On recent advances of spectral analysis for systems arising from fully-implicit RK methods

Michal Outrata[1,*]

[1] Sokolovská 83, Praha 8, 186 75, Czechia

This work deals with two groups of spectral analysis results for matrices arising in fully implicit Runge-Kutta methods used for linear time-dependent partial differential equations. These were applied for different formulations of the same problem and used different tools to arrive at results that do not immediately coincide. We show the equivalence of the results as well as the equivalence of the approaches, unifying the two directions.

1 Introduction

In recent years we have seen a renewed interest in using fully implicit Runge-Kutta (IRK) methods as numerical integrators for (linear) time-dependent partial differential equations (PDEs), see [1–7], among others and also [8, 9] for a more complete overview of Runge-Kutta methods in the context of numerical solvers for ordinary differential equations (ODEs). We will consider a time-dependent PDE in the form

$$\begin{aligned} \frac{\partial}{\partial t}u &= \mathcal{L}u + f \quad \text{in } \Omega \times (0,T), \\ \mathcal{B}u(\mathbf{x},t) = g(\mathbf{x},t) \quad &\text{on } \partial\Omega \times (0,T), \qquad u(\mathbf{x},0) = u^{(0)}(\mathbf{x}) \quad \text{in } \Omega, \end{aligned} \tag{1}$$

on a given connected, bounded domain $\Omega \subset \mathbb{R}^d$ and $(0,T) \subset \mathbb{R}$ for some given data f, g and $u^{(0)}$. Moreover, we will assume that the spatial operator $\mathcal{L}$ is bounded, coercive and self-adjoint. We note that these properties can be relaxed to other relevant cases, see [5, 7]. We first discretize in space using, e.g., a finite elements (FE) scheme, obtaining a system of n ODEs

$$\frac{\partial}{\partial t}M\mathbf{u}(t) = K\mathbf{u}(t) + \mathbf{b}^{(ST)}(t), \quad \text{with} \quad \mathbf{u}(0) = \mathbf{u}^{(0)},$$

where $M, K \in \mathbb{R}^{n\times n}$ are the mass and stiffness matrices (corresponding to the chosen FE scheme and the operator $\mathcal{L}$), the vector function $\mathbf{b}^{(ST)}(t)$ aggregates the contributions of both $f(t)$ and $g(t)$ and the vector $\mathbf{u}^{(0)}$ corresponds to the initial condition $u^{(0)}(\mathbf{x})$. Any IRK method is then given by the number of stages $s \in \mathbb{N}$ and its Butcher table, succinctly written as

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}, \quad A \in \mathbb{R}^{s\times s}, \quad \text{and} \quad \mathbf{b}, \mathbf{c} \in \mathbb{R}^s.$$

For a given timestep τ, the IRK method progresses the solution forward in time at timepoints $t_m := \tau m$ using the approximation $\mathbf{u}(t_m) \approx \mathbf{u}^{(m)}$ with

$$\mathbf{u}^{(m)} := \mathbf{u}^{(m-1)} + \tau \sum_{i=1}^{s} \mathbf{b}_i \mathbf{k}_i^{(m)},$$

where the so-called stage-functions $\mathbf{k}_i^{(m)}$ satisfy

$$M\mathbf{k}_i^{(m)} = \mathbf{b}^{(ST)}(t_{m-1} + c_i\tau) + K\mathbf{k}_i^{(m-1)} + \tau \sum_{j=1}^{s} a_{ij} K\mathbf{k}_j^{(m)}, \qquad i = 1, \dots, s. \tag{2}$$

For the so-called fully IRK methods, the Butcher matrix $A = [a_{ij}]$ is dense, i.e., $a_{ij} \neq 0$ for all $i, j = 1, \dots, s$, so that (2) becomes a rather large system. We start by rewriting (2) using a Kronecker product notation as

* Corresponding author: e-mail outrata@karlin.mff.cuni.cz, phone +420 951 553 296

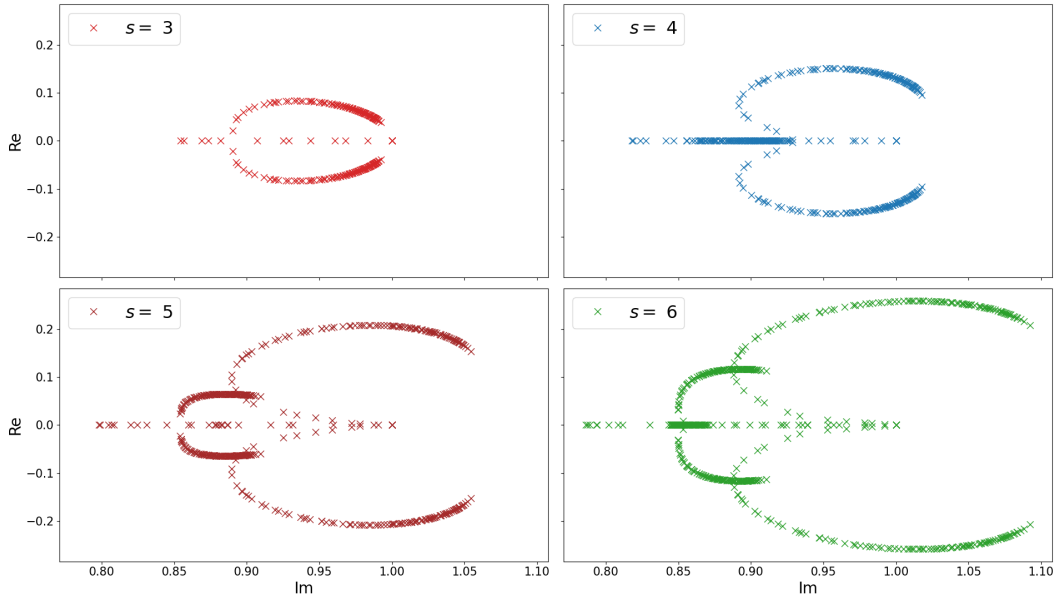

Fig. 1: The spectra of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$ for the preconditioner $\mathcal{P}$ given in (7) below, using the RadauIIA IRK method. The spatial operator is the Laplacian on an irregular domain Ω with various boundary conditions (Dirichlet, Neumann and Robin), discretized using conforming P1 FEM, see [7, Section 4, Example 2] for detailed description.

$$(I \otimes M + \tau A \otimes K)\,\mathbf{k}^{(m)} = (I \otimes M)\,\mathbf{k}^{(m-1)} + \left[\mathbf{b}^{(ST)}(t_{m-1} + c_1\tau)^T, \ldots, \mathbf{b}^{(ST)}(t_{m-1} + c_s\tau)^T\right]^T, \tag{3}$$

where $\mathbf{k}^{(m-1)} = \left[(\mathbf{k}_1^{(m-1)})^T, \cdots, (\mathbf{k}_s^{(m-1)})^T\right]^T$. Next, we transform (3) by factoring out $A \otimes I$ on the left-hand side to the left and multiplying the equation with its inverse, obtaining

$$\underbrace{\left(A^{-1} \otimes M + \tau I \otimes K\right)}_{=:\,\mathcal{A}} \mathbf{k}^{(m)} = \mathbf{b}_{\mathrm{IRK}}^{(m-1)}, \tag{4}$$

so that $\mathbf{b}_{\mathrm{IRK}}^{(m-1)}$ is defined as the right-hand side vector in (3) multiplied by $A^{-1} \otimes I$ from the left. To the best of our knowledge this was first proposed by Butcher in [10] and later became the standard, see [1, 3–5, 11–15]; the intuition on why this is useful is summarized in, e.g., [13, Section 4] or [15, Section 4.1]. The go-to solver for (4) has been the GMRES method (see [16, Sections 2.2, 5.7 and 5.10]) with a suitable preconditioner and in [17], the authors considered a generic class of block-preconditioners (there formulated for (3)) given by

$$\mathcal{P} = \tilde{A} \otimes M + \tau I \otimes K. \tag{5}$$

The authors proposed some convenient choice of $\tilde{A}$, which helped to stimulate the development of related preconditioners, notably the works [1, 11, 12, 14]. As the GMRES convergence is in practice often linked with the spectral properties of the (preconditioned) system matrix (see [16, Sections 2.2, 5.7 and 5.10], also for the limits of this analysis), the authors presented also plots showing "favorable properties"[1] of the spectrum. For illustration, we include an example in Figure 1.

As far as we are aware, there have been two independent series of works that aimed at analyzing preconditioners of the type (5) – namely their spectral properties – the first coming from the group originally led by the late Owe Axelsson ([5, 13]) and the second from the group of Martin J. Gander ([7, 19]). The purpose of this work is to analyze their overlap and show their equivalence.

Both of these works are based on the spectral properties of the matrix pencil $\tau K - \mu M$ (we replace the standard symbol λ for the generalized eigenvalue by μ), i.e., on the eigendecomposition of the matrix $\tau M^{-1}K$. As the stiffness and mass matrices come from a discretization of a coercive, self-adjoint, bounded operator $\mathcal{L}$, we assume that the pencil $\tau K - \mu M$ is

[1] In this context we would like to recall the classical result from [18] that states that *any* GMRES convergence can be observed for a system with a matrix with a given spectrum, i.e., spectrum on its own is *not* sufficient to say anything about GMRES behavior. In practice, however, this is *rarely* observed and many GMRES users use the folklore of "better-clustered spectrum suggests faster convergence", without further, case-specific justification. More details for this particular setting can be found in [7].

symmetric and positive-definite so that there exist an M-orthogonal eigenbasis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of τK, i.e.,

$$M^{-1}K = Q \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{bmatrix} Q^T =: QDQ^T, \quad \text{with} \quad Q = [\mathbf{q}_1, \dots, \mathbf{q}_n],\ D = \operatorname{diag}(\mu_1, \dots, \mu_n), \tag{6}$$

where $0 < \mu_1 \le \ldots \le \mu_n$ are the generalized eigenvalues of the pencil $\tau K - \mu M$, i.e., of the matrix $\tau M^{-1}K$, see [20, Sections 2.3 and 5] for more details and further references.

2 Formulas for the eigenproperties of the preconditioned system

2.1 Polynomial approach

This direction was initiated by the following observation of prof. Axelsson

> *Taking the RadauIIA Butcher matrix A for any $s = 2, \dots, 10$, the lower-triangular part of A^{-1} is dominating the rest of the matrix (in terms of magnitude of the entries). Similarly, the LD factor of the LDU factorization of A^{-1} is dominating the U factor (in both the spectral and the Frobenius norm).*

Heuristically, this suggests that these carry the majority of the information of A^{-1} for the RadauIIA method and as such are reasonable choices for $\tilde{A}$ for the preconditioner construction. This has been numerically observed also in [1, 11], only for A itself, rather than A^{-1}. Writing the LDU factorization $A^{-1} = \tilde{L}\tilde{D}U$ we set $L := \tilde{L}\tilde{D} = [l_{ij}]$ so that U is a unit upper-triangular matrix, i.e., we have $U = I + \hat{U}$ for some strictly upper-triangular matrix $\hat{U} = [\hat{u}_{ij}]$. Taking $\tilde{A} = L$ in (5), we obtain

$$\begin{aligned} \mathcal{P}^{-1}\mathcal{A} &= (L \otimes M + \tau I \otimes K)^{-1} \left(L \otimes M + I \otimes \tau K + (L\hat{U}) \otimes M \right) \\ &= I + \left(I + \tau L^{-1} \otimes M^{-1}K \right)^{-1} \left(\hat{U} \otimes I \right) =: I + W_1^{-1} W_2, \end{aligned} \tag{7}$$

see [13, equation (14)] or [5, Section 4] for further details. In [13], the authors consider the RadauIIA IRK and show that for $s = 2$ the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ lie inside the disc centered at 1 with the diameter $\|\hat{U}\| < 1$ and conjecture this localization for general s. In [5, Section 4], the authors improve the localization of the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ for the RadauIIA IRK, using the matrix pencil $W_2 - \lambda W_1$ (notice that W_1, W_2 are neither symmetric nor positive-definite). Namely, having a generalized eigenpair $(\lambda, \mathbf{v})$ of the pencil $W_2 - \lambda W_1$ we have

$$\mathcal{P}^{-1}\mathcal{A}\mathbf{v} = (1+\lambda)\mathbf{v} \quad \iff \quad W_2\mathbf{v} = \lambda W_1 \mathbf{v}. \tag{8}$$

The authors then work with $W_2\mathbf{v} = \lambda W_1\mathbf{v}$ as with an equation for $(\lambda, \mathbf{v})$, i.e., the aim is to solve

$$\begin{bmatrix} 0 & \hat{u}_{1,2}I & \dots & \hat{u}_{1,s}I \\ & \ddots & \ddots & \vdots \\ & & \ddots & \hat{u}_{s-1,s}I \\ & & & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \vdots \\ \mathbf{v}_s \end{bmatrix} = \lambda \left(\begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \vdots \\ \mathbf{v}_s \end{bmatrix} + \begin{bmatrix} \ell_{11} \cdot \tau M^{-1}K & & & \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ \ell_{s1} \cdot \tau M^{-1}K & \dots & \dots & \ell_{ss} \cdot \tau M^{-1}K \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \vdots \\ \mathbf{v}_s \end{bmatrix} \right) \tag{9}$$

for $(\lambda, \mathbf{v})$. In [5, Section 4.3], the authors first observe n "trivial" solutions corresponding to $\lambda = 0$. To resolve the remaining $(s-1)n$ eigenpairs, the authors proceed with a *symbolic* block backward substitution. This is presented for $s = 2, 3$ in Sections 4.1 and 4.2 and by analogy the authors argue that such a routine can be carried out for any s. As there is little space devoted to this argument in [5, Section 4.3], we derive it here ourselves, adhering to the techniques used in [5, Section 4].

Looking at the i-th ($1 \le i \le s$) block-row in (9), we can rearrange it as

$$\lambda(I + \ell_{ii} \cdot \tau M^{-1}K)\mathbf{v}_i = \sum_{j=i+1}^{s} \hat{u}_{i,j}\mathbf{v}_j - \sum_{j=1}^{i-1} \lambda \ell_{i,j} \cdot \tau M^{-1}K\mathbf{v}_j, \tag{10}$$

and notice that for $i = s$, λ factors out from both sides so that $\mathbf{v}_s$ can be expressed only in terms of $\mathbf{v}_1, \dots, \mathbf{v}_{s-1}$, i.e., independent of λ. This corresponds to the first step of the mentioned symbolic backward substitution. We proceed with the

substitution for $i = s-1, \dots, 1$. When dealing with the i-th block-row we assume to have expressions for $\mathbf{v}_s, \dots, \mathbf{v}_{i+1}$ in terms of $\mathbf{v}_1, \dots, \mathbf{v}_i$ and we insert these into (10), obtaining

$$\left[\lambda(I + \ell_{ii} \cdot \tau M^{-1}K) - H^{(i)}\right] \mathbf{v}_i = \sum_{j=1}^{i-1} \left(G_j^{(i)} - \lambda \ell_{i,j} \cdot \tau M^{-1}K\right) \mathbf{v}_j, \tag{11}$$

for some appropriate n-by-n matrices $H^{(i)}, G_j^{(i)}$ based on (10). For example,

$$\begin{aligned} H^{(s-1)} &= -\hat{u}_{s-1,s}(I + \ell_{ss} \cdot \tau M^{-1}K)^{-1} \ell_{s,s-1} \cdot \tau M^{-1}K, \\ G_j^{(s-1)} &= -\hat{u}_{s-1,s}(I + \ell_{ss} \cdot \tau M^{-1}K)^{-1} \ell_{s,j} \cdot \tau M^{-1}K. \end{aligned} \tag{12}$$

From (11) we obtain an expression for $\mathbf{v}_i$ in terms of $\mathbf{v}_{i-1}, \dots, \mathbf{v}_1$ and after s steps arrive at

$$\underbrace{\left[\lambda(I + \ell_{11} \cdot \tau M^{-1}K) - H^{(1)}\right]}_{=: \tilde{F}_{\tau M^{-1}K}(\lambda)} \mathbf{v}_1 = 0. \tag{13}$$

This process "sandwiches" the inversion of the matrices $\lambda(I + \ell_{ii} \cdot \tau M^{-1}K) - H^{(i)}$ and the multiplication with matrices $G_j^{(i)} - \lambda \ell_{i,j} \cdot \tau M^{-1}K$. In other words, $\tilde{F}_{\tau M^{-1}K}(\lambda)$ is built from the building blocks of matrices $\tau M^{-1}K$ and I by repeated applications of linear combination and inversion. This is, at least in our understanding, the key observation behind the analysis in [5] and has several relevant implications.

First, in this sense $\tilde{F}_{\tau M^{-1}K}(\lambda)$ is a rational function of $\tau M^{-1}K$ and also a rational function of λ. Second, all matrix products in the definition of $\tilde{F}_{\tau M^{-1}K}(\lambda)$ commute, for example we have

$$\begin{aligned} H^{(s-2)} &= \hat{u}_{s-2,s} G_{s-2}^{(s-1)} - \left(\hat{u}_{s-2,s-1} + \hat{u}_{s-2,s}(I + \ell_{ss} \cdot \tau M^{-1}K)^{-1} \ell_{s,s-1} \cdot \tau M^{-1}K\right) \cdot \\ &\quad \cdot \left[\lambda(I + \ell_{s-1,s-1} \cdot \tau M^{-1}K) - H^{(s-1)}\right]^{-1} \left(G_{s-2}^{(s-1)} - \lambda \ell_{s-1,s-2} \cdot \tau M^{-1}K\right) \\ &= \hat{u}_{s-2,s} G_{s-2}^{(s-1)} - \left[\lambda(I + \ell_{s-1,s-1} \cdot \tau M^{-1}K) - H^{(s-1)}\right]^{-1} \cdot \\ &\quad \cdot \left(\hat{u}_{s-2,s-1} + \hat{u}_{s-2,s}(I + \ell_{ss} \cdot \tau M^{-1}K)^{-1} \ell_{s,s-1} \cdot \tau M^{-1}K\right) \left(G_{s-2}^{(s-1)} - \lambda \ell_{s-1,s-2} \cdot \tau M^{-1}K\right). \end{aligned} \tag{14}$$

In particular, the inversed matrices "sandwiched inside" $H^{(1)}$ can freely "travel to the left" within the sandwiches, i.e., we can move all the inversed matrices in the matrix products inside $\tilde{F}_{\tau M^{-1}K}(\lambda)$ (i.e., inside $H^{(1)}$) "to the left", similarly to (14). After this rearrangement is finished, we can multiply (also from the left) with those matrices, transforming (13) so as to get rid of all inversions. For example, in (14) we would first multiply by $\lambda(I + \ell_{s-1,s-1} \cdot \tau M^{-1}K) - H^{(s-1)}$ and then with $I + \ell_{ss} \cdot \tau M^{-1}K$ (which is also present inside $H^{(s-1)}$). This corresponds to expanding back the backward substitution process and in this process we transform (13) to a new equation, let us denote it

$$F_{\tau M^{-1}K}(\lambda)\mathbf{v}_1 = 0. \tag{15}$$

Already for $s = 3$ (with $H^{(1)} = H^{(s-2)}$ given in (14)) the calculation becomes somewhat tedious and it is a downright unpleasant chore for larger number of stages. However, it allows us to focus on the quantity of interest – the matrix $F_{\tau M^{-1}K}(\lambda)$.

By construction, each of the inversed matrices consisted of linear combination of (a) terms linear in λ and (b) other inversed matrices. As we did s-step backward substitution, the above process of transforming $\tilde{F}_{\tau M^{-1}K}(\lambda)$ into $F_{\tau M^{-1}K}(\lambda)$ introduces at most s multiplications – the first $s-1$ of them will introduce a linear factor in λ, while the last (corresponding to the bottom-most block-row of (9)) is independent of λ. In other words, $F_{\tau M^{-1}K}(\lambda)$ is a polynomial function in λ of degree $s-1$. In fact, the construction showcases that $F_{\tau M^{-1}K}(\lambda)$ is the numerator of the rational function $\tilde{F}_{\tau M^{-1}K}(\lambda)$ with respect to λ. Also by construction,

$$\mathrm{Ker}\left(F_{\tau M^{-1}K}(\lambda)\right) = \mathrm{Ker}\left(\tilde{F}_{\tau M^{-1}K}(\lambda)\right), \tag{16}$$

since we always multiplied by non-singular matrices with trivial kernels. Linking this back to the preconditioned system, we observe that $(1+\lambda, \mathbf{v})$ is an eigenpair of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$ if and only if $\mathbf{v}_1 \in \mathrm{Ker}\left(F_{\tau M^{-1}K}(\lambda)\right)$. This shows there has to be a very close link between the matrix polynomial $F_{\tau M^{-1}K}(\lambda)$ and the characteristic polynomial of $\mathcal{P}^{-1}\mathcal{A}$ (after the simple change of variables $\lambda = 1 + \lambda$, see (8)).

A third consequence of this special structure of $\tilde{F}_{\tau M^{-1}K}(\lambda)$ (which extends to $F_{\tau M^{-1}K}(\lambda)$) is that $\tilde{F}_{\tau M^{-1}K}(\lambda)$ must diagonalize in the eigenbasis Q of $\tau M^{-1}K$, i.e., we can rewrite (15) as

$$Q^T\tilde{F}_{\tau M^{-1}K}(\lambda)QQ^T\mathbf{v}_1 \equiv \begin{bmatrix}\tilde{F}_{\mu_1}(\lambda) & & \\ & \ddots & \\ & & \tilde{F}_{\mu_n}(\lambda)\end{bmatrix} Q^T\mathbf{v}_1 = 0,$$

$$Q^TF_{\tau M^{-1}K}(\lambda)QQ^T\mathbf{v}_1 \equiv \begin{bmatrix}F_{\mu_1}(\lambda) & & \\ & \ddots & \\ & & F_{\mu_n}(\lambda)\end{bmatrix} Q^T\mathbf{v}_1 = 0,$$

where $\tilde{F}_{\mu_k}(\lambda)$ (or $F_{\mu_k}(\lambda)$) are now *scalar* rational (or polynomial) functions of λ with the precise structure of $\tilde{F}_{\tau M^{-1}K}(\lambda)$ (or $F_{\tau M^{-1}K}(\lambda)$), only replacing the matrices $\tau M^{-1}K$ and I by μ_k and 1. Hence, the coefficients of $\tilde{F}_{\mu_k}(\lambda)$ (or $F_{\mu_k}(\lambda)$) are themselves rational functions in μ_k. Crucially, this showcases the connection between the matrix $F_{\tau M^{-1}K}(\lambda)$ and the characteristic polynomial of $\mathcal{P}^{-1}\mathcal{A}$. By the same argument as in (16) we get

$$\det\left(F_{\tau M^{-1}K}(\lambda)\right) = \alpha_{\tau M^{-1}K}\det\left(\tilde{F}_{\tau M^{-1}K}(\lambda)\right),$$

for some scaling coefficient $\alpha_{\tau M^{-1}K} \neq 0$ and thereby we see that if $\det\left(F_{\tau M^{-1}K}(\lambda)\right) = 0$, then λ is one of the $s(n-1)$ generalized eigenvalues of the pencil $W_2 - \lambda W_1$ we are looking for. Writing

$$\hat{p}_{\mathrm{char}}(\lambda) := \det\left(F_{\tau M^{-1}K}(\lambda)\right) = \det\begin{bmatrix}F_{\mu_1}(\lambda) & & \\ & \ddots & \\ & & F_{\mu_n}(\lambda)\end{bmatrix} = \prod_{k=1}^{n} F_{\mu_k}(\lambda), \tag{17}$$

it follows that $p_{\mathrm{char}}(\lambda) := \lambda^n\hat{p}_{\mathrm{char}}(\lambda)$ is (up to a rescaling) the characteristic polynomial of the pencil $W_2 - \lambda W_1$, and therefore $p_{\mathrm{char}}(\lambda - 1)$ is (up to a rescaling) the characteristic polynomial of the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$. Due to the structure of $\hat{p}_{\mathrm{char}}(\lambda)$ in (17) it becomes natural to index the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ (i.e., the zeros of p_{char}) by the eigenvalues μ_k of $\tau M^{-1}K$.

Remark 2.1 While not stated this way, the above was clearly grasped in [5, Sections 4 and 5]. There, the above derivation is essentailly skipped by taking $\mathbf{v}_1 = \mathbf{q}_1$. This somewhat simplifies the calculation compared to (11-15) but not to an agreeable level. This is also illustrated by the fact that the numerical experiments in [5] are carried out only for $s = 2, 3$. Comparing with [5, Sections 4.3 and Theorem 5.1], the above is, in our eyes, slightly more constructive way to arrive at the same result. We also believe that it highlights more clearly the key mechanic of the derivation.

However, since the analysis in [5, Sections 4 and 5] relies on the *construction* of p_{char}, the above seems not fully satisfactory – as mentioned above, it is a laborious task to obtain the symbolic formulas beyond $s = 2, 3$ and this practical aspect was not addressed; authors simply state that this symbolic approach must work for any s.

A slight simplification can be achieved as follows. We return to (9) and pass *blockwise* into the eigenbasis Q directly there, obtaining

$$\left(\hat{U} \otimes I_n\right)\mathbf{w} = \lambda\left(I_s \otimes I_n + L^{-1} \otimes D\right)\mathbf{w}, \tag{18}$$

with $\mathbf{w} = (I \otimes Q^T)\mathbf{v}$. This transposes the block backward substitution of (9) into n independent scalar ones,

$$\hat{U}\mathbf{s}_{\mu_k} = \lambda\left(I_s + \mu_k L^{-1}\right)\mathbf{s}_{\mu_k}, \tag{19}$$

again, parametrized by $\mu_k, k = 1, \ldots, n$. By construction, the characteristic polynomial of the matrix pencil $\hat{U} - \lambda\left(I_s + \mu_k L^{-1}\right)$ is (up to a rescaling) identical to $F_{\mu_k}(\lambda)$, i.e., we can calculate the coefficients of $F_{\mu_k}(\lambda)$ using a scalar version of the symbolic backward substitution – arriving at formulas from [5, Sections 4.2] for the case $s = 3$ *without* the blockwise backward substitution in (9).

Obtaining the spectrum of $\mathcal{P}^{-1}\mathcal{A}$ (or an estimate of it) then reduces to calculating (or estimating) $\mu_1, \ldots, \mu_n$, then evaluating the symbolic formulas to get the coefficients $\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^s$ of $F_{\mu_1}(\lambda), \ldots, F_{\mu_n}(\lambda)$ and calculating the roots of these polynomials using a standard numerical software, e.g., `numpy.roots`$(\mathbf{c}_k)$.

A direct observation is that, from the numerical perspective, finding the roots of a higher-degree polynomial written in the monomial basis can be a poorly-conditioned problem with respect to the polynomial coefficients. This is a well-known problem and algorithms for approximation of the roots of a polynomial given by a coefficient vector $\mathbf{c} \in \mathbb{R}^d$, such as `numpy.roots`(), circumvent the issue by calculating eigenvalues of the so-called companion matrix $\mathbf{C}_\mathbf{c} \in \mathbb{R}^{d\times d}$. In particular, the costs of finding the roots of a polynomial with coefficients $\mathbf{c} \in \mathbb{R}^d$ are governed by the costs of solving a d-dimensional

eigenvalue problem. In our setting, this means that the costs of finding the eigenvalues $\lambda_1^{(k)}, \dots, \lambda_{s-1}^{(k)}, k = 1, \dots, n$ of the matrix pencil $W_2 - \lambda W_1$ (or their estimates) after all the symbolic work has been done still amounts to solving n independent eigenvalue problems of size $(s-1)$-by-$(s-1)$. Taking a step back, we notice that we have already expressed $\lambda_1^{(k)}, \dots, \lambda_{s-1}^{(k)}$ as the eigenvalues of n independent s-by-s generalized eigenvalue problems – in the equation (19). Moreover, their construction is *direct*, given the values (or approximations of) $\mu_1, \dots, \mu_n$ – no symbolic calculations are needed.

In other words, what we gain by the explicit construction of $F_{\mu_1}(\lambda), \dots, F_{\mu_n}(\lambda)$ compared to numerically solving (19) is the reduction of the size of each of the n independent eigenvalue problems by one and the fact that we can use eigenvalue solvers tailored to companion matrices rather than generic generalized eigenvalue solver. Given that these problems are *very* small, we believe it to be more efficient and numerically stable, to avoid the explicit construction of $F_{\mu_1}(\lambda), \dots, F_{\mu_n}(\lambda)$ and instead treat these only implicitly by using an eigenvalue solver for (19). Crucially, this is not present in [5] and it is the keystone between the analysis above and the one presented in [7] as we shall see next.

2.2 Matrix approach

This direction was initiated by prof. Gander based on the plenary talk of prof. Howle at the PRECOND conference in 2019 in Minneapolis (later, some of the results appeared in [1]). Her group has been interested in preconditioners for the systems (2) for problems like (1) and she presented some strong numerical results for the block preconditioners of the type (5) (and since has expanded the focus also to hyperbolic problems, see [11, 21]). As our analysis was framed mainly by the setup in [1], we dealt with the preconditioned system

$$\left(I \otimes M + \tau \tilde{A} \otimes K\right)^{-1} (I \otimes M + \tau A \otimes K)\, \mathbf{k}^{(m+1)} = \tilde{\mathbf{b}}^{(m)},$$

rather than (4) with the preconditioner (5) – here we will write the results in the "notation" of (4) and (5).

Similarly to [5, 13], we also first studied the case $s = 2$ in detail in [19] and only later generalized the results for arbitrary s in [7], see also [22].

The analysis is built using the same observation that allowed us to pass from (18) to (19), which is based on the Kronecker product properties. In particular, we can write

$$\begin{aligned}\mathcal{P}^{-1}\mathcal{A} &= (L \otimes M + \tau I \otimes K)^{-1} \left(A^{-1} \otimes M + I \otimes \tau K\right) \\ &= (I \otimes Q^T)^{-1} \underbrace{(L \otimes I + I \otimes D)^{-1} \left(A^{-1} \otimes I + I \otimes D\right)}_{=:\, X_{\tau M^{-1} K}} (I \otimes Q^T).\end{aligned}$$

By construction, the matrix $X_{\tau M^{-1}K}$ is an ns-by-ns block matrix, each block $X^{(ij)}_{\tau M^{-1}K} \in \mathbb{R}^{n\times n}$ being *a diagonal matrix*[2], $X^{(ij)}_{\tau M^{-1}K} = \operatorname{diag}(x_1^{(ij)}, \dots, x_n^{(ij)})$. This is a very special sparsity structure and can be transformed into one we are perhaps more used to – we can permute $X_{\tau M^{-1}K}$ into a block-diagonal matrix. That is, there exists a permutation matrix Π such that

$$\Pi^T X_{\tau M^{-1}K} \Pi \equiv \Pi^T \begin{bmatrix} X^{(11)}_{\tau M^{-1}K} & \dots & X^{(1s)}_{\tau M^{-1}K} \\ \vdots & \ddots & \vdots \\ X^{(s1)}_{\tau M^{-1}K} & \dots & X^{(ss)}_{\tau M^{-1}K} \end{bmatrix} \Pi = \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_n \end{bmatrix},$$

where $X_1, \dots, X_n$ are given as

$$X_k = \begin{bmatrix} x_k^{(11)} & \dots & x_k^{(s1)} \\ \vdots & \ddots & \vdots \\ x_k^{(s1)} & \dots & x_k^{(ss)} \end{bmatrix} = (L + \mu_k)^{-1}(A^{-1} + \mu_k) \in \mathbb{R}^{s\times s}, \tag{20}$$

for $k = 1, \dots, n$, see [7, Lemma 3.1]. Importantly, X_k contains precisely the terms in $X_{\tau M^{-1}K}$ that depend on the k-th eigenspace corresponding to μ_k and is independent of $\mu_1, \dots, \mu_{k-1}, \mu_{k+1}, \dots, \mu_n$. Hence, we will use the notation X_{μ_k} rather than X_k, similarly to Section 2.1. By construction, having an eigenpair $(\lambda, \mathbf{s})$ of X_{μ_k} we have that

$$\mathcal{P}^{-1}\mathcal{A}(I \otimes Q^T)^{-1}\Pi(\mathbf{e}_k \otimes \mathbf{s}) = (I \otimes Q^T)^{-1}\Pi(\mathbf{e}_k \otimes X_{\mu_k}\mathbf{s}) = \lambda(I \otimes Q^T)^{-1}\Pi(\mathbf{e}_k \otimes \mathbf{s}), \tag{21}$$

i.e., the eigenpairs of the preconditioned system are fully characterized by those of $X_{\mu_k} \in \mathbb{R}^{s\times s}$.

[2] We choose the subscript $\tau M^{-1}K$ by analogy to Section 2.1, as they play a similar role in the respective expositions.

Remark 2.2 We would like to highlight that the matrices X_{μ_k} combined with (21) turn out to be practically useful also for estimation of the standard GMRES bound

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S) \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq \ell}} \max_{1\leq i\leq sn} |\varphi(\lambda_i)|, \tag{22}$$

where S is the matrix of eigenvectors of $\mathcal{P}^{-1}\mathcal{A}$ and $\kappa(S)$ is its condition number. As we can see in [7, Section 4], the resulting GMRES convergence estimates are very descriptive even for larger number of stages. Crucially, though, the original version of the work contained a simple error in the last conclusion drawn from (21), namely in evaluation of $\kappa(S)$. The corrected version has been submitted since and is also present in the preprints (links are on the personal webpages of both of the authors).

2.3 Connecting the two

The first point of contact between the two groups happened at the SIAM LA meeting in Paris in 2024, where I have met Ivo Dravins who was presenting the materials from [5]. This lead to several useful discussions. Among other things, these discussions led me to write this text to clearly connect the two approaches for analyzing the spectrum of $\mathcal{P}^{-1}\mathcal{A}$ – in fact, we will see next how to map one onto the other.

The key to seeing the two approaches as one is again the Kronecker product structure and arithmetic. Comparing (20) with (19), we see two sets of n eigenvalue problems of the size s-by-s, parametrized by μ_k. A direct calculation gives

$$\begin{aligned} X_{\mu_k} &= (L+\mu_k)^{-1}(A^{-1}+\mu_k) = (L+\mu_k)^{-1}(L(I+\hat{U})+\mu_k) \\ &= I + (L+\mu_k)^{-1}(L\hat{U})) = I + (I+L^{-1}\mu_k)^{-1}\hat{U}), \end{aligned} \tag{23}$$

the analogue of the manipulations in (7), only carried out with the s-by-s blocks on the diagonal, after passing into the block basis $(I\otimes Q^T)\Pi$. Following this analogy one step further, we see that the generalized eigenvalue problems in (19) are simply a redressing of the eigenvalue problems with X_{μ_k} in (20). Relating this back to the objects featuring in [5, 7], for any $k = 1, \ldots, n$ we have

$$\det\left(\lambda I - X_{\mu_k}\right) = \det\left((\lambda-1)I - (I+L^{-1}\mu_k)^{-1}\hat{U}\right).$$

Recalling that $F_{\mu_k}(\lambda)$ is obtained by factoring out λ in the first step of the (block) backward substitution, we conclude that $(\lambda-1)F_{\mu_k}(\lambda-1)$ is the characteristic polynomial of X_{μ_k} (up to a rescaling). In other words, the approach in [5] led to assembling the characteristic polynomial of the matrices used in [7], both posed in the block eigenbasis $I\otimes Q^T$.

In both of the works, the authors recognized that this approach is fully general with respect to the chosen IRK method, i.e., with respect to A, although the results naturally do depend on this choice. Moreover, the authors also recognized, that their respective formulations of the spectrum of $\mathcal{P}^{-1}\mathcal{A}$ reveal its additional structure – the eigenvalues come from a single object – the matrix X_{μ_k} (or the polynomial p_{char}) – parametrized by μ_k. Since in many cases of interest the eigenvalues μ_k sample fairly densely some interval $(\mu_{\min}, \mu_{\max})$, each eigenvalue (zero) of X_{μ_k} (p_{char}) becomes a continuous function of μ_k. In particular, the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ necessarily form *branches* – in fact s of them – where each branch tracks out one of these "smooth functions" of μ_k.

Remark 2.3 We can observe this in Figure 1 – the complex conjugate branches are densely sampled even for a very low mesh resolution ($h \approx 0.7$). For $s = 3$ we see clearly two complex conjugate branches. There are also several real eigenvalues. In fact N of them are exactly equal to $1 + 0i$ – a theoretical result established in both [5] and [7] – and these constitute the third, branch (hence we indeed have s branches). The real eigenvalues other than $1 + 0i$ highlight that, from a certain μ_k onwards, the matrices X_{μ_k} have real spectrum. In particular, there exists a $\hat{\mu}_k$ for which $X_{\hat{\mu}_k}$ has a single real eigenvalue $\hat{\lambda}_k$ with algebraic multiplicity equal to two (based on Figure 1 we have $\hat{\lambda}_k \approx 0.89 + 0i$). In this way, Figure 1 nicely illustrates that indeed the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ are only continuous (and not, e.g., $\mathcal{C}^1$) functions of μ_k. Similar behavior is present in all of the graphs in Figure 1 – we have s branches, one of them degenerating into the single point $1 + 0i$ and the others appearing in complex conjugate pairs (if they are indeed complex); notably for $s = 4, 6$ there is an additional, purely real branch.

3 Conclusion and future work

We have demonstrated how to transpose the analysis of each of the two groups into the language of the other, spelling out in detail how these seemingly different approaches map onto each other. However, in both [5, 7], the spectral results are only a part of the story. In [5] these are used together with the block locally Toeplitz (BLT) theory to establish interesting asymptotic results (which have been independently observed also in [22, Chapter 7]). In [7], the focus is on using the obtained eigeninformation (or an estimate thereof) for estimating the GMRES behavior convergence of the preconditioned systems, using the potential theory and the Schwarz-Christoffel (SC) mapping in particular. We believe that these directions could

and should be explored in conjunction as well – the BLT theory can, in principle, provide estimates of the (fully complex) spectrum of the spatial operator, which for is the crucial piece needed for further generalization of the SC mapping approach for GMRES estimation for systems stemming from IRK applied to PDEs with more involved spatial operators. This is a work in progress and which we find quite enticing.

Acknowledgements The author would like to acknowledge the support of the PRIMUS grant `PRIMUS/25/SCI/022` and would like to thank Ivo Dravins for several inspiring discussions.

References

[1] M. M. Rana, V. E. Howle, K. Long, A. Meek, and W. Milestone, SIAM Journal on Scientific Computing **43**(5), S475–S495 (2021).
[2] P. E. Farrell, R. C. Kirby, and J. Marchena-Menéndez, ACM Transactions on Mathematical Software **47**(4) (2021).
[3] B. S. Southworth, O. Krzysik, W. Pazner, and H. De Sterck, SIAM Journal on Scientific Computing **44**(1), 416–443 (2022).
[4] B. S. Southworth, O. Krzysik, and W. Pazner, SIAM Journal on Scientific Computing **44**(2), 636–663 (2022).
[5] I. Dravins, S. Serra-Capizzano, and M. Neytcheva, SIAM Journal on Matrix Analysis and Applications **45**(2), 1007–1034 (2024).
[6] S. Leveque, L. Bergamaschi, A. Martínez, and J. W. Pearson, SIAM Journal on Matrix Analysis and Applications **45**(4), 1902–1928 (2024).
[7] M. J. Gander and M. Outrata, SIAM Journal on Scientific Computing, in press **46**(3), A2047–A2072 (2024).
[8] G. Wanner, S. P. Nørsett, and E. Hairer, Solving Ordinary Differential Equations I : Non-Stiff Problems (Springer Berlin, Heidelberg, 1987).
[9] G. Wanner and E. Hairer, Solving Ordinary Differential Equations II : Stiff and Differential-Algebraic Problems (Springer Berlin, Heidelberg, 1996).
[10] J. C. Butcher, BIT Numerical Mathematics **16**(3), 237–240 (1976).
[11] M. R. Clines, V. E. Howle, and K. R. Long, Efficient order-optimal preconditioners for implicit Runge-Kutta and Runge-Kutta-Nyström methods applicable to a large class of parabolic and hyperbolic PDEs, arXiv: https://arxiv.org/abs/2206.08991, 2022.
[12] M. Neytcheva and O. Axelsson, Numerical Solution Methods for Implicit Runge-Kutta Methods of Arbitrarily High Order, in: Proceedings of the Conference Algoritmy 2020, edited by P. Frolkovič, K. Mikula, and D. Ševčovič (VydavateĺStvo SPEKTRUM, 2020).
[13] O. Axelsson, I. Dravins, and M. Neytcheva, Numerical Linear Algebra with Applications **31**(1), e2532 (2024).
[14] P. Munch, I. Dravins, M. Kronbichler, and M. Neytcheva, SIAM Journal on Scientific Computing **46**(2), S71–S96 (2024).
[15] W. Pazner and P. O. Persson, Journal of Computational Physics **335**, 700–717 (2017).
[16] J. Liesen and Z. Strakoš, Krylov Subspace Methods: Principles and Analysis (Oxford University Press, Oxford, 2013).
[17] G. A. Staff, K. A. Mardal, and T. K. Nilssen, Modeling, Identification and Control **27**(2), 109–123 (2006).
[18] A. Greenbaum, V. Pták, and Z. Strakoš, SIAM Journal on Matrix Analysis and Applications **17**(3), 465–469 (1996).
[19] M. J. Gander and M. Outrata, Linear Algebra and its Applications (2023).
[20] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide (SIAM, Philadelphia, 2000).
[21] A. Rani, P. Ghysels, V. Howle, K. Long, and M. Outrata, Efficient solution of fully implicit Runge-Kutta methods for linear wave equations, 2025, in the 2^{nd} round of revision.
[22] M. Outrata, Schwarz methods, Schur complements, preconditioning and numerical linear algebra, PhD thesis, University of Geneva, Rue de Conseil-General 2–4, Geneva 4, Switzerland, 2022.