BIT Numerical Mathematics (2024) 64:20
https://doi.org/10.1007/s10543-024-01021-0

Semi-explicit integration of second order for weakly coupled poroelasticity

R. Altmann[1] · R. Maier[2] · B. Unger[3]

Received: 17 May 2023 / Accepted: 13 March 2024 / Published online: 7 April 2024

Abstract

We introduce a semi-explicit time-stepping scheme of second order for linear poroelasticity satisfying a weak coupling condition. Here, semi-explicit means that the system, which needs to be solved in each step, decouples and hence improves the computational efficiency. The construction and the convergence proof are based on the connection to a differential equation with two time delays, namely one and two times the step size. Numerical experiments confirm the theoretical results and indicate the applicability to higher-order schemes.

Keywords Poroelasticity · Elliptic–parabolic problem · Semi-explicit time discretization · Delay · Backward differentiation formula

Mathematics Subject Classification 65M12 · 65L80 · 76S05

Communicated by Christian Lubich.

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 467107679. B. Unger additionally acknowledges support by the Stuttgart Center for Simulation Science (SimTech). Moreover, major parts of this work were carried out while R. Altmann was affiliated with the Institute of Mathematics and the Centre for Advanced Analytics and Predictive Sciences (CAAPS) at the University of Augsburg.

✉ R. Altmann
robert.altmann@ovgu.de

R. Maier
roland.maier@kit.edu

B. Unger
benjamin.unger@simtech.uni-stuttgart.de

[1] Institute of Analysis and Numerics, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

[2] Institute for Applied and Numerical Mathematics, Karlsruhe Institute of Technology, Englerstr.2, 76131 Karlsruhe, Germany

[3] Stuttgart Center for Simulation Science (SC SimTech), University of Stuttgart, Universitätsstr.32, 70569 Stuttgart, Germany

1 Introduction

This paper is devoted to the construction and analysis of a semi-explicit time discretization scheme of second order for linear poroelasticity [16, 27]. The poroelastic equations can be characterized as a coupled system consisting of an elliptic and a parabolic equation and appear, e.g., in the field of geomechanics [10, 34]. In many applications, this coupling is rather weak in a certain sense (cf. (2.4) and (3.9) below as well as the typical poroelastic parameters stated in [16, p. 25]), which is also a central assumption in this paper to guarantee convergence. For the temporal discretization of elliptic–parabolic problems such as poroelasticity, one mainly considers *implicit* schemes such as the implicit Euler method [17] or higher-order schemes [19]. This is primarily due to the fact that a semi-discretization in space yields a differential–algebraic equation for which *explicit* time-stepping schemes cannot be used [23].

Then again, one is interested in a decoupled approach, in which the elliptic and parabolic equations can be solved sequentially. Such a decoupling does not only replace the solution of a large system by two smaller subsystems to be solved but also enables the application of standard preconditioners [24]. Moreover, the decoupling of the systems favors a co-design paradigm, allowing the usage of highly optimized software packages for the porous media flow (the parabolic equation) and the mechanical problem (the elliptic equation) separately, and, in addition, includes a linearization step if the permeability depends on the displacement, cf. [4]. One attempt in this direction are iterative decoupling methods such as the *fixed-stress*, *fixed-strain*, or *drained* splitting schemes; see, e.g., [7, 22, 25, 32]. These schemes come along with an additional inner iteration in each time step that is required to guarantee convergence [28] and, additionally, require a careful selection of tuning parameters. In [12], an alternative method based on an additional stabilization term rather than an inner iteration is proposed. Although first-order convergence in time is observed in experiments, the theory presented in [12, 13] only guarantees suboptimal convergence of order $1/2$. Moreover, an extension to a higher-order method is by far not intuitive. Similarly, extensions of the aforementioned iterative schemes would require many additional inner iterations to guarantee the prescribed accuracy, counteracting the aim of an efficient numerical method.

To combine the advantages of the monolithic and iterative coupling methods, a *semi-explicit* time-stepping scheme was introduced in [5], which decouples the equations and does not require an additional inner iteration or stabilization parameters. For a comparison of this method with the monolithic and different decoupling strategies, we refer to [26]. We emphasize that the semi-explicit scheme equals the implicit Euler discretization up to a term with a time shift in one of the equations. The perception of this scheme in terms of delay equations allows proving convergence of the method if a weak coupling condition is satisfied. This condition is independent of the step size and can be quantified explicitly.

In this paper, we extend these ideas to construct and analyze a novel higher-order decoupling time integrator for coupled elliptic–parabolic problems, which include linear poroelasticity as a special case. To the best of our knowledge, this is the first time that a rigorous convergence analysis for a higher-order decoupling time discretization scheme is presented. For the construction of our scheme, we follow the general strategy developed in [5] and first construct a nearby delay system, which is then discretized in time. Recognizing that the first-order semi-explicit scheme analyzed in [5] can be understood as zeroth-order Taylor expansion, a straightforward approach would combine higher-order Taylor expansions (for the construction of the delay system) with higher-order time-integration schemes. The resulting delay equation, however, would be of advanced type, such that a sufficient regularity of the solution cannot be guaranteed as indicated in Sect. 3.1. Instead, we proceed with an expansion including multiple delays and use a backward differentiation formula (BDF) for the time discretization of the resulting delay equation. Our main contributions are:

– A BDF-type expansion to construct a delay equation that differs with a given order from the original elliptic–parabolic problem; cf. Theorem 3.1. This results in multiple delays in the first equation that enable the decoupling of the equations without the requirement for an inner iteration or additional tuning parameters.
– A convergence proof for the second-order case in Theorems 3.2 and 4.1. For this, we solely work with the delayed parabolic equation that is obtained by resolving the elliptic equation and suitably adapt ideas from [5]. Moreover, we point out how this approach can be extended to higher orders.

As in the first-order case, our method depends on a weak coupling condition, which we explicitly quantify via the theory of delay differential–algebraic equations in Sect. 3.5. We emphasize that the coupling strength of the two equations is also of relevance for the iterative decoupling methods mentioned earlier in the sense that they require more inner iterations if the coupling is stronger. Hence, they become inefficient for strongly coupled problems.

Since we focus on time discretization, the whole convergence analysis is given on operator level, i.e., without a spatial discretization. Corresponding results for the fully discrete scheme can be obtained by the introduction of appropriate Ritz projections, cf. [3, 5]. We conclude our presentation with three numerical examples in Sect. 5.

Notation

We write $a \lesssim b$ to indicate the existence of a generic constant C, independent of spatial and temporal discretization parameters, such that $a \leq Cb$.

2 Poroelastic equations

In this section, we introduce the equations of linear poroelasticity and the corresponding abstract formulation as an elliptic–parabolic problem. We consider a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^d$, $d \in \{2, 3\}$. We seek the displacement field $u\colon [0, T] \times \Omega \to \mathbb{R}^d$ and the pore pressure $p\colon [0, T] \times \Omega \to \mathbb{R}$. For a given time horizon $T > 0$, the system equations read

$$-\nabla \cdot \sigma(u) + \nabla(\alpha p) = f \qquad \text{in } (0, T] \times \Omega, \tag{2.1a}$$

$$\partial_t\Big(\alpha \nabla \cdot u + \frac{1}{M} p\Big) - \nabla \cdot \Big(\frac{\kappa}{\nu} \nabla p\Big) = g \qquad \text{in } (0, T] \times \Omega \tag{2.1b}$$

together with initial conditions

$$u(0) = u^0, \qquad p(0) = p^0. \tag{2.1c}$$

Therein, σ denotes the stress tensor

$$\sigma(u) = \mu \left(\nabla u + (\nabla u)^T\right) + \lambda\, (\nabla \cdot u)\ \mathrm{id}$$

with Lamé coefficients λ and μ, the permeability κ, the Biot-Willis fluid-solid coupling coefficient α, the Biot modulus M, and the fluid viscosity ν; see [10, 27]. Since some of these coefficients play a central role for the analysis of our scheme, we report the coefficients for a selection of different materials in Table 1. The right-hand sides f and g are the volumetric load and the fluid source, respectively, modeling an injection or production process. Throughout this paper, we assume homogeneous Dirichlet boundary conditions, i.e., we set $u = 0$ and $p = 0$ on $(0, T] \times \partial\Omega$.

2.1 Abstract formulation

For an abstract formulation of (2.1a), we introduce the Hilbert spaces

$$\mathscr{V} := [H_0^1(\Omega)]^d, \qquad \mathscr{H}_{\mathscr{V}} := [L^2(\Omega)]^d, \qquad \mathscr{Q} := H_0^1(\Omega), \qquad \mathscr{H}_{\mathscr{Q}} := L^2(\Omega)$$

which include the assumed Dirichlet boundary conditions. With the respective dual spaces of $\mathscr{V}$ and $\mathscr{Q}$ denoted by $\mathscr{V}^*$ and $\mathscr{Q}^*$, $(\mathscr{V}, \mathscr{H}_{\mathscr{V}}, \mathscr{V}^*)$ as well as $(\mathscr{Q}, \mathscr{H}_{\mathscr{Q}}, \mathscr{Q}^*)$ form Gelfand triples with dense embeddings; see [33, Ch. 23.4] for more details. Moreover, we define the bilinear forms

$$a(u, v) := \int_\Omega \sigma(u) : \varepsilon(v)\,\mathrm{d}x, \qquad b(p, q) := \int_\Omega \frac{\kappa}{\nu}\, \nabla p \cdot \nabla q\,\mathrm{d}x,$$
$$c(p, q) := \int_\Omega \frac{1}{M}\, p\, q\,\mathrm{d}x, \qquad d(u, q) := \int_\Omega \alpha\, (\nabla \cdot u)\, q\,\mathrm{d}x$$

with the classical double dot notation, i.e., for matrices $A, B \in \mathbb{R}^{n\times m}$ we have $A : B = \mathrm{trace}(A^T B)$, and the symmetric gradient $\varepsilon(u){:} = \frac{1}{2}(\nabla u + (\nabla u)^T)$ used in continuum mechanics. With this, the weak formulation of (2.1a) can be written as follows: seek $u\colon [0, T] \to \mathscr{V}$ and $p\colon [0, T] \to \mathscr{Q}$ such that

$$a(u, v) - d(v, p) = \langle f, v\rangle, \tag{2.2a}$$
$$d(\dot{u}, q) + c(\dot{p}, q) + b(p, q) = \langle g, q\rangle \tag{2.2b}$$

for all test functions $v \in \mathscr{V}$, $q \in \mathscr{Q}$. Correspondingly, we assume that the right-hand sides satisfy $f\colon [0, T] \to \mathscr{V}^*$ and $g\colon [0, T] \to \mathscr{Q}^*$ and denote with $\langle\cdot, \cdot\rangle$ the respective duality pairings. We would like to emphasize that it is sufficient to prescribe initial data for p, since equation (2.2a) defines a consistency condition for p^0 and u^0

(which is uniquely solvable for u^0; see the forthcoming discussion on the properties of the bilinear forms).

System (2.2) may also be written in operator form in the dual spaces of $\mathscr{V}$ and $\mathscr{Q}$. For this, let $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, and $\mathscr{D}$ denote the operators corresponding to the bilinear forms a, b, c, and d, respectively. Then, (2.2) is equivalent to

$$\begin{aligned} \mathscr{A}u - \mathscr{D}^* p &= f \qquad &&\text{in } \mathscr{V}^*, \\ \mathscr{D}\dot{u} + \mathscr{C}\dot{p} + \mathscr{B}p &= g \qquad &&\text{in } \mathscr{Q}^*. \end{aligned}$$

It remains to discuss the properties of the bilinear forms. The bilinear form $a\colon \mathscr{V} \times \mathscr{V} \to \mathbb{R}$ is symmetric, elliptic, and bounded, i.e., there exist positive constants c_a, C_a such that

$$a(u, u) \geq c_a \, \|u\|^2_{\mathscr{V}}, \qquad a(u, v) \leq C_a \, \|u\|_{\mathscr{V}} \|v\|_{\mathscr{V}}$$

for all $u, v \in \mathscr{V}$. We would like to emphasize that a is well known from the theory of linear elasticity and that the ellipticity follows from Korn's inequality [14, Th. 6.3.4]. Similarly, $b\colon \mathscr{Q} \times \mathscr{Q} \to \mathbb{R}$ is symmetric, elliptic, and bounded in $\mathscr{Q}$, i.e., there exist positive constants c_b, C_b such that

$$b(p, p) \geq c_b \, \|p\|^2_{\mathscr{Q}}, \qquad b(p, q) \leq C_b \, \|p\|_{\mathscr{Q}} \|q\|_{\mathscr{Q}}$$

for all $p, q \in \mathscr{Q}$. The bilinear form $c\colon \mathscr{H}_{\mathscr{Q}} \times \mathscr{H}_{\mathscr{Q}} \to \mathbb{R}$ simply involves the multiplication by a (positive) constant and, hence, defines an inner product in the pivot space $\mathscr{H}_{\mathscr{Q}}$. In more detail, there exist positive constants c_c, C_c such that

$$c(p, p) \geq c_c \, \|p\|^2_{\mathscr{H}_{\mathscr{Q}}}, \qquad c(p, q) \leq C_c \, \|p\|_{\mathscr{H}_{\mathscr{Q}}} \|q\|_{\mathscr{H}_{\mathscr{Q}}}$$

for all $p, q \in \mathscr{H}_{\mathscr{Q}}$. The remaining bilinear form $d\colon \mathscr{V} \times \mathscr{H}_{\mathscr{Q}} \to \mathbb{R}$ models the coupling and is continuous, i.e., there exists a positive constant C_d such that

$$d(u, p) \leq C_d \, \|u\|_{\mathscr{V}} \|p\|_{\mathscr{H}_{\mathscr{Q}}}$$

for all $u \in \mathscr{V}$ and $p \in \mathscr{H}_{\mathscr{Q}}$.

Remark 2.1 System (2.2) can also be used to model linear thermoelasticity, which considers the displacement of a material due to temperature changes [11].

More generally, system (2.2) is an elliptic–parabolic system, where the elliptic part (modeled by a) and the parabolic part (modeled by b and c) are coupled through the bilinear form d. We emphasize that the forthcoming analysis does not depend on the specific application, but only on the properties of the bilinear forms introduced above.

2.2 Spatial discretization

Although this paper is mainly concerned with the temporal discretization, we shortly comment on the finite element discretization of (2.2). For more details, we refer to

[17]. In order to transfer the convergence results of this paper to the fully discrete system, one may consider spatial projection operators corresponding to the elliptic bilinear forms a and b; see [5].

Considering finite-dimensional subspaces $V_h \subseteq \mathscr{V}$ and $Q_h \subseteq \mathscr{Q}$, one seeks approximations $u_h \approx u$ and $p_h \approx p$. Here, the parameter h represents the mesh size of the triangulation used in the construction of V_h and Q_h. A direct spatial discretization of (2.2) then leads to the differential–algebraic equation

$$\begin{bmatrix} 0 & 0 \\ D & M_c \end{bmatrix} \begin{bmatrix} \dot{u}_h \\ \dot{p}_h \end{bmatrix} = \begin{bmatrix} -K_a & D^T \\ 0 & -K_b \end{bmatrix} \begin{bmatrix} u_h \\ p_h \end{bmatrix} + \begin{bmatrix} f_h \\ g_h \end{bmatrix}.$$

Therein, K_a and K_b denote the stiffness matrices corresponding to the bilinear forms a and b, respectively. Due to the assumptions discussed above, K_a and K_b can be assumed to be symmetric and positive definite. Moreover, M_c equals the mass matrix corresponding to c, which is thus also symmetric and positive definite, and D is a rectangular matrix corresponding to d.

Using standard P_1 finite elements to define V_h and Q_h, one obtains the expected convergence rates of order one in the energy norms and order two in the L^2-norms. For more precise results also on higher-order approximations, we again refer to [17].

2.3 Temporal discretization of first order

The standard way to discretize system (2.2) in time is the application of the implicit Euler scheme. This results in a time-stepping scheme of order one as shown in [17].

As already mentioned in the introduction, the differential–algebraic structure rules out the possibility of a fully explicit discretization in time. In [5], however, a semi-explicit scheme was introduced. Considering an equidistant decomposition of $[0, T]$ with step size τ, this scheme reads

$$a(u^{n+1}, v) - d(v, p^n) = \langle f^{n+1}, v \rangle, \tag{2.3a}$$

$$\tfrac{1}{\tau}\, d(u^{n+1} - u^n, q) + \tfrac{1}{\tau}\, c(p^{n+1} - p^n, q) + b(p^{n+1}, q) = \langle g^{n+1}, q \rangle \tag{2.3b}$$

for all $v \in \mathscr{V}$, $q \in \mathscr{Q}$. Here, u^n and p^n denote the approximations of $u(t^n)$ and $p(t^n)$, $t^n = n\tau$, respectively. Note that, in contrast to the implicit Euler discretization, the first equation contains p^n rather than p^{n+1}. Hence, the two equations decouple and can be solved sequentially. It is shown in [5] that this maintains the first-order convergence as long as the *weak coupling condition*

$$\alpha^2 M \le \mu + \lambda \tag{2.4}$$

is satisfied. Note that this condition is specific to the equations of poroelasticity. For general elliptic–parabolic problems defined via the bilinear forms a, b, c, and d, the weak coupling conditions reads $C_d^2 \le c_a c_c$.

For the convergence analysis, the connection of system (2.2) to a related delay system is used. This idea is also applied in the following sections to construct a semi-explicit scheme of order two, leading to a more restrictive weak coupling condition.

3 Semi-explicit integration scheme of second order

This section is devoted to the extension of the semi-explicit scheme (2.3) to second order. Following the idea in [5], we first construct a related delay equation and then discretize the delay equation with an implicit scheme of second order. For this, we need to replace the pressure in the first poroelastic equation by a time-delayed term which is second-order accurate. We first consider a Taylor expansion before we introduce discrete derivatives, leading to a system with multiple delays.

In the following, we consider a uniform partition of the time interval $[0, T]$ with step size $\tau > 0$ such that $N := T/\tau \in \mathbb{N}$. Hence, we consider time points $t^n = n\tau$ for $n = 0, \dots, N$. The approximation of a function y at time t^n is then denoted by y^n.

3.1 Related delay systems by Taylor expansion

In a first attempt, we aim to decouple system (2.2) by replacing p in the first equation by a Taylor expansion. An expansion of order k at time $t - \tau$ yields the delay system

$$a(\bar{u}, v) - d\big(v, \textstyle\sum_{j=0}^{k-1} \frac{\tau^j}{j!}\, \Delta_\tau \bar{p}^{(j)}\big) = \langle f, v\rangle, \tag{3.1a}$$

$$d(\dot{\bar{u}}, q) + c(\dot{\bar{p}}, q) + b(\bar{p}, q) = \langle g, q\rangle \tag{3.1b}$$

for test functions $v \in \mathscr{V}$ and $q \in \mathscr{Q}$. As initial condition, we set $\bar{p}(0) = p(0) = p^0$. In contrast to the original system, however, one also needs a so-called *history function* $\bar{\Phi}(t) = \bar{p}\big|_{[-\tau,0]}(t)$. Since system (3.1) is constructed by the help of a Taylor expansion, it is no surprise that the solutions (u, p) and $(\bar{u}, \bar{p})$ only differ by a term of order τ^k as long as the solution of the delay systems stays stable. We refer to Appendix A for further details.

Nevertheless, system (3.1) is not well-suited for the construction of a numerical scheme. This is due to the appearance of temporal derivatives. Already in the case of interest, namely $k = 2$, the resulting delay system (3.1) is of *advanced* type [9] and, hence, only well-posed in a distributional setting [30]. Moreover, the solution loses regularity over time. We refer to [6, 8, 31] for further details. To avoid such advanced delay systems, we consider an alternative approach and replace the derivatives by discrete derivatives. This then leads to multiple delays.

3.2 Discrete derivatives

Based on the discrete difference operator $Dy^{n+1} \colon = y^{n+1} - y^n$, we can write discrete derivatives of any order in a short way. As an example, the difference quotients of order one and two read

$$\frac{1}{\tau} Dy^{n+1} = \frac{y^{n+1} - y^n}{\tau}, \qquad \frac{1}{\tau}\bigl(Dy^{n+2} + \tfrac{1}{2} D^2 y^{n+2}\bigr) = \frac{3y^{n+2} - 4y^{n+1} + y^n}{2\tau},$$

leading, e.g., to the well-known BDF schemes. For the upcoming analysis, it is convenient to extend the definition of the difference operator also to continuous functions. More precisely, we define $Dy \colon = y - \Delta_\tau y$ with the time shift Δ_τ introduced above. The resulting order-k approximations of the derivative of a function are summarized in the following lemma.

Lemma 3.1 (discrete derivative) *Let $y \in C^k([0, T])$ and $t \in [k\tau, T]$. Then, it holds that*

$$\dot{y}(t) = \frac{1}{\tau} \sum_{j=1}^{k} \tfrac{1}{j}\, D^j\, y(t) + \mathscr{O}(\tau^k).$$

Moreover, if we have $y \in C^{k+1}([0, T])$, then there exist constants $c_1, \dots, c_k$ such that the error term can be written as $\sum_{j=1}^{k} c_j \int_{t-j\tau}^{t} (t - \xi)^j\, y^{(j+1)}(\xi)\, \mathrm{d}\xi$.

3.3 Related delay system with multiple delays

As already mentioned, we want to replace the derivatives in (3.1) by discrete derivatives. Focusing on the case $k = 2$, which will lead to a scheme of second order, we replace $\tau \Delta_\tau \dot{\bar{p}}$ by $D\Delta_\tau \bar{p} = \Delta_\tau \bar{p} - \Delta_{2\tau} \bar{p}$. This leads to the system

$$a(\tilde{u}, v) - d\bigl(v, 2\Delta_\tau \tilde{p} - \Delta_{2\tau} \tilde{p}\bigr) = \langle f, v\rangle, \tag{3.2a}$$

$$d(\dot{\tilde{u}}, q) + c(\dot{\tilde{p}}, q) + b(\tilde{p}, q) = \langle g, q\rangle \tag{3.2b}$$

for all test functions $v \in \mathscr{V}$ and $q \in \mathscr{Q}$. Note that this is a system with two delays, namely τ and 2τ. Again, we need to discuss the initial data, which includes $\tilde{p}(0) = p(0) = p^0$ and an appropriate history function $\tilde{\Phi}$ defined on $[-2\tau, 0]$. To obtain consistency in u, namely $\tilde{u}(0) = u(0) = u^0$, we assume that $\tilde{\Phi} \in C^\infty([-2\tau, 0], \mathscr{Q})$ satisfies

$$p^0 = 2\Delta_\tau \tilde{p}(0) - \Delta_{2\tau} \tilde{p}(0) = 2\,\tilde{\Phi}(-\tau) - \tilde{\Phi}(-2\tau). \tag{3.3}$$

Remark 3.1 *(approximations of higher order)* For general $k \geq 1$, one possibility is to replace the derivatives $\Delta_\tau \bar{p}^{(j)}$ in (3.1) by approximations of order τ^{k-j}. This then guarantees that the resulting expression is an approximation of the Taylor expansion $\sum_{j=0}^{k-1} \frac{\tau^j}{j!}\, \Delta_\tau \bar{p}^{(j)}$ of order k. Note, however, that this leads to a growing number of delays. For $k = 3$ this yields three delays, whereas $k = 4$ already needs five delays. The resulting scheme for $k = 3$ is presented in Sect. 5.3.

Remark 3.2 *(parabolic equation with multiple delays)* Considering the operator formulation of (3.2) and eliminating the variable $\tilde{u}$ by the first equation, we get

$$\mathscr{C}\dot{\tilde{p}} + \mathscr{B}\tilde{p} + \mathscr{D}\mathscr{A}^{-1}\mathscr{D}^*(2\Delta_\tau\dot{\tilde{p}} - \Delta_{2\tau}\dot{\tilde{p}}) = g - \mathscr{D}\mathscr{A}^{-1}\dot{f}. \tag{3.4}$$

Note that this is a parabolic equation (of neutral type) with two delays. Hence, we consider here multiple delays rather than higher derivatives.

Motivated by the approximation properties of the Taylor expansion approach, the following theorem shows that the solutions to (2.2) and (3.2) only differ by a term of order two.

Theorem 3.1 *Assume sufficiently smooth right-hand sides f and g and a history function $\tilde{\Phi}$ satisfying (3.3). Then, the solutions to (2.2) and (3.2) are equal up to a term of order τ^2, i.e., for almost all $t \in [0, T]$ we have*

$$\|\tilde{p}(t) - p(t)\|^2_{\mathscr{Q}} + \|\tilde{u}(t) - u(t)\|^2_{\mathscr{V}} \lesssim t\,\tau^4.$$

Here, the hidden constant depends on higher derivatives of the history function $\tilde{\Phi}$ as well as of $\tilde{p}$.

Proof We define $e_p := \tilde{p} - p$ and $e_u := \tilde{u} - u$. Due to the assumptions on the history function, we conclude that $e_p(0) = 0$ and $e_u(0) = 0$. Considering the difference of (3.2a) and (2.2a), we obtain

$$a(e_u, v) - d(v, e_p) = -d(v, \tilde{p} - 2\Delta_\tau\tilde{p} + \Delta_{2\tau}\tilde{p}) \le \tau^2\, C_d\, \|v\|_{\mathscr{V}}\, \|\ddot{\tilde{p}}\|_{L^\infty(-2\tau,T;\mathscr{H}_{\mathscr{Q}})},$$

where $L^\infty(-2\tau, T; \mathscr{H}_{\mathscr{Q}})$ denotes the Bochner space on the time interval $(-2\tau, T)$ with values in $\mathscr{H}_{\mathscr{Q}}$. In the same manner, we obtain by the derivatives of (3.2a) and (2.2a) that

$$a(\dot{e}_u, v) - d(v, \dot{e}_p) \le \tau^2\, C_d\, \|v\|_{\mathscr{V}}\, \|\tilde{p}^{(3)}\|_{L^\infty(-2\tau,T;\mathscr{H}_{\mathscr{Q}})}.$$

Now we can proceed as in the proof of Proposition A. 1, i.e., we consider the test function $v = \dot{e}_u$ in combination with the difference of (3.2b) and (2.2b). □

Remark 3.3 The hidden constant in Theorem 3.1 may become arbitrarily large depending on the ellipticity and continuity constants. This is discussed in more detail in Sect. 3.5.

System (3.2) yields a good starting point for the construction of higher-order discretization schemes. This is subject of the following subsection.

3.4 Semi-explicit integration scheme

In order to obtain a semi-explicit time-stepping scheme, we now apply the BDF-2 scheme to (3.2). To shorten notation, we introduce

$$\mathrm{BDF}_2 u^{n+2} := 3u^{n+2} - 4u^{n+1} + u^n = 2Du^{n+2} + D^2 u^{n+2}.$$

By Lemma 3.1, we know that $\frac{1}{2\tau}\mathrm{BDF}_2 u(t) = \dot{u}(t) + \mathscr{O}(\tau^2)$. Since the first equation does not contain any derivatives, the temporal discretization is simply given by a function evaluation at time t^{n+2} (as for the implicit Euler scheme). This discretization yields the semi-explicit scheme

$$a(u^{n+2}, v) - d\big(v, 2p^{n+1} - p^n\big) = \langle f^{n+2}, v\rangle, \tag{3.5a}$$
$$\tfrac{1}{2\tau} d\big(\mathrm{BDF}_2 u^{n+2}, q\big) + \tfrac{1}{2\tau} c\big(\mathrm{BDF}_2\, p^{n+2}, q\big) + b(p^{n+2}, q) = \langle g^{n+2}, q\rangle \tag{3.5b}$$

for test functions $v \in \mathscr{V}$ and $q \in \mathscr{Q}$. Note that this is a 2-step scheme, calling for initial data $p^0 = p(0)$ and p^1. In place of the history function, we set $p^{-2}, p^{-1} \in \mathscr{Q}$ such that

$$p^0 = 2\,p^{-1} - p^{-2}, \qquad p^1 = 2\,p^0 - p^{-1}. \tag{3.6}$$

Note that this is well-posed, since p^0 and p^1 are given. The first condition corresponds to (3.3) and gives the consistency condition for u^0. The second equation ensures that p^1 and u^1 are consistent. To be precise, this means that the resulting values $u^0, u^1 \in \mathscr{V}$ satisfy

$$a(u^0, v) - d\big(v, p^0\big) = \langle f^0, v\rangle, \qquad a(u^1, v) - d\big(v, p^1\big) = \langle f^1, v\rangle \tag{3.7}$$

for all $v \in \mathscr{V}$.

The proposed scheme (3.5) is indeed semi-explicit, since the first equation defines u^{n+2} purely by already computed values, i.e., without the knowledge of p^{n+2}. Inserting this value in the second equation, we then obtain the approximation p^{n+2}.

Remark 3.4 Scheme (3.5) may also be regarded as an implicit–explicit (α, β, γ)-BDF method; see [1, 2, 15, 18]. The particular scheme fits to the characteristic polynomials

$$\alpha(\zeta) = \tfrac{3}{2}\zeta^2 - 2\zeta + \tfrac{1}{2}, \qquad \beta(\zeta) = \zeta^2, \qquad \gamma(\zeta) = 2\zeta - 1,$$

where α is used for the discretization of the derivatives, γ for the extrapolated value (which is p in the first equation), and β for the remaining terms.

In operator form, scheme (3.5) reads

$$\mathscr{A}u^{n+2} - \mathscr{D}^*\big(2p^{n+1} - p^n\big) = f^{n+2},$$
$$\mathscr{D}\big(3u^{n+2} - 4u^{n+1} + u^n\big) + \mathscr{C}\big(3p^{n+2} - 4p^{n+1} + p^n\big) + 2\tau\mathscr{B}p^{n+2} = 2\tau g^{n+2}.$$

Using once more the invertibility of the operator $\mathscr{A}$, we can eliminate the u-variables in the second equation, leading to

$$\begin{aligned}&\mathscr{C}\big(3p^{n+2} - 4p^{n+1} + p^n\big) + 2\tau\mathscr{B}p^{n+2}\\ &\quad+\mathscr{D}\mathscr{A}^{-1}\mathscr{D}^*\big(6p^{n+1} - 11p^n + 6p^{n-1} - p^{n-2}\big)\\ &\quad= 2\tau g^{n+2} - \mathscr{D}\mathscr{A}^{-1}\big(3f^{n+2} - 4f^{n+1} + f^n\big).\end{aligned} \tag{3.8}$$

We would like to emphasize that this equals the BDF-2 discretization of the delay equation (3.4). This fact will be used in the following convergence result.

Theorem 3.2 (Second-order convergence of the semi-explicit scheme) *Assume sufficiently smooth right-hand sides f and g. Moreover, let the operators satisfy the* weak coupling condition

$$\omega := \frac{\alpha^2 M}{\mu + \lambda} \le \frac{1}{5}. \tag{3.9}$$

Then the semi-explicit scheme (3.5) converges with order two. More precisely, given $p^0 = p(0)$ and p^1 as a second-order approximation of $p(\tau)$, we can define consistent u^0 and u^1 in the sense of (3.7) such that

$$\|u(t^n) - u^n\|^2_{\mathscr{V}} + \|p(t^n) - p^n\|^2_{\mathscr{H}_\mathscr{Q}} + \tau\sum_{j=1}^{n}\|p(t^j) - p^j\|^2_{\mathscr{Q}} \lesssim t^n\tau^4$$

for all $n \ge 0$.

Proof Given p^0 and p^1, we define p^{-2} and p^{-1} satisfying (3.6) such that u^0, u^1 are consistent. Moreover, let $\tilde{\Phi}$ be a history function with $\tilde{\Phi}(-2\tau) = p^{-2}$ and $\tilde{\Phi}(-\tau) = p^{-1}$ such that (3.3) is satisfied. We can now apply Theorem 3.1 and conclude that the exact solution and the solution of the delay system (3.2) only differ by a term of order two. Hence, it is sufficient to compare the discrete solution given by (3.5) with $(\tilde{p}, \tilde{u})$.

We have seen that the presented semi-explicit scheme corresponds to the BDF-2 method applied to the delay equation (3.4). Since the operator $\mathscr{C}$ only contains a multiplicative factor, we may consider a simple rescaling leading to the question of the convergence of the BDF-2 scheme applied to the delay system

$$\dot{\tilde{p}} + \widetilde{\mathscr{B}}\tilde{p} + \widetilde{\mathscr{C}}\left(2\Delta_\tau\dot{\tilde{p}} - \Delta_{2\tau}\dot{\tilde{p}}\right) = r := \mathscr{C}^{-1}g - \mathscr{C}^{-1}\mathscr{D}\mathscr{A}^{-1}\dot{f} \tag{3.10}$$

with $\widetilde{\mathscr{B}} := \mathscr{C}^{-1}\mathscr{B}\colon \mathscr{V} \to \mathscr{V}^*$ and $\widetilde{\mathscr{C}} := \mathscr{C}^{-1}\mathscr{D}\mathscr{A}^{-1}\mathscr{D}^*\colon \mathscr{H}_\mathscr{Q} \to \mathscr{H}_\mathscr{Q}^*$. Note that these two operators are symmetric, elliptic, and continuous in the respective spaces. Moreover, the continuity constant of $\widetilde{\mathscr{C}}$ equals ω and is bounded by $1/5$ by assumption. This condition makes Theorem 4.1 of the following section applicable, providing an estimate of the form

$$\|\tilde{p}(t^n) - p^n\|^2_{\mathscr{H}_\mathscr{Q}} + \tau\sum_{j=1}^{n}\|\tilde{p}(t^j) - p^j\|^2_{\mathscr{Q}} \lesssim t^n\,\tau^4 + t^n E_{\mathrm{rhs}} + E_{\mathrm{init}}$$

for $n \geq 2$. The right-hand side error E_{rhs} appears because the approximation of the right-hand side in (3.10) involves a BDF-2 approximation of the term $\dot{f}$ rather than the nodal evaluation; see (3.8). However, due to Lemma 3.1, E_{rhs} is of order $\mathscr{O}(\tau^4)$. Further, with the assumption on p^1, we have

$$E_{\text{init}} = \|\tilde{p}(\tau) - p^1\|^2 + \tau\, \|\tilde{p}(\tau) - p^1\|_{\tilde{b}}^2 \lesssim \tau^4.$$

Hence, the above estimate holds for all $n \geq 0$ and leads to an overall error of order two. Moreover, considering the difference of equations (3.2a) and (3.5a), we get by the ellipticity of the bilinear form a that

$$\|\tilde{u}(t^{n+2}) - u^{n+2}\|_{\mathscr{V}} \lesssim 2\, \|\tilde{p}(t^{n+1}) - p^{n+1}\|_{\mathscr{H}_{\mathscr{Q}}} + \|\tilde{p}(t^n) - p^n\|_{\mathscr{H}_{\mathscr{Q}}}$$

for $n \geq 0$. Finally, due to the consistency conditions for u^0 and u^1, we further get $u(0) = \tilde{u}(0) = u^0$ and

$$\|u(\tau) - u^1\|_{\mathscr{V}} \lesssim \|p(\tau) - p^1\|_{\mathscr{H}_{\mathscr{Q}}}.$$

The combination of the previous estimates completes the proof. □

Remark 3.5 *(Initial data)* In practice, appropriate initial conditions can be realized as follows: given p^0, one first computes u^0 consistent to (2.2a). Then, p^1 and u^1 can be obtained by a single step of the implicit Euler discretization applied to (2.2). This then guarantees consistency as well as the needed accuracy for p^1. This follows, for instance, from the proof of [17, Thm. 3.1], where a slight adaptation shows the second-order rate in time for the first time step. This particularly includes the setting with a spatial discretization as well.

Before we discuss the convergence of the semi-explicit scheme, we focus on the weak coupling condition (3.9) and its meaning in terms of delay equations.

3.5 Weak coupling condition and asymptotic stability of the delay system

First, let us emphasize that there are several poroelasticity problems reported in the literature that satisfy the weak coupling condition (3.9), or almost satisfy the weak coupling condition; see Table 1. The latter will be relevant as well as the following discussion demonstrates.

To see that the weak coupling condition is not a mere technical assumption, we analyze the asymptotic stability of the related delay system constructed in Sect. 3.3 with multiple delays. To simplify the presentation, we consider here the finite-dimensional case after a semi-discretization in space (cf.Sect. 2.2), and study the neutral delay differential equation corresponding to (3.4), i.e., we study the neutral delay equation

$$\dot{\tilde{p}}_h + M_c^{-1} D K_a^{-1} D^T \left(2\Delta_\tau \dot{\tilde{p}}_h - \Delta_{2\tau} \dot{\tilde{p}}_h\right) + M_c^{-1} K_b \tilde{p}_h = \tilde{g}_h. \tag{3.11}$$

Table 1 Poroelasticity problems reported in the literature and their relation to the weak coupling condition (3.9). All examples consider porous media in combination with water [16, Tab. 4]

porous media	λ	μ	α	M	κ/ν	ω
Tennessee marble	$2.40 \cdot 10^{10}$	$2.4 \cdot 10^{10}$	0.19	$1.16 \cdot 10^{11}$	$1.0 \cdot 10^{-19}$	0.09
Charcoal granite	$2.23 \cdot 10^{10}$	$1.9 \cdot 10^{10}$	0.27	$8.50 \cdot 10^{10}$	$1.0 \cdot 10^{-19}$	0.15
Weber sandstone	$5.14 \cdot 10^{9}$	$1.2 \cdot 10^{10}$	0.64	$2.79 \cdot 10^{10}$	$1.0 \cdot 10^{-15}$	0.45
Westerly granite	$1.50 \cdot 10^{10}$	$1.5 \cdot 10^{10}$	0.47	$7.64 \cdot 10^{10}$	$4.0 \cdot 10^{-19}$	0.56
Berea sandstone	$4.00 \cdot 10^{9}$	$6.0 \cdot 10^{9}$	0.79	$1.23 \cdot 10^{10}$	$1.9 \cdot 10^{-13}$	0.76
Ruhr sandstone	$4.11 \cdot 10^{9}$	$1.3 \cdot 10^{10}$	0.65	$4.05 \cdot 10^{10}$	$2.0 \cdot 10^{-16}$	1.00

A necessary condition (cf. [20, Thm. 3.20]) for the delay-independent asymptotic stability of the unforced (i.e., $\tilde{g}_h = 0$) delay equation (3.11) is that the spectral radius of the matrix

$$N_2 := \begin{bmatrix} -2M_c^{-1} D K_a^{-1} D^T & M_c^{-1} D K_a^{-1} D^T \\ I & 0 \end{bmatrix}$$

is strictly less than one, i.e., $\rho(N_2) < 1$. Hereby, I denotes the identity matrix of suitable dimension. We thus have to compute the eigenvalues of N_2. Since M_c is symmetric and positive definite, the (principle) square root $M_c^{1/2}$ exists and is symmetric and positive definite. Thus, the matrix $M_c^{-1/2} D K_a^{-1} D^T M_c^{-1/2}$ is symmetric and hence diagonalizable, i.e., there exists a diagonal matrix Λ and an orthogonal matrix U such that

$$U M_c^{-1/2} D K_a^{-1} D^T M_c^{-1/2} U^{-1} = \Lambda.$$

Define $\Xi := \operatorname{diag}(U M_c^{1/2}, U M_c^{1/2})$. Then,

$$\Xi N_2 \Xi^{-1} = \begin{bmatrix} -2\Lambda & \Lambda \\ I & 0 \end{bmatrix}.$$

For any eigenvalue λ_Λ of Λ, it thus suffices to compute the spectral radius of the matrix

$$\begin{bmatrix} -2\lambda_\Lambda & \lambda_\Lambda \\ 1 & 0 \end{bmatrix},$$

which is given by $\lambda_\Lambda + \sqrt{\lambda_\Lambda^2 + \lambda_\Lambda}$. Since this is a monotone expression, we conclude

$$\rho(N_2) = \rho(M_c^{-1} D K_a^{-1} D^T) + \sqrt{\rho(M_c^{-1} D K_a^{-1} D^T)^2 + \rho(M_c^{-1} D K_a^{-1} D^T)},$$

and thus $\rho(N_2) < 1$ if and only if $\rho(M_c^{-1} D K_a^{-1} D^T) < \frac{1}{3}$. Consequently, we cannot expect the delay equation to be a reasonable approximation of the non-delay equation if $\rho(M_c^{-1} D K_a^{-1} D^T) > \frac{1}{3}$. In fact, in the scalar case, it is easy to see that $\rho(M_c^{-1} D K_a^{-1} D^T) < \frac{1}{3}$ is also a sufficient condition for delay-independent asymptotic stability. Using

$$\frac{\alpha^2 M}{\mu + \lambda} \leq \rho(M_c^{-1} D K_a^{-1} D^T) < \tfrac{1}{3},$$

we observe that a weak coupling condition as in (3.9) is not only a technical requirement, but indeed necessary for convergence. We discuss the details in the error analysis in the next section.

4 Convergence analysis

In this section, we prove the convergence of the BDF-2 method applied to the delay operator equation

$$\dot{z} + \widetilde{\mathscr{B}} z + \widetilde{\mathscr{C}} \big(2\Delta_\tau \dot{z} - \Delta_{2\tau} \dot{z} \big) = r. \tag{4.1}$$

Here, $\widetilde{\mathscr{B}} \colon \mathscr{Q} \to \mathscr{Q}^*$ is an operator with the same properties as $\mathscr{B}$ in the previous section and $\widetilde{\mathscr{C}} \colon \mathscr{H}_{\mathscr{Q}} \to \mathscr{H}_{\mathscr{Q}}^*$ is an operator with the same properties as $\mathscr{C}$ with continuity constant ω. Similar as in Sect. 3.3, we assume, besides the initial condition $z(0) = z^0$, a given history function $\Phi \in C^\infty([-2\tau, 0], \mathscr{Q})$ with $\Phi(0) = z^0$. Moreover, the right-hand side $r \colon [0, T] \to \mathscr{Q}^*$ is sufficiently smooth.

For the error analysis, we first present the following lemma. Note that the equality presented therein is closely connected to the G-stability of BDF-2; see [21, Ch. V.6].

Lemma 4.1 *For a symmetric and positive bilinear form* $\mathfrak{a}$ *it holds that*

$$2\,\mathfrak{a}(z^{n+2}, \mathrm{BDF}_2 z^{n+2}) = \mathrm{BDF}_2\, \|z^{n+2}\|_{\mathfrak{a}}^2 + 2\,\|Dz^{n+2}\|_{\mathfrak{a}}^2 - 2\,\|Dz^{n+1}\|_{\mathfrak{a}}^2 + \|D^2 z^{n+2}\|_{\mathfrak{a}}^2$$

with $\|\cdot\|_{\mathfrak{a}} := \sqrt{\mathfrak{a}(\cdot,\cdot)}$.

Proof Using multiple applications of the formula

$$2\,\mathfrak{a}(x, x-y) = \|x\|_{\mathfrak{a}}^2 - \|y\|_{\mathfrak{a}}^2 + \|x-y\|_{\mathfrak{a}}^2, \tag{4.2}$$

we get

$$\begin{aligned}
&2\,\mathfrak{a}(z^{n+2}, 3z^{n+2} - 4z^{n+1} + z^n)\\
&\quad= 2\,\mathfrak{a}(z^{n+2}, 3Dz^{n+2} - Dz^{n+1})\\
&\quad= 4\,\mathfrak{a}(z^{n+2}, Dz^{n+2}) + 2\,\mathfrak{a}(z^{n+2}, Dz^{n+2} - Dz^{n+1})\\
&\quad= 4\,\mathfrak{a}(z^{n+2}, Dz^{n+2}) + 2\,\mathfrak{a}(Dz^{n+2}, Dz^{n+2} - Dz^{n+1})\\
&\qquad - 2\,\mathfrak{a}(z^{n+1}, Dz^{n+1}) - 2\,\mathfrak{a}(z^{n+1}, z^{n+1} - z^{n+2})\\
&\quad= 3\,\|z^{n+2}\|_{\mathfrak{a}}^2 - 4\,\|z^{n+1}\|_{\mathfrak{a}}^2 + \|z^n\|_{\mathfrak{a}}^2 + 2\,\|Dz^{n+2}\|_{\mathfrak{a}}^2\\
&\qquad - 2\,\|Dz^{n+1}\|_{\mathfrak{a}}^2 + \|Dz^{n+2} - Dz^{n+1}\|_{\mathfrak{a}}^2,
\end{aligned}$$

which completes the proof. □

After this preparation, we are now able to formulate the main convergence theorem.

Theorem 4.1 (Convergence of BDF-2 for the delay equation (4.1)) *Let* $\widetilde{\mathscr{B}} \colon \mathscr{Q} \to \mathscr{Q}^*$ *and* $\widetilde{\mathscr{C}} \colon \mathscr{H}_{\mathscr{Q}} \to \mathscr{H}_{\mathscr{Q}}^*$ *be symmetric, elliptic, and continuous in the respective spaces. Moreover, let* ω *denote the continuity constant of* $\widetilde{\mathscr{C}}$ *satisfying* $\omega \le 1/5$. *Then, the BDF-2 scheme applied to (4.1), i.e., the scheme*

$$\mathrm{BDF}_2\, z^{n+2} + 2\tau\, \widetilde{\mathscr{B}} z^{n+2} + 2\,\widetilde{\mathscr{C}}\,\mathrm{BDF}_2\, z^{n+1} - \widetilde{\mathscr{C}}\,\mathrm{BDF}_2\, z^n = 2\tau\, \tilde{r}^{n+2}$$

yields an approximation of second order, provided that $\tilde r^{n+2}$ is a second-order approximation of $r(t^{n+2})$. To be precise, assuming a sufficiently smooth right-hand side r, a step size $\tau \le 1$, and initial data $z^0 = z(0)$, $z^{-1} = \Phi(-\tau)$, $z^{-2} = \Phi(-2\tau)$, we get

$$\|z(t^n) - z^n\|^2_{\mathscr{H}_{\mathscr{Q}}} + \tau \sum_{j=2}^{n} \|z(t^j) - z^j\|^2_{\mathscr{Q}} \lesssim \; t^n\tau^4 + t^n E_{\mathrm{rhs}} + E_{\mathrm{init}}$$

for $n \ge 2$, where $E_{\mathrm{init}} := \|z(\tau) - z^1\|^2_{\mathscr{H}_{\mathscr{Q}}} + \tau\, \|z(\tau) - z^1\|^2_{\mathscr{Q}}$ contains the initial error and $E_{\mathrm{rhs}} := \max_{j=2,\dots,n} \|r(t^j) - \tilde r^j\|^2_{\mathscr{H}_{\mathscr{Q}}}$ the right-hand side error.

Proof Inserting the exact solution of (4.1) within the numerical scheme, we obtain the defect equation

$$\begin{aligned}
&\tfrac{1}{2\tau}\mathrm{BDF}_2 z(t^{n+2}) + \widetilde{\mathscr{B}} z(t^{n+2}) + \tfrac{1}{\tau}\widetilde{\mathscr{C}}\big(\mathrm{BDF}_2 z(t^{n+1})\big) - \tfrac{1}{2\tau}\widetilde{\mathscr{C}}\big(\mathrm{BDF}_2 z(t^{n})\big) \\
&\quad = \tfrac{1}{2\tau}\mathrm{BDF}_2 z(t^{n+2}) - \dot z(t^{n+2}) + \tfrac{1}{\tau}\widetilde{\mathscr{C}}\big(\mathrm{BDF}_2 z(t^{n+1})\big) \\
&\qquad - 2\,\widetilde{\mathscr{C}}\dot z(t^{n+1}) - \tfrac{1}{2\tau}\widetilde{\mathscr{C}}\big(\mathrm{BDF}_2 z(t^{n})\big) + \widetilde{\mathscr{C}}\dot z(t^n) + r(t^{n+2}) \\
&\quad =: d^{n+2} + r(t^{n+2})
\end{aligned}$$

with $d^{n+2} = \mathscr{O}(\tau^2)$ by Lemma 3.1. With $e^n := z(t^n) - z^n$, we get

$$\mathrm{BDF}_2 e^{n+2} + 2\tau\, \widetilde{\mathscr{B}} e^{n+2} + 2\,\widetilde{\mathscr{C}}\,\mathrm{BDF}_2 e^{n+1} - \widetilde{\mathscr{C}}\,\mathrm{BDF}_2 e^{n} = 2\tau \tilde d^{n+2} \qquad (4.3)$$

with $\tilde d^{n+2} = d^{n+2} + r(t^{n+2}) - \tilde r^{n+2}$. Note that, due to the assumptions on the history function and the initial data, we have $e^{-2} = e^{-1} = e^0 = 0$. In the following, we write $\|\bullet\|_{\tilde b}$ for the norm induced by the operator $\widetilde{\mathscr{B}}$, which is equivalent to the $\mathscr{Q}$-norm, and $\|\bullet\|_{\tilde c}$ for the norm induced by $\widetilde{\mathscr{C}}$. Note that the latter is equivalent to the $\mathscr{H}_{\mathscr{Q}}$-norm with $\|\bullet\|^2_{\tilde c} \le \omega\,\|\bullet\|^2$, where we use the short notation $\|\bullet\| := \|\bullet\|_{\mathscr{H}_{\mathscr{Q}}}$.

Step 1: In the first step, we derive an auxiliary estimate for differences of e^n. If we multiply (4.3) by 2 and apply De^{n+2}, we get

$$\begin{aligned}
&\underbrace{2\,\langle \mathrm{BDF}_2 e^{n+2}, De^{n+2}\rangle}_{=:\,T_1} + \underbrace{4\tau\,\langle \widetilde{\mathscr{B}} e^{n+2}, De^{n+2}\rangle}_{=:\,T_2} \\
&\quad + \underbrace{4\,\langle \widetilde{\mathscr{C}}\,\mathrm{BDF}_2 e^{n+1}, De^{n+2}\rangle}_{=:\,T_3} - \underbrace{2\,\langle \widetilde{\mathscr{C}}\,\mathrm{BDF}_2 e^{n}, De^{n+2}\rangle}_{=:\,T_4} = 4\tau\,\langle \tilde d^{n+2}, De^{n+2}\rangle \qquad (4.4)
\end{aligned}$$

Reformulating the terms T_1 and T_2 using (4.2) yields

$$T_1 = 4\,\|De^{n+2}\|^2 + 2\,\langle D^2 e^{n+2}, De^{n+2}\rangle = 5\,\|De^{n+2}\|^2 - \|De^{n+1}\|^2 + \|D^2 e^{n+2}\|^2$$

and

$$T_2 = 2\tau\,\big(\|e^{n+2}\|^2_{\tilde b} - \|e^{n+1}\|^2_{\tilde b} + \|De^{n+2}\|^2_{\tilde b}\big).$$

With $De^{n+2} = De^{n+1} + D^2e^{n+2}$, we have

$$\begin{aligned} T_3 &= 8\,\langle \widetilde{\mathscr{C}}\,De^{n+1}, De^{n+2}\rangle + 4\,\langle \widetilde{\mathscr{C}}\,D^2e^{n+1}, De^{n+2}\rangle \\ &= 8\,\langle \widetilde{\mathscr{C}}\,De^{n+1}, De^{n+1} + D^2e^{n+2}\rangle + 4\,\langle \widetilde{\mathscr{C}}\,D^2e^{n+1}, De^{n+1} + D^2e^{n+2}\rangle \\ &= 8\,\|De^{n+1}\|_{\tilde{c}}^2 + 2\,\|De^{n+1}\|_{\tilde{c}}^2 - 2\,\|De^{n}\|_{\tilde{c}}^2 + 2\,\|D^2e^{n+1}\|_{\tilde{c}}^2 \\ &\quad + 8\,\langle \widetilde{\mathscr{C}}\,De^{n+1}, D^2e^{n+2}\rangle + 4\,\langle \widetilde{\mathscr{C}}\,D^2e^{n+1}, D^2e^{n+2}\rangle \end{aligned}$$

and

$$T_4 \le 4\,\|De^{n}\|_{\tilde{c}}\|De^{n+2}\|_{\tilde{c}} + 2\,\|D^2e^{n}\|_{\tilde{c}}\|De^{n+2}\|_{\tilde{c}}.$$

The above computations inserted in (4.4) yield

$$\begin{aligned} &4\,\|De^{n+2}\|^2 + \|D^2e^{n+2}\|^2 + 2\tau\,\|De^{n+2}\|_{\tilde{b}}^2 + 8\,\|De^{n+1}\|_{\tilde{c}}^2 + 2\,\|D^2e^{n+1}\|_{\tilde{c}}^2 \\ &\quad + \|De^{n+2}\|^2 - \|De^{n+1}\|^2 + 2\tau\,\|e^{n+2}\|_{\tilde{b}}^2 - 2\tau\,\|e^{n+1}\|_{\tilde{b}}^2 + 2\,\|De^{n+1}\|_{\tilde{c}}^2 - 2\,\|De^{n}\|_{\tilde{c}}^2 \\ &\le 4\tau\,\langle \tilde{d}^{n+2}, De^{n+2}\rangle + 8\,\|De^{n+1}\|_{\tilde{c}}\|D^2e^{n+2}\|_{\tilde{c}} + 4\,\|D^2e^{n+1}\|_{\tilde{c}}\|D^2e^{n+2}\|_{\tilde{c}} \\ &\quad + 4\,\|De^{n}\|_{\tilde{c}}\|De^{n+2}\|_{\tilde{c}} + 2\,\|D^2e^{n}\|_{\tilde{c}}\|De^{n+2}\|_{\tilde{c}} \\ &\le 4\tau\,\|\tilde{d}^{n+2}\|^2 + \tau\,\|De^{n+2}\|^2 + 4\delta\,\|De^{n+1}\|_{\tilde{c}}^2 + \tfrac{4}{\delta}\omega\,\|D^2e^{n+2}\|^2 + 2\gamma\,\|D^2e^{n+1}\|_{\tilde{c}}^2 \\ &\quad + \tfrac{2}{\gamma}\omega\,\|D^2e^{n+2}\|^2 + \tfrac{2}{\alpha}\,\|De^{n}\|_{\tilde{c}}^2 + 2\alpha\omega\,\|De^{n+2}\|^2 + \tfrac{1}{\beta}\|D^2e^{n}\|_{\tilde{c}}^2 + \beta\omega\,\|De^{n+2}\|^2, \end{aligned}$$

where we use the weighted Young inequality four times with positive constants $\alpha, \beta, \gamma, \delta$. Rearranging terms leads to

$$\begin{aligned} &(4 - \tau - 2\alpha\omega - \beta\omega)\,\|De^{n+2}\|^2 + (1 - \tfrac{2}{\gamma}\omega - \tfrac{4}{\delta}\omega)\,\|D^2e^{n+2}\|^2 + 2\tau\,\|De^{n+2}\|_{\tilde{b}}^2 \\ &\quad +(10 - 4\delta)\,\|De^{n+1}\|_{\tilde{c}}^2 - (2 + \tfrac{2}{\alpha})\,\|De^{n}\|_{\tilde{c}}^2 \\ &\quad +(2 - 2\gamma)\,\|D^2e^{n+1}\|_{\tilde{c}}^2 - \tfrac{1}{\beta}\,\|D^2e^{n+1}\|_{\tilde{c}}^2 \\ &\quad +\|De^{n+2}\|^2 - \|De^{n+1}\|^2 + 2\tau\,\|e^{n+2}\|_{\tilde{b}}^2 - 2\tau\,\|e^{n+1}\|_{\tilde{b}}^2 \\ &\le 4\tau\,\|\tilde{d}^{n+2}\|^2. \end{aligned} \tag{4.5}$$

We now set $\alpha = 7/8$, $\beta = 11/2$, $\gamma = 10/11$, and $\delta = 10/7$. This leads to

$$\begin{aligned} &(4 - \tau - \tfrac{29}{4}\omega)\,\|De^{n+2}\|^2 + (1 - 5\omega)\,\|D^2e^{n+2}\|^2 + 2\tau\,\|De^{n+2}\|_{\tilde{b}}^2 \\ &\quad + \tfrac{30}{7}\,\|De^{n+1}\|_{\tilde{c}}^2 - \tfrac{30}{7}\,\|De^{n}\|_{\tilde{c}}^2 + \tfrac{2}{11}\,\|D^2e^{n+1}\|_{\tilde{c}}^2 - \tfrac{2}{11}\,\|D^2e^{n+1}\|_{\tilde{c}}^2 \\ &\quad + \|De^{n+2}\|^2 - \|De^{n+1}\|^2 + 2\tau\,\|e^{n+2}\|_{\tilde{b}}^2 - 2\tau\,\|e^{n+1}\|_{\tilde{b}}^2 \\ &\le 4\tau\,\|\tilde{d}^{n+2}\|^2. \end{aligned}$$

Assuming $\omega \leq 1/5$ and $\tau \leq 1$, we therefore get

$$\begin{aligned}
&\|De^{n+2}\|^2 + 2\tau\,\|De^{n+2}\|^2_{\tilde{b}} + \|De^{n+2}\|^2 - \|De^{n+1}\|^2 + 2\tau\,\|e^{n+2}\|^2_{\tilde{b}} - 2\tau\,\|e^{n+1}\|^2_{\tilde{b}}\\
&\qquad + \tfrac{30}{7}\,\|De^{n+1}\|^2_{\tilde{c}} - \tfrac{30}{7}\,\|De^{n}\|^2_{\tilde{c}} + \tfrac{2}{11}\,\|D^2e^{n+1}\|^2_{\tilde{c}} - \tfrac{2}{11}\|D^2e^{n}\|^2_{\tilde{c}}\\
&\quad \leq C\,\tau^5 + C\,\tau\,\|r(t^{n+2}) - \tilde{r}^{n+2}\|^2.
\end{aligned}$$

At this point, we would like to mention that the choice of α, β, γ, and δ may be further optimized (depending on the restriction on τ) to obtain a slightly relaxed condition on ω; see Remark 4.1 below. Building the sum over n, we get with $e^{-2} = e^{-1} = e^0 = 0$ and $E_{\text{rhs}} = \max_{j=2,\ldots,n}\|r(t^j) - \tilde{r}^j\|^2$ that

$$\begin{aligned}
&\sum_{j=2}^{n}\|De^j\|^2 + 2\tau\sum_{j=2}^{n}\|De^j\|^2_{\tilde{b}} + \|De^n\|^2 + 2\tau\,\|e^n\|^2_{\tilde{b}} + \tfrac{30}{7}\,\|De^{n-1}\|^2_{\tilde{c}} + \tfrac{2}{11}\,\|D^2e^{n-1}\|^2_{\tilde{c}}\\
&\quad \leq Ct^n\tau^4 + Ct^nE_{\text{rhs}} + \|De^1\|^2 + 2\tau\,\|e^1\|^2_{\tilde{b}} + \tfrac{30}{7}\,\|De^0\|^2_{\tilde{c}} + \tfrac{2}{11}\,\|D^2e^0\|^2_{\tilde{c}}\\
&\quad \leq Ct^n\tau^4 + Ct^nE_{\text{rhs}} + 2\,\|e^1\|^2 + 2\tau\,\|e^1\|^2_{\tilde{b}}.
\end{aligned}$$

In particular, we obtain with E_{init} introduced in the statement of the theorem that $\sum_{j=2}^{n}\|De^j\|^2 \leq C\,(t^n\tau^4 + t^nE_{\text{rhs}} + E_{\text{init}})$.

Step 2: For the desired estimate of the error itself, we go back to (4.3), multiply the equation by 2, and apply e^{n+2}. This leads to

$$\begin{aligned}
&\underbrace{2\,\langle \mathrm{BDF}_2e^{n+2}, e^{n+2}\rangle}_{=:\,T_1} + \underbrace{4\tau\,\langle \widetilde{\mathscr{B}}e^{n+2}, e^{n+2}\rangle}_{=:\,T_2}\\
&\quad + \underbrace{4\,\langle \widetilde{\mathscr{C}}\,\mathrm{BDF}_2e^{n+1}, e^{n+2}\rangle}_{=:\,T_3} - \underbrace{2\,\langle \widetilde{\mathscr{C}}\,\mathrm{BDF}_2e^{n}, e^{n+2}\rangle}_{=:\,T_4} = 4\tau\,\langle \tilde{d}^{n+2}, e^{n+2}\rangle. \qquad (4.6)
\end{aligned}$$

With Lemma 4.1, we can rewrite T_1 as

$$T_1 = \mathrm{BDF}_2\|e^{n+2}\|^2 + 2\,\|De^{n+2}\|^2 - 2\,\|De^{n+1}\|^2 + \|D^2e^{n+2}\|^2.$$

For the second term, we directly get $T_2 = 4\tau\,\|e^{n+2}\|^2_{\tilde{b}}$. The third term is simplified using $e^{n+2} = e^{n+1} + De^{n+2}$ and once more Lemma 4.1, leading to

$$\begin{aligned}
T_3 &= 4\,\langle \widetilde{\mathscr{C}}\,\mathrm{BDF}_2e^{n+1}, e^{n+1}\rangle + 4\,\langle \widetilde{\mathscr{C}}\,(2D + D^2)e^{n+1}, De^{n+2}\rangle\\
&= 2\,\mathrm{BDF}_2\|e^{n+1}\|^2_{\tilde{c}} + 4\,\|De^{n+1}\|^2_{\tilde{c}} - 4\,\|De^{n}\|^2_{\tilde{c}} + 2\,\|D^2e^{n+1}\|^2_{\tilde{c}}\\
&\quad + 8\,\langle \widetilde{\mathscr{C}}De^{n+1}, De^{n+2}\rangle + 4\,\langle \widetilde{\mathscr{C}}D^2e^{n+1}, De^{n+2}\rangle.
\end{aligned}$$

Finally, using $e^{n+2} = De^{n+2} + De^{n+1} + e^n$ and Lemma 4.1, the last term can be written as

$$T_4 = \mathrm{BDF}_2\|e^{n}\|^2_{\tilde{c}} + 2\,\|De^{n}\|^2_{\tilde{c}} - 2\,\|De^{n-1}\|^2_{\tilde{c}} + \|D^2e^{n}\|^2_{\tilde{c}}$$

$$+\, 2\, \langle \widetilde{\mathscr{C}}\, \mathrm{BDF}_2 e^n,\, De^{n+2} + De^{n+1} \rangle.$$

Using the above expressions, equation (4.6) yields

$$\begin{aligned}
&\mathrm{BDF}_2 \|e^{n+2}\|^2 + 2\,\|De^{n+2}\|^2 - 2\,\|De^{n+1}\|^2 + \|D^2 e^{n+2}\|^2 + 4\tau\, \|e^{n+2}\|_{\tilde{b}}^2 \\
&\quad + 2\, \mathrm{BDF}_2 \|e^{n+1}\|_{\tilde{c}}^2 + 4\,\|De^{n+1}\|_{\tilde{c}}^2 - 4\,\|De^{n}\|_{\tilde{c}}^2 + 2\,\|D^2 e^{n+1}\|_{\tilde{c}}^2 \\
&\quad - \mathrm{BDF}_2 \|e^{n}\|_{\tilde{c}}^2 - 2\,\|De^{n}\|_{\tilde{c}}^2 + 2\,\|De^{n-1}\|_{\tilde{c}}^2 - \|D^2 e^{n}\|_{\tilde{c}}^2 \\
&\le 4\tau\, \|\tilde{d}^{n+2}\|\, \|e^{n+2}\| + 8\,\|De^{n+1}\|_{\tilde{c}} \|De^{n+2}\|_{\tilde{c}} + 4\,\|D^2 e^{n+1}\|_{\tilde{c}} \|De^{n+2}\|_{\tilde{c}} \\
&\quad + 2\,\|2De^{n} + D^2 e^{n}\|_{\tilde{c}} \|De^{n+2} + De^{n+1}\|_{\tilde{c}} \\
&\le C\tau\, \|\tilde{d}^{n+2}\|^2 + 2\tau\, \|e^{n+2}\|_{\tilde{b}}^2 + 4\,\|De^{n+1}\|_{\tilde{c}}^2 + 2\,\|D^2 e^{n+1}\|_{\tilde{c}}^2 + 6\,\|De^{n+2}\|_{\tilde{c}}^2 \\
&\quad + 8\,\|De^{n}\|_{\tilde{c}}^2 + 2\,\|D^2 e^{n}\|_{\tilde{c}}^2 + 2\,\|De^{n+2}\|_{\tilde{c}}^2 + 2\,\|De^{n+1}\|_{\tilde{c}}^2
\end{aligned}$$

for some constant C that depends on the ellipticity constant of $\widetilde{\mathscr{B}}$. With

$$2\, \mathrm{BDF}_2 \|e^{n+1}\|_{\tilde{c}}^2 - \mathrm{BDF}_2 \|e^{n}\|_{\tilde{c}}^2 = 6\,\|e^{n+1}\|_{\tilde{c}}^2 - 11\,\|e^{n}\|_{\tilde{c}}^2 + 6\,\|e^{n-1}\|_{\tilde{c}}^2 - \|e^{n-2}\|_{\tilde{c}}^2,$$

we get

$$\begin{aligned}
&\mathrm{BDF}_2 \|e^{n+2}\|^2 + \|D^2 e^{n+2}\|^2 + \|D^2 e^{n+1}\|_{\tilde{c}}^2 + 2\tau\, \|e^{n+2}\|_{\tilde{b}}^2 \\
&\quad + 2\,\|De^{n+2}\|^2 - 2\,\|De^{n+1}\|^2 + \|D^2 e^{n+1}\|_{\tilde{c}}^2 - \|D^2 e^{n}\|_{\tilde{c}}^2 + 4\,\|De^{n+1}\|_{\tilde{c}}^2 \\
&\quad - 6\,\|De^{n}\|_{\tilde{c}}^2 + 2\,\|De^{n-1}\|_{\tilde{c}}^2 + 6\,\|e^{n+1}\|_{\tilde{c}}^2 - 11\,\|e^{n}\|_{\tilde{c}}^2 + 6\,\|e^{n-1}\|_{\tilde{c}}^2 - \|e^{n-2}\|_{\tilde{c}}^2 \\
&\le C\tau\, \|\tilde{d}^{n+2}\|^2 + 8\,\|De^{n}\|_{\tilde{c}}^2 + 6\,\|De^{n+1}\|_{\tilde{c}}^2 + 8\,\|De^{n+2}\|_{\tilde{c}}^2 \\
&\quad + 2\,\|D^2 e^{n}\|_{\tilde{c}}^2 + 2\,\|D^2 e^{n+1}\|_{\tilde{c}}^2 \\
&\le C\tau\, \|\tilde{d}^{n+2}\|^2 + 4\,\|De^{n-1}\|_{\tilde{c}}^2 + 16\,\|De^{n}\|_{\tilde{c}}^2 + 10\,\|De^{n+1}\|_{\tilde{c}}^2 + 8\,\|De^{n+2}\|_{\tilde{c}}^2.
\end{aligned}$$

Dropping the terms $\|D^2 e^{n+2}\|^2$ and $\|D^2 e^{n+1}\|_{\tilde{c}}^2$ on the left-hand side, summing up, and using $e^{-2} = e^{-1} = e^0 = 0$ and $\omega \le 1/5$, we obtain

$$\begin{aligned}
&3\,\|e^{n}\|^2 - \|e^{n-1}\|^2 + 2\,\|De^{n}\|^2 + \|D^2 e^{n-1}\|_{\tilde{c}}^2 + 4\,\|De^{n-1}\|_{\tilde{c}}^2 - 2\,\|De^{n-2}\|_{\tilde{c}}^2 \\
&\quad + 6\,\|e^{n-1}\|_{\tilde{c}}^2 - 5\,\|e^{n-2}\|_{\tilde{c}}^2 + \|e^{n-3}\|_{\tilde{c}}^2 + 2\tau \sum_{j=2}^{n} \|e^{j}\|_{\tilde{b}}^2 \\
&\le C\tau \sum_{j=2}^{n} \|\tilde{d}^{j}\|^2 + \sum_{j=2}^{n} \Big(4\,\|De^{j-3}\|_{\tilde{c}}^2 + 16\,\|De^{j-2}\|_{\tilde{c}}^2 + 10\,\|De^{j-1}\|_{\tilde{c}}^2 + 8\,\|De^{j}\|_{\tilde{c}}^2 \Big) \\
&\quad + 3\,\|e^{1}\|^2 - \|e^{0}\|^2 + 2\,\|De^{1}\|^2 + \|D^2 e^{0}\|_{\tilde{c}}^2 + 4\,\|De^{0}\|_{\tilde{c}}^2 - 2\,\|De^{-1}\|_{\tilde{c}}^2 \\
&\quad + 6\,\|e^{0}\|^2 - 5\,\|e^{-1}\|^2 + \|e^{-2}\|^2 \\
&\le C\tau \sum_{j=2}^{n} \|\tilde{d}^{j}\|^2 + 38\omega \sum_{j=2}^{n} \|De^{j}\|^2 + 3\,\|e^{1}\|^2 + 8\,\|De^{1}\|^2
\end{aligned}$$

$$\leq C\tau \sum_{j=2}^{n} \|\tilde{d}^j\|^2 + 8 \sum_{j=2}^{n} \|De^j\|^2 + 11\,\|e^1\|^2.$$

Recalling $\tilde{d}^j = \mathscr{O}(\tau^2) + E_{\text{rhs}}$ and applying the estimate obtained in **Step 1** of this proof, namely $\sum_{j=2}^{n} \|De^j\|^2 \leq C\,(t^n\tau^4 + t^n E_{\text{rhs}} + E_{\text{init}})$, we obtain

$$\begin{aligned}
&3\,\|e^n\|^2 - \|e^{n-1}\|^2 + 2\,\|De^n\|^2 + \|D^2e^{n-1}\|_{\tilde{c}}^2 + 4\,\|De^{n-1}\|_{\tilde{c}}^2 - 2\,\|De^{n-2}\|_{\tilde{c}}^2 \\
&\quad + 6\,\|e^{n-1}\|_{\tilde{c}}^2 - 5\,\|e^{n-2}\|_{\tilde{c}}^2 + \|e^{n-3}\|_{\tilde{c}}^2 + 2\tau \sum_{j=2}^{n} \|e^j\|_{\tilde{b}}^2 \\
&\leq \tilde{C}\,(t^n\tau^4 + t^n E_{\text{rhs}} + E_{\text{init}}).
\end{aligned}$$

Dropping the terms $\|De^n\|^2$, $\|D^2e^{n-1}\|_{\tilde{c}}^2$, and $\|e^{n-3}\|_{\tilde{c}}^2$ on the left-hand side, we obtain

$$\begin{aligned}
&3\,\|e^n\|^2 + 4\,\|De^{n-1}\|_{\tilde{c}}^2 + 6\,\|e^{n-1}\|_{\tilde{c}}^2 + 2\tau \sum_{j=2}^{n} \|e^j\|_{\tilde{b}}^2 \\
&\leq \|e^{n-1}\|^2 + 2\,\|De^{n-2}\|_{\tilde{c}}^2 + 5\,\|e^{n-2}\|_{\tilde{c}}^2 + \tilde{C}\,(t^n\tau^4 + E_{\text{init}}) \\
&\leq \frac{5}{6}\Big(3\,\|e^{n-1}\|^2 + 4\,\|De^{n-2}\|_{\tilde{c}}^2 + 6\,\|e^{n-2}\|_{\tilde{c}}^2 + 2\tau \sum_{j=2}^{n-1} \|e^j\|_{\tilde{b}}^2\Big) \\
&\quad + \tilde{C}\,(t^n\tau^4 + t^n E_{\text{rhs}} + E_{\text{init}})
\end{aligned} \tag{4.7}$$

for all $n \geq 2$. Using the estimate in (4.7) multiple times, we get with $\sum_{j=0}^{\infty} \left(\frac{5}{6}\right)^j = 6$ that

$$\begin{aligned}
&3\,\|e^n\|^2 + 4\,\|De^{n-1}\|_{\tilde{c}}^2 + 6\,\|e^{n-1}\|_{\tilde{c}}^2 + 2\tau \sum_{j=2}^{n} \|e^j\|_{\tilde{b}}^2 \\
&\quad \leq 3\,\|e^1\|^2 + 6\,\tilde{C}\,(t^n\tau^4 + t^n E_{\text{rhs}} + E_{\text{init}}).
\end{aligned}$$

Since the first term on the right-hand side can be once again bounded in terms of E_{init}, this is the assertion. □

Remark 4.1 The choice of the parameters α, β, γ, and δ in (4.5) can be further improved, leading to a relaxed condition on ω. To balance the respective terms, we require that $5 - 2\delta = 1 + 1/\alpha$, $2 - 2\gamma = 1/\beta$, as well as

$$4 - \tau - 2\alpha\omega - \beta\omega > 0, \qquad 1 - 2/\gamma\omega - 4/\delta\omega \geq 0$$

for reasonably small values of τ. This restricts possible choices, such that the condition on ω can only be slightly improved. Nearly optimal values can be obtained by the solution of a constrained optimization problem. As an example, under the (more restrictive) assumption that $\tau \leq 1/4$, the choice $\alpha = 15/4$ and $\beta = 15/2$ (and thus $\gamma = 14/15$, $\delta = 28/15$) leads to the improved condition $\omega \leq 7/30$.

5 Numerical experiments

This section is devoted to the numerical illustration of the convergence result presented in Theorem 3.2 and the necessity of a weak coupling condition. Moreover, we present a semi-explicit method of order three based on the above construction.

5.1 Poroelastic example

In the first experiment, we investigate the convergence rates of the semi-explicit second-order scheme (3.5) and compare the results with an implicit second-order scheme based on a BDF-2 discretization. Note that the fully implicit scheme does not require any type of coupling condition, which can be seen by resolving the elliptic equation and applying standard results such as [29, Thm. 10.1] to the resulting parabolic equation. We choose $\Omega = (0, 1)^2$, $T = 1$, and consider the poroelastic parameters of *Charcoal granite* in combination with water (see Table 1 or [16, Tab. 4]), i.e., we set

$$\lambda = 2.23 \cdot 10^{10}, \quad \mu = 1.9 \cdot 10^{10}, \quad \alpha = 0.27, \quad M = 8.5 \cdot 10^{10}, \quad \kappa/\nu = 1.0 \cdot 10^{-19}.$$

Further, the right-hand sides are given by

$$f \equiv [\, 1 \;\; 2 \,]^T, \qquad g(t, x) = 30 \sin(2\pi\, t\, x_1 + 4\pi\, t)$$

and the initial condition reads $p^0(x) = 50\, x_1(1 - x_1)x_2(1 - x_2)$. Accordingly, u^0 is defined through the consistency condition (2.2a), and p^1, u^1 by an implicit Euler step as described in Remark 3.5. Note that with the above parameters, it holds that

$$\omega = \alpha^2 M/(\mu + \lambda) \approx 0.15 < 1/5$$

such that the coupling condition in Theorem 3.2 is just fulfilled.

The computations are based on a finite element implementation in FEniCS, leading to a system as described shortly in Sect. 2.2. We now investigate the convergence behavior of the semi-explicit scheme (3.5) and compare it with a second-order implicit BDF discretization. For the computation of a reference solution, we choose an implicit midpoint scheme with step size $\tau_{\text{ref}} = 2^{-11}$ and a spatial mesh width $h_{\text{ref}} = 2^{-7}$. Since we are mainly interested in the temporal discretization errors, we compute the second-order schemes for step sizes $\tau \in \{2^{-2}, \ldots, 2^{-9}\}$ with the fixed spatial parameter $h = 2^{-7}$.

The results are presented in Fig. 1. Therein, we use the notation $p(T)$ to refer to the reference solution and p_h^N to refer to the discrete solution at time $T = N\tau$ (and accordingly for u). We observe second-order convergence for both the implicit and the semi-explicit scheme. The implicit method, however, achieves slightly better results compared to the semi-explicit one. For comparison, we also included the semi-explicit

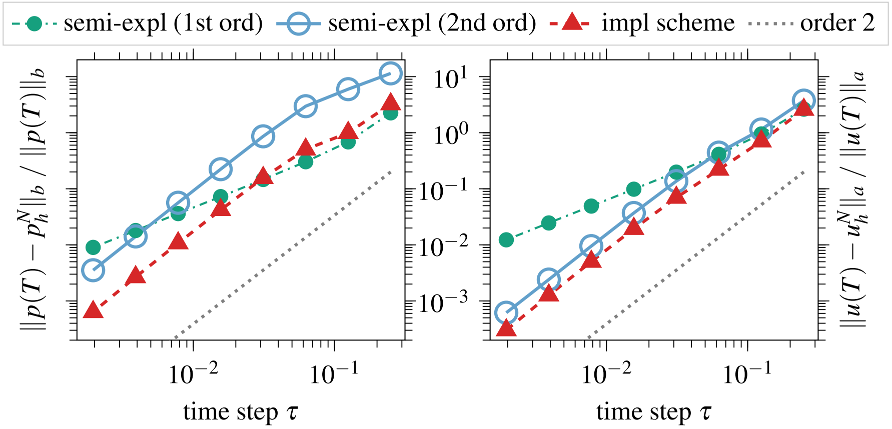

Fig. 1 Relative errors in p (left, measured in the b-norm) and u (right, measured in the a-norm) for the poroelastic example in Sect. 5.1 at the final time T for fixed $h = 2^{-7}$ and varying τ

scheme of first order; see (2.3). The main advantage of the semi-explicit scheme lies in the fact that the two poroelastic equations can be solved sequentially, which results in a computational speedup. Moreover, standard preconditioners for elliptic and parabolic systems can be used. Note, however, that the semi-explicit method is only stable if an appropriate coupling condition is fulfilled as indicated in Theorem 3.2. This is further investigated in the following subsection.

5.2 Sharpness of the weak coupling condition

We now present a numerical example to investigate the requirement of the weak coupling condition in Theorem 3.2. To this end, we consider the following toy problem of the form (2.2) with $\mathscr{V} = \mathscr{H}_{\mathscr{V}} = \mathbb{R}^3$, $\mathscr{Q} = \mathscr{H}_{\mathscr{Q}} = \mathbb{R}^1$ and bilinear forms

$$a(u,v) = v^T A u, \quad d(v,p) = \sqrt{\omega}\, p^T D v, \quad c(p,q) = q^T C p, \quad b(p,q) = q^T B p$$

with matrices

$$A\colon = \frac{1}{2-\sqrt{2}} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \qquad D\colon = \left[\tfrac{2}{3}\ \tfrac{1}{3}\ \tfrac{2}{3}\right], \qquad C\colon = 1, \qquad B\colon = 1.$$

The prefactor of A is chosen in such a way that c_a, which equals the smallest eigenvalue of A, is exactly 1. Moreover, we have $c_c = 1$ and for the continuity constant of d we get $C_d = \sqrt{\omega}$. Therefore, we consider as coupling parameter $\omega = C_d^2/(c_a c_c)$.

We test our semi-explicit scheme (3.5) with different step sizes τ and different coupling coefficients ω. The relative errors compared to a fine discretization with an implicit midpoint rule with step size $\tau_{\text{ref}} = 2^{-14}$ are computed at the final time

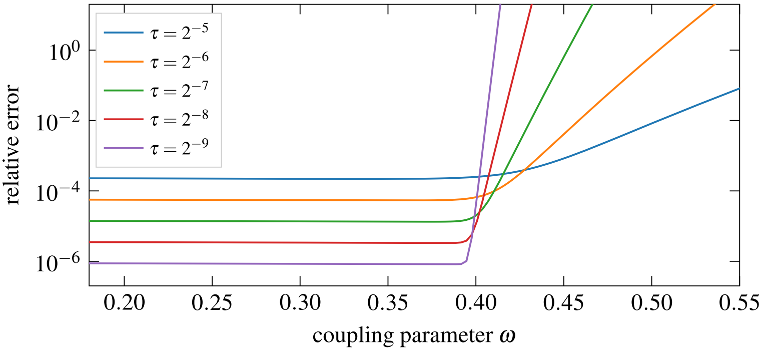

Fig. 2 Relative errors of the second-order semi-explicit method at the final time point $T = 1/2$ for different coupling parameters ω and different time step sizes τ

$T = 1/2$. For the forcing functions, we choose

$$f \equiv [\,1 \;\; 1 \;\; 1\,]^T \qquad \text{and} \qquad g(t) = \sin(t).$$

The corresponding results are presented in Fig. 2. While the sufficient condition from Theorem 3.2 reads $\omega \le 1/5$ and the delay approach from Sect. 3.5 demands $\omega < 1/3$, we observe that the critical value for stability is roughly 0.38 and therefore slightly relaxed compared to the theoretical considerations. The experiment shows that a coupling condition as in Theorem 3.2 is indeed necessary and – up to a moderate scaling factor – rather sharp.

5.3 Semi-explicit scheme of order 3

As an outlook, we go beyond the presented theory and motivate a possible extension to a semi-explicit third-order scheme. This is done by using the BDF-3 scheme for the delay system in (3.1), see the discussion in Remark 3.1. For $k = 3$, we have $p \approx 3p_\tau - 3p_{2\tau} + p_{3\tau}$, which yields the semi-explicit 3-step scheme

$$\begin{aligned} \mathscr{A}u^{n+3} - \mathscr{D}^*\big(3p^{n+2} - 3p^{n+1} + p^n\big) &= f^{n+3}, \\ \mathscr{D}\,\tfrac{11u^{n+3}-18u^{n+2}+9u^{n+1}-2u^n}{6\tau} + \mathscr{C}\,\tfrac{11p^{n+3}-18p^{n+2}+9p^{n+1}-2p^n}{6\tau} + \mathscr{B}p^{n+3} &= g^{n+3}. \end{aligned}$$

To illustrate the behavior in terms of the convergence rate and the weak coupling condition, we consider again the setting presented in Sect. 5.2. The corresponding results are shown in Fig. 3. Note that the error decreases roughly by a factor 8 when halving the step size τ, which indicates a third-order convergence rate. As before, we observe that a suitably small coupling of the two equations is necessary in order to ensure stability. The numerically observed critical point for stability is roughly $\omega \le 1/6$ and hence smaller than in the second-order case of Fig. 2. This indicates that the required coupling condition depends on the order k of the corresponding scheme.

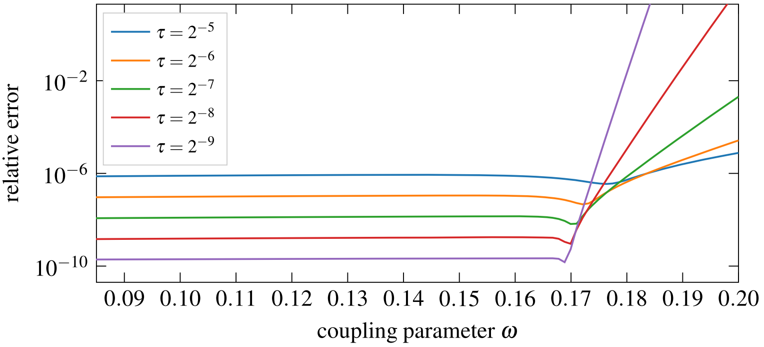

Fig. 3 Relative errors of the third-order semi-explicit method at the final time point $T = 1/2$ for different coupling parameters ω and different time step sizes τ

Performing a similar analysis of the corresponding delay equation as in Sect. 3.5 yields that delay-independent asymptotic stability is (numerically) guaranteed for $\omega < 1/7$.

6 Conclusions

Within this paper, we have constructed a semi-explicit second-order time-integration scheme for linear poroelasticity that decouples the problem and hence is suitable in a co-design paradigm where specialized legacy codes for the elliptic and parabolic equation can be used. The method is constructed by first perturbing the elastic equation with time delays and then applying BDF-2 to this delay equation. The delay times equal multiples of the time step size. We have proven convergence of this scheme under a suitable weak coupling condition. This coupling condition is, as in the first-order case [5], explicitly quantified via an asymptotic stability analysis of the delay equation. While our work focuses on the second-order scheme, we have demonstrated in a numerical example that the same idea can also be used to construct a third-order scheme, which however requires a more restrictive weak coupling condition as well as an alternative convergence proof.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

A Appendix

Proposition A. 1 *Assume sufficiently smooth right-hand sides f and g and a history function $\bar{\Phi} \in C^\infty([-\tau, 0], \mathscr{Q})$ satisfying*

$$\bar{\Phi}(-\tau) = \bar{\Phi}(0) = p^0, \qquad \bar{\Phi}^{(j)}(-\tau) = 0 \ \textit{ for } j = 1, \dots, k-1 \tag{A.1}$$

such that the solution $(\bar{u}, \bar{p})$ of the delay system (3.1) satisfies $\bar{p} \in W^{k+1,\infty}(\mathscr{H}_\mathscr{Q})$. Then, the solutions to (2.2) and (3.1) are equal up to a term of order τ^k, i.e., for almost all $t \in [0, T]$ we have

$$\|\bar{p}(t) - p(t)\|^2_\mathscr{Q} + \|\bar{u}(t) - u(t)\|^2_\mathscr{V} \lesssim t\, \tau^{2k} \Big[\|\bar{\Phi}\|^2_{W^{k+1,\infty}(-\tau,0;\mathscr{H}_\mathscr{Q})} + \|\bar{p}\|^2_{W^{k+1,\infty}(\mathscr{H}_\mathscr{Q})} \Big].$$

Proof We define $e_p := \bar{p} - p$ and $e_u := \bar{u} - u$. The conditions on the derivatives of $\bar{\Phi}$ in (A.1) ensure the consistency of the initial data, i.e., $\bar{u}(0) = u(0) = u^0$. Hence, we have $e_p(0) = 0$ and $e_u(0) = 0$. By a Taylor expansion, we know that

$$\begin{aligned} \bar{p}(t) &= \sum_{j=0}^{k-1} \tfrac{\tau^j}{j!}\, \bar{p}^{(j)}(t-\tau) + \int\limits_{t-\tau}^{t} \bar{p}^{(k)}(\xi)\, \tfrac{(t-\xi)^{k-1}}{(k-1)!}\, \mathrm{d}\xi, \\ \dot{\bar{p}}(t) &= \sum_{j=0}^{k-1} \tfrac{\tau^j}{j!}\, \bar{p}^{(j+1)}(t-\tau) + \int\limits_{t-\tau}^{t} \bar{p}^{(k+1)}(\xi)\, \tfrac{(t-\xi)^{k-1}}{(k-1)!}\, \mathrm{d}\xi. \end{aligned}$$

With this, the errors satisfy the system

$$a(e_u, v) - d(v, e_p) = -\int_{t-\tau}^{t} \tfrac{(t-\xi)^{k-1}}{(k-1)!}\, d(v, \bar{p}^{(k)}(\xi))\, \mathrm{d}\xi, \tag{A.2a}$$

$$d(\dot{e}_u, q) + c(\dot{e}_p, q) + b(e_p, q) = 0 \tag{A.2b}$$

for all test functions $v \in \mathscr{V}$, $q \in \mathscr{Q}$. Moreover, considering the derivatives of (2.2a) and (3.1a), we obtain

$$a(\dot{e}_u, v) - d(v, \dot{e}_p) = -\int_{t-\tau}^{t} \tfrac{(t-\xi)^{k-1}}{(k-1)!}\, d(v, \bar{p}^{(k+1)}(\xi))\, \mathrm{d}\xi. \tag{A.3}$$

The sum of (A.3) with test function $v = \dot{e}_u$ and (A.2b) with test function $q = \dot{e}_p$, bounding the integral, and an application of Young's inequality yield

$$\begin{aligned} \|\dot{e}_u\|^2_a + \|\dot{e}_p\|^2_c + \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|e_p\|^2_b &= -\int_{t-\tau}^{t} \tfrac{(t-\xi)^{k-1}}{(k-1)!}\, d(\dot{e}_u, \bar{p}^{(k+1)}(\xi))\, \mathrm{d}\xi \\ &\leq \tfrac{\tau^k}{(k-1)!}\, C_d\, \|\dot{e}_u\|_\mathscr{V} \|\bar{p}^{(k+1)}\|_{L^\infty(t-\tau,t;\mathscr{H}_\mathscr{Q})} \\ &\leq \tfrac{1}{2}\, \|\dot{e}_u\|^2_a + C\, \tau^{2k}\, \|\bar{p}^{(k+1)}\|^2_{L^\infty(t-\tau,t;\mathscr{H}_\mathscr{Q})} \end{aligned}$$

with the constant $C = \frac{C_d^2}{2c_a((k-1)!)^2}$. Note that we use the notion $\| \bullet \|_a$, $\| \bullet \|_b$, and $\| \bullet \|_c$ for the norms induced by the bilinear forms a, b, and c, respectively. Hence, we can eliminate $\|\dot{e}_u\|_a^2$ on the right-hand side. By the ellipticity of the bilinear forms and an integration over $[0, t]$ we conclude that

$$\int_0^t \|\dot{e}_u(s)\|_{\mathscr{V}}^2 \,\mathrm{d}s + \|e_p(t)\|_{\mathscr{Q}}^2 \lesssim \tau^{2k}\, t\, \|\bar{p}^{(k+1)}\|_{L^\infty(-\tau,t;\mathscr{H}_\mathscr{Q})}^2. \tag{A.4}$$

Note that we use here the convention that $\bar{p}$ equals the history function $\bar{\Phi}$ on $[-\tau, 0]$. In the same way, the sum of (A.2a) with test function $v = \dot{e}_u$ and (A.2b) with test function $q = e_p$ yields the estimate

$$\begin{aligned} &\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|e_u\|_a^2 + \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|e_p\|_c^2 + \|e_p\|_b^2 \\ &\quad = -\int_{t-\tau}^t \tfrac{(t-\xi)^(k-1)}{(k-1)!}\, d(\dot{e}_u, \bar{p}^{(k)}(\xi))\,\mathrm{d}\xi \lesssim \|\dot{e}_u\|_{\mathscr{V}}^2 + \tau^{2k}\, \|\bar{p}^{(k)}\|_{L^\infty(t-\tau,t;\mathscr{H}_\mathscr{Q})}^2. \end{aligned}$$

Integration over $[0, t]$ and the application of estimate (A.4) finally gives

$$\|e_u(t)\|_{\mathscr{V}}^2 \lesssim t\, \tau^{2k}\, \|\bar{p}^{(k+1)}\|_{L^\infty(-\tau,t;\mathscr{H}_\mathscr{Q})}^2,$$

which completes the proof. □

References

1. Akrivis, G., Crouzeix, M., Makridakis, C.: Implicit-explicit multistep finite element methods for nonlinear parabolic problems. Math. Comp. **67**(222), 457–477 (1998)
2. Akrivis, G., Lubich, C.: Fully implicit, linearly implicit and implicit-explicit backward difference formulae for quasi-linear parabolic equations. Numer. Math. **131**, 713–735 (2015). https://doi.org/10.1007/s00211-015-0702-0
3. Altmann, R., Chung, E., Maier, R., Peterseim, D., Pun, S.M.: Computational multiscale methods for linear heterogeneous poroelasticity. J. Comput. Math. **38**(1), 41–57 (2020)
4. Altmann, R., Maier, R.: A decoupling and linearizing discretization for poroelasticity with nonlinear permeability. SIAM J. Sci. Comput. **44**(3), B457–B478 (2022). https://doi.org/10.1137/21M1413985
5. Altmann, R., Maier, R., Unger, B.: Semi-explicit discretization schemes for weakly-coupled elliptic-parabolic problems. Math. Comp. **90**(329), 1089–1118 (2021). https://doi.org/10.1090/mcom/3608
6. Altmann, R., Zimmer, C.: On the smoothing property of linear delay partial differential equations. J. Math. Anal. Appl. **467**(2), 916–934 (2018). https://doi.org/10.1016/j.jmaa.2018.07.049
7. Armero, F., Simo, J.C.: A new unconditionally stable fractional step method for nonlinear coupled thermomechanical problems. Internat. J. Numer. Methods Engrg. **35**(4), 737–766 (1992). https://doi.org/10.1002/nme.1620350408
8. Bellen, A., Zennaro, M.: Numerical methods for delay differential equations. Oxford University Press, New York (2003). https://doi.org/10.1093/acprof:oso/9780198506546.001.0001
9. Bellman, R., Cooke, K.L.: Differential-difference equations. Academic Press, New York-London (1963)
10. Biot, M.A.: General theory of three-dimensional consolidation. J. Appl. Phys. **12**(2), 155–164 (1941)
11. Biot, M.A.: Thermoelasticity and irreversible thermodynamics. J. Appl. Phys. **27**, 240–253 (1956)
12. Chaabane, N., Rivière, B.: A sequential discontinuous Galerkin method for the coupling of flow and geomechanics. J. Sci. Comput. **74**(1), 375–395 (2018). https://doi.org/10.1007/s10915-017-0443-6
13. Chaabane, N., Rivière, B.: A splitting-based finite element method for the Biot poroelasticity system. Comput. Math. Appl. **75**(7), 2328–2337 (2018). https://doi.org/10.1016/j.camwa.2017.12.009

14. Ciarlet, P.G.: Mathematical elasticity, vol. I. North-Holland, Amsterdam (1988)
15. Crouzeix, M.: Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. Numer. Math. **35**(3), 257–276 (1980). https://doi.org/10.1007/BF01396412
16. Detournay, E., Cheng, A.H.D.: Fundamentals of poroelasticity. In: Analysis and design methods, pp. 113–171. Elsevier (1993)
17. Ern, A., Meunier, S.: A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems. ESAIM: Math. Model. Numer. Anal. **43**(2), 353–375 (2009). https://doi.org/10.1051/m2an:2008048
18. Frank, J., Hundsdorfer, W., Verwer, J.: On the stability of implicit-explicit linear multistep methods. Appl. Numer. Math. **25**(2), 193–205 (1997). https://doi.org/10.1016/S0168-9274(97)00059-7
19. Fu, G.: A high-order HDG method for the Biot's consolidation model. Comput. Math. Appl. **77**(1), 237–252 (2019)
20. Gu, K., Kharitonov, V.L., Chen, J.: Stability of Time-Delay Systems. Birkhäuser, Boston (2003). https://doi.org/10.1007/978-1-4612-0039-0
21. Hairer, E., Wanner, G.: Solving ordinary differential equations. II, *Springer Series in Computational Mathematics*, vol. 14, second edn. Springer-Verlag, Berlin (1996). https://doi.org/10.1007/978-3-642-05221-7. Stiff and differential-algebraic problems
22. Kim, J., Tchelepi, H.A., Juanes, R.: 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)
23. Kunkel, P., Mehrmann, V.: Differential-algebraic equations. Analysis and numerical solution. European mathematical society, Zürich (2006). https://doi.org/10.4171/017
24. Lee, J.J., Mardal, K.A., Winther, R.: Parameter-robust discretization and preconditioning of Biot's consolidation model. SIAM J. Sci. Comput. **39**(1), A1–A24 (2017). https://doi.org/10.1137/15M1029473
25. Mikelić, A., Wheeler, M.F.: Convergence of iterative coupling for coupled flow and geomechanics. Comput. Geosci. **17**(3), 455–461 (2013)
26. Mujahid, A.: Monolithic, non-iterative and iterative time discretization methods for linear coupled elliptic-parabolic systems. GAMM Archive for students **4**(1) (2022). https://doi.org/10.14464/gammas.v4i1.500
27. Showalter, R.E.: Diffusion in poro-elastic media. J. Math. Anal. Appl. **251**(1), 310–340 (2000)
28. Storvik, E., Both, J.W., Kumar, K., Nordbotten, J.M., Radu, F.A.: On the optimization of the fixed-stress splitting for Biot's equations. Int. J. Numer. Meth. Eng. **120**(2), 179–194 (2019)
29. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, second edn. Springer series in computational mathematics. Springer Berlin, Heidelberg (2006). https://doi.org/10.1007/3-540-33122-0
30. Trenn, S., Unger, B.: Delay regularity of differential-algebraic equations. In: 2019 IEEE 58th Conference on Decision and Control (CDC), Nice, France, pp. 989–994 (2019). https://doi.org/10.1109/CDC40024.2019.9030146
31. Unger, B.: Discontinuity propagation in delay differential-algebraic equations. Electron. J. Linear Algebr. **34**, 582–601 (2018). https://doi.org/10.13001/1081-3810.3759
32. Wheeler, M.F., Gai, X.: Iteratively coupled mixed and Galerkin finite element methods for poroelasticity. Numer. Meth. Part. D. E. **23**(4), 785–797 (2007)
33. Zeidler, E.: Nonlinear Functional Analysis and its Applications IIa: Linear Monotone Operators. Springer-Verlag, New York (1990)
34. Zoback, M.D.: Reservoir Geomechanics. Cambridge University Press, Cambridge (2010)