

DESIGN OF DIRK SCHEMES WITH HIGH WEAK STAGE ORDER

ABHIJIT BISWAS, DAVID KETCHESON, BENJAMIN SEIBOLD, AND DAVID SHIROKOFF

ABSTRACT. Runge-Kutta (RK) methods may exhibit order reduction when applied to certain stiff problems. While fully implicit RK schemes exist that avoid order reduction via high-stage order, DIRK (diagonally implicit Runge-Kutta) schemes are practically important due to their structural simplicity; however, these cannot possess high stage order. The concept of weak stage order (WSO) can also overcome order reduction, and it is compatible with the DIRK structure. DIRK schemes of WSO up to 3 have been proposed in the past, however, based on a simplified framework that cannot be extended beyond WSO 3. In this work a general theory of WSO is employed to overcome the prior WSO barrier and to construct practically useful high-order DIRK schemes with WSO 4 and above. The resulting DIRK schemes are stiffly accurate, L-stable, have optimized error coefficients, and are demonstrated to perform well on a portfolio of relevant ODE and PDE test problems.

1. INTRODUCTION

This paper focuses on Runge-Kutta (RK) methods for initial value problems

$$u'(t) = f(t, u(t)), \quad u(0) = u_0; \quad u \in \mathbb{R}^m, \quad f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m. \quad (1)$$

Let u_n and u_{n+1} denote the numerical approximations to the true solution at times t_n and $t_{n+1} = t_n + \Delta t$, respectively, where Δt is the time step size. One step of the RK method reads as

$$u_{n+1} = u_n + \Delta t \sum_{j=1}^s b_j f(t_n + c_j \Delta t, g_j), \quad (2)$$

via the stage approximations

$$g_i = u_n + \Delta t \sum_{j=1}^s a_{ij} f(t_n + c_j \Delta t, g_j), \quad i = 1, 2, \dots, s. \quad (3)$$

The parameters $A = (a_{ij})_{ij} \in \mathbb{R}^{s \times s}$ and $\vec{b} = (b_1, \dots, b_s)^T, \vec{c} = (c_1, \dots, c_s)^T \in \mathbb{R}^s$ that define the s -stage RK scheme are displayed via the Butcher tableau

$$\begin{array}{c|c} \vec{c} & A \\ \hline & \vec{b}^T \end{array}.$$

Throughout this work, we assume that the abscissas vector $\vec{c}$ is related to A via

$$\vec{c} = A\vec{e}, \quad (4)$$

where $\vec{e} \in \mathbb{R}^s$ is the vector of ones. Schemes for which A is lower-triangular are called *diagonally implicit Runge-Kutta* (DIRK) methods. Because the DIRK stage equations can be solved in sequence (whereas a fully-implicit RK method requires simultaneous solution of

2000 *Mathematics Subject Classification.* 65L04, 65L20, 65M06, 65M12, 65M22.

Key words and phrases. DIRK methods, weak stage order, order-reduction, stiffly accurate, A-stability.

all stages), these methods are of particular practical interest due to their implementation-friendly structure and cost efficiency.

One major drawback of RK methods is that they may exhibit order reduction [8, 9, 11, 12, 25, 33, 36, 17], i.e., the numerical solution of certain stiff problems [8] converges more slowly than what the formal order of the scheme would suggest. While there exist time-stepping methods that are devoid of order reduction, like linear multi-step methods (LMMs) [22], the practical importance of RK methods (as well as related approaches that are equivalent to RK methods [14]) renders the question “how can order reduction be avoided in RK methods?” central.

Order reduction may manifest in multiple shapes and forms. For explicit RK integration of mildly stiff IBVPs (e.g., advection), techniques to avoid the phenomenon have been developed in [1, 3, 4, 12, 27]. For stiff ODE problems, order reduction can be explained in terms of stiff limits [28, 17]. Order reduction in PDE IBVPs, first pointed out in [13, 15], manifests in an interesting geometric fashion, in a way that the time-stepping error produces spatial boundary layers [32]. Foundational work on the numerical analysis of order reduction includes [23, 24, 33, 36], and rigorous error analysis for RK methods applied to linear PDEs has been developed in [5, 16, 25, 34].

In the stiff setting, implicit Runge-Kutta (IRK) methods with high stage order [17] can remedy the order reduction phenomenon. Unfortunately, high stage order requires a fully implicit RK structure, while the DIRK methods [19] are limited to low stage order [20]. Approaches aimed at bridging this gap include a weaker criterion than stage order that diminishes order reduction specifically for ROW methods applied to linear problems [34]. Similar conditions were proposed in [25], albeit without providing numerical schemes that satisfy those conditions. In a similar spirit, the concept of weak stage order (WSO) was proposed in [32], which is a generalization of the conditions stated in [29]. More recently, [30, Chapter 6] and [31] extended the weak stage order conditions to generalized-structure additively partitioned Runge-Kutta (GARK) methods.

Like stage order, WSO imposes certain algebraic relations between the Runge-Kutta coefficients, but with two key differences: (a) WSO remedies order reduction only for certain problems; however, (b) the WSO conditions are compatible with the DIRK structure. A special case of WSO, called the WSO eigenvector criterion, has been studied in [20]: DIRK schemes up to order 4 and WSO 3 have been provided. At the same time, a barrier theorem was proved [20, 7]: for high-order DIRK schemes, the WSO eigenvector criterion cannot be extended beyond WSO 3.

2. THE ORDER REDUCTION PHENOMENON

In this section we review the order reduction phenomenon in the context of stiff ODEs. Prothero and Robinson [28] introduced a family of problems of the form

$$u' = \lambda(u - \phi(t)) + \phi'(t), \quad u(0) = u_0, \quad \text{with } \operatorname{Re}(\lambda) \leq 0. \quad (5)$$

Here $\phi(t)$ is any smooth function that varies at a moderate rate (i.e., $\phi'(t) = O(1)$), while λ is a parameter that allows one to make the problem (5) arbitrarily stiff. If $u(0) = \phi(0)$, then for any $\lambda \in \mathbb{C}$, (5) has the solution $u(t) = \phi(t)$. If $\operatorname{Re}(\lambda) \ll -1$, the problem (5) is stiff, and solutions different from $\phi(t)$ decay rapidly back to $\phi(t)$.

Equation (5) provides a useful model for analyzing the truncation errors of a Runge-Kutta scheme for a stiff problem. One may introduce the following *local truncation errors* (LTEs) [17]: $E_{i,\Delta t}(t_n)$ for the intermediate stages, and $E_{\Delta t}(t_n)$ for the final step update. The LTEs characterize the failure of the exact solution $u(t) = \phi(t)$ to satisfy the RK scheme, and are obtained as the residuals of substituting $g_i = \phi(t_n + c_i \Delta t)$, $u_n = \phi(t_n)$, and $u_{n+1} = \phi(t_n + \Delta t)$ into the RK scheme (2)–(3) applied to problem (5). Upon Taylor-expanding about t_n , the LTEs are [8]:

$$\vec{\mathcal{E}}(t_n) = \sum_{k \geq 1} \frac{(\Delta t)^k}{(k-1)!} \vec{\tau}^{(k)} \phi^{(k)}(t_n), \quad (6)$$

$$E_{\Delta t}(t_n) = \sum_{k \geq 1} \frac{(\Delta t)^k}{(k-1)!} \left[\sum_{j=1}^s b_j c_j^{k-1} - \frac{1}{k} \right] \phi^{(k)}(t_n), \quad (7)$$

where $\vec{\mathcal{E}}(t_n) := [E_{1,\Delta t}(t_n), E_{2,\Delta t}(t_n), \dots, E_{s,\Delta t}(t_n)]^T$. Here the vector

$$\vec{\tau}^{(k)} := A \vec{c}^{k-1} - \frac{1}{k} \vec{c}^k, \quad \text{for } k \geq 1,$$

is called the *k*th *stage order residual* (it will play an important role later), $\phi^{(k)}(t_n)$ is the *k*th derivative of ϕ at t_n , and $\vec{c}^k := [c_1^k, c_2^k, \dots, c_s^k]^T$ denotes component-wise exponentiation. Notice that (4) implies $\tau^{(1)} = 0$.

The numerical approximation error at time t_n is then defined as $\epsilon_n := u_n - \phi(t_n)$. It satisfies the same linear recursion as the RK scheme, with a forcing prescribed by the LTEs (6) and (7):

$$\epsilon_{n+1} = R(\zeta) \epsilon_n + \underbrace{\zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\mathcal{E}}(t_n)}_{=\mathcal{I}(\zeta)} + E_{\Delta t}(t_n). \quad (8)$$

Here $\zeta := \lambda \Delta t$, and $R(z)$ ($z \in \mathbb{C}$) is the *stability function*:

$$R(z) := 1 + z \vec{b}^T (I - zA)^{-1} \vec{e} = \frac{\det(I - zA + z \vec{e} \vec{b}^T)}{\det(I - zA)}. \quad (9)$$

Inspecting the expressions (6), (7) and (8) above, we see that the following conditions influence the order of the local error:

$$B(\xi) : \quad \vec{b}^T \vec{c}^{k-1} = \frac{1}{k} \quad \text{for } k = 1, 2, \dots, \xi; \quad (10)$$

$$C(\xi) : \quad \vec{\tau}^{(k)} = 0 \quad \text{for } k = 1, 2, \dots, \xi; \quad (11)$$

$$S(\xi) : \quad \vec{b}^T A^j \vec{\tau}^{(k)} = 0 \quad \text{for } k > 0, j + k < \xi; \quad (12)$$

$$T(\xi) : \quad \vec{b}^T A^{k-1} \vec{e} = \frac{1}{k!} \quad \text{for } k = 1, 2, \dots, \xi. \quad (13)$$

The conditions $B(\xi)$, $C(\xi)$ are widely used and known as *simplifying assumptions*; they determine the order of accuracy of the quadrature and subquadrature rules on which the RK method is based [10]. Notice that $B(p)$ implies $E_{\Delta t} = \mathcal{O}(\Delta t^{p+1})$. Here we have introduced notation for the additional conditions T and S since they play an important role below. The conditions $T(\xi)$ determine the order of accuracy of the method for non-stiff linear problems. The conditions $S(\xi)$ have appeared for instance in [2]. The conditions $B(p)$, $S(p)$, and $T(p)$ are necessary (though not sufficient) for a method to be of order p for general problems.

Notice that if $B(p)$ and $T(p)$ hold, then the first and last terms in (8) are $\mathcal{O}(\Delta t^{p+1})$. It remains only to bound the second term,

$$\mathcal{I}(\zeta) := \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\mathcal{E}}(t_n),$$

which is the one that causes order reduction in the stiff setting and on which we focus herein.

In the classical RK theory, i.e., in the non-stiff case, the scheme's convergence is studied in the limit $\Delta t \rightarrow 0$ with $\zeta = \mathcal{O}(\Delta t)$. A Neumann expansion in $|\zeta| \ll 1$ of $\zeta(I - \zeta A)^{-1} = \zeta I + \zeta^2 A + \zeta^3 A^2 + \dots$, then leads to the terms like $\vec{b}^T A^\ell \vec{\tau}^{(k)}$ in $\mathcal{I}(\zeta)$ with $\ell \geq 0$, so that condition $S(p)$ guarantees the one-step error is $\mathcal{O}(\Delta t^{p+1})$.

In the case of stiff problems, we are interested in time steps that are large relative to the fastest time scale of the problem dynamics, which is $\frac{1}{|\lambda|}$ for the Prothero-Robinson problem, i.e., we want $|\lambda|\Delta t \gg 1$. Hence, we study the convergence of errors under the simultaneous limits $\Delta t \rightarrow 0$ and $\zeta \rightarrow -\infty$, i.e., $\lambda \rightarrow -\infty$ faster than $\Delta t \rightarrow 0$. In this case ζ^{-1} is small and a Neumann expansion yields $\zeta(I - \zeta A)^{-1} = -A^{-1}(I - \zeta^{-1}A^{-1})^{-1} = -A^{-1} - \zeta^{-1}A^{-1} - \zeta^{-2}A^{-2} + \dots$, leading to the terms like $\vec{b}^T A^\ell \vec{\tau}^{(k)}$ but with $\ell < 0$. These quantities are not guaranteed to vanish by the order conditions, and this in general leads to order reduction.

One way to avoid order reduction is to use schemes with high stage order.

Definition 2.1 (Stage order). *The stage order of a RK scheme is $q = \min\{q_1, q_2\}$, where q_1, q_2 are the largest integers such that $B(q_1)$ and $C(q_2)$ hold.*

Stage order q implies that every stage of the scheme is an approximation accurate to at least order q , and in particular that the method itself has order at least q . Furthermore, for a scheme with stage order q , it can be shown that the local error is $\mathcal{O}(\Delta t^{q+1})$ even in the stiff regime, thus avoiding order reduction. Unfortunately, DIRK schemes are restricted to low stage order; see e.g. [20] for a proof of the following well-known result.

Theorem 2.2. *The stage order of an irreducible DIRK scheme is at most 2. The stage order of a DIRK scheme with non-singular A is at most 1.*

In the next section, we describe a criterion, called weak stage order (WSO), that is weaker than the stage order conditions but compatible with the DIRK structure. We show later that high order DIRK schemes with high WSO avoid order reduction for a certain class of problems, including the Prothero-Robinson problem (5).

3. WEAK STAGE ORDER, ORDER CONDITIONS, AND THEIR RELATIONSHIP

Since DIRK schemes cannot have high stage order, a weaker condition, referred to as *weak stage order* (WSO), was introduced [20, 32]. High weak stage order can alleviate order reduction in linear problems, and, in contrast to high stage order, is compatible with a DIRK structure.

The idea behind WSO is to prescribe conditions on $(A, \vec{b})$ that increase the accuracy of the problematic error term $\mathcal{I}(\zeta)$ in (8), via the following fact: it holds that $\mathcal{I}(\zeta) = \mathcal{O}(\Delta t^{q+1})$, if

$$\vec{b}^T (I - \zeta A)^{-1} \vec{\tau}^{(k)} = 0, \quad \text{for } k = 1, \dots, q. \quad (14)$$

Weak stage order is thus formulated to ensure that (14) holds.

Consider the following space, which is the direct sum of Krylov subspaces generated by the stage order residuals $\{\vec{\tau}^{(1)}, \vec{\tau}^{(2)}, \dots, \vec{\tau}^{(q)}\}$:

$$K_q := \text{span} \left\{ \vec{\tau}^{(1)}, A\vec{\tau}^{(1)}, \dots, A^{s-1}\vec{\tau}^{(1)}, \vec{\tau}^{(2)}, A\vec{\tau}^{(2)}, \dots, A^{s-1}\vec{\tau}^{(q)} \right\}. \quad (15)$$

Note that by definition, and through application of the Cayley-Hamilton theorem, K_q is an A -invariant subspace (that is $A\vec{v} \in K_q$ for any $\vec{v} \in K_q$). It will also be helpful to define the set of RK coefficients $(A, \vec{b})$ such that $\vec{b} \perp K_q$, namely:

$$\mathbb{W}_q := \left\{ A \in \mathbb{R}^{s \times s}, \vec{b} \in \mathbb{R}^s \mid \vec{b}^T A^j \vec{\tau}^{(k)} = 0, \text{ for } 0 \leq j \leq s-1, 1 \leq k \leq q \right\}. \quad (16)$$

Weak stage order can be defined in one of two equivalent ways.

Definition 3.1. (*Weak stage order, version 1*) The weak stage order q of an s -stage RK scheme $(A, \vec{b})$ is the largest integer for which $\vec{b} \perp K_q$, i.e., $(A, \vec{b}) \in \mathbb{W}_q$. If $\vec{b} \perp K_q$ holds for every $q \geq 1$, then $q = \infty$.

The second abstract version makes use of invariant subspaces.

Definition 3.2. (*Weak stage order, version 2*) The weak stage order q of an s -stage RK scheme $(A, \vec{b})$ is the largest integer for which there exists an A -invariant vector space V such that: $\vec{\tau}^{(k)} \in V$ for $1 \leq k \leq q$ and $\vec{b}^T y = 0$ for all $y \in V$. If $\vec{\tau}^{(k)} \in V$ for all $k \geq 1$, then $q = \infty$.

These two definitions of weak stage order are equivalent (i.e., taking $V = K_q$ in definition 3.2). Moreover, weak stage order is the most general criterion to guarantee (14), thereby avoiding order reduction in (stiff) linear problems [20].

In addition to the set of schemes that satisfy the WSO equations, we also introduce the set of p th order schemes:

$$\mathbb{V}_p := \left\{ A \in \mathbb{R}^{s \times s}, \vec{b} \in \mathbb{R}^s \mid (A, \vec{b}) \text{ satisfy all order conditions up to order } p \right\}. \quad (17)$$

A list of all order conditions up to $p = 5$ is given in Table 1.

We now discuss the polynomial equations defining RK schemes with order p and WSO q , i.e., the set $\mathbb{W}_q \cap \mathbb{V}_p$. Notice that the solutions to the WSO equations (i.e., schemes in $\mathbb{W}_q$) are of the same form as the conditions $S(\xi)$ (12), but they are required to hold for a larger set of values j, k . Since the conditions (12) appear explicitly in the RK order conditions as formulated by Albrecht [2], there is some overlap or redundancy between the conditions for WSO q that define $\mathbb{W}_q$ and the conditions for order p that define $\mathbb{V}_p$. In the more widely used formulation of RK order conditions due to Butcher (which we will also employ later), one instead has the related conditions (cf. (1))

$$\phi_{\ell,k} := \vec{b}^T A^{\ell-k} \vec{c}^k - \frac{k!}{(\ell+1)!} = 0 \quad \forall 0 \leq k \leq \ell \leq p-1.$$

The expressions appearing in the WSO conditions are just linear combinations of these $\phi_{\ell,k}$:

$$\vec{b}^T A^j \vec{\tau}^{(k)} = \vec{b}^T A^j \left(A \vec{c}^{k-1} - \frac{1}{k} \vec{c}^k \right) = \phi_{j+k,k-1} - \frac{1}{k} \phi_{j+k,k} \quad \forall j \geq 0, k \geq 1.$$

Therefore, if a method has WSO q then each of the conditions $\phi_{\ell,k} = 0$ for $1 \leq k \leq q$ is equivalent. When constructing RK schemes with high WSO, we can therefore pick just one

Order	$T(p)$	$B(p)$	Related to $S(p)$	Additional Order Conditions
$p = 1$	$\vec{b}^T \vec{e} = 1$			
$p = 2$	$\vec{b}^T A \vec{e} = \frac{1}{2}$			
$p = 3$	$\vec{b}^T A^2 \vec{e} = \frac{1}{3!}$	$\vec{b}^T \vec{e}^2 = \frac{1}{3}$		
$p = 4$	$\vec{b}^T A^3 \vec{e} = \frac{1}{4!}$	$\vec{b}^T \vec{e}^3 = \frac{1}{4}$	$\vec{b}^T A \vec{e}^2 = \frac{1}{12}$	$\vec{b}^T C A \vec{e} = \frac{1}{8}$
$p = 5$	$\vec{b}^T A^4 \vec{e} = \frac{1}{5!}$	$\vec{b}^T \vec{e}^4 = \frac{1}{5}$	$\vec{b}^T A^2 \vec{e}^2 = \frac{1}{60}$ $\vec{b}^T A \vec{e}^3 = \frac{1}{20}$	$\vec{b}^T C^2 A \vec{e} = \frac{1}{10}$ $\vec{b}^T C A \vec{e}^2 = \frac{1}{15}$ $\vec{b}^T A C A \vec{e} = \frac{1}{40}$ $\vec{b}^T D A \vec{e} = \frac{1}{20}$

Table 1. Order conditions (in Butcher’s notation): Here $D = \text{diag}(A\vec{e})$, $C = \text{diag}(\vec{e})$, and $\vec{e} = A\vec{e}$.

(p, q)	n_p	# redundant eq. $\frac{(p-1)(p-2)}{2}$	redundant $\phi_{j,k}$	$\phi_{j,k}$ kept
$(3, 2)$	4	1	$\phi_{2,1}$	$\phi_{1,1}, \phi_{2,2}$
$(3, 3)$	4	1	$\phi_{2,1}$	$\phi_{1,1}, \phi_{2,2}$
$(4, 3)$	8	3	$\phi_{2,1}, \phi_{3,2}, \phi_{3,1}$	$\phi_{1,1}, \phi_{2,2}, \phi_{3,3}$
$(4, 4)$	8	3	$\phi_{2,1}, \phi_{3,2}, \phi_{3,1}$	$\phi_{1,1}, \phi_{2,2}, \phi_{3,3}$
$(5, 4)$	17	6	$\phi_{2,1}, \phi_{3,2}, \phi_{4,3}$ $\phi_{3,1}, \phi_{4,2}, \phi_{4,1}$	$\phi_{1,1}, \phi_{2,2}, \phi_{3,3}$ $\phi_{4,4}$
$(5, 5)$	17	6	$\phi_{2,1}, \phi_{3,2}, \phi_{4,3}$ $\phi_{3,1}, \phi_{4,2}, \phi_{4,1}$	$\phi_{1,1}, \phi_{2,2}, \phi_{3,3}$ $\phi_{4,4}$

Table 2. Given order p and WSO q , the numbers n_p and $\frac{(p-1)(p-2)}{2}$ are the total, and redundant number of order condition, respectively. The last two columns show which $\phi_{\ell,k}$ we retain, vs. discard as redundant (since they are already implied by WSO and the retained conditions).

$\phi_{\ell,k} = 0$ from each equivalent family. Table 2 summarizes which order conditions we keep (and which we discard as redundant) in the construction of $\mathbb{W}_q \cap \mathbb{V}_p$.

Meanwhile, the WSO equations independently contain some redundancy; the Cayley-Hamilton theory used in definition 3.1 overestimates the set of equations required to define an invariant subspace K_q . The following section discusses how to construct a low-dimensional subspace K_q by removing redundant equations in the definition of $\mathbb{W}_q$.

4. RESULTS FROM WEAK STAGE ORDER THEORY

In this section we summarize key theoretical results from the companion paper [7]—which we use here to construct DIRK schemes with WSO greater than 3. The main results consist of (i) lower bounds on the number of stages required to obtain WSO greater than 3 (in terms of the order p); and (ii) formulas for constructing K_q .

To start, we first introduce the minimal polynomial for K_q which plays a central role in the results. Let $d = \dim(K_q)$ denote the dimension of K_q and $\vec{w}_j$, for $j = 1, \dots, d$, be a basis for K_q . Let

$$W := (\vec{w}_1 | \vec{w}_2 | \dots | \vec{w}_d) \in \mathbb{R}^{s \times d}.$$

Since the column space of W is A -invariant, there is a square matrix $B \in \mathbb{R}^{d \times d}$ such that

$$AW = WB. \quad (18)$$

Equation (18) simply states that each vector Aw_j can be expressed as a linear combination of vectors w_i , for $i = 1, \dots, d$.

The *minimal polynomial* [6, Chapter 8 & 9A] (see also [21, Chapter XIV §2]) $p(x)$ of a matrix B is the (non-zero) monic polynomial of smallest degree for which $p(B) = 0$. While it is often the case (for instance when the eigenvalues of B are distinct) that the minimal polynomial is the characteristic polynomial, in general $p(x)$ may be of lower degree than $\text{char}_B(x)$ when B has repeated eigenvalues (e.g., if $B = I$ is the $s \times s$ identity matrix then $p(x) = x - 1$ while $\text{char}_B(x) = (x - 1)^s$).

We define the *minimal polynomial* $P(x)$ of K_q as the minimal polynomial of B in (18). Note that $P(x)$ is intrinsic to the subspace K_q and remains invariant under a change of basis. That is, if $W' = WT$ for an invertible matrix T is an alternative basis for K_q , then (18) reads $AW' = W'B'$ where $B' = T^{-1}BT$ is just a conjugation of B . Since $P(B) = 0$ is equivalent to $P(B') = 0$, the $P(x)$ does not depend on the choice of basis for K_q .

The minimal polynomial $P(x)$ of K_q satisfies several important properties which follow from the linear algebra of matrices restricted to invariant subspaces. We summarize them here (without proof), along with their implications for DIRK schemes with WSO.

- a) $P(x)$ is the lowest degree (non-zero¹, monic²) polynomial that satisfies

$$P(A)\vec{w} = 0, \quad \forall \vec{w} \in K_q. \quad (19)$$

Due to the Krylov structure of K_q , relation (19) can be restated in terms of the vectors $\vec{\tau}_k$ as:

$$P(A)\vec{\tau}_k = 0, \quad \text{for } k = 2, \dots, q. \quad (20)$$

- b) $P(x)$ divides the characteristic polynomial of B . Thus,

$$\deg(P) \leq \dim(K_q).$$

- c) $P(x)$ divides the characteristic polynomial of A . Hence, every root of $P(x)$ is an eigenvalue of A . For DIRK schemes, the roots of $P(x)$ are then a subset of the diagonal entries of A , i.e., $\{a_{11}, a_{22}, \dots, a_{ss}\}$.

We now may summarize the key results from [7]. The first result is a limitation theorem on high WSO.

Theorem 4.1. (from [7]) *A DIRK scheme with invertible A and minimal polynomial satisfying $\deg(P) \leq 1$ is limited to WSO $q \leq 3$.*

The practical implication of theorem 4.1 is that WSO $q > 3$ requires a minimal polynomial $\deg(P) \geq 2$. The next theorem demonstrates how the WSO q impacts the number of stages s required to achieve a given order p .

Theorem 4.2. (from [7]) *An s -stage DIRK scheme with n_c distinct abscissa values, order $p \geq 1$, and weak stage order $q \leq 2n_c - 1$ (with K_q and $P(x)$ defined in (15) and (20)) satisfies*

$$s - p + 1 - \sigma \geq \dim(K_q) \geq \left\lfloor \frac{q}{2} \right\rfloor,$$

¹In the case when $K_q = \{0\}$, $P(x) = 1$ is the constant polynomial.

²The highest-power coefficient is 1.

where $\sigma = 1$ if the method is stiffly accurate, and $\sigma = 0$ otherwise.

A classical result in RK theory (e.g. [17, Theorem 4.18]) is that DIRK schemes have order limited to $p \leq s + 1$. Theorem 4.2 highlights that the “gap” in this bound is exactly what enables WSO q .

Our goal here is to construct schemes with $q > 3$. Motivated by the implications of theorem 4.1 and theorem 4.2, we choose $\dim(K_q)$ as small as possible, that is $\dim(K_q) = 2$ (and hence $\deg(P) \leq 2$). Theorem 4.2 then requires the number of stages to be $s \geq p + 1 + \sigma$, and limits q to $q \leq 5$ (which suffices for this work; however, the theory also allows for $q > 5$ if $\dim(K_q) > 2$). Furthermore, the following theorem (also from [7]) characterizes the roots of $P(x)$ when $q > 3$:

Theorem 4.3. *(Minimal polynomial when $\deg(P) = 2$; [7]) Consider an irreducible DIRK scheme with invertible A and WSO $q > 3$. If K_q has a minimal polynomial with $\deg(P) = 2$, then*

$$P(x) = (x - a_{11})(x - a_{22}) . \quad (21)$$

Note that $P(x)$ in (21) is valid for both cases $a_{11} = a_{22}$ and $a_{11} \neq a_{22}$.

We turn our attention to constructing spaces $K_q = \text{span}\{\vec{w}_1, \vec{w}_2\}$ with $\dim(K_q) = 2$ and $\deg(P) = 2$ (otherwise, via theorem 4.1, $\deg(P) \in \{0, 1\}$ would result in WSO 3 or less). Theorem 4.3 requires:

$$(A - a_{11}I)(A - a_{22}I)\vec{w}_j = 0, \quad j = 1, 2 . \quad (22)$$

If $\vec{w}_k$ are chosen as eigenvectors of A , then (without loss of generality) the solution to (22) is exactly one of:

$$\text{When } a_{11} \neq a_{22} : \quad A\vec{w}_1 = a_{11}\vec{w}_1 \quad \text{and} \quad A\vec{w}_2 = a_{22}\vec{w}_2 , \quad (23)$$

$$\text{When } a_{11} = a_{22} : \quad A\vec{w}_1 = a_{11}\vec{w}_1 \quad \text{and} \quad A\vec{w}_2 = a_{11}\vec{w}_2 + \vec{w}_1 . \quad (24)$$

Note that no other solution to (22) is allowed: neither of the monomials $(A - a_{11}I)$ or $(A - a_{22}I)$ in (22) can individually annihilate both vectors in K_q ; otherwise the degree of P would be 1. This forces the vectors $\vec{w}_j$ ($j = 1, 2$) to have distinct eigenvalues when $a_{11} \neq a_{22}$, or $\vec{w}_2$ to be generalized eigenvectors when $a_{11} = a_{22}$.

Finally, the space K_q with $\dim(K_q) = 2$ has the form

$$\vec{\tau}^{(k)} = \beta_1^{(k)} \vec{w}_1 + \beta_2^{(k)} \vec{w}_2, \quad \text{for } k = 2, 3, \dots, q , \quad (25)$$

where $\beta_1^{(k)}, \beta_2^{(k)}$ are unknown coefficients to be solved for (along with A). WSO then may be guaranteed if $\vec{b} \perp K_q$, that is:

$$\vec{b}^T \vec{w}_1 = 0, \quad \vec{b}^T \vec{w}_2 = 0 . \quad (26)$$

Together, we will use equations (23) and (24), as well as equations (25) and (26) as a (minimal) system of equations for weak stage order (with $\dim(K_q) = 2$).

As a final remark, we discuss a family of DIRK schemes that satisfy the WSO equations (23) and (25), yet are reducible to smaller (equivalent) schemes. Identifying and avoiding reducible DIRK schemes is important for the construction of schemes in the next section.

We say a scheme is *r-confluent* if the abscissas of its first r stages all coincide, i.e., $c_1 = \dots = c_r$. DIRK schemes that are *r-confluent* are equivalent (specifically, *S-reducible* [17,

Chapter IV.12]) to a simpler DIRK scheme where one stage replaces stages 1 through r . Specifically, consider two DIRK schemes in block form

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \text{ with } \vec{b} = \begin{pmatrix} \vec{b}_1 \\ \vec{b}_2 \end{pmatrix}, \text{ and } A^* = \begin{pmatrix} a_{11} & 0 \\ \vec{a}_{21}^* & A_{22} \end{pmatrix} \text{ with } \vec{b}^* = \begin{pmatrix} b_1^* \\ \vec{b}_2 \end{pmatrix}, \quad (27)$$

where $\vec{a}_{21}^* := A_{21}\vec{e}$, is a vector consisting of the row sums of A_{21} and $b_1^* = \vec{b}_1^T \vec{e}$ (where $\vec{e}$ here is of length $m = \text{number of columns of } A_{21}$). Then, we have

Lemma 4.4. *Let $(A, \vec{b})$ be an s -stage DIRK scheme with block structure given in (27). If $(A_{11}, \vec{b}_1)$ is an r -confluent scheme with r stages where $2 \leq r \leq s$, then $(A, \vec{b})$ is reducible to $(A^*, \vec{b}^*)$.*

Proof. Applying Definition 12.17 in [17], where the partition of equivalent stages (i.e., partition of the integers $\{1, \dots, s\}$) is taken as $S_1 = \{1, 2, \dots, r\}$ and $S_2 = \{r+1\}, \dots, S_{s-r} = \{s\}$ shows that the scheme is S -reducible. Then [18, Theorem 2.2] implies that the first r stages of A yield the same intermediate stage value solutions—and thus can be replaced by a single stage. $\square$

5. OPTIMIZATION PROBLEM FOR FINDING DIRK SCHEMES WITH DESIRABLE PROPERTIES

In this section we formulate and numerically solve the problem of constructing DIRK schemes with a prescribed order p and WSO q that are A-stable, stiffly accurate (and hence L-stable), and have an optimally small error constant. A-priori, the degrees of freedom in this optimization problem are the coefficients in the matrix $A \in \mathbb{R}^{s \times s}$ and the vector $\vec{b} \in \mathbb{R}^s$. The constraints are as follows:

- (C.0) (DIRK structure) $a_{ij} = 0$ for $j > i$;
- (C.1) (p th order conditions) $(A, \vec{b}) \in \mathbb{V}_p$, defined in (17);
- (C.2) (weak stage order q) $(A, \vec{b}) \in \mathbb{W}_q$, defined in (16);
- (C.2') (WSO if $\dim(K_q) = 2$) $(A, \vec{b})$ satisfy (25), (26), and either (23) or (24);
- (C.3) (stiff accuracy) $a_{sj} = b_j$ for $j = 1, \dots, s$;
- (C.4) (non-negative abscissas) $c_j \geq 0$ for $j = 1, \dots, s$;
- (C.5) (A-stability condition 1) $a_{ii} \geq 0$ for $1 \leq i \leq s$;
- (C.6) (A-stability condition 2) $|R(iy)| \leq 1$ for all $y \in \mathbb{R}$;
- (CR.6) (relaxation of (C.6)) $|R(iy)| \leq 1$ for $y \in \{y_1, \dots, y_m\}$ with $0 \leq y_1 < \dots < y_m$.

Condition (C.0) enforces the DIRK structure. Conditions (C.1) and (C.2) are simply the order conditions and weak stage order conditions, respectively, while (C.2') is a simplified set of WSO conditions (based on §4) for the special case $\dim(K_q) = 2$. Condition (C.3) ensures stiff accuracy and guarantees that the numerical solution is exact in the limit $\Delta t \rightarrow 0$ and $\zeta \rightarrow -\infty$ [17]. It has the effect of prescribing $\vec{b}$ in terms of A so that the degrees of freedom are the matrix A only. Constraint (C.4) ensures that evaluations of $f(t, u)$ in (1) do not occur prior to the initial time.

Lastly, conditions (C.5) and (C.6) impose A-stability (see [17, Chapter IV.3 Eqs. (3.6), (3.7)]), which, combined with (C.3), ensures L-stability. Condition (C.6) is, as written, an infinite set of constraints. While it is possible to recast (C.6) as a finite set of inequalities involving semi-definite matrices (using a connection between non-negative single-variable

polynomials and polynomials written as a sum of squares), here we take the simpler approach of approximating (C.6) by imposing it only at a finite set of values on the imaginary axis. Taking advantage of the symmetry $|R(iy)| = |R(-iy)|$ leads to the weaker set of constraints (CR.6).

For the purpose of constructing DIRK schemes with high WSO, we denote the set of equality constraints $\mathcal{C}_{\text{Eq}}$ and inequality constraints $\mathcal{C}_{\text{InEq}}$ by:

$$\begin{aligned}\mathcal{C}_{\text{Eq}} &:= \{(A, \vec{b}) \in \mathbb{R}^{s \times s} \times \mathbb{R}^s \mid (C.0), (C.1), (C.2'), \text{ and } (C.3) \text{ hold}\}, \\ \mathcal{C}_{\text{InEq}} &:= \{(A, \vec{b}) \in \mathbb{R}^{s \times s} \times \mathbb{R}^s \mid (C.4), (C.5), \text{ and } (CR.6) \text{ hold}\}.\end{aligned}$$

A *feasible scheme* is one that satisfies both sets of constraints.

To guide the construction of DIRK schemes that achieve a minimal error, we use the ℓ^2 -norm of the residuals of the $(p+1)$ st order conditions as a proxy for the error, leading to the objective function:

$$(\text{Ob}) \quad (\text{Objective function}) \quad F(A, \vec{b}) := \|(p+1)\text{st order conditions}\|_{\ell^2}^2.$$

For instance, for $p = 1$ and $p = 2$, the objective function (using $\vec{c} = A\vec{e}$) is:

$$\begin{aligned}p = 1: \quad F(A, \vec{b}) &= \left(\vec{b}^T A \vec{e} - \frac{1}{2}\right)^2, \\ p = 2: \quad F(A, \vec{b}) &= \left(\vec{b}^T A^2 \vec{e} - \frac{1}{6}\right)^2 + \left(\vec{b}^T \vec{c}^2 - \frac{1}{3}\right)^2.\end{aligned}$$

If enough degrees of freedom are allowed, then a feasible scheme of order p may satisfy $F(A, \vec{b}) = 0$, and thus be of order $p+1$. However, generally, locally optimal schemes will not satisfy $F(A, \vec{b}) = 0$ exactly. Altogether, we seek optimal DIRK schemes via the following constrained minimization problem:

$$\begin{aligned}(M) \quad & \text{Minimize } F(A, \vec{b}) \\ & \text{Subject to } (C.0), (C.1), (C.2'), (C.3) \text{ and} \\ & \quad (C.4), (C.5), (CR.6).\end{aligned}$$

5.1. Solution to Problem (M). While one can attempt to solve problem (M) directly via black-box optimization routines, numerical experiments revealed that such a direct approach becomes highly inefficient as s , p , and q are increased. Plausibly caused by the problem's lack of convexity and ill-conditioned constraints, feasible, let alone optimal, solutions are found increasingly rarely with increasing s , p , and q . In order to facilitate a more robust approach, we instead propose to solve (M) in two major steps: We first construct a feasible scheme; then, using the feasible scheme as initial guess, we apply a local optimizer to minimize the objective function.

1. Construction of a feasible scheme: We solve a sequence of sub-problems to find a feasible scheme satisfying both $\mathcal{C}_{\text{Eq}}$ and $\mathcal{C}_{\text{InEq}}$. Steps (1A) and (1B) construct a solution satisfying the equality constraints only. Step (1C) then incorporates the inequality constraints as well.

Step (1A). This step utilizes a hybrid analytical and numerical approach to find a point $(A, \vec{b})$ in $\mathcal{C}_{\text{Eq}}$. Below, the substeps (a)–(e) solve the first $(s-1)$ rows of equation (25) and make use of the theory from §4; then substep (f) solves the remaining constraints. All

numerical solutions in this step use MATLAB's `sqp` algorithm in `fmincon`, prescribing a constant objective function in order to use it simply as an algebraic solver.

- (a) Solve (analytically) the first two components of the eigenvectors $\vec{w}^{(1)}, \vec{w}^{(2)}$: there are two solution branches, one corresponding to (23) and another (24). We restrict our solutions to the branch (23) where $a_{11} \neq a_{22}$. The alternative case, $a_{11} = a_{22}$ is also possible but not pursued here.
- (b) Solve (analytically) for $\beta_1^{(k)}, \beta_2^{(k)}$ in equation (25): the first two rows of (25) uniquely define $\beta_1^{(k)}, \beta_2^{(k)}$, and are then automatically satisfied.
- (c) Solve (analytically) for the third component/row of (25): the 3 equations in six variables ($a_{11}, a_{21}, a_{22}, a_{31}, a_{32}, a_{33}$) for $q = 4$ can be solved by parameterizing a_{21}, a_{31} , and a_{32} in terms of a_{11}, a_{22} , and a_{33} . We choose the parameterized branch to avoid the reducible r -confluent schemes (see §4).
- (d) Construct (numerically) the upper 3×3 block of A with numerical entries. For $q = 4$, we select a_{11}, a_{22} , and a_{33} randomly and use the parameterization in (c) to determine a_{21}, a_{31}, a_{32} . For $q = 5$, we find a numerical solution to the third component/row of (25) for $\vec{\tau}^{(5)}$ (which via the parameterization is an equation in terms of a_{11}, a_{22}, a_{33}).
- (e) Solve (numerically) row by row (from row 4 through $s - 1$), the $(q - 1)$ equations in (25). When $q = 4, 5$, the r th row yields $(q - 1) (\leq r)$ equations in r variables. At each row, we numerically find a solution. At the end of this substep, the upper $(s - 1) \times (s - 1)$ block of A is populated with numerical entries.
- (f) To satisfy (C.3) we set $\vec{b}^T = (a_{s1}, a_{s2}, \dots, a_{ss})$, then solve (numerically) for the last row of A . Together, this amounts to solving the $(q - 1)$ equations from row s in (25) (i.e., to satisfy (C.2')) and the non-redundant (cf. §3) order conditions (C.1),

$$\vec{b}^T [\vec{w}^{(1)}, \vec{w}^{(2)}, \vec{c}, \vec{c}, \vec{c}^2, \vec{c}^3, C A \vec{c}] = [0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{8}], \quad (28)$$

for order $p = 4$. Order $p = 5$ requires, in addition to (28):

$$\vec{b}^T [\vec{c}^4, C^2 A \vec{c}, C A C \vec{c}, \text{diag}(A \vec{c}) A \vec{c}, C A^2 \vec{c}, A C A \vec{c}] = [\frac{1}{5}, \frac{1}{10}, \frac{1}{15}, \frac{1}{20}, \frac{1}{30}, \frac{1}{40}].$$

For instance, a 4th order DIRK scheme with WSO 4, yields 10 equations (7 from (28) and $q - 1 = 3$ from row s of (25)) in s variables.

This procedure generates a random scheme that satisfies the equality constraints. For robustness purposes, we choose to reject (and simply re-start the step) any scheme that has a coefficient larger than 20 in absolute value (in line with [35]) or fails to satisfy the constraints to within 10^{-10} .

Step (1B). Step (1A) uses `fmincon` as a solver with (for computational speed) the residual error tolerance set significantly larger than machine precision. This generates a first approximation $(A, \vec{b})$ to constraints (C.1), (C.2'), and (C.3). To drive the residuals down to machine precision, we use the output from step (1A) as a starting point and solve the equations defined by $\mathcal{C}_{\text{Eq}}$ (again) via a Gauss-Newton iteration.

Step (1C). In this step we reincorporate the inequality constraints: non-negative abscissae (C.4) and A-stability ((C.5) and (CR.6)) to construct (fully) feasible schemes. Using the output $(A, \vec{b})$ from step (1B), we call `fmincon` with the full constraint set and (again) a constant objective function. The resulting schemes turn out to satisfy the equality constraints to machine precision, and we observe that they tend to not lie on the boundary of the

inequality constraints $\mathcal{C}_{\text{InEq}}$. In the occurrence that the optimization solve in this step fails, we simply restart from step (1A).

Remark 5.1. (*SDIRKs and WSO*) Restricting to the solution branch defined by (23) in (1A) rules out singly diagonally implicit Runge-Kutta schemes (SDIRKs), which have all their diagonal entries identical. SDIRK schemes are of practical interest for their ease of implementation. While we defer the study of SDIRK schemes to later work, a preliminary exploration of the other solution branch (24) (albeit with $\dim(K_q) > 2$) revealed that SDIRKs with high WSO do exist, demonstrating that the SDIRK structure is compatible with high WSO.

2. *Optimization:* We will refer to schemes by the triple (s, p, q) , representing the number of stages, classical order, and weak stage order, respectively. According to theorem 4.2, a 4th order DIRK scheme with WSO 4 requires at least five stages, and a 5th DIRK scheme with WSO 4 or 5 requires at least six stages. These are lower bounds that may not be sharp, and the additional requirements we have imposed (such as A-stability and stiff accuracy) are likely to further increase the minimum viable number of stages. In numerical searches, we have found schemes of type $(7, 4, 4)$, $(12, 5, 4)$, $(12, 5, 5)$. Numerical searches failed to find methods with the corresponding p, q and fewer stages s .

We repeatedly (100,000+ times) solve (M) via local optimization, starting with initial guesses given by the output of step (1C). This yields a set of locally optimal schemes. We use MATLAB's `fmincon` with the gradient-based `sqp` algorithm. Among the locally optimal schemes found in this manner, we have selected one from each class that is close to optimal in terms of $F(A, \vec{b})$ and is Pareto-optimal in terms of minimizing $F(A, \vec{b})$ and minimizing $\max_{i,j} |a_{ij}|$. We thus provide three schemes, one for each triple (s, p, q) : DIRK-(7, 4, 4), DIRK-(12, 5, 4), and DIRK-(12, 5, 5). Since we used the relaxation (CR.6) in place of (C.6), we check *a posteriori* that the schemes are in fact A-stable. The stability regions and magnitude of the stability function along the imaginary axis, shown in Figure 1, confirm this. Scheme coefficients are given in appendix A.

6. NUMERICAL RESULTS: LINEAR PROBLEMS WITH AUTONOMOUS OPERATORS

This section presents numerical test cases for ODE and PDE problems with linear operators with time-independent coefficients (the forcing and solutions may be time-dependent). This is the class of problems for which WSO is expected to alleviate order reduction.

There are only a handful of theoretical results characterizing for which problems or PDEs the convergence rate can be guaranteed to be equal to the weak stage order. For example, Ostermann and Roche [25] examined linear boundary value problems, i.e., $u_t = \mathcal{L}u + f$ with boundary condition $\mathcal{B}u = 0$, where $\mathcal{L}$ has a complete L^2 eigenfunction basis with (point spectrum) eigenvalues satisfying $\text{Re}(\lambda) \leq 0$ (see Assumptions (3.1) in [25]). Here $\mathcal{L}$ may have coefficients that depend on space x , but not on time t ; and $f(t)$ may be time-dependent. Then RK schemes satisfying the condition

$$W_k(z) \equiv 0 \quad \text{for } 1 \leq k \leq q, \quad \text{where} \quad W_k(z) := \frac{k\vec{b}^T(I - zA)^{-1}\vec{\tau}^{(k)}}{R(z) - 1}, \quad (29)$$

along with Assumptions (2.9) in [25], overcome order reduction. Condition (29) is (essentially) implied by WSO q . In a similar spirit, condition (29) remedies order reduction for Rosenbrock methods [26] in a more abstract setting where $\mathcal{L}$ is the infinitesimal generator of an analytic

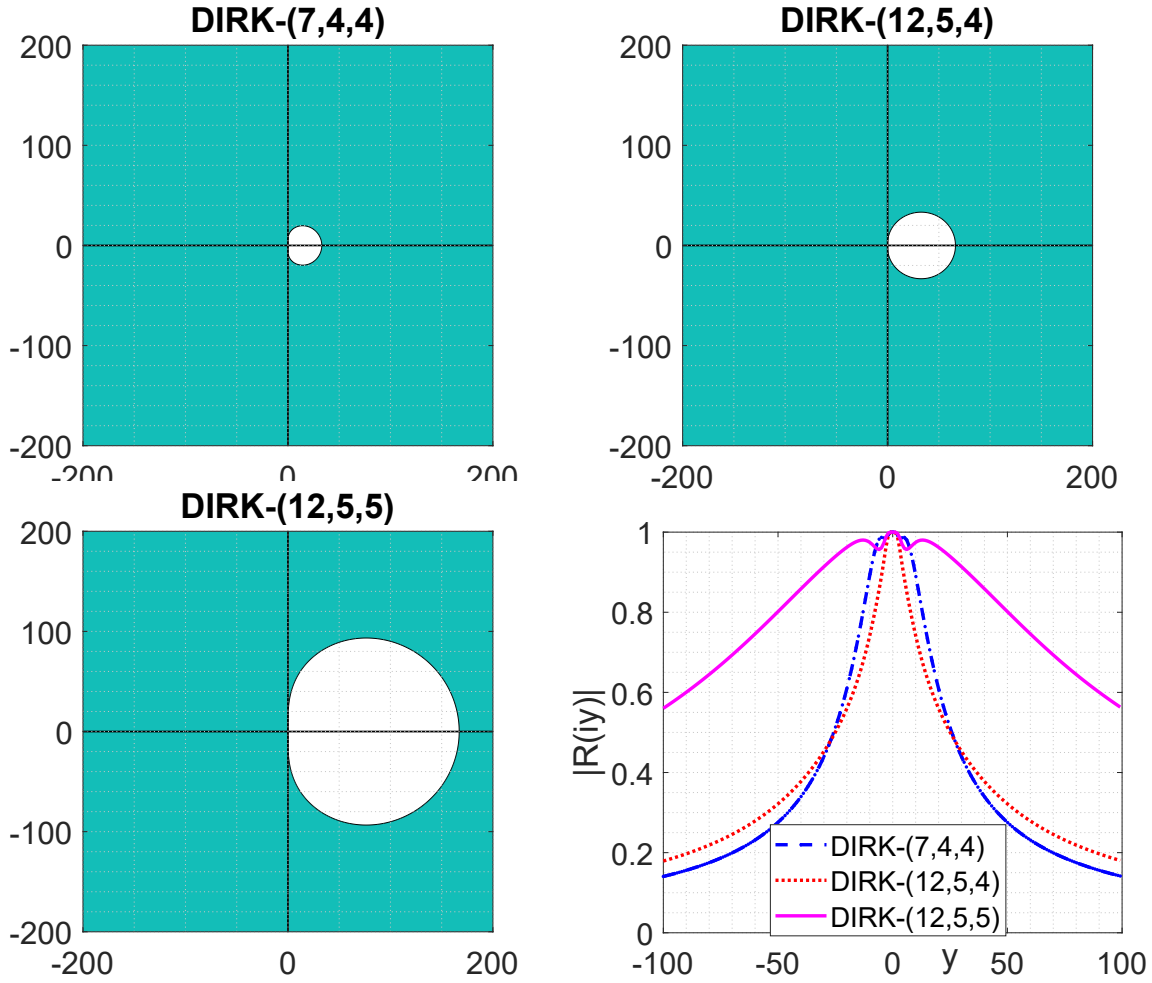


Figure 1. Stability regions (in $\mathbb{C}$) for the new schemes, DIRK-(7, 4, 4), DIRK-(12, 5, 4), and DIRK-(12, 5, 5), and $|R(iy)|$ for $y \in \mathbb{R}$ for the three schemes. Observe that $|R(iy)| \leq 1$ along the imaginary axis, hence these methods are A-stable.

semi-group. While several of the PDEs we test here fall under the framework of known convergence results in [25], some do not, such as the linear advection equation in §6.5.

Below, the new schemes are denoted by (s, p, q) , where s is the number of stages, p is the classical order, and q is the scheme's weak stage order. Methods with $q = p$ for ODEs and $q = p - 1$ for PDEs yield solutions that converge at the rate p . However, for PDEs, spatial derivatives of the solution may still exhibit order reduction if $q = p - 1$. Methods with $q = p$ for PDE problems also alleviate order reduction in the solution's derivatives [25, 32].

As references of comparison for our newly devised (high WSO) schemes, we include two schemes with WSO $q = 1$, referred to as DIRK-(5, 4, 1) [17, Chapter IV.6, Table 6.5] which is A-stable and stiffly accurate; and DIRK-(5, 5, 1) [19, Table 24, p. 98], which is A-stable but not stiffly accurate.

In each PDE test problem below, a spatial approximation is chosen so that the spatial approximation error becomes negligible relative to the temporal error. Hence, the error convergence plots below isolate the temporal error generated by the different DIRK schemes with

high (and low) WSO. Note that for the different test problems, different spatial approximation strategies and numbers of grid points are employed to achieve this objective while also balancing simplicity and computational efficiency.

6.1. Prothero-Robinson ODE test problem. We study the problem (5) with true solution $\phi(t) = e^{-t} \sin(10t) + \cos(20t)$, stiffness parameter $\lambda = -10^4$, initial condition $u(0) = \phi(0)$ and final time $T = 10$. Figure 2 contrasts high versus low WSO schemes, i.e., DIRK-(7,4,4) vs. DIRK-(5,4,1), as well as DIRK-(12,5,5) vs. DIRK-(5,5,1). For each scheme we observe convergence order p for small enough Δt . However, in line with the theoretical predictions, for the schemes with $q < p$ the convergence rate is lower (approximately equal to q) for larger values of Δt , i.e., in the stiff regime.

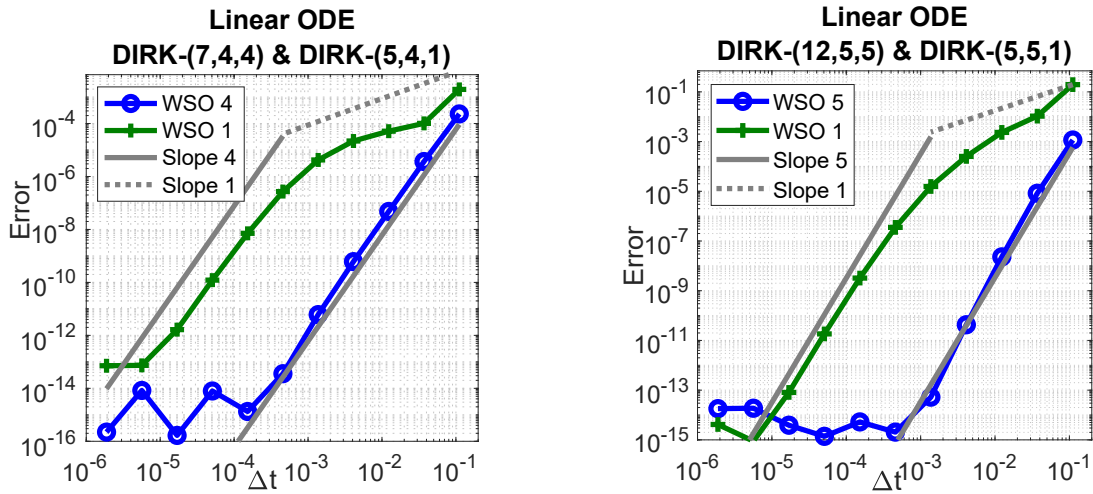


Figure 2. Convergence for the Prothero-Robinson test problem using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (blue circles) and WSO 1 (green) (left), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (blue circles) and WSO 1 (green) (right).

6.2. Heat equation. Next we consider the 1D heat equation

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

with the forcing $f(x, t)$, the boundary conditions (b.c.) and the initial condition (i.c.) chosen such that $u(x, t) = \cos(20t) \sin(10x + 10)$. To isolate the temporal error, we use a 4th order centered finite difference approximation in space on a grid with 10^4 points. Errors are computed at the final time $T = 1$ using the maximum norm in space. Figure 3 shows the convergence of function values u and derivatives u_x using the three new high WSO DIRK schemes, compared with reference WSO-1 DIRK schemes of the respective orders. In agreement with the analysis in [32], for this second-order PDE, the time stepping schemes produce spatial boundary layers (BLs) of width $\mathcal{O}(\Delta t^{0.5})$, resulting in a loss of half an order in u_x when $q < p$. The results confirm the full order of convergence in u and u_x when using DIRK-(7, 4, 4) and DIRK-(12, 5, 5), and the full order in u and half order loss in u_x with DIRK-(12, 5, 4). Note that with the given setup, the spatial approximation error is about 10^{-11} , hence the stagnation of the errors around that value.

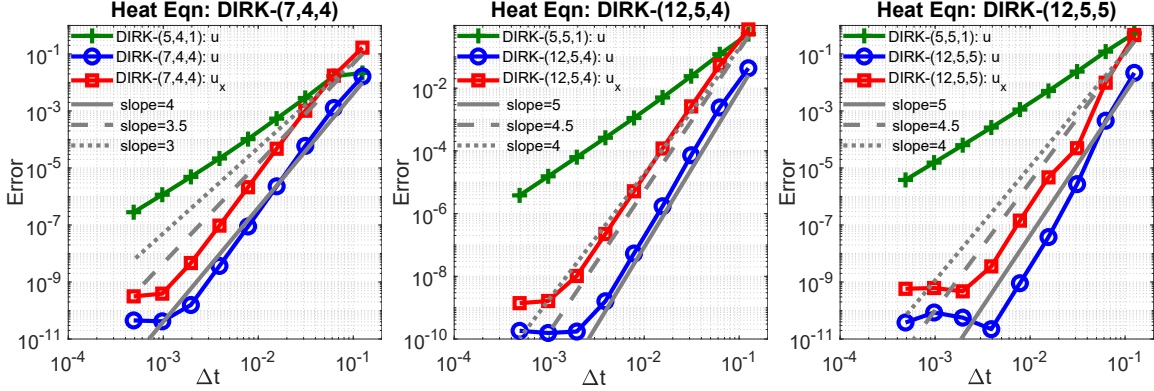


Figure 3. Convergence (u blue circles; u_x red squares) for heat equation using DIRK-(7,4,4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12,5,4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12,5,5): 5th order DIRK scheme with WSO 5 (right).

6.3. Schrödinger equation. As an example of a dispersive problem we consider

$$u_t = \frac{i\omega}{k^2} u_{xx} + f \quad \text{for } (x, t) \in (0, 1) \times (0, 1.2], \quad u = g(x, t) \quad \text{on } \{0, 1\} \times (0, 1.2],$$

with the manufactured solution $u(x, t) = \exp(-(x-t)^2) \cos(10x) \sin(t)$, where $\omega = 2\pi$ and $k = 20$. As above, u_{xx} is approximated by 4th order centered differences on a fine grid with 10^4 cells. The problem is solved up to final time $T = 1.2$ via different RK schemes, and the convergence in u and u_x is evaluated. Figure 4 shows the results obtained with DIRK-(7,4,4) (left), DIRK-(12,5,4) (middle), and DIRK-(12,5,5) (right), relative to DIRK-(5,4,1) and DIRK-(5,5,1) reference methods. Similar convergence results are observed as for the heat equation: full orders in u and u_x are recovered with DIRK-(7,4,4) and DIRK-(12,5,5), while u_x loses a half order with DIRK-(12,5,4).

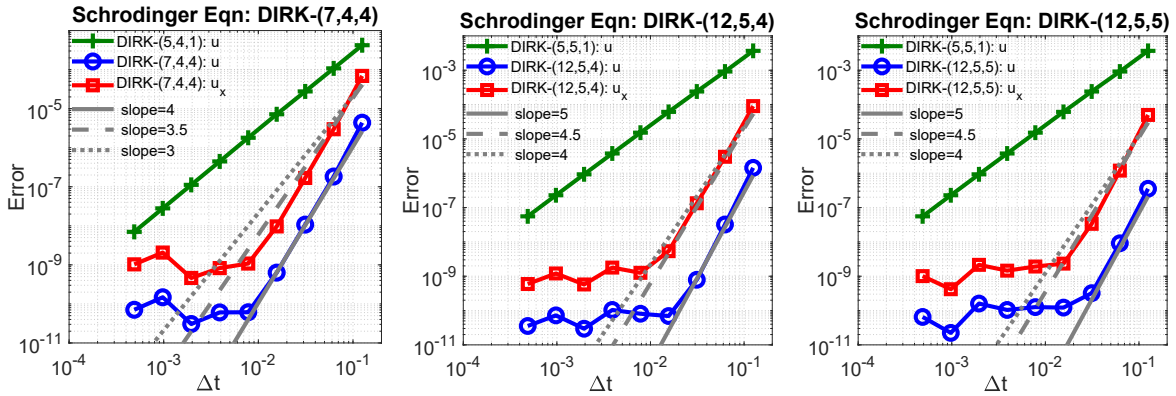


Figure 4. Convergence (u blue circles; u_x red squares) for Schrödinger equation using DIRK-(7,4,4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12,5,4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12,5,5): 5th order DIRK scheme with WSO 5 (right).

6.4. Advection-diffusion equation. This example demonstrates that DIRK schemes with high weak stage order avoid order reduction when applied to problems with physical boundary layers. We consider the 1D linear advection-diffusion equation

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

with the true solution $u(x, t) = \cos(5t) \sin(10x + 10)$, Dirichlet b.c., and viscosity $\nu = 10^{-3}$. The advection term dominates, and the outflow boundary condition at $x = 1$ leads to a physical boundary layer of width $\mathcal{O}(\nu)$. All spatial derivatives are approximated via 4th-order centered differences on a grid with 10^4 cells, and errors are evaluated at $T = 1$. The results shown in Figure 5 exhibit the expected convergence in u and u_x for DIRK-(7, 4, 4) (left), DIRK-(12, 5, 4) (middle), and DIRK-(12, 5, 5) (right). In particular, the results confirm that physical boundary layers do not interfere with the schemes' remedy of order reduction.

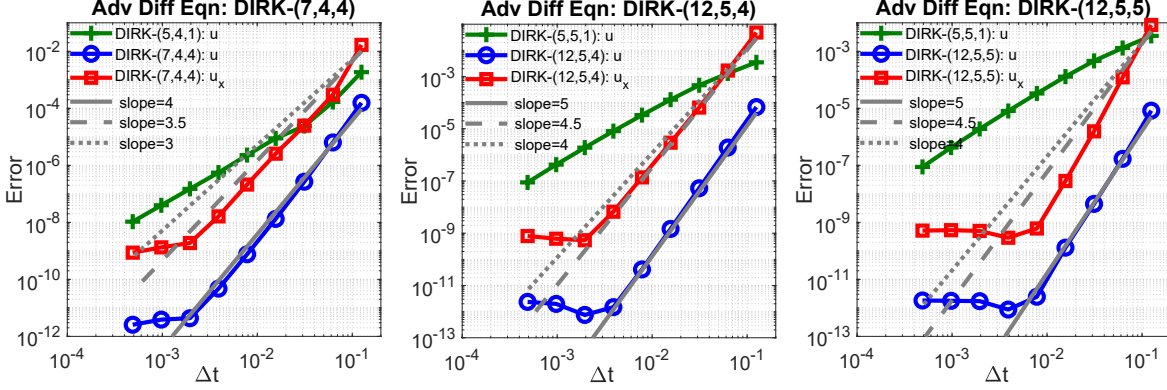


Figure 5. Convergence (u blue circles; u_x red squares) for the advection-diffusion equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

6.5. Linear advection equation. To illustrate order reduction, and its remedy, in problems with only first-order spatial derivatives, we consider

$$u_t + u_x = 0 \quad \text{for } (x, t) \in (0, 1) \times (0, 1],$$

with Dirichlet b.c. at $x = 0$, final time $T = 1$, and the true solution a traveling wave $u(x, t) = \sin(2\pi(x - t))$. Again, 4th-order centered differences with 10^4 cells are used to approximate ∂_x . For this first-order problem, the numerical boundary layer due to order reduction now is of thickness $\mathcal{O}(\Delta t)$, hence we expect a loss of a full order in u_x when $q < p$. This is demonstrated in Figure 6: DIRK-(12, 5, 4) recovers 5th order in u and 4th order in u_x . Moreover, DIRK-(7, 4, 4) and DIRK-(12, 5, 5) recover their full orders of convergence for both u and u_x . It is interesting to note that for this specific test problem, the reference schemes with $q = 1$ turn out to exhibit third-order convergence, instead of the expected second order. We do not have an explanation for this interesting behavior; however, note that this is not in contradiction to any of the theory.

6.6. Heat equation with spatially varying coefficient. The examples above are restricted to differential operators with constant coefficients. To demonstrate that our schemes remedy order reduction for more general problems, we consider the heat equation

$$u_t = (\kappa(x)u_x)_x + f \quad \text{for } (x, t) \in (0, 1) \times (0, 1], \quad u = g \quad \text{on } \{0, 1\} \times (0, 1],$$

with spatially varying diffusion coefficient $\kappa(x) = \cos(x + 0.1)$. The forcing $f(x, t)$, the b.c., and the i.c. are chosen so that the true solution is $u(x, t) = \cos(20t) \sin(10x + 10)$. We used 6th-order centered differences with 10^3 cells to approximate the spatial derivatives and the problem is solved till $T = 1$. Figure 7 confirms that the high WSO schemes recover the expected convergence orders, just as they did for the constant-coefficient heat equation.

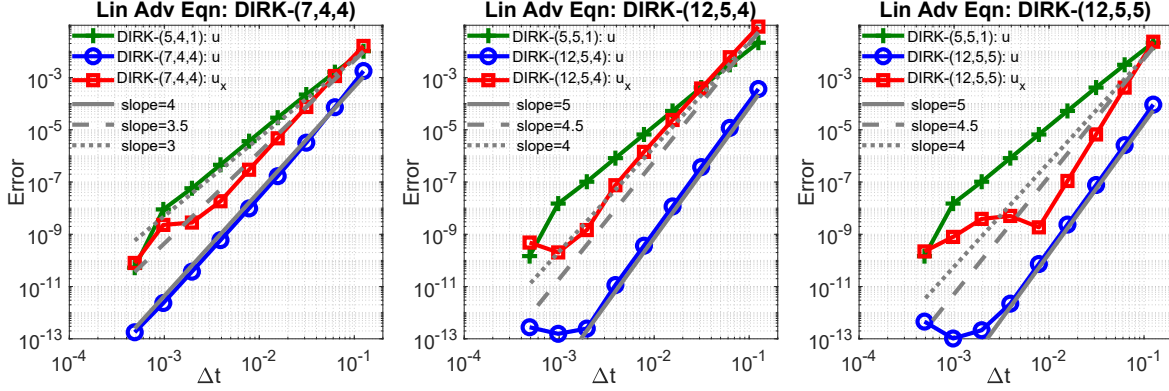


Figure 6. Convergence (u blue circles; u_x red squares) for the linear advection equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

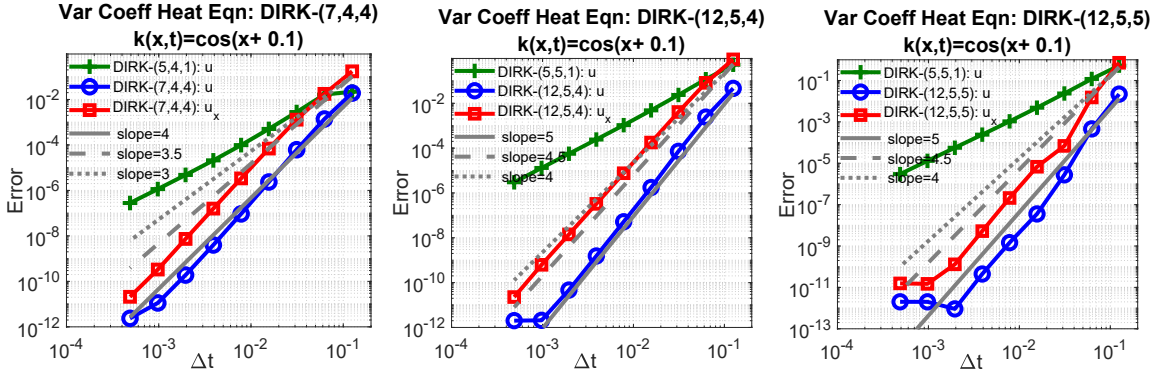


Figure 7. Convergence (u blue circles; u_x red squares) for heat equation with spatially varying diffusion coefficient $\kappa(x) = \cos(x + 0.1)$ using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

6.7. An equation with a fourth-order spatial derivative. To demonstrate that our schemes remove order reduction for PDEs with more than one boundary condition, we consider

$$u_t = -u_{xxxx} + f \text{ for } (x, t) \in (0, 1) \times (0, 1], \quad (30)$$

with both “Dirichlet” and “Neumann” boundary conditions on each side, i.e., $u(0) = g_0(t)$, $u(1) = g_1(t)$, $u_x(0) = h_0(t)$, $u_x(1) = h_1(t)$, and the forcing f such that the manufactured solution is $u(x, t) = \cos(15t)$. The final time is $T = 1$. In this equation, the 4th order spatial derivative is approximated by a 2nd-order centered finite difference on a fine grid of 10^4 cells.

The convergence results, obtained with the new high WSO DIRK schemes, as well as the reference WSO 1 schemes, are shown in Figure 8. The schemes with $q = p$ recover the full order of convergence for both u and u_x . The time-stepping schemes produce numerical boundary layers whose width scales like $\mathcal{O}(\Delta t^{\frac{1}{4}})$, leading to $\frac{1}{4}$ order loss per derivative for the schemes that have $q < p$.

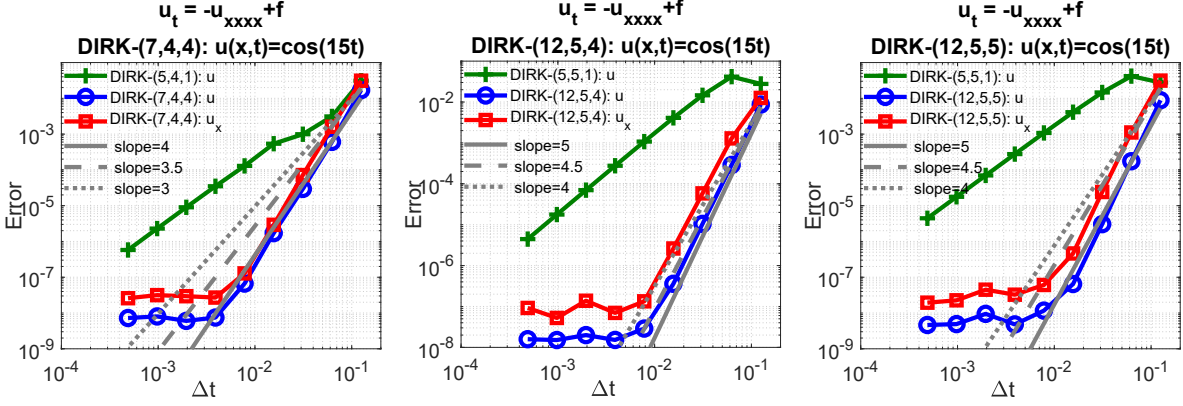


Figure 8. Error convergence (u blue; u_x red) for equation (30) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

6.8. Two-dimensional linear advection-diffusion equation. In principle, the presence of corners (non-smooth domain boundaries) may be an additional source of error that could lead to order reduction. Here we examine a two-dimensional PDE problem in a square domain and demonstrate that our schemes remedy order reduction also in this setting. We consider the advection-diffusion equation

$$u_t + u_x + u_y = \nu(u_{xx} + u_{yy}) + f(x, y), \quad (x, y) \in [-1, 1]^2,$$

with $\nu = 0.1$, the forcing f , the boundary conditions and initial condition chosen such that the manufactured solution is $u(x, y, t) = \exp(-\frac{\pi^2}{8}t) \sin(\pi x + \frac{\pi}{4}) \sin(\pi y + \frac{\pi}{4})$.

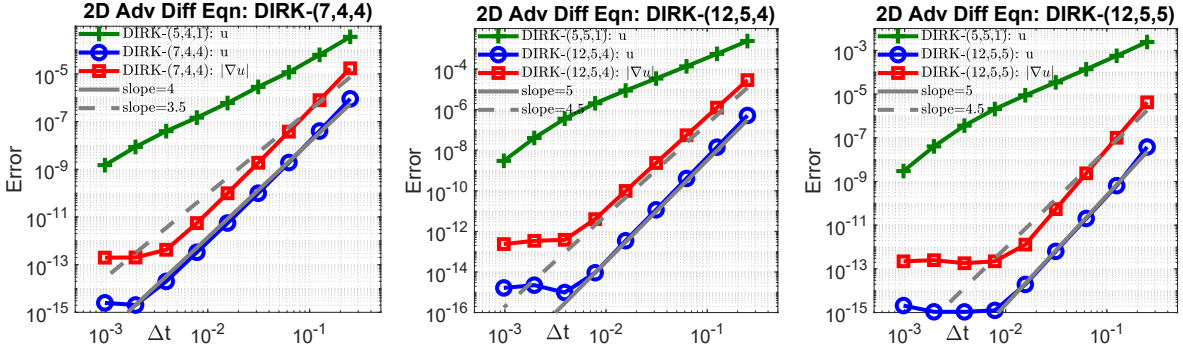


Figure 9. Error convergence (u blue; $|\nabla u|$ red) for a two-dimensional linear advection equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

To ensure the spatial error is negligible, we use a spectral method on a 2D tensor-product grid with 30 Chebyshev points in each direction. We solve the problem up to $T = 1$. Errors are plotted in Figure 9. The schemes with high WSO successfully avoid order reduction for this problem.

7. NUMERICAL RESULTS: TIME-DEPENDENT LINEAR AND NONLINEAR OPERATORS

The weak stage order conditions (see §2) are derived based on a linear problem in which the coefficient of the linear term is time-independent. Here we explore the question whether these conditions are also sufficient to alleviate order reduction for more general problems.

7.1. Heat equation with temporally varying coefficient. We revisit the variable-coefficient heat equation of §6.6, but now allowing κ to vary also in time:

$$u_t = (\kappa(x, t)u_x)_x + f \quad \text{for } (x, t) \in (0, 1) \times (0, 1], \quad u = g \quad \text{on } \{0, 1\} \times (0, 1].$$

We consider two different diffusion coefficient functions, one of which varies slowly in time, $\kappa(x, t) = \cos(0.1t + 0.2)$, and another that oscillates rapidly in time, $\kappa(x, t) = 1 + 0.5 \cos(30t + 0.1)$. In both cases, the spatial derivatives are approximated using 6th-order centered differences with 10^3 cells, and the errors are evaluated at time $T = 1$. Figure 10 shows that all high WSO schemes practically alleviate order reduction for the slowly-varying coefficient case. In contrast, for the rapidly-varying coefficient case, the schemes suffer from order reduction and do not produce clean high-order convergence results. That being said, the new high WSO schemes do turn out to yield smaller errors than the WSO-1 reference methods.

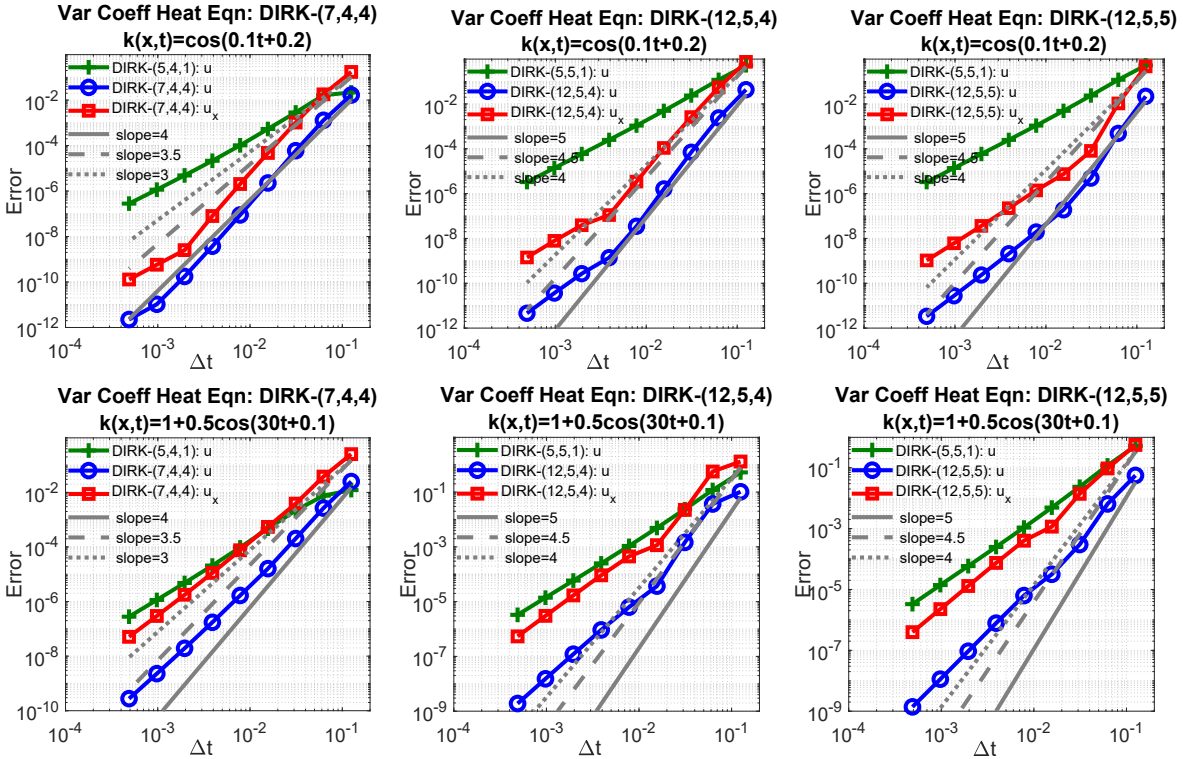


Figure 10. Convergence (u blue circles; u_x red squares) for heat equation with diffusion coefficient $\kappa(x, t) = \cos(0.1t + 0.2)$ (top) and $\kappa(x, t) = 1 + 0.5 \cos(20t)$ (bottom) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(12, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (right).

7.2. Stiff nonlinear PDE: viscous Burgers' equation. As a stiff nonlinear PDE problem, we study the viscous Burgers' equation,

$$u_t + uu_x = \nu u_{xx} + f \quad \text{for } (x, t) \in (0, 1) \times (0, 1], \quad u = g \quad \text{on } \{0, 1\} \times (0, 1], \quad (31)$$

with the true solution $u(x, t) = \cos(t)$, and the viscosity constant $\nu = 0.1$ (i.e., the main source of stiffness are the differential operators themselves). Here we choose a particularly simple manufactured solution to demonstrate that high weak stage order schemes do not fully remedy order reduction for nonlinear problems. However, we observe that these schemes still perform better than schemes with WSO 1 in terms of accuracy and convergence order. This is important in the context of the demonstration in [20] that DIRK schemes with WSO up to 3 can exhibit clean and full order of convergence, even though problem (31) is nonlinear and thus outside the class of (linear) problems for which WSO is known to improve the accuracy of the LTE.

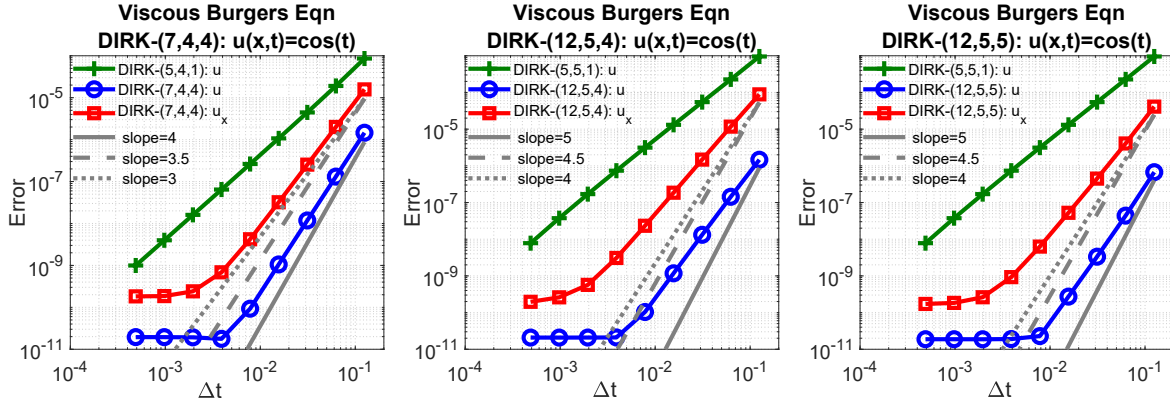


Figure 11. Convergence for the viscous Burgers' equation using DIRK-(7, 4, 4), DIRK-(12, 5, 4), and DIRK-(12, 5, 5).

Again, 6th-order centered differences with 10^3 cells are used to approximate the spatial derivatives, and the errors are evaluated at time $T = 1$. Figure 11 shows the convergence results for our new DIRK schemes with WSO 4 and 5 for the same test problem. We see that the high WSO schemes indeed turn out to generate a convergence order of 3, which is better than what schemes with WSO 1 achieve, but there remains a reduction of order for the schemes of order above 3 considered here.

7.3. Stiff nonlinear ODE: Van der Pol oscillator. To demonstrate that DIRK schemes with high weak stage order do not remedy order reduction for all types of problems, we consider, as a key benchmark example for stiff nonlinear ODE, the Van der Pol oscillator,

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = \mu(1 - x^2)y - x,$$

with stiffness parameter $\mu = 500$, initial condition $(x(0), y(0)) = (2, 0)$, and final time $T = 10$. For a range of time steps from $\Delta t = 0.5$ to $\Delta t \approx 2.44 \times 10^{-4}$, different DIRK schemes are applied, with Newton's method used to solve the nonlinear problems up to machine precision. The reference solution is calculated via the standard explicit RK4 method with time step $\Delta t = 10^{-6}$. Figure 12 shows the convergence for DIRK-(7, 4, 4) and DIRK-(12, 5, 5), clearly indicating that the high WSO in the DIRK schemes does not suffice to remove order reduction in the stiff regime ($10^{-2} < \Delta t < 10^{-1}$).

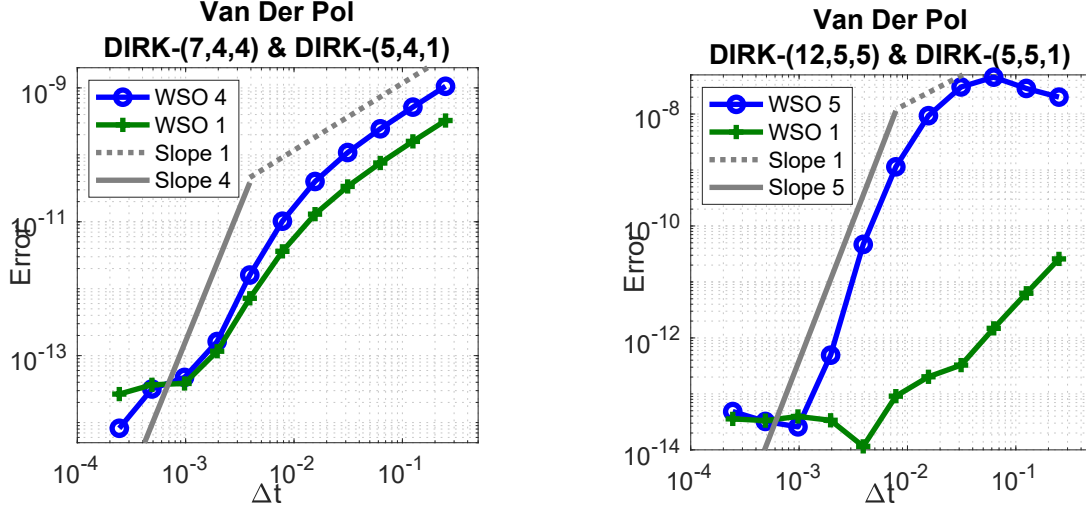


Figure 12. Convergence for the Van der Pol oscillator using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (blue circles, left), and DIRK-(12, 5, 5): 5th order DIRK scheme with WSO 5 (blue circles, right).

8. CONCLUSIONS AND OUTLOOK

The results of this work can be seen as a reinforcement of the usefulness of weak stage order, which can remove order reduction in Runge-Kutta schemes applied to linear problems with time-independent operators. A key theoretical contribution of this paper is that it has been shown that WSO can indeed be extended beyond WSO 3 (which is important because a special case of WSO had previously been shown to be limited to WSO 3 [20]). Moreover, utilizing a general theory of WSO [7], three concrete new DIRK schemes, DIRK-(7, 4, 4), DIRK-(12, 5, 4), and DIRK-(12, 5, 5), have been constructed with high WSO and other desirable properties: stiff accuracy, L-stability, and optimized error constants. These new schemes have been demonstrated to be practically useful, as they successfully address the order reduction problem in a variety of test problems, both those covered by the theory (linear problems with time-independent operators), as well as some (but not all) problems beyond the scope of the theory.

Because the new schemes have the standard form of RK methods, they can be easily incorporated into existing software and thus may be immediately useful for practitioners who seek to remedy order reduction while using DIRK time stepping.

The results presented here give rise to several further questions and research directions. First, problem (M) in §5 that characterizes optimal DIRK schemes is a polynomial optimization problem. While successfully solved via generic approaches herein, tailored modern optimization methods that yield provably globally optimal solutions represent a natural next step. Second, the number of stages used by the schemes provided here (7 and 12 stages, respectively) is larger than the theoretical minimum number of stages implied by the theorems we provide. Both sharp bounds and concrete DIRK schemes that realize the minimum number of stages remain to be found. Third, specific explorations of WSO for EDIRKs ($a_{11} = 0$), SDIRKs (all a_{ii} identical), and also explicit RK schemes, remain practically relevant open directions of research.

APPENDIX A. LIST OF NEW DIRK SCHEMES

See Table 3 for the new DIRK schemes with high weak stage order.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. DMS-2012271 (Biswas, Seibold), No. DMS-1952878 (Seibold), and No. DMS-2012268 (Shirokoff).

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.
- [2] P. Albrecht. A new theoretical approach to Runge–Kutta methods. *SIAM J. Numer. Anal.*, 24:391–406, 1987.
- [3] I. Alonso-Mallo. Runge–Kutta methods without order reduction for linear initial boundary value problems. *Numer. Math.*, 91(4):577–603, 2002.
- [4] 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.
- [5] I. Alonso-Mallo and C. Palencia. Optimal orders of convergence for Runge–Kutta methods and linear, initial boundary value problems. *Appl. Numer. Math.*, 44(1-2):1–19, 2003.
- [6] S. Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer, third edition, 2015.
- [7] A. Biswas, D. Ketcheson, B. Seibold, and D. Shirokoff. Algebraic structure of the weak stage order conditions for Runge–Kutta methods, 2022. Under review.
- [8] Abhijit Biswas. *Structure-Optimized and Order-Preserving Methods for Advection and Time Stepping*. PhD thesis, Temple University, 2021.
- [9] K. Burrage and L. Petzold. On order reduction for Runge–Kutta methods applied to differential/algebraic systems and to stiff systems of ODEs. *SIAM J. Numer. Anal.*, 27(2):447–456, 1990.
- [10] J. C. Butcher and N. Goodwin. *Numerical methods for ordinary differential equations*, volume 2. Wiley Online Library, 2008.
- [11] M.P. Calvo and C. Palencia. Avoiding the order reduction of Runge–Kutta methods for linear initial boundary value problems. *Math. Comp.*, 71(240):1529–1543, 2002.
- [12] 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.
- [13] M. Crouzeix. *Sur l’approximation des équations différentielles opérationnelles linéaires par des méthodes de Runge–Kutta*. PhD thesis, Université de Paris VI Thèse, 1975.
- [14] M. Crouzeix. Une méthode multipas implicite-explicite pour l’approximation des équations d’évolution paraboliques. *Numer. Math.*, 35(3):257–276, 1980.
- [15] M. Crouzeix and P.-A. Raviart. Approximation des problèmes d’évolution. *Lecture Notes, University Rennes*, 1980.
- [16] 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.
- [17] E. Hairer and G. Wanner. *Solving ordinary differential equations II*. Springer Berlin Heidelberg, 1996.
- [18] WH Hundsdorfer and MN Spijker. A note on B-stability of Runge–Kutta methods. *Numer. Math.*, 36(3):319–331, 1980.
- [19] C. A. Kennedy and M. H. Carpenter. *Diagonally Implicit Runge–Kutta methods for ordinary differential equations, a review*. NASA, Langley Research Center, 2016.
- [20] D. Ketcheson, B. Seibold, D. Shirokoff, and D. Zhou. DIRK schemes with high weak stage order. In *Lecture Notes in Computational Science and Engineering*, pages 453–463. Proceedings of ICOSAHOM 2018, Springer, 2020.
- [21] Serge Lang. *Algebra*, volume 1 of *Graduate Texts in Mathematics*. Springer, revised third edition, 2002.

) (middle), and DIRK-

[illegible]

- [22] C. Lubich. On the convergence of multistep methods for nonlinear stiff differential equations. *Numer. Math.*, 58(1):839–853, 1990.
- [23] C. Lubich and A. Ostermann. Runge-Kutta methods for parabolic equations and convolution quadrature. *Math. Comp.*, 60(201):105–131, 1993.
- [24] C. Lubich and A. Ostermann. Runge-Kutta approximation of quasi-linear parabolic equations. *Math. Comp.*, 64(210):601–627, 1995.
- [25] A. Ostermann and M. Roche. Runge-Kutta methods for partial differential equations and fractional orders of convergence. *Math. Comp.*, 59(200):403–420, 1992.
- [26] 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.
- [27] 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.
- [28] A. Prothero and A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.*, 28(125):145–162, 1974.
- [29] J. Rang. An analysis of the Prothero–Robinson example for constructing new DIRK and ROW methods. *J. Comput. Appl. Math.*, 262:105–114, 2014.
- [30] S. Roberts. *Multimethods for the Efficient Solution of Multiscale Differential Equations*. PhD thesis, Virginia Polytechnic Institute and State University, 2021.
- [31] S. Roberts and A. Sandu. Eliminating order reduction on linear, time-dependent ODEs with GARK methods, 2022. arxiv.org/abs/2201.07940.
- [32] R. R. Rosales, B. Seibold, D. Shirokoff, and D. Zhou. Spatial manifestations of order reduction in Runge-Kutta methods for initial boundary value problems. *arXiv preprint arXiv:1712.00897*, 2017.
- [33] 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.
- [34] S. Scholz. Order barriers for the B-convergence of ROW methods. *Computing*, 41(3):219–235, 1989.
- [35] P. W. Sharp and E. Smart. Explicit Runge–Kutta pairs with one more derivative evaluation than the minimum. *SIAM J. Sci. Comput.*, 14(2):338–348, 1993.
- [36] J. G. Verwer. Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines. In *Centre Math. Comput. Sci.*, number NM-R 8506 in CWI, 1985.

(Abhijit Biswas) COMPUTER, ELECTRICAL, AND MATHEMATICAL SCIENCES & ENGINEERING DIVISION, KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, THUWAL 23955, SAUDI ARABIA

Email address: abhijit.biswas@kaust.edu.sa

URL: <https://math.temple.edu/~tug14809>

(David Ketcheson) COMPUTER, ELECTRICAL, AND MATHEMATICAL SCIENCES & ENGINEERING DIVISION, KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, THUWAL 23955, SAUDI ARABIA

Email address: david.ketcheson@kaust.edu.sa

URL: <https://www.davidketcheson.info>

(Benjamin Seibold) DEPARTMENT OF MATHEMATICS, TEMPLE UNIVERSITY, 1805 NORTH BROAD STREET, PHILADELPHIA, PA 19122

Email address: seibold@temple.edu

URL: <http://www.math.temple.edu/~seibold>

(David Shirokoff) CORRESPONDING AUTHOR, DEPARTMENT OF MATHEMATICAL SCIENCES, NEW JERSEY INSTITUTE OF TECHNOLOGY, UNIVERSITY HEIGHTS, NEWARK, NJ 07102

Email address: david.g.shirokoff@njit.edu

URL: <https://web.njit.edu/~shirokof>