Space-time finite element discretization of the heat equation
Efficient iterative solvers for the discrete problem

Master’s thesis

submitted by
Nikolas Siccha
student number: 293707

Examiner: Prof. Dr. Arnold Reusken
Advisor: M. Sc. Igor Voulis

June 8, 2017
List of Figures

2.1 The complex eigenvalues of the characteristic temporal matrix . . . . . . . . . . . . . . . 14
2.2 Temporal convergence of the discrete solution . . . . . . . . . . . . . . . . . . . . . . . . . . 18
2.3 Condition numbers of important matrices . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.4 Dynamic vs precompiled quadrature matrices for orthogonal temporal bases . . . . . . . 20
2.5 Dynamic vs precompiled quadrature matrices for nonorthogonal temporal bases . . . . . 20
4.1 Values of the spectral function (cf. Eq. (4.37) for various discretization parameters . . . 36
4.2 Filled contour plot of $\min_{\lambda \in [1, 2]} \sigma_*(\lambda, \mu)$ . . . . . . . . . . . . . . . . . . 49
4.3 Convergence behaviour of the exact Jacobi and Gauss-Seidel preconditioner . . . . . . . . 53
4.4 Convergence rates for the exact Jacobi and Gauss-Seidel preconditioner . . . . . . . . . . 54
4.5 Convergence rates for the exact Gauss-Seidel and Schur preconditioner . . . . . . . . . . 55
4.6 Convergence rates for the exact block multigrid preconditioner . . . . . . . . . . . . . . . 55
4.7 Convergence rates for the inexact preconditioners . . . . . . . . . . . . . . . . . . . . . . . . 57
4.8 Efficiencies for the inexact preconditioners . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58
4.9 Correlation of the residuum with the accuracy of the discrete solution . . . . . . . . . . . 59
1 Introduction

Space-time finite element methods are gaining popularity in solving parabolic partial differential equations (PDE). While the usual methods like backward Euler for solving stiff ordinary partial differential equations remain applicable in the context of parabolic PDES, variational time discretization methods open up more possibilities. One of those methods is the discontinuous Galerkin (DG) time-stepping method. It has many beneficial properties, which include but are not limited to nodal superconvergence [Tho06] and strong A-stability [MS11]. For a thorough treatment in the context of parabolic problems we refer to [Tho06].

The difficulties in using the DG time-stepping method are mainly the increased complexity of its implementation and the fact that the arising linear systems are not straightforward to solve. These systems are of the same structure as the ones that arise when discretizing using implicit Runge-Kutta schemes.

Some solution strategies block-diagonalize the arising linear system of equations, for example [SS00]. The problem with this approach is twofold. While the block matrices are Euler-like systems, they are complex valued, which doubles the computational cost. The more serious problem is the badly-conditioned matrix of eigenvectors used in the diagonalization, cf. [Sme16].

For the heat equation a few methods have recently been proposed. In the context of Runge-Kutta schemes block Jacobi and block Gauss-Seidel preconditioners have been analyzed in [SMN06, MNS07]. They were found to be order-optimal with respect to all discretization parameters but the number of Runge-Kutta stages.

In [Sme16] a method has been proposed that is order-optimal with respect to all discretization parameters and robust and bounds the condition number of the symmetrically preconditioned matrix by 4, but appears to be relatively expensive.

Shortly before that another method has been proposed in [WB15] and [BB15] which is also order-optimal, but relies on the Schur complement formulation of the $2 \times 2$-block diagonalized matrix. This method bounds the condition number of the symmetrically preconditioned matrix by 2, but it appears to rely on the Schur complement matrix to be relatively well-conditioned.

In [VVL06] a method has been proposed which uses as preconditioner a block multigrid method. This appears to work well, but appears to also be more expensive than the other methods.

In this thesis we will improve the methods proposed in [SMN06, MNS07]. In the setting of the DG time-stepping method we show them to be order-optimal with respect to all discretization parameters. We will provide bounds on the convergence rates of the preconditioned methods which are equal to the ones in [BB15].

This thesis is structured as follows. In the first chapter we introduce our discretization method. This includes some standard results, but also a discussion on how to construct optimal temporal basis and test functions. Next we recall some standard convergence results of the relevant stationary iterative methods and preconditioned Krylov subspace methods. In the following chapter we use these convergence results to obtain bounds on the convergence rates of our preconditioned methods and to identify the temporal basis and test functions that are optimal with respect to the upper bound on their convergence rate. Furthermore we provide a qualitative analysis of the behaviour of the methods if we only approximate the application of the preconditioners. Finally we will perform numerical experiments to underpin our analysis. In this section we also compare the performance of our methods with the one of the methods proposed in [BB15] and in [VVL06]
2 Variational time discretization of the heat equation

In this chapter we will introduce our discretization method applied to a parabolic model problem. For the sake of simplicity and regularity we will assume the spatial domain $\Omega \subset \mathbb{R}^d$ to be open, convex and polygonal. We consider

**Problem 2.1 (Heat equation):**

Given initial condition $u_0 : \Omega \to \mathbb{R}$ and source term $f : (0, T] \times \Omega \to \mathbb{R}$, find the classical solution $u : (0, T] \times \Omega \to \mathbb{R}$ that satisfies

\[
\begin{align*}
  u' - \Delta u &= f & \text{for } (t, x) \in (0, T] \times \Omega, \\
  u &= u_0 & \text{for } (t, x) \in \{0\} \times \Omega, \\
  u &= 0 & \text{for } (t, x) \in (0, T] \times \partial \Omega.
\end{align*}
\]

We will introduce the necessary notation to formulate the weak form of this problem, and then state the discrete problem obtained by a space-time finite element method.

In doing so, it appears sensible to first discuss an elliptic model problem, so as to first introduce and use some of the notation in the context of a simple stationary problem. Some of the notation introduced in this preparatory section will also be used in the later chapters, e.g. when we discuss the geometric multigrid method. The section on the space discretization will be followed by a section focused on the variational time discretization. Again, some of the notation introduced in this section will be reused in later chapters, e.g. when we discuss the preconditioners.

Both sections are structured in a similar way. First we introduce the necessary inner products, norms and function spaces which will be needed to formulate the weak problem. Without going too much into detail concerning how exactly to obtain them, we will state the weak formulations, accompanied with some standard results concerning existence and uniqueness of the weak solution. We continue by introducing the respective discrete solution spaces and then state the discrete weak problems, again along with some standard results concerning existence, uniqueness and the discretization error. Finally we will introduce suitable bases of the discrete solution spaces to obtain the matrix equation corresponding to the discrete problems. For the space discretization, we will not define the used bases ourselves, but refer to the literature. Instead, the focus will lie on appropriate triangulations of the spatial domain and on the spectral condition numbers of the stiffness matrices. For the time discretization the focus will lie on the exact definition of the temporal bases. The choice of the temporal bases will turn out to play a major role for the iterative solvers and preconditioners, as such it appears mandatory to provide a clear definition of them.

This chapter will be concluded by a short section containing numerical experiments to confirm the correctness of our implementation. This will be done by reproducing the previously predicted convergence results for the discretization error.

2.1 The space discretization

We consider the elliptic model problem:

**Problem 2.2 (Linear reaction-diffusion equation):**

Given positive $\mu, \rho \in \mathbb{R}$ and a source term $f : \Omega \to \mathbb{R}$, find the classical solution $u : \Omega \to \mathbb{R}$ that satisfies

\[
\begin{align*}
  -\mu \Delta u + \rho u &= f & \text{for } x \in \Omega, \\
  u &= 0 & \text{for } x \in \partial \Omega.
\end{align*}
\]

We want to formulate the weak form of this problem. To this end, we introduce some standard notation. We define the $L^2(\Omega)$ inner product:

\[
\langle \cdot, \cdot \rangle_{L^2(\Omega)} : [\Omega \to \mathbb{R}] \times [\Omega \to \mathbb{R}] \to \mathbb{R},
\]

\[
\langle f, g \rangle_{L^2(\Omega)} := \int_\Omega fg \, dx
\]

(2.6)
which induces the $L^2(\Omega)$ norm:

$$\|f\|_{L^2(\Omega)} : [\Omega \to \mathbb{R}] \to \mathbb{R}, \quad \|f\|_{L^2(\Omega)} := \sqrt{\langle f, f \rangle_{L^2(\Omega)}}$$

and the space of square integrable functions:

$$L^2(\Omega) := \left\{ f : \Omega \to \mathbb{R} \mid \|f\|_{L^2(\Omega)} < \infty \right\}$$

Given a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ with norm $|\alpha| := \sum_{i=1}^{d} |\alpha_i|$ we define the multi-index derivative $D^\alpha f := \frac{\partial^{\alpha_1} f}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d} f}{\partial x_d^{\alpha_d}}$, which will be interpreted in a weak sense (cf. [Eva10, subsection 5.2.1]). Furthermore, we define the Sobolev space:

$$W^{k,p}(\Omega) := \{ f \in L^p(\Omega) \mid D^\alpha f \text{ exists and } D^\alpha f \in L^p(\Omega) \text{ for } |\alpha| \leq k \}$$

and the Hilbert space:

$$H^k(\Omega) := W^{k,2}(\Omega)$$

where the latter can also be defined as the completion of $W^{k,2}(\Omega)$ with respect to the norm induced by the $H^k(\Omega)$-inner product:

$$\langle \cdot, \cdot \rangle_{H^k(\Omega)} : H^k(\Omega) \times H^k(\Omega) \to \mathbb{R}, \quad (u, v)_{H^k(\Omega)} := \sum_{|\alpha| \leq k} \langle D^\alpha u, D^\alpha v \rangle_{L^2(\Omega)}$$

which implies that $H^k(\Omega)$ is a Hilbert space. We now define the subspace of $H^k(\Omega)$ which consists of trace-zero functions in $H^k(\Omega)$:

$$H^k_0(\Omega) := \{ u \in H^k(\Omega) \mid u_{|\partial \Omega} = 0 \}$$

where the equality $u_{|\partial \Omega} = 0$ has to be interpreted in a trace sense (cf. [Eva10, section 5.5]). This is simply the appropriate analogue to the boundary condition of problem 2.2. For the right-hand side we define the dual space to $H^1_0(\Omega)$:

$$H^{-1}(\Omega) := \{ f : H^1_0(\Omega) \to \mathbb{R} \mid f \text{ linear and bounded} \}$$

and the pairing between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$:

$$\langle \cdot, \cdot \rangle : H^{-1}(\Omega) \times H^1_0(\Omega) \to \mathbb{R}, \quad \langle u^*, u \rangle := u^*(u)$$

which are used to define the source term functional:

$$g \in H^{-1}(\Omega), \quad \langle g, w \rangle := \langle f, w \rangle_{L^2(\Omega)}$$

We define the usual bilinear forms, i.e. the stiffness bilinear form:

$$a(\cdot, \cdot) : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}, \quad a(v, w) := \langle \nabla v, \nabla w \rangle_{L^2(\Omega)}$$

the mass bilinear form:

$$m(\cdot, \cdot) : L^2(\Omega) \times L^2(\Omega) \to \mathbb{R}, \quad m(v, w) := \langle v, w \rangle_{L^2(\Omega)}$$

and the global bilinear form:

$$k(\cdot, \cdot) : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}, \quad k(v, w) := \mu a(v, w) + \rho m(v, w)$$

After the usual steps, i.e. multiplication with a smooth test function, integrating over the spatial domain $\Omega \subset \mathbb{R}^d$ and performing integration by parts, the weak problem now reads:

**Problem 2.3** (Weak form of the linear reaction-diffusion equation):

Given a source term functional $g \in H^{-1}(\Omega)$, find the weak solution $v \in H^1_0(\Omega)$ that satisfies

$$k(v, w_x) = \langle g, w_x \rangle \quad \forall \ w_x \in H^1_0(\Omega).$$
Remark 2.1. It is well-known (cf. [Eva10, section 6.2.2]), that the global bilinear form \( k(\cdot, \cdot) \) as defined above is
\[
\text{continuous in } H^1_0(\Omega): \quad \exists \alpha > 0 \text{ s.t. } |k(u, v)| \leq \alpha \|u\|_{H^1_0(\Omega)} \|v\|_{H^1_0(\Omega)} \quad \forall u, v \in H^1_0(\Omega)
\]
and
\[
\text{\( H^1_0(\Omega) \)-elliptic:} \quad \exists \beta > 0 \text{ s.t. } \beta \|u\|_{H^1_0(\Omega)} \leq k(u, u) \quad \forall u \in H^1_0(\Omega).
\]

Therefore, due to the Lax-Milgram theorem (cf. [Eva10, section 6.2.1]), the weak solution \( v \in H^1_0(\Omega) \) exists and is unique.

2.1.1 The discrete weak problem

We discretize the weak problem by restricting the search to a finite-dimensional discrete weak solution space \( V_\ell \subset H^1_0(\Omega) \). For this we introduce a family of triangulations \( T_1, \ldots, T_L \) of \( \Omega \subset \mathbb{R}^d \), consisting of sets of disjoint \( d \)-simplices \( T_\ell := \{ T_{\ell,1}, \ldots, T_{\ell,m_\ell} \} \). We define the spatial step size and assume the family of triangulations to be
\[
\text{admissible:} \quad \left\{ \begin{array}{l}
\Omega = \bigcup T_\ell \\
T_\ell \text{ has no hanging nodes}
\end{array} \right. \quad \text{for } \ell = 1, \ldots, L, \tag{2.20}
\]
\[
\text{regular:} \quad \left\{ \begin{array}{l}
\inf \{ h | T_1, \ldots, T_L \} = 0 \\
\exists \sigma : \frac{h_{T_{\ell,e}}}{\rho_{T_{\ell,e}}} \leq \sigma \forall T_{\ell,e} \in T_\ell
\end{array} \right. \quad \text{for } \ell = 1, \ldots, L, \tag{2.21}
\]
\[
\text{quasi-uniform:} \quad \exists \sigma : \frac{h}{\rho_{T_{\ell,e}}} \leq \sigma \forall T_{\ell,e} \in T_\ell \quad \text{for } \ell = 1, \ldots, L, \tag{2.22}
\]
and
\[
\text{nested:} \quad \forall T \in T_{\ell+1} \exists T' \in T_\ell \text{ s.t. } T \subset T' \quad \text{for } \ell = 1, \ldots, L - 1. \tag{2.23}
\]

We will use continuous and elementwise polynomial ansatz functions. For this we introduce the space of real polynomial functions of degree at most \( q \) in \( d \) variables:
\[
\Pi_q := \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = \sum_{|\alpha| \leq q} c_\alpha x^\alpha \right\}. \tag{2.24}
\]

With the assumptions on the triangulations (Eqs. (2.20) to (2.23)) we get a nested family of finite element spaces \( V_1, \ldots, V_L \subset H^1_0(\Omega) \) defined as
\[
V_\ell := \left\{ \phi \in C_0^0(\Omega) \mid \phi|_{T_{\ell,e}} \in \Pi_q \quad \forall T_{\ell,e} \in T_\ell \right\} \quad \text{for } \ell = 1, \ldots, L. \tag{2.25}
\]

We get

Problem 2.4 (\( \ell \)-th discrete weak problem):
Given a source term functional \( g \in H^{-1}(\Omega) \) find the discrete weak solution \( v_\ell \in V_\ell \) that satisfies
\[
k(v_\ell, w_\ell) = \langle g, w_\ell \rangle \quad \forall w_\ell \in V_\ell. \tag{2.26}
\]

Remark 2.2. Due to the discrete weak solution space \( V_\ell \) itself being a Hilbert space the Lax-Milgram theorem remains applicable, yielding existence and uniqueness of \( v_\ell \).

We now present a standard result from the literature concerning the \( \|\cdot\|_{H^1_0(\Omega)} \)-norm of the discretization error \( v - v_\ell \) due to

2.1. THE SPACE DISCRETIZATION
Lemma 2.1 (Cea’s lemma)
For the weak solution \( v \in H^1_0(\Omega) \), the discrete weak solution \( v_\ell \in V_\ell \) and the global bilinear form \( k(\cdot,\cdot) : H^1_0(\Omega) \times H^1_0(\Omega) \rightarrow \mathbb{R} \) with continuity constant \( \alpha > 0 \) and ellipticity constant \( \beta > 0 \) we get

\[
\|v - v_\ell\|_{H^1_0(\Omega)} \leq \sqrt{\frac{\beta}{\alpha}} \inf_{w_\ell \in V_\ell} \|v - w_\ell\|_{H^1_0(\Omega)} \quad \text{for } \ell = 1, \ldots, L. \tag{2.27}
\]

Proof. A proof can be found in [EG13, lemma 2.28]. \( \square \)

This implies \( v = v_\ell \) if \( v \in V_\ell \).

2.1.2 The matrix equation
The discrete weak problem can be rewritten as a matrix equation. For this we define for \( \ell = 1, \ldots, L \) the dimension of \( V_\ell \) \( n_\ell := \dim(V_\ell) \) and a family of spatial basis functions \( \phi_{\ell,1}, \ldots, \phi_{\ell,n_\ell} \in V_\ell \) that span the discrete weak solution space \( V_\ell \). We can then write the discrete weak solution \( v_\ell \in V_\ell \) as a linear combination of these basis functions using the vector of solution coefficients:

\[
v_\ell \in \mathbb{R}^{n_\ell}, \quad v_\ell := (v_{\ell,1}, \ldots, v_{\ell,n_\ell}) \tag{2.28}
\]

For this we first define the finite element isomorphism:

\[
P_\ell : \mathbb{R}^{n_\ell} \rightarrow V_\ell, \quad v_\ell \mapsto \sum_{i=1}^{n_\ell} v_{\ell,i} \phi_{\ell,i} \tag{2.29}
\]

which lets us write \( v_\ell = P_\ell v_\ell \), and then define the spatial stiffness matrix:

\[
A_\ell \in \mathbb{R}^{n_\ell \times n_\ell}, \quad [A_\ell]_{i,j} := a(\phi_{\ell,i}, \phi_{\ell,j}) \tag{2.30}
\]

the spatial mass matrix:

\[
M_\ell \in \mathbb{R}^{n_\ell \times n_\ell}, \quad [M_\ell]_{i,j} := m(\phi_{\ell,i}, \phi_{\ell,j}) \tag{2.31}
\]

and the global stiffness matrix:

\[
K_\ell \in \mathbb{R}^{n_\ell \times n_\ell}, \quad K_\ell := \mu A_\ell + \rho M_\ell \tag{2.32}
\]

with positive \( \mu, \rho \in \mathbb{R} \) as in problem 2.2. For the right-hand side we define the right-hand side vector:

\[
f_\ell \in \mathbb{R}^{n_\ell}, \quad [f_\ell]_i := \langle g, \phi_{\ell,i} \rangle \tag{2.33}
\]

which has to be evaluated using an appropriate quadrature formula. We can finally write

**Problem 2.5** (Matrix form of the discrete weak problem):
Given an \( f_\ell \in \mathbb{R}^{n_\ell} \), find the \( v_\ell \in \mathbb{R}^{n_\ell} \) that satisfies

\[
K_\ell v_\ell = f_\ell \tag{2.34}
\]

We will discuss properties of the above linear system, i.e. we provide expressions for the spectral condition numbers of the matrices \( A_\ell, M_\ell, K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \). For this we have to choose a suitable set of basis functions. For a given spatial discretization order \( q \in \mathbb{N} \), we will use the standard, nodal \( P_q \) finite element basis, i.e. for \( q = d = 1 \) we get the standard hat functions, and for other values of \( q \) and \( d \) the usual extensions. Then the matrices \( A_\ell, M_\ell \) and \( K_\ell \) are sparse, symmetric positive definite (cf. [EG13, proposition 3.64]) and with our assumptions on the triangulations we get the following relations (cf. [EG13, theorems 9.8 and 9.11]):

\[
\kappa_2(A_\ell) = \mathcal{O}(h^{-2}) \quad \text{for } \ell = 1, \ldots, L, \tag{2.35}
\]

\[
\kappa_2(M_\ell) = \mathcal{O}(1) \quad \text{for } \ell = 1, \ldots, L. \tag{2.36}
\]
2.2 The time discretization

2.2.1 The weak problem

We want to obtain the weak form of the Heat equation (cf. problem 2.1). Before we can write it down, we have to introduce some additional, standard notation.

For a Banach space $X$ with norm $\| \cdot \|_X$ we define the $L^2 (0; T; X)$ norm:

$$ \| \cdot \|_{L^2(0; T; X)} : [0, T] \to X \to \mathbb{R}, \quad f \mapsto \sqrt{\int_0^T \| f(t) \|_X^2 \, dt} \quad (2.37) $$

and the space of $L^2 (0; T; X)$ measurable functions:

$$ L^2 (0; T; X) := \left\{ f : [0, T] \to X \bigg| \| f \|_{L^2(0; T; X)} < \infty \right\} \quad (2.38) $$

and the space of $C ([0; T]; X)$ functions:

$$ C ([0; T]; X) := \left\{ f : [0, T] \to X \bigg| f \text{ continuous and } \max_{0 \leq t \leq T} \| f(t) \|_X < \infty \right\} \quad (2.39) $$

As was done in [Eva10, section 7.1], we will reinterpret the classical solution $u : (0, T) \times \Omega \to \mathbb{R}$ of the heat equation, which originally is a function in $(t, x) \in (0, T) \times \Omega$, to now be a time dependent mapping into $H^1_0 (\Omega)$:

$$ v \in L^2 (0; T; H^1_0 (\Omega)), \quad [v(t)] (x) := u ((t, x)) \quad (2.40) $$

Similarly, we reinterpret the source term $f : (0, T) \times \Omega \to \mathbb{R}$ to be a time dependent mapping into $L^2 (\Omega)$:

$$ g \in L^2 (0; T; L^2 (\Omega)), \quad [g(t)] (x) := f ((t, x)) \quad (2.41) $$

And finally we reinterpret the time derivative $u' : (0, T) \times \Omega \to \mathbb{R}$ to be a time dependent mapping into $H^{-1} (\Omega)$:

$$ v' \in L^2 (0; T; H^{-1} (\Omega)), \quad [v'(t)] (x) := u' ((t, x)) \quad (2.42) $$

**Remark 2.3.** The time derivative again has to be understood in a weak sense, (cf. [Eva10, section 5.9.2]).

Reusing the stiffness bilinear form $a (\cdot, \cdot) : H^1_0 (\Omega) \times H^1_0 (\Omega) \to \mathbb{R}$ (cf. Eq. (2.16)), we can now write

**Problem 2.6** (Weak form of the heat equation):

Given a $g \in L^2 (0; T; L^2 (\Omega))$, find the weak solution $v \in L^2 (0; T; H^1_0 (\Omega))$ with time derivative $v' \in L^2 (0; T; H^{-1} (\Omega))$ that satisfies

$$ (v', w_x) + a (v, w_x)_{L^2(\Omega)} = \langle g, w_x \rangle_{L^2(\Omega)} \quad \forall w_x \in H^1_0 (\Omega) \text{ and a.e. } t \in (0, T], \quad (2.43) $$

$$ v(0) = u_0 \quad \text{ and } \quad t \in (0, T]. \quad (2.44) $$

**Remark 2.4.** It can be shown (cf. [Eva10, section 5.9.2, theorem 3]) that the $\| \cdot \|_{L^2(0; T; X)}$-equivalence class of the weak solution $v \in L^2 (0; T; H^1_0 (\Omega))$ contains a representative that also lies in $C ([0; T]; L^2 (\Omega))$. The initial condition has to be understood to act on this representative.

Existence and uniqueness of the weak solution can also be proven (cf. [Eva10, section 7.1.2, theorems 3 and 4]).

2.2.2 The semi-discrete weak problem

We will now approximate the $v \in L^2 (0; T; H^1_0 (\Omega))$ by a piecewise polynomial approximation. For a given temporal discretization order $p \in \mathbb{N}$ we will use the

$$ \text{dG}(p)\text{-method} \quad (p\text{-th order discontinuous Galerkin time stepping method).} \quad (2.45) $$
We start by partitioning the temporal domain \((0, T) \subset \mathbb{R}\) into \(K \in \mathbb{N}\) time slabs using \(K + 1\) time points \(0 = t^{(0)} < t^{(1)} < \cdots < t^{(K-1)} < t^{(K)} = T\). For \(k = 1, \ldots, K\) we define the \(k\)-th time slab \(I^{(k)} := (t^{(k-1)}, t^{(k)})\) and the \(k\)-th time step size \(\tau_k := t^{(k)} - t^{(k-1)}\). We introduce the space of polynomials of degree less than or equal to \(p\) with coefficients in \(H^1_0(\Omega)\):

\[
\Pi_p \left[ H^1_0(\Omega) \right] := \left\{ f : \mathbb{R} \to H^1_0(\Omega) \mid f(t) = \sum_{i=0}^{p} a_i t^i \text{ with } a_0, \ldots, a_p \in H^1_0(\Omega) \right\}
\]

and define the semi-discrete weak solution space:

\[
S_p := \left\{ X : (0, T) \to H^1_0(\Omega) \mid X|_{[t_k]} \in \Pi_p \left[ H^1_0(\Omega) \right] \right\}
\]

which contains our approximation. We multiply the Weak form of the heat equation (cf. problem 2.6) with a smooth test function \(w_t \in C^\infty([0, T])\), integrate over \((0, T)\) and after integration by parts (in time) we obtain

\[
\int_0^T \langle v, w_x \rangle w_t + a(v, w_x) w_t dt = \langle vw_t|_0^T, w_x \rangle_{L^2(\Omega)} + \int_0^T \langle g, w_x \rangle_{L^2(\Omega)} w_t dt,
\]

which the weak solution \(v\) satisfies for all \(w_x \in H^1_0(\Omega)\) and \(w_t \in C^\infty([0, T])\). We now replace the time dependent mapping into \(H^1_0(\Omega) v\) by an approximation \(v_p \in S_p\) with time derivative \(v_p' \in S_{p-1}\), and the \(w := w_t w_x\) by a test function \(w_p \in S_p\). After another integration by parts and by requiring that \(v_p\) vanishes outside of each \(I^{(k)}\), we get \(K\) localized, sequential problems. To write them down, we define for an \(X \in S_p\)

\[
X^{(k,-)} := \lim_{s \downarrow 0} X(t^{(k)} + s)
\]

the limit from below at \(t^{(k)}\),

\[
X^{(k,+)} := \lim_{s \uparrow 0} X(t^{(k)} + s)
\]

and the shorthands

\[
\text{LHS}^{(k)}(v_p, w_p) := \int_{t^{(k)}}^{t^{(k+1)}} \langle v_p', w_p \rangle_{L^2(\Omega)} + a(v_p, w_p) dt + \langle g, w_p \rangle_{L^2(\Omega)}
\]

\[
\text{RHS}^{(k)}(v_p, w_p) := \langle v_p^{(k,-)}, w_p^{(k,+)} \rangle_{L^2(\Omega)} + \int_{t^{(k)}}^{t^{(k+1)}} \langle g, w_p \rangle_{L^2(\Omega)} dt
\]

to finally obtain

**Problem 2.7** (Semi-discrete weak form of the heat equation):

Given a \(g \in L^2(0, T; L^2(\Omega))\), find the semi-discrete weak solution \(v_p \in S_p\) that satisfies

\[
\text{LHS}^{(k-1)}(v_p, w_p) = \text{RHS}^{(k-1)}(v_p, w_p) \quad \forall \ w_p \in S_p, k = 1, \ldots, K,
\]

\[
v_p^{(0,+)} = u_0
\]

**Remark 2.5.** Existence and uniqueness of \(v_p\) can be proven (cf. [Tho06, chapter 12, p. 205]).

The \(v_p \in S_p\) obtained by the dG(p)-method exhibits nodal superconvergence. To present the corresponding result, we need to define the seminorm induced by \(\Delta^s\):

\[
|v|_s : H^s(\Omega) \to \mathbb{R}, \quad |v|_s := \left\| \Delta^{s/2} v \right\|_{L^2(\Omega)}
\]

Furthermore, we define the maximal time step size \(\tau_{\text{max}} := \max_{k=1, \ldots, K} \tau_k\) and denote by \(\partial_t^i v\) the \(i\)-th partial derivative of \(v\) with respect to \(t\). We then get for the \(L^2(\Omega)\) norm of the \(k\)-th nodal error \(v_p^{(k,-)} - v(t^{(k)})\) the following superconvergence result:

**Theorem 2.2** (Nodal superconvergence)

For the weak solution of the heat equation \(v \in L^2(0, T; H^1_0(\Omega))\) and the semi-discrete weak solution \(v_p \in S_p\), there exists a constant \(C > 0\) independent of \(\tau_{\text{max}}\) such that it holds

\[
\left\| v_p^{(k,-)} - v(t^{(k)}) \right\|_{L^2(\Omega)} \leq C \tau_{\text{max}}^{2p+1} \sqrt{\int_0^{t^{(k)}} \left| \partial_t^{p+1} v \right|^2_{L^{2p+1}} dt} \quad \forall \ k = 1, \ldots, K.
\]
Proof. A proof can be found in [Tho06, theorem 12.3].

We also want to mention an important property of the dG(p)-method concerning stability:

Lemma 2.3 (L-stability of the dG(p)-method)

Let \( v_p \) denote the discrete weak solution after a single timestep \( \tau_1 = \tau \) of the dG(p)-method applied to the model problem: For given \( \lambda, u_0 \in \mathbb{C} \), find a \( u : [0, T] \to \mathbb{C} \) that satisfies

\[
\begin{align*}
u' &= \lambda u & \text{for } t \in (0, T), \\
u &= u_0 & \text{for } t = 0.
\end{align*}
\]

Then, it holds

\[
\lim_{\text{Re}(\lambda) \to -\infty} \frac{|v_p(1, -)|}{|u_0|} = 0 \quad \forall \lambda, u_0 \in \mathbb{C},
\]

i.e. the dG(p)-method is L-stable.

Proof. A proof can be found in [MS11, lemma 5].

2.2.3 The fully-discrete weak problem

To obtain a fully-discrete problem that we can actually solve on a computer, we also have to discretize in space. This will be done as already discussed in Section 2.1. Reusing thus the nested family of finite element spaces \( V_1, \ldots, V_L \subset H^1_0(\Omega) \), we get the discrete weak solution space:

\[
S_{p,\ell} := \{ X : (0, T] \to V_\ell \mid X_{I[(i)]} \in \Pi_p [V_\ell] \}
\]

In complete analogy to the previous section, we get

Problem 2.8 (Fully-discrete weak form of the heat equation):

Given a \( g \in L^2 (0, T; L^2(\Omega)) \), find the discrete weak solution \( v_{p,\ell} \in S_{p,\ell} \) that satisfies

\[
\begin{align*}
\text{LHS}^{(k-1)} (v_{p,\ell}, w_{p,\ell}) &= \text{RHS}^{(k-1)} (v_{p,\ell}, w_{p,\ell}) & \forall w_{p,\ell} \in S_{p,\ell}, k = 1, \ldots, K, \\
\langle v_{p,\ell}(0, -), w_{\ell} \rangle_{L^2(\Omega)} &= \langle u_0, w_{\ell} \rangle_{L^2(\Omega)} & \forall w_{\ell} \in V_\ell.
\end{align*}
\]

Remark 2.6. If the spatial component of the weak solution is contained in \( V_\ell \), we are solving the spatial part exactly. Then, any error in the discrete solution has to stem from the temporal discretization.

2.2.4 The matrix equation

The fully-discrete problem can again be rewritten as a matrix equation. For this we reuse the discrete weak solution space \( V_\ell \subset H^1_0(\Omega) \), the dimension of \( V_\ell \) \( n_\ell := \dim (V_\ell) \) and the family of spatial basis functions \( \phi_{\ell,1}, \ldots, \phi_{\ell,n_\ell} \in V_\ell \). We define two new families of functions, the family of temporal reference basis functions:

\[
\psi_{p,1}, \ldots, \psi_{p,(p+1)} \in \Pi_p [\mathbb{R}]
\]

and the family of temporal reference test functions:

\[
\theta_{p,1}, \ldots, \theta_{p,(p+1)} \in \Pi_p [\mathbb{R}]
\]

Both of these families are defined on the unit interval \( (0, 1] \) and span \( \Pi_p [\mathbb{R}] \). They will be mapped to the \( k \)-th time slab \( I^{(k)} \) by use of the linear transformation:

\[
\zeta^{(k)} : I^{(k)} \to (0, 1], \quad t \mapsto t - t^{(k-1)} \frac{(k-1)}{\tau_k}
\]

For \( j = 1, \ldots, (p + 1) \) this yields the \( j \)-th basis function on \( I^{(k)} \):

\[
\psi_{p,j}^{(k)} : I^{(k)} \to \mathbb{R}, \quad \psi_{p,j}^{(k)} := \psi_{p,j} \circ \zeta^{(k)}
\]
and the $j$-th test function on $I^{(k)}$:

\[ \theta_{p,j}^{(k)} : I^{(k)} \rightarrow \mathbb{R}, \quad \theta_{p,j}^{(k)} := \theta_{p,j} \circ \xi^{(k)} \]  

(2.66)

We can now represent the discrete weak solution $v_{p,\ell} \in S_{p,\ell}$ on $I^{(k)}$ using the $k$-th matrix of solution coefficients:

\[ v_{p,\ell}^{(k)}(t) \in \mathbb{R}^{n_t \times (p+1)}, \quad \left[ v_{p,\ell}^{(k)} \right]_{i,j} := v_{p,\ell,i,j}^{(k)} \]  

(2.67)

or for $j = 1, \ldots, (p+1)$ the $j$-th column of $v_{p,\ell}^{(k)}$:

\[ v_{p,\ell,j}^{(k)}(t) \in \mathbb{R}^{n_t}, \quad \left[ v_{p,\ell,j}^{(k)} \right]_{i,j} := v_{p,\ell,i,j}^{(k)} \]  

(2.68)

Also reusing the finite element isomorphism $P_k : \mathbb{R}^{n_t} \rightarrow V_\ell$ (cf. Eq. (2.29)) we get the representation in terms of basis functions:

\[ v_{p,\ell}^{I^{(k)}} = \sum_{j=1}^{(p+1)} \sum_{i=1}^{n_t} \sum_{j=1}^{(p+1)} \psi_{p,j}^{(k)} P_k v_{p,\ell,j}^{(k)} = \sum_{i=1}^{n_t} \sum_{j=1}^{(p+1)} \psi_{p,\ell,i,j}^{(k)} \phi_{\ell,i} \psi_{p,j}(0) \]  

(2.69)

If we now plug in this representation into the Fully-discrete weak form of the heat equation (cf. problem 2.8) and use as discrete weak test function $w_{p,\ell}$ products of the form $\phi_{\ell,i} \theta_{p,j}^{(k)}$, we can separate the time- and space-dependent terms. After some elementary manipulations we again get a matrix equation. To write it down, we will need the spatial stiffness matrix $A_\ell \in \mathbb{R}^{n_t \times n_t}$ (cf. Eq. (2.30)), the spatial mass matrix $M_\ell \in \mathbb{R}^{n_t \times n_t}$ (cf. Eq. (2.31)) and four new matrices.

The right-hand side of the heat equation will be represented by the right-hand side matrix:

\[ f_{p,\ell}^{(k)} \in \mathbb{R}^{n_t \times (p+1)}, \quad \left[ f_{p,\ell}^{(k)} \right]_{i,j} := \langle v_{p,\ell}^{I^{(k)}}, \phi_{\ell,i} \theta_{p,j}^{(k)} (0) \rangle_{L^2(\Omega)} + \int_{I^{(k)}} \langle g, \phi_{\ell,i} \theta_{p,j}^{(k)} \rangle_{L^2(\Omega)} dt \]  

(2.70)

The first term in the above definition includes the previously computed values from the preceding time slab, or the initial conditions if $k = 1$. The second term will have to be evaluated using an appropriate quadrature formula.

For the left-hand side we define the temporal reference mass matrix:

\[ \mu_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad \left[ \mu_p \right]_{i,j} := \int_0^1 \psi_{p,i} \theta_{p,j} dt \]  

(2.71)

and the temporal reference propagation matrix:

\[ \rho_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad \left[ \rho_p \right]_{i,j} := \int_0^1 \psi_{p,i} \theta_{p,j} dt + \left( \psi_{p,i} \theta_{p,j} (0) \right) \]  

(2.72)

We can now define the global stiffness matrix:

\[ K_{p,\ell}^{(k)} \in \mathbb{R}^{[(p+1)n_t] \times [(p+1)n_t]}, \quad K_{p,\ell}^{(k)} := \tau_k \mu_p \otimes A_\ell + \rho_p \otimes M_\ell \]  

(2.73)

Using the vectorization operator:

\[ \text{vec} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{mn}, \quad \left( \begin{array}{ccc} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{array} \right) \mapsto (a_{1,1}, \ldots, a_{m,1}, \ldots, a_{1,n}, \ldots, a_{m,n})^T \]  

(2.74)

we can finally write down the matrix form of the fully-discrete heat equation:

\[ K_{p,\ell}^{(k)} \text{vec } v_{p,\ell}^{(k)} = \text{vec } f_{p,\ell}^{(k)} \]  

(2.75)
2.2.5 Temporal basis and test functions

We will now briefly discuss properties of the $\psi_{p,1}, \ldots, \psi_{p,(p+1)} \in \Pi_p[\mathbb{R}]$, the $\theta_{p,1}, \ldots, \theta_{p,(p+1)} \in \Pi_p[\mathbb{R}]$ and of the induced characteristic temporal matrix:

$$\mu_p^{-1} \rho_p \in \mathbb{R}^{(p+1) \times (p+1)} \quad (2.76)$$

As can be seen from the definition of $K_{p,t}^{(k)}$ (cf. Eq. (2.73)), the structure of the temporal reference mass matrix $\mu_p \in \mathbb{R}^{(p+1) \times (p+1)}$ and the temporal reference propagation matrix $\rho_p \in \mathbb{R}^{(p+1) \times (p+1)}$ defines the block structure of $K_{p,t}^{(k)}$. This potentially large matrix equation can be solved more cheaply if those two matrices are sparse, and even better if they possess a decoupled, block diagonal structure. As $\mu_p$ and $\rho_p$ depend on the choice of the temporal basis and test functions, we will talk about how to choose these functions to obtain a block diagonal structure, and the tradeoffs that we will have to make for that. It will turn out to be simpler to talk about the coefficients of the temporal reference functions than about the sets of functions directly. For this we first define a polynomial coefficient extractor:

$$a_i(\cdot) : \Pi_\infty[\mathbb{R}] \to \mathbb{R}, \quad \sum_{j=0}^\infty f_j \cdot t^i \mapsto f_i \quad (2.77)$$

with which we can define the matrix of basis function coefficients:

$$\psi_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad [\psi_p]_{i,j} := a_{i-1}(\psi_{p,j}) \quad (2.78)$$

and the matrix of test function coefficients:

$$\theta_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad [\theta_p]_{i,j} := a_{i-1}(\theta_{p,j}) \quad (2.79)$$

i.e. the columns of the above two matrices contain the coefficients of the polynomial (temporal) basis and test functions. We want to note, that so far we have not fixed neither the $\psi_{p,1}, \ldots, \psi_{p,(p+1)}$ nor the $\theta_{p,1}, \ldots, \theta_{p,(p+1)}$. As already mentioned, we can choose them almost completely freely, as long as they span $\Pi_p[\mathbb{R}]$. We want to discuss three distinct pairs of basis and test function families. To distinguish them, we will add another subindex to $\psi_p$, $\theta_p$, $\mu_p$ and $\rho_p$. We will use an asterisk ($[\cdot]_*$) as a placeholder for the different cases. Choosing the matrices $\psi_{p,*}$ and $\theta_{p,*}$ determines the polynomial functions, which in turn determine the matrices $\mu_{p,*}$ and $\rho_{p,*}$. We can find two matrices $W_p, P_p \in \mathbb{R}^{(p+1) \times (p+1)}$ independent of $\psi_{p,*}$ and $\theta_{p,*}$, such that they satisfy

$$\mu_{p,*} = \theta^T_{p,*} W_p \psi_{p,*} \quad \forall \psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1) \times (p+1)}, \quad (2.80)$$

$$\rho_{p,*} = \theta^T_{p,*} P_p \psi_{p,*} \quad \forall \psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1) \times (p+1)} \quad (2.81)$$

By comparing the above equations with the definitions of $\mu_p$ (cf. Eq. (2.71)) and $\rho_p$ (cf. Eq. (2.72)), we can identify $W_p$ and $P_p$ as the reference time integrator:

$$W_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad [W_p]_{i,j} := \int_0^1 t^{i+j-2} \, dt = \frac{1}{i+j-1} \quad (2.82)$$

and the reference time propagator:

$$P_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad [P_p]_{i,j} := \int_0^1 (i-1) t^{i+j-3} \, dt + (0^{i+j-2}) = \begin{cases} 1, & \text{if } i = j = 1 \\ \frac{i-1}{i+j-2}, & \text{else} \end{cases} \quad (2.83)$$

We can now easily show

**Lemma 2.4** (Eigenvalues of the characteristic temporal matrix)

Let $\psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1) \times (p+1)}$ be arbitrary but invertible, $\mu_{p,*} = \theta^T_{p,*} W_p \psi_{p,*}$ and $\rho_{p,*} = \theta^T_{p,*} P_p \psi_{p,*}$. Then $\mu_{p,*}$ and $\rho_{p,*}$ has only eigenvalues with positive real part, which only depend on $p$.

**Proof.** A proof can be found in [BB15, lemma 3.1]. Alternatively we immediately get

$$\mu_{p,*}^{-1} \rho_{p,*} = [\theta^T_{p,*} W_p \psi_{p,*}]^{-1} \theta^T_{p,*} P_p \psi_{p,*} = \psi_{p,*}^{-1} W_p^{-1} P_p \psi_{p,*} \quad (2.84)$$

which shows the independence of the eigenvalues from the polynomial functions. The positivity of the real parts follows from the L-stability of the dG(p)-method (cf. lemma 2.3). If there was an eigenvalue with non-positive real part, the method would not be L-stable. \qed
For the following discussion and also from then on we will assume the following:

\[
\begin{align*}
\mu_p^{-1} \rho_p \text{ is diagonalizable over } \mathbb{C} & \quad \text{for any } p \in \mathbb{N}, \\
\sigma(\mu_p^{-1} \rho_p) = \{\lambda_{p,i} \mid i = 1, \ldots, (p+1)/2\} & \quad \text{with } \Im(\lambda_{p,i}) \neq 0 \text{ for } i = 1, \ldots, (p+1)/2, \\
\Im(\lambda_{p,i}) > 0 & \quad \text{for } i = 1, \ldots, (p+1)/2, \\
\arg(\lambda_{p,i}) < \arg(\lambda_{p,i+1}) & \quad \text{for } i = 1, \ldots, (p+1)/2 - 1.
\end{align*}
\]

(2.85) (2.86) (2.87) (2.88)

**Remark 2.7.** For the temporal discretization order \( p \in \mathbb{N} \) that we consider \( (p \leq 20) \), the characteristic temporal matrix \( \mu_p^{-1} \rho_p \) is indeed diagonalizable over \( \mathbb{C} \). This appears to be not yet proven for any \( p \), but computational tests confirm diagonalizability for all \( p \leq 50 \) (cf. [BB15, assumption 1 and following remarks]).

The second assumption implies that \( p \) is odd and that the eigenvalues of the characteristic temporal matrix \( \mu_p^{-1} \rho_p \) only consist of complex conjugate pairs. While this again holds for all odd \( p \) that we consider, we did not prove it for any odd \( p \). The assumption that we do not have purely real eigenvalues serves mainly to simplify the presentation. Any real eigenvalue will only generate a 1 \( \times \) 1-block on the diagonal, which does not need special treatment.

The third and fourth assumptions are conditions on the ordering of the \( \lambda_{p,i} \).

![Complex eigenvalue pairs](image)

Figure 2.1: The (upper half of the) complex eigenvalues of the characteristic temporal matrix \( \mu_p^{-1} \rho_p \in \mathbb{R}^{(p+1) \times (p+1)} \) for \( p = 0, \ldots, 20 \).

Fig. 2.1 shows eigenvalues of the characteristic temporal matrix \( \mu_p^{-1} \rho_p \) for some \( p \). Each line corresponds to a specific temporal order, starting with a single point at \((1,0)\) for \( p = 0 \) and expanding outwards and to the right with increasing \( p \). The characteristic temporal matrix \( \mu_p^{-1} \rho_p \) always has distinct eigenvalues, and exactly one real eigenvalue if and only if \( p \) is even. As we increase \( p \), the (average) magnitude of the eigenvalues increases, as well as the (maximal) complex arguments of the eigenvalues.

We now want to introduce three different types of bases, which will be referred to extensively in the coming chapters. The different bases differ in the relations that hold for the \( \mu_{p,*} \) and \( \rho_{p,*} \) that they induce. To properly write them out, we need to define a special kind of 2 \( \times \) 2-block matrix.

We define the real 2 \( \times \) 2 matrix with complex eigenvalues:

\[
\gamma(\cdot) : \mathbb{C} \to \mathbb{R}^{2 \times 2}, \quad z \mapsto \begin{pmatrix} \Re(z) & \Im(z) \\ -\Im(z) & \Re(z) \end{pmatrix}
\]

(2.89)
\( \gamma(z) \) has an orthogonal eigenvector basis and eigenvalues \( z \) and \( \overline{z} \). Furthermore we define the \( 2 \times 2 \)-block diagonal matrix with complex eigenvalues:

\[
\Gamma(\cdot) : \mathbb{C}^n \to \mathbb{R}^{2n \times 2n}, \quad \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \mapsto \begin{pmatrix} \gamma(z_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \gamma(z_n) \end{pmatrix}
\] (2.90)

Finally, we define the upper half of the complex eigenvalues of \( \mu_p^{-1} \rho_p \):

\[
\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2} \in \mathbb{C}, \quad 0 < \arg(\lambda_{p,1}) < \cdots < \arg(\lambda_{p,(p+1)/2}) < \frac{\pi}{2}
\] (2.91)

and the normed conjugate square roots of \( \lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2} \):

\[
\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2} \in \mathbb{C}, \quad \tilde{\lambda}_{p,i} := \exp\left(-\frac{i}{2} \arg(\lambda_{p,i})/2\right) \quad \text{for } i = 1, \ldots, (p + 1)/2
\] (2.92)

which are normed complex numbers with half the argument of \( \lambda_{p,i} \).

Remark 2.8. The spectra of the characteristic temporal matrix \( \mu_p^{-1} \rho_p \) and \( \Gamma(\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}) \) coincide.

We will look at the following bases:

**Configuration 2.1** (Orthogonal temporal basis):

We construct the temporal basis and test functions in such a way that they satisfy

\[
\psi_{p, \text{orth}} = \theta_{p, \text{orth}} \quad \mu_{p, \text{orth}} = I_{(p+1)}
\] (2.93)

The temporal basis and test functions coincide and the polynomial basis is orthonormal on the unit interval \((0, 1]\). While the mass matrix is the identity, the propagation matrix will in general be at best block upper triangular. The condition number of the polynomial basis is fixed at one.

**Configuration 2.2** (Bi-orthogonal temporal basis):

We construct the temporal basis and test functions in such a way that they satisfy

\[
\mu_{p, \text{bi}} = I_{(p+1)} \\
\rho_{p, \text{bi}} = \Gamma\left(\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}\right)
\] (2.95)

The temporal basis and test functions are distinct and form a biorthogonal system. The mass matrix is the identity, and the propagation matrix is \( 2 \times 2 \)-block diagonal. The condition number of the polynomial basis grows exponentially with \( p \).

**Configuration 2.3** (Non-orthogonal temporal basis):

We construct the temporal basis and test functions in such a way that they satisfy

\[
\mu_{p, \text{non}} = \Gamma\left(\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2}\right) \\
\rho_{p, \text{non}} = \Gamma\left(\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2}\right) \Gamma\left(\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}\right)
\] (2.97)

The temporal basis and test functions are distinct and do not form a biorthogonal system. Instead, the (block diagonal) mass and propagation matrix are chosen such that the arguments of the eigenvalues of the mass matrix are exactly half of the negative arguments of the eigenvalues of the propagation matrix. The condition number of the polynomial basis grows exponentially with \( p \).

Finally, we provide lemmata concerning the construction of the respective bases:
Lemma 2.5 (Constructing orthogonal bases)
Let $W_p \in \mathbb{R}^{(p+1)\times(p+1)}$ be the reference time integrator as in Eq. (2.82) and
\[
\psi_{p,\text{orth}} = \theta_{p,\text{orth}} = W_p^{-1/2} Q
\] with arbitrary orthogonal $Q \in \mathbb{R}^{(p+1)\times(p+1)}.
\]
Then it holds
\[
\mu_{p,\text{orth}} = I_{(p+1)} \quad \forall \ p \in \mathbb{N}.
\]
Proof. We recall
\[
\mu_{p,*} = \theta_{p,*}^T W_p \psi_{p,*} \quad \forall \ \psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1)\times(p+1)}.
\]
$W_p$ is symmetric, and can be proven to be positive definite. Thus the inverse of its square root exists and is symmetric. Straightforward substitution yields the proposition.

Lemma 2.6 (Constructing bi-orthogonal bases)
Let $\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2} \in \mathbb{C}$ be the upper half of the complex eigenvalues of $\mu_p^{-1} \rho_p$ as in Eq. (2.91), $\Gamma (\cdot) : \mathbb{C}^n \rightarrow \mathbb{R}^{2n \times 2n}$ be the real $2 \times 2$-block diagonal matrix with complex eigenvalues as in Eq. (2.90), $W_p \in \mathbb{R}^{(p+1)\times(p+1)}$ be the reference time integrator as in Eq. (2.82), $P_p \in \mathbb{R}^{(p+1)\times(p+1)}$ be the reference time propagator as in Eq. (2.83) and $V \in \mathbb{R}^{(p+1)\times(p+1)}$ be such that $V W_p^{-1} P_p V^{-1} = \Gamma (\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2})$ holds. Then, if
\[
\psi_{p,bi} = V^{-1}
\]
\[
\theta_{p,bi} = W_p^{-T} \psi_{p,bi}^{-T}
\]
it holds
\[
\mu_{p,bi} = I_{(p+1)} \quad \forall \ p \in \mathbb{N},
\]
\[
\rho_{p,bi} = \Gamma (\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}) \quad \forall \ p \in \mathbb{N}.
\]
Proof. We recall
\[
\mu_{p,*} = \theta_{p,*}^T W_p \psi_{p,*} \quad \forall \ \psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1)\times(p+1)},
\]
\[
\rho_{p,*} = \theta_{p,*}^T P_p \psi_{p,*} \quad \forall \ \psi_{p,*}, \theta_{p,*} \in \mathbb{R}^{(p+1)\times(p+1)}.
\]
The characteristic temporal matrix $\mu_p^{-1} \rho_p$ was assumed to be diagonalizable over $\mathbb{C}$, hence $V \in \mathbb{R}^{(p+1)\times(p+1)}$ exists. Straightforward substitution yields the proposition.

Lemma 2.7 (Constructing non-orthogonal bases)
Let $\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2} \in \mathbb{C}$ be the normed conjugate square roots of $\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}$ as in Eq. (2.92), $\Gamma (\cdot) : \mathbb{C}^n \rightarrow \mathbb{R}^{2n \times 2n}$ be the real $2 \times 2$-block diagonal matrix with complex eigenvalues as in Eq. (2.90), $\psi_{p,bi}$ and $\theta_{p,bi}$ be as in lemma 2.6. Then, if
\[
\psi_{p,non} = \psi_{p,bi}
\]
\[
\theta_{p,non} = \theta_{p,bi} \Gamma (\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2})
\]
it holds
\[
\mu_{p,bi} = \Gamma (\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2}) \quad \forall \ p \in \mathbb{N},
\]
\[
\rho_{p,bi} = \Gamma (\tilde{\lambda}_{p,1}, \ldots, \tilde{\lambda}_{p,(p+1)/2}) \Gamma (\lambda_{p,1}, \ldots, \lambda_{p,(p+1)/2}) \quad \forall \ p \in \mathbb{N}.
\]
Proof. Straightforward substitution yields the proposition.
2.3 Verification of temporal superconvergence

We will now apply the dG(p)-method to a simple model problem. We want to verify the correctness of our implementation by achieving nodal superconvergence as in theorem 2.2.

We first define parameters for the heat equation, such that it has a simple classical solution.

**Configuration 2.4** (Quadratic in space and exponential in time classical solution):
Given a cuboid spatial domain $\Omega \subset \mathbb{R}^d$ and a temporal domain $[0, T] \subset \mathbb{R}$, let initial condition $u_0 : \Omega \to \mathbb{R}$, boundary condition $u_D : (0, T) \times \partial \Omega \to \mathbb{R}$ and source term $f : (0, T) \times \Omega \to \mathbb{R}$ be constructed such that the classical solution $u : (0, T) \times \Omega \to \mathbb{R}$ of the Heat equation (cf. problem 2.1) exists and satisfies

$$u(t,x) = \sum_{i=1}^{d} [x_i] \left(1 - [x_i]\right) \left[1 - \exp(t)\right] \quad \forall (t,x) \in (0, T) \times \Omega. \quad (2.112)$$

We now discretize the above problem. For this we first have to specify the used spatial triangulation:

**Configuration 2.5** (Uniform space triangulation):
Given a cuboid spatial domain $\Omega \subset \mathbb{R}^d$ and a temporal domain $[0, T] \subset \mathbb{R}$, let initial condition $u_0 : \Omega \to \mathbb{R}$, boundary condition $u_D : (0, T) \times \partial \Omega \to \mathbb{R}$ and source term $f : (0, T) \times \Omega \to \mathbb{R}$ be constructed such that the classical solution $u : (0, T) \times \Omega \to \mathbb{R}$ of the Heat equation (cf. problem 2.1) exists and satisfies

$$u(t,x) = \sum_{i=1}^{d} [x_i] \left(1 - [x_i]\right) \left[1 - \exp(t)\right] \quad \forall (t,x) \in (0, T) \times \Omega. \quad (2.112)$$

We now discretize the above problem. For this we first have to specify the used spatial triangulation:

**Configuration 2.6** ($P_q$ finite element spaces):
Given a family of triangulations $\mathcal{T}_1, \ldots, \mathcal{T}_L$ and a spatial discretization order $q \in \mathbb{N}$, we use as finite element spaces

$$V_\ell := \{ \phi \in C_0^q(\Omega) \mid \phi|_{\mathcal{T}_\ell,e} \in \Pi_q \ \forall \ T_{\ell,e} \in \mathcal{T}_\ell \} \quad \text{for } \ell = 1, \ldots, L. \quad (2.113)$$

As basis functions we will use the standard nodal finite element basis functions.

**Remark 2.9.** This family of triangulations satisfies all assumptions made in Section 2.1.1.

Next we specify the used family of spatial finite element spaces

**Configuration 2.7** (Uniform time triangulation):
For $m = 0, \ldots, M$ we partition the temporal domain $[0, T]$ uniformly into $K_m := 2^m$ time slabs and define the $m$-th time step size $\tau_m := 2^{-m} T$.

We will look at the relative $L^2(\Omega)$ norm of the nodal error at the end point:

$$\epsilon_{p,\ell,m} := \frac{\|v(T) - v_{p,\ell,m}(T)\|_{L^2(\Omega)}}{\|v(T)\|_{L^2(\Omega)}} \quad (2.114)$$

and the experimental order of convergence in time:

$$EOCT_{p,\ell,m} := \log \left( \frac{\epsilon_{p,\ell,m}}{\epsilon_{p,\ell,m-1}} \right) \bigg/ \log \left( \frac{\tau_m}{\tau_{m-1}} \right) \quad (2.115)$$

We finally introduce the experiment to confirm nodal superconvergence:

**Experiment 2.1** (Nodal superconvergence):
We solve the Heat equation (cf. problem 2.1) with $\Omega = (0, 1)^3$ and $[0, T] = (0, 1]$ and a Quadratic in space and exponential in time classical solution (cf. cfg. 2.4), using the dG(p)-method with $p = 0, \ldots, 9$, an Orthogonal temporal basis (cf. cfg. 2.1), a Uniform space triangulation (cf. cfg. 2.5) with $\ell = 1, \ldots, 10$, $P_2$ finite element spaces (cf. cfg. 2.6) with $q = 2$ and a Uniform time triangulation (cf. cfg. 2.7) with $m = 0, \ldots, 7$. At each time step we solve the matrix form of the fully-discrete heat equation (cf. Eq. (2.75)) directly. We measure the relative $L^2(\Omega)$ norm of the nodal error at the end point (cf. Eq. (2.114)), shown in Fig. 2.4b, and compute the experimental order of convergence in time (cf. Eq. (2.115)), shown in Fig. 2.2b. Both figures show the respective quantities (y-axis) over the time step size (x-axis) for fixed temporal discretization order (lines) and $\ell = 1$. 

---

2.3. VERIFICATION OF TEMPORAL SUPERCONVERGENCE
which then of course stops as the relative L²(Ω) norm of the nodal error at the end point reaches floating point precision.

2.3.1 High-order problems and remedies

As already hinted at before, we will not be able to obtain arbitrary precision for high p if we use a Bi-orthogonal temporal basis (cf. cfg. 2.2) or a Non-orthogonal temporal basis (cf. cfg. 2.3). This is due to the exponential growth of the condition number of the polynomial bases. We have to specify what this means. We follow [DR08, chapter 2]. Given a linear n-dimensional space V with basis Φ := φ₁, . . . , φₙ we define the coordinate mapping:

\[ \mathcal{L} : \mathbb{R}^n \rightarrow V, \quad \mathbf{a} \mapsto \sum_{i=1}^{n} [\mathbf{a}]_i \phi_i \]  

(2.116)

and the condition number of Φ with respect to the norms \( \| \cdot \|_X \) and \( \| \cdot \|_V \):

\[ \kappa_{X,V}(\Phi) := \min \{ C/|c| |c| \|\mathbf{a}\|_X \leq \|\mathcal{L}\mathbf{a}\|_V \leq C \|\mathbf{a}\|_X \quad \forall \mathbf{a} \in \mathbb{R}^n \} \]  

(2.117)

In our setting we have as V the space of polynomials, as \( \| \cdot \|_V \)-norm the \( L^2([0, 1]) \)-norm, as \( \mathbf{a} \) one row of the k-th matrix of solution coefficients \( \mathbf{v}_{p,k} \in \mathbb{R}^{n_2 \times (p+1)} \) (cf. Eq. (2.67)) and as \( \| \cdot \|_X \)-norm the standard Euclidean norm. Reusing the matrix of basis function coefficients \( \mathbf{v}_p \in \mathbb{R}^{(p+1) \times (p+1)} \) (cf. Eq. (2.78)) and the reference time integrator \( \mathbf{W}_p \in \mathbb{R}^{(p+1) \times (p+1)} \) (cf. Eq. (2.82)), we can observe that the equality

\[ \| \mathcal{L}\mathbf{a} \|_{L^2([0, 1])} = \left\| \mathbf{W}_p^{1/2} \psi_p \mathbf{a} \right\| \quad \forall \mathbf{a} \in \mathbb{R}^{(p+1)} \]

holds and thus

\[ \kappa_{2,L^2([0, 1])}(\psi_{p,1}, \ldots, \psi_{p,(p+1)}) = \kappa_2 \left( \mathbf{W}_p^{1/2} \psi_p \right) \quad \forall p \in \mathbb{N}. \]  

(2.118)

With the above equation we can now compute the condition numbers of the Bi-orthogonal temporal basis (cf. cfg. 2.2), shown in Fig. 2.3 over the temporal discretization order \( p \in \mathbb{N} \). We can observe the aforementioned exponential growth of the condition numbers.

Remark 2.11. From the definition of the condition number we can conclude that if we solve the matrix form of the fully-discrete heat equation (cf. Eq. (2.75)) with a relative precision \( \varepsilon \), we can not expect to get a relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)) that is much smaller than \( \kappa_{2,L^2([0, 1])}(\psi_{p,1}, \ldots, \psi_{p,(p+1)}) \varepsilon \).

2.3. VERIFICATION OF TEMPORAL SUPERCONVERGENCE
There is another source of numerical error. To talk about it, we first have to define an \( n \)-point quadrature scheme:

\[
\mathcal{Q}_{n,[a,b]} : \mathbb{I} \rightarrow \mathbb{R} \rightarrow \mathbb{R}, \quad f \mapsto \sum_{i=1}^{n} w_i f(t_i) \approx \int_{[a,b]} f(t) \, dt
\]

which can be identified with its quadrature weights \( w_1, \ldots, w_n \in \mathbb{R} \) and its quadrature points \( t_1, \ldots, t_n \in [a, b] \). We reuse the family of temporal reference test functions \( \theta_{p,1}, \ldots, \theta_{p,p+1} \in \Pi_p[\mathbb{R}] \) (cf. Eq. (2.63)) and apply a \((p+1)\)-point quadrature scheme \( \mathcal{Q}_{(p+1),[a,b]} : [a,b] \rightarrow \mathbb{R} \rightarrow \mathbb{R} \) to approximate for a given \( f : [a,b] \rightarrow \mathbb{R} \) the exact right-hand side vector:

\[
\mathbf{Q} \in \mathbb{R}^{(p+1)}, \quad [\mathbf{Q}]_i := \int_{[a,b]} f(\theta_{p,i}) \, dt
\]

by an approximated right-hand side vector:

\[
\tilde{\mathbf{Q}} \in \mathbb{R}^{(p+1)}, \quad [\tilde{\mathbf{Q}}]_i := \mathcal{Q}_{(p+1),[a,b]}(f(\theta_{p,i}))
\]

We can recast the numerical quadrature as a matrix-vector product, using the vector:

\[
\mathbf{f} \in \mathbb{R}^{(p+1)}, \quad [\mathbf{f}]_i := f(t_i)
\]

and the weighted Vandermonde matrix:

\[
\mathbf{V}_p \in \mathbb{R}^{(p+1) \times (p+1)}, \quad [\mathbf{V}_p]_{i,j} := t_i^j w_j
\]

Reusing the matrix of test function coefficients \( \theta_p \in \mathbb{R}^{(p+1) \times (p+1)} \) (cf. Eq. (2.79)) we now get

\[
\tilde{\mathbf{Q}} = \theta_p \mathbf{V}_p \mathbf{f}
\]

If we compute the approximated right-hand side vector \( \tilde{\mathbf{Q}} \in \mathbb{R}^{(p+1)} \) at runtime with finite precision, the error that we introduce depends on the spectral condition numbers of the matrix of test function coefficients \( \theta_p \) and the weighted Vandermonde matrix \( \mathbf{V}_p \) and of the quadrature matrix:

\[
\theta_p \mathbf{V}_p \in \mathbb{R}^{(p+1) \times (p+1)}
\]

While we can precompute this product, thus at least removing one possible source of error, we will not get around performing at least one matrix-vector product with the quadrature matrix \( \theta_p \mathbf{V}_p \). If this matrix is badly conditioned, we may again introduce a considerable error. We plot the condition numbers of the quadrature matrix \( \theta_p \mathbf{V}_p \) using Gauss quadrature and of the temporal reference propagation matrix \( \mathbf{Q}_p \in \mathbb{R}^{(p+1) \times (p+1)} \) (cf. Eq. (2.72)) in Fig. 2.3 alongside the condition numbers of the polynomial bases.

We now confirm the adverse effect of the increased condition numbers of the quadrature matrices and of the polynomial bases on the obtainable relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)):

**Experiment 2.2** (Condition numbers): We repeat Experiment 2.1, once using a quadrature matrix \( \theta_p \mathbf{V}_p \) evaluated at runtime (Fig. 2.4a), and once using a precompiled quadrature matrix \( \theta_p \mathbf{V}_p \) (Fig. 2.4b). We again measure the relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)), shown in Fig. 2.4.

**Results.** We observe an improvement in the best obtainable relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)) if we use a precompiled quadrature matrix \( \theta_p \mathbf{V}_p \) instead of a dynamic quadrature matrix \( \theta_p \mathbf{V}_p \).

**Experiment 2.3** (Condition numbers): We repeat Experiment 2.2, but using a Bi-orthogonal temporal basis (cf. cfg. 2.2), again measuring the relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)), shown in Fig. 2.5.

**Results.** We observe a very pronounced worsening of the best obtainable relative \( L^2(\Omega) \) norm of the nodal error at the end point (cf. Eq. (2.114)) for the dynamic quadrature matrix \( \theta_p \mathbf{V}_p \) as the temporal discretization order \( p \in \mathbb{N} \) increases. This effect is weakened if we use a precompiled quadrature matrix \( \theta_p \mathbf{V}_p \), but still slightly more marked than for the orthogonal bases (Fig. 2.4b).
CHAPTER 2. VARIATIONAL TIME DISCRETIZATION OF THE HEAT EQUATION

Figure 2.3: Condition numbers of important matrices i.e. of the quadrature matrix $\theta_p V_p$, the temporal reference propagation matrix $\rho_p$ and of the polynomial bases for a Bi-orthogonal temporal basis (cf. cfg. 2.2) (dashed) and for a Orthogonal temporal basis (cf. cfg. 2.1) (solid)

(a) Dynamic quadrature matrix $\theta_p V_p$

(b) Precompiled quadrature matrix $\theta_p V_p$

Figure 2.4: Dynamic vs precompiled quadrature matrices for orthogonal temporal bases

(a) Dynamic quadrature matrix $\theta_p V_p$

(b) Precompiled quadrature matrix $\theta_p V_p$

Figure 2.5: Dynamic vs precompiled quadrature matrices for nonorthogonal temporal bases

2.3. VERIFICATION OF TEMPORAL SUPERCONVERGENCE
3 Iterative solvers

In this chapter we give a concise overview of the iterative solvers that will be used in the following chapters. We start by introducing some general notation. A short section on stationary iterative methods follows, with a focus on geometric multigrid methods and some standard convergence results for symmetric matrices are presented. In later chapters, these multigrid methods will be used in block-preconditioners for two Krylov subspace methods, the left preconditioned conjugate gradient method and the left preconditioned generalized conjugate residual method. Consequently, the section on stationary iterative methods is followed by a brief introduction to these two Krylov subspace methods, again complete with convergence results.

Given a governing matrix $A \in \mathbb{R}^{n \times n}$ and a right-hand side vector $b \in \mathbb{R}^n$, we are looking for the solution vector $x \in \mathbb{R}^n$ which solves the linear system of equations $Ax = b$. To distinguish different iterative methods, we will again use a subscript with the asterisk (*) as a placeholder.

We will approximate the $x$ iteratively by a sequence of approximations $x^*(0), \ldots, x^*(K) \in \mathbb{R}^n$. For $k = 0, \ldots, K$ we define the $k$-th error vector:

$$e^*(k) \in \mathbb{R}^n,$$

$$e^*(k) := x - x^*(k) \quad (3.1)$$

the $k$-th residual vector:

$$r^*(k) \in \mathbb{R}^n,$$

$$r^*(k) := b - A x^*(k) = A e^*(k) \quad (3.2)$$

and the $k$-th relative error norm:

$$\gamma^*(k) := \frac{\|e^*(k)\|}{\|e^*(0)\|}. \quad (3.3)$$

Unless otherwise stated, we will denote by $\|\cdot\|$ the standard Euclidean norm. An iterative method converges if $\gamma^*(k)$ tends to zero as $k$ tends to infinity. We can quantify its convergence speed using the $k$-th convergence rate:

$$\Gamma^*(k) := \left[ \frac{\gamma^*(k)}{\gamma^*(0)} \right]^{1/k} \quad (3.4)$$

and the asymptotic convergence rate:

$$\Gamma_* := \lim_{k \to \infty} \Gamma^*(k). \quad (3.5)$$

Using the initial guess $x^*(0) \in \mathbb{R}^n$ we can identify the general form of an iterative method:

$$S_* : \mathbb{R}^{n \times n} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \{\mathbb{R}^n\}_H,$$

$$(A, b, x^*(0)) \mapsto x^*(0), \ldots, x^*(K) \quad (3.6)$$

which is nothing more than a mapping from the tuple of a governing matrix $A \in \mathbb{R}^{n \times n}$, a right-hand side vector $b \in \mathbb{R}^n$ and an initial guess $x^*(0) \in \mathbb{R}^n$ to a sequence of approximations $x^*(0), \ldots, x^*(K) \in \mathbb{R}^n$.

3.1 Stationary iterative methods

Stationary iterative methods are iterative methods, where the sequence of approximations $x^*(0), \ldots, x^*(K) \in \mathbb{R}^n$ follows a simple approximation recurrence relation:

$$x^*(k+1) = x^*(k) + M_*^{-1} r^*(k) \quad (3.7)$$

using an efficiently applicable, fixed approximation matrix:

$$M_* \in \mathbb{R}^{n \times n}, \quad M_* \approx A \quad (3.8)$$
which is an approximation of $A \in \mathbb{R}^{n \times n}$. The approximation recurrence relation (cf. Eq. (3.7)) gives rise to the iteration matrix:

$$S_* \in \mathbb{R}^{n \times n}, \quad S_* := I_n - M_*^{-1} A$$

(3.9)

which in turn describes the evolution of the $k$-th error vector $e_*^{(k)} \in \mathbb{R}^n$ using the error recurrence relation:

$$e_*^{(k+1)} = S_* e_*^{(k)}.$$  

(3.10)

We can identify any stationary iterative method $S_*$ with its approximation matrix $M_*$ or with its iteration matrix $S_*$. For basic methods we will provide an expression for approximation matrix $M_*$, while we can more naturally define composite methods using its iteration matrix $S_*$. From the error recurrence relation (cf. Eq. (3.10)) we can conclude that a stationary iterative method converges if the spectral radius:

$$\rho ( A ) := \max \{|\lambda| \mid \lambda \in \sigma ( A )\}$$  

(3.11)

of its iteration matrix $\rho ( S_* )$ is smaller than one (cf. [Saa03, theorem 4.1]).

3.1.1 Methods based on matrix splittings

The most basic stationary iterative methods make use of a matrix splitting:

$$A = A_D - A_L - A_U$$

(3.12)

which decomposes a given $A \in \mathbb{R}^{n \times n}$ into a diagonal matrix $A_D \in \mathbb{R}^{n \times n}$, a strictly lower triangular matrix $A_L \in \mathbb{R}^{n \times n}$ and an strictly upper triangular matrix $A_U \in \mathbb{R}^{n \times n}$. Using this splitting, we can define the Jacobi method $S_J$:

$$M_J := A_D$$

(3.13)

the forward Gauss-Seidel method $S_{fG}$:

$$M_{fG} := A_D - A_L$$

(3.14)

the backward Gauss-Seidel method $S_{bG}$:

$$M_{bG} := A_D - A_U$$

(3.15)

and the symmetric Gauss-Seidel method $S_{sG}$:

$$S_{sG} := S_{bG} S_{fG}$$

(3.16)

where one iteration consists of one application of the forward Gauss-Seidel method $S_{fG}$ followed by one application of the backward Gauss-Seidel method $S_{bG}$. It can be shown that the Jacobi method $S_J$, the forward Gauss-Seidel method $S_{fG}$ and the backward Gauss-Seidel method $S_{bG}$ converge for any initial guess $x_*^{(0)}$ if the governing matrix $A$ is strictly diagonally dominant (cf. [Saa03, theorem 4.9]). The symmetric Gauss-Seidel method $S_{sG}$ can be shown to converge for any initial guess $x_*^{(0)}$ if governing matrix $A$ is symmetric positive definite (cf. [Saa03, theorem 4.10]). While these assumptions hold for some of the matrices that we will look at, we will not use the above methods as standalone solvers but only as smoothers for a multigrid method. For the convergence of the multigrid method, convergence of the used smoothers is not necessary (cf. [Hac94, remark 10.6.6]).

Common to the above methods applied to matrices that typically arise during the finite element discretization of elliptic problems is a worsening of the $k$-th convergence rate $\Gamma_*^{(k)}$ as the spatial step size $h$ decreases. This behaviour is due to low-frequency errors not getting damped well enough, while high-frequency errors do experience a significant reduction.

As such, while the above methods are not very useful as standalone solvers, they are an integral part of multigrid methods, where they are used as smoothers, with the explicit purpose of mainly handling high-frequency errors.
3.1.2 The multigrid method

Multigrid methods are a special type of stationary iterative methods. In contrast to methods based on matrix splittings, the contraction rate is usually bounded by a constant (significantly) less than one, independent of the spatial stepsize of the underlying finite element discretization.

A wealth of literature exists covering scalar multigrid methods, see e.g. [Hac13, Wes95, Reu08]. We only present the main ideas behind geometric multigrid methods and their convergence analysis, as they are developed in [Reu08].

Geometric multigrid methods are intimately connected to the underlying (finite element) discretization of the (elliptic) problem. Given a nested family of finite element spaces \( V_1, \ldots, V_L \subset H^1_0(\Omega) \) (cf. ??) and for \( \ell = 1, \ldots, L \) the finite element isomorphism \( P_\ell : \mathbb{R}^{n_\ell} \rightarrow V_\ell \) (cf. Eq. (2.29)) and the global stiffness matrix \( K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \) (cf. Eq. (2.32)), we first define for \( \ell = 2, \ldots, L \) the canonical prolongation operator:

\[
P_\ell : \mathbb{R}^{n_{\ell-1}} \rightarrow \mathbb{R}^{n_\ell}, \quad x \mapsto P_\ell^{-1} P_{\ell-1} x
\]

and the canonical restriction operator:

\[
R_\ell : \mathbb{R}^{n_\ell} \rightarrow \mathbb{R}^{n_{\ell-1}}, \quad x \mapsto P_\ell^T x.
\]

These two operators fulfill the Galerkin condition:

\[
R_\ell K_\ell P_\ell = K_{\ell-1}
\]

and provide a means to either use a coarse discrete weak solution \( v_\ell \in V_\ell \) (cf. ??) as an approximation to a finer solution \( v_{\ell+1} \in V_{\ell+1} \) or to reduce the dimensionality of the finer problem. We use the canonical prolongation operator \( P_\ell \) and the canonical restriction operator \( R_\ell \) to define for \( \ell = 2, \ldots, L \) the iteration matrix of the coarse grid correction \( S_{\text{coarse},\ell} \):

\[
S_{\text{coarse},\ell} \in \mathbb{R}^{n_{\ell} \times n_{\ell}}, \quad S_{\text{coarse},\ell} := I_{n_\ell} - P_\ell M_{\text{pre},\ell-1}^{-1} K_{\ell-1} R_\ell
\]

where \( M_{\text{pre},\ell-1}^{-1} \) is the approximation matrix of a not yet specified solution method \( S_{\text{pre}} \) on the coarse grid. For the coarsest problem \( \ell = 1 \) we now define \( S_{\text{pre}} \) to solve the linear system of equations \( K_1 x[1] = b[1] \) exactly, i.e. \( M_{\text{pre},1}^{-1} = K_1 \) holds. For the rest of the problems \( \ell = 2, \ldots, L \) we need a given pre-smoothing method \( S_{\text{pre}} \), a post-smoothing method \( S_{\text{post}} \) and as parameters the pre-smoothing number \( \nu_1 \in \mathbb{N} \), the post-smoothing number \( \nu_2 \in \mathbb{N} \) and the multigrid cycle parameter \( \tau \in \mathbb{N} \). We can now define the iteration matrix of the multigrid method \( S_{\text{MG},\ell} \):

\[
S_{\text{MG},\ell} \in \mathbb{R}^{n_{\ell} \times n_{\ell}}, \quad S_{\text{MG},\ell} := S_{\text{post},\ell} S_{\text{pre},\ell} S_{\text{coarse},\ell}^{\nu_2} S_{\text{pre},\ell}^{\nu_1}
\]

which reveals \( S_{\text{pre}} \) to consist of \( \nu_1 \) applications of \( S_{\text{pre}} \), followed by \( \tau \) applications of \( S_{\text{coarse}} \) and concluded by \( \nu_2 \) applications of \( S_{\text{post}} \). We call a multigrid method with \( \tau = 1 \) a \( V \)-cycle and a multigrid method with \( \tau = 2 \) a \( W \)-cycle.

Remark 3.1. The \( M_{\text{MG},\ell} \) are symmetric if for \( \ell = 1, \ldots, L \) the \( K_\ell \) is symmetric, if it holds \( M_{\text{pre},\ell} = M_{\text{post},\ell}^T \) and if it holds \( R_\ell = P_\ell^T \). If furthermore the \( K_\ell \) are positive definite, the \( M_{\text{MG},\ell} \) are so as well in our setting (cf. [SBB12, theorem 3.6]).

The smoothing and approximation property

In the convergence analysis of a multigrid method \( S_{\text{MG}} \), one traditionally starts with a splitting of the norm of the iteration matrix of the two-grid method \( S_{\text{TG}} \) without post-smoothing (\( \nu_2 = 0 \)). The two-grid method \( S_{\text{TG}} \) is a multigrid method \( S_{\text{MG}} \) with \( L = 2 \). We then get the splitting of the norm of \( S_{\text{TG},\ell} \):

\[
\|S_{\text{TG},\ell}\| \leq \|K_\ell^{-1} - P_\ell K_{\ell-1}^{-1} R_\ell\| \|K_\ell S_{\text{pre},\ell}^{\nu_1}\|.
\]

The next step is proving for \( \ell = 1, \ldots, L \) the approximation property:

\[
\|K_\ell^{-1} - P_\ell K_{\ell-1}^{-1} R_\ell\| \leq C_A \|K_\ell\|^{-1} w
\]

with a constant \( C_A > 0 \) independent of \( \ell \) and proving the smoothing property:

\[
\|K_\ell S_{\text{pre},\ell}^{\nu_1}\| \leq \eta(\nu_1) \|K_\ell\|
\]
with monotonically decreasing function \( \eta(\nu_1) : \mathbb{N} \to \mathbb{R} \) and with \( \lim_{\nu_1 \to \infty} \eta(\nu_1) = 0 \).

From the above two properties it follows that there exists a \( \nu_1 \), such that the spectral radius \( \rho(S_{TG,\ell}) \) of the iteration matrix of the two-grid method \( S_{TG} \) is less than one, yielding convergence of the two-grid method \( S_{TG} \) for any initial guess \( x_0 \). The smaller the constant \( C_A \) and the function \( \eta(\nu_1) \), the faster the convergence.

For the smoothing property of the symmetric Gauss-Seidel method \( S_{sGS} \) applied to the global stiffness matrix \( K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \) of an elliptic problem, we introduce a theorem as presented in [Reu08].

**Theorem 3.1** (The smoothing property of the symmetric Gauss-Seidel method)
Let \( K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \) be the global stiffness matrix as in Eq. (2.32) and \( S_{post} = S_{sGS} \). Then the smoothing property (cf. Eq. (3.24)) holds with \( \eta(\nu_1) = \frac{c}{\nu_1 + 1} \) and a constant \( c \) independent of \( \nu_1 \) and \( \ell \) for all \( \nu_1 \in \mathbb{N} \).

**Proof.** A proof can be found in [Reu08, theorem 7.11].

For the approximation property, we also introduce a theorem as presented in [Reu08].

**Theorem 3.2** (The approximation property)
Let \( K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \) be the global stiffness matrix as in Eq. (2.32), \( P_\ell : \mathbb{R}^{n_\ell-1} \to \mathbb{R}^{n_\ell} \) be the canonical prolongation operator as in Eq. (3.17) and \( R_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell-1} \) be the canonical restriction operator as in Eq. (3.18). If the associated weak problem and its dual problem are \( H^2 \)-regular and if we have a typical finite element discretization error bound for the Galerkin discretization, the approximation property (cf. Eq. (3.23)) holds for all \( \ell = 1, \ldots, L \) with a constant \( C_A > 0 \) independent of \( \ell \).

**Proof.** A proof can be found in [Reu08, theorem 7.6]. The \( H^2 \)-regularity is not necessary but greatly simplifies the presentation of the proof, cf. [Reu08, remark 7.23].

Finally, we provide a result concerning the contraction number of the V-cycle multigrid method, again taken from [Reu08].

**Theorem 3.3** (The contraction number of the multigrid V-cycle)
Let \( K_\ell \in \mathbb{R}^{n_\ell \times n_\ell} \) be the global stiffness matrix as in Eq. (2.32), \( S_{MG,\ell} \in \mathbb{R}^{n_\ell \times n_\ell} \) be the iteration matrix of the multigrid method \( S_{MG} \) as in Eq. (3.21) with \( \nu_1 = \nu_2 \geq 1 \) and \( \tau = 1 \), \( S_{pre} = S_{post} = S_{sGS} \) and let \( S_{MG,\ell} \) be its iteration matrix. Then
\[
\|S_{MG,\ell}\|_{K_\ell} \leq c \leq 1
\]

holds for a constant \( c \) independent of \( \ell \).

**Proof.** A proof can be found in [Reu08, corollary 7.32].

**Remark 3.2.** While theorem 3.3 only provides a qualitative bound on the contraction number with respect to the \( K_\ell \)-norm of the multigrid V-cycle, the contraction numbers with respect to the spectral norm appear to also be bounded by a constant significantly less than one.

### 3.1.3 The block multigrid method

The block multigrid method is just the standard multigrid method, but applied to a matrix of the form
\[
K = \mu \otimes A + \rho \otimes M \in \mathbb{R}^{mn \times mn}
\]
with matrices \( \mu, \rho \in \mathbb{R}^{n \times n} \) and \( A, M \in \mathbb{R}^{n \times n} \). It exploits the tensor structure to update various temporal unknowns corresponding to the same spatial node at once. Its prolongation and restriction operators are just the operators of the standard multigrid method, but applied to all time unknowns at once, we get
\[
R_{\ell}^{(BMG)} := I \otimes R_\ell
\]
and
\[
P_{\ell}^{(BMG)} := I \otimes P_\ell.
\]
The smoothers are similarly extended. Let \( N \in \mathbb{R}^{n \times n} \) denote the approximation matrix of the smoother of the standard multigrid method. The smoother of the block multigrid method is then defined as
\[
N^{(BMG)} := I \otimes N.
\]
The block multigrid exhibits similar convergence properties as the standard multigrid method, albeit usually slightly worse.

The cost of one iteration however is roughly \( m^2/n \) times as high as the cost of one iteration of the standard multigrid method, due to the increased dimensionalty.

We will not provide an exact bound on the convergence rate of the block multigrid method. Instead, we will compare its performance with the standard methods numerically in the last chapter.

### 3.2 Krylov subspace methods

Krylov subspace methods are projection methods, i.e., iterative methods for which the \( k \)-th approximation \( \mathbf{x}_s^{(k)} \in \mathbb{R}^n \) lies in an affine subspace of dimension \( k \):

\[
\mathbf{x}_s^{(0)} + \mathcal{K}^{(k)} \subseteq \mathbb{R}^n
\]  

(3.26)

and the \( k \)-th residual vector \( \mathbf{r}_s^{(k)} \in \mathbb{R}^n \) (cf. Eq. (3.2)) satisfies a Petrov-Galerkin condition:

\[
\mathbf{r}_s^{(k)} \perp \mathcal{L}^{(k)}
\]

(3.27)

with another subspace of dimension \( k \) \( \mathcal{L}^{(k)} \subseteq \mathbb{R}^n \). A projection method is a Krylov subspace method if \( \mathcal{K}^{(k)} \) is a \( k \)-th Krylov subspace:

\[
\mathcal{K}^{(k)} \left( \mathbf{A}, \mathbf{r}_s^{(0)} \right) := \text{span} \left\{ \mathbf{r}_s^{(0)}, \mathbf{A} \mathbf{r}_s^{(0)}, \ldots, \mathbf{A}^{k-1} \mathbf{r}_s^{(0)} \right\}.
\]

(3.28)

While stationary iterative methods can be identified with their approximation matrix \( \mathbf{M}_s \in \mathbb{R}^{n \times n} \) (cf. Eq. (3.8)), Krylov subspace methods can be identified with their choice of \( \mathcal{L}^{(k)} \) combined with their method of constructing the \( k \)-th approximation \( \mathbf{x}_s^{(k)} \in \mathbb{R}^n \).

The sequence of approximations \( \mathbf{x}_s^{(0)}, \ldots, \mathbf{x}_s^{(K)} \in \mathbb{R}^n \) again follows a simple approximation recurrence relation:

\[
\mathbf{x}_s^{(k+1)} = \mathbf{x}_s^{(k)} + \alpha_s^{(k)} \mathbf{p}_s^{(k)}
\]

(3.29)

using the \( k \)-th basis coefficient \( \alpha_s^{(k)} \in \mathbb{R} \) and the \( k \)-th basis vector \( \mathbf{p}_s^{(k)} \in \mathbb{R}^n \), which are defined by a method-dependent recurrence relation. Common to all Krylov subspace methods is that for all \( k = 0, \ldots, K \), the basis vectors \( \mathbf{p}_s^{(1)}, \ldots, \mathbf{p}_s^{(k)} \in \mathbb{R}^n \) have to span the \( k \)-th Krylov subspace \( \mathcal{K}^{(k)}(\mathbf{A}, \mathbf{r}_s^{(0)}) \). Usually the basis vectors are also orthogonal with respect to a method-dependent inner product.

The two most common choices for \( \mathcal{L}^{(k)} \) are \( \mathcal{L}^{(k)} = \mathcal{K}^{(k)}(\mathbf{A}, \mathbf{r}_s^{(0)}) \) and \( \mathcal{L}^{(k)} = \mathbf{A} \mathcal{K}^{(k)}(\mathbf{A}, \mathbf{r}_s^{(0)}) \). The convergence behaviour of these types of methods can most easily be understood if we look at them as orthogonal or oblique projection methods (cf. [Saa03, chapter 5]). We present two lemmata concerning the behaviour of these types of projection methods, taken from [Saa03, section 5.2.1]. In this context the spaces \( \mathcal{K}^{(k)} \) and \( \mathcal{L}^{(k)} \) may be Krylov subspaces, but they do not have to be.

**Lemma 3.4 (Optimality of the error for orthogonal projection methods)**

Assume that \( \mathbf{A} \in \mathbb{R}^{n \times n} \) is symmetric positive definite and \( \mathcal{L}^{(k)} = \mathcal{K}^{(k)} \). Then the \( k \)-th approximation \( \mathbf{x}_s^{(k)} \in \mathbb{R}^n \) is the result of an orthogonal projection method onto \( \mathcal{K}^{(k)} \) with the initial guess \( \mathbf{x}_s^{(0)} \in \mathbb{R}^n \) if and only if it minimizes the \( \mathbf{A} \)-norm of the \( k \)-th error vector \( \mathbf{e}_s^{(k)} \in \mathbb{R}^n \) over \( \mathbf{x}_s^{(0)} + \mathcal{K}^{(k)} \), i.e., if and only if

\[
\left\| \mathbf{x}_s^{(k)} - \mathbf{x} \right\|_A = \min_{\mathbf{v} \in \mathbf{x}_s^{(0)} + \mathcal{K}^{(k)}} \| \mathbf{v} - \mathbf{x} \|_A
\]

(3.30)

holds.

**Proof.** A proof can be found in [Saa03, section 5.2.1].

**Lemma 3.5 (Optimality of the residual for oblique projection methods)**

Assume that \( \mathbf{A} \in \mathbb{R}^{n \times n} \) is square and \( \mathcal{L}^{(k)} = \mathbf{A} \mathcal{K}^{(k)} \). Then the \( k \)-th approximation \( \mathbf{x}_s^{(k)} \in \mathbb{R}^n \) is the result of an oblique projection method onto \( \mathcal{K}^{(k)} \) orthogonally to \( \mathcal{L}^{(k)} \) with the initial guess \( \mathbf{x}_s^{(0)} \in \mathbb{R}^n \) if and only if it minimizes the Euclidean norm of the \( k \)-th residual vector \( \mathbf{r}_s^{(k)} \in \mathbb{R}^n \) over \( \mathbf{x}_s^{(0)} + \mathcal{K}^{(k)} \), i.e., if and only if

\[
\left\| \mathbf{b} - \mathbf{A} \mathbf{x}_s^{(k)} \right\| = \min_{\mathbf{v} \in \mathbf{x}_s^{(0)} + \mathcal{K}^{(k)}} \| \mathbf{b} - \mathbf{A} \mathbf{v} \|_2
\]

(3.31)

holds.
Proof. A proof can be found in [Saa03, section 5.2.1].

It will turn out to be convenient to define the set of real polynomial functions of degree at most \( k \) that have value one at the origin:

\[
\mathcal{Q}^k := \{ p \in \mathbb{R}_k [x] \mid p(0) = 1 \}. \tag{3.32}
\]

If now \( \mathcal{L}^{(k)} = \mathcal{K}^{(k)} = \mathcal{K}^{(k)} (A, r_s^{(0)}) \) and if \( A \in \mathbb{R}^{n \times n} \) is symmetric positive definite, from lemma 3.4 we get for the \( A \)-norm of the \( k \)-th error vector \( e_s^{(k)} \in \mathbb{R}^n \):

\[
\| e_s^{(k)} \|_A = \min_{q_k \in \mathcal{Q}^n} \| q_k (A) e_s^{(0)} \|_A \quad \text{for } k = 0, \ldots, K. \tag{3.33}
\]

Similarly if \( \mathcal{L}^{(k)} = A \mathcal{K}^{(k)} = A \mathcal{K}^{(k)} (A, r_s^{(0)}) \) and if \( A \in \mathbb{R}^{n \times n} \) is square, from lemma 3.5 we get for the Euclidean norm of the \( k \)-th residual vector \( r_s^{(k)} \in \mathbb{R}^n \):

\[
\| r_s^{(k)} \| = \min_{q_k \in \mathcal{Q}^n} \| q_k (A) r_s^{(0)} \|_2 \quad \text{for } k = 0, \ldots, K. \tag{3.34}
\]

To evaluate the above two equations for diagonalizable matrices we define a shorthand for the minimal \( L^\infty (\sigma) \)-norm over \( \mathcal{Q}^k \):

\[
\mathcal{M}_k : \mathcal{P} (\mathbb{C}) \to \mathbb{R}, \quad \sigma \mapsto \min_{q_k \in \mathcal{Q}^k} \max_{\lambda \in \sigma} | q_k (\lambda) | \tag{3.35}
\]

using the power set \( \mathcal{P} \). The above quantity arises naturally in

**Lemma 3.6** (Optimal polynomial for diagonalizable matrices)

Let \( A \in \mathbb{R}^{n \times n} \) be diagonalizable with eigenvector matrix \( V \in \mathbb{R}^{n \times n} \). Then it holds

\[
\min_{q_k \in \mathcal{Q}^n} \| q_k (A) v \| \leq \kappa_2 (V) \mathcal{M}_k (\sigma (A)) \| v \| \quad \text{for any } v \in \mathbb{R}^n. \tag{3.36}
\]

**Proof.** With the diagonal matrix of eigenvalues \( \Lambda \in \mathbb{R}^{n \times n} \) and \( \Lambda = V^{-1} A V \) the proposition follows from

\[
\begin{align*}
\| q_k (A) v \| &\leq \| q_k (\Lambda) \| \| v \| \quad \text{for any } v \in \mathbb{R}^n, \\
\| q_k (A) \| &\leq \kappa_2 (V) \| q_k (\Lambda) \| \quad \text{for any } q_k \in \mathcal{Q}^k \text{ and} \\
\min_{q_k \in \mathcal{Q}^n} \| q_k (\Lambda) \| &\leq \mathcal{M}_k (\sigma (A)) = \mathcal{M}_k (\sigma (\Lambda)) \quad \text{for any } \Lambda = V^{-1} A V.
\end{align*}
\]

It remains to provide an estimate of \( \mathcal{M}_k \). The next lemma provides that, again taken from [Saa03, section 6.11]. We first define an ellipse in the complex plane centered at \( c \) with focal distance \( d \) and semi-major axis \( a \):

\[
E : \mathbb{C}^3 \to \mathbb{P} (\mathbb{C}), \quad (c, d, a) \mapsto \{ z \in \mathbb{C} \mid |z - c + d| + |z - c - d| \leq 2 |a| \} \tag{3.37}
\]

and now

**Lemma 3.7** (\( \mathcal{M}_k \) over an ellipse)

Let \( E (c, d, a) \) be as in Eq. (4.73). Then it holds

\[
\mathcal{M}_k (E (c, d, a)) \leq \left[ \frac{a + \sqrt{a^2 - d^2}}{c + \sqrt{c^2 - d^2}} \right]^k. \tag{3.38}
\]

**Proof.** A proof can be found in [Saa03, section 6.11].

**Remark 3.3.** For \( a = d \in \mathbb{R} \) and \( c \in \mathbb{R} \) the ellipse \( E (c, d, a) \) reduces to a line segment on the real axis. Using

\[
\lambda_{\min} := c - a, \quad \lambda_{\max} := c + a \quad \text{and} \quad \kappa := \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3.39}
\]
gives us
\[
\mathcal{M}_k (E (c, d, a)) \leq 2 \left( \frac{\sqrt{k} - 1}{\sqrt{k} + 1} \right)^k
\]  
(3.40)
which is the well-known bound for the \( k \)-th relative error norm (cf. Eq. (3.3)) measured in the \( \mathbf{A} \)-norm of the conjugate gradient method.

The task of estimating the convergence rate of a Krylov subspace method applied to a diagonalizable matrix \( \mathbf{A} \in \mathbb{R}^{n \times n} \) now reduces to finding the possible locations of the eigenvalues.

For a symmetric positive definite matrix it suffices to get an estimate of its spectral condition number, as its eigenvalues lie on the positive real axis. Then lemma 3.7 is applicable and yields an optimal estimate of the convergence rate (cf. [Saa03, section 6.11.1]).

For a non-symmetric matrix the picture looks different. Its eigenvalues may be spread across the whole complex plane. If we are lucky and the eigenvalues are clustered together, we can find a single ellipse which contains the eigenvalues and lemma 3.7 provides an asymptotically optimal estimate of the convergence rate (cf. [Saa03, section 6.11.2]). On the other hand, if the eigenvalues can be split into two well-separated clusters, the estimate may be overly pessimistic.

We are now ready to introduce the conjugate gradient method and the generalized conjugate residual method, along with some standard convergence results as presented in [Saa03, section 6.11.3] and [EES83].

3.2.1 The conjugate gradient method

The conjugate gradient method \( \mathcal{S}_{CG} \) is an orthogonal projection method onto the \( k \)-th Krylov subspace \( \mathcal{K}^{(k)} (\mathbf{A}, \mathbf{r}_*(0)) \) (cf. Eq. (3.28)) for a symmetric positive definite governing matrix \( \mathbf{A} \in \mathbb{R}^{n \times n} \). It can be identified with its basis vector recurrence relation:
\[
p_{CG}^{(k+1)} = r_{CG}^{(k+1)} + \beta_{CG}^{(k)} p_{CG}^{(k)}
\]  
(3.41)
using :
\[
\beta_{CG}^{(k)} := -\frac{\langle r_{CG}^{(k+1)}, r_{CG}^{(k+1)} \rangle}{\langle r_{CG}^{(k)}, r_{CG}^{(k)} \rangle}
\]  
(3.42)
and its basis coefficient definition:
\[
\alpha_{CG}^{(k)} = \frac{\langle r_{CG}^{(k)}, r_{CG}^{(k)} \rangle}{\langle \mathbf{A} p_{CG}^{(k)}, p_{CG}^{(k)} \rangle}.
\]  
(3.43)

For the \( \mathbf{A} \)-norm of its \( k \)-th error vector \( \mathbf{e}_{CG}^{(k)} \) we have the conjugate gradient method \( \mathcal{S}_{CG} \) convergence result:
\[
\left\| \mathbf{e}_{CG}^{(k)} \right\|_\mathbf{A} \leq 2 \left[ \frac{\sqrt{k_2(\mathbf{A})} - 1}{\sqrt{k_2(\mathbf{A})} + 1} \right]^k \left\| \mathbf{e}_{CG}^{(0)} \right\|_\mathbf{A}
\]  
(3.44)
which follows from lemmas 3.4, 3.6 and 3.7 (cf. [Saa03, section 6.11.3]).

3.2.2 The generalized conjugate residual method

The generalized conjugate residual method \( \mathcal{S}_{GCR} \) is an oblique projection method onto the \( k \)-th Krylov subspace \( \mathcal{K}^{(k)} (\mathbf{A}, \mathbf{r}_*(0)) \) (cf. Eq. (3.28)) orthogonally to \( \mathcal{L}^{(k)} = \mathbf{A} \mathcal{K}^{(k)} (\mathbf{A}, \mathbf{r}_*(0)) \) for a positive definite governing matrix \( \mathbf{A} \in \mathbb{R}^{n \times n} \). It can be identified with its basis vector recurrence relation:
\[
p_{GCR}^{(k+1)} = r_{GCR}^{(k+1)} + \sum_{l=0}^{k} \beta_{GCR}^{(k,l)} p_{GCR}^{(l)}
\]  
(3.45)
using :
\[
\beta_{GCR}^{(k,l)} := -\frac{\langle \mathbf{A} r_{GCR}^{(k+1)}, \mathbf{A} p_{GCR}^{(l)} \rangle}{\langle \mathbf{A} p_{GCR}^{(l)}, \mathbf{A} p_{GCR}^{(l)} \rangle}
\]  
(3.46)
and its basis coefficient definition:

$$a_{GCR}^{(k)} = \frac{\langle r_{GCR}^{(k)}, A p_{GCR}^{(k)} \rangle}{\langle A p_{GCR}^{(k)}, A p_{GCR}^{(k)} \rangle}.$$  

(3.47)

Assuming that $A$ is diagonalizable with eigenvector matrix $V \in \mathbb{R}^{n \times n}$ we have for the Euclidean norm of its $k$-th residual vector $r_{\ast}^{(k)}$ the generalized conjugate residual method $S_{GCR}$ convergence result:

$$\|r_{GCR}^{(k)}\| \leq \kappa_2(V) M_k(\sigma(A)) \|r_{GCR}^{(0)}\|$$  

(3.48)

which follows from lemmas 3.5 to 3.7 (cf. [EES83, theorem 3.3]).

### 3.3 Left preconditioned Krylov subspace methods

Left preconditioned Krylov subspace methods are Krylov subspace methods which instead of solving the linear system of equations $A x = b$ directly, are applied to the left preconditioned system:

$$M_x^{-1} (b - A x) = 0$$  

(3.49)

with a preconditioning matrix:

$$M_x \in \mathbb{R}^{n \times n}, \quad M_x \approx A.$$  

(3.50)

The preconditioning matrix $M_x \in \mathbb{R}^{n \times n}$ gives rise to the left preconditioned matrix:

$$P_x \in \mathbb{R}^{n \times n}, \quad P_x := M_x^{-1} A$$  

(3.51)

and to the $k$-th left preconditioned residual vector:

$$z_{\ast}^{(k)} \in \mathbb{R}^n, \quad z_{\ast}^{(k)} := M_x^{-1} r_{\ast}^{(k)}.$$  

(3.52)

The eigenvalues of $P_x \in \mathbb{R}^{n \times n}$ will determine the convergence rate of the preconditioned method. The basis vectors $p_{\ast}^{(1)}, \ldots, p_{\ast}^{(k)} \in \mathbb{R}^n$ now have to span the $k$-th Krylov subspace $K^{(k)}(P_x, z_{\ast}^{(0)})$.

#### 3.3.1 The left preconditioned conjugate gradient method

The conjugate gradient method $S_{CG}$ requires the governing matrix $A$ to be symmetric positive definite. As such, it appears to not be applicable to the left preconditioned system $M_x^{-1} (b - A x) = 0$, as the left preconditioned matrix $P_x$ is not symmetric in general. If however the preconditioning matrix $M_x$ is itself symmetric positive definite, it induces the $M_x$-inner product:

$$\langle \cdot, \cdot \rangle_{M_x} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad \langle x, y \rangle_{M_x} := \langle M_x x, y \rangle$$  

(3.53)

for which the left preconditioned matrix $P_x$ is self-adjoint. Then we can just replace the Euclidean inner product in the conjugate gradient method $S_{CG}$ with the $M_x$-inner product $\langle \cdot, \cdot \rangle_{M_x}$ to obtain the left preconditioned conjugate gradient method $S_{LPCG,\ast}$. The left preconditioned conjugate gradient method $S_{LPCG,\ast}$ is an $\langle \cdot, \cdot \rangle_{M_x}$-orthogonal projection method onto the $k$-th Krylov subspace $K^{(k)}(P_x, z_{\ast}^{(0)})$ for symmetric positive definite governing matrix $A \in \mathbb{R}^{n \times n}$ and symmetric positive definite preconditioning matrix $M_x \in \mathbb{R}^{n \times n}$. It can be identified with its basis vector recurrence relation:

$$p_{LPCG,\ast}^{(k+1)} = z_{LPCG,\ast}^{(k+1)} + \beta_{LPCG,\ast}^{(k)} p_{LPCG,\ast}^{(k)}$$  

(3.54)

using:

$$\beta_{LPCG,\ast}^{(k)} := -\frac{\langle z_{LPCG,\ast}^{(k+1)}, z_{LPCG,\ast}^{(k+1)} \rangle_{M_x}}{\langle z_{LPCG,\ast}^{(k)}, z_{LPCG,\ast}^{(k)} \rangle_{M_x}}.$$  

(3.55)

and its basis coefficient definition:

$$a_{LPCG,\ast}^{(k)} = \frac{\langle z_{LPCG,\ast}^{(k)}, z_{LPCG,\ast}^{(k)} \rangle_{M_x}}{\langle A p_{LPCG,\ast}^{(k)}, p_{LPCG,\ast}^{(k)} \rangle_{M_x}}.$$  

(3.56)
For the $\mathbf{A}$-norm of its $k$-th error vector $\mathbf{e}_{\star}^{(k)}$ we have essentially the same expression as in the conjugate gradient method $S_{CG}$ convergence result (cf. Eq. (3.44)), using the symmetrically preconditioned matrix:

$$\bar{\mathbf{P}}_{\star} \in \mathbb{R}^{n \times n}, \quad \bar{\mathbf{P}}_{\star} := \mathbf{M}_{\star}^{-1/2} \mathbf{A} \mathbf{M}_{\star}^{-1/2} \quad (3.57)$$

we get the convergence result for the left preconditioned conjugate gradient method $S_{LPCG,\star}$:

$$\|\mathbf{e}_{LPCG,\star}^{(k)}\|_{\mathbf{A}} \leq 2 \left[ \frac{\sqrt{k_2 \left( \bar{\mathbf{P}}_{LPCG,\star} \right)} - 1}{\sqrt{k_2 \left( \bar{\mathbf{P}}_{LPCG,\star} \right)} + 1} \right]^k \|\mathbf{e}_{LPCG,\star}^{(0)}\|_{\mathbf{A}} \quad (3.58)$$

which again follows from lemmas 3.4, 3.6 and 3.7.

### 3.3.2 The left preconditioned generalized conjugate residual method

Just as was done for the left preconditioned conjugate gradient method $S_{LPCG,\star}$, we will have to swap out the standard Euclidean inner product for a different one. This time we do not require the preconditioning matrix $\mathbf{M}_{\star}$ to be symmetric positive definite, but just to be invertible. Then $\mathbf{M}_{\star}^2 \mathbf{M}_{\star}$ induces the $\mathbf{M}_{\star}^2 \mathbf{M}_{\star}$-inner product:

$$\langle \cdot, \cdot \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}} := \langle \mathbf{M}_{\star} \mathbf{x}, \mathbf{M}_{\star} \mathbf{y} \rangle \quad (3.59)$$

and the left preconditioned conjugate gradient method $S_{LPGCR,\star}$ is an oblique projection method onto the $k$-th Krylov subspace $\mathcal{K}^{(k)}(\mathbf{P}_{\star}, \mathbf{z}_{\star}^{(0)}) \langle \cdot, \cdot \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}}$-orthogonally to $\mathcal{L}^{(k)} = \mathbf{P}_{\star} \mathcal{K}^{(k)}(\mathbf{P}_{\star}, \mathbf{z}_{\star}^{(0)})$ for a positive definite governing matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and an invertible preconditioning matrix $\mathbf{M}_{\star} \in \mathbb{R}^{n \times n}$. It can be identified with its basis vector recurrence relation:

$$\mathbf{p}_{LPGCR,\star}^{(k+1)} = \mathbf{z}_{LPGCR,\star}^{(k+1)} + \sum_{l=0}^{k} \beta_{LPGCR,\star}^{(k,l)} \mathbf{p}_{LPGCR,\star}^{(l)} \quad (3.60)$$

using:

$$\beta_{LPGCR,\star}^{(k,l)} := \frac{\langle \mathbf{A} \mathbf{z}_{LPGCR,\star}^{(k+1)}, \mathbf{p}_{LPGCR,\star}^{(l)} \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}}}{\langle \mathbf{A} \mathbf{p}_{LPGCR,\star}^{(l)}, \mathbf{A} \mathbf{p}_{LPGCR,\star}^{(l)} \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}}} \quad (3.61)$$

and its basis coefficient definition:

$$\alpha_{LPGCR,\star}^{(l)} = \frac{\langle \mathbf{z}_{LPGCR,\star}, \mathbf{p}_{LPGCR,\star}^{(l)} \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}}}{\langle \mathbf{A} \mathbf{p}_{LPGCR,\star}^{(l)}, \mathbf{A} \mathbf{p}_{LPGCR,\star}^{(l)} \rangle_{\mathbf{M}_{\star}^2 \mathbf{M}_{\star}}} \quad (3.62)$$

Assuming that $\mathbf{P}_{\star}$ is diagonalizable with eigenvector matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$ we have for the Euclidean norm of its $k$-th residual vector $\mathbf{r}_{\star}^{(k)}$ the convergence result for the left preconditioned conjugate gradient method $S_{LPGCR,\star}$:

$$\|\mathbf{r}_{LPGCR,\star}^{(k)}\| \leq \kappa_2 (\mathbf{V}) \mathcal{M}_k (\sigma (\mathbf{P}_{\star})) \|\mathbf{r}_{LPGCR,\star}^{(0)}\| \quad (3.63)$$

which again follows from lemmas 3.5 to 3.7.
4 Preconditioners for the discrete problem

In this chapter we present, analyze and compare our four solution strategies. We start with a short introduction of the relevant notation. We then introduce the four methods, where we want to make a distinction between the \textit{exactly preconditioned methods} and the \textit{inexactly preconditioned methods}. We make this distinction so that we can split the analysis of the convergence behaviour of the methods into two parts.

First we perform an idealized analysis in which we assume that we can invert the preconditioning matrices exactly. This has the benefit that we can clearly \textit{quantify} the impact of the different discretization parameters, and that the analysis is much simpler than for the inexact case.

After this, we perform an analysis of the convergence behaviour for the case that we approximate the inversion of the exact preconditioning matrices using a fixed number of iterations of a stationary iterative method. This will mainly be a \textit{qualitative analysis}, i.e. we are able to identify and predict trends, but we obtain no sharp bounds on the convergence behaviour for all discretization parameters.

Finally, we will present numerical experiments to validate the predictions made in the first sections of this chapter.

4.1 Convergence analysis

Let us first recall the discretization parameters and the associated notation. The spatial step size $h \in \mathbb{R}_+$ determines the \textit{Laplacian matrix}:

$A \in \mathbb{R}^{n \times n}$ (4.1)

and the \textit{spatial mass matrix}:

$M \in \mathbb{R}^{n \times n}$. (4.2)

The temporal discretization order $p \in \mathbb{N}$ determines the complex eigenvalues of the \textit{characteristic temporal matrix} $\rho^{-1} \mu \in \mathbb{R}^{(p+1) \times (p+1)}$ (cf. Eq. (2.76)), while the chosen temporal basis and test functions determine the \textit{temporal mass matrix} $\mu \in \mathbb{R}^{(p+1) \times (p+1)}$ (cf. Eq. (2.71)) and the \textit{temporal propagation matrix} $\rho \in \mathbb{R}^{(p+1) \times (p+1)}$ (cf. Eq. (2.72)). These matrices become the arguments of the mapping:

$K_m : \mathbb{R}^{m \times m} \times \mathbb{R}^{m \times m} \to \mathbb{R}^{[m n] \times [m n]}, \quad (\mu, \rho) \mapsto \mu \otimes A + \rho \otimes M$ (4.3)

to determine the global stiffness matrix using the last discretization parameter, the time step size $\tau \in \mathbb{R}_+$:

$K_{\text{global}} := K_{(p+1)} (\tau \mu, \rho)$. 

The global stiffness matrix $K_{\text{global}} \in \text{Image} (K_{(p+1)})$ consists of $(p+1) \times (p+1)$-blocks of matrices from the image of $K_1$.

Choosing the temporal basis and test functions such that the temporal matrices are block diagonal with blocks from the image of the \textit{embedding}:

$\gamma : \mathbb{C} \to \mathbb{R}^{2 \times 2}, \quad z \mapsto \begin{pmatrix} \text{Re}(z) & \text{Im}(z) \\ -\text{Im}(z) & \text{Re}(z) \end{pmatrix}$ (4.4)

furthermore causes the diagonal $2 \times 2$-block matrices to be from the image $K_2 \left[ \text{Image} (\gamma)^2 \right] \subset \mathbb{R}^{(2n) \times (2n)}$. This means that each diagonal $2 \times 2$ block matrix $K$ is of the form

$K_2 (\gamma (\mu), \gamma (\rho)) \in \mathbb{R}^{(2n) \times (2n)}$, 

i.e. it can be identified with two \textit{characteristic complex numbers}:

$\mu, \rho \in \mathbb{C}$. (4.5)
Let us now fix one block of the characteristic temporal matrix $\mu^{-1} \rho$. It is identified with its characteristic temporal eigenvalue:

$$\lambda \in \sigma(\mu^{-1} \rho).$$  \hspace{1cm} (4.6)

The characteristic complex numbers $\mu, \rho \in \mathbb{C}$ satisfy the equality

$$\mu^{-1} \rho = \tau^{-1} \lambda.$$  \hspace{1cm} (4.7)

We will assume that the real parts of both characteristic complex numbers $\mu, \rho \in \mathbb{C}$ are positive and that the characteristic complex argument:

$$\varphi \in [0, \pi/2], \quad \varphi := \text{arg}(\lambda)$$  \hspace{1cm} (4.8)

is non-negative. Without loss of generality we can assume that the equality $|\mu| = \tau$ holds.

From now on we will focus on one block matrix $K \in \mathbb{R}^{(2n) \times (2n)}$, where the blocks are from the image $\text{Image}(K_1) \subset \mathbb{R}^{n \times n}$ of the mapping $K_m$ (cf. Eq. (4.3)). We introduce notation for the diagonal block:

$$K_{\text{diag}} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n}, \quad K_{\text{diag}} := \text{Re}(\mu) A + \text{Re}(\rho) M$$  \hspace{1cm} (4.9)

and for the off-diagonal block:

$$K_{\text{off}} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n}, \quad K_{\text{off}} := \text{Im}(\mu) A + \text{Im}(\rho) M$$  \hspace{1cm} (4.10)

to define the block matrix:

$$K \in \mathbb{R}^{(2n) \times (2n)}, \quad K := K_2 (\gamma(\mu), \gamma(\rho)) = \begin{pmatrix} K_{\text{diag}} & K_{\text{off}} \\ -K_{\text{off}} & K_{\text{diag}} \end{pmatrix}$$  \hspace{1cm} (4.11)

and we use the solution vector:

$$w \in \mathbb{R}^{2n}, \quad w := \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$  \hspace{1cm} (4.12)

and the right-hand side vector:

$$f \in \mathbb{R}^{2n}, \quad f := \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$  \hspace{1cm} (4.13)

to write the block system:

$$K w \equiv \begin{pmatrix} K_{\text{diag}} & K_{\text{off}} \\ -K_{\text{off}} & K_{\text{diag}} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \equiv f.$$  \hspace{1cm} (4.14)

### 4.1.1 Solution strategies

**The block multigrid preconditioner**

We will refer to the following method as the block multigrid preconditioned method $S_{\text{BMG}}$, as the used preconditioner is just the block multigrid method.

We solve Eq. (4.14) iteratively using the left preconditioned conjugate gradient method $S_{\text{LPGCR}}$ (cf. Section 3.3.2) using the block matrix $K \in \mathbb{R}^{(2n) \times (2n)}$ itself as the exact preconditioner.

We can of course not invert the block matrix $K \in \mathbb{R}^{(2n) \times (2n)}$ exactly and cheaply, and we thus use $\kappa$ V-cycles of the block multigrid method as the inexact preconditioner.

The distinction between the exact and the inexact preconditioner serves mainly to put it into the same framework of analysis as the other methods in this chapter. The first part of this analysis is just this: If we apply the preconditioner exactly the Krylov subspace method converges after one iteration.

The more interesting analysis of this method will happen not in this section, but in the ones that follow.
The block Jacobi preconditioner

We will refer to the following method as the \textit{block Jacobi preconditioned method} \( S_J \), as the used preconditioning matrix is the block diagonal part of the block matrix \( K \in \mathbb{R}^{(2n) \times (2n)} \). It has been used e.g. in [MNS07] in the context of implicit Runge–Kutta schemes and has been shown to be order-optimal.

We solve Eq. (4.14) iteratively using the left preconditioned conjugate gradient method \( S_{LPGCR} \) (cf. Section 3.3.2) with the \textit{block Jacobi preconditioner}:

\[
K_J \in \mathbb{R}^{(2n) \times (2n)}, \quad K_J := \begin{pmatrix} K_{\text{diag}} & 0 \\ 0 & K_{\text{diag}} \end{pmatrix}
\]  

(4.15)

as preconditioning matrix.

We will show that the eigenvector basis of the \textit{block Jacobi preconditioned matrix}:

\[
K_J^{-1} \in \mathbb{R}^{(2n) \times (2n)}, \quad K_J^{-1} := \begin{pmatrix} I & K_{\text{diag}}^{-1} K_{\text{off}} \\ K_{\text{diag}}^{-1} K_{\text{off}} & I \end{pmatrix}
\]

(4.16)

is well-conditioned, and that the eigenvalues of the block Jacobi preconditioned matrix \( K_J^{-1} \) are contained in a small, compact set for all discretization parameters if we use the right temporal basis and test functions.

The block Gauss-Seidel preconditioner

We will refer to the following method as the \textit{block Gauss-Seidel preconditioned method} \( S_{GS} \), as the used preconditioning matrix is the block lower-triangular part of the block matrix \( K \in \mathbb{R}^{(2n) \times (2n)} \). It has been mentioned e.g. in [MNS07] in the context of implicit Runge–Kutta schemes and has been reported to be superior to the block Jacobi preconditioned method \( S_J \).

We solve Eq. (4.14) iteratively using the left preconditioned conjugate gradient method \( S_{LPGCR} \) (cf. Section 3.3.2) with the \textit{block Gauss-Seidel preconditioner}:

\[
K_{GS} \in \mathbb{R}^{(2n) \times (2n)}, \quad K_{GS} := \begin{pmatrix} K_{\text{diag}} & 0 \\ -K_{\text{off}} & K_{\text{diag}} \end{pmatrix}
\]  

(4.17)

as preconditioning matrix.

While the eigenvector basis of the \textit{block Gauss-Seidel preconditioned matrix}:

\[
K_{GS}^{-1} \in \mathbb{R}^{(2n) \times (2n)}, \quad K_{GS}^{-1} := \begin{pmatrix} I & K_{\text{diag}}^{-1} K_{\text{off}} \\ K_{\text{diag}}^{-1} K_{\text{off}} & I \end{pmatrix}
\]

(4.18)

is potentially badly-conditioned, we will present numerical experiments which identify the block Gauss-Seidel preconditioned method \( S_{GS} \) to be superior to the block Jacobi preconditioned method \( S_J \). This is due to the eigenvalues of the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} \) being spread out less than the eigenvalues of the block Jacobi preconditioned matrix \( K_J^{-1} \).

The preconditioned Schur complement method

We will refer to the following method as the \textit{preconditioned Schur complement method} \( S_{S} \). It has been proposed in [WB15,BB15], where it has also been shown to be \textit{globally-optimal}.

We define the \textit{Schur complement matrix}:

\[
S \in \mathbb{R}^{n \times n}, \quad S := K_{\text{diag}} K_{\text{off}}^{-1} K_{\text{diag}} + K_{\text{off}}
\]

(4.19)

We first solve Eq. (4.14) for the \textit{essential unknown} \( w_2 \in \mathbb{R}^n \) using the \textit{Schur complement equation}:

\[
Sw_2 = f_1 + K_{\text{diag}} K_{\text{off}}^{-1} f_2
\]

(4.20)

and then obtain the \textit{other unknown} \( w_1 \in \mathbb{R}^n \) by solving the second equation:

\[
K_{\text{off}} w_1 = K_{\text{diag}} w_2 - f_2
\]

(4.21)
We solve Eq. (4.20) iteratively using the left preconditioned conjugate gradient method \( S_{LGCG} \) (cf. Section 3.3.1) with the inexact Schur complement factorization:

\[
S_{\text{inexact}}^{-1} \in \mathbb{R}^{n \times n}, \quad S_{\text{inexact}} := K_1 (|\mu|, |\rho|) K_{\text{off}}^{-1} K_1 (|\mu|, |\rho|) \tag{4.22}
\]

as preconditioning matrix and any inversion of the off-diagonal block \( K_{\text{off}} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) is done using the conjugate gradient method \( S_{CG} \) (cf. Section 3.2.1). As long as the off-diagonal block \( K_{\text{off}} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) is well-conditioned, the conjugate gradient method \( S_{CG} \) is efficient.

It can be shown that both the Schur complement matrix \( S \in \mathbb{R}^{n \times n} \) and the inexact Schur complement factorization \( S_{\text{inexact}} \in \mathbb{R}^{n \times n} \) are symmetric positive definite and that the eigenvalues of the preconditioned Schur complement matrix:

\[
S_{\text{inexact}}^{-1} S \in \mathbb{R}^{n \times n} \tag{4.23}
\]

are contained in a small, compact interval independent of all discretization parameters (cf. [BB15, proposition 3.2]).

### 4.1.2 Convergence analysis for the exactly preconditioned methods

In this section we analyze the convergence behaviour of the above methods under the assumption that we invert the preconditioning matrices exactly. While we will only approximate the inversion of these matrices, the analysis of the exactly preconditioned methods is a first step in the analysis of the inexact preconditioned methods.

As was seen in the previous chapter, for the convergence analysis of the used iterative methods we need a bound on the spectrum of the preconditioned matrices and an estimate of the condition number of the eigenvector bases of the block Jacobi and block Gauss-Seidel preconditioned matrices. The following lemma will be central to our analysis, providing us with the needed building blocks.

**Lemma 4.1** (Eigendecomposition of special matrices)

Let \( A, M \in \mathbb{R}^{n \times n} \) be symmetric positive definite matrices and \( \mu, \rho \in \mathbb{C} \) be complex numbers with positive real part. Then the matrix:

\[
F \in \mathbb{R}^{n \times n}, \quad F := [\text{Re}(\mu) A + \text{Re}(\rho) M]^{-1} [\text{Im}(\mu) A + \text{Im}(\rho) M] \tag{4.24}
\]

is diagonalizable using a matrix of eigenvectors \( V \in \mathbb{R}^{n \times n} \) which is independent of \( \mu \) and \( \rho \). The condition number of the matrix of eigenvectors \( V \in \mathbb{R}^{n \times n} \) fulfills the inequality:

\[
\kappa_2 (V) \leq \sqrt{\kappa_2 (M)} \tag{4.25}
\]

and the spectrum of the matrix \( F \in \mathbb{R}^{n \times n} \) fulfills the equality:

\[
\sigma (F) = \left\{ \begin{array}{l}
\text{Im}(\mu) \lambda + \text{Im}(\rho) \lambda + \text{Re}(\rho) \lambda + \text{Re}(\mu) \\
\text{Re}(\mu) \lambda + \text{Re}(\rho) \lambda + \text{Im}(\mu) \end{array} \right\}, \lambda \in \sigma (M^{-1} A) \tag{4.26}
\]

**Proof.** We first note that due to \( M \in \mathbb{R}^{n \times n} \) being symmetric positive definite, its square root \( M^{1/2} \in \mathbb{R}^{n \times n} \) exists and is symmetric. As \( A \in \mathbb{R}^{n \times n} \) is also symmetric, the matrix:

\[
\tilde{A} \in \mathbb{R}^{n \times n}, \quad \tilde{A} := M^{-1/2} A M^{-1/2} \tag{4.27}
\]

is so as well. We now note that the equality

\[
M^{1/2} F M^{-1/2} = M^{1/2} [\text{Re}(\mu) A + \text{Re}(\rho) M]^{-1} M^{1/2} [\text{Im}(\mu) A + \text{Im}(\rho) M] M^{-1/2}
\]

holds and therefore

\[
F = M^{-1/2} \left[ \text{Re}(\mu) \tilde{A} + \text{Re}(\rho) I \right]^{-1} \left[ \text{Im}(\mu) \tilde{A} + \text{Im}(\rho) I \right] M^{1/2}.
\]

The symmetric matrix \( \tilde{A} \in \mathbb{R}^{n \times n} \) is diagonalizable as \( \tilde{A} = Q \Lambda Q^T \) with an orthogonal matrix \( Q \in \mathbb{R}^{n \times n} \) and the diagonal matrix of eigenvalues \( \Lambda \in \mathbb{R}^{n \times n} \). Therefore the matrix \( F \in \mathbb{R}^{n \times n} \) is diagonalizable as

\[
F = M^{-1/2} Q \left[ \text{Re}(\mu) A + \text{Re}(\rho) I \right]^{-1} \left[ \text{Im}(\mu) A + \text{Im}(\rho) I \right] Q^T M^{1/2},
\]
revealing one possible matrix of eigenvectors:

\[ V \in \mathbb{R}^{n \times n}, \quad V := M^{-1/2} Q. \]  

(4.28)

As \( Q \in \mathbb{R}^{n \times n} \) is orthogonal, we get Eq. (4.25). Since \( M^{-1/2} A M^{-1/2} \in \mathbb{R}^{n \times n} \) is similar to \( M^{-1} A \in \mathbb{R}^{n \times n} \) we get Eq. (4.26).

From this lemma we immediately get the following result, concerning the spectra of the preconditioned matrices:

**Lemma 4.2** (Spectra of the preconditioned matrices)

Let \( \mu, \rho \in \mathbb{C} \) be the characteristic complex numbers as in Eq. (4.5), \( S_{\text{inexact}}^{-1} \in \mathbb{R}^{n \times n} \) be the preconditioned Schur complement matrix as in Eq. (4.23), \( K_J^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Jacobi preconditioned matrix as in Eq. (4.16) and \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Gauss-Seidel preconditioned matrix as in Eq. (4.18).

Then it holds

\[
\begin{align*}
\sigma(S_{\text{inexact}}^{-1}) &= \left\{ \frac{(\text{Re}(\mu) + \text{Re}(\rho))^2 + (\text{Im}(\mu) + \text{Im}(\rho))^2}{(\mu + \rho)^2} \mid \mu \in \sigma(M^{-1} A) \right\}, \\
\sigma(K_J^{-1} K) &= \left\{ 1 \pm \frac{(\text{Im}(\mu) + \text{Im}(\rho))}{\text{Re}(\mu) + \text{Re}(\rho)} \mid \mu \in \sigma(M^{-1} A) \right\}, \\
\sigma(K_{GS}^{-1} K) &= \{1\} \cup \left\{ 1 + \frac{(\text{Im}(\mu) + \text{Im}(\rho))}{\text{Re}(\mu) + \text{Re}(\rho)} \mid \mu \in \sigma(M^{-1} A) \right\}.
\end{align*}
\]

(4.29) (4.30) (4.31)

**Proof.** For Eq. (4.29) we apply lemma 4.1 to obtain from the definitions of the Schur complement matrix \( S \in \mathbb{R}^{n \times n} \) (cf. Eq. (4.19)) and of the inexact Schur complement factorization \( S_{\text{inexact}}^{-1} \in \mathbb{R}^{n \times n} \) (cf. Eq. (4.22)) the equality

\[
S_{\text{inexact}}^{-1} = K_1 (|\mu|, |\rho|)^{-1} K_{\text{diag}} K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}} + K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}} K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}},
\]

(4.32)

as the matrices \( K_1 (|\mu|, |\rho|)^{-1} K_{\text{diag}} \in \mathbb{R}^{n \times n} \) and \( K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}} \in \mathbb{R}^{n \times n} \) commute due to the shared eigenvector basis. We diagonalize the preconditioned Schur complement matrix \( S_{\text{inexact}}^{-1} \in \mathbb{R}^{n \times n} \) using the matrix of eigenvectors \( V \in \mathbb{R}^{n \times n} \) (cf. Eq. (4.28)) from lemma 4.1 to get

\[
V^{-1} S_{\text{inexact}}^{-1} V = [|\mu| A + |\rho| I]^{-2} \left( (\text{Re}(\mu) A + \text{Re}(\rho) I)^2 + (\text{Im}(\mu) A + \text{Im}(\rho) I)^2 \right)
\]

where \( A \in \mathbb{R}^{n \times n} \) is a diagonal matrix of eigenvalues of \( M^{-1} A \), which yields Eq. (4.29).

For Eq. (4.30) and Eq. (4.31) we first diagonalize the blocks of the preconditioned matrices using the block diagonal matrix of eigenvectors:

\[
\tilde{V} \in \mathbb{R}^{(2n) \times (2n)}, \quad \tilde{V} := \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}.
\]

(4.33)

We define the diagonal matrix:

\[
A_F \in \mathbb{R}^{n \times n}, \quad A_F := [\text{Re}(\mu) A + \text{Re}(\rho) I]^{-1} [\text{Im}(\mu) A + \text{Im}(\rho) I]
\]

(4.34)

which gives us

\[
\tilde{V}^{-1} K_J^{-1} K \tilde{V} = \begin{pmatrix} I & A_F \\ -A_F & I \end{pmatrix},
\]

which can be diagonalized using an orthogonal matrix \( Q \in \mathbb{R}^{(2n) \times (2n)} \) to

\[
Q^T \tilde{V}^{-1} K_J^{-1} K \tilde{V} Q = \begin{pmatrix} I + tA_F & 0 \\ 0 & I - tA_F \end{pmatrix},
\]

(4.35)

which finally gives us Eq. (4.30).

For Eq. (4.31) we get

\[
\tilde{V}^{-1} K_{GS}^{-1} K \tilde{V} = \begin{pmatrix} I & A_F \\ 0 & I + A_F^2 \end{pmatrix}.
\]

(4.36)

As this is an upper triangular matrix, its eigenvalues are just the entries on the diagonal, which yields Eq. (4.31).

\[\square\]
There is some important information that we can extract from the above lemma.

First off, we want to note that given a characteristic temporal eigenvalue \( \lambda \in \sigma ( \mu^{-1} \rho ) \), the spectrum of the preconditioned Schur complement matrix \( S_{\text{inexact}}^{-1} S \in \mathbb{R}^{n \times n} \) does not depend on the choice of the complex arguments of the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \). We have

\[
(\Re(\mu) \lambda + \Re(\rho))^2 + (\Im(\mu) \lambda + \Im(\rho))^2 = |\mu|^2 \lambda^2 + 2 [\Re(\mu) \Re(\rho) + \Im(\mu) \Im(\rho)] \lambda + |\rho|^2
\]

and since the equality \( \mu^{-1} \rho = \tau^{-1} \lambda \) (cf. Eq. (4.7)) holds we get

\[
\Re(\mu) \Re(\rho) + \Im(\mu) \Im(\rho) = |\mu \rho| \cos (\arg(\lambda)) ,
\]

which yields the independence from the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \). Furthermore, if \( \arg(\mu) = - \arg(\rho) \) holds we have the equality

\[
\{1\} \cup \left( \frac{|\mu|}{\Re(\mu)} S_{\text{inexact}}^{-1} S \right) = \sigma ( K_{\text{GS}}^{-1} K ) ,
\]

as can be seen by comparing Eqs. (4.29) and (4.31). Therefore, if \( \arg(\mu) = - \arg(\rho) \) holds the block Gauss-Seidel preconditioned method \( S_{\text{GS}} \) performs almost as well as the preconditioned Schur complement method \( S_\Sigma \), as the spectra are almost identical. The difference between the two methods will be more pronounced if the spectrum of the preconditioned Schur complement matrix \( S_{\text{inexact}}^{-1} S \in \mathbb{R}^{n \times n} \) is well-separated from one, and less so if it includes one.

We will look at the impact of the various discretization parameters on the distribution of the eigenvalues of the preconditioned matrices \( K_J^{-1} K, K_{\text{GS}}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \). We have as discretization parameters the time step size \( \tau \in \mathbb{R}^+ \), the spatial step size \( h \in \mathbb{R}^+ \), the characteristic complex argument \( \varphi \in [0, \pi/2] \) and the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \).

We define the spectral function:

\[
\sigma_F : \mathbb{R}^+ \rightarrow \mathbb{R}, \quad \lambda \mapsto \frac{\Im(\mu) \lambda + \Im(\rho)}{\Re(\mu) \lambda + \Re(\rho)}
\]

which determines the mapping of an eigenvalue of the characteristic spatial matrix \( M^{-1} A \in \mathbb{R}^{n \times n} \) to the corresponding eigenvalue of the preconditioned matrix, compare Eqs. (4.30) and (4.31).

In Fig. 4.1 we plot \( \sigma_F (\lambda) \) over \( \lambda \in [10^{-2}, 10^2] \) for \( 10/\pi \varphi \in \{0, 1, 2, 3, 4, 5\} \), \( \arg(\mu) \in \{0, - \arg(\rho)\} \) and \( |\mu| = |\rho| \). The lines represent a fixed characteristic complex argument \( \varphi \in [0, \pi/2] \), while the left and right plot represent different temporal basis and test functions, identified by the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \).

We can see that both the characteristic complex argument \( \varphi \in [0, \pi/2] \) and the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \) determine the limits of \( \sigma_F (\lambda) \) as \( \lambda \in \mathbb{R}^+ \) goes to zero or infinity.

The plot for \( \arg(\mu) = 0 \) is highly non-symmetric, with the spectral function \( \sigma_F \) going to zero if \( \lambda \in \mathbb{R}^+ \) goes to infinity, and approaching some limit which depends on the characteristic complex argument \( \varphi \in [0, \pi/2] \) as \( \lambda \in \mathbb{R}^+ \) goes to zero. For \( \arg(\mu) = 0 \) and \( \varphi = \pi/2 \), the spectral function \( \sigma_F \) goes to infinity as \( \lambda \in \mathbb{R}^+ \) goes to zero.

The plot for \( \arg(\mu) = - \arg(\rho) \) is rotationally log-symmetric with respect to the point \( (1, 0) \). As \( \lambda \in \mathbb{R}^+ \) goes to zero or infinity, the spectral function \( \sigma_F \) approaches a limit which depends on the characteristic complex argument \( \varphi \in [0, \pi/2] \), but which is bounded by one.

Let us now look at the impact of the spatial step size \( h \in \mathbb{R}^+ \) on the spectra of the preconditioned matrices. The spatial step size \( h \in \mathbb{R}^+ \) determines the maximal ratio of eigenvalues:

\[
\kappa_\lambda : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^+, \quad A \mapsto \frac{\lambda_{\text{max}}(A)}{\lambda_{\text{min}}(A)}
\]

of the characteristic spatial matrix \( M^{-1} A \in \mathbb{R}^{n \times n} \). As the spatial step size \( h \in \mathbb{R}^+ \) decreases, i.e. as we refine our space discretization, the maximal ratio of eigenvalues \( \kappa_\lambda \) of the characteristic spatial matrix \( M^{-1} A \in \mathbb{R}^{n \times n} \) increases due to the increasing condition number of the Laplacian matrix \( A \in \mathbb{R}^{n \times n} \). In the context of Fig. 4.1 this means that the sector of the spectral function \( \sigma_F \) that is relevant to the spectra of the preconditioned matrices will increase in width.

The time step size \( \tau \in \mathbb{R}^+ \) will finally determine the width of the image:

\[
\sigma_F (\sigma (M^{-1} A) \subset (\tan (\arg(\mu)), \tan (\arg(\rho)))
\]

4.1. CONVERGENCE ANALYSIS

Page 35
of the spectrum $\sigma(M^{-1}A) \subset \mathbb{R}^+$ under the spectral function $\sigma_F$. The time step size $\tau \in \mathbb{R}^+$ will shift the relevant sector of the spectral function $\sigma_F$ to the left or to the right. In the limits $\tau \to 0$ and $\tau \to \infty$, the image $\sigma_F[\sigma(M^{-1}A)] \subset (\tan(\arg(\mu)), \tan(\arg(\rho)))$ will reduce to a single point.

We want to note that the width of the image $\sigma_F[\sigma(M^{-1}A)] \subset (\tan(\arg(\mu)), \tan(\arg(\rho)))$ is not the only thing driving the distribution of the eigenvalues of the preconditioned matrices and with it the contraction rates of the iterative solvers.

As can be seen from Eqs. (4.30) and (4.31), values of the spectral function $\sigma_F$ that are closer to zero are beneficial to the contraction rates of the iterative solvers, as those values lead to eigenvalues of the preconditioned matrices that are clustered at one. From this we can conclude that, everything else being equal, the case $\arg(\mu) = 0$ will outperform the case $\arg(\mu) = -\arg(\rho)$ for a large time step size $\tau \in \mathbb{R}^+$, as the value of the spectral function $\sigma_F$ does approach zero for the former, but not for the latter. Similarly, we can expect the advantage to be the other way around for a small time step size $\tau \in \mathbb{R}^+$, as the limit of the spectral function $\sigma_F$ as $\lambda$ approaches zero is bigger for $\arg(\mu) = 0$ than for $\arg(\mu) = -\arg(\rho)$.

It furthermore is beneficial to allow negative values of the spectral function $\sigma_F$. We can therefore predict the following behaviour for variable time step size $\tau \in \mathbb{R}^+$ and everything else fixed:

Starting with a very small time step size $\tau \in \mathbb{R}^+$, for the case $\arg(\mu) = -\arg(\rho)$ we expect the contraction rates of the iterative solvers to first worsen (i.e. grow) considerably as we increase the time step size $\tau \in \mathbb{R}^+$, until we reach a global maximum at $\sigma_F(\lambda_{\max}(M^{-1}A)) = 0$. Then it will first reach a local minimum when we have $\sigma_F(\lambda_{min}(M^{-1}A)) = -\sigma_F(\lambda_{max}(M^{-1}A))$, after which it will again worsen until we reach a global maximum at $\sigma_F(\lambda_{min}(M^{-1}A)) = 0$. Finally, it again improves for a very large time step size $\tau \in \mathbb{R}^+$.

In the same setting, for the case $\arg(\mu) = 0$ we expect the contraction rates of the iterative solvers to first worsen as the width of the image $\sigma_F[\sigma(M^{-1}A)] \subset (\tan(\arg(\mu)), \tan(\arg(\rho)))$ increases. At some point it will reach a global maximum, after which it should decrease dramatically, as the width of the the image $\sigma_F[\sigma(M^{-1}A)] \subset (\tan(\arg(\mu)), \tan(\arg(\rho)))$ decreases and its values simultaneously drift towards zero.

The hard problems are not the ones with a very small or very large time step size $\tau \in \mathbb{R}^+$, but the ones in the intermediate regime close to the maxima of the contraction rates of the iterative solvers, which
correspond to the limiting worst case:

\[ \sigma \left( \tau M^{-1} A \right) \to \mathbb{R}^+. \] (4.40)

Here, the spectrum of the preconditioned Schur complement matrix \( S^{-1}_{\text{inexact}} \in \mathbb{R}^{n \times n} \) will include one, the block Gauss-Seidel preconditioned method \( S_{\text{GS}} \) should therefore perform very similarly to the preconditioned Schur complement method \( S \) if we apply the preconditioners exactly.

We conclude the discussion of the spectra of the preconditioned matrices with a lemma providing bounds on them.

**Lemma 4.3** (Bounds on the spectra of the preconditioned matrices)

Let \( \mu, \rho \in \mathbb{C} \) be the characteristic complex numbers as in Eq. (4.5), \( \varphi \in [0, \pi/2] \) be the characteristic complex argument as in Eq. (4.8), \( S^{-1}_{\text{inexact}} \in \mathbb{R}^{n \times n} \) be the preconditioned Schur complement matrix as in Eq. (4.23), \( K^{-1}_{J} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Jacobi preconditioned matrix as in Eq. (4.16) and \( K^{-1}_{GS} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Gauss-Seidel preconditioned matrix as in Eq. (4.18). Then it holds

\[
\sigma \left( S^{-1}_{\text{inexact}} \right) \subset \left( \frac{1}{1 + \tan^2 (\varphi/2)} \right)^2, \quad (4.41) \\
\sigma \left( K^{-1}_{J} K \right) \subset \{1 \pm i \lambda \mid \lambda \in (\tan (\arg(\mu)), \tan (\arg(\rho)))\}, \quad (4.42) \\
\sigma \left( K^{-1}_{GS} K \right) \subset [1, 1 + \max \{\tan^2 (\arg(\mu)), \tan^2 (\arg(\rho))\}] \quad (4.43)
\]

for any time step size \( \tau \in \mathbb{R}^+ \) and spatial step size \( h \in \mathbb{R}^+ \).

**Proof.** A proof for Eq. (4.41) can be found in [BB15, proposition 3.2]. Both Eqs. (4.42) and (4.43) follow from the monotonicity of the spectral function \( \sigma_F \) (cf. Eq. (4.37)) and its limits as the time step size \( \tau \in \mathbb{R}^+ \) goes to zero and infinity respectively.

**Remark 4.1.** The same relations hold with an equal sign in the limiting worst case \( \sigma \left( \tau M^{-1} A \right) \to \mathbb{R}^+ \), which follows from the spectral function \( \sigma_F \) being an injection from the positive real numbers \( \mathbb{R}^+ \) to the interval \( (\tan (\arg(\mu)), \tan (\arg(\rho))) \subset \mathbb{R} \).

The above lemma reveals the following:

First, as was mentioned already, the spectrum of the preconditioned Schur complement matrix \( S^{-1}_{\text{inexact}} \in \mathbb{R}^{n \times n} \) is bounded by a small set for all discretization parameters. Next, if we choose \( \arg(\mu) = 0 \), which implies \( \arg(\rho) = \varphi \), and if the characteristic complex argument \( \varphi \in [0, \pi/2] \) goes to \( \pi/2 \), there exists no bounded set that contains the spectrum of block Jacobi preconditioned matrix \( K^{-1}_{J} K \in \mathbb{R}^{(2n) \times (2n)} \) or block Gauss-Seidel preconditioned matrix \( K^{-1}_{GS} K \in \mathbb{R}^{(2n) \times (2n)} \) for all discretization parameters. The same holds if we were to choose \( \arg(\mu) = -\varphi \) and \( \arg(\rho) = 0 \). For any other choice of the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \), there exists such a bounded set. The best choice of the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \) is the one that satisfies \( \arg(\mu) = -\arg(\rho) \), as then \( -\tan (\arg(\mu)) = \tan (\arg(\rho)) \in [0, 1] \) holds and thus

\[
\sigma \left( K^{-1}_{J} K \right) \subset \{1 \pm i \lambda \mid \lambda \in [0, 1]\}, \quad (4.44) \\
\sigma \left( K^{-1}_{GS} K \right) \subset [1, 1 + \tan^2 (\varphi/2)] \quad (4.45)
\]

holds for all discretization parameters. Note the similarity between Eqs. (4.41) and (4.45).

Now that we have provided the needed bounds on the spectra of the preconditioned matrices, we will focus our attention on the condition numbers of their eigenvector bases. Inspecting the convergence result for the left preconditioned conjugate gradient method \( S_{\text{LPCG}} \) (cf. Eq. (3.58)), we note that the condition number of the eigenvector basis of the preconditioned Schur complement matrix \( S^{-1}_{\text{inexact}} \in \mathbb{R}^{n \times n} \) does not play a role for the convergence of the exactly preconditioned Schur complement method \( S \). It will however play a role for the inexact method, we will therefore include a bound on it here. The condition number of the eigenvector basis of the preconditioned matrices \( K^{-1}_{J} K, K^{-1}_{GS} K \in \mathbb{R}^{(2n) \times (2n)} \) will influence the convergence result for the left preconditioned conjugate gradient method \( S_{\text{LPCGR}} \) (cf. Eq. (3.63)).

**Lemma 4.4** (Condition number of the eigenvector bases of the preconditioned matrices)

Let \( S^{-1}_{\text{inexact}} \in \mathbb{R}^{n \times n} \) be the preconditioned Schur complement matrix as in Eq. (4.23), \( K^{-1}_{J} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Jacobi preconditioned matrix as in Eq. (4.16), \( K^{-1}_{GS} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Gauss-Seidel preconditioned matrix as in Eq. (4.18) and assume the limiting worst case \( \sigma \left( \tau M^{-1} A \right) \to
Then there exists corresponding matrices of eigenvectors $V_S \in \mathbb{R}^{n \times n}$ and $V_J \in \mathbb{R}^{(2n) \times (2n)}$ such that the equalities

$$
\kappa_2(V_S) \leq \sqrt{\kappa_2(M)}, \\
\kappa_2(V_J) \leq \sqrt{\kappa_2(M)}
$$

hold. The condition number of the matrix of eigenvectors $V_{GS} \in \mathbb{R}^{(2n) \times (2n)}$ can get arbitrary large if $0 \in (\tan(\arg(\mu)), \tan(\arg(\rho)))$ holds.

\textbf{Proof.} We recall Eq. (4.32):

$$S_{\text{inexact}}^{-1} = K_1 (|\mu|, |\rho|)^{-1} K_{\text{diag}} K_1 (|\mu|, |\rho|)^{-1} K_{\text{diag}} + K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}} K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}}.$$}

From lemma 4.1 we know that the preconditioned Schur complement matrix $S_{\text{inexact}}^{-1} \in \mathbb{R}^{n \times n}$ is diagonalizable with a matrix of eigenvectors $V \in \mathbb{R}^{n \times n}$ (cf. Eq. (4.28)) that satisfies Eq. (4.46).

The same holds for the block Jacobi preconditioned matrix $K_J^{-1} K \in \mathbb{R}^{(2n) \times (2n)}$, as can be seen from Eq. (4.35):

$$Q^T \tilde{V}^{-1} K_J^{-1} K \tilde{V} Q = \begin{pmatrix} I + i \Lambda_F & 0 \\ 0 & I - i \Lambda_F \end{pmatrix},$$

where we reused the block diagonal matrix of eigenvectors $\tilde{V} \in \mathbb{R}^{(2n) \times (2n)}$ (cf. Eq. (4.33)) and the orthogonal matrix $Q \in \mathbb{R}^{(2n) \times (2n)}$ from lemma 4.2.

For the block Gauss-Seidel preconditioned matrix $K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)}$ we have to inspect Eq. (4.36):

$$\tilde{V}^{-1} K_{GS}^{-1} K \tilde{V} = \begin{pmatrix} I & \Lambda_F \\ 0 & 1 + \Lambda_F^2 \end{pmatrix}.$$}

As the blocks are diagonal, it suffices to look at the simplified model matrix:

$$\mathbf{R} \in \mathbb{R}^{2 \times 2}, \quad \mathbf{R} := \begin{pmatrix} 1 & \lambda_F \\ 0 & 1 + \lambda_F^2 \end{pmatrix}$$

with an eigenvalue $\lambda_F \in \sigma_F \left[ \sigma \left( M^{-1} A \right) \right]$. These blocks come up after an appropriate permutation.

Using $\varphi := \arctan(\lambda_F)$, we can identify a possible matrix of eigenvectors of the model matrix $\mathbf{R} \in \mathbb{R}^{2 \times 2}$ as the normalized matrix:

$$V_{\mathbf{R}} \in \mathbb{R}^{2 \times 2}, \quad V_{\mathbf{R}} := \begin{pmatrix} 1 & \cos(\varphi) \\ 0 & \sin(\varphi) \end{pmatrix}.$$}

Its condition number can be computed as

$$\kappa_2(V_{\mathbf{R}}) = \sqrt{\frac{1 + \cos(\varphi)}{1 - \cos(\varphi)}}.$$}

If now the eigenvalue $\lambda_F \in \sigma_F \left[ \sigma \left( M^{-1} A \right) \right]$ approaches zero, $\varphi$ does so as well, thus $\cos(\varphi)$ approaches one and the condition number $\kappa_2(V_{\mathbf{R}})$ goes to infinity and with it the condition number of the eigenvector basis of the block Gauss-Seidel preconditioned matrix $K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)}$.

While Eq. (4.47) gives us confidence that we will see immediate convergence if we use the block Jacobi preconditioned method $S_J$, the unboundedness of the eigenvector basis of the block Gauss-Seidel preconditioned matrix $K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)}$ should worry us. With an essentially arbitrarily large prefactor in the convergence result for the left preconditioned conjugate gradient method $S_{\text{LPGCR}}$ (cf. Eq. (3.63)), we could expect to see no residual reduction at all during the first iterations of the block Gauss-Seidel preconditioned method $S_{GS}$. However, the numerical experiments show no such effect! This is partly due to the following reasons: First, for a small eigenvalue $\lambda_F \in \sigma_F \left[ \sigma \left( M^{-1} A \right) \right]$, high powers $\mathbf{R}^n$ of the model matrix $\mathbf{R} \in \mathbb{R}^{2 \times 2}$ will be essentially diagonal, i.e. the offdiagonal term will quickly go to zero as the magnitude of the eigenvalue $|\lambda_F|$ goes to zero or as the power $n$ grows. Second, while the condition number of the matrix $V_{\mathbf{R}} \in \mathbb{R}^{2 \times 2}$ grows as the eigenvalue $\lambda_F \in \sigma_F \left[ \sigma \left( M^{-1} A \right) \right]$ goes to zero, the two eigenvalues of the model matrix $\mathbf{R} \in \mathbb{R}^{2 \times 2}$ approach each other even more quickly.

We can quantify these effects and show that the unboundedness of the condition number of the eigenvector matrix has only a negligible effect on the convergence of the block Gauss-Seidel preconditioned method $S_{GS}$. We state our results in the next lemma.
Lemma 4.5 (Contraction rates of the exactly preconditioned iterative solvers)
Let $\mu, \rho \in \mathbb{C}$ be the characteristic complex numbers as in Eq. (4.5), $\varphi \in [0, \pi/2]$ be the characteristic complex argument as in Eq. (4.8), $S_{\text{inexact}} \in \mathbb{R}^{n \times n}$ be the preconditioned Schur complement matrix as in Eq. (4.23), $K_1^{-1} \in \mathbb{R}^{(2n) \times (2n)}$ be the block Jacobi preconditioned matrix as in Eq. (4.16), $K_2^{-1} \in \mathbb{R}^{(2n) \times (2n)}$ be the block Gauss-Seidel preconditioned matrix as in Eq. (4.18) and we define the shorthand:

$$T := \max \{|\tan (\arg (\rho))|, |\tan (\arg (\mu))|\}.$$  \hfill (4.51)

In the limiting worst case $\sigma (\tau M^{-1} A) \to \mathbb{R}^+$ the following inequalities hold:

$$\|w_1 - w^{(k)}_{1,s}\| \leq 2 \sqrt{1 + \tan^2 (\varphi/2) - 1} \|w_1 - w^{(0)}_{1,s}\|,$$  \hfill (4.52)

$$\|K [w - w^{(k)}] \| \leq \kappa_2 (M) \left|\frac{T}{\sqrt{1 + T^2} + 1}\right|^k \|K [w - w^{(0)}]\|,$$  \hfill (4.53)

$$\|K [w - w^{(k)}] \| \leq \left[1 + \frac{k\pi}{T}\right] \kappa_2 (M) \frac{\sqrt{1 + T^2} - 1}{\sqrt{1 + T^2} + 1} \|K [w - w^{(0)}]\|.$$  \hfill (4.54)

Proof. For Eqs. (4.52) and (4.53) we just apply the standard convergence results presented in the previous chapter (cf. Eqs. (3.58) and (3.63)), combined with the bound on the eigenvector basis of the block Jacobi preconditioned matrix $K_1^{-1} \in \mathbb{R}^{(2n) \times (2n)}$ from lemma 4.1 and the bounds on the spectra of the preconditioned matrices from lemma 4.2.

For Eq. (4.54) we have to do a little bit more. First, note that in Eq. (3.63), the condition number of the eigenvector matrix of the preconditioned matrix as a prefactor comes from the application of lemma 3.6. Here we diagonalized the preconditioned matrix and pulled the matrix of eigenvectors out of the polynomial to be able to focus solely on minimizing the polynomial of the diagonal matrix of eigenvalues. But we do not have to fully diagonalize the preconditioned matrix. Recall the set of real polynomial functions of degree at most $k$ that have value one at the origin $Q^k$ (cf. Eq. (3.32)) defined in the previous section. As per Eq. (3.34), the inequality

$$\|K [w - w^{(k)}_{GS}]\| \leq \min_{q_k \in Q^k} \|q_k (K_2^{-1} K)\| \|K [w - w^{(0)}_{GS}]\|$$  \hfill (4.55)

holds. We now diagonalize only the blocks of the block Gauss-Seidel preconditioned matrix $K_2^{-1} \in \mathbb{R}^{(2n) \times (2n)}$ as in Eq. (4.36), reusing the block diagonal matrix of eigenvectors $V \in \mathbb{R}^{(2n) \times (2n)}$ (cf. Eq. (4.33)) as in lemma 4.2, yielding the inequality

$$\|K [w - w^{(k)}_{GS}]\| \leq \kappa_2 (V) \min_{q_k \in Q^k} \|q_k \begin{pmatrix} 1 & \Lambda_F \\ 0 & 1 + \Lambda_F^2 \end{pmatrix}\| \|K [w - w^{(0)}_{GS}]\|.$$  \hfill (4.56)

Recall that all non-zero entries of the diagonal matrix $\Lambda_F \in \mathbb{R}^{n \times n}$ (cf. Eq. (4.34)) have to be from the image $\sigma_F [\sigma (M^{-1} A)] \subset (\tan (\arg (\mu)), \tan (\arg (\rho)))$.

Let us now look again at the model matrix $R \in \mathbb{R}^{2 \times 2}$ (cf. Eq. (4.48)). It satifies the equality

$$R^n = \begin{pmatrix} 1 & \frac{(1 + \lambda_F^2)^n - 1}{\lambda_F^n} \\ 0 & \frac{(1 + \lambda_F^2)^n}{\lambda_F^n} \end{pmatrix},$$  \hfill (4.57)

and thus for any polynomial $q_k \in Q^k$ we get

$$q_k (R) = \begin{pmatrix} q_k (1) \\ 0 \end{pmatrix} \frac{q_k (1 + \lambda_F^2) - q_k (1)}{q_k (1 + \lambda_F^2)}.$$  \hfill (4.58)

Expanding the numerator of $\frac{q_k (1 + \lambda_F^2) - q_k (1)}{\lambda_F}$ reveals that the lowest-order term of the eigenvalue $\lambda_F \in \sigma_F [\sigma (M^{-1} A)]$ that remains in the numerator is of order two. This means that the term which caused the catastrophic growth of the condition number of the eigenvector matrix of the block Gauss-Seidel preconditioned matrix $K_2^{-1} \in \mathbb{R}^{(2n) \times (2n)}$ vanishes.

If we now had not the above upper triangular matrix, but only its diagonal part, the minimizing polynomial would be a transformed $k$-th degree Chebyshev polynomial of the first kind $T_k$, scaled and
shifted such that it has value one at the origin and the interval \([1, 1 + T^2]\) maps to the original interval \([-1, 1]\), cf. [Saa03, section 6.11.1]. Using the mapping:

\[
\xi : [1, 1 + T^2] \rightarrow [-1, 1], \quad x \mapsto -1 + \frac{2(x - 1)}{T^2}
\]  

we define this original optimal polynomial:

\[
p_k \in \mathbb{Q}^k, \quad p_k := \frac{T_k(\xi(x))}{T_k(\xi(0))}.
\]  

One of the defining characteristics of the Chebyshev polynomials is the fact that in the interval \([-1, 1]\) they take a value of at most one in modulus, and the extrema are located at

\[
\xi_i = \cos \left( \frac{\pi i}{k} \right) \quad i = 0, \ldots, k.
\]  

Therefore the original optimal polynomial \(p_k \in \mathbb{Q}^k\) has its extrema over the interval \([1, 1 + T^2]\) at \(\xi^{-1}(\xi_i)\) for \(i = 0, \ldots, k\), and one of those extrema is \(p_k(1)\). We now get

\[
\|p_k \left( \begin{bmatrix} 1 & \Lambda_F \\ 0 & I + \Lambda_E \end{bmatrix} \right) \| \leq |p_k(1)| \left[ 1 + \max_{\lambda \in [0, T]} \left\{ \frac{p_k(1) - p_k(1 + \lambda^2)}{\lambda p_k(1)} \right\} \right].
\]  

It remains to find a bound on the term being maximized in the above inequality. Let us define the mapping:

\[
f : [0, T] \rightarrow \mathbb{R}, \quad \lambda \mapsto \frac{p_k(1) - p_k(1 + \lambda^2)}{\lambda p_k(1)}.
\]  

We compute its derivative:

\[
f' : [0, T] \rightarrow \mathbb{R}, \quad \lambda \mapsto -\frac{2p_k'(1 + \lambda^2)}{p_k(1)} + \frac{p_k(1 + \lambda^2) - p_k(1)}{\lambda^2 p_k(1)}.\]  

We are looking for the maximum of the mapping \(f\), which is located at the extremal point \(\lambda_*\). As the absolute value of the original optimal polynomial \(p_k \in \mathbb{Q}^k\) is bounded by \(|p_k(1)|\) and \(1/\lambda\) is monotonically decreasing, we conclude that the maximum of the mapping \(f : [0, T] \rightarrow \mathbb{R}\) has to lie between the value of \(\lambda\) corresponding to the leftmost and the second leftmost extremum of the original optimal polynomial \(p_k \in \mathbb{Q}^k\) in the interval \([1, 1 + T^2]\). These are \(\lambda_0 = 0\) and

\[
\lambda_1 = \sqrt{1 + \xi^{-1}(-\xi_1)} = T \sqrt{\frac{1 - \cos(\pi/k)}{2}}.
\]  

We can compute the values of the derivative \(f'\) at these points as

\[
f'(0) = -\frac{p_k'(1)}{p_k(1)} > 0
\]  

and

\[
f'({\lambda_1}) = -\frac{2}{\lambda_1^2} < 0,
\]  

confirming that \(\lambda_* \in [0, \lambda_1]\) has to hold as the value of the derivative changed its sign. We note that both terms in the definition of the derivative \(f'\) (cf. Eq. (4.64)) are monotonically decreasing in the domain \([0, \lambda_1]\). Therefore, the derivative \(f'\) is maximal at \(\lambda = 0\), which implies

\[
f(\lambda_1) \leq \lambda_1 f'(0).
\]  

We have an expression for \(\lambda_1\) (cf. Eq. (4.65)), it remains to compute the second term. For this we first note that \(\frac{1}{|T_k(\xi(0))|} = |p_k(1)|\) holds, and thus per Eqs. (4.59) and (4.60) we get

\[
f'(0) = |T_k'((-1))| \xi'(1) = \frac{2k^2}{T^2}.
\]
Combining Eqs. (4.62), (4.65), (4.68) and (4.69) we obtain
\[
\left\| p_k \left( \begin{array}{cc} 1 & \Lambda_F \\ 0 & 1 + \Lambda_F^2 \end{array} \right) \right\| \leq |p_k(1)| \left[ 1 + \frac{k^2}{T} \sqrt{2(1 - \cos(\pi/k))} \right].
\]

The term \( k\sqrt{2(1 - \cos(\pi/k))} \) is monotonically increasing over the natural numbers bounded by \( \pi \). Combining this with the above inequality and with Eq. (4.56) we finally get Eq. (4.54), which completes the proof.

Let us put the bounds on the contraction rates (cf. Eqs. (4.52) to (4.54)) in perspective. First, let us note that if we choose \( \arg(\mu) = -\arg(\rho) \), the exponential terms in the iteration index \( k \) are identical for the preconditioned Schur complement method \( S_S \) and the block Gauss-Seidel preconditioned method \( S_{GS} \) (cf. Eqs. (4.52) and (4.54)). Second, the prefactor \( \sqrt{k_2(M)} \) for the block Jacobi preconditioned method \( S_J \) and the block Gauss-Seidel preconditioned method \( S_{GS} \) should be negligible, as the spatial mass matrix \( M \in \mathbb{R}^{n \times n} \) is usually well-conditioned. Next, while the additional term \( [1 + \frac{k^2}{T}] \) in Eq. (4.54) appears to be bad news, it has not that high of an impact. The linear term in the iteration index \( k \) will quickly be dominated by the exponential term, and even though the reciprocal of the short-hand \( T \) (cf. Eq. (4.51)) is unbounded, it will also be quickly be dominated by the exponential term. In fact, as
\[
\frac{\sqrt{1+T^2} - 1}{\sqrt{1+T^2} + 1} \ll \frac{1}{T}
\]
holds for \( T < 1 \), the effect of the reciprocal term will be more than undone after just one additional iteration. And finally, while the additional term for the block Gauss-Seidel preconditioned method \( S_{GS} \) is a rather sharp bound on Eq. (4.62), i.e. on the norm of the original optimal polynomial \( p_k \in \mathcal{Q}^k \) (cf. Eq. (4.60)) applied to the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \), there may possibly exist a different polynomial which yields a bound closer to the original estimate of the preconditioned Schur complement method \( S_S \).

The two bounds on the contraction rates of the block Jacobi preconditioned method \( S_J \) and the block Gauss-Seidel preconditioned method \( S_{GS} \) ensure that, if we apply the preconditioners exactly, we will see immediate convergence of the norms of the residuals. This is a common concern with non-symmetric iterative solvers, and while the block Jacobi preconditioned matrix \( K_J^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) was demonstrated to have a well-conditioned basis of eigenvectors, (cf. lemma 4.4), the eigenvector basis of the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) was demonstrated to be infinitely badly conditioned in the limiting worst case \( \sigma(\tau M^{-1} A) \rightarrow \mathbb{R}^+ \). This could have possibly led to many initial iterations of the block Gauss-Seidel preconditioned method \( S_{GS} \) where we observe no significant reduction of the residual norm.

Let us now summarize the results of this section. First, we provided the necessary ingredients for the analysis of the spectra of the preconditioned matrices, cf. lemma 4.1. Next, we applied this lemma to obtain explicit expressions for the spectra of the preconditioned matrices cf. lemma 4.2. We then discussed the impact of the discretization parameters on the spectra of the preconditioned matrices, i.e. the impact of the time step size \( \tau \in \mathbb{R}^+ \) (cf. ??), the spatial step size \( h \in \mathbb{R}^+ \) (cf. ??), the characteristic complex argument \( \varphi \in [0, \pi/2] \) (cf. Eq. (4.8)) and the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \) (cf. Eq. (4.5)). We then looked at the limiting worst case \( \sigma(\tau M^{-1} A) \rightarrow \mathbb{R}^+ \) (cf. Eq. (4.40)), and provided bounds on the spectra of the preconditioned matrices, cf. lemma 4.3. We then shortly discussed the globally optimal choice of the characteristic complex numbers \( \mu, \rho \in \mathbb{C} \), cf. Eqs. (4.44) and (4.45), to then proceed to discuss the condition number of the eigenvector bases of the preconditioned matrices, cf. lemma 4.4. This revealed the eigenvector basis of the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) to be potentially extremely badly-conditioned, rendering the usual convergence results for the used iterative method useless. Finally, as the conclusion of this section, we provided different bounds on the contraction rates of the iterative solvers, which for the block Gauss-Seidel preconditioned method \( S_{GS} \) relied on an incomplete diagonalization of the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) to obtain a bound which circumvents the bad condition number of the eigenvector basis of the fully diagonalized preconditioned matrix.

### 4.1.3 Efficiency of inexactly preconditioned methods

In the previous section we have discussed the convergence properties of the used methods under the assumption that we inverted the preconditioning matrices exactly. This however, will first of all be almost impossible for large problems, and more importantly, it will turn out to be extremely inefficient.
CHAPTER 4. PRECONDITIONERS FOR THE DISCRETE PROBLEM

Let us specify what we mean with efficiency in this context. In the application of the block Jacobi preconditioned method \( S_J \) (cf. ??), the block Gauss-Seidel preconditioned method \( S_{GS} \) (cf. ??) and the preconditioned Schur complement method \( S_S \) (cf. ??), we will approximate the needed inversions of matrices which are linear combinations of the Laplacian matrix \( A \in \mathbb{R}^{n \times n} \) and the spatial mass matrix \( M \in \mathbb{R}^{n \times n} \) by applying \( \kappa \in \mathbb{N} \) iterations of some stationary iterative method \( S \) to them. Let now \( w^{(x,k)} \in \mathbb{R}^{2n} \) denote the \( k \)-th approximation of the solution vector \( w \in \mathbb{R}^{2n} \) (cf. Eq. (4.12)) of the block system:

\[
K w \equiv \begin{pmatrix} K_{\text{diag}} & K_{\text{off}} \\ -K_{\text{off}} & K_{\text{diag}} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \equiv f
\]

generated by one of the above iterative methods using \( \kappa \in \mathbb{N} \) iterations of the stationary iterative method \( S \) per block and per iteration of the Krylov subspace method.

For the block Jacobi preconditioned method \( S_J \), the block Gauss-Seidel preconditioned method \( S_{GS} \) and the preconditioned Schur complement method \( S_S \) we use the same stationary iterative method \( S \) in the approximation of the preconditioner.

For the block Jacobi preconditioned method \( S_J \) and the block Gauss-Seidel preconditioned method \( S_{GS} \) we will use the left preconditioned conjugate gradient method \( S_{LPCG} \) (cf. Section 3.3.2), where the computational cost of the \( k \)-th iteration is the combined cost of \( 1 \) matrix-vector multiplication with the block matrix \( K \in \mathbb{R}^{(2n) \times (2n)} \), \( k + 2 \) scalar products and vector additions with vectors \( v, w \in \mathbb{R}^{2n} \) and of one application of the preconditioner. The cost of the approximated inversion of the block Jacobi preconditioner \( K_J \in \mathbb{R}^{(2n) \times (2n)} \) is \( 2k \) times the cost of one iteration of the stationary iterative method \( S \), while the cost of the approximated inversion of the block Gauss-Seidel preconditioner \( K_{GS} \in \mathbb{R}^{(2n) \times (2n)} \) is the same plus the cost of one matrix-vector multiplication with the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \).

For the preconditioned Schur complement method \( S_S \) we will use the left preconditioned conjugate gradient method \( S_{LPCG} \) (cf. Section 3.3.1), where the computational cost of the \( k \)-th iteration is the combined cost of \( 1 \) matrix-vector multiplication with the Schur complement matrix \( S \in \mathbb{R}^{n \times n}, 2 \) scalar products and vector additions with vectors \( v, w \in \mathbb{R}^n \) and of one application of the preconditioner. The cost of the approximated inversion of the matrix \( S \) is the same as for the approximated inversion of the block Gauss-Seidel preconditioner \( K_{GS} \in \mathbb{R}^{(2n) \times (2n)} \). One matrix-vector multiplication with the Schur complement matrix \( S \in \mathbb{R}^{n \times n} \) takes three matrix-vector multiplications with matrices like the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \), and one approximated inversion of the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) using the conjugate gradient method \( S_{CG} \) (cf. Section 3.2.1). The cost of this approximated inversion of the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) is variable. The needed accuracy depends on the accuracy we want to obtain in solving the Schur complement equation (cf. Eq. (4.20)). Given a number of iterations \( k' \in \mathbb{N} \) for the conjugate gradient method \( S_{CG} \) we get a cost of \( k' \) matrix-vector multiplications with matrices like the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) plus \( 2k' \) scalar products and vector additions with vectors \( v, w \in \mathbb{R}^n \). We note that while the matrix off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) is sparse, its average number of non-zeros per row is in the order of 20, which means the computational cost of multiplying with it is roughly equivalent to the cost of 20 scalar products and vector additions with vectors \( v, w \in \mathbb{R}^n \).

From this we conclude that while the length of the recurrence relation for the left preconditioned conjugate gradient method \( S_{LPCG} \) grows linearly with the iteration \( k \), and the length of the recurrence relation of the left preconditioned conjugate gradient method \( S_{LPCG} \) is fixed at one, one iteration of the preconditioned Schur complement method \( S_S \) will be slightly costlier in terms of computational effort than one iteration of the block Jacobi preconditioned method \( S_J \) or of the block Gauss-Seidel preconditioned method \( S_{GS} \).

This is due to the cost of the approximate inversion of the off-diagonal block \( K_{off} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n} \) using the conjugate gradient method \( S_{CG} \) outweighing the cost of the additional scalar products needed in the left preconditioned conjugate gradient method \( S_{LPCG} \) for usual values of \( k, k' \in \mathbb{N} \).

For the block multigrid preconditioned method \( S_{BMG} \) (cf. ??), the cost of one iteration is the same as for the block Jacobi preconditioned method \( S_J \) and the block Gauss-Seidel preconditioned method \( S_{GS} \), but with a different preconditioner. For the block multigrid preconditioned method \( S_{BMG} \) we use as preconditioner \( \kappa \) V-cycles of a block multigrid method per iteration, while we use \( 2k \) V-cycles of a \emph{scalar} multigrid method per iteration for the other three methods. In comparing the number of floating-point operations, we note that one V-cycle of the block multigrid method takes roughly quadruple the amount needed for one V-cycle of the scalar multigrid method. Some optimization with respect to the amount
of memory which has to be fetched for one V-cycle of the block multigrid method are possible if we have blocks that are bigger than $2 \times 2$. But in the current setting, we are just looking at $2 \times 2$-blocks, and we conclude that we have to compare the efficiency $e_{k,s}^\text{(B)}$ (cf. Eq. (4.103)) of the block multigrid preconditioned method $S_{\text{BMG}}$ the the squared efficiency $e_{k,s}^2$ of the other methods.

4.1.4 Convergence analysis for the inexacty preconditioned methods

Let us now look at a simplified model problem. Let us assume we have to solve a linear system involving an arbitrary, invertible matrix $A \in \mathbb{R}^{n \times n}$. We want to solve it using a preconditioned Krylov subspace method, using the preconditioning matrix $M \approx A$. Obviously, if the preconditioning matrix $M \approx A$ were equal to the matrix $A \in \mathbb{R}^{n \times n}$, our preconditioned Krylov subspace method would converge after just one iteration. But, if we could cheaply obtain an exact inverse of the matrix $A \in \mathbb{R}^{n \times n}$ there would be no need to use an iterative method in the first place. Therefore, the question remains as to how to most efficiently obtain a sufficiently good approximation of the inverse of the matrix $A \in \mathbb{R}^{n \times n}$ to be used as the preconditioner, such that our Krylov subspace method converges immediately. The best, realistically obtainable precision will be bounded by the used floating-point precision. Therefore our best option to obtain an exact (in floating point arithmetic) approximation of the inverse of the matrix $A \in \mathbb{R}^{n \times n}$ will not be a direct method, but a stationary iterative method $S$ which has a small contraction rate independent of the problem size. This may be for example a multigrid method, yielding linear complexity to obtain our exact approximation of the inverse of the matrix $A \in \mathbb{R}^{n \times n}$.

Let us now assume we have chosen our stationary iterative method $S$ with a diagonalizable iteration matrix:

$$S \in \mathbb{R}^{n \times n}, \quad S = I - M^{-1}A. \quad (4.70)$$

Using $k$ iterations of this stationary iterative method $S$ as a preconditioner for some Krylov subspace method gives us the preconditioned matrix:

$$M_k^{-1}A \in \mathbb{R}^{n \times n}, \quad M_k^{-1}A = I - S^k \quad (4.71)$$

where we used the $k$-th approximation matrix:

$$M_k \in \mathbb{R}^{n \times n} \quad \lim_{k \to \infty} M_k = A. \quad (4.72)$$

Let us furthermore assume for the sake of argument that the eigenvalues of the iteration matrix $S \in \mathbb{R}^{n \times n}$ fully populate the boundary of an ellipse in the complex plane centered at $c \in \mathbb{C}$ with focal distance $d \in \mathbb{C}$ and semi-major axis $a \in \mathbb{C}$:

$$E : \mathbb{C}^3 \to \mathbb{P}(\mathbb{C}), \quad (c, d, a) \mapsto \{ z \in \mathbb{C} | |z - c| + |z - c - d| \leq 2|a| \} \quad (4.73)$$

with center $c = 0$, and $|d|, |a| < 1$. Then the eigenvalues of the once preconditioned matrix $M_1^{-1}A \in \mathbb{R}^{n \times n}$ are located in the same ellipse, but shifted to be centered around $\lambda = 1$.

Recall now that the convergence behaviour of a Krylov subspace method is governed by the minimal maximum-norm over the set of real polynomial functions of degree at most $k$ that have value one at the origin $Q^k$:

$$M_k : \mathbb{P}(\mathbb{C}) \to \mathbb{R}^+, \quad \sigma \mapsto \min_{q_k \in Q^k} \max_{\lambda \in \sigma} |q_k(\lambda)|. \quad (4.74)$$

In the last chapter we introduced an asymptotically sharp bound for the value of this term if its argument is from the image of $E : \mathbb{C}^3 \to \mathbb{P}(\mathbb{C})$:

$$M_k(E(c, d, a)) \leq \left[ \frac{a + \sqrt{a^2 - d^2}}{c + \sqrt{c^2 - d^2}} \right]^k. \quad (4.75)$$

Let us now define the set:

$$E_k := \{ 1 - \lambda^k | \lambda \in S \} \subset \mathbb{C} \quad (4.76)$$

containing the eigenvalues of the $k$ times preconditioned matrix $M_k^{-1}A \in \mathbb{R}^{n \times n}$. Estimating the efficiency of $k$ times applying the stationary iterative method $S$ now comes down to estimating the effect this has on the set $E_k$ and thusly on $M_1(E_k)$, and comparing this with the behaviour of $M_k(E_1)$.
It will be illustrative to look at the two limiting cases $E_1 = E(1, 0, a)$ with $a \in \mathbb{C}$, which corresponds to a perfect circle, and $E_1 = E(1, a, a)$ with $a \in \mathbb{R}$, which corresponds to a line segment on the real axis.

The first case yields
\[ \mathcal{M}_1(E_\kappa) = \mathcal{M}_\kappa(E_1) = |a|^k, \]

i.e. it does not make a difference how many times we apply the stationary iterative method $\mathcal{S}$, with respect to the contraction rate.

The second case highlights the reason why it generally is more efficient to just apply the stationary iterative method $\mathcal{S}$ once. For $k = 2, \ldots$, the squared ellipse becomes $E_\kappa = E(1 - a^k/2, a^k/2, a^k/2)$. and we get
\[ \mathcal{M}_1(E_\kappa) \approx \left| \frac{a^k}{2 - a^k + 2\sqrt{1 - 2a^k}} \right| > \left| \frac{a}{1 + \sqrt{1 - a^2}} \right|^k \approx \mathcal{M}_\kappa(E_1) \]
for all $k = 2, \ldots$ and $|a| < 1$.

The reason why the once preconditioned method has to be more efficient is the following. Applying the preconditioner more than once is equivalent to picking one particular polynomial $q_\kappa \in \mathcal{Q}^k$ for the once preconditioned method. If the ellipse is a circle, this polynomial just happens to be the optimal one. But it can never be better than the optimal polynomial minimizing Eq. (4.75), and in fact if the ellipse reduces to a line segment on the real axis, it is very far from being the optimal one.

The above explanation can be made more rigorous using the definition of $\mathcal{M}_\kappa$ (cf. Eq. (4.74)). It gives us
\[ \mathcal{M}_\kappa(E_m) \equiv \min_{q_\kappa \in \mathcal{Q}^k \lambda \in \sigma(E)} \max_{q_{km} \in \mathcal{Q}^{km} \lambda \in \sigma(E)} |q_k(1 - \lambda^m)| \geq \min_{q_{km} \in \mathcal{Q}^{km} \lambda \in \sigma(E)} \max_{q_\kappa \in \mathcal{Q}^k} |q_{km}(1 - \lambda)| \equiv \mathcal{M}_{km}(E_1), \]
as for any polynomial $q_\kappa \in \mathcal{Q}^k$ there exists a polynomial $q_{km} \in \mathcal{Q}^{km}$ such that $q_k(1 - \lambda^m) = q_{km}(1 - \lambda)$ holds for all $\lambda \in \mathbb{C}$. We conclude the following:

If we use $k$ applications of a stationary iterative method $\mathcal{S}$ as the preconditioner for a Krylov subspace method, the minimal total number of applications of the stationary iterative method $\mathcal{S}$ to obtain a given residual reduction is achieved for $k = 1$.

This is mainly relevant for the block multigrid preconditioned method $\mathcal{S}_{BMG}$ (cf. ??). For the other three methods, i.e. the preconditioned Schur complement method $\mathcal{S}_\mathcal{S}$ (cf. ??), the block Jacobi preconditioned method $\mathcal{S}_J$ (cf. ??) and the block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ (cf. ??) the difference in efficiency for the inexact ($\kappa = 1$) and for the exact ($\kappa \to \infty$) application of the preconditioner is even more pronounced. The reason for this appears to be that, as we increase the accuracy of the application of the preconditioner, the preconditioned matrices do not converge to the identity matrix.

To illustrate this, let us inspect the block Jacobi preconditioned method $\mathcal{S}_J$ applied to the block matrix $\mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$ (cf. Eq. (4.11)). The preconditioning matrix is the block Jacobi preconditioner:
\[ \mathbf{K}_J \in \mathbb{R}^{(2n) \times (2n)}, \quad \mathbf{K}_J := \begin{pmatrix} \mathbf{K}_{\text{diag}} & 0 \\ 0 & \mathbf{K}_{\text{diag}} \end{pmatrix}. \]

We now approximate the inversion of this matrix by applying a stationary iterative method $\mathcal{S}$ to the diagonal block $\mathbf{K}_{\text{diag}}$ in Image($\mathbf{K}_J$) $\in \mathbb{R}^{n \times n}$ (cf. Eq. (4.9)). We again use the iteration matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ (cf. Eq. (4.70)) and the approximation matrix $\mathbf{M}_\kappa$, and get
\[ \lim_{\kappa \to \infty} \mathbf{M}_\kappa = \mathbf{K}_{\text{diag}}. \]

This stands in contrast to Eq. (4.72), where the preconditioning matrix converges to the governing matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, which would be the block matrix $\mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$ in our case.

We define the inexact block Jacobi preconditioner:
\[ \mathbf{K}_{J,\kappa} \in \mathbb{R}^{(2n) \times (2n)}, \quad \mathbf{K}_{J,\kappa} := \begin{pmatrix} \mathbf{M}_\kappa & 0 \\ 0 & \mathbf{M}_\kappa \end{pmatrix}, \]

and the inexact block Jacobi preconditioned matrix:
\[ \mathbf{K}_{J,\kappa}^{-1} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}, \quad \mathbf{K}_{J,\kappa}^{-1} \mathbf{K} = \begin{pmatrix} \mathbf{I} - \mathbf{S}^\kappa & 0 \\ 0 & \mathbf{I} - \mathbf{S}^\kappa \end{pmatrix} \mathbf{K}_{J,\kappa}^{-1} \mathbf{K}. \]

This again stands in contrast to Eq. (4.71), where we only had the first term. This time, the inexact block Jacobi preconditioned matrix $\mathbf{K}_{J,\kappa}^{-1} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$ converges towards the block Jacobi preconditioned matrix $\mathbf{K}_{J}^{-1} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$, i.e. the second term in Eq. (4.79) as $\kappa$ goes to infinity.
CHAPTER 4. PRECONDITIONERS FOR THE DISCRETE PROBLEM

Contrary to the previous setting, where we could provide an explicit expression for the spectrum of the preconditioned matrix $M^{-1} \mathbf{A} \in \mathbb{R}^{n \times n}$ (cf. Eq. (4.76)), we do not believe it to be possible to provide a similar expression for the spectra of the inexact block Jacobi preconditioned matrix $K^{-1}_{j,k} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$. What we can do, however, is provide bounds on the possible locations of the eigenvalues of the inexact block Jacobi preconditioned matrix $K^{-1}_{j,k} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$, even though these bounds will only be informative for large values of $\kappa$.

For this, let us first introduce a standard result from perturbation theory, which provides a bound on the $\varepsilon$-pseudospectrum:

$$
\sigma_\varepsilon (\mathbf{A}) := \left\{ \mu \in \mathbb{C} \mid \exists \Delta_z \in \mathbb{C}^{n \times n} \text{ with } \|\Delta_z\| \leq \varepsilon \text{ such that } \mu \in \sigma (\mathbf{A} + \Delta_z) \right\}.
$$

**Lemma 4.6** (The Bauer-Fike Theorem) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a diagonalizable matrix with matrix of eigenvectors $\mathbf{V} \in \mathbb{R}^{n \times n}$ and let $\varepsilon := \|\Delta_z\|$ denote the spectral norm of a perturbation matrix $\Delta_z \in \mathbb{R}^{n \times n}$. Then for each perturbed eigenvalue $\mu \in \sigma (\mathbf{A} + \Delta_z)$ there exists an eigenvalue $\lambda \in \sigma (\mathbf{A})$ such that the inequality

$$
|\lambda - \mu| \leq \varepsilon \kappa_2 (\mathbf{V})
$$

holds.

**Proof.** A proof can be found in [BF60].

Recall that there exists a matrix of eigenvectors of the block Jacobi preconditioned matrix $K^{-1}_{j,k} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$ whose condition number is bounded by $\sqrt{\kappa_2 (\mathbf{M})}$ (cf. lemma 4.4), with the spatial mass matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$. We can now apply lemma 4.6 to bound the $\varepsilon$-pseudospectrum $\sigma_\varepsilon (K^{-1}_{j,k} \mathbf{K})$ with $\varepsilon = \|S^n\| \|K^{-1}_{j,k} \mathbf{K}\|$, which gives us

$$
\sigma (K^{-1}_{j,k} \mathbf{K}) \subset \left\{ \mu \in \mathbb{C} \mid \exists \lambda \in \sigma (K^{-1}_{j,k} \mathbf{K}) \text{ such that } |\lambda - \mu| \leq \|S^n\| \|K^{-1}_{j,k} \mathbf{K}\| \sqrt{\kappa_2 (\mathbf{M})} \right\}.
$$

From the above equation we can already conclude that the equality

$$
\lim_{\kappa \to \infty} \sigma (K^{-1}_{j,k} \mathbf{K}) = \sigma (K^{-1}_{j,k} \mathbf{K})
$$

holds, as the norm of the $\kappa$-th power of the iteration matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ converges towards zero. However, if we want to apply this to the analogue of the block Gauss-Seidel preconditioned method $S_{GS}$, the potentially infinitely badly conditioned basis of eigenvectors of the block Gauss-Seidel preconditioned matrix $K^{-1}_{GS} \mathbf{K} \in \mathbb{R}^{(2n) \times (2n)}$ renders the bound less than ideal.

As such, we want to provide a slightly modified and adapted version of lemma 4.6, which provides a bound on the multiplicative $\varepsilon$-pseudospectrum:

$$
\tilde{\sigma}_\varepsilon (\mathbf{A}) := \left\{ \mu \in \mathbb{C} \mid \exists \Delta_z \in \mathbb{C}^{n \times n} \text{ with } \|\Delta_z\| \leq \varepsilon \text{ such that } \mu \in \sigma ([I + \Delta_z] \mathbf{A}) \right\}.
$$

**Lemma 4.7** (The modified Bauer-Fike Theorem) Let $\mathbf{A}, \mathbf{V}, \mathbf{A} \in \mathbb{R}^{n \times n}$ be invertible matrices, let the equality $\mathbf{A} = \mathbf{V} \Delta \mathbf{V}^{-1}$ hold and let $\varepsilon := \|\Delta_z\|$ denote the spectral norm of a perturbation matrix $\Delta_z \in \mathbb{R}^{n \times n}$. Then for each perturbed eigenvalue $\mu \in \sigma ([I + \Delta_z] \mathbf{A})$ the inequality

$$
\sigma_{\min} (\Delta^{-1} [\mathbf{A} - \mu I]) \leq \varepsilon \kappa_2 (\mathbf{V})
$$

holds, where $\sigma_{\min} (\mathbf{A})$ denotes the minimal singular value of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

**Proof.** We can assume that $\mu \notin \sigma (\mathbf{A})$ holds, otherwise Eq. (4.83) is trivially true as the minimal singular value would be zero. We know that the perturbed eigenvalue $\mu \in \sigma ([I + \Delta_z] \mathbf{A})$ is a root of the characteristic polynomial of the multiplicatively perturbed matrix $[I + \Delta_z] \mathbf{A} \in \mathbb{R}^{n \times n}$, ergo

$$
0 = \det ([I + \Delta_z] \mathbf{A} - \mu I)
$$

holds. We can pull out the non-zero factor $\det (\mathbf{A} - \mu I)$ and apply the similarity transformation with $\mathbf{V} \in \mathbb{R}^{n \times n}$ to obtain

$$
0 = \det \left(I + \mathbf{V}^{-1} \Delta_z \mathbf{V} \mathbf{A} [\mathbf{A} - \mu I]^{-1}\right).
$$

4.1. CONVERGENCE ANALYSIS
Therefore \(-1 \in \sigma \left( V^{-1} \Delta \mu A \right)^{-1} \) holds and thus the inequality
\[
1 \leq \left\| V^{-1} \Delta \mu A \right\|^{-1} \leq \varepsilon \kappa_2 (V),
\]
where the identity \( \sigma_{\text{max}} (A^{-1}) = \sigma_{\text{min}} (A)^{-1} \) finally yields Eq. (4.83).

We can now apply lemma 4.7 to the analogue of the block Gauss-Seidel preconditioned method \( S_{GS} \), again only partially diagonalizing the block Gauss-Seidel preconditioned matrix \( K_{GS}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \).

Before we do this, let us first define the *inexact block Gauss-Seidel preconditioner*:
\[
K_{GS,k} \in \mathbb{R}^{(2n) \times (2n)}, \quad K_{GS,k} := \begin{pmatrix} M_k & 0 \\ -K_{off} & M_k \end{pmatrix}
\tag{4.84}
\]
and the *inexact Schur complement preconditioner*:
\[
S_{S,k} \in \mathbb{R}^{n \times n}, \quad S_{S,k} := M_k K_{off}^{-1} M_k.
\tag{4.85}
\]
For the preconditioned Schur complement method \( S_k \) the approximation matrix \( M_k \) converges towards the preconditioning matrix \( K_1 (|\mu|, |\rho|) \in \text{Image} (K_1) \), which is usually a somewhat different linear combination of the Laplacian matrix \( A \in \mathbb{R}^{n \times n} \) and the spatial mass matrix \( M \in \mathbb{R}^{n \times n} \) than the diagonal block \( K_{\text{diag}} \in \text{Image} (K_1) \subset \mathbb{R}^{n \times n} \) (cf. Eq. (4.9)). In fact, both the approximation matrix \( M_k \) and the iteration matrix \( S \in \mathbb{R}^{n \times n} \) will be different for the preconditioned Schur complement method \( S_k \), but we will nevertheless use the same symbols as for the respective matrices of the block Jacobi preconditioned method \( S_j \) and for the block Gauss-Seidel preconditioned method \( S_{GS} \).

**Lemma 4.8** (Bounds on the spectra of the inexactly preconditioned matrices)
Let \( K_{J,k}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Jacobi preconditioned matrix as in Eq. (4.16), \( K_{GS,k}^{-1} K \in \mathbb{R}^{(2n) \times (2n)} \) be the block Gauss-Seidel preconditioned matrix as in Eq. (4.18), \( S_{S,k}^{-1} S \in \mathbb{R}^{n \times n} \) be the preconditioned Schur complement matrix as in Eq. (4.23), let \( K_{J,k} \in \mathbb{R}^{(2n) \times (2n)} \) be the inexact block Jacobi preconditioner as in Eq. (4.78), \( K_{GS,k} \in \mathbb{R}^{(2n) \times (2n)} \) be the inexact block Gauss-Seidel preconditioner as in Eq. (4.84), \( S_{S,k} \in \mathbb{R}^{n \times n} \) be the inexact Schur complement preconditioner as in Eq. (4.85) and let the iteration matrix \( S \in \mathbb{R}^{n \times n} \) be the iteration matrix of the respective iterative solver used to approximate the inexact Schur complement (cf. Eqs. (4.78), (4.84) and (4.85)). We reuse the shorthand \( T := \max \{ |\tan (|\rho|)|, |\tan (|\mu|)| \} \) (cf. Eq. (4.51)) and define the functions
\[
\begin{align*}
\sigma_J (\lambda, \mu) &:= \left| \frac{\lambda - \mu}{\lambda} \right|, \\
\sigma_{GS} (\lambda, \mu) &:= \frac{1}{2} \sqrt{2 \min \{|1 - |\mu|, |1 - |\mu|/\lambda|\}^2 + \frac{1}{\lambda^2} |1 - \lambda| - |\mu| \sqrt{1 - \lambda}} \lambda, \\
\Delta_{GS} (\varepsilon) &:= \varepsilon \left[ 1 + (1 + \varepsilon) \sqrt{\kappa_2 (M) T} \right], \\
\Delta_S (\varepsilon) &:= \varepsilon \left[ 1 + (1 + \varepsilon) \kappa_2 \left( K_1 (|\mu|, |\rho|) \right)^{-1} K_{off} \right].
\end{align*}
\tag{4.86-4.89}
\]

Then the following bounds hold for the spectra of the inexactly preconditioned matrices:
\[
\begin{align*}
\sigma (K_{J,k}^{-1} K) &\subset \left\{ \mu \in \mathbb{C} \mid \exists \sigma \in (K_{J,k}^{-1} K) \text{ such that } \sigma_J (\lambda, \mu) \leq \| S^\sigma \| \sqrt{\kappa_2 (M)} \right\}, \\
\sigma (K_{GS,k} K) &\subset \left\{ \mu \in \mathbb{C} \mid \exists \sigma \in (K_{GS,k}^{-1} K) \text{ such that } \sigma_{GS} (\lambda, \mu) \leq \Delta_{GS} (\| S^\sigma \|) \sqrt{\kappa_2 (M)} \right\}, \\
\sigma (S_{S,k} S) &\subset \left\{ \mu \in \mathbb{C} \mid \exists \sigma \in (S_{S,k}^{-1} S) \text{ such that } \sigma_J (\lambda, \mu) \leq \Delta_S (\| S^\sigma \|) \sqrt{\kappa_2 (M)} \right\}.
\end{align*}
\tag{4.90-4.92}
\]

**Remark 4.2.** Before we present the proof, let us comment on the use of the functions \( \sigma_J (\lambda, \mu) \) and \( \sigma_{GS} (\lambda, \mu) \) and \( \Delta_{GS} (\varepsilon) \) and \( \Delta_S (\varepsilon) \). We want to apply lemma 4.7 to bound the multiplicative \( \varepsilon \)-pseudospectrum \( \tilde{\sigma}_\varepsilon (A) \) of the preconditioned matrices. The functions \( \sigma_J (\lambda, \mu) \) and \( \sigma_{GS} (\lambda, \mu) \) correspond to lower bounds on the left hand side of Eq. (4.83), while the functions \( \Delta_{GS} (\varepsilon) \) and \( \Delta_S (\varepsilon) \) are needed for upper bounds on the norm of the perturbation matrix \( \Delta \in \mathbb{R}^{n \times n} \).
Proof. For the block Jacobi preconditioned method $S_J$ (Eq. (4.90)) we have for the perturbation matrix $\Delta_\varepsilon \in \mathbb{R}^{n \times n}$ the equality

$$
\Delta_\varepsilon = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} M_\kappa & 0 \\ 0 & M_\kappa \end{pmatrix}^{-1} \begin{pmatrix} K_{\text{diag}} & 0 \\ 0 & K_{\text{diag}} \end{pmatrix} = \begin{pmatrix} S_\kappa & 0 \\ 0 & S_\kappa \end{pmatrix}
$$

and we choose as $V \in \mathbb{R}^{n \times n}$ a matrix of eigenvectors of the block Jacobi preconditioned matrix $K_J^{-1} K \in \mathbb{R}^{(2n)\times (2n)}$, which yields as $A \in \mathbb{R}^{n \times n}$ a diagonal matrix of eigenvalues of the block Jacobi preconditioned matrix $K_J^{-1} K \in \mathbb{R}^{(2n)\times (2n)}$. For a diagonal matrix $A \in \mathbb{R}^{n \times n}$, the equality

$$
\sigma_{\min} (A^{-1} (A - \mu I)) = \min_{\lambda \in \sigma(A)} \left| \frac{\lambda - \mu}{\lambda} \right| \equiv \min_{\lambda \in \sigma(A)} \sigma_J (\lambda, \mu).
$$

holds which gives us $\sigma_J (\lambda, \mu)$.

For the preconditioned Schur complement method $S_S$ (Eq. (4.92)) we again choose a matrix of eigenvectors of the preconditioned Schur complement matrix $S_{\text{inexact}}^T S \in \mathbb{R}^{n \times n}$, which also gives us $\sigma_J (\lambda, \mu)$ on the left hand side. For the norm of the perturbation matrix $\Delta_\varepsilon \in \mathbb{R}^{n \times n}$ we first have to get an expression for it which depends on the iteration matrix $S \in \mathbb{R}^{n \times n}$:

$$
\Delta_\varepsilon = I - \begin{pmatrix} M_\kappa & 0 \\ 0 & M_\kappa \end{pmatrix}^{-1} K_1 (|\mu|, |\rho|) K_{\text{off}}^{-1} K_1 (|\mu|, |\rho|).
$$

As here we have the equality $M_\kappa^{-1} K_1 (|\mu|, |\rho|) = I - S$ we get

$$
\Delta_\varepsilon = S + [I - S] K_1 (|\mu|, |\rho|)^{-1} K_{\text{off}} S K_1 (|\mu|, |\rho|)
$$

and thus the rather coarse bound

$$
\|\Delta_\varepsilon\| \leq \|S\| \left[ 1 + (1 + \|S\|) \kappa_2 \left( K_1 (|\mu|, |\rho|)\right)^{-1} K_{\text{off}} \right] \equiv \Delta_S (\|S\|)
$$

which gives us $\Delta_S (\varepsilon)$.

For the block Gauss-Seidel preconditioned method $S_{\text{GS}}$ (Eq. (4.91)) we again have to do a little bit more due to its potentially badly-conditioned eigenvector basis. Let us first get an estimate for the norm of the perturbation matrix $\Delta_\varepsilon \in \mathbb{R}^{n \times n}$. We have

$$
\Delta_\varepsilon = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} M_\kappa & 0 \\ -K_{\text{off}} & M_\kappa \end{pmatrix}^{-1} \begin{pmatrix} K_{\text{diag}} & 0 \\ -K_{\text{off}} & K_{\text{diag}} \end{pmatrix} = \begin{pmatrix} S_\kappa & 0 \\ 0 & S_\kappa \end{pmatrix}.
$$

This yields

$$
\|\Delta_\varepsilon\| \leq \|S_\kappa\| \left[ 1 + \|I - S_\kappa\| K_{\text{diag}}^{-1} K_{\text{off}} \right] \leq \|S_\kappa\| \left[ 1 + [1 + S_\kappa] \sqrt{\kappa_2 (M)^T} \right] \equiv \Delta_{\text{GS}} (\|S_\kappa\|),
$$

where the second inequality is due to the condition number of the eigenvector basis of $K_{\text{diag}}^{-1} K_{\text{off}}$ (cf. lemma 4.1) and the bound on $\lambda_{\max} (K_{\text{diag}}^{-1} K_{\text{off}}) \leq T$ (cf. ??).

The left hand side is only slightly harder to obtain. For this, we choose in the application of lemma 4.7 for the transformation matrix the block diagonal matrix of eigenvectors $V \in \mathbb{R}^{(2n)\times (2n)}$ (cf. Eq. (4.33)), i.e. we again only diagonalize the blocks of the block Gauss-Seidel preconditioned matrix $K_{\text{GS}}^{-1} K \in \mathbb{R}^{(2n)\times (2n)}$. We now have to get a lower bound on

$$
\sigma_{\min} \left( \begin{pmatrix} I & A_F \\ 0 & I + A_F^2 \end{pmatrix}^{-1} \begin{pmatrix} I & A_F \\ 0 & I + A_F^2 \end{pmatrix} - \mu \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \right),
$$

where we reused the diagonal matrix $A_F \in \mathbb{R}^{n \times n}$ (cf. Eq. (4.34)). The above expression is equal to

$$
\min_{\lambda \in \sigma(A_F)} \sigma_{\min} \left( \begin{pmatrix} 1 - \mu & \frac{\mu \lambda}{1 + \mu} \\ 0 & 1 - \mu \lambda \end{pmatrix} \right) = \min_{\lambda \in \sigma(K_{\text{off}} K)} \sigma_{\min} \left( \begin{pmatrix} 1 - \mu & \frac{\mu \lambda - 1}{\lambda} \\ 0 & 1 - \frac{\mu}{\lambda} \end{pmatrix} \right).
$$

We can estimate the smallest singular value of a matrix using a relation obtained from the Gershgorin circle theorem, cf. [HP92, lemma 0]. Applied to an upper triangular complex $2 \times 2$ matrix

$$
A := \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in \mathbb{C}^{2 \times 2}
$$

4.1. CONVERGENCE ANALYSIS
we get
\[
\sigma_{\text{min}}(A) \geq \frac{1}{2} \left( \sqrt{4 \min \{|a|,|d|\}^2 + |b|^2 - |b|^2} \right).
\]

Using the above inequality and rearranging some terms gives us
\[
\sigma_{\text{min}} \left( \begin{pmatrix} 1 - \mu & -\mu \lambda J - (\frac{\lambda}{2} \Delta_1) \end{pmatrix} \right) \geq \frac{1}{2} \left( \lambda \sqrt{4 \min \{|1-\mu|,|1-\mu/\lambda|\}^2 + \left| \frac{\lambda}{2} \right|^2 (1-\lambda) - |\mu| \sqrt{1-\lambda}} \right),
\]
which is just the definition of \(\sigma_{\text{GS}}(\lambda, \mu)\) (cf. Eq. (4.87)).

From The modified Bauer-Fike Theorem (cf. lemma 4.7) we have now gotten inequalities of the form
\[
\min_{\lambda \in \sigma(A)} f(\lambda, \mu) \leq g(||\Delta_e||) \quad \text{for all} \quad \mu \in \sigma([1 + \Delta_e] A)
\]
for each inexactly preconditioned method. This implies that if for a \(\mu \in \mathbb{C}\) there is no unperturbed eigenvalue \(\lambda \in \sigma(A)\) such that \(f(\lambda, \mu) \leq g(||\Delta_e||)\) holds, then that \(\mu \in \mathbb{C}\) cannot be contained in the multiplicative \(\varepsilon\)-pseudospectrum \(\sigma_e(A)\) (cf. Eq. (4.82)), which is just the the implication of Eqs. (4.90) to (4.92).

We want to discuss the above lemma and its implications. Let us start with identifying the settings where the lemma can provide us with useful information and the ones were it cannot.

It gives us useful information about the possible locations of the eigenvalues of the inexactly preconditioned matrices, if the norm of the perturbation matrix \(\Delta_e \in \mathbb{R}^{n \times n}\) is much smaller than one. This implies in our setting \(||S^*|| \ll 1\) which in turn implies \(\kappa \gg 1\). If that is so, the lemma bounds the possible location of the eigenvalues of the inexactly preconditioned matrices by a set that is reasonably close to the spectra of the exactly preconditioned matrices. It also tells us how sensitive each exactly preconditioned matrix is to small perturbations.

The block Jacobi preconditioned matrix \(K_1^{-1} K \in \mathbb{R}^{(2n) \times (2n)}\) is the least sensitive, independent of all discretization parameters. The preconditioned Schur complement matrix \(S_{\text{inexact}}^1 \in \mathbb{R}^{n \times n}\) is also relatively insensitive, but it depends on the time step size \(\tau \in \mathbb{R}^+\) and the spatial step size \(h \in \mathbb{R}^+\), as these influence the condition number \(\kappa_2 \left( |K_1|, |\rho\right)^{-1} K_{\text{off}} \). The block Gauss-Seidel preconditioned matrix \(K_{\text{GS}}^{-1} K \in \mathbb{R}^{(2n) \times (2n)}\) is the most sensitive to small perturbations, but its sensitivity depends only very weakly on the discretization parameters. Its increased sensitivity is due to the potentially decreased effect of small perturbations in \(\mu\) on \(\sigma_{\text{GS}}(\lambda, \mu)\) (cf. Eq. (4.87)). Nevertheless, if the norm of the perturbation matrix \(\Delta_e \in \mathbb{R}^{n \times n}\) is moderately small, the eigenvalues of the inexactly block Gauss-Seidel preconditioned matrix can be guaranteed to be close to the eigenvalues of the exactly preconditioned matrix.

In Fig. 4.2 we show filled contour plots of the mapping:
\[
\sigma_{\text{min*,}} : \mathbb{C} \rightarrow \mathbb{R}^+,
\]
\[
\lambda \mapsto \min_{\lambda \in [1,2]} \sigma_*(\lambda, \mu)
\]
over \(\mu \in \mathbb{C}\) for \(\sigma_{\text{GS}}(\lambda, \mu)\) (cf. Eq. (4.87)) (left) and for \(\sigma_3(\lambda, \mu)\) (cf. Eq. (4.86)) (right). Each level (i.e. colour) corresponds to a value range \(\sigma_{\text{min*,}}(\mu) \in [2^i, 2^{i+1}]\) for \(i = -20, \ldots, -2\). This differently shaded areas correspond to possible locations of the perturbed eigenvalues for different accuracies of the preconditioner, where the unperturbed eigenvalues fill in the interval \([1, 2]\). The lightest colour (white) marks the area where \(\sigma_{\text{min*,}}(\mu) > 0.25\) holds, the second lightest colour corresponds to the area where \(\sigma_{\text{min*,}}(\mu) \in (0.125, 0.25]\) holds and so on. Each colour step can be understood as a doubling of the accuracy of the iterative method used in the inexact preconditioner. We can observe the decreased sensitivity of \(\sigma_{\text{GS}}(\lambda, \mu)\) with respect to perturbations in \(\mu\) compared to \(\sigma_3(\lambda, \mu)\), and thus the enlarged set of possible locations of the perturbed eigenvalues.

Let us now discuss the case where lemma 4.8 does not help us. This is first and foremost the case \(\kappa = 1\), i.e. the coarsest possible application of the preconditioner. This is because even if the used stationary iterative method exhibits good asymptotic convergence properties, identified by a small spectral radius of its iteration matrix \(S \in \mathbb{R}^{n \times n}\), the norm of this matrix is not necessarily small or monotonically decreasing in the number of iterations. This may happen for example if all eigenvalues of the iteration matrix \(S \in \mathbb{R}^{n \times n}\) are small in modulus, but its eigenvector basis is badly-conditioned.

For the same reason, the lemma does not tell us what happens if we increase \(\kappa\). It provides us with bounds on the multiplicative \(\varepsilon\)-pseudospectrum \(\sigma_e(A)\) for an arbitrary perturbation matrix \(\Delta_e \in \mathbb{R}^{n \times n}\),
but the iteration matrix $S \in \mathbb{R}^{n \times n}$ of the stationary iterative method $S$ and powers thereof are anything but arbitrary. Similarly, The modified Bauer-Fike Theorem (cf. lemma 4.7) provides us with bounds on the multiplicative $\epsilon$-pseudospectrum $\tilde{\sigma}_\epsilon (A)$ that hold for any unperturbed matrix $A \in \mathbb{R}^{n \times n}$ and similarity transformation $V \in \mathbb{R}^{n \times n}$, but again the unperturbed matrices, their matrices of eigenvectors and the perturbation matrix $\Delta_\epsilon \in \mathbb{R}^{n \times n}$ are anything but arbitrary or disconnected from each other in our setting. As such, lemma 4.8 tells us where the eigenvalues of the inexacty preconditioned matrices are guaranteed to not be, but we should not assume that these eigenvalues populate the whole of the set the lemma tells us they could be in. Take as an example the case where the right hand side of the inequality Eq. (4.90) equals one. Then, the set in Eq. (4.90) contains the origin, but the inexacty preconditioned matrix has to be invertible.

What follows now is an educated guess about what should happen to the eigenvalues of the inexacty preconditioned matrices as we increase the accuracy of the iterative solver used in the approximation. For this let us look at the inexacty block Jacobi preconditioned matrix:

$$K_{J,\kappa}^{-1} K \in \mathbb{R}^{(2n) \times (2n)}, \quad K_{J,\kappa}^{-1} K = \begin{pmatrix} I - S^\kappa & 0 \\ 0 & I - S^\kappa \end{pmatrix} K_{J}^{-1} K.$$

Here $\kappa = 1, 2, \ldots$ is implicitly assumed to be a positive integer. If the iteration matrix $S \in \mathbb{R}^{n \times n}$ is diagonalizable, we may extend the above expression to be defined on $\kappa \in [1, \infty)$, i.e. for any real number greater than or equal to one. The most reasonable extension appears to be the mapping:

$$P : [1, \infty) \to \mathbb{R}^{(2n) \times (2n)}, \quad \kappa \mapsto \begin{pmatrix} I - S^\kappa & 0 \\ 0 & I - S^\kappa \end{pmatrix} K_{J}^{-1} K.$$  

We would now expect that there exists a continuous mapping:

$$\lambda_i : [1, \infty) \to \mathbb{C}, \quad \kappa \mapsto \lambda \in \sigma (P (\kappa))$$

for each unperturbed eigenvalue:

$$\lambda_{i,\infty} \in \sigma (K_{J}^{-1} K), \quad \lambda_{i,\infty} = \lim_{\kappa \to \infty} \lambda_i (\kappa)$$

Figure 4.2: Filled contour plot of $\min_{\lambda \in [1,2]} \sigma_{GS} (\lambda, \mu)$. 

Figure 4.2: Filled contour plot of $\min_{\lambda \in [1,2]} \sigma_{J} (\lambda, \mu)$. 

4.1. CONVERGENCE ANALYSIS
and furthermore that the relative distance \( \left| \frac{\lambda_i(\kappa) - \lambda_{i,\infty}}{\lambda_{i,\infty}} \right| \in \mathbb{R}^+ \) is monotonically decreasing in \( \kappa \), and that it decreases exponentially.

These expectations appear relatively reasonable, if somewhat optimistic. In fact, they were fulfilled for the model problem at the beginning of this section, and we would expect the convergence of the perturbed eigenvalues towards the original eigenvalues in this setting to be similar to their convergence for the model problem, albeit somewhat worse.

Even if the above assumptions are valid we should see a stark drop in efficiency as we increase \( \kappa \). To see why, let us first define the mapping:

\[
E : [1, \infty) \rightarrow \text{Image}(E), \quad \kappa \mapsto \arg \min_{\{e \in \text{Image}(E) \mid e \in \mathbb{C} \}} \mathcal{M}_\kappa(e)
\] (4.97)

where we again reused the ellipse in the complex plane centered at \( c \in \mathbb{C} \) with focal distance \( d \in \mathbb{C} \) and semi-major axis \( a \in \mathbb{C} \):

\[
E : \mathbb{C}^3 \rightarrow \mathbb{P}(\mathbb{C}), \quad (c, d, a) \mapsto \{ z \in \mathbb{C} \mid |z - c + d| + |z - c - d| \leq 2|a| \}
\]

and the minimal maximum-norm over the set of real polynomial functions of degree at most \( k \) that have value one at the origin \( Q^k \):

\[
\mathcal{M}_\kappa : \mathbb{P}(\mathbb{C}) \rightarrow \mathbb{R}^+, \quad \sigma \mapsto \min_{q_\kappa \in Q^k} \max_{\lambda \in \sigma} |q_\kappa(\lambda)|.
\]

The mapping \( E \) provides us with the ellipse containing all eigenvalues of the inexactly preconditioned matrix which minimizes the term dominating the convergence rate of the used Krylov subspace method. Let the tuple \( (c(\kappa), d(\kappa), a(\kappa)) \in \mathbb{C}^3 \) denote the center, the focal distance and the semi-major axis of the ellipse \( E(\kappa) \). If the relative distance \( \left| \frac{\lambda_i(\kappa) - \lambda_{i,\infty}}{\lambda_{i,\infty}} \right| \in \mathbb{R}^+ \) is indeed exponentially decreasing for all perturbed eigenvalues we can expect the tuple \( (c(\kappa), d(\kappa), a(\kappa)) \in \mathbb{C}^3 \) to also exponentially converge towards its limit. However, if we put these arguments into Eq. (4.75) we can see that

\[
\mathcal{M}_\kappa(E(c(\kappa), d(\kappa), a(\kappa))) \leq \left[ c(\kappa) + \sqrt{c(\kappa)^2 - d(\kappa)^2} \right]
\]

will not decrease exponentially in \( \kappa \) towards zero, unless both \( a(\kappa) \) and \( d(\kappa) \) converge exponentially towards zero.

In our setting, this will not be the case, as the ellipse \( E(\kappa) \) does not converge towards a single point, but towards a line segment.

### 4.2 Experimental results

In this section we present the results of numerical experiments, carried out to compare the four methods and to validate the predictions made in the last chapters.

For this we solved the heat equation with the open unit cube \( \Omega = (0, 1)^3 \) as spatial domain and with the time domain \( (0, T] = (0, \tau) \). We constructed initial conditions and Dirichlet boundary conditions such that we have as solution to the heat equation either the vector valued \textit{monomial solution}:

\[
u_\text{mon} : \Omega \times (0, T] \rightarrow \mathbb{R}^3, \quad (t, x) \mapsto \left[ \sum_{i=1}^3 [x_i]_+ (1 - [x_i]) \right] \left( \begin{array}{c} t^2 \\ t^3 \end{array} \right)
\] (4.98)

or the scalar \textit{sinusoidal solution}:

\[
u_\text{sin} : \Omega \times (0, T] \rightarrow \mathbb{R}, \quad (t, x) \mapsto \prod_{i=1}^3 \sin(10 \pi \cdot [x_i]) \left[ \exp(-t) - 1 \right].
\] (4.99)

The former has the benefit that its spatial part gets solved exactly in the chosen discrete solution spaces, and its temporal parts get solved exactly if the temporal discretization order \( p \in \mathbb{N} \) is high enough. We can thus use the \( L^2 \)-norm of the error of the discrete solution at the end point of the time interval as a measure of the accuracy of the solution obtained using the iterative methods.
The experimental convergence results for the first case could however mislead us, as its spatial component is very smooth. The multigrid methods should therefore be extraordinarily efficient, which might not be the case in realistic applications.

We therefore repeated all experiments using a different analytical solution, namely the latter case. All results were identical.

For the space discretization we use a sequence of uniform, nested, tetrahedral grids and a corresponding sequence of nested $P_2$ finite element spaces $V_l$ with $l = 1, \ldots, 5$. The dimensions of the finite element spaces $V_l$ are $27,343,3375,29791,250047$ respectively.

For the time discretization we use the $dG(p)$-method (cf. Eq. (2.45)) with temporal discretization order $p \in \{1,3,5,7,9\}$ and using a single time step with time step size $\tau \in \{10^{-8},10^{-7},\ldots,10^0,10^3\}$. We choose temporal basis and test functions such that the temporal mass matrix $\mu \in \mathbb{R}^{(p+1)\times(p+1)}$, (cf. Eq. (2.71)) and the temporal propagation matrix $\rho \in \mathbb{R}^{(p+1)\times(p+1)}$ (cf. Eq. (2.72)) are two $2 \times 2$-block diagonal with blocks from the image of the embedding $\gamma$ (cf. Eq. (4.4)), which again enables us to identify the temporal basis and test functions with the arguments of the two characteristic complex numbers $\mu,\rho \in \mathbb{C}$.

For a given temporal discretization order $p \in \mathbb{N}$ and temporal basis and test functions, we will focus on the block matrix $K \in \mathbb{R}^{(2n)\times(2n)}$ of the global matrix corresponding to the characteristic temporal eigenvalue $\lambda \in \sigma(\mu^{-1}\rho)$ (cf. Eq. (4.6)) with the largest complex argument. This block will converge the slowest and is therefore indicative of the performance of an iterative solver applied to the global matrix.

For the block multigrid preconditioned method $\mathcal{S}_{BMG}$ and for the preconditioned Schur complement method $\mathcal{S}_S$ we choose $\text{arg}(\mu) = 0$ and for the other two methods we choose $\text{arg}(\mu) \in \{0,-\text{arg}(\rho),-\text{arg}(\lambda)\}$. Recall that $\text{arg}(\mu) = 0$ implies biorthogonal temporal basis and test functions and thus a diagonal temporal mass matrix $\mu \in \mathbb{R}^{(p+1)\times(p+1)}$, while $\text{arg}(\mu) = -\text{arg}(\lambda)$ implies $\text{arg}(\rho) = 0$ and thus a diagonal temporal propagation matrix $\rho \in \mathbb{R}^{(p+1)\times(p+1)}$.

For each block matrix $K \in \mathbb{R}^{(2n)\times(2n)}$ we solve its block system:

\[
K w \equiv \begin{pmatrix} \text{K}_{\text{diag}} & \text{K}_{\text{off}} \\ -\text{K}_{\text{off}} & \text{K}_{\text{diag}} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \equiv f
\]

using either the block multigrid preconditioned method $\mathcal{S}_{BMG}$ (cf. ??), the block Jacobi preconditioned method $\mathcal{S}_J$ (cf. ??), the block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ (cf. ??) or the preconditioned Schur complement method $\mathcal{S}_S$ (cf. ??) with a maximum of 20 Krylov subspace method iterations.

For the preconditioned Schur complement method $\mathcal{S}_S$ we perform all needed inversions of the off-diagonal block $K_{\text{off}} \in \text{Image}(K_1) \subset \mathbb{R}^{n \times n}$ up to machine precision to prevent errors made there from influencing the convergence rate of the preconditioned Schur complement method $\mathcal{S}_S$. For the block multigrid preconditioned method $\mathcal{S}_{BMG}$ we use $\kappa \in \{1,2,4,16\}$ V-cycles of the block multigrid method as the preconditioner and for the other methods we use $2\kappa$ V-cycles of the scalar multigrid method per iteration of the Krylov subspace method. Each multigrid method uses the canonical restriction and prolongation operators and one pre- and one post-smoothing iteration of the scalar or block symmetric Gauss-Seidel method. Recall that the computational cost of one iteration of the block multigrid method is roughly four times that of one iteration of the scalar multigrid method and that the computational cost of applying the preconditioner dominates the overall cost.

For the block multigrid preconditioned method $\mathcal{S}_{BMG}$, the block Jacobi preconditioned method $\mathcal{S}_J$ and the block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ we use the relative residual norm:

\[
\gamma_{\kappa,K}^{(k)} := \frac{\|K [w - w^{(\kappa,k)}]\|}{\|f\|} \quad (4.100)
\]

of its $k$-th approximation $w^{(\kappa,k)} \in \mathbb{R}^{2n}$ to the solution vector $w \in \mathbb{R}^{2n}$ as a measure of convergence and for the preconditioned Schur complement method $\mathcal{S}_S$ we use the relative residual norm:

\[
\gamma_{\kappa,S}^{(k)} := \frac{\|S [w_2 - w_2^{(\kappa,k)}]\|}{\|f_1 + K_{\text{diag}} K_{\text{off}}^{-1} f_2\|} \quad (4.101)
\]

of its $k$-th approximation $w_2^{(\kappa,k)} \in \mathbb{R}^n$ to the essential unknown $w_2 \in \mathbb{R}^n$. We define as a robust measure of the expected convergence rate of an iterative solver the median measured Krylov convergence rate:

\[
\Gamma_{\kappa,s} := \text{med}_k \left( \gamma_{\kappa,s}^{(k)} \right)^{1/k} \quad (4.102)
\]

4.2. EXPERIMENTAL RESULTS
CHAPTER 4. PRECONDITIONERS FOR THE DISCRETE PROBLEM

As a measure of efficiency which is comparable across the methods we choose for the block Jacobi preconditioned method $S_J$, the block Gauss-Seidel preconditioned method $S_{GS}$ and for the preconditioned Schur complement method $S_S$ the efficiency:

$$e_{\kappa,*}^{(k)} := [\Gamma_{\kappa,*}]^{1/(\kappa)}$$  \hspace{1cm} (4.103)

and for the block multigrid preconditioned method $S_{BMG}$ the square root of $e_{\kappa,*}^{(k)}$.

Let us now briefly summarize all the used discretization parameters. We have the four methods $S_{BMG}$, $S_J$, $S_{GS}$, $S_S$, the analytical solution $u \in \{u_{\text{nom}}, u_{\text{sin}}\}$, the spatial level $\ell \in \{1, \ldots, 5\}$, the temporal discretization order $p \in \{1, 3, 5, 7, 9\}$, the time step size $\tau \in \{10^{-8}, 10^{-7}, \ldots, 10^{2}, 10^{3}\}$, the temporal test and basis functions, identified by the arguments of their characteristic complex numbers $\mu, \rho \in \mathbb{C}$, which can be $\arg(\mu) \in \{0, -\arg(\rho), -\arg(\lambda)\}$ and finally the number of V-cycles per Krylov subspace method iteration $\kappa \in \{1, 2, 4, 16\}$. As the metrics of convergence we use the relative residual norm $\gamma_{\kappa,K}$ (cf. Eq. (4.100)) and the relative residual norm $\gamma_{\kappa,S}$ (cf. Eq. (4.101)), as a measure of convergence speed we use the Krylov convergence rate $\Gamma_{\kappa,*}$ (cf. Eq. (4.102)) and as a measure of efficiency we use the efficiency $e_{\kappa,*}^{(k)}$ (cf. Eq. (4.103)).

The following sections are structured as follows.

The first section is dedicated to confirming the predictions we made in the previous chapter regarding the influence of the discretization parameters on the convergence behaviour of the exactly preconditioned methods.

In the second section we compare the four methods in terms of their efficiency $e_{\kappa,*}^{(k)}$, and experimentally verify our prediction that the case $\kappa = 1$ should be the most efficient for all solution methods.

4.2. EXPERIMENTAL RESULTS

We will not actually apply the preconditioners exactly, but we will approximate the necessary inversions using $\kappa = 16$ V-cycles of the respective multigrid method. This appears to be a sufficiently good approximation of the exact preconditioners and numerical experiments in fact reveal that moderately lowering $\kappa$ does not significantly impact the measured convergence speeds. For this first section we will thus assume that the predictions made for the exact preconditioners hold also for the inexact preconditioners with $\kappa = 16$.

One important prediction was that the condition number of the eigenvector basis has only a negligible effect on the convergence behaviour of the relative residual norm $\gamma_{\kappa,K}$ for the block Jacobi preconditioned method $S_J$ and for the block Gauss-Seidel preconditioned method $S_{GS}$, cf. lemma 4.4.

To confirm this, we plot in Fig. 4.3 for the case $\arg(\mu) = 0$ the relative residual norm $\gamma_{\kappa,K}^{(k)}$ (y-axis) over the Krylov subspace method iteration $k$ (x-axis) with fixed temporal discretization order $p \in \mathbb{N}$ (lines) for the almost exactly ($\kappa = 16$) block Jacobi preconditioned method $S_J$ (top) and block Gauss-Seidel preconditioned method $S_{GS}$ (bottom) and three different time step sizes $\tau \in \{10^{-6}, 10^{-2}, 10^{2}\}$ (left to right). These values of the discretization parameters are only a small sample, but the behaviour for all discretization parameters was identical.

In Fig. 4.3 we already notice the influence of the time step size $\tau \in \mathbb{R}^+$ and the temporal discretization order $p \in \mathbb{N}$ on the Krylov convergence rate $\Gamma_{\kappa,*}$. We want to confirm the predictions we made regarding the impact of the discretization parameters on the convergence rates of the block Jacobi preconditioned method $S_J$ and the block Gauss-Seidel preconditioned method $S_{GS}$:

For $\arg(\mu) = 0$ and $\tau \to \infty$ we expect the eigenvalues of the block Jacobi preconditioned matrix $K_{J}^{-1}K \in \mathbb{R}^{(2n)\times(2n)}$ and of the block Gauss-Seidel preconditioned matrix $K_{GS}^{-1}K \in \mathbb{R}^{(2n)\times(2n)}$ to be clustered around $\lambda = 1$, and for $\tau \to 0$ we expect the eigenvalues to be located at two clusters, around $\lambda = 1 \pm \tau \tan(\arg(\rho))$ for the block Jacobi preconditioned method $S_J$ and around $\lambda \in \{1, 1 + \tan^2(\arg(\rho))\}$ for the block Gauss-Seidel preconditioned method $S_{GS}$, cf. lemma 4.4.

For $\arg(\rho) = 0$ we expect the limits in the time step size $\tau \in \mathbb{R}^+$ to be swapped, i.e. the methods perform better for smaller time step sizes. This relation can also be gleaned from the definition of the global matrix:

$$K_{\text{global}} \in \text{Image} \left( K_{(p+1)} \right), \quad K_{\text{global}} := K_{(p+1)} \left( \tau, \mu, \rho \right).$$

If the time step size $\tau \in \mathbb{R}^+$ is large, the term involving the temporal mass matrix $\mu \in \mathbb{R}^{(p+1)\times(p+1)}$ and by extension the characteristic complex mass coefficient $\mu \in \mathbb{C}$ dominates. If this term is better
approximated by the preconditioning matrix, i.e. if it is diagonal \((\arg(\mu) = 0)\), the block matrix \(K \in \mathbb{R}^{(2n) \times (2n)}\) as a whole is better approximated by the preconditioner. The same but reversed relation holds if the time step size \(\tau \in \mathbb{R}^+\) is small.

For \(\arg(\mu) = -\arg(\rho)\) we expect the bounds on the image of the spectral function \(\sigma_F : \mathbb{R}^+ \to \mathbb{R}\) to be smaller than the bounds for the other cases. This comes to show for the intermediate value of the time step size \(\tau \in \mathbb{R}^+\), where most of the image of the spectral function \(\sigma_F : \mathbb{R}^+ \to \mathbb{R}\) will be activated. While this basis will thus perform better in the intermediate regime, for neither of the limits of the time step size \(\tau \in \mathbb{R}^+\) it will perform as well as the best of the other two methods. This is due to its spectrum being located around two clusters for both limits, as opposed to one cluster for one side and two clusters for the other side.

For this we show in Fig. 4.4 the Krylov convergence rate \(\Gamma_{\kappa,\tau}(\tau)\) over the time step size \(\tau\) (x-axis) with fixed temporal basis functions (lines) for the almost exactly \((\kappa = 16)\) block Jacobi preconditioned method \(\mathcal{S}_J\) (top) and block Gauss-Seidel preconditioned method \(\mathcal{S}_{GS}\) (bottom) and three different temporal discretization orders \(p \in \{1, 5, 9\}\) (left to right).

Next we want to verify that the block Gauss-Seidel preconditioned method \(\mathcal{S}_{GS}\) with \(\arg(\mu) = -\arg(\rho)\) performs just as well as the preconditioned Schur complement method \(\mathcal{S}_S\) does for the hard problems, but worse for the cases \(\tau \to 0\) and \(\tau \to \infty\).

For this we show in Fig. 4.5 the Krylov convergence rate \(\Gamma_{\kappa,\tau}(\tau)\) over the time step size \(\tau\) (x-axis) for the almost exactly \((\kappa = 16)\) block Gauss-Seidel preconditioned method \(\mathcal{S}_{GS}\) with \(\arg(\mu) = -\arg(\rho)\) and preconditioned Schur complement method \(\mathcal{S}_S\) (lines) and two different temporal discretization orders \(p \in \{3, 7\}\) (left to right). We can observe the same effect for all temporal discretization orders.

Finally we want to confirm that the exactly preconditioned block multigrid preconditioned method \(\mathcal{S}_{BMG}\) converges after just one iteration. For this we show in Fig. 4.6 the Krylov convergence rate \(\Gamma_{\kappa,\tau}(\tau)\) over the time step size \(\tau\) (x-axis) for the almost exactly \((\kappa = 16)\) block multigrid preconditioned method \(\mathcal{S}_{BMG}\) for all temporal discretization orders \(p\) (lines).
Figure 4.4: Convergence rates for the exact Jacobi and Gauss-Seidel preconditioner

\[ \kappa = 16, \ell = 5, u = u_{\text{mon}}, S_\ast \in \{S_J, S_{GS}\}, \arg(\mu) \in \{0, -\arg(\rho), -\arg(\lambda)\}, \tau \in \{10^{-8}, \ldots, 10^{-3}\}, p \in \{1, 5, 9\} \].

We can clearly identify the different regimes described above. All observations hold for each temporal discretization order \( p \in \mathbb{N} \) and for both the block Jacobi preconditioned method \( S_J \) and the block Gauss-Seidel preconditioned method \( S_{GS} \).

For \( \arg(\mu) = 0 \) (blue line) we observe a skewed dependency of the convergence rate of the iterative solver on the time step size \( \tau \in \mathbb{R}^+ \). The convergence rate has a global maxium around \( \tau = 10^{-2} \). Going towards smaller \( \tau \) from there improves the convergence rate more slowly than going towards larger \( \tau \).

This behaviour is reversed for \( \arg(\mu) = -\arg(\rho) \) (green line).

For \( \arg(\mu) = -\arg(\rho) \) (orange line) we observe two global maxima and the local minimum in between them, which was predicted in the first section of this chapter. The convergence rate is roughly symmetric to this local minimum, as was also predicted. Furthermore, the maximal convergence rate is smaller than the maximal convergence rate of the other bases, but in the limits of the time step size \( \tau \in \mathbb{R}^+ \) one of the other bases converges quicker. Finally, the advantage over the other bases in the worst case grows with the temporal discretization order \( p \in \mathbb{N} \).

### 4.2.2 Convergence behaviour of the inexactly preconditioned methods

We will now compare the convergence rates of the inexact preconditioners and the efficiencies of the respective methods. For this we will focus on the block Gauss-Seidel preconditioned method \( S_{GS} \) with \( \arg(\mu) = -\arg(\rho) \), the preconditioned Schur complement method \( S_S \) and the block multigrid preconditioned method \( S_{BMG} \). The other bases were demonstrated to perform worse for the interesting problems, and the block Jacobi preconditioned method \( S_J \) was demonstrated to always perform worse than the block Gauss-Seidel preconditioned method \( S_{GS} \).

First, we want to measure the effect of not applying the preconditioners exactly on the Krylov convergence rate \( \Gamma_{\kappa,\ast} \).

We expect small perturbations to have no visible effect, i.e. if we change \( \kappa = 16 \) to \( \kappa = 15 \) we should observe no significant change in the Krylov convergence rate \( \Gamma_{\kappa,\ast} \) of the block Gauss-Seidel preconditioned method \( S_{GS} \) and of the block multigrid preconditioned method \( S_{BMG} \). This has been confirmed in numerical experiments, but will not be shown here.

Instead we focus on the effect of moderately large perturbations, which we will identify with the cases \( \kappa \in \{1, 2, 4\} \).
Figure 4.5: Convergence rates for the exact Gauss-Seidel and Schur preconditioner $\kappa = 16, \ell = 5, u = u_{mon}, S_\kappa \in \{S_{BMG}\}, \tau \in \{10^{-8}, \ldots, 10^{-3}\}, p \in \{1, 3, 5, 7, 9\}$.

We observe the preconditioned Schur complement method $S_\kappa$ performing better for small and large time step sizes. The upper bound of the Krylov convergence rate $\Gamma_{\kappa, *}$ for both methods is almost identical and is reached for similar values of $\tau$.

Figure 4.6: Convergence rates for the exact block multigrid preconditioner $\kappa = 16, \ell = 5, u = u_{mon}, S_\kappa \in \{S_{BMG}\}, \tau \in \{10^{-8}, \ldots, 10^{-3}\}, p \in \{1, 3, 5, 7, 9\}$.

We observe the almost immediate convergence of the block multigrid preconditioned method $S_{BMG}$. The values for the Krylov convergence rate $\Gamma_{\kappa, *}$ imply convergence after at most two Krylov subspace method iterations for all temporal discretization parameters. We want to note that the delayed convergence for some parameters is not due to an error in our prediction, but due to $\kappa = 16$ being not quite large enough for all discretization parameters.

We expect the effect of the perturbations to be larger if the exact methods perform better, as a small Krylov convergence rate $\Gamma_{\kappa, *}$ implies that all eigenvalues of the preconditioned matrix are clustered in a small area.

For the block multigrid preconditioned method $S_{BMG}$ all eigenvalues are one in the limit $\kappa \to \infty$ for all discretization parameters. We therefore expect the relative effect of decreasing $\kappa$ on the Krylov convergence rate $\Gamma_{\kappa, *}$ to be largest for the block multigrid preconditioned method $S_{BMG}$ and to be independent of all discretization parameters.

For the block Gauss-Seidel preconditioned method $S_{GS}$ and the preconditioned Schur complement method $S_\kappa$ we expect the effect to be larger for small or large $\tau$, but very small for the intermediate regime.
where the exact methods perform worse. This is again due to the eigenvalues of the preconditioned matrices being clustered around a single point for the preconditioned Schur complement method $\mathcal{S}_S$ in the limits $\tau \to 0$ and $\tau \to \infty$. On the other hand, if the Krylov convergence rate $\Gamma_{\kappa,*}$ of the block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ and the preconditioned Schur complement method $\mathcal{S}_S$ reaches its maximum over $\tau$, then the eigenvalues of the preconditioned matrices populate the whole of the set bounding them.

To confirm this, we show in Fig. 4.7 the Krylov convergence rate $\Gamma_{\kappa,*}$ (y-axis) over the time step size $\tau$ (x-axis) for different $\kappa \in \{1, 2, 4, 16\}$ (lines) and for the block multigrid preconditioned method $\mathcal{S}_{BMG}$ (top), block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ (center), preconditioned Schur complement method $\mathcal{S}_S$ (bottom) and two different temporal discretization orders $p \in \{3, 7\}$ (left to right). We can observe the same effect for all temporal discretization orders.

Finally we want to compare the efficiency $e^{(k)}_{\kappa,*}$ of each method for varying $\kappa$. We conjectured that the choice $\kappa = 1$ should be the optimal choice for all methods and for all discretization parameters.

We show in Fig. 4.7 the efficiency $e^{(k)}_{\kappa,*}$ (y-axis) over the time step size $\tau$ (x-axis) for different $\kappa \in \{1, 2, 4, 16\}$ (lines) and for the block multigrid preconditioned method $\mathcal{S}_{BMG}$ (top), block Gauss-Seidel preconditioned method $\mathcal{S}_{GS}$ (center), preconditioned Schur complement method $\mathcal{S}_S$ (bottom) and two different temporal discretization orders $p \in \{3, 7\}$ (left to right). We can observe the same effect for all temporal discretization orders.

4.2.3 A note on the accuracy of the Schur complement methods

Before we compare the efficiency of the methods let us comment on one issue with the preconditioned Schur complement method $\mathcal{S}_S$: The preconditioned Schur complement method $\mathcal{S}_S$ appears to only yield an accurate solution in the $L^2$ sense, if the time step size is not too large. If it is, the Schur complement matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ is extremely badly conditioned, which seems to impact the best obtainable accuracy. To visualize this we scatter plot in Fig. 4.9 the measured $L^2$-norm of the error of the discrete solution at the end point of the time interval (y-axis) over the measured relative residual norm $\gamma^{(k)}_{\kappa,*}$ after convergence (left) and the measured relative residual norm $\gamma^{(k)}_{\kappa,K}$ after convergence (left) (x-axis) for all discretization parameters.
We observe the predicted behaviour for the block multigrid preconditioned method $S_{BMG}$, where the Krylov convergence rate $\Gamma_{\kappa,*}$ goes from the previous values $\Gamma_{\kappa,*} \in [10^{-16}, 10^{-8}]$ (cf. Fig. 4.6) to values in the order of $10^{-1}$ for $\kappa = 1$. Note that we can already see that $\kappa = 1$ is the most efficient choice, as the values for $\kappa = 2$ are larger than the square of the values for $\kappa = 1$.

For both the block Gauss-Seidel preconditioned method $S_{GS}$ and the preconditioned Schur complement method $S_S$ we can also observe the predicted behaviour. The effect of decreasing $\kappa$ has the largest effect for small or large $\tau$, but almost none in the intermediate regime. Note that the values of the Krylov convergence rate $\Gamma_{\kappa,*}$ for $\tau \in \{10^{-4}, \ldots, 10^{4}\}$ are almost identical for varying $\kappa \in \{4, 16\}$.

We note that the lines of the Krylov convergence rate $\Gamma_{\kappa,*}$ for $\kappa = 1$ look very similar for all methods. This should correspond to a change in the efficiency of the multigrid method as a stationary iterative method, which depends slightly on $\tau$. 

**Figure 4.7:** Convergence rates for the inexact preconditioners

\[ p=3 \]

\[ \ell = 5, u = u_{\text{mon}}, S_* \in \{S_{BMG}, S_{GS}, S_S\}, \tau \in \{10^{-8}, \ldots, 10^{-3}\}, p \in \{3, 7\}. \]
We observe the predicted behaviour for almost all settings. For both the block multigrid preconditioned method $S_{BMG}$ and the block Gauss-Seidel preconditioned method $S_{GS}$ the case $\kappa = 1$ is always the most efficient one, and by a large margin. For the preconditioned Schur complement method $S_S$ the case $\kappa = 1$ is almost always the most efficient one, with a similar margin as the block Gauss-Seidel preconditioned method $S_{GS}$. For some reason this does not hold for large time step sizes.

Figure 4.8: Efficiencies for the inexact preconditioners $\ell = 5, u = u_{\text{mon}}, S_* \in \{S_{BMG}, S_{GS}, S_S\}, \tau \in \{10^{-8}, \ldots, 10^{-3}\}, p \in \{3, 7\}$. 

We observe the predicted behaviour for almost all settings. For both the block multigrid preconditioned method $S_{BMG}$ and the block Gauss-Seidel preconditioned method $S_{GS}$ the case $\kappa = 1$ is always the most efficient one, and by a large margin. For the preconditioned Schur complement method $S_S$ the case $\kappa = 1$ is almost always the most efficient one, with a similar margin as the block Gauss-Seidel preconditioned method $S_{GS}$. For some reason this does not hold for large time step sizes.
For all methods but for the preconditioned Schur complement method $S_S$, we observe the relative residual norm $\gamma^{(k)}_{\kappa, S}$ to be correlated with the error of the discrete solution except for very small values of the relative residual norm $\gamma^{(k)}_{\kappa, S}$.

For small values of relative residual norm $\gamma^{(k)}_{\kappa, S}$, the condition number of the eigenvector basis of the characteristic temporal matrix $\mu^{-1} \rho \in \mathbb{R}^{(p+1) \times (p+1)}$ (cf. Eq. (2.76)) introduces a slight error which increases with the temporal discretization order $p \in \mathbb{N}$.

We conclude that for these methods, the relative residual norm $\gamma^{(k)}_{\kappa, S}$ is a good indicator of the accuracy of our discrete solution. *This is not necessarily the case for the relative residual norm $\gamma^{(k)}_{\kappa, S}$. In the worst case, the error in the discrete solution is roughly eight orders of magnitude larger than the reported relative residual norm $\gamma^{(k)}_{\kappa, S}$.*

Luckily, this effect is only relevant for large values of the time step size $\tau \in \mathbb{R}^+$. It does increase however as the space discretization is refined and it should therefore be assumed to become relevant for smaller time step sizes. It would have been interesting to know whether this can happen for all orders of magnitude of $\gamma^{(k)}_{\kappa, S}$, other whether this effect is restricted to small ones such as with the effect of the badly-conditioned temporal basis.
5 Conclusion and outlook

In this thesis we analyzed four different methods to solve the linear systems that arise by discretizing the heat equation using a tensor-product space-time finite element method.

We improved the block Jacobi and block Gauss-Seidel preconditioners presented in [SMN06, MNS07] to be order-optimal also with respect to the temporal discretization order for the DG time-stepping method. This was accomplished by analyzing the impact of the temporal basis and test functions on the convergence rate of the preconditioned methods. This allowed us to identify choices of temporal basis and test functions that yield order-optimality.

We compared our analysis with the one of the method proposed in [WB15] and [BB15] and qualitatively predicted the impact of all discretization parameters on the convergence rate. The improved Gauss-Seidel method shares the same bound on the convergence rates if applied exactly.

On top of the standard analysis of the preconditioned methods under the assumption that we apply the preconditioners exactly, we also provided a coarse bound on the performance of the methods if we only approximate the application of the preconditioners.

Finally we performed numerical experiments with the aim of corroborating our results. We observed behaviours that support our claims, but were not able to identify any one method which performed best for all discretization parameters.

While our analysis of the exactly preconditioned methods appears to be able to capture the qualitative behaviour of the preconditioned methods, it does not enable us to predict which one will be the best one for a given set of discretization parameters.

Additionally, the bounds on the convergence rates of the inexactly preconditioned methods are not sharp. It remains to be seen whether it is possible to obtain better ones.
Bibliography


