dc_212_11

Operator methods for the numerical solution of elliptic PDE problems

Summary of the D.Sc. dissertation

Karátson János

ELTE, Budapest, 2011

I. The goal of the research

The field of this thesis is the numerical solution of linear and nonlinear elliptic partial differential equations. These classes of equations are widespread in modelling various phenomena in science, hence their numerical solution has continuously been a subject of extensive research. The common way is to discretize the problem, which leads to an algebraic system normally of very large size, then usually a suitable iterative solver is applied. An important measure of efficiency is the optimality property, which requires that the computational cost should be of $O(n)$ where n denotes the degrees of freedom in the algebraic system. This holds for some special PDE problems, which can then be used as preconditioners to more general problems. Then a crucial property of the iteration is mesh independence, i.e. the number of iterations to achieve prescribed accuracy should be bounded independently of n in order to preserve the optimality.

The numerical study of elliptic PDEs has often relied on Hilbert space theory, to name e.g. the finite element method and the Lax-Milgram approach as fundamental examples. In fact, it has been held since a famous paper of Kantorovich that the methods of functional analysis can be used to develop practical algorithms with as much success as they have been used for the theoretical study of these problems. Thus one can often incorporate the properties of the continuous PDE problem, from the Hilbert space in which it is posed, into the numerical procedure. The importance of this is expressed by the law of J.W. Neuberger, stating that analytical and numerical difficulties always come paired.

A fundamental approach here is the Sobolev gradient theory of J.W. Neuberger, which was shown to give a prospect for a unified theory of PDEs with extensively wide numerical applications. Sobolev gradients enable us to define preconditioned problems with significantly improved convergence via auxiliary operators in Sobolev space. In the linear case, a strongly related approach comes from the theory of equivalent operators by Manteuffel and his co-authors, which gives an organized treatment of mesh independent linear convergence based on Hilbert space theory. Moreover, they have shown that for a preconditioner arising from an operator, equivalence is essentially necessary for producing mesh independence, further, that this approach is competitive with multigrid and other state-of-the-art solvers.

The primary goal of this thesis is to complete the above theories such that an organized framework is obtained for treating a wide class of iterative methods for both linear and nonlinear problems. A particular attention is paid first to mesh independent superlinear convergence for linear problems, which is a counterpart of Manteuffel's results. For nonlinear problems our goal is to give a unified framework for treating gradient and Newton type methods. A common concept in both studies is the preconditioning operator, whose role is to produce a cheap approximation of the original operator in the linear case and of the current Jacobian operator in the nonlinear case. Our next goal is to show that this treatment results in various efficient computational algorithms that exploit the structure of the continuous PDE problem and in general produce mesh independence.

In addition, it will be shown that operator theory can be applied to study the reliability of the numerical solution. New results on the discrete maximum principle, which is an important measure of the qualitative reliability of the numerical scheme, will be given in a common Hilbert space framework. Then sharp a posteriori error estimates will be

established for nonlinear operator equations in Banach space, and shown to be applicable to several types of elliptic PDEs.

II. The methods of the research

In general, we study a linear or nonlinear operator equation

$$Lu = g \qquad \text{or} \qquad F(u) = b \tag{0.1}$$

(in a Hilbert or, more generally, Banach space) that will then model an elliptic PDE including boundary conditions. A Galerkin discretization yields a respective finite dimensional problem

$$L_h u_h = g_h \qquad \text{or} \qquad F_h(u_h) = b_h\,.$$

In the major part of this work we consider iterative methods. A one-step iterative method reads as

$$u_{i+1} = u_i - S_h^{-1}(L_h u_i - g_h), \qquad \text{or} \qquad u_{i+1} = u_i - B_h^{-1}(F_h(u_i) - b_h),$$

where S_h resp. B_h should be properly chosen, and B_h is in general allowed to depend on i. In the considered methods the basic idea is to obtain S_h resp. B_h as the discretization of a suitable operator in the given space.

Our study naturally uses the basic theory of iterative methods, including various types of widespread CG iterations instead of the above one-step method for linear problems. For instance, the CG method for symmetric linear systems $Au = b$ reads as

$$u_{k+1} = u_k + \alpha_k d_k, \text{ where } \alpha_k = -\tfrac{\langle r_k, d_k\rangle}{\langle Ad_k, d_k\rangle}\,; \quad d_{k+1} = r_{k+1} + \beta_k d_k, \text{ where } \beta_k = \tfrac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

where $r_k := Au_k - b$ are the residuals. Suitable generalizations for nonsymmetric systems (GCG-LS, GCG-LS(0), CGN iterations) will be applied together with their preconditioned versions, and their linear and superlinear convergence estimates will be used to start with. Similarly we rely on gradient and Newton type methods for nonlinear problems.

The study of iterative methods and then of the qualitative reliability will involve basic or special operator theoretical tools for the equations (0.1):

- Hilbert space calculus for linear operators: bilinear forms, energy spaces, energy norms, coercivity properties, weak forms of unbounded operators.
- Theory of singular values of compact operators.
- Estimates for elliptic operators: Lebesgue and Sobolev norm calculations, Sobolev embeddings.
- Galerkin discretizations in Hilbert spaces, finite element methods for elliptic problems.
- Nonlinear potential operators, monotone operators, minimization, norm estimates for nonlinear operators, weak forms of unbounded nonlinear operators.
- Banach space calculations, dual spaces, energy functionals.

Some basic references related to the topic are [5, 14, 15, 16, 18, 34].

III. A brief summary of the main results

The results are twofold. On the one hand, this work is theoretically oriented in the sense that many of the new results are related to Hilbert space theory, such as the introduction of new concepts in order to derive a general framework for certain classes and properties of iterative methods. On the other hand, the goal of this theory is to present efficient computational algorithms producing mesh independent convergence, which is illustrated with various examples: to this end, altogether fifteen subsections of the thesis are devoted to such applications.

The main results of this thesis can be grouped as follows.

- We introduce the notion of compact-equivalent linear operators, which expresses that preconditioning one of them with the other yields a compact perturbation of the identity, and prove the following principle for Galerkin discretizations: if the two operators (the original and preconditioner) are compact-equivalent then the preconditioned CGN method provides mesh independent superlinear convergence. This completes the analogous results of Manteuffel et al. on linear convergence. Mesh independence of superlinear convergence has not been established before.

  We characterize compact-equivalence for elliptic operators: if they have homogeneous Dirichlet conditions on the same portion of the boundary, then two elliptic operators are compact-equivalent if and only if their principal parts coincide up to a constant factor.

- We show that the introduction of the concept of S-bounded and S-coercive operators also gives a simplified framework for mesh independent linear convergence. In fact, the required uniform equivalence for the Galerkin discretizations is obtained here as a straightforward consequence.

- We also derive mesh independent superlinear convergence for the GCG-LS method for normal compact perturbations, and introduce the notion of weak symmetric part so that we can apply the abstract result to symmetric part preconditioning for general boundary conditions.

- Based on the above described theory, we present various efficient preconditioners that mostly produce mesh independent superlinear convergence for FEM discretizations of linear PDEs, including some computer realizations with symmetric preconditioners for nonsymmetric equations, parallelizable decoupled preconditioners for coupled systems, preconditioning operators with constant coefficients including nonsymmetric preconditioners.

- We introduce the concept of variable preconditioning, and show that this gives a unified framework to treat gradient and Newton type methods for monotone nonlinear problems. Applied in Sobolev spaces, we thus extend the Sobolev gradient theory of J.W. Neuberger to variable gradients. A general convergence theorem, which puts a quasi-Newton method in this context, enables us to achieve the quadratic convergence of Newton's method via potentially cheaper subproblems than those with Jacobians.

- Two theoretical contributions to Newton's method are given. First, related to the above-mentioned variable Sobolev gradients, we prove that Newton's method is an optimal variable gradient method in the sense that the descents in Newton's method are asymptotically steepest w.r. to both different directions and inner products. Second, we show via a suitable characterization that the theory of mesh independence is restricted in some sense: for elliptic problems, the quadratic convergence of Newton's method is mesh independent if and only if the elliptic equation is semilinear.

- We also give some new Sobolev gradient results for variational problems. These results, the variable preconditioning theory, and suitable combinations of inexact Newton iterations with our above-mentioned methods for linear problems form together a framework of preconditioning operators as a common approach to provide nonlinear solvers with mesh independent convergence. Based on these, we present various numerical applications of our iterative solution methods for nonlinear elliptic PDEs, including computer realizations for some real-life problems.

- Operator approach is used to establish results on the reliability of the numerical solution. First, a discrete maximum principle (DMP) is established in Hilbert space for proper Galerkin stiffness matrices, which allows us to derive DMPs for general nonlinear elliptic equations with mixed boundary conditions and then for nonlinear elliptic systems, for which classes no DMP has been established before. The results are applied to achieve the desired nonnegativity of the FEM solution of some real model problems. Finally, a sharp a posteriori error estimate is given in Banach space and then derived for nonlinear elliptic problems.

IV. A detailed summary of the results

1 Linear problems

In what follows, let H denote a Hilbert space. It will be assumed real unless explicitly stated to be complex. We are interested in solving an operator equation

$$Lu = g \tag{1.1}$$

for an unbounded linear operator L in H, where $g \in H$. Our main interest is superlinear convergence [10], which expresses that – roughly speaking – the same improvement in accuracy needs less effort in the final phase of the iteration than in the initial phase.

1.1 Compact-equivalent operators and superlinear convergence

1.1.1 S-bounded and S-coercive operators

The notion of compact-equivalent operators needs a preliminary notion of weak form of unbounded operators. This also clarifies in which space equation (1.1) is well-posed. For this, we will use an auxiliary (also unbounded) linear symmetric operator S in H which is coercive, i.e., there exists $p > 0$ such that $\langle Su, u\rangle \geq p\|u\|^2$ $(u \in D(S))$. Recall that the

energy space H_S is the completion of $D(S)$ under the inner product $\langle u, v\rangle_S = \langle Su, v\rangle$, and the coercivity of S implies $H_S \subset H$. The corresponding S-norm is denoted by $\|u\|_S$, and the space of bounded linear operators on H_S by $B(H_S)$.

Definition 1.1 Let S be a linear symmetric coercive operator in H. A linear operator L in H is said to be *S-bounded and S-coercive*, and we write $L \in BC_S(H)$, if the following properties hold:

(i) $D(L) \subset H_S$ and $D(L)$ is dense in H_S in the S-norm;

(ii) there exists $M > 0$ such that $|\langle Lu, v\rangle| \le M\|u\|_S\|v\|_S \quad (u, v \in D(L))$;

(iii) there exists $m > 0$ such that $\langle Lu, u\rangle \ge m\|u\|_S^2 \quad (u \in D(L))$.

Definition 1.2 For any $L \in BC_S(H)$, let $L_S \in B(H_S)$ be defined by

$$\langle L_S u, v\rangle_S = \langle Lu, v\rangle \qquad (u, v \in D(L)). \tag{1.2}$$

Such an L_S exists and is unique, and satisfies

$$|\langle L_S u, v\rangle_S| \le M\|u\|_S\|v\|_S, \qquad \langle L_S u, u\rangle_S \ge m\|u\|_S^2 \qquad (u, v \in H_S). \tag{1.3}$$

The Lax-Milgram lemma provides a *weak solution* of equation (1.1) defined by

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \tag{1.4}$$

1.1.2 Coercive elliptic operators

Now the corresponding class is described for elliptic problems. Let us define the elliptic operator

$$Lu \equiv -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu \qquad \text{for } u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \tag{1.5}$$

where $\frac{\partial u}{\partial \nu_A} = A\,\nu\cdot\nabla u$ denotes the weighted form of the normal derivative. For the formal domain of L to be used in Definition 1.1, we consider those $u \in H^2(\Omega)$ that satisfy the above boundary conditions and for which Lu is in $L^2(\Omega)$.

The following properties are assumed to hold:

Assumptions 1.1.1.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$;

(ii) $A \in (L^\infty \cap PC)(\overline{\Omega}, \mathbf{R}^{d\times d})$ and for all $x \in \overline{\Omega}$ the matrix $A(x)$ is symmetric; further, $\mathbf{b} \in W^{1,\infty}(\Omega)^d$ (i.e. $\partial_i b_j \in L^\infty(\Omega)$ for all $i, j = 1, ..., d$), $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_N)$;

(iii) we have the following properties which will imply coercivity: there exists $p > 0$ such that

$A(x)\xi\cdot\xi \ge p\,|\xi|^2$ for all $x \in \overline{\Omega}$ and $\xi \in \mathbf{R}^d$; $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b} \ge 0$ in Ω and $\hat{\alpha} := \alpha + \frac{1}{2}\,(\mathbf{b}\cdot\nu) \ge 0$ on Γ_N;

(iv) either $\Gamma_D \neq \emptyset$, or $\hat{c}$ or $\hat{\alpha}$ has a positive lower bound.

Let us also define a symmetric elliptic operator on the same domain Ω with otherwise analogous properties:

$$Su \equiv -\mathrm{div}\,(G\,\nabla u) + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0,\; \frac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{1.6}$$

which generates the energy space $H^1_D(\Omega) := \{u \in H^1(\Omega) : \, u_{|\Gamma_D} = 0\}$, under the following

Assumptions 1.1.2.

(i) Substituting G for A, Ω, Γ_D, Γ_N and G satisfy Assumptions 1.1.1;

(ii) $\sigma \in L^\infty(\Omega)$ and $\sigma \geq 0$; $\beta \in L^\infty(\Gamma_N)$ and $\beta \geq 0$; further, if $\Gamma_D = \emptyset$ then σ or β has a positive lower bound.

Proposition 1.1 *If Assumptions 1.1.1-2 hold, then the operator L is S-bounded and S-coercive in $L^2(\Omega)$, i.e., $L \in BC_S(L^2(\Omega))$.*

1.1.3 Compact-equivalent operators

Now the notion of compact-equivalent operators can be introduced.

Definition 1.3 Let L and N be S-bounded and S-coercive operators in H. We call L and N *compact-equivalent in H_S* if

$$L_S = \mu N_S + Q_S \tag{1.7}$$

for some constant $\mu > 0$ and compact operator $Q_S \in B(H_S)$.

We can characterize compact-equivalence for elliptic operators:

Theorem 1.1 *Let the elliptic operators L_1 and L_2 satisfy Assumptions 1.1.1 with the same Γ_N and Γ_D. Then L_1 and L_2 are compact-equivalent in $H^1_D(\Omega)$ if and only if their principal parts coincide up to some constant $\mu > 0$, i.e. $A_1 = \mu A_2$.*

1.1.4 Mesh independent superlinear convergence in Hilbert space

Equation (1.1) can be solved numerically using a Galerkin discretization in a subspace $V_h = span\{\varphi_1, \ldots, \varphi_n\} \subset H_S$. Finding the discrete solution requires solving an $n \times n$ system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{b}_h \tag{1.8}$$

with $\mathbf{b}_h = \{\langle g, \varphi_j \rangle\}_{j=1}^n$.

Now we present mesh independent superlinear convergence estimates in the case of compact-equivalent preconditioning. Bounds on the rate of superlinear convergence are given in the form of a sequence which is mesh independent and is determined only by the underlying operators.

For simplicity, in what follows, we will consider compact-equivalence with $\mu = 1$ in (1.7). This is clearly no restriction, since if a preconditioner N_S satisfies $L_S = \mu N_S + Q_S$ then we can consider the preconditioner μN_S instead.

1.1.5 Symmetric compact-equivalent preconditioners

Let us consider operators L and S as in Definition 1.1, and assume in addition that L and S are compact-equivalent with $\mu = 1$. Then (1.7) holds with $N_S = I$:

$$L_S = I + Q_S \tag{1.9}$$

with a compact operator Q_S. We apply the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.8). By (1.9), letting $\mathbf{Q}_h = \left\{\langle Q_S\varphi_j, \varphi_i\rangle_S\right\}_{i,j=1}^n$, the preconditioned system takes the form

$$(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \tilde{\mathbf{b}_h} \tag{1.10}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix.

In order to have mesh independent bounds, one must estimate the sums of eigenvalues of the perturbation matrices (which appear in the standard superlinear estimate) by those of the corresponding operator.

Proposition 1.2 *Let H be a complex Hilbert space. If Q_S is a normal compact operator in H_S and the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal, then*

$$\sum_{i=1}^{k}\left|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h)\right| \le \sum_{i=1}^{k}\left|\lambda_i(Q_S)\right| \qquad (k = 1, 2, \ldots, n).$$

If H is a real Hilbert space (as it is in this paper) then H and H_S can be extended to a complex Hilbert space. From Proposition 1.2 and the standard estimate we can then derive

Theorem 1.2 *Under the conditions of Proposition 1.2, the GCG-LS algorithm for system (1.10) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k = 1, ..., n), \quad \textit{where} \quad \varepsilon_k := \frac{2}{km}\sum_{j=1}^{k}\left|\lambda_j(Q_S)\right| \to 0 \quad \textit{as} \quad k \to \infty$$

and ε_k is a sequence independent of V_h.

The most important special case here is symmetric part preconditioning, when both normality assumptions are readily satisfied, in fact, Q_S is antisymmetric in H_S. Then the GCG-LS algorithm reduces to the truncated GCG-LS(0) version, the $\mathbf{L}_h$-norm equals the $\mathbf{S}_h$-norm and we have $m = 1$.

In the general case without normality, we have the following bounds and convergence:

Proposition 1.3 *Any compact operator Q_S in H_S satisfies the following relations:*

$$(a) \qquad \sum_{i=1}^{k}\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{S}_h^{-1}\mathbf{Q}_h) \le \sum_{i=1}^{k} s_i(Q_S)^2 \qquad (k = 1, 2, \ldots, n),$$

$$(b) \qquad \sum_{i=1}^{k}\left|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)\right| \le \sum_{i=1}^{k}\left|\lambda_i(Q_S^* + Q_S)\right| \qquad (k = 1, 2, \ldots, n),$$

Theorem 1.3 *The CGN algorithm for system (1.10) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \qquad (k=1,2,...,n), \tag{1.11}$$

where

$$\varepsilon_k := \frac{2}{km^2}\ \sum_{i=1}^{k}\Big(\big|\lambda_i(Q_S^*+Q_S)\big| + \lambda_i(Q_S^*Q_S)\Big)\ \to 0 \quad \textit{as}\ \ k\to\infty \tag{1.12}$$

and ε_k is a sequence independent of V_h.

Recall that a self-adjoint compact operator C is called a Hilbert-Schmidt operator if $|||C|||^2 \equiv \sum \lambda_i(C)^2 < \infty$ (see e.g. [16]). Then we obtain a more explicit rate $O(k^{-1/2})$:

Theorem 1.4 *If Q_S is a Hilbert-Schmidt operator, then the CG method yields*

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \left(\frac{3}{2k}\right)^{1/2} |||Q_S|||\,. \tag{1.13}$$

1.1.6 Nonsymmetric compact-equivalent preconditioners

Now let N be a nonsymmetric S-bounded and S-coercive operator which is compact-equivalent to L with $\mu = 1$, i.e., (1.7) becomes $L_S = N_S + Q_S$. We apply the stiffness matrix $\mathbf{N}_h$ of N_S as preconditioner for the discretized system (1.8). Since N is nonsymmetric, in order to define an inner product on $\mathbf{R}^n$ we endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle \mathbf{c}, \mathbf{d}\rangle_{\mathbf{S}_h} := \mathbf{S}_h\,\mathbf{c}\cdot\mathbf{d}$ as earlier. The preconditioned system $\mathbf{N}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h}$ takes the form

$$(\mathbf{I}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \hat{\mathbf{b}}_h \tag{1.14}$$

where $\mathbf{I}_h$ is the $n\times n$ identity matrix. With a proper estimation of sums of singular values, we obtain

Theorem 1.5 *Let L and N be compact-equivalent S-bounded and S-coercive operators, i.e. $L_S = N_S + Q_S$ with a compact operator Q_S on H_S. Let $s_i(Q_S)$ $(i=1,2,\dots)$ denote the singular values of Q_S. Then the CGN algorithm for system (1.14) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \qquad (k=1,2,...,n) \tag{1.15}$$

$$\textit{where}\quad \varepsilon_k = \frac{2M_N^2}{km_L^2}\ \sum_{i=1}^{k}\Big(\frac{2}{m_N}\,s_i(Q_S) + \frac{1}{m_N^2}\,s_i(Q_S)^2\Big)\ \to 0 \qquad (\textit{as}\ k\to\infty) \tag{1.16}$$

and ε_k is a sequence independent of V_h.

1.1.7 Mesh independent superlinear convergence for elliptic equations

In this section we consider nonsymmetric elliptic problems

$$\begin{cases} Lu := -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g \\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \end{cases} \tag{1.17}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, where $\frac{\partial u}{\partial \nu_A} = A\,\nu\cdot\nabla u$ denotes the weighted normal derivative. We assume that the operator L satisfies Assumptions 1.1.1, that is, L is of the type (1.5), and further, that $g \in L^2(\Omega)$.

Using FEM to solve (1.17), we define a subspace $V_h = span\{\varphi_1, \ldots, \varphi_n\} \subset H^1_D(\Omega)$ and seek the FEM solution $u_h \in V_h$, which requires solving an $n \times n$ system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h. \tag{1.18}$$

First we consider symmetric preconditioners. In order to obtain superlinear convergence based on Theorem 1.1, we define S to have the same principal part as L, i.e.,

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \sigma u \qquad \text{for } \ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{1.19}$$

assumed to satisfy Assumptions 1.1.2. We introduce the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.18), and then solve the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} \equiv (\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \tilde{\mathbf{g}_h} \tag{1.20}$$

(with $\tilde{\mathbf{g}_h} = \mathbf{S}_h^{-1}\mathbf{g}_h$) with a CG method. Let us decompose $L_S = I + Q_S$, where

$$\langle Q_S u, v\rangle_S = \int_\Omega \Big((\mathbf{b}\cdot\nabla u)v + (c-\sigma)uv\Big) \ + \int_{\Gamma_N} (\alpha-\beta)uv\,d\sigma \qquad (u, v \in H^1_D(\Omega)). \tag{1.21}$$

Theorem 1.6 *If Q_S is normal, then the GCG-LS algorithm for system (1.20) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k = 1, ..., n), \quad \textit{where} \quad \varepsilon_k := \frac{2}{km}\ \sum_{j=1}^{k}\big|\lambda_j(Q_S)\big| \ \to 0 \quad \textit{as} \ \ k\to\infty \tag{1.22}$$

and ε_k is a sequence independent of V_h.

Theorem 1.7 *The CGN algorithm for system (1.20) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \ \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{1.23}$$

where

$$\varepsilon_k := \frac{2}{km^2}\ \sum_{i=1}^{k}\Big(\big|\lambda_i(Q_S^* + Q_S)\big| + \lambda_i(Q_S^* Q_S)\Big) \ \to 0 \quad \textit{as} \ \ k\to\infty \tag{1.24}$$

and ε_k is a sequence independent of V_h.

Efficient solvers arise from symmetric preconditioners such as e.g. the symmetric part, Laplacian or Helmholtz operators [13, 32, 33], and in the general case one can use multigrid solvers for S [41].

The magnitude in which $\varepsilon_k \to 0$ can be determined in certain cases. The following estimates are available when the asymptotics for symmetric eigenvalue problems

$$Su = \mu u, \quad u_{|\Gamma_D} = 0, \quad r\left(\tfrac{\partial u}{\partial \nu_A} + \beta u\right)_{|\Gamma_N} = \mu u \tag{1.25}$$

are known, as is the case for Dirichlet problems where $\mu_i = O(i^{2/d})$. (A similar result in 2D will be seen later for symmetric part preconditioning for GCG-LS.)

Theorem 1.8 *The sequence ε_k in (1.24) satisfies $\varepsilon_k \le (4s/k)\sum_{i=1}^{k}(1/\mu_i)$ for some constants $s, r > 0$, where μ_i ($i \in \mathbf{N}^+$) are the solutions of (1.25). When the asymptotics $\mu_i = O(i^{2/d})$ holds, in particular, for Dirichlet boundary conditions,*

$$\varepsilon_k \le O\Big(\frac{\log k}{k}\Big) \quad \text{if} \quad d = 2 \quad \text{and} \quad \varepsilon_k \le O\Big(\frac{1}{k^{2/d}}\Big) \quad \text{if} \quad d \ge 3. \tag{1.26}$$

Nonsymmetric preconditioners may be needed if the original problem has large first-order terms, when the symmetric approach may not work satisfactorily and it may be advisable to include first-order terms in the preconditioning operator too. Let us consider the nonsymmetric elliptic equation (1.17) with Laplacian principal part. As before, we are interested in FEM discretization. Let us introduce the following type of nonsymmetric preconditioning operator:

$$Nu := -\Delta u + \mathbf{w}\cdot\nabla u + zu \qquad \text{for} \quad u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_K} + \eta u_{|\Gamma_N} = 0$$

for some properly chosen functions $\mathbf{w}, z, \eta$, such that N satisfies Assumptions 1.1.1 in the obvious sense.

Theorem 1.9 *The CGN algorithm for the preconditioned system $\mathbf{N}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h}$ yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \qquad (k = 1, 2, ..., n) \tag{1.27}$$

$$\text{where} \quad \varepsilon_k = \frac{2M_N^2}{km_L^2}\sum_{i=1}^{k}\Big(\frac{2}{m_N}\,s_i(Q_S) + \frac{1}{m_N^2}\,s_i(Q_S)^2\Big) \to 0 \qquad (\text{as } k \to \infty) \tag{1.28}$$

and ε_k is a sequence independent of V_h.

In general, the operator L has variable coefficients $\mathbf{b}$ and c, and one can well approximate it with a preconditioning operator with constant coefficients:

$$Nu = -\Delta u + \mathbf{w}\cdot\nabla u + zu \qquad \text{for} \quad u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \tag{1.29}$$

where $\mathbf{w} \in \mathbf{R}^d$, $z, \eta \ge 0$ are constants such that $z > 0$ or $\eta > 0$ if $\Gamma_D = \emptyset$. Then separable solvers are available for N, see [32, 33]. The preconditioning operator (1.29) can be further

simplified if one convection coefficient, say $b_1(x)$, is dominating. Then one can include only one nonsymmetric coefficient, i.e. propose the preconditioning operator

$$Nu = -\Delta u + w_1 \frac{\partial u}{\partial x_1} + zu \qquad \text{for } u \in H^2(\Omega): \ u_{|\Gamma_D} = 0, \ \frac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \tag{1.30}$$

where $w_1, z, \eta \in \mathbf{R}$ have the same properties as required for (1.29). The presence of the term $w_1 \frac{\partial u}{\partial x_1}$ itself may turn N into a much better approximation of L. Nevertheless, since this term is one-dimensional, the solution of the auxiliary problems remains considerably simpler than that of the original one, e.g. via local 1D Green's functions [8].

1.1.8 Mesh independent superlinear convergence for elliptic systems

We consider convection-diffusion type systems, coupled via the zeroth order terms. (Stokes type systems will be mentioned in subsection 1.4.6.) Here an important advantage of the equivalent operator idea is that one can define decoupled (that is, independent) operators for the preconditioner, thereby reducing the size of auxiliary systems to that of a single elliptic equation. The decoupled preconditioners allow efficient parallelization for the solution of the auxiliary systems.

We consider an elliptic system

$$\left.\begin{aligned} L_i u \equiv \ -\mathrm{div}\,(A_i \nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + \sum_{j=1}^{l} V_{ij} u_j = g_i \\ u_{i\,|\Gamma_D} = 0, \qquad \frac{\partial u_i}{\partial \nu_A} + \alpha_i u_{i\,|\Gamma_N} = 0 \end{aligned}\right\} \qquad (i = 1, \ldots, l) \tag{1.31}$$

where Ω, A_i and α_i are as in Assumptions 1.1.1, $\mathbf{b}_i \in W^{1,\infty}(\Omega)^d$, $g_i \in L^2(\Omega)$, $V_{ij} \in L^\infty(\Omega)$. We assume that $\mathbf{b}_i$ and the matrix $V = \left\{V_{ij}\right\}_{i,j=1}^{l}$ satisfy the coercivity property

$$\lambda_{min}(V + V^T) - \max_i \mathrm{div}\, \mathbf{b}_i \geq 0 \tag{1.32}$$

pointwise on Ω, where λ_{min} denotes the smallest eigenvalue; then system (1.31) has a unique weak solution $u \in H_D^1(\Omega)^l$. Such systems arise e.g. from suitable time discretization and Newton linearization of transport systems.

The preconditioning operator $S = (S_1, \ldots, S_l)$ will be the l-tuple of independent operators

$$S_i u_i := -\mathrm{div}\,(A_i \nabla u) + h_i u \qquad \text{for } u_{i\,|\Gamma_D} = 0, \ \frac{\partial u_i}{\partial \nu_A} + \beta_i u_{i\,|\Gamma_N} = 0 \qquad (i = 1, \ldots, l)$$

such that each S_i satisfies Assumptions 1.1.2. The preconditioner for the discrete system is defined as the stiffness matrix $\mathbf{S}_h$ of S in $H_D^1(\Omega)^l$, and we apply the CGN algorithm for the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$.

Theorem 1.10 *The CGN algorithm for the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \ \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{1.33}$$

$$\textit{where} \qquad \varepsilon_k := \frac{2}{km^2} \ \sum_{i=1}^{k} \Big(\big|\lambda_i(Q_S^* + Q_S)\big| + \lambda_i(Q_S^* Q_S) \Big) \ \to 0 \quad \textit{as} \ \ k \to \infty \tag{1.34}$$

and ε_k is a sequence independent of V_h.

If Q_S is normal, then one can apply the GCG-LS algorithm and obtain

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \quad (k=1,...,n), \qquad \text{where} \quad \varepsilon_k := \frac{2}{km}\sum_{j=1}^{k}\left|\lambda_j(Q_S)\right| \to 0 \quad \text{as} \quad k\to\infty$$

and ε_k is a sequence independent of V_h. As in the scalar case, our theory for GCG-LS only covers symmetric part preconditioners here (besides the practically uninteresting case of an original L with constant coefficients in L); however, the experiments in [25] show a wider validity of the mesh independent superlinear convergence result.

The proposed preconditioner has inherent parallelism, owing to the independence of the operators S_i that also implies a block diagonal form of the preconditioning matrices. Parallelization on a cluster of computers will be discussed in subsection 1.4.5. We finally note that these results can be obviously extended to uncoupled nonsymmetric preconditioners of the form (1.29).

1.2 Equivalent S-bounded and S-coercive operators and linear convergence

1.2.1 Mesh independent linear convergence in Hilbert space

Let us consider the operator equation (1.1), where L is S-bounded and S-coercive in the sense of Definition 1.1, and $g \in H$. Using a Galerkin discretization, we want to solve the arising $n \times n$ system (1.8). If no compact-equivalence is assumed then one can obtain general results on linear convergence from the S-bounded and S-coercive framework [11].

(a) Symmetric preconditioners. Let S be the symmetric coercive operator from Definition 1.1, and introduce the stiffness matrix of S as preconditioner for system (1.8). To solve the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h}, \tag{1.35}$$

one can apply a CG method using the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$.

Proposition 1.4 *If the operator L satisfies (1.3), then for any subspace $V_h \subset H_S$ the stiffness matrix $\mathbf{L}_h$ satisfies*

$$m\,(\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c}) \leq \mathbf{L}_h\,\mathbf{c}\cdot\mathbf{c}, \qquad |\mathbf{L}_h\,\mathbf{c}\cdot\mathbf{d}| \leq M\,\|\mathbf{c}\|_{\mathbf{S}_h}\|\mathbf{d}\|_{\mathbf{S}_h} \qquad (\mathbf{c},\mathbf{d}\in\mathbf{R}^n), \tag{1.36}$$

where m and M come from (1.3) and hence are independent of V_h.

Theorem 1.11 *Let the operator L satisfy (1.3). Then the GCG-LS method for for system (1.35) provides*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \left(1-\left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k=1,2,...,n) \tag{1.37}$$

and the CGN algorithm satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq 2^{1/k}\,\frac{M-m}{M+m} \qquad (k=1,2,...,n), \tag{1.38}$$

both independently of V_h.

We note that (1.37) holds as well for the GCR and Orthomin methods together with their truncated versions. We mention as a special case when L itself is a symmetric operator: then its S-coercivity and S-boundedness simply turns into a spectral equivalence relation, which immediately implies that $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \leq \frac{M}{m}$.

(b) Relation to previous conditions. Now we can clarify the relation of our setting to that by Faber, Manteuffel and Parter in [14]. Thereby they consider a more general situation than ours, similar to the Babuška lemma for well-posedness, which would mean with our terms that coercivity (the second inequality in (1.3)) can be replaced by the two weaker statements

$$\sup_{v\in H_S} \frac{\langle L_S u, v\rangle_S}{\|v\|_S} \geq m\|u\|_S \quad (u \in H_S), \qquad \sup_{u\in H_S} \langle L_S u, v\rangle_S > 0 \quad (v \in H_S). \tag{1.39}$$

However, in contrast to (1.3), the above inequalities are not automatically inherited in general subspaces V_h with the same constants, i.e., no analogue of Proposition 1.4 holds. Instead, the corresponding uniform relations for the discrete operators had to be assumed there, see (3.37)-(3.38) in [14]; with our notations, this means that one has to assume

$$\sup_{\mathbf{d}\in\mathbf{R}^n} \frac{\mathbf{L}_h\,\mathbf{c}\cdot\mathbf{d}}{\|\mathbf{d}\|_{\mathbf{S}_h}} \geq \tilde{m}\|\mathbf{c}\|_{\mathbf{S}_h} \quad (\mathbf{c} \in V_h), \qquad \sup_{\mathbf{c}\in\mathbf{R}^n} \mathbf{L}_h\,\mathbf{c}\cdot\mathbf{d} > 0 \quad (\mathbf{d} \in \mathbf{R}^n)$$

with a uniform constant $\tilde{m} > 0$ to obtain mesh independent linear convergence. (The first bound is an LBB type condition.) Although our assumptions (1.3) are more special, they hold for rather general elliptic operators as shown by Proposition 1.1, and provide mesh independent linear convergence for arbitrary subspaces $V_h \subset H_S$ without any further assumption.

(c) Nonsymmetric preconditioners. Let us consider a nonsymmetric preconditioning operator N for equation (1.1). We assume that N is S-bounded and S-coercive for the same symmetric operator S as is L. Then we introduce the stiffness matrix of N_S as preconditioner for the discretized system (1.8). To solve the preconditioned system

$$\mathbf{N}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h} \tag{1.40}$$

(with $\tilde{\mathbf{b}_h} = \mathbf{N}_h^{-1}\mathbf{b}_h$), we apply the CGN method under the $\mathbf{S}_h$-inner product $\langle .,.\rangle_{\mathbf{S}_h}$, which converges as follows (where $\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)$ is the condition number):

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq 2^{1/k}\,\frac{\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)-1}{\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)+1} \qquad (k = 1, 2, ..., n). \tag{1.41}$$

In the convergence analysis of nonsymmetric preconditioners, we must distinguish between the bounds of L and N, i.e., (1.3) is replaced by

$$\begin{aligned} &m_L\|u\|_S^2 \leq \langle L_S u, u\rangle_S, \quad |\langle L_S u, v\rangle_S| \leq M_L\|u\|_S\|v\|_S, \\ &m_N\|u\|_S^2 \leq \langle N_S u, u\rangle_S, \quad |\langle N_S u, v\rangle_S| \leq M_N\|u\|_S\|v\|_S \end{aligned} \tag{1.42}$$

for all $u, v \in H_S$.

Theorem 1.12 *If the operators L and N satisfy (1.42), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \le \frac{M_L M_N}{m_L m_N} \qquad \textit{and} \qquad \kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \le \Big(1 + \frac{m_L + m_N}{2m_L m_N}\,\|L_S - N_S\|\Big)^2 \tag{1.43}$$

independently of V_h.

Hence, by (1.41), the CGN algorithm converges with a ratio bounded independently of V_h.

1.2.2 Mesh independent linear convergence for elliptic problems

Let us consider again the nonsymmetric elliptic problem (1.17). Its FEM solution in an n-dimensional subspace $V_h \subset H_D^1(\Omega)$ requires solving the $n \times n$ system (1.18). As a preconditioning operator, we consider in general a symmetric elliptic operator S as in (1.6):

$$Su \equiv -\mathrm{div}\,(G\,\nabla u) + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0,\; \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0 \tag{1.44}$$

assumed to satisfy Assumptions 1.1.2, but in general $A \neq G$. We introduce the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.18), and then solve the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ with a CG algorithm. The basic conditioning estimate is as follows:

Proposition 1.5 *For any subspace $V_h \subset H_D^1(\Omega)$,*

$$\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \le M/m \tag{1.45}$$

independently of V_h, where

$$\begin{aligned} M &:= p_1 + C_{\Omega,S}\,q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C_{\Omega,S}^2\|c\|_{L^\infty(\Omega)} + C_{\Gamma_N,S}^2\|\alpha\|_{L^\infty(\Gamma_N)}\ , \\ m &:= \Big(p_0^{-1} + C_{\Omega,L}^2\|\sigma\|_{L^\infty(\Omega)} + C_{\Gamma_N,L}^2\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}. \end{aligned} \tag{1.46}$$

Theorem 1.13 *For the system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$, the GCG-LS algorithm satisfies*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \left(1 - \left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k = 1, 2, ..., n), \tag{1.47}$$

which holds as well for the GCR and Orthomin methods together with their truncated versions; further, the CGN algorithm satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le 2^{1/k}\ \frac{M-m}{M+m} \qquad (k = 1, 2, ..., n), \tag{1.48}$$

where both ratios are independent of V_h.

Efficient solvers arise for symmetric preconditioners such as e.g. Laplacian, Helmholtz, separable or piecewise constant coefficient operators or in general MG solvers [32, 33, 41]. The results can be extended to suitable systems, see as an example the Navier system (1.72).

1.3 Symmetric part preconditioning

Let us consider an algebraic system $\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h$ arising from a given elliptic FEM problem, and, as usual, we look for a preconditioner to provide a suitable preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$. A famous particular strategy is symmetric part preconditioning, introduced by Concus and Golub (see further analysis in [9]). Here

$$\mathbf{S}_h := \frac{1}{2}(\mathbf{L}_h + \mathbf{L}_h^T), \quad \mathbf{Q}_h := \frac{1}{2}(\mathbf{L}_h - \mathbf{L}_h^T), \tag{1.49}$$

that is, the symmetric and antisymmetric parts of $\mathbf{L}_h$, respectively. The main advantage is a simplified algorithm: the full GCG-LS algorithm then reduces to the truncated version GCG-LS(0) that uses a single, namely the current search direction [4].

We are interested in the mesh independent convergence of CG iterations. In order to apply the theory of the previous sections, we must identify the underlying operators. The elliptic problem is represented, as usual, by an operator equation $Lu = g$ for an unbounded linear operator L in H, where $g \in H$. On the other hand, we must find the operator S whose stiffness matrix is the symmetric part of $\mathbf{L}_h$, further, the operators L and S must fit in the framework developed in section 1.1. We assume for the discussion that H is complex and there exists $p > 0$ such that

$$\mathrm{Re}\langle Lu, u\rangle \geq p\|u\|^2 \quad (u \in D := D(L)). \tag{1.50}$$

1.3.1 Strong symmetric part and mesh independent convergence

Let us consider equation $Lu = g$ under the conditions $D(L) = D(L^*) =: D$, and let S and Q be the symmetric and antisymmetric parts of L:

$$Su = \frac{1}{2}(Lu + L^*u), \qquad Qu := \frac{1}{2}(Lu - L^*u) \qquad (u \in D). \tag{1.51}$$

Further, we impose the following conditions:

Assumptions 1.3.1. We have $R(S) = H$, and the operator Q can be extended to the energy space H_S, and then $S^{-1}Q$ is a bounded operator on H_S.

Theorem 1.14 *Let H be a complex Hilbert space. Let L satisfy (1.50) and $D(L) = D(L^*)$, further, assume that Assumptions 1.3.1 hold, and consider the GCG-LS(0) algorithm for the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$.*

(1) Then

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{\|S^{-1}Q\|}{\sqrt{1 + \|S^{-1}Q\|^2}} \qquad (k = 1, 2, ..., n). \tag{1.52}$$

(2) If, in addition, $S^{-1}Q$ is a compact operator on H_S, then

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \quad (k = 1, ..., n), \quad \textit{where} \quad \varepsilon_k := \frac{2}{k}\sum_{j=1}^{k}\left|\lambda_j(S^{-1}Q)\right| \to 0 \quad \textit{as} \quad k \to \infty \tag{1.53}$$

and ε_k is a sequence independent of V_h.

The above situation is applicable to Dirichlet problems as a special case of (1.17):

$$\begin{cases} Lu := -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g \\ u_{|\partial\Omega} = 0, \end{cases} \tag{1.54}$$

where we assume that L satisfies Assumptions 1.1.1. and the Kadlec conditions (i.e. Ω is C^2-diffeomorphic to a convex domain and $A \in Lip(\Omega, \mathbf{R}^{d\times d})$). Then an easy calculation shows that the symmetric part of L is the operator

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \hat{c}u \qquad \text{for } \; u_{|\partial\Omega} = 0, \tag{1.55}$$

where $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}$.

Theorem 1.15 *Let the operator L in (1.54) satisfy Assumptions 1.1.1. and the Kadlec conditions. Let S be the operator (1.55). Then the GCG-LS(0) algorithm for system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ converges superlinearly according to (1.53).*

The superlinear convergence rate can be determined for 2D problems with constant coefficients. Let us consider a special case of problem (1.54) on a bounded domain $\Omega \subset \mathbf{R}^2$, namely, $Lu = -\Delta u + \mathbf{b}\cdot\nabla u + cu$, where $\mathbf{b} = (b_1, b_2) \in \mathbf{R}^2$, $c \in \mathbf{R}^+$ and $g \in L^2(\Omega)$. The symmetric part becomes the preconditioning operator $Su = -\Delta u + cu$.

Theorem 1.16 *Then for any FEM subspace $V_h \subset H_0^1(\Omega)$, the GCG-LS algorithm for the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{C}{\sqrt{k}}$$

for some constant $C > 0$ independent of h and k.

1.3.2 Weak symmetric part and mesh independent convergence

Let us consider the operator equation $Lu = g$ again. If $D(L) \neq D(L^*)$, then the symmetric part operator S defined in (1.51) may have no real meaning. Therefore the symmetric part and its relation to L have to be handled in a more general weak sense using suitable sesquilinear (i.e. conjugate bilinear) forms [21]. This is the case when an elliptic problem has mixed boundary conditions.

(a) Construction of the weak symmetric part. We define the weak symmetric part as an inner product:

$$\langle u, v\rangle_S := \frac{1}{2}\Big(\langle Lu, v\rangle + \langle u, Lv\rangle\Big) \qquad (u, v \in D(L)). \tag{1.56}$$

Here assumption (1.50) implies the positivity of $\langle .,.\rangle_S$. The space H_S is the completion of $D(L)$ w.r.t. the inner product $\langle .,.\rangle_S$. Further, we can define the operator Q_S as follows:

Proposition 1.6 *Let the form $u, v \mapsto \langle Lu, v\rangle$ be bounded in H_S-norm. Then*
(1) there exists a unique bounded sesquilinear form on H_S satisfying

$$\langle u, v\rangle_L = \langle Lu, v\rangle \qquad (u, v \in D(L)); \tag{1.57}$$

(2) there exists a unique operator $Q_S : H_S \to H_S$, defined for given $u \in H_S$ by the expression

$$\langle Q_S u, v\rangle_S := \frac{1}{2}\Big(\langle u, v\rangle_L - \overline{\langle v, u\rangle_L}\Big) \qquad (\forall v \in H_S). \tag{1.58}$$

Further, we have

$$\langle u, v\rangle_L = \langle u, v\rangle_S + \langle Q_S u, v\rangle_S \qquad (u, v \in H_S). \tag{1.59}$$

(b) Preconditioning by the weak symmetric part. Now the weak form of $Lu = g$ is

$$\langle u, v\rangle_L = \langle g, v\rangle \qquad (\forall v \in H_S). \tag{1.60}$$

Using (1.59), if $f \in H_S$ is such that $\langle f, v\rangle_S \equiv \langle g, v\rangle$ $(\forall v \in H_S)$, then (1.60) becomes

$$(I + Q_S)u = f. \tag{1.61}$$

We now summarize our conditions.

Assumptions 1.17. L satisfies (1.50) and the form $u, v \mapsto \langle Lu, v\rangle$ is bounded in H_S-norm, further, the operator $Q_S : H_S \to H_S$, defined in (1.58), is compact on H_S.

Theorem 1.17 *Let Assumptions 1.17 hold. Then the the GCG-LS(0) algorithm applied for the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}}_h$ yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k = 1, ..., n) \quad \textit{where} \quad \varepsilon_k := \frac{2}{k}\sum_{j=1}^{k}\big|\lambda_j(Q_S)\big| \to 0 \quad \textit{as} \quad k \to \infty \tag{1.62}$$

and ε_k is a sequence independent of V_h.

(c) Symmetric part preconditioning for mixed boundary value problems. Let us consider again the nonsymmetric elliptic problem (1.17). Then one can calculate easily the weak symmetric part of L, which is the inner product generated by the preconditioning operator

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \hat{c}u \qquad \text{for} \quad u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_G} + \hat{\alpha}u_{|\Gamma_N} = 0, \tag{1.63}$$

where $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}$ and $\hat{\alpha} := \alpha + \frac{1}{2}\,(\mathbf{b}\cdot\nu)$. One can also calculate Q_S and prove that it is compact, and thus derive

Theorem 1.18 *Let the operator L in (1.17) satisfy Assumptions 1.1.1., and S be the operator (1.63). Then the GCG-LS(0) algorithm for the corresponding preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}}_h$ converges superlinearly according to (1.62).*

1.4 Applications to efficient computational algorithms

1.4.1 Helmholtz preconditioner for regular convection-diffusion equations

A regularly perturbed convection-diffusion process is described by the elliptic problem

$$\begin{cases} Lu \equiv -\Delta u + \mathbf{b} \cdot \nabla u + cu = g \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu}_{|\Gamma_N} = 0\,, \end{cases} \tag{1.64}$$

where L satisfies Assumptions 1.1.1. Using the Helmholtz preconditioning operator

$$Su \equiv -\Delta u + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0, \;\; \frac{\partial u}{\partial \nu}_{|\Gamma_N} = 0$$

where $\sigma > 0$ is a constant, Theorem 1.7 yields *mesh independent superlinear convergence* for the PCGN algorithm. The auxiliary Helmholtz problems can be solved by some *fast solver* such as multigrid or a parallel direct solver [13]. Numerical experiments confirm the theoretical convergence estimates [29].

1.4.2 Convection problems for viscous fluids

The study of the discrete steady-state of an incompressible viscous flow leads to the Oseen equations as a linearized form of the Navier-Stokes equations, where $\nu = O(1)$. The widespread Uzawa iteration for the Oseen equations defines the consecutive systems

$$\begin{cases} -\nu\Delta \mathbf{u}_k + \mathbf{w} \cdot \nabla \mathbf{u}_k + \nabla p_k = \mathbf{f}, \qquad \mathbf{u}_{k\,|\partial\Omega} = 0 \\ p_{k+1} = p_k + \alpha_k \operatorname{div} \mathbf{u}_k = 0 \end{cases} \tag{1.65}$$

(for $k = 1, 2, \ldots$, and where $\operatorname{div} \mathbf{w} = 0$), that is, one must stepwise solve uncoupled auxiliary problems for $\mathbf{u}_k$ which are special convection-diffusion type equations. Since their solution error accumulates during the outer Uzawa iteration, they require an accurate solution. The symmetric part preconditioning operator is

$$Sz \equiv -\nu\Delta z \qquad \text{for } \; z_{|\partial\Omega} = 0,$$

and Theorem 1.14 yields *mesh independent superlinear convergence* for the GCG-LS(0) iteration. The auxiliary Possion equations can be solved by a variety of *fast Possion solvers* [32, 33].

1.4.3 Scaling for problems with variable diffusion coefficients

If the diffusion is space-dependent, then the Laplacian is replaced by a variable coefficient diffusion operator. We are then led to the problem

$$Lu \equiv -\operatorname{div}(a\,\nabla u) + \mathbf{b} \cdot \nabla u + cu = g, \qquad u_{\,|\partial\Omega} = 0\,, \tag{1.66}$$

where L satisfies Assumptions 1.1.1 and we assume that $a \in C^2(\overline{\Omega})$, $a(x) \geq p > 0$.

If a fast Poisson or Helmholtz solver is available, then one can still achieve *mesh independent superlinear convergence* by applying the method of scaling. Using the new unknown function $v := a^{1/2}u$, (1.66) adopts the form

$$Nv \equiv -\Delta v + \hat{\mathbf{b}} \cdot \nabla v + \hat{c}v = \hat{g}\,, \qquad v_{\,|\partial\Omega} = 0\,, \tag{1.67}$$

hence the previously seen Poisson or Helmholtz preconditioners can be applied.

1.4.4 Decoupled preconditioners for linearized air pollution systems

Air pollution processes are described by compound nonlinear transport systems involving diffusion, convection, reaction and deposition terms [43], and such systems may consist of a huge number of equations. Properly using a standard time discretization and then a suitable linearization, one gets a linear elliptic system

$$\left.\begin{array}{l} -\mathrm{div}\,(K_i\,\nabla u_i) + \mathbf{w}_i \cdot \nabla u_i + \sum\limits_{j=1}^{l} V_{ij} u_j = g_i \\ u_{i\,|\partial\Omega} = 0, \end{array}\right\} \qquad (i = 1, \ldots, l) \tag{1.68}$$

which is a special case of system (1.31). To solve this system using FEM and PCG iteration, the equivalent operator idea can be employed very efficiently. One can define a decoupled (that is, independent) l-tuple of operators for the preconditioner, thereby reducing the size of auxiliary systems to that of a single elliptic equation:

$$S_i u_i := -\mathrm{div}\,(K_i\,\nabla u) + h_i u \qquad \text{for} \quad u_{i\,|\Gamma_D} = 0,\ \frac{\partial u_i}{\partial \nu_A} + \beta_i u_{i\,|\Gamma_N} = 0 \qquad (i = 1, \ldots, l) \tag{1.69}$$

such that each S_i satisfies Assumptions 1.1.2. Then Theorem 1.10 yields *mesh independent superlinear convergence*.

We have run numerical tests for a model problem based on [43], involving 10 equations. The mesh independent superlinear convergence was observed, and our solver was considerably faster compared to the direct solution with the original system matrix.

1.4.5 Parallelization on a cluster of computers

The proposed preconditioner in the previous subsection has inherent parallelism, hence the preconditioning step can be implemented without any communications between processors. Indeed, a considerable speed-up has been obtained in the tests in [26]. Here the GCG-LS iteration was used and mesh independent convergence was obtained again, but now the main interest was the parallelization. The tests were realized in the Institute for Parallel Processing of the Bulgarian Academy of Sciences, executed on a Linux cluster.

The obtained parallel time T_p on p processors, relative parallel speed-up $S_p = \frac{T_1}{T_p} \leq p$ and relative efficiency $E_p = \frac{S_p}{p} \leq 1$ were analyzed. Figure 1 shows the speed-up S_p of the full version of the algorithm obtained for $h^{-1} = 128$ and $l = 3, 4, \ldots 10$. As was expected, when the number of equations l is divisible by the number of processors p then the parallel efficiency of the parallel algorithm is higher.

1.4.6 Regularized flow and elasticity problems

(a) Viscous flow: the Stokes problem. A fundamental model of viscous flow is the system of Stokes equations

$$\left\{\begin{array}{l} -\Delta \mathbf{u} + \nabla p = \mathbf{f}, \qquad \mathbf{u}_{|\partial\Omega} = 0\,, \\ \mathrm{div}\,\mathbf{u} = 0 \end{array}\right. \tag{1.70}$$

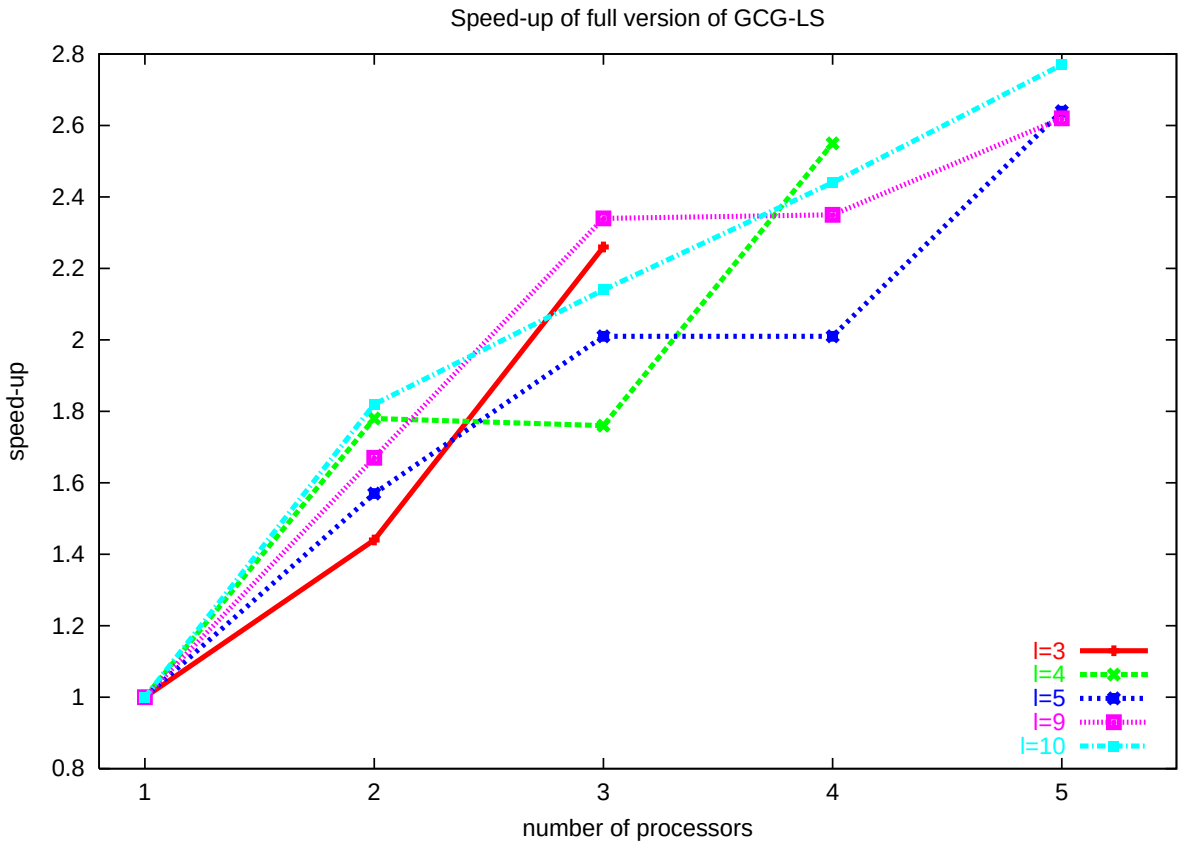

Figure 1: Speed-up of the GCG-LS algorithm for an elliptic system.

in a bounded domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3) with $\mathbf{f} \in L^2(\Omega)^d$. The numerical solution of this system has been widely investigated. Since the crucial LBB-condition restricts the suitable possible pairs of subspaces, an important effort has also been done to circumvent the LBB-condition via regularization. We consider a regularized version studied in [6]:

$$\mathbf{L}_h \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} \equiv \begin{pmatrix} diag_d(-\Delta_h^0) & \sigma^{-1/2}\nabla_h \\ \sigma^{-1/2}\,\mathrm{div}_h & -\Delta_h^\nu \end{pmatrix} \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \sigma^{-1/2}\,\mathrm{div}\,\mathbf{f}_h \end{pmatrix} \tag{1.71}$$

where σ is the regularization parameter. Then Theorem 1.14 yields *mesh independent superlinear convergence* using symmetric part preconditioning for GCG-LS(0) algorithm:

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \quad (k = 1, ..., n) \quad \text{where} \quad \varepsilon_k := \frac{2}{\sigma^{1/2}k} \sum_{j=1}^{k} \left|\lambda_j(Q_S)\right| \to 0 \quad \text{as} \quad k \to \infty$$

and ε_k depends only on σ but not on h.

(b) Linear elasticity: Navier's system of equations. Let us consider an isotropic elastic body Ω subject to a body force $\mathbf{f}$ in the case of pure displacement. A mixed formulation of the elasticity model leads to a form much similar to the Stokes problem:

$$\begin{cases} -\Delta\mathbf{u} + \nabla p = \frac{1}{\mu}\,\mathbf{f}, \qquad \mathbf{u}_{|\partial\Omega} = 0\,, \\ \\ \mathrm{div}\,\mathbf{u} + (1-2\nu)p = 0. \end{cases} \tag{1.72}$$

Now symmetric part preconditioning yields *mesh independent linear convergence* of the GCG-LS(0) algorithm for the preconditioned FEM system of the Navier equations. Using Theorem 1.14, one can derive

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{1}{\sqrt{2(1-\nu)}} \qquad (k=1,\ldots,n). \tag{1.73}$$

We note that one can regularize this system similarly to the Stokes problem, and obtain superlinear convergence for symmetric part preconditioning.

1.4.7 Nonsymmetric preconditioning for convection-dominated problems

Convection-dominated problems arise when the magnitude $|\mathbf{b}|$ of the convection coefficient is large, or equivalently, with a small coefficient ν of the Laplacian (singularly perturbed problem). There exist various approaches, mostly based on some stabilization, but here only linear convergence can be achieved. In contrast to this, our main interest is superlinear convergence. In estimating superlinear convergence, however, one cannot achieve independence of ν. Our numerical results give instead a milder deterioration of the convergence rate with increasing ν using a properly chosen preconditioning operator.

We consider the convection-dominated problem

$$Lu \equiv -\nu\Delta u + \mathbf{b}\cdot\nabla u = g, \qquad u_{|\partial\Omega} = 0\,, \tag{1.74}$$

where L satisfies Assumptions 1.1.1. A nonsymmetric preconditioning operator is chosen as

$$Nu := -\nu\,\Delta u + \mathbf{w}\cdot\nabla u \qquad \text{for } u_{|\partial\Omega} = 0\,,$$

where $\mathbf{w}$ is a constant vector function. Then systems with $\mathbf{N}_h$ are much cheaper to solve with some fast solver than systems with $\mathbf{L}_h$. We have chosen $|\mathbf{w}| = O(|\mathbf{b}|)$.

We have derived mesh independent superlinear convergence, and numerical experiments have shown a mild deterioration of the convergence rate with ν: in a range of $\nu = 1$ to $\nu = 0.05$, the number of iterations for $\|r_k\|_{\mathbf{S}_h} \leq 10^{-8}$ grew from 4 to 26.

2 Nonlinear problems

2.1 Sobolev gradients for variational problems

2.1.1 Gradient iterations in Hilbert space

We present iterative methods that model the situation to be discussed in subsection 2.1.2 on Sobolev gradients. This relates to preconditioning via the spectral notion of condition number, which can be extended in a natural way from symmetric and positive definite matrices to nonlinear operators. The condition number is infinite for differential operators in strong form, which explains the phenomenon that $cond(T_h)$ is unbounded as $h \to 0$ from proper discretizations of T. The first theorem provides preconditioning of a nonlinear operator T by a linear operator S such that $cond(S^{-1}T) \leq \frac{M}{m}$. It extends a classical result of Dyakonov, involves a weak form of an unbounded nonlinear operator in a similar

manner as we did in the linear case, see (1.2), and will connect it to the Sobolev gradient context, see (2.7). The iteration in Hilbert space mainly serves as a background to construct iterations in finite dimensional subspaces as suitable projections of the theoretical sequence in a straightforward manner. We note, however, that one can use the theoretical iteration itself in a few cases such that a sequence is constructed in the corresponding function space via Fourier or spectral type methods.

Definition 2.1 The nonlinear operator $F : H \to H$ has a *bihemicontinuous symmetric Gateaux derivative* if F is Gateaux differentiable, F' is bihemicontinuous, and for any $u \in H$ the operator $F'(u)$ is self-adjoint. (If these hold then F is a potential operator.)

Theorem 2.1 *Let H be a real Hilbert space, $D \subset H$ a dense subspace, $T : D \to H$ a nonlinear operator. Assume that $S : D \to H$ is a symmetric linear operator with lower bound $p > 0$, such that there exist constants $M \geq m > 0$ satisfying*

$$m\langle S(v-u), v-u\rangle \leq \langle T(v) - T(u), v-u\rangle \leq M\langle S(v-u), v-u\rangle \qquad (u, v \in D). \quad (2.1)$$

Then the identity

$$\langle F(u), v\rangle_S = \langle T(u), v\rangle \qquad (u, v \in D) \quad (2.2)$$

defines an operator $F : D \to H_S$. Further, if F can be extended to H_S such that it has a bihemicontinuous symmetric Gateaux derivative, then

(1) for any $g \in H$ the equation $T(u) = g$ has a unique weak solution $u^* \in H_S$, i.e.

$$\langle F(u^*), v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \quad (2.3)$$

(2) For any $u_0 \in H_S$ the sequence

$$\begin{aligned} u_{n+1} &= u_n - \tfrac{2}{M+m} z_n\,, \\ \text{where } \ \langle z_n, v\rangle_S &= \langle F(u_n), v\rangle_S - \langle g, v\rangle \quad (v \in H_S), \end{aligned} \quad (2.4)$$

converges linearly to u^*, namely,

$$\|u_n - u^*\|_S \leq \frac{1}{m}\|F(u_0) - b\|_S \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N})\,, \quad (2.5)$$

where $\langle b, v\rangle_S = \langle g, v\rangle$ $(v \in H_S)$.

(3) Under the additional condition $R(S) \supset R(T)$, if $g \in R(S)$ and $u_0 \in D$, then for any $n \in \mathbf{N}$ the element z_n in (2.4) can be expressed as $z_n = S^{-1}(T(u_n) - g)$, that is, the auxiliary problem becomes $Sz_n = T(u_n) - g$.

Now we can formulate the discrete counterpart of the above theorem. Let the conditions of Theorem 2.1 hold, let $g \in H$ and let $V_h \subset H_S$ be a given subspace. Then there exists a unique solution $u_h \in V_h$ to the problem

$$\langle F(u_h), v\rangle_S = \langle g, v\rangle \qquad (v \in V_h), \quad (2.6)$$

and the same convergence result holds:

Theorem 2.2 *For any $u_0 \in V_h$ the sequence $(u_n) \subset V_h$, defined by replacing all $v \in H_S$ in (2.4) by all $v \in V_h$, converges to u_h according to the same estimate (2.5), i.e. with a rate independent of V_h.*

More generally, it readily follows that if the constant M in assumption (2.1) is replaced by $M\big(\max\{\|u\|_S, \|v\|_S\}\big)$ for some increasing function $M : \mathbf{R}^+ \to \mathbf{R}^+$, then Theorem 2.2 holds in a modified form such that the constant M is replaced by M_0 depending on u_0:

$$M_0 := M\Big(\|u_0\| + \frac{1}{m}\|F(u_0) - b\|\Big).$$

2.1.2 Sobolev gradients and preconditioning

Theorem 2.1 relates to Sobolev gradients developed by J.W. Neuberger. Let $cond(T) = \infty$. The operator $F : H_S \to H_S$ in (2.2) has a potential $\phi_S : H \to \mathbf{R}$, then ϕ'_S denotes the gradient of ϕ w.r. to the inner product $\langle ., . \rangle_S$. On the other hand, for $\phi_{|D}$ as a functional in H w.r. to the original inner product $\langle ., . \rangle$, the gradient is denoted by ϕ'. Then

$$\phi'_S(u) = F(u) \qquad (u \in H_S) \quad \text{and} \qquad \phi'(u) = T(u) \qquad (u \in D). \tag{2.7}$$

The steepest descent iteration corresponding to the gradient ϕ'_S is the preconditioned sequence in (2.4), whereas using the gradient ϕ' one would have a steepest descent iteration $u_{n+1} = u_n - \tilde{\alpha}(T(u_n) - g)$ whose convergence could not be ensured.

Altogether, the change of the inner product yields the change of the gradient of ϕ, namely as a formally preconditioned version of the original one. For elliptic problems, the space H_S is a Sobolev space corresponding to the given problem, and the above gradient ϕ'_S plays the role of the Sobolev gradient. Whereas the latter was applied by Neuberger mostly to least-square minimization, our problems below will be variational.

2.1.3 Dirichlet problems for second order equations

First we illustrate the method on a very simple problem

$$\begin{cases} T(u) \equiv -\mathrm{div}\, f(x, \nabla u) = g(x) \\ u_{|\partial\Omega} = 0 \end{cases} \tag{2.8}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, such that the following assumptions are satisfied:

Assumptions 2.3.

(i) The function $f \in C^1(\Omega \times \mathbf{R}^d,\ \mathbf{R}^d)$ has bounded derivatives w.r.t. all x_i, further, its Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$ w.r.t. η are symmetric and their eigenvalues λ satisfy

$$0 < \mu_1 \le \lambda \le \mu_2$$

with constants $\mu_2 \ge \mu_1 > 0$ independent of (x, η).

(ii) $g \in L^2(\Omega)$.

We look for the FEM solution $u_h \in V_h$ in a given FEM subspace $V_h \subset H_0^1(\Omega)$. (For standard FEM subspaces, u_h is well-known to converge to the unique weak solution as $h \to 0$.)

Let $G \in C^1(\overline{\Omega}, \mathbf{R}^{d\times d})$ be a symmetric matrix-valued function for which there exist constants $M \geq m > 0$ such that

$$m\ G(x)\xi \cdot \xi \leq \frac{\partial f(x,\eta)}{\partial \eta}\,\xi \cdot \xi\ \leq M\ G(x)\xi \cdot \xi \qquad ((x,\eta) \in \Omega \times \mathbf{R}^d, \xi \in \mathbf{R}^d). \tag{2.9}$$

We introduce the linear preconditioning operator

$$Su \equiv\ -\mathrm{div}\,(G(x)\nabla u) \qquad \text{for}\ \ u_{|\partial\Omega} = 0. \tag{2.10}$$

The corresponding energy space is $H_0^1(\Omega)$ with the weighted inner product

$$\langle u, v\rangle_G := \int_\Omega G(x)\,\nabla u \cdot \nabla v. \tag{2.11}$$

Theorem 2.3 *Let Assumptions 2.3 hold. Then for any $u_0 \in V_h$ the sequence*

$$\begin{gathered} u_{n+1} := u_n - \tfrac{2}{M+m} z_n\ \in V_h \\ \textit{where}\ \ \int_\Omega G(x)\,\nabla z_n \cdot \nabla v = \int_\Omega f(x, \nabla u_n)\cdot \nabla v - \int_\Omega gv \qquad (v \in V_h), \end{gathered} \tag{2.12}$$

converges linearly to u_h according to

$$\|u_n - u_h\|_G \leq \frac{1}{m}\|F(u_0) - b\|_G \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N}), \tag{2.13}$$

where F and b are the weak forms of T and g.

The sequence (2.12) requires the stepwise FEM solution of problems of the type

$$Sz \equiv -\mathrm{div}\,(G(x)\nabla z) = r, \qquad z_{|\partial\Omega} = 0,$$

in V_h, where $r = T(u_n) - g$ is the current residual. Various examples of efficient choices for the preconditioning operator S will be given in subsection 2.1.5.

The method can be extended to similar but more general equations, such as mixed boundary value problems or fourth order equations.

2.1.4 Second order symmetric systems

Now we consider symmetric nonlinear elliptic systems on a bounded domain in the form

$$\left.\begin{aligned} -\mathrm{div}\ f_i(x, \nabla u_i) + q_i(x, u_1, \ldots, u_l) &= g_i \\ u_{i\,|\Gamma_D} = 0, \qquad f_i(x, \nabla u_i)\cdot \nu + \alpha_i u_{i\,|\Gamma_N} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, l). \tag{2.14}$$

Assumptions 2.4.

(i) (Domain:) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(ii) (Smoothness:) The functions $f_i : \Omega \times \mathbf{R}^d \to \mathbf{R}^d$ and $q = (q_1, \dots, q_l) : \Omega \times \mathbf{R}^l \to \mathbf{R}^l$ are measurable and bounded w.r. to the variable $x \in \Omega$ and C^1 in their second variables $\xi \in \mathbf{R}^l$ resp. $\eta \in \mathbf{R}^d$; further, $\alpha_i \in L^\infty(\Gamma_N)$ and $g_i \in L^2(\Omega)$ $(i = 1, \dots, l)$.

(iii) (Coercivity:) for all $i = 1, \dots, l$, the Jacobians $\frac{\partial f_i(x,\eta)}{\partial \eta}$ are symmetric and their eigenvalues λ satisfy $0 < \mu_1 \le \lambda \le \mu_2$ with constants $\mu_2, \mu_1 > 0$ independent of x, η and i. Further, the Jacobians $\frac{\partial q(x,\xi)}{\partial \xi}$ are symmetric and positive semidefinite for any $(x, \xi) \in \Omega \times \mathbf{R}^l$ and $\eta \in \mathbf{R}^l$. Finally, $\alpha_i \ge 0$ $(i = 1, \dots, l)$, and either $\Gamma_D \ne \emptyset$ or $\inf_{i,\Omega} \alpha > 0$.

(iv) (Growth:) let $p \ge 2$ (if $d = 2$) or $p \le \frac{2d}{d-2}$ (if $d \ge 3$), then there exist constants $c_1, c_2 \ge 0$ such that for any $(x, \xi) \in \Omega \times \mathbf{R}^l$, $\big\|q'_\xi(x, \xi)\big\| \le c_1 + c_2|\xi|^{p-2}$.

The coercivity and growth assumptions imply that problem (2.14) has a unique weak solution in the product Sobolev space $H_0^1(\Omega)^l$. Let $V_h \subset H_0^1(\Omega)$ be a given FEM subspace. We look for the FEM solution $u_h = (u_{h,1}, .., u_{h,l})$ in V_h^l.

Let $G_i \in L^\infty(\Omega, \mathbf{R}^{d\times d})$ be symmetric matrix-valued functions $(i = 1, \dots, l)$ for which there exist constants $m' \ge m > 0$ such that each G_i satisfies (2.9) with M replaced by m'. We introduce a linear preconditioning operator $S = (S_1, \dots, S_l)$ as an independent l-tuple of operators

$$S_i u_i \equiv -\mathrm{div}\,(G_i(x)\nabla u_i) \qquad \text{for } \; u_{i\ |\partial\Omega} = 0, \qquad \frac{\partial u_i}{\partial \nu_{G_i}}_{\ |\Gamma_N} = 0.$$

The corresponding energy space is $H_D^1(\Omega)^l$ with a G-inner product which now denotes the sum of the G_i-inner products. We introduce the real function

$$M(r) := m' + c_1\varrho^{-1} + d_1 K_{2,\Gamma_N}^2 + c_2 K_{p,\Omega}^{p_1} r^{p-2} \quad (r > 0), \tag{2.15}$$

where $d_1 := \max_i \|\alpha_i\|_{L^\infty}$ and $K_{p,\Omega}$, K_{2,Γ_N} are the Sobolev embedding constants, further, $\varrho > 0$ denotes the smallest eigenvalue of the operators S_i.

Theorem 2.4 *Let Assumptions 2.4 be satisfied. Let $u_0 \in V_h^l$ and*

$$M_0 := M\Big(\|u_0\|_{H_D^1(\Omega)} + \frac{1}{m}\|F(u_0) - b\|_{H_D^1(\Omega)}\Big), \tag{2.16}$$

where $M(r)$ is from (2.15) and F and b are the weak forms of $T = (T_1, \dots, T_l)$ and $g = (g_1, \dots, g_l)$. Let the sequence $(u_n) = (u_{n,1}, .., u_{n,l}) \subset V_h^l$ be defined as follows: for $n \in \mathbf{N}$ let

$$u_{n+1} = u_n - \frac{2}{M_0 + m} z_n \ , \tag{2.17}$$

where $z_n = (z_{n,1}, .., z_{n,l}) \in V_h^l$ and its coordinates satisfy

$$\int_\Omega G_i(x)\,\nabla z_{n,i} \cdot \nabla v_i = \int_\Omega \Big(f_i(x, \nabla u_{n,i}) \cdot \nabla v_i + q_i(u_{n,1}, .., u_{n,l})v_i\Big) + \int_{\Gamma_N} \alpha_i u_{n,i} v_i - \int_\Omega g_i v_i \tag{2.18}$$

$(v = (v_1, \dots, v_l) \in V_h^l)$. Then (u_n) converges linearly to u_h according to

$$\|u_n - u_h\|_G \le \frac{1}{m}\|F(u_0) - b\|_G \left(\frac{M_0 - m}{M_0 + m}\right)^n \qquad (n \in \mathbf{N})\,. \tag{2.19}$$

The sequence (u_n) requires the stepwise FEM solution of independent linear elliptic equations of the type

$$\begin{cases} S_i z_i \equiv -\mathrm{div}\,(G_i(x)\nabla z_i) = r_i \\ z_{i\,|\Gamma_D} = 0, \qquad \frac{\partial z_i}{\partial \nu_{G_i}}{}_{|\Gamma_N} = \varrho_i \end{cases} \qquad (i = 1, \ldots, l) \tag{2.20}$$

in V_h, where $r_i = T(u_{n,i}) - g_i$ and $\varrho_i = f_i(x, \nabla u_{n,i}) \cdot \nu + \alpha_i u_{n,i}$ are the current interior and boundary residuals. Thus the proposed preconditioning operator to the original system involves a cost proportional to a single equation when solving these auxiliary equations.

2.1.5 Some examples of preconditioning operators

Discrete Laplacian preconditioner. The most straightforward preconditioning operator for problem (2.8) is the minus Laplacian (i.e. with coefficient matrix $G(x) \equiv I$):

$$S = -\Delta, \qquad \text{satisfying} \quad M = \mu_2, \quad m = \mu_1$$

for the constants in (2.9) independently of V_h. The solution of the linear auxiliary systems containing the discrete Laplacian preconditioner can rely on fast Poisson solvers [32, 33].

Separable preconditioners. Let us assume that the Jacobians of f are uniformly diagonal dominant, i.e. that introducing the functions

$$\delta_i^{\pm}(x,\eta) := \tfrac{\partial f_i(x,\eta)}{\partial \eta_i} \pm \sum_{\substack{j=1 \\ j\neq i}}^{d} \left| \tfrac{\partial f_i(x,\eta)}{\partial \eta_j} \right|, \quad \text{we have} \qquad \delta_i^{-}(x,\eta) \geq \mu_1 > 0 \tag{2.21}$$

(for all $x \in \Omega,\ \eta \in \mathbf{R}^d,\ i = 1, ..., d$) for some constant μ_1 independent of x, η and i. Now, for any $x \in \Omega$ and $1 \leq s \leq d$, let $\Omega_s = \{z \in \Omega :\ z_s = x_s\}$ and

$$a_s(x_s) = \inf_{\substack{x\in\Omega_s \\ \eta\in\mathbf{R}^d}} \delta_i^{-}(x,\eta), \qquad b_s(x_s) = \sup_{\substack{x\in\Omega_s \\ \eta\in\mathbf{R}^d}} \delta_i^{+}(x,\eta).$$

Then one can propose the separable preconditioning operator

$$Su := -\sum_{s=1}^{d} \tfrac{\partial}{\partial x_s}\left(a_s(x_s)\tfrac{\partial u}{\partial x_s}\right) \quad \text{satisfying} \quad M = \sup_{x\in\Omega} \max_{s=1,..,d} b_s(x_s), \quad m = \inf_{x\in\Omega} \min_{s=1,..,d} a_s(x_s)$$

for the constants in (2.9) independently of V_h: i.e. the bounds from the Laplacian are thus improved. The solution of the auxiliary problems relies on fast separable solvers [32, 33].

Modified Newton preconditioner. The popular modified Newton method involves a preconditioning operator arising from the initial derivative of the differential operator:

$$Sz = -\mathrm{div}\left(\frac{\partial f}{\partial \eta}(x, \nabla u_0)\,\nabla z\right), \quad \text{satisfying} \quad \frac{M}{m} \leq \left(\frac{1 + \tilde{\gamma}\|F(u_0) - b\|_{H_0^1}}{1 - \tilde{\gamma}\|F(u_0) - b\|_{H_0^1}}\right)^2$$

under our conditions, assuming the Lipschitz continuity of F' and a small enough initial residual, and with $\tilde{\gamma} = L\mu_1^{-3}\mu_2$ where L is the Lipschitz constant of F'.

Some other natural choices of preconditioning operators are e.g. the biharmonic operator for fourth order equations and independent Laplacians for second order systems.

2.2 Variable preconditioning

2.2.1 Variable preconditioning via quasi-Newton methods in Hilbert space

We give two theorems on general iterations that include the gradient and Newton methods as special cases [22]. Namely, the choice $B_n = I$ below in (2.22) reproduces the gradient method and its well-known linear convergence rate, whereas $B_n := F'(u_n)$ can reproduce Newton's method and shows (since M and m can be arbitrarily close) that convergence is faster that any linear rate. More subtle estimates will be given in Theorem 2.6.

Theorem 2.5 *Let H be a real Hilbert space. Assume that the nonlinear operator $F : H \to H$ has a symmetric Gateaux derivative satisfying the following properties:*

(i) (Ellipticity.) There exist constants $\Lambda \geq \lambda > 0$ satisfying

$$\lambda\|h\|^2 \leq \langle F'(u)h, h\rangle \leq \Lambda\|h\|^2 \qquad (u, h \in H).$$

(ii) (Lipschitz continuity.) There exists $L > 0$ such that

$$\|F'(u) - F'(v)\| \leq L\|u - v\| \qquad (u, v \in H).$$

Let $b \in H$ and denote by u^* the unique solution of equation $F(u) = b$. We fix constants $M > m > 0$. Then there exists a neighbourhood $\mathcal{V}$ of u^* such that for any $u_0 \in \mathcal{V}$, the sequence

$$u_{n+1} = u_n - \frac{2}{M+m} B_n^{-1}(F(u_n) - b) \qquad (n \in \mathbf{N}), \tag{2.22}$$

with properly chosen self-adjoint linear operators B_n satisfying

$$m\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\ h \in H), \tag{2.23}$$

converges linearly to u^*. Namely,

$$\|u_n - u^*\| \leq C \cdot \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N}) \tag{2.24}$$

with some constant $C > 0$.

Now we turn to the more general version of Theorem 2.5. We will use the following norms:

$$\|h\|_n = \langle F'(u_n)^{-1}h, h\rangle^{1/2} \quad (n \in \mathbf{N}), \qquad \|h\|_* = \langle F'(u^*)^{-1}h, h\rangle^{1/2}. \tag{2.25}$$

Using damped iteration and variable spectral bound preconditioning, the theorem gives a variant of quasi-Newton method that provides global convergence up to second order.

Theorem 2.6 *Let H be a real Hilbert space. Let the operator $F : H \to H$ have a symmetric Gateaux derivative satisfying the properties (i)-(ii) of Theorem 2.5.*

Denote by u^* the unique solution of equation $F(u) = b$. For arbitrary $u_0 \in H$ let (u_n) be the sequence defined by

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} B_n^{-1}(F(u_n) - b) \qquad (n \in \mathbf{N}), \tag{2.26}$$

where the following conditions hold:

(iii) $M_n \geq m_n > 0$ *and the properly chosen self-adjoint linear operators* B_n *satisfy*

$$m_n\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M_n\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\ h \in H), \tag{2.27}$$

further, using notation $\omega(u_n) = L\lambda^{-2}\|F(u_n) - b\|$, *there exist constants* $K > 1$ *and* $\varepsilon > 0$ *such that* $M_n/m_n \leq 1 + 2/(\varepsilon + K\omega(u_n))$;

(iv) we define $\tau_n = \min\{1, \frac{1-Q_n}{2\rho_n}\}$, *where* $Q_n = \frac{M_n - m_n}{M_n + m_n}(1+\omega(u_n))$, $\rho_n = 2LM_n^2\lambda^{-3/2}(M_n + m_n)^{-2}\|F(u_n) - b\|_n(1+\omega(u_n))^{1/2}$, $\omega(u_n)$ *is as in condition (iii) and* $\|\,.\,\|_n$ *is defined in (2.25). (This value of* τ_n *ensures optimal contractivity in the* n*-th step in the* $\|\,.\,\|_*$*-norm.)*

Then $\|u_n - u^*\| \leq \lambda^{-1}\|F(u_n) - b\| \to 0$, *namely,*

$$\limsup \frac{\|F(u_{n+1}) - b\|_*}{\|F(u_n) - b\|_*} \leq \limsup \frac{M_n - m_n}{M_n + m_n} < 1\,. \tag{2.28}$$

Moreover, if in addition we assume $M_n/m_n \leq 1 + c_1\|F(u_n) - b\|^\gamma$ $(n \in \mathbf{N})$ *with some constants* $c_1 > 0$ *and* $0 < \gamma \leq 1$, *then*

$$\|F(u_{n+1}) - b\|_* \leq d_1\|F(u_n) - b\|_*^{1+\gamma} \qquad (n \in \mathbf{N}) \tag{2.29}$$

with some constant $d_1 > 0$.

Owing to the equivalence of the norms $\|\,.\,\|$ and $\|\,.\,\|_*$, the orders of convergence corresponding to the estimate (2.29) can be formulated with the original norm:

Corollary 2.1 (Rate of convergence in the original norm.) *Let the variable bounds satisfy* $M_n/m_n \leq 1 + c_1\|F(u_n - b)\|^\gamma$ *with some constants* $c_1 > 0$, $0 < \gamma \leq 1$. *Then*

$$\|F(u_{n+1}) - b\| \leq d_1\|F(u_n) - b\|^{1+\gamma} \qquad (n \in \mathbf{N})$$

for some $d_1 > 0$, *and consequently*

$$\|u_n - u^*\| \leq \lambda^{-1}\|F(u_n) - b\| \leq const. \cdot \rho^{(1+\gamma)^n}$$

with some constant $0 < \rho < 1$.

Remark 2.1 Theorem 2.6 can be generalized by only assuming Hölder continuity instead of Lipschitz: $\|F'(u) - F'(v)\| \leq L\|u - v\|^\alpha$ $(u, v \in H)$ with some constants $L > 0$, $0 < \alpha < 1$ independent of u, v. Then the same results hold with $0 < \gamma \leq 1$ replaced by $0 < \gamma \leq \alpha$ for (2.29), i.e. the fastest feasible convergence is of order $1 + \alpha$.

Remark 2.2 The above results can be considered in the context of Sobolev gradients, similarly to (2.7). Now, using a variable preconditioning operator, one obtains the variable Sobolev gradient

$$\phi'_{B_n}(u) = B_n^{-1}F(u) \qquad (u \in H).$$

2.2.2 Variable preconditioning for elliptic problems

2.2.3 Problems with nonlinear principal part

Let us consider problem (2.8) again:

$$\begin{cases} T(u) \equiv -\mathrm{div}\, f(x, \nabla u) = g(x) \\ u_{|\partial\Omega} = 0. \end{cases} \tag{2.30}$$

Assumptions 2.7. Assumptions 2.3 imposed for (2.8) are satisfied and, in addition, the Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$ are Lipschitz continuous w.r.t η.

A general iteration with variable preconditioning. Let $V_h \subset H_0^1(\Omega)$ be a given FEM subspace; we look for the FEM solution u_h again. First we derive convergence when general preconditioning operators are used. Some efficient particular choices will be given afterwards. The main idea is that the preconditioning operator (2.10) is modified with stepwise redefined diffusion coefficient matrices.

Theorem 2.7 *Let $u_0 \in V_h$ be arbitrary, and let $(u_n) \subset V_h$ be the sequence defined as follows. If, for $n \in \mathbf{N}$, u_n is obtained, then we choose constants $M_n \geq m_n > 0$ and a symmetric matrix-valued function $G_n \in L^\infty(\Omega, \mathbf{R}^{N\times N})$ for which there holds*

$$m_n\, G_n(x)\xi \cdot \xi \leq \frac{\partial f}{\partial \eta}(x, \nabla u_n(x))\, \xi \cdot \xi \leq M_n\, G_n(x)\xi \cdot \xi \qquad (x \in \Omega,\, \xi \in \mathbf{R}^N), \tag{2.31}$$

further, M_n/m_n and τ_n satisfy the conditions (iii)-(iv) in Theorem 2.6. We define

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n}\, z_n\,, \tag{2.32}$$

where $z_n \in V_h$ is the solution of the problem

$$\int_\Omega G_n(x)\, \nabla z_n \cdot \nabla v = \int_\Omega \big(f(x, \nabla u_n) \cdot \nabla v \; - gv\big) \qquad (v \in V_h). \tag{2.33}$$

Then u_n converges to u_h according to the estimates of Theorem 2.6.

Piecewise constant coefficient operators. An efficient choice for variable preconditioners is obtained if the Jacobians are replaced by the discretizations of piecewise constant coefficient preconditioning operators, motivated by the case of nearly singular Jacobians. Formally we write

$$S_n u := -\mathrm{div}\,\big(w_n(x)\nabla u\big), \quad \text{where} \qquad w_{n\,|\Omega_i} \equiv c_i \; > 0 \tag{2.34}$$

on proper subdomains Ω_i $(i = 1, \ldots, s)$. Let us now introduce the spectral bounds m_i and M_i of $J_n := \partial_\eta f(., \nabla u_n)$ relative to Ω_i. If c_i is some (arithmetic, geometric or harmonic) mean of m_i and M_i, then we obtain the improved bounds $M_n/m_n = \max_i M_i/m_i$. The numerical performance of such preconditioners will be illustrated in subsection 2.5.1.

General scalar coefficient preconditioning operators. One can more generally define any operator S with a scalar diffusion coefficient, i.e. w_n is replaced by $k_n \in L^\infty(\Omega)$ such that $k_n \geq k_0 > 0$. Then the discretized operator still has a better sparsity pattern. A useful choice for $k_n(x)$ is the diagonal of $J_n(x)$.

2.2.4 Variable preconditioning for semilinear problems

Let us consider a semilinear equation with mixed boundary conditions

$$\begin{cases} -\mathrm{div}\,(k(x)\,\nabla u) + q(x,u) = g(x) \\ u_{|\Gamma_D} = 0, \qquad k(x)\,\frac{\partial u}{\partial \nu} + \alpha u_{|\Gamma_N} = 0 \end{cases} \tag{2.35}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3) under the following assumptions:

Assumptions 2.2.

(i) the domain Ω and the functions q, α and g satisfy the corresponding parts of Assumptions 2.4, further, $k \in L^\infty(\Omega)$ and $k \geq k_0 > 0$.

(ii) (Lipschitz condition) There exists $3 \leq p$ (if $d = 2$) or $3 \leq p \leq 6$ (if $d = 3$), and there exist constants $c_1, c_2 \geq 0$ such that for any (x, ξ_1) and $(x, \xi_2) \in \Omega \times \mathbf{R}$,

$$\left\| q'_\xi(x,\xi_1) - q'_\xi(x,\xi_2) \right\| \leq \Big(c_1 + c_2\,(\max\,|\xi_1|, |\xi_2|)^{p-3} \Big) |\xi_1 - \xi_2|.$$

We introduce the variable preconditioning operator

$$S_n v \equiv -\kappa\,\Delta v + c_n v \qquad \text{for} \quad v_{|\Gamma_D} = 0, \qquad \tfrac{\partial v}{\partial \nu}_{|\Gamma_N} = 0$$

with given constants $\kappa > 0$ and $c_n > 0$. If C_Ω denotes the Poincaré-Friedrichs constant then one can derive (2.27) with $M_n := 1 + (C_\Omega/k_0) \max q'_\xi(x, u_n)$, $m_n := 1/\big(1 + (C_\Omega/k_0)\,c_n\big)$.

Corollary 2.2 *If M_n/m_n and τ_n satisfy the conditions (iii)-(iv) in Theorem 2.6, then u_n converges to u_h according to the estimates of Theorem 2.6.*

Since S_n has constant coefficients, its updating is much faster than for $F'(u_n)$, and fast solvers are available for the auxiliary problems. The inclusion of the variable coefficient c_n allows to follow the variation of the magnitude of the lower order term during the iteration.

2.3 Newton's method and operator preconditioning

2.3.1 Newton's method as optimal variable gradients

In this subsection we study the relation of the gradient and Newton's method. The usual gradient method defines an optimal descent direction when a fixed inner product is used. In contrast, let us now extend the search for an optimal descent direction by allowing the stepwise change of inner product. Whereas the descents in the gradient method are steepest w.r. to different directions, we prove that *the descents in Newton's method are steepest w.r. to both different directions and inner products* up to a second order approximation in a neighbourhood of the solution.

We study an operator equation $F(u) = 0$ in a Hilbert space H under

Assumptions 2.8. The operator $F : H \to H$ is Gateaux differentiable, uniformly monotone and F' is locally Lipschitz continuous.

Let $u_0 \in H$ and let a variable steepest descent iteration be constructed in the form

$$u_{n+1} = u_n - B_n^{-1} F(u_n), \tag{2.36}$$

where we look for B_n in the class

$$\mathcal{B} \equiv \{B \in L(H) \text{ self-adjoint} : \exists p > 0 \quad \langle Bh, h\rangle \geq p\|h\|^2 \quad (h \in H)\}. \tag{2.37}$$

Let $n \in \mathbf{N}$ and assume that the nth term of the sequence (2.36) is constructed. Then the next step yields the functional value

$$m(B_n) := \phi(u_n - B_n^{-1} F(u_n))\,. \tag{2.38}$$

We wish to choose B_n such that this step is optimal, i.e. $m(B_n)$ is minimal. We verify that

$$\min_{B_n \in \mathcal{B}} m(B_n) \;=\; m(F'(u_n)) \qquad \text{up to second order} \tag{2.39}$$

as $u_n \to u^*$, i.e. the Newton iteration realizes asymptotically the stepwise optimal steepest descent among different inner products in the neighbourhood of u^*. (Clearly, the asymptotic result cannot be replaced by an exact one, this can be seen for fixed u_n by an arbitrary nonlocal change of ϕ along the descent direction.)

We can give an exact formulation in the following way. First, for any $\nu_1 > 0$ let

$$\mathcal{B}(\nu_1) \equiv \{B \in L(H) \text{ self-adjoint} : \quad \langle Bh, h\rangle \geq \nu_1\|h\|^2 \quad (h \in H)\}, \tag{2.40}$$

i.e. the subset of $\mathcal{B}$ with operators having the common lower bound $\nu_1 > 0$.

Theorem 2.8 *Let F satisfy Assumptions 2.8. Let $u_0 \in H$ and let the sequence (u_n) be given by (2.36) with operators $B_n \in \mathcal{B}$. Let $n \in \mathbf{N}$ be fixed and*

$$\hat{m}(B_n) := \beta + \frac{1}{2}\left\langle H_n(B_n^{-1}g_n - H_n^{-1}g_n),\; B_n^{-1}g_n - H_n^{-1}g_n\right\rangle, \tag{2.41}$$

where $\beta := \phi(u^*)$, $g_n := F(u_n)$, $H_n := F'(u_n)$. Then

(1) $\min\limits_{B_n \in \mathcal{B}} \hat{m}(B_n) \;=\; \hat{m}(F'(u_n))$;

(2) $\hat{m}(B_n)$ *is the second order approximation of $m(B_n)$, i.e., for any $B_n \in \mathcal{B}(\nu_1)$*

$$|m(B_n) - \hat{m}(B_n)| \leq C\|u_n - u^*\|^3 \tag{2.42}$$

where $C = C(u_0, \nu_1) > 0$ depends on u_0 and ν_1, but does not depend on B_n or u_n.

That is, up to second order, the descents in Newton's method are steepest w.r. to both different directions and inner products.

2.3.2 Inner-outer iterations: inexact Newton plus preconditioned CG

When the Jacobians are ill-conditioned, it is advisable to use inner iterations to solve the linearized equations. Hereby one can use preconditioning operators for the latter. The convergence of such inner-outer (Newton plus PCG) iterations relies on standard estimates. We give two classes of efficient preconditioners for the inner iterations.

(a) Symmetric problems with nonlinear principal part. In general, we have seen in section 1.2.1 that the spectral bounds m and M of a self-adjoint operator L_S imply $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \leq \frac{M}{m}$ independently of the given subspace V_h. Let a nonlinear Gateaux differentiable potential operator $F : H_S \to H_S$ satisfy the uniform ellipticity property

$$m\|v\|_S^2 \leq \langle F'(u)v, v\rangle_S \leq M\|v\|_S^2 \qquad (u, v \in H_S) \tag{2.43}$$

with $M, m > 0$, which ensures well-posedness of equation $F(u) = 0$. If u_n is the nth outer Newton iterate and $L_S := F'(u_n)$, then an inner CG iteration thus converges with a mesh independent convergence rate.

The following class of operators forms the most common special case to satisfy (2.43). Let H_S be a given Sobolev space over some bounded domain $\Omega \subset \mathbf{R}^d$, such that its inner product is expressed as

$$\langle h, v\rangle_S = \int_\Omega B(h, v) \tag{2.44}$$

for some given bilinear mapping $B : H_S \times H_S \to L^1(\Omega)$. Let the operator $F : H_S \to H_S$ have the form

$$\langle F(u), v\rangle_S = \int_\Omega \Big(a(B(u,u))\, B(u,v) - fv\Big) \qquad (u, v \in H_S), \tag{2.45}$$

where $f \in L^2(\Omega)$, further, $a : \mathbf{R}^+ \to \mathbf{R}^+$ is a scalar C^1 function for which there exist constants $M \geq m > 0$ such that

$$0 < m \leq a(r) \leq M, \qquad 0 < m \leq \tfrac{d}{dr}\Big(a(r^2)r\Big) \leq M \qquad (r \geq 0), \tag{2.46}$$

Proposition 2.1 *Under assumptions (2.45)–(2.46), the operator F satisfies (2.43).*

For a corresponding boundary value problem, the inner iterations for the linearized FEM systems converge with a mesh independent rate. The above bounds can be sharpened to depend on n, which can be more efficient in practice: we have

$$m_n \int_\Omega B(v, v) \leq \langle F'(u_n)v, v\rangle_S \leq M_n \int_\Omega B(v, v) \tag{2.47}$$

where, using notations $p(r^2) = \min\Big\{a(r^2),\ \frac{d}{dr}\Big(a(r^2)r\Big)\Big\}$, $q(r^2) = \max\Big\{a(r^2),\ \frac{d}{dr}\Big(a(r^2)r\Big)\Big\}$ $(r \geq 0)$, we have $m_n := \inf_\Omega p(B(u_n, u_n)) \geq m$, $\quad M_n := \sup_\Omega q(B(u_n, u_n)) \leq M$.

For example, various second order nonlinear elliptic problems (elasto-plastic torsion, magnetic potential, subsonic flow) lead to the weak form

$$\int_\Omega a(|\nabla u|^2)\nabla u \cdot \nabla v \;=\; \int_\Omega gv \qquad (v \in H_0^1(\Omega)),$$

where the given coefficient a satisfies (2.46). This falls into the above type where (2.44) is the standard $H_0^1(\Omega)$-inner product. Then Proposition 2.1 implies mesh independent convergence of the inner CG iterations such that one has to solve inner Poisson equations.

However, for strongly nonlinear $a(r)$ a much better preconditioning operator is the piecewise constant coefficient operator (2.34). Then one can derive the improved bounds

$$m_n = \min_i \Big(\inf_{\Omega_i} p(|\nabla u_n|^2)\,/c_i \Big), \qquad M_n = \max_i \Big(\sup_{\Omega_i} q(|\nabla u_n|^2)\,/c_i \Big)$$

determined only by the values of $|\nabla u_n|$ and the given scalar function a. In practice, for a magnetic potential problem, favourable condition numbers have been achieved [7]: e.g. 6 subdomains reduce the convergence factor from $Q = 0.9785$ to $Q = 0.6711$.

The elasto-plastic bending of clamped plates is described by a fourth order problem, whose weak formulation falls again into the above type where $[u, v] := \frac{1}{2}(D^2u : D^2v + \Delta u\, \Delta v)$. Using fixed preconditioners generated by this inner product, we are led to auxiliary biharmonic problems, for which fast solvers are available. For highly varying material nonlinearities, one can construct piecewise constant coefficient operators in an analogous way. A similar description holds for elasticity systems (see subsection 2.5.5).

(b) Semilinear problems. We consider nonsymmetric systems on a bounded domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3), involving second, first and zeroth order terms as well:

$$\left.\begin{aligned} -\mathrm{div}\,(k_i\,\nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + q_i(x, u_1, \ldots, u_l) &= g_i \\ u_{i\,|\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, l). \tag{2.48}$$

Assumptions 2.3.2.

(i) (Smoothness:) $k_i \in L^\infty(\Omega)$, $\mathbf{b}_i \in C^1(\overline{\Omega})^d$ and $g_i \in L^2(\Omega)$ $(i = 1, \ldots, l)$, further, the function $q = (q_1, \ldots, q_l) : \Omega \times \mathbf{R}^l \to \mathbf{R}^l$ is measurable and bounded w.r. to the variable $x \in \Omega$ and C^1 in the variable $\xi \in \mathbf{R}^l$.

(ii) (Coercivity:) there is $m > 0$ such that $k_i \geq m$ holds for all $i = 1, \ldots, l$, further, using the notation $q'_\xi(x, \xi) := \frac{\partial q(x,\xi)}{\partial \xi}$,

$$q'_\xi(x, \xi)\,\eta \cdot \eta - \frac{1}{2}\,\big(\max_i \mathrm{div}\,\mathbf{b}_i(x)\big)\,|\eta|^2 \geq 0 \tag{2.49}$$

for any $(x, \xi) \in \Omega \times \mathbf{R}^l$ and $\eta \in \mathbf{R}^l$.

(iii) (Local Lipschitz continuity:) let $3 \leq p$ (if $d = 2$) or $3 \leq p < 6$ (if $d = 3$), then there exist constants $c_1, c_2 \geq 0$ such that for any (x, ξ_1) and $(x, \xi_2) \in \Omega \times \mathbf{R}^l$,

$$\big\| q'_\xi(x, \xi_1) - q'_\xi(x, \xi_2) \big\| \leq \Big(c_1 + c_2\,(\max\,|\xi_1|, |\xi_2|)^{p-3} \Big) |\xi_1 - \xi_2|.$$

The FEM discretization and Newton linearization of this system leads to the FEM solution of linear elliptic systems of the form (1.31). We use the PCGN method based on a preconditioning operator S, which is the independent l-tuple of elliptic operators

$$S_i u_i := -\mathrm{div}\,(k_i\,\nabla u_i) + \beta_i u_i \quad \text{for} \quad u_{i\,|\partial\Omega} = 0 \qquad (i = 1, \ldots, l), \tag{2.50}$$

where $\beta_i \in L^\infty(\Omega)$ and $\beta_i \geq 0$.

The following theorem provides *superlinear convergence independently of both the mesh size h and the outer iterate $\mathbf{u}_n$*. To formulate the result, we denote

$$s_i^{(p)} := \min_{H_{i-1} \subset H_0^1(\Omega)^l} \max_{\mathbf{v} \perp H_{i-1}} \frac{\|\mathbf{v}\|^2_{L^p(\Omega)^l}}{\|\mathbf{v}\|^2_S},$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace and orthogonality is understood in S-inner product. (These are related to the Gelfand numbers of the compact Sobolev embeddings.)

Theorem 2.9 *Let Assumptions 2.3.2 hold. The CGN algorithm with $\mathbf{S}_h$-inner product, applied for the $n \times n$ preconditioned FEM system at linearization $\mathbf{u}_n$, yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \hat{\varepsilon}_k \quad (k = 1, ..., n) \quad \textit{with} \quad \hat{\varepsilon}_k := \frac{2}{km^2} \sum_{i=1}^{k} \Big(C_1 s_i^{(2)} + C_2 s_i^{(p)} \Big) \to 0 \quad (2.51)$$

as $k \to \infty$, and here the sequence $(\hat{\varepsilon}_k)_{k \in \mathbf{N}^+}$ is independent of V_h and $\mathbf{u}_n$.

Remark 2.3 (i) One can give explicit asymptotics using the related Gelfand numbers and eigenvalues. In particular, when the $\mathbf{u}_n$ are uniformly bounded as $h \to 0$, then (1.26) holds.

(ii) Instead of the above Dirichlet problem, one could include mixed boundary conditions or interface conditions, see [3] and the numerical tests in subsection 2.5.6.

2.4 Newton's method: a characterization of mesh independence

A missing part of the theory so far is the mesh independence of quadratic convergence of Newton's method for general elliptic problems. A related property, the classical mesh independence principle (MIP) has been established on a general level in [1], and then a lot of important work has been done, see [42] and refs. there. The MIP states that the number of required iterations for some tolerance remains essentially the same as the mesh is refined. The real strength of the result is that this common convergence is quadratic. (Mesh independent linear convergence can be produced by much cheaper methods.)

This and all later results are based on the underlying Lipschitz continuity for the derivatives of the operator. However, in the mentioned works this Lipschitz continuity appears only as an assumption in general, and it is only proved for semilinear problems.

The goal of this section is to clarify this phenomenon for a general class of second order elliptic problems solved by FEM discretization. It will be shown that mesh uniform quadratic estimates in fact cannot be produced unless the principal part is linear. For this study, the '*mesh independence principle for quadratic convergence*' (MIPQC) is introduced, which only requires that the quadratic convergence rate is uniformly bounded as the mesh is refined.

Briefly, our result then states that the MIPQC holds if and only if the elliptic equation is semilinear. Moreover, this is an inherent property for this class of problems, not due to too little smoothness etc. The underlying property is that in the case of a nonlinear

principal part the derivative F' of the differential operator is not Lipschitz continuous in the corresponding Sobolev space.

We consider second order nonlinear elliptic boundary value problems of the form

$$\left\{\begin{array}{rcll} -\operatorname{div} f(x,\nabla u)+q(x,u) &=& g(x) & \text{in } \Omega \\ f(x,\nabla u)\cdot\nu+s(x,u) &=& \gamma(x) & \text{on } \Gamma_N \\ u &=& 0 & \text{on } \Gamma_D\,. \end{array}\right. \tag{2.52}$$

We impose the following conditions:

Assumptions 2.10.

(i) (Domain.) $\Omega\subset\mathbf{R}^d$, $d=2$ or 3, is a bounded domain with piecewise smooth boundary, $\Gamma_N,\Gamma_D\subset\partial\Omega$ are measurable open subsurfaces, $\Gamma_N\cap\Gamma_D=\emptyset$, $\overline{\Gamma}_N\cup\overline{\Gamma}_D=\partial\Omega$ and $\Gamma_D\neq\emptyset$.

(ii) (Smoothness.) The functions $f:\Omega\times\mathbf{R}^d\to\mathbf{R}^d$, $q:\Omega\times\mathbf{R}\to\mathbf{R}$ and $s:\Gamma_N\times\mathbf{R}\to\mathbf{R}$ are measurable and bounded w.r. to the variable $x\in\Omega$ resp. $x\in\Gamma_N$ and C^1 in the other variables. Further, $g\in L^2(\Omega)$ and $\gamma\in L^2(\Gamma_N)$.

(iii) (Ellipticity.) The Jacobians $f'_\eta(x,\eta):=\frac{\partial f(x,\eta)}{\partial\eta}$ are symmetric and have eigenvalues between constants $M\geq m>0$ independent of (x,η); further, for any $x\in\Omega$ resp. $x\in\Gamma_N$ and $\xi\in\mathbf{R}$, we have $0\leq q'_\xi(x,\xi)$ and $0\leq s'_\xi(x,\xi)$.

(iv) (Lipschitz derivatives for the principal part.) The Jacobians f'_η are Lipschitz continuous w.r. to η, i.e., there exists a constant $l_f>0$ such that for all $(x,\eta_1),(x,\eta_2)\in\Omega\times\mathbf{R}^d$ we have $\|f'_\eta(x,\eta_1)-f'_\eta(x,\eta_2)\|\leq l_f|\eta_1-\eta_2|$.

(v) (Lipschitz derivatives for the lower order terms.) Let $3\leq p_1$ (if $d=2$) or $3\leq p_1\leq 6$ (if $d=3$), then there exist constants $c_1,c_2\geq 0$ such that for any (x,ξ_1) and $(x,\xi_2)\in\Omega\times\mathbf{R}$,

$$\left|q'_\xi(x,\xi_1)-q'_\xi(x,\xi_2)\right| \leq \Big(c_1+c_2\,(\max\,|\xi_1|,|\xi_2|)^{p_1-3}\Big)|\xi_1-\xi_2|. \tag{2.53}$$

Further, let $3\leq p_2$ (if $d=2$) or $3\leq p_2\leq 4$ (if $d=3$), then there exist constants $d_1,d_2\geq 0$ such that for any (x,ξ_1) and $(x,\xi_2)\in\Gamma_N\times\mathbf{R}$,

$$\left|s'_\xi(x,\xi_1)-s'_\xi(x,\xi_2)\right| \leq \Big(d_1+d_2\,(\max\,|\xi_1|,|\xi_2|)^{p_2-3}\Big)|\xi_1-\xi_2|. \tag{2.54}$$

The Sobolev space $H^1_D(\Omega):=\{u\in H^1(\Omega):\,u_{|\Gamma_D}=0\}$, corresponding to the Dirichlet boundary Γ_D, is endowed with the standard inner product.

Definition 2.2 Problem (2.52) satisfies the *mesh independence principle for quadratic convergence (MIPQC)* of Newton's method for admissible discretizations if under Assumptions 2.10, there exist constants $h_0>0$ and $\delta>0$ *independent of* V_h with the following property:

taking into account admissible FEM subspaces $V_h \subset H^1_D(\Omega)$ with mesh parameter h, and initial guesses $u_0 = u_0^h \in V_h$, the Newton iterates satisfy

$$\sup\Big\{ \frac{\|F_h(u_{n+1})\|_{H^1_D}}{\|F_h(u_n)\|^2_{H^1_D}} : \quad h < h_0, \quad \|u_0 - u_h\|_{H^1_D} < \delta, \quad n \in \mathbf{N} \Big\} \; < \; \infty. \tag{2.55}$$

Theorem 2.10 *Let Assumptions 2.10 hold and $f \in C^2(\Omega \times \mathbf{R}^d,\ \mathbf{R}^d)$. Problem (2.52) satisfies the MIPQC of Definition 2.2 if and only if $\eta \mapsto f(x,\eta)$ is linear, i.e. the elliptic equation is semilinear.*

We note that Assumption $f \in C^2(\Omega \times \mathbf{R}^d,\ \mathbf{R}^d)$ is only required to prove the 'only if' part, the 'if' part holds under Assumptions 2.10 themselves.

2.5 Applications to efficient computational algorithms

2.5.1 Nonlinear stationary Maxwell equations: the electromagnetic potential

The 2D stationary electromagnetic field in the cross-section of a device $\Omega \subset \mathbf{R}^2$ under nonlinear dependence between the magnetic field H and induction B is described by the nonlinear Maxwell equations

$$\mathrm{rot} H = \rho \quad \text{and} \quad \mathrm{div}\, B = 0 \quad \text{in } \Omega, \qquad B \cdot \nu \; = \; 0 \quad \text{on } \partial\Omega$$

and in general the relation $H \; = \; b(x, |B|)\, B$. The electromagnetic potential u is defined by $\mathrm{curl}\, u \; = \; B$, for which one is led to a special case of (2.30). An example of arising nonlinearity is $b(x,r) \equiv a(r) = \frac{1}{\mu_0}\Big(\alpha + (1-\alpha)\frac{r^8}{r^8+\beta}\Big)$ $(r \geq 0)$, see [18]; the realistic values $\alpha = 0.0003$ and $\beta = 16000$ show that the problem is almost singular. We consider this nonlinearity a and solve the problem (2.30) with $f(x, \nabla u) \equiv a(|\nabla u|)\nabla u$.

We have run experiments [22] by applying the variable preconditioning procedure with piecewise constant coefficient preconditioning operators, developed in Theorem 2.7. We considered the unit square domain Ω and piecewise linear elements. The experiment was made using 2^k node points of the mesh with $k = 6$, 8 and 10. Table 1 summarizes the number of iterations that decrease the residual error $\|F(u_n)\|$ below 10^{-4} and 10^{-8}, respectively. The results exhibit *mesh independence*, i.e. the number of iterations remains the same when the number of node points is increased.

node points:	2^6	2^8	2^{10}
# iterations for $\varepsilon = 10^{-4}$:	10	10	10
# iterations for $\varepsilon = 10^{-8}$:	16	16	16

Altogether, our method is less costly than either a Newton or a frozen coefficient iteration due to the structure of the stiffness matrix, which only slightly increases the complexity of a discrete Laplacian. *Comparison* with some related results for such problems showed that our method required the smallest number of iterations [22].

2.5.2 Elasto-plastic torsion of a hardening rod

Let us consider a hardening rod with cross-section $\Omega \subset \mathbf{R}^2$, the lower end of the rod being clamped in the (x, y)-plane. The Saint-Venant model leads to the problem

$$-\frac{\partial}{\partial x}\left(\overline{g}(T)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\overline{g}(T)\frac{\partial u}{\partial y}\right) = 2\omega\,, \qquad u_{|\partial\Omega} = 0 \tag{2.56}$$

where $\tau = (\tau_x, \tau_y)$ is the tangential stress vector, u is the stress function, $T := |\tau| = |\nabla u|$ and the stress-strain nonlinearity $\overline{g} \in C^1[0, T_*]$ satisfies $0 < \mu_1 \le \overline{g}(T) \le (\overline{g}(T)T)' \le \mu_2$ $(T \in [0, T_*])$ with suitable constants μ_1, μ_2 independent of T. This problem is a special case of (2.30).

We have solved problem (2.56) numerically using a FEM discretization and then Sobolev gradient preconditioning with the discrete Laplacian preconditioner. The mesh independent convergence of the algorithm follows from Theorem 2.3.

We have run numerical experiments for a copper rod with a square cross-section 10 mm $\times$ 10 mm, heat treated at the temperature $600°C$ for 1 hour. The numerical tests used $\omega = 0.3613$. We applied C^1-elements, based on the hp-FEM [40] for qualitative aspects, since the continuity of the tangential stress field τ is thus reproduced by the numerical approximations without postprocessing. The computations were executed up to accuracy 10^{-4}, and we have determined the regions of elastic state, plastic state and crack. It took 16 iterations to achieve the prescribed accuracy.

2.5.3 The electrostatic potential equation

The electrostatic potential in a bounded domain $\Omega \subset \mathbf{R}^3$ is described by the problem

$$-\Delta u + e^u = 0, \qquad u_{|\partial\Omega} = 0, \tag{2.57}$$

see e.g. [18]. We have solved this problem numerically on a ball with radius $R = 2$ using Sobolev gradient preconditioning, see [27]. The main feature is that one can realize Theorem 2.1 directly in the Sobolev space $H_0^1(B)$ by keeping the iterates in the class of radially symmetric polynomials where the Laplacian can be inverted exactly.

The advantages of this method is the simplicity of the algorithm that allows straightforward coding, and the obtained fast linear convergence: the residuals achieved accuracy 10^{-6} in 9 steps.

2.5.4 Some other semilinear problems

Here we briefly mention some further applicability of our Sobolev and variable gradient methods to semilinear problems.

Nonlocal boundary-value problems. Such models arise when the flux on the boundary is influenced by the behaviour on the whole surface. A detailed model was elaborated and Sobolev gradient preconditioning was applied by properly adapting Theorem 2.4 in [19]. Numerical experiments for the problem

$$-\Delta u + u^3 = g(x, y) \quad \text{in } \Omega, \qquad \frac{\partial u}{\partial \nu} + \int_{\partial\Omega} u\, d\sigma = 0 \quad \text{on } \partial\Omega \tag{2.58}$$

were executed via truncated Fourier series, and accuracy 10^{-4} was achieved in 21 iterations.

Gradient systems. Reaction-diffusion systems where the reactions form a gradient vector function are described by a system of boundary value problems that is a special case of (2.14). We apply the iteration (2.17)–(2.18) whose convergence is ensured by Theorem 2.4. The iteration requires the solution of independent linear elliptic problems. Numerical experiments were run in [20] in the same spirit as for the nonlocal problem above: the system

$$\begin{cases} -\Delta u + u - v + u^3 \;=\; g_1(x,y) \\ -\Delta v + v - u + v^3 \;=\; 0 \\ u_{|\Gamma_1} = v_{|\Gamma_1} = 0,\;\; \partial_\nu u_{|\Gamma_2} = \partial_\nu v_{|\Gamma_2} = 0 \end{cases} \tag{2.59}$$

was solved numerically by solving the auxiliary Poisson equations via truncated Fourier series, and accuracy 10^{-4} was achieved in 18 iterations.

Radiative cooling. The steady-state temperature $u \geq 0$ in a radiating body $\Omega \subset \mathbf{R}^3$ is described by the problem

$$-\mathrm{div}\,(\kappa(x)\,\nabla u) + \sigma(x) u^4 = 0 \quad \text{in } \Omega, \qquad \kappa(x)\tfrac{\partial u}{\partial \nu} + \alpha(x)\,(u - \tilde{u}(x)) = 0 \quad \text{on } \partial\Omega \tag{2.60}$$

where $\kappa(x) > 0$ is the thermal conductivity, $\sigma(x) > 0$ is the Boltzmann factor, $\alpha(x) > 0$ is the heat transfer coefficient, $\tilde{u}(x) > 0$ is the external temperature. Problem (2.60) is a special case of problem (2.35), hence Corollary 2.2 provides convergence of the variable preconditioning procedure using constant coefficient operators with stepwise redefined coefficient of u. We can cite the numerical tests executed in [28], which show that this variable preconditioning iteration is faster w.r.t. run time compared to Newton's method, due to the lack of updating the coefficients.

2.5.5 Nonlinear elasticity systems

The description of an elastic body in structural mechanics leads to an elliptic system of three equations with mixed boundary conditions:

$$\left\{ \begin{array}{l} -\,\mathrm{div}\; T_i(x, \varepsilon(\mathbf{u})) \;=\; \varphi_i(x) \quad \text{in } \Omega \\ \\ u_i = 0 \;\; \text{on } \Gamma_D\,, \qquad T_i(x,\varepsilon(\mathbf{u}))\cdot\nu \;=\; \gamma_i(x) \;\; \text{on } \Gamma_N \end{array} \right\} \quad (i = 1,2,3), \tag{2.61}$$

where the vector function $\mathbf{u} : \Omega \to \mathbf{R}^3$ represents displacement, and the tensor T is expressed with the scalar nonlinear bulk modulus k and Lamé's coefficient μ having fixed spectral bounds $\Lambda_0 \geq \lambda_0 > 0$.

One can solve this problem by an outer-inner iteration as described in paragraph (a) of subsection 2.3.2. Then a crucial step is the choice of preconditioner for the inner linear systems which consist of three equations. An efficient choice of inner preconditioning operator is the triplet of independent Laplacians:

$$Sz = \big(-\Delta z_1,\, -\Delta z_2,\, -\Delta z_3\big),$$

called separate displacement preconditioner. Then the corresponding stiffness matrix is block diagonal, and hence the three subproblems can be solved in parallel.

Theorem 2.11 *The separate displacement preconditioner satisfies*

$$cond(S_h^{-1} L_h^{(n)}) \leq \kappa \, \frac{\Lambda_0}{\lambda_0} \tag{2.62}$$

where κ is the Korn constant and λ_0 and Λ_0 are the spectral bounds of k and μ.

Consequently, the inner PCG iteration converges with a ratio independent of both the mesh size h and the outer Newton iterate.

2.5.6 Interface problems for localized reactions

Chemical reaction-diffusion equations may involve reactions that take place in a localized way on a surface (interface), leading to so-called interface conditions. We consider compound nonlinear interface problems that involve reaction terms both inside the domain and on the interface:

$$\begin{cases} \quad -\Delta u + q(x,u) = f(x) \quad \text{in } \Omega \setminus \Gamma, \\ [u]_\Gamma = 0 \ \text{ on } \Gamma, \qquad \left[\frac{\partial u}{\partial \nu}\right]_\Gamma + s(x,u) = \gamma(x) \ \text{ on } \Gamma, \qquad u = g(x) \ \text{ on } \partial\Omega, \end{cases} \tag{2.63}$$

where $[u]_\Gamma$ and $\left[\frac{\partial u}{\partial \nu}\right]_\Gamma$ denote the jump (i.e. the difference of the limits from the two sides of the interface Γ) of u and $\frac{\partial u}{\partial \nu}$, respectively. The weak form and corresponding iterations can be described in an analogous way to mixed boundary conditions, therefore an analogue of Theorem 2.9 can be derived for outer-inner iterations [3]. Thereby, we have run experiments on a test-problem on the domain $\Omega = [0,1]\times[0,1]$ with $\Gamma = [0,1]\times\{\frac{1}{2}\}$, and we have chosen polynomials $q(x,\xi) := 1+\xi^3$ and $s(x,\xi) := 1+\xi^5$. We used Courant elements for the FEM discretization using uniform mesh. The stopping criterion $\|F_h(u_{nh}) - f_h\|_S \leq 10^{-10}$ was reached in 15 outer Newton iterations for either $h = 1/64$, $1/128$ or $1/192$ points, and also the number of inner CG iterations was mesh independent.

2.5.7 Nonsymmetric transport systems

Various steady-state transport (convection-reaction-diffuson) problems are described by a system

$$-\Delta u_i + \mathbf{b}_i \cdot \nabla u_i + f_i(u_1, \ldots, u_l) = g_i\,, \qquad u_{i\,|\partial\Omega} = 0 \qquad (i = 1, \ldots, l), \tag{2.64}$$

where $\mathbf{b}_i$ represents convection and the f_i characterize the rate of reaction between the components. Such systems satisfy suitable coercivity conditions that are typically special cases of Assumptions 2.3.2, in which case the system becomes of the form (2.48). One can solve this problem by an outer-inner iteration as described in paragraph (b) of subsection 2.3.2. For an inner preconditioning operator one can propose the l-tuple of independent diffusion operators as in (1.69). The solution of the linearized systems admits efficient parallelization, as mentioned in subsection 1.4.5. For such preconditioning both the outer and inner iterations produce superlinear convergence.

We have made experiments on the test system on the domain $\Omega = [0,1]\times[0,1]$, where $\mathbf{b}_i = (1,1)^T$ for all i, and $f(\mathbf{u}) = 4\mathbf{A}\,|\mathbf{u}|^2\mathbf{u}$ where $\mathbf{A}$ is the lower triangular part of the constant 1 matrix. The auxiliary problems were solved with FFT. The stopping criterion $\|F_h(\underline{u}_n) - b_h\| \leq 10^{-5}$ was always reached in 11 Newton iterations for mesh sizes ranging from $h = 1/16$ to $1/96$, and also the number of inner CG iterations was mesh independent.

2.5.8 Parabolic air pollution systems

The modelling of air pollution leads to a parabolic system which is a compound nonlinear transport system involving diffusion, convection, reaction and deposition terms [43], and often consists of a huge number of equations:

$$\left.\begin{aligned} &\frac{\partial u_i}{\partial t} - \operatorname{div}\left(k_i(x)\,\nabla u_i\right) + \mathbf{b}_i(x)\cdot\nabla u_i + c_i(x)u + f_i(x,t,u_1,\ldots,u_l) = 0 \\ &u_i(x,0) = \varphi_i(x) \quad (x\in\Omega), \qquad\qquad u_{i\,|\partial\Omega\times\mathbf{R}^+} = 0 \end{aligned}\right\} \quad (i=1,\ldots,l).$$

(A linearized form was studied in subsection 1.4.4.) Such problems are normally solved by time discretization, Newton linearization and inner PCG iteration. The nonlinear systems arising after time discretization are similar to (2.64), studied in the previous section.

Now we were interested in the convergence in time and the behaviour of the overall algorithm, so as to demonstrate that the so far developed elliptic solvers are suitable to be a subroutine to a parabolic solution process. Numerical tests were run on the unit square domain for a system consisting of 10 equations, with chemical reactions arising from the air pollution model in [43].

Regarding space discretization, the number of outer DIN iterations (executed in every time step) and the number of outer PCG iterations (carried out in each DIN step) were found mesh independent, using stopping criterion $\|F_h(u) - b_h\| < 10^{-8}$. Regarding time discretization, since no exact solution was available, we compared $u_h^{(\tau)}$ and $u_h^{(\tau/2)}$. We found that the error $\|u_h^{(\tau)} - u_h^{(\tau/2)}\| \to 0$ numerically as $\tau \to 0$, which shows numerical convergence of the method w.r.t. time.

3 Discrete maximum principles

The discrete maximum principle is an important measure of the qualitative reliability of the numerical scheme, including FEM, see [24] and the references there. Whereas the most classical DMP has the form $\max_{\overline{\Omega}} u_h = \max_{\partial\Omega} u_h$, we will study the case

$$\max_{\overline{\Omega}} u_h \le \max\{0, \max_{\partial\Omega} u_h\} \tag{3.1}$$

which arises when $L_h u_h \le 0$ for operators including lower order terms.

3.1 Algebraic background

Let us now consider a system of equations of order $(k+m)\times(k+m)$ with the following structure:

$$\begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \tilde{\mathbf{b}} \end{bmatrix}. \tag{3.2}$$

The goal here is to establish the algebraic analogue of (3.1):

$$\max_{i=1,\ldots,k+m} c_i \;\le\; \max\{0, \max_{i=k+1,\ldots,k+m} c_i\}. \tag{3.3}$$

Definition 3.1 [12] A $(k+m)\times(k+m)$ matrix $\bar{\mathbf{A}}$ with the structure (3.2) is said to be of *generalized nonnegative type* if the following properties hold:

(i) $a_{ii} > 0, \quad i = 1, ..., k,$

(ii) $a_{ij} \le 0, \quad i = 1, ..., k, \ j = 1, ..., k+m \quad (i \ne j),$

(iii) $\sum_{j=1}^{k+m} a_{ij} \ge 0, \quad i = 1, ..., k,$

(iv) There exists an index $i_0 \in \{1, \dots, k\}$ for which $\sum_{j=1}^{k} a_{i_0,j} > 0.$

(v) $\mathbf{A}$ is irreducible.

Then, if $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0, \ i = 1, ..., k$, then (3.3) holds [12].

The irreducibility of $\mathbf{A}$ is sometimes difficult to check, hence we will use the following

Definition 3.2 A $(k+m)\times(k+m)$ matrix $\bar{\mathbf{A}}$ with the structure (3.2) is said to be of *generalized nonnegative type with irreducible blocks* if properties (i)–(iii) of Definition 3.1 hold, further, property (iv) therein is replaced by the following stronger one:

(iv') For each irreducible component of $\mathbf{A}$ there exists an index $i_0 = i_0(l) \in N_l = \{s_1^{(l)}, \dots, s_{k_l}^{(l)}\}$ for which $\sum_{j=1}^{k} a_{i_0,j} > 0.$

Theorem 3.1 *Let $\bar{\mathbf{A}}$ be a $(k+m)\times(k+m)$ matrix with the structure (3.2), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.*

If the vector $\bar{\mathbf{c}} = (c_1, ..., c_{k+m})^T \in \mathbf{R}^{k+m}$ is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0, \ i = 1, ..., k$, then (3.3) holds.

3.2 A matrix maximum principle in Hilbert space

First we describe the operator equation and its discretization. Let H be a real Hilbert space and $H_0 \subset H$ a given subspace. We consider the following operator equation: for given vectors $\psi, g^* \in H$, find $u \in H$ such that

$$\langle A(u), v\rangle = \langle \psi, v\rangle \qquad (v \in H_0) \tag{3.4}$$

$$\text{and} \quad u - g^* \in H_0 \tag{3.5}$$

with an operator $A : H \to H$ satisfying the following conditions:

Assumptions 3.2.1.

(i) The operator $A : H \to H$ has the form $A(u) = B(u)u + R(u)u$, where B and R are given operators mapping from H to $\mathcal{B}(H)$.

(ii) There exists a constant $m > 0$ such that $\langle B(u)v, v\rangle \ge m\,\|v\|^2 \quad (u \in H, \ v \in H_0).$

(iii) There exist subsets of 'positive vectors' $D, P \subset H$ such that for any $u \in H$ and $v \in D$, we have $\langle R(u)w, v\rangle \geq 0$ provided that either $w \in P$ or $w = v \in D$.

(iv) There exists a continuous function $M_R : \mathbf{R}^+ \to \mathbf{R}^+$ and another norm $|||.|||$ on H such that

$$\langle R(u)w, v\rangle \leq M_R(\|u\|)\, |||w|||\, |||v||| \qquad (u, w, v \in H). \tag{3.6}$$

Now we turn to the numerical solution of our operator equation using Galerkin discretization. Let $n_0 \leq n$ be positive integers and $\phi_1, ..., \phi_n \in H$ be given linearly independent vectors such that $\phi_1, ..., \phi_{n_0} \in H_0$. We consider the finite dimensional subspaces

$$V_h = \text{span}\{\phi_1, ..., \phi_n\} \subset H, \qquad V_h^0 = \text{span}\{\phi_1, ..., \phi_{n_0}\} \subset H_0 \tag{3.7}$$

with a real positive parameter $h > 0$.

We formulate here some connectivity type properties for these subspaces that we will need later. For this, certain pairs $\{\phi_i, \phi_j\} \in V_h \times V_h$ are called *neighbouring basis vectors*, and then i, j are called *neighbouring indices*. The only requirement for the set of these pairs is that they satisfy Assumptions 3.2.3 below, given in terms of the *graph of neighbouring indices*, by which we mean the following. The corresponding indices $\{1, \ldots, n_0\}$ or $\{1, \ldots, n\}$, respectively, are represented as vertices of the graph, and the ith and jth vertices are connected by an edge iff i, j are neighbouring indices.

Assumptions 3.2.3. The set $\{1, \ldots, n\}$ can be partitioned into disjoint sets $S_1, \ldots, S_r$ such that for each $k = 1, \ldots, r$,

(i) both $S_k^0 := S_k \cap \{1, \ldots, n_0\}$ and $\tilde{S}_k := S_k \cap \{n_0 + 1, \ldots, n\}$ are nonempty;

(ii) the graph of all neighbouring indices in S_k^0 is connected;

(iii) the graph of all neighbouring indices in S_k is connected.

(In later PDE applications, these properties are meant to express that the supports of basis functions cover the domain, both its interior and the boundary.)

Now let $\tilde{g} = \sum_{j=n_0+1}^{n} g_j \phi_j \in V_h$ be a given approximation of the component of g^* in $H \setminus H_0$. To find the Galerkin solution of (3.4)–(3.5) in V_h, we solve the following problem: find $u^h \in V_h$ such that

$$\langle A(u^h), v\rangle = \langle \psi, v\rangle \qquad (v \in V_h^0) \tag{3.8}$$

$$\text{and} \quad u^h - \tilde{g} \in V_h^0. \tag{3.9}$$

Using Assumption 3.2.1. (i), we can rewrite (3.8) as

$$\langle B(u^h)u^h, v\rangle + \langle R(u^h)u^h, v\rangle = \langle \psi, v\rangle \qquad (v \in V_h^0). \tag{3.10}$$

Let us now formulate the nonlinear algebraic system corresponding to (3.10). We set

$$u^h = \sum_{j=1}^{n} c_j \phi_j, \tag{3.11}$$

and look for the coefficients $c_1, \ldots, c_n$. Using a partition of coefficients corresponding to V_h^0 and its complement, we obtain a system

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} \equiv \begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \tilde{\mathbf{g}} \end{bmatrix} . \tag{3.12}$$

Now we formulate and prove a *maximum principle* for the abstract discretized problem. The following notion will be crucial for our study:

Definition 3.3 A set of subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ in H is said to be a *family of subspaces* if for any $\varepsilon > 0$ there exists $V_h \in \mathcal{V}$ with $h < \varepsilon$.

First we give sufficient conditions for the generalized nonnegativity of the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$.

Theorem 3.2 *Let Assumptions 3.2.1 and 3.2.3 hold. Let us consider the discretization of operator equation (3.4)–(3.5) in a family of subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ with bases as in (3.7). Let $u^h \in V_h$ be the solution of (3.10) and let the following properties hold:*

(a) For all $\phi_i \in V_h^0$ and $\phi_j \in V_h$, one of the following holds: either

$$\langle B(u^h)\phi_j, \phi_i\rangle = 0 \quad \textit{and} \quad \langle R(u^h)\phi_j, \phi_i\rangle \le 0, \tag{3.13}$$

$$\textit{or} \qquad \langle B(u^h)\phi_j, \phi_i\rangle \le -M_B(h) \tag{3.14}$$

with a proper function $M_B : \mathbf{R}^+ \to \mathbf{R}^+$ (independent of h, ϕ_i, ϕ_j) such that, defining

$$T(h) := \sup\{|||\phi_i||| : \ \phi_i \in V_h)\}\,, \tag{3.15}$$

we have

$$\lim_{h\to 0} \frac{M_B(h)}{T(h)^2} = +\infty. \tag{3.16}$$

(b) If, in particular, $\phi_i \in V_h^0$ and $\phi_j \in V_h$ are neighbouring basis vectors (as defined for Assumptions 3.2.3), then (3.14)–(3.16) hold.

(c) $M_R(\|u^h\|)$ is bounded as $h \to 0$, where M_R is the function in Assumption 3.2.1 (iv).

(d) For all $u \in H$ and $h > 0$, $\sum_{j=1}^{n} \phi_j \in ker\, B(u)$.

(e) For all $h > 0$, $i = 1, ..., n$, we have $\phi_i \in D$ and $\sum_{j=1}^{n} \phi_j \in P$ for the sets D, P introduced in Assumption 3.2.1 (iii).

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.12) is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.

By Theorem 3.1, we immediately obtain the corresponding *matrix maximum principle* (or *algebraic discrete maximum principle*):

Corollary 3.1 *Let the assumptions of Theorem 3.2 hold. For sufficiently small h, if $d_i \le 0$ $(i = 1, ..., n_0)$ and $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ is the solution of (3.12), then*

$$\max_{i=1,\ldots,n} c_i \ \le\ \max\{0, \max_{i=n_0+1,\ldots,n} c_i\}. \tag{3.17}$$

3.3 Discrete maximum principles for nonlinear elliptic problems

3.3.1 Nonlinear elliptic equations

We consider a nonlinear boundary value problem of the following type:

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) + q(x,u) \;=\; f(x) \quad \text{in } \Omega, \\ \qquad\qquad b(x,\nabla u)\frac{\partial u}{\partial \nu} + s(x,u) \;=\; \gamma(x) \quad \text{on } \Gamma_N, \\ \qquad\qquad\qquad\qquad u \;=\; g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{3.18}$$

where Ω is a bounded domain in $\mathbf{R}^n$, under the following assumptions:

(A1) Ω has a piecewise smooth and Lipschitz continuous boundary $\partial\Omega$; $\Gamma_N, \Gamma_D \subset \partial\Omega$ are measurable open sets, such that $\Gamma_N \cap \Gamma_D = \emptyset$ and $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$.

(A2) The scalar functions $b : \overline{\Omega} \times \mathbf{R}^n \to \mathbf{R}$, $q : \overline{\Omega} \times \mathbf{R} \to \mathbf{R}$ and $s : \overline{\Gamma}_N \times \mathbf{R} \to \mathbf{R}$ are continuously differentiable in their domains of definition. Further, $f \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$ and $g = g^*{}_{|\Gamma_D}$ with $g^* \in H^1(\Omega)$.

(A3) The function b satisfies

$$0 < \mu_0 \le b(x,\eta) \le \mu_1 \tag{3.19}$$

with positive constants μ_0 and μ_1 independent of (x,η), further, the diadic product matrix $\eta \cdot \frac{\partial b(x,\eta)}{\partial \eta}$ is symmetric positive semidefinite and bounded in any matrix norm by some positive constant μ_2 independent of (x,η).

(A4) Let $2 \le p_1$ if $d = 2$, or $2 \le p_1 \le \frac{2d}{d-2}$ if $d > 2$, further, let $2 \le p_2$ if $d = 2$, or $2 \le p_2 \le \frac{2d-2}{d-2}$ if $d > 2$. There exist functions $\alpha_1 \in L^{d/2}(\Omega)$, $\alpha_2 \in L^{d-1}(\Gamma_N)$ and a constant $\beta \ge 0$ such that for any $x \in \Omega$ (or $x \in \Gamma_N$, resp.) and $\xi \in \mathbf{R}$

$$0 \le q'_\xi(x,\xi) \le \alpha_1(x) + \beta|\xi|^{p_1-2}, \qquad 0 \le s'_\xi(x,\xi) \le \alpha_2(x) + \beta|\xi|^{p_2-2}.$$

(A5) Either $\Gamma_D \ne \emptyset$, or q increases strictly and at least linearly at ∞ in the sense that $q(x,\xi) \ge c_1|\xi| - c_2(x)$ (with a constant $c_1 > 0$ and a function $c_2 \in L^1(\Omega)$) $\forall (x,\xi) \in \Omega \times \mathbf{R}$, or s increases strictly and at least linearly at ∞ in the same sense.

Theorem 3.3 *Let (A1)–(A5) hold, and let us consider a family of simplicial triangulations $\mathcal{T}_h$ $(h > 0)$ satisfying the following properties:*

(i) for any $i = 1, ..., n,\ j = 1, ..., \bar{n}$ $(i \ne j)$

$$\nabla\phi_i \cdot \nabla\phi_j \le -\frac{\sigma_0}{h^2} < 0 \tag{3.20}$$

on supp $\phi_i \cap$ supp ϕ_j with $\sigma_0 > 0$ independent of i, j and h.

(ii) The triangulations $\mathcal{T}_h$ are regular, i.e., there exist constants $m_1, m_2 > 0$ such that for any $h > 0$ and any simplex $T_h \in \mathcal{T}_h$

$$m_1 h^d \le meas(T_h) \le m_2 h^d \tag{3.21}$$

(where meas(T_h) denotes the d-dimensional measure of T_h).

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type in the sense of Definition 3.1.

More generally, triangulations $\mathcal{T}_h$ are allowed to be only quasi-regular such that the left-hand side of (3.21) is replaced by $c_1 h^\gamma \le meas(T_h)$ for some $\gamma = \gamma(d, p_1, p_2) \ge d$.

Theorem 3.4 *Let the conditions of Theorem 3.3 hold, and let*

$$f(x) - q(x,0) \le 0,\ x \in \Omega, \qquad \text{and} \qquad \gamma(x) - s(x,0) \le 0,\ x \in \Gamma_N. \tag{3.22}$$

Then

$$\max_{\overline{\Omega}} u_h \le \max\{0, \max_{\Gamma_D} g_h\}. \tag{3.23}$$

In particular, if $\Gamma_D \ne \emptyset$ and $g \ge 0$ then $\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} \tilde{g}$, and if $\Gamma_D \ne \emptyset$ and $g \le 0$, or if $\Gamma_D = \emptyset$, then we have the nonpositivity property $\max_{\overline{\Omega}} u_h \le 0$.

Remark 3.1 (i) One can verify in the same way the *discrete minimum principle* under reversed sign conditions.

(ii) In the special case $q \equiv 0$ and $s \equiv 0$, the equality $\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} \tilde{g}$ holds without assuming $g \ge 0$. Moreover, the strict negativity in (3.20) can be replaced by the weaker condition $\nabla\phi_i \cdot \nabla\phi_j \le 0$, and no regularity of the mesh needs to be assumed.

3.3.2 Cooperative elliptic systems with nonlinear coefficients

We consider cooperative and weakly diagonally dominant systems, for which the (continuous) maximum principle holds.

Formulation of the problem. We consider nonlinear elliptic systems of the form

$$\left.\begin{aligned} -\mathrm{div}\left(b_k(x,u,\nabla u)\,\nabla u_k\right) + \sum_{l=1}^{s} V_{kl}(x,u,\nabla u)\,u_l \;&=\; f_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x,u,\nabla u)\tfrac{\partial u_k}{\partial \nu} = \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \qquad u_k \;&=\; g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \quad (k = 1, \dots, s) \tag{3.24}$$

with unknown function $u = (u_1, \dots, u_s)^T$, under the following assumptions. Here ∇u denotes the $s \times d$ tensor with rows ∇u_k $(k = 1, \dots, s)$, further, 'a.e.' means Lebesgue almost everywhere and inequalities for functions are understood a.e. pointwise for all possible arguments.

Assumptions 3.5.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \ne \emptyset$.

(ii) (Smoothness and boundedness.) For all $k, l = 1, \dots, s$ we have $b_k \in (C^1 \cap L^\infty)(\Omega \times \mathbf{R}^s \times \mathbf{R}^{s\times d})$ and $V_{kl} \in L^\infty(\Omega \times \mathbf{R}^s \times \mathbf{R}^{s\times d})$.

(iii) (Ellipticity.) There exists $m > 0$ such that $b_k \geq m$ holds for all $k = 1, \ldots, s$.

(iv) (Cooperativity.) We have $V_{kl} \leq 0 \quad (k, l = 1, \ldots, s,\ k \neq l)$.

(v) (Weak diagonal dominance.) We have $\sum_{l=1}^{s} V_{kl} \geq 0 \quad (k = 1, \ldots, s)$.

(vi) For all $k = 1, \ldots, s$ we have $f_k \in L^2(\Omega)$, $\gamma_k \in L^2(\Gamma_N)$, $g_k = g^*_{k|\Gamma_D}$ with $g^*_k \in H^1(\Omega)$.

Finite element discretization. We define the finite element discretization of problem (3.24) in the following way. First, let $\bar{n}_0 \leq \bar{n}$ be positive integers and let us choose basis functions

$$\varphi_1, \ldots, \varphi_{\bar{n}_0} \in H^1_D(\Omega), \qquad \varphi_{\bar{n}_0+1}, \ldots, \varphi_{\bar{n}} \in H^1(\Omega) \setminus H^1_D(\Omega), \tag{3.25}$$

which correspond to homogeneous and inhomogeneous boundary conditions on Γ_D, respectively. (For simplicity, we will refer to them as 'interior basis functions' and 'boundary basis functions', respectively, thus adopting the terminology of Dirichlet problems even in the general case.) These basis functions are assumed to be continuous and to satisfy

$$\varphi_p \geq 0 \quad (p = 1, \ldots, \bar{n}), \qquad \sum_{p=1}^{\bar{n}} \varphi_p \equiv 1, \tag{3.26}$$

further, that there exist node points $B_p \in \Omega$ $(p = 1, \ldots, \bar{n}_0)$ and $B_p \in \Gamma_D$ $(p = \bar{n}_0+1, \ldots, \bar{n})$ such that

$$\varphi_p(B_q) = \delta_{pq} \tag{3.27}$$

where δ_{pq} is the Kronecker symbol. (These conditions hold e.g. for standard linear, bilinear or prismatic finite elements.) Finally, we assume that any two interior basis functions can be connected with a chain of interior basis functions with overlapping support. By its geometric meaning, this assumption obviously holds for any reasonable FE mesh.

We in fact need a basis in the corresponding product spaces, which we define by repeating the above functions in each of the s coordinates and setting zero in the other coordinates. That is, let $n_0 := s\bar{n}_0$ and $n := s\bar{n}$. First, for any $1 \leq i \leq n_0$,

$$\begin{gathered}\text{if } \ i = (k-1)\bar{n}_0 + p \quad \text{for some } 1 \leq k \leq s \text{ and } 1 \leq p \leq \bar{n}_0, \ \text{ then} \\ \phi_i := (0, \ldots, 0, \varphi_p, 0, \ldots, 0) \qquad \text{where } \ \varphi_p \ \text{ stands at the } k\text{-th entry,}\end{gathered} \tag{3.28}$$

that is, $(\phi_i)_m = \varphi_p$ if $m = k$ and $(\phi_i)_m = 0$ if $m \neq k$. From these, we let

$$V^0_h := \operatorname{span}\{\phi_1, ..., \phi_{n_0}\} \ \subset H^1_D(\Omega)^s. \tag{3.29}$$

Similarly, for any $n_0 + 1 \leq i \leq n$,

$$\begin{gathered}\text{if } \ i = n_0 + (k-1)(\bar{n} - \bar{n}_0) + p - \bar{n}_0 \quad \text{for some } 1 \leq k \leq s \text{ and } \bar{n}_0 + 1 \leq p \leq \bar{n}, \ \text{ then} \\ \phi_i := (0, \ldots, 0, \varphi_p, 0, \ldots, 0)^T \qquad \text{where } \ \varphi_p \ \text{ stands at the } k\text{-th entry,}\end{gathered} \tag{3.30}$$

that is, $(\phi_i)_m = \varphi_p$ if $m = k$ and $(\phi_i)_m = 0$ if $m \neq k$. From (3.29) and these, we let

$$V_h := \operatorname{span}\{\phi_1, ..., \phi_n\} \ \subset H^1(\Omega)^s. \tag{3.31}$$

Using the above FEM subspaces, the discretization of problem (3.24) leads to a system of the form (3.12). In what follows, the (patch-)regularity of the considered meshes used in Theorem 3.3 will be usually weakened in some way. The following notions will be used:

Definition 3.4 Let $\Omega \subset \mathbf{R}^d$ and let us consider a family of FEM subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ constructed as above. Here $h > 0$ is the mesh parameter, proportional to the maximal diameter of the supports of the basis functions $\phi_1, ..., \phi_n$. The corresponding family of meshes will be called

(a) *regular from above* if there exists a constant $c_0 > 0$ such that for any $V_h \in \mathcal{V}$ and basis function $\varphi_p \in V_h$,

$$meas(\operatorname{supp} \varphi_p) \le c_0 h^d \tag{3.32}$$

(where *meas* denotes d-dimensional measure and supp denotes the support, i.e. the closure of the set where the function does not vanish);

(b) *quasi-regular* if (3.21) is replaced by

$$c_1 h^\gamma \le meas(\operatorname{supp} \varphi_p) \le c_2 h^d \tag{3.33}$$

for some fixed constant

$$d \le \gamma < d + 2, \tag{3.34}$$

and *regular* if $\gamma = d$.

We can then prove the following nonnegativity result for the stiffness matrix:

Theorem 3.5 *Let problem (3.24) satisfy Assumptions 3.5. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ satisfying the following property: there exists a real number γ satisfying*

$$d \le \gamma < d + 2$$

(where d is the space dimension) such that for any $p = 1, ..., \bar{n}_0,\ t = 1, ..., \bar{n}$ $(p \ne t)$, if $meas(\operatorname{supp} \varphi_p \cap \operatorname{supp} \varphi_t) > 0$ then

$$\nabla\varphi_t \cdot \nabla\varphi_p \le 0 \quad on\ \Omega \quad and \quad \int_\Omega \nabla\varphi_t \cdot \nabla\varphi_p \le -K_0\, h^{\gamma-2} \tag{3.35}$$

with some constant $K_0 > 0$ independent of p, t and h. Further, let the family of associated meshes be quasi-regular according to Definition 3.4.

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.

Theorem 3.6 *Let the assumptions of Theorem 3.5 hold and let*

$$f_k \le 0, \qquad \gamma_k \le 0 \qquad (k = 1, \dots, s).$$

Let the basis functions satisfy (3.26)–(3.27). Then for sufficiently small h, if $u^h = (u_1^h, \dots, u_s^h)^T$ is the FEM solution of system (3.24), then

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \le \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} g_k^h\}. \tag{3.36}$$

Remark 3.2 (i) The above result implies in particular that if $f_k \le 0$, $\gamma_k \le 0$ and $g_k \le 0$ on Γ_D for all k, then we obtain the nonpositivity property $u_k^h \le 0$ on Ω for all k.

(ii) Analogously, by reversing signs of all f_k, γ_k and g_k, we obtain the famous *nonnegativity property*: $u_k^h \ge 0$ on Ω for all k.

3.3.3 Systems with non-coefficient type reaction terms

The terms $\sum V_{kl}(x,u,\nabla u)\,u_l$ can be replaced by terms of the form $q_k(x,u_1,\dots,u_s)$, i.e. not using coefficients of u_l, if they do not depend on ∇u. If q_k grow at most linearly, then the same results hold if the assumptions on V_{kl} are replaced by the same assumptions on $\frac{\partial q_k}{\partial \xi_l}$. We do not formulate this separately. On the other hand, here one may allow the q_k to grow superlinearly, which needs strengthened assumptions. These are studied below.

Let us consider the system

$$-\mathrm{div}\left(b_k(x,\nabla u_k)\,\nabla u_k\right) + \; q_k(x,u_1,\dots,u_s) \;=\; f_k(x) \quad \text{a.e. in } \Omega \qquad (k=1,\dots,s) \tag{3.37}$$

with the boundary conditions of (3.24), under the following assumptions:

Assumptions 3.7.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(ii) (Smoothness and growth.) For all $k,l = 1,\dots,s$ we have $b_k \in (C^1 \cap L^\infty)(\Omega \times \mathbf{R}^d)$ and $q_k \in C^1(\Omega \times \mathbf{R}^s)$. Further, let

$$2 \le p < p^*, \quad \text{where } p^* := \tfrac{2d}{d-2} \text{ if } d \ge 3 \text{ and } p^* := +\infty \text{ if } d = 2; \tag{3.38}$$

then there exist constants $\beta_1, \beta_2 \ge 0$ such that

$$\left|\frac{\partial q_k}{\partial \xi_l}(x,\xi)\right| \le \beta_1 + \beta_2 |\xi|^{p-2} \qquad (k,l=1,\dots,s;\; x \in \Omega,\; \xi \in \mathbf{R}^s).$$

(iii) (Ellipticity.) There exists $m > 0$ such that $b_k \ge m$ holds for all $k = 1,\dots,s$. Further, defining $a_k(x,\eta) := b_k(x,\eta)\eta$ for all k, the Jacobian matrices $\frac{\partial}{\partial \eta}\, a_k(x,\eta)$ are uniformly spectrally bounded from both below and above.

(iv) (Cooperativity.) We have $\frac{\partial q_k}{\partial \xi_l}(x,\xi) \le 0 \quad (k,l=1,\dots,s,\; k \ne l;\; x \in \Omega,\; \xi \in \mathbf{R}^s)$.

(v) (Weak diagonal dominance for the Jacobians w.r.t. rows and columns.) We have for all $k = 1,\dots,s$, $x \in \Omega$ and $\xi \in \mathbf{R}^s$: $\displaystyle\sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x,\xi) \ge 0, \quad \sum_{l=1}^{s} \frac{\partial q_l}{\partial \xi_k}(x,\xi) \ge 0.$

(vi) For all $k = 1,\dots,s$ we have $f_k \in L^2(\Omega)$, $\gamma_k \in L^2(\Gamma_N)$, $g_k = g^*_{k|\Gamma_D}$ with $g^* \in H^1(\Omega)$.

We note that one might include additional terms $s_k(x,u_1,\dots,u_s)$ on the Neumann boundary Γ_N, which we omit here for technical simplicity; then s_k must satisfy similar properties as assumed for q_k.

When considering a FEM discretization developed as in subsection 3.3.2, we need a strengthened assumption for the quasi-regularity of the mesh such that (3.34) for γ is replaced by

$$d \le \gamma < \gamma^*_d(p) := 2d - \frac{(d-2)p}{2} \tag{3.39}$$

with p from Assumption 3.7 (ii).

Theorem 3.7 *Let problem (3.37) satisfy Assumptions 3.7, and let the assumptions of Theorem 3.5 hold except that the mesh quasi-regularity is understood with γ satisfying (3.39).*

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.

Accordingly, if $f_k \le q_k(x,0)$, $\gamma_k \le 0$ $(k = 1,\dots,s)$ and the basis functions satisfy (3.26)–(3.27), then for sufficiently small h, if $u^h = (u_1^h,\dots,u_s^h)^T$ is the FEM solution of system (3.37), then (3.36) holds.

3.3.4 Sufficient conditions and their geometric meaning

The key assumption for the FEM subspaces V_h and the associated meshes has been property (3.35). A classical way to satisfy such conditions is a pointwise inequality like (3.20) together with suitable mesh regularity. However, one can ensure (3.35) with less strong conditions as well, for instance:

- there exists $0 < \varepsilon \le \gamma - d$ such that the basis functions satisfy

$$\nabla\varphi_t \cdot \nabla\varphi_p \le -\frac{\sigma}{h^{2-\varepsilon}} < 0 \quad \text{on } \Omega_{pt}, \tag{3.40}$$

  but in the quasi-regularity assumption γ is replaced by $\gamma - \varepsilon$;

- there exist subsets $\Omega_{pt}^+ \subset \Omega_{pt}$ for all p,t such that $\inf\limits_{p,t} \frac{meas(\Omega_{pt}^+)}{meas(\Omega_{pt})} > 0$ and the basis functions satisfy

$$\nabla\varphi_t \cdot \nabla\varphi_p \le -\frac{\sigma}{h^2} < 0 \ \text{ on } \ \Omega_{pt}^+, \quad \nabla\varphi_t \cdot \nabla\varphi_p \le 0 \ \text{ on } \ \Omega_{pt} \setminus \Omega_{pt}^+ . \tag{3.41}$$

These weaker conditions allow in theory easier refinement procedures than for the classical strict acuteness: (3.40) allows the acute mesh angles to deteriorate (i.e. tend to 90°) as $h \to 0$, and (3.41) allows some right mesh angles, only requiring that the measure of elements with acute mesh angles must not asymptotically vanish.

3.3.5 Nonsymmetric systems with linear convection coefficients

Finally we consider systems including first order terms. First, we may include linear convection terms in each problem considered in the previous subsection. We only formulate this for the first problem. Thus we consider systems of the following form, with the boundary conditions of (3.24), where $k = 1,\dots,s$:

$$-\mathrm{div}\left(b_k(x,u,\nabla u)\,\nabla u_k\right) + \mathbf{w}_k(x)\cdot\nabla u_k + \sum_{l=1}^{s} V_{kl}(x,u,\nabla u)\,u_l \;=\; f_k(x). \tag{3.42}$$

Assumptions 3.8. The convection coefficients satisfy $\mathbf{w}_k \in W^{1,\infty}(\Omega)$, $\mathrm{div}\,\mathbf{w}_k \le 0$ on Ω and $\mathbf{w}_k \cdot \nu \ge 0$ on Γ_N $(k = 1,\dots,s)$. The domain Ω and the other coefficients satisfy Assumptions 3.5.

When considering a FEM discretization developed as in subsection 3.3.2, we need again a strengthened assumption for the quasi-regularity of the mesh such that (3.34) for γ is now replaced by

$$d \le \gamma < \frac{d(d+2)}{d+1}\,. \tag{3.43}$$

Theorem 3.8 *Let problem (3.42) satisfy Assumptions 3.8, and let assumptions of Theorem 3.5 hold except that the mesh quasi-regularity is understood with γ satisfying (3.43).*

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.

Hence, if $f_k \le 0,\ \gamma_k \le 0$ $(k = 1, \dots, s)$ and the basis functions satisfy (3.26)–(3.27), then for sufficiently small h the FEM solution of system (3.42), satisfes (3.36).

3.3.6 Nonsymmetric systems with nonlinear convection coefficients

Finally we study a system containing nonlinear convection terms. The required strengthening in the other assumptions is the strong uniform diagonal dominance (3.45) and the homogeneity of the Dirichlet data.

Let us consider the system (for $k = 1, \dots, s$):

$$\left.\begin{aligned} -\mathrm{div}\left(b_k(x, \nabla u)\,\nabla u_k\right) + \mathbf{w}_k(x,u)\cdot\nabla u_k + q_k(x, u_1, ..., u_s) \;=\; f_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x,\nabla u)\tfrac{\partial u_k}{\partial \nu} = \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \qquad u_k = 0 \quad \text{a.e. on } \Gamma_D\,. \end{aligned}\right\} \tag{3.44}$$

Assumptions 3.9. The convection coefficients satisfy $\mathbf{w}_k \in L^\infty(\Omega \times \mathbf{R})$. The domain Ω and the other coefficients satisfy Assumptions 3.7, except that item (v) in the latter is strengthened as follows: there exists $\mu > 0$ such that

$$\sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x,\xi) \ge \mu, \qquad \sum_{l=1}^{s} \frac{\partial q_l}{\partial \xi_k}(x,\xi) \ge \mu \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s), \tag{3.45}$$

moreover, $\mu > \|\mathbf{w}\|^2_{L^\infty(\Omega)^s}/4m$ where $m > 0$ is the lower bound of the b_k.

When considering a FEM discretization developed as in subsection 3.3.2, we need a strengthened assumption for the quasi-regularity of the mesh such that (3.34) for γ is replaced by

$$d \le \gamma < \min\left\{\gamma_d^*(p),\ \ \frac{d(d+2)}{d+1}\right\} \tag{3.46}$$

with p from (3.38) and $\gamma_d^*(p)$ from (3.39).

Theorem 3.9 *Let problem (3.44) satisfy Assumptions 3.9, and let the assumptions of Theorem 3.5 hold except that the mesh quasi-regularity is understood with γ satisfying (3.46).*

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.

Accordingly, if $f_k \ge q_k(x, 0)$ and $\gamma_k \ge 0$ $(k = 1, \dots, s)$, then for sufficiently small h, the FEM solution $u^h = (u_1^h, \dots, u_s^h)^T$ of system (3.44) satisfies

$$u_k^h \ge 0 \quad \text{on } \Omega \qquad (k = 1, \dots, s). \tag{3.47}$$

3.3.7 Some applications

The above obtained DMP result can be used for various models, mostly where discrete nonnegativity is required. Using the mesh conditions descibed in subsection 3.3.4, one may e.g. reproduce the following properties of the true solution:

- nonnegativity of the desired quantity in semilinear reaction-diffusion equations in chemistry (e.g. concentration in autocatalytic chemical reactions), biology (e.g. concentration in enzyme-substrate reaction by the Michaelis-Menten rule) and physics (e.g. temperature in radiative cooling).
- boundary maximum of the subsonic flow potential;
- nonnegativity of the concentration in chemical reaction-diffusion systems with cross-catalysis and autoinhibition;
- nonnegativity of the agents (pollutants) in transport type (diffusion-convection-reaction) processes.

4 A posteriori error estimates

4.1 A sharp global error estimate in Banach space

Based on [23], sharp error estimates are given for an operator equation

$$F(u) + l = 0 \tag{4.1}$$

in a Banach space V with a given nonlinear operator $F : V \to V^*$ and a given bounded linear functional $l \in V^*$. Later we will impose conditions ensuring that equation (4.1) has a unique solution $u^* \in V$.

In this section we consider some approximate solution $u \in V$ of equation (4.1), i.e. $u \approx u^*$ where u^* is the exact solution, and our goal is to estimate the error arising from this approximation. For this purpose, we will use the following (energy type) error functional for equation (4.1):

$$E(u) := \langle F(u) + l,\, u - u^* \rangle \;\equiv\; \langle F(u) - F(u^*),\, u - u^* \rangle \qquad (u \in V). \tag{4.2}$$

Since F will be assumed uniformly monotone, we have $E(u) \geq m\|u - u^*\|_V^2$, in particular $E(u) \geq 0 = E(u^*) \quad (u \in V)$.

Assumptions 4.1.

(i) Let V and Y be Banach spaces and $\Lambda : V \to Y$ a linear operator for which

$$\|\Lambda u\|_Y = \|u\|_V \qquad (u \in V). \tag{4.3}$$

(ii) The operator $A : Y \to Y^*$ has a bihemicontinuous symmetric Gateaux derivative (according to Definition 2.1).

(iii) There exists constants $M, m > 0$ such that

$$m \|p\|_Y^2 \leq \langle A'(y)p, p\rangle \leq M \|p\|_Y^2 \qquad (y, p \in Y). \tag{4.4}$$

(iv) The operator $F : V \to V^*$ has the form

$$\langle F(u), v\rangle = \langle A(\Lambda u), \Lambda v\rangle \qquad (u, v \in V).$$

(v) There exists a subspace $W \subset Y$ with a new norm $\|.\|_W$ such that A' is Lipschitz continuous as an operator from Y to $B(W, Y^*)$.

Theorem 4.1 *Let Assumptions 4.1 hold and $u^* \in V$ be the solution of (4.1). Let $u \in V$ be an approximation of u^* such that $\Lambda u \in W$. Then for arbitrary $z^* \in W$ and $k \in V$,*

$$E(u) \leq \; E\tilde{S}T(u; z^*, k) := \Big(m^{-1/2} \, |\Lambda^* A(z^*) + l| \; + \; \tfrac{L}{2} \, m^{-3/2} \, \tilde{D}(u; z^*, k) \tag{4.5}$$

$$+ \Big(\langle A(\Lambda u) - A(z^*), \; \Lambda u - z^*\rangle \; + \; \tfrac{L}{2m} \, \tilde{D}(u; z^*, k) \, \|\Lambda u - z^*\|_Y\Big)^{1/2}\Big)^2$$

where

$$\tilde{D}(u; z^*, k) := \Big(M \, \|z^* - \Lambda k\|_Y + \; |\Lambda^* A(z^*) + l|\Big) \, \|\Lambda u - z^*\|_W \, . \tag{4.6}$$

Proposition 4.1 *Estimate (4.5) is sharp in the following sense: if $\Lambda u^* \in W$ then*

$$\min_{\substack{z^* \in W, \\ k \in V}} E\tilde{S}T(u; z^*, k) = E(u).$$

When Y is a Hilbert space, one can find the optimal k for the above estimate via a kind of 'adjoint' equation: let k_{opt} be the solution of the problem

$$\langle \Lambda k_{opt}, \Lambda v\rangle = \langle z^*, \Lambda v\rangle \qquad (v \in V), \tag{4.7}$$

i.e., k_{opt} is the orthogonal projection of z^* on the range of Λ. Then for all $k \in V$ one has $\|z^* - \Lambda k_{opt}\|_Y \leq \|z^* - \Lambda k\|_Y$, i.e., k_{opt} provides the smallest value of $\|z^* - \Lambda k\|_Y$ in (4.6).

4.2 Applications to nonlinear elliptic problems

Let us consider a problem

$$\begin{cases} -\mathrm{div}\, f(x, \nabla u) = g \\ \quad u_{|\Gamma_D} = 0, \qquad f(x, \nabla u) \cdot \nu \, _{|\Gamma_N} = \gamma \end{cases} \tag{4.8}$$

as a special case of (2.52) under Assumptions 2.10. Then Theorem 4.1 holds with the following choices: $V := H_0^1(\Omega)$, $Y := L^2(\Omega)^d$, $W := L^\infty(\Omega)^d$ and the operator $\Lambda := \nabla$.

One can determine the optimal y^* and w in $EST(u_h; y^*, w)$ in the following way. First, the optimal value of the parameter z^* should be a sufficiently accurate approximation of ∇u^*. For finite element solutions, a common way is to use an averaging procedure, i.e.,

to replace the unknown gradient ∇u^* of the exact solution by $z^* := G_h(\nabla u_h)$, where G_h is some averaging operator: for piecewise linear finite elements, $G_h(\nabla u_h)$ is closer to ∇u^* than is ∇u_h by an order of magnitude. Next, the optimal k for this z^* is given as the solution of the 'adjoint' problem (4.7), which now amounts to finding $k_{opt} \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla k_{opt} \cdot \nabla v = \int_\Omega z^* \cdot \nabla v \qquad (v \in H_0^1(\Omega)), \tag{4.9}$$

that is, the weak solution of a Poisson problem. The latter is linear, hence its numerical solution costs much less than for the original one. When obtained from a piecewise linear FEM, its right-hand side is constant on each element, hence it requires minimal numerical integration and is therefore a cheap auxiliary problem. Now using a finer mesh for (4.9) than the one used for u_h may considerably increase the accuracy of the estimate.

Remark 4.1 One can apply the error estimate in the same way for problems with the same structure as (4.8).

(i) For fourth order problems

$$\mathrm{div}^2 B(x, D^2 u) \;=\; g, \qquad u_{|\partial\Omega} = \tfrac{\partial u}{\partial \nu}\big|_{\partial\Omega} = 0$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, defined via a matrix-valued nonlinearity B with analogous properties to f in (4.8), Theorem 4.1 holds with the following choices: the spaces $V := H_0^2(\Omega)$, $Y := L^2(\Omega)^{d\times d}$, $W := L^\infty(\Omega)^{d\times d}$ and the Hessian operator $\Lambda := D^2$.

(ii) For the second order elasticity system (2.61), Theorem 4.1 holds with the spaces $V := H_D^1(\Omega)^3$, $Y := L^2(\Omega)^{3\times 3}_{symm}$ (symmetric matrix-valued functions with entries in $L^2(\Omega)$), $W := L^\infty(\Omega)^{3\times 3}_{symm}$ and the operator $\Lambda := \varepsilon$ where $\varepsilon(u) := \frac{1}{2}(\nabla u + \nabla u^t)$.

References

[1] Allgower, E.L., Böhmer, K., Potra, F.A., Rheinboldt, W.C., A mesh-independence principle for operator equations and their discretizations, *SIAM J. Numer. Anal.* 23 (1986), no. 1, 160–169.

[2] Antal I., Karátson J., A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems, *Comput. Math. Appl.* 55 (2008), 2185-2196.

[3] Antal I., Karátson J., Mesh independent superlinear convergence of an inner-outer iterative method for semilinear elliptic interface problems, *J. Comp. Appl. Math.* 226 (2009), 190-196.

[4] Axelsson, O., A generalized conjugate gradient least square method, *Numer. Math.* 51 (1987), 209-227.

[5] Axelsson, O., *Iterative Solution Methods,* Cambridge University Press, 1994.

[6] Axelsson, O., Barker, V. A., Neytcheva, M., Polman, B., Solving the Stokes problem on a massively parallel computer, *Math. Model. Anal.* 6 (2001), no. 1, 7–27.

[7] Axelsson, O., Faragó I., Karátson J., Sobolev space preconditioning for Newton's method using domain decomposition, *Numer. Lin. Alg. Appl.*, **9** (2002), 585-598.

[8] Axelsson, O., Gololobov, S. V., A combined method of local Green's functions and central difference method for singularly perturbed convection-diffusion problems, *J. Comput. Appl. Math.* 161 (2003), no. 2, 245–257.

[9] Axelsson, O., Karátson J., Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators, *Numer. Math.* 99 (2004), No. 2, 197-223.

[10] Axelsson, O., Karátson J., Mesh independent superlinear PCG rates via compact-equivalent operators, *SIAM J. Numer. Anal.*, 45 (2007), No.4, pp. 1495-1516

[11] Axelsson, O., Karátson J., Equivalent operator preconditioning for elliptic problems, *Numer. Algorithms*, 50 (2009), Issue 3, p. 297-380.

[12] Ciarlet, P. G., Discrete maximum principle for finite-difference operators, *Aequationes Math.* 4 (1970), 338–352.

[13] Erlangga, Y. A., Advances in iterative methods and preconditioners for the Helmholtz equation, *Arch. Comput. Methods Eng.* (2008) 15: 3766.

[14] Faber, V., Manteuffel, T., Parter, S.V., On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations, *Adv. in Appl. Math.,* 11 (1990), 109-163.

[15] Faragó, I., Karátson, J., *Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators. Theory and Applications.* Advances in Computation, Volume 11, NOVA Science Publishers, New York, 2002.

[16] Gohberg, I., Goldberg, S., Kaashoek, M. A., *Classes of Linear Operators*, Vol. I., Operator Theory: Advances and Applications, 49, Birkhäuser Verlag, Basel, 1990.

[17] Goldstein, C. I., Manteuffel, T. A., Parter, S. V., Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular values, *SIAM J. Numer. Anal.* 30 (1993), no. 2, 343–376.

[18] Křižek, M., Neittaanmäki, P., *Mathematical and numerical modelling in electrical engineering: theory and applications*, Kluwer Academic Publishers, 1996.

[19] Karátson J., Gradient method in Sobolev space for nonlocal boundary value problems, *Electron. J. Diff. Eqns.*, Vol. 2000 (2000), No. 51, pp. 1-17.

[20] Karátson J., Constructive Sobolev gradient preconditioning for semilinear elliptic systems, *Electron. J. Diff. Eqns.* Vol. 2004(2004), No. 75, pp. 1-26.

[21] Karátson J., Superlinear PCG algorithms: symmetric part preconditioning and boundary conditions, *Numer. Funct. Anal.* 29 (2008), No. 5-6, pp. 1-22.

[22] Karátson J., Faragó I., Variable preconditioning via quasi-Newton methods for nonlinear problems in Hilbert space, *SIAM J. Numer. Anal.* **41** (2003), No. 4, 1242-1262.

[23] Karátson J., Korotov, S., Sharp upper global a posteriori error estimates for nonlinear elliptic variational problems, *Appl. Math.* (Prague), 54 (2009), No. 4, pp. 297-336.

[24] Karátson J., Korotov, S., Discrete maximum principles for FEM solutions of nonlinear elliptic systems, in: *Computational Mathematics: Theory, Methods and Applications*, NOVA Science Publishers, New York, 2010; pp. 213-260.

[25] Karátson J., Kurics T., Superlinearly convergent PCG algorithms for some nonsymmetric elliptic systems, *J. Comp. Appl. Math.* 212 (2008), No. 2, pp. 214-230.

[26] Karátson J., Kurics T., Lirkov, I., A parallel algorithm for systems of convection-diffusion equations, in: *NMA 2006*, eds. T. Boyanov et al., *Lecture Notes Comp. Sci.* 4310, pp. 65-73, Springer, 2007.

[27] Karátson J., Lóczi L., Sobolev gradient preconditioning for the electrostatic potential equation, *Comput. Math. Appl.* 50 (2005), pp. 1093-1104.

[28] Kovács B., Comparison of efficient numerical methods for an elliptic problem (in Hung.), TDK report, ELTE University, Budapest, 2010.

[29] Kurics T., Operator preconditioning in Hilbert space, PhD dissertation, ELTE University, Budapest, 2010.

[30] Manteuffel, T., Otto, J., Optimal equivalent preconditioners, *SIAM J. Numer. Anal.,* 30 (1993), 790-812.

[31] Manteuffel, T., Parter, S. V., Preconditioning and boundary conditions, *SIAM J. Numer. Anal.* 27 (1990), no. 3, 656–694.

[32] Martinsson, P.G., A fast direct solver for a class of elliptic partial differential equations, *J. Sci. Comput.* (2009) 38: 316330.

[33] Meurant, G., *Computer solution of large linear systems,* North-Holland, 1999.

[34] Neuberger, J. W., *Sobolev Gradients and Differential Equations*, Lecture Notes in Math., No. 1670, Springer, 1997.

[35] Neuberger, J. W., Prospects for a central theory of partial differential equations, *Math. Intell.* 27, No. 3, 47-55 (2005).

[36] Neuberger, J. W., Renka R.J., Sobolev gradients and the Ginzburg-Landau equations, *SIAM J. Sci. Comput.* 20 (1998), 582-590.

[37] Neuberger, J. W., Renka R.J., Sobolev gradients: Introduction, applications, problems, *Contemporary Mathematics* 257 (2004), 85-99.

[38] Richardson, W.B., Sobolev gradient preconditioning for image-processing PDEs, *Commun. Numer. Methods Eng.* 24, No. 6, 493-504 (2008).

[39] Samaey, G., Vanroose, W., An analysis of equivalent operator preconditioning for equation-free Newton-Krylov methods, *SIAM J. Numer. Anal.* 48: (2) 633-658 (2010).

[40] Szabó, B., Babuška, I., *Finite Element Analysis*, J.Wiley and Sons, 1991

[41] Trottenberg U., Oosterlee C.W., Schueller A., *Multigrid*, Academic Press, 2001

[42] Weiser, M.D., Schiela, A., Deuflhard, P., Asymptotic mesh independence of Newton's method revisited, *SIAM J. Numer. Anal.* 42 (2005), no. 5, 1830–1845.

[43] Zlatev, Z., *Computer Treatment of Large Air Pollution Models*, Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.