COMMUN. MATH. SCI.
Vol. 22, No. 3, pp. 613–653

SPATIAL MANIFESTATIONS OF ORDER REDUCTION IN RUNGE-KUTTA METHODS FOR INITIAL BOUNDARY VALUE PROBLEMS*

RODOLFO RUBEN ROSALES†, BENJAMIN SEIBOLD‡, DAVID SHIROKOFF§, AND DONG ZHOU¶

Abstract. This paper studies the spatial manifestations of order reduction that occur when time-stepping initial-boundary-value problems (IBVPs) with high-order Runge-Kutta methods. For such IBVPs, geometric structures arise that do not have an analog in ODE IVPs: boundary layers appear, induced by a mismatch between the approximation error in the interior and at the boundaries. To understand those boundary layers, an analysis of the modes of the numerical scheme is conducted, which explains under which circumstances boundary layers persist over many time steps. Based on this, two remedies to order reduction are studied: first, a new condition on the Butcher tableau, called weak stage order, that is compatible with diagonally implicit Runge-Kutta schemes; and second, the impact of modified boundary conditions on the boundary layer theory is analyzed.

Keywords. Initial-Boundary-Value problem; time-stepping; Runge-Kutta; order reduction; boundary layer; stage order; weak stage order; modified boundary conditions.

AMS subject classifications. 65L20; 65M15; 34E05.

1. Introduction

Runge-Kutta (RK) methods advance a time-dependent differential equation forward in time by means of multiple stages. Each stage corresponds to one right-hand side evaluation or solve, and appropriate linear combinations of those evaluations generate a high order of accuracy. Two particular advantages of RK schemes over alternatives, such as multistep schemes, are their locality in time and their stability properties [19]. In particular, for stiff problems, many types of high-order implicit RK (IRK) methods exist that are A-stable.

Drawbacks of RK methods are their computational cost per time step, as well as order reduction: when applied to certain stiff problems, the observed order of accuracy of the numerical solution may be lower than the (formal) order of the scheme. While order reduction can be rationalized for ordinary differential equations (ODEs) in terms of stiff limits [41,51], for initial boundary value problems (IBVPs) geometric features in the spatial error play a key role. The specific focus of this paper is: (a) a modal analysis and geometric (via singular perturbation theory) understanding of the global-in-time spatial error, including the accuracy of gradients; (b) the impact of weak stage order (WSO)—a new condition on RK schemes that remedies order reduction—and modified boundary conditions on the spatial error and by what means these properties remedy

*Received: October 26, 2019; Accepted (in revised form): August 11, 2023. Communicated by Giovanni Russo.

†Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 USA (rrr@math.mit.edu).

‡Department of Mathematics, Temple University, 1805 North Broad Street, Philadelphia, PA 19122 USA (seibold@temple.edu). http://www.math.temple.edu/~seibold

§Department of Mathematical Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 USA (david.g.shirokoff@njit.edu). https://web.njit.edu/~shirokof

¶Department of Mathematics, California State University Los Angeles, 5151 State University Drive, Los Angeles, CA 90032 USA (dzhou11@calstatela.edu).

or alleviate order reduction. Specifically, we consider problems of the form

$$\begin{cases} u_t = \mathcal{L}u + f & \text{in} \quad \Omega \times (0, t_{\mathrm{f}}), \\ \mathcal{B}u = g & \text{on} \quad \partial\Omega \times [0, t_{\mathrm{f}}], \\ u = u_0 & \text{on} \quad \Omega \times \{t = 0\}, \end{cases} \tag{1.1}$$

where $\mathcal{L}$ is a linear differential operator, and $\mathcal{B}$ is a boundary operator. Most of the presentation/analysis in this paper focuses on (1.1) with Dirichlet boundary conditions (b.c.), a linear, second-order operator $\mathcal{L}$ (e.g., diffusion), and $\Omega = (0,1)$, because those simple situations suffice to establish the fundamental spatial manifestation of order reduction. However, the order reduction phenomenon arises similarly in higher dimensions, for other types of b.c. (see Section 6), and other differential operators (see Subsection 5.5), albeit with additional effects that are not studied here. Note, though, that many of the structural results and techniques developed in this paper (particularly Section 3) transfer to more general situations.

Order reduction for IBVPs incurs some fundamental differences to the stiff ODE case, most prominently: (i) time discretizations of (1.1) are formally infinitely stiff (i.e., eigenvalues of $\mathcal{L}$ may be arbitrarily large in magnitude); (ii) for IBVPs, spatial derivatives of the solution may be important and also exhibit order reduction; and (iii) boundary conditions play a crucial role in the manifestation of order reduction for IBVPs. In particular, the naive thing to do for a RK method is to impose the b.c. for the PDE at the times t_i associated with the stages, i.e., $u_i = g_i = g(t_i)$ in the case of Dirichlet b.c.. These *conventional b.c.* let the error vanish at the boundary, yet lead to the paradoxical situation that for IBVPs, RK schemes may *lose* accuracy because the approximation is *too accurate* near the boundary. As we will show below, the effect of conventional b.c. will give rise to a singularly perturbed problem for the spatial numerical error and generate boundary layers (BLs).

A crucial property of the order reduction phenomenon studied here is that the loss of convergence order is caused solely by the time discretization. Therefore, the analysis in this paper focuses on semi-discrete problems, where only time is discretized, but space is left continuous; and likewise, all numerical examples are conducted with an extremely fine spatial resolution. This is feasible, as we restrict to schemes that are unconditionally stable when they are applied to problem (1.1). The restriction to the semi-discrete case has an important implication: the order reduction phenomenon cannot be simply overcome by the choice of a specific spatial discretization; any spatial discretization that converges (as $\Delta x \to 0$) to the semi-discrete limit will encounter the order reduction phenomenon studied here.

1.1. A simple example IBVP. Here we demonstrate (a) that order reduction can occur with straightforward schemes (e.g., DIRK), applied to simple problems (e.g., the 1D heat equation); and (b) how it manifests spatially. The only aspect that is strictly needed is that the problem has time-dependent forcing or b.c.; autonomous problems do not incur order reduction (see [44–46] for fully discrete schemes; [36] for discrete-in-time schemes).

Consider the IBVP (1.1) with $\mathcal{L} = \partial_{xx}$, $\Omega = (0,1)$, and forcing f, Dirichlet b.c. g, and initial conditions (i.c.) u_0 chosen so that the exact solution is $u(x,t) = \cos(t)$. We discretize the problem in space, on a uniform grid with 10000 points, using standard second order centered differences (so that spatial errors are negligible relative to temporal errors). Finally, the resulting system is advanced forward in time using standard first to fourth order DIRK schemes (see Appendix A for the schemes used). Table 1.1

	DIRK1=BE	DIRK2	DIRK3	DIRK4
convergence order of u	1	2	2	2
convergence order of u_x	1	1.5	1.5	1.5
convergence order of u_{xx}	1	1	1	1
convergence order of u_{xxx}	1	0.5	0.5	0.5

TABLE 1.1. *Observed convergence order (in time) for DIRK 1 to 4. DIRK 1 is backward Euler; DIRK 2 to 4 can be found in Appendix A.*

shows the resulting convergence orders for the solution and its spatial derivatives (which are frequently important in IBVPs for body forces, Dirichlet-to-Neumann maps, etc.), all measured in the maximum norm. Backward Euler (BE=DIRK1) shows no order reduction in function value or derivatives. DIRK2 shows a reduction of half an order per spatial derivative (u_x converges with $O(\Delta t^{1.5})$; u_{xx} with $O(\Delta t)$, etc.). More severe order reduction arises for DIRK3 and DIRK4: they converge at the same orders as DIRK2. This highlights two important messages. First, order reduction can arise already in very simple problems. Second, it manifests in two ways: (i) spatial derivatives may be less accurate than function values; and (ii) schemes of order higher than two may drop to second order (less for spatial derivatives). A geometric explanation for these observations follows.

1.2. Geometric explanation of order reduction via boundary layers. The cause for the observations in Table 1.1 can be illustrated by studying the shape of the truncation errors. Figure 1.1 shows the local (single time step) and global (fixed final time) errors in space, for the 1D heat equation problem considered in Subsection 1.1, using backward Euler (DIRK1), DIRK2, and DIRK3, respectively. In each panel, results for three choices of Δt are shown, with successive ratios of 2. For *all* schemes, boundary layers appear locally. However, for DIRK1 the boundary layers vanish globally, while for DIRK2 and DIRK3 they persist globally.

The error in the interior of the domain always scales like the order of the method, but the boundary layer amplitudes scale like $O(\Delta t^2)$. For DIRK3, this results in an order reduction of 1 for u. Moreover, any boundary layer has a thickness of $O(\sqrt{\Delta t})$, resulting in a/an (additional) reduction of half an order per spatial derivative.

Why boundary layers arise in the approximation error can be understood as follows. Every stage of a DIRK scheme is a backward Euler-type solve. Therefore, it is useful to first examine the BE scheme, applied to the 1D heat equation in the unit interval

$$u^{n+1} - \Delta t\, u_{xx}^{n+1} = u^n + \Delta t\, f^{n+1} \quad \text{for} \quad 0 < x < 1, \tag{1.2}$$

with a smooth forcing, and conventional Dirichlet b.c. applied, i.e., $u^{n+1} = g^{n+1}$ for $x \in \{0,1\}$. Let u^* be the exact solution. Then the approximation error at time t_{n+1}, defined as $\epsilon_0^{n+1} = u^{n+1} - u^*(t_{n+1})$, satisfies the BVP

$$\epsilon_0^{n+1} - \Delta t\, \partial_{xx} \epsilon_0^{n+1} = \epsilon_0^n + F^{n+1} \quad \text{for} \quad 0 < x < 1, \tag{1.3}$$

with homogeneous Dirichlet b.c.. Here $F^{n+1} = -(u^*(t_{n+1}) - u^*(t_n) - \Delta t\, u_{xx}^*(t_{n+1}) - \Delta t\, f^{n+1}) = O(\Delta t^2)$.

Problem (1.3) is a singularly perturbed BVP, where the time step Δt is the small parameter. Standard boundary layer (BL) theory [9] implies that, generally, the solution exhibits a BL of thickness $O(\sqrt{\Delta t})$, and amplitude determined by $\epsilon_0^n + F^{n+1}$. If the i.c.

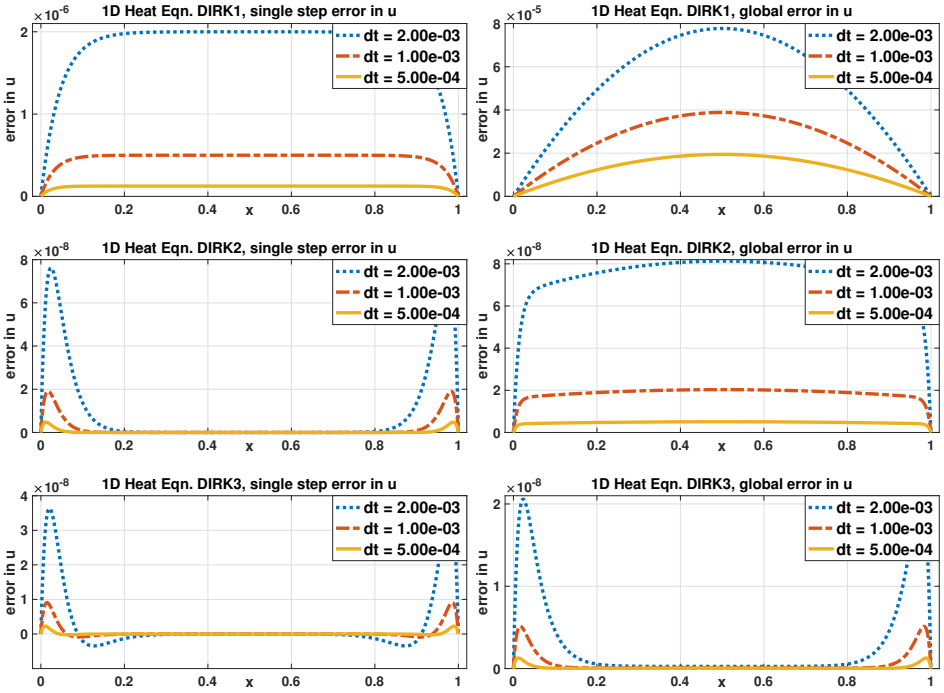

FIG. 1.1. *Local (left) and global (right) errors as functions of x for BE (top), DIRK2 (middle), and DIRK3 (bottom) with three Δt choices.*

are captured exactly, i.e., $\epsilon_0^0 = 0$, the BL amplitude is $O(\Delta t^2)$. This explains exactly the top left panel in Figure 1.1.

Higher order DIRK schemes combine intermediate stage solutions, each of which arises from a backward Euler-type (and thus singularly perturbed) problem. The BLs of the stage solutions do generally not cancel out, thus yielding a composite layer in the numerical approximation. How the composition of boundary layers from the stages works, and why they may vanish globally in time (cf. BE), is the focus of the study in Section 2; reducing the size of the BL is the focus of Section 3 and Section 4.

1.3. Current paper in context of prior research.

1.3.1. Error analysis for RK order reduction in IBVPs. Early work [11] highlighted that RK order conditions are in general not sufficient to ensure classical (p-th order) convergence for IBVP. For linear constant coefficient PDEs, they observed that additional conditions on the RK scheme, referred to as "strictly accurate" (equivalent to stage order) of order $p-1$, were sufficient to obtain a p-th order global error.

Subsequent studies in the 1980s [44–46, 50] examined the fully discrete (space and time) error incurred by RK methods applied to method of lines discretizations for IBVPs. The studies revealed that for IBVP with time dependent boundary conditions: (i) RK schemes may drop to second order temporal error [46, 50]; (ii) the global error may be smaller than the local truncation error [45, 46, 50]; (iii) order reduction may be improved if the solution happens to satisfy additional (but not natural) compatibility conditions at the boundaries [44–46]. For instance, [50] derived global ℓ^2 error convergence rates for linear problems with time-dependent Dirichlet b.c. and demonstrated

numerically that 3rd and 4th order DIRK schemes perform no better than 2nd order schemes. Many of these results are also included in [22, Chapter II.2].

Rigorous error analyses for RK schemes applied to linear PDEs in the semi-discrete setting (in time only, space continuous) [6, 17, 30, 36] have also shown: (i) the RK convergence order may be limited by the scheme's stage order [6, 36]; (ii) the order (reduction) is in general fractional and depends on the b.c. and regularity of the solution [36]; and (iii) for parabolic equations, full convergence order is attained sufficiently far from boundaries [31] and that order reduction is localized to the boundaries. Similar estimates have also been given for quasilinear parabolic [32] and nonlinear equations [38]. Order reduction in the context of singular perturbation problems, and the interaction between the time step and the small parameter, have been examined in [10].

Prior work has focused primarily on RK convergence rates quantifying the size of the error norms, and providing qualitative information on the error incurred in RK schemes (i.e., [31] and [22, Chapter II.2] show errors are localized to domain boundaries). *The work here is complementary:* we do not focus on direct estimates that bound the norms of the RK spatial error, but rather on the shape of the spatial error and its implications on the accuracy of quantities of practical interest (e.g., derivatives of the solution at the boundary). The key result in Theorem 2.1 characterizes the RK spatial error as a singular perturbation problem with boundary layers. When that error is measured in norms, the results of prior work are recovered—however, the shape of the error provides additional important insight into how to remove order reduction, and what limitations stand in the way of removing it.

1.3.2. Avoiding order reduction in IBVPs. There are several known approaches for remedying order reduction, with this work focusing on the following two.

(i) Modified boundary conditions. Approaches to overcome order reduction for explicit RK schemes, applied to advective problems without forcing, have been proposed based on modifying the intermediate stage b.c. [1, 15]. Further improvements have been developed in the context of conservation laws [39]. For the linear problems (1.1), [3,5] derived high order modified b.c. and proved that they remedy order reduction to arbitrary order ([3] for autonomous $\mathcal{L}$ and [5] for time-dependent $\mathcal{L}(t)$). Those papers also provide convergence results for $\mathcal{L}u$, demonstrating that derivatives of u can be less accurate than u. In Section 4, we provide a more general approach to deriving modified b.c. based on insights gained from viewing the RK error as a singular perturbation problem.

(ii) Time stepping coefficients with extra conditions. High stage order is the most straightforward condition on RK coefficients that will avoid order reduction—yet it is restrictive, and not compatible with high order DIRK schemes (see [20, Chapter IV.15], and Subsection 2.1). For Rosenbrock-Wanner (ROW) methods applied to linear problems, [48] devised conditions weaker than stage order that alleviate order reduction (cf. [37]). For RK schemes applied to linear IBVPs, similar conditions were stated in [36] (see Remark 3.2), but no corresponding RK schemes were provided. In Section 3 we introduce new conditions on the Butcher tableau, referred to as *weak stage order* (WSO), that are sufficient for, yet simpler than, the conditions in [36]. We then obtain new schemes that satisfy WSO and demonstrate that they alleviate order reduction. Note that the WSO conditions may appear formally equivalent to the ROW method conditions in [37, Equations (3.11')]; however, they apply to RK methods and are derived under a more general condition that does not assume an SDIRK structure as in [37]. We also note that

conditions in [42] provide, in general, a subset of the WSO conditions that may improve convergence in the stiff limit. Note that related work [25] further develops (and characterizes the limitations of) WSO schemes satisfying the eigenvector criterion introduced in Definition 3.2 below; however, [25] does not discuss the spatial manifestations of order reduction.

In addition to the above approaches, the works [45] (for explicit schemes) and [14] (implicit schemes) provide a conceptually simple, yet practically complicated, methodology for avoiding order reduction: decompose the solution into one part from an IBVP that does not exhibit order reduction, and another part obtained directly from the data. We do not examine the spatial manifestation of such approaches here.

Finally, methods equivalent to multistage methods are prone to order reduction. Specifically, deferred correction methods [10, 33] (see also [34], and references within, for a review of order reduction in deferred correction and Gauss quadrature methods) exhibit order reduction since they can be recast as RK methods [18, 26]; similarly for extrapolation methods [26], Runge-Kutta-Nyström methods [4], and ROW methods [48], etc.. Multistep methods, as implied by the error estimates in [29] and [20, Chapter IV.15], do *not* exhibit order reduction.

This paper is organized as follows. In Section 2, based on a characterization of the spatial behavior of the global-in-time error we show that the error arises as a singular perturbation problem with BLs. Sections 3 and 4 focus on remedies to order reduction: in Section 3 the concept of *weak stage order* is introduced, which (i) makes the BLs that affect the final result as accurate as the scheme's order, and (ii) is compatible with diagonally IRK (DIRK) schemes; and in Section 4, the impact of *modified boundary conditions* (MBC) on BLs and order reduction is studied. Numerical results are shown in Section 5, demonstrating the spatial manifestations of order reduction, and its remedies, in various examples. Generalizations are discussed in Section 6.

2. Boundary layers in the global error and order reduction

As seen in the prior section, RK schemes can yield singular behavior, such as BLs, in the numerical solution—and this behavior serves as a root mechanism of order reduction for IBVPs. However, the existence of a singular perturbation problem in each RK stage does not strictly imply the formation of a BL, and order reduction, in the global truncation error. For example, DIRK2 and DIRK3 can produce BLs in the global error, while backward Euler (BE) does not. This section provides an analysis that characterizes the global error behavior in space, and derives conditions under which order reduction does or does not occur.

2.1. Review of implicit Runge-Kutta time-stepping for IBVP. Here we briefly collect the key notation and results for implicit RK schemes used in the paper. The time-stepping coefficients for a general RK scheme may be represented by the Butcher notation

$$\begin{array}{c|c} \vec{c} & A \\ \hline & \vec{b}^T \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array},$$

where the entries of $\vec{c}$ are the row sums of A, i.e., $\vec{c} = A\vec{e}$, and $\vec{e} = (1, \ldots, 1)^T$ denotes the s-dimensional vector of all ones.

Diagonally implicit Runge-Kutta (DIRK) schemes are an important sub-class of implicit RK schemes. For these schemes the matrix A is lower triangular; and has

non-vanishing diagonal entries when A is non-singular. DIRK schemes are particularly simple, because the stages can be solved sequentially, with each solve being a small modification of a backward Euler step.

An unconditionally stable RK scheme (so that the semi-discrete limit is justified) applied to the IBVP (1.1) takes the form of a BVP problem for the stage values

$$u_i^{n+1} = u^n + \Delta t \sum_{j=1}^{s} a_{ij} \left(\mathcal{L} u_j^{n+1} + f_j^{n+1} \right) \quad \text{with b.c. for } u_i^{n+1}, \tag{2.1}$$

followed by an explicit update rule for the new value

$$u^{n+1} = u^n + \Delta t \sum_{j=1}^{s} b_j \left(\mathcal{L} u_j^{n+1} + f_j^{n+1} \right), \tag{2.2}$$

for which no b.c. are required. Here $\Delta t > 0$ is the time step, and u^n denotes the numerical solution at time $t_n = n\Delta t$. The i-th stage solution u_i^{n+1} is associated with time $t_n + c_i \Delta t$, as are the corresponding forcing terms $f_i^{n+1} = f(x, t_n + c_i \Delta t)$.

A scheme is said to have *(classical) order* p if for sufficiently smooth solutions, the error obtained from a single RK step is $O(\Delta t^{p+1})$ (cf. [19]). This imposes a set of constraints on the RK coefficients, known as the *order conditions* [12, 19]. Since we consider linear differential operators, we list here the RK order conditions for linear problems (see [19, Chapter II.2] for nonlinear problems)

$$\vec{b}^T A^j \vec{c}^k = \frac{1}{(j+k+1)\dots(k+1)} \quad \text{for} \quad 0 \le j+k \le p-1 \text{ and } j,k \ge 0. \tag{2.3}$$

Here a power of a vector applies to each component, i.e., $\vec{c}^k = ((c_1)^k, \dots, (c_s)^k)^T$. For example, first order schemes require $\vec{b}^T \vec{e} = 1$, second order schemes additionally require $\vec{b}^T \vec{c} = \frac{1}{2}$, while the third order conditions impose two further constraints: $\vec{b}^T \vec{c}^2 = \frac{1}{3}$ and $\vec{b}^T A \vec{c} = \frac{1}{6}$. We now introduce the *stage order residuals*, and the definition of stage order [51, Chapter IV.5]:

$$\text{(Stage order residuals)} \qquad \vec{\tau}^{(j)} = A\vec{c}^{j-1} - \frac{1}{j}\vec{c}^j, \quad j = 1, 2, \dots. \tag{2.4}$$

Definition 2.1 (Stage order). *Condition $B(\tilde{p})$: let $\tilde{p}$ be the largest number such that the quadrature condition holds $\vec{b}^T \vec{c}^{j-1} = j^{-1}$ for $j = 1 \dots \tilde{p}$. Condition $C(\tilde{q})$: let $\tilde{q}$ be the largest number such that $\vec{\tau}^{(j)} = \vec{0}$ for $j = 1 \dots \tilde{q}$. The* stage order *of a RK scheme is $q = \min\{\tilde{p}, \tilde{q}\}$.*

It is well-known that schemes with high stage order avoid order reduction in stiff ODEs. Unfortunately, high stage order is a restrictive property for DIRK schemes:

Remark 2.1. (Stage order in DIRKs) DIRK schemes with nonzero diagonal entries are limited to stage order $q = 1$ [19]. Moreover, DIRK schemes with singular A may have stage order $q = 2$ (but not higher) [25]. Examples are EDIRK schemes, such as Crank-Nicolson, or TR-BDF2 [7].

The stage order residuals $\vec{\tau}^{(j)}$ for $1 \le j \le q$ become important later, even when they are nonzero. Their significance makes use of the following orthogonality property, which follows immediately from the order conditions (2.3):

PROPOSITION 2.1. *For a p-th order RK scheme, the stage order residuals satisfy* $\vec{b}^T A^j \vec{\tau}^{(k)} = 0$ *for* $1 \le j+k \le p-1$, *with* $j \ge 0$ *and* $k \ge 1$. *In particular,* $\vec{b}^T \vec{\tau}^{(k)} = 0$ *for* $1 \le k \le p-1$.

Lastly we introduce notation relevant to numerical stability. The *stability function* of the RK scheme is given by $R(\zeta) = 1 + \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{e}$. The value $R(\zeta)$ measures the growth u^{n+1}/u^n in one step Δt, when applying the RK scheme to the test equation $u'(t) = \lambda u$, where $\zeta = \lambda \Delta t$. A RK scheme is called *A-stable*, if it is stable for all stable solutions of $u'(t) = \lambda u$ (i.e., $|R(\zeta)| \le 1$ for $\mathrm{Re}(\zeta) \le 0$); and *L-stable* if also $R(\zeta) \to 0$ as $\zeta \to -\infty$. A RK scheme is called *stiffly accurate* [41], if the last row of A equals the vector $\vec{b}^T$, i.e., if $a_{sj} = b_j$ for $j = 1, \ldots, s$. A stiffly accurate RK scheme with invertible coefficient matrix A, which is A-stable, is also L-stable [51], seen by evaluating the $\zeta \to -\infty$ limit of the stability function.

2.2. Equations for the approximation error. In this subsection we derive equations for the discrete-in-time RK error incurred by the IBVP (1.1). The full analysis using the Mellin/z-transform has been presented in [30,31], for normed-error estimates. Our focus is to set the stage for the study of BLs via asymptotic analysis for singular perturbation problems [9,27]. Using asymptotics, we show that order reduction (OR) is restricted to certain regions in space, and that elsewhere no OR occurs. For simplicity, we restrict the presentation to periodic-in-time solutions of the error equations, because those suffice to capture crucial OR mechanisms. It is important to emphasize that *we do not claim that order reduction happens solely for periodic solutions,* but rather that periodic solutions suffice to provide an intuitive/geometrical visualization of OR for PDEs. Nevertheless, as shown in Appendix B, under some conditions the periodic solutions contain the full OR phenomenon.

As pointed out earlier, the asymptotic analysis in this subsection is for the case where $\mathcal{L}$ in (1.1) is a second-order operator, with $\Omega = (0,1)$ and Dirichlet b.c., i.e., $u = g$ on $\partial\Omega$. However, the concepts generalize to other differential operators and boundary conditions. The numerical examples focus on the one-dimensional case as well.

Below, let $W_e(\theta)$ denote the wedge in the left complex half plane defined by: $\lambda \in W_e(\theta) \Longleftrightarrow |\arg(-\lambda)| < \theta$. We further assume that both the PDE and the scheme are well-defined and stable, in the following sense:

(a) There are constants K, $0 < \theta_1 < \pi/2$ such that: For any $\lambda \notin W_e(\theta_1)$ the operator $(I - \lambda\mathcal{L})$ with homogeneous boundary conditions has a uniformly bounded inverse: $\|(I - \lambda\mathcal{L})^{-1} u\|_{L^\infty} \le K \|u\|_{L^\infty}$;

(b) The scheme's stability region includes $W_e(\theta_2)$, for some $\theta_2 > \theta_1$.

(c) The eigenvalues of A have non-negative real parts.

(d) There are constants $\delta_d, c_d > 0$ such that: For any complex $|z| < \delta_d$, the matrix $\vec{e}\vec{b}^T + zA$ is diagonalizable, and the family of eigenvector matrices $T(z)$ can be selected so that their condition number satisfies $\|T(z)\| \, \|T^{-1}(z)\| < c_d$ for $|z| < \delta_d$.

(e) The matrix A is invertible, and $\vec{b}^T A^{-1} \vec{e} \neq 0$.

(2.5)

Condition (a) is a property of the operator $\mathcal{L}$ only; condition (b) is a property of the scheme in relation to the operator $\mathcal{L}$; and conditions (c–e) are properties solely of the numerical scheme. Condition (c) guarantees that the scheme Equations (2.1–2.2), equivalently (2.6), have a well defined solution at each step. It is rather natural for RK

schemes, and most commonly used methods (incl. most DIRK, Gauss, Radau, Lobatto) satisfy it. Condition (d) is a technical assumption on the RK scheme that will be used to estimate and bound the numerical errors. Requiring a uniform bound on the condition number of $T(z)$ avoids a situation in which two of the eigenvectors (i.e., columns of $T(z)$) become parallel as $z \to 0$. Condition (d) may be alternatively stated using perturbation theory, via conditions on A and an $(s-1)\times(s-1)$ matrix determined by A and $\vec{b}$. For brevity, however, we leave (d) in its current form. Due to its important role below, we introduce notation for the subspace spanned by the vectors orthogonal to $\vec{b}$:

$$\vec{b}^{\perp} = \{\vec{v} : \vec{b}^T \vec{v} = 0\}.$$

Condition (e) is also a technical assumption on the RK scheme. Most commonly used RK schemes satisfy $\vec{b}^T A^{-1} \vec{e} \neq 0$ (in particular, all stiffly accurate schemes do so, because $\vec{b}^T A^{-1} \vec{e} = 1$). However, some schemes do violate it, for example the 2-stage 4-th order Gauss method [19]. Note that some unconditionally stable schemes, such as EDIRK schemes [28], also do not have invertible A. Condition (a) is required to estimate the magnitude of the RK numerical error; it is also a numerical stability condition because it guarantees that the spectrum of $\mathcal{L}$ is contained within $W_e(\theta_1)$. Condition (a) is satisfied when $\mathcal{L}$ is (strongly) elliptic and (1.1) is parabolic [40, Chapter 2] (e.g., heat equation)[1]. In fact, the inverses of differential operators $(I - \lambda\mathcal{L})$ are generally given by Green's functions, which are singular at the spectrum of $\mathcal{L}$ and continuous functions of λ away from the spectrum. Specifically, when $(I - \lambda\mathcal{L})^{-1}$ can be written in terms of Green's functions, then condition (a) requires that the Green's function be uniformly bounded, in L^1, for λ outside $W_e(\theta_1)$. Condition (a) does not hold for dispersive equations where $\mathcal{L}$ has eigenvalues on the imaginary axis.

We first formulate equations for the RK error, and then examine the equations when the exact solution to (1.1) is time-periodic. One step (i.e., see [2]) of a RK scheme (2.1–2.2) can be written using $\vec{u}^{n+1} := (u_1^{n+1}, \ldots, u_s^{n+1})^T$ as:

$$\text{("internal" stages)} \quad \vec{u}^{n+1} = u^n \vec{e} + \Delta t A \left(\mathcal{L}\vec{u}^{n+1} + \vec{f}^{n+1} \right), \quad \text{with b.c. for } \vec{u}^{n+1}, \tag{2.6}$$

$$\text{("last" stage)} \quad u^{n+1} = u^n + \Delta t \vec{b}^T \left(\mathcal{L}\vec{u}^{n+1} + \vec{f}^{n+1} \right), \quad \text{with no b.c. for } u^{n+1}, \tag{2.7}$$

where $\vec{f}^{n+1} := (f_1^{n+1}, \ldots, f_s^{n+1})^T$, and $\mathcal{L}\vec{u}^{n+1} := (\mathcal{L}u_1^{n+1}, \ldots, \mathcal{L}u_s^{n+1})^T$. Denote the exact solution to (1.1) by $u^*(x,t)$. To obtain an equation for the propagation of the numerical error, let $\epsilon_0^n(x) := u^n(x) - u^*(x,t_n)$ be the error, and $\epsilon_i^{n+1}(x) := u_i^{n+1}(x) - u^*(x, t_n + c_i \Delta t)$ be the stage error. Substituting these expressions for the error into (2.6–2.7), yields:

$$\text{("stage" error)} \quad \vec{\epsilon}^{n+1} = \epsilon_0^n \vec{e} + \Delta t A \mathcal{L} \vec{\epsilon}^{n+1} + \vec{\delta}^n, \quad \text{with b.c. for } \vec{\epsilon}^{n+1}, \tag{2.8}$$

$$\text{("last" error)} \quad \epsilon_0^{n+1} = \epsilon_0^n + \Delta t \vec{b}^T \mathcal{L} \vec{\epsilon}^{n+1} + \delta_0^n, \quad \text{with no b.c. for } \epsilon_0^{n+1}, \tag{2.9}$$

$$\text{where} \quad \vec{\epsilon}^{n+1} = \left(\epsilon_1^{n+1}, \ldots, \epsilon_s^{n+1}\right)^T, \quad \text{and} \quad \vec{\delta}^n = \left(\delta_1^n, \ldots, \delta_s^n\right)^T.$$

Here $\vec{\delta}^n$ and δ_0^n are the local truncation errors (LTEs), and involve only u^* and $\vec{f}$. Formulas for the LTEs can then be obtained using the PDE (1.1), and Taylor expanding

[1] In fact, (2.5a) is closely related to the condition required for solutions of (1.1) to be defined by an analytic semigroup.

u^* (at t_n):

$$\vec{\delta}^n(x)=\sum_{j\geq 1}\frac{\partial_t^j u^*(x,t_n)\,\Delta t^j}{(j-1)!}\,\vec{\tau}^{(j)},\quad \delta_0^n(x)=\sum_{j\geq 1}\frac{\partial_t^j u^*(x,t_n)\,\Delta t^j}{(j-1)!}\,\Big(\vec{b}^T\vec{c}^{\,j-1}-\frac{1}{j}\Big). \tag{2.10}$$

Here $\vec{\tau}^{(j)}$ are the stage order residuals defined in (2.4). One should stress that Equations (2.8–2.10) hold for linear problems only, i.e., (1.1). For a p-th order Runge-Kutta scheme with stage order q, the first $q\leq p$ summands in $\vec{\delta}^n$ vanish; thus $\delta_j^n=O(\Delta t^{q+1})$ $(1\leq j\leq s)$. Meanwhile, the p-th order conditions guarantee that $\delta_0^n=O(\Delta t^{p+1})$ (and hence $\delta_s^n=\delta_0^n=O(\Delta t^{p+1})$ for stiffly accurate schemes). For the remainder of Section 2, we assume conventional b.c., i.e., $g_i^{n+1}=g(t_n+c_i\,\Delta t)$; equivalently, this yields homogeneous b.c. for the error $\vec{\epsilon}^n$ in (2.8). Time-periodic solutions to the IBVP (1.1) can be obtained when the forcing and b.c. have the form $f=\hat{f}\,e^{\imath\omega t}$ and $g=\hat{g}\,e^{\imath\omega t}$, where $\hat{f}$ and $\hat{g}$ are functions defined on Ω and $\partial\Omega$, respectively, and ω is a (real-valued) constant. Then $u^*(x,t)=U^*(x)e^{\imath\omega t}$ is the periodic solution to (1.1), where U^* is the (unique because of (2.5a)) solution to the BVP $(\imath\omega-\mathcal{L})U^*=\hat{f}$ with b.c. $\mathcal{B}U^*=\hat{g}$. Since RK schemes are linear, time-harmonic forcings and boundary data will also yield periodic numerical solutions. To obtain periodic solutions for $\vec{\epsilon}^n$ and ϵ_0^n when $u^*(x,t)=U^*(x)e^{\imath\omega t}$, we seek an ansatz of the form (with a slight abuse of notation):

$$\vec{\epsilon}^n(x)=\vec{\epsilon}(x)\,z^n,\qquad \epsilon_0^n(x)=\epsilon_0(x)\,z^n,\qquad \vec{\delta}^n(x)=\vec{\delta}(x)\,z^n,\qquad \delta_0^n(x)=\delta_0(x)\,z^n, \tag{2.11}$$

where $z:=e^{\imath\omega\Delta t}$, and

$$\vec{\delta}(x)=U^*(x)\sum_{j\geq 1}\frac{(\imath\omega)^j\Delta t^j}{(j-1)!}\vec{\tau}^{(j)},\qquad \delta_0(x)=U^*(x)\sum_{j\geq 1}\frac{(\imath\omega)^j\Delta t^j}{(j-1)!}\left(\vec{b}^T\vec{c}^{\,j-1}-\frac{1}{j}\right). \tag{2.12}$$

Substituting (2.11) into (2.8) and (2.9) yields the coupled system for $(\epsilon_0,\vec{\epsilon})$:

$$\underbrace{\begin{pmatrix}1 & 0\\ -z^{-1}\vec{e} & I\end{pmatrix}}_{C}\begin{pmatrix}\epsilon_0\\ \vec{\epsilon}\end{pmatrix}-\underbrace{\begin{pmatrix}0 & (z-1)^{-1}z\Delta t\vec{b}^T\\ \vec{0} & \Delta t\,A\end{pmatrix}}_{\Gamma}\begin{pmatrix}\mathcal{L}\epsilon_0\\ \mathcal{L}\vec{\epsilon}\end{pmatrix}=\begin{pmatrix}(z-1)^{-1}\delta_0\\ z^{-1}\vec{\delta}\end{pmatrix}. \tag{2.13}$$

Equation (2.13) is supplemented with s boundary conditions for $\vec{\epsilon}$, and no boundary conditions for ϵ_0. Hence, (2.13) is actually a differential algebraic equation. The block components in (2.13) for the differential equation (which involve only the stage errors $\vec{\epsilon}$) may be separated from the algebraic equation (which couple the stage and function errors $\vec{\epsilon}$, ϵ_0) by simultaneously transforming C and Γ into upper block triangular form. To do this, we first multiply (2.13) through on the left by a matrix S_Γ, which block-diagonalizes Γ, followed by the matrix D, where

$$S_\Gamma=\begin{pmatrix}1 & -z\vec{\gamma}^T\\ 0 & I\end{pmatrix},\quad D=\begin{pmatrix}1 & 0\\ 0 & I+\vec{e}\vec{\gamma}^T\end{pmatrix}\begin{pmatrix}1 & 0\\ z^{-1}\vec{e} & I\end{pmatrix}\begin{pmatrix}1+\vec{\gamma}^T\vec{e} & 0\\ 0 & I\end{pmatrix}^{-1},\quad \vec{\gamma}^T:=\frac{\vec{b}^TA^{-1}}{z-1}.$$

Here we have used that A is invertible to define $\vec{\gamma}^T$. Multiplication yields

$$\underbrace{\begin{pmatrix}1 & -\vec{\alpha}^T\\ 0 & I\end{pmatrix}}_{DS_\Gamma C}\begin{pmatrix}\epsilon_0\\ \vec{\epsilon}\end{pmatrix}-\underbrace{\begin{pmatrix}0 & 0\\ \vec{0} & M\end{pmatrix}}_{DS_\Gamma\Gamma}\begin{pmatrix}\mathcal{L}\epsilon_0\\ \mathcal{L}\vec{\epsilon}\end{pmatrix}=\begin{pmatrix}\psi_0\\ \vec{h}\end{pmatrix}, \tag{2.14}$$

where the *derivative coefficient matrix*

$$M := \Delta t\, A + \frac{\Delta t}{z-1}\,\vec{e}\vec{b}^{\,T}, \qquad \text{and} \qquad \vec{\alpha}^{\,T} := \frac{z\vec{b}^{\,T}A^{-1}}{z-1+\vec{b}^{\,T}A^{-1}\vec{e}}, \tag{2.15}$$

appear in the block matrices of (2.14), while

$$\vec{h}(x) := \frac{1}{z}\left(\vec{\delta}(x) + \frac{\delta_0(x)}{z-1}\vec{e}\right), \quad \psi_0(x) := \frac{1}{z-1+\vec{b}^{\,T}A^{-1}\vec{e}}\left(-\vec{b}^{\,T}A^{-1}\vec{\delta}(x) + \delta_0(x)\right). \tag{2.16}$$

In (2.14), multiplication by S_Γ converts $S_\Gamma\Gamma$ into a block diagonal matrix; multiplication by D converts $DS_\Gamma C$ into row echelon form while preserving the block structure of $S_\Gamma\Gamma$. Equation (2.14) is significant since it allows one to extract the spatial RK error $\epsilon_0(x)$. Working out the components of (2.14), the bottom block row yields an s-dimensional partial differential equation for the stage errors:

$$\vec{\epsilon} - M\mathcal{L}\vec{\epsilon} = \frac{1}{z}\left(\vec{\delta} + \frac{\delta_0}{z-1}\vec{e}\right), \quad \text{with b.c. } \vec{\epsilon} = 0, \tag{2.17}$$

while the top row yields one algebraic expression for the global error ϵ_0 in terms of the stage error vector:

$$\epsilon_0(x) = \vec{\alpha}^{\,T}\vec{\epsilon} + \psi_0 = \frac{1}{z-1+\vec{b}^{\,T}A^{-1}\vec{e}}\left(z\,\vec{b}^{\,T}A^{-1}\vec{\epsilon} - \vec{b}^{\,T}A^{-1}\vec{\delta} + \delta_0\right). \tag{2.18}$$

To recap, the spatial error vector for the RK stages $\vec{\epsilon}$ (corresponding to the time-periodic response of the error equation) satisfies the BVP system (2.17), in which the derivative coefficient matrix M pre-multiplies the operator term $\mathcal{L}\vec{\epsilon}$. The error of the RK scheme $\epsilon_0(x)$ is then computed by evaluating the update rule (2.18).

For schemes with singular A (see Remark 2.1), one could still obtain (2.14) by first row-reducing Γ. Singular A, however will yield multiple algebraic equations analogous to (2.17) coupling $\vec{\epsilon}$ and ϵ_0, and a lower dimensional PDE system analogous to (2.17).

2.3. Spectral properties of the Derivative Coefficient Matrix. Here we examine the periodic-in-time error Equation (2.17) in the eigenbasis defined by the matrix M. We carry out an asymptotic analysis and demonstrate that the RK error generically satisfies a singular perturbation problem. We first estimate the eigenvalues and eigenvectors of M for $z = e^{\imath\omega\Delta t}$ on the unit circle, where ω is real and Δt is small. Of particular relevance is the existence of small eigenvalues for M, as they give rise to singularly perturbed BVPs (producing BLs and, potentially, other effects associated with singular BVPs). We call an eigenvalue $\lambda \neq 0$ of M *small* if $\lambda \to 0$ as $\Delta t \to 0$.

Since $z = e^{\imath\omega\Delta t}$, for $\Delta t \ll 1$, expanding $\frac{\Delta t}{z-1}$ in powers of Δt yields

$$M = \frac{1}{\imath\omega}\vec{e}\vec{b}^{\,T} + \Delta t\,(A - \frac{1}{2}\vec{e}\vec{b}^{\,T}) + O(\Delta t^2). \tag{2.19}$$

Because $\vec{e}\vec{b}^{\,T}$ is a rank-1 matrix with eigenvalue 1 (since $\vec{b}^{\,T}\vec{e} = 1$), all eigenvalues of M vanish as $\Delta t \to 0$, except for one that (due to numerical consistency) is equal to $(\imath\omega)^{-1}$. How these zero eigenvalues are approached as $\Delta t \to 0$ depends on the structure of A and $\vec{b}$.

THEOREM 2.1 (Asymptotic eigenvalues of M). For $0 < \Delta t \ll 1$ and a fixed $\omega \in \mathbb{R}$ such that $z = e^{\imath\omega\Delta t} \neq 1$, the matrix M, defined in (2.15) satisfies:

(1) It has one $O(1)$ eigenvalue, which is at most $O(\Delta t)$ away from $\frac{1}{\imath\omega}$. The leading order part for the corresponding right and left eigenvectors are $\vec{e}$ and $\vec{b}$.

(2) It has $s-1$ (including multiplicity) small, but nonzero eigenvalues, of magnitude at most $O(\Delta t)$. They have the form $\lambda=\Delta t\mu_0+o(\Delta t)$. Here μ_0 are the eigenvalues of the matrix QA restricted to $\vec{b}^{\perp}$ (denoted by B), where $Q=I-\vec{e}\vec{b}^T$ is the projection onto $\vec{b}^{\perp}$ along $\vec{e}$.

(3) It has no zero eigenvalues.

Proof. We first prove (3). Assume M has a zero eigenvalue. Then there exists a nonzero vector $\vec{\ell}$ such that $M\vec{\ell}=\Delta t A\vec{\ell}+\frac{\Delta t}{z-1}(\vec{b}^T\vec{\ell})\vec{e}=\vec{0}$, hence $A\vec{\ell}\,\|\,\vec{e}$. But because A is non-singular, there is a uniquely defined (up to scaling) eigenvector: $\vec{\ell}^*=A^{-1}\vec{e}$. Hence,

$$M\vec{\ell}^*=\Delta t\Big(1+\frac{\vec{b}^TA^{-1}\vec{e}}{z-1}\Big)\vec{e}=\vec{0},\qquad\Longrightarrow\qquad z=1-\vec{b}^TA^{-1}\vec{e}. \tag{2.20}$$

The constraint (2.20) can only occur for $\vec{b}^TA^{-1}\vec{e}=2$ and $z=-1$, or $\vec{b}^TA^{-1}\vec{e}=0$ and $z=1$ (because $\vec{b}^TA^{-1}\vec{e}$ is real and $|z|=1$). Neither case is possible when $\Delta t\ll 1$ since $0<|\omega\Delta t|<\pi$ implies $z\neq\pm 1$.

Item (1) results from viewing M as a perturbed rank-1 matrix. The matrix $\frac{1}{\imath\omega}\vec{e}\vec{b}^T$ is diagonalizable and therefore the Bauer-Fike theorem [8] implies that M has one eigenvalue $O(\Delta t)$ away from $(\imath\omega)^{-1}$, and $(s-1)$ eigenvalues of size $O(\Delta t)$. Since the $(\imath\omega)^{-1}$ eigenvalue is simple (non-repeated), the corresponding eigenvector is, to leading order, $\vec{v}^{(0)}=\vec{e}$ [49, Chapter V 2.3]. The remaining parts for (2) are not as straightforward as (1) since $\lambda=0$ is a degenerate eigenvalue of M when $\Delta t=0$.

To show (2), we write $M=\frac{\Delta t}{z-1}M^*$, where $M^*:=\vec{e}\vec{b}^T+\delta A$ and $\delta=z-1$, so that when $\Delta t\ll 1$, we have $|\delta|\ll 1$. In addition, we write the eigenvalues of M^* as $\lambda^*=\delta\mu$ (with the corresponding eigenvalues of M being $\Delta t\mu$). We now work in a coordinate basis defined by the eigenvectors of the unperturbed matrix $\vec{e}\vec{b}^T$. Let $O_b\in\mathbb{R}^{s\times(s-1)}$ be a matrix whose columns form an orthonormal basis for $\vec{b}^{\perp}$. Set $T_0=[\vec{e},O_b]$. Then $T_0^{-1}=[\vec{b},Q^TO_b]^T$, because $\vec{b}^T\vec{e}=1$, $\vec{b}^TO_b=0$, $(O_b)^TQ\vec{e}=0$, and $(O_b)^TQO_b=(O_b)^TO_b=I$. Using T_0 as a similarity transformation, the transformed rank-1 matrix becomes $(T_0^{-1}\vec{e}\vec{b}^TT_0)_{ij}=\delta_{i1}\delta_{j1}$, where δ_{ij} is the Kronecker delta. The characteristic equation for M^* then follows from a direct computation of the corresponding determinant:

$$\begin{aligned}\det(M^*-\lambda^*I)&=\det(T_0^{-1}M^*T_0-\delta\mu I)=\det\left(\delta_{i1}\delta_{j1}+\delta\,T_0^{-1}(A-\mu I)T_0\right)\\&=\delta^{s-1}\left(\det(B-\mu I)+\delta\det(A-\mu I)\right).\end{aligned} \tag{2.21}$$

Here $B=(O_b)^TQAO_b$ is the bottom right $(s-1)\times(s-1)$ block of $T_0^{-1}AT_0$. The computation shows that the $(s-1)$ eigenvalues μ are to within $o(1)$ of μ_0, where μ_0 denotes the roots of $\det(B-\mu_0 I)=0$. ∎

Without loss of generality, we label the eigenvalues of M in such a way that $\lambda_1=O(1)$, and λ_2 through λ_s are small but nonzero. The properties below on the matrix M will be used to estimate the size and shape of the RK error in Subsection 2.4.

PROPOSITION 2.2 (Eigenvectors of M). *For a p-th order RK scheme, let $\vec{\ell}_i^T$ and $\vec{r}_i$ be the left and right eigenvectors of M associated with λ_i, normalized so that $\vec{\ell}_i^T\vec{r}_j=\delta_{ij}$ for $1\le i,j\le s$. Then $\vec{\alpha}^T\vec{r}_j$ is $O(1)$ or smaller; and $\vec{\ell}_1^T\vec{h}=O(\Delta t^p)$. Here $\vec{\alpha}$ and $\vec{h}$ are defined in (2.15–2.16).*

Proof. To show $\vec{\alpha}^T\vec{r}_j = O(1)$ or smaller, note that $z-1+\vec{b}^T A^{-1}\vec{e} = O(1)$ due to Assumption (2.5e). Second: to leading order in Δt, the $\vec{r}_j$ are the right eigenvectors of the rank-1 matrix $\vec{e}\vec{b}^T$, so that $\vec{b}^T A^{-1}\vec{r}_j$ (and hence $\vec{\alpha}^T\vec{r}_j$) is $O(1)$ or smaller. For $\vec{\ell}_1^T\vec{h}$: Theorem 2.1(2) implies $\vec{\ell}_1$ has the form $\vec{\ell}_1 = \vec{b} + \sum_{j=1}^{p-1}\Delta t^j\vec{\beta}_j + O(\Delta t^p)$, with a similar expansion for the eigenvalue. Substituting into $\vec{\ell}_1^T M = \lambda_1\vec{\ell}_1^T$ and collecting powers of Δt reveals that $\vec{\beta}_j^T$, $1\le j\le p-1$, is a linear combination of the vectors $\vec{b}^T A^m$, $0\le m\le j$. Now:

$$\vec{\ell}_1^T\vec{h} = \frac{1}{z}\vec{\ell}_1^T\sum_{k=2}^{p-1}\frac{(\omega)^k\Delta t^k}{(k-1)!}\vec{\tau}^{(k)}\,U^*(x) + \frac{\delta_0}{z(z-1)}\vec{\ell}_1^T\vec{e} + O(\Delta t^p). \tag{2.22}$$

However, $\vec{\ell}_1^T\vec{\tau}^{(k)} = \vec{b}^T\vec{\tau}^{(k)} + \sum_{j=1}^{p-1}\Delta t^j\vec{\beta}_j^T\vec{\tau}^{(k)} + O(\Delta t^p)$, where each $\vec{\beta}_j^T\vec{\tau}^{(k)}$ term is a linear combination of $\vec{b}^T A^m\vec{\tau}^{(k)}$, $0\le m\le j$. Since Proposition 2.1 implies that $\vec{b}^T\vec{\tau}^{(k)} = 0$ for $k\le p-1$ and $\vec{\beta}_j^T\vec{\tau}^{(k)} = 0$ for $j\le p-1-k$, it follows that

$$\vec{\ell}_1^T\vec{\tau}^{(k)} = \vec{b}^T\vec{\tau}^{(k)} + \sum_{j=p-k}^{p-1}\Delta t^j\vec{\beta}_j^T\vec{\tau}^{(k)} + O(\Delta t^p) = O(\Delta t^{p-k}). \tag{2.23}$$

Combining (2.22–2.23) with $\delta_0 = O(\Delta t^{p+1})$ and $\frac{\delta_0}{z-1} = O(\Delta t^p)$ yields $\vec{\ell}_1^T\vec{h} = O(\Delta t^p)$. ∎

We conclude the section by proving that the eigenvalues of M lie in the RK stability region, which guarantees solutions to (2.17) are stable (small right-hand sides yield small solutions).

THEOREM 2.2 (Eigenvalues of M are not in $W_e(\theta_2)$). *Fix $|z| = 1, z\neq 1, \Delta t > 0$, and let $\lambda\neq 0$ be an eigenvalue of M; set $\zeta = \Delta t/\lambda$. Then (at least) one of the two statements applies:*

(1) *$1/\zeta$ is an eigenvalue of A; or*

(2) *$R(\zeta) = z$, where R is the scheme's stability function.*

In particular:

(3) *λ is not in the interior of the wedge $W_e(\theta_2)$, introduced in (2.5b).*

(4) *If the scheme is A-stable, then $Re(\lambda)\ge 0$.*

If, conversely, $1/\zeta$ is an eigenvalue of A, with eigenvector in $\vec{b}^\perp$, then $\lambda = \Delta t/\zeta$ is an eigenvalue of M.

Proof. We first justify items (1–2). Let $\vec{w}\neq 0$ be such that $M\vec{w} = \lambda\vec{w}$. Now if ζ^{-1} is an eigenvalue of A, then (1) holds and we are done. Assume, instead that ζ^{-1} is not an eigenvalue of A, and hence $(I-\zeta A)$ is invertible. Then $M\vec{w} = \lambda\vec{w}$ is equal to

$$\frac{\Delta t}{z-1}\vec{e}\vec{b}^T\vec{w} + \Delta t A\vec{w} = \lambda\vec{w}, \qquad \text{and thus} \qquad \vec{w} = \frac{\zeta}{z-1}(\vec{b}^T\vec{w})(I-\zeta A)^{-1}\vec{e},$$

where we have used $\lambda = \Delta t/\zeta$. The right equation shows $\vec{b}^T\vec{w}\neq 0$ because $\vec{w}\neq 0$. Multiplying the right equation through by $((z-1)/(\vec{b}^T\vec{w}))\vec{b}^T$ and rearranging leads to the following identity:

$$R(\zeta) = 1 + \zeta\vec{b}^T(I-\zeta A)^{-1}\vec{e} = z.$$

Items (3) and (4) follow from (2.5c) when (1) applies. On the other hand, when (2) applies, ζ must be on the boundary of the stability region, since $|z| = 1$. Then (3) follows

from (2.5b), and (4) from the definition of A-stability. Finally, the converse statement follows from the definition of M in (2.15). ∎

COROLLARY 2.1. *The operator $I-M\mathcal{L}$ has an L^∞-bounded inverse. Furthermore, $(I-M\mathcal{L})^{-1}$ has an L^∞ bound which is uniform for Δt small enough.*

Proof. By Assumption (2.5d), M is diagonalizable. Let $M=TDT^{-1}$, where the columns of T are eigenvectors of M, and D is the diagonal matrix with the corresponding eigenvalues. Then $I-M\mathcal{L}=T(I-D\mathcal{L})T^{-1}$. Theorem 2.2, and Assumption (2.5a) guarantee that $(1-\lambda\mathcal{L})^{-1}$ is bounded in L^∞, independent of Δt, simultaneously for all eigenvalues of M. Assumption (2.5d) implies that both $\|T(\Delta t)\|$ and $\|T^{-1}(\Delta t)\|$ are bounded, and remain bounded (uniformly) as $\Delta t\to 0$. Hence, $(I-M\mathcal{L})^{-1}$ is bounded, and is uniform as $\Delta t\to 0$. ∎

Corollary 2.1 is used in Subsection 2.4 and Subsection 3.2 to estimate the magnitude of the errors (i.e., the amplitude of the numerical BLs) incurred by the scheme.

2.4. Qualitative behavior of the global error. We now use the spectral decomposition of the derivative coefficient matrix M, derived in Subsection 2.3, to analyze the behavior of Equation (2.17) for $|z|=1$, thus characterizing the spatial approximation error for numerical solutions that are periodic in time.

Let $\psi_i=\vec{\ell}_i^T\vec{\epsilon}$ be the component of the error $\vec{\epsilon}$ in the eigenmode corresponding to the eigenvalue λ_i. Then, left-multiplying Equation (2.17) by the left eigenvectors of M (note that $\vec{\ell}_i^T\mathcal{L}=\mathcal{L}\vec{\ell}_i^T$, because $\mathcal{L}$ is linear), we obtain the following set of decoupled BVPs:

$$\psi_i-\lambda_i\mathcal{L}\psi_i=\vec{\ell}_i^T\vec{h}\quad\text{with b.c. }\psi_i=0,\quad\text{for}\quad 1\le i\le s. \tag{2.24}$$

When λ_i is small, (2.24) is a singular perturbation problem that can be analyzed with standard methods [9]. We can thus conclude:

(I) $\boxed{\psi_0}$. The function $\psi_0=O(\Delta t^{q+1})$ or smaller, is comprised of δ_0 and $\vec{\delta}$. It has no singular behavior: spatial derivatives of ψ_0 are (generally) of the same order as ψ_0. For stiffly accurate schemes, $\vec{b}^TA^{-1}=(0,\ldots,0,1)$ and $\delta_s=\delta_0$ imply that $\psi_0\equiv 0$.

(II) $\boxed{\psi_1;\ \lambda_1=O(1)\neq 0.}$ As shown in Subsection 2.3, the matrix M has one $O(1)$ eigenvalue which is close to $(\imath\omega)^{-1}$. By Corollary 2.1, the magnitude of the eigenmode ψ_1 is determined by $\vec{\ell}_1^T\vec{h}$ in the BVP, which is $O(\Delta t^p)$ by Proposition 2.2. Thus $\psi_1=O(\Delta t^p)$. Further, ψ_1 has no singular behavior: the spatial derivatives of ψ_1 are of the same order as ψ_1 (provided $U^*(x)$ is smooth enough).

(III) $\boxed{\psi_j;\ \lambda_j \text{ is small but nonzero } (2\le j\le s).}$ Then (2.24) is a singularly perturbed BVP, with λ the small parameter. The solution ψ generally has singular behavior, often in the form of boundary layers (BLs) (see Lemma 2.1 regarding other possible effects). From Corollary 2.1, the BL amplitude in ψ is determined by the right-hand side of the BVP, $\vec{\ell}^T\vec{h}$ — thus, in general, $\psi_j=O(\Delta t^{q+1})$ $(2\le j\le s)$ or smaller. Spatial derivatives of ψ_j will lose orders of accuracy, where the exact loss of accuracy depends on $\mathcal{L}$. For example, the heat equation will introduce BLs in ψ_i that scale as $x/\sqrt{\Delta t}$, hence each derivative introduces a 1/2 order loss. The occurrence of singular behavior and BLs in the solutions of ψ are unavoidable for generic time-dependent b.c. and forcing.

Expanding $\vec{\epsilon}(x)$ in the eigenbasis of M:

$$\vec{\epsilon}(x) = \sum_{i=1}^{s} \vec{r}_i \vec{\ell}_i^T \vec{\epsilon}(x) = \sum_{i=1}^{s} \vec{r}_i \psi_i(x), \quad \text{using that } \psi_i(x) = \vec{\ell}_i^T \vec{\epsilon}(x),$$

the RK error (2.18) can then be expressed as:

$$\epsilon_0(x) = \psi_0(x) + \sum_{i=1}^{s} \vec{\alpha}^T \vec{r}_i \psi_i(x) = \psi_0(x) + \sum_{i=1}^{s} \frac{z \vec{b}^T A^{-1} \vec{r}_i}{z - 1 + \vec{b}^T A^{-1} \vec{e}} \psi_i(x). \tag{2.25}$$

Equation (2.25) shows that the global error, $\epsilon_0(x)$, is composed of errors that are of the scheme's order (ψ_1), or of the scheme's stage order (ψ_0; ψ_2 through ψ_s); and that the error may have singular behavior. Note that if Assumption (2.5e) is violated in that $\vec{b}^T A^{-1} \vec{e} = 0$, then the coefficients $\vec{\alpha}^T \vec{r}_i$ of ψ_i in (2.25) scale like $O(\Delta t^{-1})$, resulting in an additional loss of convergence order. We conclude this section with several remarks.

REMARK 2.2 (Boundary mismatch for non-stiffly accurate schemes). Equation (2.25) shows that the error ϵ_0 evaluated at the domain boundary is: $\epsilon_0 = \psi_0$. Stiffly accurate schemes guarantee that conventional b.c. (i.e., $g_i^{n+1} = g(t_n + c_i \Delta t)$) yield $\psi_0 = \epsilon_0 = 0$ and hence exactly enforce the b.c. $u^{n+1} = g^{n+1}$. For non-stiffly accurate schemes, ϵ_0 is in general non-vanishing yielding $u^{n+1} = g^{n+1} + O(\Delta t^{q+1})$.

REMARK 2.3 (Slowly decaying modes). The calculation in this section is restricted to periodic in-time modes, which (see Appendix B) is sufficient to capture the order reduction phenomena, *provided that the normal modes for both the equation and the scheme decay in time, and do so sufficiently fast as their space frequency grows.* Here we describe two situations where this condition is violated:

Schemes with growth factor such that $|R(\zeta)| \to 1$ as $\zeta \to -\infty$. Then, the numerical solution may contain transient artifacts. For example, in the heat equation those artifacts resemble BLs, but they thin out in width slowly over time (and thus can compromise the observed order for the solution and its derivatives). The artifacts can be triggered by BLs produced in the initial step, via the mechanism outlined in Subsection 1.2. The introduced high frequency modes then die arbitrarily slowly—slower the higher the frequency, which is why the artifacts tend to become narrower as time grows. An important RK scheme exhibiting this behavior is the implicit mid-point rule, defined by the Butcher tableau $A = \vec{c} = [1/2]$ with $\vec{b} = [1]$. Because it has only one stage, the matrix M has no small eigenvalues, and the scheme has no time-periodic numerical modes with BLs. The implicit mid-point rule is the simplest case of a Gauss method, which achieve order $2s$ with s stages. These methods have $R(\zeta) \to (-1)^s$ as $\zeta \to -\infty$, thus they are all examples for the issue described here. In addition, for $s \geq 2$, they also exhibit OR in the time-periodic sense, due to existence of small eigenvalues.

A second example of "slowly decaying modes" occurs when the operator $\mathcal{L}$ in (1.1) is purely dispersive. In this case the normal modes for the equation itself are time-periodic, with no decay. An accurate numerical scheme will approximate this behavior, with normal modes that decay very slowly—at least as long as their frequencies are not too high. Just as for the schemes where $|R(\zeta)| \to 1$ as $\zeta \to -\infty$, this can lead to long-lived transients in the numerical solution (also triggered by BL effects) which compromise the observed order for the solution. An example of this situation in provided in Subsection 5.2.

REMARK 2.4 (Jordan blocks). The eigen-Equation (2.24), error expansion (2.25), and general discussion in this section, are formulated for matrices M that are diagonalizable.

However, they can be modified to include the general case in which M has Jordan blocks. To see this, assume that $\vec{\ell}_{i,j+1}^T M = \lambda_i \vec{\ell}_{i,j+1}^T + \vec{\ell}_{i,j}^T$ for $1 \le j \le J_i - 1$, where $\vec{\ell}_{i,1}$ is the eigenvector associated with λ_i and J_i denotes the size of the Jordan block corresponding to λ_i. Then Equation (2.24) is modified to

$$\psi_{i,j+1} - \lambda_i \mathcal{L}\psi_{i,j+1} = \vec{\ell}_{i,j+1}^T \vec{h} + \mathcal{L}\psi_{i,j}, \text{ with b.c. } \psi_{i,j+1} = 0, \tag{2.26}$$

where $\psi_{i,j} = \vec{\ell}_{i,j}^T \vec{\epsilon}$. Note that the occurrence of $\mathcal{L}$ in the right-hand side of (2.26) provides a BL feedback mechanism through the derivatives of the BL in the prior generalized eigenfunction. This can potentially trigger worse OR effects than in the diagonal case.

2.5. Asymptotic analysis of the boundary layers. In this subsection, we conduct an asymptotic analysis ($\Delta t \ll 1$) of the singular functions $\psi_i(x)$ and the global RK numerical error (2.25), i.e., $\epsilon_0(x)$. The modes $\psi_i(x)$ solve the BVP

$$\psi_i - \lambda_i \mathcal{L}\psi_i = H_i(\Delta t) U^*(x), \quad \text{with b.c.} \quad \psi_i = 0, \tag{2.27}$$

where we have introduced $H_i(\Delta t)$ (independent of x) by writing $\vec{\ell}_i^T \vec{h} = H_i(\Delta t) U^*(x)$ using $\vec{h}$ in (2.16). Further, by Proposition 2.2: $H_1(\Delta t) = O(\Delta t^p)$, while $H_i(\Delta t) = O(\Delta t^{q+1})$ or smaller for $2 \le i \le s$.

As mentioned in Subsection 2.4, the solution to (2.27), for $i = 1$, is non-singular since $\lambda_1 = O(1)$ (and as shown below, will not contribute to order reduction). For $2 \le i \le s$, $\lambda_j = O(\Delta t)$, and as we show here, the solution to (2.27) can be described using standard matched asymptotic expansions [9, 21, 23, 27]. Below, we work out the analysis for $\mathcal{L}$ a second order operator in one space dimension. This restriction allows us to showcase how the spatial structures that arise due to OR can be constructed in a concrete fashion. Note that in higher dimensions, such concrete constructions are not as easily done (for example, when corner layers arise); however, the general nature of OR, namely its manifestations via singularly perturbed problems and its resulting asymptotic structures, persists in any dimension.

When $U^*(x)$ is smooth, ψ_i has two BLs (one at each boundary) of thickness $O(\sqrt{\Delta t})$. Away from the BLs, the solution is described by an "outer" expansion that will not contribute to OR. Inside the BLs, the solution is described by the "inner" expansion, and together the inner and outer expansions generate a "composite" expansion valid on the whole domain.

2.5.1. Outer expansion, $2 \le i \le s$. Valid away from the boundaries (i.e., $\sqrt{\Delta t} \ll x$ and $x \ll 1 - \sqrt{\Delta t}$) is a "regular" expansion based on Taylor expanding the solution ψ_i in powers of the small parameter λ_i up to order m. Namely, $\psi_i \sim \Phi_i^m$ where Φ_i^m is the truncated (Neumann) series for $(I - \lambda_i \mathcal{L})^{-1} H_i(\Delta t) U^*$:

$$\Phi_i^m(x) = H_i(\Delta t) \left(U^* + \lambda_i \mathcal{L} U^* + \lambda_i^2 \mathcal{L}^2 U^* + \ldots + \lambda_i^m \mathcal{L}^m U^* \right), \qquad \text{for } 2 \le i \le s. \tag{2.28}$$

Equation (2.28) is an $m-$th order expansion in powers of $\lambda_i = \mu_i \Delta t + o(\Delta t)$; how large one can take m depends on how many derivatives U^* has. Note that the focus here is on m fixed and $\Delta t \to 0$. No statement is made about Δt fixed and $m \to \infty$ in (2.28).

In the following, we discuss, first in a special case and then in the general setting, the situation where $U^*(x)$ is smooth and OR occurs due to BLs in ψ_i, as well as one situation where $U^*(x)$ is not smooth and generates an internal *interface* layer in ψ_i that leads to OR inside the domain. We will make use of the following:

a. Rescaled spatial variables: $X := \frac{x}{\sqrt{\Delta t}}$, $Y := \frac{1-x}{\sqrt{\Delta t}}$ and $Z := \frac{x-1/2}{\sqrt{\Delta t}}$.

b. Exponential function: $S(x) := e^{-x}$.

c. Eigenvalues $\lambda_i = \mu_i \Delta t + o(\Delta t)$, $2 \le i \le s$: From Theorem 2.1, $\mu_i \neq 0$. Theorem 2.2(3) implies $\mu_i \notin W(\theta_2)$, so that μ_i is not a negative real number. Hence,[2] $\mathrm{Re}(\sqrt{\mu_i}) > 0$ so $S(x/\sqrt{\lambda_i}) \approx e^{-x/\sqrt{\Delta t \mu_i}}$ is exponentially (in x) and rapidly decaying (in Δt).

2.5.2. Composite solution when $\mathcal{L} = \partial_x^2$, $U^*(x)$ is smooth. The composite solution has the form $\psi_i(x) \sim \Phi_i^m(x) + \Psi_{L,i}(X) + \Psi_{R,i}(Y)$, with $\Psi_{L,i}(x)$ localized near $x = 0$. It arises from constructing an inner solution consisting of a BL function $\Psi_{L,i}(X)$ that connects the b.c. $\psi_i(0) = 0$ to the outer solution $\psi_i(x) \sim \Phi_i^m(x)$ when $\sqrt{\Delta t} \ll x \ll 1 - \sqrt{\Delta t}$. After rescaling space, the left BL (the one near $x = 1$ is analogous) function $\Psi_{L,i}(X)$ solves the ODE $\Psi_{L,i} - \mu_i \Psi''_{L,i} = 0$, with b.c. $\Psi_{L,i}(0) = -\Phi_i^m(0)$ and $\Psi_{L,i}(+\infty) = 0$. The general solution of this ODE is a superposition of the exponentials $S(\pm X/\sqrt{\mu_i})$, which after matching b.c. and using property c. (i.e., only the $+X$ exponential contributes) yields $\Psi_{L,i}(X) = -\Phi_i^m(0) S(X/\sqrt{\mu_i})$. By a similar argument, the right BL has the form $\Psi_{R,i}(Y) = -\Phi_i^m(1) S(Y/\sqrt{\mu_i})$, and together the m-th order composite expansion valid in the whole interval is:

$$\psi_i \sim \Phi_i^m(x) - \Phi_i^m(0)\, e^{-\frac{X}{\sqrt{\mu_i}}} - \Phi_i^m(1)\, e^{-\frac{Y}{\sqrt{\mu_i}}}, \qquad \text{for } 2 \le i \le s. \tag{2.29}$$

2.5.3. Composite solution when $\mathcal{L} = \alpha_2(x)\partial_x^2 + \alpha_1(x)\partial_x + \alpha_0(x)$, $U^*(x)$ is smooth. Here $\alpha_2(x) > 0$ is positive. The asymptotics of the variable coefficient $\mathcal{L}$ are only a minor modification of the case $\mathcal{L} = \partial_x^2$. After rescaling space near the left BL:

$$\lambda_i \mathcal{L} = \alpha_2(X\Delta t) \frac{\lambda_i}{\Delta t} \partial_X^2 + \alpha_1(X\Delta t) \frac{\lambda_i}{\sqrt{\Delta t}} \partial_X + \alpha_0(X\Delta t) = \mu_i \alpha_2(0)\, \partial_X^2 + o(\Delta t),$$

so that at leading order in Δt, $\Psi_{L,i}(X)$ solves $\Psi_{L,i} - \mu_i \alpha_2(0) \Psi''_{L,i} = 0$, with b.c. $\Psi_{L,i}(0) = -\Phi_i^m(0)$ and $\Psi_{L,i}(+\infty) = 0$. The solution (2.29) is modified to:

$$\psi_i \sim \Phi_i^m(x) - \Phi_i^m(0)\, e^{-\frac{X}{\sqrt{\mu_i \alpha_2(0)}}} - \Phi_i^m(1)\, e^{-\frac{Y}{\sqrt{\mu_i \alpha_2(1)}}}, \qquad \text{for } 2 \le i \le s. \tag{2.30}$$

2.5.4. Composite solution when $\mathcal{L} = \alpha_2(x)\partial_x^2 + \alpha_1(x)\partial_x + \alpha_0(x)$, $U^*(x)$ is piecewise smooth. Here $\alpha_2(x) > 0$ is positive. Assume $U^*(x)$ is C^∞ on $[0, \frac{1}{2})$ and also on $(\frac{1}{2}, 1]$, and both $U^*(x)$ and all its derivatives have finite one-sided limits at $\frac{1}{2}$ (where $x = \frac{1}{2}$ is chosen without loss of generality). In addition, there is some $1 \le \kappa < m$ where $\mathcal{L}^\kappa U^*$ does not exist at $x = \frac{1}{2}$. The expression (2.30) fails near $x = \frac{1}{2}$ since $\Phi_i^m(\frac{1}{2})$ does not exist, however (2.30) still remains valid on $0 \le x \ll \frac{1}{2} - \sqrt{\Delta t}$ and separately on $\frac{1}{2} + \sqrt{\Delta t} \ll x \le 1$. In the vicinity of $x = \frac{1}{2}$, the function ψ_i has an internal layer $\Psi_{I,i}(Z)$ that connects $\psi_i(x)$ to the two sides of the outer solution $\psi_i(x) \sim \Phi_i^m(x)$ as $|x - \frac{1}{2}| \gg \Delta t$. After rescaling space, $\Psi_{I,i}(Z)$ solves $\Psi_{I,i} - \mu_i \alpha_2(\frac{1}{2}) \Psi''_{I,i} = 0$ with b.c. $\Psi_{I,i}(\pm\infty) = 0$ and interface conditions that enforce continuity of ψ_i and ψ_i' across $x = \frac{1}{2}$. Let $\Phi_\pm := \lim_{\tau \to 0^+} \Phi_i^m(\frac{1}{2} \pm \tau)$ and $\Phi'_\pm := \lim_{\tau \to 0^+} \left(\frac{d\Phi_i^m}{dx}(\frac{1}{2} \pm \tau) \right)$ and introduce the jumps across $x = \frac{1}{2}$ as $[\Phi] := \Phi_+ - \Phi_-$ and $[\Phi'] := \Phi'_+ - \Phi'_-$. We have

$$\Psi_{I,i}(Z) = \left(-\frac{\mathrm{sgn}(Z)}{2} [\Phi] + \frac{\sqrt{\Delta t \mu_i \alpha_2(\frac{1}{2})}}{2} [\Phi'] \right) e^{-\frac{|Z|}{\sqrt{\mu_i \alpha_2(\frac{1}{2})}}}, \tag{2.31}$$

[2] Using the principal branch of the square root.

where $\mathrm{sgn}(Z)$ is the sign of Z. The composite solution (2.29), is modified to ($\Psi_{L,i}$, $\Psi_{R,i}$ are the same as (2.29))

$$\psi_i \sim \begin{cases} \Phi_i^m(x)+\Psi_{L,i}(X)+\Psi_{R,i}(Y)+\Psi_{I,i}(Z), & \text{for } x\neq\frac{1}{2}, \\ \frac{1}{2}(\Phi_++\Phi_-)+\frac{\sqrt{\Delta t\mu_i\alpha_2(\frac{1}{2})}}{2}[\Phi'] & \text{for } x=\frac{1}{2} \end{cases} \quad \text{for} \quad 2\le i\le s. \tag{2.32}$$

2.5.5. Shape of the RK error inside a boundary layer. We turn our attention to the global error $\epsilon_0(x)$. When the functions ψ_i are well approximated by Taylor expansions in powers of Δt (about $\Delta t=0$) for $\Delta t\ll 1$, the RK scheme is designed so that both the modes ψ_i, and coefficients $\vec{\alpha}^T\vec{r}_i$, may be expanded via Taylor series in (2.25), and cancel out to order $O(\Delta t^p)$. The regular solution Φ_i^m is exactly the part of ψ_i that can be expanded via Taylor series (the boundary or interface layers, i.e., $\Psi_{L,i}(X)$, can not). Note that, while we present the analysis here for the special case of a singularity as in §2.5.4, the formulation (2.30), and thus also (2.32), is the most general form if $U^*(x)$ is smooth. Substituting the form (2.32) into (2.25), we group the terms as follows:

$$\epsilon_0(x)=\underbrace{\left[\vec{\alpha}^T\vec{r}_1\psi_1(x)+\psi_0(x)+\sum_{i=2}^{s}\vec{\alpha}^T\vec{r}_i\Phi_i^m(x)\right]}_{\text{Bracket } 1=O(\Delta t^p)}+\underbrace{\left[\sum_{i=2}^{s}\vec{\alpha}^T\vec{r}_i\left(\Psi_{L,i}(X)+\Psi_{R,i}(Y)+\Psi_{I,i}(Z)\right)\right]}_{\text{Bracket } 2}. \tag{2.33}$$

Note that the terms in Bracket 1 of (2.33) are individually $O(\Delta t^{q+1})$ (or smaller). However, all those terms can be Taylor-expanded in powers of Δt, and by consistency of the RK scheme, they sum together to $O(\Delta t^p)$ (or smaller); see Appendix C for a formal proof. Bracket 2 has terms $\Psi_{\beta,i}$, $\beta=\{L,R,I\}$ that are potentially $O(\Delta t^{q+1})$, however do not have Taylor expansions in Δt and generally do not cancel to high order. Note that the magnitude of $\Psi_{L,i}$, $\Psi_{R,i}$ always occurs at the boundary, which is in general $\Phi_i^m=O(\Delta t^{q+1})$, unless the leading order terms in the regular expansion Φ_i^m, i.e., U^*, $\mathcal{L}U^*$, vanish on the boundary. Similarly, if $U^*(x)$ has a singularity at $x=\frac{1}{2}$, the magnitude of $\Psi_{I,i}$ is determined by the jumps in the regular solution $[\Phi]$ and $\sqrt{\Delta t}[\Phi']$ at $x=\frac{1}{2}$ and is determined by the largest value of κ for which $\mathcal{L}^\kappa U^*(\frac{1}{2})$ exists. The fact that a loss of spatial regularity in U^* can result in order reduction has been discussed in [36]. However, the results in §2.5.4 and (2.33) characterize the precise asymptotic shape of the error in the vicinity of a singularity in U^*.

We now examine (2.33) in the vicinity of the left boundary $x=0$ (the right boundary, or near a point x where $U^*(x)$ is not smooth is similar). Taylor-expanding $\vec{\alpha}^T\vec{r}_i$ in Δt and using the fact that Bracket 1 is $O(\Delta t^p)$, the error near the left boundary is:

$$\epsilon_0(x)=\Delta t^{q+1}\sum_{i=2}^{s}P_i(\Delta t)\,S\left(\frac{x}{\sqrt{\Delta t\mu_i\alpha_2(0)}}\right)+O(\Delta t^p), \qquad \text{for } 0\le x\ll 1, \tag{2.34}$$

where $P_i(\Delta t)$ $(2\le i\le s)$ are polynomials of degree $p-q-2$. Equation (2.34) reveals the structure in the BLs, and explains why *(generically) RK schemes incur order reduction in BLs.* Specifically: the functions $S(\frac{x}{\sqrt{\Delta t\mu_i\alpha_2(0)}})$ are singular in Δt, and are linearly independent when the μ_i are distinct. Hence, Taylor-based cancellations in the summation (2.34) (on which the scheme relies to achieve its order) do not occur. The convergence order in $\epsilon_0(x)$ is controlled by the coefficients $P_i(\Delta t)$ of the $O(1)$ functions $S(\frac{x}{\sqrt{\Delta t\mu_i\alpha_2(0)}})$. An alternative viewpoint when the μ_i in (2.34) are distinct, is that the functions $\Psi_{L,i}$ have different widths of $O(\sqrt{\Delta t})$, see Subsection 2.5. Hence they

generally do not cancel through a linear combination and result in a "composite" BL in $\epsilon_0(x)$. Note that: if $\sqrt{\mu_i}$ is not a real number, the BL includes high frequency oscillations triggered by $\mathrm{Im}(\sqrt{\mu_i})$. How visible these oscillations are depends on the ratio $\mathrm{Im}(\sqrt{\mu_i})/\mathrm{Re}(\sqrt{\mu_i})$. The larger this ratio, the larger the role of the oscillations.

Finally, away from the BLs ($\Delta t \ll x \ll 1-\Delta t$) or any point where U^* is singular, the functions $\Psi_{\beta,i}$ for $\beta=\{L,R,I\}$ are exponentially small (so that Bracket 2 in (2.33) is exponentially small); hence $\psi_i \sim \Phi_i^m$ is just the regular solution and consequently $\epsilon_0(x)$ does not suffer from order reduction.

Equation (2.34) also highlights a crucial structural property of the approximation error of RK schemes for IBVPs. Aside from a few special cases (e.g., backward Euler), the singular functions $S(\frac{x}{\sqrt{\Delta t \mu_i \alpha_2(0)}})$ can generally not be avoided. Instead, methods that overcome order reduction (cf. Section 3 and Section 4) render the singular functions in $\epsilon_0(x)$ to be of the scheme's formal order p, or higher. While this remedies OR in the solution u, the persistence of BLs (or internal layers) implies that (sufficiently high) spatial derivatives of the solution generally still incur OR. The fact that derivatives generally are less accurate follows because the BL or internal layer functions $\Psi(x)$ (in unscaled variables) always satisfy the homogeneous equation $\mathcal{L}\Psi(x) \propto \Delta t^{-1}\Psi(x)$, which shows that $\mathcal{L}\Psi(x)$ is one order less than $\Psi(x)$ (see also, [3]).

2.5.6. Beyond second order operators $\mathcal{L}$. The asymptotic analysis leading to (2.33) reveals that the BLs' error contributions amplify by $\Delta t^{-1/\ell}$ per spatial derivative, if $\mathcal{L}$ is an ℓ-th order differential operator. These considerations are of particular importance for any practical problem in which gradients of the solution at/near the boundary are needed.

REMARK 2.5 (High order equations and composite BLs). For general $\mathcal{L}$, a version of (2.34) holds and the structure of the RK error can be obtained via asymptotic analysis; however the BLs in (2.34) are determined by the highest order derivative in $\mathcal{L}$. For example, let $\mathcal{L}=\alpha_\ell(x)\partial_x^\ell+\alpha_{\ell-1}(x)\partial_x^{\ell-1}+\ldots+\alpha_0(x)$, where $\alpha_\ell(x)$ is a nonvanishing function. Then the modes $\Psi_{R,i}(x)$ and $\Psi_{L,i}(x)$ contain a superposition of exponentials $\exp(x/(\Delta t^{1/\ell}\rho_{ij}))$, where ρ_{ij} $(1\le j\le \ell)$ are the ℓ roots of $\rho_{ij}^\ell=\alpha_\ell(0)\mu_i$. Values of ρ_{ij} with negative (resp. positive) real part correspond to an exponentially (in x) decaying (resp. growing) function, and contribute to a BL near $x=0$ (resp. $x=1$). Values of ρ_{ij} on the imaginary axis correspond to purely oscillatory functions.

Remark 2.5 highlights that in principle the roots ρ_{ij} of the singular Equation (2.24) could be purely imaginary, leading to modes $\psi_i \sim \exp(x/(\Delta t^{1/\ell}\rho_{ij}))$ that have *high frequency oscillations* (HFO) extending over the whole domain. (This is the same type of behavior that arises in WKB theory [9]). For constant coefficient PDEs, however, dissipation (i.e., the spectrum of $\mathcal{L}$ is contained within the wedge in (2.5a)) eliminates the possibility of HFO:

LEMMA 2.1 (High frequency oscillations). *Under the assumptions in* (2.5), *HFO cannot occur for constant coefficient differential operators* $\mathcal{L}$.

Proof. Suppose that the constant coefficient $\mathcal{L}=\alpha\partial_x^\ell+$lower order terms, where $\ell\in\mathbb{N}$ and $\alpha\neq 0$, were to produce HFO in the numerical error. That means, in the limit as $\Delta t\to 0$, there is at least one mode $\psi_i\sim\exp(x/(\Delta t^{1/\ell}\imath r))$ that solves the ODE $(1-\Delta t\mu_i\mathcal{L})\psi_i=0$, where r is purely real. Substituting $\mathcal{L}$ and the exponential ansatz for ψ_i into the ODE, yields the relationship $\alpha=(\imath r)^\ell/\mu_i$. This allows us to write the

eigenvalues of $\mathcal{L}$ (found by substituting $e^{\imath kx}$ into $\mathcal{L}u = \lambda u$) as

$$\lambda = (rk)^\ell/\mu_i + (\text{lower order terms in } k), \tag{2.35}$$

where k is real. In general only certain values of k are allowed, but those (infinitely many) include arbitrarily large k. We now show that (2.35) violates the hypothesis (2.5) when $k \to \infty$. Specifically:

(i) By Theorem 2.2(3): $|\arg(1/\mu_i)| \le \pi - \theta_2$ (θ_2 defines the wedge in (2.5b)).

(ii) Assumption (2.5a) asserts that the eigenvalues λ of $\mathcal{L}$ satisfy $|\arg(\lambda)| > \pi - \theta_1$.

In the limit as $k \to \infty$, the eigenvalues (2.35) grow with a slope that approaches $1/\mu_i$ (i.e., $\mathrm{Im}(\lambda)/\mathrm{Re}(\lambda) \to \mathrm{Im}(\mu_i^{-1})/\mathrm{Re}(\mu_i^{-1})$ as $k \to \infty$). Since $1/\mu_i$ has a slope with angle $\le \pi - \theta_2$, the eigenvalues must cross every line with a larger slope angle, including the line with angle $\pi - \theta_1$. However, this violates item (ii). ∎

The assumptions in (2.5) were introduced to (in particular) guarantee that the scheme equations can be solved at each step, and avoid numerical instabilities. But they are in no way necessary (particularly, they exclude non-dissipative systems to simplify the analysis), and less constraining assumptions are possible. Thus, in the general situation we cannot rule out HFO—even though we have not observed them in actual examples. That being said, even if HFO situations are not possible, BLs that include oscillations do occur. Further, because the BL thickness scales like $\Delta t^{1/\ell}$, unless Δt is very small, these BLs with oscillations can be quite thick, for example see Figure 5.2 in Subsection 5.2.

REMARK 2.6 (Multiple dimensions). Clearly, OR occurs in higher dimensions as well. Along smooth parts of the domain boundary, BLs should arise, similarly to the 1D BLs studied here. In addition, at corners, a breakdown in solution smoothness can also effect OR; see [37] for error estimates.

2.5.7. Numerical example. We illustrate the modal analysis with the same 1D heat equation example used in Subsection 1.1: a time-varying, constant-in-space, exact solution, approximated with two DIRK schemes: a 3-stage stiffly accurate DIRK3 and a 2-stage non-stiffly accurate DIRK3. The errors are computed with the eigenmodes, obtained by solving (2.24) with standard second order finite differences on a fixed uniform grid with 2000 points.

Figure 2.1 shows the eigenmodes $\psi_i(x)$, and the spatial part of the global-in-time error $\epsilon_0(x)$, for the stiffly accurate scheme (left panel), and the non-stiffly accurate scheme (right panel). Three different choices of time step Δt are used.

For the 3-stage stiffly accurate DIRK3, we have

$$\psi_1 = O(\Delta t^3), \quad \psi_2 = O(\Delta t^2), \quad \psi_3 = O(\Delta t^2), \quad \text{and} \quad \psi_0 \equiv 0 \,(\text{not shown}).$$

The modes ψ_2 and ψ_3 solve singularly perturbed BVPs, and produce BLs in the global-in-time error ($\epsilon_0 = \epsilon_3$, since the scheme is stiffly accurate). All modes vanish at the boundary, hence the numerical solution has no error in the boundary values.

For the 2-stage non-stiffly accurate DIRK3, we have

$$\psi_1 = O(\Delta t^3), \quad \psi_2 = O(\Delta t^2), \quad \text{and} \quad \psi_0 = O(\Delta t^2).$$

The mode ψ_2 solves a singularly perturbed BVP, and produces a BL in the global error. The constant-in-space mode ψ_0 (associated to the zero eigenvalue) produces a mismatch in the boundary values for ϵ_0. This is reflecting the fact that non-stiffly accurate schemes

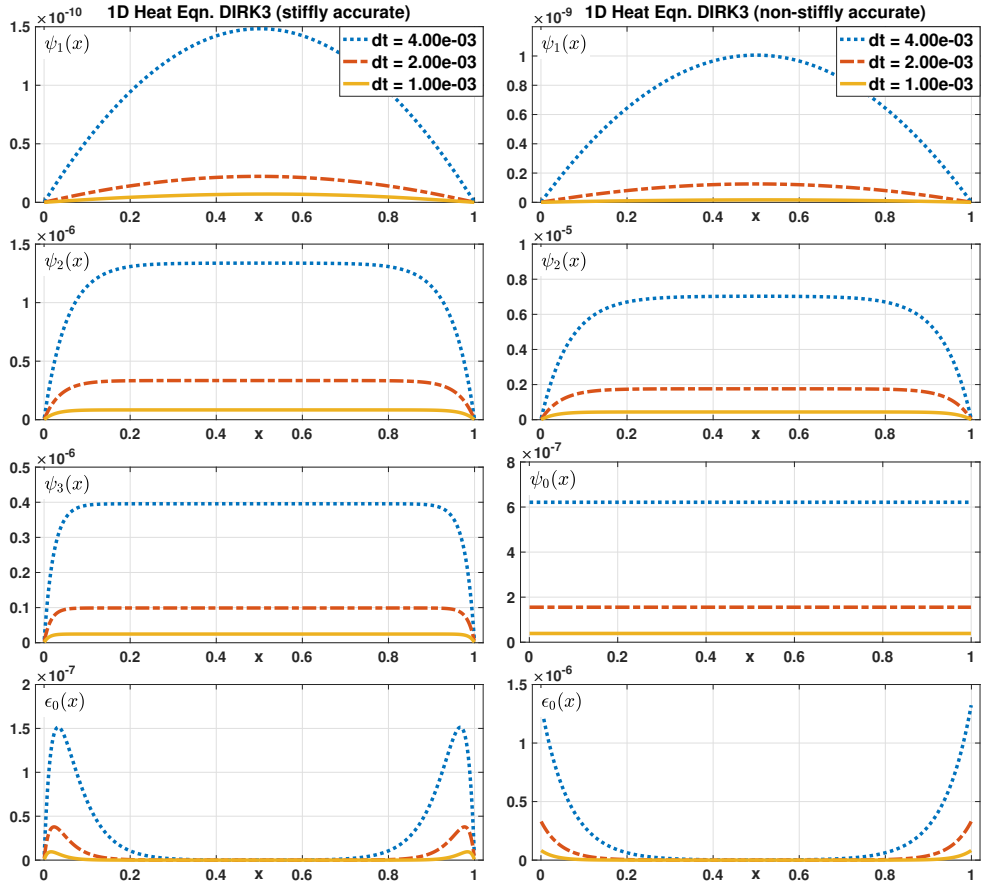

FIG. 2.1. *Eigenmodes $\psi_i(x)$ (real part) and error $\epsilon_0(x)$ for a 3-stage, 3rd order, stiffly accurate DIRK (left), and a 2-stage, 3rd order, non-stiffly accurate DIRK (right). Three Δt choices used.*

with conventional boundary conditions do not guarantee that the numerical solution satisfies the exact boundary conditions, see Remark 2.2.

For both schemes, (i) the mode ψ_1 associated with the $O(1)$ eigenvalue has no BLs, and (ii) the global-in-time error exhibits the full 3rd order accuracy away from the BLs.

3. Weak stage order

High stage order, i.e., $q>1$, is not possible with DIRK schemes (see Remark 2.1). In this section we introduce new conditions on $(A,\vec{b},\vec{c})$ that relax the stage order condition to a more general one that is both: (i) devoid of order reduction (OR); and (ii) compatible with a DIRK structure. We therefore refer to the condition as *weak stage order* (WSO).

Weak stage order addresses OR by means of only the RK time-stepping coefficients, and it is compatible with a DIRK structure (cf. [48] for a similar in spirit approach for ROW methods). The concept of WSO generalizes the work by [42] and provides the sharpest condition on the Butcher coefficients to alleviate OR in linear ODEs. For the time-stepping of PDEs, WSO can provide a practical approach for avoiding OR when using conventional b.c.. It should be stressed that neither high stage order, nor high weak stage order fully remove BLs from the numerical solution. Rather, they reduce

the size of the BLs (to the order of the scheme). Thus, in general OR still occurs in the solution's derivatives.

3.1. Definition of weak stage order. The idea behind WSO is to find conditions on $(A,\vec{b},\vec{c})$ independent of Δt, that decrease the amplitude (from $O(\Delta t^{q+1})$) of the singular terms $\vec{\alpha}^T\vec{\epsilon}$ appearing in the error $\epsilon_0(x)$ from (2.18); equivalently viewed as decreasing the contributions $\vec{\alpha}^T\vec{r}_i\,\psi_i$ of the singular functions ψ_i in (2.25). We show below that the invariant subspaces of the Butcher tableau matrix A play a key role and define the WSO condition:

DEFINITION 3.1 (Weak stage order). *A Runge-Kutta scheme $(A,\vec{b},\vec{c})$ has weak stage order $\tilde{q}\geq 1$, if there exists a vector space $\mathcal{V}\subseteq\mathbb{R}^s$ that contains the stage order residuals $\vec{\tau}^{(j)}\in\mathcal{V}$, for $1\leq j\leq\tilde{q}$, and also satisfies the following two properties:*

(i) (Orthogonality property.) *$\mathcal{V}\subset\vec{b}^{\perp}$, i.e., $\vec{b}^T\vec{v}=0$ for all $\vec{v}\in\mathcal{V}$.*

(ii) (Invariant subspace property.) *$\mathcal{V}$ is A-invariant, i.e., $A\vec{v}\in\mathcal{V}$ for all $\vec{v}\in\mathcal{V}$.*

REMARK 3.1 (Weak stage order as order conditions). By the Cayley-Hamilton theorem, Definition 3.1 is equivalent to (see [25] for a short proof): The vector $\vec{b}$ is orthogonal to the column space $C(G)$ of the (controllability) matrix

$$G:=\left(\vec{\tau}^{(1)},A\vec{\tau}^{(1)},\ldots,A^{s-1}\vec{\tau}^{(1)},\vec{\tau}^{(2)},A\vec{\tau}^{(2)},\ldots,A^{s-1}\vec{\tau}^{(\tilde{q})}\right)\in\mathbb{R}^{s\times s\tilde{q}}. \tag{3.1}$$

The standard order conditions already imply, via Proposition 2.1, that $\vec{b}$ is orthogonal to a subset of the columns of G. Hence, WSO can be viewed as imposing extra order conditions $\vec{b}\perp C(G)$.

Definition 3.1 generalizes the notion of stage order, and is automatically satisfied by a scheme with classical stage order q. Every RK scheme has both classical stage order $q\geq 1$ and WSO $\tilde{q}\geq 1$, since $A\vec{e}=\vec{c}$ guarantees that $\vec{\tau}^{(1)}=\vec{0}$. The (abstract) Definition 3.1 is helpful in simplifying proofs involving WSO, while the alternative viewpoint in Remark 3.1 is useful in practice to construct schemes satisfying high WSO. Note that WSO $\tilde{q}$ implies WSO $\tilde{q}-1$, which follows directly from the construction of G in Remark 3.1. A simplifying criterion for WSO arises when the stage order residuals $\vec{\tau}^{(j)}$ are eigenvectors of A. We refer to this situation as the weak stage order eigenvector criterion:

DEFINITION 3.2 (WSO eigenvector criterion). *A RK scheme satisfies the eigenvector criterion of order $\tilde{q}_e$ if for each $1\leq j\leq\tilde{q}_e$ there exists ζ_j such that $A\vec{\tau}^{(j)}=\zeta_j\vec{\tau}^{(j)}$, and $\vec{b}^T\vec{\tau}^{(j)}=0$.*

Weak stage order is a linear concept, and the analysis in this section shows that it remedies order reduction for linear IBVPs (1.1). In contrast, for problems in which the root cause of order reduction itself is nonlinear, or time dependent, WSO may not achieve the same benefit that classical stage order does. See also [25], which further devises additional schemes that satisfy the WSO eigenvector criterion, as well as limitations of the WSO eigenvector criterion.

3.2. Impact of weak stage order on error convergence and boundary layers. We show that weak stage order, paired with Assumptions (2.5), can avoid order reduction in RK schemes for the periodically forced solutions examined in Section 2, i.e., that $\epsilon_0=O(\Delta t^{\min\{\tilde{q}+1,p\}})$. The following proposition demonstrates how solutions to (2.18) with a right-hand side proportional to $\vec{\tau}^{(j)}$ for $j\leq\tilde{q}$ do not contribute to error $\epsilon_0(x)$.

PROPOSITION 3.1. *Consider a Runge-Kutta scheme* $(A,\vec{b},\vec{c})$ *with WSO* $\tilde{q}$*, and let* M *be given by* (2.15)*. Then for any smooth function* $f(x)$ *and stage order residuals* $\vec{\tau}^{(j)}$*,* $1\le j\le\tilde{q}$*, the following quantities vanish for any* $x\in\Omega$*:* $\vec{b}^T\vec{v}(x)=0$ *and* $\vec{b}^TA^{-1}\vec{v}(x)=0$*, where* $\vec{v}(x)$ *solves*

$$(I-M\mathcal{L})\vec{v}=f(x)\vec{\tau}^{(j)}\quad \text{with b.c. } \vec{v}=0. \tag{3.2}$$

Proof. Let $\mathcal{V}$ denote the A-invariant subspace in Definition 3.1. It suffices to show that $\vec{v}(x)\in\mathcal{V}$ for all $x\in\Omega$. Then $\vec{b}^T\vec{v}(x)=0$ follows by property (i) in Definition 3.1, and property (ii) implies that $\mathcal{V}$ is also A^{-1}-invariant, so that $A^{-1}\vec{v}(x)\in\mathcal{V}$, and thus $\vec{b}^TA^{-1}\vec{v}(x)=0$.

We show that $\vec{v}(x)\in\mathcal{V}$ by working in a coordinate basis defined by $\mathcal{V}$ and its orthogonal space $\mathcal{V}^\perp$. Let $\sigma=\dim(\mathcal{V})$. Note that no assumption is made on σ relative to $\tilde{q}$. If some of the vectors $\vec{\tau}^{(j)}$ are linearly dependent or vanish, $\sigma<\tilde{q}$ is possible; and $\sigma\ge\tilde{q}$ is possible if the vectors $\vec{\tau}^{(j)}$ are only contained in a larger A-invariant space, but do not span an A-invariant space themselves.

Let $\{\vec{v}_1,\cdots,\vec{v}_\sigma\}$ and $\{\vec{v}_{\sigma+1},\ldots,\vec{v}_s\}$ form two orthonormal bases for $\mathcal{V}$ and $\mathcal{V}^\perp$, respectively. Moreover, define the matrices $V=\left(\vec{v}_1,\ldots,\vec{v}_\sigma\right)\in\mathbb{R}^{s\times\sigma}$ and $V_\perp=\left(\vec{v}_{\sigma+1},\ldots,\vec{v}_s\right)\in\mathbb{R}^{s\times(s-\sigma)}$, and denote the full orthogonal matrix $P=\left(V,V_\perp\right)\in\mathbb{R}^{s\times s}$. We have $\vec{b}^TV=0$, thus (and similarly for $\vec{b}^TA^{-1}\vec{v}(x)$):

$$\vec{b}^T\vec{v}(x)=\vec{b}^TPP^T\vec{v}(x)=\vec{b}^TVV^T\vec{v}(x)+\vec{b}^TV_\perp V_\perp^T\vec{v}(x)=\vec{b}^TV_\perp V_\perp^T\vec{v}(x).$$

We now show that $V_\perp^T\vec{v}(x)=0$, which will complete the proof. First, observe that $\mathcal{V}$ is also an M-invariant subspace: for any $\vec{v}\in\mathcal{V}$ we have

$$M\vec{v}=\left(\frac{\Delta t}{z-1}\vec{e}\vec{b}^T+\Delta tA\right)\vec{v}=\frac{\Delta t}{z-1}\vec{e}\left(\vec{b}^T\vec{v}\right)+\Delta tA\vec{v}=\Delta t(A\vec{v})\in\mathcal{V},$$

which follows from the fact that $\mathcal{V}$ is A-invariant and orthogonal to $\vec{b}$.

Because $\mathcal{V}$ is M-invariant, the matrix P^TMP (which is M written in the coordinate basis $\{\vec{v}_j\}_{j=1}^s$) is block upper-triangular [35, Chapter 8.6]. Multiplying (3.2) by P^T, and using the block structure of P^TMP, we obtain

$$\left[\begin{pmatrix}I&0\\0&I\end{pmatrix}-\begin{pmatrix}V^TMV&V^TMV_\perp\\0&V_\perp^TMV_\perp\end{pmatrix}\mathcal{L}\right]\begin{pmatrix}V^T\vec{v}(x)\\V_\perp^T\vec{v}(x)\end{pmatrix}=f(x)\begin{pmatrix}V^T\vec{\tau}^{(j)}\\V_\perp^T\vec{\tau}^{(j)}\end{pmatrix}.$$

Hence the vector field $V_\perp^T\vec{v}(x)$ decouples from the $V^T\vec{v}(x)$ components. Moreover, the corresponding right-hand side vanishes, $V_\perp^T\vec{\tau}^{(j)}=0$, because $\vec{\tau}^{(j)}\in\mathcal{V}$. Hence,

$$\left[I-V_\perp^TMV_\perp\mathcal{L}\right](V_\perp^T\vec{v}(x))=0,\quad \text{with b.c. } V_\perp^T\vec{v}(x)=0, \tag{3.3}$$

where the coordinate transformation does not modify the homogeneous b.c.. If $V_\perp^T\vec{v}(x)\neq 0$ were to solve Equation (3.3), then Equation (3.2) would not have a unique solution, in contradiction to Corollary 2.1. Therefore, $V_\perp^T\vec{v}(x)=0$ is the unique solution to (3.3). ∎

The importance of Proposition 3.1 is that it does not depend on either z or Δt. We now state the main theorem demonstrating that weak stage order avoids OR in IBVP.

THEOREM 3.1. *Consider an* s*-stage,* p*-th order implicit Runge-Kutta scheme with weak stage order* $\tilde{q}\ge 1$*, satisfying the assumptions in* (2.5)*. Then the convergence order for periodic solutions with conventional b.c. is* $\min\{p,\tilde{q}+1\}$*, i.e.,* $\epsilon_0=O(\Delta t^{\min\{p,\tilde{q}+1\}})$.

Proof. Using the definition of the LTEs (2.12), write $\vec{\delta} = \vec{\varphi}(x) + O(\Delta t^{\tilde{q}+1})$ where $\vec{\varphi}(x)$ is a linear combination of stage order residuals: $\vec{\varphi}(x) := U^*(x) \sum_{j=2}^{\tilde{q}} \frac{(\imath\omega)^j \Delta t^j}{(j-1)!} \vec{\tau}^{(j)}$. Next, we expand the error $\epsilon_0(x)$ in (2.18) in terms of $\Delta t \ll 1$, and use the fact that $\delta_0 = O(\Delta t^{p+1})$, and $z = 1 + O(\Delta t)$, to obtain:

$$\epsilon_0(x) = \left(\frac{1}{\vec{b}^T A^{-1} \vec{e}} + O(\Delta t) \right) \left(\vec{b}^T A^{-1} \vec{\epsilon}(x) - \vec{b}^T A^{-1} \vec{\varphi}(x) + O(\Delta t^{\min\{p+1, \tilde{q}+1\}}) \right). \quad (3.4)$$

In (3.4), $\vec{b}^T A^{-1} \vec{e} \neq 0$ by Assumption (2.5e). Furthermore, the term $\vec{b}^T A^{-1} \vec{\varphi}(x) = 0$, because $\varphi(x) \in \mathcal{V}$ is a linear combination of $\vec{\tau}^{(j)}$ (see Proposition 3.1). Therefore we need to estimate $\vec{b}^T A^{-1} \vec{\epsilon}(x)$. Proposition 3.1 implies that $\vec{b}^T A^{-1} \vec{\epsilon}_\varphi = 0$, where $\vec{\epsilon}_\varphi$ solves

$$\vec{\epsilon}_\varphi - M\mathcal{L}\vec{\epsilon}_\varphi = z^{-1} \vec{\varphi}(x), \quad \text{with b.c. } \vec{\epsilon}_\varphi = 0.$$

Since $\vec{\epsilon} - \vec{\epsilon}_\varphi$ solves a BVP similar to (2.17) with right-hand side $z^{-1}(\vec{\delta} + \frac{\delta_0 \vec{e}}{z-1} - \vec{\varphi}) = O(\Delta t^{\tilde{q}+1}) + O(\Delta t^p)$, Corollary 2.1 implies that $\vec{\epsilon} - \vec{\epsilon}_\varphi = O(\Delta t^{\min\{p, \tilde{q}+1\}})$. We may then subtract $\vec{\epsilon}_\varphi$ from $\vec{\epsilon}$ and compute

$$\vec{b}^T A^{-1} \vec{\epsilon} = \vec{b}^T A^{-1} (\vec{\epsilon} - \vec{\epsilon}_\varphi) = O(\Delta t^{\tilde{q}+1}) + O(\Delta t^p) = O(\Delta t^{\min\{p, \tilde{q}+1\}}),$$

which finalizes the proof. ∎

Theorem 3.1 demonstrates that order reduction in function value (for periodic solutions) can be avoided with a weak stage order $\tilde{q} = p - 1$. The following remark, and numerical calculations in the following sections, indicate that high WSO removes order reduction for non-periodic solutions as well.

REMARK 3.2. Ostermann and Roche [36] proved (under assumptions similar to (2.5)) that if a RK scheme applied to (1.1) with homogeneous boundary conditions satisfies:

$$W_k(z) \equiv 0, \quad \text{for } 1 \leq k \leq \tilde{q} \qquad \text{where} \quad W_k(z) := \frac{k \vec{b}^T (I - zA)^{-1} \vec{\tau}^{(k)}}{R(z) - 1}, \quad (3.5)$$

then the scheme converges in L^r $(1 < r < \infty)$, with order $\min\{p, \tilde{q} + 2 + \nu\}$, where ν depends[3] on $\mathcal{L}$ and r. The focus of [36] was to establish L^r convergence results, and not to investigate what conditions would guarantee (3.5). It is easy to verify that WSO $\tilde{q}$ immediately implies (3.5) since $\mathcal{V}$ is an invariant subspace of $(I - zA)$, and $\vec{b}^T (I - zA)^{-1} \vec{\tau}^{(k)} = 0$ for all $1 \leq k \leq \tilde{q}$.

Note that (i) homogeneous b.c. on u^* increase the convergence rate to $\min\{p, \tilde{q} + 2\}$; while (ii) the constant ν stems from measuring the BL size of the ψ_i's in the L^r norm, i.e., ψ_i has BL width $O(\sqrt{\Delta t})$ for $\mathcal{L} = \partial_{xx}$. Items (i–ii) imply that the convergence rate of $\min\{p, \tilde{q} + 2 + \nu\}$ proved in [36] is consistent with Theorem 3.1 and Section 2.

3.3. A DIRK scheme with high weak stage order. An important advantage of WSO is that it allows DIRK schemes to avoid order reduction (cf. Remark 2.1). Here we present a stiffly accurate, L-stable, 4-stage, 3rd order DIRK scheme with WSO 2. This scheme is constructed using the eigenvector criterion in Definition 3.2, i.e., the stage order residual $\vec{\tau}^{(2)}$ is a right eigenvector of A. The coefficients $A = (a_{ij}) \in \mathbb{R}^{4 \times 4}$,

[3] For $\mathcal{L} = \partial_{xx}$, and L^2 convergence, $\nu = \frac{1}{4} - \varepsilon$ for any $\varepsilon > 0$, so effectively one can take $\nu = \frac{1}{4}$.

and $\vec{b}$, $\vec{c}\in\mathbb{R}^4$ are given by

$$A=\begin{bmatrix} 0.019000728905359 & & & \\ 0.404346056017447 & 0.384357175123333 & & \\ 0.064879084117003 & -0.163896402946036 & 0.515452312221597 & \\ 0.023435493738931 & -0.412078778885435 & 0.966611612813460 & 0.422031672333044 \end{bmatrix}, \tag{3.6}$$

$$b_i=a_{4i} \quad \text{and} \quad c_i=\sum_{j=1}^{i} a_{ij} \quad \text{for} \quad 1\le i\le 4.$$

In line with [25], this scheme has been found by searching the parameter space of stiffly accurate 4-stage DIRK schemes (with nonzero diagonal entries), while imposing the order conditions (2.3), the WSO eigenvector criterion (Definition 3.2), and A-stability (verified by evaluating the stability function $R(\zeta)$ along the imaginary axis) as constraints. MATLAB's `fmincon` (with default settings) is employed, minimizing the L^2 norm of the residual of the 4th order conditions, starting from thousands of randomly chosen initial points, and selecting the scheme with the smallest objective function. It has generally been observed that this optimization problem is non-convex and not well-conditioned; hence, the scheme (3.6) should not be expected to be optimal. However, it does satisfy all constraints up to machine precision and yields good convergence results for various test problems, as shown in Section 5.

Weak stage order reduces the magnitude of the coefficients $\vec{\alpha}^T\vec{r}_i$ in front of the singular functions ψ_i $(2\le i\le 3)$ in (2.25). This decreases the amplitude of the boundary layers that contribute to the error expansion for ϵ_0. For example, in the scheme (3.6): $A\vec{\tau}^{(2)}=a_{11}\vec{\tau}^{(2)}$ implies $\vec{\tau}^{(2)}$ is a right eigenvector of M (for any Δt). Without loss of generality, setting $\vec{r}_2=\vec{\tau}^{(2)}$, renders the coefficient $\vec{\alpha}^T\vec{r}_2\propto\vec{b}^T A^{-1}\vec{\tau}^{(2)}=0$ so that ψ_2 does not contribute to the error ϵ_0. Furthermore, one can work out that $\vec{\alpha}^T\vec{r}_3\psi_3=\vec{\alpha}^T\vec{r}_4\psi_4=O(\Delta t^3)$ so that the singular modes ψ_3 and ψ_4 contribute one order less to the global error ($\psi_0=0$ due to stiff accuracy). The BL amplitude in the error ϵ_0 is then reduced (but not eliminated) to $O(\Delta t^3)$. One will still observe a further order reduction in the solution derivatives.

4. Modified boundary conditions

This section presents an alternative approach for avoiding order reduction by modifying the prescribed RK b.c.—hereon referred to as *modified boundary conditions* (MBC). The concept of MBC itself is not new (see Subsection 1.3), with the most general formulation given in [3, 5]. The purpose of this section is to show how MBC can be systematically derived by removing the boundary layers in the RK spatial approximation error. The advantage of MBC (over weak stage order) is that they do not restrict the RK scheme that is used; the disadvantage is that they are more complicated to implement. In Subsection 4.1 we derive MBC via a power series expansion; and in Subsection 4.2 we show that they suitably reduce the magnitude of BLs, and also reduce any boundary mismatch for non-stiffly accurate schemes.

4.1. Derivation of MBC via power series expansion. In this subsection we choose the b.c. for (2.6) so that the solution $\vec{u}^{n+1}$ may be expanded in formal powers of Δt (uniformly across the entire domain) up to the order of the RK scheme. This will, effectively, suppress the BLs in $\vec{u}^{n+1}$ up to the scheme's order (but not to all orders) and alleviate order reduction.

In the absence of BLs, the stage vector $\vec{u}^{n+1}$ can be written via a formal power series expansion as: $\vec{u}_p^{n+1}\sim\sum_{j\ge 0}\Delta t^j\vec{U}_j$. Substituting this expansion into (2.6) and collecting equal powers of Δt leads to expressions for $\vec{U}_j$. When $\mathcal{L}$ is linear, the power

series expansion for $\vec{u}_p^{n+1}$ reduces to the Neumann series expansion [47, Chapter 6], and results in a recursive formula for $\vec{U}_j$:

$$\vec{U}_j = A\mathcal{L}\vec{U}_{j-1} \text{ for } j \geq 2 \quad \text{with} \quad \vec{U}_1 = A\mathcal{L}\vec{U}_0 + A\vec{f}^{n+1} \text{ and } \vec{U}_0 = u^n\vec{e}.$$

Thus $\vec{U}_j = (A\mathcal{L})^j u^n\vec{e} + (A\mathcal{L})^{j-1}(A\vec{f}^{n+1})$ for $j \geq 1$. Hence, $\vec{u}_p^{n+1}$ takes the form

$$\vec{u}_p^{n+1} \sim u^n\vec{e} + \sum_{j\geq 1} \Delta t^j \left(A^j \mathcal{L}^j u^n\vec{e} + A^j \mathcal{L}^{j-1} \vec{f}^{n+1} \right). \tag{4.1}$$

We refer to the expansion (4.1) (also known in the matched asymptotic expansions theory [9,16,21,27]) as the *regular solution* to (2.6). If $\vec{f}^{n+1}$ is (infinitely) smooth, the regular solution is a particular solution to (2.6) and has no boundary layers—but does not satisfy homogeneous boundary conditions $\vec{u}_p^{n+1} \neq 0$. Hence, to avoid BL in $\vec{u}^{n+1}$ (to some order), we need the b.c. of (2.6) to match the values of $\vec{u}_p^{n+1}$.

Truncating the series (4.1) up to the scheme's order p, and evaluating at the boundary, yields a set of b.c. that match (4.1) up to order p. The PDE (1.1) can then be used to replace terms involving high powers of $\mathcal{L}$, in terms of the data on the boundary $g(t)$. A technical detail is that, in (4.1), the operator $\mathcal{L}^j$ is not applied to the exact solution, but to the numerical solution which does not satisfy the PDE exactly. Hence we (i) express the numerical solution $u^n = u^*(t_n) + \epsilon_0^n$ in terms of the exact solution u^* and discretization error ϵ_0^n at time t_n; and (ii) use the PDE $u_t^* = \mathcal{L}u^* + f$ to replace $\mathcal{L}^j u^*(t_n)$ by $\partial_t^j u^*(t_n)$ and the forcing f at time t_n. Taylor-expanding $\vec{f}^{n+1}$ at t_n, the truncated expansion yields

$$\begin{aligned}\vec{u}_p^{n+1} = u^n\vec{e} + \sum_{j=1}^{p} \Delta t^j \Bigg[\partial_t^j u^*(t_n) A^{j-1}\vec{c} + \sum_{k=2}^{j-1} \left(\mathcal{L}^{j-k-1}\partial_t^k f^n\right) \\ \times \left(\tfrac{1}{k!} A^{j-k}\vec{c}^k - A^{j-1}\vec{c}\right) \Bigg] + \sum_{j=1}^{p} \Delta t^j A^{j-1}\vec{c}\mathcal{L}^j \epsilon_0^n. \end{aligned} \tag{4.2}$$

In Equation (4.2), ϵ_0^n is the error incurred by the formal expansion, so that by construction it is assumed to be $O(\Delta t^p)$. The MBC are then obtained by neglecting the error term ϵ_0^n, and evaluating the truncated series (4.2) at the boundary:

$$\vec{g}_{\text{MBC}} := g^n\vec{e} + \sum_{j=1}^{p} \Delta t^j \left[\partial_t^j g^n A^{j-1}\vec{c} + \sum_{k=2}^{j-1} \left(\mathcal{L}^{j-k-1}\partial_t^k f^n\right) \left(\frac{A^{j-k}\vec{c}^k}{k!} - A^{j-1}\vec{c} \right) \right]. \tag{4.3}$$

Here u^n at the boundary was set to g^n. By construction, $\vec{g}_{\text{MBC}}$ matches (4.1) up to the scheme's order p, and incorporates b.c. information. The MBC are unique up to order p, i.e., any other b.c. that suppress the singular behavior up to the same order, can differ from the MBC only by $O(\Delta t^{p+1})$ terms. As an example, the 3rd order MBC (MBC3), i.e., $p = 3$, take the following form:

$$\vec{g}_{\text{MBC}} = g^n\vec{e} + \Delta t\partial_t g^n\vec{c} + \Delta t^2\partial_t^2 g^n A\vec{c} + \Delta t^3 \left(\partial_t^3 g^n A^2\vec{c} + \left(\partial_t^2 f^n\right) \left(\tfrac{A\vec{c}^2}{2} - A^2\vec{c} \right) \right).$$

This derivation shows that any b.c. one prescribes that agrees with $\vec{g}_{\text{MBC}}$ up to the order of the method, will remove order reduction—for instance those obtained by [3].

REMARK 4.1. In two important special cases, the MBC (4.3) simplify. First, when the boundary data g and the forcing f at the boundary are time-independent, the summation in (4.3) vanishes. This reflects the fact that order reduction does not arise for autonomous problems. Second, the MBC $\vec{g}_{\text{MBC}}$ can also be written as the conventional b.c., modified by a sum involving only the stage order residuals $\vec{\tau}^{(j)}$. Thus, when the RK scheme's stage order satisfies $q \geq p$, all terms involving the stage order residuals vanish, implying that the MBC $\vec{g}_{\text{MBC}}$ agree with the conventional b.c. up to the scheme's order p.

4.2. Boundary value mismatch. Although the MBC (4.3) can be used to avoid order reduction, they may still result in a small mismatch of the numerical solution at the boundary with the exact prescribed boundary data, i.e. $u^{n+1} \neq g^{n+1}$, even for stiffly accurate schemes (see Remark 2.2). Enforcing $u^{n+1} = g^{n+1}$, however, may be of practical interest. In this subsection, we first show that the MBC yields a boundary mismatch error, $\epsilon_0^{n+1} = u^{n+1} - g^{n+1}$, that is always of the scheme's order (which is good). Moreover, we provide a recipe to further modify $\vec{g}_{\text{MBC}}$ to ensure that $u^{n+1} = g^{n+1}$ while still avoiding order reduction (which is even better). Note that the MBC in [3] reduces the boundary mismatch to the scheme's order, however, they did not investigate enforcing the exact boundary conditions.

4.2.1. Quantification of the boundary error in the MBCs. To obtain an expression for the boundary error, we use (2.6) to rewrite the update (2.7) in terms of $\vec{u}^{n+1}$ in lieu of $\mathcal{L}\vec{u}^{n+1}$:

$$u^{n+1} = u^n + \vec{b}^T A^{-1}(\vec{u}^{n+1} - u^n \vec{e}). \tag{4.4}$$

Evaluating (4.4) at the boundary and subtracting the true boundary value yields an expression for the error at the boundary

$$\epsilon_0^{n+1} = (1 - \vec{b}^T A^{-1}\vec{e})\epsilon_0^n + \vec{b}^T A^{-1}(\vec{u}^{n+1} - g^n \vec{e}) + g^n - g^{n+1}. \tag{4.5}$$

We may quantify the boundary mismatch ϵ_0^{n+1} introduced by the MBC, by (i) substituting the MBC $\vec{u}^{n+1} = \vec{g}_{\text{MBC}}$ from Equation (4.3) into (4.5), and (ii) Taylor-expanding g^{n+1} at t_n to obtain

$$\begin{aligned}\epsilon_0^{n+1} = &(1 - \vec{b}^T A^{-1}\vec{e})\epsilon_0^n + \sum_{j=1}^{p} \Delta t^j \left[\partial_t^j g^n \left(\vec{b}^T A^{j-2}\vec{c} - \frac{1}{j!} \right) \right. \\ &\left. + \sum_{k=2}^{j-1} \mathcal{L}^{j-k-1} \partial_t^k f^n \left(\frac{1}{k!} \vec{b}^T A^{j-k-1}\vec{c}^k - \vec{b}^T A^{j-2}\vec{c} \right) \right] + O(\Delta t^{p+1}).\end{aligned}$$

In the above expression for ϵ_0^{n+1}, the first p terms in the summation vanish due to the order conditions (2.3). Hence $\epsilon_0^{n+1} = O(\epsilon_0^n) + O(\Delta t^{p+1})$, which implies that the global error at the boundary is at most $O(\Delta t^p)$. In order words: the MBC generally introduce an error in u^{n+1} at the boundary, but order reduction is avoided.

4.2.2. Eliminating boundary mismatch. Equation (4.5) can be used to further modify $\vec{g}_{\text{MBC}}$ to ensure that the numerical solution satisfies the true b.c. at every time step. If $\epsilon_0^n = 0$ for all n, then the numerical solution $\vec{u}^{n+1}$ at the boundary satisfies

$$\vec{b}^T A^{-1}\vec{u}^{n+1} = (\vec{b}^T A^{-1}\vec{e} - 1)g^n + g^{n+1}. \tag{4.6}$$

Conversely, if Equation (4.6) holds and the initial data satisfy the true b.c., i.e., $\epsilon_0^0 = 0$, then $\epsilon_0^n = 0$ for all $n > 0$. Equation (4.6) defines one linear constraint on the values of $\vec{u}^{n+1}$ at the boundary. Hence to ensure that $\epsilon_0^{n+1} = 0$, one only needs to modify the component of $\vec{g}_{\text{MBC}}$ in the direction of $\vec{b}^T A^{-1}$ to satisfy the constraint (4.6), while keeping components orthogonal to $\vec{b}^T A^{-1}$ unchanged. This leads to a new set of MBC

$$\vec{g}^*_{\text{MBC}} = \vec{g}_{\text{MBC}} - \left[\vec{b}^T A^{-1}(\vec{g}_{\text{MBC}} - g^n\vec{e}) + g^n - g^{n+1}\right] \frac{(\vec{b}^T A^{-1})^T}{\|\vec{b}^T A^{-1}\|^2}. \tag{4.7}$$

By construction, the b.c. (4.7) $\vec{u}^{n+1} = \vec{g}^*_{\text{MBC}}$ satisfy the linear constraint (4.6), and hence ensure that $\epsilon_0^n = 0$ for every n (provided that $\epsilon_0^0 = 0$). In addition, the modification $\vec{b}^T A^{-1}(\vec{g}_{\text{MBC}} - g^n\vec{e}) + g^n - g^{n+1}$ in formula (4.7) is only an $O(\Delta t^{p+1})$ correction to $\vec{g}_{\text{MBC}}$. Hence, the boundary conditions $\vec{g}^*_{\text{MBC}}$ still suppress the singular behavior in the numerical solution, to order p, and thereby avoids order reduction. For stiffly accurate RK schemes (i.e., $\vec{b}^T A^{-1} = (0,\ldots,0,1)$), $\vec{g}^*_{\text{MBC},i} = \vec{g}_{\text{MBC},i}$ for stages $1 \le i \le s-1$, while $\vec{g}^*_{\text{MBC},s} = g^{n+1}$.

4.3. Limitations of MBC. The MBC formulas derived above hold for linear problems, where the power series solution for $\vec{u}_p^{n+1}$ matches a Neumann series expansion. A key part of the derivation is to use the PDE to express the MBC in terms of $\partial_t^j g(t_n)$ and $\mathcal{L}^i \partial_t^j f(t_n)$, which are computable from the data g and f. Consequently, MBC are challenging to apply when the data g and f are given in a way that their derivatives are difficult to obtain/compute.

Another fundamental difficulty arises when $\mathcal{L}$ is nonlinear. In this case, the power series expansion of the solution to (2.6) involves, in general, terms that are not directly computable from the known data g and f. Such a limitation may seriously hinder the practical use of MBC for nonlinear problems. For example, consider the viscous Burgers' equation (see Subsection 5.4) where $\mathcal{L}u$ is replaced by the nonlinear operator $\mathcal{N}u = \nu u_{xx} - uu_x$. When evaluating the truncated power series expansion at the boundary, the terms up to 2nd order in Δt *can* be expressed in terms of $g(t_n)$, $\partial_t g(t_n)$ and $\partial_{tt} g(t_n)$. However, the 3rd order term contains the boundary evaluation of $(\partial_t u^*(t_n))\partial_x(\partial_t u^*(t_n))$, which requires knowledge of spatial derivatives of the exact solution.

5. Numerical examples

In this section we illustrate the order reduction phenomenon, and the two remedies developed above (weak stage order (Section 3) and modified b.c. (Section 4)), in several numerical examples: heat Equation (5.1), Schrödinger (Subsection 5.2), advection-diffusion (Subsection 5.3), viscous Burgers' (Subsection 5.4), linear advection and Airy's Equation (5.5). The method of manufactured solutions [43] is used to construct a solution in each case. The spatial approximation is conducted via fourth-order centered differences on a fixed grid with 10000 cells. This renders spatial approximation errors negligible, thus isolating temporal discretization errors (measured in the maximum norm, unless noted otherwise), in line with the analysis in Section 2.

We focus on stiffly accurate DIRK schemes here due to their practical interest. In each example, we first demonstrate OR incurred with the standard 3-stage, 3rd-order DIRK scheme with weak stage order (WSO) 1 (A.3). Then we show how MBC, applied to the same RK scheme, recover the full order of convergence. Finally, we show that the new 4-stage, 3rd-order DIRK scheme with WSO 2 (Subsection 3.3) recovers the full convergence order (for function values) as well. The examples also highlight some limitations of the new approaches, such as arising in high-order MBC for nonlinear problems or the recovery of the full order in derivatives.

5.1. Heat equation. This test case has been considered already in Subsection 1.1 to introduce the order reduction phenomenon. Here we show that: (i) MBC recover full convergence order for function values as well as some derivatives; (ii) the full order for derivatives is only recovered when the MBC are carried out to the appropriate order; and (iii) DIRK schemes with high WSO recover the full order for function values, but generally not for derivatives. Consider the 1D heat equation

$$u_t = u_{xx} + f \quad \text{for} \quad (x,t) \in (0,1) \times (0,1], \tag{5.1}$$

with solution $u^*(x,t) = \cos(15t)\sin(5x+5)$, and forcing f, Dirichlet b.c. $u = g$ on $\{0,1\} \times (0,1]$, and initial condition $u = u_0$ chosen accordingly.

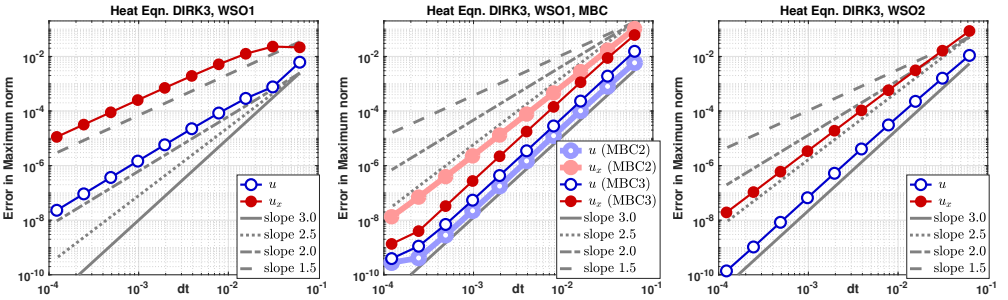

FIG. 5.1. *Error convergence (u blue; u_x red) for 3rd order schemes for the heat Equation (5.1). WSO 1 scheme with conventional b.c. (left); same scheme with MBC (middle); WSO 2 scheme (right).*

Figure 5.1 shows the convergence orders of u and u_x for the WSO 1 scheme with conventional b.c. (left), for the same scheme with MBC (middle), and for the WSO 2 scheme with conventional b.c. (right). In the middle panel, MBC are carried out up to the 2nd (MBC2) and 3rd (MBC3) order terms, respectively. Order reduction renders $u = O(\Delta t^2)$, $u_x = O(\Delta t^{1.5})$, and $u_{xx} = O(\Delta t^1)$ for the WSO 1 scheme. For the same scheme, the full MBC3 recover $u = O(\Delta t^3)$, $u_x = O(\Delta t^3)$, and $u_{xx} = O(\Delta t^3)$, while the MBC2 recover $u = O(\Delta t^3)$, but yield reduced orders in $u_x = O(\Delta t^{2.5})$, and $u_{xx} = O(\Delta t^2)$. The same orders are obtained with the WSO 2 scheme (with conventional b.c.). Note that the errors in u_{xx} are not displayed in the figure, but the convergence orders are equally clear as for u and u_x.

5.2. Schrödinger equation. As an example of a dispersive problem (which fails assumption (2.5a)), we consider the Schrödinger equation

$$u_t = \frac{\imath\omega}{k^2} u_{xx} \quad \text{for} \quad (x,t) \in (0,1) \times (0,1.2], \tag{5.2}$$

with $k = 5$ and $\omega = 2\pi$, solution $u^*(x,t) = \exp(\imath(kx - \omega t))$, Dirichlet b.c. $u = g$ on $\{0,1\} \times (0,1.2]$, and initial condition $u = u_0$ chosen accordingly.

Figure 5.2 shows that in addition to a time-periodic error with BLs, the RK scheme produces transient dispersive waves in the error far from the domain boundaries. And even more: these dispersive waves may produce order-reduction-like effects in the interior of the domain. The total RK error can be understood as a superposition of the time-periodic error (having $O(\Delta t^2)$ BLs and $O(\Delta t^3)$ error away from the BLs outlined in Section 2), and a transient dispersive wave that solves the RK scheme applied to the homogeneous Equation (5.2) (i.e., $f = 0$ and $g = 0$).

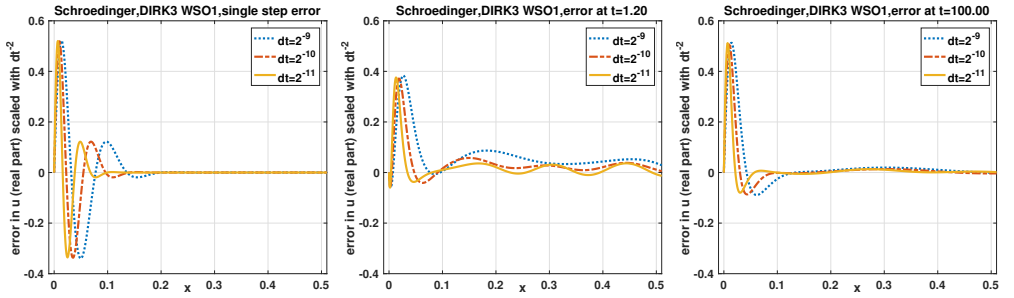

FIG. 5.2. *Errors (real part, scaled with $1/\Delta t^2$) as functions of x for the Schrödinger Equation (5.2), solved with a WSO 1 scheme with conventional b.c., after a single step (left), at a transient time (middle), and at a large time (right). Shown is the left half of the domain (the right half looks similar). The transient error component away from the BLs is clearly visible.*

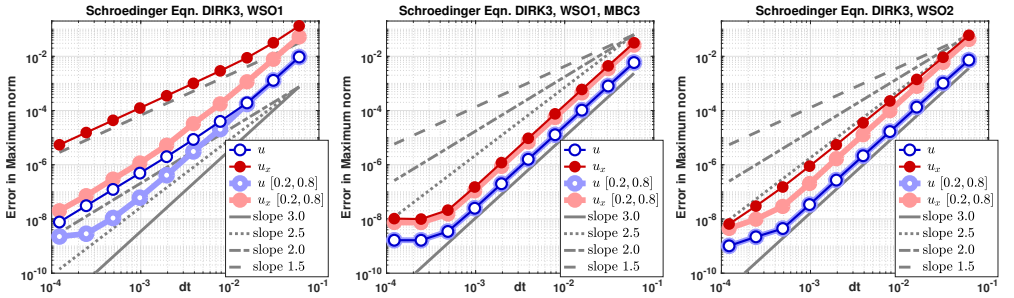

FIG. 5.3. *Error convergence (u blue; u_x red) for 3rd order schemes for the Schrödinger Equation (5.2). WSO 1 scheme with conventional b.c. (left); same scheme with MBC (middle); WSO 2 scheme (right). The error convergence measured away from the BLs is shown for u (light blue) and u_x (light red).*

Figure 5.2 shows the shape (in x, only half of the domain is shown) of the error (here re-scaled with $1/\Delta t^2$) at three different times: after one step (left panel), at a transient time (middle), and after a long time (right). Except for the $O(\Delta t)$ time, the BL is dominated by the time-periodic component, while the domain's interior is dominated by the transient component. One can clearly see that the transient component decays slowly in time (because the RK scheme is asymptotically stable for any imaginary eigenvalue). However, for transient times (middle panel), it yields a noticeable contribution to the error away from the BLs. The plots indicate that the transient component (a) scales (roughly) like $O(\Delta t^{2.5})$ in amplitude, and (b) has an $O(\Delta t^{0.5})$ wave length. This observed scaling occurs because the transient component has an i.c. with BLs of width $O(\Delta t^{0.5})$, and thus its dominant Fourier modes occur at wave numbers $O(\Delta t^{-0.5})$ and with magnitude $O(\Delta t^{0.5})$.

Figure 5.3 shows the error convergence results. When errors are considered over the full spatial domain (i.e., including the BLs), precisely the same results as for the heat equation (see Figure 5.1) are obtained. In addition, we consider errors evaluated away from the BLs (light colors). As expected from the results above, these exhibit a more interesting behavior. Without any remedies to order reduction, we observe (roughly) an error scaling of $u = O(\Delta t^{2.5})$, $u_x = O(\Delta t^2)$, and $u_{xx} = O(\Delta t^{1.5})$, thus indicating order reduction effects away from the BLs. In addition, MBC and high WSO remove the order

reduction, not only in the BLs, but also inside the domain. It should be re-iterated that the transient effects vanish after sufficiently long times (which, for most RK schemes, are $O(\Delta t)$, with a very large constant).

The observations collected here highlight that order reduction effects need not necessarily be limited to thin zones (i.e., BLs) near the domain boundaries, but can propagate into the interior of the domain, if for instance the PDE is a dispersive wave equation.

5.3. Advection-diffusion equation. As revealed by the analysis in Section 2, order reduction for IBVPs is intricately linked to boundary layers (BLs) produced by the time-stepping. A natural question is therefore: what happens in problems that possess a physical BL? To answer this question, we consider the linear advection-diffusion equation

$$u_t = \nu u_{xx} - u_x + f \quad \text{for} \quad (x,t) \in (0,1) \times (0,1.2], \tag{5.3}$$

with manufactured solution $u^*(x,t) = \sin(2\pi(x-t))$, and the nondimensional viscosity $\nu = 10^{-3}$. When ν is small, the equation becomes advection-dominated, and prescribing Dirichlet b.c. at the outflow boundary $x = 1$ results in a BL of width $O(\nu)$ in the error (note that even though our u^* does not have a BL, the error does).

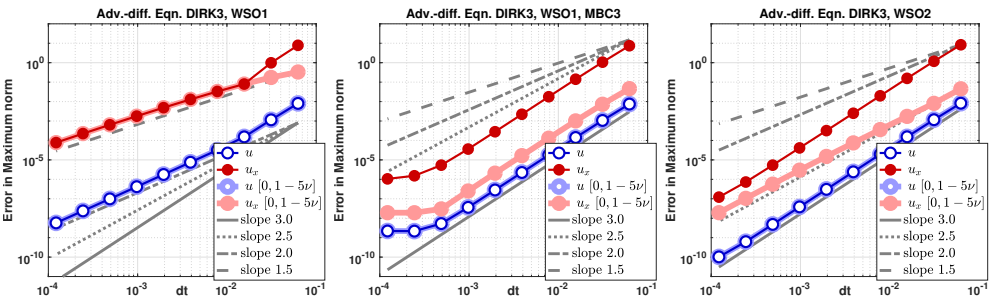

FIG. 5.4. *Error convergence (u blue; u_x red) for 3rd order schemes for the advection-diffusion Equation (5.3). WSO 1 scheme with conventional b.c. (left); same scheme with MBC (middle); WSO 2 scheme (right). The error measured away from the outflow boundary ($[0,1-5\nu]$) is shown for u (light blue) and u_x (light red).*

Figure 5.4 shows the convergence results. In addition to measuring errors (in the maximum norm) over the whole domain, we also study the error over the domain $x \in [0,1-5\nu]$, i.e., away from the BL (light colors). The results show L-shaped transitions in error behavior, depending on which types of BLs dominate. The ν-BL is of magnitude $O(\Delta t^3)$ and width $O(\nu)$, thus its effect on the u_x error is $O(\Delta t^3/\nu)$. In turn, the order reduction BLs (for the WSO 1 scheme) are of magnitude $O(\Delta t^2)$ and of width $O(\Delta t^{0.5})$, thus affecting the u_x error with $O(\Delta t^{1.5})$. Balancing these expressions explains why the kink in the u_x error (left panel) occurs at $\Delta t = O(\nu^{2/3})$. Likewise, the same error experiences a kink at $\Delta t = O(\nu^2)$ for the WSO 2 scheme (right panel).

Thus, for large Δt values the ν-BL dominates the error, and the scheme appears to not exhibit order reduction. However, for Δt sufficiently small, the error behaves the same way it does for the heat Equation (5.1), and order reduction becomes apparent. In line with our theory, the L-shapes in the errors are removed when MBC3 (middle panel) are applied. Those recover the full 3rd order convergence throughout the full range of Δt values. The WSO 2 scheme (right panel) recovers a clean third order for

u. However, it still loses an order in u_x, even though this becomes visible only for very small Δt values.

5.4. Viscous Burgers' equation. With this example we demonstrate that order reduction, as well as some remedies, also apply in nonlinear problems. We consider the viscous Burgers' equation

$$u_t + u u_x = \nu u_{xx} + f \quad \text{for} \quad (x,t) \in (0,1) \times (0,1], \tag{5.4}$$

with Dirichlet b.c. and $\nu = 0.1$. The manufactured solution is $u^*(x,t) = \cos(2 + 10t)\sin(0.2 + 20x)$. The nonlinearity yields a nonlinear implicit equation at each stage, with $\mathcal{N}u = \nu u_{xx} - u u_x$, which is solved via standard Newton iteration.

A crucial limitation is that the third order term in the MBC, obtained from the expansion in Section 4, contains terms that are not accessible without knowledge of the exact solution (see Subsection 4.3). Hence, MBC3 cannot be formulated via the procedure introduced in Section 4. However, MBC2 *can* be formulated in terms of the data, and they coincide with the corresponding expression obtained for linear problems.

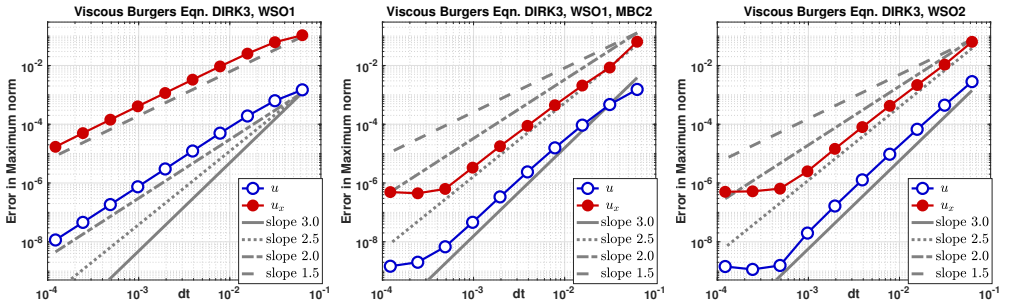

FIG. 5.5. *Error convergence (u blue; u_x red) for 3rd order schemes for the viscous Burgers' Equation (5.4). WSO 1 scheme with conventional b.c. (left); same scheme with MBC2 (middle); WSO 2 scheme (right).*

Figure 5.5 shows that the same type of order reduction arises as for linear problems (left panel). The MBC2 recover the full 3rd order for u, but lose orders for derivatives (middle). The same results are obtained with the WSO 2 scheme (right).

5.5. Linear advection equation and Airy's equation. All examples above have $\mathcal{L}$ a second-order differential operator. To demonstrate that the order reduction results, as well as the remedies, apply more generally, we consider the following two problems:

(1) The linear advection equation: $u_t = u_x$ for $(x,t) \in (0,1) \times (0,1.2]$ with Dirichlet b.c. at $x = 0$ and manufactured solution $u^*(x,t) = \sin(2\pi(x-t))$.

(2) Airy's equation: $u_t = u_{xxx} + f$ for $(x,t) \in (0,1) \times (0,1]$ with b.c. $u(0) = g(t)$, $u_x(0) = h_0(t)$, $u_x(1) = h_1(t)$, and manufactured solution $u^*(x,t) = \cos(15t)$.

The respective results are shown in Figures 5.6 and 5.7. In line with the theory in Section 2, conventional b.c. render function values one order more accurate than the scheme's WSO, and $1/m$ orders per derivative are lost, where m is the order of $\mathcal{L}$. MBC3 recover the full 3rd order for u, as well as derivatives up order m. Hence, one obtains $u = O(\Delta t^3)$, $u_x = O(\Delta t^3)$, and $u_{xx} = O(\Delta t^2)$ for $m = 1$, and $u = O(\Delta t^3)$, $u_x = O(\Delta t^3)$,

and $u_{xx} = O(\Delta t^3)$ for $m = 3$. In contrast, the WSO 2 scheme recovers the full order in u only. Hence $u = O(\Delta t^3)$, $u_x = O(\Delta t^2)$, and $u_{xx} = O(\Delta t^1)$ for $m = 1$, and $u = O(\Delta t^3)$, $u_x = O(\Delta t^{2.67})$, and $u_{xx} = O(\Delta t^{2.33})$ for $m = 3$.

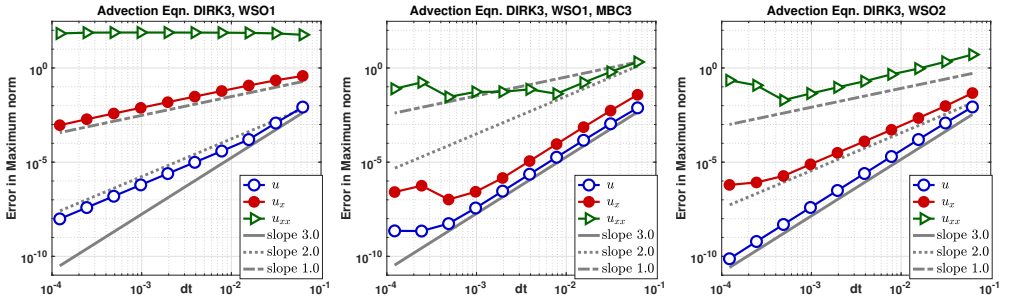

FIG. 5.6. *Error convergence (u blue; u_x red; u_{xx} green) for 3rd order schemes for the linear advection equation. WSO 1 scheme with conventional b.c. (left); same scheme with MBC (middle); WSO 2 scheme (right).*

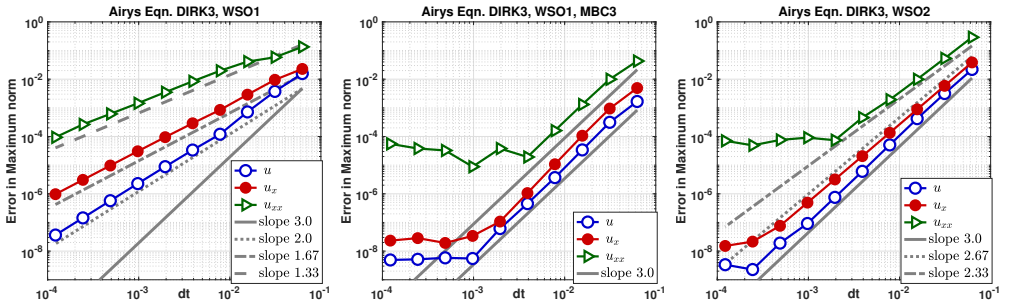

FIG. 5.7. *Error convergence (u blue; u_x red; u_{xx} green) for 3rd order schemes for the Airy's equation. WSO 1 scheme with conventional b.c. (left); same scheme with MBC (middle); WSO 2 scheme (right).*

6. Conclusions and outlook

We have demonstrated that order reduction is a generic phenomenon in implicit RK time-stepping for IBVPs (with time-dependent data) that manifests in the formation of spatial boundary layers in the numerical solution. These originate because the stage update equations are singularly perturbed boundary value problems, where most types of b.c. generate a mismatch between the boundary and the interior. The global-in-time behavior of these boundary layers has been studied via modal and asymptotic analysis; and in that light, two different approaches to overcome order reduction have been examined: (a) new conditions on the RK coefficients (weak stage order); and (b) modified b.c. that render the boundary data mismatch as small as the order of the RK scheme. Modified b.c. have the advantage that they work for a wide range of RK schemes; however, they may be complicated to compute, and difficult to implement for non-linear problems. In turn, schemes with weak stage order require no modification of the spatial approximation; however, these new conditions rule out many existing RK methods.

Because order reduction is caused by boundary layers, one could be inclined to think that it is only a minor concern, because its effects can be "felt" only near the boundary. That is not the case, because:

(i) In many applications, the solution (and particularly its derivatives) at the boundary are important, such as stresses at boundaries (lift and drag in CFD).

(ii) In many multi-physics problems, information near the boundaries feeds back to the interior of the domain (non-local terms, fast waves, etc.).

(iii) The analysis in Section 2 only studies the $t\to\infty$ behavior of the error. Many problems (e.g., the Schrödinger equation, see Subsection 5.2) exhibit transient features that reduce the observed order away from the boundary at $O(1)$ times.

(iv) The boundary layers' thickness scales like some power of the time step: $\Delta t^{1/m}$, where m is the order of the space operator. Hence, unless the time step is very small, their thickness may be considerable.

In this context (particularly point (i)), it must be stressed that the presence of boundary layers that multistage methods almost always develop generally implies order reduction in (sufficiently high) spatial derivatives, even if the numerical solution itself is devoid of order reduction (due to weak stage order or modified b.c.).

Although the analysis in this paper focuses on problems with Dirichlet b.c., the order reduction phenomenon also arises with other types of b.c. (such as Neumann), and the analysis in Subsection 2.5 carries over with minor adaptations. With Neumann b.c., the obtained convergence orders are slightly different. For instance, for the heat equation, one obtains $O(\Delta t^{\tilde{q}+1.5})$, where $\tilde{q}$ is the weak stage order, for the error in function value (i.e., half an order better than with Dirichlet conditions), and half an order loss per derivative [32].

The analysis in Section 2 also applies to RK schemes with a singular coefficient matrix A, such as Crank-Nicolson (CN) and EDIRK schemes [28]. The CN scheme is an example of a second-order scheme devoid of boundary layers, because the matrix M has no small eigenvalues (one is $O(1)$ and the other is zero). However, CN is not L-stable, and it incurs the same problem as the implicit midpoint rule (see Subsection 2.4), namely the growth factor approaches -1 as $\zeta\to-\infty$.

It should be stressed that order reduction arises in explicit RK schemes as well [15, 45]. However, the semi-discrete analysis in this paper does not directly apply.

Finally, the weak stage order and the modal analysis presented in this paper apply beyond IBVPs and RK schemes. In Subsection 6.1 we briefly outline the role of weak stage order in avoiding order reduction in stiff ODEs, and in Subsection 6.2 we employ the modal analysis to show that linear multistep methods (LMMs) do *not* exhibit order reduction.

6.1. Order reduction and weak stage order for stiff linear ODEs. The concept of weak stage order, introduced in Section 3, has been studied in terms of its impact on boundary layers in IBVPs. Because the concept is also of interest to stiff ODEs (without a spatial interpretation), we examine the model ODE proposed by Prothero and Robinson [41]:

$$y' = \lambda(y-\phi(t))+\phi'(t) \tag{6.1}$$

with i.c. $y(0)=\phi_0$ and $\mathrm{Re}\,\lambda<0$. The exact solution $y(t)=\phi(t)$ is assumed analytic.

When a RK scheme (with coefficients A, $\vec{b}^T$, and $\vec{c}$) is applied to (6.1), the error

ϵ_0^{n+1}, at time t_{n+1}, can be computed (see [51, Chapter IV.15]) to be

$$\epsilon_0^{n+1} = R(\zeta)\epsilon_0^n + \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\delta}^{n+1} + \delta_0^{n+1}, \tag{6.2}$$

where $\zeta = \lambda \Delta t$. The vector $\vec{\delta}^{n+1}$ and scalar δ_0^{n+1} denote the truncation errors incurred at the intermediate stages, and at the end of the time step, respectively. Written in terms of derivatives of ϕ, and the stage order residuals $\vec{\tau}^{(j)}$, they read as

$$\vec{\delta}^{n+1} = \sum_{j \geq q+1} \frac{\Delta t^j}{(j-1)!} \vec{\tau}^{(j)} \phi^{(j)}(t_n), \; \delta_0^{n+1} = \sum_{j \geq p+1} \frac{\Delta t^j}{(j-1)!} \left(\vec{b}^T \vec{c}^{\,j-1} - \frac{1}{j} \right) \phi^{(j)}(t_n).$$

Using these expressions, we obtain the RK scheme's stiff ODE order as follows:

PROPOSITION 6.1. *Suppose the Runge-Kutta scheme $(A, \vec{b}, \vec{c})$, with A invertible, has weak stage order $\tilde{q}$. Then in the stiff limit $\zeta \ll -1$, the local truncation error ϵ_0^1 for Equation (6.1), which is obtained by setting $\epsilon_0^0 = 0$, is of order* $\min\{\tilde{q}+1, p+1\}$.

Proof. Setting $\epsilon_0^0 = 0$ in (6.2), and substituting the formula for $\vec{\delta}^{n+1}$, we have

$$\epsilon_0^1 = \sum_{j \geq q+1} \frac{\Delta t^j}{(j-1)!} \phi^{(j)}(t_0) \left(\zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\tau}^{(j)} \right) + \delta_0^1. \tag{6.3}$$

Weak stage order $\tilde{q}$ means that the stage order residuals $\vec{\tau}^{(j)}$ for $1 \leq j \leq \tilde{q}$ lie in an A-invariant space $\mathcal{V}$ that is orthogonal to $\vec{b}$. Thus $\mathcal{V}$ is also $(I - \zeta A)^{-1}$-invariant, and $(I - \zeta A)^{-1} \vec{\tau}^{(j)} \in \mathcal{V}$, hence

$$\vec{b}^T (I - \zeta A)^{-1} \vec{\tau}^{(j)} = 0, \quad \text{for } 1 \leq j \leq \tilde{q}.$$

As a result, the first $j \leq \tilde{q}$ terms in (6.3) vanish, resulting in the sum to be over $j \geq \tilde{q}+1$. In the stiff ODE limit, i.e., $\Delta t \ll 1$ and $\zeta \ll -1$, the term $\zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\tau}^{(j)}$ can be bounded in terms of $\|A^{-1}\|$, $\|\vec{\tau}^{(j)}\|$, and an $O(1)$ constant. Hence the expression for ϵ_0^1 in (6.3) is $O(\Delta t^{\tilde{q}+1})$, while $\delta^1 = O(\Delta t^{p+1})$, which together yields $\epsilon_0^1 = O(\Delta t^{\min\{\tilde{q}+1, p+1\}})$. ∎

REMARK 6.1. By construction, weak stage order satisfies $\tilde{q} \geq q$. Hence Proposition 6.1 improves the stiff error bound given by the stage order q.

REMARK 6.2. For stiff ODEs, the global truncation error is of order $\tilde{q}$. This is a difference to PDE IBVPs, for which the global error is of order $\tilde{q}+1$. Hence, to avoid OR for stiff ODEs, the RK scheme must have $\tilde{q} = p$. In contrast, for IBVPs the choice $\tilde{q} = p - 1$ suffices.

6.2. Order reduction in non-Runge-Kutta time-stepping methods. This paper shows that order reduction for IBVPs, due to boundary layers in the spatial error, arises rather generically in Runge-Kutta methods. It also occurs in other multistage schemes, or any scheme that can be recast as a multistage scheme, see Subsection 1.3. In contrast, schemes that achieve high order via a single BVP solve per step are devoid of the order reduction mechanism. This includes linear multistep methods (LMMs) (see [29] and [20, Chapter IV.15] for LMM error estimates), for example backward differentiation formula (BDF) methods.

To illustrate that LMMs do not result in a singular perturbation problem, we use the framework introduced in Section 2. A general s-step, implicit, LMM for solving the

IBVP (1.1) with Dirichlet b.c. takes the form

$$u^{n+s} = \sum_{j=0}^{s-1} \alpha_j u^{n+j} + \Delta t \sum_{j=0}^{s} \beta_j (\mathcal{L} u^{n+j} + f^{n+j}) \quad \text{with b.c. } u^{n+s} = g(t_n + s\Delta t), \tag{6.4}$$

where $\beta_s \neq 0$. The scheme (6.4) defines a linear recursion relation for u^{n+s}, and can be written using matrix notation on the vector solution $(u^{n+s},\dots,u^{n+1})^T$, see [19]. To characterize the error, denote $\epsilon^n = u^n - u^*(t_n)$, and let $\vec{\epsilon}^{\,n+1} = (\epsilon^{n+s},\dots,\epsilon^{n+1})^T$. One can then obtain an equation for the error vector $\vec{\epsilon}^{\,n+1}$, similar to (2.8), by considering time-periodic solutions $\vec{\epsilon}^{\,n} = z^n \vec{\epsilon}(x)$ and $\vec{\delta}^{\,n} = z^n \vec{\delta}(x)$:

$$\vec{\epsilon} - M\mathcal{L}\vec{\epsilon} = N\vec{\delta} \quad \text{with homogeneous b.c. for } \vec{\epsilon}. \tag{6.5}$$

Here $M = \frac{\Delta t}{z-1} AB + \Delta t B$ and $N = E + \frac{1}{z-1} AE$ with

$$A = \begin{pmatrix} \alpha_{s-1} & \alpha_{s-2} & \cdots & \alpha_0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \beta_s & \cdots & \beta_2 & \beta_1 + \frac{1}{z}\beta_0 \\ 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

$E = \vec{e}_1 \vec{e}_1^{\,T}$ and $\vec{e}_1 = (1,0,\dots,0)^T$. Note that M has exactly $s-1$ zero eigenvalues with right eigenvectors corresponding to the $s-1$ dimensional null space of B. Moreover, $\frac{\Delta t}{z-1} AB$ is a rank-1 matrix with one $O(1)$ eigenvalue. As a result, M has exactly one $O(1)$ eigenvalue and $s-1$ zero eigenvalues. Therefore, Equation (6.5) does not incur a singularly perturbed BVP, and hence there are no BLs in the global solution $\vec{\epsilon}(x)$. Another way to interpret these results is as follows. Every time-stepping scheme has one $O(1)$ eigenvalue in the error propagation matrix M, due to consistency. Time-stepping schemes that require only a single solve per time-step are devoid of order reduction.

We conclude by mentioning general linear methods (GLMs) [24, Chapter 2], which are schemes with multiple steps and multiple stages. Although more complex, GLMs may inherit many of the good properties of Runge-Kutta and multistep methods, but also some of their drawbacks. Specifically, having multiple stages triggers the mechanism for boundary layers. Conversely, the added flexibility of GLMs allows for the construction of diagonally implicit schemes with desirable stability properties and high stage order [13, 52].

Acknowledgments. The authors would like to thank David Ketcheson and Adrian Sandu for helpful conversations and suggestions. This material is based upon work supported by the National Science Foundation under Grants DMS–1719637 (Rosales), DMS–1719640 (Zhou), DMS–2012271 and DMS–2309728 (both Seibold), DMS–2012268 and DMS–2309727) (both Shirokoff). D. Shirokoff was supported by a grant from the Simons Foundation (#359610).

Appendix A. Implicit Runge-Kutta schemes used in this paper. All the DIRK schemes listed here are from [51, Chapter IV.6]. Let s be the number of stages, and p, q, and $\tilde{q}$ denote the order of the scheme, stage order, and WSO, respectively.
Stiffly accurate DIRK with $s=2$, $p=2$, $q=1$, $\tilde{q}=1$ (DIRK2):

$$\begin{array}{c|cc} \gamma & \gamma & \\ 1 & 1-\gamma & \gamma \\ \hline & 1-\gamma & \gamma \end{array} \quad \text{for} \quad \gamma = 1 - \frac{\sqrt{2}}{2} \tag{A.1}$$

Non stiffly accurate DIRK with $s=2$, $p=3$, $q=1$, $\tilde{q}=1$:

$$\begin{array}{c|cc} \gamma & \gamma & \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \text{for} \quad \gamma = \frac{3+\sqrt{3}}{6} \tag{A.2}$$

Stiffly accurate DIRK with $s=3$, $p=3$, $q=1$, $\tilde{q}=1$ (DIRK3):

$$\begin{array}{c|ccc} 0.4358665215 & 0.4358665215 & & \\ 0.7179332608 & 0.2820667392 & 0.4358665215 & \\ 1 & 1.208496649 & -0.644363171 & 0.4358665215 \\ \hline & 1.208496649 & -0.644363171 & 0.4358665215 \end{array} \tag{A.3}$$

Stiffly accurate DIRK with $s=5$, $p=4$, $q=1$, $\tilde{q}=1$ (DIRK4):

$$\begin{array}{c|ccccc} 1/4 & 1/4 & & & & \\ 3/4 & 1/2 & 1/4 & & & \\ 11/20 & 17/50 & -1/25 & 1/4 & & \\ 1/2 & 371/1360 & -137/2720 & 15/544 & 1/4 & \\ 1 & 25/24 & -49/48 & 125/16 & -85/12 & 1/4 \\ \hline & 25/24 & -49/48 & 125/16 & -85/12 & 1/4 \end{array} \tag{A.4}$$

Appendix B. Order reduction and periodic solutions. Here we outline a class of problems where OR originates through periodically forced solutions. Assume the following applies to (1.1) and (2.1–2.2):

(i) *The operator $\mathcal{L}$, with homogeneous boundary conditions, has a complete set of normal mode eigenfunctions, and corresponding eigenvalues, $\{\lambda_\ell, \varphi_\ell(x)\}_{\ell=1}^{\infty}$.*

(ii) *The eigenvalues, as well as the corresponding scheme growth factors $R_\ell = R(\lambda_\ell \Delta t)$, satisfy $|e^{\lambda_\ell \Delta t}| \le R_*$ and $|R_\ell| \le R_*$, where $R_* \le 1$ is a constant.*[4]

(iii) *The scheme's growth factor satisfies $D_g(\zeta) = |\frac{e^\zeta - R(\zeta)}{\zeta^{p+1}}| \le B_*$ for $Re(\zeta) \le 0$, where p is the scheme's order and B_* is a constant.*

(iv) *The initial conditions $u_{ic} = \sum \alpha_\ell \varphi_\ell$ are "smooth enough", in the sense that the terms in $\sum \alpha_\ell \lambda_\ell^\alpha \varphi_\ell$ satisfy the Weierstrass M-test for all $0 \le \alpha \le p+1$ so that the series converges uniformly and absolutely to a continuous function (in $\overline{\Omega}$).*

Then OR can occur only due to the "periodic component" (defined below) of the solution.

First, we split the solution of the scheme equations into: (1) a homogeneous component u_h^n, which satisfies the initial conditions, with homogeneous b.c. and no forcing; and (2) a "periodic component" u_{p}^n, which satisfies the scheme equations (with forcing and full b.c.) with zero initial conditions (why we call this the "periodic component" is explained below).

First we show that u_h^n exhibits no OR; hence *any OR that occurs must do so solely due to the periodic component.* Clearly $u_h^n = \sum \alpha_\ell R_\ell^n \varphi_\ell$. The corresponding PDE solution $u_h = \sum \alpha_\ell e^{\lambda_\ell t} \varphi_\ell$. It follows that

$$\left\| \frac{u_h(t_n) - u_h^n}{\Delta t^p} \right\|_\infty \le n \Delta t R_*^{n-1} \sum_\ell |\alpha_\ell| D_g(\lambda_\ell \Delta t) |\lambda_\ell|^{p+1} \|\varphi_\ell\|_\infty \le C, \tag{B.1}$$

[4] Example: heat equation in $0 < x < \pi$, with Dirichlet b.c., and a stable scheme. Then $\varphi_\ell = \sin(\ell x)$ and $\lambda_\ell = -\ell^2$.

where C is a constant independent of n and Δt. The second inequality in (B.1) follows since: $n\Delta t R_*^n$ is bounded independent of n; while (iii-iv) imply the summation converges uniformly to a continuous function (on $\overline{\Omega}$). Hence, $\|u_h(t_n)-u_h^n\| \le C\Delta t^p = O(\Delta t^p)$.

We argue that the periodic component $u_{\rm p}^n$ is a linear superposition of periodic solutions (hence the name). We can solve for $u_{\rm p}^n$ using the Mellin/z-transform (correspondingly, the Laplace transform for the PDE). Because of (ii), the transform is analytic for $|z| \ge 1$, so the inverse Mellin transform can be written as an integral over the unit circle (correspondingly, the inverse Laplace transform can be written as an integral over the imaginary axis). However, because the initial data vanish, the integrands in these inverse transforms are actually periodic solutions to the scheme/equation, with forcing and b.c. provided by the transforms of the forcing and b.c. of the problem defining the periodic component.

Finally, if (iv) does not apply, then u_h may exhibit OR. However, for dissipative PDEs, u_h decays exponentially in time, and hence eventually, any OR that occurs will be dominated by the error in the periodic component $u_{\rm p}$.

Appendix C. Proof that the square bracket in (2.16) is $O(\Delta t^p)$. Here we compute the first bracket $B_1(x)$ in (2.16), and show that $B_1(x) = O(\Delta t^p)$, where:

$$B_1(x) := \underbrace{\left[\vec{\alpha}^T \vec{r}_1 \psi_1(x) + \psi_0(x) + \sum_{i=2}^{s} \vec{\alpha}^T \vec{r}_i \Phi_i^m(x) \right]}_{\text{Bracket 1}}. \tag{C.1}$$

The intuitive underlying reason that the first square bracket $B_1(x)$ in (2.16) satisfies $B_1(x) = O(\Delta t^p)$ is because this bracket represents the error contribution from the regular part of the asymptotic expansion—which is a Taylor expansion in the small parameter Δt. Because the RK scheme order arises from satisfying the equation up to $O(\Delta t^p)$ upon expanding the solution in powers of Δt, this result should not be surprising, even though the proof is not immediate. Generally, one can expect (and prove, though we do not do it here) that, wherever a regular expansion for the numerical solution applies, OR does not occur.

In the calculation below we make use of the following formulas:

(ai) From Proposition 2.2, the term $\vec{\alpha}^T \vec{r}_1 \psi_1(x) = O(\Delta t^p)$.

(aii) From (2.15–2.16) we have: $\psi_0 = -\frac{1}{z}\vec{\alpha}^T \vec{\delta} + O(\Delta t^p)$.

(aiii) From (2.16) the function $\vec{h}(x) = \frac{1}{z}\vec{\delta} + O(\Delta t^p)$ (since $\delta_0 = O(\Delta t^p)$), hence:

$$H_i(\Delta t)U^*(x) := \vec{l}_i^T \vec{h} = \frac{1}{z}\vec{\ell}_i^T \vec{\delta} + O(\Delta t^p). \tag{C.2}$$

(aiv) From the Definition in (2.28):

$$\begin{aligned} \Phi_i^m(x) &= H_i(\Delta t)\left(U^* + \lambda_i \mathcal{L}U^* + \lambda_i^2 \mathcal{L}^2 U^* + \ldots + \lambda_i^m \mathcal{L}^m U^*\right), \qquad \text{for } 2 \le i \le s. \\ &= \left(I + \lambda_i \mathcal{L} + \lambda_i^2 \mathcal{L}^2 + \ldots + \lambda_i^m \mathcal{L}^m\right) \frac{1}{z}\vec{\ell}_i^T \vec{\delta} + O(\Delta t^p) \qquad \text{(using (C.2))} \end{aligned}$$

(av) From Proposition 2.2, $\vec{\ell}_1^T \vec{\delta} = O(\Delta t^p)$ so that

$$\frac{1}{z}\vec{\alpha}^T \vec{r}_1 \left[I + \lambda_1 \mathcal{L} + \lambda_1^2 \mathcal{L}^2 + \ldots + \mathcal{L}^m\right] \vec{\ell}_1^T \delta = O(\Delta t^p). \tag{C.3}$$

Substituting the expressions from (ai), (aii) and (aiv) into Equation (2.33) yields:

$$B_1(x) = -\frac{1}{z}\vec{\alpha}^T \sum_{i=1}^{s} \vec{r}_i \vec{\ell}_i^T \vec{\delta} + \frac{1}{z}\vec{\alpha}^T \sum_{i=2}^{s} \vec{r}_i \left(I + \lambda_i \mathcal{L} + \ldots + \lambda_i^m \mathcal{L}^m\right) \vec{\ell}_i^T \vec{\delta} + O(\Delta t^p). \tag{C.4}$$

In (C.4) we have inserted a resolution of identity $I = \sum_{i=1}^{s} \vec{r}_i \vec{\ell}_i^T$ into the first term. Finally, adding (av) (which is an $O(\Delta t^p)$ correction) to (C.4) yields:

$$\begin{aligned} B_1(x) &= -\frac{1}{z}\vec{\alpha}^T \sum_{i=1}^{s} \vec{r}_i \vec{\ell}_i^T \vec{\delta} + \frac{1}{z}\vec{\alpha}^T \sum_{i=2}^{s} \vec{r}_i \left(I + \lambda_i \mathcal{L} + \ldots + \lambda_i^m \mathcal{L}^m\right) \vec{\ell}_i^T \vec{\delta} + O(\Delta t^p) \\ &= \frac{1}{z}\vec{\alpha}^T \left(\sum_{j=1}^{m} \sum_{i=1}^{s} \mathcal{L}^j \lambda_i^j \vec{r}_i \vec{\ell}_i^T \right) \vec{\delta} + O(\Delta t^p) = \frac{1}{z} \left(\sum_{j=1}^{m} \mathcal{L}^j \vec{\alpha}^T M^j \right) \vec{\delta} + O(\Delta t^p) \end{aligned} \tag{C.5}$$

We now use the following two identities

$$\vec{\alpha}^T M^j \vec{\tau}^{(i)} = 0, \qquad \text{for } 2 \le i + j \le p. \tag{C.6}$$

$$M^j = \sum_{i=1}^{s} \lambda_i^j \vec{r}_j \vec{\ell}_j^T = \lambda_1^j \vec{r}_1 \vec{\ell}_1^T + O(\Delta t^j). \tag{C.7}$$

Equation (C.6) follows from a direct calculation: $\vec{\alpha}^T M^j$ is spanned by vectors of the form $\vec{b}^T A^v$ for $0 \le v \le j-1$ so that Proposition 2.1 implies (C.6). The second identity (C.7) follows from a spectral expansion of M with $\lambda_i = O(\Delta t)$ for $2 \le i \le s$. Thus:

$$\begin{aligned} B_1(x) &= \frac{1}{z} \sum_{i=1}^{p-1} \sum_{j=1}^{m} \frac{\mathcal{L}^j (\Delta t)^i}{(i-1)!} \vec{\alpha}^T M^j \vec{\tau}^{(i)} + O(\Delta t^p) \\ &= \frac{1}{z} \sum_{i=1}^{p-1} \sum_{j=p-i+1}^{m} \frac{\mathcal{L}^j (\Delta t)^i}{(i-1)!} \vec{\alpha}^T M^j \vec{\tau}^{(i)} + O(\Delta t^p) \\ &= \frac{1}{z} \sum_{i=1}^{p-1} \sum_{j=p-i+1}^{m} \frac{\mathcal{L}^j (\Delta t)^i}{(i-1)!} \vec{\alpha}^T \left(\lambda_1^j \vec{r}_1 \underbrace{\vec{\ell}_1^T \vec{\tau}^{(i)}}_{=O(\Delta t^{p-i})} + O(\Delta t^{p-i+1}) \right) + O(\Delta t^p) = O(\Delta t^p). \end{aligned}$$

The identity $\vec{\ell}_1^T \vec{\tau}^{(i)} = O(\Delta t^{p-i})$ follows from (2.23) in Proposition 2.2.

REFERENCES

[1] S. Abarbanel, D. Gottlieb, and M.H. Carpenter, *On the removal of boundary errors caused by Runge-Kutta integration of nonlinear partial differential equations*, SIAM J. Sci. Comput., 17(3):777–782, 1996. i

[2] P. Albrecht, *The Runge-Kutta theory in a nutshell*, SIAM J. Numer. Anal., 33(5):1712–1735, 1996. 2.2

[3] I. Alonso-Mallo, *Runge-Kutta methods without order reduction for linear initial boundary value problems*, Numer. Math., 91(4):577–603, 2002. i, 2.5.5, 4, 4.1, 4.2

[4] I. Alonso-Mallo, B. Cano, and M.J. Moreta, *Order reduction and how to avoid it when explicit Runge-Kutta-Nyström methods are used to solve linear partial differential equations*, J. Comput. Appl. Math., 176(2):293–318, 2005. 1.3.2

[5] I. Alonso-Mallo and B. Cano, *Avoiding order reduction of Runge-Kutta discretizations for linear time-dependent parabolic problems*, BIT Numer. Math., 44(1):1–20, 2004. i, 4

[6] I. Alonso-Mallo and C. Palencia, *Optimal orders of convergence for Runge-Kutta methods and linear, initial boundary value problems*, Appl. Numer. Math., 44(1):1–19, 2003. 1.3.1

[7] R.E. Bank, W.M. Coughran, W. Fichtner, E.H. Grosse, D.J. Rose, and R.K. Smith, *Transient simulation of silicon devices and circuits*, IEEE Trans. Electron Devices, 32(10):1992–2007, 1985. 2.1

[8] F.L. Bauer and C.T. Fike, *Norms and exclusion theorems*, Numer. Math., 2(1):137–141, 1960. 2.3

[9] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978. 1.2, 2.2, 2.4, 2.5, 2.5.6, 4.1

[10] S. Boscarino, J.-M. Qiu, and G. Russo, *Implicit-explicit integral deferred correction methods for stiff problems*, SIAM J. Sci. Comput., 40(2):A787–A816, 2018. 1.3.1, 1.3.2

[11] P. Brenner, M. Crouzeix, and V. Thomée, *Single step methods for inhomogeneous linear differential equations in Banach space*, RAIRO Anal. Numér., 16(1):5–26, 1982. 1.3.1

[12] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations, 2nd Edition*, Wiley, New York, 2008. 2.1

[13] J.C. Butcher and Z. Jackiewicz, *Diagonally implicit general linear methods for ordinary differential equations*, BIT Numer. Math., 33:452–472, 1993. 6.2

[14] M. Calvo and C. Palencia, *Avoiding the order reduction of Runge-Kutta methods for linear initial boundary value problems*, Math. Comput., 71(240):1529–1543, 2002. 1.3.2

[15] M.H. Carpenter, D. Gottlieb, S. Abarbanel, and W.S. Don, *The theoretical accuracy of Runge-Kutta time discretizations for the initial boundary value problem: a study of the boundary error*, SIAM J. Sci. Comput., 16(6):1241–1252, 1995. i, 6

[16] G. Carrier and C. Pearson, *Partial Differential Equations: Theory and Technique*, Academic Press, Second Edition, 1988. 4.1

[17] C. González and A. Ostermann, *Optimal convergence results for Runge-Kutta discretizations of linear nonautonomous parabolic problems*, BIT Numer. Math., 39(1):79–95, 1999. 1.3.1

[18] S. Gottlieb, D.I. Ketcheson, and C.W. Shu, *High order strong stability preserving time discretizations*, J. Sci. Comput., 38(3):251–289, 2009. 1.3.2

[19] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I (2nd Revised Ed.): Nonstiff Problems*, Springer-Verlag, New York, 1993. 1, 2.1, 2.1, 2.2, 6.2

[20] E. Hairer and G. Wanner, *Stiff differential equations solved by Radau methods*, J. Comput. Appl. Math., 111(1):93–111, 1999. ii, 1.3.2, 6.2

[21] M.H. Holmes, *Introduction to Perturbation Methods*, Texts in Applied Mathematics, Springer Verlag, New York, 20, 1995. 2.5, 4.1

[22] W. Hundsdorfer and J.G. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Series in Comput. Math., Springer, 33, 2003. 1.3.1

[23] J.K. Hunter, *Asymptotic Analysis and Singular Perturbation Theory*, J.K. Hunter, UC Davis, 2004. 2.5

[24] Z. Jackiewicz, *General Linear Methods for Ordinary Differential Equations*, John Wiley & Sons, Inc., 2009. 6.2

[25] D. Ketcheson, B. Seibold, D. Shirokoff, and D. Zhou, *DIRK schemes with high weak stage order*, in S.J. Sherwin, D. Moxey, J. Peiró, P.E. Vincent, and C. Schwab (eds.), Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018, Springer, Cham., 134:453–463, 2019. ii, 2.1, 3.1, 3.1, 3.3

[26] D.I. Ketcheson and U. bin Waheed, *A comparison of high order explicit Runge-Kutta, extrapolation, and deferred correction methods in serial and parallel*, Commun. Appl. Math. Comput. Sci., 9(2):175–200, 2014. 1.3.2

[27] J. Kevorkian and J.D. Cole, *Multiple Scale and Singular Perturbation Methods*, Springer-Verlag, New York, 1996. 2.2, 2.5, 4.1

[28] A. Kværnø, S.P. Nørsett, and B. Owren, *Runge-Kutta research in Trondheim*, Appl. Numer. Math., 22(1):263–277, 1996. 2.2, 6

[29] C. Lubich, *On the convergence of multistep methods for nonlinear stiff differential equations*, Numer. Math., 58:839–853, 1991. 1.3.2, 6.2

[30] C. Lubich and A. Ostermann, *Runge-Kutta methods for parabolic equations and convolution quadrature*, Math. Comput., 60(201):105–131, 1993. 1.3.1, 2.2

[31] C. Lubich and A. Ostermann, *Interior estimates for time discretizations of parabolic equations*, Appl. Numer. Math., 18(1):241–251, 1995. 1.3.1, 2.2

[32] C. Lubich and A. Ostermann, *Runge-Kutta approximation of quasi-linear parabolic equations*, Math. Comput., 64(210):601–627, 1995. 1.3.1, 6

[33] M.L. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci., 1(3):471–500, 2003. 1.3.2

[34] M.L. Minion and R.I. Saye, *Higher-order temporal integration for the incompressible Navier-*

Stokes equations in bounded domains, J. Comput. Phys., 375:797–822, 2018. 1.3.2

[35] K. Nicholson, *Linear Algebra with Applications*, McGraw-Hill Ryerson Limited, Fourth Edition, 2003. 3.2

[36] A. Ostermann and M. Roche, *Runge-Kutta methods for partial differential equations and fractional orders of convergence*, Math. Comput., 59(200):403–420, 1992. 1.1, 1.3.1, ii, 2.5.5, 3.2, 3.2

[37] A. Ostermann and M. Roche, *Rosenbrock methods for partial differential equations and fractional orders of convergence*, SIAM J. Numer. Anal., 30(4):1084–1098, 1993. ii, 2.6

[38] A. Ostermann and M. Thalhammer, *Convergence of Runge-Kutta methods for nonlinear parabolic equations*, Appl. Numer. Math., 42(1):367–380, 2002. 1.3.1

[39] D. Pathria, *The correct formulation of intermediate boundary conditions for Runge-Kutta time integration of initial boundary value problems*, SIAM J. Sci. Comput., 18(5):1255–1266, 1997. i

[40] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag New York, Inc., USA, 44, 1983. 2.2

[41] A. Prothero and A. Robinson, *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*, Math. Comput., 28(125):145–162, 1974. 1, 2.1, 6.1

[42] J. Rang, *An analysis of the Prothero-Robinson example for constructing new DIRK and ROW methods*, J. Comput. Appl. Math., 262:105–114, 2014. ii, 3

[43] K. Salari and P. Knupp, *Code verification by the method of manufactured solutions*, Technical Report SAND2000-1444, Sandia National Laboratories, 2000. 5

[44] J.M. Sanz-Serna and J.G. Verwer, *Stability and convergence at the PDE / stiff ODE interface*, Appl. Numer. Math., 5:117–132, 1989. 1.1, 1.3.1

[45] J.M. Sanz-Serna, J.G. Verwer, and W.H. Hundsdorfer, *Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations*, Numer. Math., 50(4):405–418, 1986. 1.1, 1.3.1, 1.3.2, 6

[46] J.M. Sanz-Serna and J.G. Verwer, *Convergence analysis of one-step schemes in the method of lines*, Appl. Math. Comput., 31:183–196, 1989. 1.1, 1.3.1

[47] M. Schechter, *Principles of Functional Analysis*, Graduate Studies in Mathematics, Amer. Math. Soc., Second Edition, 36, 2000. 4.1

[48] S. Scholz, *Order barriers for the B-convergence of ROW methods*, Computing, 41:219–235, 1989. ii, 1.3.2, 3

[49] G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory*, Academic Press Inc., 1990. 2.3

[50] J.G. Verwer, *Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines*, Numer. Anal., 220–237, 1986. 1.3.1

[51] G. Wanner and E. Hairer, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1, 1991. 1, 2.1, 2.1, 6.1, A

[52] H. Zhang, A. Sandu, and S. Blaise, *Partitioned and implicit-explicit general linear methods for ordinary differential equations*, J. Sci. Comput., 61:119–144, 2014. 6.2