Spectral analysis of implicit 2 stage block Runge-Kutta preconditioners

Martin J. Gander*, Michal Outrata†

Abstract

We analyze the recently introduced family of preconditioners in [15] for the stage equations of implicit Runge-Kutta methods for two stage methods. We give explicit formulas for the eigenvalues and eigenvectors of the preconditioned systems for a general method and use these to give explicit convergence estimates of preconditioned GMRES for some common choices of the implicit Runge-Kutta methods. This analysis also allows us to qualitatively predict and explain the main observed features of the GMRES convergence behavior, not only bound it. We illustrate our analysis with numerical experiments. We also consider the direction of *numerical optimization for improving the preconditioners performance*, as suggested in [15]. We consider two different ways – both distinct to the one introduced in [15] – and numerically optimize these, using the explicit bounds obtained beforehand.

Keywords: implicit Runge-Kutta methods, two stages, preconditioners, GMRES, bounds
Classification: 65F08, 65F10

1 Introduction

Runge-Kutta methods are a well-established family of one-step solvers for systems of ordinary differential equations (ODEs; see [22, 21] for an overview and further references). For implicit methods (IRK), their efficiency relies on a solver for the so-called *stage equations* – in general a system of ns non-linear equations, where n is the number of scalar ODEs in the system and s is the number of stages of the Runge-Kutta method. An important application arises from the space discretization of time-dependent partial differential equations (PDEs), resulting in a system of ODEs with *very* large n. If the spatial operator is *linear*, then the stage equations also become a system of linear algebraic equations, which are often solved by an iterative solver, e.g., a Krylov method. In [15], the authors introduced a family of preconditioners for GMRES for the stage equations, numerically showing that these preconditioners give an *outstanding* performance, especially under refinement of the spatial mesh, i.e., as n grows. Recently, there has been also other contributions in the direction of preconditioning the *fully implicit* Runge-Kutta stage equations for PDEs, see [18, 17] and [4], but also [12] and [2], introducing new ideas and testing these numerically on a variety of test problems.

*Section de Mathématiques, Université de Genève; this work was partially supported by the SNF grant number 178752.

†Section de Mathématiques, Université de Genève; The author acknowledges support from the Federal Commission for Scholarships for Foreign Students for the Swiss Government Excellence Scholarship (ESKAS No. 2019.0384) for the academic year(s) 2019-22.

We focus on the setting considered by Rana et al. in [15] and analyze the convergence of the preconditioned GMRES method, giving a theoretical background for the performance observed for two stage methods. The general s-stage case will be treated in a follow-up manuscript. In Section 2 we summarize some preliminary knowledge and introduce the problem and the preconditioners followed by some general results for Kronecker-like matrices in Section 3. We then give the analysis of each of the block type of the preconditioners in Section 4.1 – 4.3 and also comment on the possibility of numerical optimization to improve the solution process in Section 4.4.

2 Model problem and preliminaries

As our model problem we consider the heat equation on the unit square and a time interval $(0, T_{\text{end}})$, i.e.,

$$\begin{aligned} \frac{\partial}{\partial t}u = \Delta u + f \quad &\text{in } \Omega \times (0, T_{\text{end}}), \\ u = g \quad \text{on } \partial\Omega \times (0, T_{\text{end}}) \qquad &\text{and} \qquad u = u_0 \quad \text{in } \Omega \times \{0\}, \end{aligned} \tag{1}$$

where Δ is the Laplace operator, f, g, u_0 are given functions and Ω is the unit square $\Omega := (0,1) \times (0,1)$. We discretize in space using finite difference scheme on an equidistant grid with $N+1$ rows and columns and with the mesh size $h = 1/N$ as in Figure 1. The values at the interior grid points become unknown functions of time, which are governed by the system of ODEs,

$$\frac{\partial}{\partial t}u_i(t) = \frac{u_{i-N}(t) + u_{i-1}(t) - 4u_i(t) + u_{i+1}(t) + u_{i+N}(t)}{h^2} + b_i^{(ST)}(t), \tag{2}$$

for $i = N+1, \ldots, N(N-1)-1$, where $b_i^{(\text{ST})}(t)$ collects the known values from the source terms, given by g and f, at the given point. Combining the unknowns in each grid column into one vector denoted by $\mathbf{u}_k(t)$, i.e.,

$$\mathbf{u}_k(t) := \begin{bmatrix} u_{Nk+2} & u_{Nk+3} & \cdots & u_{N(k+1)-1} \end{bmatrix}^T(t), \quad \mathbf{u}(t) := \begin{bmatrix} \mathbf{u}_1(t) & \cdots & \mathbf{u}_{N-1}(t) \end{bmatrix}^T,$$

and also analogously for $\mathbf{b}_k(t)$ and $\mathbf{b}(t)$, we rewrite (2) as

$$\frac{\partial}{\partial t}\mathbf{u}(t) = \frac{1}{h^2}L\mathbf{u}(t) + \mathbf{b}^{(\text{ST})}(t), \tag{3}$$

with

$$L = \begin{bmatrix} T & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & T \end{bmatrix}, \quad T = \begin{bmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \tag{4}$$

where L is of dimension n^2 with $n := (N-1)$ and the blocks T, I are of dimension n. We discretize $[0, T_{\text{end}}]$ with $M_{T_{\text{end}}} + 1$ equidistant time points with time step $\tau = T_{\text{end}}/M_{T_{\text{end}}}$, i.e.,

$$\{0 = t_0 < t_1 \cdots < t_{M_{T_{\text{end}}}-1} < t_{M_{T_{\text{end}}}} = T_{\text{end}}\}, \quad \tau = \frac{T_{\text{end}}}{M_{T_{\text{end}}}} \quad \text{and} \quad t_m = \tau \cdot m,\ m = 0, \ldots, M_{T_{\text{end}}}.$$

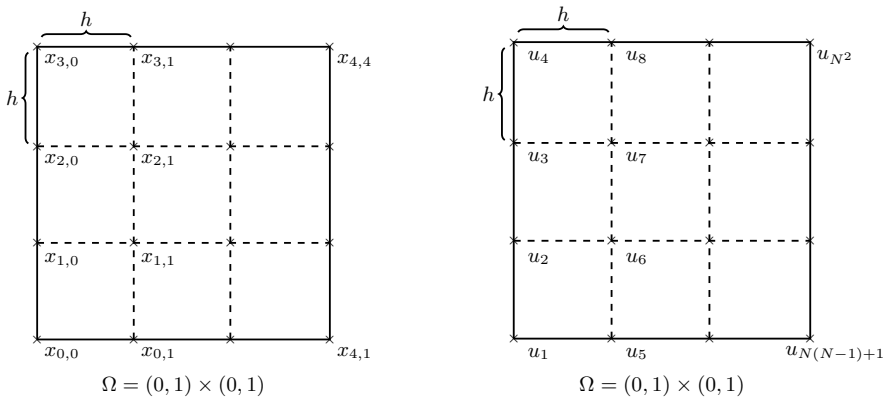

Figure 1: Left: grid points for $N+1=4$; right: lexicographical ordering of the unknowns for $N+1=4$.

Having a *Butcher tableau*

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b} \end{array} := \begin{array}{c|ccc} c_1 & a_{1,1} & \dots & a_{1,s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s,1} & \dots & a_{s,s} \\ \hline & b_1 & \dots & b_s \end{array}, \tag{5}$$

the corresponding IRK method applied to (3) at the m-th time step gives the approximation $\mathbf{u}^m \approx \mathbf{u}(t_m)$ as

$$\mathbf{u}^m = \mathbf{u}^{m-1} + \tau \sum_{i=1}^{s} b_i \mathbf{k}_i^m, \tag{6}$$

where the vectors $\mathbf{k}_1^m, \dots, \mathbf{k}_s^m \in \mathbb{R}^n$ are the solutions of the linear system

$$\underbrace{\left(\begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} - \frac{\tau}{h^2} \begin{bmatrix} a_{1,1}L & \dots & a_{1,s}L \\ \vdots & \ddots & \vdots \\ a_{s,1}L & \dots & a_{s,s}L \end{bmatrix} \right)}_{\equiv I_s \otimes I_n - \frac{\tau}{h^2}(A \otimes L) =: M} \mathbf{k}^m = \begin{bmatrix} \frac{1}{h^2} L\mathbf{u}^{m-1} + \mathbf{b}^{(\mathrm{ST})}(t_{m-1} + c_1\tau) \\ \vdots \\ \frac{1}{h^2} L\mathbf{u}^{m-1} + \mathbf{b}^{(\mathrm{ST})}(t_{m-1} + c_s\tau) \end{bmatrix}, \tag{7}$$

with

$$\mathbf{k}^m := \begin{bmatrix} \mathbf{k}_1^m & \cdots & \mathbf{k}_s^m \end{bmatrix}^T \in \mathbb{R}^{ns}.$$

The symbol $\otimes$ stands for the Kronecker product (see [20] and references therein) and we would like to note here that (7) can be reformulated into a *matrix equation*, which is in general better suited for using a Krylov solver (see [14]). Here we focus on the analysis of the results in [15] and thus we do not address this any further but a study of the preconditioners from [15] in the matrix equations setting seems worthwhile. Having $p \leq 2s$ as the order of convergence of the IRK method we assume that it is balanced with the spatial discretization error, i.e., that $h^2 = C_{\mathrm{e}}\tau^p$ for some $C_{\mathrm{e}} > 0$.

The problem (7) with the sparse system matrix M can be very large for h (and τ) small, suggesting an iterative solver such as GMRES, BiCG or GCR should be used which in turn usually requires

a preconditioner to become truly efficient. In [15], the authors introduce the block preconditioners

$$\begin{aligned} P^{\mathrm{d}} &= I_s \otimes I_n - \frac{\tau}{h^2}\mathrm{diag}(A) \otimes L, \\ P^{\mathrm{u}} = I_s \otimes I_n - \frac{\tau}{h^2} D_A U_A \otimes L \quad &\text{and} \quad P^{\mathrm{l}} = I_s \otimes I_n - \frac{\tau}{h^2} L_A D_A \otimes L, \end{aligned} \tag{8}$$

where L_A, D_A, U_A are the LDU factors of the Butcher tableau matrix A. In addition, the authors also consider the block triangular preconditioners

$$P^{\mathrm{GSL}} = I_s \otimes I_n - \frac{\tau}{h^2} A_L \otimes L \quad \text{and} \quad P^{\mathrm{GSU}} = I_s \otimes I_n - \frac{\tau}{h^2} A_U \otimes L, \tag{9}$$

where *GSL/GSU* stands for *Gauss-Seidel lower/upper*, and $A_{L,U}$ is the lower/upper triangular part of A, i.e.,

$$(A_L)_{ij} = \begin{cases} a_{ij} & \text{if } i \geq j \\ 0 & \text{otherwise} \end{cases}, \quad (A_U)_{ij} = \begin{cases} a_{ij} & \text{if } i \leq j \\ 0 & \text{otherwise} \end{cases}.$$

Notice that if $a_{ii} > 0$ for all $i = 1, \ldots, s$, then the preconditioners are invertible as L is symmetric negative-definite. More general conditions for non-singularity of the preconditioners can be also derived analogously to [18, Lemma 1].

Some of these – P^{d} and P^{GSL} – were considered already in [19], and the authors in [15] observed numerically that the newly proposed preconditioners $P^{\mathrm{u,l}}$ outperform the previously proposed $P^{\mathrm{d}}, P^{\mathrm{GSL}}$ as well as P^{GSU} but without a clear understanding why that is the case and, more generally, without any insight into the actual GMRES convergence behavior of these preconditioners. We aim to remedy this below.

Using GMRES for a linear system $C\mathbf{x} = \mathbf{f}$ with C being diagonalizable, i.e., $C = S\Lambda S^{-1}$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, a standard convergence bound for the residuals $\mathbf{r}_\ell$ reads

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S) \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq \ell}} \max_{1\leq i\leq d} |\varphi(\lambda_i)|, \tag{10}$$

where $\kappa(S)$ is the 2-norm condition number of the matrix S, see, e.g., [10, Section 5.7.2]. We would like to highlight some aspects of this bound that is often used to study GMRES convergence behavior.

Remark 1. *As indicated above, the spectral information of the system matrix in GMRES (in our case of the preconditioned system) does not generally govern the convergence (see [6], [7] and [1] and also [10, Chapter 2 and 5.7] and the references therein). If the system matrix is normal, i.e., it is diagonalizable with S unitary, then the spectral information is enough to use the ideal GMRES bound. However, if C is non-normal, then a convincing argument needs to be put forward to validate linking spectral information with the convergence behavior of GMRES as the authors in [10, p. 303, Remark 1] point out.*

Moreover, particular knowledge of the interaction of S and the initial residual $\mathbf{r}_0$ can lead to a qualitative *and* quantitative *improvement on (10), see, e.g., [9]. However, studying GMRES bahavior with the bound (10), this interaction is completely lost.*

We use the bound (10) in the above sense and do not address the aspect of the interaction of the initial residual with the eigenbasis. It turns out that the bound (10) is still a fine enough tool to help us understand the preconditioned GMRES behavior in our case.

3 Analysis of the block preconditioners

We start by transforming the calculations into the eigenbasis of the spatial operator. Denoting the eigenpairs of L by $(\lambda_k, \mathbf{v}_k)$, we organize the eigenvectors into an n-by-n matrix V and define the block transformation matrix Q,

$$V := [\mathbf{v}_1, \ldots, \mathbf{v}_n], \quad \text{and} \quad Q := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix} \in \mathbb{R}^{sn \times sn}. \tag{11}$$

Transforming M blockwise into the V basis gives $\tilde{M} := QMQ^T$,

$$\tilde{M} = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} - \frac{\tau}{h^2} \begin{bmatrix} a_{1,1}\Lambda & \ldots & a_{1,s}\Lambda \\ \vdots & \ddots & \vdots \\ a_{s,1}\Lambda & \ldots & a_{s,s}\Lambda \end{bmatrix}, \tag{12}$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. With the preconditioners proposed in (8-9) we write the spectrum of the preconditioned system as

$$\mathrm{sp}(MP^{-1}) = \mathrm{sp}(Q^T MP^{-1}Q) = \mathrm{sp}(Q^T MQQ^T P^{-1}Q) = \mathrm{sp}\left(\tilde{M}\tilde{P}^{-1}\right),$$

where $\tilde{P} := Q^T PQ$ stands for one of the right-preconditioners $P^{\mathrm{d,GSU,u}}$ and an analogous formulation follows also for the left-preconditioners $P^{\mathrm{l,GSL}}$. As the preconditioners are defined blockwise as scalar multiplications of L and I, their blockwise transformation into the eigenbasis of L is a straight-forward calculation - replacing L with Λ (and keeping I)[1]. Next, such matrices – block matrices with each block being a square, diagonal matrix – can be permuted into classical block-diagonal matrices as the following lemma shows.

Lemma 1. *Let $C \in \mathbb{R}^{ns \times ns}$ be a real matrix with block structure such that every block is a square diagonal matrix, i.e.,*

$$C = \begin{bmatrix} \Lambda_{11} & \ldots & \Lambda_{1s} \\ \vdots & \ddots & \vdots \\ \Lambda_{s1} & \ldots & \Lambda_{ss} \end{bmatrix}, \quad \text{with } \Lambda_{ij} = \mathrm{diag}\left(\lambda_1^{(ij)}, \ldots, \lambda_n^{(ij)}\right) \quad \forall ij. \tag{13}$$

Then there exists a permutation matrix $\Pi \in \mathbb{R}^{ns \times ns}$ such that

$$\Pi^T C \Pi = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_n \end{bmatrix} \quad \text{with } C_\ell = \begin{bmatrix} \lambda_\ell^{(11)} & \ldots & \lambda_\ell^{(1s)} \\ \vdots & \ddots & \vdots \\ \lambda_\ell^{(s1)} & \ldots & \lambda_\ell^{(ss)} \end{bmatrix} \in \mathbb{R}^{s \times s}, \tag{14}$$

for any $\ell = 1, \ldots, n$.

[1] This has also been observed in [18, Lemnma 1 and below] and can be formally stated as M being *a matrix over the commutative ring of linear combinations of L and I.*

Hence, C is diagonalizable if and only if C_ℓ are diagonalizable for all $\ell = 1, \dots, n$ and if $C_\ell = V_\ell^{-1} D_\ell V_\ell$ is the eigendecomposition of C_ℓ with $D_\ell = \operatorname{diag}(\mu_\ell^{(1)}, \dots, \mu_\ell^{(s)})$, then

$$\operatorname{sp}(C) = \bigcup_{\ell=1}^{n} \bigcup_{i=1}^{s} \mu_\ell^{(i)}.$$

If $(\mu, \mathbf{v})$ is an eigenpair of some C_ℓ, then $\left(\mu, \Pi^T (\mathbf{v} \otimes \mathbf{e}_\ell)\right)$ is an eigenpair of C. As a result, if C is diagonalizable with $C = V^{-1} D V$, then

$$\kappa(V) = \frac{\max\limits_{\ell=1,\dots,n} \sigma_1^{(\ell)}}{\max\limits_{\ell=1,\dots,n} \sigma_s^{(\ell)}},$$

where $\kappa(\cdot)$ is the 2-norm condition number and the matrices V_ℓ have the singular values $\sigma_1^{(\ell)} \geq \dots \geq \sigma_s^{(\ell)} \geq 0$.

Proof. Setting $E_\ell = \operatorname{diag}(0, \dots 0, 1, 0, \dots, 0) \in \mathbb{R}^{n \times n}$ as the matrix with the only non-zero entry being at the position (ℓ, ℓ) with value one, we observe that

$$C = \sum_{\ell=1}^{n} C_\ell \otimes E_\ell.$$

Using the Kronecker product permutation property from [20, Eqn. (1) and below], we take Π such that

$$\Pi^T C \Pi = \sum_{\ell=1}^{n} E_\ell \otimes C_\ell = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_n \end{bmatrix},$$

proving the first part of the statement. The rest follows by a direct calculation and the properties of block-diagonal matrices. □

Using Lemma 1 we can analyse the eigenproperties of $\tilde{M}\tilde{P}^{-1}$ directly and thereby evaluate the GMRES bound.

Remark 2. *We note that an analogous lemma to Lemma 1 can also be formulated for non-normal matrices (replacing Q^T by Q^{-1}). Considering the Jordan canonical (or the Schur decomposition form) of C_ℓ, Lemma 1 can be reformulated to obtain a block upper bidiagonal (or block upper-triangular) matrix.*

To shorten the notation we set

$$\theta_k := \frac{\tau}{h^2} \lambda_k \quad \text{and} \quad \Theta := \frac{\tau}{h^2} \Lambda, \tag{15}$$

as these quantities appear always together in the computations. By a direct calculation (see [13, Appendix B.8]) we get the limit behavior of θ_k as $\tau, h \to 0$,

$$\underbrace{\begin{array}{l} (\theta_n, \theta_1) \to (-\frac{8}{C_\mathrm{e}}, 0), \\ (\theta_1^{-1}, \theta_n^{-1}) \to \left(-\infty, -\frac{C_\mathrm{e}}{8}\right), \end{array}}_{(\mathrm{LIM})_{p=1}} \qquad \underbrace{\begin{array}{l} (\theta_n, \theta_1) \to (-\infty, 0), \\ (\theta_1^{-1}, \theta_n^{-1}) \to (-\infty, 0), \end{array}}_{(\mathrm{LIM})_{p>1}} \tag{16}$$

and continue by explicit calculations for the two stage methods.

4 Two stage methods

For the case $s = 2$ we have

$$\tilde{P}^{\mathrm{d}} = \begin{bmatrix} I - a_{11}\Theta & 0 \\ 0 & I - a_{22}\Theta \end{bmatrix}, \quad \tilde{P}^{\mathrm{GSL}} = \begin{bmatrix} I - a_{11}\Theta & 0 \\ -a_{21}\Theta & I - a_{22}\Theta \end{bmatrix}, \quad \tilde{P}^{\mathrm{GSU}} = \begin{bmatrix} I - a_{11}\Theta & -a_{12}\Theta \\ 0 & I - a_{22}\Theta \end{bmatrix},$$
$$\tilde{P}^{\mathrm{u}} = \begin{bmatrix} I - a_{11}\Theta & -a_{12}\Theta \\ 0 & I - \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right)\Theta \end{bmatrix}, \quad \tilde{P}^{\mathrm{l}} = \begin{bmatrix} I - a_{11}\Theta & 0 \\ -a_{21}\Theta & I - \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right)\Theta \end{bmatrix}, \tag{17}$$

and we start with a useful lemma summarizing some direct calculations.

Lemma 2 ([13, Appendix B.8]). *Let $C \in \mathbb{R}^{2\times 2}$.*

(i) The eigenvalues $\mu_{1,2}$ are given by $\mu_{1,2} = \frac{c_{11}+c_{22}}{2} \pm \frac{1}{2}\sqrt{D}$ with $D = (c_{11} - c_{22})^2 + 4c_{12}c_{21}$. In particular, if $D \neq 0$, then C is diagonalizable.

(ii-a) If $c_{12} \neq 0$ and $D \neq 0$, then C is diagonalizable and the eigenvectors $\mathbf{v}_{1,2}$ are given by $\mathbf{v}_{1,2} = \frac{1}{\|\tilde{\mathbf{v}}_{1,2}\|}\tilde{\mathbf{v}}_{1,2}$ with $\tilde{\mathbf{v}}_{1,2} = \begin{bmatrix} 1 \\ \alpha_{1,2} \end{bmatrix}$ and $\alpha_i = \begin{cases} 0 & \text{if} \quad c_{11} = \lambda_i, \\ -\frac{c_{11}-c_{22}\mp\sqrt{D}}{2c_{12}} & \text{if} \quad c_{11} \neq \lambda_{1,2}. \end{cases}$

(ii-b) If $c_{21} \neq 0$ and $D \neq 0$, then C is diagonalizable and the eigenvectors $\mathbf{v}_{1,2}$ are given by $\mathbf{v}_{1,2} = \frac{1}{\|\tilde{\mathbf{v}}_{1,2}\|}\tilde{\mathbf{v}}_{1,2}$ with $\tilde{\mathbf{v}}_{1,2} = \begin{bmatrix} \alpha_{1,2} \\ 1 \end{bmatrix}$ and $\alpha_i = \begin{cases} 0 & \text{if} \quad c_{22} = \lambda_i, \\ -\frac{-c_{11}+c_{22}\mp\sqrt{D}}{2c_{21}} & \text{if} \quad c_{22} \neq \lambda_{1,2}. \end{cases}$

(iii) If $c_{12}, c_{21} \neq 0$ and $D \neq 0$, then the singular values of the matrix of eigenvectors V of C are given by $\sigma_{1,2} = \sqrt{\|\tilde{v}_1\|\|\tilde{v}_2\| \pm \sqrt{(1 + \overline{\alpha_1}\alpha_2)\,(1 + \alpha_1\overline{\alpha_2})}}$, where α_i is given as in $(ii - a)$.

(iv-a) If $c_{12} = 0, c_{21} \neq 0$ and $c_{11} \neq c_{22}$, then C is diagonalizable with real eigenvalues c_{11}, c_{22} and eigenvectors $\mathbf{v}_1, \mathbf{e}_2$ and the formula from (iii) simplifies to $\sigma_{1,2} = \sqrt{\|\tilde{\mathbf{v}}_1\| \pm 1}$, where $\mathbf{v}_2$ is the vector with no zero entries from (ii-b).

(iv-b) If $c_{21} = 0, c_{12} \neq 0$ and $c_{11} \neq c_{22}$, then C is diagonalizable with real eigenvalues c_{11}, c_{22} and eigenvectors $\mathbf{e}_1, \mathbf{v}_2$ and the formula from (iii) simplifies to $\sigma_{1,2} = \sqrt{\|\tilde{\mathbf{v}}_2\| \pm 1}$, where $\mathbf{v}_1$ is the vector with no zero entries from (ii-a) and α_1 is its first component.

We analyze first the block diagonal preconditioners, and then continue with the block triangular ones. The calculations below give insight into the results presented in [15], e.g., give explicit formulas for the results in Figure 4.1 and 4.3, Table 4.3 and Table 5.1, 5.2. and 5.3 from [15] for $s = 2$.

4.1 Block diagonal preconditioner

A direct calculation gives

$$\tilde{M}(\tilde{P}^{\mathrm{diag}})^{-1} = \begin{bmatrix} I & -a_{12}\Theta\,(I - a_{22}\Theta)^{-1} \\ -a_{21}\Theta\,(I - a_{11}\Theta)^{-1} & I \end{bmatrix}, \tag{18}$$

and using Lemma 1 the eigen-information of the preconditioned system can be obtained from the 2-by-2 matrices

$$X_k^{\mathrm{d}} := \begin{bmatrix} 1 & -\frac{a_{12}\theta_k}{1-a_{22}\theta_k} \\ -\frac{a_{21}\theta_k}{1-a_{11}\theta_k} & 1 \end{bmatrix}.$$

We immediately notice that X_k^{d} is diagonalizable if and only if[2]

$$a_{12} = 0 \iff a_{21} = 0. \tag{19}$$

Assuming $a_{12}, a_{21} \neq 0$, we calculate the characteristic polynomial $p_{X_k^d}(\lambda)$ of X_k^{d},

$$p_{X_k^{\mathrm{d}}}(\lambda) = \lambda^2 - 2\lambda + 1 - \frac{a_{12}a_{21}\theta_k^2}{(1-a_{11}\theta_k)(1-a_{22}\theta_k)},$$

and therefore the eigenvalues $\xi_{1,2}^{(k)}$ of X_k^{d} are given by

$$\xi_{1,2}^{(k)} = 1 \pm \sqrt{D_k} \quad \text{with} \quad D_k = \frac{a_{12}a_{21}}{(|\theta_k^{-1}| + a_{11})(|\theta_k^{-1}| + a_{22})}. \tag{20}$$

We write $\xi_{1,2}^{(k)}$ as functions of $|\theta_k|^{-1}$,

$$\xi_1^{(k)} = 1 \pm \sqrt{\phi(|\theta_k|^{-1})} \quad \text{with} \quad \phi(\alpha) = \frac{a_{12}a_{21}}{(\alpha + a_{11})(\alpha + a_{22})},$$

and $\alpha \in (|\theta_1|^{-1}, |\theta_n|^{-1})$ – an interval converging towards the limit interval in (16). If $a_{11}, a_{22} \geq 0$ (e.g., Gauss, Radau or Lobatto methods), then $\xi_{1,2}^{(k)}$ lie on a line segment in $\mathbb{C}$ going through the point 1, which is either a part of the real axis (if $\operatorname{sign}(a_{12}a_{21}) \geq 0$) or on the line $1 + \beta i$, $\beta \in \mathbb{R}$ (otherwise). We have

$$|\xi_{1,2}^{(k)} - 1| = \left|\sqrt{\phi(|\theta_k^{-1}|)}\right| = \sqrt{\left|\frac{a_{12}a_{12}}{(|\theta_k^{-1}| + a_{11})(|\theta_k^{-1}| + a_{22})}\right|},$$

and hence[3] the maximum of $|\xi_{1,2}^{(k)} - 1|$ as a function of $|\theta_k^{-1}|$ is attained either at one of the endpoints of the interval in (16) or at an interior stationary point. Calculating the derivative, we get

$$(|\phi|)'(\alpha) = -\operatorname{sign}(\phi(\alpha))\, a_{12}a_{21} \frac{2\alpha + a_{11} + a_{22}}{(\alpha + a_{11})^2(\alpha + a_{22})^2},$$

and thus the only candidate for a stationary point is $-(a_{11} + a_{22})/2$ assuming it belongs to the domain of ϕ (as mentioned above, this is not the case for the commonly used Gauss, Radau or Lobatto methods). Assuming it does not, e.g., because $a_{11}, a_{22} > 0$, the maximum is attained at the left endpoint of the interval in (16), bounded from above by the value at $\alpha = 0$ which gives

$$|\xi_{1,2}^{(k)} - 1| \leq \sqrt{\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right|}. \tag{21}$$

[2] If $a_{12} = a_{21} = 0$, then A is in fact diagonal and hence $M = P^{\mathrm{d}}$, making this case uninteresting.

[3] We assumed at the beginning of the section that the inverse $(\tilde{P}^{\mathrm{d}})^{-1}$ exists and hence the denominator of ϕ is non-zero. Hence $|\phi(\alpha)|$ is a smooth function.

Thus making this quantity small will make the eigenvalues cluster tightly around 1. Notice that the bound above is suggesting to make the diagonal entries large compared to the off-diagonal ones, making the matrix diagonal in the limit (and hence making the preconditioner exact). Assuming $a_{ij} \neq 0$ for $i, j = 1, 2$ and $a_{11}, a_{22} \geq 0$ we use Lemma 2, and the condition number of the matrix of eigenvectors S_k^{d} of X_k^{d} is given by

$$\kappa\left(S^{\mathrm{d}}\right) = \frac{\max\limits_{k=1,\dots,n} \sigma_1^{(k)}}{\min\limits_{k=1,\dots,n} \sigma_2^{(k)}} = \sqrt{\frac{\max\limits_{k=1,\dots,n} 1 + \left|\frac{a_{21}(|\theta_k|^{-1}+a_{22})}{a_{12}(|\theta_k|^{-1}+a_{11})}\right| + \left|1 - \left|\frac{a_{21}(|\theta_k|^{-1}+a_{22})}{a_{12}(|\theta_k|^{-1}+a_{11})}\right|\right|}{\min\limits_{k=1,\dots,n} 1 + \left|\frac{a_{21}(|\theta_k|^{-1}+a_{22})}{a_{12}(|\theta_k|^{-1}+a_{11})}\right| - \left|1 - \left|\frac{a_{21}(|\theta_k|^{-1}+a_{22})}{a_{12}(|\theta_k|^{-1}+a_{11})}\right|\right|}}. \tag{22}$$

Recalling (16), we see that $\kappa\left(S^{\mathrm{d}}\right)$ has no singularities in it and the limits at 0 and $+\infty$ are also real and bounded. Therefore the conditioning of the eigenbasis of the preconditioned system will be uniformly bounded with respect to mesh refinement in space and time – just as the clustering diameter.

Remark 3. *Some authors call such preconditioners, i.e., preconditioners such that the eigenproperties of the preconditioned system can be bounded independently of h and τ,* order-optimal *and the general way to show order-optimality for these kind of preconditioners has been laid out in [11]. However, we want to emphasize that this independence does* not *mean, practically speaking, that the given preconditioner is in some sense optimal or even well-performing (and this is even more pronounced if the order-optimality is considered only with respect to the spectrum, omitting the conditioning of the eigenbasis of the preconditioned system) – but in particular settings with additional reasoning this might be a useful property. Perhaps a better-suited name would be "mesh-" or "discretization-independence".*

Given a Butcher tableau A, the bound (10) can be further approximated using the Joukowsky bound with the Chebyshev polynomials in the complex plane, see [16, Section 6.11 and Corollary 6.33] and also [10, Section 5.7.2] and references therein. A relatively direct and elementary calculation then allows us to evaluate the bound (10), and in Figure 2 we show the GMRES convergence behavior together with these bounds for different Butcher tableaus for P^{d} on the left (for validation of the above formulas and detailed calculations we refer the interested reader to [13, Section 7.3.1]). Notice that as h is fixed for all methods we obtain different τ for each of the IRK methods and therefore the scaling factor τ/h^2 becomes $h^{2(1/p-1)}$. Hence, the difference is not only in the choice of A but also in the interval spanned by θ_k and how close (or far) these are to the limit in (16).

We also want to address the notable staircase-like behavior. Since GMRES is invariant (in exact arithmetic) to an orthogonal transformation we can focus on GMRES applied to a problem with the block-diagonal matrix $X^{\mathrm{d}} := \mathrm{diag}(X_1^{\mathrm{d}}, X_2^{\mathrm{d}}, \dots, X_n^{\mathrm{d}})$. As noted in [3], the *optimal polynomial*[4] that realizes the min-max part of the bound

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S^{\mathrm{d}}) \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq\ell}} \max_{1\leq k\leq n} \|\varphi(X_k^{\mathrm{d}})\|,$$

equioscillates over the blocks in the sense of the above norm (see [3, Section 2.2, e.g., Table 2.1]) – but only for even degree polynomials, i.e., for $\ell = 2l$ for some $l \in \mathbb{N}$. Similarly, in [10, Section

[4] In [3], the authors call these the Chebyshev polynomials of the given matrix – of X^{d} in this case.

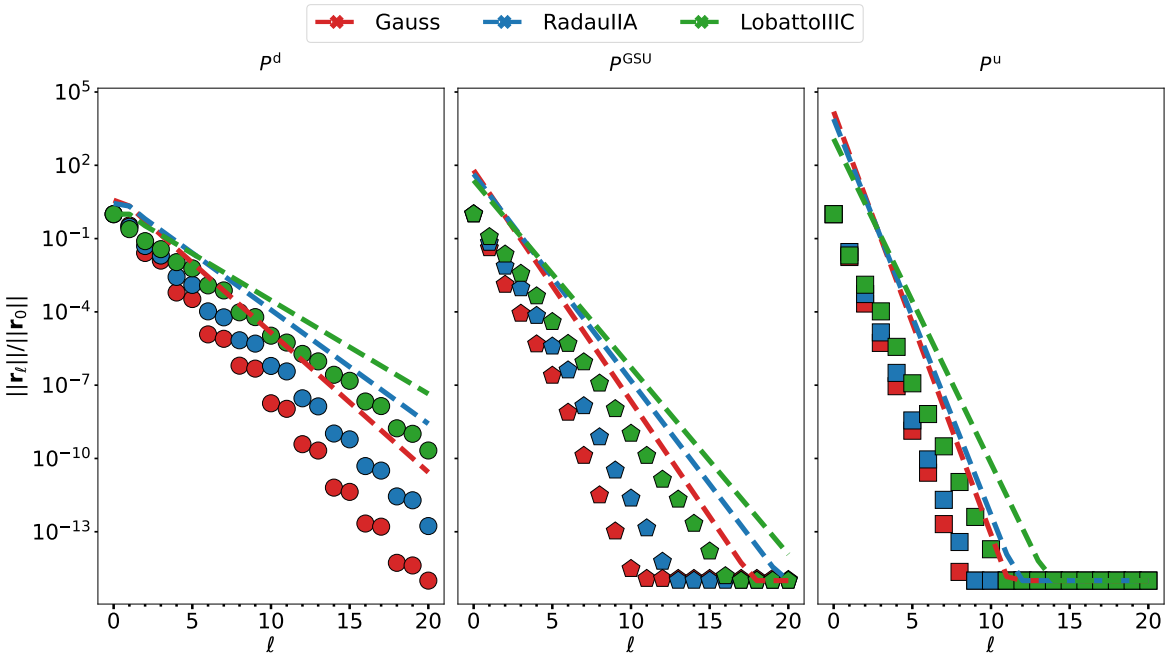

Figure 2: We show the GMRES performance (markers) together with the evaluated bound (10) (dashed lines) for different preconditioners and Butcher tableaus. We set $N = 100$, $\tau = h^{2/p}$ (p is the order of convergence of the IRK method; $p_{\mathrm{Gauss}} = 4, p_{\mathrm{RadauIIA}} = 3, p_{\mathrm{LobattoIIIC}} = 2$) and use a random right-hand side.

5.7.2, p.291] the GMRES convergence bound (10) is adapted to a symmetric indefinite problem with spectrum in two intervals $I^- \cup I^+ \equiv [-1, \nu] \cup [\nu, 1]$. Then (10) can be simplified to

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq 2\left(\frac{\kappa - 1}{\kappa + 1}\right)^{[\ell/2]}, \tag{23}$$

where $\kappa = 1/\nu$ is the condition number of the given system matrix and $[\ell/2]$ is the integer part of $\ell/2$. Unfortunately, adapting this bound to our setting becomes very quickly *very* complicated and goes beyond the scope of this text[5]. Nonetheless, both of these results suggest that GMRES can be more natural to analyze over "double-iterations" by merging two consecutive iterations into a single unit and analyzing the convergence behavior of these. This is supported by our results and illustrated in Figure 2, where these units become the mentioned stairs, e.g., iteration 14 and 15, then 16 and 17 etc. This does not, however, give any insight into what happens *within* these iteration units, i.e., to the speed-up observed at even iterations compared to the odd ones.

Recalling that the 2-by-2 blocks have eigenvalues $\xi_1^{(k)} \neq \xi_2^{(k)}$, we notice that for all of the considered Butcher tableau we have $\xi_1^{(k)} = \overline{\xi_2^{(k)}}$, i.e., the eigenvalues are in complex conjugate pairs on the line segment connecting $1 + i\sqrt{|(a_{12}a_{21})/(a_{11}a_{22})|}$ and $1 - i\sqrt{|(a_{12}a_{21})/(a_{11}a_{22})|}$. By

[5]The main difficulty, in our eyes, is in the treatment of the optimal polynomials *with real coefficients* on two separated line-segments in $\mathbb{C}$. For more details on the real case and the derivation of (23) we refer also to [5, Chapter 3].

definition, the real GMRES polynomial[6] aims to be small (in the maximum norm) over these points. However, polynomials with real coefficients of odd degree have at least one real root. Hence, progressing from an even iteration to an odd one, the additional degree of freedom of the GMRES polynomial is restricted so that the extra root is real. This condition is not restrictive if and only if the value of the GMRES polynomial is relatively large at some $\hat{\xi}$ *close* to the real line, which in our case is the same as being close to the point 1. Therefore, starting at a given even iteration, for the next GMRES iteration – odd one – the GMRES polynomial is made small at $\hat{\xi}$ (and $\overline{\hat{\xi}}$) by virtue of placing the extra root at 1. In contrast to that, at the following even iteration, the GMRES polynomial can be made small by two complex conjugate roots placed at $\hat{\xi}$ and $\overline{\hat{\xi}}$. Clearly the later is much more suitable than the former and hence we expect the even iterations to decrease the residual significantly more than the odd ones[7] – precisely as observed in Figure 2, explaining the staircase. As a result, for the preconditioner P^{d} with $\mathrm{sign}(a_{12}a_{21} = -1$ we should stop GMRES only after even iterations[8]. Also, as $h \to 0$ the scaled spectrum $\{\theta_k\}$ will represent the intervals in (16) more accurately and we expect that the convergence behavior will, under mesh refinement, tend to the bound. In the limit, the gap between the two complex conjugate branches of the spectrum will be closed, the real root will always be as useful as any other and we would expect essentially linear convergence.

Also, notice that this behavior is not quite as pronounced at the beginning. The above gives the following reason – at the first couple of iterations, the GMRES polynomial is largest at the most outlying (complex conjugate) parts of the spectrum. Hence, no matter the parity of the iteration, there is no effect of the value of the GMRES polynomial at $\hat{\xi}$ on the max norm of the GMRES polynomial and as a result, at the beginning of the GMRES convergence curve, there is no reason for a staircase-like convergence behavior.

4.2 Block upper-triangular preconditioner

We consider the preconditioners P^{GSU}, P^{u} and where necessary we join the quantities corresponding to the preconditioners by writing, e.g., $X_k^{\mathrm{GSU,u}}$ instead of X_k^{GSU} and X_k^{u}. By analogous calculations to Section 4.1, we obtain the formulas

$$X_k^{\mathrm{GSU}} = \begin{bmatrix} 1 & 0 \\ -\frac{a_{21}\theta_k}{1-a_{11}\theta_k} & \xi_k^{\mathrm{GSU}} \end{bmatrix} \quad \text{and} \quad X_k^{\mathrm{u}} = \begin{bmatrix} 1 & 0 \\ -\frac{a_{21}\theta_k}{1-a_{11}\theta_k} & \xi_k^{\mathrm{u}} \end{bmatrix},$$

[6] That is, the polynomial realizing the min-max part of the bound (10).

[7] Note that GMRES chooses globally optimal placement of the GMRES polynomial roots by possibly changing all roots placed at the previous iteration. That is, the "additional root" is not meant as an addition of an extra root to those chosen previously but rather an additional root to be placed anew together with the previous ones. The roots for two consecutive iterations will, surely, be *different* – the odd iteration has the restriction of having one real root and the other complex conjugate roots are chosen with that in mind due to the optimality condition of GMRES. For even iterations, all roots (can) come in complex conjugate pairs and in order to preserve the GMRES optimality condition, all of them will be different compared to the previous iteration.

[8] For different spectrum $\mathrm{sp}(L) = \{\lambda\}$ this can change but can be analyzed analogously to the above.

hence obtaining the spectrum of the preconditioned systems as the union of $\left\{1, \xi_k^{\mathrm{GSU,u}}\right\}_{k=1}^{n}$ with

$$\begin{aligned}\xi_k^{\mathrm{GSU}} &= \frac{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}) + 1}{|\theta|^2 a_{11}a_{22} + |\theta|(a_{11}+a_{22}) + 1},\\ \xi_k^{\mathrm{u}} &= \frac{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}) + 1}{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}-\frac{a_{21}a_{12}}{a_{11}}) + 1}.\end{aligned} \tag{24}$$

Moreover, Lemma 2 shows that for $X_k^{\mathrm{GSU},d} = (S_k^{\mathrm{GSU},d})^{-1}\mathrm{diag}(1, \xi_k^{\mathrm{GSU,u}})S_k^{\mathrm{GSU},d}$ we have

$$\kappa(S^{\mathrm{GSU,u}}) = \frac{\max\limits_{k=1,\dots,n}\sigma_1^{(k)}}{\min\limits_{k=1,\dots,n}\sigma_2^{(k)}} = \sqrt{\frac{\max\limits_{k=1,\dots,n}\sqrt{1+(\beta_{\mathrm{GSU,u}}(k))^2}+1}{\min\limits_{k=1,\dots,n}\sqrt{1+(\beta_{\mathrm{GSU,u}}(k))^2}-1}}, \tag{25}$$

where the scalars $\beta_{\mathrm{GSU,u}}(k)$ are given as functions of θ_k,

$$|\beta_{\mathrm{GSU}}| = \frac{|a_{12}|}{||\theta_k|^{-1}+a_{22}|}, \quad \text{and} \quad |\beta_{\mathrm{u}}| = \frac{|a_{12}/a_{11}|}{\left||\theta_k|\frac{\det(A)}{a_{11}}+a_{11}\right|}.$$

Further calculations give

$$\begin{aligned}\xi_k^{\mathrm{GSU}} - 1 &= \frac{|\theta_k|^2 a_{12}a_{21}}{|\theta_k|^2 a_{11}a_{22} + |\theta_k|(a_{11}+a_{22}) + 1},\\ \xi_k^{\mathrm{u}} - 1 &= \frac{|\theta_k| a_{12}a_{21}/a_{11}}{|\theta_k|^2 \det(A) + |\theta_k|(a_{11}+a_{22}-\frac{a_{21}a_{12}}{a_{11}}) + 1},\end{aligned}$$

and assuming $a_{11}, a_{22} \geq 0$, elementary calculus reveals that the cluster diameter for P^{GSU} is bounded from above by $|a_{12}a_{21}/(a_{11}a_{22})|$ – the limit as $|\theta_k| \to +\infty$. For P^{u} analogous calculations show that the cluster diameter is maximized at $|\theta_k| = \det(A)^{-1/2}$ (assuming $\det(A) \in (|\theta_n|, |\theta_1|)$) with the value

$$|\xi_k^{\mathrm{u}} - 1| \leq \left|\frac{a_{21}a_{12}}{a_{11}^2 + \det(A) + \frac{2a_{11}}{\sqrt{\det(A)}}}\right|.$$

The conditioning of the eigenbasis $S_k^{\mathrm{GSU,u}}$ of $X_k^{\mathrm{GSU,u}}$ can be treated similarly, first observing that $\kappa(S_k^{\mathrm{GSU,u}})$ is a decreasing function of $|\beta_{\mathrm{GSU,u}}| \in (0, +\infty)$ with a singularity at 0. Hence, as $h, \tau \to 0$, the matrix X_1^{GSU} (and X_n^{u}) becomes non-diagonalizable[9] as $|\theta_1| \to 0$ (and $|\theta_n| \to +\infty$). We show the preconditioned GMRES convergence behavior and the evaluated bounds in Figure 2; for detailed calculations and validation of the above formulas we refer the interested reader to [13, Section 7.3.2]. Also, notice that the stair-like behavior of P^{d} is *not* present as the entire spectrum is real and covers reasonably uniformly an interval on a real line and a single point 1. As the authors point out in [10, Section 5.6.2 and 5.7.2], as long as the condition number $\kappa(S^{\mathrm{GSU,u}})$ is not too large, the classical linear bound based on the condition number of the preconditioned system matrix (i.e., the condition number of $M\left(P^{\mathrm{GSU,u}}\right)^{-1}$) can be quite descriptive for the worst-case GMRES behavior.

[9] We can see this already in (24) since in the limits considered we get $\xi_k^{\mathrm{GSU,u}} = 1$ thus obtaining a 2-by-2 Jordan block.

If the preconditioned system is not far away from being normal, then this explains why we only see the linear convergence – the spectrum is populating the interval considered densely enough so that the superlinear convergence argument used in exact arithmetic is not applicable (i.e., we cannot single out any outliers if there are none, see [10, Section 5.6.4 (Figure 5.7 in particular) and also Section 5.6.4]). Hence, linear convergence is, in principle, to be expected. Also, this explains the observation that the number of GMRES iterations does not grow as $h \to 0$ as observed in [15] (an analogous argument applies also to the block-diagonal preconditioner). Notice that the accuracy of the bound (23) is usually supported by the same kind of arguments. By analogy, this also qualitatively explains the *linear* in the linear-over-two-iterations convergence behavior of the block-diagonal preconditioned system.

In light of the above, we can also explain the observed difference between the results of the preconditioners P^{GSU} and P^{u} for any particular choice of the method, i.e., for any given A, by simply evaluating the above formulas and comparing. In more general terms, we notice that the outliers of the spectrum $\{\xi_k^{\mathrm{GSU}}\}$ come from the outliers of the scaled spectrum $\{\theta_k\}$, in contrast to the outliers of the spectrum $\{\xi_k^{\mathrm{u}}\}$, which come from around the point $\theta_k \approx \det(A)^{-1/2}$. If the scaled spectrum $\{\theta_k\}$ is more sparse at the edges than in the interior of the interval $(|\theta_1|, |\theta_n|)$, as for our model problem, then we expect (for any not too large h) that a) the *slope* of the bound for P^{u} will be *larger* in absolute value than the one for P^{GSU}, i.e., predicting faster convergence, and b) the *slope* of the bound for P^{u} will be *more descriptive* than the one for P^{GSU} of the actual GMRES behavior – both of these qualities are clearly confirmed in Figure 2. However, the price paid is visible in the conditioning of the eigenbasis, as under the same conditions we observe that

$$\min \beta_{\mathrm{GSU}} \approx \frac{1}{\max |\theta_k|^{-1}} = \mathcal{O}(\tau) \quad \text{and} \quad \min \beta_{\mathrm{u}} \approx \frac{1}{\max |\theta_k|^{-1}} = \mathcal{O}(\tau^{p-1}),$$

i.e., the conditioning of the eigenbasis (and hence the initial offset of the corresponding bound) is asymptotically notably worse for P^{u} than for P^{GSU} based on (25) – also confirmed in Figure 2.

4.3 Block lower-triangular preconditioner

The results for $P^{\mathrm{GSL}}, P^{\mathrm{l}}$ are completely analogous to the ones from Section 4.2 and hence we just present these without much comment; more details can be found in [13, Section 7.3.3]. We have

$$X_k^{\mathrm{GSL}} = \begin{bmatrix} 1 & -\frac{a_{12}\theta_k}{1-a_{11}\theta_k} \\ 0 & \xi_k^{\mathrm{GSL}} \end{bmatrix} \quad \text{and} \quad X_k^{\mathrm{l}} = \begin{bmatrix} 1 & -\frac{a_{12}\theta_k}{1-a_{11}\theta_k} \\ 0 & \xi_k^{\mathrm{l}} \end{bmatrix},$$

hence obtaining the spectrum of the preconditioned systems as the union of $\left\{1, \xi_k^{\mathrm{GSL,l}}\right\}_{k=1}^{n}$ with

$$\begin{aligned} \xi_k^{\mathrm{GSL}} = \xi_k^{\mathrm{GSU}} &= \frac{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}) + 1}{|\theta|^2 a_{11}a_{22} + |\theta|(a_{11}+a_{22}) + 1}, \\ \xi_k^{\mathrm{l}} = \xi_k^{\mathrm{u}} &= \frac{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}) + 1}{|\theta|^2 \det(A) + |\theta|(a_{11}+a_{22}-\frac{a_{21}a_{12}}{a_{11}}) + 1}. \end{aligned} \tag{26}$$

Moreover, Lemma 2 shows that for $X_k^{\mathrm{GSL},l} = (S_k^{\mathrm{GSL},l})^{-1}\mathrm{diag}(1, \xi_k^{\mathrm{GSL,l}})S_k^{\mathrm{GSL},l}$ we have

$$\kappa(S^{\mathrm{GSU,u}}) = \frac{\max\limits_{k=1,\dots,n} \sigma_1^{(k)}}{\min\limits_{k=1,\dots,n} \sigma_2^{(k)}} = \sqrt{\frac{\max\limits_{k=1,\dots,n} \sqrt{1+(\beta_{\mathrm{GSL,l}}(k))^2}+1}{\min\limits_{k=1,\dots,n} \sqrt{1+(\beta_{\mathrm{GSL,l}}(k))^2}-1}},$$

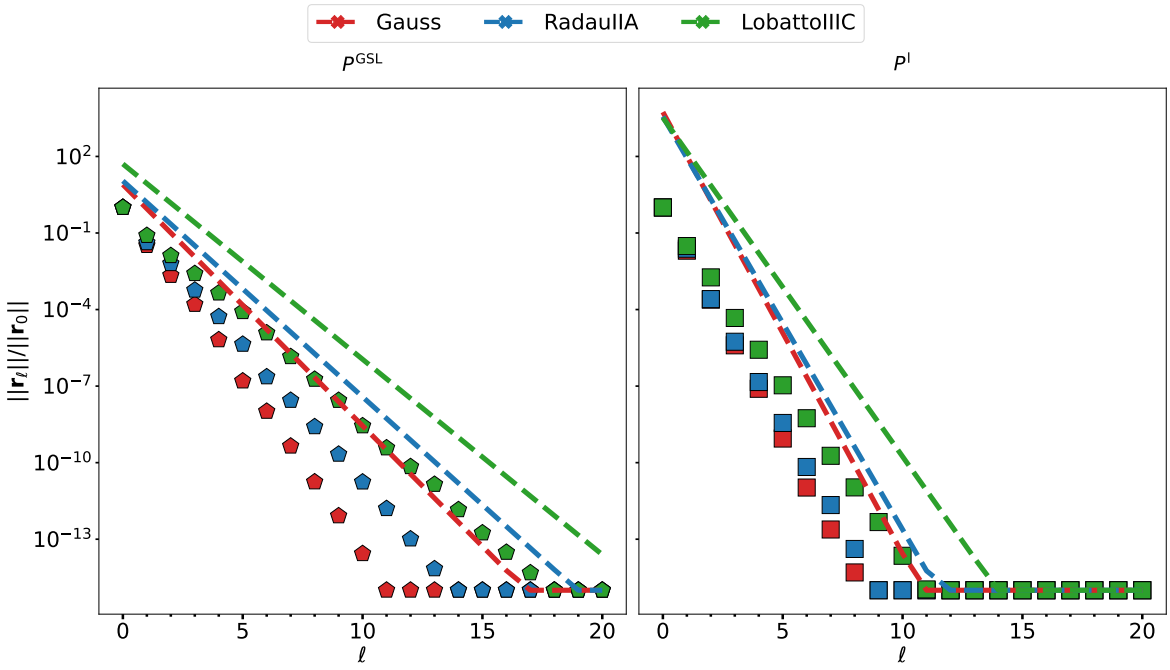

Figure 3: We show the GMRES performance (markers) together with the evaluated bound (10) (dashed lines) for different preconditioners and Butcher tableaus. We set $N = 100$, $\tau = h^{2/p}$ (p is the order of convergence of the IRK method; $p_{\mathrm{Gauss}} = 4, p_{\mathrm{RadauIIA}} = 3, p_{\mathrm{LobattoIIIC}} = 2$) and use a random right-hand side.

where the scalars $\beta_{\mathrm{GSL,l}}$ are given as functions of θ_k,

$$|\beta_{\mathrm{GSL}}| = \frac{|a_{21}|}{||\theta_k|^{-1} + a_{22}|}, \quad \text{and} \quad |\beta_{\mathrm{l}}| = \frac{|a_{21}/a_{11}|}{\left||\theta_k|\frac{\det(A)}{a_{11}} + a_{11}\right|}.$$

Notice that the results are either identical or very similar to the ones obtained with $P^{\mathrm{GSU}}, P^{\mathrm{u}}$ and a comparison of the convergence behavior as well as of the bounds in Figure 2 and Figure 3 reflects this.

4.4 Optimized Butcher tableaus

In [15], the authors consider taking a different A, let us denote it by $\tilde{A}$, for the construction of the preconditioner, motivated by [19]. $\tilde{A}$ is chosen as a result of an optimization routine seeking to minimize $\kappa(\tilde{A}^{-1}A)$ (or $\kappa(A\tilde{A}^{-1})$ for a right preconditioner) subject to a particular non-zero pattern and $\mathrm{diag}(\tilde{A}) = \mathrm{diag}(A)$. This optimization thus evaluates *only* quantities corresponding to the s-by-s Butcher tableaus, which should be negligible in cost compared to the solution process of the stage equations. Numerical results in [15, Table 4.3] show that the resulting preconditioner lowers the condition number of the preconditioned system. Having explicit formulas from Section 4.1–4.3 we can move from minimizing the quantity $\kappa(\tilde{A}^{-1}A)$ (or $\kappa(A\tilde{A}^{-1})$) to minimizing quantities in the bound (10) *so that a provable bound can be obtained* based on the result.

First, adapting the eigenpair formulas in Section 4.1 – 4.3 for the case $A \neq \tilde{A}$ we could optimize the eigenproperties *of the preconditioned system* and thus obtain a theoretical bound on the resulting preconditioner. Using Lemma 1, the matrices X_k are easily accessible with any A fixed.

Second, having $A = \tilde{A}$ as in Section 4.1 – 4.3, we can still consider the same optimization problem – but naturally with extra constraints so that the resulting IRK method has some desired order of convergence and stability properties.

We numerically test these approaches, defining the objective function to be minimized as a weighted sum of the cluster diameter[10] $|\xi_k - 1|$ and the conditioning of the eigenbasis of the preconditioned system matrix $\kappa(S)$,

$$f_{\text{obj}}(A, \tilde{A}) := \max_k |\xi_k - 1| + \omega\kappa(S),$$

with some positive small weight $\omega > 0$. Minimizing f_{obj} then aims to minimize the min-max part of (10) while keeping the eigenbasis conditioning under control. In case that the clustering is very tight we expect to get good GMRES convergence. Before showing the results we comment on the usefulness of the optimization approach for the case $A = \tilde{A}$.

Remark 4. *In order to optimize over $A = \tilde{A}$ we have to keep in mind that we are at the same time changing the IRK method we are using to solve the system of ODEs. This introduces some optimization constraints to ensure good order of convergence (equality constraints on $A, \mathbf{b}, \mathbf{c}$) and stability (equality and inequality constraints on $A, \mathbf{b}, \mathbf{c}$). In our experience, in order to obtain a notable improvement in minimization of f_{obj} over choosing $A = A_{\text{Gauss,RadauIIA,...}}$, we need to* relax *some of the qualities of the IRK methods, e.g., decrease the order of convergence or give up some level of stability.*

In our eyes, this makes the optimization for $A = \tilde{A}$ better suited for larger s (e.g., $s \geq 5$), where decreasing the order of convergence would usually not pose a significant drawback as the order of convergence of the standard IRK methods is often excessively high; see [13, Section 7.5].

Note that even though Remark 4 suggests that the optimization with $A = \tilde{A}$ is not really of interest here, it is still reasonable to start studying it for $s = 2$, to get some insight in the case where the formulas are explicitly available. We show the numerical results in Figure 4 and the resulting matrices in Table 1 and note that the optimization was done without any fine-tuning of the optimization routine itself towards our application[11].

We see that for the setting $A = \tilde{A}$ in the first row of Figure 4 the optimized A are such that we converge after two to five iterations, basically turning GMRES into a direct solver[12]. This seems natural looking at Table 1 – we observe that the Butcher tableau matrix A adapts the non-zero structure of the preconditioner, e.g., A becomes close to diagonal for P^{d}, so that the preconditioner

[10] The clustering point is for all of our preconditioners naturally equal to 1.

[11] We used the python implementation of the *Sequential Least SQuares Programming* method `scipy.optimize.minimize(method='SLSQP')`, see [8], but we observed similar results for the commonly used alternatives, e.g., with *Constrained Optimization BY Linear Approximation* (`method='COBYLA'`) or trust region methods (`method='trust-constr'`). We used the weight value $\omega = 10^{-5}$ but, again, the goal at this point is not to find the best performing parameters but rather give a generic comparison of the two approaches. We point out that in our experience taking ω much larger or much smaller resulted in comparable but somewhat worse results.

[12] Notice that in practice the solve accuracy also needs to be balanced with the discretization errors, hence even five iterations to obtain a relative residual smaller than machine precision might be considered a direct solver (i.e., converging after one or two steps) in some applications.

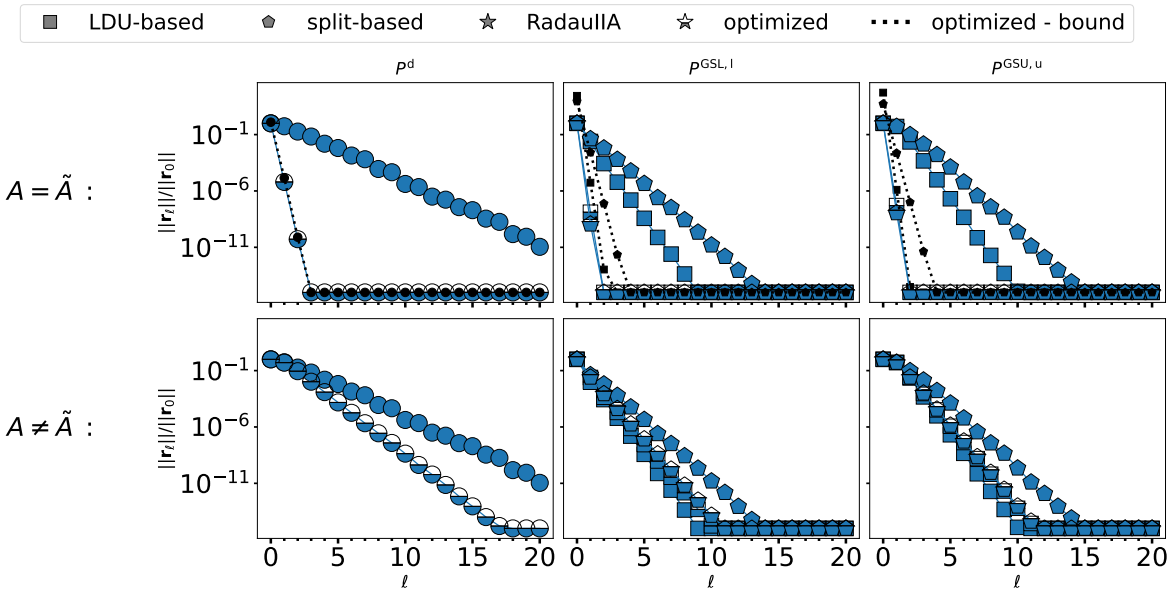

Figure 4: GMRES convergence (and bounds) for block preconditioners P^{d} (left column), $P^{\mathrm{GSL}}, P^{\mathrm{l}}$ (middle column) and $P^{\mathrm{GSU}}, P^{\mathrm{u}}$ (right column) for the two stage RadauIIA Butcher tableau (filled markers, same results as in Figure 2 and 3) and also the optimized setting considered above (half-filled markers). In the second row the optimization process changes *only* the matrix $\tilde{A}$ used to construct the preconditioner (and not M) in contrast to the first row where the optimized is used for construction of both the system matrix and the preconditioner (in this case we imposed A to be such that the resulting IRK method is at least of order two and is L-stable, the constraints can be found in [13, Section 7.3.4]). We used $\omega = 10^{-5}, N = 100$ and $\tau = h^{2/p}$ as above.

then becomes almost identical to M, reinforcing the point of Remark 4. Notably, this is achieved while not exploding the condition number of the eigenbasis of the preconditioned system.

For the second row, i.e., $A \neq \tilde{A}$ and optimizing $\tilde{A}$ that is used to construct the preconditioner, we see that for the block-diagonal preconditioner we can still obtain a considerable speed-up, as well as for the preconditioners $P^{\mathrm{GSL,GSU}}$. This is not the case, however, for the preconditioners $P^{\mathrm{l,u}}$. The reason is that with $A = \tilde{A}$, the spectrum of $M(P^{\mathrm{l,u}})^{-1}$ was *real* but this property is lost in the general case $A \neq \tilde{A}$. Hence, even though we have tightened the clustering of the eigenvalues by optimizing $\tilde{A}$ this wasn't significant enough to off-set the introduction of the complex eigenvalues of the preconditioned system.

We would like to note that we obtained similar results to the first row when considering $p = 3$ and A-stability (i.e., giving up the more restrictive property of L-stability) but for $p = 3$ and L-stable methods there seemed to be next to no gains from the extra optimization (see [13, Section 7.5]).

Last but not least we comment on the computational costs. In order to evaluate the *complete* eigenproperties of the preconditioned system $M(P^\star)^{-1}$ (or $(P^\star)^{-1}M$) we first calculate the eigen-decomposition of L (which is prohibitively costly) and then we calculate eigendecompositions of n matrices $X_k^\star$, each of dimension s, which can be done in parallel. Even without any parallelization the cost is linear in n with the constant corresponding to s^3. In practice, we often have (or can reasonably cheaply obtain) some estimate of the eigeninformation of L, e.g., estimates μ_1, μ_n of

	block diagonal	block lower-triangular		block upper-triangular	
	P^{d}	P^{GSL}	P^{l}	P^{GSU}	P^{u}
$A = \tilde{A}$	$\begin{bmatrix} 1.8 \cdot 10^{-1} & -3.8 \cdot 10^{-6} \\ 2.3 \cdot 10^{-6} & 3.9 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 3.5 \cdot 10^{-1} & -4.5 \cdot 10^{-9} \\ 2.5 \cdot 10^{-6} & 2.3 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 3.2 \cdot 10^{-1} & -3.0 \cdot 10^{-8} \\ 9.6 \cdot 10^{-1} & 2.7 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 3.5 \cdot 10^{-1} & 4.2 \cdot 10^{-1} \\ -5.8 \cdot 10^{-9} & 2.3 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 4.9 \cdot 10^{-1} & -5.2 \cdot 10^{-2} \\ 6.3 \cdot 10^{-8} & 1.7 \cdot 10^{-2} \end{bmatrix}$
$A \neq \tilde{A}$	$\begin{bmatrix} 9.3 \cdot 10^{-1} & 4.8 \cdot 10^{-2} \\ 4.0 \cdot 10^{-1} & 1.3 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 5.1 \cdot 10^{-1} & -6.6 \cdot 10^{-2} \\ 8.0 \cdot 10^{-1} & 3.0 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 4.1 \cdot 10^{-1} & -8.4 \cdot 10^{-2} \\ 7.4 \cdot 10^{-1} & 2.2 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 4.1 \cdot 10^{-1} & -8.2 \cdot 10^{-2} \\ 7.0 \cdot 10^{-1} & 3.7 \cdot 10^{-1} \end{bmatrix}$	$\begin{bmatrix} 4.1 \cdot 10^{-1} & -8.1 \cdot 10^{-2} \\ 7.2 \cdot 10^{-1} & 2.3 \cdot 10^{-1} \end{bmatrix}$

Table 1: The resulting matrices $A = \tilde{A}$ (first row) or $\tilde{A}$ (second row) corresponding to the results presented in Figure 4.

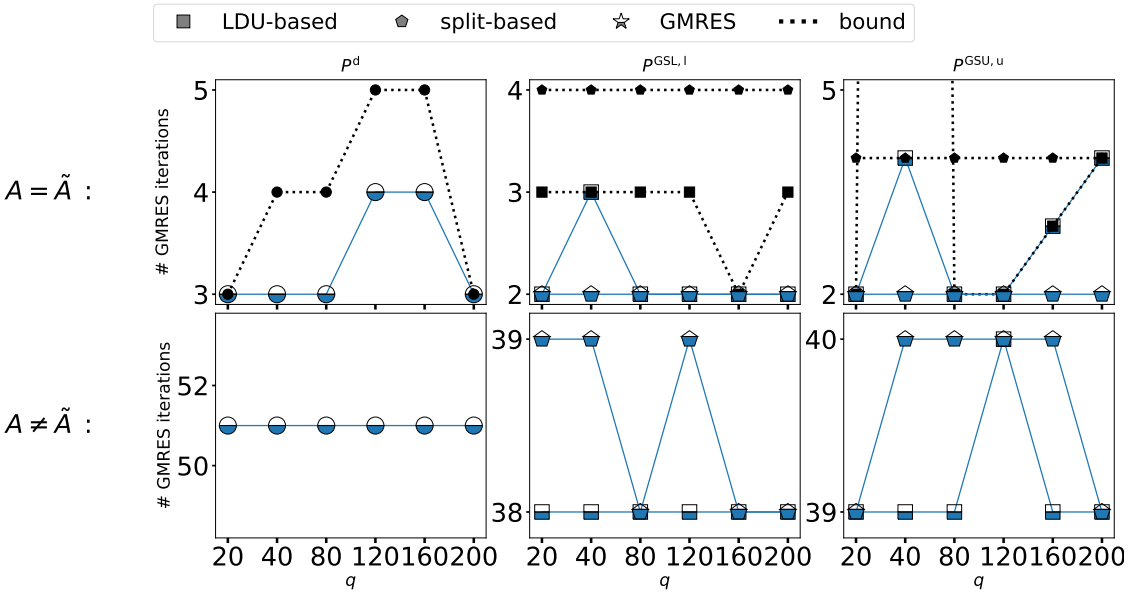

Figure 5: The number of GMRES iterations and the corresponding bound for the IRK method given by the result of optimization routine where we replace the true spectrum $\{\lambda_k\}_k$ with only q artificially spanned values in the interval spanned by the spectrum.

the true eigenvalues λ_1, λ_n or on the conditioning of V. The eigendecomposition of L can then be replaced by, e.g., considering only q "fake" λ_k sampled from $(\mu_{\min}, \mu_{\max})$ (assuming we know the spectrum of L is real).

We show a numerical illustration, considering the same setting as for Figure 4 and sample only q distinct λ_k (and thus θ_k) in the interval $\mu_{\min} := \theta_{\min}, \mu_{\max} := \theta_{\max}$. We show the evolution of the number of preconditioned GMRES with the optimized preconditioners depending on q in Figure 5. We see that for the spatial dimension $n = 99^2$ we already get comparable results to Figure 4 by sampling *only very few* "fake" θ_k compared to optimizing over the actual roughly ten thousands of them. Also, the bounds stay still quite descriptive – only for $q = 40$ and P^{GSU} have we found that running the optimization as a black-box does not give a useful descriptive bound[13]. This observation is key for either of the optimization approaches to be viable and further study of q for more involved settings (and its dependency on n) and/or sampling strategies for complex spectra is necessary. However, as the main interest is in s large, this will be presented in the follow-up work.

[13]In fact for that particular setting the optimization routine sets off in a direction of *inadmissible* A (singular) and does not find its way back. Fine tuning the parameters of the optimization routine does, however, fix this issue.

5 Generalizations and conclusion

We have shown that for two stage IRK methods the preconditioners from [15] can be analyzed explicitly using spectral techniques and the GMRES convergence behavior can be reasonably predicted using the worst-case GMRES bound. This bound can be evaluated directly before the computations and gives a theoretical background to the results observed in [15]. These results are given for a simple test problem but the analysis clearly extends to any diagonalizable spatial operator L – not just the Laplacian. The same is true also for the discretization scheme used – a finite elements scheme analysis can be done completely analogously (see [13, Section 7.7]). Some of the above can be generalized to s-stage IRK methods but due to the space restrictions these results will be presented in an upcoming manuscript.

We would also like to mention that both in[12] and [18], the authors multiplied M with $A^{-1} \otimes I_n$ from the left. The analysis above for this case is analogous and reveals that for some preconditioners we obtain much better properties and for others the performance deteriorates. This is naturally very dependent on the choice of A as well; for more details see [13, Section 7.6].

Last but not least, we would like to thank the anonymous reviewers for the careful reading of the manuscript and their insightful comments that helped us further improve it.

References

[1] M. Arioli, V. Pták, and Z. Strakoš. Krylov sequences of maximal length and convergence of GMRES. *BIT Numerical Mathematics*, 38(4):636–643, 1998.

[2] 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.

[3] V. Faber, J. Liesen, and P. Tichý. On Chebyshev polynomials of matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2205–2221, 2010.

[4] P. E. Farrell, R. C. Kirby, and J. Marchena-Menendez. Irksome: Automating Runge–Kutta time-stepping for finite element methods. *ACM Transactions on Mathematical Software*, 47 (4):1–26, 2021.

[5] B. Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Springer, Wiesbaden, 1996.

[6] A. Greenbaum and Z. Strakoš. *Matrices that generate the same Krylov residual spaces*. Springer, 1994.

[7] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17(3):465–469, 1996.

[8] D. Kraft. A software package for sequential quadratic programming. Technical report, DLR German Aerospace Center -– Institute for Flight Mechanics, 1988.

[9] J. Liesen and Z. Strakoš. GMRES convergence analysis for a convection-diffusion model problem. *SIAM Journal on Scientific Computing*, 26(6):1989–2009, 2005.

[10] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Oxford University Press, Oxford, 2013.

[11] K.-A. Mardal, T. K. Nilssen, and G. A. Staff. Order-optimal preconditioners for implicit Runge–Kutta schemes applied to parabolic PDEs. *SIAM Journal on Scientific Computing*, 29 (1):361–375, 2007.

[12] M. Neytcheva and O. Axelsson. Numerical Solution Methods for Implicit Runge-Kutta Methods of Arbitrarily High Order. In P. Frolkovič, K. Mikula, and D. Ševčovič, editors, *Proceedings of the Conference Algoritmy 2020*. Slovak University of Technology in Bratislava, Vydavateĺstvo SPEKTRUM, 2020.

[13] M. Outrata. *Schwarz methods, Schur complements, preconditioning and numerical linear algebra*. PhD thesis, University of Geneva, Math Department, 2022.

[14] D. Palitta and V. Simoncini. Optimality properties of Galerkin and Petrov–Galerkin methods for linear matrix equations. *Vietnam Journal of Mathematics*, 48(4):791–807, 2020.

[15] M. M. Rana, V. E. Howle, K. Long, A. Meek, and W. Milestone. A New Block Preconditioner for Implicit Runge-Kutta Methods for Parabolic PDE Problems. *SIAM Journal on Scientific Computing*, 43(5):S475–S495, 2021.

[16] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Other Titles in Applied Mathematics. SIAM, Philadelphia, Second edition, 2003. ISBN 978-0-89871-534-7.

[17] B. S. Southworth, O. Krzysik, and W. Pazner. Fast solution of fully implicit Runge–Kutta and discontinuous Galerkin in time for numerical PDEs, Part II: nonlinearities and DAEs. *SIAM Journal on Scientific Computing*, 44(2):636–663, 2022.

[18] B. S. Southworth, O. Krzysik, W. Pazner, and H. De Sterck. Fast solution of fully implicit Runge–Kutta and discontinuous Galerkin in time for numerical PDEs, Part I: The linear setting. *SIAM Journal on Scientific Computing*, 44(1):416–443, 2022.

[19] G. A. Staff, K.-A. Mardal, and T. K. Nilssen. Preconditioning of fully implicit Runge-Kutta schemes for parabolic PDEs. *Modeling, Identification and Control*, 27(2):109–123, 2006.

[20] C. F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, 123(1–2):85–100, 2000.

[21] G. Wanner and E. Hairer. *Solving Ordinary Differential Equations II : Stiff and Differential-Algebraic Problems*. Springer Berlin, Heidelberg, 1996.

[22] G. Wanner, S. P. Nørsett, and E. Hairer. *Solving Ordinary Differential Equations I : Non-Stiff Problems*. Springer Berlin, Heidelberg, 1987.