

1 **SPECTRAL ANALYSIS OF IMPLICIT S -STAGE BLOCK**
2 **RUNGE-KUTTA PRECONDITIONERS***

3 MARTIN J. GANDER[†] AND MICHAL OTRATA[‡]

4 **Abstract.** We analyze the recently introduced family of preconditioners in [21] for the stage
5 equations of implicit Runge-Kutta methods for two stage methods. We simplify the formulas for
6 the eigenvalues and eigenvectors of the preconditioned systems for a general s -stage method and use
7 these to obtain convergence rate estimates for preconditioned GMRES for some common choices of
8 the implicit Runge-Kutta methods. This analysis also allows us to qualitatively predict and explain
9 the main observed features of the GMRES convergence behavior and we illustrate our analysis with
10 numerical experiments.

11 **Key words.** implicit Runge-Kutta methods, stage equations, preconditioned GMRES, conver-
12 gence estimates, conformal maps

13 **MSC codes.** 65L06, 65F10, 65E05

14 **1. Introduction.** Runge-Kutta methods are a well-established family of one-
15 step solvers for systems of ordinary differential equations (ODEs; see [28, 27] for an
16 overview and further references). For implicit methods (IRK), their efficiency depends
17 on the efficiency of a solver for the so-called *stage equations* – in general a system
18 of ms non-linear equations, where m is the number of scalar ODEs in the system
19 and s is the number of stages of the Runge-Kutta method. An important application
20 arises from the space discretization of time-dependent partial differential equations
21 (PDEs), resulting in a system of ODEs with *very* large m . If the spatial operator is
22 *linear*, then the stage equations also form a system of linear algebraic equations and
23 are often solved by an iterative solver, e.g., a Krylov method. In [21], the authors
24 introduced a family of preconditioners for GMRES for the stage equations, numerically
25 showing that these preconditioners give an *outstanding* performance, especially under
26 refinement of the spatial mesh, i.e., as m grows. Recently, there have also been other
27 contributions in the direction of preconditioning the *fully implicit* Runge-Kutta stage
28 equations for PDEs, see [24, 23] but also [18, 17] and [2], introducing new ideas in
29 terms of implementation as well as formulation and testing these numerically on a
30 variety of test problems.

31 We focus on the setting considered in [21], expand the 2-stage method analysis
32 given in [9], and consider the general s -stage case, giving a theoretical background for
33 the performance and spectral properties observed. First, we recall some important
34 preliminaries in Section section 2 so that we can deliver the analysis, based on the
35 spectral analysis of the preconditioned system, in Section section 3. We support the
36 analysis by considering more involved examples in Section section 4.

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

39 (2.1)
$$\begin{aligned} \frac{\partial}{\partial t} u &= \Delta u + f && \text{in } \Omega \times (0, T_{\text{end}}), \\ u &= g && \text{on } \partial\Omega \times (0, T_{\text{end}}) \quad \text{and} \quad u = u_0 && \text{in } \Omega \times \{0\}, \end{aligned}$$

*Submitted to the editors DATE.

Funding: This work was partially supported by the SNF grant number 178752 and by the FCS Swiss Excellence PhD Fellowship program of the Swiss Federation (ESKAS No. 2019.0384).

[†]Section de Mathématiques, Université de Genève

[‡]Section de Mathématiques, Université de Genève

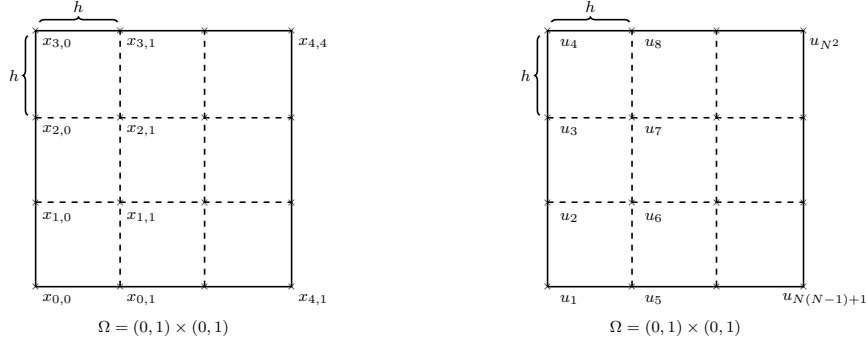


FIG. 1. Left: grid points for $N + 1 = 4$; right: lexicographical ordering of the unknowns for $N + 1 = 4$.

where Δ is the Laplace operator, f, g, u_0 are given functions and Ω is the unit square
 $\Omega := (0, 1) \times (0, 1)$. As in [9] we discretize in space using a finite difference scheme
 an equidistant grid with $N + 1$ rows and columns, and with 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

$$(2.2) \quad \frac{\partial}{\partial t} u_i(t) = \frac{u_{i-N}(t) + u_{i-1}(t) - 4u_i(t) + u_{i+1}(t) + u_{i+N}(t)}{h^2} + b_i^{(ST)}(t),$$

for $i = N + 1, \dots, N(N - 1) - 1$, where $b_i^{(ST)}(t)$ 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.,

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

and also analogously for $\mathbf{b}_k(t)$ and $\mathbf{b}(t)$, we rewrite (2.2) as

$$(2.3) \quad \frac{\partial}{\partial t} \mathbf{u}(t) = \frac{1}{h^2} L \mathbf{u}(t) + \mathbf{b}^{(ST)}(t),$$

with
 (2.4)

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

where L is of dimension $n := (N - 1)^2$ and the blocks T, I are of dimension $N - 1$.
 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.,

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

58 Having a *Butcher tableau*

$$59 \quad (2.5) \quad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b} \end{array} := \frac{\begin{array}{c|ccc} c_1 & a_{1,1} & \dots & a_{1,s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s,1} & \dots & a_{s,s} \\ \hline & b_1 & \dots & b_s \end{array}}{\quad},$$

60 the corresponding IRK method applied to (2.3) at the m -th time step gives the ap-
61 proximation $\mathbf{u}^m \approx \mathbf{u}(t_m)$ as

$$62 \quad (2.6) \quad \mathbf{u}^m = \mathbf{u}^{m-1} + \tau \sum_{i=1}^s b_i \mathbf{k}_i^m,$$

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

$$64 \quad (2.7) \quad \underbrace{\left(\begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} - \frac{\tau}{h^2} \begin{bmatrix} a_{1,1}L & \dots & a_{1,s}L \\ \vdots & \ddots & \vdots \\ a_{s,1}L & \dots & a_{s,s}L \end{bmatrix} \right)}_{\equiv I_s \otimes I_n - \frac{\tau}{h^2} (A \otimes L) =: M} \mathbf{k}^m = \begin{bmatrix} \frac{1}{h^2} L \mathbf{u}^{m-1} + \mathbf{b}^{(\text{ST})}(t_{m-1} + c_1 \tau) \\ \vdots \\ \frac{1}{h^2} L \mathbf{u}^{m-1} + \mathbf{b}^{(\text{ST})}(t_{m-1} + c_s \tau) \end{bmatrix},$$

65 with

$$66 \quad \mathbf{k}^m := [\mathbf{k}_1^m \quad \dots \quad \mathbf{k}_s^m]^T \in \mathbb{R}^{ns}.$$

67 The symbol \otimes stands for the Kronecker product (see [26] and references therein) and
68 we note that (2.7) can be reformulated into a *matrix equation*, which is in general
69 better suited for using a Krylov solver (see [20]). Here we focus on the analysis of
70 the results in [21] and thus we do not address this any further but a study of the
71 preconditioners from [21] in the matrix equations setting seems worthwhile. Having
72 $p \leq 2s$ as the order of convergence of the IRK method we assume that it is balanced
73 with the spatial discretization error, i.e., that $h^2 = C_e \tau^p$ for some $C_e > 0$.

74 The problem (2.7) with the sparse system matrix M can be very large for h (and
75 τ) small, suggesting an iterative solver such as GMRES, BiCG or GCR should be
76 used, which in turn requires a preconditioner to attain efficiency. In [21], the authors
77 introduce the block preconditioners

$$78 \quad (2.8) \quad \begin{aligned} P^d &= I_s \otimes I_n - \frac{\tau}{h^2} \text{diag}(A) \otimes L, \\ P^u &= I_s \otimes I_n - \frac{\tau}{h^2} D_A U_A \otimes L \quad \text{and} \quad P^l = I_s \otimes I_n - \frac{\tau}{h^2} L_A D_A \otimes L, \end{aligned}$$

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

$$81 \quad (2.9) \quad P^{\text{GSL}} = I_s \otimes I_n - \frac{\tau}{h^2} A_L \otimes L \quad \text{and} \quad P^{\text{GSU}} = I_s \otimes I_n - \frac{\tau}{h^2} A_U \otimes L,$$

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

$$84 \quad (A_L)_{ij} = \begin{cases} a_{ij} & \text{if } i \geq j \\ 0 & \text{otherwise} \end{cases}, \quad (A_U)_{ij} = \begin{cases} a_{ij} & \text{if } i \leq j \\ 0 & \text{otherwise} \end{cases}.$$

85 Some of these – P^d and P^{GSL} – were considered already in [25]. Notice that if $a_{ii} > 0$
 86 for all $i = 1, \dots, s$, then the preconditioners are invertible as L is symmetric, negative-
 87 definite. More general conditions for non-singularity of the preconditioners can be also
 88 derived analogously to [24, Lemma 1].

89 Using GMRES for a linear system $C\mathbf{x} = \mathbf{f}$ with C being diagonalizable, i.e.,
 90 $C = S\Lambda S^{-1}$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$, a standard convergence bound for the residuals
 91 \mathbf{r}_ℓ reads

$$92 \quad (2.10) \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S) \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{1 \leq i \leq d} |\varphi(\lambda_i)|,$$

93 where $\kappa(S)$ is the 2-norm condition number of the matrix S , see, e.g., [16, Section
 94 5.7.2]. We highlight some aspects of the bound (2.10) that are often used to study
 95 GMRES convergence behavior.

96 *Remark 2.1.* As indicated above, the spectral information of the system matrix
 97 in GMRES (in our case of the preconditioned system) does not generally govern the
 98 convergence (see [11], [10] and [1] and also [16, Chapter 2 and 5.7] and the references
 99 therein). If the system matrix is normal, i.e., it is diagonalizable with S unitary,
 100 then the spectral information is enough to evaluate the ideal GMRES bound (2.10).
 101 However, if C is non-normal, then a convincing argument needs to be put forward to
 102 validate linking spectral information with the convergence behavior of GMRES as the
 103 authors in [16, p. 303, Remark 1] point out.

104 Moreover, particular knowledge of the interaction of S and the initial residual \mathbf{r}_0
 105 can lead to a *qualitative* and *quantitative* improvement on (2.10), see, e.g., [15]. How-
 106 ever, studying GMRES behavior with the bound (2.10), this interaction is completely
 107 lost.

108 In cases where (2.10) is justifiable, the next step is usually to bound from above
 109 the mixed¹ min-max problem in the right-hand side of (2.10) by replacing the discrete
 110 set over which we take the maximum, let us denote it by σ^{discr} , by a non-discrete one,
 111 which we denote by $\sigma^{\text{non-discr}}$, so that we have $\sigma^{\text{discr}} \subset \sigma^{\text{non-discr}}$. We highlight two
 112 important aspects of this step:

- 113 (a) It is *functional* only if we can further bound or evaluate the solution of the
 114 min-max problem over $\sigma^{\text{non-discr}}$ and obtain a reasonably fast convergence
 115 estimate.
- 116 (b) It is *appropriate* only if² $\partial_{\mathbb{C}}\sigma^{\text{non-discr}}$ is reasonably uniformly covered by
 117 σ^{discr} .³ In case of clusters, we should consider having $\sigma^{\text{non-discr}}$ as a union
 118 of separate non-discrete sets $\sigma_i^{\text{non-discr}}$ each of which captures one of the
 119 clusters, i.e., is covered by one of the clusters reasonably uniformly.

120 For example, in (2.10) we can replace the spectrum $\sigma^{\text{discr}} = \{\lambda_1, \dots, \lambda_d\}$ by a disc
 121 containing all of the eigenvalues $\sigma^{\text{non-discr}} = \{z \in \mathbb{C} \mid |z - c| \leq \rho\}$. Assuming $|c| > \rho$,

¹In the sense of the minimum is over a non-discrete set while the maximum is over a discrete one.

²We denote the boundary of a set $S \subset \mathbb{C}$ in \mathbb{C} by $\partial_{\mathbb{C}}S$.

³Intuitively, we could expect that the bound will be appropriate only if σ^{discr} covers the entirety of $\sigma^{\text{non-discr}}$ but because polynomials of complex variables are harmonic we can conclude that the maximum of the modulus of a polynomial over the set $\sigma^{\text{non-discr}}$ is attained along $\partial_{\mathbb{C}}\sigma^{\text{non-discr}}$ and therefore only the relation of $\partial_{\mathbb{C}}\sigma^{\text{non-discr}}$ and σ^{discr} is important for the GMRES bound, see [4, Section 2].

122 a crude but sometimes useful approximation of the original bound is available,

$$123 \quad (2.11) \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S) \left(\frac{\rho}{|c|} \right)^k,$$

124 see [22, Section 6.11.2, Corollary 6.33 and Lemma 6.26 and below]. Here, $\sigma^{\text{non-discr}} =$
 125 $\{z \in \mathbb{C} \mid |z - c| \leq \rho\}$ was clearly chosen with the *functionality* aspect in mind as we
 126 know the polynomial that realizes the bound (see [22, Lemma 6.26]) and it gives
 127 a good convergence bound as long as $\rho \not\approx |c|$. However, it is usually far from being
 128 *appropriate* if the eigenvalues don't spread uniformly over the circle bounding the disc.
 129 One notable exception is the case of tightly clustered eigenvalues around a single point
 130 c – in this case the clustering usually makes this bound appropriate as we can choose ρ
 131 *very* small. We emphasize that the adjectives *functional* and *appropriate* make sense
 132 only if the original bound (2.10) was itself descriptive of the GMRES convergence
 133 bound, i.e., only if the system matrix is either close to normal or the initial residual is
 134 restricted to a subspace on which the system matrix is not too far from being normal.

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

$$139 \quad (3.1) \quad V := [\mathbf{v}_1, \dots, \mathbf{v}_n], \quad \text{and} \quad Q := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix} \in \mathbb{R}^{sn \times sn}.$$

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

$$141 \quad (3.2) \quad \tilde{M} = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} - \frac{\tau}{h^2} \begin{bmatrix} a_{1,1}\Lambda & \dots & a_{1,s}\Lambda \\ \vdots & \ddots & \vdots \\ a_{s,1}\Lambda & \dots & a_{s,s}\Lambda \end{bmatrix},$$

142 with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. With the preconditioners proposed in (2.8-2.9) we write
 143 the spectrum of the preconditioned system as

$$144 \quad \text{sp}(MP^{-1}) = \text{sp}(Q^T MP^{-1}Q) = \text{sp}(Q^T MQQ^T P^{-1}Q) = \text{sp}(\tilde{M}\tilde{P}^{-1}),$$

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

152 **LEMMA 3.1** (see [9, Lemma 1]). *Let $C \in \mathbb{R}^{ns \times ns}$ be a real matrix with block*
 153 *structure such that every block is a square diagonal matrix, i.e.,*

$$154 \quad (3.3) \quad C = \begin{bmatrix} \Lambda_{11} & \dots & \Lambda_{1s} \\ \vdots & \ddots & \vdots \\ \Lambda_{s1} & \dots & \Lambda_{ss} \end{bmatrix}, \quad \text{with} \quad \Lambda_{ij} = \text{diag}(\lambda_1^{(ij)}, \dots, \lambda_n^{(ij)}) \quad \forall ij.$$

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

$$156 \quad (3.4) \quad \Pi^T C \Pi = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_n \end{bmatrix} \quad \text{with} \quad C_\ell = \begin{bmatrix} \lambda_\ell^{(11)} & \dots & \lambda_\ell^{(1s)} \\ \vdots & \ddots & \vdots \\ \lambda_\ell^{(s1)} & \dots & \lambda_\ell^{(ss)} \end{bmatrix} \in \mathbb{R}^{s \times s},$$

157 for any $\ell = 1, \dots, n$.

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

$$161 \quad \text{sp}(C) = \bigcup_{\ell=1}^n \bigcup_{i=1}^s \mu_\ell^{(i)},$$

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

$$164 \quad \kappa(V) = \max_{\ell=1, \dots, s} \kappa(V_\ell),$$

165 where $\kappa(\cdot)$ is the 2-norm condition number.

166 *Remark 3.2.* We note that an analogous lemma to Lemma 3.1 can also be for-
 167 mulated for non-normal matrices (replacing Q^T by Q^{-1}). Considering the Jordan
 168 canonical (or the Schur decomposition form) of C_ℓ , Lemma 3.1 can be reformulated
 169 to obtain a block upper bi-diagonal (or block upper-triangular) matrix.

170 To shorten the notation we set

$$171 \quad (3.5) \quad \theta_k := \frac{\tau}{h^2} \lambda_k \quad \text{and} \quad \Theta := \frac{\tau}{h^2} \Lambda,$$

172 as these quantities appear always together in the computations. By a direct calcula-
 173 tion (see [19, Appendix B.8]) we get the limit behavior of θ_k as $\tau, h \rightarrow 0$,

$$174 \quad (3.6) \quad \underbrace{\begin{aligned} (\theta_n, \theta_1) &\rightarrow \left(-\frac{8}{C_e}, 0\right), \\ (\theta_1^{-1}, \theta_n^{-1}) &\rightarrow \left(-\infty, -\frac{C_e}{8}\right), \end{aligned}}_{(\text{LIM})_{p=1}} \quad \underbrace{\begin{aligned} (\theta_n, \theta_1) &\rightarrow (-\infty, 0), \\ (\theta_1^{-1}, \theta_n^{-1}) &\rightarrow (-\infty, 0). \end{aligned}}_{(\text{LIM})_{p>1}}$$

175 Next we define the s -by- s matrices

$$176 \quad M_k := \begin{bmatrix} 1 - a_{11}\theta_k & -a_{12}\theta_k & \dots & -a_{1s}\theta_k \\ -a_{21}\theta_k & 1 - a_{22}\theta_k & & \vdots \\ \vdots & & \ddots & \vdots \\ -a_{s1}\theta_k & \dots & \dots & 1 - a_{ss}\theta_k \end{bmatrix} \quad \text{and} \quad P_k^* := \begin{bmatrix} 1 - \alpha_{11}\theta_k & -\alpha_{12}\theta_k & \dots & -\alpha_{1s}\theta_k \\ -\alpha_{21}\theta_k & 1 - \alpha_{22}\theta_k & & \vdots \\ \vdots & & \ddots & \vdots \\ -\alpha_{s1}\theta_k & \dots & \dots & 1 - \alpha_{ss}\theta_k \end{bmatrix},$$

177 where α_{ij} are the entries of the replacement for A in M , e.g., taking $\star = d$ we have
 178 $\alpha_{ij} = a_{ij}$ for $i = j$ and $\alpha_{ij} = 0$ otherwise, while taking $\star = u$ we have $\alpha_{ij} = (D_A U_A)_{ij}$
 179 where $A = L_A D_A U_A$ is the LDU factorization of A and so on. Using Lemma 3.1, we
 180 obtain the following result.

181 PROPOSITION 3.3. Take M as in (2.7) and a preconditioner P from (2.8, 2.9).
 182 Assuming P is invertible, the spectrum of MP^{-1} (or $P^{-1}M$) is given as the union of
 183 the spectra of the matrices X_k given by

$$184 \quad (3.7) \quad X_k^* := M_k (P_k^*)^{-1} \quad (\text{or } P_k^{-1}M_k),$$

185 for $k = 1, \dots, n$. If all X_k^* are diagonalizable with

$$186 \quad (3.8) \quad (S_k^*)^{-1} X_k^* S_k^* = \text{diag}(\xi_1^{(k)}, \dots, \xi_s^{(k)}),$$

187 then the condition number of the matrix of the eigenvectors of the preconditioned
 188 system is given by

$$189 \quad \kappa(W) \cdot \max_{k=1, \dots, n} \kappa(S_k^*).$$

190 If the θ_k have multiplicity at most m , then the eigenvalues of the preconditioned system
 191 have algebraic multiplicity at most ms . In particular, the preconditioned system can
 192 be non-diagonalizable but the longest Jordan vector chain has length at most ms .

193 *Proof.* Transforming MP^{-1} (or $P^{-1}M$) into the basis of Q we use Lemma 3.1
 194 for the matrix $\tilde{M}\tilde{P}^{-1}$ (see (3.2)) and obtain the result. \square

195 Now we are ready to generalize the results shown in [9] for $s = 2$ to a general s -stage
 196 method.

197 COROLLARY 3.4 ([19, Proposition 7.5]). Under the assumptions of Proposi-
 198 tion 3.3, we have for the right-preconditioner P^d the formula

$$199 \quad (3.9) \quad X_k^d = \begin{bmatrix} 1 & -\frac{a_{12}\theta_k}{1-a_{22}\theta_k} & \cdots & -\frac{a_{1s}\theta_k}{1-a_{ss}\theta_k} \\ -\frac{a_{21}\theta_k}{1-a_{11}\theta_k} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{a_{1s}\theta_k}{1-a_{11}\theta_k} & \cdots & \cdots & 1 \end{bmatrix},$$

200 with the characteristic polynomial

$$201 \quad p_k^{(s)}(\lambda) = (1 - \lambda)^s + \beta_{s-2}(1 - \lambda)^{s-2} + \beta_{s-3}(1 - \lambda)^{s-3} + \dots + \beta_1(1 - \lambda) + \beta_0,$$

202 where β_j are continuous functions of θ_k and a_{ii} for $i = 1, \dots, s$. Hence, the eigenvalues
 203 become $1 - \mu$, where μ is a root of the parametrized polynomial

$$204 \quad \tilde{p}_k^{(s)}(t) = t^s + \beta_{s-2}t^{s-2} + \beta_{s-3}t^{s-3} + \dots + \beta_1t + \beta_0.$$

205 COROLLARY 3.5 ([19, Proposition 7.6]). Under the assumptions of Proposi-
 206 tion 3.3, the block upper-triangular preconditioners $P^{\text{GSU},u}$ give

$$(3.10) \quad 207 \quad X_k^{\text{GSU},u} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \star & & & & & \\ \vdots & (M_k(P_k^{\text{GSU},u})^{-1})_{2:s,2:s} & & & & \\ \star & & & & & \end{bmatrix}, \quad X_k^{\text{GSL},l} = \begin{bmatrix} 1 & \star & \cdots & \cdots & \cdots & \star \\ 0 & & & & & \\ \vdots & ((P_k^{\text{GSL},l})^{-1}M_k)_{2:s,2:s} & & & & \\ 0 & & & & & \end{bmatrix},$$

208 and hence have one eigenvalue equal to one for each k . The entries replaced by \star
 209 above do not affect the spectrum, only the eigenbasis.

210 These results suggest 1 as a natural “central point” of the spectrum of the pre-
 211 conditioned system, generalizing the observations made for $s = 2$. We note that using
 212 these results we get both quantitative and qualitative insight into the spectra shown
 213 in [21, Figure 4.1 – 4.4], e.g., we see that for $s = 3$ the eigeninformation of $M(P^u)^{-1}$
 214 and $(P^l)^{-1}M$ can still be obtained explicitly (see also [19, Section 7.4]) and on the
 215 other hand for $s \geq 6$ there is no hope for these in general – but any bound on the
 216 eigeninformation of L can be used to obtain a bound on the eigeninformation of the
 217 preconditioned system by calculating with X_k , see [9, Section 4].

218 We show the spectra of the preconditioned systems and the corresponding GM-
 219 RES convergence behavior in Figure 2 and 3, demonstrating observations and results
 220 from above. Notably, the bounds leave something to be desired, especially for P^d
 221 where they are not descriptive at all. Moreover, increasing s seems to noticeably af-
 222 fect the quality of the preconditioners – see also [21] for further numerical tests with
 223 various s and h . These numerical examples (as well as the ones in [2, 9]) are, as far as
 224 we can tell, representative of the general experience with these preconditioners. We
 225 highlight several key features illustrated in Figures 2 and 3 that remained true in all
 226 of our experiments:

- 227 1. For s small, we have observed the staircase-like convergence behavior visible
 228 in the left upper-most plot in Figure 3 and this was most pronounced for the
 229 preconditioner P^d .
- 230 2. We have usually not observed the desired *superlinear* convergence behavior,
 231 except for a speed-up after an initial stagnation (or slower speed convergence)
 232 phase.
- 233 3. In the vast majority of cases, the number of GMRES iterations to reach a
 234 certain tolerance grows only very moderately under mesh refinement and for
 235 P^u, P^l it remains almost constant.
- 236 4. In all of the experiments the spectra had the characteristic arc-like structure
 237 that we see in Figure 2.

238 Our goal is to explain all these features here as well as to investigate other bounds
 239 or estimates that would be more descriptive of the convergence behavior. This insight
 240 is of clear interest on its own but can be also used to further improve the used
 241 methods, e.g., looking at *numerical optimization* of the Butcher tableau in the spirit
 242 of [9, Section 4]. We also note that the above results translate in a straight-forward
 243 fashion to the *transformed system* after we multiply (2.7) with $(A^{-1} \otimes I_n)$ from the
 244 left, obtaining

$$245 \quad \underbrace{\left(A^{-1} \otimes I_n - \frac{\tau}{h^2} I_s \otimes L \right)}_{=: M^{\text{transf}}} \mathbf{k}^m = (A^{-1} \otimes I_n) \begin{bmatrix} \frac{1}{h^2} L \mathbf{u}^{m-1} + \mathbf{b}^{(\text{BC})}(t_{m-1} + c_i \tau) \\ \vdots \\ \frac{1}{h^2} L \mathbf{u}^{m-1} + \mathbf{b}^{(\text{BC})}(t_{m-1} + c_i \tau) \end{bmatrix},$$

246 and getting analogously the preconditioners,

$$247 \quad R^d = \text{diag}(A^{-1}) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$

$$R^l = (D_{A^{-1}} U_{A^{-1}}) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L \quad \text{and} \quad R^u = (L_{A^{-1}} D_{A^{-1}}) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$

$$R^{\text{GSL}} = (A^{-1})_L \otimes I_n - \frac{\tau}{h^2} I_s \otimes L \quad \text{and} \quad R^{\text{GSU}} = (A^{-1})_U \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$

248 where A^{-1} has the LDU factorization $A^{-1} = L_{A^{-1}} D_{A^{-1}} U_{A^{-1}}$ and $(A^{-1})_{L,U}$
 249 are defined analogously to (2.9). These preconditioners were proposed in [18] and then

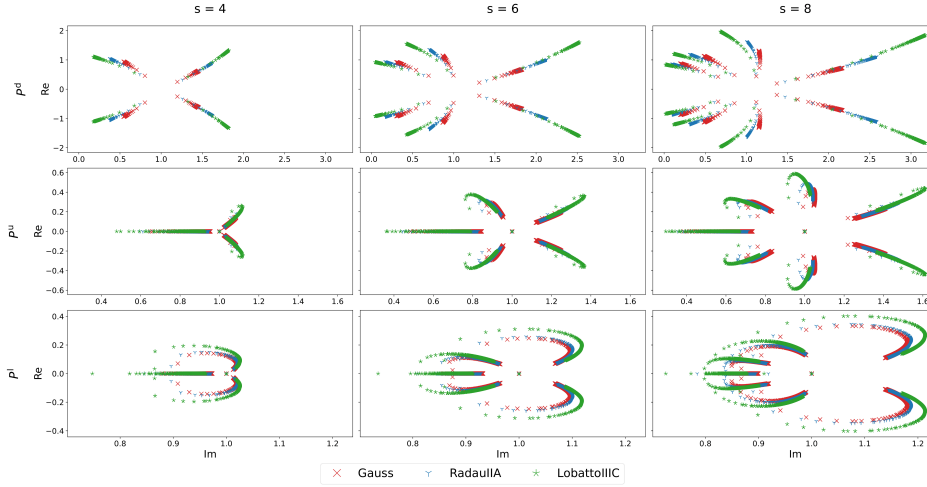


FIG. 2. The spectra of the preconditioned systems $M(P^{u,d})^{-1}$ and $(P^l)^{-1}M$ for $s = 4, 6, 8$ and for three classical choices of fully implicit Runge-Kutta schemes - Gauss, RadauIIA and LobattoIIIC. The spectra seemingly assemble in s “branches” in the first row and into $s - 1$ “branches” in the other two with a central point at $1 + 0i$. We set $N = 50$.

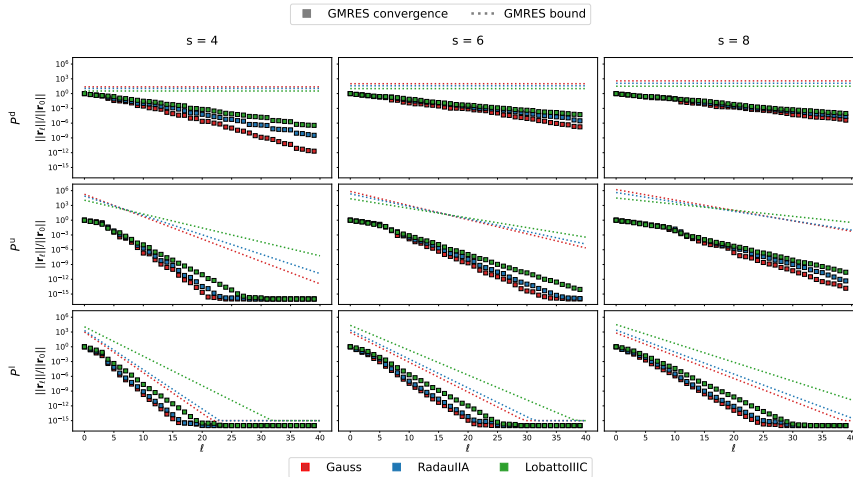


FIG. 3. The preconditioned GMRES convergence behavior for the preconditioned systems $M(P^{u,d})^{-1}$ and $(P^l)^{-1}M$ for $s = 4, 6, 8$ and three classical choices of fully implicit Runge-Kutta schemes - Gauss, RadauIIA and LobattoIIIC - together with the GMRES bound (2.11) with $c = 1$ (we set the values to 1 if $\rho \geq 1$). We set $N = 50$.

250 used further in [17] but also [24, 23]. For a *general* Butcher tableau, it is not possible to
 251 say whether the preconditioned transformed system gives a better performance than
 252 the original one. However, in [24, 23] the authors propose different preconditioners
 253 and this analysis within this framework is going to be considered elsewhere. Also,
 254 we note that the extension of the above analysis for FEM discretization is a straight-
 255 forward task – more details on both of these topics can be found in [19, Sections 7.6
 256 and 7.7].

257 **3.1. Spectral analysis.** Next we turn to the spectral analysis, keeping in mind
 258 its limitation in the sense of Remark 2.1. For block-diagonal problems we obtain

$$259 \quad (3.11) \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{j=1, \dots, n} \|\varphi(X_j)\|,$$

260 which was studied in [8], where the authors showed that the extremal polynomials
 261 (i.e., the polynomial realizing the above bound) satisfies the equioscillation property
 262 but only every s iterations, where s is the size of the diagonal blocks. Relabeling the
 263 blocks in (3.11) we get

$$264 \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{j=1, \dots, n} \|\varphi(X_j)\| = \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{\theta_j \in \text{sp}(\frac{\tau}{h^2} L)} \|\varphi(X_{\theta_j})\|.$$

265 Assuming each X_{θ_j} is diagonalizable as in Proposition 3.3, we notice that $\{\theta_j\}$ covers
 266 reasonably well the intervals $I_{h,\tau,\dots}$ as $h \rightarrow 0$ (see (3.6)) and, in the spirit of Section 2,
 267 the natural bound of (3.11) becomes

$$268 \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{\theta \in I_{h,\tau,\dots}} \|\varphi(X_\theta)\|.$$

269 First, let us assume there is a uniform bound $\kappa(S_\theta) \leq \kappa_S$ for all $\theta \in I_{h,\tau,\dots}$, which
 270 experimentally seems to be the case (see [19]) and can be confirmed analytically for
 271 $s = 2, 3$ (see [9]) – this is an important and non-trivial assumption and a proper justifi-
 272 cation is an open problem. Next, we notice that the matrices X_θ depend *smoothly*⁴
 273 on θ and as a result so do their eigenproperties. In particular, the eigenvalues $\xi_\theta^{(i)}$ of
 274 X_θ will – by definition – form an *algebraic curve*⁵ with s *arcs* (sometimes also called
 275 *branches*) some of which can be degenerate, e.g., reduced to just a point (incidentally,
 276 this is the case for at least one arc of the algebraic curve for any of the triangular
 277 preconditioners due to Corollary 3.5). Denoting the algebraic curve for the given
 278 Butcher tableau A and a choice of preconditioner P^* by Γ , we obtain

$$279 \quad (3.12) \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{\theta \in I_{h,\tau,\dots}} \kappa(S_\theta) \max_{i=1, \dots, s} \left| \varphi\left(\xi_\theta^{(i)}\right) \right| \leq \kappa_S \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{\xi \in \Gamma} |\varphi(\xi)|.$$

280 Notice that if we replace in (3.12) the interval $I_{h,\tau,\dots}$ with its limit I_{lim} as $h, \tau \rightarrow 0$
 281 (see (3.6)), we obtain a bound for all mesh sizes. Noticing that, in our case, the
 282 preconditioned system matrix has a limit as θ tends to either of the endpoints of
 283 I_{lim} , it follows that the arcs of the corresponding algebraic curve correspond to the
 284 eigenvalues of these limit matrices. Hence, the effect of mesh refinement becomes
 285 sampling more points along Γ and stretching it towards these fixed endpoints (and
 286 possibly in increasing κ_S). This suggests that from a certain mesh size onward, the
 287 mesh refinement will have little effect on Γ and hence will not affect the min-max part
 288 of (3.12), shedding some light on why these preconditioners are quite robust under
 289 mesh refinement.

⁴That is, for our model problem. This changes if we consider, e.g., an indefinite spatial operator L instead of the negative-definite Laplacian.

⁵We say that Γ is an algebraic curve provided there exists a bi-variate polynomial $p(\theta, \xi)$ such that $\Gamma = \{(\theta, \xi) \mid p(\theta, \xi) = 0\}$. Locally, this can also be viewed through the lens of perturbation theory, see [12, Chapter 2 Section 1.1].

290 *Remark 3.6.* Note that the numerical experiments in [21, 2] as well as in [19]
 291 and in Section 4 clearly show that the spectra of the preconditioned systems cover
 292 reasonably well an algebraic curve. For two-stage methods, this behavior has been
 293 observed, proved and used to obtain descriptive GMRES bounds in [9]. Moreover, for
 294 any algebraic curve Γ we have $\Gamma = \partial_{\mathbb{C}}\Gamma$, which is convenient from the point of view
 295 of choosing $\sigma^{\text{non-discr}}$, see Remark 2.1 and below.

296 We also emphasize that, in general, these preconditioners do not cluster eigenval-
 297 ues (that is, any more than the $\theta \in I_{h,\tau,\dots}$ already are) but rather place them along a
 298 particular algebraic curve $\Gamma \subset \mathbb{C}$. Hence, in general, we can reasonably expect linear
 299 convergence as opposed to superlinear, which can often be linked with clusters and
 300 numbers of outliers, in the sense of [16, Section 5.6.4].

301 Remark 3.6 also explains that the bound (2.11) is unlikely to be very descriptive or
 302 even usable. Indeed, the algebraic curves can reach into the left half-plane $\{\text{Re}(z) < 0\}$
 303 (making the bound useless due to 0 being included in the bounding circle) or, in the
 304 more favorable case, the arcs of the algebraic curve are *extremely* unlikely to align with
 305 the circle so that the bound have some resemblance of being what we earlier called
 306 *appropriate*. Naturally, the bound on the right-hand side of (3.12) is constructed to
 307 remedy that but the key question becomes if this bound is also *functional*, namely if
 308 we can (approximately) evaluate it.

309 To this end, we follow the excellent paper [4] on this topic and start by looking at
 310 the *asymptotic* convergence rate (justified by Remark 3.6 above). Considering (3.12)
 311 we are led to look at the so-called *logarithmic capacity* of Γ , denoted by $\text{cap}(\Gamma)$,
 312 which can be viewed as a measure of a compact set without isolated points in \mathbb{C} .
 313 In fact it is known to asymptotically correspond to the maximal modulus of the
 314 *extremal polynomials* (sometimes also called Chebyshev polynomials) associated with
 315 Γ , namely

$$316 \quad (3.13) \quad \left(\min_{\deg(\varphi) \leq \ell} \max_{z \in \Gamma} |\varphi(z)| \right)^{1/\ell} \rightarrow \text{cap}(\Gamma), \quad \text{as } \ell \rightarrow +\infty,$$

317 where the quantity on the left-hand side relates to the quantities we have seen in the
 318 GMRES bounds. There are two important caveats to using $\text{cap}(\Gamma)$. The first one,
 319 which has been also highlighted as a caveat for using the analysis in [4] overall, is
 320 the fact that that (3.13) only provides some information about the *limit behavior* as
 321 $\ell \rightarrow +\infty$, whereas we are interested in the behavior for relatively small values of ℓ ,
 322 say $\ell \leq 50$ or 100 . To large extend this issue is addressed by Remark 3.6 that states
 323 that we expect a linear convergence rate throughout the iteration. The second one
 324 is the fact that (3.13) describes the limit scaling of the maximal modulus over *all*
 325 *polynomials* – it lacks the crucial scaling $\varphi(0) = 1$ of Krylov methods. This issue can
 326 be fixed by re-scaling (see [4, Section 2]), shifting our attention from the logarithmic
 327 capacity to *Green's functions associated with* Γ , as long as Γ is compact and without
 328 any isolated points.

329 Things simplify considerably if we assume that Γ is connected as then the nor-
 330 malized quantity

$$331 \quad \left(\min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{z \in \Gamma} |\varphi(z)| \right)^{1/\ell}$$

332 can be evaluated directly using conformal maps, in particular the Schwarz-Christoffel

333 maps. Without going into the details (the interested reader can find these in [4,
334 Sections 2 and 3]), we obtain the *asymptotic convergence factor estimate* ρ_{est} as

$$335 \quad \rho_{\text{est}} := \lim_{\ell \rightarrow +\infty} \left(\min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \max_{z \in \Gamma} |\varphi(z)| \right)^{1/\ell} = \frac{1}{|\Phi(0)|},$$

336 where $\Phi(z)$ is the Schwarz-Christoffel map that maps the exterior of Γ to the ex-
337 terior of the unit circle. In [4, Section 3, Theorem 2 and below], the authors put this as

338 *“... if Γ is connected, the estimated asymptotic convergence factor*
339 *for a matrix iteration depends on how far the origin is from Γ –*
340 *provided that this distance is measured by level curves associated with*
341 *the exterior conformal map.”*

342 We would like to emphasize the word *estimate* when talking about ρ_{est} because we
343 truly do not get a bound anymore – in fact we get an *underestimate* as highlighted
344 also in [4, Section 5, equation (STEP1) and also Table 1]. However, we expect this
345 estimate to be descriptive as explained above.

346 For not too complicated connected, compact sets the map Φ and its value at the
347 origin can be calculated using the Schwarz-Christoffel MATLAB toolbox [3], but we
348 immediately notice that in Figure 2 the set of eigenvalues along Γ is not connected
349 and the actual algebraic curve Γ itself is also not available in an easy form, i.e., neither
350 of these can be directly given as an input to the SC toolbox. We take the natural next
351 step and approximate Γ by its linear interpolation based on the available eigenvalues
352 $\xi_{\theta}^{(i)}$. The linear interpolation gives us a good approximation of the arcs of Γ and we
353 use the point $1 + 0i$ as the natural point to join them (also by linear interpolation)
354 and denote the resulting set Γ_h . Recalling the limit behavior in (3.6), we also see that
355 Γ_h will tend towards Γ as $h \rightarrow 0$ for our model problem.

356 The calculation of $\xi_{\theta}^{(i)}$ is independent for each $k = 1, \dots, n$ but for large n the SC
357 toolbox can suffer numerically when calculating with Γ_h that is densely populated by
358 the interpolation points – both in the sense of large computational complexity as well
359 as in the sense of numerical issues (called *over-crowding*, see [3] but also [5, Section
360 2.6]). Moreover, we usually have only rough estimates on the extremal eigenvalues
361 θ_{\min} and θ_{\max} of L rather than its full spectrum. To this end, we recall the idea in [9,
362 Section 4] and instead of calculating Γ_h we use the information about $\theta_{\min, \max}$ and
363 artificially sample a fixed number of “fake” points ϑ_k between them, say q of them.
364 Then we replace θ_k by ϑ_k in the definition of Γ_h , obtaining Γ_q – an approximation of
365 Γ_h (and a further approximation of Γ) based on the linear interpolation given by the
366 eigenvalues of the matrices X_{ϑ_k} . We illustrate these points in Figure 4.

367 Another key point is that using the SC toolbox⁶ – namely the functions `externmap`
368 and `evalinv` – has difficulties (as far as we understand it) when the arcs of Γ_q inter-
369 sect, e.g., as is the case for $s = 8$ and the preconditioner P^1 , see Figure 2. Intuitively,
370 this makes sense as the exterior of Γ_q then has multiple components, making the
371 original set-up more complicated (a theoretical treatment of such problems could be
372 approached based on [7]). We address this issue by taking the “envelope” of the arcs
373 – if two arcs intersect, we follow the one staying outwards, e.g., in the case of $s = 6$
374 and the preconditioner P^1 we would exclude some portion of the arcs with smaller

⁶In our case, Γ_q qualifies as a degenerate polygon acceptable by the toolbox.

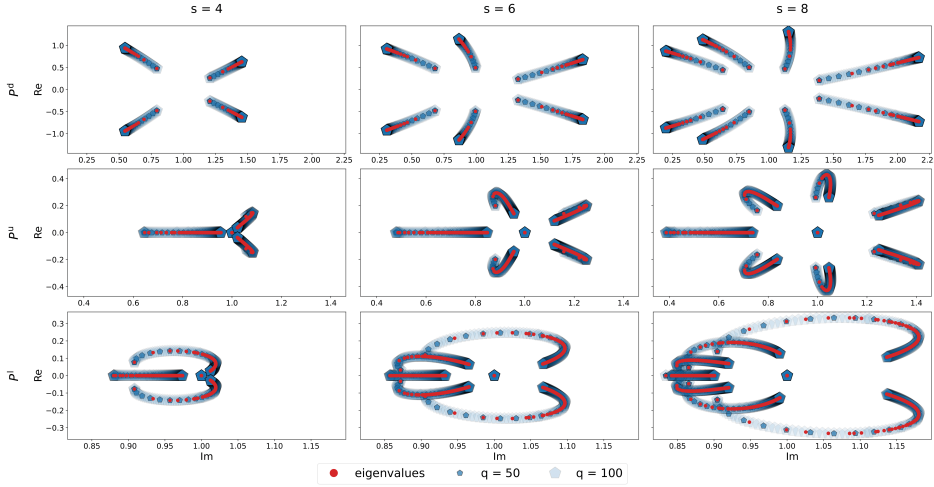


FIG. 4. The eigenvalues of the matrices X_{θ_k} (red) and X_{ϑ_k} (blue, for different values of q), using the preconditioner P^d . Joining these together with line segments would yield the curves Γ_h (red) and Γ_q (blue).

375 imaginary part (the densely populated portions) as these lie “inward” relative to the
 376 arcs with the larger imaginary part, see Figure 5. Finally, we illustrate the calculated
 377 Schwarz-Christoffel maps – or rather their contours – in Figure 5 together with the
 378 used inputs Γ_q (with the exception of $s = 6, 8$ and the preconditioner P^l , where we
 379 used the “envelopes”) and also the asymptotic convergence factor estimate ρ_{est} in
 380 Figure 6. First, we see that the results in Figure 6 fully support the arguments in
 381 Remark 3.6 for considering ρ_{est} as *the* descriptive quantity for the convergence factor.
 382 Including an estimate for κ_S then gives also an estimate for GMRES convergence – not
 383 just its rate, see Section 4. Second, we note that for $s = 8$ and the preconditioner P^u ,
 384 the arcs turned so that the right-most arcs almost intersect themselves. This causes
 385 problems for the toolbox, which during the calculations raises a flag stating that the
 386 calculated map did not converge as expected. Although the predicted ρ_{est} seems ac-
 387 curate, we see in Figure 5 that contours have ripples, confirming that the calculated
 388 results should be taken with caution. This can be fixed by a similar “envelope-like”
 389 approach we described for $s = 6, 8$ and the preconditioner P^l , see Section 4, obtaining
 390 a further approximation. Although there are a few similar caveats concerning the
 391 implementation of the above ideas, we have always found that a simple solution (such
 392 as considering the envelope or pruning the fake points in order to alleviate the crowd-
 393 ing) can be used to fix them *and* still give an *appropriate* insight into the GMRES
 394 convergence factor. As long as κ_S does not completely dominate the ideal GMRES
 395 bound (2.10) this then translates to descriptive GMRES convergence estimates, see
 396 Section 4.

397 The above analysis also gives insight into the staircase-like behavior, which has
 398 been observed and explained for $s = 2$ and the preconditioner P^d in [9] working with
 399 the minimal residual polynomial φ_ℓ^{MR} (sometimes also called the GMRES polynomial;
 400 see [16, Section 5.7.1]). The arguments used in [9] remain valid as long as the branches
 401 are not very close to each other⁷ – as long as the branches are far apart, the maximum

⁷In [9], the branches are two line segments parallel to the imaginary axis that are, moreover, reasonably well separated along the real line, i.e., a natural case of being “not very close to each

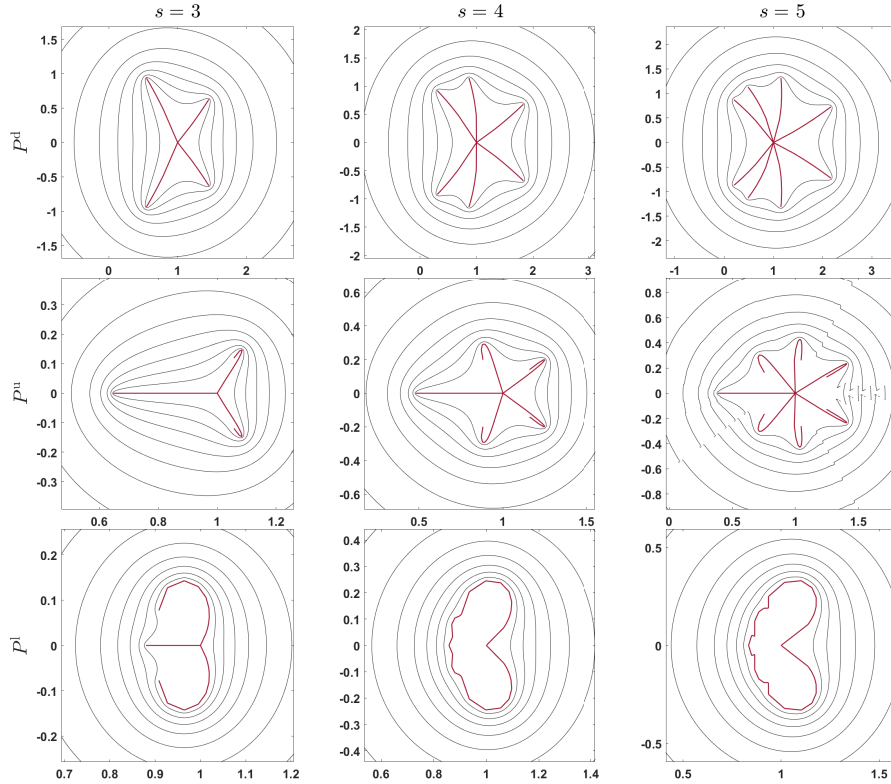


FIG. 5. In red: the curves Γ_q (first plots 1 to 7) and their “envelopes” (plots 8 and 9) for the Gauss Butcher tableau, taking $q = 15$. In black: the contours of the corresponding Schwarz-Christoffel map of the exterior of these curves (or envelopes) mapped to the exterior of the unit circle, see `extermap` in [3].

402 of the polynomial φ_ℓ^{MR} will decrease significantly more at the steps $\ell = s \cdot j$
 403 $j = 1, 2, \dots$ because only then each branch can get some attention. If the branches
 404 become close, then we do not expect this extra jump because keeping the absolute
 405 value of the polynomial small along one of the branches naturally translates into
 406 keeping the absolute value of the polynomial also small enough along another one.
 407 This is most pronounced in the first s iterations of GMRES, as we can see in Figure 6,
 408 where the convergence curves begin with a slower convergence phase – *precisely* s
 409 *steps* – for P^d and P^u , in contrast to the ones of P^l , where the arcs intersect and
 410 are, in general, closer to each other. We illustrate this further in Figure 7 for the
 411 preconditioner P^d for $s = 4, 8$ by looking at the polynomial φ_ℓ^{MR} and its roots (called
 412 harmonic Ritz values). We see that in the first row (4 branches, far apart) the
 413 possibility of “placing” one root along each of the branches was much more crucial
 414 (resulted in a more significant decrease of the modulus of the polynomial over the
 415 spectrum of the preconditioned system) than for the second row (8 branches with two
 416 complex conjugate pairs of branches that are close to each other). We note that an
 417 example of explanation (and prediction) of a *complete* staircase behavior of GMRES
 418 can be found in [4, Figure 9 and below].

other”.

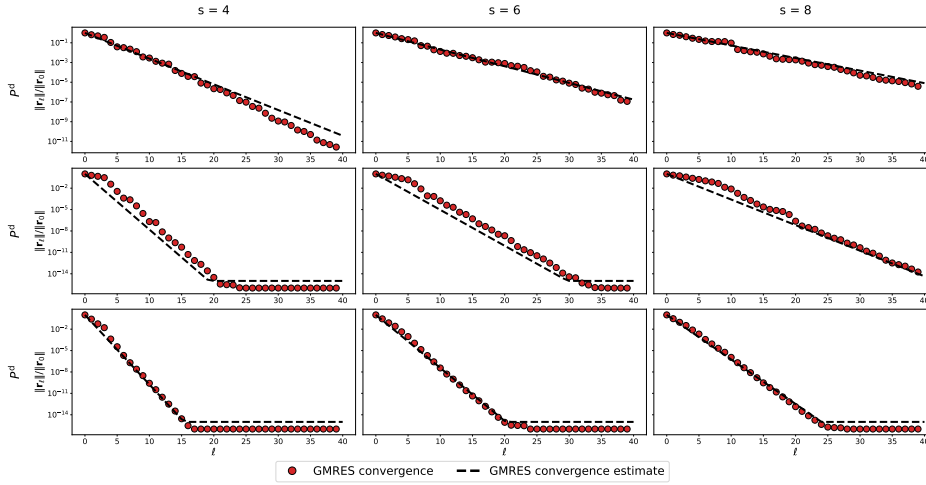


FIG. 6. The convergence behavior of preconditioned GMRES, using the Gauss Butcher tableau, together with the convergence estimate based on the calculated asymptotic convergence factor estimate ρ_{est} .

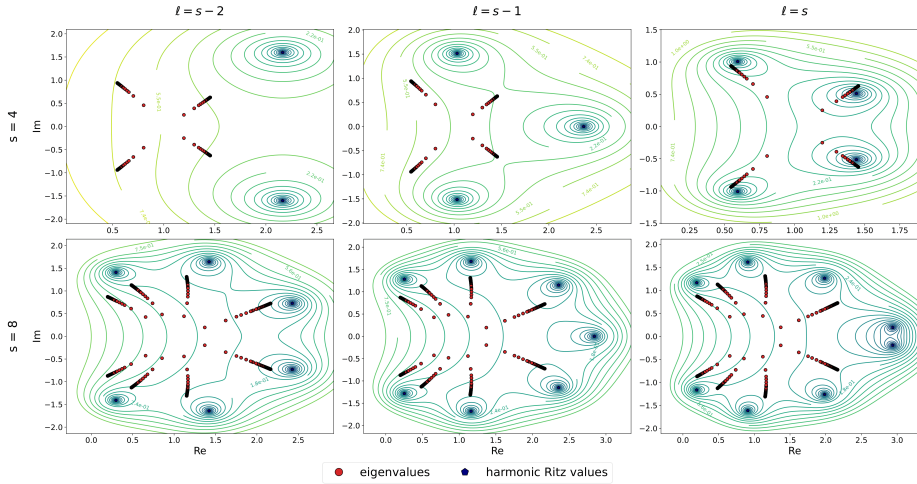


FIG. 7. The level curves of the GMRES polynomials φ_ℓ^{MR} for the preconditioned system $(P^\ell)^{-1}M$ together with the spectrum of this system as well as the roots of φ_ℓ^{MR} (so-called harmonic Ritz values). We set $N = 50$.

419 We also want to comment on a similarity with the results in [14, 15]. There, the
 420 authors addressed the question of *delay of convergence* by using similar formulations
 421 to ours, also obtaining a GMRES problem reformulated as for a block-diagonal ma-
 422 trix using Kronecker-product-like techniques as in Lemma 3.1. In particular, in [15,
 423 Section 3.1] the authors use the equality

$$424 \quad \|\mathbf{r}_\ell\| = \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \left\| \varphi \left(\begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_n \end{bmatrix} \right) \mathbf{r}_0 \right\| = \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \sqrt{\sum_{j=1}^n \|\varphi(X_j) \mathbf{s}_0^{(j)}\|^2},$$

425 where $\mathbf{s}_0^{(i)}$ is the i -th subvector of length s of $Q^T \Pi \mathbf{r}_0$, to obtain a lower bound

$$426 \quad (3.14) \quad \|\mathbf{r}_\ell\|^2 = \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \sum_{j=1}^n \left\| \varphi(X_j) \mathbf{s}_0^{(j)} \right\|^2 \geq \sum_{j=1}^n \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \leq \ell}} \left\| \varphi(X_j) \mathbf{s}_0^{(j)} \right\|^2$$

427 on the GMRES convergence behavior, explaining the initial stagnation phase in an
 428 advection-diffusion problem. This way they bound the *global* minimization problem
 429 (corresponding to solving a problem with the block-diagonal matrix $\text{diag}(X_1, \dots, X_n)$)
 430 by the sum of the *local* minimization problems (each given by the small s -by- s matrix
 431 X_j). By careful analysis of the interplay of the right-hand side (or initial residual)
 432 and the diagonal blocks in [15, Section 3.1] (there the diagonal blocks are, moreover,
 433 tridiagonal and Toeplitz), the authors conclude

434 *“... the presence of at least one system with tridiagonal Toeplitz matrix $T_j = \text{tridiag}(\gamma_j, \lambda_j, \mu_j)$ that is ‘close to the Jordan block’ (cf. [15, Section 3.3] but see also [14]), and with l representing the index of the first significant entry of the corresponding right-hand side, prevents fast convergence of GMRES for the first $N - l$ steps (N being the size of the blocks T_j) ...*
 435 *... As explained in Section 3.1, the lower bound is useless for analyzing the convergence behavior after the step $N - l$, possibly even earlier. Hence the above approach cannot be used for quantifying any possible acceleration of convergence after the initial phase.”*
 436

437 We see that the approach is *fundamentally* different – both in the intended direction
 438 as well as in the results it can deliver – in spite of the fact that it works with the same
 439 technique.

440 We finalize this section with a remark on the *field of values* (sometimes also
 441 called the numerical range) and *pseudospectra*, which sometimes are *extremely* useful
 442 to understand and predict GMRES convergence behavior, especially if the eigenbasis
 443 of the system matrix is ill-conditioned, see, e.g., [6] and also [16, Section 5.7.3, pp.
 444 296] and the references therein.

445 *Remark 3.7.* Another commonly used bound for GMRES uses the *field of values*
 446 $\nu(C)$ or the δ -*pseudospectrum* $\sigma_\delta(C)$ of the system matrix C . By a direct calculation
 447 we obtain, for our model problem, the field of values as
 448

$$449 \quad \nu(MP^{-1}) = \sum_{i=1}^n \nu(X_k) \quad (\text{and analogously for } \nu(P^{-1}M)),$$

450 where the X_k are given as in (3.7) and the set addition is understood element-wise,
 451 i.e., $\nu(X_1) + \nu(X_2) = \{\alpha_1 + \alpha_2 \mid \alpha_1 \in \nu(X_1), \alpha_2 \in \nu(X_2)\}$, or, more generally

$$452 \quad \nu(MP^{-1}) \subset \kappa(Q) \sum_{i=1}^n \nu(X_k) \quad (\text{and analogously for } \nu(P^{-1}M)).$$

453 For the pseudospectrum we obtain an analogous formula, namely

$$454 \quad \sigma_\delta(MP^{-1}) \subset \kappa(Q) \sum_{i=1}^n \sigma_\delta(X_k) \quad (\text{and analogously for } \nu(P^{-1}M)).$$

455 In other words, the principle of working with the small matrices X_k instead of the large
 456 matrix MP^{-1} naturally applies also to the other standard techniques for analyzing
 457 GMRES convergence behavior. However, adapting and using bounds based on field
 458 of values or the pseudospectrum of the preconditioned system for this set-up remains
 459 a topic for future research.

460 **4. Numerical Examples.** In this section we use the above analysis for more
 461 involved settings and, more importantly, also demonstrate the convergence estimates
 462 (instead of only the convergence rate estimates). We use a FEM discretization in
 463 space⁸ for different geometries in Example 1 and 2, see Figure 8. We also fix the
 464 number of time steps to balance the spatial and time error (see the (L2) definition in
 465 Section 2), namely we fix

$$466 \quad \tau = h^{\frac{2}{p}},$$

467 where the 2 in the numerator is the order of the spatial error (since we use linear
 468 Lagrange polynomials in the FEM discretization) and p is the order of convergence
 469 of the Runge-Kutta method. We show for both examples the GMRES convergence
 470 together with the *convergence estimates*, namely

$$471 \quad \frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \approx \min \{ \kappa_S^{\text{est}} \rho_{\text{est}}^\ell, 1 \},$$

472 where the estimate κ_S^{est} of κ_S is computed from the eigenbasis condition numbers of
 473 the “fake sampled” matrices X_{ϑ_k} for $k = 1, \dots, q$. In our experience, the best results
 474 are obtained with $q \approx 15 - 20$, as increasing q further leads to crowding problems
 475 in the SC toolbox and eventually to problems with the convergence of the Schwarz-
 476 Christoffel map. We also found that spacing the fake points ϑ_k *logarithmically* in
 477 the corresponding interval somewhat alleviates this issue and leads to more accurate
 478 predictions of the arcs of the given algebraic curve. We also recall that the seeming
 479 independence of the preconditioner quality on the spatial mesh size h was sufficiently
 480 documented elsewhere (see [21, 18, 2, 9, 19]) and explained in Section 3 so that in our
 481 eyes, there is no need to address this direction here. Illustration of the solutions as
 482 well as further numerical experiments can be found in [19, Chapter 7].

483 Last but not least, we have not set a relative residual tolerance criterion for
 484 stopping GMRES, meaning that GMRES went on until either the relative residual
 485 was on the level of machine precision or the maximum number of iterations was
 486 reached. This is not a good choice from the point of view of the solution process
 487 efficiency but since our primary focus is on studying the preconditioners, we found
 488 this reasonable.

489 *Example 1: Cookies in the oven .* The first problem is a simulation of baking
 490 cookies in an electrical oven projected in 2D, an idea borrowed from [13]. The cookies
 491 have a worse heat conductivity than the surrounding air (piecewise constant in space
 492 and constant in time) and the setting demands various boundary conditions, resulting

⁸Wherever we talk about a FEM discretization, we use linear Lagrange polynomials on conforming triangular meshes. Those are refined by the standard quadrisection of a triangle, with additional post-smoothing of the mesh.

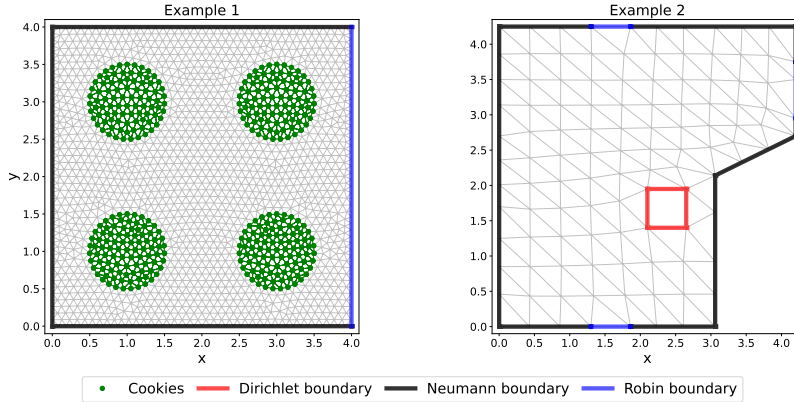


FIG. 8. The initial triangulations for Example 2 and 3 together with the boundary condition types and, for Example 2, also with highlighting the points with lower heat conductivity.

493 in

$$\begin{aligned}
 & \frac{\partial u}{\partial t} u = \operatorname{div}(\sigma \nabla u) + f \quad \text{in } \Omega \times (0, T], \\
 494 \quad & \frac{\partial u}{\partial \mathbf{n}} u = 0 \quad \text{on } \Gamma_N \times (0, T], \quad \frac{\partial u}{\partial \mathbf{n}} u + pu = 0 \quad \text{on } \Gamma_R \times (0, T], \\
 & u = 0 \quad \text{at } \Omega \times \{0\},
 \end{aligned}$$

495 with $\Omega = (0, 4) \times (0, 4)$ and the boundary of Ω is split into the Neumann and Robin
 496 parts Γ_N, Γ_R . We set the data as

$$\begin{aligned}
 & \Gamma_N = \{x = 0\} \cup \{y = 0\} \cup \{y = 4\}, \quad \Gamma_R = \{x = 4\}, \quad p = 1, \sigma = \begin{cases} 10^3 & \text{if } (x, y) \in \text{Cookie}, \\ 1 & \text{otherwise,} \end{cases} \\
 497 \quad & f(x, y, t) = \begin{cases} 3 & \text{if } \|(x, y) - (2, 2)\| \leq 1, \\ 0 & \text{otherwise,} \end{cases}
 \end{aligned}$$

498 and show the GMRES convergence behavior with the estimates in Figure 9 as well as
 499 the sampling of the algebraic curves in Figure 10.

500 *Example 2: The cabin heating*. The second problem uses the 2D projection of
 501 an attic room of a cabin in the western Bohemia region, whose primary heating is the
 502 chimney (bottom-right corner, modeled with a Dirichlet boundary condition changing
 503 in time), with two windows (top and bottom) and a door (right), modeled with
 504 Robin boundary conditions with Robin parameters p_w and p_d , and a good insulation
 505 otherwise, modeled with a Neumann condition. We obtain the problem

$$\begin{aligned}
 & \frac{\partial u}{\partial t} u = \operatorname{div}(\sigma \nabla u) \quad \text{in } \Omega \times (0, T], \\
 506 \quad & \frac{\partial u}{\partial \mathbf{n}} u = 0 \quad \text{on } \Gamma_N \times (0, T], \quad \frac{\partial u}{\partial \mathbf{n}} u + pu = 0 \quad \text{on } \Gamma_R \times (0, T], \\
 & u = 0 \quad \text{at } \Omega \times \{0\},
 \end{aligned}$$

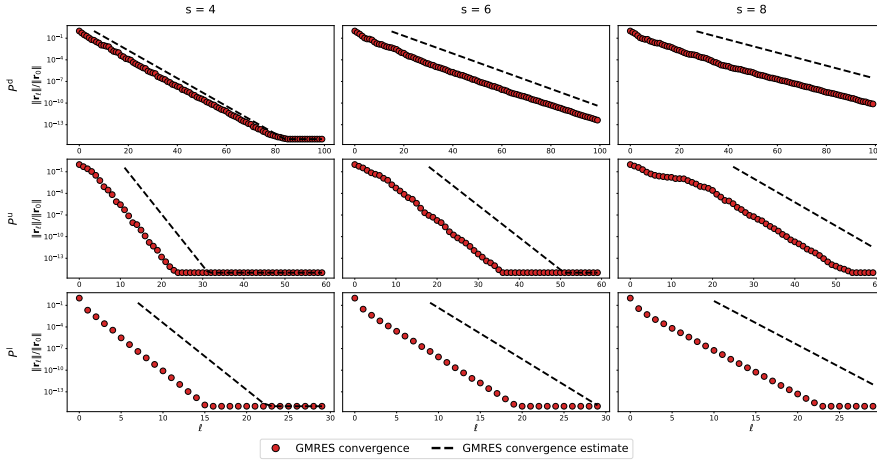


FIG. 9. The GMRES convergence behavior with the convergence estimates based on ρ_{est} for Example 1 with $n = 26985$.

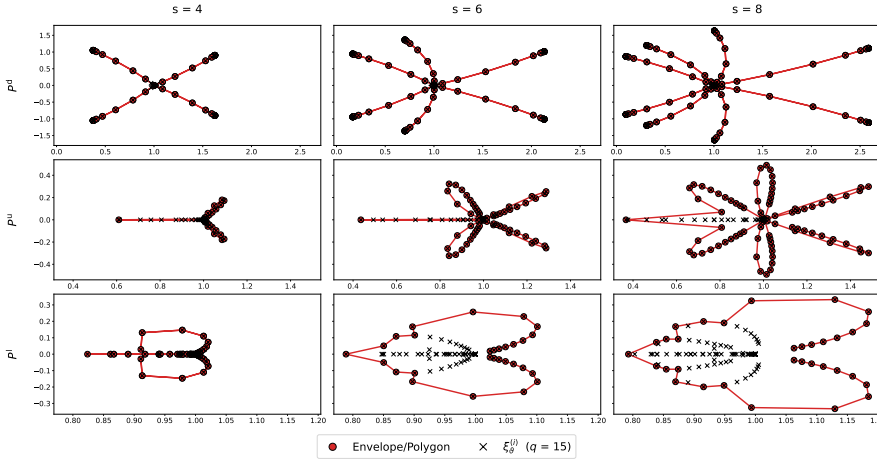


FIG. 10. The algebraic curve polygon approximations that are used in the Schwarz-Christoffel MATLAB toolbox to calculate ρ_{est} for Example 1 – for some settings these correspond to the eigenvalues $\xi_g^{(i)}$ and in some these only enclose these.

507 and take the data as

$$508 \quad p_w = 0.1, \quad p_d = 10, \quad g_D(x, y, t) = \begin{cases} \min\{t, 0.7\} & \text{if } (x, y) \in \Gamma_D, \\ 0 & \text{otherwise,} \end{cases}$$

509 and show the GMRES convergence behavior with the estimates in Figure 11 as well
510 as the sampling of the algebraic curves in Figure 12.

511 *Summary.* Overall, we observe that the convergence factor estimates delivered
512 very accurate results even for these more involved problems but the conditioning of
513 the eigenbasis of the matrices X_{ϑ_k} notably deteriorated as we increased s . The fact
514 that this does not show up in the GMRES convergence behavior suggests that more
515 delicate bounds, such as mentioned in Remark 3.7 could give a more detailed insight

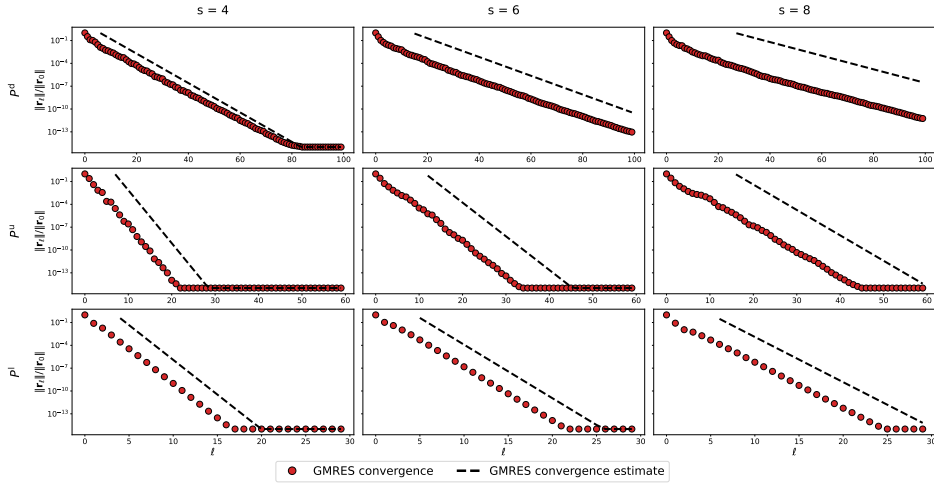


FIG. 11. The GMRES convergence behavior with the convergence estimates based on ρ_{est} for Example 2 with $n = 26985$.

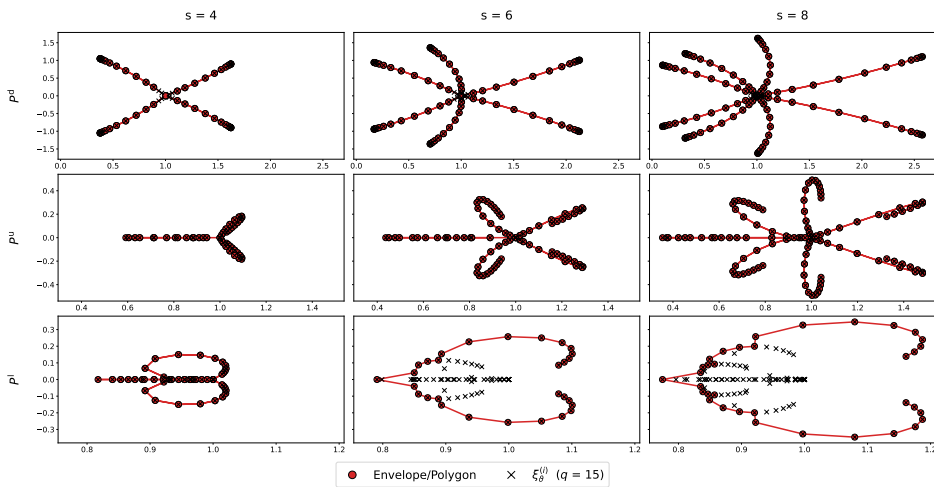


FIG. 12. The algebraic curve polygon approximations that are used in the Schwarz-Christoffel MATLAB toolbox to calculate ρ_{est} for Example 2 – for some settings these correspond to the eigenvalues $\xi_q^{(i)}$ and in some these only enclose these.

516 into the matter. However, in all cases the bounds lag behind the actual convergence
 517 behavior by 10-20 iterations, which is often still considered to be reasonably accurate.
 518 We also showed the polygons used in the Schwarz-Christoffel toolbox – notice that
 519 in many of the plots we excluded part of the arcs, mainly because either (a) the
 520 arcs intersected and we took the envelope of the algebraic curve (usually for the
 521 preconditioner P^l) or (b) the points sampled along the arcs crowded sections of the
 522 arcs, which caused issues for the toolbox. In such cases we sparsified these regions
 523 by dropping some of these points. As a result, the Schwarz-Christoffel external map
 524 converged better and faster than for the problem in Section 3.1 and the contours were
 525 “ripple-free” for all of our problems, otherwise looking almost precisely as the ones in

526 Figure 5.

527 **5. Concluding remarks.** Our main goal has been to understand the block
 528 preconditioners considered in [21, 2, 18] in more detail and to try to explain their
 529 success and/or limitations. This goal was, in our eyes, mostly achieved but could
 530 be further improved in the sense of Remark 3.7 or by considering a more refined
 531 version of the bound (2.10), see [6, Section 2.1, equations (2.1) and (EV')] – this
 532 remains an area of interest for us for the future. Moreover, the above analysis can
 533 be directly used to try to *optimize* the Runge-Kutta methods, following the ideas
 534 in [21, 19, 9]. We also note that in practice, solving with either of the matrices
 535 $P^{d,u,l,GSU,GSU,\dots}$ is often done with some level of *inaccuracy*, e.g., using a multigrid
 536 method. The question of interaction of this inaccuracy with the overall GMRES
 537 convergence is an important one and to the best of our knowledge has been addressed
 538 only numerically in [19, Chapter 7]. We also note that adapting the above analysis
 539 to the framework presented in [24, 23], or reformulating it from the vector equation
 540 to the matrix equation as suggested in [20], and to study in detail the comparison of
 541 these approaches for the IRK setting are attractive directions for future research.

542 **Acknowledgements.** Some of the ideas were stimulated by conversations with
 543 Mark Embree, Patrick Farrell, Miroslav Tůma and Petr Tichý and we would like to
 544 thank them for their inspiring comments and suggestions.

545

REFERENCES

- 546 [1] M. ARIOLI, V. PTÁK, AND Z. STRAKOŠ, *Krylov sequences of maximal length and convergence*
 547 *of GMRES*, BIT, 38 (1998), pp. 636–643.
- 548 [2] M. R. CLINES, V. E. HOWLE, AND K. R. LONG, *Efficient order-optimal preconditioners for*
 549 *implicit Runge-Kutta and Runge-Kutta-Nyström methods applicable to a large class of*
 550 *parabolic and hyperbolic PDEs*. arXiv: <https://arxiv.org/abs/2206.08991>, 2022, <https://doi.org/10.48550/ARXIV.2206.08991>.
- 551 [3] T. A. DRISCOLL, *A MATLAB toolbox for Schwarz-Christoffel mapping*, Tech. Report 2, 1996.
- 552 [4] T. A. DRISCOLL, K.-C. TOH, AND L. N. TREFETHEN, *From potential theory to matrix iterations*
 553 *in six steps*, SIAM Rev., 40 (1998), pp. 547–578.
- 554 [5] T. A. DRISCOLL AND L. N. TREFETHEN, *Schwarz-Christoffel mapping*, Cambridge University
 555 Ppress, Cambridge, First ed., 2002.
- 556 [6] M. EMBREE, *How descriptive are GMRES convergence bounds?*, 2023, <https://arxiv.org/pdf/2209.01231.pdf>. arXiv preprint: 2209.01231.
- 557 [7] M. EMBREE AND L. N. TREFETHEN, *Green's functions for multiply connected domains via*
 558 *conformal mapping*, SIAM Rev., 41 (1999), pp. 745–761.
- 559 [8] V. FABER, J. LIESEN, AND P. TICHÝ, *On Chebyshev polynomials of matrices*, SIAM J. on
 560 Matrix Anal Appl., 31 (2010), pp. 2205–2221.
- 561 [9] M. J. GANDER AND M. OUSRATA, *Spectral analysis of implicit 2-stage block Runge-Kutta pre-*
 562 *conditioners*, Linear Algebra Appl., (2023), <https://doi.org/10.1016/j.laa.2023.07.008>.
- 563 [10] A. GREENBAUM, V. PTÁK, AND Z. STRAKOŠ, *Any nonincreasing convergence curve is possible*
 564 *for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
- 565 [11] A. GREENBAUM, Z. STRAKOŠ, M. J. GANDER, AND M. OUSRATA, *Matrices that generate the*
 566 *same Krylov residual spaces*, in Recent Advances in Iterative Methods, G. H. Golub,
 567 A. Greenbaum, and M. Luskin, eds., vol. 60 of IMA Volumes in Mathematics and its
 568 Applications, Springer, 1994, pp. 95–118.
- 569 [12] T. KATO, *Perturbation Theory for Linear Operators*, vol. 132, Springer Berlin, Heidelberg,
 570 2013.
- 571 [13] D. KRESSNER AND C. TOBLER, *Low-rank tensor Krylov subspace methods for parametrized*
 572 *linear systems*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1288–1316.
- 573 [14] J. LIESEN AND Z. STRAKOŠ, *Convergence of GMRES for tridiagonal Toeplitz matrices*, SIAM
 574 J. on Matrix Anal. Appl., 26 (2004), pp. 233–251.
- 575 [15] J. LIESEN AND Z. STRAKOŠ, *GMRES convergence analysis for a convection-diffusion model*
 576 *problem*, SIAM J. Sci. Comput., 26 (2005), pp. 1989–2009.

- 579 [16] J. LIESEN AND Z. STRAKOŠ, *Krylov Subspace Methods: Principles and Analysis*, Oxford Uni-
580 versity Press, Oxford, 2013.
- 581 [17] P. MUNCH, I. DRAVINS, M. KRONBICHLER, AND M. NEYTCHEVA, *Stage-parallel fully implicit*
582 *Runge–Kutta implementations with optimal multilevel preconditioners at the scaling limit*,
583 *SIAM J. Sci. Comput.*, (2023), pp. S71–S96.
- 584 [18] M. NEYTCHEVA AND O. AXELSSON, *Numerical Solution Methods for Implicit Runge-Kutta*
585 *Methods of Arbitrarily High Order*, in Proceedings of the Conference Algoritmy 2020,
586 P. Frolkovič, K. Mikula, and D. Ševčovič, eds., Slovak University of Technology in
587 Bratislava, Vydavateľstvo SPEKTRUM, 2020.
- 588 [19] M. OUTRATA, *Schwarz methods, Schur complements, preconditioning and numerical linear*
589 *algebra*, PhD thesis, University of Geneva, Math Department, 2022.
- 590 [20] D. PALITTA AND V. SIMONCINI, *Optimality properties of Galerkin and Petrov–Galerkin methods*
591 *for linear matrix equations*, *Vietnam J. Math.*, 48 (2020), pp. 791–807.
- 592 [21] M. M. RANA, V. E. HOWLE, K. LONG, A. MEEK, AND W. MILESTONE, *A New Block Pre-*
593 *conditioner for Implicit Runge-Kutta Methods for Parabolic PDE Problems*, *SIAM J. Sci.*
594 *Comput.*, 43 (2021), pp. S475–S495.
- 595 [22] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, Other Titles in Applied Mathematics,
596 SIAM, Philadelphia, Second ed., 2003.
- 597 [23] B. S. SOUTHWORTH, O. KRZYSIK, AND W. PAZNER, *Fast solution of fully implicit Runge–Kutta*
598 *and discontinuous Galerkin in time for numerical PDEs, Part II: nonlinearities and DAEs*,
599 *SIAM J. Sci. Comput.*, 44 (2022), pp. 636–663.
- 600 [24] B. S. SOUTHWORTH, O. KRZYSIK, W. PAZNER, AND H. DE STERCK, *Fast solution of fully*
601 *implicit Runge–Kutta and discontinuous Galerkin in time for numerical PDEs, Part I:*
602 *The linear setting*, *SIAM J. Sci. Comput.*, 44 (2022), pp. 416–443.
- 603 [25] G. A. STAFF, K.-A. MARDAL, AND T. K. NILSSEN, *Preconditioning of fully implicit Runge-*
604 *Kutta schemes for parabolic PDEs*, *Modeling, Identification and Control*, 27 (2006),
605 pp. 109–123.
- 606 [26] C. F. VAN LOAN, *The ubiquitous Kronecker product*, *J. Comput. Appl. Math.*, 123 (2000),
607 pp. 85–100.
- 608 [27] G. WANNER AND E. HAIRER, *Solving Ordinary Differential Equations II : Stiff and Differential-*
609 *Algebraic Problems*, Springer Berlin, Heidelberg, 1996.
- 610 [28] G. WANNER, S. P. NØRSETT, AND E. HAIRER, *Solving Ordinary Differential Equations I :*
611 *Non-Stiff Problems*, Springer Berlin, Heidelberg, 1987.