IMEX extensions of linear multistep methods with general monotonicity and boundedness properties

Willem Hundsdorfer [a], Steven J. Ruuth [b],[*],[1]

[a] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*
[b] *Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6*

Abstract

For solving hyperbolic systems with stiff sources or relaxation terms, time stepping methods should combine favorable monotonicity properties for shocks and steep solution gradients with good stability properties for stiff terms. In this paper we consider implicit–explicit (IMEX) multistep methods. Suitable methods will be constructed, based on explicit methods with general monotonicity and boundedness properties for hyperbolic equations. Numerical comparisons are made with several implicit–explicit Runge–Kutta methods.
© 2007 Elsevier Inc. All rights reserved.

1. Introduction

1.1. IMEX linear multistep methods

There are many applications that lead to initial value problems for systems of ordinary differential equations (ODEs) in $\mathbb{R}^M$ of the form

$$u'(t) = F(u(t)) + G(u(t)), \qquad u(0) = u_0, \tag{1.1}$$

where F represents a non-stiff (or mildly stiff) part of the equation, and G is a stiff term, requiring implicit integration. To solve such systems we consider implicit–explicit (IMEX) schemes, producing numerical approximations $u_n \approx u(t_n)$ at the time levels $t_n = n\Delta t$. We shall deal in particular with combinations of k-step explicit and implicit linear multistep methods

---

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \sum_{j=1}^{k} \hat{b}_j \Delta t F(u_{n-j}) + \sum_{j=0}^{k} b_j \Delta t G(u_{n-j}), \tag{1.2}$$

with parameters a_j, $\hat{b}_j$ and b_j. For brevity we will usually denote $F_n = F(u_n)$, $G_n = G(u_n)$; for non-autonomous systems this will become $F_n = F(t_n, u_n)$, $G_n = G(t_n, u_n)$.

Combinations (1.2) of explicit and implicit multistep methods were introduced in [7,32]. In this paper we shall base such IMEX combinations on explicit methods with favorable monotonicity and boundedness properties, in order to avoid numerical oscillations for problems with steep solution gradients. The corresponding implicit methods should provide sufficient stability for stiff terms. The construction of the IMEX schemes in the present paper and the discussion of the linear stability properties is related to [3]; for this discussion, it will be assumed that the implicit term $G(u)$ is either a discretized symmetric diffusion term or a linear decay term λu with $\lambda \leqslant 0$ real. Results for complex λ with $\mathrm{Re}\lambda \leqslant 0$ can be found in [10,12]. Stability for nonlinear parabolic equations in Hilbert spaces has been studied in [1,7].

1.2. Hyperbolic equations with stiff source and relaxation terms

An interesting class of problems suitable for IMEX methods is formed by systems of hyperbolic equations with stiff sources or relaxation terms

$$u_t + \nabla \cdot f(u) = \frac{1}{\epsilon} g(u), \tag{1.3}$$

where $u = u(x, t) \in \mathbb{R}^m$. Suitable initial and (inflow) boundary conditions are assumed. In the limit $\epsilon \to 0$ the solution of (1.3) will in general satisfy $g(u) = 0$, which defines a manifold of dimension $\bar{m} < m$. This can lead to a reduced system of conservation laws, with $\bar{u} \in \mathbb{R}^{\bar{m}}$,

$$\bar{u}_t + \nabla \cdot \bar{f}(\bar{u}) = 0. \tag{1.4}$$

Examples and more detailed information can be found in [22,28], for instance; see also Section 6 for some additional examples.

To solve (1.3) one thus needs a time stepping method that is capable of treating nonlinear convection terms efficiently while also providing sufficient stability for the stiff terms. This can be obtained by IMEX schemes. In view of the reduction (1.4) we want our explicit method in the IMEX combination to possess suitable monotonicity and boundedness properties, to suppress numerical oscillations. Such explicit methods have been constructed in [21,30,29].

Finally we note that there can be of course additional terms causing stiffness, such as diffusion. Stiff terms can also be introduced by chemical reactions or by mechanical stiffness in fluid–structure interactions, for example.

1.3. Outline

This paper has two main objectives. The first of these is the construction of IMEX linear multistep methods suited for (1.3) and related problems. The second objective is to study the numerical behavior of these schemes and to make comparisons with IMEX Runge–Kutta methods.

The outline of this paper is as follows. In Section 2 we present some preliminaries and background material. The construction of suitable IMEX linear multistep methods, together with a discussion of their linear stability properties, is given in Section 3. Some state-of-the-art IMEX Runge–Kutta methods from the literature [3,20,28] that are to be used in our numerical comparisons are introduced in Section 4. Section 5 contains a discussion of the discretization errors for the IMEX methods with a focus on the possible order reduction effects that can arise with Runge–Kutta schemes. The results of several numerical tests with the IMEX multistep and Runge–Kutta schemes are presented in Section 6. Finally, Section 7 contains a summary and conclusions.

2. Preliminaries

2.1. Monotonicity and TVD with arbitrary starting values

Spatial discretization of hyperbolic equations with limiters can lead to non-stiff ODE systems $u'(t) = F(u(t))$ where the function F is such that there is a $\Delta t_{FE} > 0$ for which

$$\|v + \Delta t_{FE} F(v)\| \leqslant \|v\| \quad \text{for all } v \in \mathbb{R}^M, \tag{2.1}$$

where $\|\cdot\|$ is a given semi-norm or convex functional. Due to convexity, it is easily seen that then also $\|v + \Delta t F(v)\| \leqslant \|v\|$ whenever $\Delta t \leqslant \Delta t_{FE}$. Hence Δt_{FE} can be viewed as the maximal step size for the forward Euler method such that the monotonicity property $\|u_n\| \leqslant \|u_0\|$ is valid for an arbitrary starting value u_0.

For some of the explicit or implicit linear multistep methods there exists a $C > 0$ such that (2.1) implies

$$\|u_n\| \leqslant \max_{0 \leqslant j \leqslant k-1} \|u_j\| \quad \text{for all } n \geqslant 1 \quad \text{and} \quad \Delta t \leqslant C \Delta t_{FE}. \tag{2.2}$$

Originally [30], this was called the TVD (total variation diminishing) property, with $\|\cdot\|$ the total variation semi-norm for one-dimensional conservation laws. More recently, with arbitrary semi-norms, it is often called strong stability preservation (SSP). In view of the generalization below it will be referred to as a monotonicity property, and the step size coefficient C will be called the monotonicity threshold.

2.2. Boundedness and TVB

Property (2.2) requires that all $a_j, \hat{b}_j \geqslant 0$, and then $C = \min_{1 \leqslant j \leqslant k} a_j / \hat{b}_j$. However, there are only a few methods for which all coefficients are non-negative. More relaxed conditions can be derived for linear multistep methods combined with a starting procedure that generates $u_1, \ldots, u_{k-1}$ from the given u_0. For such combinations we can consider the less demanding property

$$\|u_n\| \leqslant K \|u_0\| \quad \text{for all } n \geqslant 1 \quad \text{and} \quad \Delta t \leqslant C \Delta t_{FE}, \tag{2.3}$$

with some constant $K \geqslant 1$. If $\|\cdot\|$ is the total variation semi-norm, this is a TVB (total variation boundedness) property. For general semi-norms or convex functionals we will simply refer to it as boundedness and C will be called the boundedness threshold.

In contrast to (2.2), there are many linear multistep methods for which (2.3) can hold with a positive threshold C; see [16,29]. For theoretical purposes it is important that the size of the bound K does not depend on the problem at hand, but only on the coefficients of the multistep method and the starting procedure. In fact, in numerical experiments [16,17] it appears that K is usually very close to one, but theoretically this is still not well understood.

As noted before, the basic assumption (2.1) on F can be satisfied for instance for semi-discrete ODE problems arising from conservation laws with spatial TVD discretizations based on limiters; examples can be found e.g. in [22,18]. However, also for the so-called WENO spatial discretizations, which do not strictly satisfy (2.1), it is a useful assumption to find suitable time stepping methods [31].

One can also try to obtain monotonicity and boundedness results for the total IMEX scheme, by making an assumption similar to (2.1) for G. However, this leads to severe time step restrictions because the threshold factors C for implicit methods are not much larger than for explicit methods [16,17]; see also [9,15] for related Runge–Kutta results. From a practical point of view these step size restrictions are not acceptable in general. Therefore we shall only regard the monotonicity properties of explicit methods and require stability of the implicit method for stiff problems, so that the limit from 1.3 and 1.4 can be taken.

2.3. Linear stability and scalar test equations

We now consider the stability of the original ODE system (1.1). The most simple equation of that form used to study linear stability properties of IMEX schemes is the scalar test equation

$$u'(t) = \hat{\lambda} u(t) + \lambda u(t), \quad \hat{\lambda}, \lambda \in \mathbb{C}, \tag{2.4}$$

with $F(u) = \hat{\lambda}u$ and stiff term $G(u) = \lambda u$. We denote in the following $\hat{z} = \Delta t \hat{\lambda}$, $z = \Delta t \lambda$. The characteristic equation for the IMEX multistep method applied to (2.4) is given by

$$\rho(\zeta) - \hat{z}\hat{\sigma}(\zeta) - z\sigma(\zeta) = 0, \tag{2.5}$$

where

$$\rho(\zeta) = \zeta^k - \sum_{j=1}^{k} a_j \zeta^{k-j}, \quad \hat{\sigma}(\zeta) = \sum_{j=1}^{k} \hat{b}_j \zeta^{k-j}, \quad \sigma(\zeta) = \sum_{j=0}^{k} b_j \zeta^{k-j}. \tag{2.6}$$

It is well known that stability of the scheme requires that (i) all roots of (2.5) have modulus less than or equal to one, and (ii) multiple roots have modulus strictly less than one. Multiple roots of modulus one lead to polynomial growth; such roots usually appear only on the boundaries of stability domains.

In applications, the $\hat{\lambda}$ and λ stand for eigenvalues of linearizations of F and G, respectively. Theoretical results are difficult to obtain if these linearizations do not commute, but in many practical situations stability considerations based on the simple scalar test equations are (surprisingly) accurate for predicting maximal step sizes for stability.

Particular attention will be given to the test equation (2.4) where $\hat{\lambda}$ is purely imaginary and z is real and non-positive, i.e.,

$$\hat{z} = i\eta, \qquad z = \xi \leqslant 0. \tag{2.7}$$

This case is relevant, for example, for advection–diffusion equations if central finite differences or spectral approximations are used in space.

The damping properties of the scheme in the limit case $\xi \to -\infty$ are determined by the *damping factor*

$$D \equiv \max\{|\zeta| : \sigma(\zeta) = 0\}. \tag{2.8}$$

This limit case corresponds to the damping of the high-frequency Fourier modes for advection–diffusion equations.

Related to $A(\alpha)$ stability for implicit methods, we can require that the IMEX scheme is stable whenever $\xi \leqslant 0$ and $|\eta/\xi| \leqslant \arctan(\alpha)$ with angle $\alpha \leqslant \frac{1}{2}\pi$. Actually, for advection–diffusion equations, stability within a parabola $\xi \leqslant 0$, $|\eta^2/\xi| \leqslant \arctan(\beta)$ can be more relevant than for a wedge with angle α; see [6] for instance. However, for the methods considered in this paper, a large angle α will correspond to a large β, as can be seen from the figures below for the stability domains $\mathscr{D}_{AR}(2)$.

2.4. Stability restrictions for advection–diffusion–reaction equations

In the next sections we will graphically present the stability properties of the IMEX methods for advection–reaction equations

$$u_t + au_x = -cu \tag{2.9}$$

and for advection–diffusion equations

$$u_t + au_x = du_{xx}, \tag{2.10}$$

with constant coefficients $a, c, d \geqslant 0$. The spatial operators will be discretized by some well-known finite-difference formulas. In both cases (2.9 and 2.10), Fourier transformation gives rise to a scalar test Eq. (2.4) with $\hat{\lambda}$ the eigenvalues for the advection terms, which are assumed to be taken explicitly in the IMEX schemes, and λ those of the (stiff) reaction or diffusion terms. In fact, we could also deal with a non-scalar reaction term $-Ku$ with positive definite matrix K, in which case c would represent an eigenvalue of K.

Consider, as an example, the standard second-order centered discretizations for advection and diffusion with mesh width Δx. Let $v = a\Delta t/\Delta x$ be the Courant number and denote the cell Péclet number for advection–diffusion by $\mu = a\Delta x/d$. Then the eigenvalues for the advection term, multiplied by Δt, are

$$\hat{z} = iv\sin(2\omega), \tag{2.11}$$

where ω is the frequency of the corresponding Fourier mode. For (2.9) we have, of course,

$$z = -c\Delta t, \tag{2.12}$$

whereas the eigenvalues for the implicit diffusion term in (2.10) are given by

$$z = -4\frac{\mu}{v}\sin^2\omega. \tag{2.13}$$

Stability for the advection–reaction equation (2.9) therefore requires that the root condition for the characteristic equation (2.5) is satisfied for $\hat{z} = i\eta$, $|\eta| \leqslant v$ and $z = \xi \leqslant 0$. For this condition we present plots of stability domains in the (ξ, v)-plane. Actually, only the boundaries of these domains will be displayed, separating the region of stability (lower-left part) from the region of instability (upper-right part). As noted before, on the boundary itself there may be multiple roots of modulus one, leading to a weak (polynomial) instability.

For the advection–diffusion Eq. (2.10) stability is slightly more complicated because $\hat{z}$ and z are then coupled through the frequency ω. It is then required that the root condition for (2.5) is satisfied for all $\hat{z}, z$ given by (2.11), (2.13) with $\omega \in [0, 2\pi]$. Stability domains for this case will be presented in the (μ, v)-plane with cell Péclet number μ horizontally and Courant number v vertically, on logarithmic scale.

Since we are particularly interested in advection discretizations with good shape-preserving properties, we will also present the corresponding stability domains for spatial advection discretization by the first-order upwind formula, where

$$z = -2v\sin^2\omega + iv\sin(2\omega) \tag{2.14}$$

and by the third-order upwind-biased formula, where

$$z = -\frac{4}{3}v\sin^4\omega - iv\sin(2\omega)\left(1 + \frac{2}{3}\sin^2\omega\right). \tag{2.15}$$

In all cases the diffusion discretization is second-order central.

The different advection discretizations are indicated by their order $q = 1, 2, 3$, that is, $q = 1$ for first-order upwind, $q = 2$ for second-order central and $q = 3$ for the third-order upwind-biased formula. The stability domains in the (ξ, v)-plane for advection–reaction will be denoted by $\mathscr{D}_{AR}(q)$. These will be displayed in the figures below on a linear scale with growth rate ξ on the horizontal axis and Courant number v on the vertical axis. Likewise, $\mathscr{D}_{AD}(q)$ will be the stability domain for advection–diffusion in the (μ, v)-plane, and for those figures a logarithmic scale is used with Courant number $v \in [0.1, 10]$ and Péclet number $\mu \in [10^{-4}, 10^4]$. Finally we note that closely related figures for the central difference case $q = 2$ have been presented in [3] for some IMEX multistep schemes and in [2] for some IMEX Runge–Kutta schemes.

2.5. Order conditions and error constants

The order conditions and the structure of local errors for IMEX linear multistep methods are quite simple, see for instance [3,7] and also Section 5.2. Let $q_0 = 1 - \sum_{j=1}^{k} a_j$ and

$$q_l = \frac{(-1)^l}{l!}\sum_{j=0}^{k}(-j^l a_j + l j^{l-1} b_j), \qquad \hat{q}_l = \frac{(-1)^l}{l!}\sum_{j=0}^{k}(-j^l a_j + l j^{l-1}\hat{b}_j), \tag{2.16}$$

where we can set $a_0 = \hat{b}_0 = 0$. It will be assumed that

$$q_0 = 0, \qquad q_l = \hat{q}_l = 0 \quad (1 \leqslant l \leqslant p). \tag{2.17}$$

These are just the conditions for the individual explicit and implicit methods to be of order p, and that is sufficient for the IMEX combination to be of order p as well. It will always be assumed that $p \geqslant 1$.

The *error constants* of the IMEX scheme are given by

$$E = \frac{q_{p+1}}{\sigma(1)}, \qquad \widehat{E} = \frac{\hat{q}_{p+1}}{\hat{\sigma}(1)}. \tag{2.18}$$

These are the constants that will appear in the global error for non-stiff problems, as can be seen by following the reasoning given in [13, p. 373]. The discretization errors for stiff problems will be discussed in Section 5. Note that for methods of order $p \geqslant 1$ we have $\sum_{j=0}^{k} b_j = \sum_{j=1}^{k} \hat{b}_j$, that is, $\sigma(1) = \hat{\sigma}(1)$.

3. The IMEX linear multistep methods

3.1. General construction issues

Our approach to develop implicit–explicit methods with good nonlinear stability properties is to start from an explicit linear multistep method with optimal or near-optimal monotonicity or boundedness properties, and then construct a compatible implicit method. Notice that once we have specified an explicit k-step method for an order-p IMEX scheme, we end up with $k - p + 1$ parameter family of compatible implicit methods. This leads us to a two-step approach for constructing IMEX schemes. First, we select an explicit method with good monotonicity properties. Second, we optimize over the free parameters to find a good, compatible implicit method.

The explicit methods will be selected from the class of optimal monotonicity-preserving methods (or TVD multistep methods) of Shu [30], or methods with favorable boundedness properties. Schemes of that type have been determined in [17,29].

The determination of the implicit method is more subtle, and requires the choice of some optimization criteria. Many properties are desirable, however, we shall mainly focus on methods that have a strong damping of high-frequency error modes. Methods of this type can be more efficient when an iterative treatment of the implicit equations is undertaken, and they are less prone to aliasing errors and tend to minimize spurious oscillations [3]. Based on this choice, we obtain an optimization problem which we refer to as the *optimal damping criterion*:

$$\min_{\{b_0, b_1 \ldots, b_k\}} D(b_0, b_1 \ldots, b_k), \tag{3.1}$$

where D is the damping factor for the scheme defined by (2.8).

Alternatively, we may want to emphasize schemes with large linear stability domains while still maintaining some damping of high-frequency error modes. While it is not obvious how to obtain optimal linear stability properties for general problems, we have found that minimizing the average damping over part of the (ξ, η)-plane usually gives good overall stability. We therefore also search for good implicit methods by finding IMEX schemes which minimize the product of the damping parameter and such an average damping. The optimization criteria, which will be referred to as the *combined criteria*, is given by

$$\min_{\{b_0, b_1 \ldots, b_k\}} S(b_0, b_1 \ldots, b_k) D(b_0, b_1 \ldots, b_k), \tag{3.2}$$

where S is the average damping over a selected part of the (ξ, η)-plane. In our optimizations we take $S \equiv \int_\Omega r(\xi, \eta) \mathrm{d}\xi \mathrm{d}\eta$ with $\Omega = [-10, 0] \times [0, 1]$ and $r(\xi, \eta)$ the maximal modulus root for (2.5), (2.7). The integral is treated using a piecewise constant approximation of r sampled over a 10×10 grid that is equispaced in each coordinate direction.

All optimizations have been carried out using the Matlab Optimization Toolkit. Maple was used to find fractional approximations of the coefficients which satisfy the order conditions exactly.

3.2. Order-two schemes

3.2.1. A case study: conflicting optimization criteria

Instead of damping properties, there are other criteria one could take for optimization. Let us consider as an example the second-order explicit three-step method of Shu [30], augmented by an implicit method

$$u_n = \frac{3}{4} u_{n-1} + \frac{1}{4} u_{n-3} + \frac{3}{2} \Delta t F_{n-1} + \sum_{j=0}^{3} b_j \Delta t G_{n-j}. \tag{3.3}$$

These schemes have a monotonicity threshold value of $C = \frac{1}{2}$. The requirement of order two leaves us two free parameters b_j, which we can take to be b_0 and b_1.

We can estimate the angle α such that the IMEX scheme is stable for the test Eqs. (2.4), (2.7) whenever $\xi \leqslant 0$ and $|\eta/\xi| \leqslant \arctan(\alpha)$. This angle can be estimated by studying the root locus curve for $\xi \to -\infty$. It
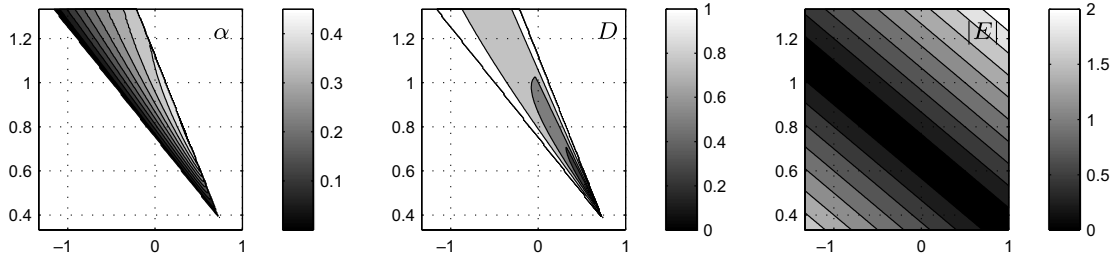
Fig. 1. Estimated angles α, damping factors D and absolute error constants $|E|$ for (3.3), with parameter $b_0 \in \left[\frac{1}{3}, \frac{4}{3}\right]$ on the vertical axis and $b_1 \in \left[-\frac{4}{3}, 1\right]$ horizontally.

gives only an estimate because the boundary of the stability domain for moderate negative ξ values needs to be verified separately, but for the schemes presented below the estimates provide accurate values for the actual angle α.

Fig. 1 shows these estimated angles α, together with the damping factors and error constants of the implicit methods as functions of b_0, b_1. The contour lines for D and α are only displayed for the parameter region where $|D| \leqslant 1$, because otherwise the implicit method is not A_0-stable. It is obvious from the figure that for this particular explicit method optimization of α does not combine with a good damping factor or small error constant.

Optimizing the damping factor D or the combined criteria leads to

$$u_n = \frac{3}{4} u_{n-1} + \frac{1}{4} u_{n-3} + \frac{3}{2} \Delta t F_{n-1} + \frac{4}{9} \Delta t G_n + \frac{2}{3} \Delta t G_{n-1} + \frac{1}{3} \Delta t G_{n-2} + \frac{1}{18} \Delta t G_{n-3}, \tag{3.4}$$

to which we will refer as the IMEX-Shu(3, 2) scheme. As mentioned above, this scheme has threshold value $C = 0.5$, and it has a rather small angle $\alpha = 0.06$. Further characteristic values of the scheme are $D = 0.5$, $\widehat{E} = 0.333$ and $E = 0$.

Second-order implicit–explicit schemes based on explicit TVD linear multistep methods were first considered by Gjesdal [12]. A scheme with good linear stability properties arising from this analysis is

$$u_n = \frac{3}{4} u_{n-1} + \frac{1}{4} u_{n-3} + \frac{3}{2} \Delta t F_{n-1} + \Delta t G_n + \frac{1}{2} \Delta t G_{n-3}. \tag{3.5}$$

We will refer to this scheme as the IMEX-SG(3, 2) scheme. It has $\alpha = 0.38$, $D = 0.794$, $\widehat{E} = 0.333$ and $E = -0.667$. The linear stability properties are better than for the IMEX-Shu(3, 2) scheme (see Fig. 2), while the damping factor D and error constant E are less favorable.

For this example with the three-step scheme (3.3) optimizing the damping properties or the angles α leads to a substantial difference in stability properties of the resulting IMEX schemes for the advection–diffusion–reaction test problems. We note that for most of the other schemes considered in this paper these differences were far less pronounced.

Since our primary goal is to provide good schemes for nonlinear systems of the type (1.3), for which damping in the limit $\epsilon \to 0$ is important, we will give some preference to optimizing the damping properties. Of

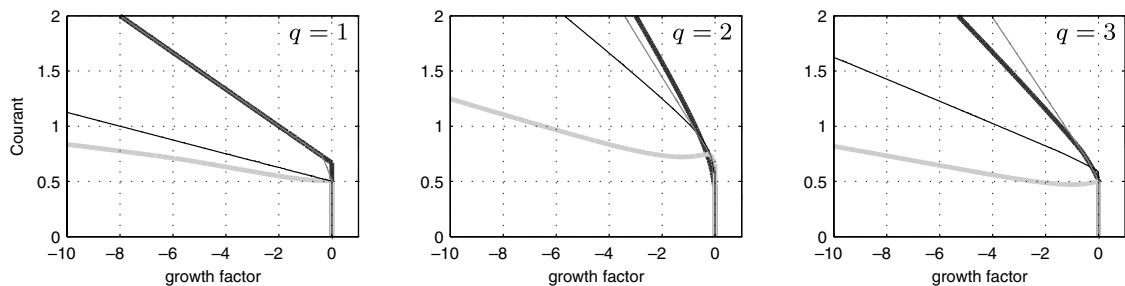Fig. 2. Boundaries of $\mathscr{D}_{\mathrm{AR}}(q)$, $q = 1, 2, 3$, for the second-order methods IMEX-BDF2 (dark gray), IMEX-Adams2 (thin black) and the IMEX-Shu(3,2) extensions (3.4) (light gray) and (3.5) (thin gray).

course, this should not go at all costs. The combined criteria will usually provide sensible choices. For all methods the stability domains $\mathscr{D}_{AR}$ for advection–reaction and $\mathscr{D}_{AD}$ for advection–diffusion will be shown graphically.

3.2.2. Other order-two schemes

Comparisons against standard two-step IMEX schemes of order two can be made. The IMEX-BDF2 scheme

$$u_n = \frac{4}{3}u_{n-1} - \frac{1}{3}u_{n-2} + \frac{4}{3}\Delta t F_{n-1} - \frac{2}{3}\Delta t F_{n-2} + \frac{2}{3}\Delta t G_n \tag{3.6}$$

is a particularly favorable choice, as it has a large threshold value $C = \frac{5}{8}$ and optimal damping $D = 0$ for high-frequency error modes. The error constants for the explicit and implicit methods are $\widehat{E} = 0.667$ and $E = -0.333$, respectively.

Another possibility is to base the IMEX scheme on the explicit two-step Adams method, resulting in the IMEX-Adams2 scheme

$$u_n = u_{n-1} + \frac{3}{2}\Delta t F_{n-1} - \frac{1}{2}\Delta t F_{n-2} + \frac{9}{16}\Delta t G_n + \frac{3}{8}\Delta t G_{n-1} + \frac{1}{16}\Delta t G_{n-1}. \tag{3.7}$$

This scheme has $C = \frac{4}{9}$ and the strongest damping of high-frequency error modes $D = \frac{1}{3}$ among the methods based on the explicit two-step Adams method. The error constants are given by $\widehat{E} = 0.417$ and $E = -0.146$. This method first appeared in [3] and may be viewed as a variant of the popular Crank–Nicholson Adams–Bashforth combination, but with superior damping properties.

The boundaries of the stability domains $\mathscr{D}_{AR}(q)$, $q = 1, 2, 3$, for advection–reaction equations are presented in Fig. 2, with $z = \xi \leqslant 0$ horizontally and Courant number v vertically. Fig. 3 shows the boundaries of the $\mathscr{D}_{AD}(q)$ domains for advection–diffusion with Péclet number μ horizontally and Courant number v vertically, both on logarithmic scale. It is clear from these figures that the linear stability properties of the IMEX-SG(3,2) and the IMEX-BDF2 scheme are very similar. These are more favorable than for either the IMEX-Shu(3,2) scheme or the IMEX-Adams2 scheme.

Finally we note that starting with the explicit two-step Adams method or the extrapolated BDF2 method, corresponding implicit methods can be found that provide larger stability domains, but this would increase the damping factors D and absolute error constants $|E|$.

3.3. Order-three schemes

New schemes arise when orders three and higher are considered. We first consider methods based on the explicit TVD multistep methods of Shu [30]. There are no three-step, third-order explicit methods which are monotone for arbitrary starting values [21,30], but, optimal four- and five-step methods are known [30]. We can construct the corresponding IMEX schemes by optimizing the damping criterion or the combined criteria as described in Section 3.1. The latter approach gives larger stability domains with only a small increase in the asymptotic damping factor D. Therefore we prefer the combined criteria for these schemes.
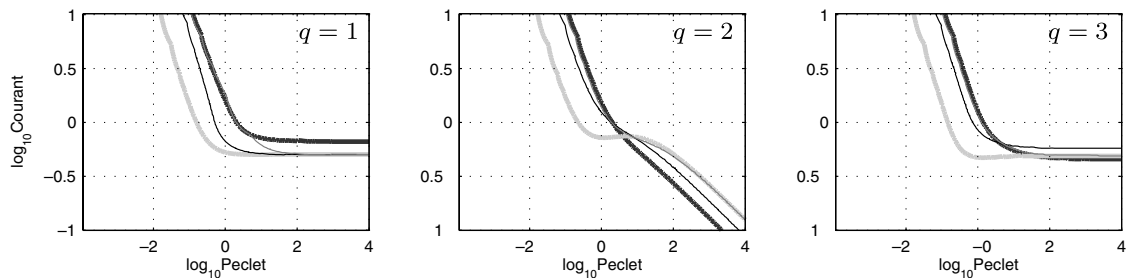

Fig. 3. Boundaries of $\mathscr{D}_{AD}(q)$, $q = 1, 2, 3$, for the second-order methods IMEX-BDF2 (dark gray), IMEX-Adams2 (thin black) and the IMEX-Shu(3,2) extensions (3.4) (light gray) and (3.5) (thin gray; largely coinciding with the light gray line for large μ and the dark gray line for small μ.)

The corresponding third-order four-step IMEX-Shu$(4, 3)$ scheme is given by

$$u_n = \frac{16}{27}u_{n-1} + \frac{11}{27}u_{n-4} + \frac{16}{9}\Delta t F_{n-1} + \frac{4}{9}\Delta t F_{n-4} + \frac{9035}{19683}\Delta t G_n + \frac{13541}{19683}\Delta t G_{n-1} + \frac{1127}{2187}\Delta t G_{n-2}$$
$$+ \frac{7927}{19683}\Delta t G_{n-3} + \frac{3094}{19683}\Delta t G_{n-4}. \tag{3.8}$$

Its characteristic values are $C = 0.333$, $D = 0.779$ and $\widehat{E} = -0.3$, $E = 0.036$.

The third-order five-step IMEX-Shu$(5, 3)$ scheme is

$$u_n = \frac{25}{32}u_{n-1} + \frac{7}{32}u_{n-5} + \frac{25}{16}\Delta t F_{n-1} + \frac{5}{16}\Delta t F_{n-5} + \frac{15863}{32768}\Delta t G_n + \frac{1159}{2048}\Delta t G_{n-1} + \frac{5019}{16384}\Delta t G_{n-2}$$
$$+ \frac{899}{4096}\Delta t G_{n-3} + \frac{6811}{32768}\Delta t G_{n-4} + \frac{187}{2048}\Delta t G_{n-5}. \tag{3.9}$$

For this method we have $C = 0.5$, $D = 0.717$ and $\widehat{E} = -0.556$, $E = 0.64$.

Alternatively, we may start from the explicit TVB$_0(3, 3)$ method from [29], which has an optimal threshold factor among the three-step methods of order three, and construct a compatible implicit method by numerically optimizing D over the corresponding class of implicit three-step linear multistep methods of order three. This yields the IMEX-TVB$_0(3, 3)$ scheme

$$u_n = \frac{3909}{2048}u_{n-1} - \frac{1367}{1024}u_{n-2} + \frac{873}{2048}u_{n-3} + \frac{18463}{12288}\Delta t F_{n-1} - \frac{1271}{768}\Delta t F_{n-2} + \frac{8233}{12288}\Delta t F_{n-3}$$
$$+ \frac{1089}{2048}\Delta t G_n - \frac{1139}{12288}\Delta t G_{n-1} - \frac{367}{6144}\Delta t G_{n-2} + \frac{1699}{12288}\Delta t G_{n-3}. \tag{3.10}$$

In the explicit case $G = 0$ it satisfies the boundedness property (2.3) provided $\Delta t \leqslant C\Delta t_{\text{FE}}$ with $C = 0.536$, and it gives a good damping of high-frequency error modes, $D = 0.639$. The error constants are $\widehat{E} = -0.832$ and $E = 0.195$. Optimization with the combined criteria gave only a slight perturbation of (3.10).

Another IMEX scheme with a favorable boundedness property is the IMEX-BDF3 scheme

$$u_n = \frac{18}{11}u_{n-1} - \frac{9}{11}u_{n-2} + \frac{2}{11}u_{n-3} + \frac{18}{11}\Delta t F_{n-1} - \frac{18}{11}\Delta t F_{n-2} + \frac{6}{11}\Delta t F_{n-3} + \frac{6}{11}\Delta t G_n, \tag{3.11}$$

with $C = \frac{7}{18}$ and $D = 0$. The error constants are $\widehat{E} = -0.75$, $E = 0.25$.

As part of our comparison we also consider the IMEX-Adams3 scheme

$$u_n = u_{n-1} + \frac{23}{12}\Delta t F_{n-1} - \frac{4}{3}\Delta t F_{n-2} + \frac{5}{12}\Delta t F_{n-3} + \frac{4661}{10000}\Delta t G_n + \frac{15551}{30000}\Delta t G_{n-1} + \frac{1949}{30000}\Delta t G_{n-2} - \frac{1483}{30000}\Delta t G_{n-3}, \tag{3.12}$$

with $C = 84/529$, $D = 0.674$, $\widehat{E} = -0.375$ and $E = 0.091$. This is a slight modification of a method introduced in [3].

Among these three-step schemes, the scheme IMEX-TVB$_0(3, 3)$ has the largest C value by a wide margin: it is 38% and 138% larger than the corresponding values for IMEX-BDF3 and IMEX-Adams3. It is also more favorable than for either IMEX-Shu$(4, 3)$ or IMEX-Shu$(5, 3)$, in particular when the number of steps is taken into account, where it should be noted that the threshold value C for the latter schemes refers to the strict monotonicity property (2.2) rather than the boundedness property (2.3).

Another important property of IMEX-TVB$_0(3, 3)$ is that its explicit method has a relatively large linear stability region $\mathscr{S} \in \mathbb{C}$ (where linear stability is valid for the test equation $u' = \hat{\lambda}u$, $\hat{z} = \Delta t\hat{\lambda} \in \mathbb{C}$) that includes a part of the imaginary axis. See Fig. 1 in [29] for a comparison of the stability region of this explicit method with the extrapolated BDF3 method.

The boundaries of the stability domains $\mathscr{D}_{\text{AR}}(q)$ and $\mathscr{D}_{\text{AD}}(q)$, $q = 1, 2, 3$, of the above order-three schemes are shown in Figs. 4 and 5. The best results are obtained for the three-step schemes IMEX-TVB$_0(3, 3)$ and IMEX-BDF3. The stability domains for IMEX-Shu$(5, 3)$ are comparable, but this is already a five-step scheme.
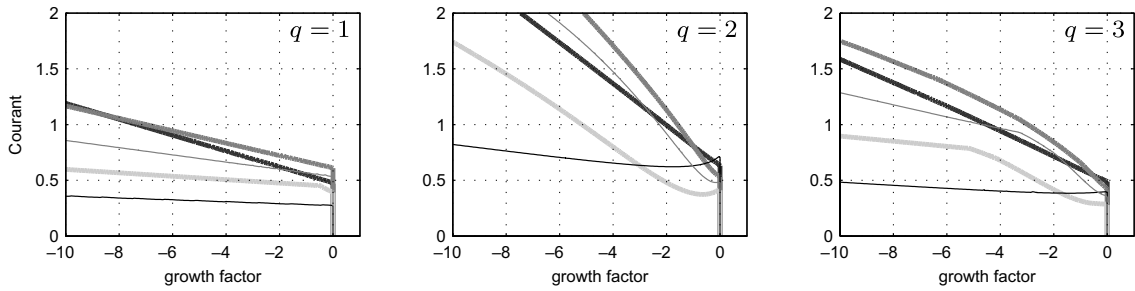
Fig. 4. Boundaries of $\mathscr{D}_{AR}(q)$ for the third-order methods IMEX-BDF3 (dark gray), IMEX-TVB(3,3) (gray), IMEX-Shu(4,3) (light gray), IMEX-Shu(5,3) (thin gray) and IMEX-Adams3 (thin black).
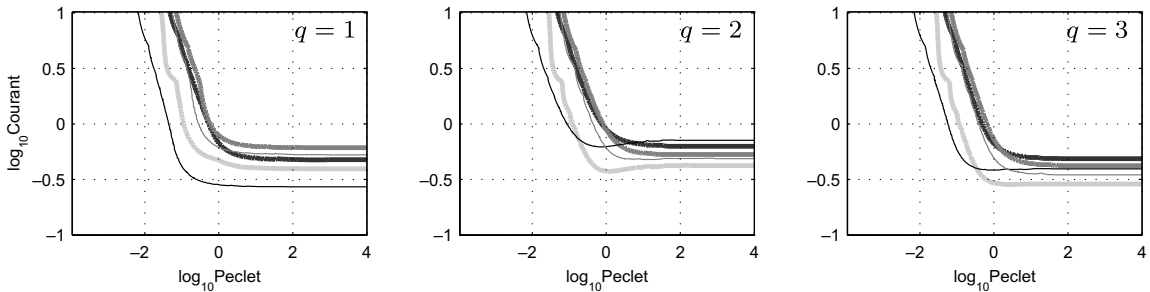

Fig. 5. Boundaries of $\mathscr{D}_{AD}(q)$ for the third-order methods IMEX-BDF3 (dark gray), IMEX-TVB(3,3) (gray), IMEX-Shu(4,3) (light gray), IMEX-Shu(5,3) (thin gray) and IMEX-Adams3 (thin black).

3.4. Order-four schemes

IMEX schemes based on explicit methods with the strict monotonicity property (2.2) become more cumbersome for orders greater than three, since more steps are required to produce reasonable threshold values C. For example, taking six steps, the largest possible C value is about 0.164. Constructing an IMEX scheme with this explicit component by an optimization of the combined criteria leads to the following IMEX-Shu(6,4) scheme,

$$
\begin{aligned}
u_n &= \frac{137}{400} u_{n-1} + \frac{959}{5000} u_{n-4} + \frac{8781}{94000} u_{n-5} + \frac{87487}{235000} u_{n-6} + \frac{976903}{470000} \Delta t F_{n-1} + \frac{136757}{117500} \Delta t F_{n-4} \\
&+ \frac{266997}{470000} \Delta t F_{n-5} + \frac{237}{500} \Delta t G_n + \frac{7547}{10000} \Delta t G_{n-1} + \frac{299}{400} \Delta t G_{n-2} + \frac{4513}{5875} \Delta t G_{n-3} + \frac{118099}{235000} \Delta t G_{n-4} \\
&+ \frac{174527}{470000} \Delta t G_{n-5} + \frac{90349}{470000} \Delta t G_{n-6}.
\end{aligned}
\tag{3.13}
$$

Here $D = 0.880$, $\widehat{E} = 0.236$ and $E = -0.088$. Optimization of the damping factor D alone lead to unacceptably small linear stability domains and only a small decrease in D.

As an alternative, we may take the TVB(4,4) scheme from [29] as the explicit method, and search for an implicit method that minimizes the decay of high-frequency error modes, D. This leads to the scheme

$$
\begin{aligned}
u_n &= \frac{21531}{8192} u_{n-1} - \frac{22753}{8192} u_{n-2} + \frac{12245}{8192} u_{n-3} - \frac{2831}{8192} u_{n-4} + \frac{13261}{8192} \Delta t F_{n-1} - \frac{75029}{24576} \Delta t F_{n-2} + \frac{54799}{24576} \Delta t F_{n-3} \\
&- \frac{15245}{24576} \Delta t F_{n-4} + \frac{4207}{8192} \Delta t G_n - \frac{3567}{8192} \Delta t G_{n-1} + \frac{697}{24576} \Delta t G_{n-2} + \frac{4315}{24576} \Delta t G_{n-3} - \frac{41}{384} \Delta t G_{n-4}.
\end{aligned}
\tag{3.14}
$$

We note that optimization over the combined criteria only lead to a perturbation of this scheme. Key properties of the scheme are that $C = 0.458$, the asymptotic damping factor is given by $D = 0.685$ and the error

constants are $\widehat{E} = 2.386$, $E = -0.544$. It should be noted that these error constants are quite large compared to the other fourth-order schemes. However, for higher-order methods the error constants are less significant than for low-order methods, since the error declines more rapidly as Δt is reduced for a high-order method.

It is natural to compare this IMEX-TVB(4,4) scheme against the fourth-order IMEX-BDF4 scheme

$$u_n = \frac{48}{25}u_{n-1} - \frac{36}{25}u_{n-2} + \frac{16}{25}u_{n-3} - \frac{3}{25}u_{n-4} + \frac{48}{25}\Delta t F_{n-1} - \frac{72}{25}\Delta t F_{n-2} + \frac{48}{25}\Delta t F_{n-3} - \frac{12}{25}\Delta t F_{n-4} + \frac{12}{25}\Delta t G_n.$$

(3.15)

The characteristic values are $C = \frac{7}{32}$, $D = 0$ and $\widehat{E} = 0.8$, $E = -0.2$.

With respect to boundedness (2.3), the IMEX-TVB(4,4) scheme provides for a 109% improvement in allowable step size in the limit $G = 0$ compared to IMEX-BDF4. We further remark that the stability region of the explicit method includes part of the imaginary axis and it compares favorably against that of the extrapolated BDF4 method; see [29, Fig. 1].

Finally we consider the fourth-order explicit Adams method. Optimization of the damping criterion gives the IMEX-Adams4 scheme

$$u_n = u_{n-1} + \frac{55}{24}\Delta t F_{n-1} - \frac{59}{24}\Delta t F_{n-2} + \frac{37}{24}\Delta t F_{n-3} - \frac{9}{24}\Delta t F_{n-4} + \frac{5}{12}\Delta t G_n + \frac{5}{8}\Delta t G_{n-1} + \frac{1}{24}\Delta t G_{n-2}$$
$$- \frac{1}{8}\Delta t G_{n-3} + \frac{1}{24}\Delta t G_{n-4}.$$

(3.16)

For this method we have the unfavorable threshold value $C = 0$. Moreover $D = 1$ only, that is, there is no damping for high-frequency Fourier components. Also the linear stability properties are quite poor. On the other hand, the error constants are relatively small with $\widehat{E} = 0.349$ and $E = -0.068$, but that is unfortunately not sufficient to make it a good scheme. We include this scheme in our tests mainly for comparison.

The boundaries of the stability domains $\mathcal{D}_{AR}(q)$ and $\mathcal{D}_{AD}(q)$, $q = 1, 2, 3$, are shown in the Figs. 6 and 7. We see that also the linear stability properties of the IMEX-TVB(4,4) scheme are the best among these fourth-order schemes, followed by IMEX-BDF4. Due to lack of damping in its implicit method, the IMEX-Adams4 scheme requires small step sizes, i.e. small Courant numbers v, in the diffusion dominated case $\mu \to 0$.

3.5. Order-five schemes

Five-step schemes of order five may also be constructed. Since the step size restrictions that arise from requiring monotonicity for arbitrary starting values become even more severe than for the lower-order schemes we only consider IMEX extensions for explicit methods that satisfy the boundedness property (2.3).

First, consider the TVB$_0$(5,5) method from [29]. A numerical search is carried out over the corresponding class of implicit, five-step, fifth-order methods to determine the coefficients with the best asymptotic damping factor D. This yields the IMEX-TVB$_0$(5,5) scheme
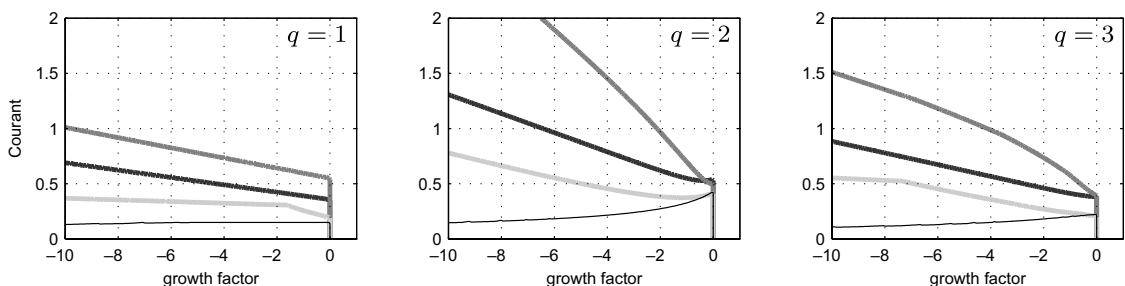

Fig. 6. Boundaries of $\mathcal{D}_{AR}(q)$ for the fourth-order methods IMEX-BDF4 (dark gray), IMEX-TVB(4,4) (gray), IMEX-Shu(6,4) (light gray) and IMEX-Adams4 (thin black).

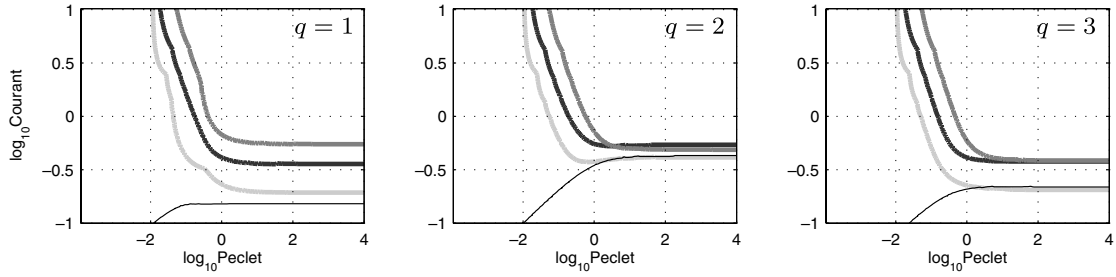Fig. 7. Boundaries of $\mathscr{D}_{AD}(q)$ for the fourth-order methods IMEX-BDF4 (dark gray), IMEX-TVB(4,4) (gray), IMEX-Shu(6,4) (light gray) and IMEX-Adams4 (thin black).

$$u_n = \frac{13553}{4096}u_{n-1} - \frac{38121}{8192}u_{n-2} + \frac{7315}{2048}u_{n-3} - \frac{6161}{4096}u_{n-4} + \frac{2269}{8192}u_{n-5} + \frac{10306951}{5898240}\Delta t F_{n-1}$$
$$- \frac{13656497}{2949120}\Delta t F_{n-2} + \frac{1249949}{245760}\Delta t F_{n-3} - \frac{7937687}{2949120}\Delta t F_{n-4} + \frac{3387361}{5898240}\Delta t F_{n-5} + \frac{4007}{8192}\Delta t G_n$$
$$- \frac{4118249}{5898240}\Delta t G_{n-1} + \frac{768703}{2949120}\Delta t G_{n-2} + \frac{47849}{245760}\Delta t G_{n-3} - \frac{725087}{2949120}\Delta t G_{n-4} + \frac{502321}{5898240}\Delta t G_{n-5}. \quad (3.17)$$

Note that this scheme has a threshold value $C = 0.376$ and a damping parameter $D = 0.709$. The error constants are $\widehat{E} = -4.740$ and $E = 0.976$. Similar to the fourth-order case, a search with the combined criteria only produced a perturbation of this scheme.

As a basis for comparison, we consider the fifth-order IMEX-BDF5 scheme, which is given by

$$u_n = \frac{300}{137}u_{n-1} - \frac{300}{137}u_{n-2} + \frac{200}{137}u_{n-3} - \frac{75}{137}u_{n-4} + \frac{12}{137}u_{n-5} + \frac{300}{137}\Delta t F_{n-1} - \frac{600}{137}\Delta t F_{n-2}$$
$$+ \frac{600}{137}\Delta t F_{n-3} - \frac{300}{137}\Delta t F_{n-4} + \frac{60}{137}\Delta t F_{n-5} + \frac{60}{137}\Delta t G_n. \quad (3.18)$$

For this scheme we have $C = 0.0867$, $D = 0$ and error constants $\widehat{E} = -0.833$, $E = 0.167$.

In terms of the step size restriction for boundedness (2.3), the scheme IMEX-TVB$_0$(5,5) gives a 335% improvement in allowable step size over the IMEX-BDF5 scheme. The linear stability region of the explicit method is also favorable when compared against the extrapolated BDF5 method; see [29, Fig.2]. These explicit methods do not include the imaginary axis near the origin in their stability regions; by a fundamental result of [19] it is known that this is impossible for any five-step method of order five. This explains the behavior in Fig. 9 for $\mathscr{D}_{AD}(q)$ for the central difference case $q = 2$ in the advection dominated case (large cell Péclet numbers μ). Further it is again clear from the Figs. 8 and 9 that the linear stability properties of IMEX-TVB$_0$(5,5) are significantly better than those of IMEX-BDF5.
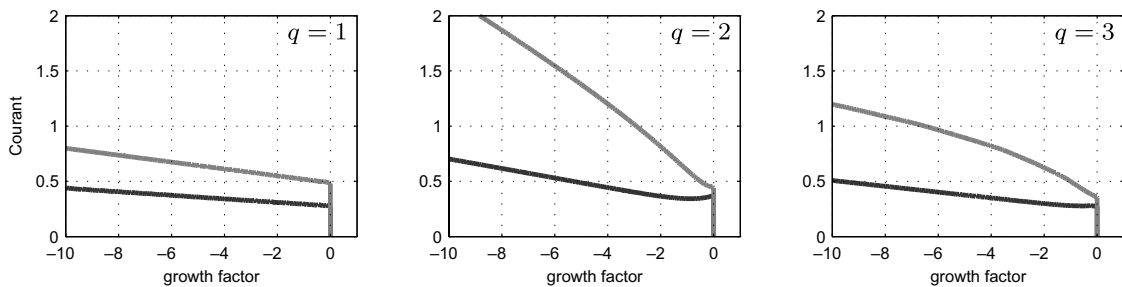

Fig. 8. Boundaries of $\mathscr{D}_{AR}(q)$ for the fifth-order methods IMEX-BDF5 (dark gray) and IMEX-TVB$_0$(5,5) (gray).
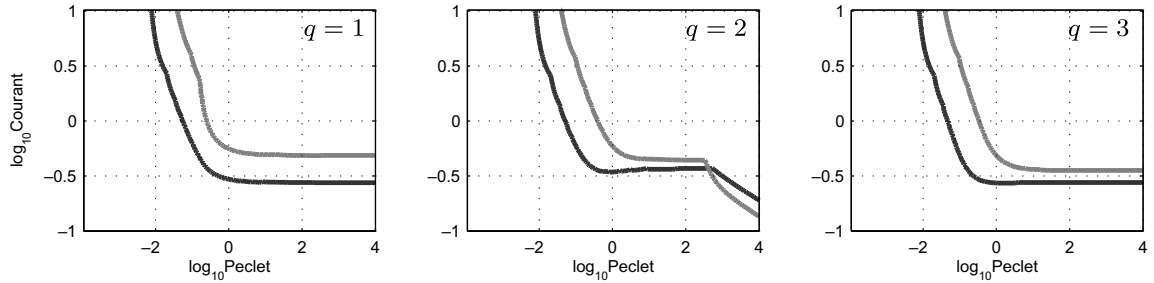
Fig. 9. Boundaries of $\mathscr{D}_{\mathrm{AD}}(q)$ for the fifth-order methods IMEX-BDF5 (dark gray) and IMEX-TVB$_0(5,5)$ (gray).

4. IMEX Runge–Kutta methods

For comparisons we will also consider combinations of s-stage explicit and implicit Runge–Kutta methods, with internal stage vectors $u_{n,i}$, $i = 1, \ldots, s$,

$$
\begin{aligned}
u_{n,i} &= u_{n-1} + \sum_{j=1}^{s} \hat{a}_{ij} \Delta t F(u_{n,j}) + \sum_{j=1}^{s} a_{ij} \Delta t G(u_{n,j}), \\
u_n &= u_{n-1} + \sum_{j=1}^{s} \hat{b}_j \Delta t F(u_{n,j}) + \sum_{j=1}^{s} b_j \Delta t G(u_{n,j}).
\end{aligned}
\tag{4.1}
$$

Similar to the multistep methods, the coefficients of the explicit method are denoted by hats, therefore $\hat{a}_{ij} = 0$ for $j \geqslant i$. Usually the implicit method is chosen from the class of diagonally implicit (DIRK) methods, so then $a_{ij} = 0$ for $j > i$. For compatibility with the explicit method, the DIRK method can be padded with zero coefficients as in [2], so that $u_{n,1} = u_n$ becomes a trivial stage.

Let $c_i = \sum_j a_{ij}$ and $\hat{c}_i = \sum_j \hat{a}_{ij}$. For non-autonomous systems the function evaluations $F(u_{n,j})$, $G(u_{n,j})$ in (4.1) are replaced by $F(t_{n-1} + \hat{c}_j \Delta t, u_{n,j})$ and $G(t_{n-1} + c_j \Delta t, u_{n,j})$. For many IMEX Runge–Kutta methods we have $c_i = \hat{c}_i$ for all $i = 1, \ldots, s$, and then the internal vectors $u_{n,i}$ are consistent approximations to $u(t_{n-1} + c_i \Delta t)$.

However, not all IMEX Runge–Kutta methods in the literature are such that $c_i = \hat{c}_i$. In particular for the so-called asymptotic-preserving methods of Pareschi and Russo [27,28] it is required that $a_{11} \neq 0$ so that the limit from 1.3,1.4 remains valid for any starting value. For these methods $c_1 \neq \hat{c}_1 = 0$.

Test results and comparisons with the IMEX linear multistep schemes will be presented for several state-of-the-art Runge–Kutta combinations from [3,20,28]. The methods are denoted by the triple (s_I, s_E, p), where s_I, s_E are the effective number of stages (not counting trivial stages) of the explicit and implicit methods, respectively, and p is the order of the IMEX Runge–Kutta scheme. For implicit methods with a trivial stage, IMEX combinations with $s_I = s - 1$ occur. Moreover, for some methods $\hat{b}_s = 0$, in which case $s_E = s - 1$. Compared to the IMEX linear multistep schemes, the amount of work per step for an IMEX Runge–Kutta scheme will be roughly s_E times larger if the work of performing F evaluations dominates, and s_I times larger if the work for solving the implicit relations with G dominate. We will take $s_{\mathrm{av}} = \frac{1}{2}(s_E + s_I)$ as an average measure.

The IMEX Runge–Kutta methods used in our comparisons are:

- the methods $(2,2,2)$ and $(3,4,3)$ from Ascher, Ruuth and Spiteri [2], which we refer to as ARS$(2,2,2)$ and ARS$(3,4,3)$;
- the methods IMEX-SSP2$(2,2,2)$, IMEX-SSP3$(4,3,3)$ of Pareschi and Russo [28, Tables 2, 6], based on monotone (SSP) explicit methods, which we refer to as PR$(2,2,2)$ and PR$(4,3,3)$;
- the fourth- and fifth-order methods ARK4(3)6L[2]SA and ARK5(4)8L[2]SA from Kennedy and Carpenter [20, pp. 176,177], which we refer to as KC$(5,6,4)$ and KC$(7,8,5)$.

In Figs. 10 and 11 the scaled stability domains for these schemes are presented. Scaling is done to make the figures comparable to those for the multistep schemes, by taking the amount of work per step into consideration. We therefore replace the Courant number ν by ν/s_{av} and the growth rate $\xi = -c\Delta t$ by ξ/s_{av}.

Fig. 10. Boundaries of scaled domains $\mathscr{D}_{AR}(q)$ for the IMEX Runge–Kutta methods PR(2,2,2) (dark gray), ARS(3,4,3) (gray), PR(4,3,3) (light gray), KC(5,6,4) (thin black) and KC(7,8,5) (thin gray). The scaled domains for ARS(2,2,2) (not shown) are close to PR(2,2,2).
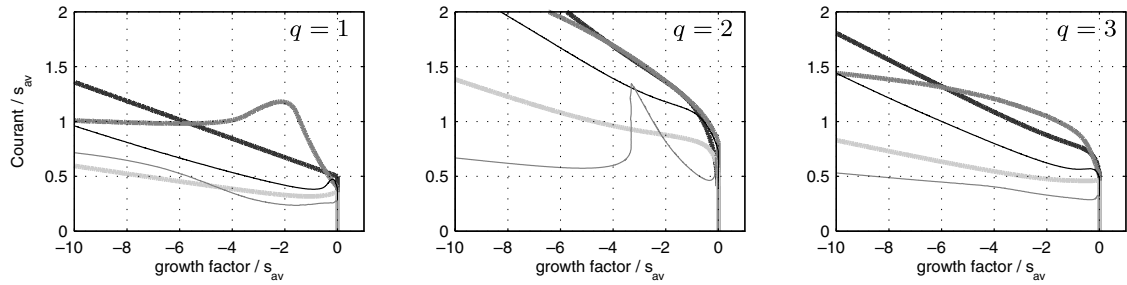

Fig. 11. Boundaries of scaled domains $\mathscr{D}_{AD}(q)$ for the IMEX Runge–Kutta methods PR(2,2,2) (dark gray), ARS(3,4,3) (gray), PR(4,3,3) (light gray), KC(5,6,4) (thin black) and KC(7,8,5) (thin gray). The scaled domains for ARS(2,2,2) (not shown) are close to PR(2,2,2).
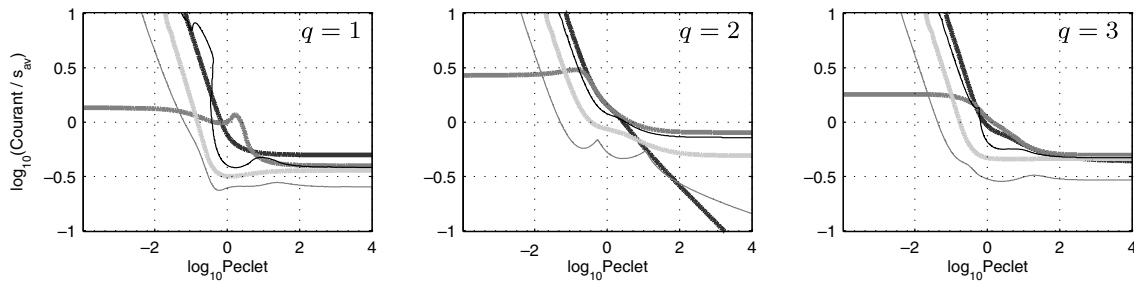
The scaled stability domains for these IMEX Runge–Kutta schemes are in general somewhat smaller than for the optimal multistep schemes of the same order. The linear stability properties of the fifth-order KC(7,8,5) scheme are rather disappointing. The domains $\mathscr{D}_{AD}(q)$ for KC(5,6,4) have a somewhat irregular shape for $q = 1, 3$, but the actual stability bounds are satisfactory for this scheme.

The stability domains for the ARS(3,4,3) scheme differ from all others considered here when large (negative) damping factors or small cell Péclet numbers are considered, in the sense that there remains a stability bound on the Courant numbers v. In the figure for $\mathscr{D}_{AR}(2)$ that is not clear, but it would become visible on larger scale; see also the comments in [6]. In the limit, for damping factor $\xi \to -\infty$ we get for this ARS(3,4,3) scheme the stability restrictions $v \leqslant 6.24$ if $q = 1$, $v \leqslant 12.72$ if $q = 2$, and $v \leqslant 8.40$ if $q = 3$. For the other schemes the stability requirement on v becomes more and more relaxed with increasing $|\xi|$. For KC(7,8,5) the increase of allowable v is quite slow, however.

5. Discretization errors for IMEX methods

5.1. Local and global errors

For an analysis of discretization errors, certain smoothness assumptions are to be made. The discretization errors of the IMEX methods will be expressed in terms of derivatives of $u(t)$ and $\varphi(t) = F(u(t))$. It will be assumed that these derivatives exist and are moderately sized (independent of the stiffness of the problem).

The global errors $e_n = u(t_n) - u_n$ are the quantities of principle interest. For these, a recursion of the form

$$e_n = \sum_{j=1}^{k} R_{n,j} e_{n-j} + d_n \tag{5.1}$$

can be derived, where $k = 1$ for the (one-step) Runge–Kutta methods. Here the matrices $R_{n,j}$ determine the propagation of previous errors and d_n is a new error introduced in this step, commonly called the local discretization error.

In this section we derive expressions for the errors that are not affected by the stiffness of the problem. For this reason, partial derivatives $G^{(j)}(u(t))$ will not be used, only total derivatives such as $\frac{\mathrm{d}^j}{\mathrm{d}t^j}G(u(t)) = u^{(j+1)}(t) - \varphi^{(j)}(t)$.

5.2. Linear multistep methods

For comparison with the structure of errors with Runge–Kutta methods, we first discuss the discretization errors for the IMEX linear multistep methods. Let the coefficients q_l, $\hat{q}_l$ be given by (2.16).

Insertion of the exact solution values in the scheme (1.2) gives a residual r_n, commonly called the local truncation error,

$$u(t_n) = \sum_{j=1}^{k} a_j u(t_{n-j}) + \sum_{j=1}^{k} \hat{b}_j \Delta t F(u(t_{n-j})) + \sum_{j=0}^{k} b_j \Delta t G(u(t_{n-j})) + r_n. \tag{5.2}$$

By Taylor expansion around $t = t_n$ it is seen that

$$r_n = \sum_{l \geqslant p+1} (q_l \Delta t^l u^{(l)}(t_n) + (\hat{q}_l - q_l)\Delta t^l \varphi^{(l-1)}(t_n)). \tag{5.3}$$

Hence, the order conditions (2.17) imply that the truncation error of the IMEX scheme is $\mathcal{O}(\Delta t^{p+1})$ without any matching conditions between the explicit and implicit method. Moreover, the local truncation error is naturally expressed in terms of derivatives of u and φ, only. Therefore, for the linear multistep methods, stiffness does not affect these local errors.

We are of course interested in the global errors $e_n = u(t_n) - u_n$. If we define matrices Z_n, $\widehat{Z}_n$, e.g. by the mean value theorem, such that

$$\Delta t(G(u(t_n)) - G(u_n)) = Z_n e_n, \qquad \Delta t(F(u(t_n)) - F(u_n)) = \widehat{Z}_n e_n,$$

the global errors are found to satisfy the recursion (5.1) with local discretization error

$$d_n = (I - b_0 Z_n)^{-1} r_n. \tag{5.4}$$

Under natural conditions on the ODE system it can be concluded that the norm of this inverse matrix is bounded in a suitable norm by a constant C uniformly in the stiffness of the problem (often $C \leqslant 1$), and then the bound for the residual errors directly leads to a bound on the local discretization errors d_n that is not affected by stiffness.

The amplification matrices in (5.1) are given by

$$R_{n,j} = (I - b_0 Z_n)^{-1}(a_j I + \hat{b}_j \widehat{Z}_{n-j} + b_j Z_{n-j}).$$

If F is a genuinely non-stiff term, then $\widehat{Z}_{n-j} = \mathcal{O}(\Delta t)$, and for stability these contributions can be neglected, in which case stability considerations of the IMEX schemes are the same as for the implicit method; see [14] for instance. In the mildly stiff case we will have $\widehat{Z}_m = \mathcal{O}(1)$ only, and then stability for general nonlinear systems becomes very complicated. In general stability is only studied for linear problems with normal commuting matrices, in which case it is sufficient to consider the scalar linear test Eq. (2.4).

5.3. Runge–Kutta methods

To derive local error expressions for the Runge–Kutta methods it is convenient to view the stage vectors $u_{n,i}$ as approximations to $u(t_{n,i})$, $t_{n,i} = t_{n-1} + c_i \Delta t$ for $i = 1, \ldots, s$. Inserting these exact solution values in the stages we obtain residual errors $r_{n,i}$ and $r_{n,0}$,

$$u(t_{n,i}) = u(t_{n-1}) + \sum_{j=1}^{i-1} \hat{a}_{ij} \Delta t F(u(t_{n,j})) + \sum_{j=1}^{i} a_{ij} \Delta t G(u(t_{n,j})) + r_{n,i},$$

$$u(t_n) = u(t_{n-1}) + \sum_{j=1}^{s} \hat{b}_j \Delta t F(u(t_{n,j})) + \sum_{j=1}^{s} b_j \Delta t G(u(t_{n,j})) + r_{n,0}. \tag{5.5}$$

Let the matrices $A = (a_{ij})$, $\widehat{A} = (\hat{a}_{ij}) \in \mathbb{R}^{s \times s}$ and the vectors $b = (b_i)$, $\hat{b} = (\hat{b}_i)$ and $c = (c_i) \in \mathbb{R}^s$ contain the coefficients of the method. Further, we denote $c^l = (c_i^l) \in \mathbb{R}^s$ for $l \geqslant 1$ and $e = (1, \ldots, 1)^T \in \mathbb{R}^s$. Finally, let $r_n = (r_{n,i}) \in \mathbb{R}^{sM}$. Assume for the moment that $M = 1$ (scalar ODE) for convenience of notation. Then Taylor expansion yields

$$r_n = \sum_{l \geqslant 1} \frac{1}{l!}((c^l - lAc^{l-1})\Delta t^l u^{(l)}(t_{n-1}) + l(Ac^{l-1} - \widehat{A}c^{l-1})\Delta t^l \varphi^{(l-1)}(t_{n-1})), \tag{5.6}$$

where c^0 is defined as $c^0 = e$. In the same way we find

$$r_{n,0} = \sum_{l \geqslant 1} \frac{1}{l!}((1 - lb^T c^{l-1})\Delta t^l u^{(l)}(t_{n-1}) + l(b^T c^{l-1} - \hat{b}^T c^{l-1})\Delta t^l \varphi^{(l-1)}(t_{n-1})). \tag{5.7}$$

Let $Z_n = \mathrm{diag}(Z_{n,i})$, $\widehat{Z}_n = \mathrm{diag}(\widehat{Z}_{n,i})$ with

$$Z_{n,i}(u(t_{n,i}) - u_{n,i}) = \Delta t(G(u(t_{n,i}) - G(u_{n,i})),$$
$$\widehat{Z}_{n,i}(u(t_{n,i}) - u_{n,i}) = \Delta t(F(u(t_{n,i}) - F(u_{n,i})).$$

By subtracting (4.1) from (5.5), and eliminating the internal quantities $u(t_{n,i}) - u_{n,i}$, it follows that $e_n = R_n e_{n-1} + d_n$ with

$$R_n = 1 + s_n^T e, \qquad s_n^T = (\hat{b}^T \widehat{Z}_n + b^T Z_n)(I - \widehat{A}\widehat{Z}_n - AZ_n)^{-1} \tag{5.8}$$

and local error

$$d_n = s_n^T r_n + r_{0,n}. \tag{5.9}$$

In the following it will be tacitly assumed that the norm of s_n^T can be bounded by a constant that is not affected by the stiffness.

For ODE systems (1.1) with dimension $M > 1$ the above formulas remain valid if we replace A, $\widehat{A}$, b^T, $\hat{b}^T$ and e by their Kronecker products with the $M \times M$ identity matrix, to scale them up to the correct dimension.

The *stage order q* of the method is the largest integer such that

$$c^l = lAc^{l-1} = l\widehat{A}c^{l-1} \quad (l = 1, \ldots, q), \tag{5.10}$$

where as before $c^0 = e$. If $Ae = \widehat{A}e$, that is $c = \hat{c}$, then the stage order is one.[2] For the methods in [28] we have $Ae \neq \widehat{A}e$, and then the stage order is only zero.

Let us first assume that $c = \hat{c}$ (stage order one). The usual conditions for order p of the implicit and explicit method imply that

$$1 = lb^T c^{l-1} = l\hat{b}^T c^{l-1} \quad (l = 1, \ldots, p). \tag{5.11}$$

If $p \geqslant 2$ the leading terms in the local discretization error are given by

$$d_n = s_n^T \left(\frac{1}{2}c^2 - Ac\right)\Delta t^2 u''(t_{n-1}) + s_n^T(Ac - \widehat{A}c)\Delta t^2 \varphi'(t_{n-1}) + \mathcal{O}(\Delta t^3). \tag{5.12}$$

This gives consistency of order one, that is, an error $\mathcal{O}(\Delta t^2)$ after one step. Due to damping and cancellation effects, similar as for standard (explicit or implicit) Runge–Kutta methods [5,18,23], this usually leads to convergence of order two. This order will indeed be observed for systems (1.1) where $\widehat{Z}_n = \mathcal{O}(1)$ and $\|Z_n\| \gg 1$. For linear non-stiff problems where both $\widehat{Z}_n, Z_n = \mathcal{O}(\Delta t)$ the classical order conditions for IMEX methods can be recovered by expanding

$$s_n^T = (\hat{b}^T \widehat{Z}_n + b^T Z_n)(I + \widehat{A}\widehat{Z}_n + AZ_n + (\widehat{A}\widehat{Z}_n + AZ_n)^2 + \cdots),$$

together with (5.6) and (5.7). For high-order methods and nonlinear problems it is more convenient, however, to use Butcher trees to derive the local errors in a systematic way; see [6,20].

---

[2] It cannot be larger than one because $c_1 = 0$ and therefore $c_2^2 \neq \hat{a}_{21}c_1$. This is due to the fact that the first non-trivial stage in an explicit method consists of a scaled forward Euler step.

For stiff problems the above derivation shows that we will often have *order reduction*. It should be noted, however, that the error constants in front of the $\mathcal{O}(\Delta t^2)$ global error terms are often quite small, and therefore the classical order p for non-stiff problems is still important if the accuracy requirements are not too strict. This is similar to the situation for implicit Runge–Kutta methods; see for instance [5,18,23,24].

If $\hat{b}^{\mathrm{T}}e = b^{\mathrm{T}}e = 1$ but $\hat{c} \neq c$ (stage order zero), such as occurs for the PR(2, 2, 2) and PR(4, 3, 3) schemes, then the leading term in the local error is

$$d_n = s_n^{\mathrm{T}}(c - \hat{c})\Delta t \varphi(t_{n-1}) + \mathcal{O}(\Delta t^2). \tag{5.13}$$

This first-order local error will already be present for *stationary solutions* of (1.1) for which one normally would not expect any error at all.[3] This can lead to errors which are many orders of magnitude larger than one might otherwise expect. Illustrations will be presented in Section 6.3.

6. Numerical tests

6.1. Population dynamics: positivity test

Our first example evolves a population density P according to

$$P_t = f(t, x) + b(x, P)P - r_d P + dP_{xx} \tag{6.1}$$

on the unit domain $[0, 1]$ using periodic boundary conditions and zero initial conditions $P(0, x) = 0$, $x \in [0, 1]$. In fact, the solution can be extended for negative time by setting $P(t, x) = 0$, $t \leqslant 0$.

This differential equation models the evolution of an ecological population using birth, death and migration. It assumes that the death rate is proportional to the number of individuals. This contributes a term $-r_d P$ to the dynamics where r_d is a constant independent of the number of individuals. The birth rate, on the other hand, is assumed to depend nonlinearly on the number of individuals, as might occur in a system with competition. We take the birth rate, $b(x, P)$, to decline nonlinearly with the population density according to a model appearing in [8]

$$b(x, P) = r_b(x) \frac{\epsilon}{\epsilon + P}$$

where we set $\epsilon = 0.005$. Also appearing in our model is a diffusion term to model migration, and a (marginally resolved) forcing term $f(t, x)$ which could represent additive noise or some forcing effect. All calculations set the mesh spacing equal to $\Delta x = 1/100$ and use second order centered differences to discretize the diffusion term. The stiff diffusion term is treated implicitly while the other terms are treated explicitly.

In our computations the death rate is set equal to one, ie $r_d \equiv 1$. The birth rate r_b varies by position according to

$$r_b(x) = \begin{cases} 1 & \text{if } x \in [0, 1/2], \\ 100 & \text{otherwise.} \end{cases}$$

The marginally resolved forcing term, $f(t, x)$, is taken to be zero for all times $t \neq 0$. At time $t = 0$, and for each grid point, f is assigned a random value in the interval $[0.8, 1.2]$ according to a uniform distribution. Finally, note that we treat models with diffusion ($d \equiv 0.01, 0.04$) and without ($d \equiv 0$). Without diffusion, the population eventually must tend to zero in the first half of the domain. Conversely, the diffusive model tends to an interesting non-zero profile as t increases; see Fig. 12. In all cases, the analytical solution is non-negative for all times $t \geqslant 0$.

The preservation of positivity is particularly desirable in this problem since it leads to more biologically meaningful population densities and avoids the possibility of having reaction terms grow unboundedly (which would happen for P tending to $-\epsilon$). For example, taking $d = 0.01$ and the IMEX-Adams4 method, the unfortunate choice of $\Delta t = 0.0133075079$ leads to a numerical error exceeding 10^4.

---

[3] Note that for a stationary solution u^* of (1.1) we have $F(u^*) + G(u^*) = 0$, but $\varphi = F(u^*) \neq 0$ in general.
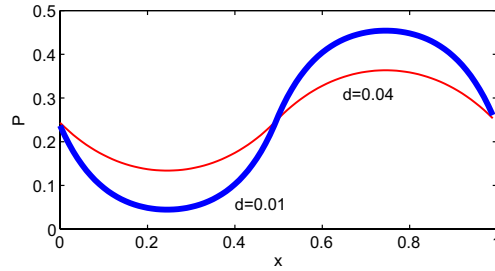
Fig. 12. Steady state population density for the population dynamics model with diffusion constants $d = 0.01$ and $d = 0.04$.

We have experimentally determined the largest step size for which the solution maintains positivity. The results for the multistep schemes are given in Table 1. For comparison, also the IMEX-BDF1 scheme is included, which consists of the forward Euler, backward Euler combination. We find that in each case (both with and without diffusion), the observed critical time step value is in good agreement with the coefficient of monotonicity C. In particular, the proposed schemes IMEX-BDF2, IMEX-TVB$_0(3,3)$, IMEX-TVB$(4,4)$ and IMEX-TVB$_0(5,5)$ are the schemes which exhibit the best positivity for orders 2–5, respectively. Note that all non-negative results also gave very good approximations of the steady state. Specifically, all calculations which maintained positivity varied from the steady state value by less than 1% at time $t = 10$.

The correspondence between the critical time step values and the coefficient of monotonicity is not as predictable for the IMEX Runge–Kutta methods when diffusion is present. Results based on the (non-monotone) Runge–Kutta methods ARS$(2,2,2)$, ARS$(3,4,3)$, KC$(5,6,4)$ and KC$(7,8,5)$ give critical time steps of zero for both the diffusive and non-diffusive cases. Schemes built on SSP Runge–Kutta methods, PR$(2,2,2)$ and PR$(4,3,3)$, also have the property that the coefficient of monotonicity and the critical timestep value are in agreement for $d = 0$. Unfortunately, and as shown in Table 2, these schemes are unable to maintain positivity for time steps comparable to the coefficient of monotonicity when diffusion is present. Given that the multistep schemes only require one function evaluation per step it is clear that the IMEX-TVB schemes are more efficient methods for preserving positivity in this example.

6.2. Van der Pol equation: accuracy test

The van der Pol equation,

Table 1
Critical time step values for positivity: multistep methods

Order	Method	C	Δt ($d = 0$)	Δt ($d = 0.01$)	Δt ($d = 0.04$)
1	IMEX-BDF1	1.000	1.004	1.048	1.145
2	IMEX-Adams2	0.444	0.447	0.445	0.478
	IMEX-SG$(3,2)$	0.500	0.503	0.513	0.563
	IMEX-BDF2	0.625	0.628	0.636	0.686
3	IMEX-Adams3	0.159	0.161	0.152	0.163
	IMEX-BDF3	0.389	0.391	0.390	0.414
	IMEX-Shu$(4,3)$	0.333	0.335	0.330	0.348
	IMEX-Shu$(5,3)$	0.500	0.502	0.502	0.531
	IMEX-TVB$_0(3,3)$	0.537	0.540	0.541	0.575
4	IMEX-Adams4	0	0	0	0
	IMEX-BDF4	0.219	0.221	0.214	0.226
	IMEX-Shu$(6,4)$	0.164	0.166	0.139	0.167
	IMEX-TVB$(4,4)$	0.459	0.461	0.460	0.487
5	IMEX-BDF5	0.087	0.088	0.074	0.082
	IMEX-TVB$_0(5,5)$	0.377	0.379	0.376	0.397

Table 2
Critical time step values for positivity: Runge–Kutta methods

Order	Method	C	Δt $(d = 0)$	Δt $(d = 0.01)$	Δt $(d = 0.04)$
2	PR(2,2,2)	1.000	1.004	0.745	0.745
3	PR(4,3,3)	1.000	1.004	0.498	0.572

$$
\begin{aligned}
y_1' &= y_2, \\
y_2' &= \frac{1}{\epsilon}(-y_1 + (1 - y_1)^2)y_2)
\end{aligned}
\tag{6.2}
$$

is a useful test problem for investigating the order reduction of numerical methods. See, for example, [4,20,25,26] for numerical studies on this equation which illustrate order reduction for a variety of semi-implicit methods. See also [4] for a related error analysis for IMEX Runge–Kutta methods.

A selection of IMEX multistep and Runge–Kutta schemes was applied to this system over the integration interval $[0, 0.5]$. In all cases we set $\epsilon = 10^{-6}$ and treat the first (non-stiff) equation explicitly and the second (stiff) equation implicitly. Initial conditions were chosen to be $y_1(0) = 2, y_2(0) = -0.66666654321$ since this gives errors which are not dominated by the first few steps of the method when $\epsilon = 10^{-6}$ [26]. Starting values for multistep schemes were obtained using the Radau-5 method. This fifth-order implicit Runge–Kutta method treats the van der Pol equation accurately (without order reduction) and in a very stable fashion; see [14] for details.

Plots of the absolute error in y_2 at time $t = 0.5$ against the time-step size appear in Fig. 13(left) for some second-order methods. The corresponding multistep schemes (IMEX-BDF2, IMEX-Shu(3,2), IMEX-SG(3,2) and IMEX-Adams2) all gave a clear second-order convergence. Of the Runge–Kutta schemes considered, ARS(2,2,2) gave the best results. It produced smaller errors than the multistep schemes and did not lead to order reduction. The performance of the PR(2,2,2) scheme, however, was disappointing: It only yielded a first-order convergence and produced much larger errors than any other scheme that was considered.

Results for order-three appear in Fig. 13(right). The multistep schemes all gave third-order convergence, with IMEX-Shu(4,3) giving a somewhat smaller error than other schemes. The IMEX-TVB$_0$(3,3) and IMEX-BDF3 results essentially overlap. Both Runge–Kutta methods exhibited error reduction for this test problem: second-order convergence was observed for ARS(3,4,3) and first-order convergence was observed for PR(4,3,3).

Fig. 14(left) gives the results for fourth-order. As anticipated, the multistep methods all gave fourth-order convergence. Of the schemes considered, the IMEX-Shu(6,4) scheme lead to the smallest errors. Order reduction was observed for the IMEX Runge–Kutta scheme, KC(5,6,4). This method had a first-order error for time steps below 10^{-3}.
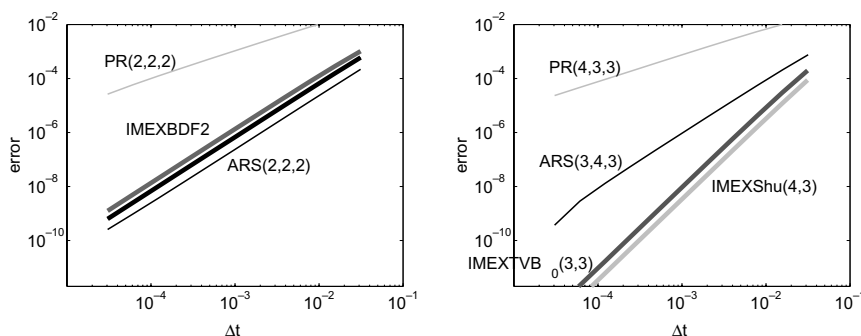

Fig. 13. Errors vs. scaled step size for the van der Pol equation. Left: Second-order methods. The thick black line (unlabeled to avoid cluttering the plot) corresponds to IMEX-Shu(3,2) and IMEX-SG(3,2). The IMEX-Adams2 results lie midway between those for IMEX-Shu(3,2) and IMEX-TVB$_0$(3,3). Right: Third-order methods. IMEX-BDF3 essentially overlaps the results for IMEX-TVB$_0$(3,3). IMEX-Shu(5,3) and IMEX-Adams3 results lie between the results for IMEX-TVB$_0$(3,3) and IMEX-Shu(4,3).
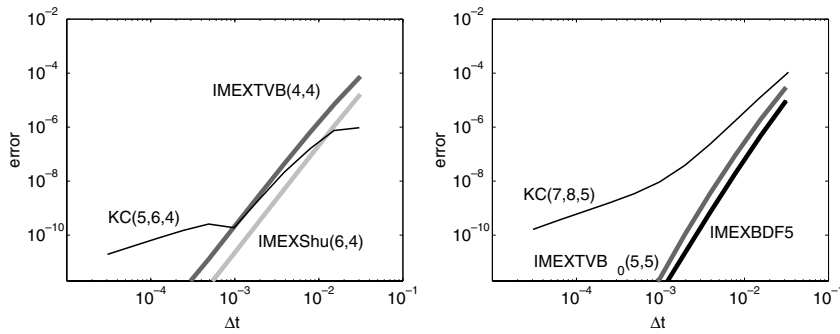
Fig. 14. Errors vs. scaled step size for the van der Pol equation. Left: Fourth-order methods. The IMEX-BDF4 results lie midway between IMEX-TVB(4,4) and IMEX-Shu(6,4). Right: Fifth-order methods.

Fifth-order results appear in Fig. 14(right). Both IMEX-TVB$_0$(5,5) and IMEX-BDF5 gave a clear fifth-order convergence and very small errors. The only Runge–Kutta scheme that was considered, KC(7,8,5), produced a relatively large error that dropped to first-order for time steps below 10^{-3}. Even for large time steps this method gave weaker convergence than the multistep methods.

Taken together, these results suggest that multistep IMEX schemes are particularly appropriate choices for problems that exhibit order reduction. It is also noteworthy that the performance of the proposed IMEX-TVB schemes is competitive with traditional multistep IMEX schemes, even though monotonicity is not an important numerical property for solving this van der Pol problem.

6.3. Linear advection–reaction: accuracy test

As a PDE test for the accuracy of the schemes we consider the linear constant coefficient problem

$$
\begin{aligned}
u_t + \alpha_1 u_x &= -k_1 u + k_2 v + s_1, \\
v_t + \alpha_2 v_x &= k_1 u - k_2 v + s_2
\end{aligned}
\tag{6.3}
$$

for $0 < x < 1$, $0 < t < 1$, with parameters $\alpha_1 = 1$, $\alpha_2 = 0$, $k_1 = 10^6$, $k_2 = 2k_1$ and $s_1 = 0$, $s_2 = 1$. The initial and boundary values are

$$
u(x,0) = 1 + s_2 x, \qquad v(x,0) = \frac{k_1}{k_2} u(x,0) + \frac{1}{k_2} s_2, \qquad u(0,t) = \gamma_1(t).
$$

For γ_1 we will consider two different choices to illustrate possible order reduction effects. Since α_2 is taken to be zero there are no boundary conditions for v.

Note that in the limit $k_1 \to \infty$, k_1/k_2 fixed, we have $k_1 u = k_2 v$, which for $\bar{u} = u + v$ leads to a reduced equation $\bar{u}_t + \bar{a}\bar{u}_x = s_1 + s_2$ with advective velocity $\bar{a} = (k_2\alpha_1 + k_1\alpha_2)/(k_1 + k_2)$.

The spatial discretization in this test is performed on a uniform grid, $x_i = i\Delta x$, $i = 1, \ldots, m$ with $\Delta x = 1/m$. The errors are measured in the discrete L_1-norm ($\|v\|_1 = h\sum_i |v_i|$) at the final time $t = 1$. In the IMEX schemes the advection term is treated explicitly, the stiff reaction implicitly.

First consider $\gamma_1 \equiv 1$. Then the initial values provide a stationary solution. We consider $m = 100$ and first-order upwind spatial discretization, which is exact here. Table 3 gives the L_1-errors of the v-component at the output time $t = 1$ for various step sizes. We see that for this stationary solution there is a strong order reduction for the two PR schemes: they converge with order one only.

For comparison we included results for the ARS(2,2,2) and IMEX-BDF2 methods in this table; the other IMEX schemes exhibit a similar behavior. The results for these methods are not exact. This is due to round-off errors which cause $k_1 u - k_2 v + s_2$ to deviate from the exact value of zero (it is of the order 10^{-10} in Matlab). However, this round-off is rather harmless in comparison to the very disappointing behavior of the PR(2,2,2) and PR(4,3,3) schemes.

With respect to the behavior of these two PR schemes, the convergence results in [4] should be mentioned. In that work, an analysis of IMEX Runge–Kutta methods for a class of singular perturbation problems

Table 3
Stationary solution, $m = 100$, L_1-errors versus step size

Δt	1.00×10^{-2}	5.00×10^{-3}	2.50×10^{-3}	1.25×10^{-3}
PR(2,2,2)	2.36×10^{-3}	1.18×10^{-3}	5.89×10^{-4}	2.93×10^{-4}
PR(4,3,3)	9.47×10^{-4}	4.74×10^{-4}	2.37×10^{-4}	1.18×10^{-4}
ARS(2,2,2)	5.46×10^{-13}	2.06×10^{-13}	1.30×10^{-13}	8.01×10^{-14}
IMEX-BDF2	1.74×10^{-11}	9.40×10^{-12}	1.49×10^{-11}	1.35×10^{-11}

showed that the second-order three-stage method of [28, Table 4] retains its order-two behavior if the stiffness parameter tends to zero, due to stiff accuracy of the implicit method. The linear reaction in this test is also of singular perturbation type, possessing eigenvalues $+0$ and $-(k_1 + k_2)$, and the same stationary test also showed a second-order convergence for this method. However, by reducing the stiffness, setting $k_1 = 10^4$, $k_2 = 10^3$, we find the reaction does not dominate the advection term by much anymore, and then again a slower convergence of order near one is observed. Of course, since this is still for the trivial stationary problem, even order-two convergence would not be very satisfactory.

Next, consider the time dependent Dirichlet data $\gamma_1(t) = 1 - \sin(12t)^4$ at the left boundary, propagating with speed $\bar{a} = 0.66$, approximately, as seen in Fig. 15. Since the initial and boundary values are close to chemical equilibrium, the solution of this problem is smooth. For the spatial discretization we therefore consider fourth-order central differences in the interior domain. At the boundaries we can take third-order finite differences, maintaining an overall accuracy of order four (also in the maximum norm), similar to [18, pp. 155, 156]. Starting values for the multistep schemes are taken to be $u_n = u_0$ for $n < 0$, which is allowed here because the initial solution still can be viewed as a solution with $\gamma_1(t) = 1$ for $t < 0$. The temporal errors are displayed in Fig. 16 for a spatial grid with $m = 400$, giving spatial errors of 1.5×10^{-5}. This spatial error level is indicated by a horizontal dashed line. The time stepping errors are mainly of interest if they are larger than 10% of this spatial error; if they are smaller their contribution to the total error becomes negligible.

For this example we see some order reduction for the IMEX Runge–Kutta methods. In the left picture of Fig. 16 both the ARS(2,2,2) and ARS(3,4,3) scheme show a second-order behavior. For ARS(3,4,3) the smaller error constant is apparent, but, nevertheless, these results are unfavorable in comparison to the IMEX-BDF3 and IMEX-TVB$_0$(3,3) schemes which do not suffer from order reduction effects.

Also the asymptotic error behavior for the two KC schemes is not entirely according to their classical orders. However, in this example with $m = 400$ the order reduction appears on an error level far below the spatial error. (Temporal errors under 10% of the spatial error will not contribute much to the total PDE error.) We did observe, however, that on finer spatial grids the temporal errors remain almost the same provided the step size is sufficiently small to have stability, and therefore this order reduction would be an issue for very high-accuracy computations.

The same test for the IMEX-Adams schemes gave somewhat smaller errors than the corresponding IMEX-BDF and IMEX-TVB schemes, but the IMEX-Adams schemes become unstable sooner for increasing Δt. This effect was particularly pronounced for the four-step scheme, as can be expected from the figures for the stability domains. Also the IMEX-Shu schemes yielded somewhat smaller errors than the IMEX-BDF and
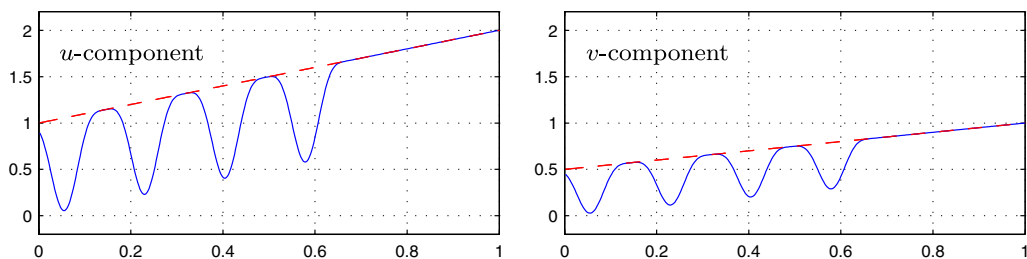

Fig. 15. Components u (left) and v (right) at $t = 1$ for the linear advection–reaction problem with $\gamma_1(t) = 1 - \sin(12t)^4$. The initial solutions are indicated by dashed lines.
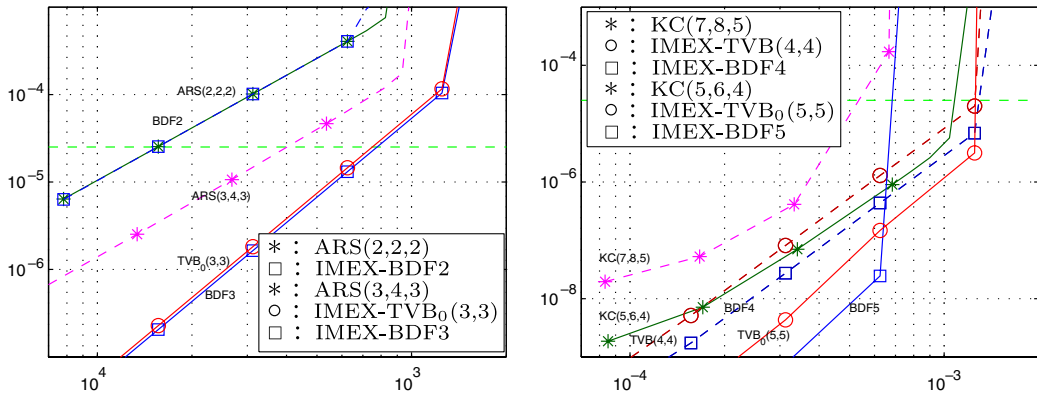
Fig. 16. Temporal L_1-errors vs. scaled step sizes, $m = 400$, spatial error 1.5×10^{-5}. Left: second- and third-order methods ARS(2,2,2), ARS(3,4,3), IMEX-BDF2, IMEX-BDF3 and IMEX-TVB$_0$(3,3). Right: high-order methods KC(5,6,4), KC(7,8,5), IMEX-BDF4, IMEX-BDF5, IMEX-TVB(4,4) and IMEX-TVB$_0$(5,5). Markers $*$ for the IMEX Runge–Kutta methods, $\square$ for the IMEX-TVB methods and ($\bigcirc$) for the IMEX-BDF methods.

IMEX-TVB schemes of the same order, but clearly lagged in performance when compared to IMEX-BDF and IMEX-TVB schemes with the same step-number k. Finally we note that both extensions (3.4) and (3.5) of the explicit Shu(3,2) method were tried, and their results were nearly identical, indicating that the accuracy of the explicit methods in the IMEX combinations is of primary relevance here.

6.4. Adsorption–desorption: accuracy test with spatial WENO discretization

As a last test, we consider a simplified adsorption–desorption problem with a dissolved concentration u and adsorbed concentration v. The equations are given by

$$
\begin{aligned}
u_t + a u_x &= \kappa(v - \phi(u)), \\
v_t &= -\kappa(v - \phi(u))
\end{aligned}
\tag{6.4}
$$

for $0 < x < 1$ and $0 < t < T$, with $\phi(u) = k_1 u/(1 + k_2 u)$. The initial values are $u = v = 0$, and the boundary values are

$$
\begin{cases}
u(0,t) = 1 - \cos^2(6\pi t) & \text{if } a > 0, \\
u(1,t) = 0 & \text{if } a < 0.
\end{cases}
$$

The parameters are taken as $\kappa = 10^6$, $k_1 = 50$, $k_2 = 100$, the final time is $T = \frac{5}{4}$, and the velocity is set to be $a = -3 \arctan(100(t-1))/\pi$, giving approximately $a = 1.5$ for $t < 1$ (adsorption phase) and $a = -1.5$ for $t > 1$ (desorption phase). A similar problem was considered in [18] with constant inflow conditions. Here we take an oscillatory inflow condition to get some smooth variations in the solution, along with the shocks, as shown in Fig. 17.

Again, this problem is of the form (1.3). In limit $\kappa \to \infty$ it holds that $v = \phi(u)$, which gives a reduced (scalar) equation for $\bar{u} = u + \phi(u)$ of the form $\bar{u} + a\bar{f}(\bar{u})_x = 0$, with $\bar{f}$ implicitly defined by $\bar{f}(u + \phi(u)) \equiv u$.

For the spatial discretization we use the WENO5 scheme, as given for instance in the review paper [31] (formulas (2.58)–(2.63) with parameter $\epsilon = 10^{-12}$), on a uniform (cell centered) grid, $x_i = (i - \frac{1}{2})\Delta x$, $i = 1, \ldots, m$, with mesh width $\Delta x = 1/m$. This WENO5 spatial scheme provides high accuracy in smooth regions together with good monotonicity properties near shocks.

In the IMEX schemes the advection has been treated explicitly, the stiff relaxation term implicitly (with a Newton iteration, where care must be taken to end up on the correct, positive branch). We compare the numerical values for the total concentration at the final time $t = T$ with an accurate reference solution. As before, the errors are measured in the discrete L_1-norm. The starting values for the multistep schemes were taken as $u_n = u_0$ for $n < 0$. The alternative, using an IMEX Runge–Kutta method of the appropriate order
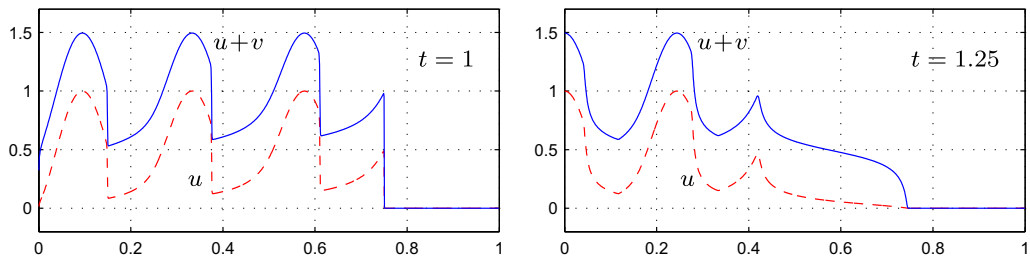
Fig. 17. Dissolved concentration u (dashed lines) and total concentration $u + v$ (solid lines) for the adsorption–desorption problem at time $t = 1$ (left) and $t = \frac{5}{4}$ (right).

to compute $u_1, \ldots, u_{k-1}$ gave nearly identical results in the error range considered here. The step size sequence in this test was taken as $\Delta t = 2^{-j}\Delta x$ (results indicated by large markers) and near instabilities extra step sizes were added (small markers). The step sizes for the Runge–Kutta schemes were scaled as before. Fig. 18 shows the temporal errors for the second- and third-order IMEX schemes as function of the (scaled) step size. The results for the higher-order schemes are shown in Fig. 19.

In Fig. 18, the results for the PR(2, 2, 2), ARS(2, 2, 2) and IMEX-BDF2 schemes largely coincide. The errors for the IMEX-Adams2 and IMEX-Shu(3, 2) schemes are somewhat smaller, due to smaller error constants. As before, the two extensions (3.4) and (3.5) of the explicit Shu(3, 2) method gave nearly identical results.

Compared to IMEX-TVB(3, 3) and IMEX-BDF3, the third-order IMEX-Shu(4, 3) and IMEX-Shu(5, 3) schemes again yielded somewhat smaller errors but they are more quickly unstable for increasing step sizes. Moreover, since these IMEX-Shu schemes are four- and five-step schemes, their errors could be compared to the corresponding higher-order IMEX-TVB and IMEX-BDF schemes and then there would be a clear advantage for the latter ones.

The high accuracy of the ARS(3, 4, 3) scheme in this test is remarkable. That is due to the fact that the underlying explicit method is of order four for linear problems. Even though the WENO5 formulas will produce a nonlinear advection discretization, the fourth-order accuracy is well maintained. For easier comparison the ARS(3, 4, 3) results are again displayed in Fig. 19, where it can be seen that the asymptotic error behavior is not entirely fourth-order; cf. the nearby results of IMEX-BDF4 and IMEX-TVB(4, 4).

The IMEX-Adams4 scheme needs small step sizes to prevent instabilities, and even then the error behavior is somewhat irregular. We did observe that this irregular behavior can be improved by using a high-order Runge–Kutta starting procedure together with a re-start at $t = 1$, where the velocity rapidly changes sign.

Further is it clear in Fig. 19 that both KC schemes are not doing too well in this test, due to instabilities. Once the step sizes are small enough for stability, then the errors become immediately very small, in particular for the KC(7, 8, 5) scheme. Nevertheless, comparison with IMEX-TVB$_0$(5, 5) and IMEX-BDF5 is unfavorable for KC(7, 8, 5) in this test.

Finally we note that the spatial error is here 1.2×10^{-3}. Therefore temporal errors below 10^{-4} ($\approx 10\%$ of the spatial error) are not very relevant anymore from the PDE point of view. In this respect, the results for IMEX-TVB$_0$(5, 5) are slightly better than those for the IMEX-BDF5 scheme.

6.5. Remarks on starting procedures and variable step sizes

In the above tests Runge–Kutta starting procedures were occasionally used for the linear multistep methods. Another popular procedure is to embed the multistep methods within a self-starting collection of multistep methods. For this we would start with very small step sizes and compute u_1 with a +1-step method, say the IMEX-BDF1 scheme consisting of the forward Euler, backward Euler combination, then u_2 by a +2-step method, and so on, while increasing the step size. This procedure requires the use of variable step sizes. Variable steps are also important of course when the temporal smoothness of the solution changes.

The fixed step size formulas can easily be adjusted for variable steps $\Delta t_n = t_n - t_{n-1}$ by using suitable interpolation. Consider for the step from $t_{n-k}, \ldots, t_{n-1}$ to t_n the fixed step size formula
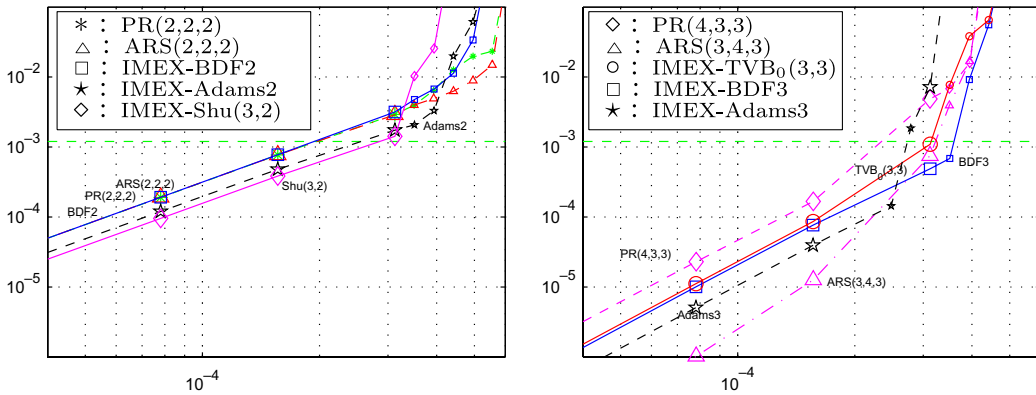
Fig. 18. Temporal L_1-errors vs. scaled step sizes $\in (4 \times 10^{-5}, 6 \times 10^{-4})$, $m = 800$, spatial error 1.2×10^{-3}. Left: second-order methods ARS(2,2,2), PR(2,2,2), IMEX-BDF2, IMEX-Adams2 and IMEX-Shu(3,2). The results for ARS(2,2,2), PR(2,2,2), IMEX-BDF2 are virtually the same for small step sizes. Right: third-order methods ARS(3,4,3), PR(4,3,3), IMEX-BDF3, IMEX-TVB$_0$(3,3) and IMEX-Adams3.
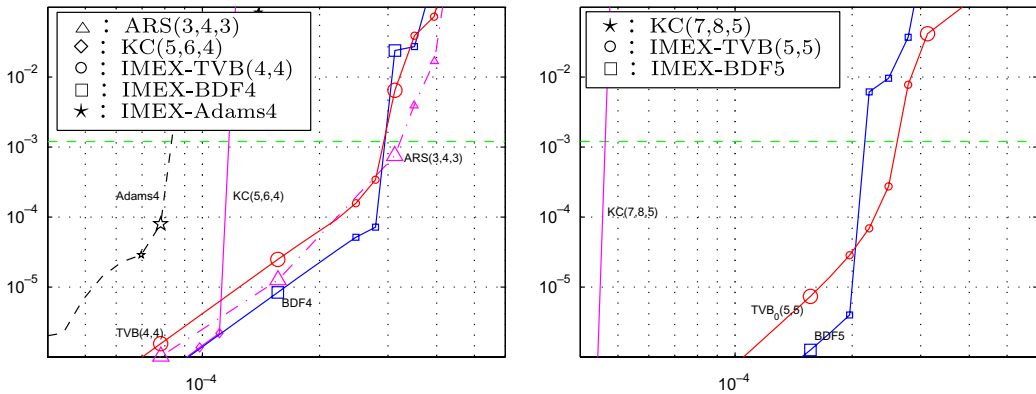

Fig. 19. Temporal L_1-errors vs. scaled step sizes $\in (4 \times 10^{-5}, 6 \times 10^{-4})$, $m = 800$, spatial error 1.2×10^{-3}. Left: fourth-order methods KC(5,6,4), IMEX-BDF4, IMEX-TVB(4,4) and IMEX-Adams4. For comparison, results of the third-order method ARS(3,4,3) are added. Right: fifth-order methods KC(7,8,5), IMEX-BDF5 and IMEX-TVB$_0$(5,5). Markers for KC(7,8,5) are outside the displayed range.

$$u_n - b_0 \Delta t_n G_n = \sum_{j=1}^{k} (a_j u_{n-j}^* + \hat{b}_j \Delta t_n F_{n-j}^* + b_j \Delta t_n G_{n-j}^*), \tag{6.5}$$

together with interpolation formulas to map the values u_{n-j}, F_{n-j} and G_{n-j} from the non-uniform grid to the uniformly distributed points $t_{n-j}^* = t_n - j\Delta t_n$, $j = 1, \ldots, k$. Actually, for a growing step size sequence, some of the t_{n-j}^* will be smaller than t_{n-k}, so formally we are then applying extrapolation.

For the interpolation of u we use the $(k+1)$st-order formula involving all u_{n-j}, including the (still unknown) value u_n. For F and G we use the kth-order formula using the values F_{n-j}, G_{n-j} with $j = 1, \ldots, k$. It is easily seen that if the fixed step size method has order k, then this variable step size implementation will have the same order.

Example 6.1. Consider a non-uniform grid with step size ratios $\theta_n = \Delta t_n / \Delta t_{n-1}$. The two-step method with fixed step size coefficients

$$u_n = a_1 u_{n-1} + a_2 u_{n-2}^* + \hat{b}_1 \Delta t_n F_{n-1} + \hat{b}_2 \Delta t_n F_{n-2}^* + b_0 \Delta t_n G_n + b_1 \Delta t_n G_{n-1} + b_2 \Delta t_n G_{n-2}^* \tag{6.6}$$

can be used with (linear interpolation)

$$F^*_{n-2} = (1 - \theta_n)F_{n-1} + \theta_n F_{n-2}, \qquad G^*_{n-2} = (1 - \theta_n)G_{n-1} + \theta_n G_{n-2} \tag{6.7}$$

and (quadratic interpolation)

$$u^*_{n-2} = -\frac{1 - \theta_n}{1 + \theta_n}u_n + 2\frac{1 - \theta_n^2}{1 + \theta_n}u_{n-1} + 2\frac{\theta_n^2}{1 + \theta_n}u_{n-2}. \tag{6.8}$$

The resulting scheme can be rewritten in the variable coefficient form

$$u_n = a_{1,n}u_{n-1} + a_{2,n}u_{n-2} + \hat{b}_{1,n}\Delta t_n F_{n-1} + \hat{b}_{2,n}\Delta t_n F_{n-2} + b_{0,n}\Delta t_n G_n + b_{1,n}\Delta t_n G_{n-1} + b_{2,n}\Delta t_n G_{n-2}. \tag{6.9}$$

Denoting $c_n = (1 + \theta_n) + (1 - \theta_n)a_2$, these variable coefficients are given by

$$a_{1,n} = \frac{(1 + \theta_n)a_1 + 2(1 - \theta_n^2)a_2}{c_n}, \quad a_{2,n} = \frac{2\theta_n^2 a_2}{c_n}, \quad \hat{b}_{1,n} = \frac{(1 + \theta_n)\hat{b}_1 + (1 - \theta_n)\hat{b}_2}{c_n},$$

$$\hat{b}_{2,n} = \frac{\theta_n(1 + \theta_n)\hat{b}_2}{c_n}, \quad b_{0,n} = \frac{(1 + \theta_n)b_0}{c_n}, \quad b_{1,n} = \frac{(1 + \theta_n)b_1 + (1 - \theta_n)b_2}{c_n}, \quad b_{2,n} = \frac{\theta_n(1 + \theta_n)b_2}{c_n}.$$

For a detailed study of variable step size IMEX formulas we refer to [33]. A proper implementation of an IMEX multistep package will involve many operational strategies. It is part of our future research plans to construct such packages based on the IMEX-BDF and IMEX-TVB schemes discussed in this paper.

7. Conclusions

In this paper, high-order implicit–explicit linear multistep methods have been constructed based on explicit methods which exhibit good monotonicity and boundedness properties. The implicit methods have been chosen to give a strong decay of high-frequency error modes, and mild step size restrictions for the linear test problems for advection, diffusion and reactions. Our methods have been studied for a variety of examples, and compared with recent implicit–explicit Runge–Kutta methods.

The k-step IMEX-Adams schemes, based on the explicit Adams–Bashforth methods, have small error constants, but are not generally recommended because of their relatively poor linear stability and monotonicity properties. This poor behavior is particularly pronounced with increasing step number k. For example the IMEX-Adams4 method has a monotonicity threshold of 0, and we were unable to maintain positivity in the population dynamics model for any positive step size. In the adsorption–desorption problem, this method became unstable for much smaller timesteps than any other fourth-order IMEX multistep scheme we considered.

Much better results were obtained for the IMEX-BDF schemes. These schemes give a strong (indeed optimal) decay of high-frequency error modes with moderate error constants, and good linear stability. The relative performance in terms of monotonicity varies by k. The second-order IMEX-BDF2 scheme gave the mildest monotonicity restriction of any of the second-order schemes under consideration. This property, combined with its strong decay of high-frequency error modes, good linear stability properties and moderate error constants make IMEX-BDF2 a very versatile scheme. The IMEX-BDF3 and IMEX-BDF4 schemes also had reasonable monotonicity properties, although their monotonicity thresholds are significantly smaller than for the corresponding IMEX-TVB$_0$(3,3) and IMEX-TVB(4,4) schemes. The IMEX-BDF5 scheme may be of somewhat more limited use, however, as both its monotonicity and linear stability properties were far more restrictive than in the third- and fourth-order cases.

The IMEX schemes based on the explicit SSP/TVD methods of Shu often gave good performance when compared to methods of the same order. However, when compared to methods using the same number of steps one finds that a far better performance can be obtained by choosing either an IMEX-BDF or IMEX-TVB scheme of higher order.

For orders three to five, the optimal monotonicity was obtained using the IMEX-TVB schemes. These schemes allowed timesteps which were 38%, 109%, and 335% larger than the corresponding third-, fourth- and fifth-order IMEX-BDF schemes. The schemes also give some damping of high-frequency error modes,

and exhibit very good linear stability (typically superior to the IMEX-BDF schemes). The error constants of the IMEX-TVB schemes of order four and five are somewhat larger than with the IMEX-BDF schemes, however.

Finally, we note that our numerical tests show that IMEX multistep schemes have a number of advantages over recent IMEX Runge–Kutta schemes. Low stage orders can degrade the formal order of accuracy in IMEX Runge–Kutta methods, which is a behavior not exhibited by IMEX multistep methods. Moreover, the complicated nature of the local discretization errors with the Runge–Kutta schemes makes it difficult to provide good error estimations for stiff problems.

References

[1] G. Akrivis, M. Crouzeix, C. Makridakis, Implicit–explicit multistep methods for quasilinear parabolic equations, Numer. Math. 82 (1999) 521–541.
[2] U.M. Ascher, S.J. Ruuth, R.J. Spiteri, Implicit–explicit Runge–Kutta methods for time-dependent partial differential equations, Appl. Numer. Math. 25 (1997) 151–167.
[3] U.M. Ascher, S.J. Ruuth, B. Wetton, Implicit–explicit methods for time-dependent PDE's, SIAM J. Numer. Anal. 32 (1995) 797–823.
[4] S. Boscarino, Error analysis of IMEX Runge–Kutta methods derived from differential-algebraic systems. Preprint, 2006.
[5] P. Brenner, M. Crouzeix, V. Thomée, Single step methods for inhomogeneous linear differential equations, RAIRO Anal. Numer. 16 (1982) 5–26.
[6] M.P. Calvo, J. de Frutos, J. Novo, Linearly implicit Runge–Kutta methods for advection–diffusion–reaction problems, Appl. Numer. Math. 37 (2001) 535–549.
[7] M. Crouzeix, Une méthode multipas implicite–explicite pour l'approximation des équations d'évolution paraboliques, Numer. Math. 35 (1980) 257–276.
[8] R. de Boer, Theoretical Biology, Department of Theoretical Biology, Utrecht University, Netherlands, 2006. <http://theory.bio.uu.nl/rdb/books/tb.pdf>.
[9] L. Ferracina, M.N. Spijker, Computing optimal monotonicity preserving Runge–Kutta methods. Preprint University of Leiden, 2005.
[10] J. Frank, W. Hundsdorfer, J.G. Verwer, On the stability of implicit–explicit linear multistep methods, Appl. Numer. Math. 25 (1997) 193–205.
[12] T. Gjesdal, Implicit–explicit methods based on strong stability preserving multistep time discretizations. Preprint, revised version, 2005.
[13] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I – Nonstiff Problems, Springer Series in Computational Mathematics, second edition, 8, Springer, 1993.
[14] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems, Springer Series in Computational Mathematics, second edition, vol. 14, Springer, 1996.
[15] I. Higueras, Strong stability for additive Runge–Kutta methods, SIAM J. Numer. Anal. 44 (2006) 1735–1758.
[16] W. Hundsdorfer, S.J. Ruuth, On monotonicity and boundedness properties of linear multistep methods, Math. Comp. 75 (2006) 655–672.
[17] W. Hundsdorfer, S.J. Ruuth, R.J. Spiteri, Monotonicity-preserving linear multistep methods, SIAM J. Numer. Anal. 41 (2003) 605–623.
[18] W. Hundsdorfer, J.G. Verwer, Numerical Solution of Time-Dependent Advection–Diffusion–Reaction Equations, Springer Series in Computational Mathematics, 33, Springer, 2003.
[19] R. Jeltsch, O. Nevanlinna, Stability of explicit time discretizations for solving initial value problems, Numer. Math. 37 (1981) 61–91.
[20] C.A. Kennedy, M.H. Carpenter, Additive Runge–Kutta schemes for convection–diffusion–reaction equations, Appl. Numer. Math. 44 (2003) 139–181.
[21] H.W.J. Lenferink, Contractivity preserving explicit linear multistep methods, Numer. Math. 55 (1989) 213–223.
[22] R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002.
[23] Ch. Lubich, A. Ostermann, Runge–Kutta approximation of quasilinear parabolic equations, Math. Comp. 64 (1995) 601–627.
[24] Ch. Lubich, A. Ostermann, Interior estimates for time discretization of parabolic equations, Appl. Numer. Math. 18 (1995) 241–251.
[25] M.L. Minion, Semi-implicit spectral deferred correction methods for ordinary differential equations, Appl. Numer. Math. 48 (2003) 369–387.
[26] A.T. Layton, M.L. Minion, Implications of the choice of quadrature nodes for Picard Integral Deferred Corrections Methods for Ordinary Differential Equations, BIT 45 (2005) 341–373.
[27] L. Pareschi, G. Russo, High order asymptotically strong-stability-preserving methods for hyperbolic systems with stiff relaxation, Hyperbolic Problems: Theory, Numerics, Applications, Springer, 2003, pp. 241–251.
[28] L. Pareschi, G. Russo, Implicit–explicit Runge–Kutta methods and applications to hyperbolic systems with relaxation, J. Sci. Comp. 25 (2005) 129–155.
[29] S.J. Ruuth, W. Hundsdorfer, High-order linear multistep methods with general monotonicity and boundedness properties, J. Comput. Phys. 209 (2005) 226–248.

[30] C.-W. Shu, Total-variation-diminishing time discretizations, SIAM J. Sci. Stat. Comp. 9 (1988) 1073–1084.
[31] C.-W. Shu, High order ENO and WENO schemes for computational fluid dynamics, in: High-Order Methods for Computational Physics, in: T.J. Barth, H. Deconinck (Eds.), Lecture Notes in Computational Science Engineering, vol. 9, Springer, 1999, pp. 439–582.
[32] J.M. Varah, Stability restrictions on second order, three-level finite-difference schemes for parabolic equations, SIAM J. Numer. Anal. 17 (1980) 300–309.
[33] D. Wang, S. Ruuth, Variable step-size implicit–explicit linear multistep methods for time-dependent partial differential equations. Preprint, 2006.