{\mathscr{Q}}}\\
&\qquad+\sum_{m=1}^{n}\tau\Big\|p(t_m)-p^{m,J_m}\Big\|^2_{\mathscr{Q}}\\
&\qquad\lesssim e^{\frac{1}{1+\delta}t_n}\Big(\Big(\tfrac{1}{\tau}+\tfrac{1}{\tau^2}\Big)\mathrm{TOL}^2+t_n\tau^2\Big)
\end{aligned}
$$

which is the sought a priori estimate. □

5.4 Elliptic–parabolic iterative method

The implicit discretization of the decoupled system (16) yields

$$
\begin{aligned}
&\begin{bmatrix}\mathscr{A}+L_{\mathscr{U}}\mathscr{D}\mathscr{U} & 0\\ \mathscr{D}^* & \mathscr{C}+\tau\mathscr{B}\end{bmatrix}\begin{bmatrix}u^{n+1,i+1}\\ p^{n+1,i+1}\end{bmatrix}\\
&\quad=\begin{bmatrix}f^{n+1}\\ \tau g^{n+1}\end{bmatrix}+\begin{bmatrix}0 & 0\\ \mathscr{D}^* & \mathscr{C}\end{bmatrix}\begin{bmatrix}u^{n,J_n}\\ p^{n,J_n}\end{bmatrix}\\
&\qquad+\begin{bmatrix}L_{\mathscr{U}}\mathscr{D}\mathscr{U} & \mathscr{D}\\ 0 & 0\end{bmatrix}\begin{bmatrix}u^{n+1,i}\\ p^{n+1,i}\end{bmatrix}. \qquad (34)
\end{aligned}
$$

The discrete problem reads: for every n in $\{0,\dots,N-1\}$, given (u^{n,J_n},p^{n,J_n}), set $(u^{n+1,0},p^{n+1,0}):=(u^{n,J_n},p^{n,J_n})$ as initial iterates, and solve for $(u^{n+1,i+1},p^{n+1,i+1})$ from (34) with $i\ge 0$ until the termination criterion is satisfied. The following lemma states the solvability of (34).

Lemma 25. *Given $(u^{n,J_n},p^{n,J_n}),(u^{n+1,i},p^{n+1,i})\in\mathscr{V}\times\mathscr{Q}$ and $(f^n,g^n)\in\mathscr{V}^*\times\mathscr{Q}^*$, the elliptic–parabolic discrete system (34) has a unique solution $(u^{n+1,i+1},p^{n+1,i+1})\in\mathscr{V}\times\mathscr{Q}$.*

Proof. Since, $L_{\mathscr{U}}\mathscr{D}\mathscr{U}$ is monotone, $\mathscr{A}+L_{\mathscr{U}}\mathscr{D}\mathscr{U}$ is invertible. A unique solution is given by block solves of (34). □

The iterative block solves is presented in Algorithm 2, which is called at all time steps. The a priori convergence estimate of the elliptic–parabolic iterative method is similar to that of the parabolic–elliptic iterative method and is stated below without proof.

Algorithm 2 Elliptic–parabolic iterative method

```
Set the initial iterates $u^{n+1,0}=u^{n,J_n}$, $p^{n+1,0}=p^{n,J_n}$, TOL and $i=0$.
while ERROR > TOL do
    Given $u^{n+1,i}$, $p^{n+1,i}$, $u^{n,J_n}$, $p^{n,J_n}$ compute $u^{n+1,i+1}$ from the first row of (34).
    Compute $p^{n+1,i+1}$ using $p^{n+1,i}$, $u^{n+1,i+1}$, $u^{n,J_n}$, $p^{n,J_n}$ from second row of (34).
    Compute ERROR from (23).
    $i\leftarrow i+1$
end while
$u^{n+1,J_{n+1}}\leftarrow u^{n+1,i}$ and $p^{n+1,J_{n+1}}\leftarrow p^{n+1,i}$
```

Theorem 26. *Let (u,p) be the solution of the original system (1) given by Theorem 9 for sufficiently smooth right-hand sides $f\colon[0,T]\to\mathscr{V}^*$, $g\colon[0,T]\to\mathscr{Q}^*$ and consistent initial data $u^0\in\mathscr{V}$, $p^0\in\mathscr{H}_{\mathscr{Q}}$. Furthermore, let the conditions of Theorem 16 hold. Then for all $n\le N$, the solution $(u^{n,J_n},p^{n,J_n})\in\mathscr{V}\times\mathscr{Q}$ of the discretized system (34) satisfies*

$$
\begin{aligned}
&\Big\|u(t_n)-u^{n,J_n}\Big\|_{\mathscr{V}}^2+\Big\|p(t_n)-p^{n,J_n}\Big\|^2_{\mathscr{H}_{\mathscr{Q}}}\\
&\qquad+\sum_{m=1}^{n}\tau\Big\|p(t_m)-p^{m,J_m}\Big\|^2_{\mathscr{Q}}\\
&\qquad\lesssim t_n\tau^2+\mathrm{TOL}^2.
\end{aligned}
$$

The proof of Theorem 26 is essentially the same as that of Theorem 24.

6 Numerical experiments

For demonstrating the convergence in time for the four methods, we consider the simulation of the following two poroelasticity problems, on a square domain $\Omega:=(0,1)^2$ and final time $T:=1$.

Test case 1

This test case is similar to the one considered in [1]. The source terms and the initial condition are chosen as

$$
\begin{aligned}
f &\equiv 0,\\
g(t) &= 30\sin(\pi x)\mathrm{e}^{-t},\\
p^0(x,y) &= 50\,x(1-x)\,y(1-y).
\end{aligned}
$$

Test case 2

This is a manufactured problem with the analytical solution

$$
u(t,x,y) = -\frac{e^{-C_{\mathrm{p}}t}}{2\pi}\begin{bmatrix}10x(1-x)y(1-y)\\10x(1-x)y(1-y)\end{bmatrix}, \tag{35a}
$$

$$
p(t,x,y) = e^{-C_{\mathrm{p}}t}10x(1-x)y(1-y) \tag{35b}
$$

where $C_{\mathrm{p}} = \frac{2\pi^2 M}{M\alpha+1}\frac{\kappa}{\nu}$ incorporates all the poroelasticity parameters.

The values of the poroelasticity parameters in test case 1 and test case 2 are given in Table 1. The choice of the parameters in test case 1 is such that it describes a flow-dominated problem with an estimate for the coupling strength $\sqrt{\omega} = 1.8 \times 10^{-9}$. In test case 2, the values chosen are such that the coupling strength is sufficiently high but satisfies the weak coupling condition, *i.e.*, $\sqrt{\omega} = 0.18$.

Table 1 – Poroelasticity parameters

Parameter	Test case 1	Test case 2
λ	7.8×10^{8}	0.5
μ	1.8×10^{9}	1.25×10^{-1}
$\frac{\kappa}{\nu}$	8.0×10^{-10}	5.0×10^{-2}
$\frac{1}{M}$	1.4×10^{-10}	3.7
α	0.85	0.75

For the FEM discretization in space on a shape regular triangular mesh, we use P_2 finite elements for the displacement and P_1 for the pressure. For the iterative methods, we set the maximum number of iterations MAX_ITER := 20 with a specified tolerance TOL := 1×10^{-5} for the iterative termination criterion at each time step for both the test cases. Note that to make the errors with respect to time dominate, TOL is chosen such that TOL $< \min(\tau, \tau^2)$.

The popular FEM computing platform *FEniCS* is used for the implementation. For test case 1, a reference solution is computed by the implicit Euler method with $\tau = 2^{-8}$ and a spatial discretization parameter $h = 2^{-8}$. Fixing the spatial discretization parameter $h = 2^{-8}$, the relative errors for the field values at final time $T = 1$ for $\tau \in \{2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}\}$ are computed in different norms. The relative errors for the four methods in relevant norms are plotted in Figure 1 for test case 1 and Figure 2 for test case 2, respectively. It can be seen that in both the test cases, a convergence of first-order in time is observed for the displacement as well as the pressure measured in respective $\mathscr{H}^1$ and $\mathscr{L}^2$ norms.

For test case 1, an average of two iterations are required for both the parabolic–elliptic and the elliptic–parabolic iterative methods irrespective of discretization parameters. However, in test case 2, an average of 5 and 6 iterations are required for the parabolic–elliptic and the elliptic–parabolic iterative methods, respectively, for coarser time discretizations. The average number of iterations reduce to 4 and 5, respectively for finer time discretizations.

Finally, in Table 2, we note the run-time comparison for the four methods when not using the specialized preconditioners for the decoupled block solves. Naturally, since in the semi-explicit method, smaller decoupled systems are solved non-iteratively, it has about 1.8 times speed-up over the monolithic implicit method. However, in the iterative methods, since a number of iterations are required at each time step, the total run-times are larger than the implicit method.

7 Summary

In this article, we have shown the existence of fixed-points for the iterative methods for the abstract coupled linear elliptic–parabolic PDE system under certain weak coupling conditions. We have derived the a priori convergence result for the parabolic–elliptic iterative decoupling method. The a priori convergence result of the elliptic–parabolic iterative method is similar, hence is stated without proof. All four time integration methods considered in this article show first-order convergence in time.

Code Availability: The commented source code and the input scripts are available under:

10.14464/gammas.v4i1.500.

Acknowledgements: The author gratefully acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project 467107679 and is thankful for support by the Stuttgart Center for Simulation Science (SimTech). The author would like to thank his supervisors R. Altmann and B. Unger for their guidance and encouragement. The author would also like to thank the reviewers

Table 2 – Runtime comparison for test case 2. The run-times T_i for the implicit method is in *seconds*, while the run-times for other methods is a multiple of T_i.

τ	implicit (T_i)	semi-explicit ($\times T_i$)	parabolic–elliptic ($\times T_i$)	elliptic–parabolic ($\times T_i$)
2^{-2}	1.21×10^2	0.56	2.83	3.43
2^{-3}	2.42×10^2	0.56	2.84	3.40
2^{-4}	4.83×10^2	0.57	2.85	2.86
2^{-5}	9.70×10^2	0.56	2.30	2.84
2^{-6}	1.93×10^3	0.57	2.30	2.85
2^{-7}	3.88×10^3	0.56	2.30	2.30

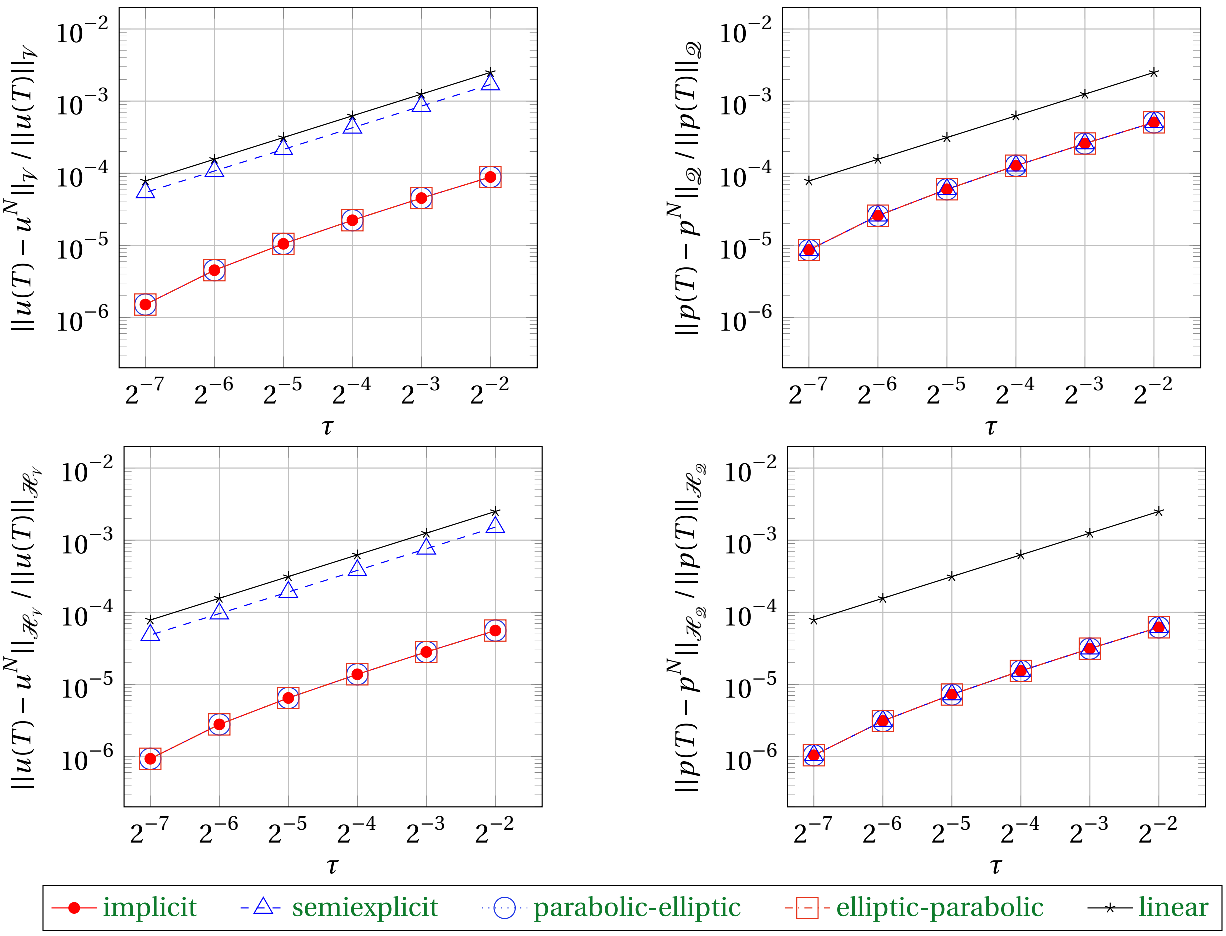

Figure 1 – Test case 1: Relative errors at final time $T = 1$ for the implicit, the semi-explicit, the parabolic–elliptic, and the elliptic–parabolic methods for a fixed spatial mesh size $h = 2^{-8}$ and a reference time mesh size $\tau = 2^{-8}$. Left: displacement u. Right: pressure p.

R. Maier and C. Torres for providing constructive feedback that helped improve the article.

References

[1] R. Altmann and R. Maier. A decoupling and linearizing discretization for poroelasticity with nonlinear permeability. *to appear in SIAM J. Sci. Comput. 2022.*

[2] R. Altmann, R. Maier, and B. Unger. Semi-explicit discretization schemes for weakly-coupled elliptic-parabolic problems. *Math. Comp.*, 90:1089–1118, 2021.

[3] R. Altmann, R. Maier, and B. Unger. A semi-explicit integration scheme for weakly-coupled poroelasticity with nonlinear permeability. *PAMM*, 20(1), January 2021. ISSN 1617-7061, 1617-7061.

[4] M. A. Biot. General theory of three-dimensional consolidation. *J. Appl. Phys.*, 12(2):155–164, 1941.

[5] M. A. Biot. Thermoelasticity and Irreversible Thermodynamics. *J. Appl. Phys.*, 27(3):240–253, 1956.

[6] A. F. Bower. *Applied Mechanics of Solids*. CRC Press, New York, October 2009. ISBN 9781439802489.

[7] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York ; London, 2011. ISBN 9780387709130.

[8] R. M. Brooks and K. Schmitt. The contraction mapping princi-

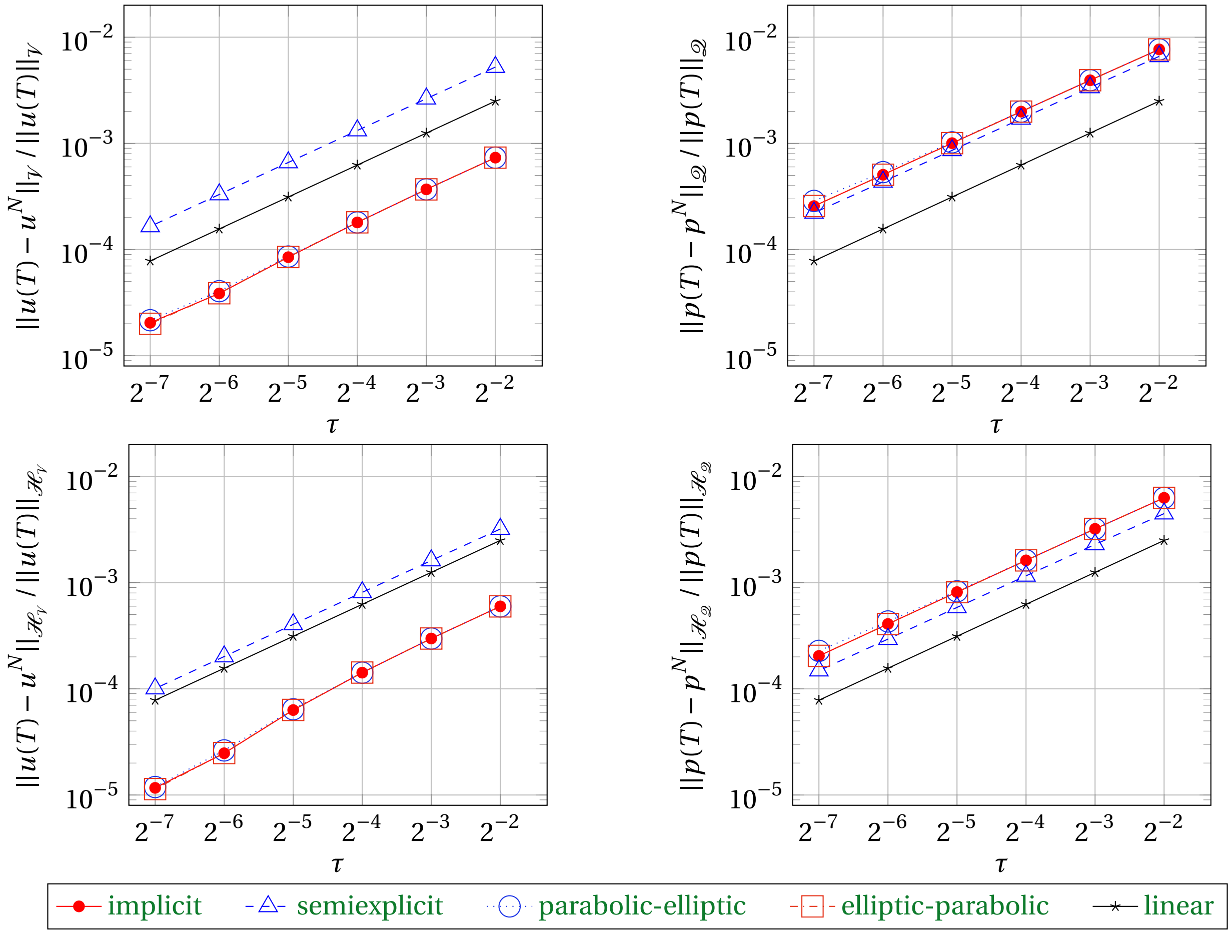

Figure 2 – Test case 2: Relative errors at final time $T = 1$ for the implicit, the semi-explicit, the parabolic–elliptic, and the elliptic–parabolic methods for a fixed spatial mesh size $h = 2^{-8}$. Left: displacement u. Right: pressure p.

ple and some applications. *Electron. J. Differ. Eq.*, 2009. ISSN 1072-6691.

[9] E. Emmrich. Discrete versions of Gronwall's lemma and their application to the numerical analysis of parabolic problems. *Preprint 637, Technische Universität Berlin, Germany*, 1999.

[10] A. Ern and S. Meunier. A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems. *ESAIM: Math. Model. Numer. Anal.*, 43(2):353–375, 2009.

[11] L. C. Evans. *Partial differential equations.* American Mathematical Society, Providence, RI, USA, 2010. ISBN 9780821849743 0821849743.

[12] J. Kim, H. A. Tchelepi, and R. Juanes. Stability and convergence of sequential methods for coupled flow and geomechanics: fixed-stress and fixed-strain splits. *Comput. Methods Appl. Mech. Engrg.*, 200(13-16):1591–1606, 2011.

[13] J. L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*, volume I. Springer, Berlin, Heidelberg, 1972.

[14] A. Mikelić and M. F. Wheeler. Convergence of iterative coupling for coupled flow and geomechanics. *Comput. Geosci.*, 17(3): 455–461, 2013.

[15] E. Zeidler. *Nonlinear functional analysis and its applications. 2A: Linear monotone operators.* Springer, New York Berlin Heidelberg, 1990.