

## ALGEBRAIC STRUCTURE OF THE WEAK STAGE ORDER CONDITIONS FOR RUNGE–KUTTA METHODS\*

ABHIJIT BISWAS<sup>†</sup>, DAVID KETCHESON<sup>†</sup>, BENJAMIN SEIBOLD<sup>‡</sup>,  
AND DAVID SHIROKOFF<sup>§</sup>

**Abstract.** Runge–Kutta (RK) methods may exhibit order reduction when applied to stiff problems. For linear problems with time-independent operators, order reduction can be avoided if the method satisfies certain weak stage order (WSO) conditions, which are less restrictive than traditional stage order conditions. This paper outlines the first algebraic theory of WSO, and establishes general order barriers that relate the WSO of a RK scheme to its order and number of stages for both fully-implicit and DIRK schemes. It is shown in several scenarios that the constructed bounds are sharp. The theory characterizes WSO in terms of orthogonal invariant subspaces and associated minimal polynomials. The resulting necessary conditions on the structure of RK methods with WSO are then shown to be of practical use for the construction of such schemes.

**Key words.** weak stage order, Runge–Kutta, order-reduction, B-convergence, DIRK methods

**MSC codes.** 65L04, 65L20, 65M12

**DOI.** 10.1137/22M1483943

**1. Introduction.** Runge–Kutta (RK) methods may exhibit a convergence rate lower than the (classical) order of the scheme. This *order reduction* phenomenon often occurs when integrating stiff ODEs, stiff PDEs, or initial-boundary value problems (IBVPs) with time-dependent boundary conditions [9, 13, 39, 42, 27, 24, 29]. In the worst case, the observed convergence rate is governed by the *stage order* of the method [32, 17]. Unfortunately, RK methods with high stage order must be fully implicit; diagonally implicit RK (DIRK) schemes cannot have stage order above 2.

By considering a restricted class of problems, one can obtain conditions weaker than stage order that are sufficient to ensure high-order convergence. For instance, in some cases where order reduction stems from time-dependent PDE boundary conditions, it can be avoided by various problem- and method-dependent modifications [12, 1, 2, 11, 3]. A set of conditions ensuring high-order convergence for linear problems has been developed [27, 19, 8, 40] and recently referred to by the term *weak stage order (WSO)* [37, 21]. The conditions for WSO are, as the name suggests, weaker than those for stage order; notably, DIRK methods can have high WSO. The conditions are closely related to the classical order conditions for RK methods. In this work we

\*Received by the editors April 7, 2022; accepted for publication (in revised form) July 25, 2023; published electronically January 4, 2024.

<https://doi.org/10.1137/22M1483943>

**Funding:** This material is based upon work supported by the National Science Foundation under grants DMS-2012271 (first and third authors), DMS-1952878 (third author), and DMS-2012268 (fourth author).

<sup>†</sup>Computer, Electrical, and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia (abhijit.biswas@kaust.edu.sa, david.ketcheson@kaust.edu.sa).

<sup>‡</sup>Department of Mathematics, Temple University, Philadelphia, PA 19122 USA (seibold@temple.edu).

<sup>§</sup>Corresponding author. Department of Mathematical Sciences, NJIT, Newark, NJ 07102-1982 USA (david.g.shirokoff@njit.edu).

establish the algebraic structure of WSO and use it to prove rigorous relationships between the achievable order, the WSO, and the number of stages of a method.

**1.1. The WSO conditions.** We consider the initial value problem

$$(1.1) \quad u'(t) = f(t, u(t)), \quad u(0) = u_0; \quad u \in \mathbb{R}^m, \quad f: \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m,$$

and its numerical approximation  $u_n \approx u(t_n)$ , where  $t_n = n\Delta t$ , via an  $s$ -stage RK scheme, where the one-step update rule

$$(1.2a) \quad g_i = u_n + \Delta t \sum_{j=1}^s a_{ij} f(t_n + c_j \Delta t, g_j), \quad i = 1, 2, \dots, s,$$

$$(1.2b) \quad u_{n+1} = u_n + \Delta t \sum_{j=1}^s b_j f(t_n + c_j \Delta t, g_j),$$

is concisely represented via the Butcher tableau

$$\begin{array}{c|c} \vec{c} & A \\ \hline & \vec{b}^T \end{array}, \quad \text{where } A = (a_{ij})_{i,j=1}^s, \quad \vec{b} = (b_1, \dots, b_s)^T, \quad \vec{c} = (c_1, \dots, c_s)^T.$$

Method (1.2) is said to be of classical order  $p$  if the one-step error [16] is of  $\mathcal{O}(\Delta t^{p+1})$  as  $\Delta t \rightarrow 0$ . This requires that the coefficients of the method satisfy certain algebraic relations known as order conditions, which are derived under the assumption that  $f$  has Lipschitz constant  $\lambda$ , and that  $\Delta t \lambda \ll 1$ .

If one does not assume that  $\Delta t \lambda$  is small (i.e., in the study of stiff problems), a careful expansion of the numerical error leads to additional algebraic conditions on the scheme's coefficients. A necessary condition for convergence of order  $\xi$  for general stiff ODEs (cf. [17, Theorem 15.3]) is that

$$(1.3) \quad \vec{\tau}^{(k)} := A \vec{c}^{k-1} - \frac{1}{k} \vec{c}^k = \left( A C^{k-1} - \frac{1}{k} C^k \right) \vec{e} = 0 \quad \text{for } 1 \leq k \leq \xi.$$

Here  $\vec{c}^k := (c_1^k, c_2^k, \dots, c_s^k)^T$  denotes componentwise exponentiation, the vector  $\vec{e} = (1, \dots, 1)^T \in \mathbb{R}^s$ , and the diagonal matrix  $C = \text{diag}(c_1, c_2, \dots, c_s)$  so that  $\vec{c} = C \vec{e}$ . Throughout this paper we assume that  $\vec{c} = A \vec{e}$ , which implies that  $\tau^{(1)} = 0$ . We refer to the vectors  $\vec{\tau}^{(k)}$  as *stage order residuals*.

If one assumes that  $f$  is a linear (possibly time-dependent) operator, then in the convergence analysis for stiff problems one can replace the stage order with the largest integer  $\xi$  such that

$$(1.4) \quad S(\xi): \quad \vec{b}^T A^j \vec{\tau}^{(k)} = 0 \quad \text{for all } 0 \leq j \leq s-1, 1 \leq k \leq \xi.$$

**DEFINITION 1.1 (WSO).** *The WSO  $q$  of an  $s$ -stage RK scheme  $(A, \vec{b})$  is the largest integer such that condition  $S(q)$  in (1.4) holds. If  $S(\xi)$  holds for every  $\xi \geq 1$ , then  $q = \infty$ .*

Conditions (1.4) were first introduced in the context of Rosenbrock W- (ROW) methods, and referred to as *parabolic order conditions* [27, 28]; see section 2.4. In section 2, we review the derivation of condition (1.4) as well as what is known about the class of problems for which this condition ensures convergence of order  $q$ . We also provide a review of the literature relating to WSO and similar conditions.

**1.2. Contribution of this work.** While much is known about RK order conditions, less is known about the solvability of the WSO conditions, and more generally, the simultaneous solution of both sets of order conditions. Since high stage order methods must be fully implicit, the solvability of the WSO and order conditions for DIRK schemes is of particular interest.

Previous attempts at constructing DIRK schemes with WSO [37, 21] relied upon the simplifying assumption that the vectors  $\tau^{(k)}$  in (1.4) be eigenvectors of  $A$ . For instance, DIRK schemes up to order 4 and WSO 3 have been provided in [21]. This simplifying “eigenvector criterion” limits the construction of WSO schemes for DIRKs (with invertible  $A$ ) to  $q \leq 3$ , as shown below. A key motivating factor for the present work is the need to construct DIRK schemes with high WSO  $q > 3$ ; in order to enable this, herein the vectors  $\tau^{(k)}$  are not assumed to be eigenvectors of  $A$ . Although we are particularly interested in DIRK schemes, some of the main results herein apply to general RK schemes.

This paper provides the first known bounds relating a RK scheme’s WSO  $q$  to the order  $p$  and number of stages  $s$ . After reviewing the background on WSO (section 2), we recast WSO in terms of orthogonal invariant subspaces related to the matrix  $A$  (section 3). From there we establish general bounds on  $s$ ,  $q$ ,  $p$  for both fully implicit RK schemes and refinements for DIRKs. In section 4 we introduce minimal polynomials corresponding to the orthogonal invariant subspaces. The minimal polynomials are then combined with a family of polynomials (orthogonal with respect to a linear functional) to yield new formulas (section 5) for the stability function, relevant for schemes with high WSO. The study also reveals necessary conditions for WSO  $q > 3$  (section 6), which provides the theoretical foundation for devising new schemes in the companion paper [7]. We conclude with examples (section 7) demonstrating the sharpness of the bounds.

**2. Relevance of WSO.** This section provides background information relevant for this work: mathematical notation and fundamentals (section 2.1), the manifestations of order reduction and error formulas (section 2.2), and a literature review of prior work related to WSO (section 2.3). While the expert reader familiar with the literature may skip this section, it provides a motivation for the relevance of this paper.

**2.1. Fundamentals and notation.** The following definitions and notation about RK schemes are used in this manuscript.

Given a RK method, we let  $n_c$  denote the number of distinct abscissas  $c_j$ , e.g., if  $c_1 = \dots = c_s$ , then  $n_c = 1$ ; in turn, if all  $c_j$  are distinct (i.e., the scheme is nonconfluent), then  $n_c = s$ .

Schemes for which  $A$  is lower-triangular ( $a_{ij} = 0$  for  $j > i$ ) are called *diagonally implicit Runge–Kutta* (DIRK) methods. If, furthermore, all diagonal entries are equal, i.e.,  $a_{ii} = \gamma$  for  $i = 1, \dots, s$ , then the scheme is called *singly diagonally implicit* (SDIRK). We refer to DIRK methods with  $a_{11} = 0$  as EDIRKs. Implicit schemes that are not diagonally implicit are referred to as *fully implicit*.

The following definition generalizes the concept of an EDIRK, and will be useful in providing some results below (such as Theorem 3.2) that are less strict for certain schemes with an explicit stage.

**DEFINITION 2.1 (GEDIRK).** *We call a DIRK scheme a generalized EDIRK (or GEDIRK) if  $\vec{c}$  contains at least one zero and  $a_{\ell\ell} = 0$ , where  $\ell = \min\{j \mid c_j = 0\}$  is the index of the first zero in  $\vec{c}$ .*

When a RK scheme is applied to the scalar ODE  $u'(t) = \lambda u$ , it results in the iteration  $u_{n+1} = R(z)u_n$ , where  $z := \lambda \Delta t$ , and

$$(2.1) \quad R(z) := 1 + z \vec{b}^T (I - zA)^{-1} \vec{e} = \frac{\det(I - zA + z \vec{e} \vec{b}^T)}{\det(I - zA)}$$

is the *stability function* of the scheme.

A RK scheme is called *stiffly accurate* [32] if the last row of  $A$  equals  $\vec{b}^T$ , i.e.,  $a_{sj} = b_j$  for  $j = 1, \dots, s$ . When  $A$  is invertible, stiff accuracy implies  $\vec{b}^T A^{-1} \vec{e} = 1$ , and hence the desirable property  $\lim_{z \rightarrow \infty} R(z) = 0$  results.

The (*classical*) *order*  $p$  of a RK scheme (see section 1.1) imposes the well-known *order conditions* on the Butcher coefficients [16]. We will also make use of the conditions

$$(2.2) \quad B(\xi) : \quad \vec{b}^T \vec{c}^{k-1} = \frac{1}{k} \quad \text{for } k = 1, 2, \dots, \xi,$$

$$(2.3) \quad C(\xi) : \quad \vec{\tau}^{(k)} = 0 \quad \text{for } k = 1, 2, \dots, \xi,$$

which determine the accuracy of the quadrature and subquadrature rules, respectively, on which the RK method is based [10].

A RK scheme is said to have *stage order*  $\tilde{q} := \min\{q_1, q_2\}$  where  $q_1, q_2$  are the largest integers such that  $B(q_1)$  and  $C(q_2)$  hold. Stage order  $\tilde{q}$  guarantees that each RK stage is accurate to order  $\tilde{q}$ , and that the method itself has order at least  $\tilde{q}$ .

**2.2. Order reduction and the role of WSO.** As the name suggests, the conditions for WSO  $q$  are a relaxation of those for stage order  $\tilde{q}$ , i.e.,  $\tilde{q} \leq q$  holds for any method. The relevance of WSO in the context of order reduction can be understood by examining the error for the (stiff) Prothero–Robinson problem [32, 17],

$$(2.4) \quad u' = \lambda(u - \phi(t)) + \phi'(t), \quad u(0) = \phi(0), \quad \text{with } \operatorname{Re}(\lambda) < 0,$$

with solution  $u(t) = \phi(t)$ . For solutions starting from a different initial value, the difference  $u(t) - \phi(t)$  decays exponentially to zero with time. Problem (2.4) can be made arbitrarily stiff by choosing  $|\lambda|$  large. As in [17, eq. (15.9), Chap. IV.15], one can derive exact formulas for the RK scheme's *global error* (with equal time-steps):

$$u_{n+1} - \phi(t_{n+1}) = (R(z))^{n+1} (u_0 - \phi(0)) + \sum_{j=0}^n (R(z))^{n-j} \vec{\delta}_{\Delta t}(t_j).$$

Here  $z = \lambda \Delta t$  and  $R(z)$  are defined as in (2.1). The *local error*  $\vec{\delta}_{\Delta t}(t_n)$  has, for a  $p$ th order method, the form

$$(2.5) \quad \vec{\delta}_{\Delta t}(t_n) = - \sum_{k \geq 1} \frac{(\Delta t)^k}{(k-1)!} W_k(z) \phi^{(k)}(t_n) + \mathcal{O}(\Delta t^{p+1}),$$

where  $\phi^{(k)}(t_n)$  is the  $k$ th derivative of  $\phi$  at  $t_n$  and  $W_k(z)$  is the function

$$W_k(z) := z \vec{b}^T (I - zA)^{-1} \vec{\tau}^{(k)} \quad \text{for } k \geq 1.$$

In the classical RK convergence theory one examines the nonstiff limit where both the time scale  $1/|\lambda|$  and the solution  $\phi(t)$  are fully resolved. If one assumes that  $\Delta t \rightarrow 0$  with  $z = \mathcal{O}(\Delta t)$ , then one can show that the local error  $\vec{\delta}_{\Delta t}(t_n) = \mathcal{O}(\Delta t^{p+1})$  by using a Neumann series for  $(I - zA)^{-1}$  and applying the order conditions.

In a stiff problem setting, one fully resolves the slow time scale defined by the solution  $\phi(t)$  (i.e.,  $\Delta t|\phi'(t)| \ll 1$ ) but under-resolves the fast time scale  $1/|\lambda|$  (i.e.,  $|z| = \Delta t|\lambda| > \mathcal{O}(1)$ ). One can then investigate the local error  $\tilde{\delta}_{\Delta t}(t_n)$  uniformly in  $z$ . As a special case this, in particular, covers the *stiff limit* given by  $\Delta t \rightarrow 0$  while simultaneously  $\lambda \rightarrow -\infty$  such that  $z \rightarrow -\infty$ . Then the local error  $\tilde{\delta}_{\Delta t}(t_n)$  is at least order  $\tilde{q} + 1$  because  $\tilde{\tau}^{(k)} = 0$  for  $k \leq \tilde{q}$ , but, in general, it will fail to be of order  $p + 1$ . The algebraic condition of WSO ensures that the functions  $W_k(z)$  vanish, yielding a local error  $\tilde{\delta}_{\Delta t}(t_n) = \mathcal{O}(\Delta t^{q+1})$ . WSO is the least restrictive condition to guarantee  $W_k(z) = 0$  for  $k \leq q$  [21, Thm. 2].

While the Prothero–Robinson problem (2.4) provides an intuitive example where the convergence of the local truncation error is facilitated by WSO, both the order reduction phenomenon and its remedy via WSO apply much more generally.

A general class of problems for which WSO improves the convergence order are effectively established by the results in [4]: consider any linear PDE

$$(2.6) \quad u_t = \mathcal{L}u + f(t), \quad \text{with boundary data } \mathcal{B}u = g(t),$$

and suitable initial data, where  $(\mathcal{L}, \mathcal{B})$  generate a  $C_0$ -semigroup. While the theorems in [4] are stated assuming that the RK scheme used to approximate (2.6) has stage order  $q$ , the proofs, in fact only, use the conditions for WSO  $q$ . Hence, WSO is of potential relevance for any PDE problem generated by a  $C_0$ -semigroup. The results in [4] build on previous work that focused on analytic semigroups [27, 28], extending their applicability to a broader class of PDEs—notably wave equations.

For PDE problems, order reduction and WSO may manifest in interesting ways that are not encountered in the ODE problem (2.4). For instance, the convergence order is controlled by the WSO of the method, plus a correction that may be fractional depending on the details of the PDE [27]. The fractional convergence can also be understood through the creation of spatial boundary layers in the RK error [37]. While WSO guarantees the remedy of order reduction only for linear problems, it is also observed numerically that WSO can improve the convergence order for certain nonlinear PDEs where the highest derivatives are linear [21].

**2.3. Prior work on avoiding order reduction.** The prior section highlights that for certain classes of stiff problems the order of the local error may be reduced, in the worst case, to the WSO of the method.

Ostermann and Roche [27] showed that, for linear problems, time-stepping schemes that satisfy

$$(2.7) \quad \tilde{W}_k(z) := \frac{k \tilde{b}^T (I - zA)^{-1} \tilde{\tau}^{(k)}}{R(z) - 1} \equiv 0 \quad \text{for } k = 1, 2, \dots, q,$$

along with other technical assumptions, avoid order reduction when applied to linear parabolic PDEs (see [27] for RK methods and [28] for the related ROW methods). The analysis considers PDE problems (2.6) where  $\mathcal{L}$  defines an analytic semigroup [28] or has a point spectrum in the left half plane with a basis of eigenfunctions [27]. The resulting error convergence rate is  $\min\{p, q + 2 + \bar{\nu}\}$ , where  $\bar{\nu}$  depends on the PDE; see [27, Thm. 2].

Hundsdorfer showed that a generalization of (2.7) implies a convergence rate of  $\mathcal{O}(\Delta t^{q+1})$  for linear-implicit RK schemes applied to linear ODEs [19, Thm. 3.3].<sup>1</sup>

<sup>1</sup>See also [8, Thm. 3.3] which used  $\tilde{W}_k(z)$  to achieve improved convergence estimates.

Montijano showed that a necessary condition for convergence at order  $q$  on the problem (2.4) is  $b^T A^{-1} c^j = 1$  for  $0 \leq j \leq q$  [25, 14].

Because (2.7) contains the variable  $z$ , it is somewhat impractical as a means to design RK schemes. Scholz [40] converted (2.7) into direct “order conditions” on the scheme coefficients. By expanding terms like  $z \vec{b}^T (I - zA) \vec{\tau}^{(k)}$  in the variable  $H = z/(1 - z\gamma)$  (which is possible only for SDIRK methods), Scholz developed B-convergence order barriers (up to order 4) for ROW methods and demonstrated that the weaker conditions are compatible with ROW (SDIRK) structures by constructing schemes with up to  $p = q = 4$  in (2.7). The complexity of Scholz’s conditions for ROW methods were subsequently simplified by Ostermann and Roche [28, eq. (3.12’)] to (effectively) the conditions (1.4), where the  $A$  and  $\vec{\tau}^{(k)}$  in (1.4) are ROW analogues of the corresponding RK scheme quantities. For further discussion of order reduction in the context of stiff problems we refer to [6, Chapter 4], [14, Chapter 7], and [17, Chapter IV.15] and references therein.

Although originally derived only for SDIRK methods, the “order conditions” (1.4) are, in fact, necessary conditions for any RK scheme to avoid order reduction in (2.4); see section 2.2 and [33, eqs. (20)–(21)], [34, Thm. 3], [41], and also [37, 21]. A geometric interpretation of (1.4) is provided in [37]. Recently, [35, Chap. 6] and [36] extended (1.4) to generalized-structure additively partitioned Runge–Kutta (GARK) methods.

**3. Bounds on WSO via  $A$ -invariant subspaces.** We now state the main results of this paper.

**THEOREM 3.1** (main result). *An RK scheme with  $s$  stages,  $n_c$  distinct abscissas, order  $p$ , and WSO  $q$  satisfies the following:*

1. *If the abscissa values are all nonzero ( $c_j \neq 0$ ), then  $q \leq 2n_c - 1$ .*
2. *If some of the abscissa values are zero ( $c_j = 0$ ), then either (i)  $q \leq 2n_c - 2$ , or (ii)  $q = \infty$  in which case  $p = 1$ .*
3. *If  $q \leq 2n_c - 1$ , then*

$$q + \left\lfloor \frac{p+1+\sigma}{2} \right\rfloor \leq s + n_c,$$

where  $\sigma = 1$  for stiffly accurate methods with invertible  $A$ , and  $\sigma = 0$  otherwise.

A stricter bound can be obtained for DIRK schemes.

**THEOREM 3.2** (main result for DIRK). *Let a DIRK scheme be given with  $s$  stages,  $n_c$  distinct abscissas, order  $p$ , and WSO  $q \leq 2n_c - 1$ . Then*

$$\left\lfloor \frac{q + \kappa}{2} \right\rfloor - \kappa + p \leq s + 1 - \sigma,$$

where  $\sigma = 1$  if  $A$  is invertible and the scheme is stiffly accurate, and  $\sigma = 0$  otherwise, and  $\kappa = 1$  if  $A$  is a GEDIRK scheme, and  $\kappa = 0$  otherwise.

Here  $\lfloor x \rfloor$  is the standard floor function for the real number  $x$ . Note that Theorem 3.1 still applies when  $q \geq 2n_c - 1$ .

**Remark 3.3** (order barriers). When  $p = s + 1 - \sigma$ , Theorem 3.2 implies that DIRK schemes are limited to WSO  $q \leq 3$ , with the exception of GEDIRK schemes which are limited to WSO  $q \leq 4$ . For a fixed  $s$ , decreasing  $p$  by 1 increases the upper bound on  $q$  by 2. There exist schemes that satisfy these barriers sharply (see section 7).

In fact, the main results appearing here can be viewed as improvements to the classical bounds on the RK order  $p$  via the number of stages  $s$ . For instance, the classical bound  $p \leq s + 1$  for DIRK methods can be recovered by setting  $q = 0$  ( $\kappa = 0$ ) in Theorem 3.2. Moreover, the main results show that the maximum WSO  $q$  is controlled by the difference in the classical bound (e.g.,  $s + 1 - p$  for DIRKs).

**3.1. A pair of orthogonal  $A$ -invariant subspaces.** The WSO conditions (1.4) can be written equivalently in terms of orthogonal subspaces. Given an  $s$ -stage RK method, let

$$(3.1) \quad K_m := \text{span} \left\{ \vec{\tau}^{(1)}, A\vec{\tau}^{(1)}, \dots, A^{s-1}\vec{\tau}^{(1)}, \vec{\tau}^{(2)}, A\vec{\tau}^{(2)}, \dots, A^{s-1}\vec{\tau}^{(m)} \right\} \quad (m \geq 1),$$

$$(3.2) \quad Y := \text{span} \left\{ \vec{b}, A^T \vec{b}, \dots, (A^T)^{s-1} \vec{b} \right\}.$$

Note that due to the Cayley–Hamilton theorem,  $K_m$  is an  $A$ -invariant subspace, while  $Y$  is left  $A$ -invariant, i.e.,  $A\vec{v} \in K_m$  for any  $\vec{v} \in K_m$  and  $A^T \vec{v} \in Y$  for all  $\vec{v} \in Y$ . In fact,  $K_m$  and  $Y$  are the smallest  $A$ -invariant spaces containing the  $\vec{\tau}^{(j)}$  ( $j \leq m$ ) and  $\vec{b}$ , respectively. This allows for the natural generalization to  $m = \infty$  in (3.1) by defining  $K_\infty$  as the smallest  $A$ -invariant space containing  $\tau^{(k)}$  for all  $k \geq 1$ .

Then (1.4) is equivalent to the condition that  $\vec{v}^T \vec{w} = 0$  for every  $\vec{v} \in Y$  and  $\vec{w} \in K_q$ . Since  $K_m$  and  $Y$  are subspaces of  $\mathbb{R}^s$ , we have the following lemma.

**LEMMA 3.4.** *Given an RK scheme defined by  $A, \vec{b}$ , let  $q \geq 1$  be an integer or  $\infty$ , and let  $K_q$  and  $Y$  be the subspaces (3.1) and (3.2) of  $\mathbb{R}^s$ . Then the scheme has WSO of at least  $q$  if and only if  $Y$  and  $K_q$  are orthogonal, in which case*

$$(3.3) \quad \dim(Y) + \dim(K_q) \leq s.$$

RK methods with (classical) stage order  $q$  have  $\dim(K_q) = 0$  and  $\dim(K_{q+1}) > 0$ . Except for schemes that were specifically designed with high WSO, almost all irreducible schemes in the literature have  $\dim(Y) = s$ , so Lemma 3.4 implies that their WSO is no greater than their stage order. However, some existing schemes have  $\dim(Y) < s$ ; for example, explicit methods with one or more entries on the first subdiagonal equal to zero. These include the 8th order method of Prince and Dormand [31], as well as some of the fifth-order SSP methods of Ruuth and Spiteri [38].

**Remark 3.5.** In the proof of Theorem 3.2 (to follow) the bound (3.7) for DIRK schemes is used. However, if  $p = 1$ , (3.6) for fully implicit schemes gives  $\dim Y \geq 1$  which is tighter than the DIRK bound (3.7) ( $\dim Y \geq 0$ ) and yields a slight improvement to Theorem 3.2.

**Remark 3.6.** For simplicity we have written “order  $p$ ” in Theorems 3.1 and 3.2; however, both results hold under the weaker condition that  $R(z) = e^z + \mathcal{O}(z^{p+1})$  is a  $p$ th order approximation as  $z \rightarrow 0$ .

The proof of the main results follows from the inequality (3.3) along with lower bounds on the dimensions of  $Y$  and  $K_q$ . The proof can be divided into three steps. Step 1 (section 3.2) bounds the dimension of  $Y$  for both fully implicit and DIRK schemes. Step 2 (section 3.3) bounds the dimension of  $K_m$  for fully implicit schemes which is then refined in Step 3 (section 3.4) for DIRK schemes.

**3.2. (Step 1) A lower bound on the dimension of the subspace  $Y$ .** Here we show that the degrees of the numerator and the denominator of the RK stability function are bound by the dimension of the subspace  $Y$ . When combined with classical

results on rational approximations to  $e^z$ , we obtain a lower bound on  $\dim Y$  in terms of  $p$ .

Throughout, we denote a rational function that approximates  $e^z$  to order  $p$  as

$$(3.4) \quad N(z)/D(z) = e^z + \mathcal{O}(z^{p+1}), \quad \text{as } z \rightarrow 0,$$

where  $N$  and  $D$  are polynomials with no common factors.

**THEOREM 3.7** (rational approximations of  $e^z$ , [17, Thms. IV.3.11, and IV.4.18], [26, Thm. 2.1]). *Let polynomials  $N(z), D(z)$  satisfy (3.4). Then  $p \leq \deg N + \deg D$ . If, further,  $N/D$  has only real poles, then  $p \leq \deg N + 1$ .*

For an  $s$ -stage RK method, the stability function is a rational approximation of  $e^z$  with numerator and denominator of degree at most  $s$ . In fact, we can replace  $s$  in this statement with  $\dim Y$ .

**THEOREM 3.8.** *Let an RK method  $(A, b)$  be given, with  $Y$  defined as in (3.2). Then the stability function of the method is  $R(z) = N(z)/D(z)$ , where  $\deg N \leq \dim Y$  and  $\deg D \leq \dim Y$ .*

*Proof.* Define an orthogonal matrix  $L = (U \mid V) \in \mathbb{R}^{s \times s}$ , where the columns of  $U \in \mathbb{R}^{s \times d}$ , with  $d := \dim Y$ , are a basis of  $Y$  and the columns of  $V$  are a basis of its orthogonal complement. Then  $\vec{b}^T V = \vec{0}$  and  $\vec{b}^T (I - zA)^{-1} V = \vec{0}$ .

So the stability function (2.1) is

$$(3.5) \quad R(z) = 1 + z \vec{b}^T L L^T (I - zA)^{-1} L L^T \vec{e} = 1 + z \vec{b}_U^T M U^T \vec{e},$$

where  $\vec{b}_U^T = \vec{b}^T U \in \mathbb{R}^d$  and  $M = U^T (I - zA)^{-1} U \in \mathbb{R}^{d \times d}$ . Using this and Cramer's rule it can be shown that the numerator and denominator of  $R(z)$  are given by determinants of  $d \times d$  matrices whose entries are linear functions of  $z$ . Thus  $R(z)$  is a rational function with numerator and denominator of degree at most  $d$ .  $\square$

For stiffly accurate schemes with  $A$  invertible,  $R(z)$  is further constrained to satisfy  $\lim_{z \rightarrow \infty} R(z) = \lim_{z \rightarrow \infty} N(z)/D(z) = 0$ , so that  $\deg N \leq \deg D - 1 \leq \dim Y - 1$ . For DIRK schemes, the stability function  $R(z)$  has real poles since the eigenvalues of  $A$  are confined to the diagonal entries—which are real. Combining these observations with Theorems 3.7 and 3.8, we have proved the following theorem.

**THEOREM 3.9.** *Let an RK scheme  $(A, b)$  be given with  $Y$  defined as in (3.2) and stability function  $R(z) = e^z + \mathcal{O}(z^{p+1})$ , as  $z \rightarrow 0$ . Then*

$$(3.6) \quad \left\lfloor \frac{p+1+\sigma}{2} \right\rfloor \leq \dim Y.$$

*If the scheme is diagonally implicit, (3.6), we have the stronger bound*

$$(3.7) \quad p \leq \dim(Y) + 1 - \sigma.$$

*where  $\sigma = 1$  if the scheme is stiffly accurate and  $A$  is invertible, and  $\sigma = 0$ , otherwise.*

**3.3. (Step 2) A lower bound on the dimension of the subspace  $K_m$  for general RK schemes.** Here we provide a lower bound on  $\dim(K_m)$  in terms of  $m$  and the number of distinct abscissas  $n_c$ . The key idea is that  $K_m$  contains linear combinations of the vectors  $\vec{\tau}^{(k)}$ , and for large enough  $m$  the addition of each new vector increases the dimension of the space  $K_m$ .



We first denote the column space of the Vandermonde matrix with columns  $\vec{c}$

$$V := \begin{cases} \text{span}\{\vec{c}, \vec{c}^2, \dots, \vec{c}^{n_c-1}\} & \text{if all } c_j \neq 0, \\ \text{span}\{\vec{c}, \vec{c}^2, \dots, \vec{c}^{n_c-1}\} & \text{if at least one } c_j = 0. \end{cases}$$

Now define subspaces of  $V$  for values of  $m \geq n_c + 1$  as follows:

$$W_m := \left\{ w(C)\vec{c} \mid \text{for all } w(x) = \int_0^x \alpha(x)P_C(x) dx \text{ and } \alpha(x) \in \Pi_{m-1-n_c} \right\},$$

$$\Pi_d = \{ \alpha_0 + \dots + \alpha_d x^d \mid \alpha_j \in \mathbb{R} \text{ for } j = 0, \dots, d \}.$$

Here  $P_C(x)$  is the minimal polynomial for the matrix  $C$  (so that  $P_C(C) = 0$ ) and  $\Pi_d$  is the space of polynomials of degree at most  $d$ . By construction  $W_m \subseteq W_{m+1}$ .

We denote the smallest  $A$ -invariant subspace containing a vector space  $X$  by  $\mathcal{K}(X)$ ; namely  $\mathcal{K}(X)$  can be constructed explicitly by taking the span of all the  $A$ -Krylov subspaces of the basis vectors of  $X$  (similar to (3.1)). In particular,  $\mathcal{K}(V)$  and  $\mathcal{K}(W_m)$  are the smallest  $A$ -invariant spaces containing  $V$  and  $W_m$ , respectively. With these notations, we now have the following lemma.

LEMMA 3.10 (the  $K_m$  sandwich). *For any  $m \geq n_c + 1$  (or  $m = \infty$ ) it holds that*

$$W_m \subseteq \mathcal{K}(W_m) \subseteq K_m \subseteq \mathcal{K}(V).$$

*Proof.* The fact that  $K_m \subseteq \mathcal{K}(V)$  follows from the observation that every  $\vec{r}^{(k)}$ , for arbitrary  $k$ , is a linear combination of  $A\vec{c}^k \in \mathcal{K}(V)$  and  $\vec{c}^{k+1} \in \mathcal{K}(V)$ .

Next, we show that  $K_m$  contains the subspace  $W_m$ . Due to the fact that  $\mathcal{K}(W_m)$  and  $K_m$  are both  $A$ -invariant, this implies  $\mathcal{K}(W_m) \subseteq K_m$ .

Let  $w'(x) = w_{m-1}x^{m-1} + w_{m-2}x^{m-2} + \dots + w_0$  denote an arbitrary (real) polynomial and let  $w(x) := \int_0^x w'(s) ds$ . The space  $K_m$  then contains the linear combination

$$(3.8) \quad \sum_{j=1}^m w_{j-1} \vec{r}^{(j)} = \left( \sum_{j=1}^m A w_{j-1} C^{j-1} - \frac{1}{j} w_{j-1} C^j \right) \vec{c}, \quad \text{using (1.3),}$$

$$= (Aw'(C) - w(C))\vec{c} \in K_m.$$

Since  $C$  is diagonal, its minimal polynomial  $P_C(x)$  has simple roots (one for each distinct entry of  $\vec{c}$ ) and so  $\deg P_C = n_c$ . Thus, for  $m \geq n_c + 1$ , we can set  $w'(x) = \alpha(x)P_C(x)$  in (3.8) (so that  $w'(C) = 0$ ), showing that  $w(C)\vec{c} \in K_m$ . Hence  $K_m$  contains the subspace  $W_m$ .  $\square$

The next theorem follows directly from properties of  $W_m$ .

THEOREM 3.11 (properties of  $K_m$ ). *Consider an RK scheme with  $n_c$  distinct abscissas and the space  $K_m$  defined in (3.1). Then we have the following:*

- (a) *If  $m \leq 2n_c - 1$ , then  $\dim K_m \geq \max\{m - n_c, 0\}$ .*
- (b) *If  $m \geq 2n_c$  and all  $c_j \neq 0$  (i.e.,  $P_C(0) \neq 0$ ), then  $K_m = \mathcal{K}(V)$ ; in particular,  $\vec{c} \in K_m$  and  $\dim K_m \geq n_c$ .*
- (c) *If  $m \geq 2n_c - 1$  and  $c_j = 0$  for at least one value of  $j$  (i.e.,  $P_C(0) = 0$ ), then  $K_m = \mathcal{K}(V)$ ; in particular,  $\vec{c} \in K_m$  and  $\dim K_m \geq n_c - 1$ .*

*Proof.* For (a), we compute  $\dim W_m$  which bounds  $\dim K_m$  from below. First, note that vectors of the form  $w(C)\vec{c}$  (for any polynomial  $w(x)$ ) are in one-to-one correspondence with their associated polynomial remainders<sup>2</sup>  $w(x) \bmod P_C(x)$

<sup>2</sup>Here  $w(x) \bmod P_C(x)$  is the remainder of  $w(x)$  when divided by  $P_C(x)$ ; the set of all such remainders can be identified with the vector space  $\Pi_{n_c-1}$ .

(i.e., as vector spaces they are isomorphic).<sup>3</sup> Thus,  $W_m$  has the same dimension as the range of  $I : \Pi_{m-1-n_c} \rightarrow \Pi_{n_c-1}$  defined by

$$I[\alpha] := \left( \int_0^x \alpha(x) P_C(x) dx \right) \mod P_C(x),$$

where  $I[\cdot]$  is the linear (integral) operator that maps  $\Pi_{m-1-n_c}$  (which has dimension  $m - n_c$ ) into polynomials modulo  $P_C(x)$  (which has dimension  $n_c$ ). We now claim that  $I[\cdot]$  has a trivial null space whenever  $m \leq 2n_c - 1$ . Suppose not—then there are polynomials  $\alpha(x) \neq 0$  ( $\deg \alpha \leq n_c - 2$ ) and  $\beta(x)$  ( $\deg \beta \leq n_c - 1$ ) such that

$$\int_0^x \alpha(x) P_C(x) dx = \beta(x) P_C(x).$$

Differentiating both sides yields

$$(3.9) \quad (\alpha(x) - \beta'(x)) P_C(x) = \beta(x) P'_C(x).$$

The minimal polynomial  $P_C(x)$  contains  $n_c \geq 1$  distinct roots, and hence  $P'_C(x)$  has no common root with  $P_C(x)$  (i.e.,  $P_C(x)$  and  $P'_C(x)$  are relatively prime). Thus, (3.9) requires that  $P'_C(x)$  divide  $(\alpha(x) - \beta'(x))$  and  $P_C(x)$  divide  $\beta(x)$ , which is impossible in light of the degrees of  $\alpha(x)$ ,  $\beta(x)$ . Thus no  $\alpha(x)$  exists, and whenever  $n_c + 1 \leq m \leq 2n_c - 1$  one has the following dimension:

$$(3.10) \quad \dim W_m = \dim \Pi_{m-1-n_c} = (m - n_c) \leq \dim K_m.$$

Combining (3.10) with the trivial  $\dim K_m \geq 0$  when  $m < n_c + 1$  proves (a).

For (b), when  $m = 2n_c$ , (3.9) has the (unique) family of solutions

$$\beta(x) = \gamma P_C(x) \quad \text{and} \quad \alpha(x) = 2\gamma P'_C(x) \quad (\gamma \in \mathbb{R}).$$

Substituting this value of  $\alpha(x)$  into

$$I[\alpha] = \gamma P_C(x)^2 - \gamma P_C(0)^2 = -\gamma P_C(0)^2 \mod P_C(x),$$

shows that  $\gamma = 0$  is still the only solution to  $I[\alpha] = 0$ . Hence,  $I[\cdot]$  has a trivial null space,  $W_m$  contains the vector  $\vec{e}$  (e.g., take  $\gamma = -P_C(0)^{-2}$ ), and  $\dim W_m = n_c$ . Thus,  $W_m$  has the same dimension as  $V$  which forces  $W_m = V$  showing that  $K_m = \mathcal{K}(V)$ . For  $m \geq 2n_c$ ,  $K_m = \mathcal{K}(V)$  holds trivially since  $K_m \subseteq K_{m+1}$ .

For (c), if  $P_C(0) = 0$ , then every  $w(x)$  appearing in  $W_m$  is divisible by  $x^2$  so that  $w(x)$  and  $P_C(x)$  share a common factor of  $x$ . Thus  $x$  divides the remainder  $w(x) \mod P_C(x)$  as well as every polynomial in the range of  $I[\cdot]$ , i.e., for every  $\alpha \in \Pi_{m-1-n_c}$ ,  $I[\alpha] = xr(x)$  for some  $r(x)$  with  $\deg r \leq n_c - 2$ . When  $m = 2n_c - 1$  the range of  $I[\cdot]$  has dimension  $n_c - 1$  and includes all monomials  $x^j$  for  $1 \leq j \leq n_c - 1$ . Hence  $W_m = V$  showing that  $K_m = \mathcal{K}(V)$ .

The vector  $\vec{e}$  has an explicit construction: since  $P_C(x)$  and  $P'_C(x)$  are relatively prime, polynomials  $\beta(x)$  ( $\deg \beta \leq n_c - 1$ ) and  $\gamma(x)$  ( $\deg \gamma(x) \leq n_c - 2$ ) exist such that

$$\gamma(x) P_C(x) = \beta(x) P'_C(x) + 1.$$

<sup>3</sup>For any polynomial  $w(x) \mod P_C(x)$ , the identification of  $w(x) \mapsto w(C)\vec{e}$  is surjective and preserves the operations of the vector space. It is also injective: since  $w(C)$  is a diagonal matrix, the vector  $w(C)\vec{e} = 0$  if and only if  $w(C) = 0$ , and  $w(C) = 0$  if and only if  $w(x)$  is divisible by  $P_C(x)$ . Hence, the vectors  $w_1(C)\vec{e} = w_2(C)\vec{e}$  if and only if  $w_1(x) - w_2(x) = \alpha(x)P_C(x)$  for some  $\alpha(x)$ , i.e.,  $w_1(x) \equiv w_2(x) \mod P_C(x)$ .

Set  $\alpha(x) := \gamma(x) - \beta'(x)$  ( $\deg \alpha \leq n_c - 2$ ), so that  $\alpha(x)P_C(x) = \frac{d}{dx}(\beta(x)P_C(x)) + 1$ . Since  $P_C(0) = 0$ , integrating both sides shows that  $I[\alpha] = x$  and hence:  $\vec{c} \in W_m$ .  $\square$

By combining the lower bounds, we obtain a proof of the main result.

*Proof of the main result, Theorem 3.2.* Parts (1) and (2) of Theorem 3.2 follow from Theorem 3.11(b) and (c), respectively, with the fact that the order conditions  $\vec{b}^T \vec{c} = 1$  (for  $p = 1$ ) and  $\vec{b}^T \vec{c} = 1/2$  (for  $p = 2$ ) are incompatible with the WSO requirement that  $\vec{b} \perp K_m$  whenever  $\vec{c} \in K_m$  or  $\vec{c} \in K_m$ .

Theorem 3.1(3) follows from combining the inequality (3.3); the lower bound for  $\dim Y$  from Theorem 3.9 for fully implicit schemes; and the lower bound for  $\dim K_m$  in Theorem 3.11(a).  $\square$

**3.4. (Step 3) A lower bound on the dimension of the subspace  $K_m$  for DIRK schemes.** We now turn to obtaining lower bounds on the dimension of  $K_m$  in the more restrictive setting when  $A$  is DIRK. The results in this section build on Theorem 3.11. The difference of 1 (dimension) in parts (b) and (c) in Theorem 3.11 creates a somewhat bothersome issue for directly obtaining the lower bound in this section. However, this difference of 1 can be avoided for DIRK schemes that do not admit a GEDIRK structure.

**LEMMA 3.12** (extra dimension in  $K_m$  when  $c_j = 0$ ). *For a DIRK scheme that is not a GEDIRK scheme, if  $m \geq 2n_c$ , then  $\dim K_m \geq n_c$  (where the space  $K_m$  is defined in (3.1)).*

*Proof.* If  $\vec{c}$  has no zero, then Theorem 3.11(b) applies and we are done. Let  $\ell$  be the first occurrence of  $c_\ell = 0$  in  $\vec{c}$  (i.e.,  $c_j \neq 0$  for  $j < \ell$ ). Note that  $\ell > 1$  since  $c_1 = a_{11} \neq 0$  (because  $A$  is not a GEDIRK). Theorem 3.11(c) shows that  $W_m \subseteq K_m$  has dimension  $n_c - 1$ , that the vector  $\vec{c} \in K_m$ , and that  $W_m$  is orthogonal to  $\vec{e}_\ell$ , i.e.,  $\vec{e}_\ell^T \vec{u} = 0$  for all  $\vec{u} \in W_m$ . In addition,  $W_m$  contains a vector of the form (since  $\ell$  is the first occurrence of 0)

$$\vec{y} := (1, \dots, 1, 0, \star) \in W_m,$$

where the first  $\ell - 1$  entries of  $\vec{y}$  are 1 and  $\star$  is unimportant. Since  $K_m$  is  $A$ -invariant, we have  $\vec{c} - A\vec{y} \in K_m$ ; however,  $\vec{e}_\ell^T(\vec{c} - A\vec{y}) = a_{\ell\ell} \neq 0$  (where  $a_{\ell\ell} \neq 0$  since  $A$  is not a GEDIRK), showing that  $\dim K_m \geq n_c$ .  $\square$

To expedite the remaining proofs below, we define a truncation map  $[\cdot]_j$  which acts on matrices as  $[\cdot]_j : \mathbb{R}^{s \times s} \rightarrow \mathbb{R}^{j \times j}$  to isolate the upper  $j \times j$  block of  $A$ , and acts on vectors as  $[\cdot]_j : \mathbb{R}^s \rightarrow \mathbb{R}^j$  to isolate the top  $s$  components of  $\vec{c}$ :

$$[A]_j := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} \\ a_{21} & \ddots & & \vdots \\ \vdots & & & \\ a_{j1} & a_{j2} & \cdots & a_{jj} \end{pmatrix} \quad \text{and} \quad [\vec{c}]_j := \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_j \end{pmatrix}.$$

For a lower-triangular matrix  $A$ , the first  $j$  ( $\leq s$ ) components of the vectors  $\vec{\tau}^{(k)}$  and  $\vec{c}$  are functions of the upper  $j \times j$  sub-block of the matrix  $A \in \mathbb{R}^{s \times s}$  only. Moreover,

$$(3.11) \quad [A^n]_j = ([A]_j)^n \quad \text{and} \quad [\vec{c}^n]_j = ([\vec{c}]_j)^n.$$

We will also refer to the vectors  $\tilde{\tau}^{(k)}$  that arise when  $A$  is replaced by  $[A]_j$  in (1.3):

$$(3.12) \quad \tilde{\tau}^{(k)}([A]_j) := [A]_j [\tilde{c}^{k-1}]_j - \frac{1}{k} [\tilde{c}^k]_j \quad \text{for } k \geq 1,$$

which commute with the map in the sense that for every  $1 \leq j \leq s$ ,  $n \geq 0$ , and  $k \geq 1$ ,

$$(3.13) \quad \left[ \tilde{\tau}^{(k)}(A) \right]_j = \tilde{\tau}^{(k)}([A]_j), \quad \left[ A^n \tilde{\tau}^{(k)}(A) \right]_j = [A]_j^n \tilde{\tau}^{(k)}([A]_j).$$

We also extend  $[\cdot]$  to vector subspaces  $U \subseteq \mathbb{R}^s$  as  $[U]_j := \text{span}\{[\tilde{u}]_j \mid \tilde{u} \in U\}$ , which implies  $\dim U \geq \dim [U]_j$ .

LEMMA 3.13 (dimension of  $K_m$  for a DIRK scheme). *Consider a DIRK scheme with  $n_c$  distinct abscissas, and the corresponding space  $K_m$  defined in (3.1). Then*

$$(3.14) \quad \dim K_m \geq \min \left\{ \left\lfloor \frac{m + \kappa}{2} \right\rfloor, n_c \right\} - \kappa,$$

where  $\kappa = 1$  if  $A$  is a GEDIRK scheme, and  $\kappa = 0$ , otherwise.

The main idea in the proof is that the commutation property (3.13) implies that the matrix  $[A]_j$  and vectors  $\tilde{\tau}^{(k)}([A]_j)$  define an  $[A]_j$ -invariant space. Specifically,  $[K_m]_j$  is exactly the smallest  $[A]_j$ -invariant space containing  $\tilde{\tau}^{(k)}([A]_j)$  for  $k = 2, \dots, m$ . Hence, Theorem 3.11 and Lemma 3.12 apply<sup>4</sup> to  $[A]_j$  for each  $1 \leq j \leq s$ , where  $n_c$  is replaced with the number of distinct values in  $\{c_1, \dots, c_j\}$ . The dimension of the associated  $[A]_j$ -invariant space (for any  $j$ ) bounds is  $\dim K_m \geq \dim [K_m]_j$ . We then pick the “worst case”  $j$ . For simplicity, let

$$T(j) := \max\{r \mid \#\{c_1, \dots, c_r\} = j\},$$

where  $\#$  denotes the number of distinct values in a set. Then  $T(\cdot)$  is strictly increasing,  $T(1) \geq 1$ ,  $T(n_c) = s$ , and the set  $\{c_1, \dots, c_{T(j)}\}$  contains  $j$  distinct values.

*Proof of Lemma 3.13.* Assume  $m > 1$  (it holds trivially for  $m = 1$ ). If  $A$  is not a GEDIRK, set  $r := \min\{\lfloor \frac{m}{2} \rfloor, n_c\}$ . Apply Lemma 3.12 to  $[A]_{T(r)}$ , which has  $r$  distinct  $\tilde{c}$  values, to obtain  $\dim K_m \geq r$  (which is (3.14) when  $\kappa = 0$ ).

If  $A$  is a GEDIRK, we use the more general bound Theorem 3.11(a)–(c) applied to a DIRK matrix  $[A]_{T(r')}$  with  $r' (\leq m)$  distinct abscissas values:

$$(3.15) \quad \dim K_m \geq \min\{r' - 1, m - r'\}.$$

Set  $r' := \min\{\frac{m}{2}, n_c\}$  if  $m$  is even and  $r' := \min\{\frac{m+1}{2}, n_c\}$  if  $m$  is odd. Applying (3.15), and observing that (in all four cases)  $\min\{r' - 1, m - r'\} = r' - 1$  yields (3.14) with  $\kappa = 1$ .  $\square$

We conclude this section with a proof of the main result for DIRKs.

*Proof of the main result for DIRKs, Theorem 3.2.* When  $q \geq 2n_c - 1$ , Lemma 3.13 gives a lower bound on  $\dim K_m$  of  $\lfloor \frac{m+\kappa}{2} \rfloor - \kappa$ . Substituting this and the lower bound on  $\dim Y$  in Theorem 3.9 for DIRKs into (3.3) yields the result.  $\square$

<sup>4</sup>The new space  $W_m$  as well as the definition  $I[\cdot]$  must use the minimal polynomial of  $[C]_j$ .

**4. Minimal polynomials for the spaces  $Y$  and  $K_q$ .** Our results in the next two sections will be based on the linear algebra of orthogonal left and right invariant subspaces applied to  $Y$  and  $K_m$ . We therefore first review and establish key results (e.g., eigenspace, eigenvalue, and minimal polynomials) pertaining to the action of a matrix on its orthogonal invariant subspaces. We employ the standard terminology: a *monic polynomial*  $p(x)$  is defined as having a leading (i.e., highest degree term) coefficient of one; the *characteristic polynomial* of a matrix is defined as the (monic) polynomial  $\text{char}_A(x) := \det(xI - A)$ ; the *minimal polynomial*  $p(x)$  of a matrix  $A$  is the monic polynomial of smallest degree for which  $p(A) = 0$ . While it is often the case that (for instance, when the eigenvalues of  $A$  are distinct) the minimal polynomial is the characteristic polynomial, i.e.,  $p(x) = \text{char}_A(x)$ . In general,  $p(x)$  is of lower degree than  $\text{char}_A(x)$  when  $A$  has repeated eigenvalues.

For a real matrix  $A$  [5, Chaps. 8 and 9A], the minimal polynomial  $p(x)$  of  $A$  is (i) unique; (ii) has real coefficients;<sup>5</sup> (iii) divides the characteristic polynomial of  $A$  (thus every root of  $p(x)$  is an eigenvalue of  $A$  and the  $\deg p \leq \deg \text{char}_A(x)$ ); moreover, (iv) every eigenvalue of  $A$  (not including multiplicity) is a root of  $p(x)$ ; and (v) every polynomial  $\tilde{p}(x)$  satisfying  $\tilde{p}(A) = 0$  is divisible by  $p(x)$ . The following two theorems concern minimal polynomials for matrices restricted to invariant subspaces and will be employed in our study of WSO.

**THEOREM 4.1** (*A*-invariant subspaces). *Let  $A \in \mathbb{R}^{s \times s}$ , and let  $U \subseteq \mathbb{R}^s$  be an  $A$ -invariant subspace with dimension  $d := \dim(U)$  ( $d = 0$  is possible). Then there exists a unique monic polynomial  $p(x)$  of minimal degree and coefficients in  $\mathbb{R}$ , such that*

$$(4.1) \quad p(A)\vec{u} = 0 \quad \forall \vec{u} \in U.$$

*This polynomial has the following properties:*

- (a)  $p(x)$  divides the characteristic polynomial  $\text{char}_A(x)$ ;
- (b) every root of  $p(x)$  is an eigenvalue of  $A$ ; and
- (c)  $\deg p \leq \dim(U)$ .

*If, in addition,  $U$  has the form  $U = \text{span}\{\vec{v}, A\vec{v}, \dots, A^{d-1}\vec{v}\}$  (or  $U = \{0\}$  when  $\dim(U) = 0$ ) where  $A^j\vec{v}$  are linearly independent for  $j = 0, \dots, d-1$ , then*

- (d)  $\deg p = \dim(U)$ .
- (e) Condition (4.1) is equivalent to  $p(A)\vec{v} = 0$ .

We include a proof of Theorem 4.1 in Appendix A, using straightforward generalizations of textbook arguments:  $p(x)$  is the minimal polynomial for  $A$  restricted to the subspace  $U$  (from which properties (a)–(e) follow). Theorem 4.1 can then be extended to the case when  $A$  has both a left and right orthogonal invariant subspace.

**THEOREM 4.2** (left and right orthogonal  $A$ -invariant subspaces). *Let  $A \in \mathbb{R}^{s \times s}$  be a real matrix, and let  $U \subseteq \mathbb{R}^s$  and  $V \subseteq \mathbb{R}^s$  be orthogonal subspaces, where  $U$  is  $A$ -invariant and  $V$  is  $A^T$ -invariant. Denote the minimal polynomials in (4.1) from Theorem 4.1 as  $p(x)$  for  $U$  and  $q(x)$  for  $V$ , i.e.,  $p(A)\vec{u} = 0 \quad \forall \vec{u} \in U$  and  $q(A^T)\vec{v} = 0 \quad \forall \vec{v} \in V$ . Then the product  $p(x)q(x)$  divides the characteristic polynomial  $\text{char}_A(x)$ .*

Again, a proof of Theorem 4.2 is included in Appendix A for completeness. These results have a direct consequence for RK schemes with high WSO.

<sup>5</sup>More precisely, when  $A$  is real with  $p_r(x)$  the minimal polynomial with coefficients in  $\mathbb{R}$ , and  $p_c(x)$  is the minimal polynomial with coefficients over  $\mathbb{C}$ , then  $p_r(x) = p_c(x)$ .

*Example 4.3.* To highlight Theorems 4.1 and 4.2, consider

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & -1 & 1 & 0 \\ -2 & 0 & 1 & 2 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \quad \vec{v} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}.$$

The matrix  $A$  has an  $A$ -invariant subspace spanned by the columns of  $U$ , and an orthogonal  $A^T$ -invariant space spanned by  $\vec{v}$ . The associated minimal polynomials are  $p(x) = (x-1)^3$ , s.t.  $p(A)U = 0$ , and  $q(x) = (x-2)$ , s.t.  $\vec{v}^T q(A) = 0$ . The product  $p(x)q(x) = (x-1)^3(x-2) = \text{char}_A(x)$  equals, and thus divides, the characteristic polynomial. The example here will also be used to illustrate the lemmas in section 6.

LEMMA 4.4. *Let  $(A, \vec{b})$  be an RK method. Then there exists a unique (nonzero, real coefficient) monic polynomial  $Q(x)$  of minimal degree such that*

$$(4.2) \quad \vec{b}^T Q(A) = 0.$$

*The polynomial  $Q(x)$  has  $\deg Q = \dim Y$  and divides  $\text{char}_A(x)$ .*

*If, in addition, the RK method has WSO  $q \geq 2$ , then there exists a unique (nonzero, real coefficient) monic polynomial  $P(x)$  of minimal degree, such that*

$$(4.3) \quad P(A)\vec{\tau}^{(k)} = 0 \quad \text{for } k = 2, \dots, q.$$

*Moreover,  $\deg P \leq \dim K_q$ , and the product  $P(x)Q(x)$  divides  $\text{char}_A(x)$ .*

In (4.2)–(4.3),  $K_q$  and  $Y$  are the subspaces defined by (3.1) and (3.2); and  $\vec{\tau}^{(k)}$  are the stage order residuals. We refer to  $Q(x)$  and  $P(x)$  in Lemma 4.4 as the *minimal polynomials* for  $Y$  and  $K_q$ , respectively.

*Proof.* For the existence of  $Q$  and its properties, set  $U = Y$  and apply Theorem 4.1(d)–(e). For the existence of  $P$  and its properties, set  $U = K_q$  and  $V = Y$  and apply Theorem 4.2, where we denote the polynomials  $p(x)$ ,  $q(x)$  in Theorem 4.2 as  $P(x)$ ,  $Q(x)$ , respectively. The only point to prove is that the condition

$$(4.4) \quad P(A)\vec{v} = 0 \quad \forall \vec{v} \in K_q,$$

from Theorem 4.2 is equivalent to (4.3) where  $\vec{v}$  is restricted to be the stage order residuals  $\vec{\tau}^{(k)}$  for  $k = 1, \dots, q$ . Clearly, (4.4) implies (4.3) since  $K_q$  trivially includes  $\vec{\tau}^{(k)}$  for  $k = 1, \dots, q$ . Conversely, (4.3) implies that every basis vector in  $K_q$ , i.e.,  $A^j \vec{\tau}^{(k)}$ , satisfies  $P(A)A^j \vec{\tau}^{(k)} = 0$ . Hence, (4.3) may be used in lieu of (4.4) to define the minimal polynomial  $P(x)$ .  $\square$

Combined with (3.3) and Theorem 3.9, Lemma 4.4 implies the following corollary.

COROLLARY 4.5. *For an  $s$ -stage RK scheme with order  $p \geq 1$  and WSO  $q$  (with  $K_q$  and  $P(x)$  defined in (3.1) and (4.3)), one has*

$$\deg(P) \leq \dim(K_q) \leq \begin{cases} s - \lfloor \frac{p+1+\sigma}{2} \rfloor & \text{for fully implicit schemes,} \\ s - p - 1 + \sigma & \text{for diagonally implicit schemes,} \end{cases}$$

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

Remark 4.6 (eigenvalues of  $A$ ). A consequence of Lemma 4.4 is that the eigenvalues of  $A$  (including multiplicity) for a RK scheme can be partitioned into

three sets: the roots of  $P(x)$ , the roots of  $Q(x)$ , and the roots of  $N(x)$ , where  $\text{char}_A(x) = P(x)Q(x)N(x)$ .

*Remark 4.7.* For an SDIRK with diagonal entries of  $\gamma$  in  $A$ , the polynomials  $P, Q$  have the form  $Q(x) = (x - \gamma)^d$  ( $d = \dim Y$ ), and  $P(x) = (x - \gamma)^\mu$  ( $\mu \leq \dim K_q$ ).

Broadly speaking, the polynomials  $Q(x)$  and  $P(x)$  are algebraic objects associated with the corresponding geometric (orthogonal) spaces  $K_q$  and  $Y$ . In the next sections, we use the polynomial  $P(x)$  to obtain an expression for the stability function of RK schemes (particularly useful for schemes with WSO), and use  $Q(x)$  to obtain necessary conditions for high WSO schemes (useful in the construction of WSO schemes).

**5. Stability function in a basis for  $Y$ .** In general, the RK stability function has a restricted form when  $p$  is close to (the optimal)  $2s$ , and also when  $p \geq s$  (e.g., see [17, Chap. IV.3, p. 47]). In this section, we extend this constrained structure for  $R(z)$  to the more general setting  $p \geq \dim Y$  and to when  $p$  is close to  $2\dim Y$  (which is a setting applicable for WSO). The key idea is to write  $Q(x)$ , and  $Y$ , in a basis of orthogonal polynomials related to the (*tall-tree*)  $p$ th order conditions

$$(5.1) \quad \vec{b}^T A^j \vec{e} = \frac{1}{(j+1)!}, \quad 0 \leq j \leq p-1.$$

Specifically, define the linear functional  $\mathcal{L}(\cdot)$  on the space of monomials  $x^n$  with *moments*  $(\mu_n)_{n \geq 0}$  as

$$(5.2) \quad \mathcal{L}(x^n) := \mu_n, \text{ where } \mu_n := \frac{1}{(n+1)!} \text{ for } n \geq 0.$$

Associated to  $\mathcal{L}$  are the Hankel moment matrices  $\mathbf{H}_n = (\mu_{i+j-2})_{i,j=1}^n$ . Appendix B shows that  $\det \mathbf{H}_n \neq 0$  for all  $n \geq 0$ , which is a sufficient condition to construct a basis of polynomials  $(Q_j)_{j \geq 0}$  satisfying the orthogonality relation

$$\mathcal{L}(Q_i Q_j) = \zeta_i \delta_{ij}, \text{ where } \zeta_0 = 1, \text{ and } \zeta_i = \frac{\det(\mathbf{H}_{i+1})}{\det(\mathbf{H}_i)} \text{ if } i \geq 1.$$

Here  $\delta_{ij}$  is the Kronecker delta. For the moments in (5.2), the first two orthogonal polynomials are  $Q_0(x) = 1$  and  $Q_1(x) = x - 1/2$ ; subsequent polynomials  $Q_j(x)$  satisfy the three term recursion relation given by (see Appendix B for details):

$$(5.3) \quad Q_{n+1}(x) = xQ_n(x) + \xi_n^2 Q_{n-1}(x), \quad \xi_n^2 = \frac{1}{4(4n^2 - 1)} \text{ for } n \geq 1.$$

For instance, the next two polynomials are

$$Q_2(x) = x^2 - \frac{1}{2}x + \frac{1}{12} \quad \text{and} \quad Q_3(x) = x^3 - \frac{1}{2}x^2 + \frac{1}{10}x - \frac{1}{120}.$$

Any degree  $d$  polynomial can then be written in terms of the basis  $\{Q_1, \dots, Q_d\}$ . We use  $\alpha_j$  to denote the coefficients of the minimal polynomial for  $Y$  in this basis:

$$(5.4) \quad Q(x) = Q_d(x) + \alpha_{d-1}Q_{d-1}(x) + \dots + \alpha_0Q_0(x), \quad \text{where } d := \dim Y.$$

The order conditions then constrain (5.4) as follows.

**LEMMA 5.1** ( $Q$  is “orthogonal” to  $Q_j$ ). *Consider an RK method with coefficients  $(A, \vec{b})$ , and let  $Q$  be the minimal polynomial for the subspace  $Y$ , as defined in (3.2), written in the form (5.4). If the method’s stability function is  $R(z) = e^z + \mathcal{O}(z^{p+1})$  as  $z \rightarrow 0$ , then  $\alpha_j = 0$  for  $j \leq p - d - 1$ .*

*Proof.* Set  $N := p - d - 1 \geq 0$  (if  $N < 0$  there is nothing to prove). The  $p$ th order conditions (5.1) imply  $\vec{b}^T Q(A) A^\ell \vec{e} = \mathcal{L}(x^\ell Q(x))$  for all  $0 \leq \ell \leq N$ . On the other hand,  $\vec{b}^T Q(A) = 0$ . Hence,  $\mathcal{L}(Q_\ell Q) = \zeta_\ell \alpha_\ell = 0$  for all  $0 \leq \ell \leq N$ , so  $\alpha_j = 0$  for  $j \leq N$  (since  $\zeta_j \neq 0$ ). If  $N \geq d$ , this would imply  $\mathcal{L}(Q Q_d) = \zeta_d = 0$ , which is a contradiction. Note that  $p \leq 2d$  or  $\lfloor \frac{p+1}{2} \rfloor \leq \dim(Y)$  when  $p \geq 1$  and  $d \geq 0$  are integers.  $\square$

Utilizing the basis  $(Q_j)_{j \geq 0}$ , we now work out an expression for the stability function  $R(z)$ . We first expand  $\vec{b}^T (I - zA)^{-1}$  in the Krylov basis  $\vec{b}^T Q_0(A), \dots, \vec{b}^T Q_{d-1}(A)$  with (to be determined) coefficients  $\beta_j = \beta_j(z)$  functions of  $z$ :

$$(5.5) \quad \vec{b}^T (I - zA)^{-1} = \vec{b}^T (\beta_{d-1} Q_{d-1}(A) + \beta_{d-2} Q_{d-2}(A) + \dots + \beta_0 Q_0(A)).$$

Right-multiplying (5.5) by  $I - zA$  and using the fact that  $\vec{b}^T Q(A) = 0$ , leads to the algebraic equation for the coefficients:

$$(5.6) \quad (1 - zx)(\beta_{d-1} Q_{d-1}(x) + \beta_{d-2} Q_{d-2}(x) + \dots + \beta_0 Q_0(x)) \equiv 1 \pmod{Q(x)}.$$

To solve for  $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_{d-1})^T$ , write  $\vec{Q}(x) = (Q_0(x), Q_1(x), \dots, Q_{d-1}(x))^T$ . Using relation (5.3), multiplication  $x\vec{Q}$  can be written as a matrix multiplication  $\mathbf{S}\vec{Q}$ :

$$x\vec{Q} = \begin{pmatrix} Q_1 + \frac{1}{2}Q_0 \\ Q_2 - \xi_1^2 Q_0 \\ \vdots \\ Q_d - \xi_{d-1}^2 Q_{d-2} \end{pmatrix} = \mathbf{S}\vec{Q}, \text{ where } \mathbf{S} := \begin{pmatrix} \frac{1}{2} & 1 & 0 & \dots & 0 \\ -\xi_1^2 & 0 & 1 & & \vdots \\ 0 & -\xi_2^2 & 0 & & \\ \vdots & & & \ddots & 1 \\ 0 & \dots & & -\xi_{d-1}^2 & 0 \end{pmatrix} - \vec{e}_d \vec{\alpha}^T.$$

Here  $\vec{e}_j$  is the  $j$ th unit vector and  $\vec{\alpha} = (\alpha_0, \dots, \alpha_{d-1})^T$  are the coefficients of  $Q(x)$  in (5.4). Using these notations and relations, (5.6) becomes

$$\begin{aligned} (1 - zx)\vec{\beta}^T \vec{Q} &\equiv 1 \pmod{Q(x)}, \\ \left( (I - z\mathbf{S}^T) \vec{\beta} - \vec{e}_1 \right)^T \vec{Q} &\equiv 0 \pmod{Q(x)}. \end{aligned}$$

The vector  $\vec{\beta}$  then has the solution

$$(5.7) \quad (I - z\mathbf{S}^T) \vec{\beta} = \vec{e}_1, \quad \text{and via Cramer's rule: } \beta_{j-1}(z) = \frac{\det((I - z\mathbf{S}^T)_j)}{\det(I - z\mathbf{S})},$$

where  $(I - z\mathbf{S}^T)_j$  is the matrix  $I - z\mathbf{S}^T$  with column  $j$  replaced by  $\vec{e}_1$ . Substituting the expression for  $\vec{b}^T (I - zA)^{-1}$  into  $R(z)$  yields

$$(5.8) \quad R(z) = 1 + \sum_{j=0}^{d-1} \left( \vec{b}^T Q_j(A) \vec{e} \right) z \beta_j(z).$$

Finally, the orthogonality property Lemma 5.1 implies that  $\vec{b}^T Q_j(A) \vec{e}$  agrees with  $\mathcal{L}(Q_j) = \delta_{0j}$  for all  $j \leq p-1$ . This (significantly) simplifies the summation in (5.8).

LEMMA 5.2 (stability function when  $p \geq \dim Y$ ). *If  $R(z) = e^z + \mathcal{O}(z^{p+1})$  is a  $p$ th order approximation as  $z \rightarrow 0$  and  $p \geq \dim Y$ , then*

$$(5.9) \quad R(z) = 1 + z\beta_0(z),$$



where  $\beta_0(z)$  is given by (5.7) and depends only on the expansion of  $Q(x)$  from (5.4), i.e.,  $R(z)$  is a function of  $\alpha_j$  (for  $j = p - \dim Y, \dots, \dim Y - 1$ ).

To the best of the authors' knowledge, the orthogonal polynomials  $(Q_j)_{j \geq 0}$  have not previously been introduced in the context of RK schemes. However, they are connected with existing ideas in the (extensive) literature on RK methods. For instance, the values  $\xi_j$  defined in (5.3) appear in the  $W$ -transform and in recurrence relations for the shifted Legendre polynomials [17, Chap. IV.5] commonly used in the construction of high stage order schemes. Lemma 5.1 is also similar in spirit to the orthogonality relation Lemma 5.15 in [17, Chap. IV.5] which expands the characteristic polynomial of  $C$  (when  $n_c = s$ ) in the shifted Legendre polynomial basis.

*Remark 5.3* (orthogonal polynomials with respect to a linear functional). In the (semi)classical theory of orthogonal polynomials, the Hankel (moment) matrices are positive definite and  $\mathcal{L}(\cdot)$  defines an inner product as an integral with respect to a (nonnegative) measure [15, 20]. In contrast, for the matrices  $\mathbf{H}_d$  considered here, both  $\det \mathbf{H}_d \neq 0$  and  $\zeta_i \neq 0$ . However,  $\zeta_i < 0$  for some values of  $i$ , so  $\mathcal{L}(\cdot)$  does not have an inner product representation. While much of the (semi)classical theory of orthogonal polynomials still holds [15] (e.g., the three-term recurrence), some properties do not. For instance, the basis polynomials  $Q_j(x)$  have complex roots, whereas in the classical theory of orthogonal polynomials the roots are real and simple.

*Remark 5.4* (moment-generating function of  $\mathcal{L}$ ). The moments  $(\mu_n)_{n \geq 0}$  do not arise by accident; their generator function is  $G(z) = (e^z - 1)/z = \sum_{n \geq 0} \mu_n z^n$ , whose Taylor expansion agrees up to a suitable order with  $(R(z) - 1)/z$ . The Hankel determinants, or coefficients of the recurrence relation (5.3) appear in the  $J$ -fraction (continued fraction) of the generating function. For  $G(z)$ , this continued fraction appears in the RK literature through Padé approximations of  $e^z$  (see, e.g., Exercise 4 in Chap. IV.3 or Theorem 5.18 of [17]).

*Remark 5.5* (stiff accuracy). Equation (5.9) can be extended to incorporate the structure imposed by stiff accuracy (for invertible  $A$ ) using an alternative polynomial basis  $(\tilde{Q})_{n \geq 0}$ . Specifically, let  $\tilde{\mu}_0 = 1$ ,  $\tilde{\mu}_{n+1} := \mu_n$  for  $n \geq 0$  and  $\tilde{\mathcal{L}}(x^n) := \tilde{\mu}_n$ . The Hankel matrices with moments  $\tilde{\mu}_n$  are shown in (B.1) to have nonzero determinants and define a basis orthogonal with respect to  $\tilde{\mathcal{L}}$ , i.e.,  $\tilde{Q}_0(x) = 1$ ,  $\tilde{Q}_1(x) = x - 1$ , etc. The stability function (having one less degree of freedom) for stiffly accurate schemes can then be obtained in this basis.

**6. Necessary conditions on  $P(x)$  for DIRK schemes.** Theorem 3.2 provides a bound on the WSO for a method with order  $p$  and number of stages  $s$ . However, the theorem does not explain how one might go about constructing schemes with high WSO. In this section we examine the solvability of the equations  $P(A)\tilde{\tau}^{(k)} = 0$  for the matrices  $[A]_j$ . The results impose necessary conditions on  $P(x)$  (e.g., constraints on the roots) for the construction of high WSO schemes. The necessary conditions also restrict how the spectrum of  $A$  is partitioned into the minimal polynomials  $Q(x)$ ,  $P(x)$ . We focus in this section on schemes that are not GEDIRK schemes.

Here we write  $p_j(x)$  to denote the minimal polynomial of  $[A]_j$  when  $j \geq 1$  (we have  $p_j(x) = 1$  if  $j = 0$ ). The main result is that  $p_r$  must divide  $P(x)$  when the first  $r$  abscissas are distinct.

**LEMMA 6.1** (necessary condition on  $P(x)$  for a DIRK). *Let a DIRK scheme that is not a GEDIRK be given with WSO  $q \geq 2$ . If the abscissas  $\{c_1, c_2, \dots, c_r\}$  are distinct for  $r \leq \lfloor \frac{q}{2} \rfloor$ , then  $p_r(x)$  divides  $P(x)$ .*

*Proof.* Apply the map  $[\cdot]_r$  to (4.3) and use (3.11) to obtain

$$(6.1) \quad P([A]_r) \tau^{(k)}([A]_r) = 0 \text{ for } k = 2, \dots, q.$$

By Lemma 3.12 we have that  $[K_q]_r = \mathbb{R}^r$ , which when combined with (6.1) implies  $P([A]_r) = 0$ . Thus  $p_r(x)$  divides  $P(x)$ .  $\square$

**COROLLARY 6.2.** *For any nonconfluent DIRK scheme (that is not a GEDIRK scheme),  $p_r(x)$  divides  $P(x)$  where  $r = \lfloor \frac{q}{2} \rfloor$ .*

Example 4.3 demonstrates that the assumption of distinct abscissas in Lemma 6.1 is necessary. It provides a confluent scheme  $(\vec{c} = (1, 2, 2, 1)^T)$  that is  $S$ -irreducible and for which the minimal polynomial of  $A$  (i.e.,  $p_4(x) = (x-1)^3(x-2)$ ) is equal to the characteristic polynomial and does not divide  $P(x) = (x-1)^3$ .

One might wonder (I) whether the assumption of distinct abscissas in Lemma 6.1 can be relaxed; and (II) whether  $p_j(x)$  must ever be the characteristic polynomial of  $[A]_j$ . In general, the answer to both questions is no (see Example 4.3). However, due to the small number of variables in the top block of  $[A]_j$  the answer to both questions is yes when  $j \leq 3$  in (I) and  $j \leq 2$  in (II), provided that the scheme is  $S$ -irreducible.

**DEFINITION 6.3** ( $S$ -reducible, Def. 12.17 in [17]). *An RK method is  $S$ -reducible if for some partition  $(S_1, \dots, S_{\tilde{r}})$  of  $\{1, \dots, s\}$  ( $\tilde{r} < s$ ) the indicator vectors<sup>6</sup>  $\vec{S}^{(m)} = \sum_{j \in S_m} \vec{e}_j$  satisfy for all  $\ell, m$*

$$(6.2) \quad (\vec{e}_i - \vec{e}_j)^T A \vec{S}^{(m)} = 0 \quad \text{if } i, j \in S_\ell.$$

An  $S$ -reducible scheme is equivalent to a smaller  $\tilde{r}$ -stage RK scheme  $(A^*, \vec{b}^*)$  where each partition  $S_j$  is replaced by a single stage (see eq. (12.24) of [17, Chapt. IV.12]; the smaller scheme generates the same stage solutions  $g_i$  in (1.2a)). For  $S$ -reducible DIRKs, the new scheme can also be made a DIRK.<sup>7</sup>

**Remark 6.4** (a simple  $S$ -reducibility observation). Any DIRK scheme with  $c_1 = c_2$  is  $S$ -reducible: apply Def. 12.17 in [17], where the partition of equivalent stages (i.e., partition of the integers  $\{1, \dots, s\}$ ) is taken as  $S_1 = \{1, 2\}$  and  $S_2 = \{3\}, \dots, S_{s-1} = \{s\}$ . Then [18, Thm. 2.2] implies that the first two stages of  $A$  yield the same intermediate stage value solutions and can be replaced by a single stage.

**LEMMA 6.5** (a general necessary condition on  $P(x)$ ). *Let an  $S$ -irreducible DIRK scheme (that is not a GEDIRK) be given with coefficient matrix  $A$  and WSO  $q$ , such that  $[A]_3$  is invertible.*

- (a) *If  $q > 1$ , then  $\deg P \geq 1$  and  $a_{11}$  is a root of  $P$ .*
- (b) *If  $q > 3$ , then  $\deg P \geq 2$ , and  $a_{11}, a_{22}$  are roots of  $P$ .*
- (c) *If  $q > 5$ , then  $P(x) = p_3(x)\tilde{P}(x)$  where  $p_3(x)$  is the minimal polynomial of  $[A]_3$  and is divisible by  $(x - a_{11})(x - a_{22})$ .*

*Proof.* (a) Since the set  $\{c_1\}$  contains one distinct element (trivially), we can apply Lemma 6.1 with  $r = 1$  when  $q > 1$ , so that  $p_1(x) = (x - a_{11})$  divides  $P(x)$ .

For (b), Remark 6.4 implies that  $c_1 \neq c_2$  for any  $S$ -irreducible method. Applying Lemma 6.1 with  $r = 2$  when  $q > 3$  implies that  $p_2(x)$  divides  $P(x)$ . If  $p_2(x) = (x - a_{11})$

<sup>6</sup>In a slight abuse of notation we use  $\vec{S}_j$  to denote both the set and the indicator vector. The notation  $\vec{S}_j$  is nonstandard—the definition typically writes (6.2) as a summation.

<sup>7</sup>Formally, if we order the partitions so that  $\min\{x \in S_i\} < \min\{y \in S_j\}$  whenever  $i < j$ , then for  $1 \leq i \leq \tilde{r}$  define the new scheme (which is a DIRK) as  $a_{ij}^* = \sum_{k \in S_j} a_{ik}$ ,  $b_j^* = \sum_{k \in S_j} b_k$ .

has only one root (which must be  $a_{11}$  by (a)), then  $p_2([A]_2) = 0$  implies  $[A]_2 = a_{11}I$ , whence  $c_1 = c_2 = a_{11}$  (which is a contradiction). Thus,  $p_2(x) = (x - a_{11})(x - a_{22})$  is the characteristic polynomial of  $[A]_2$ .

For (c), suppose by way of contradiction that  $p_3(x)$  does not divide  $P(x)$ , so by (b),  $p_2(x)$  is the minimal polynomial of  $[A]_3$  restricted to  $[K_q]_3$ . When  $q > 5$ , we claim that

$$(6.3) \quad [K_q]_3 = \text{span}\{\vec{e}_3, [\vec{c}]_3\},$$

where either  $c_1 = c_3 \neq c_2$  or  $c_2 = c_3 \neq c_1$ . The reasons are as follows: (i) Lemma 6.1 implies that  $\vec{c}$  is confluent (with  $c_1 \neq c_2$ ) so  $n_c = 2$  and (by the same argument in the proof)  $\dim [K_q]_3 \leq 2$ ; (ii) if  $c_j = 0$ , (6.3) follows immediately from Theorem 3.11(b); (iii) if  $c_j = 0$ , the construction in Lemma 3.12 implies  $[K_q]_3 = \text{span}\{[\vec{c}]_3, [A\vec{c}]_3\}$  which is (6.3) when  $[A]_3$  is invertible (i.e., multiply by  $[A]_3^{-1}$ ). For  $[K_q]_3$  to be  $A$ -invariant requires

$$(6.4) \quad A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{11}-a_{33} & 0 & a_{33} \end{pmatrix} \text{ if } c_1 = c_3, \quad A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{21} & a_{22}-a_{33} & a_{33} \end{pmatrix} \text{ if } c_2 = c_3.$$

In both cases  $A$  is  $S$ -reducible with partitions  $S_1 = \{1, 3\}$ ,  $S_2 = \{2\}$  ( $c_1 = c_3$ ) or  $S_1 = \{1\}$ ,  $S_2 = \{2, 3\}$  ( $c_2 = c_3$ ). Hence  $p_2(x)$  divides  $p_3(x)$  which divides  $P(x)$ .  $\square$

While this section has focused entirely on  $P(x)$ , the final remark concerns a similar-in-spirit result for  $Q(x)$ .

*Remark 6.6* ( $a_{ss}$  is a root of  $Q(x)$ ). A DJ-irreducible DIRK scheme has  $\vec{e}_s^T \vec{b} \neq 0$ ; otherwise, the scheme is independent of the last stage. Since  $0 = \vec{b}^T Q(A) \vec{e}_s = Q(a_{ss}) \vec{b}^T \vec{e}_s$ , the entry  $a_{ss}$  must be a root of  $Q$ , i.e.,  $Q(x) = (x - a_{ss})\tilde{Q}(x)$ .

**7. Examples.** This section provides some concrete examples that highlight how the theory established above can be used to reduce the number of degrees of freedom when constructing RK schemes with high WSO. In Examples 7.1 and 7.2 we parameterize DIRK schemes with invertible  $A$  and WSO  $q = 3$ . Based on Remark 4.6 and Lemma 6.1, we can determine that certain diagonal entries of  $A$  (those closest to the top left) are roots of  $P(x)$ , while those near the bottom right are roots of  $Q(x)$ .

*Example 7.1.* The theory enables a complete characterization of DIRK schemes with  $(s, p, q) = (2, 2, 3)$ . Theorem 3.9 and Lemma 3.13 require the following:  $\dim Y = 1$ ,  $\dim K_q = 1$ . Via Lemma 6.1 the associated polynomials are  $P(x) = x - a_{11}$  and  $Q(x) = x - a_{22}$ . For a  $p = 2$  method, the orthogonality property from Lemma 5.1 implies that  $Q(x) = Q_1(x) = x - \frac{1}{2}$ , so  $a_{22} = \frac{1}{2}$ . The first row of equation  $P(A)\vec{\tau}^{(k)} = 0$  for  $k = 2, 3$  is automatically satisfied; the second row yields two equations, the solution of which determines  $A$ . The last order condition  $\vec{b}^T \vec{e} = 1$  determines  $\vec{b}$ , resulting in two schemes:

$$(7.1) \quad A = \begin{pmatrix} 1 \mp \frac{\sqrt{2}}{2} & 0 \\ \frac{1}{2} \pm \frac{\sqrt{2}}{2} & \frac{1}{2} \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} \frac{1}{2} \pm \frac{\sqrt{2}}{4} \\ \frac{1}{2} \mp \frac{\sqrt{2}}{4} \end{pmatrix}, \quad R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}.$$

Either choice of signs leads to a  $(2, 2, 3)$  scheme. In both cases,  $\vec{b}^T A = \frac{1}{2} \vec{b}^T$  is a left eigenvector of  $A$  and  $K_3 = \text{span}\{2\vec{e} - \vec{e}\} = W_3$ .

*Example 7.2* (DIRK schemes with  $(s, p, q) = (3, 3, 3)$ ). The chosen values of  $(s, p, q)$  satisfy Theorem 3.2 sharply and require that  $\dim Y = 2$  and  $\deg P = \dim K_q = 1$  with

$P(x) = x - a_{11}$ ,  $Q(x) = (x - a_{22})(x - a_{33})$ . Since  $[A]_2$  satisfies (6.1) it must be a scalar multiple (call the ratio  $a$ ) of the Butcher matrix in (7.1). The polynomial  $Q(x) = Q_2(x) + \alpha_1 Q_1(x)$  then has  $a_{22} = a/2$  as a root so that  $\alpha_1 = -Q_2(\frac{a}{2})/Q_1(\frac{a}{2})$ , which uniquely determines (i)  $a_{33}$  as the second root of  $Q(x)$ ; and (ii) the stability function in terms of  $a$ . The matrix  $A$  and  $R(z)$  are constrained to be

$$A = \begin{pmatrix} (1 \mp \frac{\sqrt{2}}{2})a & 0 & 0 \\ (\frac{1}{2} \pm \frac{\sqrt{2}}{2})a & \frac{1}{2}a & 0 \\ a_{31} & a_{32} & \frac{3a-2}{6(a-1)} \end{pmatrix}, \quad R(z) = \frac{(1 + 6\alpha_1)z^2 + (6 + 12\alpha_1)z + 12}{(1 - 6\alpha_1)z^2 + (12\alpha_1 - 6)z + 12}.$$

The values  $(a_{31}, a_{32})$  can be parameterized in terms of  $a$  via the 3rd row equations of  $P(A)\vec{\tau}^{(j)} = 0$  for  $j = 2, 3$ . For each choice of  $\pm$  in  $A$ , we have the following. There are three solution branches for  $(a_{31}, a_{32})$ , two of which yield S-reducible schemes with the structures given in (6.4). The one irreducible solution branch has that  $\vec{\tau}^{(2)}$  and  $\vec{\tau}^{(3)}$  are parallel. The vector  $\vec{b}$  is determined via  $\vec{b}^T(\vec{c}, \vec{c}, \vec{\tau}^{(2)}) = (1, \frac{1}{2}, 0)$ . By construction,  $\vec{b}^T Q(A) = 0$  so that  $\vec{b}^T \vec{c} = 1$  and  $\vec{b}^T \vec{c} = \frac{1}{2}$  ensure  $\vec{b}^T A^2 \vec{c} = \frac{1}{6}$ . Finally, combining  $\vec{b}^T (A\vec{\tau}^{(1)} + \vec{\tau}^{(2)}) = 0$  with  $\vec{b}^T A^2 \vec{c} = \frac{1}{6}$  yields  $\vec{b}^T \vec{c}^2 = \frac{1}{3}$ .

As expected, with  $s = 3$  there is no choice of  $a$  that satisfies the additional conditions imposed by order  $p = 4$  or  $q = 4$ . Finally,  $0 < a < \frac{2}{3}$  or  $a > 1$  is required for  $A$  to have positive eigenvalues and  $R(z)$  to have all its poles in the right half plane.

*Example 7.3.* The bound in Theorem 3.1 can be sharp for both stiffly accurate schemes and EDIRKs. Specifically, stiffly accurate DIRK schemes with  $(s, p, q) = (4, 3, 3)$  were constructed in an ad-hoc fashion in [21], where a posteriorily  $s$  is observed to be sharp. Setting  $\dim K_q = 0$  in Lemma 3.13, the fact  $\tilde{q} \leq q$  recovers the known result that stage order is limited to  $\tilde{q} = 2$  for EDIRKs and  $\tilde{q} = 1$  for DIRKs.

*Example 7.4* (schemes with high stage order). The Gauss–Legendre RK methods satisfy the bound in Theorem 3.1 sharply since  $p = 2s$ ,  $q = p$  (so that  $\dim(K_q) = 0$ ) and  $n_c = s$  (and are not stiffly accurate). As an example of Corollary 4.5, we obtain the bound  $p \leq 2s$ , well known already from the theory of numerical quadrature.

*Remark 7.5* (guide to constructing DIRK schemes with high WSO). For schemes with  $q = p$ , one can take  $r = \lfloor q/2 \rfloor$ , set  $P(x) = (x - a_{11}) \cdots (x - a_{rr})$  and then solve (4.3) as a sufficient condition (and use Theorem 3.2 to guide the number of stages).

**8. Conclusions and outlook.** Weak stage order can be viewed as a geometric condition that, when satisfied, can remove order reduction in RK schemes applied to linear problems with time-independent operators. The general theory of WSO provided here relates geometric objects (WSO invariant subspaces) to associated algebraic objects (minimal polynomials). This relationship allows us to establish order barriers for WSO that generalize the well-known bounds on the RK order  $p$  in terms of  $s$  (both for general RK schemes and DIRKs). Along the way, we also provide new formulas for the RK stability function in terms of a family of polynomials which are “orthogonal” with respect to a linear functional. The new necessary conditions show how one needs to split the spectrum of  $A$  into the roots of  $P(x)$  and  $Q(x)$ —which is of practical use when constructing high WSO schemes (i.e., beyond 3). Indeed, the necessary conditions were used in the companion work [7] in the construction of new DIRK schemes with WSO 4 and 5 (and satisfy the full set of order conditions).

Since SDIRK methods are a subset of DIRK methods, and ERK methods are a subset of GEDIRK methods, the bounds in Theorem 3.2 apply to these classes as well. It is natural to ask whether stricter bounds can be found for these smaller classes.

Based on Theorem 3.7, it seems that no further improvement can be obtained in the bounds on  $\dim(Y)$  for SDIRK methods compared to DIRK methods, but it might be possible to obtain tighter bounds by further exploiting the structure of  $K_q$ . Stricter bounds on ERK methods are an area of current research.

The results presented here give rise to several research directions. First, is Theorem 3.2 sharp for all  $p, q, s$ ? And, is it further constrained by the nontall tree order conditions? There is also the related (practical) issue of constructing DIRK schemes with  $q = p$  (or  $q = p - 1$ ) and the fewest stages  $s \sim \frac{2}{3}p$  satisfying Theorem 3.2 (or better yet, analytically parameterizing such schemes). While the main results in this work do apply to all (E)DIRKs, DIRKs with additional constraints such as SDIRKs or ERKs are expected to further impact the bounds in Theorem 3.2.

**Appendix A. Proofs of Theorems 4.1 and 4.2.** We collect the linear algebra proofs here, which are adapted from various materials (e.g., [30, Chap. 2.6]).

*Proof of Theorem 4.1,  $A$ -invariant subspaces.* Let  $\mathbf{U} \in \mathbb{R}^{s \times d}$  form an orthogonal basis for  $U$ . Since  $U$  is  $A$ -invariant, there is a square matrix  $A_{UU} := \mathbf{U}^T A \mathbf{U} \in \mathbb{R}^{d \times d}$  such that (each column of  $\mathbf{U}$  is mapped back into the column space of  $\mathbf{U}$ )

$$(A.1) \quad A\mathbf{U} = \mathbf{U}A_{UU},$$

For any polynomial  $p(x)$ , (A.1) then implies that  $p(A)\mathbf{U} = \mathbf{U}p(A_{UU})$ . Thus,  $p(x)$  in (4.1) is the minimal polynomial of  $A_{UU}$  (which has real coefficients since  $A_{UU}$  is real) because it satisfies  $p(A)\mathbf{U} = 0$  and any other smaller degree polynomial fails to satisfy  $p(A)\mathbf{U} = \mathbf{U}p(A_{UU}) \neq 0$ .

For (a),  $p(x)$  divides the characteristic polynomial of  $A_{UU}$ , which divides  $\text{char}_A(x)$ ; (c) also follows since  $\deg p \leq \deg \text{char}_{A_{UU}}(x) = \dim(U)$ .

For (b), every root  $\lambda$  of  $p(x)$  is an eigenvalue of  $A_{UU}$  and hence of  $A$ .

For (d), set  $\ell = \deg p$ . The condition  $p(A)\vec{v} = 0$  (where  $\vec{v}$  is the vector defining  $U$ ) implies that  $A^\ell \vec{v}$  is a linear combination of  $\vec{v}, \dots, A^{\ell-1} \vec{v}$ , which is only possible if  $\dim(U) \leq \deg p$ . Combined with part (c), this shows  $\dim U = \deg p$ .

For (e), when  $U = \{\vec{v}, A\vec{v}, \dots, A^{d-1}\vec{v}\}$ , condition (4.1) implies that  $p(A)\vec{v} = 0$  since  $\vec{v} \in U$ . Conversely, any element  $\vec{u} \in U$  is a linear combination of  $A^j \vec{v}$  for  $j = 0, \dots, d-1$ . But then,  $p(A)\vec{v} = 0$  implies (since  $A^j$  and  $p(A)$  commute) that  $p(A)A^j \vec{v} = 0$  for any  $j$ , and hence  $p(A)\vec{u} = 0$  for any  $\vec{u} \in U$ .  $\square$

*Proof of Theorem 4.2, left and right orthogonal  $A$ -invariant subspaces.* Similar to the proof of Theorem 4.1, introduce the orthogonal matrix  $\mathbf{O} = (\mathbf{U} \mid \mathbf{W} \mid \mathbf{V}) \in \mathbb{R}^{s \times s}$ , where the columns of  $\mathbf{U} \in \mathbb{R}^{s \times d_u}$  and  $\mathbf{V} \in \mathbb{R}^{s \times d_v}$  form an orthogonal basis for  $U$  and  $V$ , respectively (the columns of  $\mathbf{W}$  span the remaining space). Now  $A$  block-diagonalizes as

$$\mathbf{O}^T A \mathbf{O} = \begin{pmatrix} A_{UU} & A_{UW} & A_{UV} \\ 0 & A_{WW} & A_{WV} \\ 0 & 0 & A_{VV} \end{pmatrix},$$

where  $A_{\Sigma\Theta} = \Sigma^T A \Theta$  where  $\Sigma, \Theta \in \{\mathbf{U}, \mathbf{V}, \mathbf{W}\}$ . Hence,

$$(A.2) \quad \text{char}_A(x) = \text{char}_{A_{UU}}(x) \text{char}_{A_{WW}}(x) \text{char}_{A_{VV}}(x).$$

Via Theorem 4.1 the polynomial  $p(x)$  is the minimal polynomial of  $A_{UU}$ . Theorem 4.1 applies to  $A^T$  with space  $\mathbf{V}$ , so that  $q(x)$  is the minimal polynomial of  $A_{VV}^T$  (and is the same as the minimal polynomial of  $A_{VV}$ ). Hence,  $p(x)q(x)$  divides  $\text{char}_{A_{WW}}(x)\text{char}_{A_{VV}}(x)$ , which, from (A.2), divides  $\text{char}_A(x)$ .  $\square$

It is generally not true that  $p(x)q(x)$  divides the minimal polynomial of  $A$ .

**Appendix B. Proofs of Hankel matrix determinant and orthogonal polynomial recurrence relation.** Here we provide details for the determinant computation ( $m \in \mathbb{Z}_{\geq 0}$ ) of

$$(B.1) \quad \mathbf{H}_n(m) = \begin{pmatrix} \frac{1}{m!} & \frac{1}{(m+1)!} & \cdots & \frac{1}{(m+n-1)!} \\ \frac{1}{(m+1)!} & \ddots & & \frac{1}{(m+n)!} \\ \vdots & & \ddots & \vdots \\ \frac{1}{(m+n-1)!} & \cdots & \cdots & \frac{1}{(m+2n-2)!} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Let  $M_n(m) := \det(\mathbf{H}_n(m))$ . A formula (without proof) for  $M_n(m)$  is given in [23]:

$$(B.2) \quad M_n(m) = \sigma(n) \frac{c(n)c(m+n-1)}{c(m+2n-1)}, \quad \text{where } c(n) := \prod_{i=1}^{n-1} i!,$$

and  $\sigma(n) := \begin{cases} 1 & \text{if } r = 0 \text{ or } 1, \\ -1 & \text{if } r = 2 \text{ or } 3, \end{cases} \quad \text{where } n \equiv r \pmod{4}.$

Since we could not find a proof of (B.2) in the literature, we provide a brief one here. Following the approach in [22, Method 2 in section 4], for any square matrix  $A$ , a determinant formula due to Jacobi reads

$$(B.3) \quad \det A \cdot \det A_{1,n}^{1,n} = \det A_1^1 \cdot \det A_n^n - \det A_1^n \cdot \det A_n^1,$$

where  $A_i^j$  denotes the submatrix of  $A$  with row  $i$  and column  $j$  removed. Applying (B.3) to the matrix  $M_n(m)$  and using the symmetry of the Hankel matrix yields the recursion:

$$(B.4) \quad M_n(m)M_{n-2}(m+2) = M_{n-1}(m+2)M_{n-1}(m) - M_{n-1}(m+1)^2.$$

Formula (B.4) relates  $M_n(m)$  to matrices of size  $n-1$  and  $n-2$ , and can be used to prove (B.2) via induction. Note that for  $M_1(m) = 1/m!$  and  $M_2(m) = -[(m+1)!(m+2)!]^{-1}$ , formula (B.2) is readily verified (and holds  $\forall m \geq 0$ ). Assume (B.2) holds for  $1 \leq k < n$ . We can then divide (B.4) through by  $M_{n-1}(m+1)M_{n-2}(m+2)$  (which are nonzero by assumption) to obtain

$$(B.5) \quad \frac{M_n(m)}{M_{n-1}(m+1)} = \frac{M_{n-1}(m+2)M_{n-1}(m)}{M_{n-2}(m+2)M_{n-1}(m+1)} - \frac{M_{n-1}(m+1)}{M_{n-2}(m+2)}.$$

It is then a matter of substituting the formulas from (B.2) into (B.5) to verify (after several lines of factorial cancellations) that the following holds:<sup>8</sup>

$$\frac{(n-1)!}{(m+2n-2)!} = - \left( \frac{(n-2)!(m+n-1)!(m+2n-3)!}{(m+2n-2)!(m+2n-3)!(m+n-2)!} - \frac{(n-2)!}{(m+2n-3)!} \right).$$

Since (for every  $n$ ) the Hankel matrix determinants  $M_n(m)$  do not vanish, the entries of  $\mathbf{H}_n(m)$  define a *quasi-definite linear functional* which then have associated orthogonal polynomials. Monic orthogonal polynomials (always) satisfy a three term linear recurrence of the general form

$$(B.6) \quad Q_{n+1}(x) = (x + \gamma_n)Q_n(x) + \beta_n Q_{n-1}(x),$$

<sup>8</sup>Note that  $\sigma(n)/\sigma(n-1)$  always has opposite sign to  $\sigma(n-1)/\sigma(n-2)$ .

where  $\mathbf{H}_n(m)$  is initialized via  $Q_0(x) = 1$ ,  $Q_1(x) = x - 1/(m+1)$ . We now obtain formulas for  $\beta_n$ ,  $\gamma_n$  via determinant computations. From [22, Thm. 29] we have

$$\beta_n(m) = -\frac{M_{n-1}(m)}{M_n(m)} \cdot \frac{M_{n+1}(m)}{M_n(m)} = \frac{n(m+n-1)}{(m+2n)(m+2n-1)^2(m+2n-2)},$$

When  $m=1$ :  $\beta_n = \frac{1}{4(4n^2-1)}.$

The  $\gamma_n$  depend only on the second leading term  $\lambda_n(m)$  of  $Q_n(x) = x^n + \lambda_n(m)x^{n-1} +$  low order terms. That is, evaluating the coefficient of  $x^n$  in (B.6) yields  $\gamma_n(m) = \lambda_{n+1}(m) - \lambda_n(m)$ .

Orthogonality of  $Q_j(x)$  demands that the coefficients  $\vec{\lambda} = (\lambda_0(m), \dots, \lambda_n(m))^T$  satisfy  $\mathbf{H}_n(m)\vec{\lambda} = \vec{h}_n$  with  $\vec{h}_n = -(1/(m+n)!, \dots, 1/(m+2n+1)!)^T$ . Using Cramer's rule for  $\lambda_n(m)$  and (B.3), one obtains the two term recursion for  $\lambda_n(m)$

$$\begin{aligned} \lambda_n(m)M_n(m)M_{n-2}(m+2) &= \lambda_{n-1}(m+2)M_{n-1}(m+2)M_{n-1}(m) \\ &\quad - \lambda_{n-1}(m+1)M_{n-1}(m+1)^2, \end{aligned}$$

which simplifies to

$$\lambda_n(m) = \frac{(m+2n-2)}{(n-1)}\lambda_{n-1}(m+1) - \frac{(m+n-1)}{(n-1)}\lambda_{n-1}(m+2).$$

Induction then shows that

$$\lambda_n(m) = -\frac{n}{m+2n-1} \quad \forall n \geq 1, m \geq 0.$$

The orthogonal polynomials in this paper use the values  $m=0, 1$ . In the case when  $m=1$ , the coefficient  $\lambda_n(1) = -\frac{1}{2}$  is constant, in which case  $\gamma_n(1) = 0 \quad \forall n \geq 1$ .

**Acknowledgments.** The authors gratefully acknowledge the feedback from two anonymous reviewers whose suggestions led to improvements in this paper. Shirokoff gratefully acknowledges the hospitality of Robert Pego and the Department of Mathematical Sciences at Carnegie Melon University where portions of this work were completed.

## 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 (1996), pp. 777–782, <https://doi.org/10.1137/S1064827595282520>.
- [2] I. ALONSO-MALLO, *Runge-Kutta methods without order reduction for linear initial boundary value problems*, Numer. Math., 91 (2002), pp. 577–603.
- [3] I. ALONSO-MALLO AND B. CANO, *Avoiding order reduction of Runge-Kutta discretizations for linear time-dependent parabolic problems*, BIT, 44 (2004), pp. 1–20.
- [4] I. ALONSO-MALLO AND C. PALENCIA, *Optimal orders of convergence for Runge-Kutta methods and linear, initial boundary value problems*, Appl. Numer. Math., 44 (2003), pp. 1–19.
- [5] S. AXLER, *Linear Algebra Done Right*, 3rd ed., Undergraduate Texts in Mathematics, Springer, 2015.
- [6] A. BISWAS, *Structure-Optimized and Order-Preserving Methods for Advection and Time Stepping*, Ph.D. thesis, Temple University, Philadelphia, PA, 2021.
- [7] A. BISWAS, D. KETCHESON, B. SEIBOLD, AND D. SHIROKOFF, *Design of DIRK schemes with high weak stage order*, Commun. Appl. Math. Comput. Sci., 18 (2023), pp. 1–28.
- [8] K. BURRAGE, W. HUNSDORFER, AND J. VERWER, *A study of B-Convergence of Runge-Kutta methods*, Computing, 36 (1986), pp. 17–34.

- [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 (1990), pp. 447–456, <https://doi.org/10.1137/0727027>
- [10] J. C. BUTCHER AND N. GOODWIN, *Numerical Methods for Ordinary Differential Equations*, 2nd ed., John Wiley & Sons, Chichester, 2008.
- [11] M. CALVO AND C. PALENCIA, *Avoiding the order reduction of Runge–Kutta methods for linear initial boundary value problems*, Math. Comp., 71 (2002), pp. 1529–1543.
- [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 (1995), pp. 1241–1252, <https://doi.org/10.1137/0916072>.
- [13] M. CROUZEIX AND P.-A. RAVIART, *Approximation des problèmes d'évolution*, Lecture Notes, University Rennes, Rennes, France, 1980.
- [14] K. DEKKER AND J. G. VERWER, *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*, CWI Monographs 2, North-Holland Publishing, Amsterdam, 1984.
- [15] J. GARCÍA-ARDILA, F. MARCELLÁN, AND M. MARRIAGA, *From standard orthogonal polynomials to Sobolev orthogonal polynomials: The role of semiclassical linear functionals*, in *Orthogonal Polynomials*, AIMS/SVSW 2018, M. Foupouagnigni and W. Koepf, eds., Birkhäuser, Cham, 2020, pp. 245–292.
- [16] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations I (2nd Revised. Ed.): Nonstiff Problems*, Springer-Verlag, New York, 1993.
- [17] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II*, Springer, Berlin Heidelberg, 1996.
- [18] W. HUNSDORFER AND M. SPIJKER, *A note on B-stability of Runge–Kutta methods*, Numer. Math., 36 (1980), pp. 319–331.
- [19] W. H. HUNSDORFER, *Stability and B-Convergence of linearly implicit Runge–Kutta methods*, Numer. Math., 50 (1986), pp. 83–95.
- [20] M. ISMAIL, *Classical and Quantum Orthogonal Polynomials in One Variable*, Cambridge University Press, New York, 2005.
- [21] D. KETCHESON, B. SEIBOLD, D. SHIROKOFF, AND D. ZHOU, *DIRK schemes with high weak stage order*, in *Lect. Notes Comput. Sci. Eng. 134, Proceedings of ICOSAHOM 2018*, Springer, 2020, pp. 453–463.
- [22] C. KRATTENTHALER, *Advanced determinant calculus: A complement*, Linear Algebra Appl., 411 (2005), pp. 68–166.
- [23] S. LANDSBURG, *Some Hankel determinants*, MathOverflow, <https://mathoverflow.net/q/249838> (version: 2017-04-13).
- [24] C. LUBICH AND A. OSTERMANN, *Runge–Kutta approximation of quasi-linear parabolic equations*, Math. Comp., 64 (1995), pp. 601–627.
- [25] J. MONTIJANO, *Estudio de los metodos SIRK para la resolucion numérica de ecuaciones diferenciales de tipo stiff*, Ph.D. thesis, University of Zaragoza, Zaragoza, Spain, 1983.
- [26] S. P. NØRSETT AND A. WOLFBRANDT, *Attainable order of rational approximations to the exponential function with only real poles*, BIT Numer. Math., 17 (1977), pp. 200–208.
- [27] A. OSTERMANN AND M. ROCHE, *Runge–Kutta methods for partial differential equations and fractional orders of convergence*, Math. Comp., 59 (1992), pp. 403–420.
- [28] A. OSTERMANN AND M. ROCHE, *Rosenbrock methods for partial differential equations and fractional orders of convergence*, SIAM J. Numer. Anal., 30 (1993), pp. 1084–1098, <https://doi.org/10.1137/0730056>.
- [29] A. OSTERMANN AND M. THALHAMMER, *Convergence of Runge–Kutta methods for nonlinear parabolic equations*, Appl. Numer. Math., 42 (2002), pp. 367–380.
- [30] P. PETERSEN, *Linear Algebra*, Springer, 2021.
- [31] P. PRINCE AND J. DORMAND, *High order embedded Runge–Kutta formulae*, J. Comput. Appl. Math., 7 (1981), pp. 67–75.
- [32] 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 (1974), pp. 145–162.
- [33] J. RANG, *An analysis of the Prothero–Robinson example for constructing new DIRK and ROW methods*, J. Comput. Appl. Math., 262 (2014), pp. 105–114.
- [34] J. RANG, *The Prothero and Robinson example: Convergence studies for Runge–Kutta and Rosenbrock–Wanner methods*, Appl. Numer. Math., 108 (2016), pp. 37–56.
- [35] S. ROBERTS, *Multimethods for the Efficient Solution of Multiscale Differential Equations*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2021.
- [36] S. ROBERTS AND A. SANDU, *Eliminating Order Reduction on Linear, Time-Dependent ODEs with GARK Methods*, preprint, <https://arxiv.org/abs/2201.07940>, 2022.



- [37] R. R. ROSALES, B. SEIBOLD, D. SHIROKOFF, AND D. ZHOU, *Spatial manifestations of order reduction in Runge-Kutta methods for initial boundary value problems*, Comm. Math. Sci., (2023), (in press).
- [38] S. J. RUUTH AND R. J. SPITERI, *High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations*, SIAM J. Numer. Anal., 42 (2004), pp. 974–996, <https://doi.org/10.1137/S0036142902419284>.
- [39] 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 (1986), pp. 405–418.
- [40] S. SCHOLZ, *Order barriers for the B-convergence of ROW methods*, Computing, 41 (1989), pp. 219–235.
- [41] L. M. SKVORTSOV, *How to avoid accuracy and order reduction in Runge-Kutta methods as applied to stiff problems*, Comput. Math. Math. Phys., 57 (2017), pp. 1124–1139.
- [42] J. VERWER AND J. SANZ-SERNA, *Convergence of method of lines approximations to partial differential equations*, Computing, 33 (1984), pp. 297–313.