Electronic Transactions on Numerical Analysis.
Volume 65, pp. 226–253, 2026.

ISSN 1068–9613.
DOI: 10.1553/etna_vol65s226

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ITERATIVE SOLVERS FOR PARTIAL DIFFERENTIAL EQUATIONS WITH DISSIPATIVE STRUCTURE: OPERATOR PRECONDITIONING AND OPTIMAL CONTROL*

VOLKER MEHRMANN†, MANUEL SCHALLER‡, AND MARTIN STOLL‡

Abstract. This work considers the iterative solution of large-scale problems subject to non-symmetric matrices or operators arising in discretizations of (port-)Hamiltonian partial differential equations. We consider problems governed by an operator $\mathcal{A} = \mathcal{H} + \mathcal{S}$ with a symmetric part $\mathcal{H}$ that is positive (semi-)definite and a skew-symmetric part $\mathcal{S}$. Prior work has shown that the structure and sparsity of the associated linear system enables Krylov subspace solvers such as the generalized minimal residual method (GMRES) or short recurrence variants such as Widlund's or Rapoport's method using the symmetric part $\mathcal{H}$, or an approximation of it, as preconditioner. In this work, we analyze the resulting condition numbers, which are crucial for fast convergence of these methods, for various partial differential equations (PDEs) arising in diffusion phenomena, fluid dynamics, and elasticity. We show that preconditioning with the symmetric part leads to a condition number uniform in the mesh size in the case of elliptic and parabolic PDEs, where $\mathcal{H}^{-1}\mathcal{S}$ is a bounded operator. Further, we employ the tailored Krylov subspace methods in optimal control by means of a condensing approach and a constraint preconditioner for the optimality system. We illustrate the results by various large-scale numerical examples and discuss efficient evaluations of the preconditioner such as the incomplete Cholesky factorization or the algebraic multigrid method.

Key words. dissipative Hamiltonian systems, preconditioning, partial differential equations, optimal control, Krylov subspace methods

AMS subject classifications. 65F08, 65F10, 65N22, 49M05, 49M41

1. Introduction.

In this work, we investigate iterative solvers for problems governed by linear operators $\mathcal{A} : \mathrm{dom}(\mathcal{A}) \supset X \to X$ on a Hilbert space X subject to the (formal) splitting

$$\mathcal{A} = \mathcal{H} + \mathcal{S}, \tag{1.1}$$

where $\mathcal{S} \colon X \supset \mathrm{dom}(\mathcal{S}) \to X$ is skew-symmetric (or skew-adjoint) and $\mathcal{H} \colon X \supset \mathrm{dom}(\mathcal{H}) \to X$ is symmetric (or self-adjoint) and nonnegative. This structure is ubiquitous in stationary or time-dependent partial differential equations (PDEs) governed by dissipative operators such as port-Hamiltonian systems. Therein, the positive definiteness of the symmetric part $\mathcal{H}$ corresponds to coercivity of a governing even-order differential operator such as in diffusion problems, fluid dynamics, or elasticity with strong damping.

Previous work [15, 25, 32, 33, 34] has suggested leveraging a skew-symmetric/symmetric splitting (1.1) for the efficient solution of matrix problems, i.e., via left-preconditioning in the generalized minimal residual method (GMRES) (see, e.g., [41]) using the symmetric part, or via Widlund's or Rapoport's method, [39, 54], which constitute short-term recurrence Krylov subspace methods. As usual for iterative methods, the convergence rate strongly depends on the condition number, respectively the spectral width, of the preconditioned system; see, e.g., [25, 49]. In the case of an underlying infinite-dimensional problem (e.g., a partial differential equation) governed by an operator (1.1) with invertible symmetric part $\mathcal{H}$, this

*Received October 18, 2025. Accepted May 6, 2026. Published online on June 22, 2026. Recommended by Daniel B. Szyld. Manuel Schaller acknowledges support by the German Research Foundation (DFG) under project number 519323897.
†Institute of Mathematics, Technische Universität Berlin, Berlin.
‡Corresponding author: M. Schaller, Faculty of Mathematics, Chemnitz University of Technology, Chemnitz (manuel.schaller@mathematik.tu-chemnitz.de).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

condition number is determined by Galerkin projections of

$$\mathcal{H}^{-1}\mathcal{A} = I + \mathcal{H}^{-1}\mathcal{S}$$

onto a suitable finite element space. In this work, we analyze this approach through the lens of operator preconditioning and provide applications in large-scale optimal control of PDEs.

Operator preconditioning is a powerful framework to assess the conditioning of Galerkin projections by means of the analysis of the underlying partial differential equation in function spaces. There is a rich literature for this topic, and we refer to the overview articles [6, 28] focusing on elliptic problems and the work [35] for an abstract and operator-theoretic approach to problems in Hilbert spaces. In [26], the equivalence between choosing a preconditioner and choosing a suitable inner product for conjugate gradient and minimal residual methods is presented. For an operator preconditioning approach to equality-constrained optimization, we refer to [44]. Preconditioning with the symmetric part for the Navier-Stokes equation was analyzed in [11]. For a general overview of preconditioning, in particular in the context of partial differential equations, we refer to the overview article [53] (see, in particular, Section 6.4 there for preconditioning with the symmetric part), the book [16], the overview article [31] focusing particularly on the conjugate gradient method, or the book [50] for domain decomposition-based preconditioners.

In this work, we leverage the structure of the underlying operator equation and its discretized version, in particular for preconditioning in the simulation and optimal control of partial differential equations. In this context, the accretivity of the underlying equation (or equivalently, the dissipativity) plays a crucial role and, in a wide range of applications, is induced by a dissipative Hamiltonian or port-Hamiltonian structure; see [1, 2, 36, 37, 55] for an analysis of the spectral properties of matrix and operator pencils with dissipative Hamiltonian structure. We also briefly mention other works where a dissipative Hamiltonian structure is used in (numerical) linear algebra. In [15, 25, 32, 33] short-term recurrence Krylov subspace methods for matrices occurring in dissipative Hamiltonian problems are discussed; see also the recent work [34] for a treatment of systems with saddle-point structure. In [8, 9], splitting and dynamic iteration methods leveraging the modular port-Hamiltonian structure were suggested, where the dissipativity particularly provides a monotonic bound for convergence. Peaceman–Rachford-type splitting methods in function spaces were suggested in [20] for simulation and in [21] for optimal control.

Contribution and outline. This paper features two main contributions, the first is an analysis of particular PDEs through the lens of operator preconditioning using the symmetric part, and the second focuses on structure-exploiting numerical methods for the solution of linear systems arising in optimal control. In the first part, in Section 3, we investigate the conditioning for a wide range of example problems such as advection-diffusion-type problems, fluid dynamics problems such as Stokes and Oseen equations, or problems from mechanics such as the wave equation and a beam equation. We observe that, depending on the functional analytic structure of the equation, in particular if $\mathcal{H}^{-1}\mathcal{S}$ is a bounded operator, the preconditioning leads to a condition number that is uniform in the mesh size. As a second contribution, in Section 4, we illustrate how iterative solvers (such as GMRES preconditioned with the symmetric part, Widlund's, or Rapoport's method) may be implemented in numerical optimal control of PDEs. In particular, as the adjoint equation inherits the structure of the state equation (e.g., accretive, dissipative, or port-Hamiltonian), we leverage the structure-exploiting methods for solving the state and the adjoint equation. Using this, we first suggest an elimination of the control, in which the corresponding optimality system is solved with a conjugate gradient method and in which the inner evaluation of the optimality system is performed with state and adjoint equation solves, including the above-mentioned methods.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Second, we use the methods in a constraint preconditioner in a conjugate gradient method applied to the full optimality system. By proposing efficient approximations of $\mathcal{H}^{-1}$, such as incomplete Cholesky factorizations or multigrid methods, we show that the proposed preconditioning technique may be efficiently implemented within the optimal control problem.

Notation. We adopt standard notation from operator theory (see [17]), functional analysis, and partial differential equations (see [3]). Let $(X, \langle\cdot,\cdot\rangle)$ be a Hilbert space. We say that an operator $\mathcal{A} : \mathrm{dom}(A) \subset X \to X$ is *accretive* if $\langle \mathcal{A}x, x\rangle \geq 0$ for all $x \in \mathrm{dom}(\mathcal{A})$, and *strictly accretive* if $\langle \mathcal{A}x, x\rangle \geq \mu\|x\|^2$ for some $\mu > 0$ and all $x \in \mathrm{dom}(\mathcal{A})$. We call $\mathcal{A}$ *(strictly) dissipative*, if $-\mathcal{A}$ is (strictly) accretive. If $\mathrm{dom}(\mathcal{A})$ is dense in X, then the *adjoint of* $\mathcal{A}$ is defined by $\mathcal{A}^* :\ \mathrm{dom}(A^*) \subset X \to X$ with

$$\mathrm{dom}(\mathcal{A}^*) \coloneqq \{y \in X \mid \text{there is } z \in X : \langle y, \mathcal{A}x\rangle_X = \langle z, x\rangle_X \text{ for all } x \in \mathrm{dom}(A)\}\,. \tag{1.2}$$

Due to the density of $\mathrm{dom}(\mathcal{A})$ in X, the element z in this set is uniquely determined, and we set $\mathcal{A}^*y \coloneqq z$. We say that $\mathcal{A}$ is *symmetric (respectively skew-symmetric)*, denoted by $\mathcal{A} \subset \mathcal{A}^*$ (respectively $\mathcal{A} \subset -\mathcal{A}^*$) if for all $x \in \mathrm{dom}(\mathcal{A})$, $\mathcal{A}x = \mathcal{A}^*x$ (respectively $\mathcal{A}x = -\mathcal{A}^*x$). We say that $\mathcal{A}$ is *self-adjoint (respectively skew-adjoint)*, denoted by $\mathcal{A} = \mathcal{A}^*$ (respectively $\mathcal{A} = -\mathcal{A}^*$), if $\mathcal{A}$ is symmetric (respectively skew-symmetric) and $\mathrm{dom}(\mathcal{A}) = \mathrm{dom}(\mathcal{A}^*)$.

Let $\Omega \subset \mathbb{R}^d$. For a function $f : \Omega \to \mathbb{R}$, we denote its gradient by ∇f. For a vector field $g : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$, we denote its divergence by $\mathrm{div}\, g = \sum_{i=1}^d \frac{\partial g_i}{\partial \omega_i}$. We denote by $L^2(\Omega;\mathbb{R}^d)$ (respectively $L^\infty(\Omega;\mathbb{R}^d)$) the space of square-integrable (respectively measurable and essentially bounded) (equivalence classes of) functions $f : \Omega \to \mathbb{R}^d$. By $H^1(\Omega;\mathbb{R}^d)$ we denote the standard Sobolev space of (equivalence classes of) square-integrable functions $f : \Omega \to \mathbb{R}^d$ with square-integrable first derivatives, and the subspace $H_0^1(\Omega;\mathbb{R}^d)$ consists of functions in $H^1(\Omega;\mathbb{R}^d)$ with vanishing trace on the boundary. In case $d = 1$, we abbreviate by omitting the second argument, e.g., $L^2(\Omega) = L^2(\Omega;\mathbb{R})$. Moreover, $H(\mathrm{div},\Omega)$ denotes the space of vector fields $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ with components in $L^2(\Omega)$ whose divergence (in the weak sense) also belongs to $L^2(\Omega)$.

Code availability and reproducibility. The code for all algorithms and numerical experiments conducted in this work as well as the involved matrices and the files to generate the discretizations via the finite element library FEniCS [4] are provided in the repository https://github.com/maschaller/indefinite_solvers.

2. Specialized solvers leveraging symmetric/skew-symmetric splittings. In the following, we briefly recall the solvers discussed in [15, 25] for a linear system

$$(H + S)x = b, \tag{2.1}$$

in which $H \in \mathbb{R}^{n\times n}$, $S \in \mathbb{R}^{n\times n}$, that typically result from Galerkin projections of operators $\mathcal{H}$ and $\mathcal{S}$ from (1.1), respectively. In particular, we assume in the following that $H = H^\top > 0$ and $S = -S^\top$. The case of a possibly non-invertible $H = H^\top \geq 0$ will be discussed later by a Schur complement approach.

To design iterative methods for the finite-dimensional non-symmetric problem (2.1), one could choose a very general approach such as GMRES [41] endowed with a suitable preconditioner. The downside of GMRES and the motivation for its variants is that new vectors must be orthogonalized against all basis vectors within the Krylov subspace. For purely symmetric or skew-symmetric system matrices, one can rely on short-term recurrence methods such as the MINRES or CG methods that only require orthogonalization against a small number of basis vectors. We now briefly introduce methods that share this favorable

ETNA
Kent State University and
Johann Radon Institute (RICAM)

property and which rely on the fact that the system matrix is symmetric or skew-symmetric in a non-standard inner product [18, 48]. In our case, the matrix $H^{-1}S$ is skew-symmetric in the inner product defined by the symmetric part H, that is,

$$\begin{aligned}\langle H^{-1}Sx, x\rangle_H &= x^\top (H^{-1}S)^\top Hx = x^\top S^\top H^{-1}Hx \\ &= x^\top HH^{-1}S^\top x = -x^\top HH^{-1}Sx = -\langle x, H^{-1}Sx\rangle_H.\end{aligned} \tag{2.2}$$

Methods that are tailored to exploit this skew-symmetry for a short-term recurrence were introduced by Widlund [54] and Rapoport [39]. The main idea is to build up a Krylov subspace via a Lanczos relation

$$H^{-1}SV_k = V_{k+1}T_{k+1,k},$$

where V_k and V_{k+1} store the basis vectors for the Krylov subspace and $T_{k+1,k} \in \mathbb{R}^{k+1,k}$ is a tridiagonal coefficient matrix. Here, in view of (2.2), the Lanczos method is based on the orthogonality in the inner product defined by H. The methods by Widlund and Rapoport are then obtained by minimizing a desired quantity such as the residual r_k or the error $x - x_k$ in a norm induced by H. In more detail, setting $\hat{b} = H^{-1}b$, Widlund's method corresponds to solving

$$(I_k + T_{k,k})y_k = \|\hat{b}\|_H e_1,$$

which yields the iterate $x_k = V_k y_k$. For Rapoport's method, a residual minimization in the H^{-1} norm results in the solution of the $k \times k$ system

$$T_{k+1,k}^\top T_{k+1,k} y_k = \|\hat{b}\|_H T_{k+1,k}^\top e_1,$$

where again $x_k = V_k y_k$. We note that similar search spaces are used when applying GMRES[1] to the left-preconditioned system $(I + H^{-1}S)x = H^{-1}b$. GMRES, however, requires an orthogonalization against *all* previous basis vectors in the Krylov space and thus can become quite slow if the preconditioner is not very effective. To this end, restarted variants of GMRES or methods with lower memory usage such as Loose GMRES (LGMRES) have been proposed [7]. If the preconditioner is evaluated inexactly, e.g., when using only a few steps of a multigrid method, flexible solvers provide a framework to ensure convergence and low-recurrence in the Lanczos methods despite inexactness. We refer the reader to the work [15], which suggests various flexible variants for preconditioning the system (2.1) with the symmetric part.

All the discussed methods, i.e., Widlund's and Rapoport's method as well as GMRES, are iterative Krylov subspace methods. As such, their convergence strongly depends on the condition number, respectively the spectral width, of the governing matrix $I + H^{-1}S$ as illustrated in [25, Section 5]. To this end, let $\lambda \in \mathbb{R}$ be such that the spectral inclusion $\sigma(H^{-1}S) \subset i[-\lambda, \lambda]$ holds (note that $H^{-1}S$ is H-skew-symmetric due to (2.2) and thus $H^{-1}S$ has imaginary eigenvalues). For Widlund's method, the rate is given by

$$\frac{\|x - x_{2k}\|_H}{\|x\|_H} \leq 2\left(\frac{\sqrt{1+\lambda^2} - 1}{\sqrt{1+\lambda^2} + 1}\right)^k. \tag{2.3}$$

For Rapoport's method, we have convergence of the residual of the form

$$\frac{\|b - (H+S)x_k\|_{H^{-1}}}{\|b\|_{H^{-1}}} \leq 2\left(\frac{\lambda}{\sqrt{1+\lambda^2} + 1}\right)^k. \tag{2.4}$$

[1] In this work, we will always consider GMRES with respect to the standard scalar product in $\mathbb{R}^n$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

In both estimates, the convergence speed strongly depends on the spectral width of $H^{-1}S$, that is, if $\lambda \to \infty$, then the convergence deteriorates. For discretizations of partial differential equations, it is thus desirable that this spectral width is bounded uniformly with respect to the mesh, which is closely associated with boundedness of the underlying (preconditioned) operator $\mathcal{H}^{-1}\mathcal{S}$. In the following section, we inspect this boundedness property and the associated conditioning for various problem classes.

3. Spectral properties in Hermitian preconditioning of partial differential equations. In this section, we briefly recall the idea of operator preconditioning in Krylov subspace methods and refer to [6, 26, 28, 35] for more details. Let $\mathcal{A} : X \supset \mathrm{dom}(\mathcal{A}) \to X$ be a closed and densely defined but unbounded linear operator. Here we call a densely defined operator $\mathcal{A} : X \supset \mathrm{dom}(\mathcal{A}) \to X$ *unbounded* on $(X, \|\cdot\|)$ if there is no $c \geq 0$ such that $\|\mathcal{A}x\| \leq c\|x\|$ for all $x \in \mathrm{dom}(\mathcal{A})$.

Then, for a given $b \in X$, we aim to find $x \in \mathrm{dom}(\mathcal{A})$ such that

$$\mathcal{A}x = b. \tag{3.1}$$

Since $\mathcal{A}$ is unbounded on X, the Krylov subspace $\mathcal{K}_m(\mathcal{A}, b) := \{b, \mathcal{A}b, \ldots, \mathcal{A}^{m-1}b\}$ is not well-defined, as applying powers of $\mathcal{A}$ is not feasible due to $\mathcal{A}^2 b = \mathcal{A}\mathcal{A}b$ but, in general, $\mathcal{A}b \notin \mathrm{dom}(\mathcal{A})$. To alleviate this, in operator (left-)preconditioning one introduces a mapping $\mathcal{P} \in L(X, \mathrm{dom}(\mathcal{A}))$ and considers the problem

$$\mathcal{P}\mathcal{A}x = \mathcal{P}x$$

such that $\mathcal{P}\mathcal{A} \in L(\mathrm{dom}(\mathcal{A}), \mathrm{dom}(\mathcal{A}))$. We note that since $\mathcal{A}$ is closed, we may render $\mathrm{dom}(\mathcal{A})$ a Banach space when endowed with the graph norm $\|x\|_{\mathrm{dom}(\mathcal{A})} := \|x\| + \|Ax\|$. Consequently, the Krylov subspace $\mathcal{K}_m(\mathcal{P}\mathcal{A}, \mathcal{P}b) = \{\mathcal{P}b, \mathcal{P}\mathcal{A}\mathcal{P}b, \ldots, (\mathcal{P}\mathcal{A})^{m-1}\mathcal{P}b\} \subset \mathrm{dom}(\mathcal{A})$ is well-defined.

A simple illustration of operator preconditioning may be obtained by the second-order operator governing the advection-diffusion-reaction equation on $\Omega \subset \mathbb{R}^d$, which we discuss in detail and in a more general setting in Section 3.1. Consider the PDE

$$\mathcal{A}x = -\Delta x + \mathbf{b}^\top \nabla x + \mathbf{c}x = f \tag{3.2}$$

for $\mathbf{c} > 0$ and $\mathbf{b} : \Omega \to \mathbb{R}^d$. Endowing the system with homogeneous Dirichlet boundary conditions, we may define the operator $\mathcal{A}$ on $X = H_0^1(\Omega)^*$ with $\mathrm{dom}(\mathcal{A}) = H_0^1(\Omega)$. Here, a suitable preconditioner is $(-\Delta + \mathbf{c}I)^{-1} : X \to \mathrm{dom}(\mathcal{A})$, which corresponds to the Riesz isomorphism in $H^1(\Omega)$ [35]. This may be understood as choosing a suitable scalar product for the computation of, e.g., gradients in a (conjugate) gradient method [26]. A similar argument applies when considering $X = L^2(\Omega)$ with domain $\mathrm{dom}(\mathcal{A}) = H_0^1(\Omega) \cap H^2(\Omega)$.

Besides being crucial for the well-definedness of the Krylov subspace for the governing infinite-dimensional operator equation (3.1), preconditioning is a central aspect in ensuring fast convergence of iterative methods. In the conjugate gradient method for symmetric problems, convergence strongly depends on the condition number, while in Rapoport's and Widlund's method, we observe a dependence on the spectral width; see (2.3) and (2.4). If the underlying problem is subject to an unbounded operator, and hence has an unbounded spectrum, this implies that for Galerkin projections onto suitable spaces, such as finite element spaces $V_{h,k}$ of order $k \in \mathbb{N}$ and mesh size $h > 0$, the conditioning severely deteriorates when refining the mesh. This may be seen easily with the Cauchy interlacing theorem (or min-max theorem) [12, Theorem 9.12] in the Babuška–Osborn theory, which states that (assuming real eigenvalues)

$$\lambda_i \leq \mu_i \leq \lambda_i + C(i)\mathcal{O}(h^k) \qquad \text{for all } 1 \leq i \leq N := \dim V_{h,k}, \tag{3.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where the λ_i are the eigenvalues of $\mathcal{A}$ and the μ_i are the eigenvalues of the Galerkin projection $P_{V_{h,k}}\mathcal{A}P_{V_{h,k}}$, both sorted in ascending order, and where $P_{V_{h,k}}$ is the orthogonal projection onto the finite element subspace $V_{h,k}$. If $\mathcal{A}$ is unbounded, then the spectrum of $\mathcal{A}$ is, as well, such that $|\lambda_i| \to \infty$, for $i \to \infty$. Consequently, considering the condition number corresponding to the spectral norm, $\kappa_2(P_{V_{h,k}}\mathcal{A}P_{V_{h,k}}) = \mu_N/\mu_1$ diverges as $|\mu_N| \to \infty$ as $h \to 0$ due to (3.3) and the unboundedness of $\mathcal{A}$. This divergence may be made precise when considering particular settings. For Galerkin projections of second-order operators such as (3.2) and piecewise affine linear finite elements, the condition number is of order h^{-2}; see [16, Theorem 1.32]. In the numerical examples presented later in this work, we observe also different orders of the condition number, closely related to the order of the underlying differential operator $\mathcal{A}$ and its preconditioned variant $\mathcal{H}^{-1}\mathcal{A}$.

In the following sections we inspect the suitability of self-adjoint preconditioning for problems governed by operators that admit a self-adjoint/skew-adjoint splitting as in (1.1). In view of the discussion above, to ensure that the self-adjoint part is a suitable preconditioner that in particular renders the condition number, respectively the spectral width, uniformly bounded in terms of the mesh width $h > 0$, we have to ensure that

$$\mathcal{H}^{-1}\mathcal{A} = I + \mathcal{H}^{-1}\mathcal{S} \in L(X, X). \tag{3.4}$$

We observe that this boundedness condition usually holds for parabolic or elliptic problems as illustrated in Section 3.1 for an advection-diffusion-type problem, in Section 3.2 for the Stokes equation, in Section 3.3 for an Oseen equation, or in Section 3.5 for a beam equation with structural Kelvin–Voigt damping.

However, for the wave equation with momentum damping, where the governing derivative operator of highest order is skew-symmetric, we observe in Section 3.4 that Hermitian preconditioning is not suitable if one aims for a condition number bounded uniformly in the mesh size.

For the sake of clarity of presentation, we present results for stationary problems. However, we stress that all considerations also carry over to the case of a semi-discretization of dissipative Hamiltonian problems of the form

$$\dot{x} = -\mathcal{M}x, \tag{3.5}$$

where $\mathcal{M} : X \supset \mathrm{dom}(\mathcal{M}) \to X$ is an accretive operator, i.e., $\langle x, \mathcal{M}x\rangle \geq 0$ for all $x \in \mathrm{dom}(\mathcal{M})$, which we may formally split via $\mathcal{M} = \mathcal{H}_\mathcal{M} + \mathcal{S}_\mathcal{M}$ into a symmetric and accretive operator $\mathcal{H}$ and a skew-symmetric operator $\mathcal{S}$. Note that for general operators, such a splitting can be very intricate, in particular the naive definition $\mathcal{H} = \frac{1}{2}(\mathcal{M} + \mathcal{M}^*)$ may lead to a trivial domain [5]. However, for particular applications in PDEs we make this splitting mathematically precise in the subsequent sections. Semi-discretization in time, e.g., by the implicit midpoint rule with step size $\delta t > 0$, leads to the implicit iteration

$$(I + \tfrac{\delta t}{2}\mathcal{M})x^+ = (I - \tfrac{\delta t}{2}\mathcal{M})x. \tag{3.6}$$

Due to the accretivity of $\mathcal{M}$, the operator on the left-hand side of (3.6) is strictly accretive, i.e., $\langle x, I + \frac{\delta t}{2}\mathcal{M}x\rangle \geq \|x\|^2$ and (at least formally) may be split into the symmetric and accretive operator $\mathcal{H} = I + \frac{\delta t}{2}\mathcal{H}_\mathcal{M}$ and the skew-symmetric operator $\mathcal{S} = \frac{\delta t}{2}\mathcal{H}_\mathcal{S}$, where $\mathcal{H}_\mathcal{M}$ and $\mathcal{H}_\mathcal{S}$ are the symmetric, respectively skew-symmetric, parts of $\mathcal{M}$. Thus, the governing operator is simply a scaled variant of the accretive operator $\mathcal{M}$ shifted by the identity, which of course does not influence its boundedness and thus neither the conditioning nor the spectral width of Galerkin projections.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We also briefly note that in the remainder of this work, we sometimes verify the (skew-) symmetry of operators (in contrast to the stronger property of skew-adjointness and self-adjointness). However, when considering Galerkin methods, one usually chooses the ansatz space $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_N\} \subset \operatorname{dom}(\mathcal{A})$ and performs a projection of the operator by defining $A \in \mathbb{R}^{N\times N}$ by

$$A_{ij} := \langle \mathcal{A}\varphi_i, \varphi_j \rangle_{L^2(\Omega)} \qquad \text{for all } i, j = 1, \dots, N,$$

hence only testing against elements in $\operatorname{dom}(\mathcal{A})$. Consequently, skew-symmetry $\mathcal{A}x = -\mathcal{A}^*x$ for all $x \in \operatorname{dom}(\mathcal{A})$ or symmetry $\mathcal{A}x = \mathcal{A}^*x$ for all $x \in \operatorname{dom}(\mathcal{A})$ (hence only the condition $\operatorname{dom}(\mathcal{A}^*) \subset \operatorname{dom}(\mathcal{A})$ and not necessarily $\operatorname{dom}(\mathcal{A}^*) = \operatorname{dom}(\mathcal{A})$) is enough to ensure (skew)-symmetry of the resulting matrix. A different domain of the adjoint can be included using Petrov–Galerkin methods, in which the test space differs from the ansatz space.

3.1. Stationary advection-diffusion-reaction equation. As a first model, we consider a stationary advection-diffusion-reaction equation with homogeneous Dirichlet boundary conditions given by the PDE

$$\begin{aligned} -\operatorname{div}(\nu \nabla x) + \mathbf{b}^\top \nabla x + \mathbf{c}x &= f \qquad \text{on } \Omega, \\ x &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{3.7}$$

where $x : \Omega \to \mathbb{R}$ models the scalar-valued concentration of a species on a spatial domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, and $f : \Omega \to \mathbb{R}$ is a given source term. The *diffusivity* $\nu \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ is a pointwise symmetric and uniformly positive matrix-valued mapping, i.e., $\nu(\omega)^\top = \nu(\omega)$ for all $\omega \in \Omega$, and there is $\underline{\nu} > 0$ such that $v^\top \nu(\omega) v \geq \underline{\nu}\|v\|^2_{\mathbb{R}^d}$ for all $v \in \mathbb{R}^d$ and $\omega \in \Omega$. To ensure accretivity of the governing operator, we assume that $\mathbf{b} \in L^\infty(\Omega; \mathbb{R}^d)$ has a vanishing weak divergence, i.e., $\mathbf{b} \in H(\operatorname{div}, \Omega)$, $\operatorname{div} \mathbf{b} = 0$ a.e. on Ω, and that $\mathbf{c} \in L^\infty(\Omega; \mathbb{R})$ is pointwise nonnegative, i.e., $\mathbf{c}(\omega) \geq 0$ for a.e. $\omega \in \Omega$.

In the following lemma we analyze the structure of the associated operator.

LEMMA 3.1. *Let* $D = \{x \in H^1_0(\Omega) \mid \nu\nabla x \in H(\operatorname{div}, \Omega)\}$, *and define*

$$\begin{aligned} &\mathcal{A} : L^2(\Omega) \supset D \to L^2(\Omega), && \mathcal{A}x = -\operatorname{div}(\nu\nabla x) + \mathbf{b}^\top \nabla x + \mathbf{c}x, \\ &\mathcal{H} : L^2(\Omega) \supset D \to L^2(\Omega), && \mathcal{H}x = -\operatorname{div}(\nu\nabla x) + \mathbf{c}x, \qquad \textit{and} \\ &\mathcal{S} : L^2(\Omega) \supset D \to L^2(\Omega), && \mathcal{S}x = \mathbf{b}^\top \nabla x. \end{aligned}$$

Then $\mathcal{H}$ *is boundedly invertible, and* $\mathcal{A}$ *can be split as*

$$\mathcal{A} = \mathcal{H} + \mathcal{S}, \quad \textit{with} \quad \mathcal{H} = \mathcal{H}^* \quad \textit{and} \quad \mathcal{S} \subset -\mathcal{S}^*,$$

where we use the notation as introduced after (1.2).

Proof. We first prove the bounded invertibility of $\mathcal{H}$. First, for all $x \in D$,

$$\begin{aligned} \langle \mathcal{H}x, x \rangle_{L^2(\Omega)} &= -\langle \operatorname{div}(\nu\nabla x), x \rangle_{L^2(\Omega)} + \langle \mathbf{c}x, x \rangle_{L^2(\Omega)} \\ &= \langle \nu\nabla x, \nabla x \rangle_{L^2(\Omega;\mathbb{R}^d)} + \langle \mathbf{c}x, x \rangle_{L^2(\Omega)} \\ &\geq \tfrac{1}{2} \min\{\alpha, 1\} \underline{\nu} \|x\|^2_{H^1(\Omega;\mathbb{R}^d)}, \end{aligned}$$

where $\alpha > 0$ is the Poincaré constant, i.e., $\|\nabla x\|^2_{L^2(\Omega)} \geq \alpha \|x\|^2_{L^2(\Omega)}$ for all $x \in D \subset H^1_0(\Omega)$. Hence, by the Lax–Milgram theorem and maximal elliptic regularity [24, Section 3], $\mathcal{H}$ is boundedly invertible. We proceed to show self-adjointness of $\mathcal{H}$. Let $x_1, x_2 \in D$. Then, due to the symmetry of $\nu : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and by the previous computations, we have

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$$\begin{aligned}\langle \mathcal{H}x_1, x_2\rangle_{L^2(\Omega)} &= \langle \nu\nabla x_1, \nabla x_2\rangle_{L^2(\Omega)} + \langle \mathbf{c}x_1, x_2\rangle_{L^2(\Omega)}\\ &= \langle \nabla x_1, \nu\nabla x_2\rangle_{L^2(\Omega)} + \langle x_1, \mathbf{c}x_2\rangle_{L^2(\Omega)}\\ &= -\langle \operatorname{div}(\nu\nabla x_2), x_1\rangle_{L^2(\Omega)} + \langle \mathbf{c}x_2, x_1\rangle_{L^2(\Omega)} = \langle \mathcal{H}x_2, x_1\rangle_{L^2(\Omega)},\end{aligned}$$

showing that $\mathcal{H}$ is symmetric. As $\mathcal{H}$ is boundedly invertible, it is also self-adjoint. This follows from $\mathcal{H} = (\mathcal{H}^{-1})^{-1} = (\mathcal{H}^{-*})^{-1} = \mathcal{H}^*$, as the bounded inverse of a symmetric operator is always self-adjoint.

We now compute the adjoint of $\mathcal{S}$. Let $x_1, x_2 \in D$. Then

$$\langle \mathcal{S}x_1, x_2\rangle_{L^2(\Omega)} = \langle \mathbf{b}^\top \nabla x_1, x_2\rangle_{L^2(\Omega)} = -\langle x_1, \operatorname{div}(\mathbf{b}x_2)\rangle_{L^2(\Omega)} = \langle x_1, \mathcal{S}^* x_2\rangle_{L^2(\Omega)} \tag{3.8}$$

by the definition of the adjoint. This implies that $\operatorname{dom}(\mathcal{S}^*) = \{x \in L^2(\Omega) \,|\, \mathbf{b}x \in H(\operatorname{div}, \Omega)\}$. Hence $\operatorname{dom}(\mathcal{S}^*) \supsetneq \operatorname{dom}(\mathcal{S}) = D$. For $x \in \operatorname{dom}(\mathcal{S}) = D \subset H_0^1(\Omega)$, we may proceed and use the chain rule for the weak divergence and obtain

$$\mathcal{S}^* x = -\operatorname{div}(\mathbf{b}x) = -(\operatorname{div}\mathbf{b})x - \mathbf{b}^\top \nabla x = -\mathbf{b}^\top \nabla x$$

as $\mathbf{b}$ is divergence-free. Together with (3.8) this implies that $\mathcal{S}$ is skew-symmetric. ∎

We note that in the previous result, we could also define $\mathcal{S}$ on the space $H^1(\Omega)$, which would be the canonical extension as $\mathcal{S}$ involves only a gradient term.

PROPOSITION 3.2 (Preconditioning). *The preconditioned operator*

$$I + \mathcal{H}^{-1}\mathcal{S} : H_0^1(\Omega) \to H_0^1(\Omega)$$

is bounded.

Proof. As $\mathcal{S} : H_0^1(\Omega) \to L^2(\Omega)$ is linear and bounded and $\mathcal{H} : D \to L^2(\Omega)$ is boundedly invertible, $\mathcal{H}^{-1}\mathcal{S} : H_0^1(\Omega) \to \operatorname{dom}(\mathcal{H}) = \{x \in H_0^1(\Omega) \,|\, \nu\nabla x \in H(\operatorname{div}, \Omega)\} \hookrightarrow H_0^1(\Omega)$ continuously such that $\mathcal{H}^{-1}\mathcal{S} \in L(H_0^1(\Omega), H_0^1(\Omega))$. This yields the assertion. ∎

In the following remark we briefly discuss compactness and weak formulations.

REMARK 3.3.

1) If the boundary is smooth enough and $\nu \in C^1(\Omega) \cap C(\bar{\Omega})$, then it holds that $\operatorname{dom}(\mathcal{H}) = \{x \in H_0^1(\Omega) \,|\, \nu\nabla x \in H(\operatorname{div}, \Omega)\} = H^2(\Omega) \cap H_0^1(\Omega)$ such that the embedding $\operatorname{dom}(\mathcal{H}) \hookrightarrow H_0^1(\Omega)$ is compact. Consequently, $\mathcal{H}^{-1}\mathcal{S}$ is also compact. A similar argument may also be employed for less regular boundaries and coefficients, where we have an embedding $\operatorname{dom}(\mathcal{H}) \hookrightarrow H_0^1(\Omega) \cap H^{1+\varepsilon}(\Omega)$ for some $\varepsilon > 0$, which is sufficient for compactness. The reader is referred to [24, Section 2]. However, we note that this compactness does not carry over to the operator $I + \mathcal{H}^{-1}\mathcal{S}$, which is the governing operator of the problem; see (3.4).

2) The operator $\mathcal{S}$ admits a unique extension $\mathcal{S}_\mathrm{e} : H^{-1}(\Omega) \supset \operatorname{dom}(\mathcal{S}_\mathrm{e}) = L^2(\Omega) \to H^{-1}(\Omega)$ defined by

$$\langle \mathcal{S}_\mathrm{e} x, y\rangle_{H^{-1}(\Omega), H^1(\Omega)} = -\langle x, \operatorname{div}(\mathbf{b}x)\rangle_{L^2(\Omega)}$$

for $x \in \operatorname{dom}(\mathcal{S}_\mathrm{e})$ and $y \in H_0^1(\Omega)$. The same applies to $\mathcal{H}$, which admits a unique extension $\mathcal{H}_\mathrm{e} : H^{-1}(\Omega) \supset \operatorname{dom}(\mathcal{H}_\mathrm{e}) = H_0^1(\Omega)$ defined by

$$\langle \mathcal{H}x, y\rangle_{H^{-1}(\Omega), H^1(\Omega)} = \langle \nu\nabla x, \nabla y\rangle_{L^2(\Omega)} + \langle \mathbf{c}x, y\rangle_{L^2(\Omega)}$$

for $x \in \operatorname{dom}(\mathcal{H}_\mathrm{e})$ and $y \in H_0^1(\Omega)$. These extensions correspond to the *weak form* of the PDE (3.7). By straightforward adaptation of the proof of Proposition 3.2, we may analogously obtain boundedness of $\mathcal{H}_\mathrm{e}^{-1}\mathcal{S}_\mathrm{e} : L^2(\Omega) \to L^2(\Omega)$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We now present a numerical illustration of Proposition 3.2 and compute the condition numbers of a Galerkin discretization using piecewise affine linear finite elements. We set $\Omega = [0,1] \times [0,5] \times [0,1]$ with diffusivity $\nu = 0.001$, advection field $\mathbf{b} = (0.5, 0, 0)^\top$, and $\mathbf{c} \equiv 0$. On the left-hand side of Figure 3.1, we depict the condition numbers of the Galerkin projection of $\mathcal{A}$ denoted by A, of its symmetric part $\mathcal{H}$ denoted by H, and its skew-symmetric part $\mathcal{S}$ denoted by S. We observe in the left plot that the condition number of A is dominated by that of the stiffness matrix of the diffusion term in the symmetric part, behaving like h^{-2} (due to the presence of two derivatives), while the condition number of the skew-symmetric part corresponding to the advection scales like h^{-1} as it only involves one derivative. In the right plot, we observe that the condition number of the preconditioned matrix $H^{-1}A = I + H^{-1}S$ is bounded uniformly in the mesh width. This is due to the fact that, as proven in Proposition 3.2, $\mathcal{H}^{-1}\mathcal{S}$ is a bounded operator and hence has bounded spectrum such that $I + \mathcal{H}^{-1}\mathcal{S}$ has a bounded spectrum as well. Correspondingly, the condition number for Galerkin projections is uniformly bounded. The same can be observed when approximating H^{-1} with an incomplete Cholesky factorization using a fill-in tolerance of 10^{-2}; see, e.g., [46].

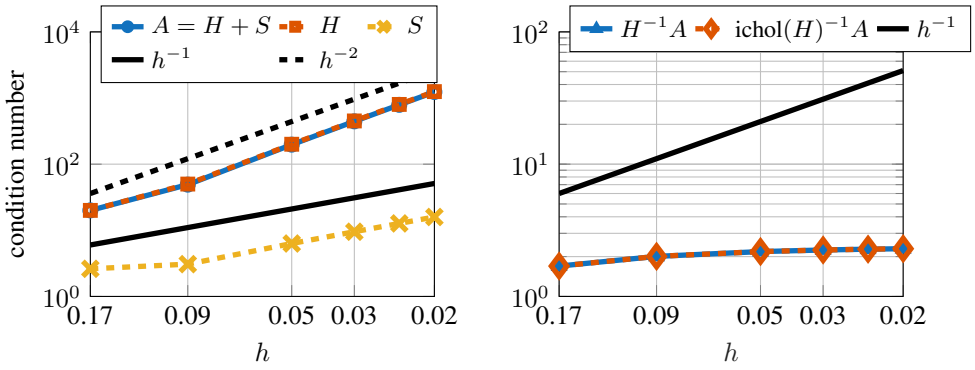

FIG. 3.1. *Condition numbers before (left) and after preconditioning (right) for the advection-diffusion equation over a varying mesh size $h > 0$.*

3.2. Stokes equation. As a second example, we consider the stationary Stokes equation given by

$$\begin{aligned} -\operatorname{div}(\nu \nabla v) + \nabla P &= f && \text{on } \Omega, \\ \operatorname{div} v &= 0 && \text{on } \Omega, \\ v &= 0 && \text{on } \partial\Omega, \end{aligned}$$

for $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. Here, $v : \Omega \to \mathbb{R}^d$ is the vector field modeling the velocity of the fluid, and the scalar field $P : \Omega \to \mathbb{R}$ corresponds to the pressure. The differential operators $\nabla : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $\operatorname{div} : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ are to be understood row-wise, and the kinematic viscosity $\nu : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is again a pointwise symmetric and uniformly positive matrix-valued mapping.

To include pressure-regularized formulations, as commonly used in finite element methods, we introduce $s_1, s_2 \in \mathbb{R}$ and consider the governing operator

$$\mathcal{A} = \begin{bmatrix} -\operatorname{div}(\nu\nabla v) & \nabla \\ \operatorname{div} & s_1 - s_2\Delta \end{bmatrix}, \tag{3.9}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which we may consider as an operator

$$\mathcal{A} : L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega) \supset D \to L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega),$$
$$D = \{x \in H_0^1(\Omega;\mathbb{R}^d) \,|\, \nu\nabla x \in H(\operatorname{div},\Omega;\mathbb{R}^d)\} \times \{x \in H^1(\Omega) \,|\, s_2\Delta x \in L^2(\Omega)\}.$$

For $s_2 \neq 0$, we have $\{x \in H^1(\Omega) \,|\, s_2\Delta x \in L^2(\Omega)\} \cong H^2(\Omega)$ (here $\cong$ denotes that the spaces are isomorphic), and for $s_2 = 0$ we have $\{x \in H^1(\Omega) \,|\, s_2\Delta x \in L^2(\Omega)\} = H^1(\Omega)$. The operators then have the following structure:

PROPOSITION 3.4. *The operator $\mathcal{A}$ in (3.9) may be decomposed into*

$$\mathcal{H} = \begin{bmatrix} -\operatorname{div}(\nu\nabla\cdot) & 0 \\ 0 & s_1 - s_2\Delta \end{bmatrix}, \qquad \mathcal{S} = \begin{bmatrix} 0 & \nabla \\ \operatorname{div} & 0 \end{bmatrix},$$

where $\mathcal{H} : L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega) \supset D \to L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega)$ is symmetric and $\mathcal{S} : L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega) \supset D \to L^2(\Omega;\mathbb{R}^d) \times L^2(\Omega)$ is skew-symmetric. If $s_2 > 0$, then $\mathcal{H}$ is self-adjoint.

Proof. Let us start with the top-left block of $\mathcal{H}$. The self-adjointness follows from the same arguments as in the proof of Lemma 3.1. For $s_2 \neq 0$, the same applies to the bottom-right block. For $s_2 = 0$, we have the multiplication operator $s_1 I$ acting on $L^2(\Omega)$, which is clearly self-adjoint.

The skew-symmetry of $\mathcal{S}$ follows since for all $(v,P),(w,Q) \in D \subset H_0^1(\Omega;\mathbb{R}^d) \times H^1(\Omega)$,

$$\begin{aligned} \langle \mathcal{S}(v,P),(w,Q)\rangle_{L^2(\Omega;\mathbb{R}^d)\times L^2(\Omega)} &= \langle \nabla P, w\rangle_{L^2(\Omega;\mathbb{R}^d)} + \langle \operatorname{div} v, Q\rangle_{L^2(\Omega)} \\ &= -\langle P, \operatorname{div} w\rangle_{L^2(\Omega)} - \langle v, \nabla Q\rangle_{L^2(\Omega;\mathbb{R}^d)} \end{aligned} \tag{3.10}$$

due to the homogeneous Dirichlet boundary conditions. ∎

We briefly comment on the functional analytic framework of Proposition 3.4. The considered spaces only allow for skew-symmetry of $\mathcal{S}$, which is in particular not closed on D. Loosely speaking, this results from domains that are not maximally chosen. Correspondingly, choosing suitable maximal domains, one may obtain also skew-adjointness. However, we stress that in usual finite element implementations [13], one would assemble a Galerkin discretization of $\mathcal{A}$ on a finite element subspace $V \subset \operatorname{dom}(\mathcal{A})$ and then extract symmetric and skew-symmetric parts on the matrix level. In this regard, we also note that one may consider a weak formulation in the sense of Remark 3.3 leading to bilinear forms. This is illustrated in the next section by means of the Oseen equation, which is structurally similar to the Stokes equation. We have the following result for the preconditioned operator:

PROPOSITION 3.5. *Consider the operator $\mathcal{A}$ in (3.9). Let $s_1 \in \mathbb{R} \setminus \{0\}$, and set $X := H(\operatorname{div};\Omega) \times H^1(\Omega)$. If $s_2 = 0$, then $\mathcal{H}^{-1}\mathcal{S} : X \to X$ is unbounded. If $s_2 > 0$, then $\mathcal{H}^{-1}\mathcal{S} : X \to X$ is bounded. In particular, $I + \mathcal{H}^{-1}\mathcal{S} \in L(X,X)$.*

Proof. Denote by $\mathcal{H}_{11}$ and $\mathcal{H}_{22}$ the diagonal blocks of $\mathcal{H}$ endowed with the canonical domain inherited from the cross-product structure of $D = D_1 \times D_2$. Then, $\mathcal{H}_{11}$ is self-adjoint by a similar argument as in the proof of Proposition 3.2 applied to vector-valued functions. Moreover, $\mathcal{H}_{11}^{-1} \in L(L^2(\Omega;\mathbb{R}^d), D_1)$, which implies the boundedness of $\mathcal{H}_{11}^{-1}\nabla \in L(H^1(\Omega), H^1(\Omega;\mathbb{R}^d))$ due to $\nabla \in L(H^1(\Omega), L^2(\Omega;\mathbb{R}^d))$ and $D_1 \hookrightarrow H^1(\Omega;\mathbb{R}^d)$ (where $\hookrightarrow$ denotes the canonical embedding).

Regarding the second block $\mathcal{H}_{22}$, we distinguish the two cases. First, let $s_2 = 0$ such that $\mathcal{H}_{22} = s_1 I$, and hence

$$\mathcal{H}^{-1}\mathcal{S} = \begin{bmatrix} 0 & \mathcal{H}_{11}^{-1}\nabla \\ \frac{1}{s_1}\operatorname{div} & 0 \end{bmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The bottom-left block is clearly unbounded as an operator from $H(\mathrm{div};\Omega)$ to $H^1(\Omega)$, as the weak divergence applied to an element from $H(\mathrm{div};\Omega)$ is (by definition of $H(\mathrm{div};\Omega)$) only square integrable and not necessarily weakly differentiable.

In case $s_1 > 0$, $\mathcal{H}_{22}$ is closed such that $\mathcal{H}_{22} = s_1 - s_2\Delta$ is boundedly invertible (as a mapping from $L^2(\Omega)$ to $\mathrm{dom}(\mathcal{H}_{22}) = H^2(\Omega) \hookrightarrow H^1(\Omega)$) due to the Lax–Milgram theorem and a similar argument as in the proof of Proposition 3.2. In particular, we have $\mathcal{H}_{22}^{-1}\,\mathrm{div} \in L(H(\mathrm{div};\Omega), H^1(\Omega))$. Hence $\mathcal{H}^{-1}\mathcal{S}$ is bounded as all blocks are bounded. ∎

The numerical illustration of Proposition 3.5 is presented in Figure 3.2, where we have chosen $\Omega = [0,1]^2$, $\nu \equiv 1$, and discretized the system with P2 vector-valued finite elements for the velocity field $v : \Omega \to \mathbb{R}^2$ and P1 finite elements for the pressure $P : \Omega \to \mathbb{R}$. We note that for this choice of finite elements, one usually does not employ a stabilization. Here, we do so regardless to obtain invertibility of $\mathcal{H}$. The unstabilized case will be considered in the next section.

In the left plot of Figure 3.2, we observe an increasing condition number if we only employ an L^2-pressure stabilization, corresponding to the choice $s_2 = 0$ in Proposition 3.5. This is due to the fact that $\mathcal{H}^{-1}\mathcal{S}$ contains a (first-order) differential operator in the bottom-left block, leading to a behavior of the condition number like $\mathcal{O}(h^{-1})$. In the right plot of Figure 3.2, we see that an H^1-regularization ($s_2 > 0$) leads to a uniformly bounded (in the mesh size) condition number, which reflects the boundedness of $\mathcal{H}^{-1}\mathcal{S}$ as proven in Proposition 3.5. Furthermore, the inversion of the block-wise elliptic operator H may be approximated with an incomplete Cholesky factorization, where we, however, observe an increase of the condition number for small mesh sizes (depending on the fill-in tolerance). Note that this example was also considered in [35, Examples 7.1 and 7.2], where the authors observed a similar behavior of the condition number and discussed suitable inf-sup conditions for finite element discretizations.

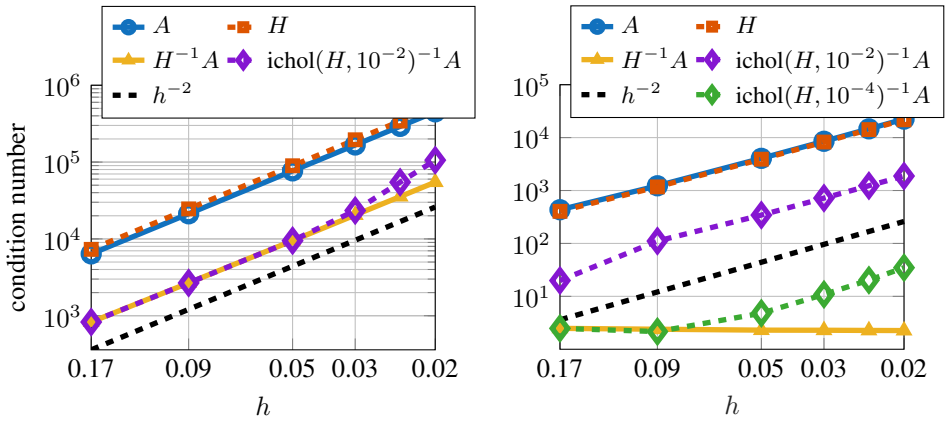

FIG. 3.2. *Condition numbers for the Stokes equation with L^2-pressure regularization $(s_1, s_2) = (1, 0)$ (left) and H^1-pressure regularization $(s_1, s_2) = (1, 1)$ (right).*

3.3. Oseen equation. As third example we consider the Oseen equation given by

$$\begin{aligned} -\,\mathrm{div}\,\sigma(v,P) + \mathbf{b}^\top \nabla v &= f && \text{on } \Omega,\\ \mathrm{div}\, v &= 0 && \text{on } \Omega,\\ v &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{3.11}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which arises when linearizing the Navier-Stokes equation at a stationary profile. Again, we choose $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, and with the dynamic viscosity $\mu > 0$, we denote the symmetric stress tensor by

$$\sigma(v,P) = \mu(\nabla v + \nabla v^\top) - PI_d.$$

Writing (3.11) in a block-wise fashion, we consider the operator

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & 0 \end{bmatrix} := \begin{bmatrix} -\operatorname{div} \nu(\nabla \cdot + \nabla^\top \cdot) + \mathbf{b}^\top \nabla \cdot & \nabla \\ \operatorname{div} & 0 \end{bmatrix}$$

as a mapping

$$\begin{aligned} &\mathcal{A} : H^{-1}(\Omega;\mathbb{R}^d) \times H^{-1}(\Omega) \supset \operatorname{dom}(\mathcal{A}) \to H^{-1}(\Omega;\mathbb{R}^d) \times H^{-1}(\Omega), \\ &\operatorname{dom}(\mathcal{A}) = H_0^1(\Omega;\mathbb{R}^d) \times L^2(\Omega). \end{aligned}$$

Note that here (and in contrast to the advection-diffusion-reaction equation in Section 3.1), we interpret the second-order differential operator in divergence form in the top-left block in weak form, i.e., for $v, w \in H^1(\Omega;\mathbb{R}^d)$,

$$\begin{aligned} &\langle \mathcal{A}_{11} v, w \rangle_{H^{-1}(\Omega;\mathbb{R}^d), H^1(\Omega;\mathbb{R}^d)} \\ &\qquad = \langle \nu(\nabla v + \nabla^\top v), \nabla w \rangle_{L^2(\Omega;\mathbb{R}^{d\times d})} + \langle \mathbf{b}^\top \nabla v, w \rangle_{L^2(\Omega;\mathbb{R}^d)}. \end{aligned} \tag{3.12}$$

Similarly, $\mathcal{A}_{12}$ is to be understood as the operator

$$\langle \mathcal{A}_{12} P, w \rangle_{H^{-1}(\Omega;\mathbb{R}^d), H^1(\Omega;\mathbb{R}^d)} = -\langle P, \operatorname{div} w \rangle_{L^2(\Omega)}, \tag{3.13}$$

for all $P \in H^1(\Omega)$, $w \in H_0^1(\Omega)$. In contrast to the Stokes equation considered in Section 3.2, we do not employ a pressure regularization here. Hence, the bottom-right block of $\mathcal{A}$ is zero such that the symmetric part is not invertible. To deal with this case, we consider the Schur complement approach as suggested in [25], i.e., we consider the Schur complement of $\mathcal{A}_{11}$ in $\mathcal{A}$ given by

$$\mathcal{W} = -\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \tag{3.14}$$

and obtain the following structure and boundedness result:

PROPOSITION 3.6. *The Schur complement $\mathcal{W} : L^2(\Omega) \supset H^1(\Omega) \to L^2(\Omega)$ is bounded, that is, $\mathcal{W} \in L(L^2(\Omega), L^2(\Omega))$, and accretive.*

Proof. First we show that $\mathcal{A}_{11}$ as defined in (3.12) induces a coercive and bounded bilinear form in $H_0^1(\Omega;\mathbb{R}^d)$. The boundedness is clear by definition. To show the coercivity, note that for all $v \in H_0^1(\Omega;\mathbb{R}^d)$ we have

$$\langle \mathbf{b}^\top \nabla v, v \rangle_{L^2(\Omega;\mathbb{R}^d)} = -\langle v, \mathbf{b}^\top \nabla v \rangle_{L^2(\Omega;\mathbb{R}^d)} \tag{3.15}$$

as $\mathbf{b}$ is divergence free by definition, and due to the no-slip boundary condition we have $v = 0$. Thus, $\langle \mathbf{b}^\top \nabla v, v \rangle_{L^2(\Omega;\mathbb{R}^d)} = 0$. Moreover, by Korn's inequality [14, Theorem 6.3.4], there exists a constant $c > 0$ such that

$$\langle \nu(\nabla v + \nabla^\top v), \nabla v \rangle_{L^2(\Omega;\mathbb{R}^{d\times d})} \geq c \|v\|^2_{H^1(\Omega;\mathbb{R}^d)}.$$

Together with (3.15), this implies that the bilinear form is coercive, i.e.,

$$\langle \mathcal{A}_{11} v, v \rangle_{H^{-1}(\Omega;\mathbb{R}^d), H^1(\Omega;\mathbb{R}^d)} \geq c \|v\|^2_{H^1(\Omega;\mathbb{R}^d)} \qquad \text{for all } v \in H_0^1(\Omega;\mathbb{R}^d).$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Thus, $\mathcal{A}_{11}$ is invertible and $\mathcal{A}_{11}^{-1} : H^{-1}(\Omega;\mathbb{R}^d) \to H_0^1(\Omega;\mathbb{R}^d)$. Hence, in view of (3.13),

$$\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \in L(L^2(\Omega), H^1(\Omega;\mathbb{R}^d))$$

such that, applying the divergence operator $\mathcal{A}_{21} = \operatorname{div} \in L(H^1(\Omega;\mathbb{R}^d), L^2(\Omega))$, we obtain boundedness of the Schur complement, that is, $\mathcal{W} = \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12} \in L(L^2(\Omega), L^2(\Omega))$.

To see that $\mathcal{W}$ is accretive, we show that $\mathcal{A}_{21} = \mathcal{A}_{12}^*$ (where we consider $\mathcal{A}_{21} \in L(H^1(\Omega;\mathbb{R}^d), L^2(\Omega))$ and $\mathcal{A}_{12} \in L(L^2(\Omega)), H^{-1}(\Omega;\mathbb{R}^d))$ and that $\mathcal{A}_{11}^{-1}$ is accretive. This follows due to

$$\langle \mathcal{A}_{12}P, w\rangle_{H^{-1}(\Omega;\mathbb{R}^d),H^1(\Omega;\mathbb{R}^d)} = -\langle P, \operatorname{div} w\rangle_{L^2(\Omega)} = -\langle P, \mathcal{A}_{21}w\rangle_{L^2(\Omega)}$$

by the no-slip boundary condition $w = 0$ on $\partial\Omega$. For the accretivity of $\mathcal{A}_{11}^{-1}$, let $v \in H^{-1}(\Omega;\mathbb{R}^d)$. Then, by the invertibility of $\mathcal{A}_{11} : H_0^1(\Omega;\mathbb{R}^d) \to H^{-1}(\Omega;\mathbb{R}^d)$, there is a unique $w \in H_0^1(\Omega;\mathbb{R}^d)$ such that $\mathcal{A}_{11}w = v$. Hence,

$$\begin{aligned}\langle \mathcal{A}_{11}^{-1}v, v\rangle_{H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d)} &= \langle \mathcal{A}_{11}^{-1}\mathcal{A}_{11}w, \mathcal{A}_{11}w\rangle_{H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d)}\\ &= \langle w, \mathcal{A}_{11}w\rangle_{H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d)}\\ &\geq c\|w\|^2_{H^1(\Omega;\mathbb{R}^d)} = c\|\mathcal{A}_{11}^{-1}v\|^2_{H^1(\Omega;\mathbb{R}^d)} \geq \tilde{c}\|v\|^2_{H^{-1}(\Omega;\mathbb{R}^d)},\end{aligned}$$

where the second last inequality follows from the accretivity of $\mathcal{A}_{11}$ and the last inequality follows with $\tilde{c} = \frac{c}{\bar{c}}$ from the boundedness, i.e., from $\|\mathcal{A}_{11}\|_{L(H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d))} \leq \bar{c}$. Consequently,

$$\begin{aligned}\langle \mathcal{W}v, v\rangle_{L^2(\Omega)} &= -\langle \mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}v, v\rangle_{L^2(\Omega)}\\ &= -\langle \mathcal{A}_{11}^{-1}\mathcal{A}_{12}v, \mathcal{A}_{21}^*v\rangle_{H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d)}\\ &= \langle \mathcal{A}_{11}^{-1}\mathcal{A}_{12}v, \mathcal{A}_{12}v\rangle_{H^1(\Omega;\mathbb{R}^d),H^{-1}(\Omega;\mathbb{R}^d)} \geq \tilde{c}\|\mathcal{A}_{12}v\|^2_{H^{-1}(\Omega;\mathbb{R}^d)} \geq 0,\end{aligned}$$

which shows the accretivity of $\mathcal{W}$. ∎

We briefly comment on the last part of the proof of Proposition 3.6. When considering $L^2(\Omega)$-functions with zero mean, we may deduce strict accretivity by means of an extension of the Poincaré inequality $\|\mathcal{A}_{12}v\|_{H^{-1}(\Omega;\mathbb{R}^d)} = \|\nabla v\|_{H^{-1}(\Omega;\mathbb{R}^d)} \geq C\|v\|_{L^2(\Omega;\mathbb{R}^d)}$ for some constant $C > 0$ and all $v \in L^2(\Omega;\mathbb{R}^d)$.

We depict the condition numbers associated with the Oseen equation in Figure 3.3. We choose the same parameters and domain as for the Stokes equation of Section 3.2 and the advection term $\mathbf{b} = [1, 1]^\top$. On the left-hand side of Figure 3.3, and as to be expected, we observe the same behavior for the top-left block as for the advection-diffusion equation, that is, an increase of the condition number proportional to $\frac{1}{h^2}$. On the right-hand side of Figure 3.3 we depict the condition number of the Schur complement W (being a discretization of $\mathcal{W}$ from (3.14)) and its preconditioned versions. We observe that already the unpreconditioned Schur complement has a uniformly bounded condition number (due to its boundedness proven in Proposition 3.6), while a preconditioning with a pressure mass matrix M_p decreases this condition number by one order of magnitude, and preconditioning with the symmetric part of the Schur complement H_W decreases it by another order of magnitude. Note that, however, for the latter it is necessary to form the Schur complement, while a preconditioning with the mass matrix of the pressure can be done very efficiently, e.g., only using its diagonal part [52].

3.4. Wave equation with momentum damping. As a next example, we consider a hyperbolic PDE given by a wave equation. The stationary system corresponding to the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

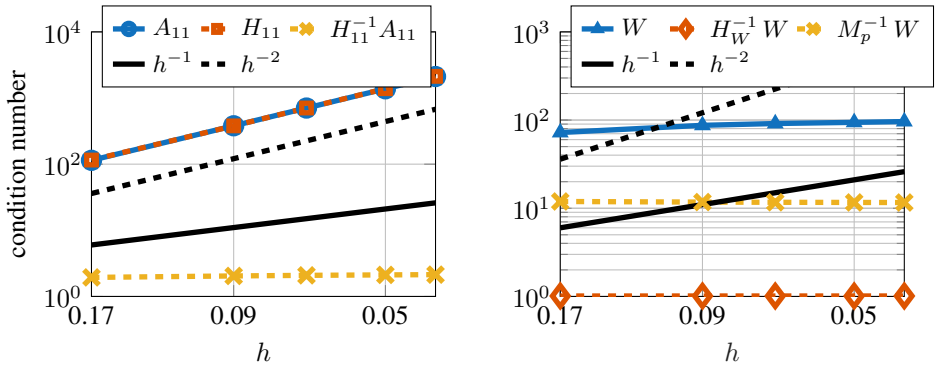

FIG. 3.3. *Condition numbers for Oseen equation for the original system (left) and the Schur complement (right).*

first-order port-Hamiltonian formulation reads

$$\begin{aligned} -\operatorname{div} q + \rho p &= f_1 && \text{on } \Omega, \\ -\nabla p + \eta q &= f_2 && \text{on } \Omega, \\ n^\top q &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here, $p\colon \Omega \to \mathbb{R}$ is the momentum, $q\colon \Omega \to \mathbb{R}^d$ is the vector-valued strain, and $\rho, \eta \geq 0$ are scalar friction parameters. The boundary condition models a vanishing strain in normal direction. The governing operator of this system is given by

$$\mathcal{A} = \begin{bmatrix} \rho & -\operatorname{div} \\ -\nabla & \eta \end{bmatrix}, \tag{3.16}$$

which we consider as a mapping

$$\mathcal{A} : L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d) \supset H^1(\Omega) \times H(\operatorname{div};\Omega) \to L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d).$$

PROPOSITION 3.7. *The operator $\mathcal{A}$ in (3.16) may be decomposed as*

$$\mathcal{H} + \mathcal{S} = \begin{bmatrix} \rho & 0 \\ 0 & \eta \end{bmatrix} + \begin{bmatrix} 0 & -\operatorname{div} \\ -\nabla & 0 \end{bmatrix},$$

where $\mathcal{S} : L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d) \supset H^1(\Omega) \times H(\operatorname{div};\Omega) \to L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d)$ is skew-adjoint and $\mathcal{H} : L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d) \to L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d)$ is self-adjoint.

Proof. The self-adjointness of $\mathcal{H}$ as a multiplication operator is obvious. The skew-adjointness of $\mathcal{S}$ follows from a computation analogous to (3.10). ∎

The next proposition follows straightforwardly, as the symmetric part is a bounded operator, and hence its inverse cannot compensate for the unboundedness of the derivative operators.

PROPOSITION 3.8. *Consider the operator $\mathcal{A}$ in (3.16) and its decomposition from Proposition 3.7. Let $X \in \{L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d), H^1(\Omega) \times H(\operatorname{div};\Omega)\}$. Then $\mathcal{H}^{-1}\mathcal{S} : X \to X$ is unbounded.*

We illustrate the unboundedness in Proposition 3.8 in Figure 3.4, where we choose $\Omega = [0,1]$ and $\rho = \nu = 1$. In the left plot, we see that the condition number of the discretization A of $\mathcal{A}$ and also that of the skew-symmetric part consisting of the first-order differential operators in the off-diagonal blocks behaves like $\mathcal{O}(h^{-1})$. In the right plot we observe that the preconditioning has no effect, which is to be expected as $\mathcal{H} = I \in L(L^2(\Omega), L^2(\Omega))$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

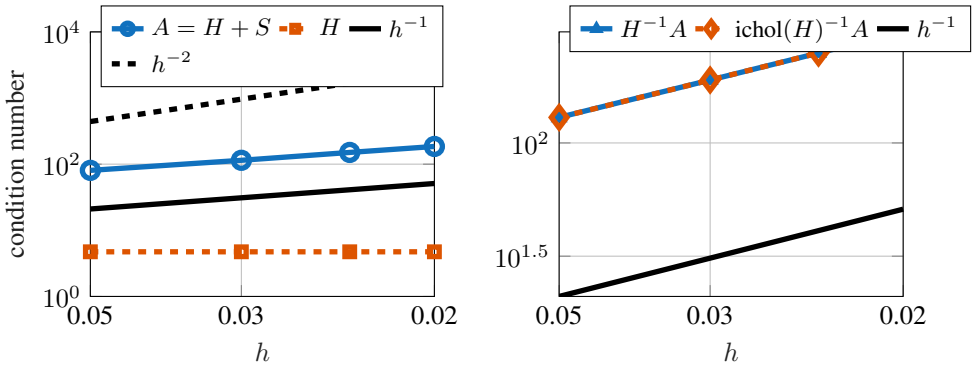

FIG. 3.4. *Condition numbers for wave equation with momentum damping in the first-order formulation.*

In this section we have demonstrated that the preconditioning with the symmetric part does not lead to a mesh-independent order of the condition number if the skew-symmetric part is dominant in the sense that it has a higher order of the differential operator.

In this example, it would hence clearly be more suitable to choose the (unbounded) skew-symmetric part $\mathcal{S}$ or an approximation thereof as a preconditioner. In this regard, we refer to the work [22] focusing on preconditioning with the skew-symmetric part. An alternative would be to multiply the equation with the imaginary unit such that $i\mathcal{S}$ is self-adjoint.

3.5. Beam equation with structural (Kelvin–Voigt) damping. As a last example, we consider a beam equation that is subject to strong (structural) damping on the domain $\Omega = (0,1)$. Denoting by $x : \mathbb{R}_{\geq 0} \times (0,1) \to \mathbb{R}$ the transverse displacement of the beam, the dynamics of the beam is given by the PDE

$$\ddot{x}(t,r) + \frac{\partial^2}{\partial r^2}\left(E\frac{\partial^2}{\partial r^2}x(t,r) + C\frac{\partial^3}{\partial r^2 \partial t}x(t,r)\right) = f(t,r),$$

with boundary conditions

$$x(t,0) = \frac{\partial}{\partial r}x(t,1) = \frac{\partial^2}{\partial r^2}x(t,0) = \frac{\partial^3}{\partial r^3}x(t,1) = 0. \tag{3.17}$$

For an in-depth analytical treatment, we refer to [29]. Note that although at first sight, the beam equation seems hyperbolic, it was proven in [29] that it gives rise to an analytic semigroup, hence behaving more like a parabolic equation such as the advection-diffusion-reaction equation. The reason for this is the strong damping. This parabolic nature will also be observed in the boundedness of the preconditioned operator and hence in uniform condition numbers of Galerkin projections.

To obtain a first-order dissipative formulation, we define the variables $p = \dot{x}$, $q = \frac{\partial^2}{\partial r^2}x$, which yield the dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} -\frac{\partial^4}{\partial r^4} & -\frac{\partial^2}{\partial r^2} \\ \frac{\partial^2}{\partial r^2} & 0 \end{bmatrix}\begin{bmatrix} p \\ q \end{bmatrix} + \begin{bmatrix} f \\ 0 \end{bmatrix}.$$

As the boundary conditions for the position (3.17) also translate to the momentum, we have, for $p_1, p_2 : (0,1) \to \mathbb{R}$ satisfying the boundary conditions in (3.17), the (formal) equality

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$$\begin{aligned}\langle p_1^{(4)}, p_2\rangle_{L^2(0,1)} &= \langle p_1^{(2)}, p_2^{(2)}\rangle_{L^2(0,1)} + p_1^{(3)}(1)p_2(1) - p_1^{(3)}(0)p_2(0)\\ &\qquad - p_1^{(2)}(1)p_2^{(1)}(1) + p_1^{(2)}(0)p_2^{(1)}(0).\\ &= \langle p_1^{(2)}, p_2^{(2)}\rangle_{L^2(0,1)},\end{aligned}$$

and, for momentum p and curvature q respecting the boundary conditions (3.17), we have

$$\begin{aligned}\langle p^{(2)}, q\rangle_{L^2(0,1)} &= \langle p, q^{(2)}\rangle_{L^2(0,1)} + p^{(1)}(1)q(1) - p^{(1)}(0)q(0)\\ &\qquad - p(1)q^{(1)}(1) + p(0)q^{(1)}(0)\\ &= \langle p, q^{(2)}\rangle_{L^2(0,1)}.\end{aligned}$$

Hence, the corresponding operator in weak form (readily accessible for finite element methods) may be formulated via

$$\mathcal{A} = \begin{bmatrix}\mathcal{A}_{11} & \mathcal{A}_{12}\\ \mathcal{A}_{21} & 0\end{bmatrix}, \tag{3.18}$$

where, setting $\operatorname{dom}(\mathcal{A}_{11}) := \{p \in H^2(0,1) \,|\, p(0) = p'(1) = 0\} \subset H^2(0,1)$ and $H^{-2}(0,1) := \operatorname{dom}(\mathcal{A}_{11})^*$, with the dual space taken with respect to the Pivot space $L^2(0,1)$, we consider the operators

$$\begin{aligned}\mathcal{A}_{11} : H^{-2}(0,1) \supset \operatorname{dom}(\mathcal{A}_{11}) &\to H^{-2}(0,1),\\ \mathcal{A}_{12} : L^2(0,1) &\to H^{-2}(0,1),\\ \mathcal{A}_{21} : \operatorname{dom}(\mathcal{A}_{11}) &\to L^2(0,1),\end{aligned}$$

which are, for $p_1, p_2 \in \operatorname{dom}(\mathcal{A}_{11})$ and $q \in L^2(0,1)$, defined by

$$\begin{aligned}\langle \mathcal{A}_{11}p_1, p_2\rangle_{H^{-2}(0,1),H^2(0,1)} &:= \langle p_1^{(2)}, p_2^{(2)}\rangle_{L^2(\Omega)},\\ \langle \mathcal{A}_{12}q, p_1\rangle_{H^{-2}(0,1),H^2(0,1)} &:= \langle q, p_1^{(2)}\rangle_{L^2(0,1)},\\ \langle \mathcal{A}_{21}p_1, q\rangle_{L^2(0,1)} &:= -\langle p_1^{(2)}, q\rangle_{L^2(0,1)}.\end{aligned}$$

As in the Oseen example discussed in Section 3.3, we form the Schur complement

$$\mathcal{W} := -\mathcal{A}_{21}\mathcal{A}_{11}^{-1}\mathcal{A}_{12}, \tag{3.19}$$

and we obtain an analogous result to Proposition 3.6.

PROPOSITION 3.9. *Consider the operator $\mathcal{A}$ in (3.18) and the associated Schur complement $\mathcal{W}$ in (3.19). Then $\mathcal{W}$ is bounded, i.e., $\mathcal{W} \in L(L^2(0,1), L^2(0,1))$, and accretive.*

Proof. The proof is analogous to the proof of Proposition 3.6. First, we show that $\mathcal{A}_{11} : H^{-2}(0,1) \supset \operatorname{dom}(\mathcal{A}_{11}) \to H^{-2}(0,1)$ gives rise to a bounded and bilinear form $a : \operatorname{dom}(\mathcal{A}_{11}) \times \operatorname{dom}(\mathcal{A}_{11}) \to \mathbb{R}$ defined by

$$a(p_1, p_2) := \langle \mathcal{A}_{11}p_1, p_2\rangle_{H^{-2}(0,1),H^2(0,1)}$$

in order to invoke the Lax–Milgram theorem. To this end, for $p_1, p_2 \in \operatorname{dom}(\mathcal{A}_{11})$, boundedness follows by

$$a(p_1, p_2) = \langle p_1^{(2)}, p_2^{(2)}\rangle_{L^2(0,1)} \le \|p_1\|_{H^2(0,1)}\|p_2\|_{H^2(0,1)}$$

from the Cauchy–Schwarz inequality. Moreover, for $p \in \operatorname{dom}(\mathcal{A}_{11})$,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$$a(p,p) = \|p^{(2)}\|^2_{L^2(0,1)} = \frac{1}{3}\|p^{(2)}\|^2_{L^2(0,1)} + \frac{2}{3}\|p^{(2)}\|^2_{L^2(0,1)}$$
$$\geq \frac{1}{3}\|p^{(2)}\|^2_{L^2(0,1)} + \frac{2}{3}c\|p^{(1)}\|^2_{L^2(0,1)}$$

as a result of applying the Poincaré inequality to $p^{(1)}$, taking into account the boundary condition $p^{(1)}(1) = 0$ for $p \in \operatorname{dom}(\mathcal{A}_{11})$. Again applying the Poincaré inequality for p, using that $p(0) = 0$ for $p \in \operatorname{dom}(\mathcal{A}_{11})$, we deduce

$$a(p,p) \geq \frac{1}{3}\left(\|p^{(2)}\|^2_{L^2(0,1)} + c\|p^{(1)}\|^2_{L^2(0,1)} + c^2\|p\|^2_{L^2(0,1)}\right) = \tilde{c}\|p\|^2_{H^2(0,1)}$$

with $\tilde{c} = \frac{1}{3}\min\{1, c^2\}$. Thus, by the Lax–Milgram theorem $\mathcal{A}_{11} : \operatorname{dom}(\mathcal{A}_{11}) \to H^{-2}(0,1)$ has a bounded inverse $\mathcal{A}_{11}^{-1} \in L(H^{-2}(0,1), H^2(0,1))$ with $\operatorname{im}\mathcal{A}_{11}^{-1} \subset \operatorname{dom}(\mathcal{A}_{11})$. Consequently, as a concatenation of bounded linear maps, the Schur complement $\mathcal{W}$ defined in (3.19) is bounded, as well.

By the above computations it follows also that $\mathcal{A}_{11}$ is coercive. Moreover, $\mathcal{A}_{21} = -\mathcal{A}_{12}^*$ follows by definition and integration by parts using the boundary conditions included in $\operatorname{dom}(\mathcal{A}_{11})$. Consequently, by an analogous argument as in the proof of Proposition 3.6, the Schur complement is dissipative. ∎

The numerical results for this example are presented in Figure 3.5. While the condition number of the top-left block of the operator behaves like $\mathcal{O}(h^{-4})$ due to the presence of fourth-order derivatives and the condition number of the top-right block scales like $\mathcal{O}(h^{-2})$ due to the presence of second-order derivatives, the condition number of the Schur complement is bounded uniformly in the mesh width. This reflects the boundedness proven in Proposition 3.9.

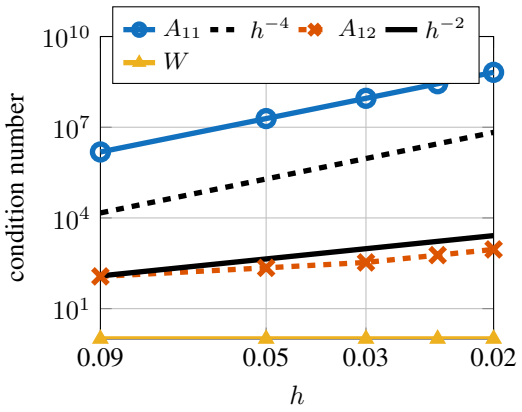

FIG. 3.5. *Condition numbers for the beam equation with Kelvin–Voigt damping.*

So far we have studied operator preconditioning for linear operators which have a symmetric part that is semidefinite. In the next section we apply these techniques to the solution of optimal control problems.

4. Application in optimal control. Having discussed the condition numbers in the context of operator preconditioning for partial differential equations, in this section we discuss two implementations of the structure-exploiting iterative schemes in the context of optimal control problems. We consider the prototypical problem

$$\min_{(x,u)\in\operatorname{dom}(\mathcal{A})\times U} \frac{1}{2}\|\mathcal{C}x - y_{\text{ref}}\|^2_Y + \frac{\lambda}{2}\|u - u_{\text{ref}}\|^2_U \qquad \text{s.t.} \quad \mathcal{A}x - \mathcal{B}u = f, \tag{4.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for Hilbert spaces X, Y, U. Here, $\mathcal{A} : \mathrm{dom}(\mathcal{A}) \subset X \to X$ is a densely defined linear operator, $\mathcal{B} \in L(U, X)$ is an input operator, $\mathcal{C} \in L(X, Y)$ is an output operator, $\lambda > 0$ is a regularization parameter, $f \in X$ is a source term, and $y_{\mathrm{ref}} \in Y$ as well as $u_{\mathrm{ref}} \in U$ are a reference output and a reference control.

Throughout the following we assume that $\mathcal{A}$ is an accretive operator giving rise to a splitting

$$\mathcal{A} = \mathcal{H} + \mathcal{S}$$

with symmetric $\mathcal{H}$ and skew-symmetric $\mathcal{S}$. For particular applications in stationary problems, we refer to the examples provided in Sections 3.1–3.5. Note that $\mathcal{A}$ may be the model of a semi-discretized time-dependent problem, as long as the time discretization is dissipativity-preserving, as discussed after (3.5). Using, e.g., an implicit midpoint discretization of a dissipative evolution equation $\dot{x} = -\mathcal{M}x$ with dissipative operator $-\mathcal{M}$ as in (3.5), with time step size $\delta t > 0$, the operator $\mathcal{A}$ in the optimal control problem (4.1) reads

$$\mathcal{A} = \begin{bmatrix} I & 0 & 0 & 0 & \cdots & 0 \\ I - \delta t/2\mathcal{M} & I + \delta t/2\mathcal{M} & 0 & 0 & \cdots & 0 \\ 0 & I - \delta t/2\mathcal{M} & I + \delta t/2\mathcal{M} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & I - \delta t/2\mathcal{M} & I + \delta t/2\mathcal{M} & 0 \\ 0 & 0 & 0 & 0 & I - \delta t/2\mathcal{M} & I + \delta t/2\mathcal{M} \end{bmatrix},$$

where the first block line assigns a given initial condition. Denoting by $\mathcal{H}_{\mathcal{M}}$ and $\mathcal{S}_{\mathcal{M}}$ the symmetric, respectively skew-symmetric, part of $\mathcal{M}$, we may split

$$I + \delta t/2\mathcal{M} = \underbrace{I + \delta t/2\mathcal{H}_{\mathcal{M}}}_{:=\mathcal{H}_d} \underbrace{-\delta t/2\mathcal{S}_{\mathcal{M}}}_{:=\mathcal{S}_d}.$$

Clearly, $\mathcal{H}_d$ is symmetric positive definite and $\mathcal{S}_d$ is skew-symmetric. Consequently, we set

$$\mathcal{A} = \mathcal{H} + \mathcal{S}$$

with

$$\mathcal{H} = \frac{1}{2}\begin{bmatrix} I & I - \delta t/2\mathcal{M}^* & 0 & 0 & \cdots & \\ I - \delta t/2\mathcal{M} & 2\mathcal{H}_d & I - \delta t/2\mathcal{M}^* & 0 & \cdots & 0 \\ 0 & I - \delta t/2\mathcal{M} & 2\mathcal{H}_d & I - \delta t/2\mathcal{M}^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & I - \delta t/2\mathcal{M} & 2\mathcal{H}_d & I - \delta t/2\mathcal{M}^* \\ 0 & 0 & 0 & 0 & I - \delta t/2\mathcal{M} & 2\mathcal{H}_d \end{bmatrix}$$

and

$$\mathcal{S} = \frac{1}{2}\begin{bmatrix} 0 & -I + \delta t/2\mathcal{M}^* & 0 & 0 & \cdots & \\ I - \delta t/2\mathcal{M} & 2\mathcal{S}_d & -I + \delta t/2\mathcal{M}^* & 0 & \cdots & 0 \\ 0 & I - \delta t/2\mathcal{M} & 2\mathcal{S}_d & -I + \delta t/2\mathcal{M}^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & I - \delta t/2\mathcal{M} & 2\mathcal{S}_d & -I + \delta t/2\mathcal{M}^* \\ 0 & 0 & 0 & 0 & I - \delta t/2\mathcal{M} & 2\mathcal{S}_d \end{bmatrix}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

such that $\mathcal{H} = \mathcal{H}^* \geq I$ and $\mathcal{S} = -\mathcal{S}^*$. Note that one would usually not apply a factorization-based solver to the full block operator $\mathcal{A}$ but rather solve the resulting system block-wise, recovering the iteration (3.6).

REMARK 4.1. In principle, we could also consider time-dependent problems in function spaces. There, $X = L^2(0,T;\bar{X})$ for another Hilbert space $\bar{X}$, $\mathcal{A} = \frac{\mathrm{d}}{\mathrm{d}t} - \mathcal{M}$ with dissipative $\mathcal{M} : \mathrm{dom}(\mathcal{M}) \subset \bar{X} \to \bar{X}$ being the generator of a strongly continuous semigroup, hence closed such that $\mathrm{dom}(\mathcal{M})$ becomes a Banach space when endowed with the graph norm. Correspondingly, we may set $\mathrm{dom}(\mathcal{A}) = H^1(0,T;\bar{X}) \cap L^2(0,T;\mathrm{dom}(\mathcal{M}))$ such that $\mathcal{A}$ is densely defined in $L^2(0,T;\bar{X})$; see, e.g., [21]. For parabolic equations where $\mathcal{M}$ gives rise to an analytic semigroup as the ones considered in Section 3 with the exception of the wave equation, $\mathcal{A}$ is closed due to the maximal parabolic regularity; see [10, Section 3.6]. For example, if $-\mathcal{M}$ is a second-order elliptic operator on $\bar{X} = H^{-1}(\Omega)$ in weak form, then $D(\mathcal{A}) = H^1(0,T;H^{-1}(\Omega)) \cap L^2(0,T;H^1(\Omega))$ corresponds to the usual $W(0,T)$-space used in variational theory [51].

Optimality conditions. Before going to the structure-exploiting linear solvers, we briefly recall the optimality conditions for (4.1), assuming that $\mathcal{A}$ has closed range; see [42, Theorem 1.1 and Remark 1.2]. Note that this closed-range assumption is satisfied, if, e.g., $\mathcal{A}$ is surjective, which is the case for differential operators in a suitable functional analytic setting (see, e.g., the operators in Section 3.1–3.5).

Let $(x,u) \in \mathrm{dom}(\mathcal{A}) \times U$ be optimal. Then there is $p \in \mathrm{dom}(\mathcal{A}^*)$ such that

$$\begin{bmatrix} \mathcal{C}^*\mathcal{C} & 0 & \mathcal{A}^* \\ 0 & \lambda I & -\mathcal{B}^* \\ \mathcal{A} & -\mathcal{B} & 0 \end{bmatrix} \begin{bmatrix} x \\ u \\ p \end{bmatrix} = \begin{bmatrix} \mathcal{C}^* y_{\mathrm{ref}} \\ \lambda\, u_{\mathrm{ref}} \\ f \end{bmatrix}. \tag{4.2}$$

We now discuss three different routes to leverage symmetric/skew-symmetric splittings in the optimality system (4.2).

4.1. Direct splitting of the optimality system. In a first approach, we do not require any structure of the constraint operator $\mathcal{A}$. Multiplying the last block row in (4.2) by minus one, we may split the optimality system into its symmetric and skew-symmetric part via

$$\mathcal{H} = \begin{bmatrix} \mathcal{C}^*\mathcal{C} & 0 & 0 \\ 0 & \lambda I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{S} = \begin{bmatrix} 0 & 0 & \mathcal{A}^* \\ 0 & 0 & -\mathcal{B}^* \\ -\mathcal{A} & \mathcal{B} & 0 \end{bmatrix}.$$

Here, the symmetric part is not invertible. If $\mathcal{C}^*\mathcal{C}$ is invertible, then we may apply the Schur complement to the top-left 2x2 block of the symmetric part. This leads to a problem governed by the matrix

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \end{bmatrix} \begin{bmatrix} \mathcal{C}^*\mathcal{C} & 0 \\ 0 & \lambda I \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}^* \\ \mathcal{B}^* \end{bmatrix} = \mathcal{A}(\mathcal{C}^*\mathcal{C})^{-1}\mathcal{A}^* + \frac{1}{\lambda}\mathcal{B}\mathcal{B}^*. \tag{4.3}$$

This operator is (formally) symmetric and semidefinite. If $\mathcal{A}$ is an isomorphism, as in many PDE applications, then the first term is even definite. Hence, the method of choice would be a suitable variant of the conjugate gradient method. Note, however, that the invertibility of $\mathcal{C}^*\mathcal{C}$, which in many applications with partial differential equations implies observation on the whole domain, would be very restrictive. This problem could be resolved by a further partitioning of $\mathcal{C}$ into the observable and unobservable part and then using another Schur complement—again an approach that is often not feasible for large-scale problems. In the design of preconditioners, the goal is to reflect both terms on the right-hand side of (4.3)

ETNA
Kent State University and
Johann Radon Institute (RICAM)

equally well; typically this is done for the case when $\mathcal{C}^*\mathcal{C}$ is invertible; see [38, 40, 45] among others. In case of non-invertible observation term, for the sake of preconditioning, a perturbation rendering $\mathcal{C}^*\mathcal{C}$ invertible could be used but still would only be useful when dealing with the full saddle-point formulation.

4.2. Condensed formulation. A different but also well-known approach in optimal control of partial differential equations is the reduction to the control variable via the state-to-control map; see, e.g., [51]. For this, we assume that the operator $\mathcal{A}$ in the optimal control problem (4.1) is boundedly invertible such that we may eliminate the state via $x = \mathcal{A}^{-1}\mathcal{B}u + \mathcal{A}^{-1}f$. This leads to the reduced unconstrained optimization problem

$$\min_{u\in U} f(u) := \frac{1}{2}\|\mathcal{C}(\mathcal{A}^{-1}\mathcal{B}u + \mathcal{A}^{-1}f) - y_{\mathrm{ref}}\|_Y^2 + \frac{\lambda}{2}\|u - u_{\mathrm{ref}}\|_U^2.$$

Due to the strict convexity of the cost function, solving this problem amounts to solving the first-order necessary (and sufficient) optimality condition at the optimal control $u^* \in U$

$$0 = \nabla f(u^*) = \left((\mathcal{C}\mathcal{A}^{-1}\mathcal{B})^*\mathcal{C}\mathcal{A}^{-1}\mathcal{B} + \lambda I\right) u^* + (\mathcal{C}\mathcal{A}^{-1}\mathcal{B})^*\mathcal{C}(\mathcal{A}^{-1}f - y_{\mathrm{ref}}). \tag{4.4}$$

Here, the governing operator

$$(\mathcal{C}\mathcal{A}^{-1}\mathcal{B})^*\mathcal{C}\mathcal{A}^{-1}\mathcal{B} + \lambda I \tag{4.5}$$

is symmetric and strictly accretive such that we can apply the conjugate gradient method. However, an application of this matrix requires in each step the solution of the state and the adjoint equation, i.e., evaluating $\mathcal{A}^{-1}$ and $\mathcal{A}^{-*}$.

To implement these solution operators, we use the accretivity and the symmetric/skew-symmetric splitting of the governing operator $\mathcal{A} = \mathcal{H} + \mathcal{S}$, which clearly also carries over to the adjoint operator $\mathcal{A}^* = \mathcal{H}^* + \mathcal{S}^*$. More precisely, we solve both the state and the adjoint equation via GMRES, Rapoport's, or Widlund's method endowed with preconditioning by the symmetric part or an approximation of it such as an incomplete Cholesky factorization or an algebraic multigrid method. While we ignore a possible source of inexactness in these approximations, we would like to again mention the work [15], where flexible methods that tolerate inexact evaluations of the preconditioner are discussed.

In the following we denote again by A, H, and S the Galerkin projections of the operators $\mathcal{A}$, $\mathcal{H}$, and $\mathcal{S}$ onto a suitable finite element space, respectively. We then use the following methods and their abbreviations.:

Methods for solving $\mathcal{A}^{-1}$ and $\mathcal{A}^{-*}$. In the following, the methods considered to evaluate state and adjoint equation in the outer CG loop applied to the system governed by the (discretization of the) operator (4.5) are:

- Direct: Direct solution by means of an LU factorization of A.
- ILU: Incomplete LU factorization of A with prescribed drop tolerance.
- GMRES: Generalized Minimal Residual Method for $Ax = b$ and $A^*p = b$ without preconditioner.
- GMRES(IC): GMRES with an incomplete Cholesky factorization $LL^\top \approx H$ as left preconditioner.
- GMRES(MG): GMRES with algebraic multigrid method for H as a left preconditioner. Here, we use the algebraic multigrid method as part of the HSL library [47].
- Rapoport: Rapoport's method [39] as described in Section 2.
- Widlund: Widlund's method [54] as described in Section 2.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We solve the reduced optimality system (4.4) with a conjugate gradient method up to a relative tolerance $\mathrm{cgtol} > 0$. In each step, we solve state and adjoint equation with one of the methods described by the columns of Table 4.1. Note that for inner solvers requiring a stopping criterion, we use the inner relative tolerance $\mathrm{cgtol}/10$ adapted to the tolerance of the outer CG loop. The full algorithm is given in Algorithm 1. The inner solvers and their respective parameters with respect to the outer CG tolerance are summarized in Table 4.1.

Algorithm 1 Condensed approach.

Require: OCP (4.1), outer tolerance cgtol, inner tolerance $\mathrm{innertol}$, initial outer iterate u^0

1: Assemble the method to evaluate A^{-1} and A^{-*} up to tolerance $\mathrm{innertol}$ according to Table 4.1.
2: Apply a conjugate gradient method to (4.4) up to tolerance cgtol.
3: **return** Approximation u^k of optimal control u^*.

TABLE 4.1
Used inner solver for the outer CG iteration with tolerance cgtol.

	direct	ILU	GMRES	GMRES(IC)	GMRES(MG)	Rapoport	Widlund
assembly	lu(A)	ilu(A,cgtol/100)	-	ichol(H,10^{-1})	multigrid (2c)	multigrid (2c)	multigrid (2c)
tolerance	-	-	cgtol/10	cgtol/10	cgtol/10	cgtol/10	cgtol/10

In the following, we evaluate the suggested approach by means of two applications, namely the advection-diffusion-reaction equation from Section 3.1 and the Stokes equation from Section 3.2. The code can be found at

https://github.com/maschaller/indefinite_solvers.

All computations are performed on a compute server with two AMD EPYC 9534 64-Core processors and 1.5 TB RAM.

4.2.1. Advection-diffusion-reaction equation. As a first example, consider the three-dimensional stationary advection-diffusion-reaction equation introduced in Section 3.1 on the unit cube $\Omega = [0,1]^3$. We choose a constant diffusivity $\nu \equiv 1$, an advection term $\mathbf{b} \equiv \begin{bmatrix}-0.5 & 0 & 0\end{bmatrix}^\top$, a reaction term $\mathbf{c} \equiv 1$, and a source term $f \equiv 10$. For the input-output configuration, we choose full observation and control, that is, $X = Y = U = L^2(\Omega)$, $B = C = I$, and vanishing reference functions $y_{\mathrm{ref}} = u_{\mathrm{ref}} \equiv 0$.

The numerical results for varying regularization parameters λ and varying mesh sizes are displayed in Figure 4.1. Therein, a smaller λ implies a larger condition number of the operator (4.5), hence leading to a higher number of outer CG iterations.

In the left column of Figure 4.1, we depict the total time required for the full outer iteration. This includes the assembly of preconditioners and factorizations. We observe that the direct methods (LU and ILU factorizations) fail after already a few refinements of the grid, as is to be expected for the three-dimensional setting. The iterative inner solvers all perform similarly in this log scale, therefore we provide a comparison with a linear scale in Figure 4.2. Therein, observe that the multigrid-preconditioned GMRES performs best (being an optimal method), while GMRES without preconditioner performs worst. Widlund's and Rapoport's method lead to very similar iteration numbers, which is to be expected due to their strong similarity; see Section 2.

In the middle column of Figure 4.1, we display the total number of outer CG iterations. We observe that GMRES and the incomplete-Cholesky-preconditioned GMRES lead to a

ETNA
Kent State University and
Johann Radon Institute (RICAM)

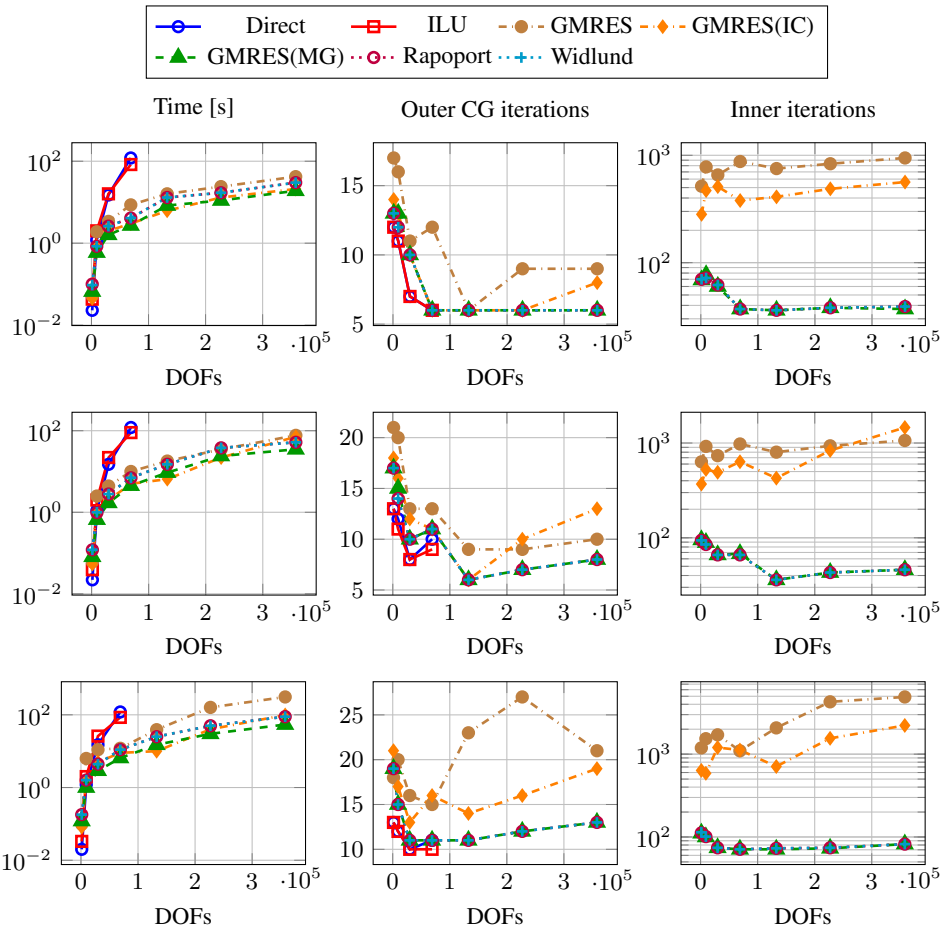

FIG. 4.1. *Advection-diffusion-reaction equation: solution of the condensed system with outer tolerance* $\mathrm{cgtol} = 10^{-4}$ *and* $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ *(top to bottom).*

higher number of outer iterations. We expect this to be due to an error estimator in the inner solves leading to termination before meeting the specified tolerance of $\mathrm{cgtol}/10$.

The right column of Figure 4.1 provides the total number of inner iterations for state and adjoint solves. Unpreconditioned GMRES has the highest iteration numbers (as to be expected), followed by the incomplete-Cholesky-preconditioned GMRES. The reason for this is that the incomplete Cholesky factorization with fixed fill-in does not represent a mesh-independent preconditioner (in contrast to the algebraic multigrid method), hence performing significantly worse for a larger number of spatial grid points. We briefly comment on the storage requirements of the discussed methods. Complete factorizations (such as LU or Cholesky factorizations) typically suffer from significant fill-in as the row operations introduce additional entries. This may lead to significantly higher storage compared to the original sparse matrices obtained by finite element discretization. Our choice of incomplete factorizations alleviates this fill-in. While this is desirable in terms of storage requirements, it comes at the cost of losing mesh-independence of the preconditioned operator. For algebraic multigrid methods, storage requirements are more difficult to measure as it involves a matrix operator defined on various levels. Thus, one often compares the number of nonzeros on all levels

ETNA
Kent State University and
Johann Radon Institute (RICAM)

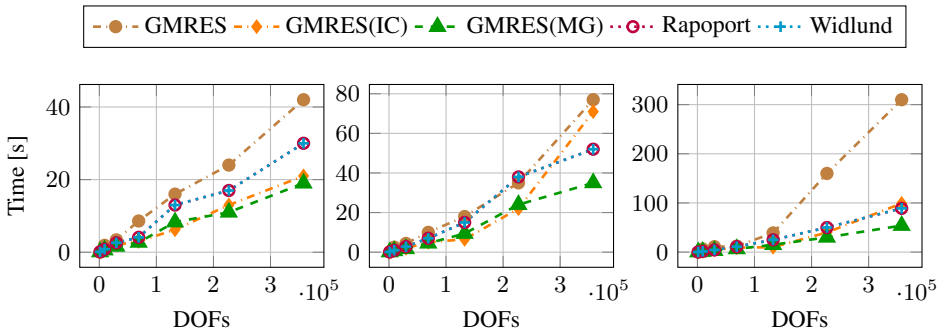

FIG. 4.2. *Time of the total outer iteration for indirect inner solvers. Outer tolerance* $\mathrm{cgtol} = 10^{-4}$ *and* $\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ *(left to right).*

versus the number on the finest (original) level. The author in [19] reports this ratio to be less than two for problems of the type we consider here.

Thus, we may conclude that GMRES, Rapoport's, and Widlund's method preconditioned with an algebraic multigrid method for the symmetric part yield the best inner solvers, both in terms of required iterations and the total time. GMRES with incomplete-Cholesky preconditioner performs well in view of total required time, however, it requires significantly more inner iterations.

4.2.2. Stokes equation. As a second example, we consider the two-dimensional pressure-stabilized Stokes equation of Section 3.2. We choose again full observations and a source term $f \equiv \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$. We use the same parameters as in Table 4.1 with the difference that we now use four multigrid V-cycles and an incomplete Cholesky factorization with drop tolerance 10^{-2}.

The results are depicted in Figure 4.3 for the stabilized case. The upper row displays the case of pressure regularization in the full H^1-norm, while the lower row only includes the gradient term. Again, we observe that the direct methods fail after a few refinements. The unpreconditioned GMRES method and GMRES preconditioned by an incomplete Cholesky factorization perform well in terms of outer CG iterations, however, they require a high amount of computation time due to the high number of inner iterations. Again, GMRES endowed with an AMG preconditioner for the symmetric part performs best. Rapoport's and Widlund's method fail due to lack of symmetry after a few refinement steps due to the inexact evaluation of the preconditioner by the multigrid method, motivating future research using flexible methods as suggested in [15].

4.3. Projected conjugate gradients. As another method, we suggest a conjugate gradient variant applied to the full optimality system (4.2) without eliminating the state variable via the state-control map.

As the governing operator is not positive definite (being a saddle-point matrix and due to the zero in the bottom-right block), we use the constraint preconditioner

$$\mathcal{P} := \begin{bmatrix} 0 & 0 & \mathcal{A}^* \\ 0 & \lambda I & -\mathcal{B}^* \\ \mathcal{A} & -\mathcal{B} & 0 \end{bmatrix},$$

which ensures that the iterates $\begin{bmatrix} x_k & u_k \end{bmatrix}^\top$ evolve in the kernel of the equality constraint. Thus, the proposed method falls under the class of projected preconditioned conjugate gradient (PPCG) methods. Such a preconditioner was used, e.g., in [30] in the context of affine covariant composite step methods. This preconditioner shares two important features:

ETNA
Kent State University and
Johann Radon Institute (RICAM)

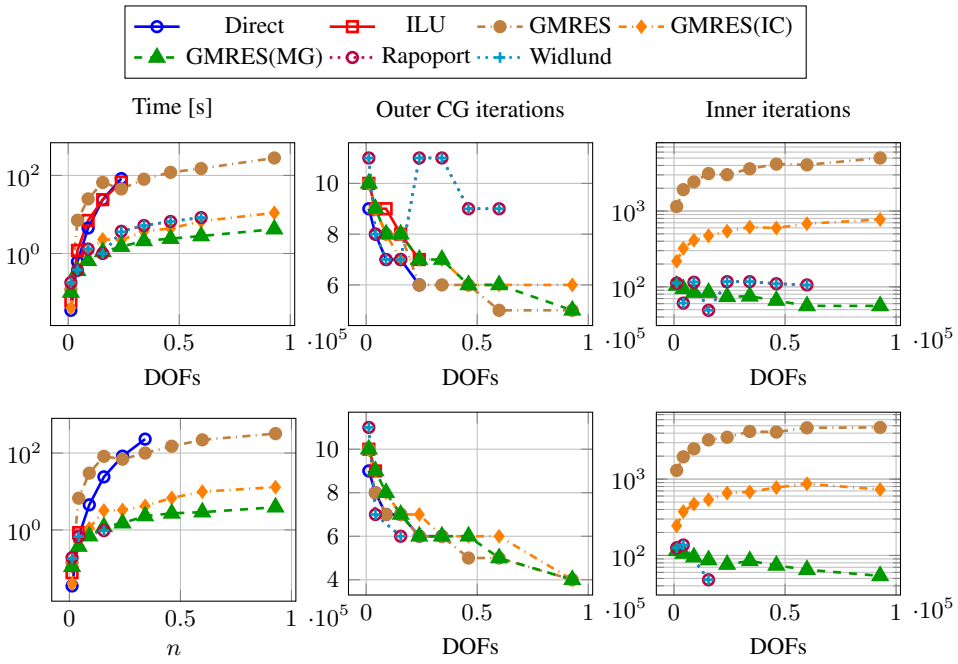

FIG. 4.3. *Stabilized Stokes equation: solution of the condensed system with outer tolerance* 10^{-4} *and* $\lambda = 10^{-1}$. *Top:* $s_1 = s_2 = 1$. *Bottom:* $s_1 = 0$, $s_2 = 1$.

1. If the conjugate gradient iteration is started with $\begin{bmatrix} x^0 & u^0 & p^0 \end{bmatrix}^\top$ such that the primal variables are admissible, i.e., that $\mathcal{A}x^0 + \mathcal{B}u^0 = 0$, then all residuals vanish in the last block component such that the preconditioned system gives rise to search directions $\begin{bmatrix} d_x & d_u & d_p \end{bmatrix}^\top$ with $\mathcal{A}d_x + \mathcal{B}d_u = 0$. Hence, the iteration yields a gradient method fully evolving in the kernel of the equality constraint, hence removing indefiniteness.

2. If a solver for $\mathcal{A}$ and $\mathcal{A}^*$ is available, then this preconditioner can be efficiently evaluated in a block-row-wise fashion starting with the top-left block, i.e., by the preconditioner

$$\mathcal{P}^{-1} = \begin{bmatrix} \frac{1}{\lambda}\mathcal{A}^{-1}\mathcal{B}\mathcal{B}^*\mathcal{A}^{-*} & \frac{1}{\lambda}\mathcal{A}^{-1}\mathcal{B} & \mathcal{A}^{-1} \\ \frac{1}{\lambda}\mathcal{B}^*\mathcal{A}^{-*} & \frac{1}{\lambda}I & 0 \\ \mathcal{A}^{-*} & 0 & 0 \end{bmatrix}. \tag{4.6}$$

Thus, the preconditioned version of the optimality system (4.2) is given by

$$\mathcal{P}^{-1}\begin{bmatrix} \mathcal{C}^*\mathcal{C} & 0 & \mathcal{A}^* \\ 0 & \lambda I & -\mathcal{B}^* \\ \mathcal{A} & -\mathcal{B} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda}\mathcal{A}^{-1}\mathcal{B}\mathcal{B}^*\mathcal{A}^{-*}\mathcal{C}^*\mathcal{C} + I & 0 & 0 \\ \frac{1}{\lambda}\mathcal{B}^*\mathcal{A}^{-1}\mathcal{C}^*\mathcal{C} & I & 0 \\ \mathcal{A}^{-1}\mathcal{C}^*\mathcal{C} & 0 & I \end{bmatrix},$$

which is a bounded block operator as it only involves inverses of differential operators that have a smoothing effect. Further, the efficient numerical evaluation of this preconditioner requires only one solve with $\mathcal{A}$ and $\mathcal{A}^*$.

In view of the point above, we again use the solvers tailored to the structure of $\mathcal{A} = \mathcal{H} + \mathcal{S}$ such as the (preconditioned) GMRES, Widlund's, or Rapoport's method. As these approaches constitute iterative methods, inexact evaluations of the preconditioner are inevitable. In this work, to focus on the main contribution, we set a very low solver tolerance such that

ETNA
Kent State University and
Johann Radon Institute (RICAM)

the inner solves are performed with high accuracy; see the setting described in Table 4.2. However, various works have considered inexact CG methods. A general framework for inexact evaluations of the preconditioner building upon a bound on the corresponding residual was developed in [23]. Moreover, inexact PPCG methods in the context of inexact SQP methods were discussed in [27], and in [43] a primal-dual projection method to alleviate inexact solves is presented.

In Algorithm 2, we summarize our proposed method. The specifications of the inner solvers are given in Table 4.2.

Algorithm 2 Projected CG.

Require: Optimal control problem (4.1), outer tolerance cgtol, initial outer iterate (x^0, u^0, p^0).

1: Assemble the method to evaluate A^{-1} and A^{-*} up to tolerance $\mathrm{innertol}$ according to Table 4.2.
2: Apply a conjugate gradient method to (4.2) with preconditioner (4.6) up to tolerance cgtol.
3: **return** An approximation (x^k, u^k, p^k) of the optimal state, control, and adjoint (x, u, p).

TABLE 4.2
Used inner solver for the evaluation of the constraint preconditioner for the outer CG iteration with tolerance cgtol.

	direct	ILU	GMRES	GMRES(IC)	GMRES(MG)	Rapoport	Widlund
assembly	lu(A)	ilu(A)	-	ichol(H,10^{-1})	multigrid (2c)	multigrid	multigrid
tolerance	-	cgtol/10 (drop)	1e-6	1e-6	1e-6	1e-6	1e-6

In Figure 4.4, we display the resulting runtime and iteration numbers for the outer CG loop applied to the optimality system (4.2) endowed with the proposed constraint preconditioner.

In the top row of Figure 4.4, we provide the results for the advection-diffusion-reaction equation. We see that the total number of outer iterations is robust with respect to the refinements. The reason for this is that the constraint preconditioned operator is bounded, and hence, the discretization has a uniformly bounded condition number. However, as in the approach with state elimination suggested in Section 4.2, the time and inner iterations required by GMRES and the incomplete-Cholesky-preconditioned variant are significantly higher as for the other methods. Thus, we conclude that also in this approach using a constraint preconditioner, again the tailored methods, being GMRES, Rapoport's methods, or Widlund's method endowed with an algebraic multigrid preconditioner for the symmetric part, perform best.

5. Conclusion. We have analyzed and evaluated tailored methods and preconditioning for accretive systems occurring in partial differential equations such as dissipative and port-Hamiltonian systems. Therein, the preconditioner is obtained by the symmetric part of the underlying operator, allowing for short-recurrence Krylov subspace methods using suitable inner products such as Widlund's and Rapoport's method. In the first part of this work, we have analyzed this approach in view of operator preconditioning for a wide range of differential operators occurring in advection-diffusion-reaction equations, fluid mechanics, wave propagation, or elasticity. In the second part we have proposed two approaches to include these methods in a large-scale optimal control solver, i.e., a reduced formulation and a

ETNA
Kent State University and
Johann Radon Institute (RICAM)

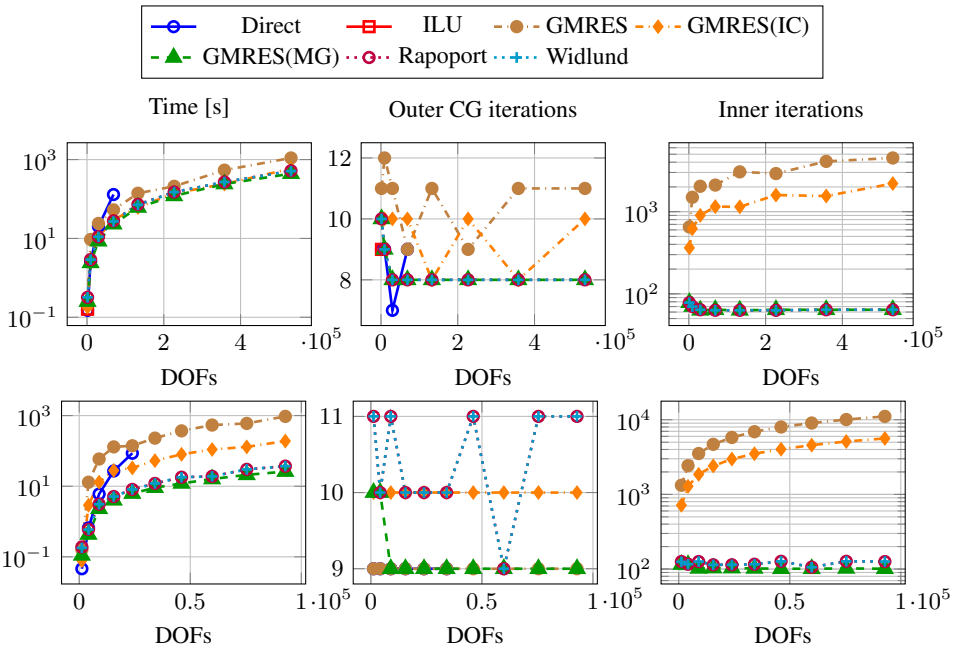

FIG. 4.4. *PPCG: Solution of optimality system with PPCG for advection-diffusion-reaction equation with outer tolerance* 10^{-4} *and* $\lambda = 10^{-4}$ *(top) and Stokes equation with outer tolerance* 10^{-3} *and* $\lambda = 10^{-5}$ *(bottom).*

projected preconditioned CG method for the full optimality system. We have illustrated that approximating the symmetric preconditioner via algebraic multigrid methods or incomplete Cholesky factorizations leads to highly efficient and robust solvers. As future work we plan to analyze this approach for nonlinear partial differential operators and abstract differential-algebraic operators.

REFERENCES

[1] F. ACHLEITNER, A. ARNOLD, AND V. MEHRMANN, *Hypocoercivity in algebraically constrained partial differential equations with application to Oseen equations*, J. Dynam. Differential Equations, 37 (2025), pp. 1747–1786.

[2] F. ACHLEITNER, A. ARNOLD, V. MEHRMANN, AND E. A. NIGSCH, *Hypocoercivity in Hilbert spaces*, J. Funct. Anal., 288 (2025), Paper No. 110691, 51 pages.

[3] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, 2nd ed., Elsevier, Amsterdam, 2003.

[4] M. ALNÆS, J. BLECHTA, J. HAKE, A. JOHANSSON, B. KEHLET, A. LOGG, C. RICHARDSON, J. RING, M. E. ROGNES, AND G. N. WELLS, *The FEniCS project version 1.5*, Arch. Num. Softw., 3 (2015), pp. 9–23.

[5] Y. ARLINSKIĬ AND C. TRETTER, *Everything is possible for the domain intersection* $\operatorname{dom} T \cap \operatorname{dom} T^*$, Adv. Math., 374 (2020), Paper No. 107383, 46 pages.

[6] O. AXELSSON AND J. KARÁTSON, *Equivalent operator preconditioning for elliptic problems*, Numer. Algorithms, 50 (2009), pp. 297–380.

[7] A. H. BAKER, E. R. JESSUP, AND T. MANTEUFFEL, *A technique for accelerating the convergence of restarted GMRES*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 962–984.

[8] A. BARTEL, M. DIAB, A. FROMMER, M. GÜNTHER, AND N. MARHEINEKE, *Splitting techniques for DAEs with port-Hamiltonian applications*, Appl. Numer. Math., 214 (2025), pp. 28–53.

[9] A. BARTEL, M. GÜNTHER, B. JACOB, AND T. REIS, *Operator splitting based dynamic iteration for linear differential-algebraic port-Hamiltonian systems*, Numer. Math., 155 (2023), pp. 1–34.

[10] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, 2nd ed., Birkhäuser, Boston, 2007.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[11] M. BENZI, M. NG, Q. NIU, AND Z. WANG, *A relaxed dimensional factorization preconditioner for the incompressible Navier-Stokes equations*, J. Comput. Phys., 230 (2011), pp. 6185–6202.
[12] D. BOFFI, *Finite element approximation of eigenvalue problems*, Acta Numer., 19 (2010), pp. 1–120.
[13] D. BRAESS AND R. SARAZIN, *An efficient smoother for the Stokes problem*, Appl. Numer. Math., 23 (1997), pp. 3–19.
[14] P. G. CIARLET, *Mathematical Elasticity. Volume I. Three-Dimensional Elasticity*, SIAM, Philadelphia, 2022.
[15] M. DIAB, A. FROMMER, AND K. KAHL, *A flexible short recurrence Krylov subspace method for matrices arising in the time integration of port-Hamiltonian systems and ODEs/DAEs with a dissipative Hamiltonian*, BIT Numer. Math., 63 (2023), Paper No. 57, 21 pages.
[16] H. C. ELMAN, D. J. SILVESTER, AND A. J. WATHEN, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, 2nd ed., Oxford University Press, Oxford, 2014.
[17] K.-J. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Springer, New York, 2000.
[18] V. FABER, J. LIESEN, AND P. TICHÝ, *The Faber-Manteuffel theorem for linear operators*, SIAM J. Numer. Anal., 46 (2008), pp. 1323–1337.
[19] R. D. FALGOUT, *An introduction to algebraic multigrid*, Tech. Rep., Lawrence Livermore National Laboratory (LLNL), Livermore, 2006.
[20] B. FARKAS, B. JACOB, T. REIS, AND M. SCHMITZ, *Operator splitting based dynamic iteration for linear infinite-dimensional port-Hamiltonian systems*, Preprint on arXiv, 2023. https://arxiv.org/abs/2302.01195
[21] B. FARKAS, B. JACOB, M. SCHALLER, AND M. SCHMITZ, *Dissipativity-based time domain decomposition for optimal control of hyperbolic PDEs*, Preprint on arXiv, 2025. https://arxiv.org/abs/2507.07812
[22] G. H. GOLUB AND D. VANDERSTRAETEN, *On the preconditioning of matrices with skew-symmetric splittings*, Numer. Algorithms, 25 (2000), pp. 223–239.
[23] G. H. GOLUB AND Q. YE, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999/00), pp. 1305–1320.
[24] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, SIAM, Philadelphia, 2011.
[25] C. GÜDÜCÜ, J. LIESEN, V. MEHRMANN, AND D. B. SZYLD, *On non-Hermitian positive (semi)definite linear algebraic systems arising from dissipative Hamiltonian DAEs*, SIAM J. Sci. Comput., 44 (2022), pp. A2871–A2894.
[26] A. GÜNNEL, R. HERZOG, AND E. SACHS, *A note on preconditioners and scalar products in Krylov subspace methods for self-adjoint problems in Hilbert space*, Electron. Trans. Numer. Anal., 41 (2014), pp. 13–20. https://etna.ricam.oeaw.ac.at/vol.41.2014/pp13-20.dir/pp13-20.pdf
[27] M. HEINKENSCHLOSS AND D. RIDZAL, *A matrix-free trust-region SQP method for equality constrained optimization*, SIAM J. Optim., 24 (2014), pp. 1507–1541.
[28] R. HIPTMAIR, *Operator preconditioning*, Comput. Math. Appl., 52 (2006), pp. 699–706.
[29] B. JACOB, C. TRUNK, AND M. WINKLMEIER, *Analyticity and Riesz basis property of semigroups associated to damped vibrations*, J. Evol. Equ., 8 (2008), pp. 263–281.
[30] L. LUBKOLL, A. SCHIELA, AND M. WEISER, *An affine covariant composite step method for optimization with PDEs as equality constraints*, Optim. Methods Softw., 32 (2017), pp. 1132–1161.
[31] J. MÁLEK AND Z. STRAKOŠ, *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*, SIAM, Philadelphia, 2015.
[32] M. MANGUOĞLU AND V. MEHRMANN, *A robust iterative scheme for symmetric indefinite systems*, SIAM J. Sci. Comput., 41 (2019), pp. A1733–A1752.
[33] ———, *A two-level iterative scheme for general sparse linear systems based on approximate skew-symmetrizers*, Electron. Trans. Numer. Anal., 54 (2021), pp. 370–391. https://etna.ricam.oeaw.ac.at/vol.54.2021/pp370-391.dir/pp370-391.pdf
[34] ———, *Robust iterative methods for linear systems with saddle point structure*, Preprint on arXiv, 2025. https://arxiv.org/abs/2502.21174v1
[35] K.-A. MARDAL AND R. WINTHER, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.
[36] C. MEHL, V. MEHRMANN, AND M. WOJTYLAK, *Matrix pencils with coefficients that have positive semidefinite Hermitian parts*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 1186–1212.
[37] ———, *Spectral theory of infinite dimensional dissipative Hamiltonian systems*, J. Dynam. Differential Equations, (2025), 27 pages. https://doi.org/10.1007/s10884-025-10420-y
[38] J. W. PEARSON, M. STOLL, AND A. J. WATHEN, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 1126–1152.
[39] D. RAPOPORT, *A nonlinear Lanczos algorithm and the stationary Navier-Stokes equation*, PhD. Thesis, New York University, New York, 1978.
[40] T. REES, H. S. DOLLAR, AND A. J. WATHEN, *Optimal solvers for PDE-constrained optimization*, SIAM J. Sci. Comput., 32 (2010), pp. 271–298.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[41] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[42] A. SCHIELA, *A concise proof for existence and uniqueness of solutions of linear parabolic PDEs in the context of optimal control*, Systems Control Lett., 62 (2013), pp. 895–901.
[43] A. SCHIELA, M. STÖCKLEIN, AND M. WEISER, *A primal-dual projection algorithm for efficient constraint preconditioning*, SIAM J. Sci. Comput., 43 (2021), pp. A4095–A4120.
[44] A. SCHIELA AND S. ULBRICH, *Operator preconditioning for a class of inequality constrained optimal control problems*, SIAM J. Optim., 24 (2014), pp. 435–466.
[45] J. SCHÖBERL AND W. ZULEHNER, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 752–773.
[46] J. SCOTT AND M. TŮMA, *Algorithms for Sparse Linear Systems*, Birkhäuser/Springer, Cham, 2023.
[47] STFC COMPUTATIONAL MATHEMATICS GROUP, *The HSL mathematical software library*, version HSL 2023. http://www.hsl.rl.ac.uk/.
[48] M. STOLL AND A. WATHEN, *Combination preconditioning and the Bramble–Pasciak$^+$ preconditioner*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 582–608.
[49] D. B. SZYLD AND O. B. WIDLUND, *Variational analysis of some conjugate gradient methods*, East-West J. Numer. Math., 1 (1993), pp. 51–74.
[50] A. TOSELLI AND O. WIDLUND, *Domain Decomposition Methods—Algorithms and Theory*, Springer, Berlin, 2005.
[51] F. TRÖLTZSCH, *Optimal Control of Partial Differential Equations*, American Mathematical Society, Providence, 2010.
[52] A. J. WATHEN, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.
[53] A. J. WATHEN, *Preconditioning*, Acta Numer., 24 (2015), pp. 329–376.
[54] O. WIDLUND, *A Lanczos method for a class of nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 801–812.
[55] H. ZWART AND V. MEHRMANN, *Abstract dissipative Hamiltonian differential-algebraic equations are everywhere*, DAE Panel, 2 (2024), 36 pages. https://doi.org/10.52825/dae-p.v2i.957