SMAI-JCM
SMAI Journal of Computational Mathematics

Decoupling multistep schemes for elliptic–parabolic problems

Robert Altmann, Abdullah Mujahid & Benjamin Unger

Volume 12 (2026), p. 371-391.
https://doi.org/10.5802/smai-jcm.152

The SMAI Journal of Computational Mathematics is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
Submissions at https://smai-jcm.centre-mersenne.org/ojs/submission
e-ISSN: 2426-8399

SMAI Journal of Computational Mathematics
Vol. 12, 371-391 (2026)

Decoupling multistep schemes for elliptic–parabolic problems

Robert Altmann [1]
Abdullah Mujahid [2]
Benjamin Unger [3]

[1] Institute of Analysis and Numerics, Otto von Guericke University Magdeburg, Magdeburg, Germany
E-mail address: robert.altmann@ovgu.de
[2] Stuttgart Center for Simulation Science (SC SimTech), University of Stuttgart, Stuttgart, Germany
E-mail address: abdullah@hawksims.com
[3] Institute for Applied and Numerical Mathematics, Karlsruhe Institute of Technology, Karlsruhe, Germany
E-mail address: benjamin.unger@kit.edu.

Abstract. We study the construction and convergence of decoupling multistep schemes of higher order using the backward differentiation formulae for an elliptic–parabolic problem, which includes multiple-network poroelasticity as a special case. These schemes were first introduced in [Altmann, Maier, Unger, BIT Numer. Math., 64:20, 2024], where a convergence proof for the second-order case is presented. Here, we present a slightly modified version of these schemes using a different construction of related time delay systems. We present a novel convergence proof relying on concepts from G-stability applicable for any order and providing a sharper characterization of the required weak coupling condition. The key tool for the convergence analysis is the construction of a weighted norm enabling a telescoping argument for the sum of the errors.

2020 Mathematics Subject Classification. 65M12, 65J10, 76S05.

Keywords. poroelasticity, decoupling, higher-order discretization, backward differentiation formulae.

1. Introduction

This paper explores semi-explicit time integration schemes of higher order for the equations of poroelasticity [11] or, more generally, for linear elliptic–parabolic problems. Among important applications of this problem are biomechanics, where the human brain and heart are modeled as poroelastic medium with multiple fluid networks [17, 31], and geomechanics, where poroelasticity serves as a model problem [33].

The well-posedness of the considered elliptic–parabolic problem is studied in [26]. To reduce the computational effort resulting from the coupling, several decoupling strategies exist [12, 21, 22, 23, 27], most of which are based on a fixed-point iteration. However, higher-order methods in the iterative framework need many inner iteration steps and hence call for a stronger contraction [8].

Alternatively, a semi-explicit method of first order, as proposed and analyzed in [5], decouples the system whenever a weak coupling condition is satisfied. The design of the semi-explicit method is based on constructing a related partial differential equation with a time delay, approximating the solution of the original system up to a desired order. An extension to second order is considered in [6] and to

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 467107679. Parts of this work were carried out while RA and AM were affiliated with the Institute of Mathematics, University of Augsburg. RA was also affiliated with the Centre for Advanced Analytics and Predictive Sciences (CAAPS). BU was affiliated with the Stuttgart Center for Simulation Science (SimTech) at the University of Stuttgart. Moreover, AM and BU acknowledge support by the Stuttgart Center for Simulation Science (SimTech).
https://doi.org/10.5802/smai-jcm.152

nonlinear problems in [4]. Unlike the step size restriction in explicit methods, semi-explicit methods require a weak coupling condition, which restricts the problem class given by the material parameters. To lift this restriction, schemes with a fixed number of relaxation steps are introduced in [2, 3].

In this article, we provide closed-form expressions for the construction of the delay systems and develop the necessary tools to show the convergence of the semi-explicit schemes using the *backward differentiation formulae* (BDF). In the standard analysis of BDF-k methods, one can construct a weighted norm with a symmetric positive definite $k \times k$ matrix G. The weighting allows for setting up a telescoping sum of the errors and, hence, is the key tool for the convergence analysis. Moreover, it exploits the equivalence between A–stability and G–stability [15]; see also [20, Ch. V. 6]. For methods that are not A-stable, one constructs additional multipliers [25]; see also [20, Ch. V. 8]. In practice, explicitly constructing the G matrix is often tricky. Hence, the classical way is to prove that the method is A-stable (potentially with appropriate multipliers) and then infer the existence of G. In this paper, we proceed differently and explicitly construct the weighted norm. Our main results are:

(1) We present a novel way to construct the related delay equation of a given order, which, in contrast to the approach presented in [6], is explicit and requires fewer delays.

(2) We use the norm defined through G (assuming that it exists) to present a novel convergence proof that works for any BDF order and provides a sharper characterization of the weak coupling condition compared to [6].

(3) We explicitly construct the weighting matrix G for convergence up to order three.

The remainder of this paper is organized as follows. After this introduction, the abstract model problem is introduced in Section 2 with particular examples from poroelasticity. The design of semi-explicit methods via delay systems and their well-posedness is discussed in Section 3. The main result of this article is Theorem 4.9, which combines the convergence analysis of the decoupling multistep methods based on BDF schemes from Theorem 4.3 with the verification of Assumption 1, which is the basis for the telescoping argument used in the proof. Finally, we verify the theoretical results numerically in Section 5.

Notation

Throughout the paper, we write $a \lesssim b$ to indicate that there exists a generic constant $C > 0$, independent of spatial and temporal discretization parameters, such that $a \leq Cb$. Moreover, we denote the Bochner spaces on the time interval $[0, T]$ for a Banach space $\mathcal{X}$ by $L^p(0, T; \mathcal{X})$ and $W^{k,p}(0, T; \mathcal{X})$, where $p \geq 1$, $k \in \mathbb{N}$. The L^2 norm is denoted by $\|\cdot\|$.

2. Abstract Formulation and Examples

Consider the Hilbert spaces

$$\mathcal{V} := [H^1_0(\Omega)]^m, \qquad \mathcal{Q} := H^1_0(\Omega), \qquad \mathcal{H}_\mathcal{V} := [L^2(\Omega)]^m, \qquad \mathcal{H}_\mathcal{Q} := L^2(\Omega),$$

resulting in the Gelfand triples $\mathcal{V} \hookrightarrow \mathcal{H}_\mathcal{V} \simeq \mathcal{H}_\mathcal{V}^* \hookrightarrow \mathcal{V}^*$ and $\mathcal{Q} \hookrightarrow \mathcal{H}_\mathcal{Q} \simeq \mathcal{H}_\mathcal{Q}^* \hookrightarrow \mathcal{Q}^*$, cf. [32, §23.4]. The system of interest is the following abstract coupled elliptic–parabolic problem: given sufficiently smooth source terms $f\colon [0, T] \to \mathcal{V}^*$ and $g\colon [0, T] \to \mathcal{Q}^*$, seek abstract functions $u\colon [0, T] \to \mathcal{V}$ and $p\colon [0, T] \to \mathcal{Q}$ such that for almost every $t \in (0, T]$ it holds that

$$a(u, v) - d(v, p) = \langle f, v\rangle, \tag{2.1a}$$

$$d(\dot{u}, q) + c(\dot{p}, q) + b(p, q) = \langle g, q\rangle \tag{2.1b}$$

for all test functions $v \in \mathcal{V}$, $q \in \mathcal{Q}$. As initial data, we assume

$$u(0) = u^0 \in \mathcal{V}, \qquad p(0) = p^0 \in \mathcal{H}_{\mathcal{Q}} \tag{2.1c}$$

to be *consistent*, 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 forms $a\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, $b\colon \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}$, and $c\colon \mathcal{H}_{\mathcal{Q}} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ are assumed to be symmetric, continuous, and elliptic in their respective spaces. The corresponding ellipticity and continuity constants of $\mathfrak{a} \in \{a, b, c\}$ are denoted by $c_{\mathfrak{a}}, C_{\mathfrak{a}} > 0$, respectively. For notational convenience we introduce the b-norm $\|\cdot\|_b := b(\,\cdot\,,\cdot\,)^{1/2}$ and the c-norm $\|\cdot\|_c := c(\,\cdot\,,\cdot\,)^{1/2}$ satisfying

$$\tfrac{1}{C_b}\|\cdot\|_b^2 \le \|\cdot\|_{\mathcal{Q}}^2 \le \tfrac{1}{c_b}\|\cdot\|_b^2 \quad \text{and} \quad \tfrac{1}{C_c}\|\cdot\|_c^2 \le \|\cdot\|_{\mathcal{H}_{\mathcal{Q}}}^2 \le \tfrac{1}{c_c}\|\cdot\|_c^2,$$

respectively. Furthermore, $d\colon \mathcal{V} \times \mathcal{H}_{\mathcal{Q}} \to \mathbb{R}$ is bounded, i.e., we assume that there exists a positive constant $C_d > 0$ such that $d(u, p) \le C_d\|u\|_{\mathcal{V}}\|p\|_{\mathcal{H}_{\mathcal{Q}}}$ for all $u \in \mathcal{V}$, $p \in \mathcal{H}_{\mathcal{Q}}$. Without the coupling term represented by the bilinear form d, equation (2.1a) is elliptic in u and equation (2.1b) is parabolic in p. This explains why system (2.1) is called an elliptic–parabolic problem. At this point, we would like to emphasize that $c(\dot p, q)$ – and similarly $d(\dot u, q)$ – should be understood as $\frac{\mathrm{d}}{\mathrm{d}t}c(p, q)$, see also [24].

Example 2.1 (linear poroelasticity)**.** Let $\Omega \subseteq \mathbb{R}^m$, $m \in \{2, 3\}$, be a bounded Lipschitz domain and $[0, T]$ a time interval with $T > 0$. We consider the quasi-static Biot poroelasticity model with homogeneous Dirichlet boundary conditions as introduced in [11], where one seeks the deformation $u\colon [0, T] \times \Omega \to \mathbb{R}^m$ and the pressure $p\colon [0, T] \times \Omega \to \mathbb{R}$ satisfying

$$-\nabla\cdot\sigma(u) + \alpha\nabla p = \widehat{f} \qquad \text{in } (0, T] \times \Omega, \tag{2.2a}$$

$$\partial_t\big(\alpha\nabla\cdot u + \tfrac{1}{M}p\big) - \nabla\cdot(\kappa\nabla p) = \widehat{g} \qquad \text{in } (0, T] \times \Omega. \tag{2.2b}$$

Here, σ denotes the stress tensor

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

with Lamé coefficients λ and μ. The permeability is denoted with κ. Moreover, α denotes the Biot–Willis fluid–solid coupling coefficient and M the Biot modulus. The right-hand sides $\widehat{f}$ and $\widehat{g}$ are the volumetric load and the fluid source terms, respectively, modeling an injection or production process. We refer to [5] for the definition of the corresponding bilinear forms resulting in (2.1).

Example 2.2 (multiple network poroelasticity)**.** A generalization of the Biot poroelasticity model is given by the multiple network poroelasticity theory, where we seek the deformation $u\colon [0, T]\times\Omega \to \mathbb{R}^m$ and pressure variables $p_i\colon [0, T] \times \Omega \to \mathbb{R}$ for $i = 1, \dots, J$ where $J \in \mathbb{N}$ denotes the number of fluid networks, satisfying the coupled system

$$\begin{aligned} -\nabla\cdot\sigma(u) + \sum_{i=1}^{J}\alpha_i\nabla p_i &= \widehat{f} \qquad \text{in } (0, T] \times \Omega,\\ \partial_t\Big(\alpha_i\nabla\cdot u + \tfrac{1}{M_i}p_i\Big) - \nabla\cdot(\kappa_i\nabla p_i) + \sum_{j\neq i}^{J}\beta_{ij}(p_i - p_j) &= \widehat{g}_i \qquad \text{in } (0, T] \times \Omega. \end{aligned}$$

Given some additional assumptions on the exchange rates between the fluid networks β_{ij}, this model has a similar structure as (2.1). In practice, the exchange rates are usually symmetric and sufficiently small such that the ellipticity property of the bilinear form b is satisfied [7].

To describe the stability of the decoupling schemes introduced in this article, we define

$$\omega := \tfrac{C_d^2}{c_a c_c}, \tag{2.3}$$

as the *coupling strength* between the elliptic and the parabolic equation. This value plays a crucial role for the convergence analysis and depends on the physical coefficients of the application. We refer to [16, Tab. 4] for concrete values appearing in poroelasticity; see also [6, Tab. 3.1].

The abstract formulation (2.1) can also be written in terms of operators in the respective dual spaces. For this, let $\mathcal{A}\colon \mathcal{V} \to \mathcal{V}^*$, $\mathcal{B}\colon \mathcal{Q} \to \mathcal{Q}^*$, $\mathcal{C}\colon \mathcal{H}_\mathcal{Q} \to \mathcal{H}_\mathcal{Q}^*$, and $\mathcal{D}\colon \mathcal{V} \to \mathcal{H}_\mathcal{Q}^*$ be the operators associated with the bilinear forms a, b, c, and d respectively. This then leads to the equivalent formulation

$$\mathcal{A}u - \mathcal{D}^*p = f \qquad \text{in } \mathcal{V}^*, \tag{2.4a}$$

$$\mathcal{D}\dot{u} + \mathcal{C}\dot{p} + \mathcal{B}p = g \qquad \text{in } \mathcal{Q}^*. \tag{2.4b}$$

Since $\mathcal{A}$ is invertible, we can eliminate the variable u, leading to the parabolic equation

$$(\mathcal{M} + \mathcal{C})\dot{p} + \mathcal{B}p = r, \tag{2.5}$$

where $r := g - \mathcal{D}\mathcal{A}^{-1}\dot{f}$ and $\mathcal{M} := \mathcal{D}\mathcal{A}^{-1}\mathcal{D}^*$ is a self-adjoint and non-negative operator.

3. Related Delay Equations

The key idea for the construction as well as convergence analysis of a decoupling time integration scheme in [5] is to introduce an approximation of the coupled elliptic–parabolic problem (2.4) by adding a time delay $\tau > 0$ in (2.4a), where the time delay equals the time step size. In more detail, one considers the delay system

$$\mathcal{A}\bar{u} - \mathcal{D}^*\widehat{p}(t;\tau) = f \qquad \text{in } \mathcal{V}^*, \tag{3.1a}$$

$$\mathcal{D}\dot{\bar{u}} + \mathcal{C}\dot{\bar{p}} + \mathcal{B}\bar{p} = g \qquad \text{in } \mathcal{Q}^*, \tag{3.1b}$$

where $\widehat{p}(t;\tau)$ is an approximation of $\bar{p}$ using the time delay τ. In the first-order case studied in [5] we have $\widehat{p}(t;\tau) = \Delta_\tau\bar{p}(t)$ with the shift operator $\Delta_\tau p := p(\,\cdot - \tau)$. Let us emphasize that, in contrast to the original system (2.4), such a delay system calls for a *history function*, i.e., a prescription of $\bar{p}$ for $-\tau \le t \le 0$.

If an implicit time discretization scheme of order k with time step size τ is used to discretize (3.1), then the equations (3.1a) and (3.1b) are decoupled in the sense that one can first solve (3.1a) for $\bar{u}$ and then (3.1b) for $\bar{p}$. This was done and analyzed for the implicit Euler method in [5] and for BDF-2 with a different $\widehat{p}$ in [6]. It goes without saying that a higher-order scheme is only reasonable if the delay approximation (3.1) is of the same order.

3.1. Higher-order approximation via Taylor series expansion

For a higher-order approximation, a general time delayed approximation of $p(t)$ is sought. This means that we consider a delay term $\widehat{p}(t;\tau)$ with

$$p(t) = \widehat{p}(t;\tau) + \mathcal{O}(\tau^\delta).$$

One straightforward possibility is to consider a Taylor series expansion of $p(t)$ around $t - \tau$. This, however, leads to a *delay differential equations* (DDEs) of advanced type — see [10] for a precise definition — as shown in the following example, which is motivated from [29, Ex. 2.1].

Example 3.1. Consider two terms in the Taylor series, i.e., $\bar{p} \approx \Delta_\tau\bar{p} + \tau\Delta_\tau\dot{\bar{p}}$. Solving (3.1a) for $\bar{u}$ and substituting this in (3.1b) results the inherent delay differential equation

$$\mathcal{C}\dot{\bar{p}} + \mathcal{B}\bar{p} = r - \mathcal{M}(\Delta_\tau\dot{\bar{p}} + \tau\Delta_\tau\ddot{\bar{p}}).$$

If we consider the scalar case where we replace all operators by the constant one and set $f \equiv 0$, $g \equiv 0$, then we obtain $\dot{\bar{y}} + \bar{y} = \Delta_\tau\dot{\bar{y}} + \tau\Delta_\tau\ddot{\bar{y}}$. The corresponding first-order formulation

$$\dot{\bar{y}} = \bar{z}, \qquad \bar{z} + \bar{y} = \Delta_\tau\bar{z} + \tau\Delta_\tau\dot{\bar{z}}$$

Table 3.1. Coefficients $c_{\delta,\ell}$, $\ell = 1, \dots, \delta$, for different values of δ.

δ	$c_{\delta,1}$	$c_{\delta,2}$	$c_{\delta,3}$	$c_{\delta,4}$
1	1			
2	2	-1		
3	3	-3	1	
4	4	-6	4	-1

equals a DDE of advanced type, since the second equation involves $\Delta_\tau \dot{z}$ but not $\dot{z}$. To illustrate the inherent difficulties of DDEs of advances type, we set $\tau = 1$ and consider

$$\phi(t) = y^0 + \tfrac{1}{n}\sin(n\pi t)$$

for $t \in [-1, 0]$as history function. The solution can be obtained using Bellman's method of steps [9, Ch. 3.4] and yields

$$\bar{y}(t) = \begin{cases} y^0 + \frac{1}{n}\sin(n\pi t), & -1 \le t \le 0,\\ y^0 e^{-t} + \pi\cos(n\pi)e^{-t} - \pi\cos(n\pi(t-1)), & 0 \le t \le 1,\\ y^0 e^{-t} + \pi\cos(n\pi)e^{-t} - \pi e^{-t+1} - n\pi^2\sin(n\pi(t-2)), & 1 \le t \le 2,\\ y^0 e^{-t} + \pi\cos(n\pi)e^{-t} - \pi e^{-t+1} + n^2\pi^2\cos(n\pi)e^{-t+2} - n^2\pi^3\cos(n\pi(t-3)), & 2 \le t \le 3. \end{cases}$$

Although the history function is uniformly bounded for all n, the solution in each successive interval grows with an increasing factor of n. In this sense, the delay system is ill-posed.

Solving DDEs of advanced type requires a distributional solution concept even in the finite-dimensional setting [28, 30]. To overcome this difficulty, we instead aim for a higher-order approximation via multiple delays.

3.2. Higher-order approximation via multiple delays

We choose $\widehat{p}(t;\tau)$ in (3.1) as the Lagrange interpolation polynomial of degree $\delta - 1$ for the interpolation points $s_\ell := t - \ell\tau$ given by

$$\widehat{p}(t;\tau) = \sum_{\ell=1}^{\delta} \bar{p}(s_\ell) L_\ell(t), \qquad L_\ell(t) = \prod_{j=1,\, j\neq\ell}^{\delta} \frac{t - s_j}{s_\ell - s_j}.$$

Substituting the particular form of the interpolation points, we obtain

$$L_\ell(t) \equiv (-1)^{(\ell-1)}\binom{\delta}{\ell} =: c_{\delta,\ell} \quad \ell = 1, \dots, \delta.$$

We would like to emphasize that $\bar{p}(s_\ell) = \bar{p}(t - \ell\tau)$ is not a constant but depends on t and τ, whereas the polynomial basis loses its dependence on t and are constants, which we call *delay coeffients*; see the examples in Table 3.1. Moreover, we have $\sum_{\ell=1}^{\delta} c_{\delta,\ell} = 1$ and obtain

$$\widehat{p}(t;\tau) := \sum_{\ell=1}^{\delta} c_{\delta,\ell} \Delta_{\ell\tau}\bar{p}(t) \tag{3.2}$$

as the delay term. The proof that this is indeed an approximation of $p(t)$ of order δ is given in Lemma A.1.

Example 3.2. Following Table 3.1, the delay approximations for $\delta = 1, 2, 3$ are given by

$$\widehat{p}(\,\cdot\,;\tau) = \Delta_\tau \bar{p}, \qquad \widehat{p}(\,\cdot\,;\tau) = 2\Delta_\tau \bar{p} - \Delta_{2\tau}\bar{p}, \qquad \widehat{p}(\,\cdot\,;\tau) = 3\Delta_\tau \bar{p} - 3\Delta_{2\tau}\bar{p} + \Delta_{3\tau}\bar{p},$$

respectively.

Remark 3.3. The recipe in [6] to obtain the delay system is to replace the time derivatives of the delayed terms in the Taylor series expansion with finite differences of appropriate order such that the resulting approximation is still of order δ but of neutral delay type. For $\delta \leq 3$, this approach resembles the schemes we obtain here, but for an approximation of order τ^4 already five delays are required. In contrast, we obtain a closed form expression for the delay term $\widehat{p}(\,\cdot\,;\tau)$ with δ delays.

Substituting the expression for $\widehat{p}(\,\cdot\,;\tau)$ from (3.2) in (3.1) yields the system

$$\mathcal{A}\bar{u} - \mathcal{D}^*\Bigg(\sum_{\ell=1}^{\delta} c_{\delta,\ell}\Delta_{\ell\tau}\bar{p}\Bigg) = f \qquad \text{in } \mathcal{V}^*, \tag{3.3a}$$

$$\mathcal{D}\dot{\bar{u}} + \mathcal{C}\dot{\bar{p}} + \mathcal{B}\bar{p} = g \qquad \text{in } \mathcal{Q}^* \tag{3.3b}$$

with δ delays. For $\delta = 1$, we recover the already mentioned delay system of order one introduced in [5], and for $\delta \in \{2, 3\}$ the delayed systems from [6]. In (3.3), we seek solutions $\bar{u}\colon [0,T] \to \mathcal{V}$ and $\bar{p}\colon [-\delta\tau, T] \to \mathcal{Q}$ given sufficiently smooth right-hand sides, initial data, and a history function $\Phi \in C^\infty([-\delta\tau, 0], \mathcal{H}_\mathcal{Q})$ for $\bar{p}$ that ensures the consistency of initial conditions of the original problem (2.4). For this, a sufficient condition reads

$$\Phi(-\ell\tau) = \Phi(0) = p^0 \qquad \text{for all } \ell = 1, \cdots, \delta, \tag{3.4}$$

since this implies with (3.3a) that $\mathcal{A}\bar{u}(0) - \mathcal{D}^* p^0 = \mathcal{A}\bar{u}(0) - (\sum_{\ell=1}^{\delta} c_{\delta,\ell})\mathcal{D}^* p^0 = f(0)$ and hence, $\bar{u}(0) = u^0$.

Proposition 3.4 (Distance to delay solution)**.** *Assume sufficiently smooth right-hand sides f, g, consistent initial data* (2.1c)*, and a history function $\Phi \in C^\infty([-\delta\tau, 0], \mathcal{H}_\mathcal{Q})$ which satisfies* (3.4) *such that the solution $(\bar{u}, \bar{p})$ of the delay problem* (3.3) *satisfies $\bar{p} \in W^{\delta+1,\infty}(0, T; \mathcal{H}_\mathcal{Q})$. Then the solutions of* (2.4) *and* (3.3) *only differ by a term of order δ, i.e., for almost every $t \in [0, T]$ it holds that*

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

Proof. For details on the existence and stability of $\bar{p}$, we refer to [5, Prop. A.4]. Now define $e_p := \bar{p} - p$ and $e_u := \bar{u} - u$. Taking the difference of (2.4) and (3.1) as well as the difference of the time derivatives of (2.4a) and (3.1a), Lemma A.1 yields

$$a(e_u, v) - d(v, e_p) = \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\zeta)^{\delta-1} d\Big(v, \bar{p}^{(\delta)}(\zeta)\Big)\,\mathrm{d}\zeta, \tag{3.5a}$$

$$a(\dot{e}_u, v) - d(v, \dot{e}_p) = \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\zeta)^{\delta-1} d\Big(v, \bar{p}^{(\delta+1)}(\zeta)\Big)\,\mathrm{d}\zeta, \tag{3.5b}$$

and

$$d(\dot{e}_u, q) + c(\dot{e}_p, q) + b(e_p, q) = 0 \tag{3.5c}$$

for all test functions $v \in \mathcal{V}$ and $q \in \mathcal{Q}$. The sum of (3.5c) and (3.5b) tested with $v = \dot e_u, q = \dot e_p$ results in

$$\begin{aligned}\|\dot e_u\|_a^2 + \|\dot e_p\|_c^2 + \tfrac12 \tfrac{\mathrm{d}}{\mathrm{d}t}\|e_p\|_b^2 &= \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\zeta)^{\delta-1} d\Big(\dot e_u, \bar p^{(\delta+1)}(\zeta)\Big)\,\mathrm{d}\zeta \\ &\le \frac{\tau^\delta}{(\delta-1)!}\sum_{\ell=1}^{\delta}\binom{\delta}{\ell}\ell^\delta C_d \|\dot e_u\|_{\mathcal{V}}\|\bar p^{(\delta+1)}\|_{L^\infty(t-\ell\tau,t;\mathcal{H}_\mathcal{Q})} \\ &\le \frac12\|\dot e_u\|_{\mathcal{V}}^2 + \frac{\delta}{2}\frac{\tau^{2\delta}}{((\delta-1)!)^2}C_d^2\sum_{\ell=1}^{\delta}\binom{\delta}{\ell}\binom{\delta}{\ell}\ell^{2\delta}\|\bar p^{(\delta+1)}\|^2_{L^\infty(t-\ell\tau,t;\mathcal{H}_\mathcal{Q})}.\end{aligned}$$

Integrating over $[0,t]$ and using the ellipticity properties of the bilinear forms yields

$$\int_0^t \|\dot e_u\|_{\mathcal{V}}^2\,\mathrm{d}s + \int_0^t \|\dot e_p\|_{\mathcal{H}_\mathcal{Q}}^2\,\mathrm{d}s + \|e_p\|_{\mathcal{Q}}^2 \lesssim \tau^{2\delta} t\Big(\|\bar p^{(\delta+1)}\|^2_{L^\infty(0,t;\mathcal{H}_\mathcal{Q})} + \|\Phi^{(\delta+1)}\|^2_{L^\infty(-\delta\tau,0;\mathcal{H}_\mathcal{Q})}\Big). \tag{3.6}$$

Similarly the sum of (3.5a) and (3.5c) with the choice of test functions $v = \dot e_u, q = e_p$ results in

$$\frac12\tfrac{\mathrm{d}}{\mathrm{d}t}\|e_u\|_a^2 + \frac12\tfrac{\mathrm{d}}{\mathrm{d}t}\|e_p\|_c^2 + \|e_p\|_b^2 \le \frac12\|\dot e_u\|_{\mathcal{V}}^2 + \frac{\delta}{2}\frac{\tau^{2\delta}}{((\delta-1)!)^2}C_d^2\sum_{\ell=1}^{\delta}\binom{\delta}{\ell}\binom{\delta}{\ell}\ell^{2\delta}\|\bar p^{(\delta)}\|^2_{L^\infty(t-\ell\tau,t;\mathcal{H}_\mathcal{Q})}.$$

Integrating once more over $[0,t]$ yields

$$\|e_u\|_{\mathcal{V}}^2 + \|e_p\|_{\mathcal{H}_\mathcal{Q}}^2 + \int_0^t \|e_p\|_{\mathcal{Q}}^2\,\mathrm{d}s \lesssim \int_0^t \|\dot e_u\|_{\mathcal{V}}^2\,\mathrm{d}s + \tau^{2\delta} t\,\|\bar p^{(\delta)}\|^2_{L^\infty(0,t;\mathcal{H}_\mathcal{Q})} + \tau^{2\delta} t\,\|\Phi^{(\delta)}\|^2_{L^\infty(-\delta\tau,0;\mathcal{H}_\mathcal{Q})}. \tag{3.7}$$

From (3.6) and (3.7) we arrive at the assertion. ■

Similar to (2.5), we mention here the inherent delay parabolic equation

$$\mathcal{C}\dot{\bar p} + \mathcal{M}\Big(\sum_{\ell=1}^{\delta} c_{\delta,\ell}\Delta_{\ell\tau}\dot{\bar p}\Big) + \mathcal{B}\bar p = r, \tag{3.8}$$

which is obtained from (3.3) by solving for $\bar u$ in the first equation.

4. Summation Assumption and Convergence Analysis

As discussed before, to obtain higher-order schemes for (2.4) which decouple, we apply an implicit discretization (of higher order) to the delay equation (3.3) with a constant time step size equal to the delay. More precisely, we discuss the application of the well-known BDF schemes. For a sequence $(y^n)_{n\in\mathbb{Z}}$ and a constant step size τ, we define the BDF-k operator

$$\Xi_k(y^n) := \frac{1}{\tau}\sum_{\ell=0}^{k}\xi_{k,\ell}\, y^{n-\ell}, \tag{4.1}$$

where the coefficients $\xi_{k,\ell}$ are listed in Table 4.1 (cf. [19, Ch. III.3]). Recall that BDF methods up to order $k = 6$ are stable [19, Ch. III.3]. The decoupling time discretization scheme for (2.4) is thus obtained by replacing the time derivatives in (3.3) with the BDF-k operator, i.e., we consider the scheme

$$\mathcal{A}u^n - \mathcal{D}^*\Psi_\delta(p^n) = f^n \qquad \text{in } \mathcal{V}^*, \tag{4.2a}$$

$$\mathcal{D}\Xi_k(u^n) + \mathcal{C}\Xi_k(p^n) + \mathcal{B}p^n = g^n \qquad \text{in } \mathcal{Q}^* \tag{4.2b}$$

Table 4.1. BDF-k coefficients $\xi_{k,\ell}$, $\ell = 0, \dots, k$, for different values of k.

k	$\xi_{k,0}$	$\xi_{k,1}$	$\xi_{k,2}$	$\xi_{k,3}$
1	1	-1		
2	$\frac{3}{2}$	-2	$\frac{1}{2}$	
3	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$

with the discrete delay operator

$$\Psi_\delta(p^n) := \sum_{\ell=1}^{\delta} c_{\delta,\ell}\, p^{n-\ell}. \tag{4.3}$$

Note that (4.2) defines a multistep scheme with $\max(k, \delta)$ steps.

Remark 4.1. Scheme (4.2) can also be interpreted as an (α, β, γ)-BDF discretization of (2.4). In [1, 14] the combination of BDF-k with an explicit multistep method of order k is considered, based on the characteristic polynomials

$$\alpha(\zeta) = \sum_{j=1}^{k} \tfrac{1}{j}\, \zeta^{k-j}(\zeta-1)^j, \qquad \beta(\zeta) = \zeta^k, \qquad \gamma(\zeta) = \zeta^k - (\zeta-1)^k.$$

This means that the standard BDF-k method – represented by (α, β) – is applied to all but the extrapolated values. For the latter, which in our case corresponds to p in (2.4a), a multistep discretization defined by γ is used instead. Since the coefficients of γ coincide with the delay coefficients given in Table 3.1, this then yields (4.2). Unfortunately, known convergence results for (α, β, γ) schemes do not apply for the here-considered elliptic–parabolic systems.

4.1. **Equivalent formulation**

To simplify the convergence analysis, we make the following observation. Solving (4.2a) for u^n, substituting the expression into (4.2b), and exploiting that the discrete delay operator and the BDF-k operator commute, we obtain

$$\mathcal{C}\Xi_k(p^n) + \mathcal{M}\Psi_\delta\Xi_k(p^n) + \mathcal{B}p^n = r^n \tag{4.4}$$

with $\mathcal{M} = \mathcal{D}\mathcal{A}^{-1}\mathcal{D}^*$ as before and $r^n := g^n - \mathcal{D}\mathcal{A}^{-1}\dot{f}^n$. We immediately notice that (4.4) equals the BDF-k discretization with step size τ of the operator delay equation (3.8). In particular, we conclude that instead of showing that the solution of (4.2) converges to the solution of (2.4), we can equivalently show that the solution of (4.4) converges to the solution of (3.8). Note, however, that the combination of the delay and the BDF approximation yields a $(\delta + k)$-step method for p. Afterwards, u can be reconstructed.

Example 4.2. For the specific choice $\delta = k$ we obtain the schemes

$$\begin{aligned}
k = 1: &\quad \mathcal{C}p^n - (\mathcal{C} - \mathcal{M})p^{n-1} - \mathcal{M}p^{n-2} + \tau\mathcal{B}p^n = \tau r^n,\\
k = 2: &\quad 3\mathcal{C}p^n - (4\mathcal{C} - 6\mathcal{M})p^{n-1} + (\mathcal{C} - 11\mathcal{M})p^{n-2} + 6\mathcal{M}p^{n-3} - \mathcal{M}p^{n-4} + 2\tau\mathcal{B}p^n = 2\tau r^n,\\
k = 3: &\quad 11\mathcal{C}p^n - (18\mathcal{C} - 33\mathcal{M})p^{n-1} + (9\mathcal{C} - 87\mathcal{M})p^{n-2} + (-2\mathcal{C} + 92\mathcal{M})p^{n-3}\\
&\qquad - 51\mathcal{M}p^{n-4} + 15\mathcal{M}p^{n-5} - 2\mathcal{M}p^{n-6} + 6\tau\mathcal{B}p^n = 6\tau r^n.
\end{aligned}$$

4.2. **Summation assumption**

For the ease of presentation, we introduce the following notation. Given a real Hilbert space $\mathcal{X}$, $s \in \mathbb{N}$, and a matrix $G \in \mathbb{S}^s_\succ$, where $\mathbb{S}^s_\succ$ is the set of real-valued symmetric positive definite matrices of size $s \times s$, we define the weighted inner product

$$\|Y\|_G^2 := \langle GY, Y\rangle_{\mathcal{X}^s} \qquad \text{for } Y \in \mathcal{X}^s.$$

Moreover, for a sequence $(y^n)_{n\in\mathbb{Z}}$ with $y^n \in \mathcal{X}$ and numbers $k, \delta \in \mathbb{N}$, we define

$$Y^n = \begin{bmatrix} y^n \\ \vdots \\ y^{n-k-\delta+1} \end{bmatrix} \in \mathcal{X}^{k+\delta}. \tag{4.6}$$

For the convergence analysis, we need the following algebraic criterion that leads to G–stability. Here, we follow the multiplier technique of Nevanlinna and Odeh [25] but formulate the criterion as an assumption which needs to be verified afterwards.

Assumption 1 (Summation assumption)**.** *Let $\mathcal{X}$ be a Hilbert space and $\omega, \tau > 0$. Then there exist a parameter $\eta \in [0, 1)$ and for any $\mu \in [0, \omega]$ a matrix $G \in \mathbb{S}^{k+\delta}_\succ$ and a vector $\gamma = [\gamma_0, \dots, \gamma_{k+\delta}] \in \mathbb{R}^{k+\delta+1}$ such that for any sequence $\{y^n\}_{n\in\mathbb{Z}}$ with $y^n \in \mathcal{X}$ for $n \in \mathbb{Z}$ we have*

$$\Big\langle \tau\Xi_k(y^n) + \mu\Psi_\delta(\tau\Xi_k(y^n)), y^n - \eta y^{n-1}\Big\rangle_{\mathcal{X}} = \|Y^n\|_G^2 - \|Y^{n-1}\|_G^2 + \Big\|\sum_{i=0}^{k+\delta} \gamma_i y^{n-i}\Big\|_{\mathcal{X}}^2, \tag{4.7}$$

where Y^n is defined as in (4.6) *and Ψ_δ is the discrete delay operator defined in* (4.3)*.*

Let us emphasize that Assumption 1 only formally depends on τ, since $\tau\Xi_k(\,\cdot\,)$ is independent of τ, and that the matrix G as well as the vector γ depend on μ. In the forthcoming convergence proof, we need to estimate the matrix G in terms of its smallest and largest eigenvalues and thus make the following assumption.

Assumption 2. *Let the notation be as in Assumption 1. Then*

$$c_G := \inf_{\mu\in[0,\omega]} \lambda(G(\mu)) > 0 \qquad \textit{and} \qquad C_G := \sup_{\mu\in[0,\omega]} \lambda(G(\mu)) < \infty, \tag{4.8}$$

where $\lambda(G)$ is the set of eigenvalues of the symmetric matrix G.

4.3. **Convergence analysis**

In the following, we use Assumptions 1 and 2 to prove the convergence of the BDF discretization for the delay system (3.8). The verification of these assumptions is then subject of the upcoming Section 4.4.

Theorem 4.3. *Consider the solution $\bar p$ of the delay equation* (3.8) *for a sufficiently smooth right-hand side $r\colon [0, T] \to \mathcal{Q}^*$. Let p^n be its approximation at time step t^n given by* (4.4)*. If Assumptions 1 and 2 hold, then we get*

$$\|p^n - \bar p(t^n)\|^2_{\mathcal{H}_\mathcal{Q}} + \tau\|p^n - \bar p(t^n)\|^2_\mathcal{Q} \lesssim \tau^{2k} + \sum_{\ell=-\delta-1}^{k-1} \|p^\ell - \bar p(t^\ell)\|^2_{\mathcal{H}_\mathcal{Q}} + \tau\,\|p^{k-1} - \bar p(t^{k-1})\|^2_\mathcal{Q}.$$

Proof. We define the error and the defect by

$$e^n := p^n - \bar p(t^n), \qquad d^n := \frac{1}{\tau}\left(\sum_{\ell=0}^{k} \xi_{k,\ell}\bar p(t^{n-\ell})\right) - \dot{\bar p}(t^n).$$

Due to the construction of the BDF operator, we immediately conclude that the defect satisfies the estimate $\|d^n\|_{\mathcal{H}_\mathcal{Q}} \lesssim \tau^k$. Considering the difference between (3.8) and (4.4) and testing with $e^n - \eta e^{n-1}$, where η is the multiplier from Assumption 1, we obtain

$$\Big\langle \mathcal{C}\tau\Xi_k(e^n) + \mathcal{M}\Psi_\delta(\tau\Xi_k(e^n)), e^n - \eta e^{n-1}\Big\rangle + \tau \left\langle \mathcal{B}e^n, e^n - \eta e^{n-1}\right\rangle = \tau \Big\langle d^n + \mathcal{M}\Psi_\delta(d^n), e^n - \eta e^{n-1}\Big\rangle. \tag{4.9}$$

Since $\mathcal{B}$ is self-adjoint, we have

$$2\left\langle \mathcal{B}e^n, e^n - \eta e^{n-1}\right\rangle = \|e^n\|_b^2 - \eta^2\|e^{n-1}\|_b^2 + \|e^n - \eta e^{n-1}\|_b^2 \geq \|e^n\|_b^2 - \eta^2\|e^{n-1}\|_b^2. \tag{4.10}$$

We now consider the spectral decomposition of the operator $\mathcal{M}$ with respect to the inner product defined by the bilinear form c. This means that we seek functions $\phi_i \in \mathcal{H}_\mathcal{Q}$ satisfying

$$\mathcal{M}\phi_j = \mu_j\mathcal{C}\phi_j \qquad \text{and} \qquad \langle \mathcal{C}\phi_j, \phi_i\rangle = \delta_{ji}.$$

Following the spectral result in [13, Thm. 6.11], this yields an *orthonormal basis* of $\mathcal{H}_\mathcal{Q}$ with respect to the c-norm. This conclusion relies on the fact that $\mathcal{M}$ is compact when interpreted as an operator mapping from $\mathcal{Q}$ to its dual. The compactness then follows from the compact embedding $\mathcal{V} \hookrightarrow \mathcal{H}_\mathcal{V}$. For the eigenvalues, we obtain the estimate $0 \leq \mu_j \leq \omega$, since the operator $\mathcal{C}^{-1}\mathcal{M}$ is bounded by ω. Now define $\epsilon_i^n := \langle \mathcal{C}e^n, \phi_i\rangle \in \mathbb{R}$ and $E_i^n := [\epsilon_i^n, \dots, \epsilon_i^{n-k-\delta+1}]^\mathsf{T} \in \mathbb{R}^{k+\delta}$, leading to the basis expansion $e^n = \sum_{i=1}^\infty \epsilon_i^n\phi_i$. Using Assumption 1 and that all eigenvalues are bounded above by ω, the first term in (4.9) can be rewritten as

$$\begin{aligned}\Big\langle \mathcal{C}\tau\Xi_k(e^n) + \mathcal{M}\Psi_\delta(\tau\Xi_k(e^n)), e^n - \eta e^{n-1}\Big\rangle_{\mathcal{H}_\mathcal{Q}} &= \sum_{i=1}^\infty \Big\langle \tau\Xi_k(\epsilon_i^n) + \mu_i\Psi_\delta(\tau\Xi_k(\epsilon^n)), \epsilon_i^n - \eta\epsilon_i^{n-1}\Big\rangle_\mathbb{R} \\ &\geq \sum_{i=1}^\infty \Big(\|E_i^n\|_{G(\mu_i)}^2 - \|E_i^{n-1}\|_{G(\mu_i)}^2\Big).\end{aligned}$$

Due to the orthonormality of the ϕ_i functions, we have $\|e^n\|_c^2 = \sum_{i=1}^\infty(\epsilon_i^n)^2$. In the same way, it holds that

$$\|E^n\|_c^2 = \bigg\langle \sum_{i=1}^\infty \mathcal{C}E_i^n\phi_i, \sum_{j=1}^\infty E_j^n\phi_j\bigg\rangle_{\mathcal{H}_\mathcal{Q}^{k+\delta}} = \sum_{i=1}^\infty\sum_{j=1}^\infty \langle \mathcal{C}E_i^n\phi_i, E_j^n\phi_j\rangle_{\mathcal{H}_\mathcal{Q}^{k+\delta}} = \sum_{i=1}^\infty \|E_i^n\|^2,$$

where $E_i^n \in \mathbb{R}^{k+\delta}$ and $E^n \in \mathcal{H}_\mathcal{Q}^{k+\delta}$ (see (4.6)). Using the weighted Young's inequality for (4.9), we obtain with (yet to be determined constants) $\delta_1, \delta_2, \delta_3, \delta_4 > 0$,

$$\begin{aligned}&\bigg\langle d^n + \mathcal{M}\bigg(\sum_{\ell=1}^\delta c_{\delta,\ell}d^{n-\ell}\bigg), e^n - \eta e^{n-1}\bigg\rangle \\ &\quad= \langle d^n, e^n\rangle - \eta\left\langle d^n, e^{n-1}\right\rangle + \sum_{\ell=1}^\delta c_{\delta,\ell}\Big(\langle \mathcal{M}d^{n-\ell}, e^n\rangle - \eta\left\langle \mathcal{M}d^{n-\ell}, e^{n-1}\right\rangle\Big) \\ &\quad\leq \big(\tfrac{\delta_1}{2}\|e^n\|^2 + \tfrac{1}{2\delta_1}\|d^n\|^2\big) + \eta\big(\tfrac{\delta_2}{2}\|e^{n-1}\|^2 + \tfrac{1}{2\delta_2}\|d^n\|^2\big) \\ &\qquad+ \sum_{\ell=1}^\delta |c_{\delta,\ell}|\Big(\tfrac{\delta_3}{2}\|e^n\|^2 + \tfrac{1}{2\delta_3}\|\mathcal{M}d^{n-\ell}\|^2\Big) + \eta\sum_{\ell=1}^\delta |c_{\delta,\ell}|\Big(\tfrac{\delta_4}{2}\|e^{n-1}\|^2 + \tfrac{1}{2\delta_4}\|\mathcal{M}d^{n-\ell}\|^2\Big) \\ &\quad\leq \big(\tfrac{\delta_1}{2} + \tfrac{\bar{c}_\delta\delta_3}{2}\big)\|e^n\|^2 + \eta\big(\tfrac{\delta_2}{2} + \tfrac{\bar{c}_\delta\delta_4}{2}\big)\|e^{n-1}\|^2 + \big(\tfrac{1}{2\delta_1} + \tfrac{\eta}{2\delta_2}\big)\|d^n\|^2 \\ &\qquad+ c_c^2\,\omega^2\big(\tfrac{1}{2\delta_3} + \tfrac{\eta}{2\delta_4}\big)\sum_{\ell=1}^\delta |c_{\delta,\ell}|\|d^{n-\ell}\|^2,\end{aligned}$$

where $\bar{c}_\delta := \sum_{\ell=1}^{\delta} |c_{\delta,\ell}|$. Note that we have made use of $\|\mathcal{M}\cdot\| \leq c_c\,\omega\|\cdot\|$ in the last inequality. With the choice

$$\delta_1 = \frac{c_b}{2}(1-\eta^2), \qquad \delta_2 = \frac{\eta}{2}, \qquad \delta_3 = \frac{c_b}{2}\frac{1}{\bar{c}_\delta}(1-\eta^2), \qquad \delta_4 = \frac{\eta}{2}\frac{1}{\bar{c}_\delta},$$

we obtain together with (4.10) the estimate

$$\begin{aligned}\sum_{i=1}^{\infty}\Big(\|E_i^n\|^2_{G(\mu_i)} - \|E_i^{n-1}\|^2_{G(\mu_i)}\Big) &+ \tfrac{1}{2}\tau\eta^2\|e^n\|_b^2 - \tfrac{1}{2}\tau\eta^2\|e^{n-1}\|_b^2 - \tfrac{1}{2}\tau\eta^2\|e^{n-1}\|^2 \\ &\leq \Big(1+\tfrac{1}{c_b(1-\eta^2)}\Big)\Big(\tau\,\|d^n\|^2 + \tau c_c^2\,\omega^2\bar{c}_\delta\sum_{\ell=1}^{\delta}|c_{\delta,\ell}|\|d^{n-\ell}\|^2\Big) \\ &\lesssim (1+\omega^2)\,\tau^{2k+1}.\end{aligned}$$

Using Assumption 2, we observe that

$$\|e^{n-1}\|^2 \leq \tfrac{1}{c_c}\|e^{n-1}\|_c^2 = \tfrac{1}{c_c}\sum_{i=1}^{\infty}(\epsilon_i^{n-1})^2 \leq \tfrac{1}{c_c}\sum_{i=1}^{\infty}\tfrac{1}{c_{G(\mu_i)}}\|E_i^{n-1}\|^2_{G(\mu_i)} \leq \tfrac{1}{c_c c_G}\sum_{i=1}^{\infty}\|E_i^{n-1}\|^2_{G(\mu_i)}.$$

Overall, this yields the perturbed telescope sum

$$\sum_{i=1}^{\infty}\Big(\|E_i^n\|^2_{G(\mu_i)} - (1+\tau C)\|E_i^{n-1}\|^2_{G(\mu_i)}\Big) + \tfrac{1}{2}\tau\eta^2\Big(\|e^n\|_b^2 - \|e^{n-1}\|_b^2\Big) \lesssim (1+\omega^2)\,\tau^{2k+1},$$

where $C = \eta^2\frac{1}{2c_c c_G}$. To see that the Grönwall-type inequality from [8, Lem. 2.5] can be applied, define $a^n := \sum_{i=1}^{\infty}\|E_i^n\|^2_{G(\mu_i)}$. This then leads to

$$\sum_{i=1}^{\infty}\|E_i^n\|^2_{G(\mu_i)} + \tau\,\|e^n\|_b^2 \lesssim e^{Cn\tau}\Big(\sum_{i=1}^{\infty}\|E_i^{k-1}\|^2_{G(\mu_i)} + \tau\,\|e^{k-1}\|_b^2 + t^n\tau^{2k}\Big).$$

Together with

$$\sum_{i=1}^{\infty}\|E_i^n\|^2_{G(\mu_i)} \geq c_G\sum_{i=1}^{\infty}\|E_i^n\|^2 = c_G\|E^n\|_c^2 \geq c_G\|e^n\|_c^2$$

and

$$\sum_{i=1}^{\infty}\|E_i^{k-1}\|^2_{G(\mu_i)} \leq C_G\sum_{i=1}^{\infty}\|E_i^{k-1}\|^2 = C_G\,\|E^{k-1}\|_c^2 = C_G\,\Big(\|e^{k-1}\|_c^2 + \cdots + \|e^{-\delta-1}\|_c^2\Big),$$

we obtain the desired estimate. ∎

Remark 4.4. Suppose that the assumptions from Proposition 3.4 and Assumptions 1 and 2 hold. Let (u,p) be the solution of the original system (2.4) and (u^n,p^n) the numerical approximation at time t^n given by (4.2). Then the difference satisfies

$$\|u^n - u(t^n)\|^2_{\mathcal{V}} + \|p^n - p(t^n)\|^2_{\mathcal{H}_Q} \lesssim \tau^{2\delta} + \tau^{2k} + \sum_{\ell=0}^{k-1}\|p^\ell - p(t^\ell)\|^2_{\mathcal{H}_Q},$$

indicating that we obtain an optimal (balanced) error for the choice $\delta = k$.

4.4. **Verification of the summation assumption**

It remains to validate Assumptions 1 and 2. Since our main goal is the error analysis for the decoupling integration scheme, we follow Remark 4.4 and consider $k = \delta$.

4.4.1. *Case $k=\delta=1$*

Rewriting (4.7) for $k=\delta=1$, we get

$$\begin{aligned}\langle y^n - y^{n-1} &+ \mu(y^{n-1}-y^{n-2}), y^n - \eta y^{n-1}\rangle \\ &= \left\langle \begin{bmatrix} g_{11} & g_{12} \\ g_{12} & g_{22}\end{bmatrix}\begin{bmatrix} y^n \\ y^{n-1}\end{bmatrix}, \begin{bmatrix} y^n \\ y^{n-1}\end{bmatrix}\right\rangle - \left\langle \begin{bmatrix} g_{11} & g_{12} \\ g_{12} & g_{22}\end{bmatrix}\begin{bmatrix} y^{n-1} \\ y^{n-2}\end{bmatrix}, \begin{bmatrix} y^{n-1} \\ y^{n-2}\end{bmatrix}\right\rangle \\ &\quad + \left\|\gamma_0 y^n + \gamma_1 y^{n-1} + \gamma_2 y^{n-2}\right\|^2,\end{aligned}$$

which on rearranging gives

$$\begin{aligned}(g_{11} + \gamma_0^2 - 1)\langle y^n, y^n\rangle &+ (\eta\mu - \eta - g_{11} + g_{22} + \gamma_1^2)\langle y^{n-1}, y^{n-1}\rangle \\ + (-g_{22} + \gamma_2^2)\langle y^{n-2}, y^{n-2}\rangle &+ (\eta + 2g_{12} + 2\gamma_0\gamma_1 - \mu + 1)\langle y^n, y^{n-1}\rangle \\ + (2\gamma_0\gamma_2 + \mu)\langle y^n, y^{n-2}\rangle &+ (-\eta\mu - 2g_{12} + 2\gamma_1\gamma_2)\langle y^{n-1}, y^{n-2}\rangle = 0.\end{aligned}$$

Equating the coefficients of each distinct inner product to zero, we obtain the system of equations

$$g_{11} + \gamma_0^2 = 1, \qquad -g_{11} + g_{22} + \gamma_1^2 = \eta(1-\mu), \qquad 2g_{12} + 2\gamma_0\gamma_1 = -1 + \mu - \eta, \tag{4.11a}$$

$$-g_{22} + \gamma_2^2 = 0, \qquad 2\gamma_0\gamma_2 = -\mu, \qquad -2g_{12} + 2\gamma_1\gamma_2 = \eta\mu. \tag{4.11b}$$

Solving for the variables g_{11}, g_{22}, g_{12} in terms of $\gamma_0, \gamma_1, \gamma_2$ from (4.11a), we obtain

$$G = \begin{bmatrix} 1-\gamma_0^2 & -\gamma_0\gamma_1 + \frac{1}{2}(\mu - \eta - 1) \\ -\gamma_0\gamma_1 + \frac{1}{2}(\mu-\eta-1) & -\gamma_0^2 - \gamma_1^2 + 1 - \eta\mu + \eta\end{bmatrix}$$

and the γ_i's have to be solved from

$$2\gamma_0\gamma_2 = -\mu, \tag{4.12a}$$

$$2\gamma_0\gamma_1 + 2\gamma_1\gamma_2 = \eta\mu - \eta + \mu - 1, \tag{4.12b}$$

$$\gamma_0^2 + \gamma_1^2 + \gamma_2^2 = 1 - \eta\mu + \eta. \tag{4.12c}$$

Summing up the equations in (4.12) yields $(\sum_{i=0}^{3}\gamma_i)^2 = 0$ and we conclude that

$$\gamma_0 + \gamma_1 + \gamma_2 = 0. \tag{4.13}$$

From (4.12b) and (4.13) we further obtain

$$\gamma_1^2 = \tfrac{1}{2}(1+\eta)(1-\mu).$$

Hence, there cannot exist a (real-valued) solution for $\mu > 1$, i.e., we recover the weak coupling condition from [5]. From (4.12a) and (4.13) we solve for γ_0 and γ_2 to obtain

$$\begin{aligned}\gamma_0 &= -\tfrac{1}{2}\sqrt{\tfrac{1}{2}(1+\eta)(1-\mu)} + \sqrt{\tfrac{1}{2}\mu + \tfrac{1}{8}(1+\eta)(1-\mu)}, \\ \gamma_2 &= -\tfrac{1}{2}\sqrt{\tfrac{1}{2}(1+\eta)(1-\mu)} - \sqrt{\tfrac{1}{2}\mu + \tfrac{1}{8}(1+\eta)(1-\mu)}, \\ \gamma_1 &= \sqrt{\tfrac{1}{2}(1+\eta)(1-\mu)}\end{aligned}$$

as a solution. Setting $\eta = 0$ in the above expressions, the eigenvalues of the G matrix are

$$\begin{aligned}\lambda_1 &= \tfrac{1}{4}\sqrt{1-\mu}\sqrt{3\mu+1} - \tfrac{1}{4}\sqrt{(3+\mu)(1-\mu) + 2(1-\mu)\sqrt{1-\mu}\sqrt{3\mu+1}} + \tfrac{1}{2}, \\ \lambda_2 &= \tfrac{1}{4}\sqrt{1-\mu}\sqrt{3\mu+1} + \tfrac{1}{4}\sqrt{(3+\mu)(1-\mu) + 2(1-\mu)\sqrt{1-\mu}\sqrt{3\mu+1}} + \tfrac{1}{2}\end{aligned}$$

and a necessary and sufficient condition for $\lambda_1, \lambda_2 \in \mathbb{R}_+$ is $0 \le \mu \le 1$. We conclude that Assumptions 1 and 2 are satisfied. We refer to Figure 4.1a for a visualization of the eigenvalues.

4.4.2. *Case $k = \delta = 2$*

Equating the coefficients of all distinct inner products in (4.7) results in a system with 15 unknowns and 15 equations (while treating μ and η as free parameters). Similarly as in the previous subsection, we can write all entries of G as functions of the γ_i. The details are given in Appendix B.1. In turn, $\gamma_0, \dots, \gamma_4$ have to satisfy the following five nonlinear equations

$$2\gamma_0\gamma_4 = -\tfrac{1}{2}\mu, \tag{4.14a}$$
$$2\gamma_0\gamma_3 + 2\gamma_1\gamma_4 = \tfrac{1}{2}\eta\mu + 3\mu, \tag{4.14b}$$
$$2\gamma_0\gamma_2 + 2\gamma_1\gamma_3 + 2\gamma_2\gamma_4 = -3\eta\mu - \tfrac{11}{2}\mu + \tfrac{1}{2}, \tag{4.14c}$$
$$2\gamma_0\gamma_1 + 2\gamma_1\gamma_2 + 2\gamma_2\gamma_3 + 2\gamma_3\gamma_4 = \tfrac{11}{2}\eta\mu - 2\eta + 3\mu - 2, \tag{4.14d}$$
$$\gamma_0^2 + \gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \gamma_4^2 = -3\eta\mu + 2\eta + \tfrac{3}{2}. \tag{4.14e}$$

Lemma 4.5. *Consider* (4.14) *and assume $\eta \geq 0$. Then,* (4.14) *is not solvable for $\mu > \frac{1}{3}$. For $\mu \leq \frac{1}{3}$ a solution is given by*

$$\gamma_0 = \tfrac{\theta}{2} + \tfrac{\sqrt{\mu+\theta^2}}{2}, \qquad \gamma_2 = -\tfrac{3\sqrt{1-3\mu}}{4} + \tfrac{\sqrt{5\mu+1}}{4}, \qquad \gamma_4 = \tfrac{\theta}{2} - \tfrac{\sqrt{\mu+\theta^2}}{2}, \tag{4.15a}$$
$$\gamma_1 = \tfrac{\sqrt{1-3\mu}}{2} - \tfrac{3\mu-\sqrt{1-3\mu}\theta}{2\sqrt{\mu+\theta^2}}, \qquad \gamma_3 = \tfrac{\sqrt{1-3\mu}}{2} + \tfrac{3\mu-\sqrt{1-3\mu}\theta}{2\sqrt{\mu+\theta^2}} \qquad \eta = 0 \tag{4.15b}$$

with $\theta = -\frac{1}{4}\big(\sqrt{5\mu+1} + \sqrt{1-3\mu}\big)$. For this particular solution, the matrix G is positive definite for every $\mu \in [0, \frac{1}{3}]$. Hence, Assumptions 1 and 2 are true for $k = \delta = 2$ and $\omega \leq \frac{1}{3}$.

Proof. Summing up the equations in (4.14) yields $(\sum_{i=0}^4 \gamma_i)^2 = 0$ and, hence, $\sum_{i=0}^4 \gamma_i = 0$. Factoring the sum of equations (4.14b) and (4.14d) gives

$$2(\gamma_1 + \gamma_3)(\gamma_0 + \gamma_2 + \gamma_4) = 6\eta\mu + 6\mu - 2\eta - 2 = 2(1+\eta)(3\mu - 1).$$

Together with the fact that the γ_i's sum up to zero, we obtain

$$(\gamma_1 + \gamma_3)^2 = (1+\eta)(1-3\mu).$$

We conclude that a (real-valued) solution cannot exist for $\mu > \frac{1}{3}$, which establishes the first claim. The second claim is obtained by directly verifying (4.14). In more detail, we immediately see that (4.15) satisfies (4.14a). For (4.14b), we compute

$$2\gamma_0\gamma_3 + 2\gamma_1\gamma_4 = 4\tfrac{\theta}{2}\tfrac{\sqrt{1-3\mu}}{2} + 4\tfrac{\sqrt{\mu+\theta^2}}{2}\tfrac{3\mu-\theta\sqrt{1-3\mu}}{2\sqrt{\mu+\theta^2}} = 3\mu.$$

Observe that

$$\begin{aligned}\frac{(3\mu - \theta\sqrt{1-3\mu})^2}{\mu+\theta^2} &= \frac{\frac{1}{8}(1+10\mu+33\mu^2) + \frac{1}{8}(1+9\mu)\sqrt{1+5\mu}\sqrt{1-3\mu}}{\frac{1}{8}(1+9\mu) + \frac{1}{8}\sqrt{1+5\mu}\sqrt{1-3\mu}} \\ &= \frac{\mu(1+6\mu)(1+9\mu) + \mu(1+6\mu)\sqrt{1+5\mu}\sqrt{1-3\mu}}{2\mu(1+6\mu)} \\ &= \tfrac{1}{2}(1+9\mu) + \tfrac{1}{2}\sqrt{1+5\mu}\sqrt{1-3\mu}.\end{aligned}$$

Towards (4.14c), we obtain

$$\begin{aligned}(\gamma_0+\gamma_4)\gamma_2 &= -\tfrac{3\theta\sqrt{1-3\mu}}{4} + \tfrac{\theta\sqrt{5\mu+1}}{4} = \tfrac{1-7\mu}{8} + \tfrac{\sqrt{1-3\mu}\sqrt{5\mu+1}}{8}, \\ \gamma_1\gamma_3 &= \tfrac{1-3\mu}{4} - \tfrac{(3\mu-\theta\sqrt{1-3\mu})^2}{4(\mu+\theta^2)} = -\tfrac{\sqrt{1-3\mu}\sqrt{5\mu+1}}{8} + \tfrac{1-15\mu}{8}, \\ \gamma_0\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_4 &= \tfrac{1-7\mu}{8} + \tfrac{1-15\mu}{8} = \tfrac{1}{4} - \tfrac{11}{4}\mu.\end{aligned}$$

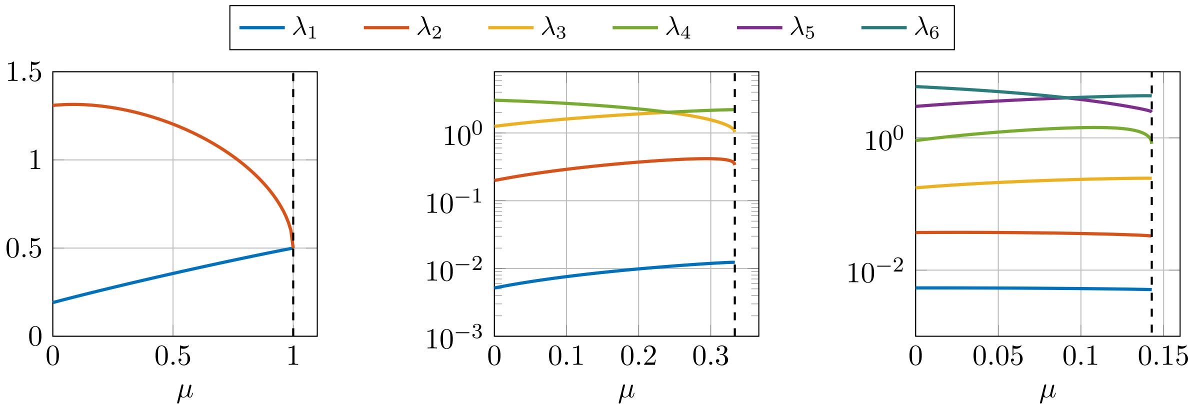
Towards (4.14d), we have

$$\begin{aligned}\gamma_2(\gamma_1+\gamma_3)+\gamma_0\gamma_1+\gamma_3\gamma_4 &= \Big(-\tfrac{3\sqrt{1-3\mu}}{4}+\tfrac{\sqrt{5\mu+1}}{4}\Big)\sqrt{1-3\mu}+\tfrac{\theta}{2}\sqrt{1-3\mu}-\tfrac{3\mu-\theta\sqrt{1-3\mu}}{2}\\ &= (-\sqrt{1-3\mu}-\theta)\sqrt{1-3\mu}+\tfrac{\theta}{2}\sqrt{1-3\mu}-\tfrac{3\mu-\theta\sqrt{1-3\mu}}{2}\\ &= -1+\tfrac{3}{2}\mu.\end{aligned}$$

Finally, for (4.14e), we have

$$\begin{aligned}\sum_{i=0}^{4}\gamma_i^2 &= 2\Big(\tfrac{\theta^2}{4}+\tfrac{\mu+\theta^2}{4}\Big)+2\Big(\tfrac{1-3\mu}{4}+\tfrac{(3\mu-\theta^2\sqrt{1-3\mu})^2}{4(\mu+\theta^2)}\Big)+\Big(-\tfrac{3\sqrt{1-3\mu}}{4}+\tfrac{\sqrt{5\mu+1}}{4}\Big)^2\\ &= \tfrac{\mu}{2}+\tfrac{1}{16}(1-3\mu+1+5\mu+2\sqrt{1-3\mu}\sqrt{5\mu+1})\\ &\quad+\tfrac{1-3\mu}{2}+\tfrac{1}{4}(1+9\mu)+\tfrac{1}{4}\sqrt{1+5\mu}\sqrt{1-3\mu}\\ &\quad+\tfrac{1}{16}(9-27\mu+1+5\mu-6\sqrt{1-3\mu}\sqrt{5\mu+1})\\ &= \tfrac{1}{2}-\mu+\tfrac{1}{16}(12-20\mu)+\tfrac{1}{4}(1+9\mu)=\tfrac{3}{2}.\end{aligned}$$

For the positive definiteness of the G matrix, we substitute the above analytical expressions for the γ_i's and numerically compute the eigenvalues of the parametrized G matrix w.r.t. μ. One can see in Figure 4.1b that the eigenvalues are all bounded from above and away from zero. ■



(A) $k=1$ (B) $k=2$ (C) $k=3$

Figure 4.1. Eigenvalues of $G(\mu)$ for different k. The dashed lines indicate the critical values from Theorem 4.9, i.e., $\omega=1$, $\omega=\frac{1}{3}$, and $\omega=\frac{1}{7}$.

Remark 4.6. The condition $\mu\le\frac{1}{3}$ improves upon the existing bound $\mu\le\frac{1}{5}$ used in the convergence proof in [6] and is in agreement with the asymptotic stability analysis for the delay equation. We refer to [6, §3.5] for further details.

Remark 4.7. There exists other solutions of (4.14) such that the eigenvalues of the G matrix are not bounded away from zero, namely

$$\begin{aligned}\gamma_0 &= \tfrac{\theta}{2}+\tfrac{\sqrt{\mu+\theta^2}}{2}, & \gamma_2 &= -\tfrac{3\sqrt{1-3\mu}}{4}-\tfrac{\sqrt{5\mu+1}}{4}, & \gamma_4 &= \tfrac{\theta}{2}-\tfrac{\sqrt{\mu+\theta^2}}{2},\\ \gamma_1 &= \tfrac{\sqrt{1-3\mu}}{2}+\tfrac{\theta\sqrt{1-3\mu}-3\mu}{2\sqrt{\mu+\theta^2}}, & \gamma_3 &= \tfrac{\sqrt{1-3\mu}}{2}-\tfrac{\theta\sqrt{1-3\mu}-3\mu}{2\sqrt{\mu+\theta^2}}, & \eta &= 0\end{aligned}$$

with $\theta=\frac{1}{4}(\sqrt{5\mu+1}-\sqrt{1-3\mu})$.

4.4.3. *Case $k = \delta = 3$*

Equating the coefficients of all distinct inner products in (4.7) results in a system of 28 unknowns and 28 equations, again treating μ and η as free parameters. For (4.7) to hold, we obtain G as in (B.1) such that γ_i, $0 \le i \le 6$, have to satisfy the following (nonlinear) system of equations

$$2\gamma_0\gamma_6 = -\tfrac{1}{3}\mu, \tag{4.16a}$$

$$2\gamma_0\gamma_5 + 2\gamma_1\gamma_6 = \tfrac{1}{3}\eta\mu + \tfrac{5}{2}\mu, \tag{4.16b}$$

$$2\gamma_0\gamma_4 + 2\gamma_1\gamma_5 + 2\gamma_2\gamma_6 = -\tfrac{5}{2}\eta\mu - \tfrac{17}{2}\mu, \tag{4.16c}$$

$$2\gamma_0\gamma_3 + 2\gamma_1\gamma_4 + 2\gamma_2\gamma_5 + 2\gamma_3\gamma_6 = \tfrac{17}{2}\eta\mu + \tfrac{46}{3}\mu - \tfrac{1}{3}, \tag{4.16d}$$

$$2\gamma_0\gamma_2 + 2\gamma_1\gamma_3 + 2\gamma_2\gamma_4 + 2\gamma_3\gamma_5 + 2\gamma_4\gamma_6 = -\tfrac{46}{3}\eta\mu + \tfrac{1}{3}\eta - \tfrac{29}{2}\mu + \tfrac{3}{2}, \tag{4.16e}$$

$$2\gamma_0\gamma_1 + 2\gamma_1\gamma_2 + 2\gamma_2\gamma_3 + 2\gamma_3\gamma_4 + 2\gamma_4\gamma_5 + 2\gamma_5\gamma_6 = \tfrac{29}{2}\eta\mu - \tfrac{10}{3}\eta + \tfrac{11}{2}\mu - 3, \tag{4.16f}$$

$$\gamma_0^2 + \gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \gamma_4^2 + \gamma_5^2 + \gamma_6^2 = -\tfrac{11}{2}\eta\mu + 3\eta + \tfrac{11}{6}. \tag{4.16g}$$

Lemma 4.8. *Consider* (4.16) *and assume $\eta \ge 0$. Then,* (4.16) *is not solvable for $\mu > \frac{1}{7}$. For $\mu \le \frac{1}{7}$,* (4.16) *is solvable and there exist a positive definite matrix G for this solution. Hence, Assumptions 1 and 2 are true for $k = \delta = 3$ and $\omega \le \frac{1}{7}$.*

Proof. Summing up the equations in (4.16) yields again $(\sum_{i=0}^{6}\gamma_i)^2 = 0$ and $\sum_{i=0}^{6}\gamma_i = 0$. Now consider the sum of equations (4.16b), (4.16d), and (4.16f), since these contain terms with product of γ_i's with odd and even indices, which on factoring gives

$$\begin{aligned} 2\,(\gamma_1+\gamma_3+\gamma_5)(\gamma_0+\gamma_2+\gamma_4+\gamma_6) &= \Big(\tfrac{1}{3}+\tfrac{17}{2}+\tfrac{29}{2}\Big)\eta\mu + \Big(\tfrac{5}{2}+\tfrac{46}{3}+\tfrac{11}{2}\Big)\mu - \tfrac{10}{3}\eta + \Big(-\tfrac{1}{3}-3\Big) \\ &= \tfrac{70}{3}\eta\mu + \tfrac{70}{3}\mu - \tfrac{10}{3}\eta - \tfrac{10}{3} \\ &= \tfrac{10}{3}(1+\eta)(7\mu-1). \end{aligned}$$

This yields

$$(\gamma_1+\gamma_3+\gamma_5)^2 = \tfrac{5}{3}\,(1+\eta)(1-7\mu),$$

which establishes the first claim. For the second claim, we verify numerically that (4.16) is solvable; see Figure 4.2 for the computed $\gamma_0,\dots,\gamma_6$ and the residual norm for solving system (4.16) numerically. The maximum of the residual norms is several orders of magnitude smaller than the minimum eigenvalue, which is reported in Figure 4.1c. Hence, the computed eigenvalues are numerically reliable and bounded away from zero, i.e., the matrix G is positive definite. ∎

4.5. **Summary**

We summarize the previous subsections concerning the k-step schemes (with $k = \delta$) in the upcoming Theorem 4.9. In the formulation including p and u, these schemes read for $k = 1$:

$$\begin{aligned} \mathcal{A}u^n - \mathcal{D}^*p^{n-1} &= f^n, \\ \tfrac{1}{\tau}\mathcal{D}(u^n - u^{n-1}) + \tfrac{1}{\tau}\mathcal{C}(p^n - p^{n-1}) + \mathcal{B}p^n &= g^n, \end{aligned}$$

for $k = 2$:

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

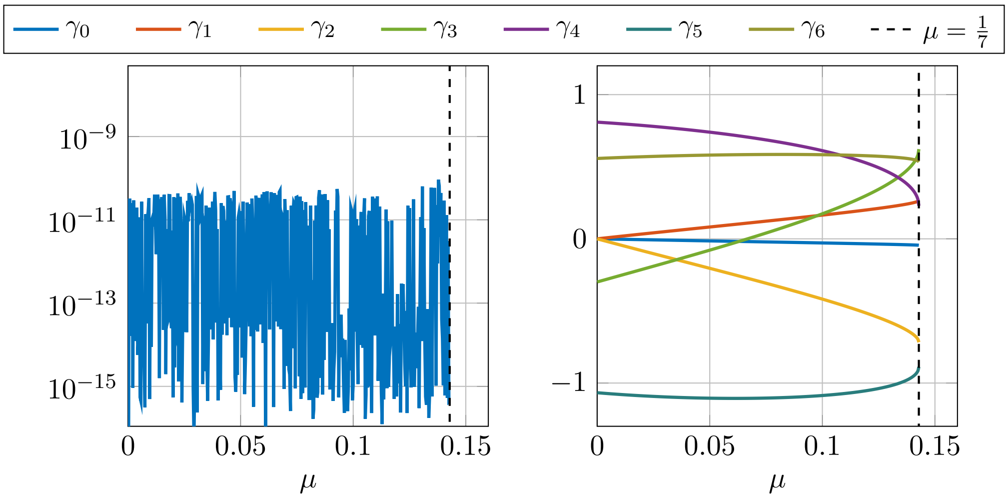


Figure 4.2. Residual norm (left) for solving the system for γ for BDF-3 with $\eta = 0.12$ and the corresponding solutions γ (right). The dashed lines indicate the critical value $\omega = \frac{1}{7}$ from Theorem 4.9.

and, for $k = 3$:

$$\mathcal{A}u^n - \mathcal{D}^*(3p^{n-1} - 3p^{n-2} + p^{n-3}) = f^n,$$
$$\tfrac{1}{\tau}\mathcal{D}(\tfrac{11}{6}u^n - 3u^{n-1} + \tfrac{3}{2}u^{n-2} - \tfrac{1}{3}u^{n-3}) + \tfrac{1}{\tau}\mathcal{C}(\tfrac{11}{6}p^n - 3p^{n-1} + \tfrac{3}{2}p^{n-2} - \tfrac{1}{3}p^{n-3}) + \mathcal{B}p^n = g^n.$$

Theorem 4.9. *Assume that one of the following applies:*

(1) $k = 1$ *and* $\omega \le 1$,

(2) $k = 2$ *and* $\omega \le \frac{1}{3}$,

(3) $k = 3$ *and* $\omega \le \frac{1}{7}$.

Then Assumptions 1 and 2 are true for $\delta = k$. *Hence, for sufficiently smooth right-hand sides* f, g *and consistent initial data, the solution of the original system* (2.5) *and its* k*-th order BDF approximation given by* (4.4) *satisfy*

$$\|u^n - u(t^n)\|_{\mathcal{V}}^2 + \|p^n - p(t^n)\|_{\mathcal{H}_\mathcal{Q}}^2 \lesssim \tau^{2k} + \sum_{\ell=0}^{k-1} \|p^\ell - p(t^\ell)\|_{\mathcal{H}_\mathcal{Q}}^2.$$

Note that the results are in line with the numerical observations in [6]. A numerical verification of the derived convergence results is subject of the following section.

5. Numerical Experiment

On the unit square $\Omega = (0,1)^2$ and $t \in [0, 10]$, we set as exact solution

$$u(t,x,y) = -10\mathrm{e}^{-\frac{5}{21}t}\begin{bmatrix}\cos(\pi x)\sin(\pi y)\\ \sin(\pi x)\cos(\pi y)\end{bmatrix}, \qquad p(t,x,y) = 10\mathrm{e}^{-\frac{5}{21}t}\sin(\pi x)\sin(\pi y)$$

and define the right-hand sides accordingly. The poroelasticity parameters are chosen as in Table 5.1, which results in a coupling strength of $\omega \approx 0.11 < 1/7$. A (uniform) mesh size of 2^{-7} is chosen for the spatial discretization. To make the errors in time dominate, it is reasonable to consider (P_3, P_2)

Table 5.1. Poroelasticity parameters for the numerical study in Section 5.

λ	μ	$\frac{\kappa}{\nu}$	M	α
0.5	0.125	0.05	0.27	0.5

finite elements for (u, p). Note that the polynomial degree for u is one higher than for p as suggested in [18]. For customary verification of the convergence result, the errors from the monolithic and the corresponding decoupled methods at the final time are plotted in Figure 5.1, showing the expected convergence rates.

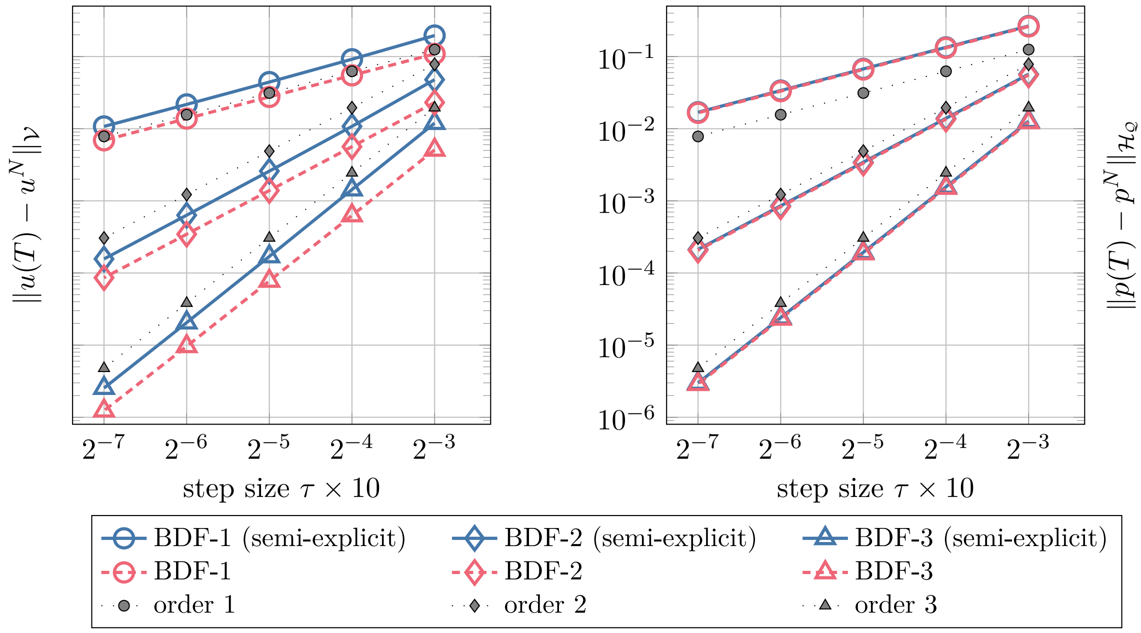


Figure 5.1. Convergence study for the numerical experiment on the unit square at time $t = T = 10$ with mesh size $h = 2^{-7}$. The solid blue curves correspond to the semi-explicit method and the dashed red curves to the fully coupled methods (original BDF schemes). Errors in u (left) and p (right).

6. **Conclusions**

The convergence of semi-explicit methods based on BDF-k schemes is proven for $k \leq 3$. The essential tool of the proof is the assumption on the G-stability of the schemes. In contrast to the standard G-stability concept, the inclusion of delay terms leads to a dependence on the spectrum of the involved operators of the problem. The evidence for this assumption is verified in this article.

In future work, we aim to show that the semi-explicit schemes are A-stable for appropriate multipliers, which would directly imply G-stability. Moreover, we plan to extend the convergence proof to the semi-explicit methods based on Runge–Kutta schemes.

Appendix A. Auxiliary Results

Within the analysis of the elliptic–parabolic system of Section 3 we use the following approximation property of the delay approach.

Lemma A.1. *Consider* $\bar{p} \in C^{\delta+1}([-\delta\tau, T]; \mathcal{H}_Q)$ *and* $\widehat{p}(\,\cdot\,;\tau) = \sum_{\ell=1}^{\delta} c_{\delta,\ell}\Delta_{\ell\tau}\bar{p}$*. Then,*

$$\bar{p}(t) - \widehat{p}(t;\tau) = -R_T^{\delta,\delta} = \mathcal{O}(\tau^\delta) \qquad \text{and} \qquad \dot{\bar{p}}(t) - \dot{\widehat{p}}(t;\tau) = -R_T^{\delta,\delta+1} = \mathcal{O}(\tau^\delta)$$

with remainder $R_T^{\delta,k} = \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\ell\tau-\zeta)^{\delta-1}\bar{p}^{(k)}(\zeta)\,\mathrm{d}\zeta$.

Proof. The Taylor series expansion of $\Delta_{\ell\tau}\bar{p}$ around t is

$$\Delta_{\ell\tau}\bar{p}(t) = \bar{p}(t-\ell\tau) = \sum_{i=0}^{\delta-1}\frac{(-\ell\tau)^i}{i!}\bar{p}^{(i)}(t) - \frac{1}{(\delta-1)!}\int_{t-\ell\tau}^{t}(t-\ell\tau-\zeta)^{\delta-1}\bar{p}^{(\delta)}(\zeta)\,\mathrm{d}\zeta.$$

Substituting this in (3.2), we obtain

$$\widehat{p}(t;\tau) = \sum_{\ell=1}^{\delta} c_{\delta,\ell}\Delta_{\ell\tau}\bar{p}(t) = \sum_{i=0}^{\delta-1}\left(\sum_{\ell=1}^{\delta}\ell^i(-1)^{(\ell-1)}\binom{\delta}{\ell}\right)\frac{(-\tau)^i}{i!}\bar{p}^{(i)}(t) + R_T^{\delta,\delta}, \tag{A.1}$$

where

$$R_T^{\delta,\delta} = \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\ell\tau-\zeta)^{\delta-1}\bar{p}^{(\delta)}(\zeta)\,\mathrm{d}\zeta. \tag{A.2}$$

The first summand in (A.1), i.e., for $i=0$, we obtain

$$\sum_{\ell=1}^{\delta}\binom{\delta}{\ell}(-1)^{(\ell-1)}\bar{p}(t) = \left(-(1-1)^\delta + 1\right)\bar{p}(t) = \bar{p}(t).$$

Observe that for $0 < i < \delta$, we can obtain the inner sum in (A.1) by substituting $x=-1$ in

$$\begin{aligned}
\sum_{\ell=1}^{\delta}\binom{\delta}{\ell}\ell^i\,x^{(\ell-1)} &= \frac{1}{x}\left[x\frac{\mathrm{d}}{\mathrm{d}x}(\,\cdot\,)\right]^i((1+x)^\delta - 1) = \frac{1}{x}\left[(1+x)\frac{\mathrm{d}}{\mathrm{d}x}(\,\cdot\,) - \frac{\mathrm{d}}{\mathrm{d}x}(\,\cdot\,)\right]^i(1+x)^\delta \\
&= \frac{1}{x}\sum_{\ell=0}^{i}\binom{i}{\ell}\left(\left[(1+x)\frac{\mathrm{d}}{\mathrm{d}x}(\,\cdot\,)\right]^{i-\ell}(1+x)^\delta\right)\left(\left[-\frac{\mathrm{d}}{\mathrm{d}x}(\,\cdot\,)\right]^{\ell}(1+x)^\delta\right) \\
&= \frac{1}{x}\sum_{\ell=0}^{i}\binom{i}{\ell}\left(\delta^{(i-\ell)}(1+x)^\delta\right)\left((-1)^\ell\binom{\delta}{\ell}(1+x)^{\delta-\ell}\right),
\end{aligned}$$

which gives 0 for all $0 < i < \delta$. Therefore, $\bar{p}(t) - \widehat{p}(t;\tau) = -R_T^{\delta,\delta} = \mathcal{O}(\tau^\delta)$. Similarly, it can be shown that $\dot{\bar{p}}(t) - \dot{\widehat{p}}(t;\tau) = -R_T^{\delta,\delta+1} = \mathcal{O}(\tau^\delta)$ with

$$R_T^{\delta,\delta+1} = \frac{1}{(\delta-1)!}\sum_{\ell=1}^{\delta}(-1)^{(\ell-1)}\binom{\delta}{\ell}\int_{t-\ell\tau}^{t}(t-\ell\tau-\zeta)^{\delta-1}\bar{p}^{(\delta+1)}(\zeta)\,\mathrm{d}\zeta. \qquad \blacksquare$$

Appendix B. G matrices

In this section, we gather expressions of the G matrix for the cases $k=2$ and $k=3$.

B.1. **Case** $k = \delta = 2$

The G matrix is given by

$$G = \begin{bmatrix} \frac{3}{2} & -1 & \frac{1}{4} & 0 \\ -1 & \frac{3}{2} & -1 & \frac{1}{4} \\ \frac{1}{4} & -1 & \frac{3}{2} & -1 \\ 0 & \frac{1}{4} & -1 & \frac{3}{2} \end{bmatrix} + \mu \begin{bmatrix} 0 & \frac{3}{2} & -\frac{11}{4} & \frac{3}{2} \\ \frac{3}{2} & 0 & \frac{3}{2} & -\frac{11}{4} \\ -\frac{11}{4} & \frac{3}{2} & 0 & \frac{3}{2} \\ \frac{3}{2} & -\frac{11}{4} & \frac{3}{2} & 0 \end{bmatrix}$$
$$+ \eta \begin{bmatrix} 0 & -\frac{3}{4} & 0 & 0 \\ -\frac{3}{4} & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} + \eta\mu \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -3 & \frac{11}{4} & -\frac{3}{2} \\ 0 & \frac{11}{4} & -3 & \frac{11}{4} \\ 0 & -\frac{3}{2} & \frac{11}{4} & -3 \end{bmatrix} - LL^\top$$

where

$$L = \begin{bmatrix} \gamma_0 & 0 & 0 & 0 \\ \gamma_1 & \gamma_0 & 0 & 0 \\ \gamma_2 & \gamma_1 & \gamma_0 & 0 \\ \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}.$$

B.2. **Case** $k = \delta = 3$

The G matrix is given by

$$G = \begin{bmatrix} \frac{11}{6} & -\frac{3}{2} & \frac{3}{4} & -\frac{1}{6} & 0 & 0 \\ -\frac{3}{2} & \frac{11}{6} & -\frac{3}{2} & \frac{3}{4} & -\frac{1}{6} & 0 \\ \frac{3}{4} & -\frac{3}{2} & \frac{11}{6} & -\frac{3}{2} & \frac{3}{4} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{3}{4} & -\frac{3}{2} & \frac{11}{6} & -\frac{3}{2} & \frac{3}{4} \\ 0 & -\frac{1}{6} & \frac{3}{4} & -\frac{3}{2} & \frac{11}{6} & -\frac{3}{2} \\ 0 & 0 & -\frac{1}{6} & \frac{3}{4} & -\frac{3}{2} & \frac{11}{6} \end{bmatrix} + \mu \begin{bmatrix} 0 & \frac{11}{4} & -\frac{29}{4} & \frac{23}{3} & -\frac{17}{4} & \frac{5}{4} \\ \frac{11}{4} & 0 & \frac{11}{4} & -\frac{29}{4} & \frac{23}{3} & -\frac{17}{4} \\ -\frac{29}{4} & \frac{11}{4} & 0 & \frac{11}{4} & -\frac{29}{4} & \frac{23}{3} \\ \frac{23}{3} & -\frac{29}{4} & \frac{11}{4} & 0 & \frac{11}{4} & -\frac{29}{4} \\ -\frac{17}{4} & \frac{23}{3} & -\frac{29}{4} & \frac{11}{4} & 0 & \frac{11}{4} \\ \frac{5}{4} & -\frac{17}{4} & \frac{23}{3} & -\frac{29}{4} & \frac{11}{4} & 0 \end{bmatrix}$$
$$+ \eta \begin{bmatrix} 0 & -\frac{11}{12} & 0 & 0 & 0 & 0 \\ -\frac{11}{12} & 3 & -\frac{5}{3} & \frac{1}{6} & 0 & 0 \\ 0 & -\frac{5}{3} & 3 & -\frac{5}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & -\frac{5}{3} & 3 & -\frac{5}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & -\frac{5}{3} & 3 & -\frac{5}{3} \\ 0 & 0 & 0 & \frac{1}{6} & -\frac{5}{3} & 3 \end{bmatrix} + \eta\mu \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{11}{2} & \frac{29}{4} & -\frac{23}{3} & \frac{17}{4} & -\frac{5}{4} \\ 0 & \frac{29}{4} & -\frac{11}{2} & \frac{29}{4} & -\frac{23}{3} & \frac{17}{4} \\ 0 & -\frac{23}{3} & \frac{29}{4} & -\frac{11}{2} & \frac{29}{4} & -\frac{23}{3} \\ 0 & \frac{17}{4} & -\frac{23}{3} & \frac{29}{4} & -\frac{11}{2} & \frac{29}{4} \\ 0 & -\frac{5}{4} & \frac{17}{4} & -\frac{23}{3} & \frac{29}{4} & -\frac{11}{2} \end{bmatrix} - LL^\top \tag{B.1}$$

where

$$L = \begin{bmatrix} \gamma_0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_1 & \gamma_0 & 0 & 0 & 0 & 0 \\ \gamma_2 & \gamma_1 & \gamma_0 & 0 & 0 & 0 \\ \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 & 0 & 0 \\ \gamma_4 & \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 & 0 \\ \gamma_5 & \gamma_4 & \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}.$$

References

[1] G. Akrivis, M. Crouzeix, and C. Makridakis. Implicit-Explicit Multistep Finite Element Methods for Nonlinear Parabolic Problems. *Math. Comput.*, 67(222):457–477, 1998.

[2] R. Altmann and M. Deiml. A novel iterative time integration scheme for linear poroelasticity. *Electron. Trans. Numer. Anal.*, 60:256–275, 2024.

[3] R. Altmann and M. Deiml. A second-order iterative time integration scheme for linear poroelasticity. *SIAM J. Sci. Comput.*, 47(4):B875–B898, 2025.

[4] R. Altmann and R. Maier. A decoupling and linearizing discretization for poroelasticity with nonlinear permeability. *SIAM J. Sci. Comput.*, 44(3):B457–B478, 2022.

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

[6] R. Altmann, R. Maier, and B. Unger. Semi-explicit integration of second order for weakly coupled poroelasticity. *BIT Numer. Math.*, 64: article no. 20 (27 pages), 2024.

[7] R. Altmann, V. Mehrmann, and B. Unger. Port-Hamiltonian formulations of poroelastic network models. *Math. Comput. Model. Dyn. Sys.*, 27(1):429–452, 2021.

[8] R. Altmann, A. Mujahid, and B. Unger. Higher-order iterative decoupling for poroelasticity. *Adv. Comput. Math.*, 50(6): article no. 111 (22 pages), 2024.

[9] R. Bellman. On the computational solution of differential-difference equations. *J. Math. Anal. Appl.*, 2:108–110, 1961.

[10] R. Bellman and K. Cooke. *Differential-difference equations.* Academic Press Inc., 1963.

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

[12] J. W. Both, M. Borregales, J. M. Nordbotten, K. Kumar, and F. A. Radu. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Appl. Math. Lett.*, 68:101–108, 2017.

[13] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, 2011.

[14] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numer. Math.*, 35(3):257–276, 1980.

[15] G. Dahlquist. G-stability is equivalent to A-stability. *BIT Numer. Math.*, 18(4):384–401, 1978.

[16] E. Detournay and A. H. D. Cheng. Fundamentals of poroelasticity. In *Analysis and Design Methods*, pages 113–171. Elsevier, 1993.

[17] E. Eliseussen, M. E. Rognes, and T. B. Thompson. A posteriori error estimation and adaptivity for multiple-network poroelasticity. *ESAIM, Math. Model. Numer. Anal.*, 57(4):1921–1952, 2023.

[18] 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.

[19] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer, 2nd rev. ed edition, 2009.

[20] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems.* Springer, 1996.

[21] J. Kim, H. A. Tchelepi, and R. Juanes. Stability and convergence of sequential methods for coupled flow and geomechanics: Drained and undrained splits. *Comput. Meth. Appl. Mech. Eng.*, 200(23):2094–2116, 2011.

[22] 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. Meth. Appl. Mech. Eng.*, 200(13):1591–1606, 2011.

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

[24] A. Mujahid. Monolithic, non-iterative and iterative time discretization methods for linear coupled elliptic-parabolic systems. *GAMM Archive for Students*, 4(1), 2022.

[25] O. Nevanlinna and F. Odeh. Multiplier techniques for linear multistep methods. *Numer. Funct. Anal. Optim.*, 3(4):377–423, 1981.

[26] R. E. Showalter. Diffusion in poro-elastic media. *J. Math. Anal. Appl.*, 251(1):310–340, 2000.

[27] E. Storvik, J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu. On the optimization of the fixed-stress splitting for Biot's equations. *Int. J. Numer. Methods Eng.*, 120(2):179–194, 2019.

[28] S. Trenn and B. Unger. Delay regularity of differential-algebraic equations. In *Proc. 58th IEEE Conf. Decision Control (CDC) 2019, Nice, France*, pages 989–994. IEEE Press, 2019.

[29] B. Unger. Discontinuity propagation in delay differential-algebraic equations. *Electron. J. Linear Algebra*, 34:582–601, 2018.

[30] B. Unger. *Well-Posedness and Realization Theory for Delay Differential-Algebraic Equations*. Dissertation, Technische Universität Berlin, 2020.

[31] J. C. Vardakis, D. Chou, B. J. Tully, C. C. Hung, T. H. Lee, P. H. Tsui, and Y. Ventikos. Investigating cerebral oedema using poroelasticity. *Med. Eng. Phys.*, 38(1):48–57, 2016.

[32] E. Zeidler. *Nonlinear functional analysis and its applications. 2A: Linear monotone operators*. Springer, 1990.

[33] M. D. Zoback. *Reservoir geomechanics*. Cambridge University Press, 2010.