Journal of Computational Physics 376 (2019) 295–321

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

# Unconditional stability for multistep ImEx schemes: Practice

Benjamin Seibold [a], David Shirokoff [b,*], Dong Zhou [c]

[a] Department of Mathematics, Temple University, 1801 N. Broad Street, Philadelphia, PA 19122, USA
[b] Department of Mathematical Sciences, New Jersey Institute of Technology, University Heights Newark, NJ 07102, USA
[c] Department of Mathematics, California State University, Los Angeles, 5151 State University Dr. Los Angeles, CA 90032, USA

ARTICLE INFO

*Article history:*
Received 18 April 2018
Received in revised form 15 August 2018
Accepted 24 September 2018
Available online 27 September 2018

*Keywords:*
Linear multistep ImEx
Unconditional stability
ImEx stability
High order time stepping
Semi-implicit backward differentiation

ABSTRACT

This paper focuses on the question of how unconditional stability can be achieved via multistep ImEx schemes, in practice problems where both the implicit and explicit terms are allowed to be stiff. For a class of new ImEx multistep schemes that involve a free parameter, strategies are presented on how to choose the ImEx splitting and the time stepping parameter, so that unconditional stability is achieved under the smallest approximation errors. These strategies are based on recently developed stability concepts, which also provide novel insights into the limitations of existing semi-implicit backward differentiation formulas (SBDF). For instance, the new strategies enable higher order time stepping that is not otherwise possible with SBDF. With specific applications in nonlinear diffusion problems and incompressible channel flows, it is demonstrated how the unconditional stability property can be leveraged to efficiently solve stiff nonlinear or nonlocal problems without the need to solve nonlinear or nonlocal problems implicitly.



## 1. Introduction

This paper builds on the theoretical work [1] on the unconditional stability of linear multistep methods (LMMs). While [1] introduced a new unconditional stability theory for implicit–explicit (ImEx) methods, and presented a novel class of ImEx LMMs that involve a stability parameter, this paper develops strategies on how to select the time stepping parameter and the ImEx splitting in an optimal fashion. The key focus is on problems for which an ImEx splitting is warranted in which both the implicit and the explicit terms are stiff, for example because the stiff terms are difficult to treat implicitly.

Conventional ImEx splittings often treat all stiff terms implicitly to ensure that one does not encounter a stiff time step restriction (one usually accepts a time step restriction from the non-stiff explicit part). However, as demonstrated in [1], this is not always required: one may treat stiff terms explicitly and nevertheless avoid a stiff time step restriction, provided the implicit term and the scheme are properly chosen. This paper provides strategies on how to make these choices (splitting and scheme) in practical problems. We do so through the use of the unconditional stability theory from [1], which is based on geometric diagrams that play a role analogous to the absolute stability diagram in conventional ordinary differential equation (ODE) stability theory. Specifically, we present strategies on how to achieve unconditional stability via (i) choosing the splitting for a given scheme; (ii) modifying a time-stepping scheme for a given splitting; and (iii) designing the splitting and the scheme in a coupled fashion. In addition, we employ the stability theory to provide new insights on the limitations

* Corresponding author.
*E-mail addresses:* seibold@temple.edu (B. Seibold), david.g.shirokoff@njit.edu (D. Shirokoff), dong.zhou@calstatela.edu (D. Zhou).

https://doi.org/10.1016/j.jcp.2018.09.044
0021-9991/

of popular semi-implicit backward differentiation formulas (SBDF). In fact, we show that the new ImEx LMMs generalize SBDF methods, in a way that they overcome some of their fundamental stability limitations.

### 1.1. Problem setting

We are concerned with the time-evolution of linear ODEs of the form

$$\boldsymbol{u}_t = \boldsymbol{L}\boldsymbol{u} + \boldsymbol{f}, \quad \boldsymbol{u}(0) = \boldsymbol{u}_0. \tag{1.1}$$

Here $\boldsymbol{u}(t) \in \mathbb{R}^N$, $\boldsymbol{L} \in \mathbb{R}^{N\times N}$, and $\boldsymbol{f}(t) \in \mathbb{R}^N$ is an external forcing. We assume that $\boldsymbol{L}$ gives rise to asymptotically stable solutions — i.e. solutions to the homogeneous ODE $\boldsymbol{u}_t = \boldsymbol{L}\boldsymbol{u}$ decay in time (the eigenvalues of $\boldsymbol{L}$ are in the strict left-half-plane). This assumption can be relaxed; however then additional caveats are required (see §6 for when $\boldsymbol{L}$ has a zero eigenvalue, or §8 for when $\boldsymbol{L}$ has purely imaginary eigenvalues).

For the right hand side $\boldsymbol{L}$, an *ImEx splitting* $(\boldsymbol{A}, \boldsymbol{B})$ is conducted [2–5], i.e. $\boldsymbol{L}$ is split into two parts, $\boldsymbol{L} = \boldsymbol{A} + \boldsymbol{B}$, where $\boldsymbol{A}$ is treated implicitly (Im) and $\boldsymbol{B}$ is treated explicitly (Ex). Clearly, the splitting $(\boldsymbol{A}, \boldsymbol{B})$ is non-unique, and in fact, any matrix $\boldsymbol{A}$ defines a splitting by choosing $\boldsymbol{B} := \boldsymbol{L} - \boldsymbol{A}$. For this ImEx splitting, we now require the time-stepping scheme to be unconditionally stable. This is a stringent, but very practical property (especially when $\boldsymbol{L}$ is stiff) as it allows one to choose a time step as large as accuracy requirements permit.

Note that the theory in this paper is developed for linear ODEs, as this assumption allows for a rigorous geometric stability theory involving unconditional stability diagrams. However, we then extend the results, in an ad-hoc but rather natural fashion, to nonlinear problems as well.

### 1.2. Examples from partial differential equations

A crucial source of stiff problems is the method-of-lines (MOL) semi-discretization of a partial differential equation (PDE). In that situation, rather than having one single right hand side $\boldsymbol{L}$, one faces a family $\boldsymbol{L}_h$ (with $h$ the mesh size) that approximates a spatial differential operator $\mathcal{L}$. A key property of the time-stepping strategies studied here is that for many PDE problems, the choice of ImEx splitting and scheme can in fact be conduced on the level of differential operators, or equivalently, to hold for the family $\boldsymbol{L}_h$, uniformly in $h$ (see Sections 6 and 7).

An important PDE situation in which unconditional stability is important is the MOL discretization of diffusion. A fully explicit treatment of diffusion gives rise to a stiff time step restriction $k \leq Ch^2$. Hence, for problems in which diffusion represents the highest spatial derivative, a common approach is to include all of the discretization of $\frac{\partial^2}{\partial x^2}$ into the implicit part $\boldsymbol{A}_h$, and leave $\boldsymbol{B}_h$ as the remaining non-stiff terms. Such an approach will then avoid a stiff time step restriction. However, treating all stiff terms of $\boldsymbol{L}$ implicitly may in general be costly (see §6); and in fact it is not always necessary. Having new approaches that allow one to treat (some of the) stiff terms explicitly, without incurring a stiff time step restriction, can be a significant practical benefit. In problems where $\boldsymbol{L}$ is stiff and costly to treat fully implicitly, this opens the door for designing a well-chosen ImEx splitting where $\boldsymbol{A}$ contains only part of the stiff components of $\boldsymbol{L}$, and is much more efficient to treat implicitly.

### 1.3. Background and relation to other works

ImEx unconditional stability has been studied in numerous theoretical and practical works. On the theoretical side, general abstract sufficient conditions for unconditional stability and arbitrary multistep schemes are stated in [6–9]. Although these conditions have the advantage of incorporating nonlinear terms (i.e. $\boldsymbol{B}$ is allowed to be a nonlinear operator), they have the drawback that they require the implicit matrix $\boldsymbol{A}$ be larger (in the sense of an appropriate norm) than $\boldsymbol{B}$, and are overly restrictive for the problems we consider (e.g. they do not apply to Example 2).

Generally speaking, in the context of multistep methods, proofs for unconditional stability are commonplace for first and second order methods. Meanwhile, for higher order schemes, unconditional stability is usually only studied numerically, and in limited settings. This gap is likely due to the limitations that existing high-order methods encounter (see §5). Important works in which unconditional stability is proved for first or second order methods, or numerically observed in higher order schemes, are the following papers (and references therein). Some of the first applications involving unconditional stability originated in the 1970s, with alternating direction implicit (ADI) methods [10]. Others include magneto-hydrodynamics [11]; unconditional stability (also referred to as unconditionally energy stable, or as convex-concave splitting methods) for phase-field models [12–19,14,20]; applications to fluid-interface problems [21]; incompressible Navier–Stokes equations [22–27]; Stokes–Darcy systems [28], compressible Navier–Stokes equations [29,30], and PDEs with the explicit treatment of non-local terms [31,32]. One disadvantage of low (i.e. first or second) order methods is that they can also have large error constants for dissipative PDEs [33] and dispersive PDEs [34], thus further reducing their applicability for the long-time numerical simulations. We differ from these previous works in several ways:

1. We include higher order schemes as part of the study.

2. Whereas many existing works use von-Neumann analysis or energy estimates that are tailored to a specific problem, we make use of recently introduced unconditional stability diagrams [1]. The diagram approach simplifies the design of high-order unconditionally stable schemes and is applicable to a wider range of applications.
3. We include variable ImEx time stepping coefficients. It may be surprising that stability considerations for ImEx schemes do not require all stiff terms to be included in $\boldsymbol{A}$. In fact, ImEx schemes can even go far beyond such a restriction: not only can $\boldsymbol{B}$ be stiff, it can (in some sense) even be *larger* than $\boldsymbol{A}$, while still retaining unconditional stability. The underlying mechanism is that $\boldsymbol{A}$ is chosen in a way that stabilizes the numerical instabilities created by the explicit treatment of $\boldsymbol{B}$ with a suitable (simultaneous) choice of a splitting and time stepping scheme.

It should also be stressed that there are numerous time stepping approaches (not strictly multistep methods) for specific application areas that possess good stability properties. Recently, high order unconditional stable methods for ADI applications have been obtained by combining second order multistep schemes with Richardson extrapolation [35–37]. For PDE systems that have a gradient flow structure, new conditions [38] allow for the design of third order, unconditionally energy-stable Runge–Kutta (RK) methods. Other techniques include: semi-implicit deferred correction methods [39]; semi-implicit matrix exponential schemes where the linear terms are treated with an integrating factor [40–42]; and explicit RK schemes with very large stability regions for parabolic problems [43].

### 1.4. Outline of this paper

This paper is organized as follows. After introducing the key notation and definitions (§2), a self-contained review of the employed unconditional stability theory is provided that takes a different viewpoint than [1] by placing a practical emphasis on the eigenvalues of $\boldsymbol{A}$, $\boldsymbol{B}$. Section 4 and onward (including Appendix A) contain new results. Section 4 provides recipes for designing (optimal) unconditionally stable ImEx schemes that minimize the numerical error. Section 5 characterizes the limitations of the well-known SBDF methods. Section 6 uses insight from §5 to overcome the limitations of SBDF and devise optimal high order (i.e. beyond 2nd order) unconditionally stable schemes for the variable-coefficient and non-linear diffusion problems. This section includes new formulas for ImEx splittings and schemes (accompanied by rigorous proofs in Appendix A); as well as computational examples. Section 7 studies an application example that is motivated by incompressible Navier–Stokes flow in a channel and provides general insight into stability issues in computational fluid dynamics. Section 8 provides an outlook and conclusions, and Appendix B lists the specific ImEx coefficients to be used in practice.

## 2. Introduction to the ImEx schemes and unconditional stability property

This section introduces the assumptions, notations, and ImEx schemes used throughout the paper. As discussed above, we are interested in unconditional stability for ImEx splittings $\boldsymbol{L} = \boldsymbol{A} + \boldsymbol{B}$ of equation (1.1) where in general both the implicit matrix $\boldsymbol{A}$, and the explicit matrix $\boldsymbol{B}$ are allowed to be stiff.

We restrict to splittings in which $\boldsymbol{A}$ is Hermitian (symmetric in the real case) negative definite, i.e. $\boldsymbol{A}$ has strictly negative eigenvalues:

$$\boldsymbol{A}^\dagger = \boldsymbol{A}, \quad \text{and} \quad \langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{u} \rangle < 0, \quad \text{for all } \boldsymbol{u} \neq 0,\ \boldsymbol{u} \in \mathbb{C}^N. \tag{2.1}$$

Here we have adopted the standard notation on vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$ (or $\mathbb{R}^N$):

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{j=1}^{N} \overline{x}_j y_j, \quad \|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle, \quad \boldsymbol{A}^\dagger = \overline{\boldsymbol{A}}^T, \quad \boldsymbol{x} = \left(x_1, x_2, \cdots, x_N\right)^T.$$

Note that $\boldsymbol{L}$ itself is not assumed symmetric/Hermitian or negative definite. Furthermore, assumption (2.1) on $\boldsymbol{A}$ is not overly restrictive, because for any given $\boldsymbol{L}$ one can choose $\boldsymbol{A}$ symmetric negative definite, and then set $\boldsymbol{B} = \boldsymbol{L} - \boldsymbol{A}$. Note that spectral methods (for the spatial discretization of PDEs) may give rise to a complex matrix $\boldsymbol{A}$, which is why we do not restrict $\boldsymbol{A}$ to be real. It is also worth noting that much of the theory we present still persists even when $\boldsymbol{A}$ is not Hermitian and negative definite (see Section 8).

Finally, we remark that the implicit treatment of a matrix $\boldsymbol{A}$ in multistep methods (or even Runge–Kutta methods), requires one to solve linear systems with coefficient matrices of the form $(\boldsymbol{I} - \gamma k \boldsymbol{A})$, where $\gamma > 0$ is a constant and $k > 0$ is the time step. For $\boldsymbol{A}$ symmetric negative definite, those system matrices are positive definite and thus favorable for fast solvers (chapter IV, lecture 38, [44]).

We will generally assume that the problem gives rise to a preferred/natural matrix structure $\boldsymbol{A}_0$ (symmetric, negative definite) that one wishes to treat implicitly; however, its overall magnitude is up to choice. In other words, the user fixes $\boldsymbol{A}_0$ and would accept any implicit matrix of the form $\boldsymbol{A} = \sigma \boldsymbol{A}_0$ (with the *splitting parameter* $\sigma > 0$), provided that such an $\boldsymbol{A}$ yields unconditional stability. This is in a spirit similar to [10]. For example, in spatial discretizations of a variable coefficient diffusion PDE where $\boldsymbol{L}u \approx (d(x)u_x)_x$, the user may prefer an implicit treatment of the constant coefficient Laplacian $\boldsymbol{A}_0 u \approx u_{xx}$, however, would accept any constant multiple as well, i.e. $\boldsymbol{A}u \approx \sigma u_{xx}$. Writing $\boldsymbol{A} = \sigma \boldsymbol{A}_0$ where $\boldsymbol{A}_0$ is fixed, introduces

the scalar $\sigma$ as a key parameter. This paper shows how to choose $\sigma$ in a systematic fashion to obtain unconditionally stability.

We restrict our attention to ImEx versions of linear multistep methods (LMMs) [2,3]; however it is worth noting that some of the concepts developed here may extend to other time stepping schemes as well, such as Runge–Kutta (multi-stage) ImEx schemes. The general form of an $r$-step LMM applied to the ODE (1.1) with a splitting $(\boldsymbol{A}, \boldsymbol{B})$ is:

$$\frac{1}{k}\sum_{j=0}^{r} a_j \boldsymbol{u}_{n+j} = \sum_{j=0}^{r} \Big( c_j \boldsymbol{A}\boldsymbol{u}_{n+j} + b_j \boldsymbol{B}\boldsymbol{u}_{n+j} + b_j \boldsymbol{f}_{n+j} \Big). \tag{2.2}$$

Here $k > 0$ is the time step, the variable $b_r = 0$ (so that $\boldsymbol{B}$ is explicit in (2.2)), $\boldsymbol{u}_n = \boldsymbol{u}(nk)$ is the numerical solution $\boldsymbol{u}(t)$ (with a slight abuse of notation) evaluated at the $n$-th time step, and $\boldsymbol{f}_n = \boldsymbol{f}(nk)$. We refer to the values $(a_j, b_j, c_j)$, with $0 \le j \le r$ as the ImEx (time stepping) coefficients. The LMMs of the form (2.2) require $r$ initial conditions $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r-1}$. The computation of these initial conditions to sufficient accuracy is a separate problem (chapter 5.9.3, [45]), and is not considered here. When discussing stability it will be useful to define the polynomials $a(z)$, $b(z)$, $c(z)$, using the ImEx coefficients in (2.2):

$$a(z) = \sum_{j=0}^{r} a_j z^j, \quad b(z) = \sum_{j=0}^{r-1} b_j z^j, \quad c(z) = \sum_{j=0}^{r} c_j z^j. \tag{2.3}$$

Given the ImEx coefficients, one may write down the polynomials $a(z)$, $b(z)$, $c(z)$, or alternatively, given polynomials $a(z)$, $b(z)$, $c(z)$, one may read off the different coefficients in front of $z^j$ to obtain the time stepping coefficients $(a_j, b_j, c_j)$.

In this work we utilize a one-parameter family of ImEx coefficients, introduced in [1], that have desirable unconditional stability properties. The new ImEx coefficients are characterized by a parameter $0 < \delta \le 1$, i.e. they are functions of a single *ImEx parameter* $\delta$, and are defined for orders $r = 1$ through $r = 5$. Formulas for the new coefficients $(a_j, b_j, c_j)$, in terms of $\delta$, may be found in Table B.4; and substituting different values of $0 < \delta \le 1$ into these formulas yields different ImEx schemes. For example, the new one-parameter ImEx schemes for first $(r = 1)$ and second order $(r = 2)$ take the form:

$$\text{1st order:} \quad \frac{1}{k}\big(\delta\, \boldsymbol{u}_{n+1} - \delta\, \boldsymbol{u}_n\big) = \boldsymbol{A}\boldsymbol{u}_{n+1} + (\delta - 1)\boldsymbol{A}\boldsymbol{u}_n + \delta\, \boldsymbol{B}\boldsymbol{u}_n, \tag{2.4}$$

$$\text{2nd order:} \quad \frac{1}{k}\Big(\big(2\delta - \frac{1}{2}\delta^2\big)\boldsymbol{u}_{n+2} + \big(-4\delta + 2\delta^2\big)\boldsymbol{u}_{n+1} + \big(2\delta - \frac{3}{2}\delta^2\big)\boldsymbol{u}_n\Big) = \tag{2.5}$$

$$\boldsymbol{A}\boldsymbol{u}_{n+2} + 2\big(\delta - 1\big)\boldsymbol{A}\boldsymbol{u}_{n+1} + \big(\delta - 1\big)^2\boldsymbol{A}\boldsymbol{u}_n + 2\delta\, \boldsymbol{B}\boldsymbol{u}_{n+1} + \big((\delta - 1)^2 - 1\big)\boldsymbol{B}\boldsymbol{u}_n.$$

For brevity we have set $\boldsymbol{f} = 0$ in the formulas (2.4)–(2.5), however one may include it in the explicit term $\boldsymbol{B}\boldsymbol{u}$ (or even the implicit term) as in equation (2.2). Although the formulas for the coefficients might appear unruly, they have simple polynomial expressions.

**Remark 1.** (ImEx coefficients from Table B.4 written in polynomial form) For orders $1 \le r \le 5$, and $0 < \delta \le 1$, the ImEx coefficients $(a_j, b_j, c_j)$, for $0 \le j \le r$ from Table B.4 correspond to the following polynomials:

$$a(z) = \sum_{j=1}^{r} \frac{f^{(j)}(1)}{j!}(z-1)^j, \quad \text{where} \quad f(z) = (\ln z)(z - 1 + \delta)^r, \tag{2.6}$$

$$b(z) = (z - 1 + \delta)^r - (z - 1)^r, \qquad c(z) = (z - 1 + \delta)^r. \tag{2.7}$$

The relationships between the polynomials, i.e. $b(z) = c(z) - (z-1)^r$, and $a(z)$ as the $r$-th order Taylor polynomial of $\ln(z)c(z)$ ensure that the ImEx coefficients satisfy the order conditions required to define an $r$-th order scheme.

Note that in the Remark 1, the polynomial $c(z)$ has roots that approach 1 as $\delta \to 0$. This is not an accident, and it is this property that will eventually lead to good unconditional stability properties for the new schemes.

Equations (2.4)–(2.5), as well as the 3rd, 4th, 5th order schemes in Table B.4, define families of time-stepping schemes. When the value $\delta = 1$ is substituted into the coefficient formulas in equations (2.4)–(2.5), one obtains the well-known backward differentiation formulas for the coefficients of $\boldsymbol{A}$, also referred to as semi-implicit backward differentiation formulas (SBDFr, where $r$ denotes the order of the scheme):

$$\text{SBDF1 } (\delta = 1): \quad \frac{1}{k}\big(\boldsymbol{u}_{n+1} - \boldsymbol{u}_n\big) = \boldsymbol{A}\boldsymbol{u}_{n+1} + \boldsymbol{B}\boldsymbol{u}_n,$$

$$\text{SBDF2 } (\delta = 1): \quad \frac{1}{k}\Big(\frac{3}{2}\boldsymbol{u}_{n+2} - 2\boldsymbol{u}_{n+1} + \frac{1}{2}\boldsymbol{u}_n\Big) = \boldsymbol{A}\boldsymbol{u}_{n+2} + 2\boldsymbol{B}\boldsymbol{u}_{n+1} - \boldsymbol{B}\boldsymbol{u}_n.$$

Choosing values $\delta \neq 1$ yields different (new) schemes. We have only displayed orders $r = 1, 2$ in the above expressions, however coefficients are also given for orders $r = 3, 4, 5$ in Table B.4. Lastly we note that the new ImEx schemes are zero-stable for any value $0 < \delta \leq 1$, and the coefficients satisfy the order conditions [1] to guarantee that they define an $r$-th order scheme (i.e. solving (2.2) using the coefficients approximates the solution to (1.1) with an error that scales like $\mathcal{O}(k^r)$ as $k \to 0$).

Each fixed set of ImEx coefficients, such as SBDF ($\delta = 1$), or ImEx versions of Crank–Nicolson, or even schemes not considered in this paper, provide unconditional stability for only a certain set of matrix splittings $(\boldsymbol{A}, \boldsymbol{B})$ — and these may not include a practitioner's desired splitting for a given problem. Introducing the one-parameter family of ImEx schemes (parameterized by $\delta$) provides the flexibility needed to attain unconditional stability for new classes of matrices $(\boldsymbol{A}, \boldsymbol{B})$ beyond the capabilities of what is possible using a fixed set of coefficients. This point becomes particularly apparent in §5, in the discussion of the limitations of SBDF methods. This gain in unconditional stability offered by the parameter $\delta$ may come with a trade-off of increasing the numerical approximation error constants. Thus, an important discussion (see §4) is how to choose an ImEx scheme (i.e. how to choose $\delta$) for a given problem splitting (i.e. $(\boldsymbol{A}, \boldsymbol{B})$) to balance the trade off of gaining unconditional stability while minimizing the numerical error. Or, even better, how to choose the splitting and scheme in a coupled fashion.

Our goal is to avoid unnecessarily small time step restrictions in the numerical scheme (2.2). To do this we examine when (2.2) is *unconditionally stable* — i.e. the numerical scheme (2.2) with $\boldsymbol{f} = 0$ remains stable regardless of how large one chooses the time step $k > 0$. Formally, we adopt the following definition:

**Definition 2.1.** (Unconditional stability) A scheme (2.2) is unconditionally stable if: when $\boldsymbol{f} = 0$, there exists a constant $C$ such that

$$\|\boldsymbol{u}_n\| \leq C \max_{0 \leq j \leq r-1} \|\boldsymbol{u}_j\|, \quad \text{for all } n \geq r,\ k > 0 \quad \text{and } \boldsymbol{u}_j \in \mathbb{R}^N, \ \text{ where } 0 \leq j \leq r-1.$$

Note that $C$ may depend on the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and the coefficients $(a_j, b_j, c_j)$, but is independent of the time step $k$, the time index $n$, and the initial vectors $\boldsymbol{u}_j$, $0 \leq j \leq r-1$.

It is important to note that unconditional stability of an ImEx LMM like (2.2) can be difficult to determine in practice, as this question depends simultaneously on the choice of coefficients $(a_j, b_j, c_j)$ and the splitting $(\boldsymbol{A}, \boldsymbol{B})$. The purpose of introducing a new stability theory in [1] was to remedy this difficulty and formulate unconditional stability (or failure thereof) in terms of two separate computable quantities: one quantity that depends only on the coefficients $(a_j, b_j, c_j)$, and one that depends only on the splitting $(\boldsymbol{A}, \boldsymbol{B})$. The theory then allows for a variety of possibilities:

(i) Given a fixed splitting $(\boldsymbol{A}, \boldsymbol{B})$, design coefficients $(a_j, b_j, c_j)$ (by choosing $0 < \delta \leq 1$) that achieve unconditional stability — see §4, Recipe 1.
(ii) Given a fixed set of coefficients $(a_j, b_j, c_j)$ (such as SBDF when $\delta = 1$), determine how to choose a splitting $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (i.e. choose $\sigma > 0$) that guarantees unconditional stability — see §4, Recipe 2.
(iii) Offer the most flexibility by simultaneously choosing both the coefficients $(a_j, b_j, c_j)$ and the splitting $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ to achieve unconditional stability. This will involve the simultaneous choice of $(\sigma, \delta)$ and is discussed in §4, Recipe 3.

## 3. The unconditional stability theory

In this section we review the unconditional stability theory from [1] — which imposes conditions on $(\boldsymbol{A}, \boldsymbol{B})$ and the time-stepping coefficients $(a_j, b_j, c_j)$ that (when satisfied) ensure the unconditional stability of (2.2). The stability theory will then provide a guide for choosing the ImEx coefficients $(a_j, b_j, c_j)$ and/or splitting $(\boldsymbol{A}, \boldsymbol{B})$ that guarantee unconditional stability for a given problem (i.e. $\boldsymbol{L}$). The unconditional stability theory is somewhat analogous to the classical absolute stability theory (chapter 7, [45]), as it relies on a stability diagram — and we highlight the parallels with an example here:

**Example 1.** (Absolute stability theory) Given an ODE of the form $\boldsymbol{u}_t = \boldsymbol{A}\boldsymbol{u}$, the absolute stability diagram $\mathcal{A}$ is defined as

$$a_r \boldsymbol{u}_{n+r} + \ldots + a_0 \boldsymbol{u}_n = k\,(c_r \boldsymbol{A}\boldsymbol{u}_{n+r} + \ldots + c_0 \boldsymbol{A}\boldsymbol{u}_n), \tag{3.1}$$

$$\mathcal{A} = \left\{ \mu \in \mathbb{C} : a(z) = \mu c(z), \text{ has stable solutions z} \right\}.$$

The scheme (3.1) is stable with time step $k$, if and only if every eigenvalue $\lambda$ of $\boldsymbol{A}$ (i.e. $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$) satisfies $k\lambda \in \mathcal{A}$ (with the possible exception of repeated eigenvalues $\lambda$, and time steps $k$ that lie on the boundary $k\lambda \in \partial\mathcal{A}$).

A key feature of the absolute stability theory is that it *decouples* the stability criteria into (i) a property of the matrix $\boldsymbol{A}$ only (i.e. the eigenvalues), in relation to (ii) a property of the time stepping scheme only (i.e. $\mathcal{A}$). Decoupling the stability theory is extremely useful; for instance, it allows one to determine which matrices $\boldsymbol{A}$ can be solved using a given time stepping scheme. The unconditional stability theory in this section will parallel that of the absolute stability theory, and:

- Introduce the unconditional stability diagram (defined solely by $(a_j, b_j, c_j)$); and provide formulas for the diagrams to the schemes corresponding to Table B.4.
- Provide computable quantities in terms of $(\boldsymbol{A}, \boldsymbol{B})$ that are analogous to the eigenvalues of $\boldsymbol{A}$ in Example 1. Unconditional stability will then be framed in terms of the computable quantities lying inside the unconditional stability region.

### 3.1. The unconditional stability diagram $\mathcal{D}$

The absolute stability theory in Example 1 was obtained by replacing the matrix $\boldsymbol{A}$ with one of its eigenvalues $\lambda$ — resulting in a (simpler) stability analysis of a scalar ODE. In a similar spirit, if $(\boldsymbol{A}, \boldsymbol{B})$ can be simultaneously diagonalized (for instance when they are commuting and diagonalizable matrices), then $(\boldsymbol{A}, \boldsymbol{B})$ may be replaced by their eigenvalues — resulting likewise in a scalar ODE. The unconditional stability diagram can then be derived from this scalar ODE. We stress that although the diagram is derived here assuming $(\boldsymbol{A}, \boldsymbol{B})$ are simultaneously diagonalizable, the diagram is also applicable to general matrices $(\boldsymbol{A}, \boldsymbol{B})$ (i.e. that do not commute), as outlined below. Suppose $\boldsymbol{v}$ is a simultaneous eigenvector to $\boldsymbol{A}$ and $\boldsymbol{B}$ and satisfies

$$-\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \quad \boldsymbol{B}\boldsymbol{v} = \gamma \boldsymbol{v}, \quad -\mu \boldsymbol{A}\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}, \quad \text{where } \mu = \frac{\gamma}{\lambda}. \tag{3.2}$$

Here $\lambda > 0$ (and real) since $\boldsymbol{A}$ is symmetric/Hermitian and positive definite. Substituting $\boldsymbol{u}(t) = v(t)\,\boldsymbol{v}$ into the ODE (1.1) yields the scalar equation

$$v_t = -\lambda v + \gamma v. \tag{3.3}$$

One can then examine stability for the ImEx scheme (2.2), applied to equation (3.3) (with the $\lambda$ term treated implicitly and the $\gamma$ term explicitly), in the usual way: set $v_n = z^n v_0$, to obtain a polynomial equation for the growth factors $z$

$$k^{-1}a(z) = -\lambda c(z) + \gamma b(z). \tag{3.4}$$

Here $a(z), b(z), c(z)$ are the polynomials defined in (2.3). Note that the polynomial equation (3.4) was used in [2] for the purpose of determining CFL-type time step stability restrictions for advection–diffusion problems; and also in [46] in the context of computing absolute stability-type diagrams for ImEx schemes (see also [47] for a treatment of delay differential equations). In both cases, the matrices $(\boldsymbol{A}, \boldsymbol{B})$ were assumed to be simultaneously diagonalizable, and neither study was focused on unconditional stability. Equation (3.4) is also sometimes used as a (non-rigorous) model for stability in the case when $(\boldsymbol{A}, \boldsymbol{B})$ are not simultaneously diagonalizable. The study of unconditional stability digresses from prior work by re-parameterizing equation (3.4) with the substitution $y = -k\lambda$ and $\mu = \gamma\lambda^{-1}$:

$$a(z) = y\Big(c(z) - \mu b(z)\Big). \tag{3.5}$$

Note that $y$ takes on all values $y < 0$ as $k$ varies between 0 and $+\infty$; and that $\mu \in \mathbb{C}$. For a fixed mode, i.e. fixed $\lambda$ and $\gamma$, unconditional stability demands that the growth factors $z$ solving equation (3.4) are stable for all $k > 0$. Viewed in the context of (3.5), this requirement leads to the definition of the unconditional stability diagram $\mathcal{D}$: the values $\mu \in \mathbb{C}$ for which the growth factors $z$ to (3.5) are stable for all $y < 0$ (including $y \to -\infty$)

$$\mathcal{D} := \Big\{\mu \in \mathbb{C} : \text{Solutions } z \text{ to (3.5) are stable for all } y < 0\Big\}.$$

Here we say that $z$ is stable if $|z| < 1$; and for technical convenience we exclude (non-repeated) values of $|z| = 1$. Thus far, the definition for $\mathcal{D}$ is very general and may be computed for any set of ImEx LMM coefficients $(a_j, b_j, c_j)$. It is also crucial to note that $\mathcal{D}$ is defined *only* in terms of the ImEx scheme coefficients.

It was proved (Thm. 8, Prop. 9 [1]) that for the schemes in Table B.4, the value of $y \to -\infty$ (i.e. requiring stability for large time steps, $k \to \infty$) imposes the most severe restriction on the growth factors in equation (3.5). This theoretical result has the consequence that the set $\mathcal{D}$ is completely determined by setting $y \to -\infty$ in (3.5), leading to the simplification

$$\mathcal{D} = \Big\{\mu \in \mathbb{C} : c(z) - \mu b(z) \text{ has stable roots}\Big\} \quad \text{(For schemes in Table B.4)} \tag{3.6}$$

We stress that (3.6) is not necessarily a general property of ImEx LMM — but it holds for the schemes in Table B.4. Equation (3.6) is useful as it allows one to compute $\mathcal{D}$ (Thm. 8 [1]) in terms of a *boundary locus* formulation (chapter 7.6, [45]) with the polynomials $b(z)$ and $c(z)$ introduced in Remark 1:

(B1) The set $\mathcal{D}$ (for orders $1 \le r \le 5$) includes the origin (i.e. $0 \in \mathcal{D}$) and has the boundary

$$\partial\mathcal{D} = \left\{\frac{(z-1+\delta)^r}{(z-1+\delta)^r - (z-1)^r} : |z| = 1,\ \arg z_0 \le \arg z \le 2\pi - \arg z_0\right\}, \tag{3.7}$$

$$\text{with: } z_0 = 1, \text{ for } r = 1, \text{ and } z_0 = \frac{2 - \delta - 2(1-\delta)\cos(\pi/r)e^{\imath\pi/r}}{2 - \delta - 2\cos(\pi/r)e^{\imath\pi/r}}, \text{ for } 2 \le r \le 5.$$

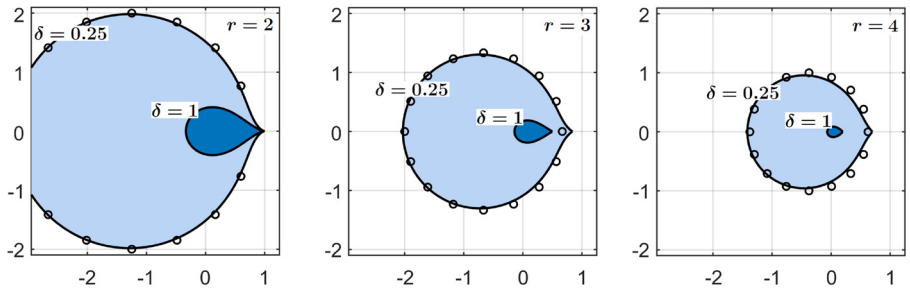


**Fig. 1.** The sets $\mathcal{D}$ for orders (left to right) $r \in \{2, 3, 4\}$. The set $\mathcal{D}$ for $\delta = 1$ (SBDF) (dark blue) is much smaller than $\mathcal{D}$ for $\delta = 0.25$ (light blue). The asymptotic circle $C$ in formula (B3) for $\delta = 0.25$ is shown in dots ($\circ$). The stability regions also decrease in size with increasing $r$. The orders $r = 1, 5$ (not plotted) exhibit a similar behavior. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

(B2) The right-most point $m_r$ and left-most point $m_l$ of $\partial\mathcal{D}$ are on the real axis with:

$$m_l = \left(1 - (1 - \delta/2)^{-r}\right)^{-1}, \quad m_r = \begin{cases} 1, & r = 1, \\ \left(1 + \left((1 - \delta/2)\sec(\pi/r)\right)^{-r}\right)^{-1}, & 2 \le r \le 5. \end{cases}$$

(B3) In the asymptotic limit $\delta \ll 1$, the set $\mathcal{D}$ approaches the circle $C$, where

$$C = \left\{ z \in \mathbb{C} : \left| z + \frac{1}{r\delta} - \frac{r+1}{2r} \right| \le \frac{1}{r\delta} \right\}.$$

Note that $C$ has a center at $\sim -\frac{1}{r\delta}$ and radius $\sim \frac{1}{r\delta}$; and hence becomes arbitrarily large as $\delta \to 0$. Therefore, $\mathcal{D}$ becomes large as $\delta \to 0$.

Fig. 1 plots the stability diagrams $\mathcal{D}$ for different orders and $\delta$ values — and also shows that $\mathcal{D}$ asymptotically approaches (as $\delta \to 0$) the large circle $C$. Having formulas for the shape and size of $\mathcal{D}$ as functions of $\delta$ will be important for designing unconditionally stable schemes (2.2), and for characterizing the limitations of well-known schemes such as SBDF. Lastly, we note that the ImEx schemes parameterized by $\delta$ bare some similarity to the non-ImEx schemes with large regions of absolute stability originally examined in [48,49]. However, we will eventually choose the parameter value $\delta$ to be as large as possible (to minimize the error), while maintaining unconditional stability. This is of a fundamentally different nature than the non-ImEx study carried out in [48,49].

We now come to our first condition for unconditional stability — which is stated in terms of the generalized eigenvalues

$$\Lambda(\boldsymbol{A}, \boldsymbol{B}) := \left\{ \mu \in \mathbb{C} : -\mu \boldsymbol{A}\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}, \boldsymbol{v} \neq \boldsymbol{0} \right\}. \tag{3.8}$$

Note that a negative sign was added, for convenience, in the definition of $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ to make $(-\boldsymbol{A})$ positive definite; and that $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ is equivalent to the eigenvalues of $(-\boldsymbol{A})^{-1}\boldsymbol{B}$.

**Condition 1.** *(Unconditional stability when $(\boldsymbol{A}, \boldsymbol{B})$ are simultaneously diagonalizable) Given time stepping coefficients $(a_j, b_j, c_j)$ with diagram $\mathcal{D}$, and simultaneously diagonalizable matrices $(\boldsymbol{A}, \boldsymbol{B})$ with generalized eigenvalues $\Lambda(\boldsymbol{A}, \boldsymbol{B})$, we have the following…*

(SC) *Sufficient conditions: The scheme (2.2) is unconditionally stable if every generalized eigenvalue $\mu \in \Lambda(\boldsymbol{A}, \boldsymbol{B})$ lies in $\mathcal{D}$, i.e. $\mu \in \mathcal{D}$.*
(NC) *Necessary conditions: If a generalized eigenvalue $\mu \in \Lambda(\boldsymbol{A}, \boldsymbol{B})$ is not in $\mathcal{D}$, i.e. $\mu \notin \mathcal{D}$, then the scheme (2.2) is not unconditionally stable.*[1]

In Condition 1, the (NC) and (SC) are essentially identical and give a sharp characterization of unconditional stability. Although Condition 1 is useful when $(\boldsymbol{A}, \boldsymbol{B})$ are simultaneously diagonalizable, we also wish to consider matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ that do not commute. The results in [1] generalize the (SC) in Condition 1 to arbitrary matrices $(\boldsymbol{A}, \boldsymbol{B})$ ($\boldsymbol{A}$ still symmetric positive definite) by replacing the set $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ with a (somewhat larger) set defined in terms of a *numerical range* (also known as the *field of values*). Specifically, let $p \in \mathbb{R}$ be any real number (different values of $p$ will eventually be useful for different problem matrices $\boldsymbol{L}$), and introduce the following sets:

$$W_p(\boldsymbol{A}, \boldsymbol{B}) := \left\{ \langle \boldsymbol{v}, (-\boldsymbol{A})^{p-1}\boldsymbol{B}\boldsymbol{v} \rangle : \langle \boldsymbol{v}, (-\boldsymbol{A})^{p}\boldsymbol{v} \rangle = 1, \boldsymbol{v} \in \mathbb{C}^N \right\}. \tag{3.9}$$

[1] Strictly speaking, the precise theorem (Proposition 10, [1]) is that if $\mu \notin \mathcal{D}$, or $\mu \notin \Gamma$ where $\Gamma = \{c(z)/b(z) : |z| = 1\}$ is the boundary locus of $\mathcal{D}$, then the scheme is not unconditionally stable. However, for practical purposes, the boundary locus can be ignored since it is a curve.

The set $W_p(\boldsymbol{A}, \boldsymbol{B})$ can also be written, using a change of variables $\boldsymbol{v} = (-\boldsymbol{A})^{\frac{p}{2}}\boldsymbol{x}$, as:

$$W_p(\boldsymbol{A}, \boldsymbol{B}) = W\left((-\boldsymbol{A})^{\frac{p}{2}-1}\,\boldsymbol{B}\,(-\boldsymbol{A})^{-\frac{p}{2}}\right), \quad \text{where} \tag{3.10}$$

$$W(\boldsymbol{X}) := \left\{\langle \boldsymbol{x}, \boldsymbol{X}\boldsymbol{x}\rangle : \|\boldsymbol{x}\| = 1,\ \boldsymbol{x} \in \mathbb{C}^N\right\}. \tag{3.11}$$

Here $W(\boldsymbol{X})$ is the definition of the *numerical range* of a matrix; and is a well-known set (chapter 1, [50]) that may be computed using a sequence of eigenvalue computations [51]. Note that $W_p(\boldsymbol{A}, \boldsymbol{B})$ depends only on the matrix splitting $(\boldsymbol{A}, \boldsymbol{B})$ and is independent of the time stepping coefficients. Condition 1 may then be modified as follows.

**Condition 2.** *((Theorem 5, [1]) Unconditional stability for a general splitting* $(\boldsymbol{A}, \boldsymbol{B})$*)*

(SC) *Sufficient conditions: The scheme (2.2) is unconditionally stable if there is a value of* $p \in \mathbb{R}$ *for which the set* $W_p(\boldsymbol{A}, \boldsymbol{B})$ *is contained in* $\mathcal{D}$*, i.e.* $W_p(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$*.*

(NC) *Necessary conditions: If a generalized eigenvalue* $\mu \in \Lambda(\boldsymbol{A}, \boldsymbol{B})$ *is not in* $\mathcal{D}$*, i.e.* $\mu \notin \mathcal{D}$*, then the scheme (2.2) is not unconditionally stable.*

Note that in Condition 2 the (NC) are the same as in Condition 1, however the (SC) are no long the same — due to the non-commuting matrices. In Conditions 1–2 the (SC) provide a target criterion that will ensure unconditional stability; while the (NC) will provide insight into when a scheme may fail to be unconditionally stable.

We provide a brief explanation here for why one should replace $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ with the sets $W_p(\boldsymbol{A}, \boldsymbol{B})$ in Condition 2. If one seeks an eigenvector solution to (2.2) of the form $\boldsymbol{u}_n = z^n\boldsymbol{v}$; and then multiplies equation (2.2) from the left by $(-\boldsymbol{A})^{p-1}\boldsymbol{v}$, then one obtains equation (3.5), with the modification that the value $\mu$ is no longer a generalized eigenvalue, but is given by a general Rayleigh quotient:

$$\mu = \frac{\langle \boldsymbol{v}, (-\boldsymbol{A})^{p-1}\boldsymbol{B}\boldsymbol{v}\rangle}{\langle \boldsymbol{v}, (-\boldsymbol{A})^{p}\boldsymbol{v}\rangle} \subseteq W_p(\boldsymbol{A}, \boldsymbol{B}).$$

Hence, ensuring $W_p(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$ guarantees that the value of $\mu$ in (3.5) lies within the unconditional stability region.

**Remark 2.** (Properties of the numerical range and $W_p(\boldsymbol{A}, \boldsymbol{B})$) Since the sets $W_p(\boldsymbol{A}, \boldsymbol{B})$ can be written in terms of a numerical range, they exhibit all the well-known properties of a numerical range. The numerical range $W(\boldsymbol{X})$ for a matrix $\boldsymbol{X}$ is convex (Hausdorf–Toeplitz theorem), bounded, and always contains the eigenvalues $\mu$ of $\boldsymbol{X}$, i.e. $\mu \in W(\boldsymbol{X})$. In the case when $\boldsymbol{X}$ is a normal matrix, $W(\boldsymbol{X})$ is the convex hull of the eigenvalues. Hence, the convex hull of $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ is contained in $W_p(\boldsymbol{A}, \boldsymbol{B})$ (for all $p \in \mathbb{R}$).

**Remark 3.** Different values of $p$ may modify the size of $W_p(\boldsymbol{A}, \boldsymbol{B})$ in the complex plane. Condition 2 only requires one value of $p$ to satisfy $W_p(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$ (even if other values of $p$ violate $W_p(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$).

## 4. How to choose the ImEx parameter $\delta$ and splitting $(\boldsymbol{A}, \boldsymbol{B})$

In this section we provide general recipes for choosing the ImEx parameter $\delta$ and the matrix splitting $(\sigma\boldsymbol{A}_0, \boldsymbol{B})$ for a problem matrix $\boldsymbol{L}$. The recipes are based on minimizing a proxy for the numerical error while ensuring that the sufficient conditions (SC) are satisfied.

Solely based on the formulas for $\mathcal{D}$, one could think that one should use ImEx coefficients with very large unconditional stability region $\mathcal{D}$, by taking $\delta \ll 1$. After all, such a choice would increase the chance of unconditional stability by ensuring that $W_p(\boldsymbol{A}, \boldsymbol{B})$ fits inside $\mathcal{D}$ thereby satisfying the (SC) in Conditions 1–2.

However, choosing $\delta$ small without any regard for the error is not a good strategy. Specifically, there is a trade-off between schemes with good unconditional stability properties (i.e. small $\delta$ and large $\mathcal{D}$) and the resulting numerical accuracy. Ideally, one would choose $\delta$ so that the scheme's numerical approximation error is minimized, while still guaranteeing unconditional stability. However, because the true error is generally not accessible, we use $\delta$ as a proxy for the approximation quality, which is justified by the following remark.

**Remark 4.** (Dependence of the global truncation error constant on $\delta$) The *global truncation error* (GTE) at time $t_n = nk$ is defined by $\max_{1\le j\le N} |\boldsymbol{u}_n - \boldsymbol{u}^*(nk)|_j$. Because the ImEx schemes in Remark 1 are formally of $r$-th order, for any fixed $0 < \delta \le 1$, the GTE scales (for $k$ small) like $C_r k^r$. The error constant $C_r$ depends on $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{f}$, and the time stepping coefficients. Formulas for the behavior of the GTE error constants in a LMM may be computed in terms of the polynomials (see equation (2.3), p. 373, in [52]) $b(z)$ and $c(z)$. In particular, one may compute two separate error constants. One error constant is obtained when the ImEx scheme is applied as a fully implicit scheme (i.e. $\boldsymbol{A} = \boldsymbol{L}$, $\boldsymbol{B} = 0$) as $C_r \propto 1/c(1) = \delta^{-r}$. A second error constant may be computed when the ImEx scheme is applied to a fully explicit splitting (i.e. $\boldsymbol{B} = \boldsymbol{L}$, $\boldsymbol{A} = 0$), where $C_r \propto 1/b(1) = \delta^{-1}$. In general, for a fixed splitting $(\boldsymbol{A}, \boldsymbol{B})$, one then has a GTE that scales like

$$\text{GTE} \sim \mathcal{O}(\delta^{-r} k^r). \tag{4.1}$$

A more detailed description, along with numerical error tests verifying the asymptotic formula (4.1) may be found in [1].

Remark 4 indicates that for a fixed splitting $(\boldsymbol{A}, \boldsymbol{B})$, the GTE error is (asymptotically) minimized by taking a maximum value of $\delta$. Moreover, as a secondary trend, if a family of ImEx splittings $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ is considered, then it is generally observed that smaller values of $\sigma$ yield a smaller GTE. Hence, one should generally choose $\delta$ as *large* as possible and $\sigma$ *small*, while still satisfying the (SC) constraint in Conditions 1–2.

We now provide recipes for three different scenarios that may arise in practice. Recipe 1 specifies how to choose the ImEx parameter $\delta$ to achieve unconditional stability when a fixed matrix splitting $(\boldsymbol{A}, \boldsymbol{B})$ is specified (i.e. this a special case where $\sigma = 1$ and $\boldsymbol{A} = \boldsymbol{A}_0$).

**Unconditional Stability Recipe 1.** *How to choose the ImEx parameter $\delta$ for a fixed matrix splitting $(\boldsymbol{A}, \boldsymbol{B})$.*

0. *Choose an order $1 \le r \le 5$; and retrieve the formulas for $\mathcal{D}$ in equation (3.7).*
1. *Compute/plot the generalized eigenvalues $\Lambda(\boldsymbol{A}, \boldsymbol{B})$ and the sets $\mathcal{D}$ for different $\delta$. Then check whether $(\boldsymbol{A}, \boldsymbol{B})$ can satisfy the (NC), either graphically or via the formulas in (3.7): is there an admissible range of $\delta$ values that guarantees $\Lambda(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$? (If not, then unconditional stability is not possible for $(\boldsymbol{A}, \boldsymbol{B})$.)*
2. *Now use the sufficient conditions (SC) to determine $\delta$.*
   - *Choose a $p \in \mathbb{R}$, (try first $p = 1$). Compute $W_p(\boldsymbol{A}, \boldsymbol{B})$ from equation (3.10), for instance, using a software such as Chebfun [53].*
   - *By varying $0 < \delta \le 1$, find the largest $\delta$ that ensures $W_p(\boldsymbol{A}, \boldsymbol{B}) \subseteq \mathcal{D}$, and guarantees unconditional stability ($\mathcal{D}$ becomes larger as $\delta$ decreases). Call this parameter $\delta^*$.*
3. *If no value $0 < \delta \le 1$ can be found in Step 2, or $\delta^*$ is prohibitively small (leading to a large error constant), try and repeat Step 2 with a different $p$.*
4. *Choose a $\delta < \delta^*$ (e.g. $\delta = 0.95\,\delta^*$, with 0.95 for robustness), and substitute it into Table B.4 to obtain the ImEx coefficients for the ODE solver.*

**Example 2.** (Simple example using the Recipe 1) Consider the ODE $u_t = -10u$, with implicit part $\boldsymbol{A}u = -u$ and explicit part $\boldsymbol{B}u = -9u$ (this ODE splitting was also examined in [1]), for which we wish to devise a 3rd order ($r = 3$) unconditionally stable scheme. For this splitting, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are (trivially) simultaneously diagonalized with $\Lambda(\boldsymbol{A}, \boldsymbol{B}) = \{-9\}$. Condition 1 then requires $\{-9\} \in \mathcal{D}$ for both the (NC) and (SC). For a 3rd order scheme, $r = 3$, we use the formulas for $m_r$ and $m_l$ in (B2) so that the constraint reads:

$$m_l < -9 < m_r \quad \Longrightarrow \quad -\frac{(2-\delta)^3}{8-(2-\delta)^3} < -9 < \frac{(2-\delta)^3}{(2-\delta)^3+1}. \tag{4.2}$$

The largest $\delta$ value that satisfies the inequality (4.2) (with $<$ replaced by $\le$) is: $\delta^* = 2 - (7.2)^{1/3}$. Any value $0 < \delta < \delta^*$ will guarantee unconditional stability — i.e. one could take a fraction $\delta = 0.95\,\delta^*$ so that $\delta \approx 0.0656$. Substituting this value into the formulas in Table B.4 yields the ImEx coefficients.

In situations where one is using a pre-programmed ODE or black-box solver, it may not be possible to modify the time stepping coefficients $(a_j, b_j, c_j)$. Instead, one may have the ability to modify the matrix splitting $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ by varying the parameter $\sigma$. Recipe 2 outlines how one may choose the parameter $\sigma$ when the scheme and the matrix $\boldsymbol{A}_0$ are fixed. The recipe uses the sets $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ and $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$, whose dependence on $\sigma$ is characterized by the following remark.

**Remark 5.** (Dependence of $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ and $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ on $\sigma$) The sets $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ and $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ are simple transformations of the $\sigma$-independent sets $W_p(\boldsymbol{A}_0, \boldsymbol{L})$ and $\Lambda(\boldsymbol{A}_0, \boldsymbol{L})$:

$$\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B}) = 1 + \sigma^{-1} \Lambda(\boldsymbol{A}_0, \boldsymbol{L}), \qquad W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B}) = 1 + \sigma^{-1} W_p(\boldsymbol{A}_0, \boldsymbol{L}). \tag{4.3}$$

Here the identities (4.3) follow from a direct calculation using $\boldsymbol{B} = \boldsymbol{L} - \sigma \boldsymbol{A}_0$:

$$(-\sigma \boldsymbol{A}_0)^{-1} \boldsymbol{B} = \boldsymbol{I} + \sigma^{-1} (-\boldsymbol{A}_0)^{-1} \boldsymbol{L}. \tag{4.4}$$

$$(-\sigma \boldsymbol{A}_0)^{\frac{p}{2}-1} \boldsymbol{B} (-\sigma \boldsymbol{A}_0)^{-\frac{p}{2}} = \boldsymbol{I} + \sigma^{-1} (-\boldsymbol{A}_0)^{\frac{p}{2}-1} \boldsymbol{L} (-\boldsymbol{A}_0)^{-\frac{p}{2}}. \tag{4.5}$$

Due to properties (4.3), one can, for fixed $\boldsymbol{A}_0$ and $\boldsymbol{L}$, pre-compute the sets $\Lambda(\boldsymbol{A}_0, \boldsymbol{L})$ and $W_p(\boldsymbol{A}_0, \boldsymbol{L})$. The range $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ and generalized eigenvalues $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ are then simply rescaled versions (w.r.t. the point 1 in the complex plane) of the corresponding range and eigenvalues using $\boldsymbol{A}_0$ and $\boldsymbol{L}$, where $\sigma$ yields the scaling parameter. This becomes important in §5 when we examine and overcome the fundamental limitations of SBDF.

**Unconditional Stability Recipe 2.** *Given a fixed ImEx scheme and matrix $\boldsymbol{A}_0$, how to choose the splitting parameter $\sigma$ for the splitting $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$.*

0. *Choose an order $1 \leq r \leq 5$; and retrieve the formulas for $\mathcal{D}$ in equation (3.7). If the time stepping scheme being used is not included as one from Table B.4, then an unconditional stability diagram $\mathcal{D}$ will need to be computed.*
1. *Compute/plot the generalized eigenvalues $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ for different $\sigma$ (see Remark 5). Then check the (NC), either graphically or via the formulas in (3.7): is there an admissible range of $\sigma$ that guarantees $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$? (If not, then unconditional stability is not possible, and a different ImEx scheme or matrix $\boldsymbol{A}_0$ must be used.)*
2. *Now use the sufficient conditions (SC) to determine $\sigma$.*
   - *Choose a $p \in \mathbb{R}$ (try first $p = 1$) and compute $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (see Remark 5).*
   - *Vary $\sigma$ to find the smallest $\sigma > 0$ that ensures $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ and guarantees unconditional stability. ($W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ becomes larger as $\sigma$ decreases).*
3. *If no value of $\sigma > 0$ can be found in Step 2, repeat Step 2 with a different $p$.*

Section 5 provides examples that illustrate Recipe 2. Recipes 1 and 2 are in line with a common perspective on ImEx schemes. Either, one has to determine the ImEx parameter $\delta$ when the matrix splitting is fixed; or choose the splitting parameter $\sigma$ when the scheme is fixed. In practice, there may be cases in which neither of these two approaches is able to achieve unconditionally stability.

We therefore advocate, whenever possible, to allow to simultaneously vary the ImEx parameter $\delta$ and the splitting parameter $\sigma$. It turns out that this yields an enormous amount of flexibility when designing unconditionally stable schemes. Many splittings of the form $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$, where $\boldsymbol{A}_0$ and $\boldsymbol{L}$ are chosen and predetermined from the problem (see Sections 6–7 for specific PDE applications), can be stabilized this way.

**Unconditional Stability Recipe 3.** *Given a matrix $\boldsymbol{A}_0$, how to simultaneously choose both the ImEx and splitting parameters $(\delta, \sigma)$ (with $0 < \delta \leq 1$, $\sigma > 0$).*

1. *Repeat Steps 0–1 in Recipes 1–2 to ensure that there is a range of values $(\delta, \sigma)$ that satisfy the necessary conditions (NC) $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$. Note: $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ depends solely on $\sigma$, while $\mathcal{D}$ depends solely on $\delta$.*
2. *Use the sufficient conditions (SC) to determine $(\delta, \sigma)$.*
   - *Choose a $p \in \mathbb{R}$ (try first $p = 1$) and compute $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (see Remark 5).*
   - *The sufficient condition $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ provides a constraint on the parameters $(\delta, \sigma)$ that achieve unconditional stability. Within this constrained set, determine the points $(\delta^*, \sigma^*)$ that maximize $\delta^*$. If there is more than one solution, choose $\sigma^*$ small.*
3. *If no value of $(\delta, \sigma)$ can be found in Step 2, repeat Step 2 with a different $p$.*

Sections 6–7 provide specific applications of Recipe 3 in PDE problems.

### 4.1. Additional details for PDEs: choosing $\boldsymbol{A}_0$

When $\boldsymbol{L}_h$ arises as the spatial discretization of a PDE with meshsize $h$, one does not have a fixed matrix splitting $(\boldsymbol{A}, \boldsymbol{B})$, or $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$, but rather a family of splittings parameterized by $h$: $(\boldsymbol{A}_h, \boldsymbol{B}_h)$, or $(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$. In this situation, it is crucial to be able to choose the parameters $(\delta, \sigma)$ independent of the meshsize $h$ — i.e. to have one and the same ImEx scheme be unconditionally stable for an entire family of splittings $(\boldsymbol{A}_h, \boldsymbol{B}_h)$, or $(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$. If, for example, unconditional stability required one to choose the ImEx parameter $\delta$ as a function of the grid size $h$ (i.e. such as $\delta = h$), then such a choice would have a deleterious effect on the GTE (GTE $\sim \mathcal{O}(h^{-r} k^r)$), and limit the benefits of unconditional stability.

To be able to choose a single set of parameters $(\delta, \sigma)$ that stabilizes the family of splittings $(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$ for all $h$, some care must be taken to ensure the matrix $\boldsymbol{A}_{0,h}$ is properly chosen relative to $\boldsymbol{B}_h$. Once a suitable choice of $\boldsymbol{A}_{0,h}$ is fixed, one may use the Recipe 3 to simultaneously choose $(\sigma, \delta)$ for unconditional stability.

**Remark 6.** (Guidelines for choosing $\boldsymbol{A}_{0,h}$ when $\boldsymbol{L}_h$ is the spatial discretization of a PDE) Generally speaking, it is a good idea to ensure that $\boldsymbol{A}_{0,h}$ has the same derivative order as $\boldsymbol{L}_h$, as backed up by the following heuristic scaling argument. Suppose

$$\boldsymbol{L}_h \approx C(x) \frac{\partial^q}{\partial x^q} + (\text{lower order derivatives}).$$

A natural choice for $\boldsymbol{A}_{0,h}$ might be

$$\boldsymbol{A}_{0,h} \approx \frac{\partial^s}{\partial x^s}$$

(one could include a variable coefficient approximation as well). Many spatial approximation methods yield the scaling $\frac{\partial}{\partial x} \propto h^{-1}$, hence one may expect some of the eigenvalues of $(\boldsymbol{A}_{0,h})^{-1}\boldsymbol{L}_h$ to scale like $\mathcal{O}(h^{s-q})$. The re-scaling formulas in Remark 5 then imply that there may be generalized eigenvalues $\mu \in \Lambda(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$ that scale like $\mu = 1 - \sigma^{-1}\mathcal{O}(h^{s-q})$. This gives rise to three cases for choosing $s$:

- If $s < q$, some of the generalized eigenvalues diverge $\mu = 1 - \sigma^{-1}\mathcal{O}(h^{s-q}) \to \infty$ as $h \to 0$. Using formula (B3) for the asymptotic behavior of $\mathcal{D}$, the ImEx parameter $\delta$ would then have to scale like $\delta \sim h^{q-s}$ as $h \to 0$ (and fixed $\sigma$) to ensure that these large eigenvalues remain inside $\mathcal{D}$ (to satisfy the (NC)). Hence, $(\delta, \sigma)$ cannot be chosen independent of the mesh $h$.
- If $s > q$, some of the generalized eigenvalues $\mu = 1 - \sigma^{-1}\mathcal{O}(h^{s-q}) \to 1$ as $h \to 0$ (and fixed $\sigma$). In this case, the formulas in (B2) show that only order $r = 1, 2$ schemes contain the point $1 \in \mathcal{D}$ (see Fig. 1). Hence, $s > q$ is generally not a good choice if one is looking for a scheme with orders $r > 2$.
- If $s = q$, then all generalized eigenvalues $\mu$ have a chance (based solely on the scaling of $h$) to be uniformly bounded (i.e. do not become arbitrarily large) as $h \to 0$; and also remain strictly bounded away from 1 as $h \to 0$. In this case, there is a chance to obtain high order by means of choosing the parameters $(\delta, \sigma)$ independent of $h$.

## 5. Limitations of unconditional stability for SBDF schemes

In §3, the unconditional stability region $\mathcal{D}$ was used to derive sufficient (SC) and necessary (NC) conditions for unconditional stability. Using these conditions, this section illustrates how the geometrical properties of $\mathcal{D}$ can be used to understand the fundamental limitations that classical SBDF methods possess with regard to unconditional stability. Specifically, two significant qualitative transitions occur: (i) moving from 1st to 2nd order schemes for non-symmetric matrices $\boldsymbol{L}$; and (ii) moving from 2nd to 3rd order for symmetric matrices $\boldsymbol{L}$. Guided by Recipe 2, we discuss under which circumstances a choice of $\sigma$ exists so that a splitting $(\sigma \boldsymbol{A}_0, \boldsymbol{L})$ is unconditionally stable with SBDF.

**Case 1: $\boldsymbol{L}$ non-symmetric.** Let $\boldsymbol{L}$ be a non-symmetric matrix that, together with $\boldsymbol{A}_0$, has both a range $\mathrm{Re}\big(W_p(\boldsymbol{A}_0, \boldsymbol{L})\big) < 0$ and eigenvalues $\mathrm{Re}\big(\Lambda(\boldsymbol{A}_0, \boldsymbol{L})\big) < 0$ with negative real part, i.e. they lie strictly in the left-half plane, but are not necessarily contained on the real line. Such a situation occurs for instance in discretizations of advection–diffusion PDEs (with an implicit diffusion, and explicit advection). The following transition arises between first and second order SBDF when the ImEx splitting is taken as $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$:

1. SBDF1 can always be made unconditionally stable, by choosing $\sigma$ suitably large. This is due to the fact that $\mathcal{D}$ for SBDF1 is a circle with its right-most point at 1. Hence one can always rescale $W_p(\boldsymbol{A}_0, \boldsymbol{L})$ (see Remark 5) so that $W_p(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$.
2. SBDF2 can, in general, *not* be made unconditionally stable by means of choosing $\sigma > 0$. This is a result of the cusp at 1 in $\mathcal{D}$ (see Fig. 2). If, for instance, the imaginary part of $\mu \in \Lambda(\boldsymbol{A}_0, \boldsymbol{L})$ is larger (in absolute value) than its real part, then the scaled eigenvalue (see Remark 5) $1 + \sigma^{-1}\mu \in \Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{L})$ will never enter $\mathcal{D}$, regardless of the value of $\sigma$.

We highlight these insights with the following simple example.

**Example 3.** (A non-symmetric $\boldsymbol{L}$) Consider the following non-symmetric matrix $\boldsymbol{L}$ and choice of matrix $\boldsymbol{A}_0$:

$$\boldsymbol{L} = \begin{pmatrix} -0.2 & 0 & 0 \\ 0 & -2 & 2 \\ 0 & -2 & -2 \end{pmatrix}, \quad \boldsymbol{A}_0 = -\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.1}$$

The generalized eigenvalues are $\Lambda(\boldsymbol{A}_0, \boldsymbol{L}) = \{-0.2, -2 + 2\imath, -2 - 2\imath\}$, and $W_1(\boldsymbol{A}_0, \boldsymbol{L}) = \mathrm{conv}\{-0.2, -2 + 2\imath, -2 - 2\imath\}$ is a triangle consisting of the convex hull[2] of the eigenvalues. Note that, for simplicity, we have chosen an example in which $\boldsymbol{A}$ and $\boldsymbol{B}$ commute. Hence, Condition 1 may be used. We also plot $W_1(\boldsymbol{A}, \boldsymbol{B})$ to illustrate how to apply Condition 2 (which is stronger than Condition 1) when one is faced with matrices $\boldsymbol{A}, \boldsymbol{B}$ that do not commute. Fig. 2 visualizes that SBDF1 can be made unconditionally stable with $\sigma = 2.5$, while SBDF2 cannot be made unconditionally stable by only varying $\sigma > 0$. However, high order schemes (i.e. $r \geq 2$) that are unconditionally stable for (5.1) *are* possible by varying both $(\delta, \sigma)$, as seen in Fig. 3.

**Case 2: $\boldsymbol{L}$ symmetric.** Let $\boldsymbol{L}$ be a symmetric negative definite matrix. Assume now that $\boldsymbol{A}_0$ is such that the range of $W_1(\boldsymbol{A}_0, \boldsymbol{L})$ and eigenvalues $\Lambda(\boldsymbol{A}_0, \boldsymbol{L})$ are real and strictly negative. Such a situation arises for instance in the discretization of a purely parabolic (gradient flow) problem. The following transition occurs between second- and third-order schemes for the splittings $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$:

[2] The matrix $\boldsymbol{L}$ is normal, which results in a simple expression for the range $W_1(\boldsymbol{A}_0, \boldsymbol{L})$.

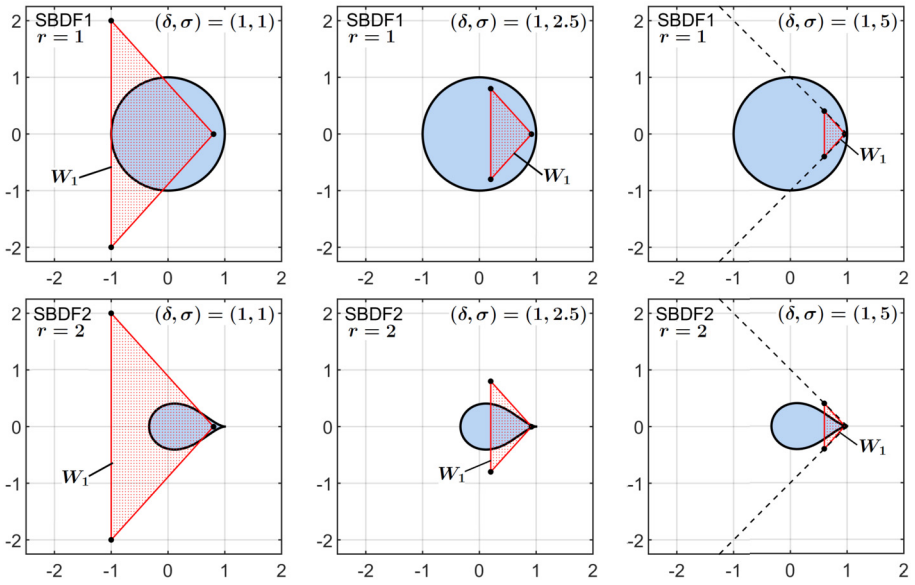


**Fig. 2.** Example for non-symmetric $\boldsymbol{L}$ in (5.1). The figures show the SBDF1 (top row) and SBDF2 (bottom row) stability diagrams $\mathcal{D}$ (blue shaded region) in relation to the sets $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (red shaded region, abbreviated as $W_1$) and $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (black dots) for (left to right) $\sigma \in \{1, 2.5, 5\}$. Note that $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ for SBDF1 with $\sigma \in \{2.5, 5\}$ guaranteeing the (SC) for unconditional stability. The bottom row highlights the fundamental limitation for SBDF2: no $\sigma > 0$ exists that can ensure $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$. Dashed lines show the effect of the rescaling by $\sigma$, outlined in Remark 5, on the set $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B})$.

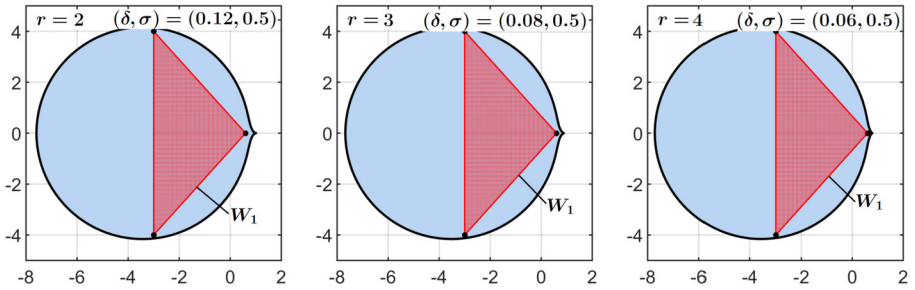


**Fig. 3.** Unconditionally stable schemes for (5.1), and orders (left to right) $r \in \{2, 3, 4\}$. The figure shows that $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ when $\sigma = 0.5$, guaranteeing the (SC) for unconditional stability. Values are (left to right) $\delta \in \{0.12, 0.08, 0.06\}$. The chosen $(\delta, \sigma)$-values are guided by Recipe 3, and are almost optimal, however other values are also possible.

1. SBDF2 can always be made unconditionally stable, by choosing $\sigma$ suitably large. This is due to the fact that the right-most point of $\mathcal{D}$ for SBDF2 is 1, and $W_1(\boldsymbol{A}_0, \boldsymbol{L})$ is real and negative, so one can always rescale (see Remark 5) $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ into $\mathcal{D}$.
2. SBDF3 can, in general, *not* be made unconditionally stable by means of choosing $\sigma > 0$. This is because the right-most point of $\mathcal{D}$ is $1/2$ (instead of 1), so that a negative real $\Lambda(\boldsymbol{A}_0, \boldsymbol{L})$ may be impossible to contain within $\mathcal{D}$ via the choice of $\sigma$.

Unconditional stability limitations of SBDF, applied to splittings $(\sigma \boldsymbol{A}_0, \boldsymbol{B})$, may be overcome by simultaneously choosing $(\delta, \sigma)$, i.e. by following Recipe 3.

**Example 4.** (A symmetric $\boldsymbol{L}$) Consider the following symmetric matrices:

$$\boldsymbol{L} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \quad \boldsymbol{A}_0 = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.2}$$

Then $\Lambda(\boldsymbol{A}_0, \boldsymbol{L}) = \{-3, -1\}$, and $W_1(\boldsymbol{A}_0, \boldsymbol{L}) = [-3, -1]$ is an interval along the real axis. Fig. 4 shows that SBDF2 can be made unconditionally stable with $\sigma = 2.5$, while SBDF3 cannot be made unconditionally stable by only varying $\sigma > 0$. In contrast, third to fifth order schemes that are unconditionally stable for (5.2) *are* possible by varying both $(\delta, \sigma)$, as seen in Fig. 5.

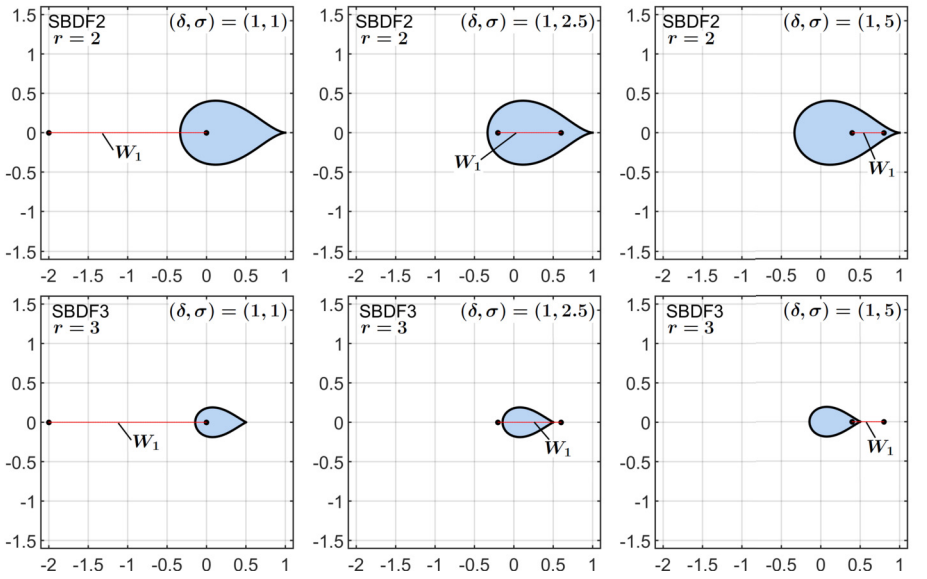


**Fig. 4.** Example for symmetric $\boldsymbol{L}$ in (5.2). The figures show the SBDF2 (top row) and SBDF3 (bottom row) stability diagrams $\mathcal{D}$ (blue shaded region) in relation to the sets $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (red region, abbreviated $W_1$) and $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B})$ (black dots) for (left to right) $\sigma \in \{1, 2.5, 5\}$. Note that $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ for SBDF2 with $\sigma \in \{2.5, 5\}$, guaranteeing unconditional stability. The bottom row highlights the fundamental limitation for SBDF3: no $\sigma > 0$ exists that can ensure $\Lambda(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$.

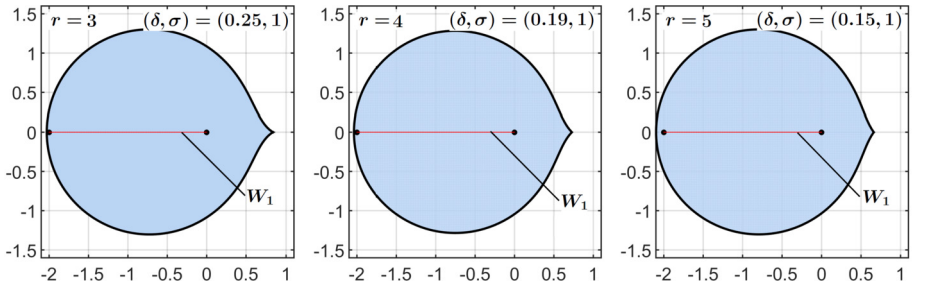


**Fig. 5.** Unconditionally stable schemes for (5.2), and orders (left to right) $r \in \{3, 4, 5\}$. The figure shows that $W_1(\sigma \boldsymbol{A}_0, \boldsymbol{B}) \subseteq \mathcal{D}$ provided $\sigma = 1$ and (left to right) $\delta \in \{0.25, 0.19, 0.15\}$. The chosen $(\delta, \sigma)$-values are guided by Recipe 3, however other values are also possible.

## 6. Examples from diffusion PDEs

In this section we apply the unconditional stability theory from §4.1 and Recipe 3 to PDE diffusion problems with spatially varying, and even non-linear diffusion coefficients. The presented methodology highlights how one can avoid a stiff time step restriction (here: of diffusive type $k \propto h^2$, where $h$ is the smallest grid size) — or any time step restriction for that matter — while inverting only simple constant coefficient matrices. This allows one to leverage *fast solvers* where an implicit treatment of $\boldsymbol{A}_h$ (i.e. using the fast Fourier transform) can be carried out much more rapidly than a fully implicit treatment of $\boldsymbol{L}_h$ (i.e. that contains all the stiff diffusive terms). The new ImEx coefficients (see §5) enable high order time stepping beyond what is possible using only SBDF methods.

### 6.1. Numerical discretization

We start by providing numerical details for the one-dimensional Fourier spectral methods used. Computations in three dimensions are then conducted by naturally extending the one-dimensional approach via Cartesian products. We use a periodic computational domain $\Omega = [0, 1]$; discretize space using a uniform grid with an even number of grid points $N$; and approximate the function $u(x)$ at $x_j$ by $u_j \approx u(x_j)$, where:

$$x_j = jh, \quad h = \frac{1}{N}, \quad \boldsymbol{u} = \left(u_1, u_2, \ldots, u_N\right)^T \in \mathbb{R}^N.$$

Because our analysis is based on the matrices $\boldsymbol{A}_{0,h}$ that are written in terms of Fourier transforms, it is useful to introduce notation for the discrete Fourier transform (DFT) matrix $\boldsymbol{F}$, and for the spectral differentiation matrix $\boldsymbol{D}$ — even

though in practice one will never use those matrices, but rather use the *fast Fourier transform* (FFT) to compute $\boldsymbol{F}\boldsymbol{u} = \mathrm{fft}(\boldsymbol{u})$. The DFT matrix $\boldsymbol{F}$ has the coefficients:

$$\boldsymbol{F}_{j\ell} = \omega^{(j-1)\times(\ell-1)}, \quad \omega = e^{-\frac{2\pi\imath}{N}}, \quad \text{so that} \quad (\boldsymbol{F}\boldsymbol{u})_j = \sum_{\ell=1}^{N} u_\ell\, \omega^{(j-1)\times(\ell-1)}.$$

The (spectral) differentiation of a function defined on the uniform grid amounts to a scalar multiplication in Fourier space, i.e. $(\boldsymbol{D}\boldsymbol{u})_j \approx u_x(x_j)$. Hence, the matrix $\boldsymbol{D}$ takes the form: $\boldsymbol{D} = \imath\, \boldsymbol{F}^{-1}\, \mathrm{diag}(\boldsymbol{\xi})\, \boldsymbol{F}$, where $\mathrm{diag}(\boldsymbol{\xi})$ denotes the matrix with diagonal entries of the vector:

$$\boldsymbol{\xi} = \big(\xi_1, \xi_2, \ldots, \xi_N\big)^T \in \mathbb{R}^N, \quad \text{where} \quad \xi_j = \begin{cases} 2\pi\,(j-1) & \text{if } 1 \le j \le \frac{N}{2}, \\ 2\pi\,(j-N) & \text{if } \frac{N}{2}+1 < j \le N, \\ N\pi & \text{if } j = \frac{N}{2}+1. \end{cases} \tag{6.1}$$

Since $\boldsymbol{F}^{-1} = N^{-1}\boldsymbol{F}^\dagger$, the matrix $\boldsymbol{D}^\dagger = -\boldsymbol{D}$ is skew-Hermitian and the matrix $\boldsymbol{D}^2$ is Hermitian. If $\boldsymbol{A}_h$ is diagonalized by $\boldsymbol{F}$, then solving for $\boldsymbol{u}_{n+r}$ in the implicit step of the evolution (2.2), i.e. $(a_r\boldsymbol{I} - kc_r\boldsymbol{A}_h)\boldsymbol{u}_{n+r} = \mathrm{RHS}$, is done via two FFTs.

### 6.2. An FFT-based treatment for the variable coefficient diffusion equation

We now devise unconditionally stable ImEx schemes for the variable coefficient diffusion equation (with diffusion coefficient $d(x) > 0$)

$$u_t = \big(d(x)u_x\big)_x + f(x,t), \qquad \text{on } \Omega \times (0,T], \tag{6.2}$$

that make use of an FFT-based treatment of the implicit matrix $\boldsymbol{A}_h$. The choice of splitting $(\boldsymbol{A}_h, \boldsymbol{B}_h)$ is guided by §4.1, and the choice of parameters $(\delta, \sigma)$ by Recipe 3. To ensure a high spatial accuracy, we adopt a spectral discretization of equation (6.2) and set:

$$\boldsymbol{L}_h = \boldsymbol{D}\big(\mathrm{diag}(\boldsymbol{d})\big)\boldsymbol{D}, \quad \text{where} \quad \boldsymbol{d} = \big(d(x_1), d(x_2), \ldots, d(x_N)\big)^T.$$

Note that $\boldsymbol{L}_h$ is a dense matrix and (due to the $x$-dependence of $d(x)$) is not diagonalized via the DFT matrix $\boldsymbol{F}$. To seek an ImEx splitting of $\boldsymbol{L}_h$, we follow the guidelines in Remark 6: the matrix $\boldsymbol{L}_h$ has two factors of $\boldsymbol{D}$ and hence the implicit matrix $\boldsymbol{A}_h$ should have two factors of $\boldsymbol{D}$ as well. This motivates the following matrix splitting:

$$\boldsymbol{A}_h = \sigma\,\boldsymbol{D}^2, \qquad \boldsymbol{B}_h = \boldsymbol{D}\,\Big(\mathrm{diag}(\boldsymbol{d}) - \sigma\,\boldsymbol{I}\Big)\,\boldsymbol{D}, \tag{6.3}$$

i.e. $\boldsymbol{A}_h\boldsymbol{u} \approx \sigma\, u_{xx}$ and $\boldsymbol{B}_h\boldsymbol{u} \approx \big((d(x)-\sigma)u_x\big)_x$.

Our goal is to determine, following Recipe 3, the parameters $(\delta, \sigma)$ that guarantee unconditional stability. Before doing so, we must discuss a caveat: the matrices $\boldsymbol{L}_h$ and $\boldsymbol{A}_h$ are not invertible — which was an assumption in the derivation of the conditions for unconditional stability. We do not provide a general treatment for when $\boldsymbol{A}_h$ is not invertible due to subtleties that may arise (for instance when the null space of $\boldsymbol{A}_h$ *interacts* with $\boldsymbol{B}_h$ through the ImEx evolution). However, for the specific splitting (6.3), the unconditional stability theory presented in §3 (and recipes in §4) can be applied with only a minor adaptation, namely: the definition/computation of the sets $W_p(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ are done on the subspace $\mathbb{V}$ where $\boldsymbol{A}_h$ is invertible, as follows.

The matrix $\boldsymbol{D}$ has a null space spanned by the constant vector $\mathbf{1} = (1,\ 1,\ \ldots,\ 1)^T$. Hence, $\boldsymbol{D}^2$ and $\boldsymbol{A}_h$ have the null space $\mathbb{1}$, and column space (range) $\mathbb{V}$ where:

$$\mathbb{V} := \{\boldsymbol{u} \in \mathbb{C}^N : \mathbf{1}^T\boldsymbol{u} = \mathbf{0}\}, \quad \mathbb{1} := \mathrm{span}\{\mathbf{1}\}, \quad \text{so that} \quad \mathbb{C}^N = \mathbb{V} \oplus \mathbb{1}.$$

Using the orthogonal projection $\boldsymbol{P} = \boldsymbol{I} - N^{-1}\mathbf{1}\,\mathbf{1}^T$ onto $\mathbb{V}$, and noting that $\boldsymbol{A}_h$ (and also $\boldsymbol{B}_h$) satisfies $\mathbf{1}^T\boldsymbol{A}_h = \boldsymbol{A}_h\,\mathbf{1} = \mathbf{0}$, so that $\boldsymbol{A}_h = \boldsymbol{P}\,\boldsymbol{A}_h = \boldsymbol{A}_h\,\boldsymbol{P}$, the evolution equation

$$\boldsymbol{u}_t = \boldsymbol{A}_h\boldsymbol{u} + \boldsymbol{B}_h\boldsymbol{u} \tag{6.4}$$

decouples into separate components that lie in the subspaces $\mathbb{1}$ and $\mathbb{V}$ (i.e. $\mathbb{1}$ and $\mathbb{V}$ are invariant subspaces of equation (6.4)):

$$\text{Dynamics in } \mathbb{1}: \qquad \big(\mathbf{1}^T\boldsymbol{u}\big)_t = \mathbf{1}^T\big(\boldsymbol{A}_h\boldsymbol{u} + \boldsymbol{B}_h\boldsymbol{u}\big) = 0. \tag{6.5}$$

$$\text{Dynamics in } \mathbb{V}: \qquad \big(\boldsymbol{P}\boldsymbol{u}\big)_t = \boldsymbol{P}\big(\boldsymbol{A}_h\boldsymbol{u} + \boldsymbol{B}_h\boldsymbol{u}\big) = \boldsymbol{A}_h\big(\boldsymbol{P}\boldsymbol{u}\big) + \boldsymbol{B}_h\big(\boldsymbol{P}\boldsymbol{u}\big). \tag{6.6}$$

Equation (6.5) shows that the mean of $\boldsymbol{u}$, i.e. $(\mathbf{1}^T\boldsymbol{u})$, remains constant. Any zero-stable ImEx scheme (such as the ones we use) applied to (6.4) automatically ensures that $(\mathbf{1}^T\boldsymbol{u})$ evolves according to (6.5) with stable growth factors (independent of $k$, given by $a(z) = 0$). Hence, the mean $(\mathbf{1}^T\boldsymbol{u})$ is unconditionally stable. In turn, equation (6.6) can be viewed as the

restriction of equation (6.4) to the space $\mathbb{V}$. Because $\boldsymbol{A}_h$ is invertible on $\mathbb{V}$, the stability theory outlined in §3 applies to equation (6.6), where the sets $W_p(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ are computed on the subspace $\mathbb{V}$ instead of $\mathbb{C}^N$. To summarize the results:

**Remark 7.** (Modification of $W_p(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ for a non-invertible $\boldsymbol{A}_h$) The splitting (6.3) with the discretization in §6.1 leads to a matrix $\boldsymbol{A}_h$ that is not invertible. This violates the assumptions for the necessary and sufficient conditions in §3. Nevertheless, Conditions 1–2 may be used, provided $W_p(\boldsymbol{A}_h, \boldsymbol{B}_h)$ is computed on the space $\mathbb{V}$:

$$W_p(\boldsymbol{A}_h, \boldsymbol{B}_h) = \left\{ \langle \boldsymbol{x}, (-\boldsymbol{A}_h)^{p-1} \boldsymbol{B}_h \boldsymbol{x} \rangle : \langle \boldsymbol{x}, (-\boldsymbol{A}_h)^p \boldsymbol{x} \rangle = 1, \boldsymbol{x} \in \mathbb{V} \right\},$$

and likewise, $\mu \in \Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ are restricted to the eigenvalues with corresponding eigenvectors $\boldsymbol{v} \in \mathbb{V}$.

With a slight abuse of notation, we continue to use $W_p(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ throughout this section with the understanding that they are computed only on the subspace $\mathbb{V}$.

Owing to the simple structure of $\boldsymbol{B}_h$ in relation to $\boldsymbol{A}_h$, we can compute (almost exactly) the (modified) set $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ described in Remark 7, as well as the minimum and maximum eigenvalues (the eigenvalues in this case are real) $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ in terms of the discrete vector $\boldsymbol{d}$ and diffusion coefficient $d(x)$. To do so, we introduce the notation

$$d_{\min} = \min_{x \in \Omega} d(x), \quad d_{\max} = \max_{x \in \Omega} d(x),$$

as well as the discrete values

$$d_{2,\min} = \left\{ \text{Second smallest element of } \boldsymbol{d} \right\}, \quad \mu_{\min} = \min \left\{ \mu : \mu \in \Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h) \right\},$$
$$d_{2,\max} = \left\{ \text{Second largest element of } \boldsymbol{d} \right\}, \quad \mu_{\max} = \max \left\{ \mu : \mu \in \Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h) \right\}.$$

The sets $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and max/min values in $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ then satisfy:

**Proposition 6.1.** *The set $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ for any splitting of the form (6.3) is strictly real and contained inside the interval:*

$$1 - \sigma^{-1} d_{\max} \leq W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \leq 1 - \sigma^{-1} d_{\min}.$$

*Moreover, the generalized eigenvalues $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ are all real, and are bounded by*

$$1 - \sigma^{-1} d_{\max} \leq \mu_{\min} \leq 1 - \sigma^{-1} d_{2,\max}, \qquad 1 - \sigma^{-1} d_{2,\min} \leq \mu_{\max} \leq 1 - \sigma^{-1} d_{\min}.$$

**Remark 8.** (Motivation based on operators) The intuition for the proof of Proposition 6.1 arises at the continuum level of differential operators. Roughly speaking, one can write $\mathcal{A} = \frac{d^2}{dx^2}$ and $\mathcal{B} = \frac{d}{dx} d(x) \frac{d}{dx}$, so one may expect $\mathcal{A}^{-\frac{1}{2}} \propto (\frac{d}{dx})^{-1}$. This yields the operator product $\mathcal{A}^{-\frac{1}{2}} \mathcal{B} \mathcal{A}^{-\frac{1}{2}} = d(x)$, which allows for the computation of $W_1(\mathcal{A}, \mathcal{B})$. The proof of Proposition 6.1 in Appendix A effectively formalizes this operator computation at the level of matrices. Moreover, due to the continuum nature of the argument, Proposition 6.1 carries over to other spatial discretizations, such as other spectral methods, finite differences, etc.

Proposition 6.1 is useful as it allows the design of unconditionally stable ImEx schemes by choosing $(\delta, \sigma)$ so that $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$. It is significant for two more reasons. First, the bounds on $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ in Proposition 6.1 do not depend on $h$. This allows one to choose a single ImEx parameter $\delta$ (independent of $h$) to stabilize an entire family of splittings $(\boldsymbol{A}_h, \boldsymbol{B}_h)$. Second, the proposition is almost exact:

**Remark 9.** (Proposition 6.1 is almost exact) Although the formulas in Proposition 6.1 are inequalities, they are almost exact. For smooth functions $d(x)$, the values $d_{2,\min}, d_{2,\max}$ are at least $\mathcal{O}(N^{-1})$ close to $d_{\min}$ and $d_{\max}$. Hence, the bounds for $\mu_{\min}$ or $\mu_{\max}$ are sharp to within $\mathcal{O}(N^{-1})$. In a similar fashion, it can be shown that the inequalities on the set $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ in Proposition 6.1 are accurate to within an error $\mathcal{O}(N^{-1})$.

We now follow Recipe 3 to choose both $(\delta, \sigma)$ to design an unconditionally stable scheme:

1. Retrieve the formulas for $\mathcal{D}$. Since both $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h)$ and $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ are real, it is sufficient to use the interval $[m_l, m_r]$ of $\mathcal{D}$ on the real line via the formulas (B2).
2. The second step is heuristic only: establish a range of $(\delta, \sigma)$-values that ensure the (NC), i.e. $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$. In this case, the upper (resp. lower) estimate for $\mu_{\max}$ (resp. $\mu_{\min}$) agrees exactly with the upper (resp. lower) estimate on $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$. Therefore, there is (essentially) no difference in trying to ensure that $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$, versus $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$.

3. Apply the (SC) to determine feasible $(\delta, \sigma)$-values. Setting $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$ requires that the endpoints of $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ lie within $\mathcal{D}$:

$$\text{Left endpoint of } W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \text{ in } \mathcal{D}: \quad m_l < 1 - \sigma^{-1} d_{\max}, \tag{6.7}$$

$$\text{Right endpoint of } W_1(\boldsymbol{A}_h, \boldsymbol{B}_h) \text{ in } \mathcal{D}: \quad 1 - \sigma^{-1} d_{\min} < m_r. \tag{6.8}$$

Equations (6.7)–(6.8) can be rewritten as:

$$(1 - m_l)^{-1} d_{\max} < \sigma \quad \text{and} \quad \sigma < (1 - m_r)^{-1} d_{\min}. \tag{6.9}$$

The inequalities (6.9), along with $\sigma > 0$, $0 < \delta \leq 1$, establish the feasible points $(\delta, \sigma)$ that guarantee unconditional stability. This feasible set is always non-empty, because as $\delta \to 0$, (6.9) yields $0 < \sigma < (1 + \cos^{-r}(\pi/r)) d_{\min}$ for $2 \leq r \leq 5$. Hence, one can always achieve unconditional stability, by choosing $\delta$ small enough.

4. The last step is to choose a value $(\delta, \sigma)$ in the feasible set that *maximizes* $\delta$ — which is a proxy for minimizing the numerical truncation error.

**Case 1:** $1 \leq r \leq 2$. Here the maximum value of $\delta^* = 1$ is feasible (i.e., one may use SBDF). The upper bound inequality (6.9) is satisfied since $m_r = 1$ yields $\sigma < \infty$. The lower bound constraint on $\sigma$ leads to a range of possible $(\delta^*, \sigma^*)$ values:

$$\text{For } r = 1\text{: } \delta^* = 1, \ \ \sigma^* > \tfrac{1}{2} d_{\max}, \qquad \text{For } r = 2\text{: } \delta^* = 1, \ \ \sigma^* > \tfrac{3}{4} d_{\max}. \tag{6.10}$$

Generally speaking, choosing $\sigma^*$ large leads to large truncation errors. This motivates a choice of $\sigma^*$ close to the minimum possible values above.

**Case 2:** $3 \leq r \leq 5$. In this case, SBDF may not be able to guarantee unconditional stability (see §4) when the value $\delta^* = 1$ is outside the inequalities (6.9). Fig. 6 displays the allowable $(\delta, \sigma)$-values defined by the inequalities in (6.9) for some representative $d_{\min}$ and $d_{\max}$ values. Note that the optimal point (i.e. maximum $\delta$) occurs at the intersection of the inequalities (6.9), and is below 1. To solve for the optimal $(\delta^*, \sigma^*)$ values, we set the upper and lower bounds in (6.9) *almost* equal to each other. Specifically, introduce a *gap parameter* $0 < \eta < 1$, and set the left and right hand inequalities for $\sigma$ in (6.9) equal to within a factor of $(1 - \eta)$ to eliminate $\sigma$:

$$\Big(1 - (1 - \delta/2)^r\Big) d_{\max} = (1 - \eta)\Big(1 + (1 - \delta/2)^r \cos^{-r}(\pi/r)\Big) d_{\min}. \tag{6.11}$$

Equation (6.11) defines an optimal (largest, up to a $(1 - \eta)$ error) $\delta^*$ value. Substituting this $\delta^*$ back into (6.9) yields a range (roughly of size $(1 - \eta)$) of feasible $\sigma$ values. Among those, we choose (somewhat arbitrarily) $\sigma^*$ as the average of the two bounds in (6.9). Formulas for the (almost) optimal solutions $(\delta^*, \sigma^*)$ are given as follows: fix a gap parameter $0 < \eta < 1$ (smaller values of $\eta$ are more optimal) and order $3 \leq r \leq 5$:

$$\delta^* = 2 - 2\Big(\frac{1 - \kappa}{1 + \kappa \cos^{-r}(\pi/r)}\Big)^{1/r}, \quad \sigma^* = d_{\min}\Big(1 - \frac{1}{2}\eta\Big)\frac{1 + \cos^{-r}(\pi/r)}{1 + \kappa \cos^{-r}(\pi/r)}, \tag{6.12}$$

$$\text{where } \kappa = \frac{d_{\min}}{d_{\max}}(1 - \eta).$$

**Remark 10.** (Failure of unconditional stability for SBDF and orders $3 \leq r \leq 5$) The necessary conditions for unconditional stability require that the generalized eigenvalues $\Lambda(\boldsymbol{A}_h, \boldsymbol{B}_h) \subseteq \mathcal{D}$. The formulas from Proposition 6.1 lead to the requirement that:

$$(1 - 2^{-r})\, d_{2,\max} \leq \sigma \quad \text{and} \quad \sigma \leq \big(1 + 2^{-r} \cos^{-r}(\pi/r)\big)\, d_{2,\min}.$$

These two inequalities cannot be simultaneously satisfied if $d_{2,\max} > D_r\, d_{2,\min}$, where $D_r$ is given by $D_3 = 2.1429$, $D_4 = 1.2667$, $D_5 = 1.0931$ for orders $r = 3, 4, 5$, respectively. Note that $d_{2,\max}/d_{2,\min}$ is (up to $\mathcal{O}(N^{-1})$) a measure of the ratio $d_{\max}/d_{\min}$. As a result, if the ratio between the maximum and minimum diffusion coefficient values exceed $D_r$, then SBDF cannot provide unconditional stability for splittings of the form (6.3).

**Remark 11.** (Overcoming limitations for SBDF and orders $3 \leq r \leq 5$) The formulas (6.12) provide a way to overcome the unconditional stability limitations encountered with SBDF methods for variable coefficient diffusion problems — regardless of the diffusion coefficient $d(x)$.

To demonstrate that the new approach works in practice, we conduct a convergence test of equation (6.2) with the variable diffusion coefficient

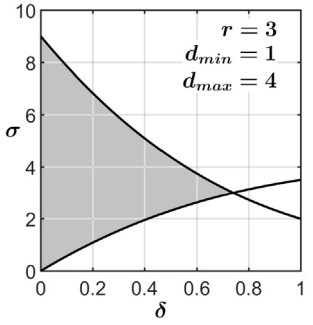


**Fig. 6.** Example of feasible $(\delta, \sigma)$-values (shaded region), that satisfy (SC) for order $r = 3$ and values $d_{\min} = 1$, $d_{\max} = 4$. The feasible point that maximizes $\delta$ is approximately $(\delta, \sigma) \approx (0.74, 3)$.

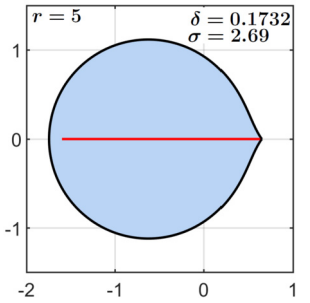


**Fig. 7.** Set $\mathcal{D}$ (blue region) containing $W_1(\boldsymbol{A}_h, \boldsymbol{B}_h)$ (shown in red) with (optimal) parameters $(\delta, \sigma) = (0.1732, 2.69)$, order $r = 5$, $N = 64$.

**Table 1**
Errors $\|u - u^*\|_{\infty,h}$ for variable coefficient diffusion test case (6.3) with (6.13), using ImEx and splitting parameters $(\delta, \sigma) = (0.1732, 2.69)$, final time $t_f = 5$, and $N = 64$ Fourier modes. The range of fourth and fifth order convergence is capped due to round-off errors, amplified by the problem's conditioning.

| Num. Steps | $k$ | Error $r=1$ | Rate | Error $r=2$ | Rate | Error $r=3$ | Rate | Error $r=4$ | Rate | Error $r=5$ | Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 7.9e+00 | – | 4.7e+01 | – | 3.4e+02 | – | 9.0e+02 | – | 8.8e+02 | – |
| 10 | $2^{-1}$ | 3.4e+00 | 1.2 | 6.7e+01 | −0.5 | 4.9e+02 | −0.5 | 2.2e+03 | −1.3 | 4.3e+03 | −2.3 |
| 20 | $2^{-2}$ | 4.3e+00 | −0.4 | 2.4e+01 | 1.5 | 5.6e+02 | −0.2 | 3.7e+03 | −0.7 | 6.3e+03 | −0.5 |
| 40 | $2^{-3}$ | 1.3e+00 | 1.7 | 3.5e+01 | −0.5 | 5.4e+02 | 0.1 | 6.3e+03 | −0.8 | 5.8e+04 | −3.2 |
| 80 | $2^{-4}$ | 6.9e−01 | 1.0 | 7.1e+00 | 2.3 | 1.3e+01 | 5.4 | 7.4e+02 | 3.1 | 6.0e+03 | 3.3 |
| 160 | $2^{-5}$ | 2.7e−01 | 1.4 | 1.0e+00 | 2.8 | 1.1e+01 | 0.2 | 5.3e+01 | 3.8 | 5.7e+01 | 6.7 |
| 320 | $2^{-6}$ | 2.2e−01 | 0.3 | 6.0e−01 | 0.8 | 2.5e+00 | 2.2 | 2.8e+00 | 4.2 | 7.1e+00 | 3.0 |
| 640 | $2^{-7}$ | 2.9e−01 | −0.4 | 5.3e−01 | 0.2 | 6.3e−01 | 2.0 | 1.5e−01 | 4.3 | 4.0e−01 | 4.1 |
| 1280 | $2^{-8}$ | 2.5e−01 | 0.2 | 2.2e−01 | 1.3 | 5.0e−02 | 3.7 | 3.6e−02 | 2.1 | 2.5e−02 | 4.0 |
| 2560 | $2^{-9}$ | 1.6e−01 | 0.6 | 5.6e−02 | 2.0 | 4.9e−03 | 3.4 | 3.5e−03 | 3.4 | 2.8e−04 | 6.4 |
| 5120 | $2^{-10}$ | 9.1e−02 | 0.8 | 1.2e−02 | 2.2 | 8.5e−04 | 2.5 | 2.0e−04 | 4.1 | 1.0e−05 | 4.8 |
| 1.0e+04 | $2^{-11}$ | 4.8e−02 | 0.9 | 2.8e−03 | 2.1 | 1.3e−04 | 2.7 | 1.1e−05 | 4.2 | 3.8e−07 | 4.7 |
| 2.0e+04 | $2^{-12}$ | 2.5e−02 | 1.0 | 6.7e−04 | 2.1 | 1.8e−05 | 2.9 | 6.1e−07 | 4.2 | 1.3e−08 | 4.9 |
| 4.1e+04 | $2^{-13}$ | 1.2e−02 | 1.0 | 1.6e−04 | 2.0 | 2.4e−06 | 2.9 | 3.6e−08 | 4.1 | 1.1e−09 | 3.5 |
| 8.2e+04 | $2^{-14}$ | 6.3e−03 | 1.0 | 4.0e−05 | 2.0 | 3.0e−07 | 3.0 | 2.2e−09 | 4.0 | 1.4e−09 | – |
| 1.6e+05 | $2^{-15}$ | 3.1e−03 | 1.0 | 9.8e−06 | 2.0 | 3.8e−08 | 3.0 | 2.3e−10 | 3.3 | 2.8e−09 | – |

$$d(x) = 4 + 3\cos(2\pi x) \quad \Longrightarrow \quad d_{\min} = 1, \quad d_{\max} = 7.$$

This ratio $d_{\max}/d_{\min} = 7$ exceeds the value that can be stabilized by SBDF (see Remark 10). Using a value of $\eta = 0.1$ in (6.12), and order $r = 5$ yields the ImEx and splitting parameters $(\delta, \sigma) = (0.1732, 2.69)$ (see Fig. 7). Since the unconditional stability region $\mathcal{D}$ becomes smaller as the order $r$ increases, using $r = 5$ automatically guarantees unconditional stability for all orders $1 \le r \le 5$. We manufacture the forcing $f(x,t)$ to generate an exact solution

$$u^*(x,t) = \sin(20t)e^{\sin(2\pi x)}, \tag{6.13}$$

run up to final time $t_f = 5$. The multistep scheme is initialized with the exact data: $u_j = u^*(jk)$ for $j = 0, -1, \ldots, -r+1$. The spatial resolution is $N = 64$.

Table 1 shows the error $\|u - u^*\|_{\infty,h}$, using the discrete maximum norm $\|u\|_{\infty,h} = \max_{1\le j\le N} |u(x_j)|$, capped at $10^{-9}$. Convergence rates for $1 \le r \le 5$ are reported. Note that with $N = 64$, the diffusive time step restriction is $k \le 2^{-18}$. Hence, time steps can be used that are orders of magnitude larger than those required by an explicit scheme. This highlights the benefits of unconditional stability when performing computations with progressively smaller grids.

### *6.3. A nonlinear example: diffusion in porous media and anomalous diffusion rates*

Thus far, the unconditional stability theory has been applied exclusively to linear problems. Now we use the linear theory as a guide for choosing $(\delta, \sigma)$ in nonlinear problems, and numerically demonstrate that the new concepts work. In spirit, the presented methodology shares some similarities with Rosenbrock methods (chapter VI.4, [54]) in that it also avoids nonlinear implicit terms by means of a properly chosen linear implicit term. A key difference — other than the fact that Rosenbrock methods are multistage schemes — is that we do not compute Jacobian matrices (which can be dense and

**Table 2**
Errors $\|\rho - \rho^*\|_{\infty,h}$ for three-dimensional nonlinear diffusion coefficient test case with manufactured solution (6.15) and $(\delta, \sigma) = (0.19166, 13.8)$. Errors are computed at the final time $t_f = 1$, with $64^3$ ($N = 64$) Fourier modes. Cancellation errors, amplified by the problem's conditioning, limit the observed range of fourth and fifth order convergence.

| Num. Steps | $k$ | Error $r=1$ | Rate | Error $r=2$ | Rate | Error $r=3$ | Rate | Error $r=4$ | Rate | Error $r=5$ | Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | $2^{-3}$ | 1.0e+00 | – | 8.3e−01 | – | 6.4e−02 | – | 9.7e−02 | – | 3.4e−04 | – |
| 16 | $2^{-4}$ | 7.7e−01 | 0.4 | 3.8e−01 | 1.1 | 3.4e−02 | 0.9 | 2.2e−02 | 2.1 | 8.1e−04 | −1.2 |
| 32 | $2^{-5}$ | 5.0e−01 | 0.6 | 8.3e−02 | 2.2 | 8.6e−03 | 2.0 | 1.9e−03 | 3.6 | 1.2e−04 | 2.7 |
| 64 | $2^{-6}$ | 2.6e−01 | 0.9 | 1.5e−02 | 2.4 | 1.4e−03 | 2.6 | 1.2e−04 | 4.0 | 7.6e−06 | 4.0 |
| 128 | $2^{-7}$ | 1.3e−01 | 1.0 | 3.6e−03 | 2.1 | 1.9e−04 | 2.8 | 6.6e−06 | 4.2 | 3.0e−07 | 4.7 |
| 256 | $2^{-8}$ | 6.4e−02 | 1.0 | 8.6e−04 | 2.1 | 2.5e−05 | 2.9 | 3.8e−07 | 4.1 | 1.3e−08 | 4.5 |
| 512 | $2^{-9}$ | 3.2e−02 | 1.0 | 2.1e−04 | 2.0 | 3.2e−06 | 3.0 | 2.2e−08 | 4.1 | 6.2e−09 | – |
| 1024 | $2^{-10}$ | 1.6e−02 | 1.0 | 5.2e−05 | 2.0 | 4.0e−07 | 3.0 | 9.8e−10 | 4.5 | 1.2e−08 | – |
| 2048 | $2^{-11}$ | 7.8e−03 | 1.0 | 1.3e−05 | 2.0 | 5.0e−08 | 3.0 | 6.9e−10 | – | 2.4e−08 | – |

time-dependent), but rather always invert a simple constant coefficient matrix determined from the theory. Hence, the new approach offers more flexibility for choosing efficiency-based implicit terms.

We consider a nonlinear model for a gas diffusing into a porous medium [55,56]:

$$\rho_t + \nabla \cdot \big(\boldsymbol{V}\,\rho\big) = 0 \quad \text{(Conservation of mass)}, \qquad \boldsymbol{V} = -\frac{\tilde{\kappa}}{\tilde{\mu}} \nabla p \quad \text{(Darcy's law)}.$$

Here $\tilde{\kappa}$ is the intrinsic permeability of the medium, and $\tilde{\mu}$ is the effective viscosity. Combined with the equation of state $p = p_0\,\rho^{\tilde{\gamma}}$, where $\tilde{\gamma}$ is the adiabatic constant ($\tilde{\gamma} = 5/3$ for an ideal monatomic gas), the porous media equation takes the form:

$$\rho_t = a\,\nabla \cdot \big(\rho^{\tilde{\gamma}}\,\nabla \rho\big), \quad \text{on } \Omega \times (0, T]. \tag{6.14}$$

The constant $a = \kappa\, p_0 \tilde{\gamma}\, \tilde{\mu}^{-1}$ may, without loss of generality, be set to any positive value by re-scaling time.

We discretize (6.14) in three space dimensions using $N^3$ Fourier modes on the periodic domain $\Omega = [0, 1]^3$. Our goal is to achieve unconditional stability by choosing the discrete matrix $\boldsymbol{A}_h \approx \sigma \nabla^2$ proportional to the constant coefficient Laplacian (which is easy to treat implicitly). This approach then avoids an implicit treatment of nonlinear terms, thereby bypassing the need for nonlinear solvers. Due to the nonlinearity in the diffusion coefficient, our choice of $(\delta, \sigma)$ via the formulas (6.12) requires estimates for the maximum and minimum values of the solution $\rho(x, y, z, t)$ over the simulation. At first glance, it may seem troubling to require time stepping parameters based on the solution; however this is not unusual — numerical simulations for nonlinear PDEs often require choosing a time step $k$ that may depend on the solution.

To test the approach for unconditional stability and accuracy, we perform convergence tests of (6.14) with $\tilde{\gamma} = 5/3$ and $a = 1$, using the manufactured solution

$$\rho^*(x, y, z, t) = 2e + e^{\sin(4\pi x)} \cos(2\pi y) \cos(2\pi z) \cos(t). \tag{6.15}$$

We initialize the ImEx scheme with the exact initial data $\rho_j(x, y, z) = \rho^*(x, y, z, jk)$ for $j = 0, -1, \ldots, -r + 1$. We estimate the maximum and minimum value of the nonlinear diffusion coefficient:

$$\max_{\boldsymbol{x}\in\Omega, t\in\mathbb{R}} \rho^*(x, y, z, t)^{\tilde{\gamma}} \le (3e)^{5/3}, \qquad \min_{\boldsymbol{x}\in\Omega, t\in\mathbb{R}} \rho^*(x, y, z, t)^{\tilde{\gamma}} \ge e^{5/3}.$$

Using formulas (6.12) with $d_{\max} = (3e)^{5/3}$, $d_{\min} = e^{5/3}$, $\eta = 0.1$, and $r = 5$ (so that the resulting scheme is stable for all orders 1 through 5), yields: $(\delta, \sigma) = (0.19166, 13.8)$.

Table 2 shows the numerical error $\|\rho - \rho^*\|_{\infty,h}$, using the discrete norm $\|u\|_{\infty,h} = \max_{\boldsymbol{x}\in\text{grid}} |u(\boldsymbol{x})|$ evaluated at the final time $t_f = 1$, using $64^3$ grid points ($N = 64$). In addition to confirming the convergence orders, the table demonstrates that the scheme is stable for $k$-values far larger than required by a fully explicit scheme. Moreover, we have confirmed the observations using other values $\tilde{\gamma} \neq 5/3$ and other manufactured solutions (not shown here).

Now we conduct a test in which the nonlinear behavior is natural to equation (6.14): a decaying/spreading profile without forcing. We use a Gaussian $\rho(x, y, z, 0) = 1 + e^{-\|\boldsymbol{x}-(0.5,0.5,0.5)\|^2/0.15^2}$ as initial data, which is not exactly periodic, but is sufficiently resolved in space using $128^3$ Fourier modes ($N = 128$) to carry out temporal convergence studies. We choose $\tilde{\gamma} = 5/3$, and set $a = 2^{-4}$, which for the given initial data will lead to dynamics that evolve on an $\mathcal{O}(1)$ time scale. Using (6.12) with $d_{\max} = 2^{\tilde{\gamma}}$, $d_{\min} = 1$, $\eta = 0.1$, and $r = 3$, we obtain $(\delta, \sigma) = (0.794, 2.616)$. To generate the initial data required to start the high-order multistep methods, we use the low order ($r = 1$ and $r = 2$) unconditionally stable schemes with many subgrid time steps.

Table 3 provides a verification for the test, by computing the convergence rate estimate

$$R_k := \log_2 \big(\|\rho_{4k}(t_f) - \rho_{2k}(t_f)\|_{\infty,h} / \|\rho_{2k}(t_f) - \rho_k(t_f)\|_{\infty,h}\big), \tag{6.16}$$

**Table 3**
Convergence test for a decaying Gaussian initial data, reporting the values $R_k$ from (6.16) at $t_f = 1$, and $128^3$ grid points.

| $k$ | $2^{-4}$ | $2^{-5}$ | $2^{-6}$ | $2^{-7}$ | $2^{-8}$ | $2^{-9}$ | $2^{-10}$ | $2^{-11}$ | $2^{-12}$ |
|---|---|---|---|---|---|---|---|---|---|
| $r = 1$, $R_k$ | 1.22 | 1.58 | 1.52 | 1.28 | 1.14 | 1.07 | 1.03 | 1.02 | 1.01 |
| $r = 2$, $R_k$ | 0.39 | 5.12 | 2.06 | 2.02 | 1.96 | 1.95 | 1.96 | 1.97 | 1.98 |
| $r = 3$, $R_k$ | 1.38 | 3.54 | 1.63 | 8.52 | 2.83 | 2.74 | 2.81 | 2.88 | 2.93 |

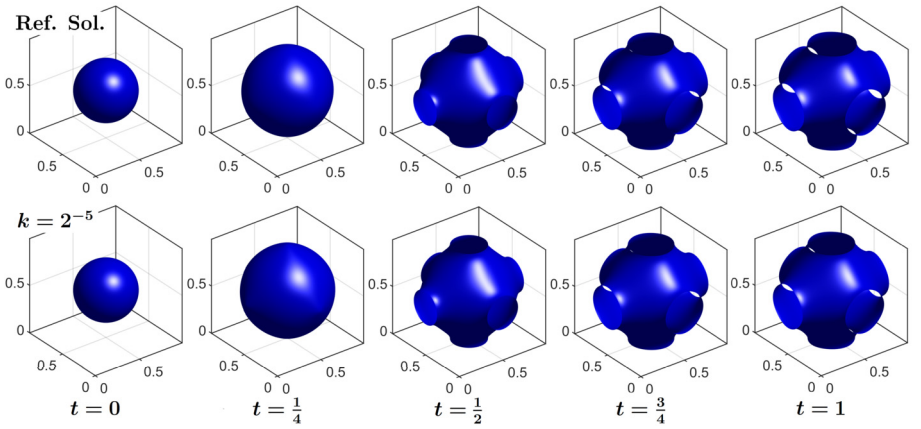


**Fig. 8.** Evolution of the solution towards $t_f = 1$ for a Gaussian initial data. Snapshots are of the level sets $\{\boldsymbol{x} : \rho(\boldsymbol{x}, t) = \bar{\rho}\}$ for the reference solution (top), compared to the solution computed using order $r = 3$ with $(\delta, \sigma) = (0.794, 2.616)$ and a large time step $k = 2^{-5}$ (bottom). Here $\bar{\rho}$ is the discrete average value of the solution (which is a conserved quantity).

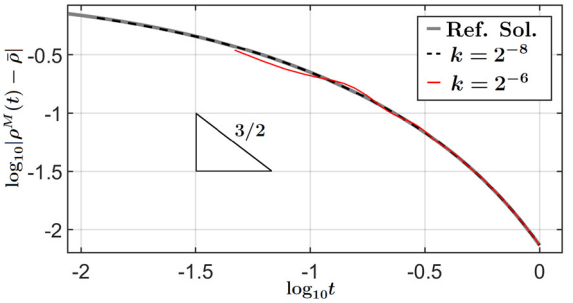


**Fig. 9.** Decay of the maximum solution $\rho^M(t)$ to a constant. For reference, the slope of $-3/2$ is shown — which is the decay rate for Gaussian initial data obtained with a constant coefficient diffusion equation. The simulation uses the order $r = 3$ scheme with $(\delta, \sigma) = (0.794, 2.616)$, and $k = 2^{-6}$ (64 time steps) as well as $k = 2^{-8}$ (256 time steps). Both are much larger than the restriction $k = 2^{-16}$ (65536 time steps) required by the fully explicit reference solution.

where $\rho_k(t_f)$ denotes the discrete solution at $t_f$ computed using time step $k$. Table 3 confirms that at small $k$ values, $R_k$ converges to the order of the scheme.

Finally, guided by the data in Table 3, we choose a time step $k = 2^{-5}$, and compute the solution towards $t_f = 1$ (i.e. merely 32 time steps). Fig. 8 visualizes level sets of the solution for different times and compares them to a reference solution (obtained using the fully explicit, second-order scheme with $\delta = 1$, i.e. Adams–Bashforth, with $k = 2^{-16}$, i.e. 65536 time steps). The unconditionally stable method is successfully capturing the solution, albeit using very large time steps.

Meanwhile, Fig. 9 highlights the nonlinear effect (anomalous diffusion) in the solution with plots of the decay rate of the peak $\rho^M(t) = \|\rho(\boldsymbol{x}, t)\|_{\infty,h}$ relative to the mean value $\bar{\rho} := h^3 \sum_{\boldsymbol{x} \in \text{grid}} \rho(\boldsymbol{x}, 0) \approx \int_0^1 \int_0^1 \int_0^1 \rho(\boldsymbol{x}, 0)\, \mathrm{d}\boldsymbol{x}$. For reference, the decay rate for a linear constant coefficient diffusion problem $(-3/2)$ is shown as well. The plot shows, against the reference solution, the ImEx schemes with $k = 2^{-6}$ (64 time steps, a small error) and $k = 2^{-8}$ (256 time steps, visually indistinguishable).

## 7. Incompressible channel flow

In this section we perform a study that zooms in on the question of unconditional stability for ImEx splittings that arise in fluid dynamics problems. Specifically, for the time-dependent Stokes equation in a channel geometry, we devise unconditionally stable ImEx schemes that treat the pressure explicitly and viscosity implicitly. High-order ImEx schemes

that are provably unconditionally stable for the incompressible Navier–Stokes equations are notoriously difficult to attain (see for instance [57,23,27]). This study highlights some peculiar challenges that arise with ImEx schemes for incompressible flows, for instance that unconditional stability may depend on model parameters such as the shape and size of the domain. The new unconditional stability theory may provide new ways to stabilize operator splittings in fluid dynamics applications that might otherwise be unstable.

It is worth mentioning that in fluid flow, the alternative to unconditional stability may not necessarily be detrimental (in fact, the situation here is of that type). For instance, when unconditional stability is not attained, an ImEx approach may still provide competitive stability benefits, by incurring a time step restriction that is $\mathcal{O}(1)$ (i.e. independent on the spatial mesh), and thus not stiff. This type of stability restriction has been recently referred to as *quasiunconditional stability* [29,30] (and has been applied to the compressible Navier–Stokes equations in an alternate direction implicit (ADI) setting). Nonetheless the question of whether unconditionally stable schemes can be devised is of interest, as in other situations the lack of unconditional stability may not be as forgiving as the quasiunconditional stability scenario.

We focus on reformulations of the incompressible Navier–Stokes equations [58,59,22,27,60], that take the form of a pressure Poisson equation (PPE) system (sometimes also referred to as an *extended Navier–Stokes* system). They replace the divergence-free constraint by a non-local pressure operator (defined via the solution of a Poisson equation), and thus allow for the application of time-stepping schemes without having to worry about constraints. In contrast to projection methods [22,27,58,60], PPE systems are not based on fractional steps, and thus allow in principle for arbitrary order in time. In turn, the use of ImEx schemes allows for an explicit treatment of the pressure — which (in contrast to fully implicit time-stepping) avoids large saddle-point problems in which velocity and pressure are coupled together. The challenge is that the explicit pressure term may become stiff (because it can be recast as a function of the viscosity term [22,60]).

For simplicity, we restrict this presentation to the linear Navier–Stokes equations (i.e. without the advection terms), because these equations already capture the key challenges arising from the interaction between viscosity and pressure. One could also investigate (unconditional) stability for incompressible flows with advection terms; however we do not pursue this here. For a two dimensional domain $\Omega \subset \mathbb{R}^2$, we use the PPE reformulation by Johnston and Liu [22,59] for problems with no-slip boundary conditions:

$$\left.\begin{aligned} u_t &= u_{xx} + u_{yy} - p_x + f_1 \quad \text{for } \mathbf{x} \in \Omega, \\ v_t &= v_{xx} + v_{yy} - p_y + f_2 \quad \text{for } \mathbf{x} \in \Omega, \\ & u = v = 0 \qquad\qquad\quad \text{for } \mathbf{x} \in \partial\Omega, \end{aligned}\right\} \tag{7.1}$$

where $p$ is the solution of

$$\left.\begin{aligned} p_{xx} + p_{yy} &= (f_1)_x + (f_2)_y \qquad\qquad\quad \text{for } \mathbf{x} \in \Omega, \\ \boldsymbol{n} \cdot \nabla p &= -\boldsymbol{n} \cdot (\nabla \times \nabla \times \boldsymbol{u}) + \boldsymbol{n} \cdot \boldsymbol{f} \quad \text{for } \mathbf{x} \in \partial\Omega, \\ \textstyle\int_\Omega p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} &= 0 \end{aligned}\right\}. \tag{7.2}$$

Equation (7.1) is the standard momentum equation for the velocity field $(u, v)$, while equation (7.2) is the PPE reformulation for the pressure, acting to keep the flow incompressible. Note that the last requirement in equation (7.2) is added to uniquely define the pressure. Here, $\boldsymbol{n}$ is the outward facing normal on $\partial\Omega$, and $\boldsymbol{f} = (f_1, f_2)$ is the body force. The viscosity terms $u_{xx} + u_{yy}$ and $v_{xx} + v_{yy}$ are the stiff terms that are treated implicitly (matrix $\boldsymbol{A}$), while the pressure terms $p_x, p_y$ are linear functions of the velocity $(u, v)$ that are treated explicitly (matrix $\boldsymbol{B}$). We consider a channel geometry, $\Omega = [0, L_x] \times (0, 1)$, that is periodic in the $x$-direction.

The simple geometry allows us to solve for the pressure $p$ analytically, and convert equations (7.1)–(7.2) into a non-local PDE for the velocity $u$. This simplifies the computation of the set $W_p(\boldsymbol{A}, \boldsymbol{B})$ and provides fundamental insight into why existing ImEx splittings that treat the viscosity terms implicitly and pressure terms explicitly may become unstable. In more general problems, one would of course need to conduct a full spatial discretization of equations (7.1)–(7.2), and apply the recipes described in §4 to the resulting large ODE system. It is important to note that, while theoretical insights are less clear in that situation, there is no fundamental problem with applying the methodology.

To derive the non-local PDE for the velocity $u$, we start off by setting $\boldsymbol{f} = 0$ (unconditional stability does not depend on the forcing). Because $\Omega$ is periodic in the $x$-direction, we conduct a Fourier expansion in the $x$-direction and set $(u, v) = (\imath u(y, t; \xi), v(y, t; \xi))e^{\imath \xi x}$ and $p = p(y; \xi)e^{\imath \xi x}$. The system (7.1)–(7.2) then becomes:

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \left(\frac{\partial^2}{\partial y^2} - \xi^2\right)\begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} \xi p \\ p_y \end{pmatrix} \text{ on } 0 < y < 1\,, \quad \text{and } \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{0} \text{ on } y = \{0, 1\}, \tag{7.3}$$

$$p_{yy} - \xi^2 p = 0 \text{ on } 0 < y < 1\,, \quad \text{and } \frac{dp}{dy} = \xi u_y \text{ on } y = \{0, 1\}. \tag{7.4}$$

The allowable wave numbers are given by $\xi = \pm 2\pi L_x^{-1} n_\xi$ and natural numbers $n_\xi \in \mathbb{N}$. The pressure equation (7.4) is uniquely solvable for all $\xi \neq 0$; while for $\xi = 0$, the integral constraint in (7.2), together with (7.4), fixes $p(y; 0) = 0$. For $\xi \neq 0$, equation (7.4) can be solved analytically to obtain:

$$p(y;\xi) = -\frac{\cosh\big(\xi(y-1)\big)}{\sinh(\xi)}\, u_y(0) + \frac{\cosh(\xi y)}{\sinh(\xi)}\, u_y(1). \quad \xi \neq 0. \tag{7.5}$$

The pressure can then be substituted back into equation (7.3) to yield a non-local PDE for the horizontal velocity $u = u(y,t;\xi)$ (for $\xi \neq 0$):

$$u_t = \Big(\frac{\partial}{\partial y^2} - \xi^2\Big)u + \xi\, \frac{\cosh\big(\xi(y-1)\big)}{\sinh(\xi)}\, u_y(0) - \xi\, \frac{\cosh(\xi y)}{\sinh(\xi)}\, u_y(1) \tag{7.6}$$
$$\text{Boundary conditions:} \quad u = 0, \text{ on } y = \{0,1\}.$$

Solving equation (7.6) for $u(y,t;\xi)$ at every wave number $\xi$ then allows one to reconstruct $u(x,y,t)$. In a similar fashion, one can use (7.5) to reconstruct $p(x,y)$. Once either $u(x,y,t)$ or $p(x,y)$ is known, the vertical velocity $v(x,y,t)$ can then be obtained by solving either (i) the $v$-component of equation (7.3) with the pressure as a prescribed forcing, or (ii) using the fact that the PPE reformulation automatically enforces the divergence constraint so that $v(y;\xi)_y = -\xi u(y;\xi)$. Collectively, the solutions $(u,v,p)$ from (7.5)–(7.6) solve the original PDEs (7.1)–(7.2) with $\boldsymbol{f} = 0$. Thus, devising unconditionally stable ImEx splittings for (7.6) can be used as a guide for stabilizing discretizations of the full equations (7.1)–(7.2).

### 7.1. Numerical discretization and ImEx splitting for equation (7.6)

In line with prior examples, we seek ImEx splittings of equation (7.6) in which the pressure terms are treated explicitly, while a portion of the viscosity is treated implicitly:

$$\boldsymbol{A}_{h,\xi}\, \boldsymbol{u} \approx \sigma\Big(\frac{\partial}{\partial y^2} - \xi^2\Big)u, \quad \boldsymbol{B}_{h,\xi}\, \boldsymbol{u} \approx (1-\sigma)\Big(\frac{\partial}{\partial y^2} - \xi^2\Big)u - \xi p\,. \tag{7.7}$$

Although equation (7.6) is a non-local PDE, the highest derivative degree is 2 so that the splitting (7.7) still adheres to the guidelines in Remark 6. As with prior discussions, the inclusion of the splitting parameter $\sigma$ in (7.7) provides additional flexibility in devising stable schemes, compared to many existing splittings that effectively fix $\sigma = 1$.

To obtain the matrices $(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$ we discretize the $y$-direction using $N_y$ equispaced grid points $y_j = jh$, and spacing $h = (N_y+1)^{-1}$, so that $\boldsymbol{u}_j = u(y_j) \in \mathbb{R}^{N_y}$, for $1 \leq j \leq N_y$. A standard 3-point finite difference stencil for $\partial_{yy}$ (with Dirichlet boundary conditions $u(0) = u(1) = 0$ at $y \in \{0,1\}$) leads to the following discretization: $\boldsymbol{A}_{h,\xi} = \sigma \boldsymbol{A}_{0,h,\xi}$, where

$$\boldsymbol{A}_{0,h,\xi} = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} - \xi^2 \boldsymbol{I}. \tag{7.8}$$

In a similar fashion, we set $\boldsymbol{B}_{h,\xi} = (1-\sigma)\boldsymbol{A}_{0,h,\xi} + \boldsymbol{Q}_{h,\xi}$ where $\boldsymbol{Q}_{h,\xi}\boldsymbol{u} \approx -\xi p$ is a matrix that computes the pressure. The matrix $\boldsymbol{Q}_{h,\xi}$ is built using equation (7.5) as

$$\boldsymbol{Q}_{h,\xi} = \boldsymbol{a}(\xi)\, \boldsymbol{d}_1^T + \boldsymbol{b}(\xi)\, \boldsymbol{d}_2^T \quad \text{for} \quad \xi \neq 0, \quad \text{and} \quad \boldsymbol{Q}_{h,0} = \boldsymbol{0} \quad \text{for} \quad \xi = 0. \tag{7.9}$$

Here the vectors $\boldsymbol{d}_1 = h^{-1}\begin{pmatrix}1, & 0, & \ldots, & 0\end{pmatrix}^T$ and $\boldsymbol{d}_2 = h^{-1}\begin{pmatrix}0, & \ldots, & 0, & -1\end{pmatrix}^T$ approximate the derivatives $\boldsymbol{d}_1^T\boldsymbol{u} \approx u_y(0)$ and $\boldsymbol{d}_2^T\boldsymbol{u} \approx u_y(1)$, and thus encode the boundary conditions $u(0) = u(1) = 0$. The vectors $\boldsymbol{a}(\xi), \boldsymbol{b}(\xi) \in \mathbb{R}^{N_y}$ are discretizations of the functions

$$\boldsymbol{a}(\xi)_j = \xi\, \mathrm{csch}(\xi) \cosh\big(\xi(y_j - 1)\big), \quad \boldsymbol{b}(\xi)_j = -\xi\, \mathrm{csch}(\xi) \cosh(\xi y_j).$$

Note that for each fixed value of $\xi$, the matrix $\boldsymbol{Q}_{h,\xi}$ is the sum of two rank-1 matrices.

### 7.2. Applying the unconditional stability theory to determine $(\delta, \sigma)$

Now we follow the guidelines in Remark 6 to determine $(\delta, \sigma)$. We directly focus on computing an appropriate set $W_p(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$ to be used in the stability theory.

Unlike prior examples, for the channel flow application considered here, the choice $p = 2$ is most useful for the analysis of the $W_p$ sets. This is because the maximum size and shape of the sets $W_2(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$ are effectively independent of $h$. In contrast, numerical experiments show that the sets $W_1(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$ arising from (7.7) tend to grow as $h \to 0$. It is worth noting that $p = 2$ is also motivated by [22], in which a version of the set $W_2(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$ was studied to prove that SBDF1 is unconditionally stable when applied to the channel flow PDEs (7.1)–(7.2). Here, we apply the full new unconditional stability theory to systematically investigate stability for high order schemes.

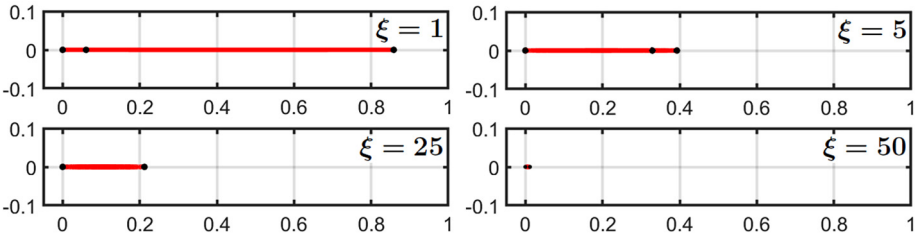


**Fig. 10.** Sets $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$ (red) in (7.10) for $N_y = 32$ and wave numbers $\xi \in \{1, 5, 25, 50\}$. The sets are computed numerically and are confined to the real axis. The sets shrink in size as $\xi$ increases.

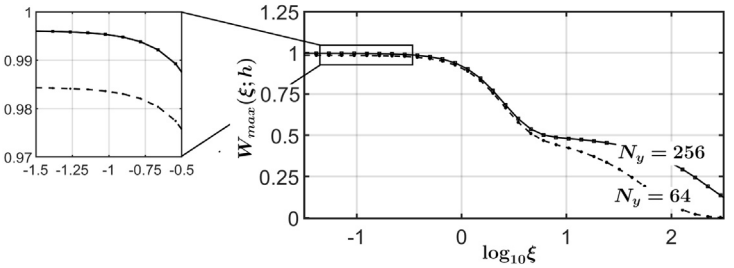


**Fig. 11.** Plot of $W_{max}(\xi; h)$ versus wave number $\xi$ for $N_y = 64$ (dashed) and $N_y = 256$ (solid). The sets $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$ (and consequently $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$) are contained in the interval $[0, 1]$ along the real axis.

To compute $W_2(\boldsymbol{A}_{h,\xi}, \boldsymbol{B}_{h,\xi})$, we use the scaling property from Remark 5 to write

$$W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi}) = 1 - \sigma^{-1} + \sigma^{-1} W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi}), \tag{7.10}$$

and use Chebfun's numerical range (field of values) routine [53] to compute $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$; from which $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ is obtained via a shift and re-scaling. Note that Chebfun employs an algorithm due to Johnson [51] that reduces the computation of $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ to a collection of eigenvalue computations.

A theoretical study of the set $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$ where continuum operators $\mathcal{A} = (-\Delta\boldsymbol{u})$ and $\mathcal{B} = (-\nabla p)$ were used instead of discrete matrices $\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi}$ (note that the set $W_2$ is still defined using operators), was carried out in part of the work [22]. They showed that the set $W_2(\mathcal{A}, \mathcal{B})$, using continuum operators, was real and contained in the interval $[0, 1]$. Our numerical computations of $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ also show that for each fixed value of $\xi$, the sets are within a discretization error (at most $\mathcal{O}(h)$) of the interval $[0, 1]$.

To quantify the region that $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$ occupies along the real axis, let

$$W_{max}(\xi; h) = \max \mathrm{Re}\Big(W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})\Big), \quad W_{min}(\xi; h) = \min \mathrm{Re}\Big(W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})\Big).$$

Fig. 10 plots the sets $W_2(\boldsymbol{A}_{0,h,\xi}, \boldsymbol{Q}_{h,\xi})$ for $\xi \in \{1, 5, 25, 50\}$ and grid spacing $N_y = 32$. Fig. 11 plots $W_{max}(\xi; h)$ for different wave numbers $\xi$ (where $\xi = 0$ is a special case and excluded from the plot) and grids $N_y \in \{64, 256\}$. The plot shows that $W_{max}(\xi; h)$ decreases (monotonically) with increasing $\xi$ – which is important for the simultaneous stabilization of all wave numbers that arise in a channel geometry. The minimum value $W_{min}(\xi; h)$ is always zero, i.e. $W_{min}(\xi; h) = 0$.

Since the sets $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ lie along the real axis, the procedure for choosing $(\delta^*, \sigma^*)$ parallels that of the diffusion example in §6.2. For instance, we can use equations (6.10) (for $r \in \{1, 2\}$) and (6.12) (for $r \in \{3, 4, 5\}$) via the substitutions $d_{\max} \to (1 - W_{min}(\xi; h)) = 1$ and $d_{\min} \to (1 - W_{max}(\xi; h))$ to ensure that $W_2(\sigma\boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi}) \subseteq \mathcal{D}$. This approach determines the parameters $(\delta^*, \sigma^*)$ that ensure unconditional stability for the PDE (7.9) for one fixed wave number $\xi$. In practice, however, stabilizing equation (7.6) for a channel geometry requires that it be unconditionally stable for all allowable wave numbers $\xi = \pm 2\pi L_x^{-1} n_x$. The following remark shows that it suffices to ensure unconditional stability for the smallest non-zero wave number (which then stabilizes all others).

**Remark 12.** (Choosing one $(\delta^*, \sigma^*)$ that works for all wave numbers $\xi$) Unconditional stability for equations (7.1)–(7.2) corresponds to ensuring that (7.6) is stable for all modes $\xi = \pm 2\pi L_x^{-1} n_\xi$. Here we argue why it suffices to ensure that only the smallest mode is stable, i.e., to require that $W_2(\sigma\boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1}) \subseteq \mathcal{D}$, where $\xi_1 = 2\pi L_x^{-1}$ is the smallest positive wave number.

To show this, we use the simple property that the sets $W_2(\cdot, \cdot)$ may be written as a numerical range. First, one can view the simultaneous solution of equation (7.6) over all allowable $\xi$, as solving one very large system of equations with matrices $\boldsymbol{A}_{0,h}$ and $\boldsymbol{B}_h$ that are written as direct sums. Specifically, write $\boldsymbol{A}_{0,h} = \bigoplus_\xi \boldsymbol{A}_{0,h,\xi}$ and $\boldsymbol{B}_h = \bigoplus_\xi \boldsymbol{B}_{h,\xi}$, where the direct sum is over the wave numbers $\xi = \pm 2\pi L_x^{-1} n_x$ and natural numbers $n_\xi \in \mathbb{N}$ (i.e. $(\boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$ are infinite block di-

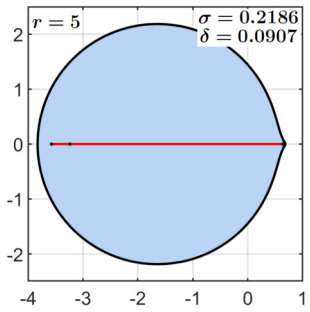


**Fig. 12.** Unconditional stability for equation (7.6) and channel length $L_x = 2\pi$. The set $W_2(\boldsymbol{A}_h, \boldsymbol{B}_h)$ (shown in red) is contained within the unconditional stability region (blue set) when $(\delta, \sigma) = (0.0907, 0.2186)$.

agonal matrices with each block being $\boldsymbol{A}_{0,h,\xi}$ or $\boldsymbol{B}_{h,\xi}$). Now use the fact that the set $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ may be written as a numerical range (see Remark 2) — and that the numerical range of the direct sum of two matrices is the convex hull of their two numerical ranges, i.e. $W(\boldsymbol{X} \bigoplus \boldsymbol{Y}) = \text{conv}\{W(\boldsymbol{X}), W(\boldsymbol{Y})\}$. As a result, the set $W_2(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h)$ is the convex hull of the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ over all allowable wave numbers. Lastly, we observe that the set $W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$ contains (up to at most an error $\mathcal{O}(h)$) each of the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ for all $\xi = \pm 2\pi L_x^{-1} n_x$ and $n_x \in \mathbb{N}$. This shows that the convex hull of the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ over all $\xi$ is approximately equal to the set $W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$. Specifically:

- For the mode $\xi = 0$, we have $\boldsymbol{Q}_h = \boldsymbol{0}$. Hence $W_2(\sigma \boldsymbol{A}_{0,h,0}, \boldsymbol{B}_{h,0}) = \{1 - \sigma^{-1}\}$ is a single point contained in the set $W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$.
- The matrices $\boldsymbol{Q}_{h,-\xi} = \boldsymbol{Q}_{h,\xi}$ and $\boldsymbol{A}_{0,h,-\xi} = \boldsymbol{A}_{0,h,\xi}$ are even functions of $\xi$. Hence the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi})$ are the same for both $\pm\xi$ values.
- Fig. 11 shows that the value $W_{max}(\xi; h)$ is a decreasing function of $\xi$. Hence, using formula (7.10) along with the definitions of $W_{max}(\xi; h)$ (and the fact that $W_{min}(0; h) = 0$), one has $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{h,\xi}) \subseteq W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$ whenever $\xi > \xi_1$.

Hence we have that $W_2(\sigma \boldsymbol{A}_{0,h}, \boldsymbol{B}_h) \approx W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$, where the $\approx$ sign (as opposed to an $=$ sign) denotes the fact that there may be an $\mathcal{O}(h)$ error.

Based on this important insight, we now choose $(\delta^*, \sigma^*)$ for a channel geometry of length $L_x = 2\pi$ and smallest positive wave number $\xi_1 = 1$. For grid size $N_y = 256$, the set $W_2(\sigma \boldsymbol{A}_{0,h,\xi_1}, \boldsymbol{B}_{h,\xi_1})$ is obtained by inserting the values $W_{min}(1; h) = 0$ and $W_{max}(1; h) = 0.93$ (see Fig. 11) into equation (7.10). A crucial observation is that the largest and smallest generalized eigenvalues $\Lambda(\boldsymbol{A}_{0,\xi_1;h}, \boldsymbol{Q}_{\xi_1;h})$ equal the maximum and minimum values of $W_{max}(\xi_1; h)$, see Fig. 11 — and the gap between these generalized eigenvalues exceeds the unconditional stability capabilities of SBDF3 (see Remark 10). Hence, the new coefficients (i.e. $\delta < 1$) must be used to achieve unconditional stability. Using a gap parameter of $\eta = 0.1$, order $r = 5$, and values $d_{min} \to 1 - 0.93 = 0.07$, $d_{max} \to 1$ in equation (6.12), leads to $(\delta, \sigma) = (0.0907, 0.2186)$. Fig. 12 verifies that this choice in fact satisfies the sufficient conditions for unconditional stability.

We conclude this section with a few important observations. For a fixed value of $\xi$, the maximum value $W_{max}(\xi; h)$ remains bounded below 1 as $h \to 0$. This implies that one value of $(\delta, \sigma)$ may be used to stabilize an entire family of splittings. On the other hand, the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{\xi,h})$ do depend on $\xi$ — and Fig. 11 implies that the sets $W_2(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{\xi,h})$ (and also the generalized eigenvalues $\Lambda(\sigma \boldsymbol{A}_{0,h,\xi}, \boldsymbol{B}_{\xi,h})$) become large when $\xi_1 \to 0$. This observation is important because it implies that designing unconditionally stable schemes requires a choice of $(\delta, \sigma)$ that depends on the domain size $L_x$, and also on the fact that the domain is a channel geometry.

A natural question then arises: what values can $W_2(\boldsymbol{A}_h, \boldsymbol{B}_h)$ take for a general fluid dynamics problem? Because the set may depend on the geometry shape (for instance whether $\Omega$ has corners, see [61]) and size of the computational domain, one would expect that numerical computations may be required to determine or estimate $W_2(\boldsymbol{A}_h, \boldsymbol{B}_h)$. For instance, one may perform a few rapid computations of $W_2(\boldsymbol{A}_h, \boldsymbol{B}_h)$ using a coarse mesh (i.e. large $h$) and thus small matrices $\boldsymbol{A}_h$ and $\boldsymbol{B}_h$, to obtain a guide for determining the parameters $(\delta, \sigma)$ for the fully resolved problem.

## 8. Conclusions and outlook

With this work on unconditionally stable ImEx multistep methods we wish to stress two key messages: first, we advocate to conduct the selection of the ImEx splitting and selection of the time-stepping scheme in a simultaneous fashion; and second, it is often possible to achieve unconditional stability in significantly more general settings than one might think at first glance.

The examples and applications discussed herein may serve as a blueprint for how to approach many other types of problems, by using the new stability theory and new ImEx schemes to determine feasible and optimal parameters $(\delta, \sigma)$ that characterize the scheme and splitting, respectively.

The theoretical foundations of this work establish necessary and sufficient conditions for unconditional stability, resulting from unconditional stability diagrams (that depend only on the scheme) and computable matrix quantities (that depend only on the ImEx splitting). This analysis is then used to explain fundamental limitations of the popular SBDF schemes. In particular, it is shown why SBDF can frequently not be extended beyond first or second order — and how the new schemes can overcome this barrier.

The variable coefficient and nonlinear diffusion examples highlight the practical impact that the new methodology can bring: being able to treat problems whose stiff terms are challenging to invert, without stiff time step restrictions *and* without having to conduct challenging solves. In addition, the theory serves to provide some fundamental insight into unconditional stability (or breakdown thereof) in incompressible fluid flow simulations.

A key limitation of this work is the formal restriction to positive definite matrices $\boldsymbol{A}$. This excludes many splittings that would be warranted for intrinsically non-symmetric problems, such as advection or dispersion. Regarding this limitation, it should first be noted that much of the theory persists when the assumptions on $\boldsymbol{A}$ are relaxed (for instance, $\boldsymbol{A}$ may be a normal matrix with eigenvalues $\lambda$ having complex arguments $|\arg(-\lambda)|$ that are not too large). Second, the extension of the theory to truly non-symmetric $\boldsymbol{A}$ is an important subject of future work.

**Acknowledgements**

The authors wish to acknowledge support by the National Science Foundation through grants DMS–1719640 (B. Seibold and D. Zhou) and DMS–1719693 (D. Shirokoff). D. Shirokoff was supported by a grant from the Simons Foundation (#359610).

## Appendix A. Proof of Proposition 6.1

Throughout this section we suppress the subscript $h$ on the matrices $(\boldsymbol{A}_h, \boldsymbol{B}_h)$, and simply write $(\boldsymbol{A}, \boldsymbol{B})$. We start with computing the set

$$W_1(\boldsymbol{A}, \boldsymbol{B}) = \left\{ \langle \boldsymbol{v}, \boldsymbol{B}\boldsymbol{v} \rangle : \langle \boldsymbol{v}, (-\boldsymbol{A})\boldsymbol{v} \rangle = 1, \boldsymbol{v} \in \mathbb{V} \right\} \tag{A.1}$$

$$= \left\{ \langle \boldsymbol{D}\boldsymbol{v}, \big(\sigma \boldsymbol{I} - \mathrm{diag}(\boldsymbol{d})\big)\boldsymbol{D}\boldsymbol{v} \rangle : \langle \boldsymbol{D}\boldsymbol{v}, \boldsymbol{D}\boldsymbol{v} \rangle = \sigma^{-1}, \boldsymbol{v} \in \mathbb{V} \right\} \tag{A.2}$$

In the expression in (A.2), we have used the fact that the derivative matrix is skew-symmetric $\boldsymbol{D}^\dagger = \overline{\boldsymbol{D}}^T = -\boldsymbol{D}$. Note that $\boldsymbol{D}$ is invertible on the space $\mathbb{V}$ (i.e. $\mathbb{V}$ is orthogonal to $\mathbf{1}$ — which is the nullspace of $\boldsymbol{D}$). Making the change of variables $\boldsymbol{y} = \sigma^{\frac{1}{2}} \boldsymbol{D}\boldsymbol{v}$ in (A.2), we observe that as $\boldsymbol{v}$ varies over $\mathbb{V}$, $\boldsymbol{y}$ varies over $\mathbb{V}$. This yields:

$$W_1(\boldsymbol{A}, \boldsymbol{B}) = \left\{ \langle \boldsymbol{y}, \big(\boldsymbol{I} - \sigma^{-1}\mathrm{diag}(\boldsymbol{d})\big)\boldsymbol{y} \rangle : \|\boldsymbol{y}\| = 1, \boldsymbol{y} \in \mathbb{V} \right\} = 1 - \sigma^{-1} \sum_{j=1}^{N} d(x_j) |y_j|^2,$$

where $\boldsymbol{y} = \big(y_1, y_2, \ldots, y_N\big)^T$. Since $\|\boldsymbol{y}\|^2 = 1$, each value $|y_j|^2$ is real and confined to the region $0 \le |y_j|^2 \le 1$ (note that because $\boldsymbol{y} \in \mathbb{V}$, not all vectors $\boldsymbol{y}$ are allowed — only those having zero mean). Combining the results leads to the following inequality:

$$d_{\min} = d_{\min} \sum_{j=1}^{N} |y_j|^2 \le \sum_{j=1}^{N} d(x_j) |y_j|^2 \le d_{\max} \sum_{j=1}^{N} |y_j|^2 = d_{\max}.$$

Hence, the set $W_1(\boldsymbol{A}, \boldsymbol{B})$ is real and bounded by:

$$1 - \sigma^{-1} d_{\max} \le W_1(\boldsymbol{A}, \boldsymbol{B}) \le 1 - \sigma^{-1} d_{\min}.$$

This concludes the first part of the proof. To prove the eigenvalue bounds on $\Lambda(\boldsymbol{A}, \boldsymbol{B})$, the upper bound estimates (i.e. the bounds overestimating the largest $\mu$ and underestimating the smallest $\mu$) follow directly from using the established bounds on $W_1(\boldsymbol{A}, \boldsymbol{B})$ with the fact that $\Lambda(\boldsymbol{A}, \boldsymbol{B}) \subseteq W_1(\boldsymbol{A}, \boldsymbol{B})$. Thus, it suffices to prove only the lower bounds. We are interested

in bounding the eigenvalues of $\mu(-\boldsymbol{A})\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}$ with eigenvectors $\mathbf{1}^T\boldsymbol{v} = 0$ restricted to $\mathbb{V}$. Substituting $\boldsymbol{A}$ and $\boldsymbol{B}$ into the eigenvalue equation yields:

$$\mu \boldsymbol{D}^2\boldsymbol{v} = \Big(\boldsymbol{D}\boldsymbol{S}\boldsymbol{D}\Big)\boldsymbol{v}, \quad \text{where} \quad \boldsymbol{S} := (\boldsymbol{I} - \sigma^{-1}\mathrm{diag}(\boldsymbol{d})).$$

As before, let $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{v}$, which is an invertible transformation on $\mathbb{V}$. Then

$$\boldsymbol{D}\Big(\boldsymbol{S}\boldsymbol{y} - \mu\boldsymbol{y}\Big) = 0, \quad \boldsymbol{y}^T\mathbf{1} = 0,$$

or alternatively

$$(\boldsymbol{S} - \mu\boldsymbol{I})\boldsymbol{y} = \alpha\mathbf{1}, \quad \boldsymbol{y}^T\mathbf{1} = 0, \tag{A.3}$$

for some $\alpha \in \mathbb{C}$. We now solve equation (A.3) for two separate cases:

**Case 1: $\boldsymbol{S}\mathbf{1} = \mathbf{0}$.** Here $\sigma = d(x_j)$ for all $d(x_j)$. This is only possible if $d(x_j) = d_0$ for all $j$, i.e. one has a constant coefficient diffusion, and thus $\boldsymbol{S} = \mathbf{0}$ identically. Dotting (A.3) through by $\mathbf{1}$ further shows that $\alpha = 0$, thereby forcing all $\mu = 0$. Hence, Proposition 6.1 is satisfied (trivially) because $\mu_{\max} \geq 1 - \sigma^{-1}d_{2,\min} = 0$ and $\mu_{\min} \leq 1 - \sigma^{-1}d_{2,\max} = 0$.

**Case 2: $\boldsymbol{S}\mathbf{1} \neq \mathbf{0}$.** In this case, we solve equations (A.3) by first writing the components of $\boldsymbol{y}$ in terms of the unknown eigenvalue $\mu$:

$$y_j = \frac{\alpha}{1 - \sigma^{-1}d(x_j) - \mu}.$$

Applying the constraint $\mathbf{1}^T\boldsymbol{y} = 0$ to the vector $\boldsymbol{y}$, shows that the eigenvalues $\mu$ are roots to the following equation:

$$g(\mu) = 0, \quad \text{where} \quad g(\mu) := \sum_{j=1}^{N} \big(1 - \sigma^{-1}d(x_j) - \mu\big)^{-1}.$$

Ordering the poles of $g(\mu)$ along the real axis from smallest to largest shows that there is at least one root of $g(\mu)$ between the smallest two values (or largest two values) of $(1 - \sigma^{-1}d(x_j))$. Hence, Proposition 6.1 follows.

## Appendix B. Formulas for the ImEx coefficients

**Table B.4**
ImEx coefficients for orders 1–5 as functions of $\delta$. To use, choose an order and determine a value $0 < \delta \leq 1$ small enough to ensure the splitting of choice $(\boldsymbol{A}, \boldsymbol{B})$ is unconditionally stable. Substitute this value $\delta$ into the table to obtain the time stepping coefficients. Coefficients reduce to SBDF when $\delta = 1$.

| Order | | $j=3$ | $j=2$ | $j=1$ | $j=0$ |
|---|---|---|---|---|---|
| 1 | $a_j$ | . | . | $\delta$ | $-\delta$ |
| | $c_j$ | . | . | 1 | ($\delta$-1) |
| | $b_j$ | . | . | 0 | $\delta$ |

| Order | | $j=3$ | $j=2$ | $j=1$ | $j=0$ |
|---|---|---|---|---|---|
| 2 | $a_j$ | . | $2\delta - \frac{1}{2}\delta^2$ | $-4\delta + 2\delta^2$ | $2\delta - \frac{3}{2}\delta^2$ |
| | $c_j$ | . | 1 | $2(\delta - 1)$ | $(\delta - 1)^2$ |
| | $b_j$ | . | 0 | $2\delta$ | $(\delta - 1)^2 - 1$ |

| Order | | $j=3$ | $j=2$ | $j=1$ | $j=0$ |
|---|---|---|---|---|---|
| 3 | $a_j$ | $3\delta - \frac{3}{2}\delta^2 + \frac{1}{3}\delta^3$ | $-9\delta + \frac{15}{2}\delta^2 - \frac{3}{2}\delta^3$ | $9\delta - \frac{21}{2}\delta^2 + 3\delta^3$ | $-3\delta + \frac{9}{2}\delta^2 - \frac{11}{6}\delta^3$ |
| | $c_j$ | 1 | $3(\delta - 1)$ | $3(\delta - 1)^2$ | $(\delta - 1)^3$ |
| | $b_j$ | 0 | $3\delta$ | $-6\delta + 3\delta^2$ | $(\delta - 1)^3 + 1$ |

| Order | | | $j=4$ | $j=3$ |
|---|---|---|---|---|
| 4 | $a_j$ | . | $4\delta - 3\delta^2 + \frac{4}{3}\delta^3 - \frac{1}{4}\delta^4$ | $-16\delta + 18\delta^2 - \frac{22}{3}\delta^3 + \frac{4}{3}\delta^4$ |
| | $c_j$ | . | 1 | $4(\delta - 1)$ |
| | $b_j$ | . | 0 | $4\delta$ |

| | | $j=2$ | $j=1$ | $j=0$ |
|---|---|---|---|---|
| | $a_j$ | $24\delta - 36\delta^2 + 18\delta^3 - 3\delta^4$ | $-16\delta + 30\delta^2 - \frac{58}{3}\delta^3 + 4\delta^4$ | $4\delta - 9\delta^2 + \frac{22}{3}\delta^3 - \frac{25}{12}\delta^4$ |
| | $c_j$ | $6(\delta - 1)^2$ | $4(\delta - 1)^3$ | $(\delta - 1)^4$ |
| | $b_j$ | $-12\delta + 6\delta^2$ | $12\delta - 12\delta^2 + 4\delta^3$ | $(\delta - 1)^4 - 1$ |

(*continued on next page*)

**Table B.4** (*continued*)

| Order | | $j=5$ | $j=4$ |
|---|---|---|---|
| 5 | $a_j$ | $5\delta - 5\delta^2 + \frac{10}{3}\delta^3 - \frac{5}{4}\delta^4 + \frac{1}{5}\delta^5$ | $-25\delta + 35\delta^2 - \frac{65}{3}\delta^3 + \frac{95}{12}\delta^4 - \frac{5}{4}\delta^5$ |
| | $c_j$ | 1 | $5(\delta - 1)$ |
| | $b_j$ | 0 | $5\delta$ |
| | | $j=3$ | $j=2$ |
| | $a_j$ | $50\delta - 90\delta^2 + \frac{190}{3}\delta^3 - \frac{65}{3}\delta^4 + \frac{10}{3}\delta^5$ | $-50\delta + 110\delta^2 - \frac{280}{3}\delta^3 + 35\delta^4 - 5\delta^5$ |
| | $c_j$ | $10(\delta - 1)^2$ | $10(\delta - 1)^3$ |
| | $b_j$ | $-20\delta + 10\delta^2$ | $30\delta + 10\delta^3 - 30\delta^2$ |
| | | $j=1$ | $j=0$ |
| | $a_j$ | $25\delta - 65\delta^2 + \frac{200}{3}\delta^3 - \frac{365}{12}\delta^4 + 5\delta^5$ | $-5\delta + 15\delta^2 - \frac{55}{3}\delta^3 + \frac{125}{12}\delta^4 - \frac{137}{60}\delta^5$ |
| | $c_j$ | $5(\delta - 1)^4$ | $(\delta - 1)^5$ |
| | $b_j$ | $-20\delta + 30\delta^2 - 20\delta^3 + 5\delta^4$ | $(\delta - 1)^5 + 1$ |

## Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jcp.2018.09.044.

## References


[1] R. Rosales, B. Seibold, D. Shirokoff, D. Zhou, Unconditional stability for multistep ImEx schemes – Theory, SIAM J. Numer. Anal. 55 (2017) 2336–2360.
[2] U. Ascher, S.J. Ruuth, B. Wetton, Implicit–explicit methods for time dependent partial differential equations, SIAM J. Numer. Anal. 32 (1995) 797–823.
[3] M. Crouzeix, Une méthode multipas implicite–explicite pour l'approximation des équations d'évolution paraboliques, Numer. Math. 35 (1980) 257–276.
[4] J.M. Varah, Stability restrictions on second order, three level finite difference schemes for parabolic equations, SIAM J. Numer. Anal. 17 (1980) 300–309.
[5] W. Hundsdorfer, J. Verwer, Numerical Solution of Time-Dependent Advection–Diffusion–Feaction Equations, Springer Series in Comput. Math., vol. 33, Springer, 2003.
[6] G. Akrivis, Implicit–explicit multistep methods for nonlinear parabolic equations, Math. Comput. 82 (2012) 45–68.
[7] G. Akrivis, M. Crouzeix, C. Makridakis, Implicit–explicit multistep finite element methods for nonlinear parabolic problems, Math. Comput. 67 (1998) 457–477.
[8] G. Akrivis, M. Crouzeix, C. Makridakis, Implicit–explicit multistep methods for quasilinear parabolic equations, Numer. Math. 82 (1999) 521–541.
[9] G. Akrivis, F. Karakatsani, Modified implicit–explicit BDF methods for nonlinear parabolic equations, BIT Numer. Math. 43 (2003) 467–483.
[10] J. Douglas, T. Dupont, Alternating-direction Galerkin methods on rectangles, in: B. Hubbard (Ed.), Numerical Solution of Partial Differential Equations, vol. II, SYNSPADE-1970, Univ. of Maryland, Academic Press, New York, College Park, Md, 1971, pp. 133–213.
[11] T. Heister, M.A. Olshanskii, L.G. Rebholz, Decoupled, unconditionally stable, higher order discretizations for MHD flow simulation, J. Sci. Comput. 71 (2017) 21–43.
[12] D. Eyre, Unconditionally gradient stable time marching the Cahn–Hilliard equation, in: J.W. Bullard, R. Kalia, M. Stoneham, L. Chen (Eds.), Computational and Mathematical Models of Microstructural Evolution, vol. 53, Materials Research Society, Warrendale, PA, USA, 1998, pp. 1686–1712.
[13] A. Bertozzi, N. Ju, J.-W. Lu, A biharmonic-modified forward time stepping method for fourth order nonlinear diffusion equations, Discrete Contin. Dyn. Syst. 29 (2011) 1367–1391.
[14] M. Elsey, B. Wirth, A simple and efficient scheme for phase field crystal simulation, Modél. Math. Anal. Numér. 47 (2013) 1413–1432.
[15] G. Sheng, T. Wang, Q. Du, K. Wang, Z. Liu, L.Q. Chen, Coarsening kinetics of a two phase mixture with highly disparate diffusion mobility, Commun. Comput. Phys. 8 (2010) 249–264.
[16] P. Smereka, Semi-implicit level set methods for curvature and surface diffusion motion, J. Sci. Comput. 19 (2003) 439–456.
[17] K. Glasner, S. Orizaga, Improving the accuracy of convexity splitting methods for gradient flow equations, J. Comput. Phys. 315 (2016) 52–64.
[18] Z. Guan, J. Lowengrub, C. Wang, S. Wise, Second-order convex splitting schemes for periodic nonlocal Cahn–Hilliard and Allen–Cahn equations, J. Comput. Phys. 277 (2014) 48–71.
[19] Y. Yan, W. Chen, C. Wang, S. Wise, A second-order energy stable BDF numerical scheme for the Cahn–Hilliard equation, Commun. Comput. Phys. 23 (2018) 572–602.
[20] V. Badalassi, H. Ceniceros, S. Banerjee, Computation of multiphase systems with phase field models, J. Comput. Phys. 190 (2003) 371–397.
[21] L. Duchemin, J. Eggers, The explicit–implicit-null method: removing the numerical instability of PDEs, J. Comput. Phys. 263 (2014) 37–52.
[22] H. Johnston, J.-G. Liu, Accurate, stable and efficient Navier–Stokes solvers based on explicit treatment of the pressure term, J. Comput. Phys. 199 (2004) 221–259.
[23] J.-G. Liu, J. Liu, R.L. Pego, Stability and convergence of efficient Navier–Stokes solvers via a commutator estimate, Commun. Pure Appl. Math. 60 (2007) 1443–1487.
[24] T. Heister, M.A. Olshanskii, L.G. Rebholz, Unconditional long-time stability method for the 2D Navier–Stokes equations, Numer. Math. 135 (2017) 143–167.
[25] G. Karniadakis, M. Israeli, S.A. Orszag, High-order splitting methods for the incompressible Navier–Stokes equations, J. Comput. Phys. 97 (1991) 414–443.
[26] J. Kim, P. Moin, Application of a fractional step method to incompressible Navier–Stokes equations, J. Comput. Phys. 59 (1985) 308–323.
[27] J.-G. Liu, J. Liu, R.L. Pego, Stable and accurate pressure approximation for unsteady incompressible viscous flow, J. Comput. Phys. 229 (2010) 3428–3453.
[28] W. Layton, C. Trenchea, Stability of two IMEX methods, CNLF and BDF2-AB2, for uncoupling systems of evolution equations, Appl. Numer. Math. 62 (2012) 112–120.
[29] O.P. Bruno, M. Cubillos, Higher-order in time quasi-unconditionally stable ADI solvers for the compressible Navier–Stokes equations in 2D and 3D curvilinear domains, J. Comput. Phys. 307 (2016) 476–495.
[30] O.P. Bruno, M. Cubillos, On the quasi-unconditional stability of BDF-ADI solvers for the compressible Navier–Stokes equations, SIAM J. Numer. Anal. 55 (2017) 892–922.

[31] M. Anitescu, W. Layton, F. Pahlevani, Implicit for local effects, explicit for nonlocal is unconditionally stable, Electron. Trans. Numer. Anal. 18 (2004) 174–187.
[32] C. Trenchea, Second order implicit for local effects and explicit for nonlocal effects is unconditionally stable, ROMAI J. 12 (2016) 163–178.
[33] A. Christlieb, J. Jones, K. Promislow, B. Wetton, M. Willoughby, High accuracy solutions to energy gradient flows from material science models, J. Comput. Phys. 257 (2014) 193–215.
[34] H. Ceniceros, A semi-implicit moving mesh method for the focusing nonlinear Schroedinger equation, Commun. Pure Appl. Anal. 1 (2002) 1–14.
[35] O.P. Bruno, E. Jimenez, Higher-order linear-time unconditionally stable ADI methods for nonlinear convection–diffusion PDE systems, J. Fluids Eng. 136 (2014) 060904.
[36] O.P. Bruno, M. Lyon, High-order unconditionally stable FC-AD solvers for general smooth domains I. Basic elements, J. Comput. Phys. 229 (2010) 2009–2033.
[37] O.P. Bruno, M. Lyon, High-order unconditionally stable FC-AD solvers for general smooth domains II. Elliptic, parabolic and hyperbolic PDEs; theoretical considerations, J. Comput. Phys. 229 (2010) 3358–3381.
[38] J. Shin, H. Lee, J.-Y. Lee, Unconditionally stable methods for gradient flow using convex splitting Runge–Kutta scheme, J. Comput. Phys. 347 (2017) 367–381.
[39] M.L. Minion, Semi-implicit spectral deferred correction methods for ordinary differential equations, Commun. Math. Sci. 1 (2003) 471–500.
[40] A.-K. Kassam, L.N. Trefethen, Fourth-order time-stepping for stiff PDEs, SIAM J. Sci. Comput. 26 (2005) 1214–1233.
[41] L. Ju, J. Zhang, L. Zhu, Q. Du, Fast explicit integration factor methods for semilinear parabolic equations, J. Sci. Comput. 62 (2015) 431–455.
[42] P.A. Milewski, E.G. Tabak, A pseudo-spectral algorithm for the solution of nonlinear wave equations, SIAM J. Sci. Comput. 21 (1999) 1102–1114.
[43] A. Abdulle, A.A. Medovikov, Second order Chebyshev methods based on orthogonal polynomials, Numer. Math. 90 (2001) 1–18.
[44] L.N. Trefethen, D. Bau, Numerical Linear Algebra, SIAM, Philadelphia, 2000.
[45] R.J. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems, first ed., SIAM, Philadelphia, 2007.
[46] J. Frank, W. Hundsdorfer, J. Verwer, On the stability of IMEX LM methods, Appl. Numer. Math. 25 (1997) 193–205.
[47] T. Koto, Stability of implicit–explicit linear multistep methods for ordinary and delay differential equations, Front. Math. China 4 (2009) 113–129.
[48] R. Jeltsch, O. Nevanlinna, Stability of explicit time discretizations for solving initial value problems, Numer. Math. 37 (1981) 61–91.
[49] R. Jeltsch, O. Nevanlinna, Stability and accuracy of time discretizations for initial value problems, Numer. Math. 40 (1982) 245–296.
[50] A. Horn, C. Johnson, Topics in Matrix Analysis, Cambridge University Press, 1991.
[51] C.R. Johnson, Numerical determination of the field of values of a general complex matrix, SIAM J. Numer. Anal. 15 (1978) 595–602.
[52] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, second ed., Springer-Verlag, Berlin, 1987.
[53] T.A. Driscoll, N. Hale, L.N. Trefethen (Eds.), Chebfun Guide, Pafnuty Publications, Oxford, 2014.
[54] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, second ed., Springer-Verlag, Berlin, 1991.
[55] L.S. Leibenzon, General problem of the movement of a compressible fluid in a porous media, Izv. Akad. Nauk SSSR, Geogr. Geophys. 9 (1945) 7–10 (Russian).
[56] M. Muskat, The Flow of Homogeneous Fluids Through Porous Media, McGrawHill, New York, 1937.
[57] J.-L. Guermond, P. Minev, High-order time stepping for the incompressible Navier–Stokes equations, SIAM J. Sci. Comput. 36 (2015) A2656–A2681.
[58] W.D. Henshaw, A fourth-order accurate method for the incompressible Navier–Stokes equations on overlapping grids, J. Comput. Phys. 113 (1994) 13–25.
[59] H. Johnston, J.-G. Liu, A finite difference method for incompressible flow based on local pressure boundary conditions, J. Comput. Phys. 180 (2002) 120–154.
[60] D. Shirokoff, R.R. Rosales, An efficient method for the incompressible Navier–Stokes equations on irregular domains with no-slip boundary conditions, high order up to the boundary, J. Comput. Phys. 230 (2011) 8619–8646.
[61] E. Cozzi, R.L. Pego, On optimal estimates for the Laplace–Leray commutator in planar domains with corners, Proc. Am. Math. Soc. 139 (2011) 1691–1706.