Spectral analysis of implicit s stage block Runge-Kutta preconditioners

Martin J. Gander[*], Michal Outrata[†]

Abstract

We analyze the recently introduced family of preconditioners in [20] for the stage equations of implicit Runge-Kutta methods for two stage methods. We simplify the formulas for the eigenvalues and eigenvectors of the preconditioned systems for a general method and use these to obtain convergence rate estimates for the 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 and we illustrate our analysis with numerical experiments.

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 [27, 26] for an overview and further references). For implicit methods (IRK), their efficiency copies the efficiency of a solver for the so-called *stage equations* – in general a system of ms non-linear equations, where m 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 m. If the spatial operator is *linear*, then the stage equations also form a system of linear algebraic equations and are often solved by an iterative solver, e.g., a Krylov method. In [20], 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 m grows. Recently, there has been also other contributions in the direction of preconditioning the *fully implicit* Runge-Kutta stage equations for PDEs, see [23, 22] but also [17] and [2], introducing new ideas and testing these numerically on a variety of test problems.

We focus on the setting considered in [20] and expand on the 2-stage method analysis given in [9] and consider the general s-stage case, giving a theoretical background for the performance and spectral properties observed. First, we recall some important preliminaries in Section 2 so that we can deliver the analysis, based on the spectral analysis of the preconditioned system, in Section 3. We support the analysis by considering more involved examples in Section 4.

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

[†]Section de Mathématiques, Université de Genève; this work was partially supported by the FCS Swiss Excellence PhD Fellowship program of the Swiss Federation.

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.,

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

where Δ is the Laplace operator, f, g, u_0 are given functions and Ω is the unit square $\Omega := (0,1) \times (0,1)$. As in [9] 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 $i = N+1, \dots, N(N-1)-1$, where $b_i^{(\mathrm{ST})}(t)$ 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

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

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 $n := (N-1)^2$, 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.,

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

Having a *Butcher tableau*

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

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

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

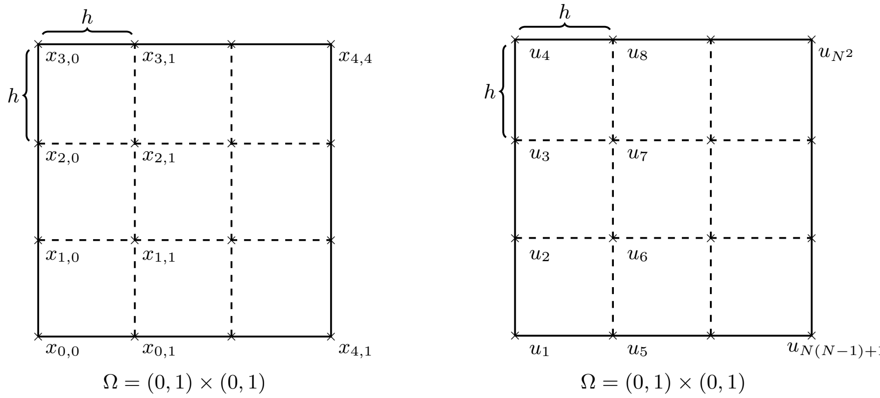

Figure 1: Left: grid points for $N+1=4$; right: lexicographical ordering of the unknowns for $N+1=4$.

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 [25] and references therein) and we note that (7) can be reformulated into a *matrix equation*, which is in general better suited for using a Krylov solver (see [19]). Here we focus on the analysis of the results in [20] and thus we do not address this any further but a study of the preconditioners from [20] 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 system matrix M sparse 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 [20], the authors introduce the block preconditioners

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

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.,

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

Some of these – P^{d} and P^{GSL} – were considered already in [24]. 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 [23, Lemma 1].

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., [16, Section 5.7.2]. We highlight some aspects of the bound (10) 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 [10], [11] and [1] and also [16, 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 evaluate 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 [16, 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., [15]. However, studying GMRES behavior with the bound* (10)*, this interaction is completely lost.*

In cases where (10) is justifiable, the next step is usually to bound from above the mixed[1] min-max problem in the right-hand side of (10) by replacing the discrete set over which we take the maximum, let us denote it by σ^{discr}, by a non-discrete one, which we denote by $\sigma^{\mathrm{non-discr}}$, so that we have $\sigma^{\mathrm{discr}} \subset \sigma^{\mathrm{non-discr}}$. We highlight two important aspects of this step:

(a) It is *functional* only if we can further bound or evaluate the solution of the min-max problem over $\sigma^{\mathrm{non-discr}}$ *and* obtain a reasonably fast convergence estimate.

(b) It is *appropriate* only if[2] $\partial_{\mathbb{C}}\sigma^{\mathrm{non-discr}}$ is reasonably uniformly covered by σ^{discr}.[3] In case of clusters, we should consider having $\sigma^{\mathrm{non-discr}}$ as a union of separate non-discrete sets $\sigma_i^{\mathrm{non-discr}}$ each of which captures one of the clusters, i.e., is covered by one of the clusters reasonably uniformly.

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

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

[3]Intuitively, we could expect that the bound will be appropriate only if σ^{discr} covers the entirety of $\sigma^{\mathrm{non-discr}}$ but because polynomials of complex variables are harmonic we can conclude that the maximum of the modulus of a polynomial over the set $\sigma^{\mathrm{non-discr}}$ is attained somewhere along $\partial_{\mathbb{C}}\sigma^{\mathrm{non-discr}}$ and therefore for the GMRES bound is key only the relation of $\partial_{\mathbb{C}}\sigma^{\mathrm{non-discr}}$ and σ^{discr}, see [5, Section 2].

For example, in (10) we can replace the spectrum $\sigma^{\text{discr}} = \{\lambda_1, \dots, \lambda_d\}$ by a disc containing all of the eigenvalues $\sigma^{\text{non-discr}} = \{z \in \mathbb{C} \,|\, |z - c| \leq \rho\}$. Assuming $|c| > \rho$, a crude but sometimes useful approximation of the original bound is available,

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \kappa(S) \left(\frac{\rho}{|c|}\right)^k, \tag{11}$$

see [21, Section 6.11.2, Corollary 6.33 and Lemma 6.26 and below]. Here, $\sigma^{\text{non-discr}} = \{z \in \mathbb{C} \,|\, |z - c| \leq \rho\}$ was clearly chosen with the *functionality* aspect in mind as we know the polynomial that realizes the bound (see [21, Lemma 6.26]) and it gives a good convergence bound as long as and $\rho \not\approx |c|$. However, it is usually far from being *appropriate* as the eigenvalues usually don't spread uniformly over the circle bounding the disc. One notable exception is the case of tightly clustered eigenvalues around a single point c – in this case the clustering usually makes this bound appropriate as we can choose ρ *very* small. We emphasize that the adjectives *functional* and *appropriate* make sense only if the original bound (10) was itself descriptive of the GMRES convergence bound, i.e., only if the system matrix is either close to normal or the initial residual is restricted to the a subspace on which the system matrix is not too far from being normal.

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,

$$V := [\mathbf{v}_1, \dots, \mathbf{v}_n], \quad \text{and} \quad Q := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix} \in \mathbb{R}^{sn \times sn}. \tag{12}$$

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

$$\tilde{M} = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \end{bmatrix} - \frac{\tau}{h^2} \begin{bmatrix} a_{1,1}\Lambda & \dots & a_{1,s}\Lambda \\ \vdots & \ddots & \vdots \\ a_{s,1}\Lambda & \dots & a_{s,s}\Lambda \end{bmatrix}, \tag{13}$$

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

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

where $\tilde{P} := Q^T PQ$ stands for one of the right-preconditioners $P^{\text{d,GSU,u}}$ and an analogous formulation follows also for the left-preconditioners $P^{\text{l,GSL}}$. 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). 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 (see [9, 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.,*

$$C = \begin{bmatrix} \Lambda_{11} & \dots & \Lambda_{1s} \\ \vdots & \ddots & \vdots \\ \Lambda_{s1} & \dots & \Lambda_{ss} \end{bmatrix}, \quad \text{with} \quad \Lambda_{ij} = \operatorname{diag}\left(\lambda_1^{(ij)}, \dots, \lambda_n^{(ij)}\right) \quad \forall ij. \tag{14}$$

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

$$\Pi^T C \Pi = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_n \end{bmatrix} \quad \text{with} \quad C_\ell = \begin{bmatrix} \lambda_\ell^{(11)} & \dots & \lambda_\ell^{(1s)} \\ \vdots & \ddots & \vdots \\ \lambda_\ell^{(s1)} & \dots & \lambda_\ell^{(ss)} \end{bmatrix} \in \mathbb{R}^{s\times s}, \tag{15}$$

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

Hence, C is diagonalizable if and only if C_ℓ is 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

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

and 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 $C = V^{-1}DV$, then

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

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

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 bi-diagonal (or block upper-triangular) matrix.*

To shorten the notation we set

$$\theta_k := \frac{\tau}{h^2}\lambda_k \quad \text{and} \quad \Theta := \frac{\tau}{h^2}\Lambda, \tag{16}$$

as these quantities appear always together in the computations. By a direct calculation (see [18, Appendix B.8]) we get the limit behavior of θ_k as $\tau, h \to 0$,

$$\underbrace{\begin{aligned} &(\theta_n, \theta_1) \to (-\frac{8}{C_\mathrm{e}}, 0), \\ &(\theta_1^{-1}, \theta_n^{-1}) \to \left(-\infty, -\frac{C_\mathrm{e}}{8}\right), \end{aligned}}_{(\mathrm{LIM})_{p=1}} \qquad \underbrace{\begin{aligned} &(\theta_n, \theta_1) \to (-\infty, 0), \\ &(\theta_1^{-1}, \theta_n^{-1}) \to (-\infty, 0). \end{aligned}}_{(\mathrm{LIM})_{p>1}} \tag{17}$$

Next we define the following s-by-s matrices

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

where α_{ij} are the entries of the replacement for A in M, e.g., taking $\star = \mathrm{d}$ we have $\alpha_{ij} = a_{ij}$ for $i = j$ and $\alpha_{ij} = 0$ otherwise while taking $\star = \mathrm{u}$ we have $\alpha_{ij} = (D_A U_A)_{ij}$ where $A = L_A D_A U_A$ id the LDU factorization of A and so on. Using Lemma 1, we obtain the following result.

Proposition 1. *Take M as in (7) and one of the preconditioners from (8 9) as P. Assuming P is invertible, the spectrum of MP^{-1} (or $P^{-1}M$) is given as the union of the spectra of the matrices X_k given by*

$$X_k^\star := M_k \left(P_k^\star\right)^{-1} \quad (\text{or } P_k^{-1} M_k), \tag{18}$$

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

$$(S_k^\star)^{-1} X_k^\star S_k^\star = \mathrm{diag}(\xi_1^{(k)}, \dots, \xi_s^{(k)}), \tag{19}$$

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

$$\kappa\left(W\right) \cdot \max_{k=1,\dots,n} \kappa\left(S_k^\star\right).$$

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

Proof. Transforming MP^{-1} (or $P^{-1}M$) into the basis given by Q we use Lemma 1 for the matrix $\tilde{M}\tilde{P}^{-1}$ (see (13)) and obtain the result. □

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

Corollary 1 ([18, Proposition 7.5])**.** *Under the assumptions of Proposition 1 we have for the right-preconditioner P^{d} the formula*

$$X_k^{\mathrm{d}} = \begin{bmatrix} 1 & -\frac{a_{12}\theta_k}{1-a_{22}\theta_k} & \dots & -\frac{a_{1s}\theta_k}{1-a_{ss}\theta_k} \\ -\frac{a_{21}\theta_k}{1-a_{11}\theta_k} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{a_{1s}\theta_k}{1-a_{11}\theta_k} & \dots & \dots & 1 \end{bmatrix}, \tag{20}$$

with the characteristic polynomial

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

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

$$\tilde{p}_k^{(s)}(t) = t^s + \beta_{s-2}t^{s-2} + \beta_{s-3}t^{s-3} + \ldots + \beta_1 t + \beta_0.$$

Corollary 2 ([18, Proposition 7.6]). *Under the assumptions of Proposition 1 the block upper-triangular preconditioners $P^{\mathrm{GSU,u}}$ give*

$$X_k^{\mathrm{GSU,u}} = \begin{bmatrix} 1 & 0 \; \dots \; \dots \; \dots \; 0 \\ \star & \\ \vdots & \left(M_k (P_k^{\mathrm{GSU,u}})^{-1}\right)_{2:s,2:s} \\ \star & \end{bmatrix}, \quad X_k^{\mathrm{GSL,l}} = \begin{bmatrix} 1 & \star \; \dots \; \dots \; \dots \; \star \\ 0 & \\ \vdots & \left((P_k^{\mathrm{GSL,l}})^{-1} M_k\right)_{2:s,2:s} \\ 0 & \end{bmatrix}, \tag{21}$$

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

These results suggest 1 as a natural "central point" of the spectrum of the preconditioned system, generalizing the observations made for $s = 2$. We note that using these results we get both quantitative and qualitative insight into the spectra shown in [20, Figure 4.1 – 4.4], e.g., we see that for $s = 3$ the eigeninformation of $M(P^{\mathrm{u}})^{-1}$ and $(P^{\mathrm{l}})^{-1}M$ can be still obtained explicitly (see also [18, Section 7.4]) and on the other hand for $s \geq 6$ there is no hope for these in general – but any bound on the eigeninformation of L can be used to obtain a bound on the eigeninformation of the preconditioned system by calculating with X_k, see [9, Section 4].

We show the spectra of the preconditioned systems and the corresponding GMRES convergence behavior in Figure 2 and 3, demonstrating observations and results from above. Notably, the bounds leave something to be desired, especially for P^{d} where they are not descriptive at all. Moreover, increasing s seems to noticeably affect the quality of the preconditioners – see also [20] for further numerical tests with various s and h. These numerical examples (as well as these in [2, 9]) are, as far as we can tell, representative of the general experience with these preconditioners. We highlight several key features illustrated in Figures 2 – 3 that remained true in all of our experiments:

1. For s small, we have observed the staircase-like convergence behavior visible in the left uppermost plot in Figure 3 and this was most pronounced for the preconditioner P^{d}.

2. We have usually not observed the desired *superlinear* convergence behavior, except for a speed-up after an initial stagnation (or slower speed convergence) phase.

3. In vast majority of cases, the number of GMRES iterations to reach a certain tolerance grows only very moderately under mesh refinement and for $P^{\mathrm{u}}, P^{\mathrm{l}}$ it remains almost constant.

4. In all of the experiments the spectra had the characteristic arc-like structure that we see in Figure 2.

We aim to explain all of these features here as well as investigate other bounds or estimates that would be more descriptive of the convergence behavior but also other possibilities for further improvement, such as *numerical optimization* in the spirit of [9, Section 4]. But before we do that, we note that the above results can be also transposed in a straight-forward fashion to the *transformed system* after we multiply (7) with $(A^{-1} \otimes I_n)$ from the left, obtaining

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

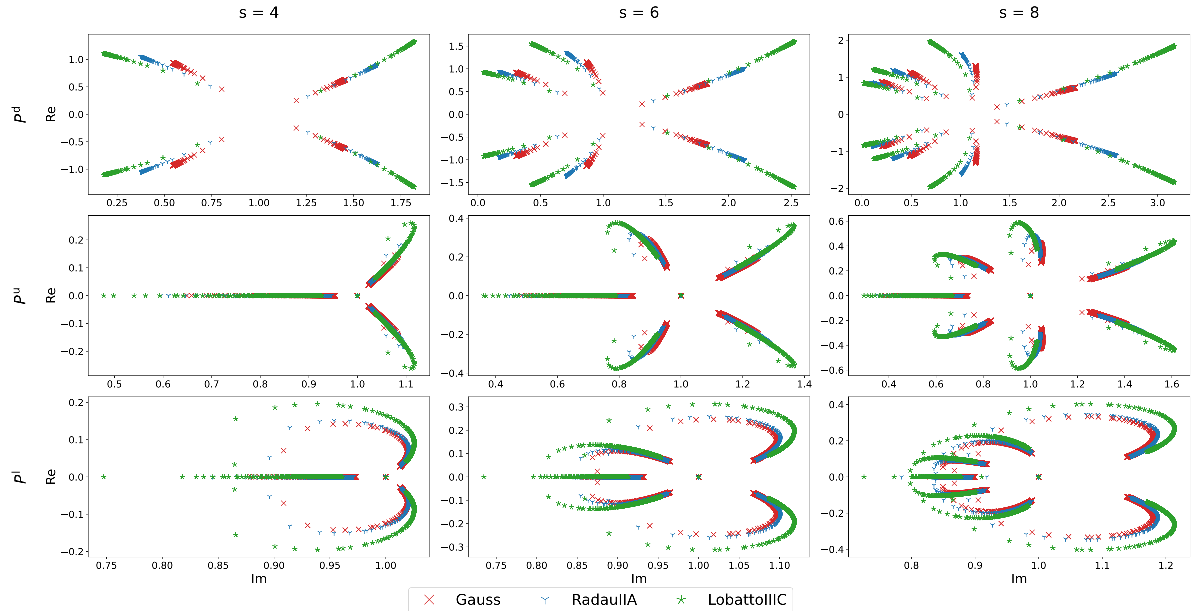

Figure 2

and getting analogously the preconditioners,

$$R^{\mathrm{d}} = \operatorname{diag}\left(A^{-1}\right) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$
$$R^{\mathrm{l}} = \left(D_{A^{-1}} U_{A^{-1}}\right) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L \quad \text{and} \quad R^{\mathrm{u}} = \left(L_{A^{-1}} D_{A^{-1}}\right) \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$
$$R^{\mathrm{GSL}} = \left(A^{-1}\right)_L \otimes I_n - \frac{\tau}{h^2} I_s \otimes L \quad \text{and} \quad R^{\mathrm{GSU}} = \left(A^{-1}\right)_U \otimes I_n - \frac{\tau}{h^2} I_s \otimes L,$$

where A^{-1} has the LDU factorization $A^{-1} = L_{A^{-1}} D_{A^{-1}} U_{A^{-1}}$ and $\left(A^{-1}\right)_{L,U}$ are defined analogously to (9). These preconditioners were proposed in [17] and then this transformation was further utilized in [23, 22]. For a *general* Butcher tableau, it is not possible to say whether the preconditioned transformed system gives a better performance than the original one. However, in [23, 22] the authors propose different preconditioners and its analysis within this framework is going to be considered elsewhere. Also, we note that the extension of the above analysis for FEM discretization is a straight-forward task – more details on both of these topics can be found in [18, Sections 7.6 and 7.7].

3.1 Spectral analysis

Next we turn to the the spectral analysis, keeping in mind its limitation in the sense of Remark 1. For block-diagonal problems we obtain

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\leq \ell}} \max_{j=1,\ldots,n} \left\|\varphi\left(X_j\right)\right\|, \tag{22}$$

which was studied in [8], where the authors showed that the extremal polynomials (i.e., the polynomial realizing the above bound) satisfies the equioscillation property but only every s iterations,

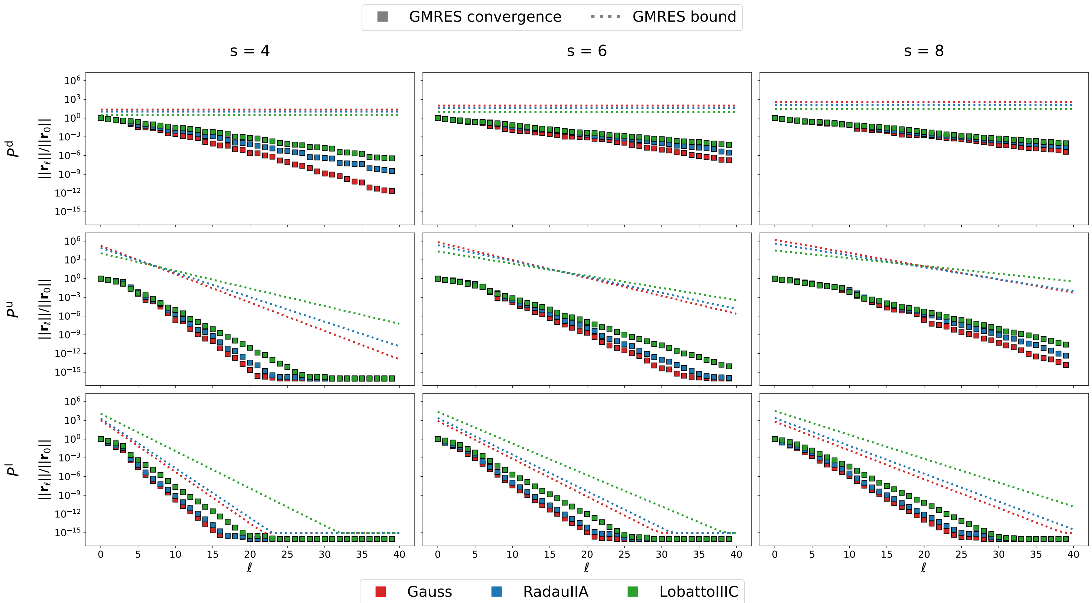

Figure 3

where s is the size of the diagonal blocks. Relabeling the blocks in (22) we get

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

Assuming each X_{θ_j} is diagonalizable as in Proposition 1, we notice that $\{\theta_j\}$ covers reasonably well the intervals $I_{h,\tau,\ldots}$ as $h \to 0$ (see (17)) and, in the spirit of Section 2, the natural bound of (22) becomes

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

First, let us assume there is a uniform bound $\kappa\left(S_\theta\right) \leq \kappa_S$ for all $\theta \in I_{h,\tau,\ldots}$, which experimentally seems to be the case (see [18]) and can be confirmed analytically for $s = 2, 3$ (see [9]) – this is an important and non-trivial assumption and a proper justification is an open problem. Next, we notice that the matrices X_θ depend *smoothly*[4] on θ and as a result so do their eigenproperties. In particular, the eigenvalues $\xi_\theta^{(i)}$ of X_θ will – by definition – form an *algebraic curve*[5] with s *arcs* (sometimes also called *branches*) some of which can be degenerate, e.g., reduced to just a point (incidentally, this is the case for at least one arc of the algebraic curve for any of the triangular preconditioners due to Corollary 2). Denoting the algebraic curve for the given Butcher tableau A

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

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

and a choice of preconditioner $P^\star$ by Γ, we obtain

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \leq \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq\ell}} \max_{\theta\in I_{h,\tau,\dots}} \kappa\left(S_\theta\right) \max_{i=1,\dots,s} \left|\varphi\left(\xi_\theta^{(i)}\right)\right| \leq \kappa_S \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq\ell}} \max_{\xi\in\Gamma} \left|\varphi\left(\xi\right)\right|. \tag{23}$$

Notice that if we replace in the above the interval $I_{h,\tau,\dots}$ with its limit $I_{\lim}$ as $h, \tau \to 0$ (see (17)), we obtain a bound for all mesh sizes. Noticing that, in our case, the preconditioned system matrix has a limit as θ tends to either of the endpoints of $I_{\lim}$, it follows that the arcs of the corresponding algebraic curve correspond to the eigenvalues of these limit matrices. Hence, the effect of mesh refinement becomes sampling more points along Γ and stretching it towards these fixed endpoints (and possibly in increasing κ_S). This suggests that from a certain mesh size onward, the mesh refinement will have little effect on Γ and hence will not affect the min-max part of (23), shedding some light on why these preconditioners are quite robust under mesh refinement.

Remark 3. *We highlight that the numerical experiments in [20, 2] as well as in [18] and in Section 4 clearly show that the spectra of the preconditioned systems cover reasonably well an algebraic curve. For two-stage methods, this behavior has been observed, proved and used to obtain descriptive GMRES bounds in [9]. Moreover, for any algebraic curve Γ we have $\Gamma = \partial_{\mathbb{C}}\Gamma$, which is convenient from the point of view of choosing $\sigma^{\text{non-discr}}$, see Remark 1 and below.*

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

Remark 3 also highlights that the bound (11) is unlikely to be very descriptive or even usable. Indeed, the algebraic curves can reach into the right half-plane $\{\mathrm{Re}(z) < 0\}$ (making the bound useless due to 0 being included in the bounding circle) or, in the more favorable case, the arcs of the algebraic curve are *extremely* unlikely to align with the circle so that the bound have some resemblance of being what we earlier called *appropriate*. Naturally, the bound on the right-hand side of (23) is constructed to remedy that but the key question becomes if this bound is also *functional*, namely if we can (approximately) evaluate it.

To this end, we follow the excellent paper [5] on this topic and start by looking at the *asymptotic* convergence rate (justified by Remark 3 above). Considering (23) we are led to look at the so-called *logarithmic capacity of* Γ, denoted by $\mathrm{cap}(\Gamma)$, which can be viewed as a measure of a compact set without isolated points in $\mathbb{C}$. In fact it is known to asymptotically correspond to the maximal modulus of the *extremal polynomials* (sometimes also called Chebyshev polynomials) associated with Γ, namely

$$\left(\min_{\deg(\varphi)\leq\ell} \max_{z\in\Gamma} |\varphi(z)|\right)^{1/\ell} \to \mathrm{cap}(\Gamma), \quad \text{as } \ell \to +\infty, \tag{24}$$

where the quantity on the left-hand side relates to the quantities we have seen in the GMRES bounds. There are two important caveats to using $\mathrm{cap}(\Gamma)$. The first one, which has been also highlighted as a caveat for using the analysis in [5] overall, is the fact that that (24) only provides some information about the *limit behavior* as $\ell \to +\infty$, whereas we are interested in the behavior for relatively small values of ℓ, say $\ell \leq 50$ or 100. To large extend this issue is addressed by Remark 3 that highlights that we expect a linear convergence rate throughout the iteration. The second one is the fact that (24) describes the limit scaling of the maximal modulus over *all polynomials* – it lack the crucial scaling $\varphi(0) = 1$ of Krylov methods. This issue can be fixed by re-scaling (see [5, Section 2]), shifting our attention from the logarithmic capacity to *Green's functions associated with* Γ, as long as Γ is compact and without any isolated points.

Things simplify considerably if we assume that Γ is connected as then the normalized quantity

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

can be evaluated directly using the conformal maps, in particular the Schwarz-Christoffel maps. Without going into the details (the interested reader can find these in [5, Sections 2 and 3]), we obtain the *asymptotic convergence rate estimate* ρ_{est} as

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

where $\Phi(z)$ is the Schwarz-Chriostoffel map that maps the exterior of Γ to the exterior of the unit circle. In [5, Section 3, Theorem 2 and below], the authors puts it as

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

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

For not too complicated connected, compact sets the map Φ and its value at the origin can be calculated using the Schwarz-Christoffel MATLAB toolbox [3] but we immediately notice that in Figure 2 the set of eigenvalues along Γ is not connected and the actual algebraic curve Γ itself is also not available in an easy form, i.e., neither of these can be directly given as an input to the SC toolbox. We take the natural next step and approximate Γ by its linear interpolation based on the available eigenvalues $\xi_\theta^{(i)}$. The linear interpolation gives us a good approximation of the arcs of Γ and we use the point $1 + 0i$ as the natural point to join them (also by linear interpolation) and denote the resulting set Γ_h. Recalling the limit behavior in (17), we also see that Γ_h will tend towards Γ as $h \to 0$ for our model problem.

The calculation of $\xi_\theta^{(i)}$ may be quite expensive (although, in principle, this can be done independently for each θ_k and thus can be heavily parallelized for large n) and on top of that the SC toolbox will suffer numerically from calculating with Γ_h obtained from the interpolation with large n – both in the sense of large computational complexity as well as in the sense of numerical issues (called *over-crowding*, see [3] but also [4, Section 2.6]). Moreover, in most applications we do not know the spectrum of L beforehand anyway and usually are content with some rough estimates on the extremal eigenvalues $\theta_{\min}$ and $\theta_{\max}$.

To this end, we recall the idea in [9, Section 4] and instead of calculating Γ_h we use only the information about $\theta_{\min,\max}$ and artificially sample a fixed number of "fake" points ϑ_k between them, say q of them. Then we replace θ_k by ϑ_k in the definition of Γ_h, obtaining Γ_q – an approximation of Γ_h (and a further approximation of Γ) based on the linear interpolation given by the eigenvalues of the matrices X_{ϑ_k}. We illustrate these points in Figure 4.

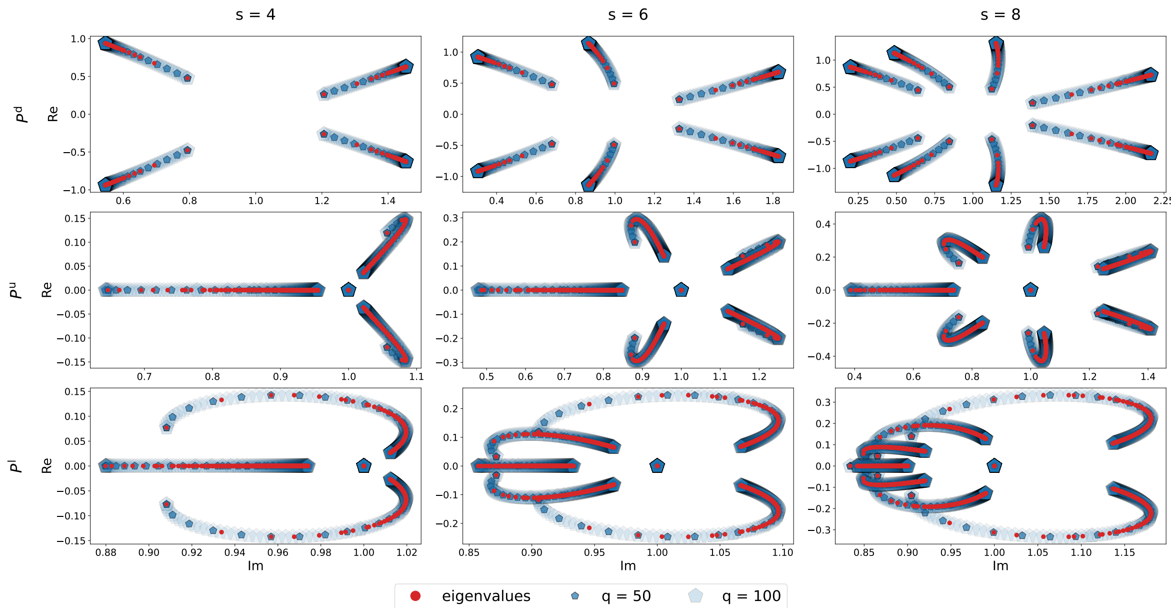

Figure 4: The eigenvalues of the matrices X_{θ_k} (red) and X_{ϑ_k} (blue, for different values of q), using the preconditioner P^{d}. Joining these together with line segments would yield the curves Γ_h (red) and Γ_q (blue).

Another key point is that using the SC toolbox[6] – namely the functions `extermap` and `evalinv` – has difficulties (as far as we understand it) when the arcs of Γ_q intersect, e.g., as is the case for $s = 8$ and the preconditioner P^{l}, see Figure 2. Intuitively, this makes sense as the exterior of Γ_q then has multiple components, making the original set-up more complicated (we assume that theoretical treatment of such problems could be approached based on the analysis in [7]). We address this issue by taking the "envelope" of the arcs – if two arcs intersect, we follow the one staying outwards, e.g., in the case of $s = 6$ and the preconditioner P^{l} we would exclude some portion of the arcs with smaller imaginary part (the densely populated portions) as these lie "inwards" relative to the arcs with the larger imaginary part, see Figure 5. Finally, we illustrate the calculated Schwarz-Christoffel maps – or rather their contours – in Figure 5 together with the used inputs Γ_q (with the exception of $s = 6, 8$ and the preconditioner P^{l}, where we used the "envelopes") and also the asymptotic convergence rate estimate ρ_{est} in Figure 6. First, we see that the results in Figure 6 fully support argument in Remark 3 for considering ρ_{est} as *the* descriptive quantity for the convergence rate. Including an estimate for κ_S then gives also an estimate for GMRES convergence – not just its rate, see Section 4. Second, we highlight that for $s = 8$ and the preconditioner P^{u}, the arcs turned so that the right-most arcs almost intersect themselves. This again causes problems for the toolbox, which raises on output the possibility that the calculated map could be incorrect. Although the predicted ρ_{est} seems accurate, we see in Figure 5 that contours have ripples, confirming that the calculated results should be taken with skepticism. This can be fixed by a similar "envelope-like"

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

approach we described for $s = 6, 8$ and the preconditioner P^{l}, see Section 4, obtaining a further approximation. Although there are couple of other similar caveats concerning the implementation of the above ideas, we have always found that a simple solutions (such as considering the envelope or pruning the fake points in order to alleviate the crowding) can be used to fix them *and* still give an *appropriate* insight into the GMRES convergence rate. As long as κ_S does not completely dominate the ideal GMRES bound (10) this then translates to descriptive GMRES convergence estimates, see Section 4.

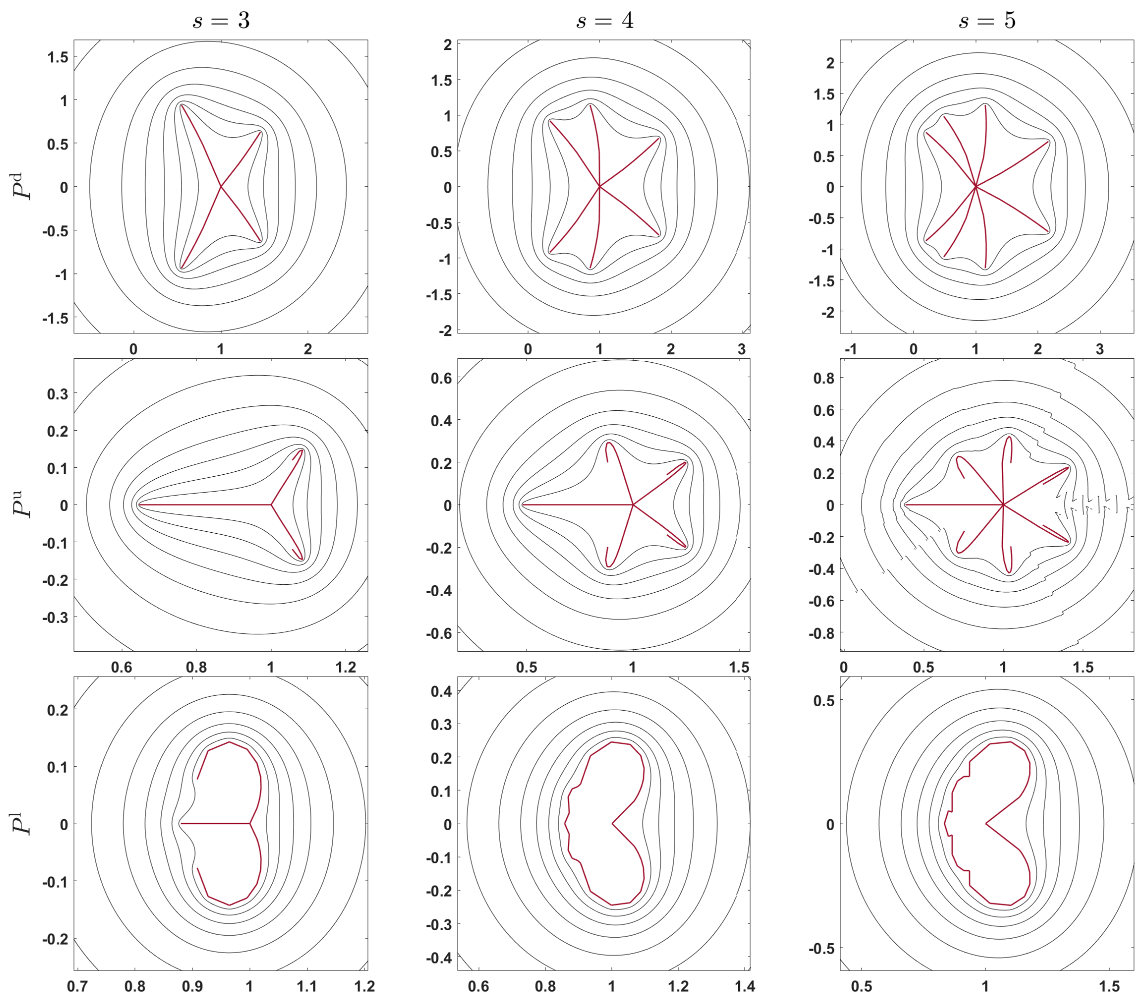

Figure 5: In red: the curves Γ_q (first plots 1 to 7) and their "envelopes" (plots 8 and 9) for the Gauss Butcher tableau, taking $q = 15$. In black: the contours of the corresponding Schwarz-Christoffel map of the exterior of these curves (or envelopes) mapped to the exterior of the unit circle, see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

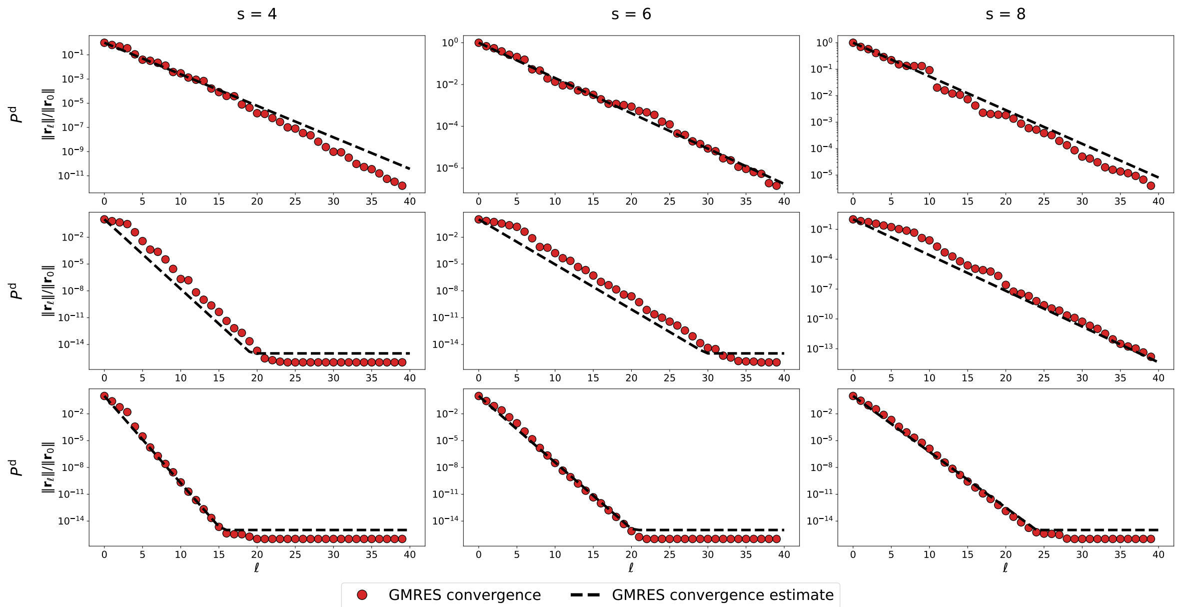

Figure 6: The convergence behavior of preconditioned GMRES, using the Gauss Butcher tableau, together with the convergence estimate based on the calculated asymptotic convergence rate estimate ρ_{est}.

The above analysis also gives insight into the staircase-like behavior, which has been observed and explained for $s = 2$ and the preconditioner P^{d} in [9] by observations about the minimal residual polynomial φ_ℓ^{MR} (sometimes also called the GMRES polynomial; see [16, Section 5.7.1]). Namely, the arguments used in [9] remain valid as long as the branches are not very close to each other[7] – as long as the branches are far apart, the maximum of the polynomial φ_ℓ^{MR} will decrease significantly more at the steps $\ell = s \cdot j$ for $j = 1, 2, \ldots$ because only then each branch can get some attention. If the branches become close, then we do not expect this extra jump because keeping the absolute value of the polynomial small along one of the branches naturally translates into keeping the absolute value of the polynomial also small enough along another one. This is the most pronounced in the first s iterations of GMRES, as we can see in Figure 6, where the convergence curves begin with a slower convergence phase – *precisely s steps* – for P^{d} and P^{u}, in contrast to these of P^{l}, where the arcs intersect and are, in general, closer to each other. We illustrate this further in Figure 7 for the preconditioner P^{d} for $s = 4, 8$ by looking at the polynomial φ_ℓ^{MR} and its roots (called harmonic Ritz values). We see that in the first row (4 branches, far apart) the possibility of "placing" one root along each of the branches was much more crucial (resulted in a more significant decrease of the modulus of the polynomial over the spectrum of the preconditioned system) than for the second row (8 branches with two complex conjugate pairs of branches that are close to each other). We note that an example of explanation (and prediction) of a *complete* staircase behavior of GMRES can be found in [5, Figure 9 and below].

We also want to comment on a similarity with the results in [14, 15]. There, the authors addressed the question of *delay of convergence* by using similar reformulations to ours, also obtain-

[7] In [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 other".

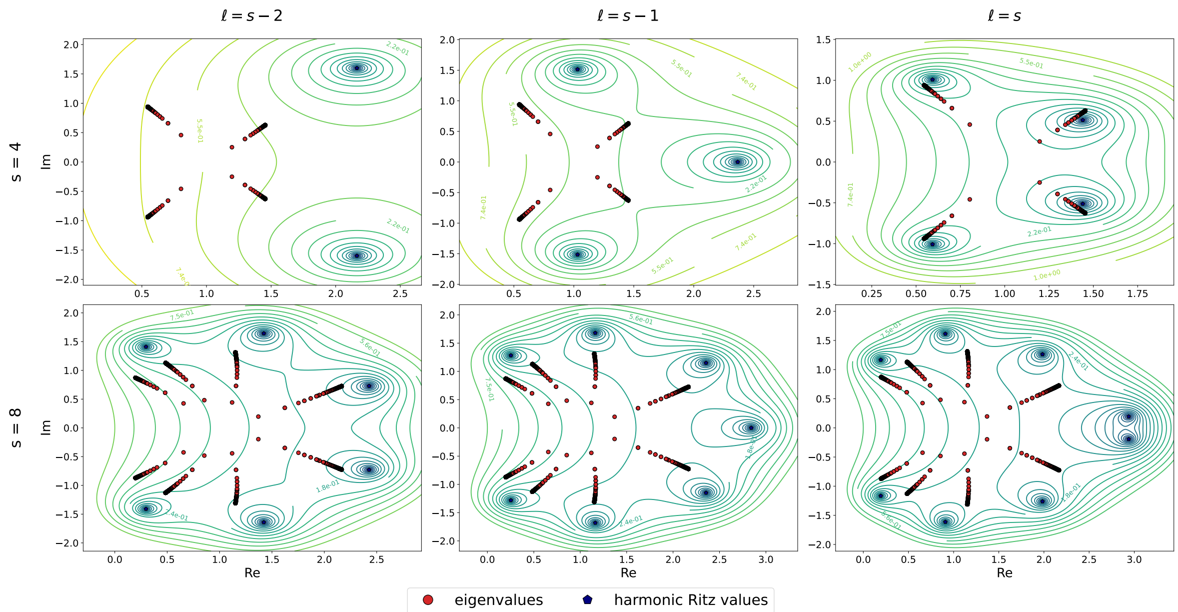

Figure 7

ing a GMRES problem reformulated as for block-diagonal matrix using a Kronecker-product-like techniques as in Lemma 1. In particular, in [15, Section 3.1] the authors use the equality

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

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

$$\|\mathbf{r}_\ell\|^2 = \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le\ell}} \sum_{j=1}^{n} \left\| \varphi\left(X_j\right)\mathbf{s}_0^{(i)} \right\|^2 \ge \sum_{j=1}^{n} \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le\ell}} \left\| \varphi\left(X_j\right)\mathbf{s}_0^{(i)} \right\|^2, \tag{25}$$

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

"... the presence of at least one system with tridiagonal Toeplitz matrix $T_j = \mathrm{tridiag}(\gamma_j, \lambda_j, \mu_j)$ that is 'close to the Jordan block' (cf. [15, Section 3.3] but see also [14]), and with l representing the index of 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) ...
... 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. "*

We see that the approach is *qualitatively* different – both in the intended direction as well as in the results it can deliver – in spite of the fact that it works with the same technique. It is tempting to try to reformulate 25 but given the above analysis, we believe that that is a more appropriate way to study the convergence behavior of these preconditioners.

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

Remark 4. *Another commonly used bounds for GMRES use the* field of values $\nu(C)$ *or the δ-*pseudospectrum $\sigma_\delta(C)$ *of the system matrix C. By a direct calculation we obtain, for our model problem, the field of values as*

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

where X_k are given as in (18) and the set addition is understood element-wise, i.e., $\nu(X_1)+\nu(X_2) = \{\alpha_1 + \alpha_2 \,|\, \alpha_1 \in \nu(X_1), \alpha_2 \in \nu(X_2)\}$, or, more generally

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

For the pseudospectrum we obtain an analogous formula, namely

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

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

4 Numerical Examples

In this section we use the above analysis for more involved setting and, more importantly, also demonstrate the convergence estimates (instead of only the convergence rate estimates). We use the FEM discretization in space[8] for different geometries in Example 1 and 2, see Figure 8. We also fix the number of time steps to balance the spatial and time error (see the (L2) definition in Section 2), namely we fix

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

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

$$\frac{\|\mathbf{r}_\ell\|}{\|\mathbf{r}_0\|} \lesssim \min\left\{\kappa_S^{\text{est}}\rho_{\text{est}}^{\ell}, 1\right\},$$

where the estimate κ_S^{est} of κ_S is computed from the eigenbasis condition numbers of the "fake sampled" matrices X_{ϑ_k} for $k = 1, \ldots, q$. In our experience, the best results are obtained with $q \approx 15 - -20$, as increasing q further leads to crowding problems in the SC toolbox and eventually to problems with the convergence of the Schwarz-Christoffel map. We also found it that spacing the fake points ϑ_k *logarithmically* in the corresponding interval somewhat alleviates this issue and leads to more accurate predictions of the arcs of the given algebraic curve. We also recall, that the seeming independence of the preconditioner quality on the spatial mesh size h was sufficiently documented elsewhere (see [20, 17, 2, 9, 18]) and explained in Section 3 so that in our eyes, there is no need to address this direction here. Illustration of the solutions as well as further numerical experiments can be found in [18, Chapter 7].

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

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

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

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

$$\Gamma_N = \{x=0\} \cup \{y=0\} \cup \{y=4\}, \quad \Gamma_R = \{x=4\}, \quad p = 1, \sigma = \begin{cases} 10^3 & \text{if } (x,y) \in \text{Cookie}, \\ 1 & \text{otherwise}, \end{cases}$$

$$f(x,y,t) = \begin{cases} 3 & \text{if } \|(x,y) - (2,2)\| \leq 1, \\ 0 & \text{otherwise}, \end{cases}$$

[8] Wherever we talk about an 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.

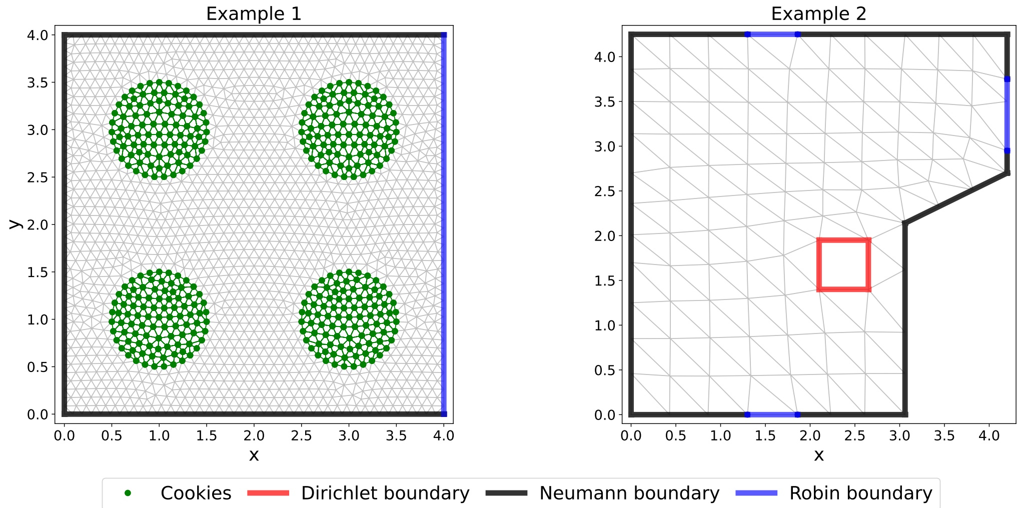

Figure 8: The initial triangulations for Example 2 and 3 together with the boundary condition types and, for Example 2, also with highlighting the points with lower heat conductivity.

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

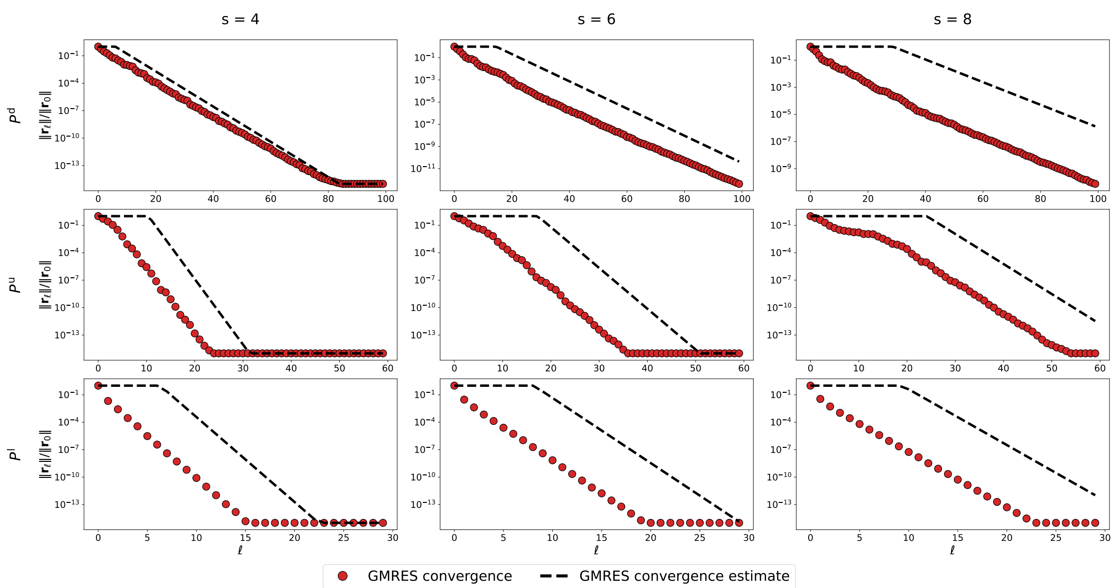

Figure 9: The GMRES convergence behavior with the convergence estimates based on ρ_{est} for Example 1.

Example 2: The cabin heating The second problem uses the 2D projection of an attic room of a cabin in the western Bohemia region, which primary heating is the chimney (bottom-right corner, modeled with a Dirichlet boundary condition changing in time), with two windows (top

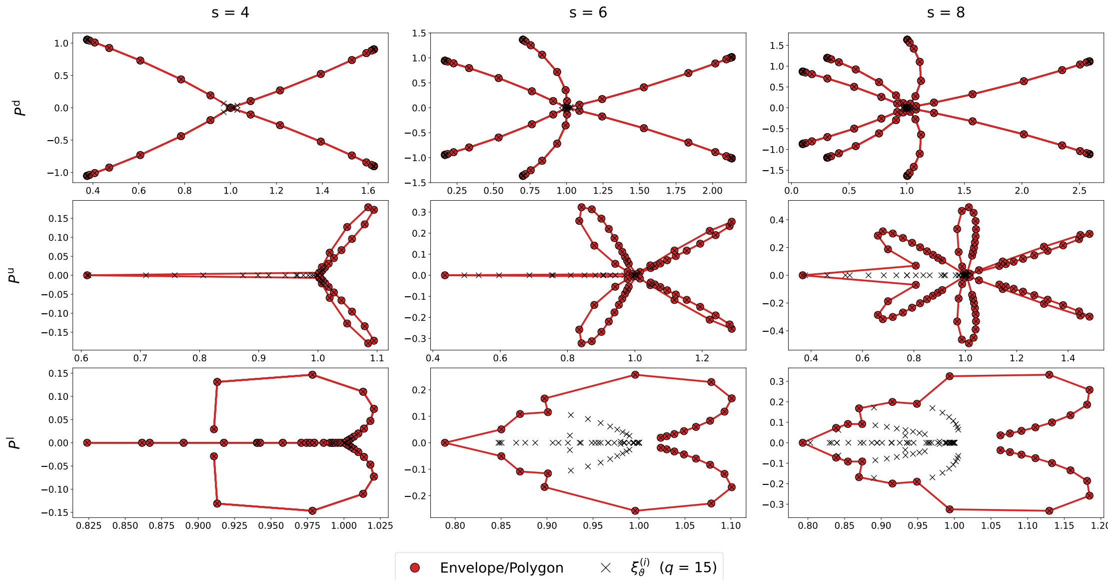

Figure 10: The algebraic curve polygon approximations that are used in the Schwarz-Christoffel MATLAB toolbox to calculate ρ_{est} – for some settings these correspond to the eigenvalues $\xi_\vartheta^{(i)}$ and in some these only enclose these – for Example 1.

and bottom) and doors (right), modeled with Robin boundary conditions with Robin parameters p_w and p_d, and a good insulation otherwise, modeled with a Neumann condition. We obtain the problem

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

and take the data as

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

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

Summary Overall, we observe that the convergence rate estimates delivered a very accurate estimate even for these more involved problems but the conditioning of the eigenbasis of the matrices X_{ϑ_k} notably deteriorated as we increased s. The fact that this has not showed in the GMRES convergence behavior suggests that more delicate bounds, such as mentioned in Remark 4 could give a more detailed insight into the matter. We also showed the polygons used in the Schwarz-Christoffel toolbox – notice that in many of the plots we excluded part of the arcs, mainly because either (a) the arcs intersected and we took the envelope of the algebraic curve (usually for the preconditioner P^{l}) or (b) the points sampled along the arcs crowded sections of the arcs, which

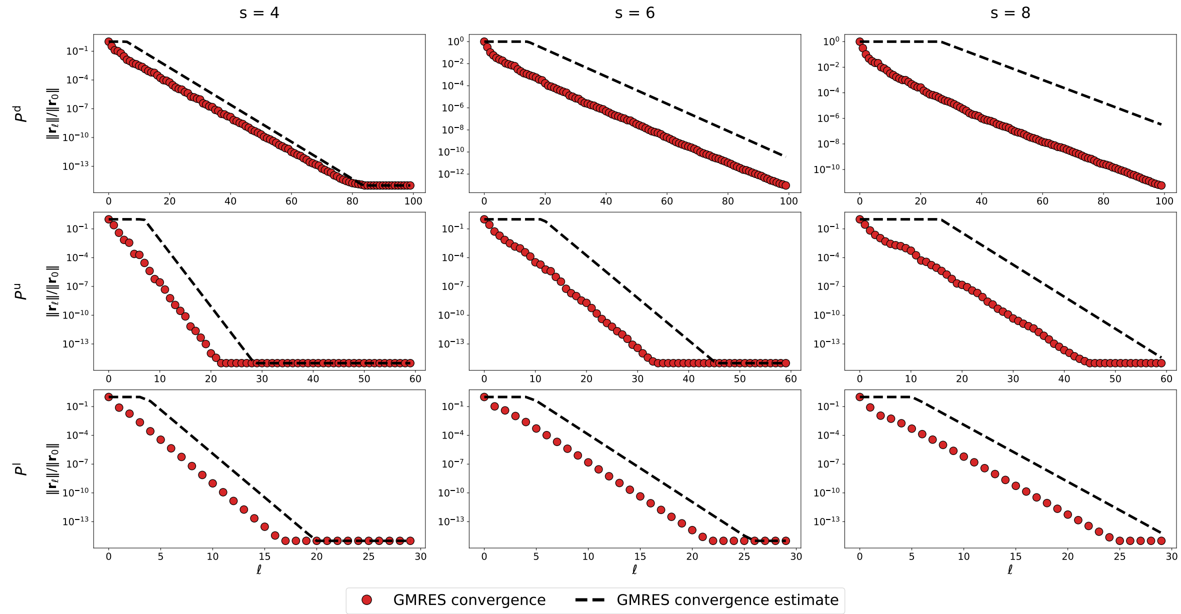

Figure 11: The GMRES convergence behavior with the convergence estimates based on ρ_{est} for Example 2.

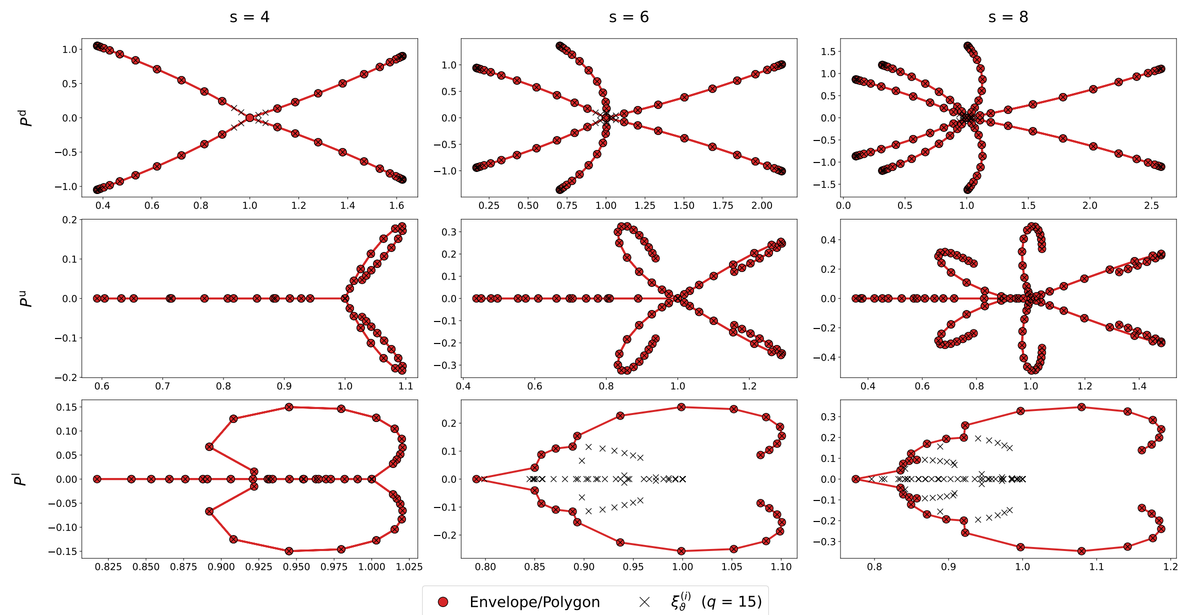

Figure 12: The algebraic curve polygon approximations that are used in the Schwarz-Christoffel MATLAB toolbox to calculate ρ_{est} – for some settings these correspond to the eigenvalues $\xi_\vartheta^{(i)}$ and in some these only enclose these – for Example 2.

caused issues for the toolbox. In such cases we sparsified these regions by dropping some of these points. As a result, the Schwarz-Christoffel external map converged better and faster than for the problem in Section 3.1 and the contours were "ripple-free" for all of our problems, otherwise looking almost precisely as the ones in Figure 5.

5 Concluding remarks

The main goal has been to understand the block preconditioners considered in [20, 2, 17] in more details and try to explain their success and/or limitations. This goal was, in our eyes, mostly achieved but could be further improved in the sense of Remark 4 or by considering a more refined version of the bound (10), see [6, Section 2.1, equations (2.1) and (EV')] – this remains an area of interest for us for the future. Moreover, the above analysis can be directly used to try to *optimize* the Runge-Kutta methods, following the ideas in [20, 18, 9]. We also note that in practice, solving with either of the matrices $P^{\mathrm{d,u,l,GSU,GSL,...}}$ is often done with some level of *inaccuracy*, e.g., using a multigrid method. The question of interaction of this inaccuracy with the overall GMRES convergence is an important one and to the best of our knowledge has been addressed only numerically in [18, Chapter 7]. We also note that adapting the above analysis to the framework presented in [23, 22] or reformulating from vector equation to the matrix equation as suggested in [19] and study in detail the comparison of these approaches for the IRK setting are both attractive directions for future research.

6 Acknowledgement

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

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] T. A. Driscoll. A MATLAB toolbox for Schwarz-Christoffel mapping. Technical Report 2, 1996.

[4] T. A. Driscoll and L. N. Trefethen. *Schwarz-Christoffel mapping.* Cambridge University Ppress, Cambridge, First edition, 2002.

[5] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen. From potential theory to matrix iterations in six steps. *SIAM Review*, 40(3):547–578, 1998.

[6] M. Embree. How descriptive are GMRES convergence bounds? 2022.

[7] M. Embree and L. N. Trefethen. Green's functions for multiply connected domains via conformal mapping. *SIAM Review*, 41(4):745–761, 1999.

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

[9] M. J. Gander and M. Outrata. Spectral analysis of implicit 2 stage block Runge-Kutta preconditioners. *Linear Algebra and its Applications*, 2023. URL https://doi.org/10.1016/j.laa.2023.07.008.

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

[11] 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.

[12] T. Kato. *Perturbation Theory for Linear Operators*, volume 132. Springer Berlin, Heidelberg, 2013.

[13] D. Kressner and C. Tobler. Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1288–1316, 2011.

[14] J. Liesen and Z. Strakoš. Convergence of GMRES for tridiagonal Toeplitz matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(1):233–251, 2004.

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

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

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

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

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