Operator methods for the numerical solution of elliptic PDE problems

D.Sc. dissertation

Karátson János

ELTE, Budapest, 2011

Contents

Introduction

The field of this dissertation 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 (the minimally necessary) order $O(n)$, where n denotes the degrees of freedom in the algebraic system. (One can in fact also do with quasi-optimality, usually of the form $O(n \log n)$.) 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 (owing to the optimality property).

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.

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 to model and real-life problems.

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.

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 under 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 certain real-life problems.

- Operator approach is used to derive results on the reliability of the numerical solution. First, a discrete maximum principle (DMP) is established in Hilbert space for proper Galerkin stiffness matrices. Then we prove DMPs for general nonlinear elliptic equations with mixed boundary conditions, and further, for several types of 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 various classes of nonlinear elliptic problems.

Part I

Iterative methods based on operator preconditioning

Chapter 1

Linear problems

1.1 Preliminaries

In this chapter we study the numerical solution of a linear operator equation

$$Lu = g \tag{1.1.1}$$

(in a Hilbert space) that will then model an elliptic PDE including boundary conditions. A Galerkin (resp. FEM) discretization yields a finite dimensional problem

$$L_h u_h = g_h. \tag{1.1.2}$$

First we briefly summarize some basic ideas from previous work that are important for our investigation.

1.1.1 Basic ideas

(a) Preconditioning using auxiliary operators

Linear elliptic partial differential equations (PDEs) are usually solved numerically using the finite element or finite difference method. Since the arising linear algebraic systems are large and sparse, they are normally solved by iteration, most commonly using a preconditioned conjugate gradient (PCG) method (see subsection 1.1.2). For special types of problems, however, there exist particular methods (such as FFT or FACR for problems with constant coefficients [114, 137, 149], or multigrid/multilevel methods for more general single symmetric equations – possibly with scalar diffusion coefficients – [69, 115]) that have the *optimality or quasi-optimality* property. This means that the computational cost is of the minimally necessary order $O(n)$ or (practically being very close to that) $O(n \log n)$, respectively, where n denotes the degrees of freedom in the algebraic system. The basic idea is that such special discrete systems can then be used as preconditioners to more general problems. This leads to the following general framework to construct preconditioners.

Instead of constructing the preconditioner directly for the given finite element (FE) or finite difference (FD) matrix, it can be more efficient to first approximate the given differential operator by some simpler differential operator, and then to use the FE or FD matrix of this operator as preconditioner, hereby using the same discretization mesh as for the original operator. Formally, to solve (1.1.2), one can take *another elliptic operator* S, in some way related to L, and propose its discretization S_h as preconditioner for (1.1.2):

$$S_h^{-1} L_h u_h = S_h^{-1} g_h\,. \tag{1.1.3}$$

Then a CG iteration involves stepwise formal multiplications with $S_h^{-1}L_h$, which in fact requires the solution of systems with S_h.

It is historically important to mention the discrete Laplacian as the first application of the equivalent operator idea for discretized elliptic problems. The Laplacian as preconditioner was first introduced in an infinite-dimensional setting by László Czách for steepest descent in his CSc. thesis [39] supervised by Kantorovich, also quoted in [79]. Then the centered finite difference discretization of an elliptic problem with scalar diffusion was studied on a rectangle [43, 68], and the Laplacian preconditioning for simple iteration was later termed as D'yakonov-Gunn iteration. Various modifications of the D'yakonov-Gunn iteration have then been given, including preconditioners resulting from scaled Laplacians, separable operators or symmetric part etc., see e.g. [25, 36, 49, 76, 129, 154], and [19] for a survey. A discrete Laplacian as preconditioner also appears in Uzawa type iterations for saddle-point problems, see e.g. [47, 141].

To obtain favourable preconditioners, one must satisfy the two well-known basic requirements for the preconditioning matrix [8]. First, solving problems with S_h should be considerably simpler than those with L_h. This clearly holds in the ideal case for the mentioned optimal or quasi-optimal solvers. More generally, one still obtains efficient preconditioners if, in contrast to L, the operator S is symmetric (or, more generally, incorporates parts of the given operator that can be solved far more easily than that); if S_h is an M-matrix or is diagonally dominant; if S_h has a favourable block structure, or if S_h has a better sparsity pattern.

On the other hand, the conditioning of $S_h^{-1}L_h$ should be considerably better than the conditioning of L_h. Here one is mostly interested in *mesh independence*, i.e. that the number of iterations to achieve prescribed accuracy should be bounded independently of n. This is a crucial property of the iteration, since one preserves in this way the optimality for the overall iteration: if, to prescribed accuracy, systems with S_h are solved with $O(n)$ operations, and one applies such solvers mesh-independently many times, then the original system is also solved with $O(n)$ operations.

The above fact shows that the theoretical study of mesh independence leads to the very practical result of constructing optimal overall iterative solvers.

(b) Concepts of equivalent operators

For the general study of mesh independent linear convergence, a natural framework to describe the related preconditioning properties is that of equivalent operators, developed rigorously by T. Manteuffel et al. in [52], see also [66, 111, 112] and the references therein to earlier applications. Under proper assumptions, roughly speaking, the condition number $\kappa(S_h^{-1}L_h)$ approaches $\kappa(S^{-1}L)$ as $h \to 0$, and hence it is bounded as $h \to 0$, in contrast to $\kappa(L_h)$ which tends to ∞. Moreover, for FEM discretizations we usually have $\kappa(S_h^{-1}L_h) \leq \kappa(S^{-1}L)$.

Briefly, if the two operators (the original and preconditioner) are equivalent then the corresponding PCG method provides mesh independent linear convergence.

We briefly outline some notions and related results from their work. Let $B : W \to V$ and $A : W \to V$ be linear operators between the Hilbert spaces W and V. For our purposes it suffices to consider the case when B and A are one-to-one and $D = D(A) \cap D(B)$ is dense. The operator A is said to be equivalent in V-norm to B on D if there exist constants $K \geq k > 0$ such that

$$k \leq \frac{\|Au\|_V}{\|Bu\|_V} \leq K \qquad (u \in D \setminus \{0\}). \tag{1.1.4}$$

If (1.1.4) holds, then under suitable density assumptions on D, the condition number of AB^{-1} in V is bounded by K/k. The W-norm equivalence of B^{-1} and A^{-1} implies this bound similarly for $B^{-1}A$.

The analogous property for the discretized problems is uniform norm equivalence defined as follows. The families of operators A_h and B_h (indexed by $h > 0$) are said to be V-norm uniformly equivalent if there exist constants $\tilde{K} \geq \tilde{k} > 0$, independent of h, such that

$$\tilde{k} \leq \frac{\|A_h u\|_V}{\|B_h u\|_V} \leq \tilde{K} \qquad (u \in D \setminus \{0\},\ h > 0). \tag{1.1.5}$$

Analogously to the above, this implies that the condition numbers of the family $A_h B_h^{-1}$ are bounded uniformly in h, and the similar uniform equivalence of B_h^{-1} and A_h^{-1} implies that the condition numbers of the family $B_h^{-1} A_h$ are bounded uniformly in h.

Using the above notions, the following general results hold. First, the V-norm equivalence of A and B is necessary for the V-norm uniform equivalence of the families A_h and B_h. Second, the former is also sufficient for the latter if the families A_h and B_h are obtained via orthogonal projections from A and B and, further, if A and B are equivalent to the families A_h and B_h. For details and various special and related cases see [52, Chap. 2].

The above setting is mostly intended to handle L_2-norm equivalence for elliptic operators. However, it is often more convenient to use H^1-norm equivalence [52, 112] based on a weak formulation, since this helps to avoid regularity requirements. The notion of H^1-norm equivalence is based on the weak form of elliptic operators as follows, see [112] for details. In a standard way, using Green's formula, one can define the bilinear form $a(.,.)$ corresponding to an elliptic operator A on a subspace $H^1_D(\Omega)$ of $H^1(\Omega)$ (associated with the boundary conditions), and this form satisfies $a(u,v) = \langle Au, v\rangle_{L^2}$ for $u, v \in D(A)$. The bounded bilinear form a gives rise to an operator A_w from $H^1_D(\Omega)$ into its dual satisfying $A_w u(v) = a(u,v)$. We note that the dual of $H^1_D(\Omega)$ can be identified with $H^1_D(\Omega)$ itself by the Riesz theorem, which will be convenient for our purposes as we can consider A_w as mapping into $H^1_D(\Omega)$ and satisfying

$$\langle A_w u, v\rangle_{H^1_D} = \langle Au, v\rangle_{L^2} \qquad (u, v \in D(A)). \tag{1.1.6}$$

The basic result on H^1-norm equivalence in [112] reads as follows: if A and B are invertible uniformly elliptic operators, then A_w^{-1} and B_w^{-1} are H^1-norm equivalent if and only if A and B have homogeneous Dirichlet boundary conditions on the same portion of the boundary.

In the sequel we will build on the above result in the sense that we will develop a simpler Hilbert space setting of equivalent operators a priori suited for invertible elliptic operators with identical Dirichlet boundary.

1.1.2 Conjugate gradient algorithms

As mentioned before, the most widespread iterative method to solve discretized linear elliptic problems is the conjugate gradient (CG) method, normally applied to a preconditioned form like (1.1.3). We briefly summarize some required well-known facts about the convergence of the main CG algorithms, see, e.g. [8, 45] or, for a brief summary, [19, Chap. 2]. The algorithms themselves are also described in these works.

Let us consider a linear algebraic system

$$Au = b \tag{1.1.7}$$

with a given nonsingular matrix $A \in \mathbf{R}^{n\times n}$. Letting $\langle .,. \rangle$ be a given inner product on $\mathbf{R}^n$, assume that A is positive definite w.r.t. $\langle .,. \rangle$. We define the following quantities:

$$\lambda_0 := \lambda_0(A) := \inf\{\langle Ax, x\rangle : \ \|x\| = 1\} > 0, \qquad \Lambda := \Lambda(A) := \|A\|, \tag{1.1.8}$$

where $\|.\|$ denotes the norm induced by the inner product $\langle .,. \rangle$.

If A is self-adjoint w.r.t. $\langle .,. \rangle$, then $\lambda_0(A) = \lambda_{min}(A)$, $\Lambda(A) = \lambda_{max}(A)$, and the standard CG method provides the linear convergence estimate

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2^{1/k}\,\frac{\sqrt{\Lambda}-\sqrt{\lambda_0}}{\sqrt{\Lambda}+\sqrt{\lambda_0}} = 2^{1/k}\,\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \qquad (k = 1, 2, ..., n), \tag{1.1.9}$$

where $\kappa(A) = \Lambda/\lambda_0$ is the standard condition number and $e_k := u - u_k$ are the error vectors. In the superlinear phase of the convergence history, one normally uses the following estimate: writing the decomposition $A = \mu I + E$ for some $\mu > 0$,

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2\|A^{-1}\|}{k}\ \sum_{j=1}^{k}\bigl|\lambda_j(E)\bigr| \qquad (k = 1, 2, ..., n). \tag{1.1.10}$$

Another approach, based on the K-condition number provides similar estimates. One often lets $\mu = 1$ without loss of generality, e.g. for symmetric part preconditioning.

For nonsymmetric matrices A, several CG algorithms exist such as the widely used GMRES and its variants. A method in general form is the GCG-LS (generalized conjugate gradient–least square) method, which provides

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \left(1 - \left(\frac{\lambda_0}{\Lambda}\right)^2\right)^{1/2} \qquad (k = 1, 2, ..., n), \tag{1.1.11}$$

where $r_k := Au_k - b$. The same estimate holds for the GCR and Orthomin methods together with their truncated versions. If A is normal, then (1.1.10) also holds for $(\|r_k\|/\|r_0\|)^{1/k}$.

Another common way to solve (1.1.7) with nonsymmetric A is the CGN method ('conjugate gradients for the normal equation'), i.e. to consider the normal equation $A^*Au = A^*b$

and apply the symmetric CG algorithm for the latter. (Here A^* is the adjoint of A w.r.t. the given inner product.) This yields the linear convergence estimate

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le 2^{1/k}\,\frac{\Lambda-\lambda_0}{\Lambda+\lambda_0} \qquad (k=1,2,...,n), \tag{1.1.12}$$

and, having the decomposition $A = I + E$, the superlinear rate

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2\|A^{-1}\|^2}{k}\sum_{i=1}^{k}\Big(\big|\lambda_i(E^*+E)\big| + \lambda_i(E^*E)\Big) \qquad (k=1,2,...,n). \tag{1.1.13}$$

Finally, using $\|A^{-1}\| \le 1/\lambda_0$, the estimates (1.1.10) and (1.1.13) become

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \le \frac{2}{k\lambda_0}\sum_{j=1}^{k}\big|\lambda_j(E)\big|, \qquad \left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2}{k\lambda_0^2}\sum_{i=1}^{k}\Big(\big|\lambda_i(E^*+E)\big| + \lambda_i(E^*E)\Big). \tag{1.1.14}$$

1.2 Compact-equivalent operators and superlinear convergence

In this section we develop our contribution that completes the mentioned results of Manteuffel et al. on linear convergence. As a motivation, recall that the convergence history of a CG iteration for a discretized elliptic problem usually consists of two pronounced phases: first a *linear* and then a *superlinear* phase of convergence takes place, see e.g. [8, 13]. This is shown on a logarithmic scale in Figure 1.1. 'Superlinear' means a fast convergence phase when the relative error decays faster than any geometric sequence, which is a desirable property when an increased accuracy is required. (Roughly speaking, each additional correct digit in the approximate solution then requires fewer iterations than the previous digit.)

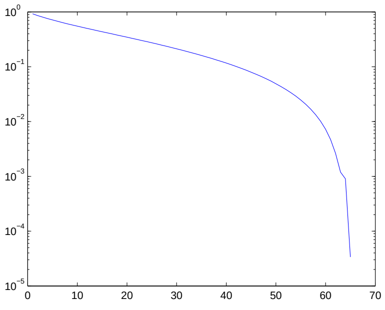

Figure 1.1: The convergence history of a CG iteration for a discretized elliptic problem

In the context of mesh independent convergence, the first (linear) phase has been properly handled by the equivalent operator theory: if the two operators (the original and preconditioner) are equivalent, then the preconditioned CGN method provides mesh independent linear convergence [52]. This raises the question how to approach the mesh independence theory of superlinear convergence.

In this section 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. As the counterpart of the results of Manteuffel et al, we 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.*

We also 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. This will enable us to derive mesh independent superlinear convergence for discretized elliptic problems such that the first and zeroth order terms are chosen freely, and we can treat both symmetric and nonsymmetric problems, both equations and systems.

The description is based on our following papers: mesh independence of superlinear convergence has first been established in special cases in [16], the compact-equivalent operator framework has been developed in [18] and further applied in [19].

1.2.1 S-bounded and S-coercive operators

The notion of compact-equivalent operators needs a preliminary notion of weak form of unbounded operators. To describe this weak form, we first develop the concept of S-bounded and S-coercive operators.

This concept is useful in other respects too. First, it provides a proper setting to define the weak solution of an operator equation when the coercive operator is nonsymmetric (and thus has no energy space itself), i.e. we can thus clarify in which space equation (1.1.1) is well-posed. Further, it will also help us to give a simplified general framework for mesh independent linear convergence in the next chapter.

(a) The Hilbert space framework

Let H be a real Hilbert space. We are interested in solving the operator equation (1.1.1). To this end, we recast the required properties of L to the energy space of a suitable auxiliary operator S, which is an (also unbounded) linear symmetric operator in H and assumed to be coercive, i.e., there exists $p > 0$ such that $\langle Su, u\rangle \geq p\|u\|^2 \ \ (u \in D(S))$.

We 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 1.2.1 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 $|\langle Lu, v\rangle| \leq M\|u\|_S\|v\|_S \quad (u, v \in D(L))$;

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

Definition 1.2.2 For any $L \in BC_S(H)$, let $L_S \in B(H_S)$ be defined by

$$\langle L_S u, v\rangle_S = \langle Lu, v\rangle \qquad (u, v \in D(L)). \tag{1.2.1}$$

Remark 1.2.1 (a) The above definition makes sense since L_S is the bounded linear operator on H_S that represents the unique extension to H_S of the densely defined S-bounded bilinear form $u, v \mapsto \langle Lu, v\rangle$.

(b) The density of $D(L)$ implies

$$|\langle L_S u, v\rangle_S| \leq M\|u\|_S\|v\|_S, \qquad \langle L_S u, u\rangle_S \geq m\|u\|_S^2 \qquad (u, v \in H_S). \tag{1.2.2}$$

Our setting leads to equivalent operators in the sense of Manteuffel et al.:

Proposition 1.2.1 *Let N and L be S-bounded and S-coercive operators for the same S. Then*
(a) N_S and L_S are H_S-norm equivalent,
(b) N_S^{-1} and L_S^{-1} are H_S-norm equivalent.

Proof. (a) By (1.1.4), we must find $K \geq k > 0$ such that

$$k\|N_S u\|_S \leq \|L_S u\|_S \leq K\|N_S u\|_S \qquad (u \in H_S). \tag{1.2.3}$$

Since $L \in BC_S(H)$, there exists constants $M_L \geq m_L > 0$ such that for all $u \in H_S$,

$$m_L\|u\|_S \leq \frac{\langle L_S u, u\rangle_S}{\|u\|_S} \leq \|L_S u\|_S = \sup_{v \in H_S \setminus \mathbf{0}} \frac{\langle L_S u, v\rangle_S}{\|v\|_S} \leq M_L\|u\|_S \tag{1.2.4}$$

and the analogous estimate holds for N with some $M_N \geq m_N > 0$. The two estimates yield (1.2.3) with $K = \frac{M_L}{m_N}$ and $k = \frac{m_L}{M_N}$.

(b) Properties (1.2.2) imply that L_S is invertible in $B(H_S)$, hence for all $v \in H_S$ we can set $u = L_S^{-1}v$ in (1.2.4) to obtain

$$m_L\|L_S^{-1}v\|_S \leq \|v\|_S \leq M_L\|L_S^{-1}v\|_S \qquad (v \in H_S).$$

This and its analogue for N yield the required estimate similarly as in (a), now with $K = \frac{M_N}{m_L}$ and $k = \frac{m_N}{M_L}$. ■

Let us now return to the operator equation (1.1.1) for $L \in BC_S(H)$.

Definition 1.2.3 For given $L \in BC_S(H)$, we call $u \in H_S$ the *weak solution* of equation (1.1.1) if

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \tag{1.2.5}$$

For all $g \in H$ the weak solution of (1.1.1) exists and is unique, which follows in a standard way from the Lax-Milgram lemma.

(b) Coercive elliptic operators

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

$$Lu \equiv\ -\text{div}\,(A\,\nabla u) +\ \mathbf{b}\cdot\nabla u + cu \qquad \text{for}\ \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \tag{1.2.6}$$

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 1.2.1, 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 1.2.1

(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

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

$$Su \equiv\ -\text{div}\,(G\,\nabla u) + \sigma u \qquad \text{for}\ \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{1.2.7}$$

which satisfies

Assumptions 1.2.2

(i) Substituting G for A, Ω, Γ_D, Γ_N and G satisfy Assumptions 1.2.1;

(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.

Here the energy space H_S of the operator S is in fact

$$H^1_D(\Omega) := \{u \in H^1(\Omega) :\ u_{|\Gamma_D} = 0\} \quad \text{with}\ \ \langle u, v\rangle_S := \int_\Omega (G\ \nabla u\cdot\nabla v + \sigma uv) + \int_{\Gamma_N} \beta uv\, d\sigma\,. \tag{1.2.8}$$

Proposition 1.2.2 *If Assumptions 1.2.1-2 hold, then the operator L is S-bounded and S-coercive in $L^2(\Omega)$, i.e., $L \in BC_S(L^2(\Omega))$.*

Proof. We must verify the properties in Definition 1.2.1 from the above assumptions. The domain of definition of L is $D(L) := \{u \in H^2(\Omega) : \ Lu \in L^2(\Omega), \ u_{|\Gamma_D} = 0, \ \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0\,\}$ in the Hilbert space $L^2(\Omega)$, so $D(L) \subset H_S = H^1_D(\Omega)$ and $D(L)$ is dense in $H^1_D(\Omega)$ in the S-inner product (1.2.8). Further, for $u, v \in D(L)$, by Green's formula, we have

$$\langle Lu, v\rangle_{L^2(\Omega)} = \int_\Omega \Big(A\,\nabla u\cdot\nabla v + (\mathbf{b}\cdot\nabla u)v + cuv\Big) \ + \int_{\Gamma_N} \alpha uv\, d\sigma\,. \tag{1.2.9}$$

Using this and (1.2.8), one can check properties (ii)-(iii) of Definition 1.2.1 with a standard calculation as follows. First, Assumptions 1.2.2 imply that the S-norm related to (1.2.8) is equivalent to the usual H^1-norm, and accordingly, there exist embedding constants $C_{\Omega,S} > 0$ and $C_{\Gamma_N,S} > 0$ such that

$$\|u\|_{L^2(\Omega)} \le C_{\Omega,S}\|u\|_S \quad \text{and} \quad \|u\|_{L^2(\Gamma_N)} \le C_{\Gamma_N,S}\|u\|_S \qquad (u \in H^1_D(\Omega)), \tag{1.2.10}$$

see, e.g., [148]. Further, the uniform spectral bounds of A and G also imply the existence of constants $p_1 \ge p_0 > 0$ such that

$$p_0\ (G(x)\xi\cdot\xi) \le A(x)\xi\cdot\xi \le p_1\ (G(x)\xi\cdot\xi) \qquad (x \in \overline{\Omega},\ \xi \in \mathbf{R}^d), \tag{1.2.11}$$

and there exists $q > 0$ such that

$$q\,\|\nabla u\|^2_{L^2(\Omega)} \le \int_\Omega G\,\nabla u\cdot\nabla u \le \|u\|^2_S \qquad (u \in H^1_D(\Omega)). \tag{1.2.12}$$

Then from (1.2.9) we obtain

$$\begin{aligned}\langle Lu, v\rangle \ &\le \ p_1\|u\|_S\|v\|_S + \|\mathbf{b}\|_{L^\infty(\Omega)^d}\|\nabla u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\ &\qquad + \|c\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\alpha\|_{L^\infty(\Gamma_N)}\|u\|_{L^2(\Gamma_N)}\|v\|_{L^2(\Gamma_N)} \\ &\le \Big(p_1 + C_{\Omega,S}\,q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C^2_{\Omega,S}\|c\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,S}\|\alpha\|_{L^\infty(\Gamma_N)}\Big)\ \|u\|_S\|v\|_S\,. \end{aligned} \tag{1.2.13}$$

On the other hand, for any $u \in H^1_D(\Omega)$, using the definition of $\hat{c}$ and $\hat{\alpha}$ from Assumptions 1.2.1 (iii), a standard calculation with Green's formula yields (see, e.g., [85]) that

$$\langle Lu, u\rangle_{L^2(\Omega)} = \int_\Omega (A\ \nabla u\cdot\nabla u + \hat{c}u^2) + \int_{\Gamma_N} \hat{\alpha}u^2\, d\sigma \ =: \|u\|^2_L\,. \tag{1.2.14}$$

Assumptions 1.2.1 imply that the L-norm, defined above on the right, is equivalent to the usual H^1-norm, hence there exist constants $C_{\Omega,L} > 0$ and $C_{\Gamma_N,L} > 0$ such that the analogue of (1.2.10) holds for the L-norm instead of the S-norm. Therefore

$$\begin{aligned}\|u\|^2_S &= \int_\Omega (G\ \nabla u\cdot\nabla u + \sigma u^2) + \int_{\Gamma_N} \beta u^2\, d\sigma \\ &\le p_0^{-1}\int_\Omega\ A\ \nabla u\cdot\nabla u + \|\sigma\|_{L^\infty(\Omega)}\int_\Omega u^2 + \ \|\beta\|_{L^\infty(\Gamma_N)}\int_{\Gamma_N} u^2\, d\sigma \\ &\le \Big(p_0^{-1} + C^2_{\Omega,L}\|\sigma\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,L}\|\beta\|_{L^\infty(\Gamma_N)}\Big)\ \langle Lu, u\rangle_{L^2(\Omega)} \qquad (u \in H^1_D(\Omega)). \end{aligned} \tag{1.2.15}$$

Summing up, estimates (1.2.13) and (1.2.15) yield that properties (ii)-(iii) of Definition 1.2.1 are valid with

$$\begin{aligned} M &:= p_1 + C_{\Omega,S}\, q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C_{\Omega,S}^2\|c\|_{L^\infty(\Omega)} + C_{\Gamma_N,S}^2\|\alpha\|_{L^\infty(\Gamma_N)} \ , \\ m &:= \Big(p_0^{-1} + C_{\Omega,L}^2\|\sigma\|_{L^\infty(\Omega)} + C_{\Gamma_N,L}^2\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1} . \qquad \blacksquare \end{aligned} \tag{1.2.16}$$

Remark 1.2.2 The constants $C_{\Omega,S}$ and $C_{\Gamma_N,S}$ in (1.2.16) can be calculated as follows. (The same holds for $C_{\Omega,L}$ and $C_{\Gamma_N,L}$.)

In order to find $C_{\Omega,S}$, first let $\Gamma_D \neq \emptyset$. Then it suffices to determine $C_\Omega > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq C_\Omega \|\nabla u\|_{L^2(\Omega)} \qquad (u \in H^1_D(\Omega)), \tag{1.2.17}$$

in which case $C_{\Omega,S} = q^{-1/2}C_\Omega$ from (1.2.12). Here such a C_Ω exists because for $\Gamma_D \neq \emptyset$, the usual H^1-norm is equivalent to $\|\nabla u\|^2_{L^2(\Omega)}$. Its sharp value satisfies $C_\Omega = \lambda_1^{-1/2}$, where λ_1 is the smallest eigenvalue of $-\Delta$ under boundary conditions $u_{|\Gamma_D} = 0$, $\frac{\partial u}{\partial \nu}_{|\Gamma_N} = 0$. For Dirichlet boundary conditions, one can use the estimate

$$C_\Omega \leq \Big(\sum_{i=1}^d \big(\frac{\pi}{a_i}\big)^2\Big)^{-1/2}$$

if Ω is embedded in a brick with edges $a_1, \dots, a_d$, see, e.g., [118]. If $\Gamma_D = \emptyset$ then similarly as above, $C_{\Omega,S} \leq p_0^{-1/2}\hat{C}_\Omega$, where $\hat{C}_\Omega$ is the smallest eigenvalue of the operator $-\Delta u + (\sigma_0/p_0)u$ under boundary conditions $u_{|\Gamma_D} = 0$, $\frac{\partial u}{\partial \nu} + (\beta_0/p_0)_{|\Gamma_N} = 0$, in which $\sigma_0 := \inf \sigma$ and $\beta_0 := \inf \beta$. Here it is advisable to choose σ to satisfy $\sigma_0 > 0$, in which case $\|u\|^2_{L^2(\Omega)} \leq \sigma_0^{-1}\int_\Omega \sigma u^2 \leq \sigma_0^{-1}\|u\|_S^2$, i.e. $C_{\Omega,S} \leq \sigma_0^{-1/2}$.

For $C_{\Gamma_N,S}$, one should first find $C_{\Gamma_N} > 0$ such that

$$\|u\|_{L^2(\Gamma_N)} \leq C_{\Gamma_N}\|u\|_{H^1(\Omega)} \qquad (u \in H^1_D(\Omega)),$$

in which case $C_{\Gamma_N,S} = \big(1 + C_\Omega^2\big)^{1/2} q^{-1/2}\, C_{\Gamma_N}$ from (1.2.12) and (1.2.17). For polygonal domains in 2D, explicit estimates for C_{Γ_N} are given in [134].

1.2.2 Compact-equivalent operators

(a) The notion of compact-equivalent operators

In this section we involve compact operators in Hilbert space, i.e., linear operators C such that the image (Cv_i) of any bounded sequence (v_i) contains a convergent subsequence. Recall that the eigenvalues of a compact self-adjoint operator cluster at the origin.

Definition 1.2.4 (i) We call $\lambda_i(F)$ $(i = 1, 2, \dots)$ the *ordered eigenvalues* of a compact self-adjoint linear operator F in H if each of them is repeated as many times as its multiplicity and $|\lambda_1(F)| \geq |\lambda_2(F)| \geq ...$

(ii) The *singular values* of a compact operator C in H are

$$s_i(C) := \lambda_i(C^*C)^{1/2}, \qquad (i = 1, 2, \dots)$$

where $\lambda_i(C^*C)$ are the ordered eigenvalues of C^*C. In particular, if C is self-adjoint then $s_i(C) = |\lambda_i(C)|$.

It follows that the singular values of a compact operator cluster at the origin. Some useful properties of compact operators are listed below:

Proposition 1.2.3 *Let C be a compact operator in H. Then*

(a) for any $k \in \mathbf{N}^+$ and any orthonormal vectors $u_1, ..., u_k \in H$,

$$\sum_{i=1}^{k} \big|\langle Cu_i, u_i\rangle\big| \le \sum_{i=1}^{k} s_i(C)\,.$$

(b) If B is bounded linear operator in H, then

$$s_i(BC) \le \|B\|\, s_i(C) \qquad (i = 1, 2, \dots).$$

(c) (Variational characterization of the eigenvalues). If C is also self-adjoint, then

$$\big|\lambda_i(C)\big| = \min_{H_{i-1}\subset H} \max_{\substack{u\perp H_{i-1} \\ u\neq 0}} \frac{\big|\langle Cu, u\rangle\big|}{\|u\|^2}\,,$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace.

(d) If a sequence $(u_i) \subset H$ satisfies $\langle u_i, u_j\rangle = \langle Cu_i, u_j\rangle = 0$ $(i \neq j)$, then

$$\inf_i \, |\langle Cu_i, u_i\rangle|/\|u_i\|^2 = 0.$$

Proof. Statements (a) and (b) are the consequences of [65, Chap. VI, Corollary 3.3 and Proposition 1.3, resp.], for statement (c) see [64, Theorem III.9.1]. To prove (d), assume to the contrary that the infimum equals $\delta > 0$. We may assume that $\langle Cu_i, u_i\rangle$ has constant sign (otherwise we consider such a subsequence only). Then the orthonormal sequence $v_i := u_i/\|u_i\|$ satisfies for all $i \neq j$

$$2\delta \le |\langle Cv_i, v_i\rangle + \langle Cv_j, v_j\rangle| = |\langle C(v_i - v_j), v_i - v_j\rangle| \le \|C(v_i - v_j)\|\,\|v_i - v_j\| = \sqrt{2}\|C(v_i - v_j)\|,$$

hence the image (Cv_i) of the bounded sequence (v_i) contains no convergent subsequence, i.e. C is not compact. ■

Now the main definition comes, which we introduce within the class of S-bounded and S-coercive operators.

Definition 1.2.5 Let L and N be S-bounded and S-coercive operators in H. We call L and N *compact-equivalent in H_S* if

$$L_S = \mu N_S + Q_S \tag{1.2.18}$$

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

It follows in a straightforward way that the property of compact-equivalence is an equivalence relation.

(b) Characterization of compact-equivalence for elliptic operators

Let us now characterize compact-equivalence for elliptic operators. For this, let us consider the class of coercive elliptic operators defined in subsection 1.2.1. That is, let us pick two operators as in (1.2.6):

$$\begin{aligned} L_1 u &\equiv \ -\mathrm{div}\,(A_1\,\nabla u) + \ \mathbf{b}_1 \cdot \nabla u + c_1 u \qquad \text{for } \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_{A_1}} + \alpha_1 u_{|\Gamma_N} = 0, \\ L_2 u &\equiv \ -\mathrm{div}\,(A_2\,\nabla u) + \ \mathbf{b}_2 \cdot \nabla u + c_2 u \qquad \text{for } \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_{A_2}} + \alpha_2 u_{|\Gamma_N} = 0 \end{aligned}$$

where we assume that L_1 and L_2 satisfy Assumptions 1.2.1. Then by Proposition 1.2.2, the operators L_1 and L_2 are S-bounded and S-coercive in $L^2(\Omega)$, where S is the symmetric operator from (1.2.7). The corresponding energy space $H_S = H^1_D(\Omega)$ with S-inner product has been given in (1.2.8). Then it makes sense to study the compact-equivalence of L_1 and L_2 in $H^1_D(\Omega)$, and the following result is available:

Theorem 1.2.1 *Let the elliptic operators L_1 and L_2 satisfy Assumptions 1.2.1. 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. $A_1 = \mu A_2$.*

Proof. We have for all $u, v \in H^1_D(\Omega)$

$$\langle (L_i)_S u, v\rangle_S \ = \int_\Omega \Big(A_i \ \nabla u \cdot \nabla v + (\mathbf{b}_i \cdot \nabla u) v + c_i uv \Big)\, dx \ + \int_{\Gamma_N} \alpha_i uv \, d\sigma \, .$$

Hence $(L_1)_S - \mu (L_2)_S = J_S + Q_S$ where, using notations $\mathbf{b} := \mathbf{b}_1 - \mu \mathbf{b}_2$, $c := c_1 - \mu c_2$ and $\alpha := \alpha_1 - \mu\alpha_2$, we have

$$\langle J_S u, v\rangle_S \ = \int_\Omega (A_1 - \mu A_2)\, \nabla u \cdot \nabla v \, dx \quad \text{and} \quad \langle Q_S u, v\rangle_S = \int_\Omega \Big((\mathbf{b} \cdot \nabla u) v + cuv \Big)\, dx + \int_{\Gamma_N} \alpha uv \, d\sigma \, . \tag{1.2.19}$$

Here Q_S is compact, which is known [66] when L_1 and L_2 have the same boundary conditions. Otherwise we use the equality

$$\int_\Omega (\mathbf{b} \cdot \nabla u) v \, dx \ = - \int_\Omega u (\mathbf{b} \cdot \nabla v)\, dx - \int_\Omega (\mathrm{div}\,\mathbf{b}) uv \, dx + \int_{\Gamma_N} (\mathbf{b} \cdot \nu)\, uv \, d\sigma \quad (u, v \in H^1_D(\Omega))$$

whence, using notations $\tilde{c} := c - \mathrm{div}\,\mathbf{b}$ and $\tilde{\alpha} := \alpha + \mathbf{b} \cdot \nu$,

$$\|Q_S u\|_S = \sup_{\substack{v \in H^1_D(\Omega) \\ \|v\|_S = 1}} |\langle Q_S u, v\rangle_S| = \sup_{\substack{v \in H^1_D(\Omega) \\ \|v\|_S = 1}} \left| - \int_\Omega u (\mathbf{b} \cdot \nabla v)\, dx + \int_\Omega \tilde{c} uv \, dx + \int_{\Gamma_N} \tilde{\alpha}\, uv \, d\sigma \right| \, .$$

Using the embedding estimates (1.2.10) and that $\|\nabla v\|_{L^2(\Omega)} \leq p^{-1/2} \|v\|_S$, and letting $K_1 := p^{-1/2} \|\mathbf{b}\|_{L^\infty(\Omega)} + C_{\Omega,S} \|\tilde{c}\|_{L^\infty(\Omega)}$, $K_2 := C_{\Gamma_N,S} \|\tilde{\alpha}\|_{L^\infty(\Gamma_N)}$, we obtain

$$\|Q_S u\|_S \leq K_1 \|u\|_{L^2(\Omega)} + K_2 \|u\|_{L^2(\Gamma_N)}. \tag{1.2.20}$$

From this we can prove that Q_S is compact. Namely, let $(u_n) \subset H^1_D(\Omega)$ be a bounded sequence in the S-norm. Since the embedding of $H^1_D(\Omega)$ into $L^2(\Omega)$ is compact, (u_n) has a convergent subsequence in L^2-norm. This sequence is also bounded in the S-norm, and since the trace mapping of $H^1_D(\Omega)$ into $L^2(\Gamma_N)$ is compact, we find that (u_n) has a convergent subsequence in both $L^2(\Omega)$-norm and $L^2(\Gamma_N)$-norm. By (1.2.20), we obtain that $(Q_S u_n)$ has a convergent subsequence in the S-norm, hence Q_S is compact.

It remains to prove that if $A_1 \neq \mu A_2$ then J_S is not compact. Using Proposition 1.2.3 (d), it suffices to find a sequence $(u_i) \subset H^1_0(\Omega) \subset H^1_D(\Omega)$ satisfying

$$\langle u_i, u_j \rangle_S = \langle J_S u_i, u_j \rangle_S = 0 \qquad (i \neq j), \tag{1.2.21}$$

$$\inf_i \, |\langle J_S u_i, u_i \rangle_S| / \|u_i\|^2_S \geq \delta > 0. \tag{1.2.22}$$

Let $A := A_1 - \mu A_2$. Since A is not identically zero, there is $x_0 \in \Omega$ such that $A_0 := A(x_0) \neq 0$. Here A_0 is symmetric, hence there is $u_0 \in H^1_0(\Omega)$ such that $\displaystyle\int_\Omega A_0 \, \nabla u_0 \cdot \nabla u_0 \neq 0$. Let

$$\varepsilon := \Big| \int_\Omega A_0 \, \nabla u_0 \cdot \nabla u_0 \Big| / \Big(\int_\Omega |\nabla u_0|^2 \Big), \qquad \Omega_{\varepsilon/2} := \{ x \in \Omega : \ \|A(x) - A_0\| < \varepsilon/2 \}$$

which contains an open set since $A \in PC(\Omega)$. Fix $z' \in \Omega$, and for any $z \in \Omega$ and $R > 0$ let $\Omega_{z,R} := \{ x \in \mathbf{R}^d : \ z' + R(x-z) \in \Omega \}$. Let $z_i \in \Omega$, $R_i > 0$ $(i \in \mathbf{N}^+)$ such that $\Omega_i := \Omega_{z_i, R_i} \subset \Omega_{\varepsilon/2}$ and $\overline{\Omega}_i$ are pairwise disjoint sets. We define $u_i \in H^1_0(\Omega)$ by $u_i(x) := u_0\big(z' + R_i(x - z_i)\big)$ for $x \in \Omega_i$ and $u_i(x) := 0$ for $x \in \Omega \setminus \Omega_i$. Since $\operatorname{supp} u_i = \overline{\Omega}_i$ are disjoint, (1.2.21) is satisfied. Further, using the fact $\Omega_i \subset \Omega_{\varepsilon/2}$ and a linear transformation $\Omega_i \to \Omega$ in the integral, we obtain

$$\frac{|\langle J_S u_i, u_i \rangle_S|}{\int_{\Omega_i} |\nabla u_i|^2} = \frac{\left| \int_{\Omega_i} A \, \nabla u_i \cdot \nabla u_i \right|}{\int_{\Omega_i} |\nabla u_i|^2} \geq \frac{\left| \int_{\Omega_i} A_0 \, \nabla u_i \cdot \nabla u_i \right|}{\int_{\Omega_i} |\nabla u_i|^2} - \frac{\varepsilon}{2} = \frac{\left| \int_{\Omega} A_0 \, \nabla u_0 \cdot \nabla u_0 \right|}{\int_{\Omega} |\nabla u_0|^2} - \frac{\varepsilon}{2} = \frac{\varepsilon}{2} \, .$$

Since for $u \in H^1_0(\Omega)$ have $\|u\|^2_S \leq C \cdot \int_\Omega |\nabla u|^2$, the above estimate yields (1.2.22) with $\delta = \frac{\varepsilon}{2C} > 0$. ■

1.2.3 Mesh independent superlinear convergence in Hilbert space

Equation (1.1.1) can be solved numerically using a Galerkin discretization in a subspace $V_h = span\{\varphi_1, \dots, \varphi_n\} \subset H_S$. Finding the discrete solution $u_h \in V_h$ in a form $u = \sum\limits_{i=1}^n c_i \varphi_i$ requires solving the $n \times n$ system

$$\mathbf{L}_h \, \mathbf{c} = \mathbf{b}_h \tag{1.2.23}$$

where $\mathbf{L}_h = \big\{ \langle L_S \varphi_j, \varphi_i \rangle_S \big\}_{i,j=1}^n$ and $\mathbf{b}_h = \{ \langle g, \varphi_j \rangle \}_{j=1}^n$. Since $L \in BC_S(H)$, the symmetric part of $\mathbf{L}_h$ is positive definite, hence system (1.2.23) has a unique solution. Moreover, if a sequence of such subspaces V_h satisfies $\inf_{v \in V_h} \|u - v\|_S \to 0$ for all $u \in H_S$, then the coercivity of L_S implies that u_h converges to the exact weak solution in H_S-norm [34].

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 (1.2.18). 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.

(a) Symmetric compact-equivalent preconditioners

Let us consider operators L and S such that L is S-bounded and S-coercive as in Definition 1.2.1. Assume in addition that L and S are compact-equivalent with $\mu = 1$. Then (1.2.18) holds with $N_S = I$:

$$L_S = I + Q_S \tag{1.2.24}$$

with a compact operator Q_S. We apply the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.2.23). By (1.2.24), letting

$$\mathbf{Q}_h = \left\{ \langle Q_S \varphi_j, \varphi_i \rangle_S \right\}_{i,j=1}^{n}, \tag{1.2.25}$$

the preconditioned system takes the form

$$(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \tilde{\mathbf{b}_h} \tag{1.2.26}$$

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

In order to have mesh independent bounds for the CG estimates in the case $A = \mathbf{S}_h^{-1}\mathbf{L}_h$, we first verify the bound $\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) \geq m$, hence the remaining task will be to find bounds for the sums of eigenvalues in the CG estimate expressions in the case $E = \mathbf{S}_h^{-1}\mathbf{Q}_h$.

Proposition 1.2.4 *The lower bounds satisfy* $\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) \geq m$, *where* λ_0 *is defined in (1.1.8) and* m *comes from (1.2.2).*

Proof. We have

$$\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) = \min_{\substack{\mathbf{c}\in\mathbf{R}^n \\ \mathbf{c}\neq\mathbf{0}}} \frac{\langle \mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c}, \mathbf{c}\rangle_{\mathbf{S}_h}}{\|\mathbf{c}\|_{\mathbf{S}_h}^2} = \min_{\substack{\mathbf{c}\in\mathbf{R}^n \\ \mathbf{c}\neq\mathbf{0}}} \frac{\mathbf{L}_h\,\mathbf{c}\cdot\mathbf{c}}{\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c}} = \min_{\substack{u\in V_h \\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2}$$

$$\geq \inf_{\substack{u\in H_S \\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2} = \inf_{\substack{u\in D(L) \\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2} = \inf_{\substack{u\in D(L) \\ u\neq 0}} \frac{\langle L u, u\rangle}{\|u\|_S^2} = m$$

where the density of $D(L)$ in H_S has been used. ■

Proposition 1.2.5 *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*

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

Proof. By Proposition 1.2.3 (a), any compact operator L and any orthonormal vectors $u_1, ..., u_k$ in the space H_S satisfy the inequality

$$\sum_{m=1}^{k} |\langle Lu_m, u_m\rangle| \leq \sum_{m=1}^{k} s_m(L) \tag{1.2.27}$$

where $|s_1(L)| \geq |s_2(L)| \geq ...$ are the singular values of L, i.e. the ordered eigenvalues of the operator $(L^*L)^{1/2}$. If L is also normal then $s_m(L) = |\lambda_m(L)|$ where $|\lambda_1(L)| \geq |\lambda_2(L)| \geq ...$ are the eigenvalues of L. Hence for the operator Q_S in H_S we obtain

$$\sum_{m=1}^{k} |\langle Qu_m, u_m\rangle| = \sum_{m=1}^{k} |\langle Q_S u_m, u_m\rangle_S| \leq \sum_{m=1}^{k} |\lambda_m(Q_S)|. \tag{1.2.28}$$

Therefore, in order to prove (1.2.45), it remains to find orthonormal vectors $u_1, ..., u_n$ in H_S such that

$$\sum_{m=1}^{k} |\lambda_m(\mathbf{S}_h^{-1}\mathbf{Q}_h)| \leq \sum_{m=1}^{k} |\langle Qu_m, u_m\rangle| \qquad (k = 1, \dots, n). \tag{1.2.29}$$

In what follows, let $\lambda_m \quad (m = 1, ..., n)$ denote the eigenvalues $\lambda_m(\mathbf{S}_h^{-1}\mathbf{Q}_h)$ of $\mathbf{S}_h^{-1}\mathbf{Q}_h$. Let $\mathbf{c}^m = (c_1^m, \dots, c_n^m) \in \mathbf{C}^n$ be corresponding eigenvectors. Then

$$\mathbf{Q}_h \mathbf{c}^m = \lambda_m \mathbf{S}_h \mathbf{c}^m \qquad (m = 1, ..., n). \tag{1.2.30}$$

Since $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is normal w.r.t the $\mathbf{S}_h$-inner product, the eigenvectors $\mathbf{c}^m$ $(m = 1, ..., n)$ are orthogonal in $\mathbf{C}^n$ w.r.t the $\mathbf{S}_h$-inner product. Let them be also orthonormal:

$$\mathbf{S}_h \, \mathbf{c}^m \cdot \mathbf{c}^l = \delta_{ml} \qquad (m, l = 1, ..., n), \tag{1.2.31}$$

where δ_{ml} is the Kronecker symbol.

Let $u_m = \sum_{i=1}^n c_i^m \varphi_i \ \in V_h \qquad (m = 1, ..., n)$. Then for all $m, l = 1, ..., n$

$$\langle u_m, u_l\rangle_S = \sum_{i,j=1}^{n} \langle \varphi_i, \varphi_j\rangle_S \ \ c_i^m \, c_j^l = \mathbf{S}_h \, \mathbf{c}^m \cdot \mathbf{c}^l, \tag{1.2.32}$$

hence (1.2.31) implies that $u_1, ..., u_n$ form an orthonormal base in V_h w.r.t the inner product of H_S. Further, (1.2.30) and (1.2.31) yield $\mathbf{Q}_h \, \mathbf{c}^m \cdot \mathbf{c}^l = \lambda_m \, \delta_{ml}$ $(m, l = 1, ..., n)$ and, together with the analogue of (1.2.32) for Q, this implies

$$\langle Qu_m, u_l\rangle = \lambda_m \, \delta_{ml} \qquad (m, l = 1, ..., n) \tag{1.2.33}$$

and hence

$$\sum_{m=1}^{k} |\lambda_m| = \sum_{m=1}^{k} |\langle Qu_m, u_m\rangle|. \tag{1.2.34}$$

■

If H is a real Hilbert space (as was originally in this chapter) then H and H_S can be extended to a complex Hilbert space in a standard way. From Proposition 1.2.5 and the standard estimate we can then derive

Theorem 1.2.2 *Under the conditions of Proposition 1.2.5, the GCG-LS algorithm for system (1.2.26) yields*

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

and ε_k is a sequence independent of V_h.

Proof. The result follows directly from the analogue of (1.1.14), mentioned in section (1.1.2), which now holds for $(\|r_k\|_{\mathbf{S}_h}/\|r_0\|_{\mathbf{S}_h})^{1/k}$ since E is normal in $\mathbf{S}_h$-inner product, and Propositions 1.2.4 and 1.2.5. Further, the property $\varepsilon_k \to 0$ follows from the fact that $|\lambda_m(Q_S)| \to 0$ (as $k \to \infty$) and ε_k is the arithmetic mean sequence of them. ■

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 $m=1$, see [16].

In the general case without normality, we have the following bounds for (1.1.14):

Proposition 1.2.6 *Any compact operator Q_S in H_S satisfies the following relations:*

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

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

Proof. (a) Let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{S}_h^{-1}\mathbf{Q}_h)$ $(i=1,...,n)$ and let $\mathbf{c}^i = (c_1^i,\dots,c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors such that

$$\mathbf{S}_h\,\mathbf{c}^i\cdot\mathbf{c}^l = \delta_{il} \qquad (i,l=1,...,n), \tag{1.2.35}$$

where $\cdot$ denotes the ordinary inner product on $\mathbf{R}^n$. Then

$$\mathbf{S}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i\cdot\mathbf{Q}_h\,\mathbf{c}^i = \lambda_i \qquad (i=1,...,n). \tag{1.2.36}$$

Let $\mathbf{d}^i := \mathbf{S}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i$ for all i, that is

$$\mathbf{S}_h\,\mathbf{d}^i = \mathbf{Q}_h\,\mathbf{c}^i \tag{1.2.37}$$

which turns (1.2.36) into

$$\mathbf{S}_h\mathbf{d}^i\cdot\mathbf{d}^i = \lambda_i\,. \tag{1.2.38}$$

Now let $u_i = \sum_{j=1}^{n} c_j^i\varphi_j \in V_h$ and $z_i = \sum_{j=1}^{n} d_j^i\varphi_j \in V_h$ $(i=1,...,n)$. Then (1.2.38) yields

$$\|z_i\|_S^2 = \lambda_i\,. \tag{1.2.39}$$

Further, for all $v = \sum_{j=1}^{n} p_j \varphi_j \ \in V_h$, with notation $\mathbf{p} = (p_1, \dots, p_n) \in \mathbf{R}^n$, (1.2.37) yields $\mathbf{S}_h\, \mathbf{d}^i \cdot \mathbf{p} = \mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{p}$, which implies

$$\langle z_i, v\rangle_S = \langle Q_S u_i, v\rangle_S \qquad (v \in V_h),$$

i.e. z_i is the orthogonal projection of $Q_S u_i \in H_S$ into V_h. Therefore $\|z_i\|_S \le \|Q_S u_i\|_S$, and (1.2.39) provides

$$\sum_{i=1}^{k} \lambda_i \le \sum_{i=1}^{k} \|Q_S u_i\|_S^2 = \sum_{i=1}^{k} \langle Q_S^* Q_S u_i, u_i\rangle_S. \tag{1.2.40}$$

Here $\langle u_i, u_l\rangle_S = \mathbf{S}_h\, \mathbf{c}^i \cdot \mathbf{c}^l$ for all $i, l = 1, ..., n$, hence by (1.2.35) the vectors u_i are orthonormal in H_S. Therefore Proposition 1.2.3 (a) for the operator $C = Q_S^* Q_S$ in the space H_S yields the desired estimate.

(b) The proof is similar to that of (a). Now let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)$ and let $\mathbf{c}^i = (c_1^i, \dots, c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (1.2.35). Then

$$(\mathbf{Q}_h^T + \mathbf{Q}_h)\, \mathbf{c}^i = \lambda_i\, \mathbf{S}_h\, \mathbf{c}^i \qquad (i = 1, ..., n)$$

and (1.2.35) yields

$$\lambda_i \ = \ (\mathbf{Q}_h^T + \mathbf{Q}_h)\, \mathbf{c}^i \cdot \mathbf{c}^i = 2\, \mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{c}^i.$$

For $u_i = \sum_{j=1}^{n} c_j^i \varphi_j \ \in V_h$ we thus obtain

$$\sum_{i=1}^{k} |\lambda_i| \ = 2 \sum_{i=1}^{k} \left|\langle Q_S u_i, u_i\rangle_S\right| = \sum_{i=1}^{k} \left|\langle (Q_S^* + Q_S) u_i, u_i\rangle_S\right|, \tag{1.2.41}$$

and Proposition 1.2.3 (a) for the operator $C = Q_S^* + Q_S$ in the space H_S yields the desired estimate. ■

In virtue of (1.1.14) and Propositions 1.2.4 and 1.2.6, we have proved

Theorem 1.2.3 *The CGN algorithm for system (1.2.26) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \ \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{1.2.42}$$

where

$$\varepsilon_k := \frac{2}{km^2} \ \sum_{i=1}^{k} \Big(\left|\lambda_i(Q_S^* + Q_S)\right| + \lambda_i(Q_S^* Q_S) \Big) \ \to 0 \quad \textit{as} \quad k \to \infty \tag{1.2.43}$$

and ε_k is a sequence independent of V_h.

A more explicit bound can be obtained for more special operators. 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. [65]). Then we can obtain a more explicit rate $O(k^{-1/2})$. First, applying the geometric-arithmetic mean estimate to (1.1.14), we obtain

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2}{\sqrt{k}\lambda_0}\,\|E\|_F \qquad (k=1,2,...,n), \tag{1.2.44}$$

where $\|E\|_F := \big(\sum\limits_{j=1}^{k} |\lambda_j(E)|^2\big)^{1/2}$ is the Frobenius norm of E. In the case $E := \mathbf{S}_h^{-1}\mathbf{Q}_h$ in (1.2.26), the Frobenius norm in (1.2.44) can be estimated as follows.

Proposition 1.2.7 *If Q_S is a Hilbert-Schmidt operator in H_S, then*

$$\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_F \leq |||Q_S|||. \tag{1.2.45}$$

Proof. It is similar to the proof of Proposition 1.2.5. We use (1.2.33) and let $u_{k+1}, u_{k+2}, ...$ be a complete orthonormal system in the orthocomplement of V_h in H_S. Then $u_1, u_2, ...$ form a complete orthonormal system in H_S. Using the invariance theorem on an arbitrary Hilbert-Schmidt operator L in some Hilbert space [65], and then (1.2.33), we obtain for Q_S in the space H_S that

$$|||Q_S|||^2 = \sum_{m,l=1}^{\infty} |\langle Q_S u_m, u_l\rangle_S|^2 = \sum_{m,l=1}^{\infty} |\langle Q u_m, u_l\rangle|^2 \geq \sum_{m,l=1}^{k} |\langle Q u_m, u_l\rangle|^2 = \|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_F^2. \quad \blacksquare$$

Then (1.2.44) and Propositions 1.2.4 and 1.2.7 yield the rate $O(k^{-1/2})$:

Corollary 1.2.1 *If Q_S is a Hilbert-Schmidt operator, then the CG method for* (1.2.26) *yields*

$$\left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}}\right)^{1/k} \leq \frac{2}{\sqrt{k}m}\,|||Q_S||| \qquad (k=1,2,...,n). \tag{1.2.46}$$

We note that the factor $2/m$ of $|||Q_S|||/\sqrt{k}$ can be improved to $\sqrt{3/2m}$, using K-condition numbers as in [83].

(b) 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., (1.2.18) becomes $L_S = N_S + Q_S$. We apply the stiffness matrix $\mathbf{N}_h$ of N_S as preconditioner for the discretized system (1.2.23). 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. Then the $\mathbf{S}_h$-adjoint of $\mathbf{N}_h^{-1}\mathbf{L}_h$ is $\mathbf{S}_h^{-1}\mathbf{L}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h$, hence we apply the CGN algorithm with $A = \mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{N}_h^{-T}\mathbf{S}_h$.

Using (1.2.25), the preconditioned system (1.3.9) takes the form

$$(\mathbf{I}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \hat{\mathbf{b}}_h \tag{1.2.47}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix. By (1.1.13), the CGN algorithm then provides

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{2}{k\nu_h} \sum_{i=1}^{k} \Big(\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h) + \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h)\Big) \tag{1.2.48}$$

$(k = 1, 2, ..., n)$, where

$$\nu_h = \frac{1}{\|\mathbf{N}_h^{-1}\mathbf{L}_h\|_{\mathbf{S}_h}} = \min_{\mathbf{c}\in\mathbf{R}^n} \frac{\|\mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}\|^2_{\mathbf{S}_h}}{\|\mathbf{c}\|^2_{\mathbf{S}_h}}\,. \tag{1.2.49}$$

Again, our goal is to give a bound on (1.2.48) that is independent of V_h.

Proposition 1.2.8 *Let L and N be S-bounded and S-coercive operators, in particular*

$$m := \inf_{\substack{u\in D(L)\\ u\neq 0}} \frac{\langle Lu, u\rangle}{\|u\|_S^2} > 0, \qquad \hat{m} := \inf_{\substack{u\in D(N)\\ u\neq 0}} \frac{\langle Nu, u\rangle}{\|u\|_S^2} > 0, \qquad \hat{M} := \sup_{\substack{u\in D(N)\\ u\neq 0}} \frac{|\langle Nu, v\rangle|}{\|u\|_S\|v\|_S} > 0,$$

and let Q_S be a compact operator on H_S. Let $\mathbf{S}_h$, $\mathbf{N}_h$ and $\mathbf{Q}_h$ be defined as above, and let $s_i(Q_S)$ $(i = 1, 2, \dots)$ denote the singular values of Q_S. Then the following relations hold:

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

(b) $$\sum_{i=1}^{k} \big|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)\big| \le \frac{2}{\hat{m}} \sum_{i=1}^{k} s_i(Q_S) \qquad (k = 1, \dots, n),$$

(c) $$\nu_h \ge \frac{m^2}{\hat{M}^2}\,.$$

Proof. (a) We proceed similarly to Proposition 1.2.6. Let

$$\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h) \qquad (i = 1, ..., n)$$

and let $\mathbf{c}^i = (c_1^i, \dots, c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (1.2.35). Then

$$\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i = \lambda_i \qquad (i = 1, ..., n). \tag{1.2.50}$$

Let $\mathbf{d}^i := \mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i$ for all i, that is

$$\mathbf{N}_h\,\mathbf{d}^i = \mathbf{Q}_h\,\mathbf{c}^i\,. \tag{1.2.51}$$

For this $\mathbf{d}^i$ and λ_i, similarly to Proposition 1.2.6, we have (1.2.38) and, letting $u_i = \sum\limits_{j=1}^{n} c_j^i\varphi_j \in V_h$ and $z_i = \sum\limits_{j=1}^{n} d_j^i\varphi_j \in V_h$ we obtain (1.2.39). Further, for all $v = \sum\limits_{j=1}^{n} p_j\varphi_j \in V_h$, with notation $\mathbf{p} = (p_1, \dots, p_n) \in \mathbf{R}^n$, (1.2.51) yields $\mathbf{N}_h\,\mathbf{d}^i \cdot \mathbf{p} = \mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{p}$, which means

$$\langle N_S z_i, v\rangle_S = \langle Q_S u_i, v\rangle_S \qquad (v \in V_h).$$

From this we have

$$\|z_i\|_S^2 \le \frac{1}{\hat{m}} \langle N_S z_i, z_i\rangle_S = \frac{1}{\hat{m}} \langle Q_S u_i, z_i\rangle_S \le \frac{1}{\hat{m}} \|Q_S u_i\|_S \|z_i\|_S \,,$$

hence $\|z_i\|_S \le \frac{1}{\hat{m}} \|Q_S u_i\|_S$. Then from (1.2.39)

$$\sum_{i=1}^{k} \lambda_i \le \frac{1}{\hat{m}^2} \sum_{i=1}^{k} \|Q_S u_i\|_S^2 = \frac{1}{\hat{m}^2} \sum_{i=1}^{k} \langle Q_S^* Q_S u_i, u_i\rangle_S, \tag{1.2.52}$$

whence the desired estimate follows in the same way as from (1.2.40) in Proposition 1.2.6.

(b) Now let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)$ and let $\mathbf{c}^i = (c_1^i, \dots, c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (1.2.35). Then

$$\lambda_i = \lambda_i\, \mathbf{S}_h\, \mathbf{c}^i \cdot \mathbf{c}^i = \mathbf{Q}_h^T\, \mathbf{N}_h^{-T} \mathbf{S}_h\, \mathbf{c}^i \cdot \mathbf{c}^i + \mathbf{S}_h \mathbf{N}_h^{-1} \mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{c}^i \;=\; 2\; \mathbf{S}_h \mathbf{N}_h^{-1} \mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{c}^i \;= 2\, \mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{e}^i$$

where $\mathbf{e}^i := \mathbf{N}_h^{-T} \mathbf{S}_h\, \mathbf{c}^i$ for all i. Here for all $v = \sum_{j=1}^{n} p_j \varphi_j \in V_h$, with notation $\mathbf{p} = (p_1, \dots, p_n) \in \mathbf{R}^n$, we obtain $\mathbf{e}^i \cdot \mathbf{N}_h\, \mathbf{p} = \mathbf{S}_h\, \mathbf{c}^i \cdot \mathbf{p}$, which means $\langle w_i, N_S v\rangle_S = \langle u_i, v\rangle_S$ for all $v \in V_h$, where $w_i = \sum_{j=1}^{n} e_j^i \varphi_j$ and $u_i = \sum_{j=1}^{n} c_j^i \varphi_j$, or

$$\langle N_S^* w_i, v\rangle_S = \langle u_i, v\rangle_S \qquad (v \in V_h). \tag{1.2.53}$$

Denote by P the orthogonal projection of H_S onto V_h. Then (1.2.53) yields $u_i = P N_S^* w_i$. Here the linear mapping $(P N_S^*)_{|V_h} : V_h \to V_h$ is one-to-one, since for all $v \in V_h$

$$\langle P N_S^* v, v\rangle_S = \langle N_S^* v, v\rangle_S = \langle N_S v, v\rangle_S \ge \hat{m} \|v\|_S^2 \,. \tag{1.2.54}$$

Therefore

$$\mathbf{Q}_h\, \mathbf{c}^i \cdot \mathbf{e}^i = \langle Q_S u_i, w_i\rangle_S = \langle Q_S u_i, (P N_S^*)_{|V_h}^{-1} u_i\rangle_S = \langle u_i,\, Q_S^* (P N_S^*)_{|V_h}^{-1} u_i\rangle_S.$$

Here the operator $(P N_S^*)_{|V_h}^{-1}$ has a norm-preserving extension $\hat{N}$ from V_h onto H_S (namely, with $\hat{N}\big|_{(V_h)\perp} := 0$), and from (1.2.54) we have $\|\hat{N}\| \le \frac{1}{\hat{m}}$. Altogether, we obtain

$$\sum_{i=1}^{k} |\lambda_i| = 2 \sum_{i=1}^{k} \big|\langle Q_S^* (P N_S^*)_{|V_h}^{-1} u_i,\, u_i\rangle_S\big| \;=\; 2 \sum_{i=1}^{k} \big|\langle Q_S^* \hat{N} u_i,\, u_i\rangle_S\big| \;\le\; 2 \sum_{i=1}^{k} s_i\big(Q_S^* \hat{N}\big)$$

$$\le \frac{2}{\hat{m}} \sum_{i=1}^{k} s_i\big(Q_S^*\big) = \frac{2}{\hat{m}} \sum_{i=1}^{k} s_i\big(Q_S\big)$$

(where, in the inequalities, statements (a) and (b) of Proposition 1.2.3 have been used, respectively).

(c) Let $\mathbf{c} \in \mathbf{R}^n$ be arbitrary, $\mathbf{d} := \mathbf{N}_h^{-1} \mathbf{L}_h \mathbf{c}$. Let $u = \sum_{j=1}^{n} c_j \varphi_j \in V_h$ and $z = \sum_{j=1}^{n} d_j \varphi_j \in V_h$. Then $m\|u\|_S^2 \le \langle L_S u, u\rangle_S = \mathbf{L}_h\, \mathbf{c} \cdot \mathbf{c} = \mathbf{N}_h\, \mathbf{d} \cdot \mathbf{c} = \langle N_S z, u\rangle_S \le \|N_S z\|_S \|u\|_S$, hence

$$m\|u\|_S \le \|N_S z\|_S$$

and
$$\frac{\|\mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}\|^2_{\mathbf{S}_h}}{\|\mathbf{c}\|^2_{\mathbf{S}_h}} = \frac{\mathbf{S}_h\,\mathbf{d}\cdot\mathbf{d}}{\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c}} = \frac{\|z\|^2_S}{\|u\|^2_S} \geq m^2\,\frac{\|z\|^2_S}{\|N_S z\|^2_S} \geq \frac{m^2}{\hat{M}^2}\,. \qquad \blacksquare$$

Theorem 1.2.4 *Using compact-equivalent operators L and N, the CGN algorithm for system (1.2.47) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \ \varepsilon_k \qquad (k=1,2,...,n) \tag{1.2.55}$$

$$\textit{where}\quad \varepsilon_k = \frac{2M_N^2}{km_L^2}\,\sum_{i=1}^{k}\Big(\frac{2}{m_N}\,s_i(Q_S) + \frac{1}{m_N^2}\,s_i(Q_S)^2\Big) \ \to 0 \qquad (\textit{as } k\to\infty) \tag{1.2.56}$$

and ε_k is a sequence independent of V_h.

Proof. It follows from (1.2.48) and Proposition 1.2.8. ■

1.2.4 Mesh independent superlinear convergence for elliptic problems

(a) Elliptic equations

In this subsection we consider nonsymmetric elliptic problems

$$\left\{\begin{array}{l} Lu := -\mathrm{div}\,(A\,\nabla u) + \,\mathbf{b}\cdot\nabla u + cu = g \\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \end{array}\right. \tag{1.2.57}$$

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 1.2.1, that is, L is of the type (1.2.6), and further, that $g \in L^2(\Omega)$. Defining the corresponding Sobolev space $H^1_D(\Omega)$ as in (1.2.8), problem (1.2.57) has a unique weak solution $u \in H^1_D(\Omega)$. Such equations typically arise in convection-diffusion problems.

We use the FEM to solve (1.2.57), we define a subspace $V_h = span\{\varphi_1,\dots,\varphi_n\} \subset H^1_D(\Omega)$ and seek the FEM solution $u_h \in V_h$, which requires solving an $n\times n$ system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h. \tag{1.2.58}$$

Based on the previous abstract results, we can readily derive efficient preconditioned algorithms that produce mesh independent superlinear convergence.

Symmetric preconditioners: general convergence. To exploit Theorem 1.2.1, we define S to have the same principal part as L:

$$Su \equiv\ -\mathrm{div}\,(A\,\nabla u) + \sigma u \qquad \text{for}\ \ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{1.2.59}$$

assumed to satisfy Assumptions 1.2.2. We introduce the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.2.58), and then solve the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h} \tag{1.2.60}$$

(with $\tilde{\mathbf{g}_h} = \mathbf{S}_h^{-1}\mathbf{g}_h$) with a CG method. Using the decomposition $\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h$, system (1.2.60) can be rewritten as in (1.2.26), and here $\mathbf{Q}_h = \Big\{\langle Q_S\varphi_j, \varphi_i\rangle_S\Big\}_{i,j=1}^n$ where Q_S is the operator on $H_D^1(\Omega)$ defined via

$$\langle Q_S u, v\rangle_S = \int_\Omega \Big((\mathbf{b}\cdot\nabla u)v + (c-\sigma)uv\Big) + \int_{\Gamma_N}(\alpha-\beta)uv\,d\sigma \qquad (u, v \in H_D^1(\Omega)), \tag{1.2.61}$$

which satisfies (1.2.24) in $H_D^1(\Omega)$ under the S-inner product.

First we consider the GCG-LS method.

Theorem 1.2.5 *If Q_S in (1.2.61) is normal, then the GCG-LS algorithm for system (1.2.60) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k=1,...,n), \quad \textit{where} \quad \varepsilon_k := \frac{2}{km}\sum_{j=1}^k \big|\lambda_j(Q_S)\big| \;\to 0 \quad \textit{as} \quad k\to\infty \tag{1.2.62}$$

(with m from (1.2.16)) and ε_k is a sequence independent of V_h.

Proof. By Proposition 1.2.2, L is S-bounded and S-coercive. Theorem 1.2.1 yields that L and S are compact-equivalent in $H_D^1(\Omega)$ if the latter is endowed with the S-inner product. Therefore Theorem 1.2.2 is valid with the compact operator Q_S defined in (1.2.61). ∎

The main application of this is symmetric part preconditioning, discussed in the separate section 1.4. Besides, the normality assumption on Q_S is only known to cover the case of constant coefficients in L, which is practically uninteresting: however, the experiments in [108] show a wider validity of mesh independent superlinear convergence.

In turn, the CGN algorithm provides similar results without any such restrictions:

Theorem 1.2.6 *The CGN algorithm for the preconditioned system (1.2.60) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \qquad (k=1,2,...,n), \tag{1.2.63}$$

where

$$\varepsilon_k := \frac{2}{km^2}\sum_{i=1}^k \Big(\big|\lambda_i(Q_S^* + Q_S)\big| + \lambda_i(Q_S^*Q_S)\Big) \;\to 0 \quad \textit{as} \quad k\to\infty \tag{1.2.64}$$

(with m from (1.2.16) and Q_S from (1.2.61)), and ε_k is a sequence independent of V_h.

Proof. It is same as that of Theorem 1.2.5, but now we use Theorem 1.2.3 instead of Theorem 1.2.2. ■

Efficient solvers arise from symmetric preconditioners such as e.g. the symmetric part, Laplacian or Helmholtz operators [113, 116], and in the general case one can use multigrid solvers for S [146]. Various examples will be given in section 1.5.

Symmetric preconditioners: the magnitude of superlinear convergence. Besides superlinear convergence in general, of greater practical interest is a more constructive form of the above sequences ε_k.

Although they are not a priori computable in practice, the magnitude in which $\varepsilon_k \to 0$ can be determined in certain cases. Consider first the CGN method and Theorem 1.2.6. We give magnitude estimates in the case when the asymptotics $\mu_i = O(i^{2/d})$ are known for symmetric eigenvalue problems

$$Su = \mu u, \quad u_{|\Gamma_D} = 0, \quad r\left(\tfrac{\partial u}{\partial \nu_A} + \beta u\right)_{|\Gamma_N} = \mu u\,, \tag{1.2.65}$$

as is the case for Dirichlet problems.

A similar result in 2D will be seen later for symmetric part preconditioning for the GCG-LS method in subsection 1.4.1.

Theorem 1.2.7 *The sequence ε_k in (1.2.64) satisfies $\varepsilon_k \le (4s/k)\sum\limits_{i=1}^{k}(1/\mu_i)$ for some constants $s, r > 0$, where μ_i ($i \in \mathbf{N}^+$) are the solutions of (1.2.65). When the asymptotics $\mu_i = O(i^{2/d})$ holds, in particular, for Dirichlet boundary conditions,*

$$\varepsilon_k \le O\Big(\frac{\log k}{k}\Big) \quad \textit{if} \;\; d = 2 \quad \textit{and} \quad \varepsilon_k \le O\Big(\frac{1}{k^{2/d}}\Big) \quad \textit{if} \;\; d \ge 3. \tag{1.2.66}$$

Proof. From (1.2.61) and the divergence theorem, letting $d = c - h$ and $\gamma = \alpha - \beta$,

$$\langle Q_S u, u\rangle_S = \int_\Omega \Big(d - \frac{1}{2}(\operatorname{div} \mathbf{b})\Big)u^2 \; + \int_{\Gamma_N} \Big(\gamma + \frac{1}{2}(\mathbf{b}\cdot\nu)\Big)u^2\, d\sigma \le C_1\|u\|^2_{L^2(\Omega)} + C_2\|u\|^2_{L^2(\Gamma_N)}.$$

We have $\big|\langle (Q_S^* + Q_S)u, u\rangle_S\big| = 2\big|\langle Q_S u, u\rangle_S\big|$, hence the variational characterization of the eigenvalues yields

$$\big|\lambda_i(Q_S^* + Q_S)\big| = \min_{H_{i-1}\subset H_S}\; \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{\big|\langle (Q_S^* + Q_S)u, u\rangle_S\big|}{\|u\|_S^2} \le 2 \min_{H_{i-1}\subset H_S}\; \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{C_1\|u\|^2_{L^2(\Omega)} + C_2\|u\|^2_{L^2(\Gamma_N)}}{\|u\|_S^2}\,,$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace. On the other hand, here Q_S falls into the type (1.2.19), hence (1.2.20) implies

$$\|Q_S u\|_S^2 \le 2K_1^2\|u\|^2_{L^2(\Omega)} + 2K_2^2\|u\|^2_{L^2(\Gamma_N)}\,.$$

Since $s_i(Q_S)^2 = \lambda_i(Q_S^* Q_S)$ and $\langle Q_S^* Q_S u, u\rangle_S = \|Q_S u\|_S^2$, we obtain as above that

$$s_i(Q_S)^2 = \min_{H_{i-1}\subset H_S}\; \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{\langle Q_S^* Q_S u, u\rangle_S}{\|u\|_S^2} \le \min_{H_{i-1}\subset H_S}\; \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{2K_1^2\|u\|^2_{L^2(\Omega)} + 2K_2^2\|u\|^2_{L^2(\Gamma_N)}}{\|u\|_S^2}\,.$$

Altogether, letting $s := \frac{C_1+K_1^2}{m^2}$, $r := \frac{C_1+K_1^2}{C_2+K_2^2}$, formula (1.2.64) implies

$$\varepsilon_k \leq \frac{4s}{k} \sum_{i=1}^{k} \hat{\mu}_i \qquad \text{where } \hat{\mu}_i = \min_{H_{i-1} \subset H_S} \max_{\substack{u \perp H_{i-1} \\ u \neq 0}} \frac{\|u\|^2_{L^2(\Omega)} + \frac{1}{r}\|u\|^2_{L^2(\Gamma_N)}}{\|u\|^2_S},$$

in which the fraction equals $1/\mu$ for (1.2.65), hence the equality $\hat{\mu}_i = \frac{1}{\mu_i}$ follows from the variational characterization of the eigenvalues.

Estimate (1.2.66) follows from $\mu_i = O(i^{2/d})$ by an elementary calculation. For Dirichlet boundary conditions, this asymptotic behaviour is found in [38]. ■

Nonsymmetric equivalent preconditioners. If the original problem has large nonsymmetric (first-order) terms, then the symmetric approach may not work satisfactorily and it may still be advisable to include nonsymmetric terms in the preconditioning operator. We briefly outline the general case and mention two examples, based on [18, 19]. Let us consider the nonsymmetric elliptic equation (1.2.57) with Laplacian principal part:

$$\left\{ \begin{array}{l} Lu := -\Delta u + \mathbf{b} \cdot \nabla u + cu = g \\ u_{|\Gamma_D} = 0, \; \frac{\partial u}{\partial \nu} + \alpha u_{|\Gamma_N} = 0 \end{array} \right. \tag{1.2.67}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, where L satisfies Assumptions 1.2.1 and $g \in L^2(\Omega)$. As before, we are interested in FEM discretization. Let us introduce the following type of nonsymmetric preconditioning operator:

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

for some properly chosen functions $\mathbf{w}, z, \eta$, such that N satisfies Assumptions 1.2.1 in the obvious sense.

Theorem 1.2.8 *The CGN algorithm for the preconditioned system $\mathbf{N}_h^{-1}\mathbf{L}_h \mathbf{c} = \tilde{\mathbf{b}}_h$ yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \varepsilon_k \qquad (k = 1, 2, ..., n) \tag{1.2.68}$$

$$\textit{where} \quad \varepsilon_k = \frac{2M_N^2}{km_L^2} \sum_{i=1}^{k} \Big(\frac{2}{m_N} s_i(Q_S) + \frac{1}{m_N^2} s_i(Q_S)^2 \Big) \to 0 \qquad (\textit{as } k \to \infty) \tag{1.2.69}$$

and ε_k is a sequence independent of V_h.

PROOF. Similar to that of Theorem 1.2.6, now Theorem 1.2.4 is applied in $H^1_D(\Omega)$.

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:

$$Nu = -\Delta u + \mathbf{w} \cdot \nabla u + zu \qquad \text{for } u \in H^2(\Omega): \; u_{|\Gamma_D} = 0, \; \frac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \tag{1.2.70}$$

where $\mathbf{w} \in \mathbf{R}^d$, $z, \eta \geq 0$ are constants such that $z > 0$ or $\eta > 0$ if $\Gamma_D = \emptyset$. Then separable solvers are available for N, see [113, 116].

The preconditioning operator (1.2.70) can be further simplified if one convection coefficient is dominating [16]. Assume that, say, $b_1(x)$ has considerably larger values than $b_j(x)$ $(j \geq 2)$. Then one can include only one nonsymmetric coefficient, i.e. propose the preconditioning operator

$$Nu = -\Delta u + w_1 \tfrac{\partial u}{\partial x_1} + zu \qquad \text{for } \ u \in H^2(\Omega): \ u_{|\Gamma_D} = 0, \ \tfrac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \qquad (1.2.71)$$

where $w_1, z, \eta \in \mathbf{R}$ have the same properties as required for (1.2.70). 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 [12].

(b) Elliptic systems

We consider convection-diffusion type systems, coupled via the zeroth order terms. (Stokes type systems will be mentioned in subsection 1.5.6.) 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.

Decoupled symmetric preconditioners for convection-diffusion-reaction systems have been developed in our paper [95], and extended to parallel computers in [96]. Let us summarize this briefly.

We consider an elliptic system

$$\left.\begin{aligned} L_i u \equiv -\mathrm{div}\,(A_i \nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + \sum_{j=1}^{l} V_{ij} u_j = g_i \\ u_{i\,|\Gamma_D} = 0, \qquad \tfrac{\partial u_i}{\partial \nu_{A_i}} + \alpha_i u_{i\,|\Gamma_N} = 0 \end{aligned}\right\} \qquad (i = 1, \dots, l) \qquad (1.2.72)$$

where Ω, A_i and α_i are as in Assumptions 1.2.1, $\mathbf{b}_i \in W^{1,\infty}(\Omega)^d$, $g_i \in L^2(\Omega)$, $V_{ij} \in L^\infty(\Omega)$. We assume that $\mathbf{b}_i$ and the matrix $V = \{V_{ij}\}_{i,j=1}^{l}$ satisfy the coercivity property

$$\lambda_{min}(V + V^T) - \max_i \mathrm{div}\, \mathbf{b}_i \geq 0 \qquad (1.2.73)$$

a.e. pointwise on Ω, where λ_{min} denotes the smallest eigenvalue. These conditions imply that the operator

$$L = (L_1, \dots, L_l)$$

is coercive in $H_D^1(\Omega)^l$, hence system (1.2.72) 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.

Let us define the preconditioning operator

$$S = (S_1, \dots, S_l)$$

as the l-tuple of independent operators

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

such that each S_i satisfies Assumptions 1.2.2. The preconditioner for the discrete system is defined as the stiffness matrix $\mathbf{S}_h$ of S in $H^1_D(\Omega)^l$, and we apply the CGN algorithm for the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}. \tag{1.2.74}$$

Theorem 1.2.9 *The CGN algorithm for the preconditioned system (1.2.74) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \ \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{1.2.75}$$

$$\textit{where} \qquad \varepsilon_k := \frac{2}{km^2}\sum_{i=1}^{k}\Big(\big|\lambda_i(Q_S^* + Q_S)\big| + \lambda_i(Q_S^* Q_S)\Big) \ \to 0 \quad \textit{as} \ \ k \to \infty \tag{1.2.76}$$

and ε_k is a sequence independent of V_h.

Proof. Similar to Theorem 1.2.6. Here Q_S arises as a sum analogous to (1.2.61). ■

If Q_S is normal, then one can apply the GCG-LS algorithm to system (1.2.74), and Theorem 1.2.2 yields

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k = 1, ..., n) \quad \text{where} \quad \varepsilon_k := \frac{2}{km}\sum_{j=1}^{k}\big|\lambda_j(Q_S)\big| \ \to 0 \quad \text{as} \quad k \to \infty \tag{1.2.77}$$

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); however, the experiments in [95] 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 1.5.5. We finally note that these results can be obviously extended to uncoupled nonsymmetric preconditioners of the form (1.2.70).

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

As we have seen in subsection 1.2.1, the weak formulation with S-bounded and S-coercive operators allows us to treat the equivalence of operators in an easy form, and also ensures well-posedness. Now we show that this concept also allows us to derive general mesh independent linear convergence results when no compact-equivalence is assumed, with no extra assumption. This is an advantage compared to the somewhat more general setting of Manteuffel et al. [52, 112], since our framework still covers all usual (Dirichlet, Neumann and Robin) boundary conditions, moreover, mesh independent linear convergence will be readily derived for general FEM discretizations. We follow our paper [19].

We note that such a uniform framework can only be given for FEM discretizations, owing to the Hilbert space background. Similar mesh independence results have also been given for FDM discretizations [37, 49, 52, 155], but these only concern rectangular domains where explicit calculations can be done, and are achieved (depending on the concrete preconditioner) with a case-by-case study.

1.3.1 Mesh independent linear convergence in Hilbert space

Let us consider the operator equation (1.1.1), where L is S-bounded and S-coercive in the sense of Definition 1.2.1, and $g \in H$. Using a Galerkin discretization, we want to solve the arising $n \times n$ system (1.2.23).

(a) Symmetric preconditioners

In general, when L is nonsymmetric, we can take again the symmetric coercive operator S from Definition 1.2.1 and introduce the stiffness matrix of S as preconditioner for system (1.2.23), i.e., $\mathbf{S}_h = \left\{ \langle \varphi_i, \varphi_j \rangle_S \right\}_{i,j=1}^{n}$. To solve the preconditioned system

$$\mathbf{S}_h^{-1} \mathbf{L}_h \, \mathbf{c} = \tilde{\mathbf{b}_h}, \tag{1.3.1}$$

one can apply the CG method using the $\mathbf{S}_h$-inner product $\langle ., . \rangle_{\mathbf{S}_h}$. As follows from (1.1.11) and (1.1.12), the convergence estimates depend on the bounds

$$\lambda_0 = \lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) := \inf\{\mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c} : \ \mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c} = 1\}, \qquad \Lambda = \Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h) := \|\mathbf{S}_h^{-1}\mathbf{L}_h\|_{\mathbf{S}_h} \, ,$$

defined as in (1.1.8). Moreover, the convergence factor is determined by the ratio Λ/λ_0.

Proposition 1.3.1 *If the operator L satisfies (1.2.2), then for any subspace $V_h \subset H_S$ the stiffness matrix $\mathbf{L}_h$ satisfies*

$$m \, (\mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c}) \le \mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c}, \qquad |\mathbf{L}_h \, \mathbf{c} \cdot \mathbf{d}| \le M \, \|\mathbf{c}\|_{\mathbf{S}_h} \|\mathbf{d}\|_{\mathbf{S}_h} \qquad (\mathbf{c}, \mathbf{d} \in \mathbf{R}^n) \tag{1.3.2}$$

where m and M come from (1.2.2) and hence are independent of V_h.

Proof. Set $u = \sum\limits_{i=1}^{n} c_i \varphi_i \ \in V_h$ and $v = \sum\limits_{j=1}^{n} d_j \varphi_j \ \in V_h$ in (1.2.2). Here

$$\langle L_S u, v \rangle_S = \sum_{i,j=1}^{n} \langle L_S \varphi_i, \varphi_j \rangle_S \, c_i d_j = \sum_{i,j=1}^{n} (\mathbf{L}_h)_{ji} c_i d_j = \mathbf{L}_h \, \mathbf{c} \cdot \mathbf{d}$$

and similarly

$$\|u\|_S^2 = \sum_{i,j=1}^{n} \langle \varphi_i, \varphi_j \rangle_S \, c_i c_j = \mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c} = \|\mathbf{c}\|_{\mathbf{S}_h}^2,$$

which show that (1.2.2) implies (1.3.2). ■

Thus we obtain that for any subspace $V_h \subset H_S$

$$\Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h) \le M, \qquad \lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) \ge m \tag{1.3.3}$$

independently of V_h. Then, using (1.1.11), we have proved

Theorem 1.3.1 *Let the operator L satisfy (1.2.2). Then the GCG-LS method for for system (1.3.1) provides*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \left(1-\left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k=1,2,...,n) \tag{1.3.4}$$

independently of V_h, and the CGN algorithm satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le 2^{1/k}\,\frac{M-m}{M+m} \qquad (k=1,2,...,n) \tag{1.3.5}$$

independently of V_h.

We note that (1.3.4) 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 the spectral equivalence relation

$$m\|u\|_S^2 \le \langle L_S u, u\rangle_S \le M\|u\|_S^2 \qquad (u \in H_S). \tag{1.3.6}$$

Then $\mathbf{L}_h$ is symmetric too. Let S be the symmetric coercive operator from Definition 1.2.1, and introduce the stiffness matrix of S. It immediately follows, see e.g. [8], that

$$\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \le \frac{M}{m}. \tag{1.3.7}$$

(b) Relation to previous conditions

Now we can clarify the relation of our setting to that by Manteuffel et al in [52]. 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 (1.2.2)) can be replaced by the two weaker statements

$$\sup_{v\in H_S}\frac{\langle L_S u, v\rangle_S}{\|v\|_S} \ge m\|u\|_S \quad (u\in H_S), \qquad \sup_{u\in H_S}\langle L_S u, v\rangle_S > 0 \quad (v\in H_S). \tag{1.3.8}$$

However, in contrast to (1.2.2), the above inequalities are not automatically inherited in general subspaces V_h with the same constants, i.e., no analogue of Proposition 1.3.1 holds. Instead, the corresponding uniform relations for the discrete operators had to be assumed there, see (3.37)-(3.38) in [52]; with our notations, this means that one has to assume

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

with a uniform constant $\tilde{m}>0$ to obtain mesh independent linear convergence. (The first bound is an LBB type condition.) Although our assumptions (1.2.2) are more special, they hold for rather general elliptic operators as shown by Proposition 1.2.2, 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 (1.1.1). We assume that N is S-bounded and S-coercive, i.e. $N \in BC_S(H)$ in the sense of Definition 1.2.1, for the same symmetric operator S as is L. Then we introduce the stiffness matrix of N_S, i.e. $\mathbf{N}_h = \left\{ \langle N_S \varphi_j, \varphi_i \rangle_S \right\}_{i,j=1}^{n}$, as preconditioner for the discretized system (1.2.23). To solve the preconditioned system

$$\mathbf{N}_h^{-1} \mathbf{L}_h \, \mathbf{c} = \tilde{\mathbf{b}_h} \tag{1.3.9}$$

(with $\tilde{\mathbf{b}_h} = \mathbf{N}_h^{-1} \mathbf{b}_h$), we apply the CGN method under the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$. By (1.1.12), this algorithm converges as

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq 2^{1/k} \, \frac{\kappa(\mathbf{N}_h^{-1} \mathbf{L}_h) - 1}{\kappa(\mathbf{N}_h^{-1} \mathbf{L}_h) + 1} \qquad (k = 1, 2, ..., n). \tag{1.3.10}$$

In the convergence analysis of nonsymmetric preconditioners, we must distinguish between the bounds of L and N, i.e., (1.2.2) is replaced by

$$\begin{aligned} m_L \|u\|_S^2 \leq \langle L_S u, u \rangle_S, \quad |\langle L_S u, v \rangle_S| \leq M_L \|u\|_S \|v\|_S, \\ m_N \|u\|_S^2 \leq \langle N_S u, u \rangle_S, \quad |\langle N_S u, v \rangle_S| \leq M_N \|u\|_S \|v\|_S \end{aligned} \tag{1.3.11}$$

for all $u, v \in H_S$.

Theorem 1.3.2 *If the operators L and N satisfy (1.3.11), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{N}_h^{-1} \mathbf{L}_h) \leq \frac{M_L M_N}{m_L m_N} \qquad \textit{and} \qquad \kappa(\mathbf{N}_h^{-1} \mathbf{L}_h) \leq \left(1 + \frac{m_L + m_N}{2 m_L m_N} \, \|L_S - N_S\| \right)^2 \tag{1.3.12}$$

independently of V_h.

Proof. (i) Let $\mathbf{c} \in \mathbf{R}^n$ be arbitrary, $\mathbf{d} := \mathbf{N}_h^{-1} \mathbf{L}_h \mathbf{c}$, i.e. $\mathbf{N}_h \mathbf{d} = \mathbf{L}_h \mathbf{c}$, further, let $u = \sum\limits_{j=1}^{n} c_j \varphi_j \ \in V_h$ and $z = \sum\limits_{j=1}^{n} d_j \varphi_j \ \in V_h$. Then

$$m_L \|u\|_S^2 \leq \langle L_S u, u \rangle_S = \mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c} = \mathbf{N}_h \, \mathbf{d} \cdot \mathbf{c} = \langle N_S z, u \rangle_S \leq \|N_S z\|_S \|u\|_S,$$

hence $m_L \|u\|_S \leq \|N_S z\|_S \leq M_N \|z\|_S$, and by exchanging L and N resp. u and z, we similarly obtain $m_N \|z\|_S \leq \|L_S u\|_S \leq M_L \|u\|_S$. Hence, altogether,

$$\frac{m_L}{M_N} \leq \frac{\|\mathbf{N}_h^{-1} \mathbf{L}_h \mathbf{c}\|_{\mathbf{S}_h}}{\|\mathbf{c}\|_{\mathbf{S}_h}} = \frac{(\mathbf{S}_h \, \mathbf{d} \cdot \mathbf{d})^{1/2}}{(\mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c})^{1/2}} = \frac{\|z\|_S}{\|u\|_S} \leq \frac{M_L}{m_N} \, . \tag{1.3.13}$$

(ii) We follow the proof of Proposition 1.3.2. Let $\mathbf{c},\mathbf{d}\in\mathbf{R}^n$ and $u,z\in V_h$ be as therein, $\mathbf{k}:=\mathbf{d}-\mathbf{c}$ and $h:=\sum\limits_{j=1}^{n}k_j\varphi_j=z-u$. Then

$$\begin{aligned}m_N\|h\|_S^2&\le\langle N_S h,h\rangle_S=\mathbf{N}_h\,\mathbf{k}\cdot\mathbf{k}=\mathbf{N}_h\,\mathbf{d}\cdot\mathbf{k}-\mathbf{N}_h\,\mathbf{c}\cdot\mathbf{k}=(\mathbf{L}_h-\mathbf{N}_h)\,\mathbf{c}\cdot\mathbf{k}\\&=\langle(L_S-N_S)u,h\rangle_S\le\|L_S-N_S\|\,\|u\|_S\|h\|_S.\end{aligned}$$

Hence

$$\|z\|_S\le\|u\|_S+\|h\|_S\le\|u\|_S\Big(1+\frac{1}{m_N}\|L_S-N_S\|\Big).$$

Exchanging L and N resp. u and z, we obtain $\|u\|_S\le\|z\|_S\Big(1+\frac{1}{m_L}\|L_S-N_S\|\Big)$. In view of (1.3.13), the obtained bounds on the ratio $\|z\|_S/\|u\|_S$ imply

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)\le\Big(1+\frac{1}{m_N}\|L_S-N_S\|\Big)\Big(1+\frac{1}{m_L}\|L_S-N_S\|\Big)\le\Big(1+\frac{m_L+m_N}{2m_Lm_N}\,\|L_S-N_S\|\Big)^2$$

where the second estimate uses the arithmetic-geometric mean inequality. ■

Hence, by (1.3.10), the CGN algorithm converges with a ratio bounded independently of V_h. Note that the above first estimate is a direct extension of the case of symmetric preconditioners: the latter is recovered by the case $N=S$, for which $M_N=m_N=1$. However, if both N and L have a large ratio M/m, then the upper bound in (1.3.12) becomes large even if N is an accurate approximation of L. In this case it is more useful to involve the difference of N and L in the bound, as done in the second estimate above.

1.3.2 Mesh independent linear convergence for elliptic problems

Let us consider again the nonsymmetric elliptic problem (1.2.57), i.e.,

$$\left\{\begin{array}{l}Lu:=-\mathrm{div}\,(A\,\nabla u)+\,\mathbf{b}\cdot\nabla u+cu=g\\ u_{|\Gamma_D}=0,\ \frac{\partial u}{\partial\nu_A}+\alpha u_{|\Gamma_N}=0\end{array}\right.\tag{1.3.14}$$

on a bounded domain $\Omega\subset\mathbf{R}^d$, and we assume that L satisfies Assumptions 1.2.1. As a preconditioning operator, we consider in general a symmetric elliptic operator S introduced in (1.2.7):

$$Su\equiv\ -\mathrm{div}\,(G\,\nabla u)+\sigma u\qquad\text{for}\ \ u_{|\Gamma_D}=0,\ \frac{\partial u}{\partial\nu_G}+\beta u_{|\Gamma_N}=0,\tag{1.3.15}$$

assumed to satisfy Assumptions 1.2.2. Now, in contrast to section 1.2.4, we allow in general $A\ne G$. We introduce the stiffness matrix $\mathbf{S}_h$ of S as preconditioner for system (1.2.58), and then solve the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c}=\tilde{\mathbf{g}_h}$ (with $\tilde{\mathbf{g}_h}=\mathbf{S}_h^{-1}\mathbf{g}_h$) with a CG algorithm. The basic conditioning estimate is as follows:

Proposition 1.3.2 *For the system* $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c}=\tilde{\mathbf{g}_h}$, *the bounds (1.1.8) satisfy*

$$\Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h)\le M,\qquad\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h)\ge m\tag{1.3.16}$$

independently of V_h, where

$$\begin{aligned} M &:= p_1 + C_{\Omega,S}\, q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C_{\Omega,S}^2\|c\|_{L^\infty(\Omega)} + C_{\Gamma_N,S}^2\|\alpha\|_{L^\infty(\Gamma_N)}\ , \\ m &:= \Big(p_0^{-1} + C_{\Omega,L}^2\|\sigma\|_{L^\infty(\Omega)} + C_{\Gamma_N,L}^2\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}\ . \end{aligned} \tag{1.3.17}$$

Proof. It follows from (1.2.16) and (1.3.3). ■

Using (1.1.11), we have thus proved

Theorem 1.3.3 *For system (1.2.60), the GCG-LS algorithm satisfies*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \left(1 - \left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k = 1, 2, ..., n), \tag{1.3.18}$$

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

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le 2^{1/k}\ \frac{M-m}{M+m} \qquad (k = 1, 2, ..., n), \tag{1.3.19}$$

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 [113, 116, 146]. The results can be extended to suitable systems, see as an example the Navier system (1.5.12) and the procedure described there.

Finally, Theorem 1.3.2 can be used when a nonsymmetric preconditioner is applied, such as an operator with constant coefficients (1.2.70) for an equation (1.3.14) with a variable diffusion coefficient.

1.4 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 [37] (see further analysis in [16, 49, 156]). Here

$$\mathbf{S}_h := \frac{1}{2}(\mathbf{L}_h + \mathbf{L}_h^T), \quad \mathbf{Q}_h := \frac{1}{2}(\mathbf{L}_h - \mathbf{L}_h^T) \tag{1.4.1}$$

are the symmetric and antisymmetric parts of $\mathbf{L}_h$, respectively. The main advantage of symmetric part preconditioning is a simplified CG algorithm. As shown in [7], the full GCG-LS algorithm then reduces to the truncated version GCG-LS(0) that requires a very simple recurrence: it uses a single, namely the current search direction.

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 1.2. We assume for the discussion that H is complex and there exists $p > 0$ such that

$$\mathrm{Re}\langle Lu, u\rangle \geq p\|u\|^2 \quad (u \in D := D(L)). \tag{1.4.2}$$

1.4.1 Strong symmetric part and mesh independent convergence

(a) Construction and general convergence results

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:

$$Su = \frac{1}{2}(Lu + L^*u), \qquad Qu := \frac{1}{2}(Lu - L^*u) \qquad (u \in D). \tag{1.4.3}$$

Further, we impose the following conditions:

Assumptions 1.4.1. 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 1.4.1 *Let H be a complex Hilbert space. Let L satisfy (1.4.2) and $D(L) = D(L^*)$, further, assume that Assumptions 1.4.1 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

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{\|S^{-1}Q\|}{\sqrt{1 + \|S^{-1}Q\|^2}} \qquad (k = 1, 2, ..., n). \tag{1.4.4}$$

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

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \quad (k = 1, ..., n), \quad \textit{where} \quad \varepsilon_k := \frac{2}{k}\sum_{j=1}^{k}\left|\lambda_j(S^{-1}Q)\right| \to 0 \quad \textit{as} \quad k \to \infty \tag{1.4.5}$$

and ε_k is a sequence independent of V_h.

Proof. (1) Let us consider the preconditioned system

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

Let $\mathbf{A}_h := \mathbf{S}_h^{-1}\mathbf{L}_h$ and $\mathbf{E}_h := \mathbf{S}_h^{-1}\mathbf{Q}_h$, i.e. we have $\mathbf{A}_h = \mathbf{I}_h + \mathbf{E}_h$. Since $\mathbf{E}_h$ is antisymmetric w.r.t the $\mathbf{S}_h$-inner product, we have $\langle \mathbf{A}_h\mathbf{c}, \mathbf{c}\rangle_{\mathbf{S}_h} = \|\mathbf{c}\|^2_{\mathbf{S}_h}$ for all $\mathbf{c}$, hence $\lambda_0 = 1$. Since $\mathbf{A}_h$ is normal and $\mathbf{E}_h$ has imaginary eigenvalues, we have $\Lambda^2 = \|\mathbf{A}_h\|^2_{\mathbf{S}_h} = |\lambda_{max}(\mathbf{A}_h)|^2 = 1 + |\lambda_{max}(\mathbf{E}_h)|^2 = 1 + \|\mathbf{E}_h\|^2_{\mathbf{S}_h}$. That is,

$$1 - (\lambda_0/\Lambda)^2 = \|\mathbf{E}_h\|^2_{\mathbf{S}_h}/(1 + \|\mathbf{E}_h\|^2_{\mathbf{S}_h}),$$

hence (1.1.11) yields that the GCG-LS(0) algorithm converges with rate $\|\mathbf{E}_h\|_{\mathbf{S}_h}/\sqrt{1+\|\mathbf{E}_h\|^2_{\mathbf{S}_h}}$. Now, similarly to (1.3.3), we have the estimate

$$\|\mathbf{E}_h\|_{\mathbf{S}_h} := \|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_{\mathbf{S}_h} \le \|Q_S\|,$$

hence we altogether obtain

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_{\mathbf{S}_h}}{\sqrt{1+\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|^2_{\mathbf{S}_h}}} \le \frac{\|Q_S\|}{\sqrt{1+\|Q_S\|^2}} \qquad (k=1,2,...,n). \tag{1.4.6}$$

Since $Q_S = S^{-1}Q$, we have obtained (1.4.4).

(2) This estimate follows from Theorem 1.2.2, since $Q_S = S^{-1}Q$ is an antisymmetric operator in H_S and $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is an antisymmetric matrix w.r.t the $\mathbf{S}_h$-inner product, hence they are also normal. ■

The above situation is applicable to Dirichlet problems as a special case of (1.3.14):

$$\begin{cases} Lu := -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g \\ u_{|\partial\Omega} = 0, \end{cases} \tag{1.4.7}$$

where we assume that L satisfies Assumptions 1.2.1. 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

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \hat{c}u \qquad \text{for} \;\; u_{|\partial\Omega} = 0, \tag{1.4.8}$$

where $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}$. Since L satisfies Assumptions 1.2.1, we just obtain that $\sigma := \hat{c} \ge 0$, and hence (together with the above assumptions) S satisfies Assumptions 1.2.2.

Theorem 1.4.2 *Let the operator L in (1.4.7) satisfy Assumptions 1.2.1. and the Kadlec conditions. Let S be the operator (1.4.8). 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 (1.4.5).*

Proof. Let $D(L) := H^2(\Omega) \cap H^1_0(\Omega)$, then $D(L) = D(L^*)$. Here L and S satisfy Assumptions 1.4.1, in particular, $R(S) = H$ follows from $D(L) = H^2(\Omega) \cap H^1_0(\Omega)$, using [78], further, the compactness of $S^{-1}Q = Q_S$ follows from the compact embedding of $H^1_D(\Omega)$ to $L^2(\Omega)$ as in the proof of Theorem 1.2.1. Hence we can apply statement (2) of Theorem 1.4.1. ■

(b) The superlinear convergence rate for problems with constant coefficients

The superlinear convergence rate can be shown to be $O\big(\frac{1}{\sqrt{k}}\big)$ for 2D problems with constant coefficients. Namely, let us consider the following special case of problem (1.4.7) on a bounded domain $\Omega \subset \mathbf{R}^2$:

$$\begin{cases} -\Delta u + \mathbf{b}\cdot\nabla u + cu = g \\ u_{|\partial\Omega} = 0 \end{cases} \tag{1.4.9}$$

where $\mathbf{b}=(b_1,b_2)\in\mathbf{R}^2$, $c\in\mathbf{R}$, $c\geq 0$ and $g\in L^2(\Omega)$. The symmetric and antisymmetric part operators become

$$Su=-\Delta u+cu, \qquad Qu=\mathbf{b}\cdot\nabla u.$$

First we prove a "domain independence principle". We consider the GCG-LS method for problems on arbitrary subdomains Ω' of a given domain Ω, and our goal is to estimate the convergence uniformly in Ω'. Therefore we must indicate the dependence on Ω' whenever necessary. In particular, for a given FEM subspace $V_h(\Omega')\subset H_0^1(\Omega')$, we denote by $\big(\mathbf{S}_h\big)(\Omega')$ and $\big(\mathbf{L}_h\big)(\Omega')$ the stiffness matrices of L and S on Ω', further, let $\mathbf{Q}_h(\Omega')=\mathbf{L}_h(\Omega')-\mathbf{S}_h(\Omega')$. Then we consider the preconditioned system

$$\mathbf{S}_h(\Omega')^{-1}\mathbf{L}_h(\Omega')\,\mathbf{c}=\tilde{\mathbf{b}}. \tag{1.4.10}$$

The FEM subspace $V_h(\Omega')$ is called a normal discretization if the corresponding matrix $\mathbf{S}_h(\Omega')^{-1}\mathbf{Q}_h(\Omega')$ is $\mathbf{S}_h(\Omega')$-normal. We first need the following

Proposition 1.4.1 [65]. *Let $B:H\to H$ be a compact linear operator. Then for all $k\in\mathbf{N}^+$*

$$\sum_{j=1}^{k}s_j(B)=\max_{U,\,u_1,\dots,u_k}\sum_{j=1}^{k}|\langle UBu_j,u_j\rangle| \tag{1.4.11}$$

where the maximum is taken for all unitary operators U on H and all orthonormal vectors $u_1,...,u_k$ in H.

Corollary 1.4.1 *Let $B:H\to H$ be a compact linear operator. Then for all $k\in\mathbf{N}^+$*

$$\sum_{j=1}^{k}s_j(B)=\max\Big\{\sum_{j=1}^{k}|\langle Bu_j,v_j\rangle|:\ u_1,...,u_k,\,v_1,...,v_k\in H,\ \langle u_i,u_j\rangle=\langle v_i,v_j\rangle=\delta_{ij}\Big\}.$$

Proof. Set $u_j=Uv_j$ in (1.4.11). ■

Lemma 1.4.1 *Let $\Omega'\subset\Omega$ be an arbitrary subdomain, and let us define the spaces*

$$H_S=H_0^1(\Omega)\ \textit{with inner product}\ \ \langle u,v\rangle_S=\int_\Omega(\nabla u\cdot\nabla\overline{v}+cu\overline{v}),$$
$$H_{S'}=H_0^1(\Omega')\ \textit{with inner product}\ \ \langle u,v\rangle_{S'}=\int_{\Omega'}(\nabla u\cdot\nabla\overline{v}+cu\overline{v})$$

and denote by $s_i(Q_S)$ and $s_i'(Q_S)$ the singular values of the operator Q_S on $H_0^1(\Omega)$ and $H_0^1(\Omega')$, respectively. Then for all $k\in\mathbf{N}^+$

$$\sum_{i=1}^{k}s_i'(Q_S)\leq\sum_{i=1}^{k}s_i(Q_S). \tag{1.4.12}$$

Proof. Note that for any $u \in H_0^1(\Omega')$ the function $\hat{u}$ defined by

$$\hat{u} :\equiv \begin{cases} u & \text{on } \Omega' \\ 0 & \text{on } \Omega \setminus \Omega' \end{cases}$$

satisfies $u \in H_0^1(\Omega)$, further, for any $u, v \in H_S$, $\langle Q_S u, v\rangle_S = \langle Qu, v\rangle_{L^2(\Omega)} = \int_\Omega (\mathbf{b} \cdot \nabla u)\overline{v}$ and similarly in $H_{S'}$. Applying Corollary 1.4.1 for the operator Q_S in the spaces H_S and $H_{S'}$, we obtain

$$\sum_{i=1}^{k} s_i'(Q_S) = \max\Big\{ \sum_{i=1}^{k} \Big| \int_{\Omega'} (\mathbf{b} \cdot \nabla u_i)\overline{v_i} \Big| : \ u_i,\, v_i \in H_0^1(\Omega'),\ \langle u_i, u_j\rangle_{S'} = \langle v_i, v_j\rangle_{S'} = \delta_{ij} \Big\}$$

$$= \max\Big\{ \sum_{i=1}^{k} \Big| \int_{\Omega} (\mathbf{b} \cdot \nabla \hat{u}_i)\overline{\hat{v}_i} \Big| : \ \hat{u}_i,\, \hat{v}_i \in H_0^1(\Omega),\ \hat{u}_{i\,|\Omega\setminus\Omega'} \equiv \hat{v}_{i\,|\Omega\setminus\Omega'} \equiv 0,\ \langle \hat{u}_i, \hat{u}_j\rangle_{S} = \langle \hat{v}_i, \hat{v}_j\rangle_{S} = \delta_{ij} \Big\}$$

$$\leq \max\Big\{ \sum_{i=1}^{k} \Big| \int_{\Omega} (\mathbf{b} \cdot \nabla u_i)\overline{v_i} \Big| : \ u_i,\, v_i \in H_0^1(\Omega),\ \langle u_i, u_j\rangle_{S} = \langle v_i, v_j\rangle_{S} = \delta_{ij} \Big\} = \sum_{i=1}^{k} s_i(Q_S). \qquad \blacksquare$$

Corollary 1.4.2 *Let $\Omega \subset \mathbf{R}^n$ be a given domain. Then for any subdomain $\Omega' \subset \Omega$ and for any normal discretization $V_h(\Omega') \subset H_0^1(\Omega')$, the GCG-LS algorithm yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h(\Omega')}}{\|r_0\|_{\mathbf{S}_h(\Omega')}} \right)^{1/k} \leq \varepsilon_k, \qquad \textit{where} \quad \varepsilon_k = \frac{2}{mk} \sum_{i=1}^{k} |\lambda_i(Q_S)| \to 0 \qquad (\textit{as } k \to \infty). \quad (1.4.13)$$

Proof. We apply Theorem 1.2.2, Lemma 1.4.1, and the fact that Q_S is normal (being antisymmetric), which implies $s_i(Q_S) = \lambda_i(Q_S)$ for all $i \in \mathbf{N}^+$. Then

$$\left(\frac{\|r_k\|_{\mathbf{S}_h(\Omega')}}{\|r_0\|_{\mathbf{S}_h(\Omega')}} \right)^{1/k} \leq \frac{2}{mk} \sum_{i=1}^{k} s_i'(Q_S) \leq \frac{2}{mk} \sum_{i=1}^{k} s_i(Q_S) = \frac{2}{mk} \sum_{i=1}^{k} \lambda_i(Q_S). \qquad \blacksquare$$

Theorem 1.4.3 *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*

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

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

Proof. We have shown the desired estimate $\frac{C}{\sqrt{k}}$ on the unit square by an elementary but tedious calculation, using that the eigenvalues satisfy

$$|\lambda_{ij}^{\pm}| = \frac{|b|}{2\sqrt{c + \pi^2(i^2 + j^2)}}$$

and estimating ε_k via the integral of the function $(c + \pi^2(x^2 + y^2))^{-\frac{1}{2}}$ on the quarter of a disc, see [16, sec. 3.4] for details. Then, for any rectangle the estimate is preserved by transforming the eigenfunctions by the linear one-to-one mapping of the unit square onto the rectangle. Finally, using Corollary 1.4.2, the estimate remains valid for all subdomains of rectangles, i.e. for all bounded domains. $\blacksquare$

1.4.2 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 (1.4.3) may have no 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. This is the case when an elliptic problem has mixed boundary conditions.

(a) Construction of the weak symmetric part

We will define a sesquilinear form that corresponds to the symmetric part of L. This is based on the following

Proposition 1.4.2 *Let* $\operatorname{Re}\langle Lu, u\rangle > 0 \quad (u \in D(L))$. *The formula*

$$\langle u, v\rangle_S := \frac{1}{2}\Big(\langle Lu, v\rangle + \langle u, Lv\rangle\Big) \qquad (u, v \in D(L)) \tag{1.4.14}$$

on $D(L)$ defines an inner product, which will be called the weak symmetric part of L.

Proof. The facts that $\langle ., .\rangle_S$ is sesquilinear (i.e., linear and conjugate linear in the first and second variables, respectively) and conjugate symmetric (i.e., $\langle v, u\rangle_S = \overline{\langle u, v\rangle_S}$) follow directly from its definition and the same properties of the inner product $\langle ., .\rangle$. Further,

$$\|u\|_S^2 = \langle u, u\rangle_S = \operatorname{Re}\langle Lu, u\rangle > 0 \qquad (u \in D(L),\ u \neq 0), \tag{1.4.15}$$

hence $\langle ., .\rangle_S$ is positive definite. ■

Consequently, we can define the corresponding Hilbert space:

Definition 1.4.1 The space H_S is the completion of $D(L)$ w.r.t. the inner product $\langle ., .\rangle_S$.

Remark 1.4.1 If there exists a dense subspace $D \subset H$ and a strongly positive operator $S : D \to H$ such that its energy space coincides with the above space H_S, then we can say that S *represents the symmetric part* of L. Clearly, such an operator is not unique, e.g. the restriction of such an operator to any H_S-dense subspace of D also generates H_S. (One may obviously define the maximal domain D to consist of those $u \in H$ for which there exists $u^* \in H$ satisfying $\langle u^*, v\rangle = \langle u, v\rangle_S$ for all $v \in H_S$, and then $Su := u^*$. However, in practice the maximal domain would be hard and unnecessary to determine.) Note that the domain D of S need not be the same as $D(L)$, hence S is not the symmetric part of L in the classical sense.

In addition to $\langle ., .\rangle_S$, we need to define the sesquilinear form corresponding to L on H_S, and the operator Q_S that will replace $S^{-1}Q$.

Proposition 1.4.3 *Assume that L is S-bounded. Then*
(1) there exists a unique bounded sesquilinear form on H_S satisfying

$$\langle u, v\rangle_L = \langle Lu, v\rangle \qquad (u, v \in D(L)); \tag{1.4.16}$$

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

$$\langle Q_S u, v\rangle_S := \frac{1}{2}\Big(\langle u, v\rangle_L - \overline{\langle v, u\rangle_L}\Big) \qquad (\forall v \in H_S). \tag{1.4.17}$$

Further, we have

$$\langle u, v\rangle_L = \langle u, v\rangle_S + \langle Q_S u, v\rangle_S \qquad (u, v \in H_S). \tag{1.4.18}$$

Proof. (1) is obvious, namely, $\langle u, v\rangle_L := \langle L_S u, v\rangle_S$. For (2), let us fix $u \in H_S$ and define the linear functional $\phi_u : H_S \to \mathbf{C}$ by $\phi_u v := \frac{1}{2}\Big(\langle u, v\rangle_L - \overline{\langle v, u\rangle_L}\Big)$. Then

$$|\phi_u v| \le \frac{1}{2}\Big(|\langle u, v\rangle_L| + |\langle v, u\rangle_L|\Big) \le M\|u\|_S\|v\|_S\,,$$

hence ϕ_u is bounded in H_S and the Riesz theorem provides an element $Q_S u \in H_S$ satisfying $\phi_u v = \langle Q_S u, v\rangle_S$.

To verify (1.4.18), we note that (1.4.14) implies $\langle u, v\rangle_S = \frac{1}{2}\Big(\langle u, v\rangle_L + \overline{\langle v, u\rangle_L}\Big)$ for all $u, v \in D(L)$. Adding (1.4.17) to this, we obtain (1.4.18). ■

(b) Preconditioning by the weak symmetric part

Our goal is to define the preconditioned form of equation $Lu = g$ by the weak symmetric part and the corresponding PCG algorithm, and then to verify the analogue of Theorem 1.4.1. First note that the weak form of $Lu = g$ is

$$\langle u, v\rangle_L = \langle g, v\rangle \qquad (\forall v \in H_S). \tag{1.4.19}$$

Using (1.4.18), if there is $f \in H_S$ such that $\langle f, v\rangle_S \equiv \langle g, v\rangle$ $(\forall v \in H_S)$, then (1.4.19) becomes

$$(I + Q_S)u = f. \tag{1.4.20}$$

Let us now summarize our conditions (some of which have already been used to develop the above setting). Here we only deal with superlinear convergence, therefore we will include the compactness of Q_S.

Assumptions 1.4.4. L satisfies (1.4.2) and is S-bounded, further, the operator $Q_S : H_S \to H_S$, defined in (1.4.17), is compact on H_S.

The truncated theoretical preconditioned GCG-LS(0) algorithm for equation (1.4.19) is defined such that the arising equations are replaced by their weak forms, i.e. equations $Sr_0 = Lu_0 - g$ and $Sz_k = Ld_k$ are replaced by

$$\langle r_0, v\rangle_S = \langle u_0, v\rangle_L - \langle g, v\rangle \qquad \text{and} \quad \langle z_k, v\rangle_S = \langle d_k, v\rangle_L \qquad (\forall v \in H_S),$$

respectively. In the discrete case that we study, there is no difference in the corresponding algebraic systems related to V_h, and it follows readily that the symmetric part of the matrix $\mathbf{L}_h$ coincides with the stiffness matrix $\mathbf{S}_h$ of S.

Theorem 1.4.4 *Let Assumptions 1.4.4 hold. Then the GCG-LS(0) algorithm applied for the preconditioned system* $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}}_h$ *yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \quad (k=1,...,n) \quad \textit{where} \quad \varepsilon_k := \frac{2}{k}\sum_{j=1}^{k}\left|\lambda_j(Q_S)\right| \to 0 \quad \textit{as} \quad k\to\infty \tag{1.4.21}$$

and ε_k *is a sequence independent of* V_h.

Proof. The proof of Theorem 1.4.2 can be repeated with Q_S. ■

(c) Symmetric part preconditioning for mixed boundary value problems

Let us consider again the nonsymmetric elliptic problem (1.2.57), where the operator L satisfies Assumptions 1.2.1, Then one can calculate easily the weak symmetric part.

Proposition 1.4.4 *The weak symmetric part of* L *is the inner product generated by the preconditioning operator*

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \hat{c}u \qquad \textit{for} \qquad u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_G} + \hat{\alpha}u_{|\Gamma_N} = 0 \tag{1.4.22}$$

in $H_S := H_D^1(\Omega)$*, where* $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}$ *and* $\hat{\alpha} := \alpha + \frac{1}{2}\,(\mathbf{b}\cdot\nu)$.

Proof. The divergence theorem implies

$$\int_\Omega (Lu)\overline{v}\,dx = \int_\Omega \Big(A\,\nabla u\cdot\overline{\nabla}v + (\mathbf{b}\cdot\nabla u)\overline{v} + cu\overline{v}\Big)\,dx + \int_{\Gamma_N} \alpha u\overline{v}\,d\sigma \qquad (u,v\in D(L)) \tag{1.4.23}$$

and

$$\int_\Omega (\mathbf{b}\cdot\nabla u)\overline{v}\,dx = -\int_\Omega u(\mathbf{b}\cdot\nabla\overline{v})\,dx - \int_\Omega (\mathrm{div}\,\mathbf{b})u\overline{v}\,dx + \int_{\Gamma_N}(\mathbf{b}\cdot\nu)\,u\overline{v}\,d\sigma \quad (u,v\in H_D^1(\Omega)). \tag{1.4.24}$$

Then (1.4.23) and (1.4.24) imply

$$\frac{1}{2}\left(\langle Lu,v\rangle_{L^2} + \langle u,Lv\rangle_{L^2}\right) = \int_\Omega \left(A\,\nabla u\cdot\overline{\nabla}v + \big(c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}\big)u\overline{v}\right)\,dx + \int_{\Gamma_N}\left(\alpha + \frac{1}{2}\,(\mathbf{b}\cdot\nu)\right)u\overline{v}\,d\sigma$$

$$= \int_\Omega (A\,\nabla u\cdot\overline{\nabla}v + \hat{c}u\overline{v})\,dx + \int_{\Gamma_N}\hat{\alpha}u\overline{v}\,d\sigma \qquad (u,v\in D(L)), \tag{1.4.25}$$

which is indeed the inner product generated by the operator(1.4.22). Here

$$\mathrm{Re}\langle Lu,u\rangle_{L^2} = \frac{1}{2}\left(\langle Lu,u\rangle_{L^2} + \langle u,Lu\rangle_{L^2}\right) = \int_\Omega (A\,|\nabla u|^2 + \hat{c}|u|^2)\,dx + \int_{\Gamma_N}\hat{\alpha}|u|^2\,d\sigma > 0 \tag{1.4.26}$$

for all $u\in D(L)$, $u\ne 0$, using Assumptions 1.2.1, hence Proposition 1.4.2 can be applied. ■

Note that in general $\hat{\alpha}\ne\alpha$, hence $D(S)\ne D(L)$, i.e. one could not have applied the strong symmetric part framework from section 1.4.1.

Theorem 1.4.5 *Let the operator L in (1.2.57) satisfy Assumptions 1.2.1., and S be the operator (1.4.22). Then the GCG-LS(0) algorithm for the corresponding preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ yields mesh independent superlinear convergence according to (1.4.21).*

Proof. The operator $Q_S : H_S \to H_S$ has the form

$$\langle Q_S u, v\rangle_S = \frac{1}{2}\left(\int_\Omega (\mathbf{b}\cdot\nabla u)\,\overline{v}\,dx\; - \int_\Omega u\,(\mathbf{b}\cdot\nabla\overline{v})\,dx\right) \tag{1.4.27}$$

$$= -\int_\Omega u\,(\mathbf{b}\cdot\nabla\overline{v})\,dx\; - \frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{b})\,u\overline{v}\,dx + \frac{1}{2}\int_{\Gamma_N}(\mathbf{b}\cdot\nu)\,u\overline{v}\,d\sigma, \tag{1.4.28}$$

where (1.4.24) has been used. Hence Q_S is compact, which follows in the same way as in the proof of Theorem 1.2.1. Since, by Proposition 1.2.2, L is S-bounded and S-coercive, hence Assumptions 1.4.4 hold and thus one can apply Theorem 1.4.4. ■

As mentioned before, the main advantage of symmetric part preconditioning is a simplified CG algorithm: the full GCG-LS algorithm reduces to the truncated version GCG-LS(0) that requires a very simple one-step recurrence. By Theorem 1.4.5, one can still achieve mesh independent superlinear convergence by avoiding the normal equation (used in the CGN method).

1.5 Applications to efficient computational algorithms

Based on the above described theory, we present various efficient preconditioners for FEM discretizations of linear PDEs that mostly produce mesh independent superlinear convergence. Computer realization is also included for some of the examples, and always confirms the theoretical convergence results.

Among the other ones let us emphasize here the applications in subsections 1.5.4-1.5.5, where linearized air pollution systems are considered. Namely, for such systems consisting of many equations, the equivalent operator idea can be employed very efficiently, since one can define independent operators for the preconditioner, thereby reducing the size of auxiliary systems to that of a single elliptic equation. Moreover, one can parallelize the computer solution of these independent auxiliary equations.

1.5.1 Helmholtz preconditioner for regular convection-diffusion equations

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

$$\left\{\begin{array}{l} Lu \equiv -\Delta u + \mathbf{b}\cdot\nabla u + cu = g \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial\nu}_{|\Gamma_N} = 0\,, \end{array}\right. \tag{1.5.1}$$

where L satisfies Assumptions 1.2.1; in particular, $\hat{c} := c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0$ in Ω and $\mathbf{b}\cdot\nu \geq 0$ on Γ_N (i.e. Neumann conditions are only imposed on the outflow boundary), further, let $\Gamma_D \neq \emptyset$.

The proposed numerical solution method is some FEM discretization and then a PCGN iteration, where the preconditioner is the stiffness matrix of the Helmholtz 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. Then Theorem 1.2.6 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 [113, 137, 146], hence with an optimal or quasi-optimal number of operations ($O(n)$ or $O(n \log n)$) and thus considerably cheaper than e.g. using MG for the original nonsymmetric problem. We note that a previous study of linear convergence [111] for a similar Dirichlet problem with constant coefficients suggests $\sigma = O(|\mathbf{b}|^2)$ as a good choice.

Numerical experiments have shown that even the GCG-LS(0) method can be applied: the tests in [108] provide mesh independent superlinear convergence.

1.5.2 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, see e.g. [46]. The widespread Uzawa iteration for the Oseen equations defines the consecutive systems

$$\begin{cases} -\nu \Delta \mathbf{u}_k + \mathbf{w} \cdot \nabla \mathbf{u}_k + \nabla p_k = \mathbf{f}, \qquad \mathbf{u}_{k\,|\partial\Omega} = 0 \\ p_{k+1} = p_k + \alpha_k \operatorname{div} \mathbf{u}_k = 0 \end{cases} \tag{1.5.2}$$

(with given initial p_0 and for $k \in \mathbf{N}$), where $\mathbf{w}$, $\mathbf{f}$ are given functions with $\operatorname{div} \mathbf{w} = 0$, and $\alpha_k > 0$ are proper stepsizes. The process can be projected in a proper FEM subspace. Here (1.5.2) means that one must stepwise solve uncoupled auxiliary problems for $\mathbf{u}_k$: namely, if (for simplicity) we neglect k and denote by z and g a given coordinate function of $\mathbf{u}_k$ and $\mathbf{f} - \nabla p_k$, resp., then the auxiliary equations have the form

$$-\nu \Delta z + \mathbf{w} \cdot \nabla z = g, \qquad z_{|\partial\Omega} = 0\,. \tag{1.5.3}$$

These are special convection-diffusion type equations ($\nu >> 0$, e.g. for water $\nu \approx 1$). Since their solution error accumulates during the outer Uzawa iteration, they require an accurate solution. It is thus desirable to have a superlinearly convergent inner iterative method to decrease the cost.

The proposed iterative method for the FEM solution of (1.5.3) is GCG-LS(0) method using symmetric part preconditioning. As mentioned at the beginning of section 1.4, the iteration then reduces to a *simple one-step recurrence form.* Since $\operatorname{div} \mathbf{w} = 0$, it follows from (1.4.8) that the symmetric part operator is

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

Then Theorem 1.4.2 yields *mesh independent superlinear convergence* for the iteration. The auxiliary Possion equations can be solved by various *fast Possion solvers* [113, 116] with an optimal or quasi-optimal number of operations.

1.5.3 Scaling for problems with variable diffusion coefficients

If the diffusion is space-dependent, then the Laplacian is replaced by a variable coefficient diffusion operator. Assuming Dirichlet boundary conditions for simplicity, the convection-diffusion problem (1.5.1) is then replaced by

$$\begin{cases} Lu \equiv -\mathrm{div}\,(a\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g \\ u_{\,|\partial\Omega} = 0\,, \end{cases} \tag{1.5.4}$$

where L satisfies Assumptions 1.2.1 and we assume that $a \in C^2(\overline{\Omega})$, $a(x) \geq m > 0$.

If a fast Poisson or Helmholtz solver is available, then it could only yield linear convergence as a preconditioner for (1.5.4). However, one can still achieve mesh independent superlinear convergence by applying the method of scaling, which was originally introduced for symmetric operators [36, 67]. Namely, let us rewrite our equation as

$$a^{-1/2}Lu = a^{-1/2}g =: \hat{g} \tag{1.5.5}$$

and introduce the new unknown function $v := a^{1/2}u$. Then, by a direct calculation [36], $a^{-1/2}\mathrm{div}\,(a\,\nabla u) + qu \;=\; \Delta v$, where $q = \Delta(a^{1/2})$, which implies that

$$a^{-1/2}Lu = -\Delta v + \text{ lower order terms,}$$

that is, (1.5.5) becomes

$$Nv \equiv -\Delta v + \hat{\mathbf{b}}\cdot\nabla v + \hat{c}v = \hat{g}\,. \tag{1.5.6}$$

Here $\hat{\mathbf{b}} = a^{-1}\mathbf{b}$ and $\hat{c} = a^{-1}c - (1/2a^2)\mathbf{b}\cdot\nabla a + a^{-1/2}\Delta(a^{1/2})$.

The relation $Nv \equiv a^{-1/2}Lu$ shows that

$$\langle Nv, v\rangle_{L^2} = \langle a^{-1/2}Lu,\; a^{1/2}u\rangle_{L^2} = \langle Lu, u\rangle_{L^2}$$

for all $u \in D(L)$ and $v := a^{1/2}u$. Further, using the uniform positivity of a, it is easy to see that the norms $\|u\|_{H^1}$ and $\|v\|_{H^1}$ are equivalent. Therefore N inherits the H^1-coercivity of L, i.e. the relation $\langle Lu, u\rangle_{L^2} \geq m\|u\|^2_{H^1}$ is replaced by $\langle Nv, v\rangle_{L^2} \geq \hat{m}\|v\|^2_{H^1}$ for some other proper constant $\hat{m} > 0$.

This implies that the scaled problem is of type (1.5.1). Hence the algorithm of subsection 1.5.1 can be applied using a Poisson or Helmholtz preconditioning operator, and we have *mesh independent superlinear convergence* of the CGN method.

1.5.4 Decoupled preconditioners for linearized air pollution systems

Air pollution processes are described by compound nonlinear transport systems involving diffusion, convection, reaction and deposition terms. In real-life situations, where there are several chemical species, such systems may consist of a huge number of equations [160].

The standard approach to solve such systems is a time discretization and then a suitable linearization. Then one gets a linear elliptic system

$$\left.\begin{aligned} &-\mathrm{div}\,(K_i\,\nabla u_i) + \mathbf{w}_i\cdot\nabla u_i + \sum_{j=1}^{l} V_{ij}u_j = g_i \\ &u_{i\,|\partial\Omega} = 0, \end{aligned}\right\} \qquad (i = 1,\dots,l) \tag{1.5.7}$$

which is a special case of system (1.2.72). Here $\mathbf{w}_i$ is the effect of wind, and V_{ij} come from the linearized reaction rates. Then the assumptions imposed for (1.2.72) are satisfied, in particular, the coercivity property (1.2.73) can be ensured by choosing a sufficiently small stepsize τ in the time discretization.

To solve this system using FEM and PCG iteration, the equivalent operator idea can be employed very efficiently. Namely, 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. This is a considerable advantage when the elliptic system consists of many equations.

The preconditioning operator is the l-tuple of independent operators

$$S_i u_i := -\mathrm{div}\,(K_i \nabla u) + q_i u \qquad \text{for} \;\; u_{i\,|\partial\Omega} = 0, \qquad (i = 1, \dots, l) \tag{1.5.8}$$

such that each S_i satisfies Assumptions 1.2.2. Then Theorem 1.2.9 is valid for the convergence of the CGN method, i.e., the *mesh independent superlinear convergence* estimates (1.2.75)–(1.2.76) hold, where Q_S now denotes the sum (from 1 to l) of the corresponding operators defined as in (1.4.27). Further, the auxiliary scalar symmetric problems can be solved by some standard direct or multigrid solver, again with an optimal or quasi-optimal number of operations.

We have run numerical tests in [95] for a model problem based on [160], involving 10 equations. The linearization used the previous time layers, and the right-hand sides of the equations also came from the results from the previous time-step. The coefficients V_{ij} arise from chemical reactions, and vary in a large range. The time-step $\tau = 0.2829e - 03$ was chosen sufficiently small to ensure the coercivity property. Further, for suitable balancing different coefficients β_i were chosen, namely, $\beta = \tau \cdot \begin{pmatrix} 1 & 100 & 1 & 10 & 1 & 1 & 1 & 1 & \frac{1}{10} & \frac{1}{100} \end{pmatrix}$. In the first phase of the algorithm the matrices $\mathbf{S}_h$ and $\mathbf{Q}_h$ are constructed. The iterative algorithm solves systems like $\mathbf{S}_h z_h = d_h$ as many times as many iteration step is chosen. To make it faster, the Cholesky decomposition was used instead of $\mathbf{S}_h$ itself.

Table 1.1: Convergence factors for the linearized air pollution system.

	1/h			
Itr.	8	16	32	64
1	0.0073	0.0076	0.0076	0.0077
2	0.0067	0.0071	0.0072	0.0072
3	0.0060	0.0065	0.0066	0.0066
4	0.0054	0.0060	0.0061	0.0061
5	0.0048	0.0054	0.0056	0.0056
6	0.0043	0.0050	0.0052	0.0053

Mesh independent superlinear convergence is seen in Table 1.1. In this experiment the time of computing has also been measured: the run-times for this system can be found below in Table 1.2. The last two colums show the difference between the direct solution and the conjugate gradient method. The numbers in the last column are the total time of the decomposition and the iteration. We may observe that the iteration with solving the block-diagonal symmetric auxiliary problems for a relevant mesh size h was considerably faster than the direct solution with the nonsymmetric full matrix.

Table 1.2: Computational times for the linearized air pollution system.

1/h	creating $\mathbf{S}_h, \mathbf{L}_h$	Cholesky	iteration	direct solution	CGM
8	0.0470	0.0470	0.5780	0.0150	0.6250
16	0.1090	0.0620	1.2350	0.3130	1.2970
32	0.4220	0.1880	3.9680	9.5780	5.8480
64	1.9070	2.3600	17.8120	177.7030	20.1720

The decoupled preconditioners allow efficient parallelization for the solution of the auxiliary systems, due to the block diagonal forms of the matrices. This will be considered in the next subsection.

1.5.5 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 following tests [96]. 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.

The experiments were executed on a Linux cluster consisting of 4 dual processor PowerPCs with G4 450 MHz processors, 512 MB memory per node. The developed parallel code has been implemented in C and the parallelization has been facilitated using the MPI library. The LAPACK library was used for computing the Cholesky factorization of the preconditioner and for solving the linear systems arising in GCG-LS. Times have been collected using the MPI provided timer.

The first test problem is a class of systems of the form (1.5.7) with $l = 2, 3, \ldots, 10$ equations, where diffusion is constant, $\mathbf{b}_i = (1, 0)^T$ and the matrix V is skew-symmetric with elements which are randomly generated constants. Our second test problem comes from the time discretization and linearization of a nonlinear reaction-convection-diffusion system of 10 equations, used in meteorological air-pollution models [160], see the previous subsection. Since the run times here have proved to be very similar to the case of a random 10×10 matrix in the first test problem, we will only present the test results for the first problem.

In our experiments we used a stopping criterion $\|r_k\| \leq 10^{-14}$. Table 1.3 shows the required number of iterations, which is mesh independent. We then studied 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 $E_p = \frac{S_p}{p} \leq 1$. Figure 1.2 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. The reason is the partitioning of the vectors onto the processors.

Table 1.3: Number of iterations in the parallel algorithm.

1/h	l									
	1	2	3	4	5	6	7	8	9	10
8	9	10	11	12	12	12	13	13	14	14
16	9	10	12	12	13	13	13	14	14	14
32	9	10	12	12	13	13	14	14	14	14
64	9	10	12	12	13	13	14	14	14	14
128	9	10	12	12	13	13	14	14	14	14

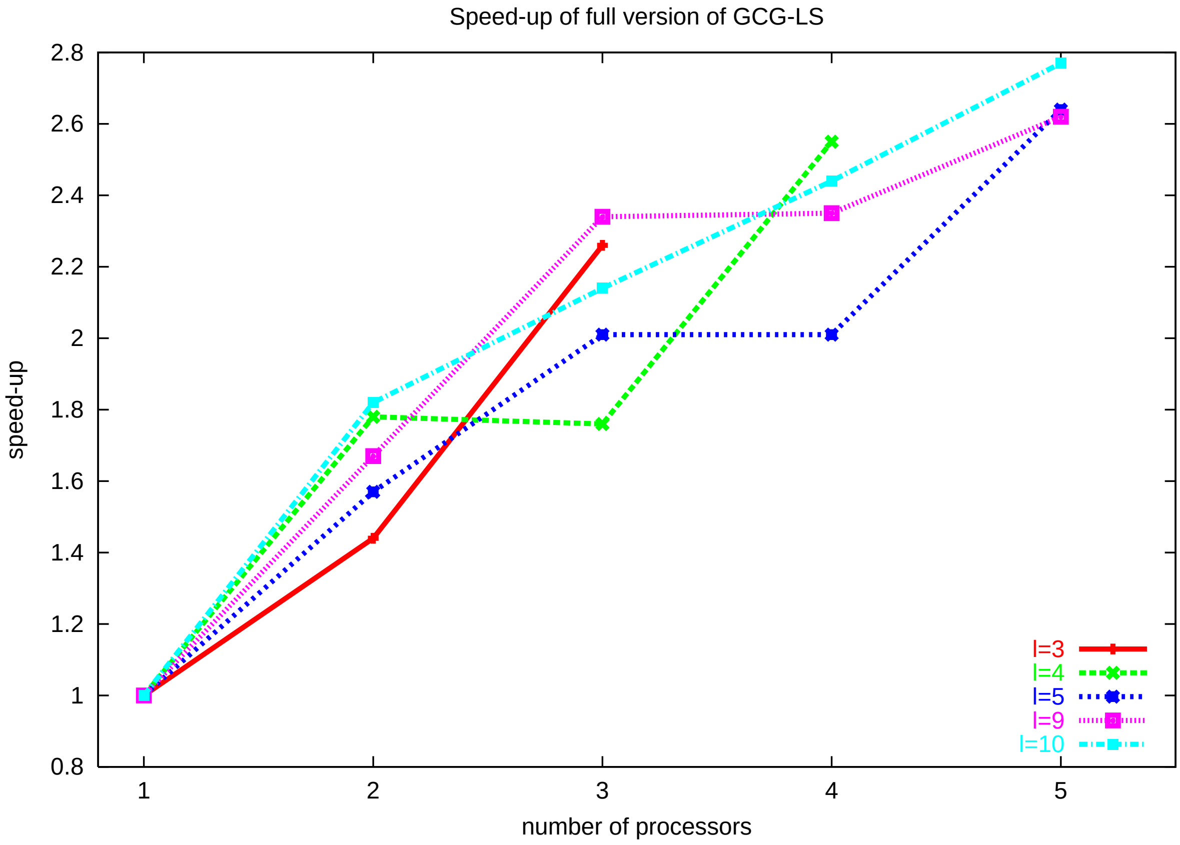

Figure 1.2: Speed-up in the parallel algorithm for different l.

1.5.6 Regularized flow and elasticity problems

(a) Viscous flow: the Stokes problem

A fundamental model of viscous flow is the system of Stokes equations

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f} \\ \operatorname{div} \mathbf{u} = 0 \\ \mathbf{u}_{|\partial\Omega} = 0 \end{cases} \tag{1.5.9}$$

in a bounded domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3) with $\mathbf{f} \in L^2(\Omega)^d$. One looks for the weak solution $(\mathbf{u}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$, where $L_0^2(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p = 0\}$. The numerical solution of this system has been widely investigated and has a vast literature, see e.g. [24].

A crucial issue in the FEM solution is to satisfy the LBB-condition, which restricts the suitable possible pairs of subspaces. Hence an important effort has been done to circumvent the LBB-condition via suitable regularization. A regularized version has been studied in [9], leading to the nonsymmetric algebraic system

$$\mathbf{L}_h \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} \equiv \begin{pmatrix} diag_d(-\Delta_h^0) & \sigma^{-1/2}\,\nabla_h \\ \sigma^{-1/2}\operatorname{div}_h & -\Delta_h^\nu \end{pmatrix} \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \sigma^{-1/2}\operatorname{div}\mathbf{f}_h \end{pmatrix} \tag{1.5.10}$$

where $\sigma > 0$ is a regularization parameter. Then the weak solution lies in $H_S := H_0^1(\Omega)^d \times \dot{H}^1(\Omega)$, where $\dot{H}^1(\Omega) := H^1(\Omega) \cap L_0^2(\Omega)$, and the FEM subspace is chosen in H_S.

Then one can obtain mesh independent superlinear convergence using symmetric part preconditioning. Namely [17], the antisymmetric part of the matrix $\mathbf{L}_h$ comes from the discretizaton of an operator $Q_S : H_S \to H_S$ which is compact and antisymmetric:

$$\left\langle Q_S \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\rangle_S = -\int_\Omega s\,(\operatorname{div}\overline{\mathbf{v}}) + \int_\Omega (\operatorname{div}\mathbf{u})\,\overline{q} \qquad \left(\forall \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \in H_S\right). \tag{1.5.11}$$

Denoting by $\mathbf{S}_h$ the symmetric part of $\mathbf{L}_h$, we can obtain *mesh independent superlinear convergence* using Theorem 1.4.1. Namely, the GCG-LS(0) algorithm for the preconditioned form of (1.5.10) using symmetric part preconditioning yields

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

and thus ε_k depends only on the chosen σ.

(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 is given using the displacement $\mathbf{u}$ and pressure p that satisfy $\operatorname{div}\mathbf{u} = -(1-2\nu)p$, and using the relation $(1-2\nu)(\lambda+\mu) = \mu$ between the Lamé coefficients λ, μ and the Poisson ratio ν, see e.g. [22, 28]. Then

$$\begin{cases} -\Delta\mathbf{u} + \nabla p = \frac{1}{\mu}\,\mathbf{f} \\ \operatorname{div}\mathbf{u} + (1-2\nu)p = 0 \\ \mathbf{u}_{|\partial\Omega} = 0\,. \end{cases} \tag{1.5.12}$$

Here $0 < \nu < \frac{1}{2}$, hence $1 - 2\nu \neq 0$. Except for this term, the system has the same form as the Stokes equations. One looks again for the weak solution $(\mathbf{u}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$.

Similarly as above, the antisymmetric part of the FEM matrix $\mathbf{L}_h$ comes from the discretizaton of an antisymmetric operator Q_S which, however, is not compact now in $H_0^1(\Omega)^d \times L_0^2(\Omega)$. A suitable calculation yields $\|Q_S\| \le (1-2\nu)^{-1/2}$. Thus symmetric part preconditioning yields *mesh independent linear convergence* of GCG-LS(0) algorithm for the preconditioned FEM discretization of the Navier system, namely, Theorem 1.4.1 yields

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{1}{\sqrt{2(1-\nu)}} \qquad (k=1,\dots,n). \tag{1.5.13}$$

This problem has been studied in [17]. We note that one can regularize the system similarly to the Stokes problem, and obtain superlinear convergence for symmetric part preconditioning, such that the sequence $\varepsilon_k \to 0$ depends on σ and Ω but is independent of ν and h (see also [17]).

1.5.7 Nonsymmetric preconditioning for convection-dominated problems

Convection-dominated problems arise when the magnitude $|\mathbf{b}|$ of the convection coefficient is large. If $\mathbf{b}$ is fixed, then this is equivalently expressed by a small coefficient ε of the Laplacian. Then the problem is called singularly perturbed, and often the case $\varepsilon \to 0$ is studied with the need of convergence independently of ε. There exist various approaches out of the scope of equivalent operators, mostly based on some stabilization [48], but here only linear convergence can be achieved. Also, Manteuffel and Otto [111] constructed an equivalent preconditioner for such a problem which is asymptotically robust w.r.t. ε, but not mesh independent; this result also concerns linear convergence.

In contrast to this, our main interest is superlinear convergence. However, in estimating superlinear convergence, one cannot achieve independence of ε. Our numerical results instead give a milder deterioration of the convergence rate with ε for a properly chosen preconditioning operator. We consider the convection-dominated problem

$$\begin{cases} Lu \equiv -\varepsilon\Delta u + \mathbf{b}\cdot\nabla u = g \\ u_{|\partial\Omega} = 0\,, \end{cases} \tag{1.5.14}$$

where L satisfies Assumptions 1.2.1. Should one choose the preconditioning operator $S := -\varepsilon\Delta$ like in subsection 1.5.1, the superlinear convergence rate would contain Q_S which comes from $\mathbf{b}$ and grows quickly as $|\mathbf{b}|/\varepsilon$ is increased.

This motivates the inclusion of a first order term in the preconditioning operator, i.e. a nonsymmetric preconditioning operator is chosen as

$$Nu := -\varepsilon\,\Delta u + \mathbf{w}\cdot\nabla u \qquad \text{for } \ u_{|\partial\Omega} = 0\,, \tag{1.5.15}$$

where $\mathbf{w}$ is a constant function. Then systems with $\mathbf{N}_h$ can be solved with a quasi-optimal number of operations using a fast direct solver for separable equations, see e.g. [143].

We can apply Theorem 1.2.8 to obtain mesh independent superlinear convergence:

Theorem 1.5.1 *For any FEM subspace $V_h \subset H^1_D(\Omega)$, using the stiffness matrix $\mathbf{N}_h$ as preconditioner to $\mathbf{L}_h$, the preconditioned CGN method converges superlinearly in a mesh independent way, i.e. the residuals satisfy* (1.2.68)–(1.2.69) *independently of n and V_h.*

On the other hand, dependence on ε is not eliminated. We will run tests to study the behaviour of the following natural choice. The definition of $\mathbf{w}$ is motivated by the consideration that N should be a good approximation of L, i.e. $\mathbf{w}$ should be a good constant approximation of $\mathbf{b}$. Then, as $\varepsilon \to 0$, the limit operators of L and N are $\mathbf{b} \cdot \nabla u$ and $\mathbf{w} \cdot \nabla u$, respectively. To obtain proportional quantities, we assume from now on that $\mathbf{b}$ satisfies the following, and $\mathbf{w}$ is chosen as follows:

$$0 < \beta_1 \le |\mathbf{b}| \le \beta_2, \qquad 0 < \beta_1 \le |\mathbf{w}| \le \beta_2, \tag{1.5.16}$$

respectively, for some constants β_1, β_2. In fact, if we have coordinatewise $\beta_1^{(i)} := \inf \mathbf{b}_i$ and $\beta_2^{(i)} := \sup \mathbf{b}_i$, then one can define $\mathbf{w}_i := \frac{1}{2}(\beta_1^{(i)} + \beta_2^{(i)})$.

Numerical experiments. For our tests, we consider problem (1.5.14) on the unit square $\Omega := [0,1]^2$ with a constant $\varepsilon > 0$ to be varied and with a piecewise constant $\mathbf{b}$:

$$\mathbf{b}(x,y) := \begin{cases} (1,1) \text{ if } 0 \le x < 0.5 \\ (2,2) \text{ if } 0.5 \le x \le 1. \end{cases}$$

The preconditioning operator (1.5.15) is N for the same Dirichlet boundary conditions with convection coefficient $\mathbf{w} := (1.5, 1.5)$. To solve (1.5.14), linear FEM was used with mesh width $1/h = 32$. The experiments were run using Matlab.

The stopping criterion was $\|r_k\|_{\mathbf{S}_h} \le 10^{-8}$, where $\mathbf{S}_h$ is the symmetric part of $\mathbf{N}_h$. The corresponding number of iterations is shown in Table 1.4. The number of iterations is increasing as ε decreases, but is still reasonable for $\varepsilon = 0.005$. The convergence ratio $Q_k := (\|r_k\|_{\mathbf{S}_h} / \|r_0\|_{\mathbf{S}_h})^{1/k}$ $(k = 1, 2, ...)$ was also measured and found to behave superlinearly as predicted, a typical behaviour (for $\varepsilon = 0.05$) is given in Table 1.5.

Table 1.4: Number of iterations, $\|r_k\|_{\mathbf{S}_h} \le 10^{-8}$.

	$1/h = 32$					
ε	1	0.5	0.1	0.05	0.01	0.005
Itr.	4	5	11	14	21	26

Table 1.5: The convergence factors Q_k, $1/h = 32$, $\varepsilon = 0.05$.

Itr.	1	2	3	4	5	6	7
Q_k	0.5680	0.4582	0.4194	0.3886	0.3800	0.3684	0.3633
Itr.	8	9	10	11	12	13	14
Q_k	0.3550	0.3517	0.3468	0.3422	0.3395	0.3377	0.3341

Chapter 2

Nonlinear problems

2.1 The general framework

In this chapter we study the numerical solution of a nonlinear operator equation

$$F(u) = b \tag{2.1.1}$$

(in a Hilbert space) that will then model a nonlinear elliptic PDE including boundary conditions. A Galerkin (resp. FEM) discretization yields a finite dimensional problem

$$F_h(u_h) = b_h. \tag{2.1.2}$$

The equivalent operator framework in Chapter 1 relies on the idea that it is sometimes more efficient to first approximate the given differential operator by some simpler differential operator, and then to use the stiffness matrix of this operator as preconditioner, than to discretize first and then construct a preconditioner algebraically.

Now we extend this idea to nonlinear problems, and develop the concept of *preconditioning operators*. It provides a general framework to discuss iterative methods, the scope of which reaches from simple iterations to Newton methods. This is strongly related to the concept of Sobolev gradients [123, 124], and in fact connects the latter with Newton-type methods. The idea of using suitable operators to derive preconditioning has also appeared in earlier works, involving the modified Newton-Kantorovich method, frozen coefficients or Gram matrices etc., see e.g. [30, 74, 76, 79], then an organized treatment was given in our book [55].

Considering one-step iterations and allowing the preconditioners to vary stepwise, this idea can be summarized as follows. Let us consider a nonlinear boundary value problem (2.1.1) and its discretization (2.1.2), respectively. The standard way of numerical solution consists of 'discretization plus iteration', whereas the preconditioning operator approach consists of 'iteration plus discretization'. Here we first define suitable linear elliptic differential operators $S^{(n)}$ and a sequence $\{u^{(n)}\}_{n\in\mathbf{N}}$ with these operators as preconditioners in the corresponding Sobolev space, providing a theoretical sequence

$$u^{(n+1)} = u^{(n)} - (S^{(n)})^{-1}(F(u^{(n)}) - b).$$

Then we propose the preconditioning matrices $S_h^{(n)}$ for the iteration in the considered FE or FD subspace V_h, which means that the preconditioning matrices are obtained using the same discretization for the operators $S^{(n)}$ as was used to obtain the system (2.1.2) from problem (2.1.1). Thus one obtains the iterative sequence

$$u_h^{(n+1)} = u_h^{(n)} - (S_h^{(n)})^{-1}(F_h(u_h^{(n)}) - b_h)$$

in the considered subspace V_h. The advantages of this idea come from exploiting the properties of the original PDE. In practice, the auxiliary linear elliptic problems can be solved by highly developed efficient and often optimal solvers, see e.g. [69, 70].

The common framework reaching from simple iterations to Newton methods is given through the idea of *variable preconditioning*, whose essence can be summarized as follows. In general, under certain conditions given in Theorem 2.3.2, if B_n are stepwise variable preconditioning operators that yield the variable spectral equivalence

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

then one can define the variably preconditioned sequence:

$$u_{n+1} = u_n - \frac{2}{M_n + m_n} B_n^{-1}(F(u_n) - b) \qquad (n \in \mathbf{N}) \tag{2.1.3}$$

(locally) or its damped version (globally), and it converges with the rate

$$q = \limsup \frac{M_n - m_n}{M_n + m_n}.$$

The iteration (2.1.3) is a quasi-Newton method based on variable spectral bounds. Using the first special choice $B_n = I$ (with $m_n = m$, $M_n = M$), this includes the gradient method and its well-known linear convergence rate. Second, the other special choice $B_n := F'(u_n)$ (with $m_n = M_n = 1$) reproduces Newton's method and shows that $q = 0$, i.e. superlinear convergence is achieved. In general, a fast (superlinear) convergence can be achieved by potentially simpler auxiliary operators than the derivatives $F'(u_n)$, such that the choice of B_n represents a compromise between I and $F'(u_n)$ (i.e. between lowest cost and greatest efficiency, just as is the case for preconditioners for linear problems).

It will also be seen that (2.1.3) is a variable gradient method corresponding to the potential of F, and can thus represent a variable Sobolev gradient iteration. We will prove that Newton's method is optimal w.r.t. local minimization among these variable descent methods. On the other hand, from computational aspect it can be more efficient to construct B_n to be more easily solvable than $F'(u_n)$, and we will show examples when this is done.

Our study is hence done in three stages, getting from simple Sobolev gradient iterations via variably preconditioned iterations to Newton's method.

2.2 Sobolev gradients for variational problems

The Sobolev gradient theory of J.W. Neuberger was shown to give a prospect for a unified theory of PDEs with extensively wide numerical applications, see e.g. [123, 124, 125, 135].

Sobolev gradients define descent methods in which the gradient is defined w.r.t. the Sobolev inner product [123]. Gradient type methods are in general less widely used in comparison with Newton-like methods, owing to the faster convergence of the latter. However, in some cases the gradient method can be altogether less costly and be therefore competitive, as observed e.g. in the numerous applications of Sobolev gradients. The GM may be competitive when updating Jacobians is costly but the required accuracy is not very high, which fact can even be simply quantified [87].

Here some new Sobolev gradient results are presented briefly for variational problems, based on a Hilbert space extension of the abstract gradient iteration which enables us to involve different choices of preconditioning operators. More details are given in [55].

2.2.1 Gradient iterations in Hilbert space

The presented iterative methods model the situation to be discussed at the beginning of subsection 2.2.2 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 $cond(S^{-1}T) \le \frac{M}{m}$. It extends a classical result of Dyakonov [44], involves a weak form of an unbounded nonlinear operator in a similar manner as we did in the linear case, see (1.2.1), and will connect it to the Sobolev gradient context, see (2.2.10). 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 2.2.1 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.)

In the following theorem we first define a weak form of an unbounded nonlinear operator T in a similar manner as we did in the linear case, see (1.2.1). This weak operator might be denoted similarly by T_S, but to make it fit simpler in the later discussion on Newton type methods, we just use another letter for the weak operator (usually F). The suitable properties of this weak operator provide the convergence of the iteration.

Theorem 2.2.1 *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 $M \ge m > 0$ satisfying*

$$m\langle S(v-u), v-u\rangle \le \langle T(v)-T(u), v-u\rangle \le M\langle S(v-u), v-u\rangle \qquad (u,v \in D). \quad (2.2.1)$$

Then the identity

$$\langle F(u), v\rangle_S = \langle T(u), v\rangle \qquad (u,v \in D) \quad (2.2.2)$$

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.

$$\langle F(u^*), v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \tag{2.2.3}$$

(If $g \in R(T)$ then $T(u^*) = g$.)

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

$$\begin{aligned} u_{n+1} &= u_n - \tfrac{2}{M+m} z_n\,, \\ \textit{where} \;\; \langle z_n, v\rangle_S &= \langle F(u_n), v\rangle_S - \langle g, v\rangle \quad (v \in H_S), \end{aligned} \tag{2.2.4}$$

converges linearly to u^*, namely,

$$\|u_n - u^*\|_S \le \frac{1}{m}\|F(u_0) - b\|_S \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N})\,, \tag{2.2.5}$$

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 (2.2.4) can be expressed as $z_n = S^{-1}(T(u_n) - g)$, that is, the auxiliary problem becomes $Sz_n = T(u_n) - g$.

Proof. Let $u \in D$ be fixed. Then the inequality $\|v\| \le p^{-1/2}\|v\|_S$ for the energy norm implies

$$|\langle T(u), v\rangle| \le p^{-1/2}\|T(u)\|\|v\|_S \qquad (v \in D),$$

hence $v \mapsto \langle T(u), v\rangle$ is a bounded linear functional on $D \subset H_S$. It has a unique bounded linear extension $\Phi_u : H_S \to \mathbf{R}$, hence the Riesz theorem defines a unique vector $F(u) \in H_S$ that satisfies

$$\langle F(u), v\rangle_S = \Phi_u\, v \qquad (v \in H_S).$$

The latter gives (2.2.2) for $v \in D$, i.e. F is the required operator. Now it is easy to verify assertions (1)–(3).

(1) Let F be extended to H_S such that it has a bihemicontinuous symmetric Gateaux derivative. This extension can be denoted also by F without confusion. Then (2.2.1) implies

$$m\|v-u\|_S^2 \le \langle F(v) - F(u), v-u\rangle_S \le M\|v-u\|_S^2 \qquad (u, v \in H_S), \tag{2.2.6}$$

i.e. the spectral bounds of F are between m and M. Thus equation $F(u) = b$ has a unique solution $u^* \in H_S$, and the equality $F(u^*) = b$ coincides with (2.2.3). If $g \in R(T)$, then (2.2.3) means

$$\langle T(u^*), v\rangle = \langle g, v\rangle \qquad (v \in H_S),$$

hence $T(u^*) = g$.

(2) Since F has a bihemicontinuous symmetric Gateaux derivative with spectral bounds between m and M, it is well-known (see e.g. [59]) that the estimate (2.2.5) holds for the sequence u_n.

(3) We have

$$\langle T(u), v\rangle = \langle S^{-1}T(u), v\rangle_S \qquad (u, v \in D),$$

hence (2.2.2) implies

$$F_{|D} = S^{-1}T. \tag{2.2.7}$$

Therefore, if $u_n \in D$, then the auxiliary equation in (2.2.4) takes the form

$$\langle z_n, v\rangle_S = \langle T(u_n) - g, v\rangle \quad (v \in H_S),$$

and is solved by

$$z_n = S^{-1}(T(u_n) - g) \in D.$$

Hence $u_0 \in D$ implies by induction that the sequence $(u_n) \subset D$ and that z_n is as above. ■

Remark 2.2.1 In the case $R(S) \supset R(T)$ we have (2.2.7), i.e. F can be considered as a preconditioned version of T.

Now we can formulate the discrete counterpart of the above theorem. Let the conditions of Theorem 2.2.1 hold, let $g \in H$ and let $V_h \subset H_S$ be a given finite-dimensional subspace. Then there exists a unique solution $u_h \in V_h$ to the projected problem

$$\langle F(u_h), v\rangle_S = \langle g, v\rangle \qquad (v \in V_h), \tag{2.2.8}$$

and the same convergence result holds:

Theorem 2.2.2 *For any $u_0 \in V_h$ the sequence $(u_n) \subset V_h$, defined by replacing all $v \in H_S$ in (2.2.4) by all $v \in V_h$, converges to u_h according to the same estimate (2.2.5), i.e. with a rate independent of V_h.*

Proof. Both the solvability and the convergence follow similarly to Theorem 2.2.1 if the space H is replaced by V_h. ■

More generally, one may allow natural weaker conditions e.g. as follows, see [55]:

Theorem 2.2.3 *If assumption (2.2.1) is replaced by*

$$\begin{aligned} m\,\|v-u\|_S^2 \le \langle T(v) - T(u), v-u\rangle \le M(r)\,\|v-u\|_S^2 \\ (u, v \in D,\ \|u\|_S, \|v\|_S \le r) \end{aligned} \tag{2.2.9}$$

for some increasing function $M : \mathbf{R}^+ \to \mathbf{R}^+$, then Theorem 2.2.2 holds in a modified form such that the constant M is replaced by M_0 depending on u_0:

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

Proof. It follows since (u_n) runs in the ball with radius $\|u_0\| + \frac{1}{m}\|F(u_0) - b\|$. ■

2.2.2 Sobolev gradients for elliptic problems

(a) Sobolev gradients and preconditioning

Theorem 2.2.1 relates to Sobolev gradients developed by J.W. Neuberger. Let *cond*$(T) = \infty$. The operator $F : H_S \to H_S$ in (2.2.2) 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

$$\phi'_S(u) = F(u) \qquad (u \in H_S) \quad \text{and} \qquad \phi'(u) = T(u) \qquad (u \in D). \tag{2.2.10}$$

The steepest descent iteration corresponding to the gradient ϕ'_S is the preconditioned sequence in (2.2.4), 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 (2.2.7) 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.

(b) Dirichlet problems for second order equations

First we illustrate the method on a very simple problem

$$\begin{cases} T(u) \equiv \; -\mathrm{div}\, f(x, \nabla u) = g(x) \\ u_{|\partial\Omega} = 0 \end{cases} \tag{2.2.11}$$

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

Assumptions 2.2.4.

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

Let $V_h \subset H^1_0(\Omega)$ be a given FEM subspace. We look for the FEM solution u_h of problem (2.2.11) in V_h. (Under the above assumptions it is well-known that (2.2.11) has a unique weak solution u^*, the FEM problem has a unique solution u_h and $\|u_h - u^*\|_{H^1_0} \to 0$ under standard assumptions on the subspaces V_h.) For a fixed V_h, we will construct a Sobolev gradient iteration to find u_h using a weighted inner product.

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

$$m\, G(x)\xi \cdot \xi \le \frac{\partial f(x,\eta)}{\partial \eta} \xi \cdot \xi \ \le M\, G(x)\xi \cdot \xi \qquad ((x,\eta) \in \Omega \times \mathbf{R}^d, \xi \in \mathbf{R}^d). \tag{2.2.12}$$

We introduce the linear preconditioning operator

$$Su \equiv -\mathrm{div}\,(G(x)\nabla u) \qquad \text{for } \ u_{|\partial\Omega} = 0. \tag{2.2.13}$$

The corresponding energy space is $H_0^1(\Omega)$ with the G-inner product (which is equivalent to the usual one):

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

Theorem 2.2.4 *Let Assumptions 2.2.4 be satisfied. Then for any $u_0 \in V_h$ the sequence $(u_n) \subset V_h$ defined by*

$$\begin{aligned} u_{n+1} &= u_n - \frac{2}{M+m} z_n\,, \\ \textit{where} \quad \int_\Omega G(x)\,\nabla z_n \cdot \nabla v &= \int_\Omega f(x, \nabla u_n) \cdot \nabla v - \int_\Omega gv \qquad (v \in V_h), \end{aligned} \tag{2.2.14}$$

converges linearly to u_h according to

$$\|u_n - u_h\|_G \le \frac{1}{m}\|F(u_0) - b\|_G \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N})\,, \tag{2.2.15}$$

where F and b are the weak forms of T and g (see below in (2.2.16)).

If $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, then $\|F(u_0) - b\|_G$ can be estimated by $\varrho^{-1/2}\|T(u_0) - g\|_{L^2(\Omega)}$, where $\varrho > 0$ is the smallest eigenvalue of S on $H^2(\Omega) \cap H_0^1(\Omega)$.

Proof. The generalized differential operator $F : H_0^1(\Omega) \to H_0^1(\Omega)$ and the weak form of the right-hand side g are given by the equalities

$$\langle F(u), v\rangle_G = \int_\Omega f(x, \nabla u) \cdot \nabla v \qquad \langle b, v\rangle_G = \int_\Omega gv \qquad (v \in H_0^1(\Omega)), \tag{2.2.16}$$

respectively. Let T be the operator in (2.2.11) with domain $D(T) = D := H^2(\Omega) \cap H_0^1(\Omega)$ in the real Hilbert space $L^2(\Omega)$. We verify that T and S satisfy the assumptions of Theorem 2.2.3. Therefore we check the conditions of Theorem 2.2.1, with (2.2.1) replaced by the weak form (2.2.9), but with $M(r) \equiv M$.

Inequality (2.2.12) implies that the eigenvalues of the matrices $G(x)$ have a uniform positive lower bound similarly to the Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$, hence the operator S in (2.2.13) is a symmetric linear operator in $L^2(\Omega)$ with some positive lower bound $\varrho > 0$. The divergence theorem yields

$$\int_\Omega T(u)v = \int_\Omega f(x, \nabla u) \cdot \nabla v \qquad (u, v \in H^2(\Omega) \cap H_0^1(\Omega)) \tag{2.2.17}$$

and condition (2.2.12) implies

$$\begin{aligned} m\,G(x)(\nabla v - \nabla u) \cdot (\nabla v - \nabla u) &\le (f(x, \nabla v) - f(x, \nabla u)) \cdot (\nabla v - \nabla u) \\ &\le M\,G(x)(\nabla v - \nabla u) \cdot (\nabla v - \nabla u), \end{aligned}$$

hence (2.2.17) gives

$$m\|v-u\|_G^2 \le \int_\Omega (T(v)-T(u))(v-u) \le M\|v-u\|_G^2 \quad (u,v\in H^2(\Omega)\cap H_0^1(\Omega)). \quad (2.2.18)$$

Further, by (2.2.17) the operator F defined in (2.2.2) now takes the form as in (2.2.16). Using that $f\in C^1$, it is easy to see that F is Gateaux differentiable,

$$\langle F'(u)h,v\rangle_G = \int_\Omega \frac{\partial f}{\partial \eta}(x,\nabla u)\,\nabla h\cdot\nabla v \quad (u,h,v\in H_0^1(\Omega)) \qquad (2.2.19)$$

and thus F' is bihemicontinuous and symmetric, hence the conditions of Theorem 2.2.1 are satisfied. The last statement follows from (2.2.2) and the Poincaré-Friedrichs inequality. ■

Remark 2.2.2 The generalized differential operator F in (2.2.16) maps from $H_0^1(\Omega)$ into $H_0^1(\Omega)$, i.e. for any $u\in H_0^1(\Omega)$ there exists the function $F(u)\in H_0^1(\Omega)$ that defines the equality (2.2.16). This differs from the more usual treatment when the same integral formula defines T as a weak form of operator from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$. However, our setting requires operators from H to H. The fact that now the element $F(u)$ is a function in $H_0^1(\Omega)$ can be seen more visually in the regular case, when the decomposition $F(u)=S^{-1}T(u)$ holds for $u\in H^2\cap H_0^1$.

The sequence (2.2.14) requires the stepwise FEM solution of a linear elliptic problem of the type

$$\begin{cases} Sz \equiv -\mathrm{div}\,(G(x)\nabla z) = r \\ z_{|\partial\Omega}=0, \end{cases} \qquad (2.2.20)$$

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 (d).

The method can be extended to similar but more general problems, such as mixed boundary conditions or fourth order equations [55]. We only deal here with certain systems in the next subsection when the upper spectral bound is not uniform.

(c) Second order symmetric systems

Now we consider more general problems: symmetric nonlinear elliptic systems of the form

$$\left.\begin{aligned} -\mathrm{div}\ f_i(x,\nabla u_i) + q_i(x,u_1,\dots,u_l) = g_i \\ u_{i\,|\Gamma_D}=0, \qquad f_i(x,\nabla u_i)\cdot\nu + \alpha_i u_{i\,|\Gamma_N} = 0 \end{aligned}\right\} \quad (i=1,\dots,l) \qquad (2.2.21)$$

on a bounded domain $\Omega\subset\mathbf{R}^d$ under the following assumptions:

Assumptions 2.2.5.

(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 $f_i : \Omega \times \mathbf{R}^d \to \mathbf{R}^d$ $(i = 1, \dots, l)$ 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 $\eta \in \mathbf{R}^d$ resp. $\xi \in \mathbf{R}^l$. 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 $\inf_{i,\Omega} \alpha_i > 0$.

(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 $(x, \xi) \in \Omega \times \mathbf{R}^l$

$$\left\| q'_\xi(x, \xi) \right\| \le c_1 + c_2 |\xi|^{p-2}.$$

The coercivity and growth assumptions imply that problem (1.2.72) has a unique weak solution in the product Sobolev space $H_0^1(\Omega)^l := H_0^1(\Omega) \times \cdots \times H_0^1(\Omega)$, see e.g. [55]. Let $V_h \subset H_0^1(\Omega)$ be a given FEM subspace. We look for the FEM solution $u_h = (u_{h,1}, .., u_{h,l})$ of problem (2.2.21) 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 (2.2.12) with M replaced by m'. We introduce a linear preconditioning operator $S = (S_1, \dots, S_l)$ as an independent l-tuple of operators

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

The corresponding energy space is $H_D^1(\Omega)^l$ with the G-inner product (which is equivalent to the usual one):

$$\langle u, v \rangle_G := \int_\Omega \sum_{i=1}^l G_i(x)\, \nabla u_i \cdot \nabla v_i.$$

We introduce the real function

$$M(r) := m' + c_1 \varrho^{-1} + d_1 K_{2,\Gamma_N}^2 + c_2 K_{p,\Omega}^{p_1} r^{p-2} \quad (r > 0), \tag{2.2.22}$$

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 2.2.5 *Let Assumptions 2.2.5 be satisfied. Let $u_0 \in V_h^l$ and*

$$M_0 := M\Big(\|u_0\|_{H_D^1(\Omega)} + \frac{1}{m} \|F(u_0) - b\|_{H_D^1(\Omega)} \Big), \tag{2.2.23}$$

where $M(r)$ is from (2.2.22) and F and b are the weak forms of $T = (T_1, \dots, T_l)$ and $g = (g_1, \dots, g_l)$, respectively. 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

$$u_{n+1} = u_n - \frac{2}{M_0 + m} z_n \ , \tag{2.2.24}$$

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

$$\int_\Omega G_i(x)\,\nabla z_{n,i}\cdot\nabla v_i = \int_\Omega \Big(f_i(x,\nabla u_{n,i})\cdot\nabla v_i + q_i(u_{n,1},..,u_{n,l})v_i\Big) + \int_{\Gamma_N}\alpha_i u_{n,i}v_i - \int_\Omega g_i v_i \tag{2.2.25}$$

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

$$\|u_n - u_h\|_G \le \frac{1}{m}\|F(u_0)-b\|_G \left(\frac{M_0-m}{M_0+m}\right)^n \qquad (n\in\mathbf{N}). \tag{2.2.26}$$

Proof. It follows from [55], Theorems 7.3-7.4. ■

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

$$\left\{\begin{array}{ll} S_i z_i \equiv -\mathrm{div}\,(G_i(x)\nabla z_i) = r_i & \\ z_{i\,|\Gamma_D} = 0, \qquad \frac{\partial z_i}{\partial\nu_{G_i}}_{\,|\Gamma_N} = \varrho_i & \end{array}\right. \qquad (i=1,\dots,l) \tag{2.2.27}$$

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.

(d) Some examples of preconditioning operators

Discrete Laplacian preconditioner. The most straightforward preconditioning operator for problem (2.2.11) 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 (2.2.12) independently of V_h. The solution of the linear auxiliary systems containing the discrete Laplacian preconditioner can rely on fast Poisson solvers [113, 116].

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

$$\delta_i^{\pm}(x,\eta) := \frac{\partial f_i(x,\eta)}{\partial\eta_i} \pm \sum_{\substack{j=1\\ j\neq i}}^{d}\left|\frac{\partial f_i(x,\eta)}{\partial\eta_j}\right|, \quad \text{we have} \quad \delta_i^-(x,\eta) \ge \mu_1 > 0 \tag{2.2.28}$$

(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 $1\le s\le d$, 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

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

independently of V_h. The solution of the linear auxiliary systems relies on fast separable solvers [113, 116].

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

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

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', see [55].

Some other cases. Let us mention very briefly some other natural choices of preconditioning operators.

(i) If we have Neumann boundary conditions [54], then we can get round the non-injectivity of the nonlinear operator by suitable factorization: the above operators S are replaced by

$$S_{|D}\,, \qquad \text{where} \quad D := \{u \in H^2(\Omega) : \tfrac{\partial u}{\partial \nu}\big|_{\Omega} = 0, \quad \int_{\Omega} u = 0\}\,.$$

(ii) In the case of 4th order problems the analogue of the discrete Laplacian preconditioner is the discrete biharmonic operator [81]. (Then one can use fast biharmonic solvers [23] or treat its higher order by suitable techniques like mixed formulation.)

(iii) For systems of PDEs an efficient choice of preconditioning operator is the r-tuple of independent Laplacians [55].

(iv) The Laplacian or biharmonic operator in (i)-(iii) can be replaced by more general operators similarly to those mentioned above (separable, initial Newton).

2.3 Variable preconditioning

2.3.1 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 [90]. Namely, the choice $B_n = I$ below in (2.3.1) 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. The more general version will be then given in Theorem 2.3.2.

Theorem 2.3.1 *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

$$u_{n+1} = u_n - \frac{2}{M+m} B_n^{-1}(F(u_n) - b) \qquad (n \in \mathbf{N}), \tag{2.3.1}$$

with properly chosen self-adjoint linear operators B_n satisfying

$$m\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\ h \in H), \tag{2.3.2}$$

converges linearly to u^*. Namely,

$$\|u_n - u^*\| \leq C \cdot \left(\frac{M-m}{M+m}\right)^n \qquad (n \in \mathbf{N}) \tag{2.3.3}$$

with some constant $C > 0$.

The proof of Theorem 2.3.1 is preceded by some required properties.

Lemma 2.3.1 [55]. *Let A and B be strongly positive bounded self-adjoint linear operators in H such that $mB \leq A \leq MB$ for some constants $M, m > 0$. Then the following properties hold:*

$$m^{1/2}\|h\|_{A^{-1}} \leq \|h\|_{B^{-1}} \leq M^{1/2}\|h\|_{A^{-1}} \qquad (h \in H), \tag{2.3.4}$$

$$\|I - \frac{2}{M+m} AB^{-1}\|_{A^{-1}} \leq \frac{M-m}{M+m}\,. \tag{2.3.5}$$

Lemma 2.3.2 *Let the conditions (i)-(ii) of Theorem 2.3.1 hold. Then for any $u, v, h \in H$,*

$$\langle F'(u)h, h\rangle \leq \langle F'(v)h, h\rangle \left(1 + L\lambda^{-2}\|F(u) - F(v)\|\right).$$

Proof. Assumption (i) implies $\|F(u) - F(v)\| \geq \lambda\|u - v\|$. Hence

$$\langle F'(u)h, h\rangle \leq \langle F'(v)h, h\rangle + L\|u-v\|\|h\|^2 \leq \langle F'(v)h, h\rangle + L\lambda^{-2}\|F(u) - F(v)\|\langle F'(v)h, h\rangle. \blacksquare$$

Applying Lemma 2.3.2 to u and u^*, we obtain

Corollary 2.3.1 *If $F(u^*) = b$, then for any fixed $u \in H$ there holds*

$$\frac{1}{1+\mu(u)} \leq \frac{\langle F'(u^*)h, h\rangle}{\langle F'(u)h, h\rangle} \leq 1 + \mu(u) \qquad (h \in H),$$

where $\mu(u) = L\lambda^{-2}\|F(u) - b\|$.

We introduce the norms

$$\|h\|_u = \langle F'(u)^{-1}h, h\rangle^{1/2} \quad (u, h \in H). \tag{2.3.6}$$

Then (2.3.4) and Corollary 2.3.1 imply directly

Corollary 2.3.2 *If $F(u^*) = b$, then for any fixed $u \in H$ there holds*

$$\frac{1}{1+\mu(u)} \leq \frac{\|h\|^2_{u^*}}{\|h\|^2_u} \leq 1 + \mu(u) \qquad (h \in H),$$

where $\mu(u)$ is from Corollary 2.3.1.

Proof of Theorem 2.3.1. We assume without loss of generality that $b = 0$, i.e. we study the equation $F(u) = 0$.

Assumption (i) and (2.3.4) imply that $\Lambda^{-1}\|h\|^2 \leq \langle F'(u)^{-1}h, h\rangle \leq \lambda^{-1}\|h\|^2$ for any $u, h \in H$. Hence the norms (2.3.6) satisfy

$$\lambda^{1/2}\|h\|_u \leq \|h\| \leq \Lambda^{1/2}\|h\|_u \qquad (u, h \in H), \tag{2.3.7}$$

and there also holds

$$\|F'(u)^{-1/2}\| \leq \lambda^{-1/2} \qquad (u \in H). \tag{2.3.8}$$

Since the assumptions imply that $\lambda M^{-1}\|h\|^2 \leq \langle B_n h, h\rangle$ for any $h \in H$, we obtain similarly to (2.3.8) that

$$\|B_n^{-1/2}\| \leq \lambda^{-1/2}M^{1/2}. \tag{2.3.9}$$

The following norms (special cases of (2.3.6)) will be used throughout the proof:

$$\| \,.\, \|_n = \| \,.\, \|_{u_n} \quad (n \in \mathbf{N}), \qquad \| \,.\, \|_* = \| \,.\, \|_{u^*} \tag{2.3.10}$$

The Lipschitz continuity of F' implies that

$$F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1} - u_n) + R(u_n), \tag{2.3.11}$$

where

$$\|R(u_n)\| \leq \frac{L}{2}\|u_{n+1} - u_n\|^2. \tag{2.3.12}$$

Here

$$F(u_n) + F'(u_n)(u_{n+1} - u_n) = F(u_n) - \frac{2}{M+m}F'(u_n)B_n^{-1}F(u_n),$$

hence (2.3.2) and (2.3.5) imply that

$$\|F(u_n)+F'(u_n)(u_{n+1}-u_n)\|_n \leq \left\| I-\frac{2}{M+m}F'(u_n)B_n^{-1}\right\|_n \|F(u_n)\|_n \leq \frac{M-m}{M+m}\|F(u_n)\|_n\,. \tag{2.3.13}$$

Further, (2.3.7) and (2.3.12) yield

$$\|R(u_n)\|_n \leq \frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2.$$

Here, using (2.3.9), (2.3.2) and (2.3.4), we have

$$\begin{aligned}\|B_n^{-1}F(u_n)\|^2 &\leq \|B_n^{-1/2}\|^2\|B_n^{-1/2}F(u_n)\|^2 \leq M\lambda^{-1}\langle B_n^{-1}F(u_n),F(u_n)\rangle\\ &\leq M^2\lambda^{-1}\langle F'(u_n)^{-1}F(u_n),F(u_n)\rangle = M^2\lambda^{-1}\|F(u_n)\|_n^2.\end{aligned}$$

Hence

$$\|R(u_n)\|_n \leq \frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2. \tag{2.3.14}$$

Altogether, (2.3.11), (2.3.13) and (2.3.14) yield

$$\|F(u_{n+1})\|_n \leq \left(\frac{M-m}{M+m}+\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\right)\|F(u_n)\|_n\,.$$

Finally, using Corollary 2.3.2 and (2.3.10), we obtain

$$\|F(u_{n+1})\|_* \leq (1+\mu(u_n))\left(\frac{M-m}{M+m}+\frac{2LM^2}{\lambda^{3/2}(M+m)^2}(1+\mu(u_n))^{1/2}\|F(u_n)\|_*\right)\|F(u_n)\|_*\,,$$

where $\mu(u_n) = L\Lambda^{1/2}\lambda^{-2}\|F(u_n)\|_*$ using (2.3.7). That is,

$$\|F(u_{n+1})\|_* \leq \varphi(\|F(u_n)\|_*)\,\|F(u_n)\|_*\,, \tag{2.3.15}$$

where

$$\varphi(t) = (1+\beta\Lambda^{1/2}t)\left(Q+M^2\beta\alpha^{-2}\lambda^{1/2}(t/2)\left(1+\beta\Lambda^{1/2}t\right)^{1/2}\right) \tag{2.3.16}$$

and the notations

$$\alpha=\frac{M+m}{2},\quad \beta=\frac{L}{\lambda^2},\quad Q=\frac{M-m}{M+m}$$

are used. Then $\varphi:\mathbf{R}^+\to\mathbf{R}^+$ is a strictly increasing continuous function and $\varphi(0)=Q$.

Estimate (2.3.15) puts us in the position to prove the required convergence estimate (2.3.3), provided that the assumption

$$r := \varphi(\|F(u_0)\|_*)\ <\ 1 \tag{2.3.17}$$

is satisfied for the initial guess.

First, we obtain by induction that

$$\|F(u_{n+1})\|_* \leq r\|F(u_n)\|_* \qquad (n\in\mathbf{N}). \tag{2.3.18}$$

Namely, $\|F(u_1)\|_* = r\|F(u_0)\|_*$. Further, the assumption $\|F(u_{k+1})\|_* \leq r\|F(u_k)\|_*$ $(k = 0, ..., n-1)$ yields $\|F(u_n)\|_* < \|F(u_0)\|_*$, hence

$$\|F(u_{n+1})\|_* \leq \varphi(\|F(u_n)\|_*)\,\|F(u_n)\|_* \leq \varphi(\|F(u_0)\|_*)\,\|F(u_n)\|_* = r\|F(u_n)\|_* \,.$$

Inequality (2.3.18) implies $\|F(u_n)\|_* \leq r^n\|F(u_0)\|_* \to 0, \quad \varphi(\|F(u_n)\|_*) \to Q$ and hence

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \leq \lim \varphi(\|F(u_n)\|_*) = Q.$$

From now on we use the notation $e_n := \|F(u_n)\|_*$. Then (2.3.15) implies

$$e_n \leq \left(\prod_{k=0}^{n-1} \varphi(e_k)\right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q}\right) Q^n e_0 \qquad (n \in \mathbf{N}). \tag{2.3.19}$$

Using (2.3.16) and the notations $c = \beta\Lambda^{1/2}$, $d = (M^2\beta\alpha^{-2}\lambda^{1/2})/2$, we have

$$\varphi(t) = (1+ct)\left(Q + dt\,(1+ct)^{1/2}\right).$$

Here

$$\frac{\varphi(e_k)}{Q} = (1+ce_k)\left(1 + \frac{d}{Q}e_k\,(1+ce_k)^{1/2}\right)$$

$$\leq (1+ce_k)\left(1 + \frac{d}{Q}e_k\left(1+\frac{c}{2}e_k\right)\right) = 1 + \left(c + \frac{d}{Q}\right)e_k + \frac{cd}{Q}e_k^2 + \frac{c^2d}{2Q}e_k^3$$

$$\leq 1 + \left(c + \frac{d}{Q}\right)e_0 r^k + \frac{cd}{Q}e_0^2 r^{2k} + \frac{c^2d}{2Q}e_0^3 r^{3k}.$$

Since for any sequence $(a_k) \subset \mathbf{R}^+$ there holds $\prod_{k=0}^{n-1}(1+a_k) \leq \prod_{k=0}^{n-1}\exp(a_k) \leq \exp(\sum_{k=0}^{\infty} a_k)$, hence we obtain

$$\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q} \leq \exp\left\{\left(c + \frac{d}{Q}\right)\frac{e_0}{1-r} + \frac{cd}{Q}\frac{e_0^2}{1-r^2} + \frac{c^2d}{2Q}\frac{e_0^3}{1-r^3}\right\} =: E\,.$$

Therefore (2.3.19) yields

$$e_n \leq e_0 E \cdot Q^n \qquad (n \in \mathbf{N}).$$

Finally, using condition (ii) and (2.3.7), this implies

$$\|u_n - u^*\| \leq \lambda^{-1}\|F(u_n)\| \leq \lambda^{-1}\Lambda^{1/2}e_0 E \cdot Q^n \qquad (n \in \mathbf{N}), \tag{2.3.20}$$

which coincides with the required convergence estimate with $C = \lambda^{-1}\Lambda^{1/2}e_0E$. ■

Now we turn to the more general version of Theorem 2.3.1. First we recall the following definitions of norms (see (2.3.10)), where (u_n) is an iterative sequence and u^* is the solution of $F(u) = b$:

$$\|h\|_n = \langle F'(u_n)^{-1}h, h\rangle^{1/2} \quad (n \in \mathbf{N}), \qquad \|h\|_* = \langle F'(u^*)^{-1}h, h\rangle^{1/2}. \tag{2.3.21}$$

The following theorem gives the main result on variable preconditioning. 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 2.3.2 *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 2.3.1.*

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

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} B_n^{-1}(F(u_n) - b) \qquad (n \in \mathbf{N}), \tag{2.3.22}$$

where the following conditions hold:

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

$$m_n\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M_n\langle B_n h, h\rangle \qquad (n \in \mathbf{N},\, h \in H), \tag{2.3.23}$$

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*

$$\tau_n = \min\{1, \frac{1 - Q_n}{2\rho_n}\}, \tag{2.3.24}$$

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 (2.3.21). (This value of τ_n ensures optimal contractivity in the n-th step in the $\| \,.\, \|_*$-norm.)

Then there holds

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

namely,

$$\limsup \frac{\|F(u_{n+1}) - b\|_*}{\|F(u_n) - b\|_*} \leq \limsup \frac{M_n - m_n}{M_n + m_n} < 1\,. \tag{2.3.25}$$

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

$$\|F(u_{n+1}) - b\|_* \leq d_1\|F(u_n) - b\|_*^{1+\gamma} \qquad (n \in \mathbf{N}) \tag{2.3.26}$$

with some constant $d_1 > 0$.

Owing to the equivalence of the norms $\| \,.\, \|$ and $\| \,.\, \|_*$, the orders of convergence corresponding to the estimate (2.3.26) can be formulated with the original norm:

Corollary 2.3.3 *(Rate of convergence in the original norm.) Let*

$$M_n/m_n \leq 1 + c_1\|F(u_n - b)\|^\gamma$$

with some constants $c_1 > 0$, $0 < \gamma \leq 1$. Then there holds

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

and consequently

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

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

Proof of Theorem 2.3.2. We assume without loss of generality (similarly to Theorem 2.3.1) that $b = 0$, i.e. we study the equation $F(u) = 0$.

Using (2.3.11) and (2.3.22), we obtain

$$F(u_{n+1}) = (1-\tau_n)F(u_n) + \tau_n\left(F(u_n) - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}F(u_n)\right) + R(u_n).$$

Hence

$$\|F(u_{n+1})\|_* \le (1-\tau_n)\|F(u_n)\|_* + \tau_n\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* + \|R(u_n)\|_* \,.$$

Here, using Corollary 2.3.2 and (2.3.5),

$$\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* \le\ (1+\mu(u_n))^{1/2}\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_n$$

$$\le\ (1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_* \,,$$

where $\mu(u_n) = L\lambda^{-2}\|F(u_n)\|$. Further, from (2.3.7) and (2.3.12) there follows

$$\|R(u_n)\|_* \le \frac{L}{2\lambda^{1/2}}\|u_{n+1}-u_n\|^2 = \tau_n^2\frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2,$$

hence, using the estimate preceding (2.3.14) and then Corollary 2.3.2, we obtain

$$\|R(u_n)\|_* \le \tau_n^2\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2 \le \tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\|F(u_n)\|_*$$

Summing up, we obtain

$$\|F(u_{n+1})\|_* \le$$
$$\left(1-\tau_n+\tau_n(1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n} + \tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\right)\ \|F(u_n)\|_* \,.$$

That is,

$$\|F(u_{n+1})\|_* \le \left(1-\tau_n(1-Q_n)+\tau_n^2\rho_n\right)\ \|F(u_n)\|_* \,, \tag{2.3.27}$$

where Q_n and ρ_n are as in condition (iv).

There exists $\tilde{Q} < 1$ such that

$$Q_n \le \tilde{Q} \qquad (n \in \mathbf{N}). \tag{2.3.28}$$

Namely, the assumption $M_n/m_n \le 1 + 2/(\varepsilon + K\mu(u_n))$ with $K > 1$ and $\varepsilon > 0$ implies

$$1+\varepsilon+K\mu(u_n) \le 1 + \frac{2}{(M_n/m_n)-1} = \frac{M_n+m_n}{M_n-m_n},$$

hence

$$1+\mu(u_n) \le \tilde{Q}\,\frac{M_n+m_n}{M_n-m_n}$$

with $\tilde{Q} := \max\{1/K, 1/(1+\varepsilon)\} < 1$.

Let us introduce the function $p : [0,1] \to \mathbf{R}$, $p(t) := 1 - (1-Q_n)t + \rho_n t^2$. Here $p'(t) = -(1-Q_n) + 2\rho_n t$ yields that τ_n defined in (2.3.24) satisfies

$$p(\tau_n) = \min_{t\in[0,1]} p(t) < 1,$$

since $p'(0) = -(1-Q_n) < 0$. Hence from (2.3.27)

$$\|F(u_{n+1})\|_* \le p(\tau_n)\|F(u_n)\|_* < \|F(u_n)\|_* \,. \tag{2.3.29}$$

Moreover, if $\tau_n = 1$ (i.e. when $1 \le (1-Q_n)/2\rho_n$), then

$$p(\tau_n) = Q_n + \rho_n \le Q_n + (1-Q_n)/2 = (1+Q_n)/2 \le (1+\tilde{Q})/2 < 1.$$

In the case $\tau_n = (1-Q_n)/2\rho_n$ we have

$$p(\tau_n) = 1 - (1-Q_n)^2/(4\rho_n) \le 1 - (1-\tilde{Q})^2/(4\sup_n \rho_n) =: Q' < 1.$$

The latter holds since by (2.3.29) $\|F(u_n)\|_*$ is bounded, hence

$$\rho_n = const. \cdot \|F(u_n)\|_n \left(1 + const. \cdot \|F(u_n)\|\right)^{1/2} \tag{2.3.30}$$

is bounded, the three norms being equivalent. Altogether, from (2.3.29) we obtain

$$\|F(u_n)\|_* \le const. \cdot r^n \to 0$$

where $r = \max\{(1+\tilde{Q})/2, Q'\}$. This also implies that $\rho_n \to 0$ and $\mu(u_n) = L\lambda^{-2}\|F(u_n)\| \to 0$. A brief calculation gives

$$p(\tau_n) = Q_n + \rho_n \left(1 - (1-\tau_n)^2\right) \tag{2.3.31}$$

(for both $\tau_n = 1$ and $\tau_n < 1$), hence (2.3.29) yields

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup Q_n = \limsup \frac{M_n - m_n}{M_n + m_n} \,.$$

The bound $M_n/m_n \le 1 + 2/\varepsilon$ in assumption (iv) implies that

$$\limsup \frac{M_n - m_n}{M_n + m_n} \le \frac{1}{1+\varepsilon} < 1 \,.$$

Finally, let $M_n/m_n \le 1 + c_1\|F(u_n)\|^\gamma$ with constants $c_1 > 0$, $0 < \gamma \le 1$. Then $M_n/m_n \le 1 + c_2\|F(u_n)\|_*^\gamma$ with $c_2 = c_1\Lambda^{1/2}$, hence

$$\frac{M_n - m_n}{M_n + m_n} < \frac{M_n - m_n}{m_n} \le c_2\|F(u_n)\|_*^\gamma \,,$$

and therefore

$$Q_n \le c_3\|F(u_n)\|_*^\gamma$$

with $c_3 = c_2(1 + \sup_n \mu(u_n))$. Also, $\rho_n \le c_4\|F(u_n)\|_*$ with some $c_4 > 0$ since $\|F(u_n)\|_*$ is bounded (cf. (2.3.30)). Using notation $e_n = \|F(u_n)\|_*$ and $d_1 := c_3 + c_4 e_0^{1-\gamma}$, we obtain from (2.3.29) and (2.3.31) that

$$e_{n+1} \le (Q_n + \rho_n)\, e_n \le (Q_n + c_4 e_n)\, e_n \le \left(c_3 e_n^\gamma + c_4 e_0 \frac{e_n}{e_0}\right) e_n \le d_1 e_n^{1+\gamma}. \qquad \blacksquare$$

Remark 2.3.1 Theorem 2.3.2 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 (2.3.26), i.e. the fastest feasible convergence is of order $1 + \alpha$.

Remark 2.3.2 Our results can be formulated in the context of Sobolev gradients, similarly to (2.2.10). Now, using a variable preconditioning operator, one obtains the *variable Sobolev gradient*

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

so in this sense we have extended the Sobolev gradient steepest descent to steepest descent w.r.t. variable inner product. It is worth mentioning that the idea of steepest descent w.r.t. variable inner product was later also extended to Lebesgue and Besov spaces [131] and Riemann manifolds in the study of differential-algebraic equations [127].

2.3.2 Variable preconditioning for elliptic problems

(a) Problems with nonlinear principal part

Let us consider problem (2.2.11) again:

$$\begin{cases} T(u) \equiv -\operatorname{div} f(x, \nabla u) = g(x) \\ u_{|\partial\Omega} = 0. \end{cases} \tag{2.3.32}$$

Assumptions 2.3.3. Assumptions 2.2.4 imposed for (2.2.11) are modified such that we only demand that $f : \Omega \times \mathbf{R}^d \to \mathbf{R}^d$ is measurable and bounded w.r. to the variable $x \in \Omega$ and C^1 w.r. to the variable $\eta \in \mathbf{R}^d$. In addition, the Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$ are Lipschitz continuous w.r.t η.

Letting $V_h \subset H_0^1(\Omega)$ be a given FEM subspace, we look for the FEM solution u_h of problem (2.3.32) in V_h.

A general iteration with variable preconditioning. 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 (2.2.13) is modified with stepwise redefined diffusion coefficient matrices.

Theorem 2.3.3 *Let Assumptions 2.3.3 hold, 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*

$$m_n\, G_n(x)\xi \cdot \xi \leq \frac{\partial f}{\partial \eta}(x, \nabla u_n(x))\, \xi \cdot \xi \leq M_n\, G_n(x)\xi \cdot \xi \qquad (x \in \Omega,\ \xi \in \mathbf{R}^N), \tag{2.3.33}$$

further, M_n/m_n and τ_n satisfy the conditions (iv)-(v) in Theorem 2.3.2. We define

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} z_n\,, \tag{2.3.34}$$

where $z_n \in V_h$ is the solution of

$$\int_\Omega G_n(x)\, \nabla z_n \cdot \nabla v = \int_\Omega \big(f(x, \nabla u_n) \cdot \nabla v \; - gv \big) \qquad (v \in V_h). \tag{2.3.35}$$

Then u_n converges to u_h according to the estimates of Theorem 2.3.2.

Proof. The weak form F of the operator has a similar form as (2.2.16) before, using the special case $G = I$. As seen in (2.2.19) and after, F fulfils property (i) of Theorem 2.3.1. Further, (2.2.19) also implies that F' inherits the Lipschitz continuity of $\frac{\partial f(x,\eta)}{\partial \eta}$ in η. For any $n \in \mathbf{N}$ let $B_n : V_h \to V_h$ denote the linear operator

$$\langle B_n v, w \rangle = \int_\Omega G_n(x)\, \nabla v \cdot \nabla w \qquad (v, w \in V_h),$$

then B_n is self-adjoint and (2.3.33) implies (2.3.23) in V_h. Therefore all the conditions of Theorem 2.3.2 are satisfied and thus the convergence results hold. ■

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, as demonstrated in [90]. This is motivated by the case of ill-posed (nearly singular) problems when the lower bound of the Jacobians of f is close to 0. In this case the discrete Laplacian, which represents a Sobolev gradient w.r.t. the usual $H_0^1(\Omega)$-inner product, would yield a convergence factor close to 1. We get round this via a suitable generalization of the discrete Laplacian to 'local Laplacians', i.e. a preconditioning operator with piecewise constant coefficients.

Formally we write

$$S_n u := -\mathrm{div}\, \big(w_n(x) \nabla u \big) \tag{2.3.36}$$

where w_n is a piecewise constant function, that is, the domain Ω is decomposed in subdomains $\Omega_i \quad (i = 1, \dots, s)$, and for all i,

$$w_{n\,|\Omega_i} \equiv c_i \; > 0. \tag{2.3.37}$$

Then $G_n(x) = w_n(x)\, I$, where I is the identity matrix. In fact we only have to use the corresponding inner product

$$\langle u, v \rangle_{S_n} = \int_\Omega w_n(x)\, \nabla u \cdot \nabla v \qquad (u, v \in H_0^1(\Omega)).$$

The estimate of the condition number follows from a proper choice of the constants c_i. Let

$$J_n(x) := \frac{\partial f}{\partial \eta}(x, \nabla u_n(x)) \tag{2.3.38}$$

denote the current Jacobian, and let us introduce the spectral bounds m_i and M_i of J_n relative to Ω_i, i.e. such that $\sigma(J_n(x)) \subset [m_i, M_i]$ for all $x \in \Omega_i$. Then one should choose c_i between m_i and M_i. The definition of m_i and M_i implies that

$$(\min_i \frac{m_i}{c_i})\; w_n(x) |\xi|^2 \leq J_n(x)\; \xi \cdot \xi \leq (\max_i \frac{M_i}{c_i})\; w_n(x) |\xi|^2 \qquad (h \in H_0^1(\Omega)).$$

Introducing

$$m_n := \min_i m_i/c_i \quad \text{and} \quad M_n := \max_i M_i/c_i \tag{2.3.39}$$

we obtain that estimate (2.3.33) holds for $G_n(x) := w_n(x)\, I$.

In particular, if c_i is some (arithmetic, geometric or harmonic) mean of m_i and M_i, then $M_n/m_n = \max\limits_i M_i/m_i$. Altogether, using these values of M_n and m_n, an improved mesh independent convergence estimate is valid. The numerical performance of such preconditioners will be illustrated in subsection 2.6.1.

General scalar coefficient preconditioning operators. One can more generally define any operator S with a scalar diffusion coefficient:

$$S_n u := -\text{div}\,\big(k_n(x)\nabla u\big), \tag{2.3.40}$$

where $k_n \in L^\infty(\Omega)$ and $k_n \ge k_0 > 0$. The solution of the auxiliary problems is more convenient with such an operator than with the full Jacobian matrix as a diffusion tensor; in particular, the discretized scalar coefficient operator has a better sparsity pattern and is an M-matrix under many reasonable discretizations.

Here $G_n(x) = k_n(x)\, I$ and the corresponding inner product is

$$\langle u, v\rangle_{S_n} = \int_\Omega k_n\, \nabla u \cdot \nabla v \qquad (u, v \in H_0^1(\Omega)).$$

One can easily derive the theoretical bounds. Let $\lambda_n(x)$ and $\Lambda_n(x)$ denote the extreme eigenvalues of the Jacobian (2.3.38) for fixed $x \in \overline{\Omega}$. Then

$$\Big(\min_{x\in\overline{\Omega}} \frac{\lambda_n(x)}{k_n(x)}\Big)\; k_n(x)|\xi|^2 \le \lambda_n(x)|\xi|^2 \le J_n(x)\,\xi\cdot\xi \le \Lambda_n(x)|\xi|^2 \le \Big(\max_{x\in\overline{\Omega}} \frac{\Lambda_n(x)}{k_n(x)}\Big)\; k_n(x)|\xi|^2, \tag{2.3.41}$$

i.e. estimate (2.3.33) holds for $G_n(x) := k_n(x)\, I$ and the bounds

$$m_n := \min_{x\in\overline{\Omega}} \frac{\lambda_n(x)}{k_n(x)} \quad \text{and} \quad M_n := \max_{x\in\overline{\Omega}} \frac{\Lambda_n(x)}{k_n(x)}.$$

We note that

$$\frac{M_n}{m_n} = \max_{x\in\overline{\Omega}} \frac{\Lambda_n(x)}{k_n(x)} \max_{x\in\overline{\Omega}} \frac{k_n(x)}{\lambda_n(x)} \ge \max_{x\in\overline{\Omega}} \frac{\Lambda_n(x)}{\lambda_n(x)}$$

and the latter is achieved when $k_n(x) = (\lambda_n(x)\Lambda_n(x))^{1/2}$. Hence, this k_n is the optimal choice concerning the possible condition numbers.

In practice, the functions λ_n and Λ_n are often not known in advance. A useful choice for $k_n(x)$ can then be simply the diagonal of $J_n(x)$.

(b) Variable preconditioning for semilinear problems

Let us consider a semilinear equation with mixed boundary conditions

$$\begin{cases} -\text{div}\,(k(x)\,\nabla u) + q(x,u) = g(x) \\[2mm] u_{|\Gamma_D} = 0, \qquad k(x)\,\frac{\partial u}{\partial \nu} + \alpha u_{|\Gamma_N} = 0 \end{cases} \tag{2.3.42}$$

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

Assumptions 2.3.42.

(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) $k \in L^\infty(\Omega)$ and the function $q : \Omega \times \mathbf{R} \to \mathbf{R}$ is measurable and bounded w.r. to the variable $x \in \Omega$ and C^1 in the second variable. Further, $\alpha \in L^\infty(\Gamma_N)$ and $g \in L^2(\Omega)$.

(iii) $k(x) \geq k_0 > 0$ ($x \in \Omega$) and $q'_\xi(x, \xi) := \frac{\partial q(x,\xi)}{\partial \xi} \geq 0$ for all $(x, \xi) \in \Omega \times \mathbf{R}$. Further, $\alpha(x) \geq 0$ ($x \in \Gamma_N$) and either $\Gamma_D \neq \emptyset$ or $\inf_\Omega \alpha > 0$.

(iv) 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 $(x, \xi_2) \in \Omega \times \mathbf{R}$,

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

Assumptions (iii)-(iv) imply that there exist constants $c_3, c_4 \geq 0$ such that for any $(x, \xi) \in \Omega \times \mathbf{R}$

$$0 \leq q'_\xi(x, \xi) \leq c_3 + c_4 |\xi|^{p-2}. \tag{2.3.43}$$

Then the generalized differential operator F is Gateaux differentiable and F' is locally Lipschitz continuous, see e.g. [55]. It satisfies

$$\langle F'(u)v, v \rangle = \int_\Omega \left(k(x)\, |\nabla v|^2 + q'_\xi(x, u) v^2 \right) \qquad (v \in H^1_D(\Omega)).$$

Let us construct an iteration as in Theorem 2.3.2 with the operators

$$\langle B_n v, z \rangle := \kappa \int_\Omega \nabla v \cdot \nabla z + c_n \int_\Omega vz$$

for some constants $\kappa > 0$ and $c_n > 0$. Then B_n is the weak form of the operator

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

For given u_n, one has

$$\langle F'(u_n)v, v \rangle \leq \int_\Omega k(x)\, |\nabla v|^2 + \max_\Omega q'_\xi(x, u_n) \int_\Omega v^2 \leq M_n \langle B_n v, v \rangle$$

and similarly

$$\langle F'(u_n)v, v \rangle \geq m_n \langle B_n v, v \rangle$$

where, using the Poincaré-Friedrichs constant C_Ω,

$$M_n := \max\{\|k\|_\infty / \kappa,\ \max q'_\xi(x, u_n)/c_n\}, \qquad m_n := k_0/(\kappa + C_\Omega c_n).$$

Corollary 2.3.4 *If M_n/m_n and τ_n satisfy the conditions (iv)-(v) in Theorem 2.3.2, then u_n converges to u_h according to the estimates of Theorem 2.3.2.*

The main point of this method is that S_n has constant coefficients, hence its updating is much faster than for $F'(u_n)$, and also fast solvers are available for the auxiliary problems. On the other hand, the inclusion of the variable coefficient c_n allows to follow the variation of the magnitude of the lower order term during the iteration. The numerical performance of such preconditioners will be mentioned in subsection 2.6.4 for problem (2.6.23).

2.4 Newton's method and operator preconditioning

In this section we study Newton's method in the context of preconditioning and gradients. First a theoretical result, then the realization of Newton's method is in focus.

2.4.1 Newton's method as optimal variable gradients

Now we study the relation of the gradient and Newton's method, as developed in [98]. 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. The main theoretical result will be the following: whereas the descents in the gradient method are steepest w.r. to different directions, *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 2.4.1.

(i) F has a bihemicontinuous symmetric Gateaux derivative (see Definition 2.2.1);

(ii) for every $R > 0$ there exist constants $P \ge p > 0$ such that

$$p\|h\|^2 \le \langle F'(u)h, h\rangle \le P\|h\|^2 \qquad (\|u\| \le R,\ h \in H); \tag{2.4.1}$$

(iii) for every $R > 0$ there exists a constant $L > 0$ such that

$$\|F'(u) - F'(v)\| \le L\|u - v\| \qquad (\|u\|, \|v\| \le R).$$

Here, by (i), F has a potential ϕ. The above conditions themselves do not ensure that equation $F(u) = 0$ has a solution, hence we impose condition

(iv) equation $F(u) = 0$ has a solution $u^* \in H$.

Then the solution u^* is unique and also minimizes ϕ. We note that the existence of u^* is already ensured if the lower bound $p = p(R)$ in condition (ii) satisfies $\lim_{R\to\infty} R\,p(R) = +\infty$, or if p does not depend on R at all (see e.g. [55, 59]).

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

$$u_{n+1} = u_n - B_n^{-1}F(u_n), \tag{2.4.2}$$

where we look for B_n in the class

$$\mathcal{B} \equiv \{B \in L(H) \text{ self-adjoint} : \ \exists\, p > 0 \quad \langle Bh, h\rangle \ge p\|h\|^2 \quad (h \in H)\}. \tag{2.4.3}$$

(The uniform positivity is needed to yield $R(B_n) = H$, by which the existence of $B_n^{-1}F(u_n)$ is ensured in the iteration.) Let $n \in \mathbf{N}$ and assume that the nth term of the sequence (2.4.2) is constructed. Then the next step yields the functional value

$$m(B_n) := \phi(u_n - B_n^{-1}F(u_n))\,. \tag{2.4.4}$$

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

$$\min_{B_n\in\mathcal{B}} m(B_n) \;=\; m(F'(u_n)) \qquad \text{up to second order} \tag{2.4.5}$$

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

$$\mathcal{B}(\nu_1) \equiv \{B \in L(H) \text{ self-adjoint}: \quad \langle Bh, h\rangle \ge \nu_1\|h\|^2 \quad (h \in H)\}, \tag{2.4.6}$$

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

Theorem 2.4.1 *Let F satisfy Assumptions 2.4.1. Let $u_0 \in H$ and let the sequence (u_n) be given by (2.4.2) with operators $B_n \in \mathcal{B}$. Let $n \in \mathbf{N}$ be fixed and*

$$\hat{m}(B_n) := \beta + \frac{1}{2}\left\langle H_n(B_n^{-1}g_n - H_n^{-1}g_n),\; B_n^{-1}g_n - H_n^{-1}g_n\right\rangle, \tag{2.4.7}$$

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

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

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

$$|m(B_n) - \hat{m}(B_n)| \le C\|u_n - u^*\|^3 \tag{2.4.8}$$

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.

Proof. (1) This part of the theorem simply follows using that $H_n = F'(u_n)$ is positive definite by assumption (ii), whence we obtain

$$\hat{m}(B_n) \ge \beta = \hat{m}(H_n) = \hat{m}(F'(u_n)).$$

(2) We verify the required estimate in four steps.

(i) First we prove that

$$\|u_n - u^*\| \le R_0 \tag{2.4.9}$$

where R_0 depends on u_0, that is, the initial guess determines an a priori bound for a ball $B(u^*, R_0)$ around u^* containing the sequence (2.4.2). For this it suffices to prove that the level set corresponding to $\phi(u_0)$ is contained in such a ball, i.e.,

$$\{u \in H:\; \phi(u) \le \phi(u_0)\} \;\subset\; B(u^*, R_0), \tag{2.4.10}$$

since u_n is a descent sequence w.r.t. ϕ.

Let $u \in H$ be fixed and consider the real function $f(t) := \phi\Big(u^* + t\,\frac{u-u^*}{\|u-u^*\|}\Big)$ $(t \in \mathbf{R})$, which is C^2, convex and has its minimum at 0. By the assumed uniform monotonicity, there exists $p_1 > 0$ such that $\langle \phi''(v)h, h\rangle \geq p_1\|h\|^2$ $(\|v-u^*\| \leq 1,\ h \in H)$, and hence $f''(t) \geq p_1$ $(|t| \leq 1)$. Then elementary calculus yields that $f'(1) \geq p_1$ and $f(1) - f(0) \geq p_1/2$, hence

$$\phi(u) - \phi(u^*) \;=\; f(\|u-u^*\|) - f(1) + f(1) - f(0)$$

$$\geq\; f'(1)(\|u-u^*\| - 1) + f(1) - f(0) \;\geq\; p_1\Big(\|u-u^*\| - \frac{1}{2}\Big).$$

This implies that if

$$\|u-u^*\| \geq \frac{1}{p_1}\Big(\phi(u_0) - \phi(u^*)\Big) + \frac{1}{2} \;\equiv R_0$$

then $\phi(u) \geq \phi(u_0)$, that is, (2.4.10) holds with this R_0.

(ii) In the sequel we omit the index n for notational simplicity, and let $u = u_n\,,\quad g = g_n\,,\quad H = H_n\,,\quad B = B_n$, where $g_n = F(u_n)$ and $H_n = F'(u_n)$. Using these notations, (2.4.4) turns into $m(B) = \phi(u - B^{-1}g)$. Further, we fix $\nu_1 > 0$ and assume that $B \in \mathcal{B}(\nu_1)$ as defined by (2.4.6).

Now we verify that

$$m(B) = \phi(u) - \langle B^{-1}g, g\rangle + \frac{1}{2}\langle HB^{-1}g, B^{-1}g\rangle + R_1 \tag{2.4.11}$$

where

$$|R_1| \leq C_1\|u-u^*\|^3 \tag{2.4.12}$$

with $C_1 > 0$ depending only on u_0 and ν_1. Let $z = B^{-1}g$. Then the Taylor expansion yields

$$m(B) = \phi(u-z) = \phi(u) - \langle \phi'(u), z\rangle + \frac{1}{2}\langle \phi''(u)z, z\rangle + R_1\,, \tag{2.4.13}$$

here the Lipschitz continuity of ϕ'' implies

$$|R_1| \leq \frac{L_0}{6}\|z\|^3 \tag{2.4.14}$$

where L_0 is the Lipschitz constant corresponding to the ball $B(u^*, R_0)$ according to assumption (iii). Here $\phi'(u) = F(u) = g$ and $\phi''(u) = F'(u) = H$, hence the definition of z and the symmetry of B yield $\langle \phi'(u), z\rangle = \langle B^{-1}g, g\rangle$, $\langle \phi''(u)z, z\rangle = \langle HB^{-1}g, B^{-1}g\rangle$ and in order to verify (2.4.12) it suffices to prove that

$$\|z\| \leq K_1\|u-u^*\| \tag{2.4.15}$$

with $K_1 > 0$ depending on u_0 and ν_1.

The Taylor expansion for ϕ' yields

$$g = \phi'(u) = \phi'(u^*) + \phi''(u^*)(u-u^*) + \varrho_1\,, \tag{2.4.16}$$

where

$$|\varrho_1| \leq \frac{L_0}{2}\|u-u^*\|^2$$

with L_0 as above. Here $\phi'(u^*) = 0$. Let P_0 be the upper spectral bound of ϕ'' on the ball $B(u^*, R_0)$, obtained from assumption (ii). Then, also using (2.4.9), we have

$$\|g\| \le P_0\|u-u^*\| + \frac{L_0}{2}\|u-u^*\|^2 \le \Big(P_0 + \frac{L_0R_0}{2}\Big)\|u-u^*\| = K_0\|u-u^*\|. \tag{2.4.17}$$

From this the assumption $B \in \mathcal{B}(\nu_1)$ yields $\|z\| = \|B^{-1}g\| \le (K_0/\nu_1)\|u-u^*\|$, hence (2.4.15) holds with $K_1 = K_0/\nu_1$ and thus (2.4.11)-(2.4.12) are verified.

(iii) Now we prove that

$$\phi(u) = \beta + \frac{1}{2}\langle H^{-1}g,^{-1} g\rangle + R_2 \tag{2.4.18}$$

where

$$|R_2| \le C_2\|u-u^*\|^3 \tag{2.4.19}$$

with $C_2 > 0$ depending only on u_0 and ν_1. Similarly to (2.4.13)-(2.4.14), we have

$$\phi(u) = \phi(u^*) + \langle\phi'(u^*), u-u^*\rangle + \frac{1}{2}\langle\phi''(u^*)(u-u^*), u-u^*\rangle + \varrho_2\,,$$

where $|\varrho_2| \le \frac{L_0}{6}\|u-u^*\|^3$. Here $\phi(u^*) = \beta$, $\phi'(u^*) = 0$ and

$$|\langle\phi''(u^*)(u-u^*), u-u^*\rangle - \langle H(u-u^*), u-u^*\rangle| \le L_0\|u-u^*\|^3$$

from $H = \phi''(u)$ and the Lipschitz condition. Hence

$$\phi(u) = \beta + \frac{1}{2}\langle H(u-u^*), u-u^*\rangle + \varrho_3\,,$$

where $|\varrho_3| \le \frac{2L_0}{3}\|u-u^*\|^3$. Therefore it remains to prove that

$$|\langle H(u-u^*), u-u^*\rangle - \langle H^{-1}g, g\rangle| \le C_3\|u-u^*\|^3. \tag{2.4.20}$$

Here (2.4.16) implies

$$g = \phi'(u) = \phi''(u^*)(u-u^*) + \varrho_1 = H(u-u^*) + (\phi''(u^*) - H)(u-u^*) + \varrho_1\,.$$

Using again the Lipschitz condition for ϕ'', we have $\|(\phi''(u^*)-H)(u-u^*)\| \le L_0\|u-u^*\|^2$, hence

$$g = H(u-u^*) + \varrho_4 \tag{2.4.21}$$

with

$$|\varrho_4| \le C_4\|u-u^*\|^2. \tag{2.4.22}$$

Setting (2.4.21) into the left-hand side expression in (2.4.20) and using the symmetry of H, we obtain

$$|\langle H(u-u^*), u-u^*\rangle - \langle H^{-1}g, g\rangle| = |\langle g-\varrho_4, H^{-1}(g-\varrho_4)\rangle - \langle H^{-1}g, g\rangle|$$

$$= |-2\langle H^{-1}g, \varrho_4\rangle + \langle H^{-1}\varrho_4, \varrho_4\rangle| \le 2|\langle H^{-1}g, \varrho_4\rangle| + |\langle H^{-1}\varrho_4, \varrho_4\rangle|\,.$$

Let p_0 be the lower spectral bound of ϕ'' on the ball $B(u^*, R_0)$, obtained from assumption (ii). Then $\|H^{-1}\| \leq 1/p_0$. Hence, using (2.4.17), (2.4.22) and (2.4.9), we have

$$|\langle H(u-u^*), u-u^*\rangle - \langle H^{-1}g, g\rangle| \leq \frac{1}{p_0}\Big(2\|g\|\|\varrho_4\| + \|\varrho_4\|^2\Big)$$

$$\leq \frac{1}{p_0}\Big(2K_0\, C_4\|u-u^*\|^3 + C_4^2\|u-u^*\|^4\Big) \leq \frac{1}{p_0}\Big(2K_0\, C_4 + R_0\, C_4^2\Big)\,\|u-u^*\|^3,$$

that is, (2.4.20) holds and thus (2.4.18)-(2.4.19) are verified.

(iv) Let us set (2.4.18) into (2.4.11) and use notation $R_3 = R_1 + R_2$:

$$m(B) = \beta + \frac{1}{2}\langle H^{-1}g,^{-1}g\rangle - \langle B^{-1}g, g\rangle + \frac{1}{2}\langle HB^{-1}g, B^{-1}g\rangle + R_3$$

$$= \beta + \frac{1}{2}\,\big\langle H(B^{-1}g - H^{-1}g),\ B^{-1}g - H^{-1}g\big\rangle + R_3 = \hat{m}(B) + R_3\,,$$

where by (2.4.12) and (2.4.19) we get $|R_3| \leq C\|u-u^*\|^3$ with $C = C_1 + C_2$. Therefore (2.4.8) is true and the proof is complete. ■

2.4.2 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 equally use preconditioning operators to define preconditioners in the inner iterations.

The convergence of such outer-inner (Newton plus PCG) iterations relies on the following two standard estimates:

(i) in the outer iteration, if the inexact Newton method contains stepwise errors δ_n, then under condition

$$\delta_n \leq const. \cdot \|F(u_n) - b\|^\gamma$$

with some constant $0 < \gamma \leq 1$, the convergence is locally of order $1+\gamma$:

$$\|F(u_{n+1}) - b\| \leq c_1\|F(u_n) - b\|^{1+\gamma}\,,$$

(ii) in the inner iteration, if the preconditioning operator yields bounds m_n and M_n, then the CG iterates $(p_n^{(k)})_{k\in\mathbf{N}}$ satisfy

$$\|F'(u_n)p_n^{(k)} + (F(u_n) - b)\|_{B_n^{-1}} \leq C_0\, Q^k\|F(u_n) - b\|_{B_n^{-1}} \qquad (k \in \mathbf{N}),$$

where $C_0 = 2$ and $Q = \frac{M_n - m_n}{M_n + m_n}$ for the CGN method, and $C_0 = 2\sqrt{M_n/m_n}$ and $Q = \frac{\sqrt{M_n}-\sqrt{m_n}}{\sqrt{M_n}+\sqrt{m_n}}$ for the symmetric CG method. This enables us to control the number of inner iterations for the prescribed outer accuracy δ_n.

In what follows, we present two classes of efficient preconditioners for the inner iterations.

(a) Symmetric problems with nonlinear principal part

In general, we have seen in section 1.3.1 that the spectral bounds m and M of a self-adjoint operator L_S imply $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \le \frac{M}{m}$ 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

$$m\|v\|_S^2 \le \langle F'(u)v, v\rangle_S \le M\|v\|_S^2 \qquad (u, v \in H_S) \tag{2.4.23}$$

with $M, m > 0$, which also 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 (2.4.23). 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

$$\langle h, v\rangle_S = \int_\Omega B(h, v) \tag{2.4.24}$$

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

$$\langle F(u), v\rangle_S = \int_\Omega \Big(a(B(u, u))\, B(u, v) - fv\Big) \qquad (u, v \in H_S), \tag{2.4.25}$$

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

$$0 < m \le a(r) \le M, \qquad 0 < m \le \tfrac{d}{dr}\Big(a(r^2)r\Big) \le M \qquad (r \ge 0). \tag{2.4.26}$$

Proposition 2.4.1 *Under assumptions (2.4.25)–(2.4.26), the operator F satisfies (2.4.23).*

Proof. Let

$$p(r^2) = \min\Big\{a(r^2),\, \tfrac{d}{dr}\Big(a(r^2)r\Big)\Big\}, \quad q(r^2) = \max\Big\{a(r^2),\, \tfrac{d}{dr}\Big(a(r^2)r\Big)\Big\} \qquad (r \ge 0), \tag{2.4.27}$$

where by (2.4.26),

$$0 < m \le p(r) \le q(r) \le M \qquad (r \ge 0). \tag{2.4.28}$$

It follows readily that for all $u, h, v \in H_S$

$$\langle F'(u)h, v\rangle_S = \int_\Omega \Big(a(B(u, u))\, B(h, v) + 2a'(B(u, u))\, B(u, h)\, B(u, v)\Big) \qquad (u, h, v \in H_S) \tag{2.4.29}$$

and hence

$$m\int_\Omega B(v, v) \le \int_\Omega p(B(u, u))\, B(v, v) \le \langle F'(u)v, v\rangle_S \le \int_\Omega q(B(u, u))\, B(v, v) \le M\int_\Omega B(v, v)\,, \tag{2.4.30}$$

which coincides with (2.4.23). ■

For a corresponding boundary value problem, the FEM solution u_h in some subspace $V_h \subset H$ must satisfy

$$\langle F(u_h), v\rangle_S = 0 \qquad (v \in V_h) \tag{2.4.31}$$

or $F(u_h) = 0$. If u_n is the nth Newton iterate, then the correction term $p_n \in V_h$ is found by solving the linearized problem

$$\langle F'(u_n)p_n, v\rangle_S = -\langle F(u_n), v\rangle_S \qquad (v \in V_h), \tag{2.4.32}$$

which now reads as follows: for all $v \in V_h$

$$\int_\Omega \Big(a(B(u_n,u_n))\, B(p_n,v) + 2a'(B(u_n,u_n))\, B(u_n,p_n)\, B(u_n,v) \Big) = -\int_\Omega \Big(a(B(u_n,u_n))\, B(u_n,v) - fv \Big) \tag{2.4.33}$$

As stated above after (2.4.23), we obtain mesh independent convergence for the inner CG iteration for problem (2.4.33).

The above bounds can be sharpened to depend on n, which can be much more efficient in practice. In fact, (2.4.30) implies

$$m_n \int_\Omega B(v,v) \le \langle F'(u_n)v, v\rangle_S \le M_n \int_\Omega B(v,v) \tag{2.4.34}$$

where

$$m_n := \inf_\Omega p(B(u_n,u_n)) \ge m, \qquad M_n := \sup_\Omega q(B(u_n,u_n)) \le M.$$

Second order equations. Various second order nonlinear elliptic problems (elasto-plastic torsion, magnetic potential, subsonic flow) lead to the following weak formulation: find $u \in H_0^1(\Omega)$ such that

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

where the given coefficient a satisfies (2.4.26). This falls into the above type where (2.4.24) is the standard $H_0^1(\Omega)$-inner product $\langle u, v\rangle_S = \int_\Omega \nabla u \cdot \nabla v$. Then estimate (2.4.23) implies mesh independent convergence of inner CG iterations such that the auxiliary problems come from the $H_0^1(\Omega)$-inner product, which means that one has to solve inner Poisson equations.

However, for some problems these bounds are too wide and convergence is very slow, as e.g. for magnetic potential problems. Then a much better preconditioning operator is the piecewise constant coefficient operator (2.3.36). The required decompositions are straightforward to define in this case when we have a scalar nonlinearity. In fact, analogously to (2.4.29), the function

$$f(x,\eta) = a(|\eta|^2)\eta$$

satisfies

$$\frac{\partial f(x,\eta)}{\partial \eta}\, \xi \cdot \xi = a(|\eta|^2)|\xi|^2 + 2a'(|\eta|^2)\ (\eta \cdot \xi)^2$$

and, using notations (2.4.27),

$$p(|\nabla u_n(x)|^2)|\xi|^2 \le \frac{\partial f}{\partial \eta}(x, \nabla u_n(x))\,\xi \cdot \xi \le q(|\nabla u_n(x)|^2)|\xi|^2\,. \tag{2.4.35}$$

This implies the local spectral bounds

$$m_i \;= \inf_{\Omega_i} p(|\nabla u_n|^2)\,, \qquad M_i \;= \sup_{\Omega_i} q(|\nabla u_n|^2)\,, \tag{2.4.36}$$

then the global bounds m_n and M_n come from (2.3.39). Altogether, these bounds are determined only by the values of $|\nabla u_n|$ and the given scalar function $a(r)$.

Conversely, prescribed condition numbers M/m can be achieved via a suitable recursive definition of the subdomains in a form

$$\Omega_i := \{x \in \Omega : \; r_{i-1} \le |\nabla u_n(x)| < r_i\} \qquad (i = 1, ..., s) \tag{2.4.37}$$

with prescribed ratios r_i/r_{i-1}, which reduces the conditioning analysis to the scalar functions p and q from (2.4.27). In practice, for a magnetic potential problem, favourable condition numbers have thus been achieved with few subdomains [10]: e.g. 6 subdomains reduced the convergence factor from $Q = 0.9785$ to $Q = 0.6711$.

Other problems. The elasto-plastic bending of clamped plates is described by a fourth order problem. Its weak formulation reads as follows: find $u \in H_0^2(\Omega)$ such that

$$\int_\Omega \overline{g}([u,u])\;[u,v] \;=\; \int_\Omega \alpha v \qquad (v \in H_0^2(\Omega)), \tag{2.4.38}$$

where $[u,v] := \frac{1}{2}(D^2u \cdot D^2v + \,\Delta u\,\Delta v)$ and the scalar material nonlinearity $\overline{g}$ satisfies (2.4.26). This falls into the above type again. Using fixed preconditioners generated by this inner product, we are led to auxiliary biharmonic problems, for which fast solvers are available [23]. For highly varying material nonlinearities, one can instead use the above described procedure to construct a piecewise constant coefficient preconditioning operator, whose weak form is

$$\langle B_n v, z\rangle = \frac{1}{2}\int_\Omega w_n(x)\;(D^2v \cdot D^2z + \,\Delta v\,\Delta z) \qquad (v, z \in H_0^2(\Omega)).$$

A similar description holds for nonlinear elasticity systems. These will be studied in subsection 2.6.5.

(b) Semilinear problems

We consider nonsymmetric systems involving second, first and zeroth order terms as well:

$$\left.\begin{aligned} &-\mathrm{div}\,(k_i\,\nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + f_i(x, u_1, \ldots, u_l) = g_i \\ &u_{i\,|\partial\Omega} = 0 \end{aligned}\right\} \qquad (i = 1, \ldots, l) \tag{2.4.39}$$

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

Assumptions 2.4.2.

(i) (Smoothness:) $k_i \in L^\infty(\Omega)$, $\mathbf{b}_i \in C^1(\overline{\Omega})^d$ and $g_i \in L^2(\Omega)$ $(i = 1, \dots, l)$, further, the function $f = (f_1, \dots, f_l) : \Omega \times \mathbf{R}^l \to \mathbf{R}^l$ 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 $i = 1, \dots, l$, further, using the notation $f'_\xi(x, \xi) := \frac{\partial f(x,\xi)}{\partial \xi}$,

$$f'_\xi(x, \xi)\, \eta \cdot \eta - \frac{1}{2} \left(\max_i \operatorname{div} \mathbf{b}_i(x) \right) |\eta|^2 \geq 0 \tag{2.4.40}$$

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 $(x, \xi_2) \in \Omega \times \mathbf{R}^l$,

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

We note that assumption (iii) implies the estimates

$$\left\| f'_\xi(x, \xi) \right\| \leq c_3 + c_4 |\xi|^{p-2}, \qquad |f(x, \xi)| \leq c_5 + c_6 |\xi|^{p-1} \tag{2.4.41}$$

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

The FEM discretization and Newton linearization of this system leads to the FEM solution of the linear elliptic problem

$$\left. \begin{array}{l} -\operatorname{div}(k_i \nabla p_i) + \mathbf{b}_i \cdot \nabla p_i + \sum\limits_{j=1}^{l} \partial_j f_i(x, \mathbf{u}_n) p_j = r_i \\ p_{i\,|\partial\Omega} = 0 \end{array} \right\} \qquad (i = 1, \dots, l) \tag{2.4.42}$$

where $r_i = g_i + \operatorname{div}(k_i \nabla u_{n,i}) - \mathbf{b}_i \cdot \nabla u_{n,i} - f_i(x, \mathbf{u}_n)$. We use the PCGN method based on a preconditioning operator S, which is the independent l-tuple of elliptic operators

$$S_i u_i := -\operatorname{div}(k_i \nabla u_i) + q_i u_i \quad \text{for} \quad u_{i\,|\partial\Omega} = 0 \qquad (i = 1, \dots, l), \tag{2.4.43}$$

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

We are interested in superlinear convergence. The following theorem, established in [4], provides this result 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 constant multiples of the squares of the so-called *Gelfand numbers* of the compact embeddings $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$, which tend to 0, see [128]. For $p = 2$, the latter are eigenvalues of the related compact operator.)

Theorem 2.4.2 *Let Assumptions 2.4.2 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*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \hat{\varepsilon}_k \quad (k=1,...,n) \quad \textit{with} \quad \hat{\varepsilon}_k := \frac{2}{km^2} \sum_{i=1}^{k} \Big(C_1 s_i^{(2)} + C_2 s_i^{(p)}\Big) \to 0 \tag{2.4.44}$$

as $k \to \infty$, and here the constants $C_1, C_2 > 0$ and hence the sequence $(\hat{\varepsilon}_k)_{k \in \mathbf{N}^+}$ are independent of V_h and $\mathbf{u}_n$.

Proof. The general superlinear convergence estimate for this problem is given by Theorem 1.2.9. To prove the desired independence result, we must show that the sequence ε_k in (1.2.76) satisfies $\varepsilon_k \leq \hat{\varepsilon}_k$ if $Q_S = Q_S^{(n)}$ corresponding to linearization at $\mathbf{u}_n$, further, that $\hat{\varepsilon}_k \to 0$. The divergence theorem yields for $\mathbf{v}, \mathbf{z} \in H_0^1(\Omega)^l$

$$\int_\Omega (\mathbf{b}_i \cdot \nabla v_i) z_i = -\int_\Omega v_i (\mathbf{b}_i \cdot \nabla z_i) - \int_\Omega (\operatorname{div} \mathbf{b}_i) v_i z_i \,, \tag{2.4.45}$$

hence from (1.2.61) and (2.4.41)

$$\|Q_S^{(n)} \mathbf{v}\|_S = \sup_{\substack{\mathbf{z} \in H_0^1(\Omega)^l \\ \|\mathbf{z}\|_S = 1}} |\langle Q_S^{(n)} \mathbf{v}, \mathbf{z} \rangle_S|$$

$$= \sup_{\substack{\mathbf{z} \in H_0^1(\Omega)^l \\ \|\mathbf{z}\|_S = 1}} \left| \sum_{i=1}^{l} \int_\Omega \Big(-v_i \left(\mathbf{b}_i \cdot \nabla z_i\right) + \big(\sum_{j=1}^{l} \partial_j f_i(x, \mathbf{u}_n) v_j - q_i v_i - \left(\operatorname{div} \mathbf{b}_i\right) v_i \big) \, z_i \Big) \right|$$

$$\equiv \sup_{\substack{\mathbf{z} \in H_0^1(\Omega)^l \\ \|\mathbf{z}\|_S = 1}} \left| \int_\Omega \Big(-\mathbf{v} \cdot (\mathbf{b} \cdot \nabla \mathbf{z}) + \big(f'_\xi(x, \mathbf{u}_n) - (\mathbf{q} + \operatorname{div} \mathbf{b}) I \,\big) \mathbf{v} \cdot \mathbf{z} \Big) \right|$$

$$\leq \sup_{\substack{\mathbf{z} \in H_0^1(\Omega)^l \\ \|\mathbf{z}\|_S = 1}} \left(\max_i \|\mathbf{b}_i\|_{L^\infty(\Omega)^l} \int_\Omega |\mathbf{v}| \, |\nabla \mathbf{z}| \; + \; \big(c_3 + \max_i \big\| q_i + \operatorname{div} \mathbf{b}_i \big\|_{L^\infty(\Omega)}\big) \int_\Omega |\mathbf{v}\mathbf{z}| \right.$$

$$\left. + \, c_4 \int_\Omega |\mathbf{u}_n|^{p-2} |\mathbf{v}\mathbf{z}| \right) \tag{2.4.46}$$

$$\leq \sup_{\substack{\mathbf{z} \in H_0^1(\Omega)^l \\ \|\mathbf{z}\|_S = 1}} \Big(\max_i \|\mathbf{b}_i\|_{L^\infty(\Omega)^l} \|\mathbf{v}\|_{L^2(\Omega)^l} \|\nabla \mathbf{z}\|_{L^2(\Omega)^{ld}} + \big(c_3 + \max_i \big\| q_i + \operatorname{div} \mathbf{b}_i \big\|_{L^\infty(\Omega)}\big) \|\mathbf{v}\|_{L^2(\Omega)^l} \|\mathbf{z}\|_{L^2(\Omega)^l}$$

$$+ \, c_4 \|\mathbf{u}_n\|_{L^p(\Omega)^l}^{p-2} \, \|\mathbf{v}\|_{L^p(\Omega)^l} \, \|\mathbf{z}\|_{L^p(\Omega)^l} \Big), \tag{2.4.47}$$

where in the last term Hölder's inequality has been used for the case $\frac{p-2}{p} + \frac{1}{p} + \frac{1}{p} = 1$. Here we have $\|\nabla \mathbf{z}\|_{L^2(\Omega)^{ld}} = \|\mathbf{z}\|_{H_0^1} \leq \frac{1}{\sqrt{m}} \cdot \|\mathbf{z}\|_S = \frac{1}{\sqrt{m}}$ and $\|\mathbf{z}\|_{L^p(\Omega)^l} \leq \frac{C_p}{\sqrt{m}} \cdot \|\mathbf{z}\|_S = \frac{C_p}{\sqrt{m}}$ for all $p \leq p^*$. Therefore

$$\|Q_S^{(n)} \mathbf{v}\|_S \leq \Big(\frac{1}{\sqrt{m}} \max_i \|\mathbf{b}_i\|_{L^\infty(\Omega)^l} \|\mathbf{v}\|_{L^2(\Omega)^l}$$

$$+ \frac{C_2}{\sqrt{m}} \left(c_3 + \max_i \big\| q_i + \operatorname{div} \mathbf{b}_i \big\|_{L^\infty(\Omega)}\right) \|\mathbf{v}\|_{L^2(\Omega)^l} + \, c_4 \frac{C_p}{\sqrt{m}} \, \|\mathbf{u}_n\|_{L^p(\Omega)^l}^{p-2} \, \|\mathbf{v}\|_{L^p(\Omega)^l} \Big),$$

moreover,

$$\|\mathbf{u}_n\|_{L^p(\Omega)^l} \le C_p \cdot \|\mathbf{u}_n\|_{H_0^1} \le C_p\, R_0\,, \tag{2.4.48}$$

hence

$$\|Q_S^{(n)}\mathbf{v}\|_S \le \mathit{const.} \cdot \|\mathbf{v}\|_{L^2(\Omega)^l} + \mathit{const.} \cdot \|\mathbf{v}\|_{L^p(\Omega)^l},$$

which implies

$$\|Q_S^{(n)}\mathbf{v}\|_S^2 \le K_1\|\mathbf{v}\|^2_{L^2(\Omega)^l} + K_2\|\mathbf{v}\|^2_{L^p(\Omega)^l} \tag{2.4.49}$$

and here K_1, K_2 are independent of h and $\mathbf{u}_n$.

Now setting $v_i = z_i$ in (2.4.45),

$$\int_\Omega (\mathbf{b}_i \cdot \nabla v_i)\, v_i = \; -\int_\Omega \frac{1}{2}\,(\operatorname{div}\mathbf{b}_i)\, v_i^2$$

hence

$$\begin{aligned}
\Big|\langle Q_S^{(n)}\mathbf{v},\mathbf{v}\rangle_S\Big| &= \left|\sum_{i=1}^l \int_\Omega \Big((\mathbf{b}_i\cdot\nabla v_i)\, v_i + \big(\sum_{j=1}^l \partial_j f_i(x,\mathbf{u}_n)v_j - q_i v_i\big)\, v_i\Big)\right| \\
&\equiv \left|\int_\Omega \Big((\mathbf{b}\cdot\nabla\mathbf{v})\cdot\mathbf{v} + (f'_\xi(x,\mathbf{u}_n) - \mathbf{q}I)\mathbf{v}\cdot\mathbf{v}\Big)\right| \\
&\le \int_\Omega \max_i \big|q_i + \frac{1}{2}\operatorname{div}\mathbf{b}_i\big|\, |\mathbf{v}|^2 + \int_\Omega \big(c_3 + c_4|\mathbf{u}_n|^{p-2}\big)\, |\mathbf{v}|^2 \\
&\le \big(c_3 + \max_i \big\|q_i + \frac{1}{2}\operatorname{div}\mathbf{b}_i\big\|_{L^\infty(\Omega)}\big)\, \|\mathbf{v}\|^2_{L^2(\Omega)^l} + c_4\|\mathbf{u}_n\|^{p-2}_{L^p(\Omega)^l}\, \|\mathbf{v}\|^2_{L^p(\Omega)^l}.
\end{aligned}$$

Using (2.4.48) again, we obtain

$$\Big|\langle Q_S^{(n)}\mathbf{v},\mathbf{v}\rangle_S\Big| \le K_3\|\mathbf{v}\|^2_{L^2(\Omega)^l} + K_4\|\mathbf{v}\|^2_{L^p(\Omega)^l} \tag{2.4.50}$$

and here K_3, K_4 are independent of h and $\mathbf{u}_n$. Now let $H_S = H_0^1(\Omega)^l$ with the S-inner product. The variational characterization of the eigenvalues yields

$$\big|\lambda_i\Big(\big(Q_S^{(n)}\big)^* + Q_S^{(n)}\Big)\big| = \min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{|\langle\Big(\big(Q_S^{(n)}\big)^* + Q_S^{(n)}\Big)\mathbf{v},\mathbf{v}\rangle_S|}{\|\mathbf{v}\|_S^2} = 2\min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{|\langle Q_S^{(n)}\mathbf{v},\mathbf{v}\rangle_S|}{\|\mathbf{v}\|_S^2}$$

and

$$s_i(Q_S^{(n)})^2 = \lambda_i\big(\big(Q_S^{(n)}\big)^* Q_S^{(n)}\big) = \min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{\langle\big(Q_S^{(n)}\big)^* Q_S^{(n)}\mathbf{v},\mathbf{v}\rangle_S}{\|\mathbf{v}\|_S^2} = \min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{\|Q_S^{(n)}\mathbf{v}\|_S^2}{\|\mathbf{v}\|_S^2}\,,$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace. Summing up and using (2.4.50) and (2.4.49), respectively, we obtain

$$\big|\lambda_i\Big(\big(Q_S^{(n)}\big)^* + Q_S^{(n)}\Big)\big| + s_i(Q_S^{(n)})^2 \le C_1 \min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{\|\mathbf{v}\|^2_{L^2(\Omega)^l}}{\|\mathbf{v}\|_S^2} + C_2 \min_{H_{i-1}\subset H_S}\; \max_{\mathbf{v}\perp H_{i-1}} \frac{\|\mathbf{v}\|^2_{L^p(\Omega)^l}}{\|\mathbf{v}\|_S^2}$$

where $C_1 = 2K_3 + K_1$, $C_2 = 2K_4 + K_2$. Here both terms on the r.h.s. tend to 0 as $i \to \infty$, owing to the compactness of the embeddings $H_0^1(\Omega)^l \subset L^2(\Omega)^l$ and $H_0^1(\Omega)^l \subset L^p(\Omega)^l$. (In particular, the first min-max term gives the reciprocal of the eigenvalues of S in $L^2(\Omega)^l$.) That is, the sequence $(\hat\varepsilon_k)$ is constant times the arithmetic means of a sequence that tends to zero, hence, as is well-known, $\hat\varepsilon_k$ itself tends to zero. ■

Remark 2.4.1 (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 (1.2.66) holds [4].

(ii) Instead of the above Dirichlet problem, one could include mixed boundary conditions or interface conditions, see [5] and the numerical tests in subsection 2.6.6.

2.5 Newton's method: a characterization of mesh independence

A missing part of the previous theory in this part 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 [3], and then a lot of important work has been done, see [20, 29, 130, 153]. 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 were based on the underlying Lipschitz continuity for the derivatives of the operator.

However, in all 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. It will be shown that mesh uniform quadratic estimates in fact cannot be produced unless the principal part is linear. For this, 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 in fact as follows: in the case of a nonlinear principal part, as we will prove in Corollary 2.5.1, *the derivative F' of the differential operator is not locally Lipschitz continuous* in the corresponding Sobolev space.

We finally mention that although the underlying property can be given very simply, the exact proofs will require very lengthy and technical calculations, following [88].

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

$$\left\{\begin{aligned} -\operatorname{div} f(x,\nabla u) + q(x,u) &= g(x) \quad \text{in } \Omega \\ f(x,\nabla u)\cdot\nu + s(x,u) &= \gamma(x) \quad \text{on } \Gamma_N \\ u &= 0 \qquad \text{on } \Gamma_D\,. \end{aligned}\right. \tag{2.5.1}$$

We impose the following conditions:

Assumptions 2.5.1.

(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 $\Lambda \geq \lambda > 0$ 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 $(x, \xi_2) \in \Omega \times \mathbf{R}$,

$$\Big|q'_\xi(x, \xi_1) - q'_\xi(x, \xi_2)\Big| \leq \Big(c_1 + c_2 \,(\max\, |\xi_1|, |\xi_2|)^{p_1-3}\Big)|\xi_1 - \xi_2|. \tag{2.5.2}$$

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 $(x, \xi_2) \in \Gamma_N \times \mathbf{R}$,

$$\Big|s'_\xi(x, \xi_1) - s'_\xi(x, \xi_2)\Big| \leq \Big(d_1 + d_2 \,(\max\, |\xi_1|, |\xi_2|)^{p_2-3}\Big)|\xi_1 - \xi_2|. \tag{2.5.3}$$

The Sobolev space

$$H^1_D(\Omega) := \{u \in H^1(\Omega) : \, u_{|\Gamma_D} = 0\}\,, \tag{2.5.4}$$

corresponding to the Dirichlet boundary Γ_D, is endowed with the inner product

$$\langle u, v\rangle := \int_\Omega \nabla u \cdot \nabla v, \qquad \|u\|_{H^1_D} = \|\nabla u\|_{L^2(\Omega)}. \tag{2.5.5}$$

Condition $\Gamma_D \neq \emptyset$ in assumption (i) ensures that (2.5.5) is positive definite. Let p_1, p_2 be real numbers as in assumption (v). Then [1] there hold the Sobolev embeddings

$$H^1_D(\Omega) \subset L^{p_1}(\Omega), \qquad \|u\|_{L^{p_1}(\Omega)} \leq K_{p_1,\Omega}\|u\|_{H^1_D} \qquad (u \in H^1_D(\Omega)) \tag{2.5.6}$$

$$H^1_D(\Omega)_{|\Gamma_N} \subset L^{p_2}(\Gamma_N), \qquad \|u\|_{L^{p_2}(\Gamma_N)} \leq K_{p_2,\Gamma_N}\|u\|_{H^1_D} \qquad (u \in H^1_D(\Omega)) \tag{2.5.7}$$

with suitable constants $K_{p_1,\Omega}, K_{p_2,\Gamma_N} > 0$.

We will consider finite element subspaces under the following assumptions, which will allow us to use standard FE theory [34].

Definition 2.5.1 A family of FEM subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$ will be called *admissible* if

(i) $V_h \subset C(\overline{\Omega}) \cap H^1_D(\Omega) \cap W^{1,\infty}(\Omega)$ for all $V_h \subset \mathcal{V}$;

(ii) V_h is a regular affine family;

(ii) V_h contains all piecewise linear polynomials, but the degrees of freedom involve at most first derivatives.

Hereby, as usual, a *regular affine family* means that all elements are affine-equivalent to a reference element, and the element diameters are bounded by constant times the maximal inscribed ball diameters [34].

The standard Newton's method reads as follows. Let $u_0 \in V_h$ be arbitrary, and let the sequence (u_n) be defined by the following iteration. If, for $n \in \mathbf{N}$, u_n is obtained, then

$$u_{n+1} = u_n + p_n \qquad (n \in \mathbf{N}), \tag{2.5.8}$$

where $p_n \in V_h$ is the solution of the linear auxiliary problem

$$F_h'(u_n)p_n = -F_h(u_n). \tag{2.5.9}$$

Definition 2.5.2 Problem (2.5.1) satisfies the *mesh independence principle for quadratic convergence (MIPQC)* of Newton's method for admissible discretizations if under Assumptions 2.5.1, 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 sequences (2.5.8)–(2.5.9) satisfy

$$\sup\left\{ \frac{\|F_h(u_{n+1})\|_{H^1_D}}{\|F_h(u_n)\|^2_{H^1_D}} : \quad h < h_0, \quad \|u_0 - u_h\|_{H^1_D} < \delta, \quad n \in \mathbf{N} \right\} < \infty. \tag{2.5.10}$$

Theorem 2.5.1 *Let Assumptions 2.5.1 hold and $f \in C^2(\Omega \times \mathbf{R}^d,\ \mathbf{R}^d)$. Problem (2.5.1) satisfies the MIPQC of Definition 2.5.2 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 2.5.1 themselves.

For both proofs, let us decompose the operator F in

$$F = G + R, \tag{2.5.11}$$

where for any $u \in H^1_D(\Omega)$,

$$\langle G(u), v\rangle_{H^1_D} \equiv \int_\Omega f(x, \nabla u) \cdot \nabla v \qquad (v \in H^1_D(\Omega)); \tag{2.5.12}$$

$$\langle R(u), v\rangle_{H^1_D} \equiv \int_\Omega q(x,u)\, v + \int_{\Gamma_N} s(x,u)\, v - \int_\Omega gv - \int_{\Gamma_N} \gamma v\, d\sigma \qquad (v \in H^1_D(\Omega)). \tag{2.5.13}$$

(a) Proof of the 'if' part

We verify the following

Proposition 2.5.1 *Let Assumptions 2.5.1 hold. If* $\eta \mapsto f(x,\eta)$ *is linear, i.e. the elliptic equation is semilinear, then problem (2.5.1) satisfies the MIPQC of Definition 2.5.2.*

Proof. It suffices to show that F_h' is locally Lipschitz continuous independently of h. The required Lipschitz continuity has been proved in many related specific situations, see e.g. [55], and a similar derivation can be used in our general setting. For completeness, we briefly summarize the proof. For brevity, we omit indices h from elements of V_h. Using the linearity of f and the decomposition (2.5.11), we have

$$\langle (F_h'(u) - F_h'(v))w, z\rangle = \int_\Omega \big(\frac{\partial q}{\partial \xi}(x,u) - \frac{\partial q}{\partial \xi}(x,v)\big)\, wz \; + \int_{\Gamma_N} \big(\frac{\partial s}{\partial \xi}(x,u) - \frac{\partial q}{\partial \xi}(x,v)\big)\, wz$$

$$=: \; \langle (R_1'(u) - R_1'(v))w, z\rangle + \langle (R_2'(u) - R_2'(v))w, z\rangle \qquad (u,v,w,z \in V_h).$$

Here R_2' satisfies

$$|\langle (R_2'(u) - R_2'(v))w, z\rangle| \le \int_{\Gamma_N} \Big(d_1 + d_2 \,(\max\, |u|, |v|)^{p_2-3} \Big) |u-v|\, |w|\, |z|$$

$$\le d_1 \, \|u-v\|_{L^3(\Gamma_N)}\, \|w\|_{L^3(\Gamma_N)} \|z\|_{L^3(\Gamma_N)}$$

$$+ \; d_2 \, \big(\max\, \|u\|_{L^{p_2}}, \|v\|_{L^{p_2}(\Gamma_N)}\big)^{p_2-3} \, \|u-v\|_{L^{p_2}(\Gamma_N)}\, \|w\|_{L^{p_2}(\Gamma_N)} \|z\|_{L^{p_2}(\Gamma_N)}$$

where Hölder's inequality has been used for the cases $\frac{1}{3}+\frac{1}{3}+\frac{1}{3} = 1$ and $\frac{p_2-3}{p_2}+\frac{1}{p_2}+\frac{1}{p_2}+\frac{1}{p_2} = 1$. Then, also using (2.5.7), we have

$$\|R_2'(u) - R_2'(v)\| = \sup_{\substack{w,z\in V_h \\ \|w\|=\|z\|=1}} |\langle (R_2'(u) - R_2'(v))w, z\rangle|$$

$$\le \Big(d_1\, K_{3,\Gamma_N}^3 \; + \; d_2\, K_{p_2,\Gamma_N}^{p_2} \, \Big(\max\, \|u\|_{H_D^1}, \|v\|_{H_D^1}\Big)^{p_2-3} \Big) \, \|u-v\|_{H_D^1}\, .$$

A similar calculation holds for R_1' with obviously replaced constants. Hence we obtain

$$\tilde{L}(r) = c_1\, K_{3,\Omega}^3 \; + \; c_2\, K_{p_1,\Omega}^{p_1}\, r^{p_1-3} + d_1\, K_{3,\Gamma_N}^3 \; + \; d_2\, K_{p_2,\Gamma_N}^{p_2}\, r^{p_2-3}$$

independently of h. ■

(b) Proof of the 'only if' part

Now we must verify the following

Proposition 2.5.2 *Let Assumptions 2.5.1 hold and* $f \in C^2(\Omega \times \mathbf{R}^d,\ \mathbf{R}^d)$. *If* $\eta \mapsto f(x,\eta)$ *is not linear, i.e. the elliptic equation is not semilinear, then problem (2.5.1) does not satisfy the MIPQC of Definition 2.5.2.*

The stated mesh-dependence will be derived in Proposition 2.5.3 after a series of lemmata. First we consider the meaning of the required negation:

Remark 2.5.1 To show the contrary of Definition 2.5.2, one must find right-hand sides $g \in L^2(\Omega)$ and $\gamma \in L^2(\Gamma_N)$ in problem (2.5.1) with the following property:

for any constants $h_0 > 0$ and $\delta > 0$, taking into account admissible FEM subspaces $V_h \subset H^1_D(\Omega)$ with mesh parameter h, and initial guesses $u_0 \in V_h$, the sequences (2.5.8)–(2.5.9) satisfy

$$\sup\left\{ \frac{\|F_h(u_{n+1})\|_{H^1_D}}{\|F_h(u_n)\|^2_{H^1_D}} : \quad h < h_0, \;\; \|u_0 - u_h\|_{H^1_D} < \delta, \;\; n \in \mathbf{N} \right\} \; = \; \infty. \tag{2.5.14}$$

Now, note that Assumption 2.5.1 (iv) implies

$$\left\| \frac{\partial^2 f(x,\eta)}{\partial \eta^2} \right\|_{\mathbf{R}^{n\times n\times n}} \le l_f \tag{2.5.15}$$

for all $(x,\eta) \in \Omega \times \mathbf{R}^n$.

Lemma 2.5.1 *Let $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$. If f is not linear then there exists a point $(x_0,\eta_0) \in \Omega \times \mathbf{R}^n$ and a vector $v \in \mathbf{R}^n$, $|v| = 1$ such that $\frac{\partial^2 f(x_0,\eta_0)}{\partial \eta^2}(v,v,v) \; > \; 0$.*

Proof. f is not linear if and only if $\partial^2_\eta f := \frac{\partial^2 f}{\partial \eta^2} \not\equiv \mathbf{0}$, i.e. there exists a point $(x_0,\eta_0) \in \Omega \times \mathbf{R}^n$ where $\partial^2_\eta f(x_0,\eta_0)$ is not the zero tensor. Owing to $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$, the tensor $\partial^2_\eta f(x_0,\eta_0)$ is symmetric, hence the absolute-value estimate [152] states that

$$\left\| \partial^2_\eta f(x_0,\eta_0) \right\|_{\mathbf{R}^{N^3}} = \sup_{|v|=1} \partial^2_\eta f(x_0,\eta_0)(v,v,v),$$

from which our lemma follows. ■

Lemma 2.5.2 *If Assumption 2.5.1 (iii) holds, then for all $u, v, z \in H^1_D(\Omega)$*

$$|\langle G'(u)v, z\rangle| \le \Lambda \|\nabla v\|_{L^2(\Omega)} \|\nabla z\|_{L^2(\Omega)} \, . \tag{2.5.16}$$

Proof. It readily follows from the assumption and the Cauchy-Schwarz inequality. ■

Lemma 2.5.3 *Let $v \in \mathbf{R}^n$ be a given vector such that $|v| = 1$, and $B_{r_0}(x_0) \subset \Omega$ be a given ball with center x_0 and radius r_0. Then there exists a subset $D \subset B_{r_0}(x_0)$ such that the following holds. For any integer $m \ge 1$ and number $M > 0$ there exists a function $p_\alpha \in C^{(2m-1)}(\Omega)$ with the following properties:*

(i) $\nabla p_\alpha(x) = \varphi(x)\, v \quad (x \in D)$ *for some nonnegative function* $\varphi : D \to \mathbf{R}^+$.

(ii) $\|\nabla p_\alpha\|_{L^3(D)} \ge M$.

(iii) $\|\nabla p_\alpha\|_{L^3(\Omega \setminus D)} \le 1$.

(iv) $\|\nabla p_\alpha\|_{L^2(\Omega)} \le 1$.

(v) p_α *has a compact support in* Ω.

Proof. Let $r > 0$ be a given number. Let $e_2, \dots, e_n \in \mathbf{R}^n$ be vectors such that $\{v, e_2, \dots, e_n\}$ is an orthonormal basis in $\mathbf{R}^n$, and $H_r := \{x_0 + \sum_{i=2}^{N} c_i e_i : \; |c_i| \le r, \; i = 2, \dots, n\}$ (a hypercube orthogonal to v). For given points $x \in \mathbf{R}^n$ and $y \in H_r$, if $x = x_0 + a_1 v + \sum_{i=2}^{N} a_i e_i$ and $y = x_0 + \sum_{i=2}^{N} b_i e_i$, then we define

$$d(x,y) := \begin{cases} d_+(x,y) := \Big(a_1^2 + \sum_{i=2}^{N} (a_i - b_i)^2\Big)^{1/2} & \text{if } \; a_1 = \langle x - x_0, v\rangle \ge 0, \\ d_-(x,y) := \Big(a_1^4 + \sum_{i=2}^{N} (a_i - b_i)^2\Big)^{1/2} & \text{if } \; a_1 = \langle x - x_0, v\rangle \le 0; \end{cases} \tag{2.5.17}$$

$$d(x, H_r) := \inf_{y \in H_r} d(x,y). \tag{2.5.18}$$

Let r be chosen so small that the set $E_r := \{x \in \mathbf{R}^n : \; d(x, H_r) \le r\} \subset B_{r_0}(x_0)$. We fix this r and define

$$D := D_r := \{x_0 + c_1 v + \sum_{i=2}^{N} c_i e_i : \; 0 \le c_i \le r, \; i = 1, \dots, n\}.$$

Let $m \ge 1$ be a fixed integer. For another given integer $\alpha \ge 1$ we define

$$S_\alpha(t) := 1 - \alpha^{-\frac{1}{4m}} \, e^{-\alpha t^m} \qquad (t \ge 0).$$

Then

$$S'_\alpha(t) = s_\alpha(t) := m\alpha^{\frac{4m-1}{4m}} \; t^{m-1} \; e^{-\alpha t^m} \qquad (t \ge 0).$$

Let us introduce

$$q_\alpha(x) := S_\alpha(d(x, H_r)^2) = 1 - \alpha^{-\frac{1}{4m}} \, e^{-\alpha \, d(x,H_r)^{2m}} \qquad (x \in \Omega).$$

We will first prove that for large enough α, the function q_α will satisfy similar properties as (i)-(iv). Then we will look for p_α in the form $p_\alpha = C q_\alpha \psi$ for some proper constant $C > 0$ and function $\psi \in C_0^\infty(\Omega)$.

(1) *The properties of q_α.* Let $k, l \ge 0$ be given numbers. It is well-known that

$$J^{(k)}(x) := \int_0^x s^k e^{-s^2} ds \to \int_0^{+\infty} s^k e^{-s^2} ds < +\infty \qquad \text{as } \; x \to +\infty,$$

therefore, for given constant $p > 0$ and positive integer $\alpha \in \mathbf{N}^+$, using the substitution $s = \sqrt{p\alpha}\varrho$,

$$I_{p\alpha}^{(k)}(r) := \int_0^r t^k e^{-p\alpha t^2} dt = (p\alpha)^{-\frac{k+1}{2}} J^{(k)}(\sqrt{p\alpha} r) = O(\alpha^{-\frac{k+1}{2}}) \qquad \text{as } \; \alpha \to +\infty. \tag{2.5.19}$$

From this, using the substitution $t = \varrho^l$,

$$J_{p\alpha}^{[k,l]}(r) := \int_0^r \varrho^k e^{-p\alpha\varrho^{2l}} d\varrho = \frac{1}{l} \int_0^{r^l} t^{\frac{k+1}{l}-1} e^{-p\alpha t^2} dt = \frac{1}{l} \, I_{p\alpha}^{(\frac{k+1}{l}-1)}(r^l) = O(\alpha^{-\frac{k+1}{2l}})$$

as $\alpha \to +\infty$.

For simplicity, we first consider in detail the 2-dimensional case, i.e. when $\Omega \subset \mathbf{R}^2$. This will also clarify the procedure in higher dimensions.

(1a) *Construction in 2D.* Now $H_r = \{x_0 + c_2 e_2 :\ |c_2| \le r\}$ is a segment. Further, we rewrite D_r and introduce five additional sets:

$$
\begin{aligned}
&D_r = \{x = x_0 + \xi_1 v + \xi_2 e_2 :\ 0 \le \varrho := \xi_1 \le r,\ |\xi_2| \le r\}.\\
&B_r^{1,+} := \{x = x_0 + \xi_1 v + (r + \xi_2) e_2 :\ 0 < \xi_i \le r\ (i = 1,2),\ \varrho := \sqrt{\xi_1^2 + \xi_2^2} \le r\},\\
&B_r^{2,+} := \{x = x_0 + \xi_1 v - (r + \xi_2) e_2 :\ 0 < \xi_i \le r\ (i = 1,2),\ \varrho := \sqrt{\xi_1^2 + \xi_2^2} \le r\},\\
&B_r^{1,-} := \{x = x_0 - \xi_1 v + (r + \xi_2) e_2 :\ 0 < \xi_i \le r\ (i = 1,2),\ \varrho := \sqrt{\xi_1^4 + \xi_2^2} \le r\},\\
&B_r^{2,-} := \{x = x_0 - \xi_1 v - (r + \xi_2) e_2 :\ 0 < \xi_i \le r\ (i = 1,2),\ \varrho := \sqrt{\xi_1^4 + \xi_2^2} \le r\}.\\
&D_r^- := \{x = x_0 - \xi_1 v + \xi_2 e_2 :\ 0 \le \xi_1,\ \varrho := \xi_1^2 \le r,\ \ |\xi_2| \le r\},
\end{aligned}
$$

Then $E_r = D_r \cup B_r^{1,+} \cup B_r^{2,+} \cup D_r^- \cup B_r^{1,-} \cup B_r^{2,-}$. In each set we have $\varrho = d(x, H_r)$, in particular, the infimum in (2.5.18) is attained for a vertex $y = x_0 \pm r e_2$ in the case of $B_r^{i,\pm}$ $(i = 1, 2)$ and attained for an interior point $y = x_0 + \xi_2 e_2$ of H_r in the case of D_r and D_r^-.

We can define $\sigma(\xi) := d(x, H_r)$ (then $\sigma(\xi) = \varrho$). Since ξ_1, ξ_2 are (local) coordinates in other orthonormal bases, we have $|\nabla(d(x, H_r)^2)| = |\nabla(\sigma(\xi)^2)|$. In detail:

$$
\begin{aligned}
&\text{on } D_r\text{:} \quad \varrho^2 = \sigma(\xi)^2 = \xi_1^2, \quad |\nabla(\sigma(\xi)^2)| = 2\xi_1;\\
&\text{on } D_r^-\text{:} \quad \varrho^2 = \sigma(\xi)^2 = \xi_1^4, \quad |\nabla(\sigma(\xi)^2)| = 4\xi_1^3;\\
&\text{on } B_r^{i,+}\ (i = 1,2)\text{:} \quad \varrho^2 = \sigma(\xi)^2 = \xi_1^2 + \xi_2^2, \quad |\nabla(\sigma(\xi)^2)| = 2\sqrt{\xi_1^2 + \xi_2^2};\\
&\text{on } B_r^{i,-}\ (i = 1,2)\text{:} \quad \varrho^2 = \sigma(\xi)^2 = \xi_1^4 + \xi_2^2, \quad |\nabla(\sigma(\xi)^2)| = 2\sqrt{4\xi_1^6 + \xi_2^2}.
\end{aligned}
$$

First we observe that q_α is the composition of an analytic function with $d(x, H_r)^{2m}$. The latter equals $\sigma(\xi)^{2m}$, which can be defined on the whole Ω by the same formula as on E_r (if we drop the condition $\varrho \le r$), and it is easy to see that $\sigma(\xi)^{2m}$ is a piecewise polynomial which has $2m - 1$ vanishing derivatives on the boundaries in those variables that change across the considered subdomains. Hence $q_\alpha \in C^{(2m-1)}(\Omega)$. Now we study some properties of q_α analogous to (i)-(iv).

If $x = x_0 + \xi_1 v + \xi_2 e_2 \in D_r$, then $d(x, H_r) = \varrho = \xi_1 = \langle x - x_0, v\rangle$, hence $q_\alpha(x) = S_\alpha(\langle x - x_0, v\rangle^2)$ and hence

$$
\nabla q_\alpha(x) = 2 s_\alpha(\langle x - x_0, v\rangle^2)\ \langle x - x_0, v\rangle\ v\ =\ \varphi(x)\, v \qquad (x \in D_r), \tag{2.5.20}
$$

where $\varphi(x) = 2 s_\alpha(\langle x - x_0, v\rangle^2) \langle x - x_0, v\rangle = 2 s_\alpha(\xi_1)\xi_1$ is a nonnegative function on D_r. That is, the analogue of property (i) holds.

For any $x \in E_r$,

$$
\begin{aligned}
|\nabla q_\alpha(x)| &= |s_\alpha(d(x, H_r)^2)|\, |\nabla(d(x, H_r)^2)|\\
&= m\alpha^{\frac{4m-1}{4m}}\ d(x, H_r)^{2m-2}\ e^{-\alpha d(x, H_r)^{2m}}\ |\nabla(d(x, H_r)^2)|\\
&= m\alpha^{\frac{4m-1}{4m}}\ \varrho^{2m-2}\ e^{-\alpha \varrho^{2m}}\ |\nabla(\sigma(\xi)^2)|.
\end{aligned}
\tag{2.5.21}
$$

Let us now calculate the magnitude of $\|\nabla q_\alpha\|_{L^p}^p$ for $p = 2, 3$ on the considered subdomains, using (2.5.21). Consider first D_r. As seen above, we have $\varrho = \xi_1$ and $|\nabla(\sigma(\xi)^2)| =$

$2\xi_1$ on D_r, hence $\varrho^{2m-2}\, e^{-\alpha\varrho^{2m}}\, |\nabla(\sigma(\xi)^2)| = 2\xi_1^{2m-1}\, e^{-\alpha\xi_1^{2m}}$. The transformation that carries the points $(\xi_1,\xi_2) \in D_r^0 := [0,r]\times[-r,r]$ to $x = x_0 + \xi_1 v + \xi_2 e_2 \in D_r$ is orthogonal, hence

$$\begin{aligned}
\|\nabla q_\alpha\|^p_{L^p(D_r)} &\equiv \int_{D_r} |\nabla q_\alpha(x)|^p dx_1 dx_2 = m^p\,\alpha^{\frac{(4m-1)p}{4m}} \int_{D_r^0} \left(2\xi_1^{2m-1}\, e^{-\alpha\xi_1^{2m}}\right)^p d\xi_1 d\xi_2 \\
&= 2^{p+1} r m^p\,\alpha^{\frac{(4m-1)p}{4m}} \int_0^r \xi_1^{(2m-1)p}\ e^{-p\alpha\xi_1^{2m}}\ d\xi_1 \\
&= c_1(r,m,p)\ \alpha^{\frac{(4m-1)p}{4m}}\, J_{p\alpha}^{[(2m-1)p,\, m]}(r) = O(\alpha^{\frac{(4m-1)p}{4m} - \frac{(2m-1)p+1}{2m}}) = O(\alpha^{\frac{p-2}{4m}})
\end{aligned}$$

as $\alpha \to +\infty$, where $c_1(r,m,p) := 2^{p+1} r m^p$. Hence

$$\|\nabla q_\alpha\|^2_{L^2(D_r)} = O(1), \text{ i.e.} \leq K \qquad \text{and} \quad \|\nabla q_\alpha\|^3_{L^3(D_r)} = O(\alpha^{\frac{1}{4m}}) \to \infty \tag{2.5.22}$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

Now consider $B_r^{1,+}$. As seen above, we have $\varrho = \sqrt{\xi_1^2 + \xi_2^2}$ and $|\nabla(\sigma(\xi)^2)| = 2\sqrt{\xi_1^2+\xi_2^2} = 2\varrho$ on $B_r^{1,+}$. Hence $\varrho^{2m-2}\, e^{-\alpha\varrho^{2m}}\, |\nabla(\sigma(\xi)^2)| = 2\varrho^{2m-1}\, e^{-\alpha\varrho^{2m}}$, and, using polar transformation,

$$\begin{aligned}
\|\nabla q_\alpha\|^p_{L^p(B_r^{1,+})} &\equiv \int_{B_r^{1,+}} |\nabla q_\alpha(x)|^p dx_1 dx_2 = \pi 2^{p-1} m^p\ \alpha^{\frac{(4m-1)p}{4m}} \int_0^r \varrho^{(2m-1)p+1}\ e^{-p\alpha\varrho^{2m}}\ d\varrho \\
&= c_2(r,m,p)\ \alpha^{\frac{(4m-1)p}{4m}}\, J_{p\alpha}^{[(2m-1)p+1,\, m]}(r) \\
&= O(\alpha^{\frac{(4m-1)p}{4m} - \frac{(2m-1)p+2}{2m}}) = O(\alpha^{\frac{p-4}{4m}}) \to 0, \quad \text{hence} \leq K \qquad \text{for } p = 2,3
\end{aligned}$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

The calculation is exactly the same on $B_r^{2,+}$.

Similarly, on $B_r^{1,-}$, using the above, we have $\varrho = \sqrt{\xi_1^4 + \xi_2^2}$ and $|\nabla(\sigma(\xi)^2)| = 2\sqrt{4\xi_1^6 + \xi_2^2}$. Here we must use the substitution $\xi_1 = \sqrt{\varrho\cos\theta}$, $\xi_2 = \varrho\sin\theta$, then $\frac{\partial(\xi_1,\xi_2)}{\partial(r,\theta)} = \frac{\sqrt{\varrho}}{2\sqrt{\cos\theta}}$. Further, we may assume that $r \leq 1$, then $\varrho^3 \leq \varrho^2$ and hence $|\nabla(\sigma(\xi)^2)| = 2\sqrt{4\varrho^3\cos^3\theta + \varrho^2\sin\theta^2}$ $\leq 2\sqrt{5}\varrho$. Thus $\varrho^{2m-2}\, e^{-\alpha\varrho^{2m}}\, |\nabla(\sigma(\xi)^2)| \leq 2\sqrt{5}\varrho^{2m-1}\, e^{-\alpha\varrho^{2m}}$, hence

$$\begin{aligned}
\|\nabla q_\alpha\|^p_{L^p(B_r^{1,-})} &\leq (2\sqrt{5}m)^p\ \alpha^{\frac{(4m-1)p}{4m}} \int_0^{\frac{\pi}{2}} \int_0^r \varrho^{(2m-1)p}\, e^{-p\alpha\varrho^{2m}} \frac{\sqrt{\varrho}}{2\sqrt{\cos\theta}}\ d\varrho\ d\theta \\
&= (2\sqrt{5}m)^p\ \alpha^{\frac{(4m-1)p}{4m}} \int_0^{\frac{\pi}{2}} \frac{1}{2\sqrt{\cos\theta}}\ d\theta \int_0^r \varrho^{(2m-1)p+\frac{1}{2}}\, e^{-p\alpha\varrho^{2m}}\ d\varrho \\
&= c_3(r,m,p)\ \alpha^{\frac{(4m-1)p}{4m}}\, J_{p\alpha}^{[(2m-1)p+\frac{1}{2},\, m]}(r) \\
&= O(\alpha^{\frac{(4m-1)p}{4m} - \frac{(2m-1)p+\frac{3}{2}}{2m}}) = O(\alpha^{\frac{p-3}{4m}}) \quad \leq K \qquad \text{for } p = 2,3
\end{aligned}$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

Finally, on D_r^- we have $\varrho = \xi_1^2$ and $|\nabla(\sigma(\xi)^2)| = 4\xi_1^3$ on D_r, hence $\varrho^{2m-2}\, e^{-\alpha\varrho^{2m}}\, |\nabla(\sigma(\xi)^2)| = 4\xi_1^{4m-1}\, e^{-\alpha\xi_1^{4m}}$. Arguing as on D_r, we obtain

$$\|\nabla q_\alpha\|^p_{L^p(D_r^-)} = 2r(4m)^p\,\alpha^{\frac{(4m-1)p}{4m}} \int_0^r \xi_1^{(4m-1)p}\ e^{-p\alpha\xi_1^{4m}}\ d\xi_1.$$

The substitution $t = \xi_1^2$ yields

$$\begin{aligned}\|\nabla q_\alpha\|^p_{L^p(D_r^-)} &= 2r(4m)^p\,\alpha^{\frac{(4m-1)p}{4m}} \int_0^{r^2} t^{\frac{(4m-1)p-1}{2}}\; e^{-p\alpha t_1^{2m}}\; dt \\ &= c_4(r,m,p)\;\alpha^{\frac{(4m-1)p}{4m}}\, J_{p\alpha}^{[\frac{(4m-1)p-1}{2},\,m]}(r^2) \\ &= O(\alpha^{\frac{(4m-1)p}{4m}-\frac{(4m-1)p+1}{4m}}) = O(\alpha^{-\frac{1}{4m}}) \to 0, \quad \text{hence} \le K \qquad \text{for } p=2,3\end{aligned}$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

Altogether, we have obtained that

$$\|\nabla q_\alpha\|_{L^3(D_r)} \to \infty, \quad \text{whereas} \quad \|\nabla q_\alpha\|_{L^3(E_r\setminus D_r)} \le K \quad \text{and} \quad \|\nabla q_\alpha\|_{L^2(E_r)} \le K \qquad (2.5.23)$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

Now we study $|\nabla q_\alpha|$ on $\Omega \setminus E_r$. We have seen above that $\sigma(\xi)^{2m}$ can be defined on the whole Ω by the same formula as on E_r, and that $|\nabla(\sigma(\xi)^2)|$ is bounded above by constant times a power of ϱ. Hence (2.5.21) implies

$$|\nabla q_\alpha(x)| \le const. \cdot \alpha^{\frac{4m-1}{4m}}\; \varrho^k\, e^{-\alpha\varrho^{2m}} \qquad (x \in \Omega)$$

for some integer k independent of α. It is elementary to see that the maximizer $\varrho_{\alpha,max}$ of the real function $\varrho \to \varrho^k\, e^{-\alpha\varrho^{2m}}$, i.e. where $\varrho^k_{\alpha,max}\; e^{-\alpha\varrho^{2m}_{\alpha,max}} = \max_{\varrho\ge 0} \varrho^k\, e^{-\alpha\varrho^{2m}}$ holds, satisfies $\varrho_{\alpha,max} \to 0$ as $\alpha \to +\infty$. Thus, for large enough α, the function $\varrho \to \varrho^k\, e^{-\alpha\varrho^{2m}}$ decreases on $[r, +\infty)$. Since $\Omega \setminus E_r = \{x \in \mathbf{R}^2 : \ \varrho = d(x, H_r) > r\}$, we obtain

$$\sup_{\Omega\setminus E_r} |\nabla q_\alpha(x)| \le const. \cdot \alpha^{\frac{4m-1}{4m}}\; r^k\, e^{-\alpha r^{2m}} \to 0, \quad \text{hence} \le K$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α. Since Ω is a bounded domain, this implies that

$$\|\nabla q_\alpha\|_{L^3(\Omega\setminus E_r)} \le K \quad \text{and} \quad \|\nabla q_\alpha\|_{L^2(\Omega\setminus E_r)} \le K$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α. Together with (2.5.23), we obtain that

$$\|\nabla q_\alpha\|_{L^3(D_r)} \to \infty, \quad \text{whereas} \quad \|\nabla q_\alpha\|_{L^3(\Omega\setminus D_r)} \le K \quad \text{and} \quad \|\nabla q_\alpha\|_{L^2(\Omega)} \le K \qquad (2.5.24)$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

(1b) *The case of 3 and more dimensions.* Let us first consider a 3D domain. Then $H_r = \{x_0 + c_2e_2 + c_3e_3 : \ |c_2| \le r\}$ is a square and $D_r = \{x = x_0 + \xi_1 v + \xi_2 e_2 + \xi_3 e_3 : \ 0 \le \varrho := \xi_1 \le r, \ |\xi_2|, |\xi_3| \le r\}$ is a brick. Here $D_r^- = \{x = x_0 - \xi_1 v + \xi_2 e_2 + \xi_3 e_3 : \ 0 \le \xi_1, \ \varrho := \xi_1^2 \le r, \ \ |\xi_2|, |\xi_3| \le r\}$, and the remainder of the set E_r can be divided into two types of subdomains, containing points x where the infimum in $d(x, H_r)$ is attained on a vertex or on an edge of the square H_r, respectively. These subdomains are subsets, namely halves or quarters, of (partly distorted) cylinders or balls, respectively, being two kinds of analogues to the parts of (distorted) discs in 2D, where the distortion comes from the term a_1^4 in the definition of $d(x, y)$ in (2.5.17).

Then one can repeat the 2D calculations, such that the case of D_r and D_r^- goes identically to the 2D case, whereas otherwise the polar type transformations in the integrals are replaced by cylindrical or 3D polar (spherical) type transformations, respectively. In the cylindrical case we obtain constant times the same magnitude of the integrals as in the 2D case, whereas in the spherical case one has ϱ to a greater power. The main point here, as seen in the integrals (2.5.19), is that greater powers of ϱ lead to smaller magnitude of the integral w.r.t. α as $\alpha \to +\infty$. That is, the integrals outside D_r remain bounded as $\alpha \to +\infty$, and behave as before on D_r and D_r^-, hence the 2D argument can be repeated.

This argument also shows that the n-dimensional case can be treated analogously as above. Then H_r is an $(n-1)$-dimensional hypercube, and (after omitting D_r and D_r^-) the remainder of the set E_r can be divided into $n-1$ types of subdomains, containing points x where the infimum in $d(x, H_r)$ is attained on a vertex/an edge/etc., i.e. a k-dimensional hypercube of the boundary of H_r, respectively ($k = 0, 1, \dots, n-1$). The corresponding integrals contain ϱ at least to the same power as in the 2D case, hence they remain bounded as $\alpha \to +\infty$, and behave as before on D_r and D_r^-. Altogether, we thus obtain the behaviour (2.5.24) as $\alpha \to +\infty$.

(2) *The function* p_α. Let us pick a function

$$\psi \in C_0^\infty(\Omega) \quad \text{such that} \quad \psi_{|E_r} \equiv 1 \quad \text{and} \quad 0 \le \psi \le 1 \quad \text{on } \Omega. \tag{2.5.25}$$

Denoting $\Sigma := \operatorname{supp} \psi$, we have $\psi_{|\Omega \setminus \Sigma} \equiv 0$. We define $\widehat{q}_\alpha := q_\alpha \psi$ on Ω. Then

$$\nabla \widehat{q}_\alpha = \begin{cases} \nabla q_\alpha & \text{on } E_r \\ (\nabla q_\alpha \psi + q_\alpha \nabla \psi) & \text{on } \Sigma \setminus E_r \\ 0 & \text{on } \Omega \setminus \Sigma. \end{cases}$$

The definition of q_α implies $0 \le q_\alpha \le 1$ on Ω, hence $|\nabla \widehat{q}_\alpha| \le |\nabla q_\alpha| + |\nabla \psi|$ on $\Sigma \setminus E_r$, but this obviously holds on E_r and $\Omega \setminus \Sigma$ as well, i.e. on the whole Ω. Together with (2.5.24), this implies that

$$\|\nabla \widehat{q}_\alpha\|_{L^3(D_r)} \to \infty, \quad \text{whereas} \quad \|\nabla \widehat{q}_\alpha\|_{L^3(\Omega \setminus D_r)} \le K \quad \text{and} \quad \|\nabla \widehat{q}_\alpha\|_{L^2(\Omega)} \le K \tag{2.5.26}$$

as $\alpha \to +\infty$, for some $K > 0$ independent of α.

Finally, let us prescribe an integer $m \ge 1$ and a number $M > 0$. We construct the functions $\widehat{q}_\alpha$ as above for all $\alpha \in \mathbf{N}^+$. Then, as seen before, $q_\alpha \in C^{(2m-1)}(\Omega)$; further, (2.5.26) holds for some proper constant $K > 0$ independent of α. Let

$$p_\alpha := \frac{1}{K} \widehat{q}_\alpha = \frac{1}{K} q_\alpha \psi,$$

then $p_\alpha \in C^{(2m-1)}(\Omega)$, and we must check properties (i)-(v). We have $p_\alpha = (1/K) q_\alpha$ on $D_r \subset E_r$, hence p_α inherits (2.5.20) from q_α. Further, from (2.5.26),

$$\|\nabla p_\alpha\|_{L^3(D_r)} \to \infty, \quad \text{whereas} \quad \|\nabla p_\alpha\|_{L^3(\Omega \setminus D_r)} \le 1 \quad \text{and} \quad \|\nabla p_\alpha\|_{L^2(\Omega)} \le 1$$

as $\alpha \to +\infty$. Clearly, p_α has a compact support in Ω, a subset of that of ψ. Thus properties (i) and (iii)–(v) hold for the p_α for all $\alpha \in \mathbf{N}^+$. We can now fix α such that $\|\nabla p_\alpha\|_{L^3(D_r)} \ge M$ to obtain that property (iv) holds too. ■

We note that property (iv) and assumption $|v| = 1$ imply $\varphi(x) = |\nabla p_\alpha|$, hence

$$\nabla p_\alpha = |\nabla p_\alpha|\, v \quad \text{on } E_r. \tag{2.5.27}$$

Lemma 2.5.4 *Assume that $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$ is not linear. Then there exist $u_* \in C_0^\infty(\Omega)$ and a number $t_0 > 0$ with the following property: for any constant $K_1 > 0$ and any integer $m \geq 1$, there exists $p_\alpha \in C^{(2m-1)}(\Omega)$ such that if $0 \leq t \leq t_0$ and $|t+s| \leq t_0$, then*

$$\Big\langle \Big(G'(u_* + (s+t)p_\alpha) - G'(u_* + sp_\alpha)\Big) p_\alpha, p_\alpha \Big\rangle_{H_D^1} \;\geq\; tK_1 \|\nabla p_\alpha\|^3_{L^2(\Omega)} \tag{2.5.28}$$

(where the operator G is from (2.5.12)).

Proof. We will use the brief notations $\partial_\eta f := \frac{\partial f}{\partial \eta}$ and $\partial^2_\eta f := \frac{\partial^2 f}{\partial \eta^2}$. Let $(x_0, \eta_0) \in \Omega \times \mathbf{R}^n$ be the point given by Lemma 2.5.1. Since $\partial^2_\eta f$ is continuous, there exists $r_0 > 0$ and $m > 0$ such that such that $\partial^2_\eta f(x, \eta_0)(v,v,v) \geq m$ on the ball $B_{r_0}(x_0)$. Letting $D \subset B_{r_0}(x_0)$ be the set obtained in Lemma 2.5.3, we have

$$\partial^2_\eta f(x, \eta_0)(v,v,v) \geq m \qquad (x \in D).$$

Now let us define

$$u_*(x) := \langle x, \eta_0 \rangle\, \psi(x). \tag{2.5.29}$$

Then $u_*(x) = \langle x, \eta_0\rangle$ in D, hence

$$\nabla u_* \equiv \eta_0 \quad \text{in } D. \tag{2.5.30}$$

From the above, we then have

$$\partial^2_\eta f(x, \nabla u_*)(v,v,v) \geq m \qquad (x \in D). \tag{2.5.31}$$

Now let us fix $M > 0$ and an integer $m \geq 1$. Let the corresponding p_α be chosen from Lemma 2.5.3. Then for any $t > 0$ and $s \in \mathbf{R}$ and some function $\theta = \theta(s,t)$,

$$\begin{aligned}
&\Big\langle \Big(G'(u_* + (s+t)p_\alpha) - G'(u_* + sp_\alpha)\Big) p_\alpha, p_\alpha \Big\rangle_{H_D^1} \\
&= \int_\Omega \Big(\partial_\eta f(x, \nabla u_* + (s+t)\nabla p_\alpha) - \partial_\eta f(x, \nabla u_* + s\nabla p_\alpha)\Big)\, \nabla p_\alpha \cdot \nabla p_\alpha \\
&= t \int_\Omega \partial^2_\eta f(x, \nabla u_* + (s+\theta t)\nabla p_\alpha)\ (\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha) \;=:\; t\,\Big(I_1(s,t) + I_2(s,t)\Big),
\end{aligned} \tag{2.5.32}$$

where $I_1(s,t)$ and $I_2(s,t)$ denote the integrals on $\Omega \setminus D$ and D, respectively.

From (2.5.15) and property (iii) of Lemma 2.5.3, respectively, we have

$$|I_1(s,t)| \leq l_f \int_{\Omega\setminus D} |\nabla p_\alpha|^3 = l_f\, \|\nabla p_\alpha\|^3_{L^3(\Omega\setminus D)} \leq l_f \tag{2.5.33}$$

independently of s,t and the given K_1.

Let us consider $I_2(s,t)$. The integrand has the following two properties: since $f \in C^2$,

$$\partial_\eta^2 f(x, \nabla u_* + (s+\theta t)\nabla p_\alpha)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha) \;\to\; \partial_\eta^2 f(x, \nabla u_*)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha)$$

almost everywhere on D as $s,t \to 0$; further,

$$|\partial_\eta^2 f(x, \nabla u_* + (s+\theta t)\nabla p_\alpha)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha)| \;\le\; l_f |\nabla p_\alpha|^3 \;\in C_0^\infty(\Omega) \subset L^1(D) \tag{2.5.34}$$

i.e. it has an integrable majorant. Then, using the Lebesgue dominated convergence theorem,

$$\int_D \partial_\eta^2 f(x, \nabla u_* + (s+\theta t)\nabla p_\alpha)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha) \;\to\; \int_D \partial_\eta^2 f(x, \nabla u_*)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha) \tag{2.5.35}$$

as $s,t \to 0$. Here, by (2.5.27), (2.5.30) and property (i) of Lemma 2.5.3, respectively, the limit satisfies

$$\int_D \partial_\eta^2 f(x, \nabla u_*)\,(\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha) = \int_D |\nabla p_\alpha|^3 \; \partial_\eta^2 f(x, \nabla u_*)(v,v,v) \ge m \int_D |\nabla p_\alpha|^3$$
$$= m \|\nabla p_\alpha\|_{L^3(D)}^3 \ge m M^3\,,$$

hence $I_2(s,t) \ge \frac{1}{2}\, m M^3$ for sufficiently small s,t. If M is sufficiently large then $\frac{1}{2}\, mM^3 - l_f \ge \frac{1}{4}\, mM^3$, where l_f in (2.5.33) is independent of M. It readily follows that

$$I_1(s,t) + I_2(s,t) \ge I_2(s,t) - |I_1(s,t)| \ge \frac{1}{4}\, mM^3 \ge K_1$$

if M is sufficiently large. By property (iv) of Lemma 2.5.3, we find

$$I_1(s,t) + I_2(s,t) \ge K_1 \ge K_1 \|\nabla p_\alpha\|_{L^2(\Omega)}^3. \tag{2.5.36}$$

Using (2.5.32), we then obtain (2.5.28). ■

Corollary 2.5.1 *Assume that $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$ is not linear. Then the operator F' (with F from (2.5.11)) is not locally Lipschitz continuous in $H_D^1(\Omega)$.*

Proof. Denote $\|.\| := \|.\|_{H_D^1}$. We must find bounded sequencees (u_n) and (v_n) in $H_D^1(\Omega)$ such that

$$\frac{\|F'(u_n) - F'(v_n)\|}{\|u_n - v_n\|} \to +\infty.$$

Let $u_* \in H_D^1(\Omega)$ be as in Lemma 2.5.4. Further, let $n \in \mathbf{N}^+$ be given. By Lemma 2.5.4, there exists $p_\alpha \in H_D^1(\Omega)$ such that for all proper s,t, inequality (2.5.28) holds with n instead of K_1. Let us choose $0 \le t_n \le t_0$ such that $t_n \|p_\alpha\| \le 1$. Now let $u_n := u_* + t_n p_\alpha$ and $v_n := u_*$, these are bounded sequences as $n \to \infty$. On the other hand, letting $s := 0$ and $t := t_n$ in (2.5.28), and using that $\|p_\alpha\| = \|\nabla p_\alpha\|_{L^2(\Omega)}$, we obtain

$$\frac{\|G'(u_n) - G'(v_n)\|}{\|u_n - v_n\|} = \frac{\|G'(u_* + t_n p_\alpha) - G'(u_*)\|}{t\|p_\alpha\|} \ge \frac{\big\langle \big(G'(u_* + t_n p_\alpha) - G'(u_*)\big)p_\alpha, p_\alpha \big\rangle_{H_D^1}}{t\|p_\alpha\|^3} \ge n.$$

Finally, since $F' = G' + R'$,

$$\frac{\|F'(u_n) - F'(v_n)\|}{\|u_n - v_n\|} \geq \frac{\|G'(u_n) - G'(v_n)\|}{\|u_n - v_n\|} - \frac{\|R'(u_n) - R'(v_n)\|}{\|u_n - v_n\|} \geq n - L(R') \to \infty,$$

where $L(R')$ is the Lipschitz constant of R' on the ball with radius $\|u_*\| + 1$ which contains the sequences (u_n) and (v_n). (The local Lipschitz continuity of R' follows in the same way as Proposition 2.5.1.) ■

Lemma 2.5.5 *Assume that $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$ is not linear. Then there exists $u_* \in C_0^\infty(\Omega)$ with the following property: for any numbers $K_0 > 0$ and $h_0 > 0$ there exists an admissible FEM subspace V_h with mesh parameter $h < h_0$, a function $p_\alpha^h \in V_h$ and a number $t_0 > 0$ such that if $0 \leq t \leq t_0$ and $|t + s| \leq t_0$, then*

$$\Big\langle \Big(F_h'(u_h + (s+t)p_\alpha^h) - F_h'(u_h + sp_\alpha^h) \Big) p_\alpha^h, p_\alpha^h \Big\rangle_{H_D^1} \ \geq\ tK_0 \|\nabla p_\alpha^h\|^3_{L^2(\Omega)} \tag{2.5.37}$$

where $F_h(u_h)$ is the projection of $F(u_*)$ into V_h.

Proof. For any subspace V_h, if we denote by w_h the projection of $F(u_*)$ into V_h, then we define u_h by $u_h := F_h^{-1}(w_h)$ (since F_h is one-to-one on V_h).

First we prove the statement for the operator G, defined in (2.5.12), instead of F. Let $K_1 > K_0$ be given and u_*, p_α be the functions defined in Lemma 2.5.4. We introduce the function

$$r(s,t) := \frac{\Big\langle \Big(G'(u_* + (s+t)p_\alpha) - G'(u_* + sp_\alpha) \Big) p_\alpha, p_\alpha \Big\rangle_{H_D^1}}{t\, \|\nabla p_\alpha\|^3_{L^2(\Omega)}} \qquad (t \neq 0).$$

Then r has a continuous extension defined by

$$r(\sigma, 0) := \lim_{\substack{s \to \sigma \\ t \to 0}} r(s,t) = \frac{1}{\|\nabla p_\alpha\|^3_{L^2(\Omega)}} \int_\Omega \partial_\eta^2 f(x, \nabla(u_* + \sigma p_\alpha))\ (\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha),$$

where the existence and the value of the limit can been derived just as in Lemma 2.5.4, see (2.5.32)–(2.5.35), if we replace D by Ω and u_* by $u_* + \sigma p_\alpha$ therein. Since by (2.5.28) we have $r(s,t) \geq K_1$ for small enough $s, t \neq 0$, we obtain

$$K_1 \leq r(0,0) = \frac{1}{\|\nabla p_\alpha\|^3_{L^2(\Omega)}} \int_\Omega \partial_\eta^2 f(x, \nabla u_*)\ (\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha).$$

If V_h is an admissible FEM subspace and $p_\alpha^h \in V_h$ is the FE approximation of p_α, then we define the analogue of r by

$$r_h(s,t) := \frac{\Big\langle \Big(G_h'(u_h + (s+t)p_\alpha^h) - G_h'(u_h + sp_\alpha^h) \Big) p_\alpha^h, p_\alpha^h \Big\rangle_{H_D^1}}{t\, \|\nabla p_\alpha^h\|^3_{L^2(\Omega)}} \qquad (t \neq 0).$$

Then, similarly as above, we can define $r_h(\sigma, 0) := \lim_{\substack{s\to\sigma \\ t\to 0}} r_h(s,t)$ and we obtain

$$r_h(0,0) = \frac{1}{\|\nabla p_\alpha^h\|^3_{L^2(\Omega)}} \int_\Omega \partial_\eta^2 f(x, \nabla u_h)\ (\nabla p_\alpha^h, \nabla p_\alpha^h, \nabla p_\alpha^h).$$

Now our goal is to find a sequence of admissible subspaces V_{h_i} such that

$$\lim_{h_i\to 0} r_{h_i}(0,0) = r(0,0). \tag{2.5.38}$$

Here Definition 2.5.1 and the coercivity of F imply [34] that

$$\|p_\alpha - p_\alpha^h\| = \|\nabla(p_\alpha - p_\alpha^h)\|_{L^2(\Omega)} \to 0 \ \text{ and } \|u_* - u_h\| = \|\nabla(u_* - u_h)\|_{L^2(\Omega)} \to 0 \quad \text{as } h \to 0. \tag{2.5.39}$$

Then, in particular, the denominator $\|\nabla p_\alpha^h\|^3_{L^2(\Omega)}$ of $r_h(0,0)$ tends to that of $r(0,0)$. We must prove the same for the numerators for a suitable subsequence. Here, for all h,

$$\begin{aligned}
&\Big| \int_\Omega \partial_\eta^2 f(x,\nabla u_h)\ (\nabla p_\alpha^h, \nabla p_\alpha^h, \nabla p_\alpha^h) - \int_\Omega \partial_\eta^2 f(x,\nabla u_*)\ (\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha)\Big| \\
&\leq \Big| \int_\Omega \partial_\eta^2 f(x,\nabla u_h)\ (\nabla(p_\alpha^h - p_\alpha), \nabla p_\alpha^h, \nabla p_\alpha^h)\Big| + \Big| \int_\Omega \partial_\eta^2 f(x,\nabla u_h)\ (\nabla p_\alpha, \nabla(p_\alpha^h - p_\alpha), \nabla p_\alpha^h)\Big| \\
&+ \Big| \int_\Omega \partial_\eta^2 f(x,\nabla u_h)\ (\nabla p_\alpha, \nabla p_\alpha, \nabla(p_\alpha^h - p_\alpha))\Big| + \Big| \int_\Omega \Big(\partial_\eta^2 f(x,\nabla u_h) - \partial_\eta^2 f(x,\nabla u_*)\Big)\ (\nabla p_\alpha, \nabla p_\alpha, \nabla p_\alpha)\Big| \\
&=: \ I_1(h) + I_2(h) + I_3(h) + I_4(h).
\end{aligned}$$

The integrals $I_k(h)$ for $k = 1,2,3$ behave quite similarly. First, (2.5.15) and Hölder's inequality yield

$$I_1(h) \leq \ l_f \int_\Omega |\nabla(p_\alpha^h - p_\alpha)|\, |\nabla p_\alpha^h|^2 \leq \ l_f \|\nabla(p_\alpha^h - p_\alpha)\|_{L^2(\Omega)} \|\nabla p_\alpha^h\|^2_{L^4(\Omega)},$$

and similarly, the term $\|\nabla p_\alpha^h\|^2_{L^4(\Omega)}$ is replaced in the same estimate for $I_2(h)$ and $I_3(h)$ by $\|\nabla p_\alpha\|_{L^4(\Omega)} \|\nabla p_\alpha^h\|_{L^4(\Omega)}$ and $\|\nabla p_\alpha\|^2_{L^4(\Omega)}$, resp . Hence $I_k(h) \to 0$ ($k = 1,2,3$) as $h \to 0$, using (2.5.39) and that the other factors are bounded [34]. Further, since $\|\nabla u_h - \nabla u_*\|_{L^2(\Omega)} \to 0$, it follows [136] that $\nabla u_{h_i} \to \nabla u_*$ almost everywhere for a subsequence u_{h_i}. Then, using that $f \in C^2$, the integrand in $I_4(h_i)$ converges to 0 almost everywhere on Ω, further, similarly to (2.5.34), the function $2l_f |\nabla p_\alpha|^3$ is an integrable majorant, hence Lebesgue's dominated convergence theorem yields that $I_4(h_i) \to 0$ as $h_i \to 0$. Altogether, we have proved that $\lim_{h_i\to 0} r_{h_i}(0,0) = r(0,0)$, and hence $\lim_{h_i\to 0} r_{h_i}(0,0) \geq K_1$.

Consequently, since $K_1 > K_0$ was chosen, there exists $h < h_0$ such that $r_h(0,0) > K_0$ for the corresponding function p_α^h. Then the relation $r_h(0,0) := \lim_{s,t\to 0} r_h(s,t)$ implies that

$$r_h(s,t) \geq K_0$$

for small enough s,t, in particular, for $0 \leq t \leq t_0$ and $|t+s| \leq t_0$ with a suitably chosen t_0. Thus altogether we have proved (2.5.37) for G_h instead of F_h.

Finally, note that if F_h in (2.5.37) is replaced by R_h, then the proof of Proposition 2.5.1 yields a uniform upper bound for the obtained expression independently of K_0. Since $F_h = G_h + R_h$, we obtain that F_h inherits the required property from G_h. ■

Similarly as Corollary 2.5.1 was obtained from Lemma 2.5.4, we can now derive

Corollary 2.5.2 *Assume that $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$ is not linear. Then the operator F_h' is not uniformly locally Lipschitz continuous.*

Now we are in the position to verify the required mesh dependence.

Proposition 2.5.3 *Assume that $f \in C^2(\Omega \times \mathbf{R}^n,\ \mathbf{R}^n)$ is not linear. Then there exist right-hand sides $g \in L^2(\Omega)$ and $\gamma \in L^2(\Gamma_N)$ in problem (2.5.1) such that for arbitrary $K > 0$, $h_0 > 0$ and $\delta > 0$, there exists a FEM subspace $V_h \subset H^1_D(\Omega)$ with mesh parameter $h < h_0$, satisfying Definition 2.5.1, and there exists $u_0 \in V_h$ with $\|u_0 - u_h\|_{H^1_D} < \delta$, such that for some $n \in \mathbf{N}$ in the iteration (2.5.8)–(2.5.9),*

$$\frac{\|F_h(u_{n+1})\|_{H^1_D}}{\|F_h(u_n)\|^2_{H^1_D}} \geq K. \tag{2.5.40}$$

Proof. We define

$$g := -\operatorname{div} f(x, \nabla u_*) + q(x, u_*) \quad \text{and} \quad \gamma := s(x, u_*), \tag{2.5.41}$$

where the function $u_* \in C_0^\infty(\Omega)$ is taken from (2.5.29). Then u_* is the exact solution of problem (2.5.1).

Let $K > 0$, $h_0 > 0$ and $\delta > 0$ be arbitrary. Our goal is to find a FEM subspace $V_h \subset H^1_D(\Omega)$ with mesh parameter $h < h_0$, satisfying Definition 2.5.1, and an initial guess $u_0 \in V_h$ with $\|u_0 - u_h\|_{H^1_D} < \delta$, such that for some $n \in \mathbf{N}$ (2.5.40) holds.

First note that Assumption 2.5.1 (iii) implies the uniform regularity

$$\|F_h'(u)z\|_{H^1_D} \geq \lambda \|z\|_{H^1_D} \tag{2.5.42}$$

for all $u, z \in V_h$. The opposite direction also holds if Λ in (2.5.16) is modified by a term coming from the coefficient $q(x,\xi)$ (see e.g. [55, Thm. 6.2]): then Λ may depend on $\|u\|_{H^1_D}$, but the latter is bounded throughout the Newton iteration (the iterates run in a neighbourhood of u_h that remains bounded as $h \to 0$). Therefore there exists $\tilde{\Lambda} > 0$ such that

$$\|F_h'(u)z\|_{H^1_D} \leq \tilde{\Lambda} \|z\|_{H^1_D}. \tag{2.5.43}$$

Similarly, the operator F_h', which is locally Lipschitz continuous in V_h, has a uniform Lipschitz constant for the Newton iterates that satisfies

$$L_h = \sup_{\substack{u \in V_h \\ \|u\| \leq R_0}} \|F_h''(u)\| \tag{2.5.44}$$

where R_0 is a bound for $\|u\|_{H^1_D}$ on the above-mentioned neighbourhood.

Let
$$K_0 := 32K\tilde{\Lambda}^2.$$

For this K_0 and the prescribed $h_0 > 0$, there exists a subspace V_h with $h < h_0$ and a function $p^h_\alpha \in V_h$ from Lemma 2.5.5 such that (2.5.37) holds. Let $u_h \in V_h$ be the FE solution of problem (2.5.1) with the above data, let $\varepsilon > 0$ be some constant and

$$\overline{u}^h := u_h - \varepsilon p^h_\alpha. \tag{2.5.45}$$

Let us consider the linearized equation at $\overline{u}^h$:

$$F'_h(\overline{u}^h)\overline{p}^h = -F_h(\overline{u}^h), \tag{2.5.46}$$

and let
$$\overline{r}^h := \overline{p}^h - \varepsilon p^h_\alpha\,. \tag{2.5.47}$$

We will prove the existence of the desired u_0 by choosing ε small enough. For this, we now proceed in five steps.

Step 1. We prove that
$$\|\overline{r}^h\|_{H^1_D} \le c_1\,\varepsilon^2. \tag{2.5.48}$$
for some $c_1 > 0$ independent of ε. Namely, by Taylor expansion and (2.5.45),

$$\begin{aligned} 0 = F_h(u_h) &= F_h(\overline{u}^h) + F'_h(\overline{u}^h)(u_h - \overline{u}^h) + \tfrac{1}{2}\,\langle F''_h(\overline{u}^h + \theta(u_h - \overline{u}^h))(u_h - \overline{u}^h), u_h - \overline{u}^h\rangle \\ &= F_h(\overline{u}^h) + F'_h(\overline{u}^h)(\varepsilon p^h_\alpha) + \tfrac{1}{2}\,\langle F''_h(\overline{u}^h + \theta(\varepsilon p^h_\alpha))(\varepsilon p^h_\alpha), \varepsilon p^h_\alpha\rangle, \end{aligned}$$

hence from (2.5.46) we have $F'_h(\overline{u}^h)\overline{p}^h = F'_h(\overline{u}^h)(\varepsilon p^h_\alpha) + \frac{1}{2}\,\langle F''_h(\overline{u}^h + \theta(\varepsilon p^h_\alpha))(\varepsilon p^h_\alpha), \varepsilon p^h_\alpha\rangle$, and (2.5.47) yields $F'_h(\overline{u}^h)\overline{r}^h = \frac{1}{2}\,\langle F''_h(\overline{u}^h + \theta(\varepsilon p^h_\alpha))(\varepsilon p^h_\alpha), \varepsilon p^h_\alpha\rangle$. Using (2.5.44) and (2.5.42), we obtain
$$\lambda\|\overline{r}^h\|_{H^1_D} \le (L_h/2)\ \|p^h_\alpha\|^2_{H^1_D}\varepsilon^2\,.$$
Here from property (ii) of Lemma 2.5.3, we have

$$\|p^h_\alpha\|_{H^1_D} \le P_0 \tag{2.5.49}$$

for $h \le h_0$, with some constant $P_0 > 0$ independent of h, ε and K_0, hence inequality (2.5.48) follows for $c_1 := L_h P_0^2/2\lambda$.

Step 2. Let $\tau > 0$ be fixed. We will prove that

$$\begin{aligned} &\Big\langle \Big(F'_h(u_h - \varepsilon p^h_\alpha + \tau\varepsilon p^h_\alpha + \tau\overline{r}^h) - F'_h(u_h - \varepsilon p^h_\alpha)\Big)\overline{p}^h, \overline{p}^h\Big\rangle_{H^1_D} \\ &\qquad = \ \Big\langle \Big(F'_h(u_h - \varepsilon p^h_\alpha + \tau\varepsilon p^h_\alpha) - F'_h(u_h - \varepsilon p^h_\alpha)\Big)\varepsilon p^h_\alpha, \varepsilon p^h_\alpha\Big\rangle_{H^1_D} \ + \ \tau R(\varepsilon), \end{aligned} \tag{2.5.50}$$
$$\text{where} \quad R(\varepsilon) = O(\varepsilon^4) \quad \text{as} \ \ \varepsilon \to 0$$

independently of τ, i.e. $|R(\varepsilon)| \le c_2\varepsilon^4$ where c_2 is independent of τ and ε. Namely, expanding the bilinear forms, we obtain

$$\tau R(\varepsilon) = B_1(\varepsilon) + B_2(\varepsilon) - B_3(\varepsilon), \qquad \text{where}$$

$$B_1(\varepsilon) := \Big\langle \Big(F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h + \tau \overline{r}^h) - F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h)\Big)\overline{p}^h, \overline{p}^h \Big\rangle_{H_D^1},$$

$$B_2(\varepsilon) := \Big\langle F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h)\overline{p}^h, \overline{p}^h \Big\rangle_{H_D^1} - \Big\langle F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h)\varepsilon p_\alpha^h, \varepsilon p_\alpha^h \Big\rangle_{H_D^1},$$

$$B_3(\varepsilon) := \Big\langle F_h'(u_h - \varepsilon p_\alpha^h)\overline{p}^h, \overline{p}^h \Big\rangle_{H_D^1} - \Big\langle F_h'(u_h - \varepsilon p_\alpha^h)\varepsilon p_\alpha^h, \varepsilon p_\alpha^h \Big\rangle_{H_D^1}.$$

First, by (2.5.44),

$$|B_1(\varepsilon)| \le L_h \tau \|\overline{r}^h\|_{H_D^1} \|\overline{p}^h\|_{H_D^1}^2.$$

Then (2.5.47) and (2.5.48) yield

$$\|\overline{p}^h\|_{H_D^1} \le \|\overline{r}^h\|_{H_D^1} + \varepsilon \|p_\alpha^h\|_{H_D^1} \le \varepsilon(c_1\varepsilon + \|p_\alpha^h\|_{H_D^1}) \le O(\varepsilon), \tag{2.5.51}$$

hence

$$|B_1(\varepsilon)| \le \tau O(\varepsilon^4).$$

Further, using $\overline{p}^h = \overline{r}^h + \varepsilon p_\alpha^h$ and the symmetry of F_h',

$$B_2(\varepsilon) = 2\Big\langle F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h)\varepsilon p_\alpha^h, \overline{r}^h \Big\rangle_{H_D^1} + \Big\langle F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h)\overline{r}^h, \overline{r}^h \Big\rangle_{H_D^1}$$

and

$$B_3(\varepsilon) = 2\Big\langle F_h'(u_h - \varepsilon p_\alpha^h)\varepsilon p_\alpha^h, \overline{r}^h \Big\rangle_{H_D^1} + \Big\langle F_h'(u_h - \varepsilon p_\alpha^h)\overline{r}^h, \overline{r}^h \Big\rangle_{H_D^1},$$

hence

$$\begin{aligned} B_2(\varepsilon) - B_3(\varepsilon) \ &= \ 2\Big\langle \Big(F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h) - F_h'(u_h - \varepsilon p_\alpha^h)\Big)\varepsilon p_\alpha^h, \overline{r}^h \Big\rangle_{H_D^1} \\ &\qquad + \Big\langle \Big(F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h) - F_h'(u_h - \varepsilon p_\alpha^h)\Big)\overline{r}^h, \overline{r}^h \Big\rangle_{H_D^1}. \end{aligned}$$

Then by (2.5.44) and (2.5.48),

$$|B_2(\varepsilon) - B_3(\varepsilon)| \le \ L_h \tau\varepsilon \, \|p_\alpha\|_{H_D^1} \Big(2\varepsilon \|p_\alpha\|_{H_D^1} \|\overline{r}^h\|_{H_D^1} + \|\overline{r}^h\|_{H_D^1}^2 \Big) \le \tau\varepsilon\Big(O(\varepsilon^3) + O(\varepsilon^4)\Big) \le \tau O(\varepsilon^4).$$

Thus (2.5.50) is proved.

Step 3. Using Lemma 2.5.5, there exists $t_0 > 0$ such that if $0 < \varepsilon \le t_0$ and $0 \le \tau \le 1$ then the choices $s := -\varepsilon$ and $t := \tau\varepsilon$ satisfy

$$\Big\langle \Big(F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h) - F_h'(u_h - \varepsilon p_\alpha^h)\Big)\varepsilon p_\alpha^h, \varepsilon p_\alpha^h \Big\rangle_{H_D^1} \ge \varepsilon^3 \tau K_0 \|p_\alpha^h\|_{H_D^1}^3.$$

Using (2.5.50) and that $|R(\varepsilon)| \le c_2 \varepsilon^4$, if $0 < \varepsilon \le \varepsilon_0 := K_0 \|p_\alpha^h\|_{H_D^1}^3 / 2c_2$ and $0 \le \tau \le 1$ then

$$\Big\langle \Big(F_h'(u_h - \varepsilon p_\alpha^h + \tau\varepsilon p_\alpha^h + \tau\overline{r}^h) - F_h'(u_h - \varepsilon p_\alpha^h)\Big)\overline{p}^h, \overline{p}^h \Big\rangle_{H_D^1} \ge \frac{1}{2}\, \varepsilon^3 \tau K_0 \|p_\alpha^h\|_{H_D^1}^3.$$

Here (2.5.51) implies $\|\overline{p}^h\|_{H_D^1} \le 2\varepsilon \|p_\alpha^h\|_{H_D^1}$ for $0 < \varepsilon \le \varepsilon_0 := \|p_\alpha^h\|_{H_D^1}/c_1$. These imply, also using (2.5.45) and (2.5.47), that

$$\Big\langle \Big(F_h'(\overline{u}^h + \tau \overline{p}^h) - F_h'(\overline{u}^h)\Big)\overline{p}^h, \overline{p}^h \Big\rangle_{H_D^1} \ge \ \frac{\tau K_0}{16} \, \|\overline{p}^h\|_{H_D^1}^3. \tag{2.5.52}$$

Step 4. Let $\varepsilon > 0$ be a constant as obtained above in Step 3, i.e. it must satisfy $\varepsilon \le \min\{K_0\|p_\alpha^h\|^3_{H^1_D}/2c_2,\ \|p_\alpha^h\|_{H^1_D}/c_1\}$. We prove that the resulting $\overline{u}^h$ and $\overline{p}^h$ satisfy

$$\frac{\|F_h(\overline{u}^h+\overline{p}^h)\|_{H^1_D}}{\|F_h(\overline{u}^h)\|^2_{H^1_D}} \ge K. \tag{2.5.53}$$

Namely, using (2.5.52),

$$\|F_h(\overline{u}^h+\overline{p}^h)\|_{H^1_D} \ge \frac{\langle F_h(\overline{u}^h+\overline{p}^h),\overline{p}^h\rangle}{\|\overline{p}^h\|} = \frac{1}{\|\overline{p}^h\|}\int_0^1 \Big\langle \Big(F_h'(\overline{u}^h+\tau\overline{p}^h)-F_h'(\overline{u}^h)\Big)\overline{p}^h,\overline{p}^h\Big\rangle d\tau \ge \frac{K_0}{32}\,\|\overline{p}^h\|^2\,.$$

From (2.5.43) and (2.5.46) we obtain

$$\|\overline{p}^h\|_{H^1_D} \ge \frac{1}{\tilde\Lambda}\|F_h'(\overline{u}^h)\overline{p}^h\|_{H^1_D} = \frac{1}{\tilde\Lambda}\|F_h(\overline{u}^h)\|_{H^1_D},$$

hence

$$\|F_h(\overline{u}^h+\overline{p}^h)\|_{H^1_D} \ge \frac{K_0}{32\tilde\Lambda^2}\,\|F_h(\overline{u}^h)\|^2_{H^1_D} = K\|F_h(\overline{u}^h)\|^2_{H^1_D}.$$

Step 5. There exists $u_0 \in V_h$ with $\|u_0-u_h\|_{H^1_D} < \delta$, such that $u_n = \overline{u}^h$ for some $n \in \mathbf{N}$ in the Newton iteration (2.5.8)–(2.5.9). This even holds, for instance, for $u_0 := \overline{u}^h$ (for this one simply has to ensure $\|\overline{u}^h - u_h\|_{H^1_D} < \delta$, which, by (2.5.45), prescribes $\varepsilon\|p_\alpha^h\|_{H^1_D} < \delta$ for ε at the beginning.) Then by (2.5.46), $u_{n+1} = \overline{u}^h + \overline{p}^h$, hence (2.5.53) becomes (2.5.40). ■

Remark 2.5.2 It is important to note that the negative part of the result is not due to improper data, assumptions or setting.

The obtained negative result has essentially followed from the non-uniformity of the Lipschitz constants of F_h' as $h \to 0$, see Lemma 2.5.5. The underlying property is that F' itself is not locally Lipschitz continuous, stated in Corollary 2.5.1. This shows that the obtained mesh dependence is an inherent property for this class of problems, and could not be prevented by changing some parts of the setting.

In particular, the concrete boundary value problem, used in the proof to provide mesh dependence (see (2.5.41)), has a C_0^∞ solution u_*, which is zero on $\partial\Omega$, and here the domain can also have a C^∞ boundary. The right-hand sides g and γ are as smooth as f, q and s are, respectively, possibly even C^∞ as well. Finally, the initial guess can be the function $\overline{u}^h$ from (2.5.45), i.e. the projection of $u_* - \varepsilon p_\alpha$, which can be chosen in C^{2m-1} for arbitrary prescribed $m \in \mathbf{N}^+$. Therefore the negative result is not due to a too little smoothness of the data. Also, the considered FEM subspaces cover a fairly large standard class, including the widespread cases such as linear simplicial or bilinear elements.

One might also pose the problem in other function spaces, such as for $F: H^1 \to H^{-1}$ or $F: H^2 \to L_2$ (or more generally $F: H^{m+2} \to H^m$), or only require uniform Hölder continuity of F_h' instead of its Lipschitz continuity, but none of these can remedy the negative result, see details in [88].

2.6 Applications to efficient computational algorithms

We apply our previously described theory to derive various efficient iterative methods for nonlinear PDEs, which include several real-life models. For most of the examples computer experiments were also run. These test results confirm the theoretical convergence results. For FEM discretizations we have obtained mesh independent convergence, moreover, for semilinear problems, applying outer-inner iterations, the inner convergence is superlinear with a rate independent of both the mesh and of the outer iterate.

2.6.1 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

$$\left.\begin{aligned} \mathrm{rot} H = \rho \\ \mathrm{div}\, B = 0 \end{aligned}\right\} \quad \text{in } \Omega$$
$$B \cdot \nu = 0 \qquad \text{on } \partial\Omega$$

and the relation

$$H = b(x, |B|)\, B\,.$$

The electromagnetic potential u is defined by $\mathrm{curl}\, u = B$, and thus we obtain the boundary value problem

$$\left\{ \begin{aligned} -\mathrm{div}\ (b(x, |\nabla u|)\, \nabla u) &= \rho(x) \\ u_{|\partial\Omega} &= 0\,, \end{aligned} \right. \tag{2.6.1}$$

where the scalar-valued function $b : \Omega \times \mathbf{R}^+ \to \mathbf{R}$ describes magnetic reluctance and $\rho \in L^2(\Omega)$ is the electric current density. The function b is measurable and bounded w.r. to x and C^1 w.r. to the variable r, further, it satisfies

$$0 < \mu_1 \le b(x,r) \le \frac{\partial}{\partial r}(r\, b(x,r)) \le \mu_2 \qquad (x \in \Omega,\ r > 0)$$

with constants $\mu_2 \ge \mu_1 > 0$ independent of (x, r). This model is described e.g. in [106].

Typically, the function b is independent of x in some subdomain and constant on the complement, where these subdomains correspond to ferromagnetic and other media, respectively. That is,

$$b(x,r) = \left\{ \begin{array}{ll} a(r) & \text{if } x \in \Omega_1 \\ \alpha & \text{if } x \in \Omega \setminus \Omega_1\,, \end{array} \right. \tag{2.6.2}$$

where $\Omega_1 \subset \Omega$ is a given subdomain, $\alpha > 0$ is a constant and $a \in C^1(\mathbf{R}^+)$ satisfies

$$0 < \mu_1 \le a(r) \le a(r) + a'(r) r \le \mu_2 \qquad (r \ge 0)$$

with constants $\mu_2 \ge \mu_1 > 0$ independent of r. In the case of the reluctance of stator sheets in the cross-sections of an electrical motor in the case of isotropic media, the following nonlinearity appears, see [106]:

$$a(r) = \frac{1}{\mu_0}\left(\alpha + (1-\alpha)\frac{r^8}{r^8+\beta}\right) \qquad (r \geq 0). \tag{2.6.3}$$

Here μ_0 is the vacuum permeability and $\alpha, \beta > 0$ are characteristic constants. An example of realistic values is $\alpha = 0.0003$ and $\beta = 16000$, which shows that our problem is almost singular.

We consider the above nonlinearity a and solve the problem

$$\left\{ \begin{array}{l} -\mathrm{div}\,(a(|\nabla u|)\nabla u) \;=\; g(x) \\ \;\; u_{|\partial\Omega} = 0\,. \end{array} \right. \tag{2.6.4}$$

That is, we focus on the ill-conditioning of the problem, but for simplicity we neglect the part where $b(x, r)$ is constant (i.e. the operator is the Laplacian).

We have run experiments by applying the variable preconditioning procedure with piecewise constant coefficient preconditioning operators [90]. For simplicity, we have chosen the unit square domain $\Omega = [0, 1] \times [0, 1]$, and V_h was the subspace of piecewise linear elements on a uniform triangulation of Ω. The coefficients are $\alpha = 3 \cdot 10^{-4}$ and $\beta = 1.6 \cdot 10^4$, $\mu_0 = 1$ is the normalized vacuum permeability, further, for the right-hand side we set $g(x) \equiv \rho = 4 \cdot 10^6$ which is a realistic value for the electric current density [106]. Then the iteration has the form

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} z_n \tag{2.6.5}$$

with $z_n \in V_h$ being the solution of problem

$$\int_\Omega w_n(x)\,\nabla z_n \cdot \nabla v = \int_\Omega \big(a(|\nabla u_n|)\,\nabla u_n \cdot \nabla v \;-\rho v\big) \qquad (v \in V_h). \tag{2.6.6}$$

The convergence of the iteration is ensured by Theorem 2.3.3.

The piecewise constant weight function w_n is constructed as given in subsection 2.3.2 and in (2.4.36)–(2.4.37). The corresponding preconditioning matrix $\mathcal{B}_n$ is the modification of the discrete Laplacian via blockwise multiplication by the corresponding constants c_i. Moreover, the matrix $\mathcal{B}_n$ can be decomposed in the product form $\mathcal{B}_n = \mathcal{C}\mathcal{W}_n\mathcal{C}^T$ where the matrices $\mathcal{C}$ and $\mathcal{C}^T$ correspond to the discretization of $-div$ and ∇, respectively, and hence are independent of n; further, $\mathcal{W}_n$ is a diagonal matrix consisting of constants c_i at the entries corresponding to the subdomains Ω_i.

Numerical experiment. We have used a decomposition to 6 subdomains in each step of the iteration. We have chosen c_i to be the arithmetic mean of λ_i and Λ_i for all i.

The error during the iteration was measured by the weighted residual errors corresponding to (2.2.16) with the inner product with weight w_n. This error is obtained from the iteration without any extra work as the weighted norm of the actual coefficient vector w.r. to the Gram matrix. (It is a computable approximation of the $*$-norm (2.3.21) that appears in the convergence estimates of Theorem 2.3.2.)

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

Table 1: the number of iterations to achieve error 10^{-4} and 10^{-8} under different number of node points using 6 subdomains.

We have repeated the experiment with 12 subdomains, and the results were the same (except that for 2^6 node points the number of iterations for $\varepsilon = 10^{-8}$ was only 15). This means that the smaller number of subdomains already suffices to achieve the available convergence speed.

The distribution of the errors behaved much similarly for the different runs. We give one of them below for illustration, where 16 iterations were done to achieve relative accuracy 10^{-8}.

1.0	0.03014214	0.00027565	0.00000033
0.32290943	0.01194232	0.00005601	0.00000006
0.14549087	0.00414995	0.00001047	0.00000001
0.06899055	0.00120266	0.00000182	0.00000000

Table 2: the sequence of errors up to 8 digits, using 2^{10} node points and 6 subdomains.

Finally, in order to compare the results in Table 1, we cite results from other papers where the same or a similar problem is studied. The same coefficients were used in [106] and a similar nonlinearity was first considered in the early paper [35]. In the latter Newton method is applied with overrelaxation for FDM on a square with 90 and 870 points, and requires 20, resp. 98, iterations to achieve a residual error $\varepsilon = 10^{-6}$. Successive overrelaxation (or Kacanov's frozen coefficient method) in op. cit. requires 18, resp. 58, iterations for the same error, and the variants of this method require 162 iterations for $\varepsilon = 10^{-5}$ with 384 node points in [63] (on a complicated domain) and 15 iterations for $\varepsilon = 10^{-6}$ with 1000 node points in [106], respectively. Compared even to this last fastest result, the iteration (2.6.5)–(2.6.6) is less costly. Namely, since the auxiliary systems in (2.6.6) come from a piecewise constant coefficient operator, their structure is simpler than either for Newton method or for frozen coefficients. That is, the matrices of the auxiliary systems are the modifications of the discrete Laplacian such that their updating consists of updating the diagonal matrix $\mathcal{W}_n$ in the decomposition $\mathcal{B}_n = \mathcal{C}\mathcal{W}_n\mathcal{C}^T$. In fact, updating only requires distributing the six constants c_i at the entries corresponding to the subdomains Ω_i, and this structure property only slightly increases the complexity of a Laplacian solver.

2.6.2 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 aim is to determine the tangential stress in the points of the rod under given torsion. The data for our runs were provided for us from ELTE, Institute of Physics [150], and we have presented our results in [57].

The Saint-Venant model of elasto-plastic torsion in the hardening state is described in [77]. One assumes that the cross-sections experience rigid rotation in their planes and are distorted in the direction of the z-axis. The tangential stress vectors τ act in cross-sections parallel to the (x, y)-plane, thus we write

$$\tau \;=\; (\tau_x, \tau_y).$$

Further, one can introduce a stress function u fulfilling

$$\tau_x = \tfrac{\partial u}{\partial y}, \quad \tau_y = -\tfrac{\partial u}{\partial x}. \tag{2.6.7}$$

The condition of the hardening state involves the single curve model, wherein the connection between strain and stress depends only on the strain and stress intensities. The latter is denoted by $T := \left(\tau_x^2 + \tau_y^2\right)^{1/2} = |\tau|$. The increasing connection function $\overline{g}$ is defined in a bounded validity interval $[0, T_*]$. We require that $\overline{g} \in C^1[0, T_*]$ and

$$0 < \mu_1 \le \overline{g}(T) \le (\overline{g}(T)T)' \le \mu_2 \qquad (T \in [0, T_*]) \tag{2.6.8}$$

with suitable constants μ_1, μ_2 independent of T. Based on these, one can derive the equation

$$\begin{aligned} -\tfrac{\partial}{\partial x}\left(\overline{g}(T)\tfrac{\partial u}{\partial x}\right) - \tfrac{\partial}{\partial y}\left(\overline{g}(T)\tfrac{\partial u}{\partial y}\right) \;&=\; 2\omega\,, \\ \text{where} \quad T = |\tau| = |\nabla u|\,. \end{aligned} \tag{2.6.9}$$

Since u is only determined up to additive constant, the constant boundary value may be chosen 0, hence the discussed model leads to the nonlinear Dirichlet boundary value problem written briefly as

$$\begin{cases} -\mathrm{div}\,(\overline{g}(|\nabla u|)\nabla u) \;=\; 2\omega \\ \qquad\qquad\qquad u_{|\partial\Omega} \;=\; 0\,. \end{cases} \tag{2.6.10}$$

If this is solved for u then the required tangential stress is obtained from (2.6.7).

We solve problem (2.6.10) numerically using a FEM discretization and then Sobolev gradient preconditioning with the discrete Laplacian preconditioner

$$(-\Delta_h)_{i,j} = \int_\Omega \nabla v_i \cdot \nabla v_j \qquad (i, j = 1, ..., k),$$

where $v_1, ..., v_k$ is a basis of V_h. We define the bounds

$$m = \overline{g}(0), \quad M = \max_{0 \le T \le T_*} (\overline{g}(T)T)'. \tag{2.6.11}$$

For simplicity, we pick $u_0 \equiv 0$ for the initial guess. Then the algorithm is as follows:

$$\begin{cases} (a) \quad u_0 \equiv 0; \\ \\ \qquad \text{for any } \ n \in \mathbf{N}: \ \text{ if } \ u_n \in V_h \ \text{ is obtained, then} \\ \\ (b1) \quad z_n \in V_h \text{ is the solution of} \\ \\ \qquad \displaystyle\int_\Omega \nabla z_n \cdot \nabla v = \int_\Omega \overline{g}(|\nabla u_n|) \nabla u_n \cdot \nabla v - 2\omega \int_\Omega v \qquad (v \in V_h); \\ \\ (b2) \qquad u_{n+1} = u_n - \dfrac{2}{M+m} z_n\,. \end{cases} \tag{2.6.12}$$

Here the constants m and M are taken from (2.6.11), further, the auxiliary linear algebraic systems in step (b1) can be solved by a fast Poisson solver.

The convergence of the algorithm (2.6.12) follows from Theorem 2.2.4:

$$\|u_n - u_h\|_{H_0^1} \le C \cdot \left(\frac{M-m}{M+m} \right)^n \qquad (n \in \mathbf{N})$$

with $C = \frac{2}{m\varrho^{1/2}} \|\omega\|_{L^2(\Omega)}$, where m and M are from (2.6.11), and $\varrho > 0$ is the smallest eigenvalue of $-\Delta$ on $H^2(\Omega) \cap H_0^1(\Omega)$. For the constant C, we note that $\|\omega\|_{L^2(\Omega)} = \omega |\Omega|^{1/2}$ since ω is constant, and one can use the estimate $\varrho \ge 2\pi^2 / diam(\Omega)^2$ from [38] where $|\Omega|$ and $diam(\Omega)$ denote the area and diameter of Ω, respectively. Hence

$$C \le \frac{diam(\Omega)\, (2|\Omega|)^{1/2} \omega}{m\pi}$$

with m from (2.6.11). Note that the obtained convergence estimate is *mesh independent*, since it only contains data from the original problem before discretization.

Numerical experiment. We enclose the numerical results from [57] for problem (2.6.10). We consider a copper rod with a square cross-section 10 mm × 10 mm. The material was heat treated at the temperature $600°C$ for 1 hour, and the corresponding strain-stress function $\overline{g}$ is then determined using data obtained from the measurements [150]. The aim is to determine the tangential stress field ∇u.

We applied the algorithm (2.6.12) with C^1-elements. The use of such a higher order FEM requires a much larger number of arithmetic operations, and therefore it is not widespread. However, the reasonability of its usage is justified in literature [159] and, in particular, it is also a basis for the *hp*-version [144]. The C^1-elements lead to higher order error estimates [144], therefore a given accuracy requires smaller h than with lower degree elements and hence the arising matrix sizes are not much larger. In our case C^1-elements were motivated by qualitative aspects, since the continuity of the tangential stress field τ is thus reproduced by the numerical approximations without postprocessing.

The numerical tests used $\omega = 0.3613$. The derived convergence quotient estimate was $\frac{M-m}{M+m} = 0.6243$. The computations were executed up to accuracy 10^{-4}. The FEM

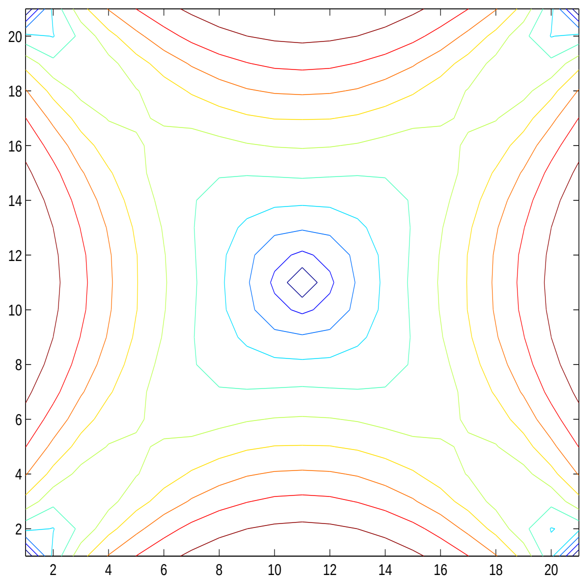

Figure 2.1: The contours of the tangential stress intensity

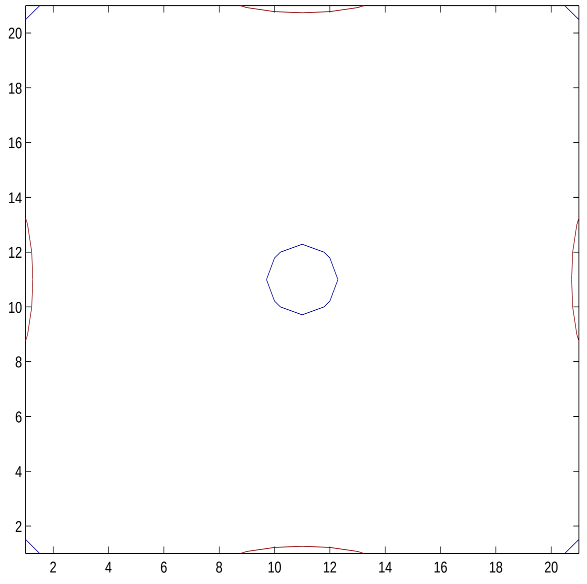

Figure 2.2: The regions of elastic state, plastic state and crack

error estimate shows that even $h = 2.5$ mm is a reasonable choice for this purpose. The convenience of this coarse mesh is due to the use of C^1-elements. Then it took 16 iterations to achieve the prescribed accuracy.

The contours of the obtained tangential stress intensity are plotted in Figure 2.1. The cross-section can be divided into three parts: the corners and a small central part are in elastic state, in the middle of the edges crack occurs, and the intermediate region is in plastic state (see Figure 2.2.)

Remark 2.6.1 (i) One could also use either Newton's method or the piecewise constant coefficient preconditioners from the previous subsection. However, these more involved methods are not necessary for the given nonlinearity: as shown by the experiment, the convergence ratio was small enough to justify the cheaper Laplacian preconditioner that produced the desired convergence in 16 steps and did not need any updating of coefficients.

(ii) The method is applicable in a similar way to related problems with the same structure, described in subsection 2.4.2. The fourth-order model of elasto-plastic bending of plates can be solved similarly: one can using biharmonic preconditioners if the nonlinearity $\overline{g}$ is as mild in the above example, whereas piecewise constant coefficient preconditioners are proposed for nonlinearities with large jumps.

2.6.3 The electrostatic potential equation

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

$$\begin{cases} T(u) \equiv -\Delta u + e^u = 0 \\ u_{|\partial\Omega} = 0, \end{cases} \tag{2.6.13}$$

see e.g. [106]. We assume that the domain is C^2-diffeomorphic to a convex one, which essentially means that it does not have concave corners.

We apply Sobolev gradient preconditioning based on subsection 2.2.2, see [97] for more details. Let $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, $u_0 \le 0$ and the sequence $(u_n) \subset H^2(\Omega) \cap H_0^1(\Omega)$ be defined by

$$\begin{gathered} u_{n+1} = u_n - \frac{2\varrho}{2\varrho+1}\, z_n\,, \\ \text{where} \quad -\Delta z_n = -\Delta u_n + e^{u_n}, \quad z_{n|\partial\Omega} = 0 \end{gathered} \tag{2.6.14}$$

and $\varrho > 0$ is the smallest eigenvalue of $-\Delta$ on $H^2(\Omega) \cap H_0^1(\Omega)$. By setting $w_n := z_n - u_n$, the iteration (2.6.14) takes the simpler form

$$u_{n+1} = \frac{1}{2\varrho+1}\,(u_n - 2\varrho\, w_n)\,, \qquad \text{where} \quad -\Delta w_n = e^{u_n}, \quad w_{n|\partial\Omega} = 0.$$

The maximum principle and induction imply that $u_n \le 0$, hence the nonlinearity remains bounded. Then (u_n) converges linearly to u^*, namely,

$$\|u_n - u^*\|_{H_0^1(\Omega)} \le \varrho^{-1/2} \| -\Delta u_0 + e^{u_0}\|_{L^2(\Omega)} \left(\frac{1}{2\varrho+1}\right)^n \qquad (n \in \mathbf{N}). \tag{2.6.15}$$

Numerical experiment. We have developed a direct realization when the domain Ω is a ball. Then the solution is radially symmetric [60] and, thanks to the special form of the problem, a direct approach becomes possible which avoids discretization. One can instead realize Theorem 2.2.1 directly in the Sobolev space $H_0^1(B)$ by keeping the iterates in the class of radially symmetric polynomials

$$\mathcal{P} = \{\sum_{m=0}^{l} a_m r^{2m} : l \in \mathbf{N}, a_m \in \mathbf{R}\}, \quad \text{with} \quad r = |x| \quad \text{for } x \in B,$$

where the Laplacian can be inverted exactly.

In each step of the iteration we have approximated e^{u_n} by a suitable Taylor polynomial, and dropped the small high-index coefficients to avoid rapid growth in the degrees of the polynomials. Then the iteration satisfies (2.6.15) up to accuracy ε, which (letting $u_0 := 0$, and using the actual data) amounts to

$$\|u_n - u^*\|_{H_0^1(B)} \leq (\frac{|B|}{\varrho})^{1/2} \left(\frac{1}{2\varrho + 1} \right)^n + \varepsilon \qquad (n \in \mathbf{N}). \tag{2.6.16}$$

By induction, the approximated right-hand sides are in $\mathcal{P}$, and if

$$p(u_n)(r) = \sum_{m=0}^{l_n} a_m r^{2m} \qquad (r \in [-R, R]), \tag{2.6.17}$$

then $-\Delta w_n = p(u_n)$ is equivalent to

$$-\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial w_n}{\partial r} \right) = \sum_{m=0}^{l_n} a_m r^{2m}, \qquad w_n(-R) = w_n(R) = 0.$$

Thus the Laplacian can be inverted exactly, since the solution $w_n \in \mathcal{P}$ of the above equation is given explicitly by

$$w_n(r) = \sum_{m=0}^{l_n} \frac{a_m}{(2m+3)(2m+2)} (R^{2m+2} - r^{2m+2}) \,. \tag{2.6.18}$$

The experiments used *Mathematica*[1] as a working environment. The test were performed on a ball with radius $R = 2$.

The residuals achieved accuracy 10^{-6} in 9 steps. Figure 1 contains graphs of the first few terms of this sequence and shows the rapid convergence. (In fact, for the sake of positivity, the functions $|u_n(r)|$ are plotted instead.)

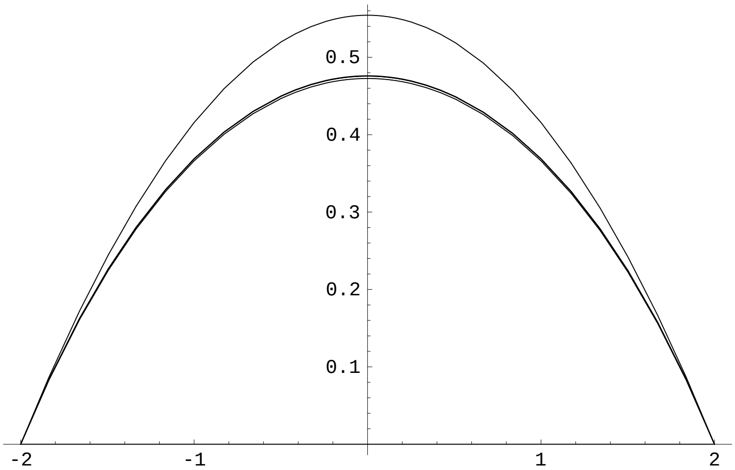

Figure 1 : *The first few terms of the sequence* $|u_n(r)|$.

[1] Copyright 1988-2000 Wolfram Research, Inc.

In order to better visualize the graph of the numerical solution, one dimension is omitted in Figure 2 by plotting the surface of the 2D function which attains the same values along the radii.

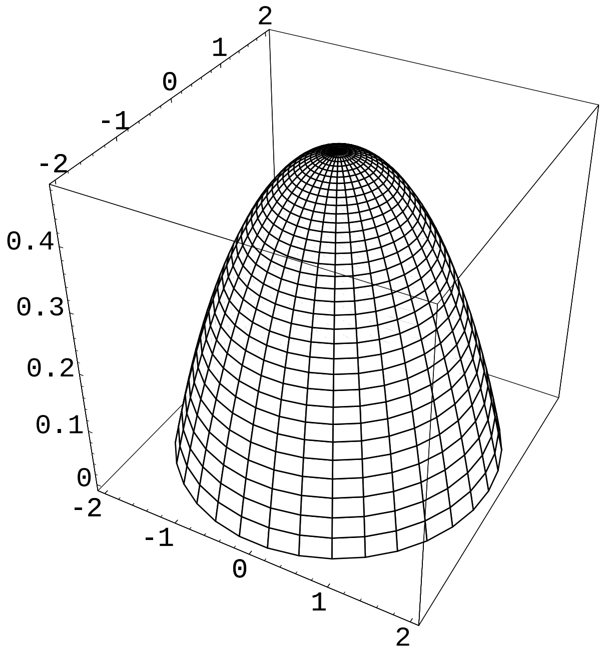

Figure 2 : *Graph of the modulus of the numerical solution* u_{12}.

Concluding the example, we have realized direct Laplacian preconditioning, due to keeping the iteration in the class of radially symmetric polynomials where the Laplacian is exactly invertible. The main advantage of this method is the simplicity of the algorithm, whose fast linear convergence has been observed.

2.6.4 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. In general, consider the quasilinear problem

$$\begin{aligned} T(u) &\equiv -\operatorname{div} f(x,\nabla u) + q(x,u) = g(x) \quad \text{in } \Omega \\ Q(u) &\equiv f(x,\nabla u)\cdot \nu + \int_{\partial\Omega} \varphi(x,y)u(y)\,d\sigma(y) = 0, \quad \text{on } \partial\Omega \end{aligned}$$

where the conditions corresponding to problem (2.2.21) are satisfied, and in addition, the nonlocal term has the following properties: the function $\varphi : \partial\Omega^2 \to \mathbb{R}$ is

(i) a positive kernel, i.e. it fulfills

$$\varphi(x,y) = \int_{\partial\Omega} \psi(x,z)\psi(z,y)\, d\sigma(z) \quad (x,y \in \partial\Omega)$$

with some $\psi \in L^2(\partial\Omega^2)$ satisfying $\psi(x,y) = \psi(y,x)$ $(x,y \in \partial\Omega)$;

(ii) regular, i.e. the function $x \mapsto \displaystyle\int_{\partial\Omega} \varphi(x,z)\, d\sigma(z)$ does not a.e. vanish on $\partial\Omega$.

Then one can define proper Sobolev gradient preconditioning by adapting Theorem 2.2.5. Let $u_0 \in H^1(\Omega)$ and compute M_0 as in (2.2.23). Assume that u_n is constructed. Then

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

where $z_n \in H^1(\Omega)$ solves the auxiliary linear nonlocal problem

$$\begin{aligned} \int_\Omega \nabla z_n \cdot \nabla v + \frac{1}{m} \iint_{\partial\Omega^2} \varphi(x,y) z_n(y) v(x)\, d\sigma(y)\, d\sigma(x) \\ = \ \langle F(u_n), v\rangle - \int_\Omega gv \quad (v \in H^1(\Omega)). \end{aligned} \tag{2.6.19}$$

Note that the auxiliary problems involve a fixed linear operator without updating.

Numerical experiments were run in [80] for the semilinear problem

$$-\Delta u + u^3 = g(x,y) \quad \text{in } \Omega, \qquad \tfrac{\partial u}{\partial \nu} + \int_{\partial\Omega} u\, d\sigma = 0 \quad \text{on } \partial\Omega\,. \tag{2.6.20}$$

The calculations 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 the system of boundary value problems

$$\left\{ \begin{array}{l} T_i(u_1,\dots,u_r) \equiv \ -\text{div}\,(a_i(x)\nabla u_i) + f_i(x,u_1,\dots,u_r) \ = \ g_i(x) \quad \text{in } \Omega \\ Qu_i \equiv (\alpha(x)u_i + \beta(x)\partial_\nu u_i)_{|\partial\Omega} = 0 \end{array} \right. \tag{2.6.21}$$

$(i = 1,\dots,r)$ on a bounded domain $\Omega \subset \mathbf{R}^N$. We impose the following conditions:

(C1) $\partial\Omega \in C^2$, $a_i \in C^1(\overline{\Omega})$, $f_i \in C^1(\overline{\Omega} \times \mathbf{R}^r)$, $g_i \in L^2(\Omega)$.

(C2) $\alpha, \beta \in C^1(\partial\Omega)$, $\alpha, \beta \geq 0$, $\alpha^2 + \beta^2 > 0$ almost everywhere on $\partial\Omega$.

(C3) There are constants $m, m' > 0$ such that $0 < m \leq a_i(x) \leq m'$ $(x \in \overline{\Omega})$, further, $\eta \equiv \sup\limits_{\Gamma_\beta} \frac{\alpha}{\beta} < +\infty$ where

$$\Gamma_\beta \equiv \{x \in \partial\Omega : \beta(x) > 0\}.$$

(C4) Let $2 \le p \le \frac{2N}{N-2}$ (if $N > 2$), $2 \le p$ (if $N = 2$). There exist constants $\kappa' \ge \kappa \ge 0$ and $\gamma \ge 0$ such that for any $(x,\xi) \in \overline{\Omega} \times \mathbf{R}^r$ the Jacobians $\partial_\xi f(x,\xi) = \{\partial_{\xi_k} f_j(x,\xi_1,\dots,\xi_r)\}_{j,k=1}^r \in \mathbf{R}^{r\times r}$ are symmetric and their eigenvalues μ fulfil

$$\kappa \le \mu \le \kappa' + \gamma \sum_{j=1}^r |\xi_j|^{p-2}.$$

Moreover, in the case $\alpha \equiv 0$ we assume $\kappa > 0$, otherwise $\kappa = 0$.

System (2.6.21) is a special case of (2.2.21), hence the convergence of the iteration (2.2.24)–(2.2.25) is ensured by Theorem 2.2.5. The iteration requires the solution of independent linear elliptic problems.

Numerical experiments were run in [82] in the same spirit as for the nonlocal problem above. The system

$$\begin{cases} -\Delta u + u - v + u^3 = g_1(x,y) \\ -\Delta v + v - u + v^3 = 0 \\ u_{|\Gamma_1} = v_{|\Gamma_1} = 0,\ \partial_\nu u_{|\Gamma_2} = \partial_\nu v_{|\Gamma_2} = 0 \end{cases} \tag{2.6.22}$$

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 $u \ge 0$ in a radiating body $\Omega \subset \mathbf{R}^3$ is described by the problem

$$\begin{cases} -\mathrm{div}\,(\kappa(x)\,\nabla u) + \sigma(x) u^4 = 0 & \text{in } \Omega, \\ \kappa(x)\frac{\partial u}{\partial \nu} + \alpha(x)\,(u - \tilde{u}(x)) = 0 & \text{on } \partial\Omega\,, \end{cases} \tag{2.6.23}$$

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 [99].

Problem (2.6.23) is a special case of problem (2.3.42), hence Corollary 2.3.4 provides convergence of the variable preconditioning procedure using constant coefficient operators with stepwise redefined coefficient of u. We cite the numerical tests executed in [105], which show that this variable preconditioning iteration can become faster w.r.t. run time compared to Newton's method, due to the lack of updating the coefficients.

2.6.5 Nonlinear elasticity systems

The description of an elastic body in structural mechanics leads to an elliptic system of three equations

$$\left.\begin{cases} -\,\mathrm{div}\; T_i(x,\varepsilon(\mathbf{u})) = \varphi_i(x) & \text{in } \Omega \\ T_i(x,\varepsilon(\mathbf{u}))\cdot\nu = \gamma_i(x) & \text{on } \Gamma_N \\ u_i = 0 & \text{on } \Gamma_D \end{cases}\right\} \quad (i = 1,2,3), \tag{2.6.24}$$

where the vector function $\mathbf{u} : \Omega \to \mathbf{R}^3$ represents displacement, and the tensor T is expressed with the bulk modulus k and Lamé's coefficient μ as

$$T(x,\varepsilon(\mathbf{u})) = 3k(x, |\mathrm{vol}\,\varepsilon(\mathbf{u})|^2)\ \mathrm{vol}\,\varepsilon(\mathbf{u}) + 2\mu(x, |\mathrm{dev}\,\varepsilon(\mathbf{u})|^2)\ \mathrm{dev}\,\varepsilon(\mathbf{u}).$$

Here

$$\begin{aligned} &0<\lambda_0\le 3k(x,s)\le\Lambda_0\,, \qquad 0<\lambda_0\le 2\mu(x,s)\le\Lambda_0\,,\\ &0<\lambda_0\le \tfrac{\partial}{\partial s}\left(3\,k(x,s^2)s\right)\le\Lambda_0\,, \qquad 0<\lambda_0\le \tfrac{\partial}{\partial s}\left(2\,\mu(x,s^2)s\right)\le\Lambda_0\,, \end{aligned}\tag{2.6.25}$$

with suitable constants $\Lambda_0\ge\lambda_0>0$ independent of (x,s). Further, the functions $\varphi:\Omega\to\mathbf{R}^3$ and $\gamma:\Gamma_N\to\mathbf{R}^3$ describe the body and boundary force vectors, respectively. See [26].

One can solve this problem by an outer-inner iteration as described in paragraph (a) of subsection 2.4.2. Then a crucial step is the choice of preconditioner for the linearized systems $L_h^{(n)}p_h^{(n)}=r_h^{(n)}$ which consist of three equations. An efficient choice of inner preconditioning operator is the triplet of independent Laplacians:

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

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

One can then derive that this preconditioner leads to condition numbers bounded independently of both V_h and n (the outer iteration number):

Theorem 2.6.1 *The separate displacement preconditioner satisfies*

$$cond(S_h^{-1}L_h)\le\kappa\,\frac{\Lambda_0}{\lambda_0}\tag{2.6.26}$$

where $\kappa>0$ is the Korn constant and λ_0 and Λ_0 are from (2.6.25).

Proof. Proposition 2.4.1 and (2.6.25) yield the bound Λ_0/λ_0 in the norm $\left(\int\limits_\Omega|\varepsilon(\mathbf{u})|^2\right)^{1/2}$, and the additional factor κ comes from the estimates involving Korn's inequality [58]

$$\int\limits_\Omega|\varepsilon(\mathbf{u})|^2\le\|\mathbf{u}\|^2_{H^1_D}\le\kappa\int\limits_\Omega|\varepsilon(\mathbf{u})|^2\qquad(\mathbf{u}\in H^1_D(\Omega)^3).\qquad\blacksquare$$

Consequently, the inner PCG iteration converges with ratio independently of both the mesh size and the outer Newton iterate. More details on this problem are found in [15].

2.6.6 Interface problems for localized reactions

Chemical reaction-diffusion equations may involve reactions that take place in a localized way on a surface (interface). This gives rise to so-called interface conditions similar to Neumann boundary conditions, but involving the jump of the solution and its normal derivative.

We consider compound nonlinear interface problems that involve reaction terms both inside the domain and on the interface. Then one has the problem

$$\left\{\begin{array}{rcll} -\Delta u+q(x,u) &=& f(x) & \text{in } \Omega\setminus\Gamma,\\ [u]_\Gamma &=& 0 & \text{on } \Gamma,\\ \left[\frac{\partial u}{\partial\nu}\right]_\Gamma+s(x,u) &=& \gamma(x) & \text{on } \Gamma,\\ u &=& g(x) & \text{on } \partial\Omega, \end{array}\right.\tag{2.6.27}$$

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, see [93] for a derivation. Therefore outer-inner (Newton plus PCG) iterations can be defined in a similar way as for standard mixed boundary value problems. Preconditioning the arising linearized problems by the Laplacian principal part, one can achieve mesh independent superlinear convergence similarly to the problems we had seen before, see [5].

Table 2.1: Outer residuals and inner iteration numbers for the interface problem

	$N = 64$		$N = 128$		$N = 192$	
n	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}
1	2.7768	1	2.7784	1	2.7787	1
2	2.5545	1	2.5562	1	2.5565	1
3	2.3322	1	2.3339	1	2.3342	1
4	2.1099	1	2.1116	1	2.1119	1
5	1.8875	1	1.8892	1	1.8895	1
6	1.6651	1	1.6668	1	1.6671	1
7	1.4426	1	1.4443	1	1.4446	1
8	1.2201	1	1.2217	1	1.2221	1
9	0.99753	1	0.99918	1	0.99949	1
10	0.77492	1	0.77657	1	0.77688	1
11	0.55228	1	0.55393	1	0.55424	1
12	0.32961	1	0.33126	1	0.33157	1
13	$7.3156 \cdot 10^{-3}$	3	$7.3741 \cdot 10^{-3}$	3	$7.3849 \cdot 10^{-3}$	3
14	$4.0382 \cdot 10^{-6}$	7	$4.0782 \cdot 10^{-6}$	7	$4.0867 \cdot 10^{-6}$	7
15	$9.5271 \cdot 10^{-12}$	15	$1.1658 \cdot 10^{-12}$	15	$1.3051 \cdot 10^{-12}$	15

Thereby, we have run experiments on a test-problem as follows. The domain was $\Omega = [0,1]\times[0,1]$ 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 code was written in Matlab, and the stopping criterion was $\|F_h(u_{nh}) - f_h\|_S \le 10^{-10}$. The result are described in Table 2.1, and show the expected mesh independence.

2.6.7 Nonsymmetric transport systems

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

$$\left.\begin{aligned} -\Delta u_i + \mathbf{b}_i \cdot \nabla u_i + f_i(u_1,\dots,u_l) &= g_i \\ u_{i\,|\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1,\dots,l), \tag{2.6.28}$$

where the $\mathbf{b}_i$ represent 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 2.4.2, in which case the system becomes of the form (2.4.39).

One can solve this problem by an outer-inner iteration as described in paragraph (b) of subsection 2.4.2. Then the outer Newton iteration consists of systems of the form (1.2.72), where one can propose as inner preconditioning operator the l-tuple of independent diffusion operators (the principal parts) as in (1.5.8). The solution of the linearized systems admits efficient parallelization as mentioned in subsection 1.5.5. For such preconditioning both the outer and inner iterations produce mesh independent 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 r.h.s. came from a given exact solution. The experiments were carried out in the following way:

- we used Courant elements for the FEM discretization using uniform triangle mesh with width h
- the stopping criterion was $\|F_h(\underline{u}_n) - b_h\| \leq 10^{-6}$;
- the auxiliary problems were solved with FFT;
- we used adaptive damping parameters τ_n;
- the code was written in Matlab and run on a PC.

We have run the code for the system with $l = 2, 4, 6$ equations, respectively. The results were much similar for different l with a slight increase in number of inner iterations and large increase in computing time.

We present the results in Table 2.2 for $l = 4$ equations, here $r_n := \|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{H_0^1}$ is the residual error at the nth outer and n_{inn} denotes the number of inner iterations. The superlinear phase of the outer DIN iteration starts around the 5th step. The mesh uniform behaviour of the convergence can be observed in both the outer and inner iterations.

The CPU times are also given. These also include the time of building the finite element matrices. Since Matlab has been used, no total time-cost analysis is carried out but the CPU times only serve for illustration.

2.6.8 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 [160]. A linearized form was studied in subsection 1.5.4. As has been mentioned, in real-life situations there may be several chemical species, leading to a huge number of equations. The system has the following form:

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

Table 2.2: Outer residuals and inner PCG steps for the transport system

	$1/h = 17$		$1/h = 33$		$1/h = 49$	
n	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}
1	7.3726	1	7.4081	1	7.4151	1
2	5.3727	1	5.3940	1	5.3982	1
3	3.4515	2	3.4790	2	3.4845	2
4	1.3288	1	1.3399	2	1.3421	2
5	$6.6101 \cdot 10^{-1}$	2	$3.5355 \cdot 10^{-1}$	2	$3.5561 \cdot 10^{-1}$	2
6	$2.3429 \cdot 10^{-1}$	2	$9.2309 \cdot 10^{-2}$	5	$9.3523 \cdot 10^{-2}$	5
7	$5.7094 \cdot 10^{-2}$	5	$1.6705 \cdot 10^{-2}$	7	$1.6983 \cdot 10^{-2}$	7
8	$3.5825 \cdot 10^{-3}$	17	$2.2688 \cdot 10^{-3}$	17	$2.3033 \cdot 10^{-3}$	17
9	$3.3643 \cdot 10^{-4}$	24	$2.8591 \cdot 10^{-4}$	24	$2.9181 \cdot 10^{-4}$	24
10	$3.5510 \cdot 10^{-5}$	23	$3.7328 \cdot 10^{-5}$	37	$3.8277 \cdot 10^{-5}$	37
11	$4.4460 \cdot 10^{-6}$	41	$4.9166 \cdot 10^{-6}$	49	$5.0674 \cdot 10^{-6}$	49
CPU time(s)	$1.1822 \cdot 10^{2}$		$8.2159 \cdot 10^{2}$		$4.1348 \cdot 10^{3}$	
	$1/h = 65$		$1/h = 81$		$1/h = 97$	
n	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}	$\|r_n\|$	n_{inn}
1	7.4176	1	7.4188	1	7.4194	1
2	5.3997	1	5.4004	1	5.4008	1
3	3.4865	2	3.4874	2	3.4879	2
4	1.3429	2	1.3433	2	1.3435	2
5	$3.5636 \cdot 10^{-1}$	2	$3.5670 \cdot 10^{-1}$	2	$3.5690 \cdot 10^{-1}$	2
6	$9.3961 \cdot 10^{-2}$	5	$9.4167 \cdot 10^{-2}$	5	$9.4280 \cdot 10^{-2}$	5
7	$1.7084 \cdot 10^{-2}$	7	$1.7132 \cdot 10^{-2}$	7	$1.7158 \cdot 10^{-2}$	7
8	$2.3158 \cdot 10^{-3}$	18	$2.3217 \cdot 10^{-3}$	18	$2.3249 \cdot 10^{-3}$	18
9	$2.9276 \cdot 10^{-4}$	24	$2.9376 \cdot 10^{-4}$	24	$2.9430 \cdot 10^{-4}$	24
10	$3.9288 \cdot 10^{-5}$	37	$3.9456 \cdot 10^{-5}$	37	$3.9548 \cdot 10^{-5}$	37
11	$5.2105 \cdot 10^{-6}$	49	$5.2372 \cdot 10^{-6}$	49	$5.2519 \cdot 10^{-6}$	49
CPU time(s)	$1.2864 \cdot 10^{4}$		$3.0766 \cdot 10^{4}$		$6.2980 \cdot 10^{4}$	

Such problems are solved by time discretization, Newton linearization and inner PCG iteration. The coercivity property can be ensured by choosing a sufficiently small stepsize τ in the time discretization. The nonlinear systems arising after time discretization are similar to (2.6.28) studied in the previous section.

Now we are interested in the convergence in time and the behaviour of the overall algorithm. This will demonstrate that the so far developed elliptic solvers are suitable to be a subroutine to a parabolic solution process.

Numerical results are presented as follows. The tests were done on the unit square domain for a system of convection-diffusion consisting of 10 equations, with chemical reactions arising from the air pollution model in [160].

Table 2.3 shows the results on the time levels. The number of outer DIN iterations (executed in every time step) and the number of inner PCG iterations (carried out in each

Table 2.3: Outer residuals and inner PCG steps for the air pollution problem on time levels

$N = h^{-1} = 32$											
$t = 0.00$			$t = 0.25$			$t = 0.50$			$t = 0.75$		
n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}
0	0.09482921	2	0	0.01575960	2	0	0.00260900	3	0	0.00043107	4
1	0.02841575	2	1	0.00472253	3	1	0.00078171	4	1	0.00012916	4
2	0.00254771	3	2	0.00042270	4	2	0.00006996	5	2	0.00001156	5
3	0.00000222	7	3	0.00000024	6	3	0.00000003	5	3	0.00000000	-
4	0.00000000	-	4	0.00000000	-	4	0.00000000	-	-	-	-

DIN step) are denoted by n and n_{inn}, respectively. The stopping criterion in the DIN method was chosen to be $\|F_h(u) - b_h\| < 10^{-8}$.

Considering time, the errors are shown in four different points in the time interval, when various spatial ($h = 1/N$) and time parameters (τ) were chosen. Since no exact solution is available, only the approximate solutions calculated in a pair of grids can be compared, when τ and $\tau/2$ are used as time parameters. The results are shown in Table 2.4. Mesh independence as N grows is also seen here, and considering time, one can observe that the error $\|u_h^{(\tau)} - u_h^{(\tau/2)}\| \to 0$ numerically as $t \to 0$, which shows numerical convergence of the method w.r.t. time.

Table 2.4: Error estimation in time for the air pollution problem

		$\|u_h^{(\tau)} - u_h^{(\tau/2)}\|$			
t	τ	$N = 8$	$N = 16$	$N = 32$	$N = 64$
0.25	1/4	5.6032e-03	5.5971e-03	5.5962e-03	5.6078e-03
	1/8	2.9354e-03	2.9157e-03	2.9357e-03	2.9311e-03
	1/16	1.3272e-03	1.3174e-03	1.3210e-03	1.3189e-03
0.50	1/4	1.5072e-03	1.4957e-03	1.4987e-03	1.4979e-03
	1/8	3.9029e-04	3.8588e-04	3.8338e-04	3.8192e-04
	1/16	8.9336e-05	8.7142e-05	8.6723e-05	8.6821e-04
0.75	1/4	3.0803e-04	3.0438e-04	3.0280e-04	3.0129e-04
	1/8	3.9768e-05	3.8658e-05	3.8191e-05	3.7851e-05
	1/16	4.5972e-06	4.3512e-06	4.2254e-06	4.1974e-06
1.00	1/4	5.7434e-05	5.6288e-05	5.5750e-05	5.5580e-05
	1/8	3.7062e-06	3.5447e-06	3.4740e-06	3.4536e-06
	1/16	2.1499e-07	1.9754e-07	1.9221e-07	1.8993e-07

Part II

Reliability of the numerical solution

Chapter 3

Discrete maximum principles

3.1 Preliminaries

The maximum principle forms an important qualitative property of second order linear or nonlinear elliptic equations [61, 133], therefore its discrete analogues, the so-called discrete maximum principles (DMPs) have drawn much attention. The DMP is in fact an important measure of the qualitative reliability of the numerical scheme, otherwise one could get unphysical numerical solutions like negative concentrations etc.

Various DMPs, including geometric conditions on the computational meshes for FEM solutions, have been given e.g. in [33, 71, 104, 140, 157]. For elliptic operators with only principal part, if the discretized operator L_h and the FEM solution u_h satisfy $L_h u_h \leq 0$, then the DMP has the simple form $\max\limits_{\overline{\Omega}} u_h = \max\limits_{\partial\Omega} u_h$. On the other hand, for operators with lower order terms as well, one has the weaker statement

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\partial\Omega} u_h\}, \tag{3.1.1}$$

which means that u_h can attain a nonnegative maximum only on the boundary. Moreover, in the latter case one can only provide the DMP for sufficiently fine mesh and needs stronger acuteness type conditions in the case of standard simplicial FEM meshes. Formula (3.1.1) always includes as a special case that $\max\limits_{\partial\Omega} u_h \geq 0$ implies the simple form $\max\limits_{\overline{\Omega}} u_h = \max\limits_{\partial\Omega} u_h$, hence we will not always formulate the latter separately in what follows. We note that significant work on the DMP was also done for stabilized discretizations of convection-dominated problems [142], but our interest here lies in regularly perturbed problems and the extension of the standard Galerkin DMP from linear equations to nonlinear equations and systems.

Previous work on the elliptic DMP was restricted to linear equations, with the exception of [107] where an equation in 3D was considered with a nonlinear coefficient. The DMP was extended for the first time to general nonlinear equations with lower order terms and mixed boundary conditions in our paper [91], and then to nonlinear systems in [92]. The latter was further generalized including first order terms in [94]. This chapter is devoted to nonlinear elliptic equations and systems of general type, based on our mentioned three papers.

The first problem with mixed boundary conditions, even for a single equation, is to clarify what to expect at all as a maximum principle. Namely, estimate (3.1.1) holds independently of boundary conditions, but it only gives real information for Dirichlet boundary conditions when the r.h.s. of (3.1.1) is a priori known. We will show that for mixed boundary conditions one can replace (3.1.1) by

$$\max_{\overline{\Omega}} u_h \leq \max\{0, \max_{\Gamma_D} u_h\} \tag{3.1.2}$$

(where Γ_D is the Dirichlet boundary) if we additionally assume that the Neumann boundary data are also nonpositive. We will also prove this for the CMP, hence (3.1.2) reflects a real property of the exact solution. The r.h.s. of (3.1.2) is a priori known for a mixed boundary value problem.

In the case of coupled systems, we consider a class with coupling which is cooperative and weakly diagonally dominant, since these conditions on the coupling also appear in the underlying continuous maximum principle [40, 119]. In the case of mixed boundary conditions and nonpositive right-hand sides, we have the counterpart of (3.1.2):

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \leq \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} u_k^h\} \tag{3.1.3}$$

where Γ_D is the Dirichlet boundary and k is the number of equations.

As a main practical consequence (also in the scalar case (3.1.2)), this relation will imply discrete nonpositivity or, by reversing signs, discrete nonnegativity under suitable sign conditions on the data.

This chapter is built up as follows. First, after giving some required algebraic background, a matrix maximum principle is established in a Hilbert space framework for proper operator equations. Then we prove discrete maximum principles for nonlinear elliptic equations and various systems. The acuteness type conditions for simplicial FE meshes are also suitably weakened. Some applications are mentioned briefly, where the DMP often reduces to the natural requirement of nonnegativity for the appropriate discrete quantities.

The main technical difficulties encountered are as follows. First, one has to get round the irreducibility criterion that is assumed in the classical algebraic background. Second, when lower order terms of polynomial growth are involved, one needs careful estimates using embedding results and quasi-regular meshes to ensure the required algebraic properties of the stiffness matrix. Whereas in the case of a single equation the DMP will be proved directly, in the case of the considered various types of systems the Hilbert space setting will be exploited to derive the corresponding results in an organized way.

Some classical algebraic results, required in the sequel, are summarized first. We recall a basic definition in the study of DMP (cf. [147]):

Definition 3.1.1 A square $k \times k$ matrix $\mathbf{A} = (a_{ij})_{i,j=1}^k$ is called *irreducible* if for any $i \neq j$ there exists a sequence of nonzero entries $\{a_{i,i_1}, a_{i_1,i_2}, \dots, a_{i_s,j}\}$ of A, where $i, i_1, i_2, \dots, i_s, j$ are distinct indices.

Definition 3.1.2 Let $\mathbf{A}$ be an arbitrary $k \times k$ matrix. The *irreducible blocks* of $\mathbf{A}$ are the matrices $\mathbf{A}^{(l)}$ $(l = 1, \dots, q)$ defined as follows.

Let us call the indices $i, j \in \{1, \ldots, k\}$ *connectible* if there exists a sequence of nonzero entries $\{a_{i,i_1}, a_{i_1,i_2}, \ldots, a_{i_s,j}\}$ of $\mathbf{A}$, where $i, i_1, i_2, \ldots, i_s, j \in \{1, \ldots, k\}$ are distinct indices. Further, let us call the indices i, j mutually connectible if both i, j and j, i are connectible in the above sense. (Clearly, mutual connectibility is an equivalence relation.) Let $N_1, \ldots, N_q$ be the equivalence classes, i.e. the maximal sets of mutually connectible indices. (Clearly, $\mathbf{A}$ is irreducible iff $q = 1$.) Letting $N_l = \{s_1^{(l)}, \ldots, s_{k_l}^{(l)}\}$ for $l = 1, \ldots, q$, we have $k_1 + \cdots + k_q = k$. Then we define for all $l = 1, \ldots, q$ the $k_l \times k_l$ matrix $\mathbf{A}^{(l)}$ by $\mathbf{A}_{pq}^{(l)} := a_{s_p^{(l)}, s_q^{(l)}} \quad (p, q = 1, \ldots, k_l)$.

Remark 3.1.1 One may prove (cf. [8, Th. 4.2]) that by a proper permutation of indices, $\mathbf{A}$ becomes a block lower triangular matrix with the irreducible diagonal blocks $\mathbf{A}^{(l)}$.

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

$$\bar{\mathbf{A}}\bar{\mathbf{c}} \equiv \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \tilde{\mathbf{b}} \end{bmatrix} \equiv \bar{\mathbf{b}}, \tag{3.1.4}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{0}$ is the $m \times k$ zero matrix. The goal here is to establish the algebraic analogue of (3.1.1):

$$\max_{i=1,\ldots,k+m} c_i \le \max\{0, \max_{i=k+1,\ldots,k+m} c_i\}. \tag{3.1.5}$$

Following [33], we introduce

Definition 3.1.3 A $(k+m) \times (k+m)$ matrix $\bar{\mathbf{A}}$ with the structure in (3.1.4) is said to be of *generalized nonnegative type* if the following properties hold:

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

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

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

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

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

Many known results on various discrete maximum principles are based on the following theorem, considered as 'matrix maximum principle' [33, Th. 3]):

Theorem 3.1.1 *Let $\bar{\mathbf{A}}$ be a $(k+m) \times (k+m)$ matrix with the structure as in (3.1.4), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type in the sense of Definition 3.1.3.*

If the vector $\bar{\mathbf{c}} = (c_1, \ldots, c_{k+m})^T \in \mathbf{R}^{k+m}$ (where $(\,.\,)^T$ denotes the transposed) is such that $(\bar{\mathbf{A}}\bar{\mathbf{c}})_i \le 0,\; i = 1, \ldots, k$, then (3.1.5) *holds.*

3.2 Algebraic background

We present some required extensions of the classical results. In fact, the irreducibility of $\mathbf{A}$ is a technical condition which is sometimes difficult to check in applications, see e.g. [42]. We now show that it can be omitted from the assumptions if (iv) is suitably strengthened. For convenient formulations, we will hence use the following

Definition 3.2.1 A $(k+m)\times(k+m)$ matrix $\bar{\mathbf{A}}$ with the structure as in (3.1.4) is said to be of *generalized nonnegative type with irreducible blocks* if properties (i)–(iii) of Definition 3.1.3 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 $\sum\limits_{j=1}^{k} a_{i_0,j} > 0$.

Remark 3.2.1 Let assumptions (i)–(iii) hold in Definitions 3.1.3 or 3.2.1. Then for a given index $i_0 \in \{1,\dots,k\}$, a sufficient condition for the positive row-sum (as in assumption (iv)) to hold is that:

$$\text{there exists an index } j_0 \in \{k+1,\dots,k+m\} \text{ for which } a_{i_0,j_0} < 0.$$

Namely, using also assumptions (ii) and (iii), respectively, we then have

$$\sum_{j=1}^{k} a_{i_0,j} > \sum_{j=1}^{k} a_{i_0,j} + a_{i_0,j_0} \ge \sum_{j=1}^{k} a_{i_0,j} + a_{i_0,j_0} + \sum_{\substack{j=k+1\\ j\neq j_0}}^{k+m} a_{i_0,j} = \sum_{j=1}^{k+m} a_{i_0,j} \ge 0.$$

Theorem 3.2.1 *Let $\bar{\mathbf{A}}$ be a $(k+m)\times(k+m)$ matrix with the structure as in (3.1.4), and assume that $\bar{\mathbf{A}}$ is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.*

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 (3.1.5) holds.

Proof. We may assume that $\mathbf{A}$ has the lower block triangular form mentioned in Remark 3.1.1. (Otherwise we can permute the indices to have this form, since the desired result is independent of the ordering of indices in the block $\mathbf{A}$.) That is, the block $\mathbf{A}$ in $\bar{\mathbf{A}}$ has the irreducible diagonal blocks $\mathbf{A}^{(l)}$ (i.e. the irreducible components defined in Definition 3.1.2), and the corresponding blocks in $\mathbf{A}$ vanish in the upper block triangular part, further, we can use an analogous column decomposition of the block $\tilde{\mathbf{A}}$ to blocks $\tilde{\mathbf{A}}^{(l)}$ $(l = 1,\dots,q)$. Using an analogous decomposition of the vectors $\mathbf{c}$ and $\mathbf{b}$, system (3.1.4) can be written as

$$\begin{bmatrix} \mathbf{A}^{(1)} & \mathbf{0} & \mathbf{0} & \dots & \tilde{\mathbf{A}}^{(1)} \\ \mathbf{A}^{(21)} & \mathbf{A}^{(2)} & \mathbf{0} & \dots & \tilde{\mathbf{A}}^{(2)} \\ \dots & & & & \dots \\ \mathbf{A}^{(q1)} & \mathbf{A}^{(q2)} & \dots & \mathbf{A}^{(q)} & \tilde{\mathbf{A}}^{(q)} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c}^{(1)} \\ \mathbf{c}^{(2)} \\ \dots \\ \mathbf{c}^{(q)} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ \dots \\ \mathbf{b}^{(q)} \\ \tilde{\mathbf{b}} \end{bmatrix} \tag{3.2.1}$$

We must prove that if $\mathbf{b}^{(1)}, ..., \mathbf{b}^{(q)} \leq 0$, then (3.1.5) holds, i.e. $\bar{\mathbf{c}} \leq \max \tilde{\mathbf{c}}$.

Step 1. First we consider the special case when $\tilde{\mathbf{b}} \leq 0$. Then $\tilde{\mathbf{c}} = \tilde{\mathbf{b}} \leq 0$, hence the statement (3.1.5) becomes $\bar{\mathbf{c}} \leq 0$. Since $\bar{\mathbf{c}} = [\mathbf{c}, \tilde{\mathbf{c}}]^T$, we in fact need to prove $\mathbf{c} \leq 0$. We prove by induction that $\mathbf{c}^{(1)}, ..., \mathbf{c}^{(q)} \leq 0$.

Note first that $\mathbf{A}^{(l)}$ $(l = 1, \dots, q)$ are of generalized nonnegative type, since they inherit Assumptions (i)–(iv') in Definition 3.1.3 from $\mathbf{A}$. Namely, this is obvious for Assumptions (i)–(ii). The nonnegativity in Assumption (iii) holds for $\mathbf{A}^{(l)}$ since we drop nonpositive elements in the row sum for $\mathbf{A}^{(l)}$ compared to the row sum for $\mathbf{A}$. Finally, Assumption (iv') for $\mathbf{A}$ just means that the original Assumption (iv) holds for each $\mathbf{A}^{(l)}$. Also, $\mathbf{A}^{(l)}$ are irreducible by definition, hence Theorem 3.1.1 can be applied to systems of the form as in (3.1.4) with left upper block $\mathbf{A}^{(l)}$. We will do this repeatedly for the case $\tilde{\mathbf{c}} \leq 0$ to obtain nonpositive solution vectors.

The first and last rows of (3.2.1) yield the system

$$\begin{bmatrix} \mathbf{A}^{(1)} & \tilde{\mathbf{A}}^{(1)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c}^{(1)} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \tilde{\mathbf{b}} \end{bmatrix}. \tag{3.2.2}$$

Here $\mathbf{b}^{(1)} \leq 0$ and $\tilde{\mathbf{c}} = \tilde{\mathbf{b}} \leq 0$, hence Theorem 3.1.1 yields $\mathbf{c}^{(1)} \leq 0$.

Now let $l \in \{2, \dots, q\}$ and assume that $\mathbf{c}^{(1)}, ..., \mathbf{c}^{(l-1)} \leq 0$. The lth and last rows of (3.2.1) yield the system

$$\begin{bmatrix} \mathbf{A}^{(l)} & \tilde{\mathbf{A}}^{(l)} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c}^{(l)} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{b}}^{(l)} \\ \tilde{\mathbf{b}} \end{bmatrix}, \tag{3.2.3}$$

where $\hat{\mathbf{b}}^{(l)} := \mathbf{b}^{(l)} - \sum\limits_{s=1}^{l-1} \mathbf{A}^{(ls)}\mathbf{c}^{(s)}$. Here $\mathbf{b}^{(l)} \leq 0$ by assumption, $\mathbf{c}^{(s)} \leq 0$ $(s = 1, \dots, l-1)$ from the inductional assumption and $\mathbf{A}^{(ls)} \leq 0$ elementwise from property (ii) of Definition 3.1.3, therefore $\hat{\mathbf{b}}^{(l)} \leq 0$. Using $\tilde{\mathbf{c}} = \tilde{\mathbf{b}} \leq 0$ and applying Theorem 3.1.1 again, we obtain $\mathbf{c}^{(l)} \leq 0$.

Step 2. Let us consider the case when $\max \tilde{\mathbf{b}} = \max \tilde{\mathbf{c}} > 0$. We must prove that if $\mathbf{b}^{(1)}, ..., \mathbf{b}^{(q)} \leq 0$ (i.e. $\mathbf{b} \leq 0$) then (3.1.5) holds, i.e. that $\bar{\mathbf{c}} \leq \max \tilde{\mathbf{c}}$.

Let $\mathbf{c}^* := \bar{\mathbf{c}} - (\max \tilde{\mathbf{c}}) \cdot \mathbf{1}_{k+m}$, where $\mathbf{1}_{k+m}$ is the constant 1 vector of length $k+m$. Since $\bar{\mathbf{A}}\bar{\mathbf{c}} = \bar{\mathbf{b}}$, therefore $\mathbf{c}^*$ is the solution of the linear system $\bar{\mathbf{A}}\mathbf{c}^* = \mathbf{b}^*$, where

$$\begin{aligned} \mathbf{b}^* := \bar{\mathbf{b}} - (\max \tilde{\mathbf{c}}) \cdot \bar{\mathbf{A}}\mathbf{1}_{k+m} &= \begin{bmatrix} \mathbf{b} \\ \tilde{\mathbf{b}} \end{bmatrix} - (\max \tilde{\mathbf{c}}) \cdot \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{1}_k \\ \mathbf{1}_m \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{b} - (\max \tilde{\mathbf{c}}) \cdot \begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \end{bmatrix} \mathbf{1}_{k+m} \\ \tilde{\mathbf{b}} - (\max \tilde{\mathbf{c}}) \cdot \mathbf{1}_m \end{bmatrix}. \end{aligned} \tag{3.2.4}$$

Here the first component in (3.2.4) is nonpositive, since $\mathbf{b} \leq 0$ and $\max \tilde{\mathbf{c}} > 0$ by assumption, further, $\begin{bmatrix} \mathbf{A} & \tilde{\mathbf{A}} \end{bmatrix} \mathbf{1}_{k+m} \geq 0$ by item (iii) of Definition 3.1.3. The second component in (3.2.4) is also nonpositive, since obviously $\tilde{\mathbf{b}} = \tilde{\mathbf{c}} \leq \max \tilde{\mathbf{c}}$. Therefore $\mathbf{b}^* \leq 0$. Thus, applying step 1 to system $\bar{\mathbf{A}}\mathbf{c}^* = \mathbf{b}^*$, we obtain $\mathbf{c}^* \leq 0$, i.e. $\bar{\mathbf{c}} - (\max \tilde{\mathbf{c}}) \cdot \mathbf{1}_{k+m} \leq 0$, which was to be proved. ■

Consequently, in what follows, our main goal is to show that the stiffness matrix of the problems considered is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.

3.3 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

$$\langle A(u), v\rangle = \langle \psi, v\rangle \qquad (v \in H_0) \tag{3.3.1}$$

$$\text{and} \quad u - g^* \in H_0 \tag{3.3.2}$$

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

Assumptions 3.3.1.

(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 $\langle B(u)v, v\rangle \geq m\,\|v\|^2 \quad (u \in H,\ v \in H_0)$.

(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

$$\langle R(u)w, v\rangle \leq M_R(\|u\|)\, |||w|||\, |||v||| \qquad (u, w, v \in H). \tag{3.3.3}$$

In practice for PDE problems (considered in section 3.4.2), g^* plays the role of boundary condition and H_0 will be the subspace corresponding to homogeneous boundary conditions, further, $B(u)$ is the principal part of A.

Assumptions 3.3.1 are not in general known to imply existence and uniqueness for (3.3.1)–(3.3.2). The following extra conditions already ensure well-posedness:

Assumptions 3.3.2.

(i) Let $F(u) := B(u)u, \quad G(u) := R(u)u \quad (u \in H)$. The operators $F, G : H \to H$ are Gateaux differentiable, further, F' and G' are bihemicontinuous (i.e. mappings $(s,t) \mapsto F'(u + sk + tw)h$ are continuous from $\mathbf{R}^2$ to H, and similarly for G').

(ii) There exists a continuous function $M_A : \mathbf{R}^+ \to \mathbf{R}^+$ such that

$$\langle A'(u)w, v\rangle \leq M_A(\|u\|)\, \|w\|\, \|v\| \qquad (u \in H,\ w, v \in H_0). \tag{3.3.4}$$

(iii) There exists a constant $m > 0$ such that $\langle F'(u)v, v\rangle \geq m\,\|v\|^2 \qquad (u \in H,\ v \in H_0)$.

(iv) We have $\langle G'(u)v, v\rangle \geq 0 \quad (u \in H,\ v \in H_0)$.

Proposition 3.3.1 *If Assumptions 3.3.1–3.3.2 hold, then problem (3.3.1)–(3.3.2) is well-posed.*

Proof. Problem (3.3.1)–(3.3.2) can be rewritten as follows:

$$\text{find } u_0 \in H: \ \langle \tilde{A}(u_0), v\rangle \equiv \langle A(u_0+g^*), v\rangle = \langle \psi, v\rangle \qquad (v \in H_0), \tag{3.3.5}$$

$$\text{and let } u := u_0 + g^*. \tag{3.3.6}$$

From assumptions (iii)–(iv) we have

$$\langle A'(u)v, v\rangle \geq m\,\|v\|^2 \qquad (u \in H,\ v \in H_0) \tag{3.3.7}$$

whence A is uniformly monotone on H_0, further, from (3.3.4), A is locally Lipschitz continuous on H_0. These properties of A are inherited by $\tilde{A}$ by the definition of the latter: that is, for all $u, v \in H_0$, we obtain

$$m\,\|u-v\|^2 \leq \langle \tilde{A}(u) - \tilde{A}(v), u-v\rangle, \qquad \|\tilde{A}(u) - \tilde{A}(v)\| \leq M_A(\max\{\|u\|, \|v\|\})\,\|u-v\|\,. \tag{3.3.8}$$

These imply well-posedness for (3.3.5), see, e.g., [55, 106]. ■

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

$$V_h = \text{span}\{\phi_1, ..., \phi_n\} \ \subset H, \qquad V_h^0 = \text{span}\{\phi_1, ..., \phi_{n_0}\} \ \subset H_0 \tag{3.3.9}$$

with a real positive parameter $h > 0$. In practice, as is usual for FEM, h is inversely proportional to n, and one will consider a family of such subspaces, see Definition 3.3.1 later.

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 3.3.3 below, given in terms of the *graph of neighbouring indices*, by which we mean the following. The corresponding indices $\{1, \dots, n_0\}$ or $\{1, \dots, n\}$, 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 3.3.3. The set $\{1, \dots, n\}$ can be partitioned into disjoint sets $S_1, \dots, S_r$ such that for each $k = 1, \dots, r$,

(i) both $S_k^0 := S_k \cap \{1, \dots, n_0\}$ and $\tilde{S}_k := S_k \cap \{n_0+1, \dots, n\}$ 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 $g_h = \sum\limits_{j=n_0+1}^{n} g_j \phi_j \ \in V_h$ be a given approximation of the component of g^* in $H \setminus H_0$. To find the Galerkin solution of (3.3.1)–(3.3.2) in V_h, we solve the following problem: find $u^h \in V_h$ such that

$$\langle A(u^h), v\rangle = \langle \psi, v\rangle \qquad (v \in V_h^0) \tag{3.3.10}$$
$$\text{and} \quad u^h - g_h \in V_h^0. \tag{3.3.11}$$

Using Assumption 3.3.1. (i), we can rewrite (3.3.10) as

$$\langle B(u^h)u^h, v\rangle + \langle R(u^h)u^h, v\rangle = \langle \psi, v\rangle \qquad (v \in V_h^0). \tag{3.3.12}$$

Let us now formulate the nonlinear algebraic system corresponding to (3.3.12). We set

$$u^h = \sum_{j=1}^{n} c_j \phi_j, \tag{3.3.13}$$

and look for the coefficients $c_1, \dots, c_n$. For any $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$, $i = 1, ..., n_0$ and $j = 1, ..., n$, we set

$$b_{ij}(\bar{\mathbf{c}}) := \langle B(u^h)\phi_j, \phi_i\rangle \qquad r_{ij}(\bar{\mathbf{c}}) := \langle R(u^h)\phi_j, \phi_i\rangle, \qquad d_i := \langle \psi, \phi_i\rangle,$$

$$a_{ij}(\bar{\mathbf{c}}) := b_{ij}(\bar{\mathbf{c}}) + r_{ij}(\bar{\mathbf{c}}).$$

Putting (3.3.13) and $v = \phi_i$ into (3.3.12), we obtain a $n_0 \times n$ system of algebraic equations which, using the notations

$$\mathbf{A}(\bar{\mathbf{c}}) := \{a_{ij}(\bar{\mathbf{c}})\},\ i, j = 1, ..., n_0, \qquad \tilde{\mathbf{A}}(\bar{\mathbf{c}}) := \{a_{ij}(\mathbf{c})\},\ i = 1, ..., n_0;\ j = n_0 + 1, ..., n,$$

$$\mathbf{d} := \{d_j\},\ \mathbf{c} := \{c_j\}, \quad j = 1, ..., n_0, \quad \text{and} \quad \tilde{\mathbf{c}} := \{c_j\}, \quad j = n_0 + 1, ..., n, \tag{3.3.14}$$

turns into

$$\mathbf{A}(\bar{\mathbf{c}})\mathbf{c} + \tilde{\mathbf{A}}(\bar{\mathbf{c}})\tilde{\mathbf{c}} = \mathbf{d}. \tag{3.3.15}$$

In order to obtain a system with a square matrix, we enlarge our system to an $n \times n$ one. Since $u^h - g_h \in V_h^0$, the coordinates c_i with $n_0 + 1 \le i \le n$ satisfy automatically $c_i = g_i$, i.e.,

$$\tilde{\mathbf{c}} = \tilde{\mathbf{g}} := \{g_j\}, \quad j = n_0 + 1, ..., n,$$

hence we can replace (3.3.15) by an equivalent system analogous to (3.1.4):

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} \equiv \begin{bmatrix} \mathbf{A}(\bar{\mathbf{c}}) & \tilde{\mathbf{A}}(\bar{\mathbf{c}}) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \tilde{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \tilde{\mathbf{g}} \end{bmatrix}. \tag{3.3.16}$$

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

Definition 3.3.1 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 3.3.1 *Let Assumptions 3.3.1 and 3.3.3 hold. Let us consider the discretization of operator equation (3.3.1)–(3.3.2) in a family of subspaces $\mathcal{V} = \{V_h\}_{h \to 0}$ with bases as in (3.3.9). Let $u^h \in V_h$ be the solution of (3.3.12) 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

$$\langle B(u^h)\phi_j, \phi_i\rangle = 0 \quad \text{and} \quad \langle R(u^h)\phi_j, \phi_i\rangle \le 0, \tag{3.3.17}$$

$$\text{or} \qquad \langle B(u^h)\phi_j, \phi_i\rangle \le -M_B(h) \tag{3.3.18}$$

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

$$T(h) := \sup\{|||\phi_i||| : \ \phi_i \in V_h)\}, \tag{3.3.19}$$

we have

$$\lim_{h\to 0} \frac{M_B(h)}{T(h)^2} = +\infty. \tag{3.3.20}$$

(b) If, in particular, $\phi_i \in V_h^0$ and $\phi_j \in V_h$ are neighbouring basis vectors (as defined for Assumptions 3.3.3), then (3.3.18)–(3.3.20) hold.

(c) $M_R(\|u^h\|)$ is bounded as $h \to 0$, where M_R is the function in Assumption 3.3.1 (iv).

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

(e) For all $h > 0$, $i = 1, ..., n$, we have $\phi_i \in D$ and $\sum\limits_{j=1}^{n} \phi_j \in P$ for the sets D, P introduced in Assumption 3.3.1 (iii).

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.3.14) is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.

Proof. Our task is to check properties (i)–(iv') of Definition 3.2.1 for

$$a_{ij}(\bar{\mathbf{c}}) = \langle B(u^h)\phi_j, \phi_i\rangle + \langle R(u^h)\phi_j, \phi_i\rangle \qquad (i, j = 1, ..., n). \tag{3.3.21}$$

(i) For any $i = 1, ..., n_0$, we have $\phi_i \in V_h^0 \subset H_0$ from (3.3.9), hence we can set $v = \phi_i$ in Assumptions 3.3.1 (ii). Further, by assumption (e), we have $\phi_i \in D$, hence we can set $v = w = \phi_i$ in Assumptions 3.3.1 (iii). These imply

$$a_{ii}(\bar{\mathbf{c}}) = \langle B(u^h)\phi_i, \phi_i\rangle + \langle R(u^h)\phi_i, \phi_i\rangle \ge m\,\|\phi_i\|^2 > 0.$$

(ii) Let $i = 1, ..., n_0$, $j = 1, ..., n$ with $i \ne j$. If (3.3.17) holds then $a_{ij}(\bar{\mathbf{c}}) \le 0$ by (3.3.21). If (3.3.18) holds then, using also (3.3.21), (3.3.3), respectively, and letting $\tilde{M} := \sup M_R(\|u^h\|)$, we obtain

$$a_{ij}(\bar{\mathbf{c}}) \le -M_B(h) + M_R(\|u^h\|)\,|||\phi_i|||\,|||\phi_j||| \le -M_B(h) + M_R(\|u^h\|)\,T(h)^2$$

$$\le T(h)^2\Big(-\frac{M_B(h)}{T(h)^2} + \tilde{M}\Big) < 0 \tag{3.3.22}$$

for sufficiently small h, since by (3.3.20) the expression in brackets tends to $-\infty$ as $h \to 0$.

(iii) For any $i = 1, ..., n_0$,

$$\sum_{j=1}^{n} a_{ij}(\bar{\mathbf{c}}) = \Big\langle B(u^h)\Big(\sum_{j=1}^{n} \phi_j\Big),\, \phi_i\Big\rangle + \Big\langle R(u^h)\Big(\sum_{j=1}^{n} \phi_j\Big),\, \phi_i\Big\rangle \geq 0,$$

since the first term equals zero by assumption (d), further, by assumption (e) we can set $w = \sum\limits_{j=1}^{n} \phi_j$ and $v = \phi_i$ in Assumption 3.3.1 (iii), hence the second term is nonnegative.

(iv') We must prove that for each irreducible component of $\mathbf{A}(\bar{\mathbf{c}})$ there exists an index $i_0 \in N_l = \{s_1^{(l)}, \dots, s_{k_l}^{(l)}\}$ for which $\sum\limits_{j=1}^{n_0} a(\bar{\mathbf{c}})_{i_0,j} > 0$. Here, with the notations of Definition 3.1.2, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ has q irreducible blocks $\mathbf{A}^{(l)}(\bar{\mathbf{c}})$ $(l = 1, \dots, q)$, and N_l denotes the indices arising in $\mathbf{A}^{(l)}(\bar{\mathbf{c}})$. Then $k_1 + \cdots + k_q = n_0$. Using Remark 3.2.1, we must prove that for all $l = 1, \dots, q$ there exist indices $i_0 \in N_l$ and $j_0 \in \{n_0 + 1, \dots, n\}$ such that $a(\bar{\mathbf{c}})_{i_0,j_0} < 0$.

From now, let $N_0 := \{1, \dots, n_0\}$, $\tilde{N} := \{n_0 + 1, \dots, n\}$ and $N := \{1, \dots, n\} = N_0 \cup \tilde{N}$.

First note that if $i \in N_0, j \in N$ are neighbouring indices then $a_{ij}(\bar{\mathbf{c}}) < 0$ for sufficiently small h. Namely, (3.3.18) holds by assumption (b), whence (3.3.22) yields $a_{ij}(\bar{\mathbf{c}}) < 0$ for sufficiently small h. Hence, it suffices to find $i_0 \in N_l$ and $j_0 \in \tilde{N}$ such that i_0, j_0 are neighbouring indices.

Now we observe that each N_l contains entire sets S_k^0, introduced in Assumptions 3.3.3. Namely, by item (ii) of Assumptions 3.3.3, the graph of all neighbouring indices in S_k^0 is connected, i.e. for all $i, j \in S_k^0$ there exists a chain (i, i_1), $(i_1, i_2), \dots, (i_r, j)$ of neighbouring indices (with all $i_m \in S_k^0$), whence by the above $a_{i,i_1}(\bar{\mathbf{c}}) < 0$, $a_{i_1,i_2}(\bar{\mathbf{c}}) < 0, \dots, a_{i_r,j}(\bar{\mathbf{c}}) < 0$. Therefore the entries of $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ with indices in S_k^0 belong to the same irreducible component, i.e. S_k^0 lies entirely in one of the sets N_l.

Consequently, it suffices to prove that for all $k = 1, \dots, r$ there exist indices $i_0 \in S_k^0$ and $j_0 \in \tilde{N}$ such that i_0, j_0 are neighbouring indices. By item (i) of Assumptions 3.3.3, there exists $i \in S_k^0$ and $j \in \tilde{S}_k$. Using that $i, j \in S_k$, by item (iii) of Assumptions 3.3.3, there exists a chain (i, i_1), $(i_1, i_2), \dots, (i_r, j)$ of neighbouring indices with all $i_m \in S_k$. If $i_1 \in \tilde{S}_k$ then we let $i_0 := i (\in S_k^0)$ and $j_0 := i_1 (\in \tilde{N})$. Otherwise, since $j \in \tilde{S}_k$, there exists a first index k in the chain such that $i_k \in S_k^0$ and $i_{k+1} \in \tilde{S}_k$, and then we let $i_0 := i_k (\in S_k^0)$ and $j_0 := i_{k+1} (\in \tilde{N})$. ■

By Theorem 3.2.1, we immediately obtain the corresponding *matrix maximum principle* (or *algebraic discrete maximum principle*):

Corollary 3.3.1 *Let the assumptions of Theorem 3.3.1 hold. For sufficiently small h, if $d_i \leq 0$ $(i = 1, ..., n_0)$ in (3.3.14) and $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ is the solution of (3.3.16), then*

$$\max_{i=1,\dots,n} c_i \leq \max\{0, \max_{i=n_0+1,\dots,n} c_i\}. \tag{3.3.23}$$

Remark 3.3.1 Assumption (c) of Theorem 3.3.1 follows in particular if Assumptions 3.3.2 are added to Assumptions 3.3.1 as done in Proposition 3.3.1, provided that the functions $g_h \in V_h$ in (3.3.11) are bounded in H-norm as $h \to 0$. (In practice, the usual choices for

g_h even produce $g_h \to g^*$ in H-norm.) In fact, in this case $\|u^h\|$ is bounded as $h \to 0$; then the continuity of M_R yields that $M_R(\|u^h\|)$ is bounded too.

Namely, using (3.3.7),

$$\langle A(u^h) - A(g_h), u^h - g_h\rangle = \langle A'(\theta u^h + (1-\theta)g_h)(u^h - g_h), u^h - g_h\rangle \geq \ m\,\|u^h - g_h\|^2$$

(where $\theta \in [0,1]$). From (3.3.10)

$$\langle A(u^h) - A(g_h), u^h - g_h\rangle = \langle f - A(g_h), u^h - g_h\rangle \tag{3.3.24}$$

and from (3.3.4)

$$\begin{aligned}\langle A(g^*) - A(g_h), u^h - g_h\rangle &= \langle A'(\theta g^* + (1-\theta)g_h)(g^* - g_h), u^h - g_h\rangle \\ &\leq M_A(\max\{\|g^*\|, \|g_h\|\})\,\|g^* - g_h\|\,\|u^h - g_h\| \end{aligned}\tag{3.3.25}$$

(where $\theta \in [0,1]$). From the above,

$$\begin{aligned}&m\,\|u^h - g_h\|^2 \leq \langle f - A(g^*), u^h - g_h\rangle + M_A(\max\{\|g^*\|, \|g_h\|\})\,\|g^* - g_h\|\,\|u^h - g_h\| \\ &\leq (\|f - A(g^*)\| + M_A(\max\{\|g^*\|, \|g_h\|\})\,\|g^* - g_h\|)\,\|u^h - g_h\|\,.\end{aligned}$$

Using the notation $\gamma := \sup_{h>0}\|g^* - g^h\|$, we obtain

$$\|u^h\| \leq \|g_h\| + \|u^h - g_h\| \leq \|g^*\| + \gamma + \tfrac{1}{m}\Big(\|f - A(g^*)\| + M_A(\|g^*\| + \gamma)\,\gamma\Big),$$

i.e. $\|u^h\|$ is bounded as $h \to 0$.

Remark 3.3.2 It is easy to see that Theorem 3.3.1 also holds for operators $A(u) = B(u)u + N(u)u + R(u)u$, if $B+N$ satisfies Assumption 3.3 (ii) and $N+R$ satisfies Assumption 3.3 (iii), further, if one substitutes $\langle B(u_h)\phi_j, \phi_i\rangle = \langle N(u_h)\phi_j, \phi_i\rangle = 0$ in (3.3.17) and $\sum_{j=1}^n \phi_j \in \ker B(u) \cap \ker N(u)$ in assumption (d) of Theorem 3.3.1; see [94]. We omit details for simplicity.

3.4 Discrete maximum principles for nonlinear elliptic problems

3.4.1 Nonlinear elliptic equations

Let us consider a nonlinear boundary value problem of the following type:

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) + q(x,u) \;=\; f(x) \quad \text{in } \Omega, \\ b(x,\nabla u)\frac{\partial u}{\partial \nu} + s(x,u) \;=\; \gamma(x) \quad \text{on } \Gamma_N, \\ u \;=\; g(x) \quad \text{on } \Gamma_D, \end{cases} \tag{3.4.1}$$

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

Assumptions 3.4.1.

(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

$$0 < \mu_0 \le b(x,\eta) \le \mu_1 \tag{3.4.2}$$

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 $\xi \in \mathbf{R}$

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

(A5) Either $\Gamma_D \ne \emptyset$, or q increases strictly and at least linearly at ∞ in the sense that

$$q(x,\xi) \ge c_1|\xi| - c_2(x) \tag{3.4.3}$$

(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.

The above assumptions ensure the well-posedness of (3.4.1) in $H^1_D(\Omega)$, as we have proved in [91], but we omit the proof for brevity. Now we follow [91] in developing the continuous and discrete maximum principles for problem (3.4.1).

(a) The continuous maximum principle

The classical form of the continuous maximum principle (CMP) states

$$\max_{\Omega} u \le \max\{0, \max_{\partial\Omega} u\} \tag{3.4.4}$$

under proper conditions if $Lu \le 0$ holds for an elliptic operator L with lower order terms, see e.g. [133]. However, as mentioned in the introduction, if mixed boundary conditions are imposed then the property (3.4.4) gives no essential information about the solution u, because it is not known on the whole boundary $\partial\Omega$. Hence we must clarify what to expect instead of (3.4.4) to get a computable bound on u.

We show that for mixed boundary conditions one can replace the r.h.s. of (3.4.4) by $\max\{0, \max_{\Gamma_D} u_h\}$ if γ satisfies a similar condition as f. Such a result has not been given before to our knowledge. An Alexandrov-Bakelman type estimate has been given for linear mixed boundary value problems in [32].

Theorem 3.4.1 *Let Assumptions 3.4.1 hold and let the weak solution u of problem (3.4.1) belong to $C^1(\Omega) \cap C(\overline{\Omega})$. If*

$$f(x) - q(x,0) \le 0,\ x \in \Omega, \quad \text{and} \quad \gamma(x) - s(x,0) \le 0,\ x \in \Gamma_N \tag{3.4.5}$$

(where the inequalities for the L^2 functions hold almost everywhere), then

$$\max_{\overline{\Omega}} u \le \max\{0, \max_{\Gamma_D} g\}. \tag{3.4.6}$$

In particular, if $\Gamma_D \neq \emptyset$ and $g \ge 0$, then $\max\limits_{\overline{\Omega}} u = \max\limits_{\Gamma_D} g$, and, if $\Gamma_D \neq \emptyset$ and $g \le 0$, or if $\Gamma_D = \emptyset$, then we have the nonpositivity property $\max\limits_{\overline{\Omega}} u \le 0$.

Proof. Let

$$r(x,\xi) := \begin{cases} \frac{q(x,\xi)-q(x,0)}{\xi}, & \text{if } \xi \neq 0, \\ \frac{\partial q}{\partial \xi}(x,0), & \text{if } \xi = 0, \end{cases} \qquad z(x,\xi) := \begin{cases} \frac{s(x,\xi)-s(x,0)}{\xi}, & \text{if } \xi \neq 0, \\ \frac{\partial s}{\partial \xi}(x,0), & \text{if } \xi = 0. \end{cases} \tag{3.4.7}$$

Here, by (A2), we have $q \in C^1(\overline{\Omega} \times \mathbf{R})$ and $s \in C^1(\Gamma_N \times \mathbf{R})$, therefore the functions r and z are continuous. Further, in view of (A4), we have

$$r(x,\xi) \ge 0, \qquad z(x,\xi) \ge 0. \tag{3.4.8}$$

We define

$$\begin{aligned} \tilde{a}(x) := b(x, \nabla u(x)), \quad \tilde{h}(x) := r(x, u(x)) \qquad (x \in \overline{\Omega}), \\ \tilde{k}(x) := z(x, u(x)) \qquad (x \in \Gamma_N). \end{aligned} \tag{3.4.9}$$

Then

$$\tilde{L}u := -\operatorname{div}\Big(\tilde{a}(x)\,\nabla u\Big) + \tilde{h}(x)u = f(x) - q(x,0) \le 0,$$

hence the nonlinear equation in (3.4.1) is recast to the setting of linear problems, and the usual techniques can be used. Using also the notations

$$\hat{f}(x) := f(x) - q(x,0) \qquad \text{and} \quad \hat{\gamma}(x) := \gamma(x) - s(x,0), \tag{3.4.10}$$

the weak formulation of problem (3.4.1) is given as

$$\int_\Omega \big(\tilde{a}\ \nabla u \cdot \nabla v + \tilde{h}uv\big)\,dx + \int_{\Gamma_N} \tilde{k}uv\,d\sigma = \int_\Omega \hat{f}v\,dx + \int_{\Gamma_N} \hat{\gamma}v\,d\sigma \qquad \forall v \in H^1_D(\Omega). \tag{3.4.11}$$

Now we let $M := \max\{0, \max\limits_{\Gamma_D} g\}$ and we introduce the piecewise C^1 function

$$v := \max\{u - M,\, 0\}.$$

Then we have $v \ge 0$ and $v_{|\Gamma_D} = 0$, further, $u(x) = v(x) + M$ for $x \in \Omega^+$ (where $v(x) \ge 0$) and $v(x) = 0$ otherwise. Hence, for this v the left-hand side of (3.4.11) satisfies

$$\int_\Omega \big(\tilde{a}\ \nabla u \cdot \nabla v + \tilde{h}uv\big)\,dx + \int_{\Gamma_N} \tilde{k}uv\,d\sigma = \int_{\Omega^+} \big(\tilde{a}\,|\nabla v|^2 + \tilde{h}(v+M)v\big)\,dx + \int_{\Gamma_N^+} \tilde{k}(v+M)v\,d\sigma \ge 0,$$

since the functions $\tilde{a}, \tilde{h}, \tilde{k}, v$ and the constant M are nonnegative. On the other hand, the assumptions $\hat{f} \le 0$, $\hat{\gamma} \le 0$ imply that for this v the right-hand side of (3.4.11) satisfies

$$\int\limits_{\Omega} \hat{f} v \, dx + \int\limits_{\Gamma_N} \hat{\gamma} v \, d\sigma \le 0,$$

hence, altogether we have

$$\int\limits_{\Omega} \left(\tilde{a} \, |\nabla v|^2 + \tilde{h}(v+M)v \right) dx + \int\limits_{\Gamma_N} \tilde{k}(v+M)v \, d\sigma = 0.$$

Here $\tilde{a}$ has a positive minimum in view of (A3), hence $|\nabla v| = 0$, i.e., v is constant and as seen above it is nonnegative: say,

$$v(x) \equiv c \ge 0 \qquad \text{on } \overline{\Omega}.$$

If $c = 0$ then $u \le M$ on Ω, i.e., (3.4.6) is proved. If $c > 0$ then $\Gamma_D = \emptyset$ (otherwise property $v_{|\Gamma_D} = 0$ would yield a contradiction). Then $M = 0$ and $v = \max\{u, 0\}$, hence $v \equiv c$ implies $u \equiv c$. Therefore, (3.4.1) reduces to $q(x,c) \equiv f(x)$ in Ω and $s(x,c) \equiv \gamma(x)$ on $\partial\Omega$. Then (3.4.5) implies $q(x,c) \le q(x,0)$ and $s(x,c) \le s(x,0)$ with $c > 0$. This is impossible since, by (A5), either q or s is strictly increasing. Altogether, we obtain that $c = 0$ and hence (3.4.6) holds. ■

In the special case $q \equiv 0$ and $s \equiv 0$, equality holds without assuming $g \ge 0$:

Theorem 3.4.2 *Consider problem (3.4.1) with* $q \equiv 0$, $s \equiv 0$ *under the assumptions of Theorem 3.4.1. That is, (A1)–(A3) are satisfied,* $u \in C^1(\Omega) \cap C(\overline{\Omega})$, *and (3.4.5) now takes the form* $f(x) \le 0,\ x \in \Omega$ *and* $\gamma(x) \le 0,\ x \in \Gamma_N$. *Then*

$$\max_{\overline{\Omega}} u = \max_{\Gamma_D} g. \tag{3.4.12}$$

Proof. If $\max\limits_{\Gamma_D} g \ge 0$ then (3.4.6) implies (3.4.12). Let $\max\limits_{\Gamma_D} g < 0$, say, $\max\limits_{\Gamma_D} g = -K$ with some $K > 0$. Then the function $w := u + K$ satisfies the same mixed problem with right-hand sides f, γ and $g + K$, respectively, hence Theorem 3.4.1 is valid for this problem as well, and (3.4.6) for w yields $\max\limits_{\overline{\Omega}} w \le \max\{0, \max\limits_{\Gamma_D} (g+K)\} = 0$. Then

$$\max_{\overline{\Omega}} u \le -K = \max_{\Gamma_D} g. \qquad ■$$

Remark 3.4.1 We note that the corresponding minimum principles and nonnegativity property hold if the sign conditions in (3.4.5) are reversed. Further, the analogues of the above theorems hold in the same way for the case $u \in H^1(\Omega)$, i.e., with no regularity assumption on the weak solution, provided that g is bounded on Γ_D. Then $\max u$ and $\max g$ are replaced by $\operatorname{ess\,sup} u$ and $\operatorname{ess\,sup} g$, respectively.

(b) The discrete maximum principle

First we briefly summarize the FE discretization of problem (3.4.1). In what follows, we assume that Ω is a polytopic domain. We define the finite element discretization of our problem using simplicial elements and continuous piecewise linear basis functions. The symbol $\mathcal{T}_h$ stands for a conforming triangulation of $\overline{\Omega}$ into tetrahedra, whose vertices are $B_1, ..., B_{\bar{n}}$. When a family of meshes are considered then h is proportional to the maximal element diameter. We denote by $\phi_1, ..., \phi_{\bar{n}}$ the piecewise linear continuous basis functions defined in a standard way, i.e., $\phi_i(B_j) = \delta_{ij}$ for $i, j = 1, ..., \bar{n}$, where δ_{ij} is the Kronecker symbol. Let V_h denote the finite element subspace spanned by the above basis functions:

$$V_h = \text{span}\{\phi_1, ..., \phi_{\bar{n}}\} \ \subset H^1(\Omega).$$

Now, let $n < \bar{n}$ be such that $B_1, ..., B_n$ are the vertices that lie in Ω or on Γ_N, and let $B_{n+1}, ..., B_{\bar{n}}$ be the vertices that lie on Γ_D. Then the basis functions $\phi_1, ..., \phi_n$ satisfy the homogeneous Dirichlet boundary condition on Γ_D, i.e., $\phi_i \in H^1_D(\Omega)$. We define

$$V_h^0 = \text{span}\{\phi_1, ..., \phi_n\} \ \subset H^1_D(\Omega).$$

Further, let $g_h \in V_h$ be the projection of g^* into the subspace $span\{\varphi_{n+1}, \dots, \varphi_{\bar{n}}\}$.

The FEM solution is defined in the usual way by setting the basis functions as test functions in the weak form. Rewriting this using (3.4.7) and (3.4.10), we obtain

$$\int\limits_\Omega \big[b(x, \nabla u_h)\ \nabla u_h \cdot \nabla v_h + \ r(x, u_h) u_h v_h\big]\, dx \ + \int\limits_{\Gamma_N} z(x, u_h) u_h v_h \, d\sigma = \int\limits_\Omega \hat{f} v_h \, dx \ + \int\limits_{\Gamma_N} \hat{\gamma} v_h \, d\sigma \tag{3.4.13}$$

$(\forall v_h \in V_h^0)$. Now we turn to the nonlinear algebraic system corresponding to (3.4.13). We look for the coefficients $c_1, \dots, c_{\bar{n}}$ of u_h. For any $\bar{\mathbf{c}} = (c_1, ..., c_{\bar{n}})^T \in \mathbf{R}^{\bar{n}}$, $i = 1, ..., n$ and $j = 1, ..., \bar{n}$, we set $\mathbf{c} = \{c_j\}$, $j = 1, ..., n$, and $\tilde{\mathbf{c}} = \{c_j\}$, $j = n+1, ..., \bar{n}$, further,

$$b_{ij}(\bar{\mathbf{c}}) = \int_\Omega b(x, \sum_{k=1}^{\bar{n}} c_k \nabla\phi_k)\ \nabla\phi_j \cdot \nabla\phi_i \, dx, \qquad r_{ij}(\bar{\mathbf{c}}) = \int_\Omega r(x, \sum_{k=1}^{\bar{n}} c_k \phi_k)\ \phi_j \phi_i \, dx,$$

$$z_{ij}(\bar{\mathbf{c}}) = \int_{\Gamma_N} z(x, \sum_{k=1}^{\bar{n}} c_k \phi_k)\ \phi_j \phi_i \, d\sigma, \qquad d_i(\bar{\mathbf{c}}) = \int_\Omega \hat{f} \phi_i \, dx \ + \int_{\Gamma_N} \hat{\gamma} \phi_i \, d\sigma\,,$$

$$a_{ij}(\bar{\mathbf{c}}) = b_{ij}(\bar{\mathbf{c}}) + r_{ij}(\bar{\mathbf{c}}) + z_{ij}(\bar{\mathbf{c}}).$$

Setting (3.3.13) and $v_h = \phi_i$ into (3.4.13), we obtain the $n \times \bar{n}$ system of algebraic equations

$$\sum_{j=1}^{\bar{n}} a_{ij}(\bar{\mathbf{c}})\, c_j = d_i, \quad i = 1, ..., n. \tag{3.4.14}$$

Using the obvious notations, system (3.4.14) turns into $\mathbf{A}(\bar{\mathbf{c}})\mathbf{c} + \tilde{\mathbf{A}}(\bar{\mathbf{c}})\tilde{\mathbf{c}} = \mathbf{d}$.

In order to obtain a system with a square matrix, we enlarge our system to an $\bar{n} \times \bar{n}$ one. Namely, since $u_h = g_h$ on Γ_D, the coordinates c_i with $n + 1 \le i \le \bar{n}$ satisfy automatically

$c_i = g_i$, i.e., $\tilde{\mathbf{c}} = \tilde{\mathbf{g}}$, where $\tilde{\mathbf{g}} = \{g_j\}$, $j = n+1, ..., \bar{n}$. That is, we can replace (3.4.14) by the equivalent system

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

Theorem 3.4.3 *Let Assumptions 3.4.1 hold, and let us consider a family of simplicial triangulations $\mathcal{T}_h$ $(h > 0)$ satisfying the following property: for any $i = 1, ..., n$, $j = 1, ..., \bar{n}$ $(i \neq j)$*

$$\nabla\phi_i \cdot \nabla\phi_j \le -\frac{\sigma_0}{h^2} < 0 \tag{3.4.15}$$

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

(1) Let the simplicial triangulations $\mathcal{T}_h$ be regular, i.e., there exist constants $m_1, m_2 > 0$ such that for any $h > 0$ and any simplex $T_h \in \mathcal{T}_h$

$$m_1 h^d \le meas(T_h) \le m_2 h^d \tag{3.4.16}$$

(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 3.1.3.

(2) More generally, the above statement is also valid if the triangulations $\mathcal{T}_h$ are only quasi-regular in the following sense: the left-hand side of (3.4.16) is replaced by

$$c_1 h^\gamma \le meas(T_h)\,, \tag{3.4.17}$$

where $\gamma \ge d$ satisfies

$$\begin{aligned} &2 \le \gamma < 3 \quad \text{if } d = 2, \qquad 3 \le \gamma < \min\{\tfrac{12}{p_1-2},\ 5 - \tfrac{p_2}{2}\} \quad \text{if } d = 3, \\ &d \le \gamma < \min\{\tfrac{4d}{(p_1-2)(d-2)},\ 3 + \tfrac{(4-p_2)(d-2)}{2}\} \quad \text{if } d > 3 \end{aligned} \tag{3.4.18}$$

with p_1, p_2 from assumption (A4) for problem (3.4.1).

Proof. We enclose our direct proof from [91]. Although this theorem will be extended to systems in the next chapter, those results do not entirely cover the present case of (3.4.1) since (for technical simplicity) there we will only consider Neumann boundary conditions on Γ_N (i.e. $s \equiv 0$), constant α_1, α_2 and will only study 2 or 3 dimensions.

We have $a_{ij}(\bar{\mathbf{c}}) = \int\limits_\Omega \Big[b(x, \nabla u_h)\ \nabla\phi_i \cdot \nabla\phi_j\ + r(x, u_h)\ \phi_i\phi_j \Big]\, dx + \int\limits_{\Gamma_N} z(x, u_h)\ \phi_i\phi_j\, d\sigma$ for any $i = 1, ..., n$. We now prove the properties (i)-(v) of Definition 3.1.3 in the more general case (2); the conditions (3.4.18) are only used in part (ii).

(i) From assumptions $b \ge \mu_0 > 0$, $r, z \ge 0$ we have $a_{ii}(\bar{\mathbf{c}}) \ge \mu_0 \int\limits_\Omega |\nabla\phi_i|^2\, dx\ > 0$.

(ii) Let $i = 1, ..., n$, $j = 1, ..., \bar{n}$ with $i \neq j$ and let Ω_{ij} denote the interior of *supp* $\phi_i \cap$ *supp* ϕ_j. If $\Omega_{ij} = \emptyset$ then $a_{ij}(\bar{\mathbf{c}}) = 0$. If $\Omega_{ij} \neq \emptyset$ then (3.4.15) and the fact $0 \le \phi_i \le 1, i = 1, ..., \bar{n}$, imply

$$a_{ij}(\bar{\mathbf{c}}) \le -\frac{\sigma_0}{h^2}\,\mu_0\,\text{meas}\,(\Omega_{ij}) + \int\limits_{\Omega_{ij}} r(x, u_h)\, dx + \int\limits_{\Gamma_{ij}} z(x, u_h)\, d\sigma, \tag{3.4.19}$$

using notation $\Gamma_{ij} = \Gamma_N \cap \overline{\Omega}_{ij}$. Here, from Assumption (A4),

$$\int\limits_{\Omega_{ij}} r(x,u_h)\,dx = \int\limits_{\Omega_{ij}} \frac{\partial q}{\partial \xi}(x,\theta u_h)\,dx \le \int\limits_{\Omega_{ij}} \big(\alpha_1(x)+\beta|\theta u_h|^{p_1-2}\big)\,dx \le \int\limits_{\Omega_{ij}} \alpha_1(x)\,dx+\beta\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx$$

(where we had some $\theta=\theta(x)\in[0,1]$), and in just the same way we have

$$\int\limits_{\Gamma_{ij}} z(x,u_h)\,d\sigma \le \int\limits_{\Gamma_{ij}} \alpha_2(x)\,d\sigma + \beta\int\limits_{\Gamma_{ij}} |u_h|^{p_2-2}\,d\sigma.$$

Now we can estimate the integrals $\int_{\Omega_{ij}} |u_h|^{p_1-2}\,dx$ and $\int_{\Gamma_{ij}} |u_h|^{p_2-2}\,d\sigma$ as follows. We define $p^* := \frac{2d}{d-2}$ and $p^{**} := \frac{2(d-1)}{d-2}$ if $d\ge 3$, and $p^* := p^{**} := +\infty$ if $d=2$. Then the Sobolev embedding estimates

$$\|v\|_{L^{p^*}(\Omega)} \le k_1\|v\|_1, \quad \|v\|_{L^{p^{**}}(\Gamma_N)} \le k_2\|v\|_1, \quad v\in H^1(\Omega) \tag{3.4.20}$$

hold with constants $k_1,k_2>0$, where $\|v\|_1 = \|v\|_{H^1(\Omega)}$ (see [1]). Assume for a while that $p_1,p_2>2$ and let us fix real numbers r and t satisfying

$$\frac{\gamma}{2} < r \le \frac{p^*}{p_1-2}, \qquad \frac{d-1}{d+1-\gamma} < t \le \frac{p^{**}}{p_2-2}\,. \tag{3.4.21}$$

Such numbers exist since for $d\ge 3$, by (3.4.18),

$$\gamma < \frac{2p^*}{p_1-2} \quad \text{and} \quad \gamma < 3+\frac{(4-p_2)(d-2)}{2} = d+1+\frac{(2-p_2)(d-2)}{2} = d+1-\frac{(p_2-2)(d-1)}{p^{**}}\,.$$

Further, $\gamma\ge 2$ implies $r\ge 1$ and $t\ge 1$. If $\frac{1}{r}+\frac{1}{s} = \frac{1}{t}+\frac{1}{l} = 1$ then Hölder's inequality implies

$$\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx \le \|1\|_{L^s(\Omega_{ij})} \Big\| |u_h|^{p_1-2} \Big\|_{L^r(\Omega_{ij})} = meas(\Omega_{ij})^{1/s}\,\|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega_{ij})}\,. \tag{3.4.22}$$

Here $(p_1-2)r\le p^*$ and (3.4.20) imply

$$\|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega_{ij})} \le \|u_h\|^{p_1-2}_{L^{(p_1-2)r}(\Omega)} \le const.\cdot \|u_h\|^{p_1-2}_{L^{p^*}(\Omega)} \le const.\cdot \|u_h\|_1^{p_1-2}\,.$$

Owing to the basic FEM convergence result [34], we have $\|u_h\|_1 \to \|u^*\|_1$, where u^* is the exact weak solution of our problem. Hence if h is less than some fixed h_0 then (3.4.22) finally turns into

$$\int\limits_{\Omega_{ij}} |u_h|^{p_1-2}\,dx \le K_1\; meas(\Omega_{ij})^{1/s} \tag{3.4.23}$$

with some constant $K_1>0$ independent of h. In just the same way we obtain

$$\int\limits_{\Gamma_{ij}} |u_h|^{p_2-2}\,dx \le K_2\; meas(\Gamma_{ij})^{1/l}\,. \tag{3.4.24}$$

Finally, if p_1 or p_2 equals 2 then the corresponding equality (3.4.23) or (3.4.24) holds with $s = 1$ or $l = 1$, respectively.

The integrals of $\alpha_1(x)$ and $\alpha_2(x)$ can be estimated with Hölder's inequality similarly to (3.4.22) by letting $\frac{2}{d} + \frac{1}{s'} = \frac{1}{d-1} + \frac{1}{l'} = 1$:

$$\int\limits_{\Omega_{ij}} \alpha_1(x)\, dx \le K_3 \ meas(\Omega_{ij})^{1/s'}, \qquad \int\limits_{\Gamma_{ij}} \alpha_2(x)\, d\sigma \le K_4 \ meas(\Gamma_{ij})^{1/l'}$$

with $K_3 = \|\alpha_1\|_{L^{d/2}(\Omega)}$ and $K_4 = \|\alpha_2\|_{L^{d-1}(\Gamma_N)}$.

Substituting all the estimates in (3.4.19), we obtain

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) \le \quad & -\frac{\sigma_0\mu_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij}) + \beta K_1 \ meas(\Omega_{ij})^{1/s} + \ K_3 \ meas(\Omega_{ij})^{1/s'} \\ & + \beta K_2 \ meas(\Gamma_{ij})^{1/l} + K_4 \ meas(\Gamma_{ij})^{1/l'} \,. \end{aligned} \tag{3.4.25}$$

We can write

$$a_{ij}(\bar{\mathbf{c}}) \le \ A_1^{ij}(h) + A_2^{ij}(h) + A_3^{ij}(h) + A_4^{ij}(h)$$

where, with suitable constants C_0, C_1, C_2, C_3, $C_4 > 0$ independent of h and i, j,

$$A_1^{ij}(h) := -\frac{C_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij}) + C_1 \, meas(\Omega_{ij})^{1/s}, \qquad A_2^{ij}(h) := -\frac{C_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij}) + C_2 \, meas(\Gamma_{ij})^{1/l},$$

$$A_3^{ij}(h) := -\frac{C_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij}) + C_3 \, meas(\Omega_{ij})^{1/s'}, \qquad A_4^{ij}(h) := -\frac{C_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij}) + C_4 \, meas(\Gamma_{ij})^{1/l'} \,.$$

We verify that for small enough h we have $A_k^{ij}(h) < 0$ $(k = 1, 2, 3, 4)$.

Using $\frac{1}{r} + \frac{1}{s} = 1$ and (3.4.55), we have

$$A_1^{ij}(h) = meas(\Omega_{ij})^{1/s} \Big(-\frac{C_0}{h^2}\,\mathrm{meas}\,(\Omega_{ij})^{1/r} \ + \ C_1 \Big) \le meas(\Omega_{ij})^{1/s} \Big(-C_5\, h^{-2+(\gamma/r)} \ + \ C_1 \Big).$$

Since (3.4.21) implies $\frac{\gamma}{r} < 2$, the term in brackets tends to $-\infty$ as $h \to 0$ and hence $A_1^{ij}(h) < 0$ for small h.

Using (3.4.55) again and the fact that $meas(\Gamma_{ij}) \le const.\cdot h^{d-1}$ (since h is the diameter of the simplices and Γ_{ij} lies on the $(d-1)$-dimensional boundary), we have

$$A_2^{ij}(h) \le -C_6\, h^{\gamma-2} + \ C_7\, h^{\frac{d-1}{l}}.$$

Since (3.4.21) implies $1 - \frac{1}{l} = \frac{1}{t} < \frac{d+1-\gamma}{d-1} = 1 - \frac{\gamma-2}{d-1}$, we obtain $\frac{d-1}{l} > \gamma - 2$, i.e. the second term tends to 0 faster and hence $A_2^{ij}(h) < 0$ for small h.

The terms $A_3^{ij}(h)$ and $A_4^{ij}(h)$ can be handled similarly, since s' and l' satisfy the same estimates as s and l. Namely, we have $\frac{d}{2} = \frac{p^*}{p^*-2}$ and $d-1 = \frac{p^{**}}{p^{**}-2}$, hence by substituting $\frac{d}{2}$ and $d-1$ for r and t, respectively, we obtain that (3.4.21) holds in the special case $p_1 = p^*$ and $p_2 = p^{**}$. Owing to the condition $\frac{2}{d} + \frac{1}{s'} = \frac{1}{d-1} + \frac{1}{l'} = 1$, the numbers s' and l' play the same role as s and l and therefore the above estimates on $A_1^{ij}(h)$ and $A_2^{ij}(h)$ can be repeated for $A_3^{ij}(h)$ and $A_4^{ij}(h)$.

Altogether, we obtain that for small enough h, $A_k^{ij}(h) < 0$ $(k = 1, 2, 3, 4)$, that is, there exists $h_0 > 0$ such that

$$a_{ij}(\bar{\mathbf{c}}) < 0 \tag{3.4.26}$$

for all $h \le h_0$ and all $i \neq j$ with $\Omega_{ij} \neq \emptyset$.

(iii) For any $i = 1, ..., n$,

$$\sum_{j=1}^{\bar{n}} a_{ij}(\bar{\mathbf{c}}) = \int\limits_{\Omega} \Big[b(x, \nabla u_h)\ \nabla\phi_i \cdot \nabla(\sum_{j=1}^{\bar{n}} \phi_j) + r(x, u_h)\ \phi_i(\sum_{j=1}^{\bar{n}} \phi_j)\Big]\, dx \tag{3.4.27}$$

$$+ \int\limits_{\Gamma_N} z(x, u_h)\ \phi_i(\sum_{j=1}^{\bar{n}} \phi_j)\, d\sigma = \int\limits_{\Omega} r(x, u_h)\, \phi_i\, dx + \int\limits_{\Gamma_N} z(x, u_h)\, \phi_i\, d\sigma \ge 0,$$

using the fact that $\sum_{j=1}^{\bar{n}} \phi_j \equiv 1$ and r, z, ϕ_i are nonnegative.

(iv) Assume for contradiction that $\sum\limits_{j=1}^{n} a_{ij}(\bar{\mathbf{c}}) = 0$ for all $i = 1, ..., n$. This means that $\mathbf{A}(\bar{\mathbf{c}})$ carries the n-tuple of ones $\{1, ..., 1\}$ into the zero vector. This is impossible since $\mathbf{A}(\bar{\mathbf{c}})$ is symmetric and positive definite, and hence one-to-one.

(v) For any $i, j = 1, ..., n$ with $i \neq j$, let us pick a sequence of neighbouring vertices B_{i_k} $(k = 1, ..., s)$ in Ω that connect B_i with B_j (i.e. $i_0 = i$ and $i_s = j$). Here (3.4.26) shows that $a_{i_k, i_{k+1}}(\bar{\mathbf{c}}) < 0$, hence by Definition 3.1.1, $\mathbf{A}(\bar{\mathbf{c}})$ is irreducible. ■

Now we can derive the discrete maximum principle. By Theorem 3.4.1, it will reflect a real property of the exact solution.

Theorem 3.4.4 *Let the conditions of Theorem 3.4.3 hold, and let*

$$f(x) - q(x, 0) \le 0,\ x \in \Omega, \qquad \textit{and} \qquad \gamma(x) - s(x, 0) \le 0,\ x \in \Gamma_N. \tag{3.4.28}$$

Then

$$\max_{\overline{\Omega}} u_h \le \max\{0, \max_{\Gamma_D} g_h\}. \tag{3.4.29}$$

In particular, if $\Gamma_D \neq \emptyset$ *and* $g \ge 0$ *then* $\max\limits_{\overline{\Omega}} u_h = \max\limits_{\Gamma_D} g_h$, *and if* $\Gamma_D \neq \emptyset$ *and* $g \le 0$, *or if* $\Gamma_D = \emptyset$, *then we have the nonpositivity property* $u_h \le 0$ *on* Ω.

Proof. We can apply Theorem 3.1.1 with $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ and $\bar{n}$ substituted for $\bar{\mathbf{A}}$ and $n + m$, respectively, since $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ is of generalized nonnegative type in the sense of Definition 3.1.3. Since, by (3.4.28), $\mathbf{d} \le 0$, we get $\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} \le 0$ and hence Theorem 3.1.1 yields $\max\limits_{i=1,...,\bar{n}} c_i \le \max\{0, \max\limits_{i=n+1,...,\bar{n}} c_i\}$. Since $c_i = g_i$ for all $i = n+1, ..., \bar{n}$, we obtain $\max\limits_{i=1,...,\bar{n}} c_i \le \max\{0, \max\limits_{i=n+1,...,\bar{n}} g_i\}$, which implies (3.4.29) for the considered piecewise linear basis functions. ■

One can verify in the same way the *minimum principle* for problem (3.4.1). We only formulate the special case of *discrete nonnegativity*:

Theorem 3.4.5 *Let the conditions of Theorem 3.4.3 hold, and let* $f(x) - q(x,0) \geq 0,\ x \in \Omega,$ *and* $\gamma(x) - s(x,0) \geq 0,\ x \in \Gamma_N$*. If* $\Gamma_D \neq \emptyset$ *and* $g \geq 0$*, or if* $\Gamma_D = \emptyset$*, then*

$$u_h \geq 0 \quad \text{on } \Omega.$$

In the special case $q \equiv 0$ and $s \equiv 0$, equality $\max_{\overline{\Omega}} u_h = \max_{\Gamma_D} g_h$ holds without assuming $g \geq 0$. This is the discrete counterpart of Theorem 3.4.2. We formulate this for both the maximum and minimum principles. Moreover, the strict negativity in (3.4.15) can be replaced by a weaker nonnegativity condition, and no special condition on the mesh like (3.4.17) needs to be assumed.

Theorem 3.4.6 *Let us consider the following special case of problem (3.4.1):*

$$\begin{cases} -\operatorname{div}\Big(b(x,\nabla u)\,\nabla u\Big) = f(x) \quad \text{in } \Omega, \\ b(x,\nabla u)\frac{\partial u}{\partial \nu} = \gamma(x) \quad \text{on } \Gamma_N, \qquad u = g(x) \quad \text{on } \Gamma_D. \end{cases} \tag{3.4.30}$$

Let (A1)–(A3) hold and $\Gamma_D \neq \emptyset$*, further, let the triangulation* $\mathcal{T}_h$ *satisfy the following property: for any* $i = 1,...,n,\ j = 1,...,\bar{n}\ (i \neq j)$

$$\nabla\phi_i \cdot \nabla\phi_j \leq 0. \tag{3.4.31}$$

(1) If $f \leq 0$ *and* $\gamma \leq 0$*, then* $\max\limits_{\overline{\Omega}} u_h = \max\limits_{\Gamma_D} g_h$*.*

(2) If $f \geq 0$ *and* $\gamma \geq 0$*, then* $\min\limits_{\overline{\Omega}} u_h = \min\limits_{\Gamma_D} g_h$*.*

(3) If $f = 0$ *and* $\gamma = 0$*, then the ranges of* u_h *and* g_h *coincide, i.e., we have* $[\min\limits_{\overline{\Omega}} u_h, \max\limits_{\overline{\Omega}} u_h] = [\min\limits_{\Gamma_D} g_h, \max\limits_{\Gamma_D} g_h]$ *for the corresponding intervals.*

Proof. Similarly to that of Theorem 3.4.3. The main difference arises in proving property (ii), i.e., $a_{ij}(\bar{\mathbf{c}}) \leq 0$, where (3.4.31) is enough, since the assumptions $q \equiv 0$ and $s \equiv 0$ imply $r \equiv 0$ and $z \equiv 0$. ■

Remark 3.4.2 (a) Note that the values $\nabla\phi_i \cdot \nabla\phi_j$ are constant on each element, hence conditions (3.4.15) and (3.4.31) are not difficult to check. Moreover, these conditions have a nice geometric interpretation, which will be discussed in detail in the next subsection.

(b) Condition (3.4.15) can be relaxed such that $\nabla\phi_i \cdot \nabla\phi_j$ need not be negative on each element, see later (3.4.119) and the discussion afterwards.

Sufficient conditions and their geometric meaning. In view of well-known results, the conditions (3.4.15) and (3.4.31) have nice geometric interpretations. Namely, in order to satisfy condition (3.4.15) in the case of a simplicial mesh, it is sufficient if the employed mesh is acute, and similarly, condition (3.4.31) is satisfied if the employed mesh is nonobtuse [107]. We note that these conditions are sufficient but not necessary: as shown by paragraph (b) of Remark 3.4.2, the DMP may still hold if some obtuse interior angles occur in the simplices of the meshes. This is analogous to the case of linear problems [104, 157].

These geometric conditions need special attention when we apply a global refinement of the initial mesh using some refinement technique. Then we must take care that the refined mesh preserves the desired acuteness or nonobtuseness property. In the two-dimensional case, using the standard “2D red refinement” [103], we obtain a mesh consisting only of acute or nonobtuse triangular elements if the initial mesh had only acute or nonobtuse triangles, respectively. If we consider a tetrahedral mesh, the task is far from being trivial since in general it is not possible to refine any tetrahedron into eight subtetrahedra similar to it using “3D red refinement” (cf. [103]). A new technique, the so-called “3D yellow refinement” was developed in [102], which allows a global refinement of a nonobtuse tetrahedral mesh so that the resulting (conforming) mesh preserves the property of nonobtuseness.

In order to save computer memory, it is often desirable to perform only local refinements of tetrahedral meshes near edges and vertices, or where the true solution or its derivatives have singularities, e.g. near Fichera corners. Algorithms allowing to do that with a preservation of nonobtuseness have been constructed and tested in detail in [27].

Condition (3.4.15) is altogether still rather strong, and will be relaxed at the end of section 3.4.2.

3.4.2 Nonlinear cooperative elliptic systems

Now we consider various systems, in which the lower order coupling terms are cooperative and form a weakly diagonally dominant system. We impose these conditions because they appear in the underlying continuous maximum principle, which we will also address briefly.

Whereas in the case of a single equation the DMP was be proved directly, in the case of systems the Hilbert space setting will be exploited to derive the results. This framework helps us in structuring the proof procedure under the technical difficulties caused by the more compound form of the FEM and the complications with the lack of irreducibility. We follow [92, 94].

(a) Systems with nonlinear coefficients

Formulation of the problem. First we consider nonlinear elliptic systems of the form

$$\left.\begin{aligned} -\mathrm{div}\Big(b_k(x,u,\nabla u)\,\nabla u_k\Big) + \sum_{l=1}^{s} V_{kl}(x,u,\nabla u)\,u_l &= f_k(x) \quad \text{a.e. in } \Omega,\\ b_k(x,u,\nabla u)\tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N,\\ u_k &= g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k=1,\dots,s) \tag{3.4.32}$$

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 3.4.7.

(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 $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \neq \emptyset$.

(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 $k = 1, \dots, s$.

(iv) (Cooperativity.) We have

$$V_{kl} \leq 0 \qquad (k, l = 1, \dots, s,\ k \neq l). \tag{3.4.33}$$

(v) (Weak diagonal dominance.) We have

$$\sum_{l=1}^{s} V_{kl} \geq 0 \qquad (k = 1, \dots, s). \tag{3.4.34}$$

(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 $g^*_k \in H^1(\Omega)$.

Remark 3.4.3 Assumptions (3.4.33)–(3.4.34) imply $V_{kk} \geq 0 \quad (k = 1, \dots, s)$.

Let us define the Sobolev space $H^1_D(\Omega) := \{z \in H^1(\Omega) : z_{|\Gamma_D} = 0\}$. The weak formulation of problem (3.4.32) then reads as follows: find $u \in H^1(\Omega)^s$ such that

$$\langle A(u), v\rangle = \langle \psi, v\rangle \qquad (\forall v \in H^1_D(\Omega)^s) \tag{3.4.35}$$

$$\text{and} \quad u - g^* \in H^1_D(\Omega)^s, \qquad \text{where} \tag{3.4.36}$$

$$\langle A(u), v\rangle = \int_\Omega \Big(\sum_{k=1}^{s} b_k(x, u, \nabla u)\, \nabla u_k \cdot \nabla v_k + \sum_{k,l=1}^{s} V_{kl}(x, u, \nabla u)\, u_l\, v_k \Big) \tag{3.4.37}$$

for given $u = (u_1, \dots, u_s) \in H^1(\Omega)^s$ and $v = (v_1, \dots, v_s) \in H^1_D(\Omega)^s$, further,

$$\langle \psi, v\rangle = \int_\Omega \sum_{k=1}^{s} f_k v_k + \int_{\Gamma_N} \sum_{k=1}^{s} \gamma_k v_k \tag{3.4.38}$$

for given $v = (v_1, \dots, v_s) \in H^1_D(\Omega)^s$, and $g^* := (g^*_1, \dots, g^*_s)$.

On continuous maximum principles. The extension of the CMP from elliptic equations to systems has attracted much interest, and has been achieved in different forms (coordinatewise or for $|u|$), but under strong restrictions only. The main class of problems where a CMP is generally valid is that of cooperative systems, and in addition, one often also assumes weak diagonal dominance of V. This is why we also impose these conditions.

Important results of this type are found e.g. in [40, 110, 132, 139], and some extensions to non-cooperative systems are also known, see [31] and references therein. However, for cooperative systems, no CMP is known at the generality of (3.4.32) to our knowledge. It is not our goal to complete this background, however, for Dirichlet problems it is easy to derive a CMP in a form analogous to (3.4.6), based on a linear result [132].

Proposition 3.4.1 *Let Assumptions 3.4.7 hold and u be a classical solution of* (3.4.32) *under assumption $\Gamma_D = \partial\Omega$. If, for all $k = 1, \dots, s$, we have $f_k \leq 0$ on Ω and $\gamma_k \leq 0$ on Γ_N, then*

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k \leq \max_{k=1,\dots,s} \max\{0, \max_{\partial\Omega} g_k\}. \tag{3.4.39}$$

Proof. Let us define the bounded functions $a_k(x) := b_k(x, u(x), \nabla u(x))$ and $Q_{kl}(x) := V_{kl}(x, u(x), \nabla u(x))$, and consider the linear system

$$\left.\begin{aligned} -\mathrm{div}\left(a_k(x)\,\nabla z_k\right) + \sum_{l=1}^{s} Q_{kl}(x)\, z_l \;&=\; f_k(x) \quad \text{a.e. in } \Omega, \\ z_k \;&=\; g_k(x) \quad \text{a.e. on } \partial\Omega \end{aligned}\right\} \qquad (k = 1, \ldots, s). \tag{3.4.40}$$

By definition, u is a solution of (3.4.40). Here (for all x and $k \neq l$) $a_k(x) \geq m > 0$ and $Q_{kl}(x) \leq 0$, $\sum\limits_{l=1}^{s} Q_{kl}(x) \geq 0$. Hence [132, Th. 3.4] states that if z is a solution of (3.4.40) and a component z_k attains a nonnegative maximum in Ω, then z_k is constant. This property then holds for u. Let $K := \max\limits_{k=1,\ldots,s} \max\limits_{\overline{\Omega}} u_k$. If $K \leq 0$ then (3.4.39) holds. Now let $K > 0$. Then $K = \max\limits_{\overline{\Omega}} u_{k^*}$ for some index k^*, and u_{k^*} attains a nonnegative maximum. By the cited property, u_{k^*} must attain this maximum on $\partial\Omega$, hence $K = \max\limits_{\partial\Omega} u_{k^*} = \max\limits_{\partial\Omega} g_{k^*}$. Thus

$$\max_{k=1,\ldots,s} \max_{\overline{\Omega}} u_k = K = \max_{\partial\Omega} g_{k^*} = \max\{0, \max_{\partial\Omega} g_{k^*}\} \leq \max_{k=1,\ldots,s} \max\{0, \max_{\partial\Omega} g_k\}. \qquad \blacksquare$$

We also enclose a proof for mixed problems under another additional assumption, using a suitable combination of the proofs in [91, 151] with diagonal dominance.

Proposition 3.4.2 *Let Assumptions 3.4.7 hold and $u \in H^1(\Omega)^s$ be a weak solution of system* (3.4.32)*, such that $u \in C(\overline{\Omega})$. Assume further that V is also weakly diagonally dominant w.r.t. columns, i.e.* (3.4.34) *also holds for summation w.r.t. the index k. If, for all $k = 1, \ldots, s$, we have $f_k \leq 0$ on Ω and $\gamma_k \leq 0$ on Γ_N, then* (3.4.39) *holds.*

Proof. Let $M := \max\limits_{k=1,\ldots,s} \max\{0, \max\limits_{\Gamma_D} g_k\}$, and introduce the functions

$$v_k^+ := \max\{u_k - M,\, 0\} \qquad (k = 1, \ldots, s).$$

Then $u_k \in H^1(\Omega)$ implies $v_k^+ \in H^1(\Omega)$ (see e.g. [62]), and $v^+_{k|\Gamma_D} = 0$, hence $v^+ \in H^1_D(\Omega)^s$ and we can set $v := v^+$ into (3.4.35). Consider first the left-hand side (3.4.37) of (3.4.35):

$$\langle A(u), v^+\rangle = \int_\Omega \sum_{k=1}^{s} b_k(x, u, \nabla u)\, \nabla u_k \cdot \nabla v_k^+ + \int_\Omega \sum_{k,l=1}^{s} V_{kl}(x, u, \nabla u)\, u_l\, v_k^+ \,.$$

Its first term is nonnegative, since all $b_k \geq 0$, and v_k^+ equals either 0 or $u_k - M$, hence $\nabla u_k \cdot \nabla v_k^+$ equals either 0 or $|\nabla u_k|^2 \geq 0$. The second term is also nonnegative. Namely, let us introduce the further notations

$$\widehat{V}_{kl}(x) := V_{kl}(x, u(x), \nabla u(x)), \qquad v_k^- := \max\{M - u_k,\, 0\}$$

$(x \in \Omega,\ k, l = 1, \ldots, s)$. Then, for all $l = 1, \ldots, s$, we have $u_l = v_l^+ - v_l^- + M$ and hence the second integrand pointwise satisfies

$$\sum_{k,l=1}^{s} \widehat{V}_{kl}\, u_l\, v_k^+ = \sum_{k,l=1}^{s} \widehat{V}_{kl}\, v_l^+\, v_k^+ - \sum_{k=1}^{s} \widehat{V}_{kk}\, v_k^-\, v_k^+ + \sum_{k\neq l=1}^{s} (-\widehat{V}_{kl})\, v_l^-\, v_k^+ + M \sum_{k=1}^{s} \Big(\sum_{l=1}^{s} \widehat{V}_{kl}\Big)\, v_k^+ \,.$$

Here the first term on the r.h.s. equals the quadratic form $\widehat{V}v^+ \cdot v^+$. The cooperativity and the weak diagonal dominance of V w.r.t. both rows and columns imply that $\widehat{V}$ is positive semidefinite, hence $\widehat{V}v^+ \cdot v^+ \geq 0$. The second term equals zero, since either v_k^- or v_k^+ vanishes for all k. The third term is nonnegative, since $\widehat{V}_{kl} \leq 0$ from (3.4.33) and v_l^-, $v_k^+ \geq 0$ by definition. The last term is also nonnegative, since $\sum\limits_{l=1}^{s} \widehat{V}_{kl} \geq 0$ from (3.4.34).

Altogether, we obtain $\langle A(u), v^+\rangle \geq 0$. On the other hand, the assumptions $f_k \leq 0$ and $\gamma_k \leq 0$ imply that the right-hand side (3.4.38) of (3.4.35) satisfies

$$\langle \psi, v^+\rangle = \int_\Omega \sum_{k=1}^{s} f_k v_k^+ + \int_{\Gamma_N} \sum_{k=1}^{s} \gamma_k v_k^+ \leq 0.$$

This implies that $\langle A(u), v^+\rangle = \langle \psi, v^+\rangle = 0$. Moreover, both integrands in $\langle A(u), v^+\rangle$ vanish. Introducing the notation $\Omega_k^+ := \{x \in \Omega : \ u_k(x) \geq M\}$, the first integrand in $\langle A(u), v^+\rangle$ satisfies

$$0 = \int_\Omega \sum_{k=1}^{s} b_k(x, u, \nabla u)\, \nabla u_k \cdot \nabla v_k^+ = \sum_{k=1}^{s} \int_{\Omega_k^+} b_k(x, u, \nabla u)\, |\nabla v_k^+|^2.$$

Using condition $b_k \geq m > 0$, we obtain that the integrals on each Ω_k^+ vanish, moreover, if Ω_k^+ has a positive measure then $\nabla v_k^+ \equiv 0$, i.e. v_k^+ is constant, and (using $v_{k|\Gamma_D}^+ = 0$ and $\Gamma_D \neq \emptyset$) we obtain $v_k^+ \equiv 0$, which means that $u_k \leq M$ on Ω. On the other hand, if Ω_k^+ has zero measure then $u_k \leq M$ on Ω again, now by the definition of v_k^+.

Altogether, we obtain $u_k \leq M$ on Ω for all k, which is equivalent to (3.4.39). ■

If $u \in C(\overline{\Omega})$ is not assumed then the same proof can be repeated, provided that g_k are bounded on Γ_D: then $\max u_k$ and $\max g_k$ in (3.4.39) are replaced by $\operatorname{ess\,sup} u_k$ and $\operatorname{ess\,sup} g_k$, respectively. In what follows, we will look for the DMP in the same form as (3.4.39).

Finite element discretization. We define the finite element discretization of problem (3.4.32) in the following way. First, let $\bar{n}_0 \leq \bar{n}$ be positive integers and let us choose basis functions

$$\varphi_1, \dots, \varphi_{\bar{n}_0} \in H_D^1(\Omega), \qquad \varphi_{\bar{n}_0+1}, \dots, \varphi_{\bar{n}} \in H^1(\Omega) \setminus H_D^1(\Omega), \tag{3.4.41}$$

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

$$\varphi_p \geq 0 \quad (p = 1, \dots, \bar{n}), \qquad \sum_{p=1}^{\bar{n}} \varphi_p \equiv 1, \tag{3.4.42}$$

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

$$\varphi_p(B_q) = \delta_{pq} \tag{3.4.43}$$

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 $1 \le i \le n_0$,

$$\text{if} \quad i = (k-1)\bar{n}_0 + p \quad \text{for some } 1 \le k \le s \text{ and } 1 \le p \le \bar{n}_0, \quad \text{then}$$

$$\phi_i := (0, \dots, 0, \varphi_p, 0, \dots, 0) \qquad \text{where} \ \ \varphi_p \ \text{ stands at the } k\text{-th entry}, \tag{3.4.44}$$

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

$$V_h^0 := \text{span}\{\phi_1, ..., \phi_{n_0}\} \ \subset H_D^1(\Omega)^s. \tag{3.4.45}$$

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

$$\text{if} \quad i = n_0 + (k-1)(\bar{n} - \bar{n}_0) + p - \bar{n}_0 \quad \text{for some } 1 \le k \le s \text{ and } \bar{n}_0 + 1 \le p \le \bar{n}, \quad \text{then}$$

$$\phi_i := (0, \dots, 0, \varphi_p, 0, \dots, 0)^T \qquad \text{where} \ \ \varphi_p \ \text{ stands at the } k\text{-th entry}, \tag{3.4.46}$$

that is, $(\phi_i)_m = \varphi_p$ if $m = k$ and $(\phi_i)_m = 0$ if $m \ne k$. From (3.4.45) and these, we let

$$V_h := \text{span}\{\phi_1, ..., \phi_n\} \ \subset H^1(\Omega)^s. \tag{3.4.47}$$

Using the above FEM subspaces, the finite element discretization of problem (3.4.32) leads to the task of finding $u^h \in V_h$ such that

$$\langle A(u^h), v\rangle = \langle \psi, v\rangle \qquad (\forall v \in V_h^0) \tag{3.4.48}$$

$$\text{and} \quad u^h - g^h \in V_h^0, \quad \text{i.e.,} \ \ u^h = g^h \text{ on } \Gamma_D \tag{3.4.49}$$

(where $g^h = \sum\limits_{j=n_0+1}^{n} g_j \phi_j \ \in V_h$ is the projection of g^* into the subspace spanned by the 'boundary vector basis functions' $\varphi_{n_0+1}, \dots, \varphi_n$). Then, setting $u^h = \sum\limits_{j=1}^{n} c_j \phi_j$ and $v = \phi_i$ $(i = 1, \dots, n_0)$ in (3.4.35) (just as in (3.3.13)) we obtain the $n_0 \times n$ system of algebraic equations

$$\sum_{j=1}^{n} a_{ij}(\bar{\mathbf{c}})\, c_j = d_i \qquad (i = 1, ..., n_0), \tag{3.4.50}$$

where for any $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ $(i = 1, ..., n_0, \ \ j = 1, ..., n)$,

$$a_{ij}(\bar{\mathbf{c}}) := \int_\Omega \Big(\sum_{k=1}^{s} b_k(x, u^h, \nabla u^h)\, (\nabla\phi_j)_k \cdot (\nabla\phi_i)_k + \sum_{k,l=1}^{s} V_{kl}(x, u^h, \nabla u^h)\, (\phi_j)_l\, (\phi_i)_k \Big) \tag{3.4.51}$$

$$\text{and} \qquad d_i := \int_\Omega \sum_{k=1}^{s} f_k (\phi_i)_k + \int_{\Gamma_N} \sum_{k=1}^{s} \gamma_k (\phi_i)_k\,. \tag{3.4.52}$$

In the same way as before, we enlarge system (3.4.50) to a square one by adding an identity block, and write it briefly as

$$\bar{\mathbf{A}}(\bar{\mathbf{c}})\bar{\mathbf{c}} = \mathbf{d}\,. \tag{3.4.53}$$

That is, for $i = 1, ..., n_0$ and $j = 1, ..., n$, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ has the entry $a_{ij}(\bar{\mathbf{c}})$ from (3.4.51).

In what follows, the (patch-)regularity of the considered meshes used in Theorem 3.4.3 will be usually weakened in some way. The following notions will be used:

Definition 3.4.1 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 $\varphi_p \in V_h$,

$$meas(\operatorname{supp}\varphi_p) \le c_0 h^d \tag{3.4.54}$$

(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 (3.4.16) is replaced by

$$c_1 h^\gamma \le meas(\operatorname{supp}\varphi_p) \le c_2 h^d \tag{3.4.55}$$

for some fixed constant

$$d \le \gamma < d + 2, \tag{3.4.56}$$

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

The discrete maximum principle for systems with nonlinear coefficients. Our goal is to apply Theorem 3.3.1 to derive a DMP for problem (3.4.32). For this, we first define the underlying operators and check Assumptions 3.3.1.

Lemma 3.4.1 *For any $u \in H^1(\Omega)^s$, let us define the operators $B(u)$ and $R(u)$ via*

$$\langle B(u)w, v\rangle = \int_\Omega \sum_{k=1}^s b_k(x, u, \nabla u)\, \nabla w_k \cdot \nabla v_k, \quad \langle R(u)w, v\rangle = \int_\Omega \sum_{k,l=1}^s V_{kl}(x, u, \nabla u)\, w_l\, v_k \tag{3.4.57}$$

($w \in H^1(\Omega)^s$, $v \in H^1_D(\Omega)^s$). Together with the operator A, defined in (3.4.37), the operators $B(u)$ and $R(u)$ satisfy Assumptions 3.3.1 in the spaces $H = H^1(\Omega)^s$ and $H_0 = H^1_D(\Omega)^s$.

Proof. Since $\Gamma_D \ne \emptyset$, we can endow $H^1(\Omega)^s$ with the norm

$$\|v\|^2 := \sum_{k=1}^s \Big(\int_\Omega |\nabla v_k|^2 + \int_{\Gamma_D} |v_k|^2\Big) \tag{3.4.58}$$

Then for $v \in H^1_D(\Omega)^s$ we have $\|v\|^2 = \sum_{k=1}^s \int_\Omega |\nabla v_k|^2$.

(i) It is obvious from (3.4.37) and (3.4.57) that $A(u) = B(u)u + R(u)u$.

(ii) Assumption 3.4.7 (iii) implies for all $u \in H^1(\Omega)^s$, $v \in H^1_D(\Omega)^s$ that

$$\langle B(u)v, v\rangle = \int_\Omega \sum_{k=1}^s b_k(x,u,\nabla u)\, |\nabla v_k|^2 \geq m \int_\Omega \sum_{k=1}^s |\nabla v_k|^2 = m\, \|v\|^2 . \tag{3.4.59}$$

(iii) Let $D \subset H^1(\Omega)^s$ consist of the functions that have only one nonzero coordinate that is nonnegative, i.e. $v \in D$ iff $v = (0, \dots, 0, z, 0, \dots, 0)^T$ with z at the k-th entry for some $1 \leq k \leq s$ and $z \in H^1(\Omega)$, $z \geq 0$. Further, let $P \subset H^1(\Omega)^s$ consist of the functions that have identical nonnegative coordinates, i.e. $v \in P$ iff $v = (y, \dots, y)$ for some $y \in H^1(\Omega)$, $y \geq 0$. Now let $u \in H^1(\Omega)^s$ and $v \in D$. If $w \in P$, then

$$\langle R(u)w, v\rangle = \int_\Omega \Big(\sum_{l=1}^s V_{kl}(x,u,\nabla u) \Big)\, yz \geq 0$$

by (3.4.34) and that $y, z \geq 0$. If $w = v \in D$, then by Remark 3.4.3

$$\langle R(u)v, v\rangle = \int_\Omega V_{kk}(x,u,\nabla u)\, z^2 \geq 0.$$

(iv) Let $\tilde{V} := \max\limits_{k,l} \|V_{kl}\|_{L^\infty}$, which is finite by Assumption 3.4.7 (ii), and let us define the new norm

$$|||v|||^2 := \|v\|^2_{L^2(\Omega)^s} = \int_\Omega \sum_{k=1}^s v_k^2 \tag{3.4.60}$$

on $H^1(\Omega)^s$. Then we have for all $u, w, v \in H^1(\Omega)^s$

$$\langle R(u)w, v\rangle \leq \tilde{V} \int_\Omega \sum_{k,l=1}^s |w_l|\, |v_k| \leq s\tilde{V} \int_\Omega \big(\sum_{k=1}^s |v_k|^2 \big)^{1/2} \big(\sum_{l=1}^s |w_l|^2 \big)^{1/2} \leq s\tilde{V}\, |||w|||\, |||v|||,$$

i.e. (3.3.3) holds with the constant function $M_R(r) \equiv s\tilde{V}$ $(r \geq 0)$.

Now we consider a finite element discretization for problem (3.4.32), developed as in (3.4.41) and afterwards. We can then prove the following nonnegativity result for the stiffness matrix:

Theorem 3.4.7 *Let problem (3.4.32) satisfy Assumptions 3.4.7. 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 $d \leq \gamma < d+2$ (where d is the space dimension) such that for any $p = 1, ..., \bar{n}_0$, $t = 1, ..., \bar{n}$ $(p \neq t)$, if $meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$ then*

$$\nabla\varphi_t \cdot \nabla\varphi_p \leq 0 \quad on\ \Omega \quad and \quad \int_\Omega \nabla\varphi_t \cdot \nabla\varphi_p \leq -K_0\, h^{\gamma-2} \tag{3.4.61}$$

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 3.4.1.

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.4.51) is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.

Proof. We wish to apply Theorem 3.3.1. With the operator A defined in (3.4.37), our problem (3.4.35)–(3.4.36) coincides with (3.3.1)–(3.3.2). The FEM subspaces (3.4.45) and (3.4.47) fall into the class (3.3.9). Using the operators $B(u)$ and $R(u)$ in (3.4.57), the discrete problem (3.4.48)–(3.4.49) turns into the form (3.3.12) such that by Lemma 3.4.1, $B(u)$ and $R(u)$ satisfy Assumptions 3.3.1 in the spaces $H = H^1(\Omega)^s$ and $H_0 = H^1_D(\Omega)^s$.

Next, we need to define neighbouring basis functions satisfying Assumptions 3.3.3. Let $\phi_i, \phi_j \in V_h$. Using definitions (3.4.44) and (3.4.46), assume that ϕ_i has φ_p at its k-th entry and ϕ_j has φ_t at its l-th entry. Then we call ϕ_i and ϕ_j neighbouring basis functions if $k = l$ and $meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$. Let $N := \{1, \dots, n\}$ as before. For any $k = 1, \dots, s$ let

$$\begin{aligned} S^0_k &:= \{i \in N : \ i = (k-1)\bar n_0 + p \text{ for some } 1 \le p \le \bar n_0\}, \\ \tilde S_k &:= \{i \in N : \ i = n_0 + (k-1)(\bar n - \bar n_0) + p - \bar n_0 \text{ for some } \bar n_0 + 1 \le p \le \bar n\}, \\ S_k &:= S^0_k \cup \tilde S_k \,, \end{aligned}$$

i.e. by (3.4.44) and (3.4.46), the basis functions ϕ_i with index $i \in S_k$ have a nonzero coordinate φ_p for some p at the k-th entry, and in particular, $i \in S^0_k$ if this φ_p is an 'interior' basis function (i.e. $1 \le p \le \bar n_0$) and $i \in \tilde S_k$ if this φ_p is a 'boundary' basis function (i.e. $\bar n_0 + 1 \le p \le \bar n$). Clearly, the set $N = \{1, \dots, n\}$ can be partitioned into the disjoint sets $S_1, \dots, S_s$, and we have to check items (i)–(iii) of Assumptions 3.3.3. Let $k \in \{1, \dots, s\}$. By definition $S^0_k = S_k \cap \{1, \dots, n_0\}$ and $\tilde S_k = S_k \cap \{n_0 + 1, \dots, n\}$, and both S^0_k and $\tilde S_k$ are nonempty, hence item (i) holds. We have assumed in the construction that that any two 'interior' basis functions φ_p, φ_t can be connected with a chain of interior basis functions with overlapping support. Defining a chain of vector basis functions by having the terms of the above chain at the k-th coordinates and zeros in all the other coordinates, the consecutive terms will be neighbouring basis functions, hence we obtain that the graph of all neighbouring indices in S^0_k is connected, i.e. item (ii) holds. Finally, it follows from (3.4.42) that arbitrary two basis functions φ_p, φ_t can be connected with a chain of basis functions with overlapping support. (Namely, take the union of the supports of the basis functions in all possible chains with overlapping supports from φ_p. If the obtained set Ω_p were not the entire Ω, then we would have $\sum_{p=1}^{\bar n} \varphi_p(x) = 0$ for $x \in \partial\Omega$ in contrast to (3.4.42). Therefore $\Omega_p = \Omega$, hence one of the chains reaches φ_t as well.) Defining again a chain of vector basis functions by having the terms of the above chain at the k-th coordinates and zeros in all the other coordinates, this just means as above that the graph of all neighbouring indices (as defined before Assumptions 3.3.3) in S_k is connected, i.e. item (iii) holds.

Our remaining task is to check assumptions (a)–(e) of Theorem 3.3.1.

(a) Let $\phi_i \in V^0_h$, $\phi_j \in V_h$, and let ϕ_i have φ_p at its k-th entry and ϕ_j have φ_t at its l-th entry. We must prove that either (3.3.17) or (3.3.18)–(3.3.20) holds. If $k \ne l$ then ϕ_i and ϕ_j have no common nonzero coordinates, hence $\langle B(u^h)\phi_j, \phi_i\rangle = 0$; further, by (3.4.33) and (3.4.42),

$$\langle R(u^h)\phi_j, \phi_i\rangle = \int_\Omega V_{kl}(x, u^h, \nabla u^h)\, \varphi_t\, \varphi_p \le 0 \tag{3.4.62}$$

i.e. (3.3.17) holds. If $k = l$, then Assumption 3.4.7 (iii) and (3.4.61) yield

$$\langle B(u^h)\phi_j, \phi_i\rangle = \int_\Omega b_k(x, u^h, \nabla u^h)\, \nabla\varphi_t \cdot \nabla\varphi_p \le m \int_{\Omega_{pt}} \nabla\varphi_t \cdot \nabla\varphi_p \tag{3.4.63}$$

where $\Omega_{pt} := \operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t$. If $meas(\Omega_{pt}) = 0$ then $\langle B(u^h)\phi_j, \phi_i\rangle = 0$ and we have (3.4.62) similarly as before, hence (3.3.17) holds again. If $meas(\Omega_{pt}) > 0$ then (3.4.61) implies

$$\langle B(u^h)\phi_j, \phi_i\rangle \le -mK_0 h^{\gamma-2} \equiv -\hat{c}_1\, h^{\gamma-2} =: -M_B(h) \tag{3.4.64}$$

and we must check (3.3.20). Here the norm (3.4.60) of the basis functions satisfies the following estimate, where ϕ_j has φ_t at its l-th entry as before, and we use (3.4.54) and that (3.4.42) implies $\varphi_t \le 1$:

$$|||\phi_j|||^2 = \|\phi_j\|^2_{L^2(\Omega)^s} = \|\varphi_t\|^2_{L^2(\Omega)} \le \int_{\operatorname{supp}\varphi_t} 1 \;= meas(\operatorname{supp}\varphi_t) \le c_2 h^d\,, \tag{3.4.65}$$

hence (3.3.19) gives $T(h)^2 \le h^d$. From this, using (3.4.64) and that $\gamma < d+2$ (as defined for (3.4.55)), we obtain

$$\lim_{h\to 0} M_B(h)/T(h)^2 \ge (\hat{c}_1/c_2)\, \lim_{h\to 0}\, h^{\gamma-2-d} = +\infty. \tag{3.4.66}$$

(b) Let $\phi_i \in V_h^0$ and $\phi_j \in V_h$ be neighbouring basis vectors, i.e, as defined before in the proof, $k = l$ and $meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$. Then, as seen just above, we obtain (3.4.64) and (3.4.66), which coincide with (3.3.18)–(3.3.20).

(c) We have obtained the constant bound $M_R(r) \equiv s\tilde{V}$ in Lemma 3.4.1 for Assumption 3.3.1 (iii), hence $M_R(\|u^h\|) \equiv s\tilde{V}$ is trivially bounded as $h \to 0$.

(d) For all $u \in H^1(\Omega)^s$ and $h > 0$, the definition of the functions ϕ_j and assumption (3.4.42) imply

$$\sum_{j=1}^{n} \phi_j = \Big(\sum_{p=1}^{\bar{n}} \varphi_p,\ \sum_{p=1}^{\bar{n}} \varphi_p,\ \ldots, \sum_{p=1}^{\bar{n}} \varphi_p\Big)^T = (1,\ 1, \ldots 1)^T =: \mathbf{1} \in \ker B(u), \tag{3.4.67}$$

since by (3.4.57), for all $v \in H^1_D(\Omega)^s$

$$\langle B(u)(\sum_{j=1}^{n} \phi_j), v\rangle = \langle B(u)\mathbf{1}, v\rangle = \int_\Omega \sum_{k=1}^{s} b_k(x, u, \nabla u)\, \nabla 1 \cdot \nabla v_k = 0.$$

(e) Let $h > 0$ and $i = 1, ..., n$ be arbitrary. We must prove that $\phi_i \in D$ and $\sum_{j=1}^{n} \phi_j \in P$ for the sets D, P defined in the proof of Lemma 3.4.1, paragraph (iii). First, by definition, ϕ_i has only one nonzero coordinate function φ_p that is nonnegative by (3.4.42), i.e. $\phi_i \in D$. Second, as seen in (3.4.67), we have $\sum_{j=1}^{n} \phi_j = \mathbf{1} \in P$. ■

Corollary 3.4.1 *Let the assumptions of Theorem 3.4.7 hold and let $f_k \le 0$, $\gamma_k \le 0$ $(k = 1, \ldots, s)$. For sufficiently small h, if $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ is the solution of (3.4.50) with matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.4.51), then*

$$\max_{i=1,\ldots,n} c_i \le \max\{0, \max_{i=n_0+1,\ldots,n} c_i\}. \tag{3.4.68}$$

Proof. By (3.4.52), $d_i \le 0 \ \ (i = 1, ..., n_0)$, hence Corollary 3.3.1 can be used. ■

The meaning of (3.4.68) is as follows. Let us split the vector $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ as before, i.e. $\bar{\mathbf{c}} = [\mathbf{c};\ \tilde{\mathbf{c}}]^T$ where $\mathbf{c} = (c_1, ..., c_{n_0})^T$ and $\tilde{\mathbf{c}} = (c_{n_0+1}, ..., c_n)^T$. Following the notions introduced after (3.4.41), the vectors $\mathbf{c}$ and $\tilde{\mathbf{c}}$ contain the coefficients of the 'interior basis functions' and 'boundary basis functions', respectively. Then (3.4.68) states that the maximal coordinate is nonpositive or arises for a boundary basis function.

Our main interest is the meaning of Corollary 3.4.1 for the FEM solution $u^h = (u_1^h, \dots, u_s^h)^T$ itself. It turns out to be the counterpart of (3.4.39):

Theorem 3.4.8 *Let the basis functions satisfy (3.4.42)–(3.4.43). If (3.4.68) holds for the FEM solution $u^h = (u_1^h, \dots, u_s^h)^T$, then u^h satisfies*

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \le \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} g_k^h\}. \tag{3.4.69}$$

Proof. Refine the above splitting $\bar{\mathbf{c}} = [\mathbf{c};\ \tilde{\mathbf{c}}]^T$ of the vector $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ as

$$\bar{\mathbf{c}} = \left(c_1^{(1)}, \dots, c_{\bar{n}_0}^{(1)};\ \dots\ ;\ c_1^{(s)}, \dots, c_{\bar{n}_0}^{(s)};\ c_{\bar{n}_0+1}^{(1)}, \dots, c_{\bar{n}}^{(1)};\ \dots\ ;\ c_{\bar{n}_0+1}^{(s)}, \dots, c_{\bar{n}}^{(s)}\right)^T,$$

that is, $\mathbf{c}$ has the $n_0 = s\bar{n}_0$ entries from $c_1^{(1)}$ to $c_{\bar{n}_0}^{(s)}$ belonging to the interior points, and $\tilde{\mathbf{c}}$ has the $n - n_0 = s(\bar{n} - \bar{n}_0)$ entries from $c_{\bar{n}_0+1}^{(1)}$ to $c_{\bar{n}}^{(s)}$ belonging to the boundary points, such that the upper index from 1 to s gives the number of coordinate in the elliptic system. Here for all $k = 1, \dots, s$ we have $u_k^h = \sum\limits_{p=1}^{\bar{n}} c_p^{(k)} \varphi_p$. Now let $k^* \in \{1, \dots, s\}$ and $p^* \in \{1, \dots, \bar{n}\}$ be indices such that $c_{p^*}^{(k^*)} = \max\limits_{i=1,\dots,n} c_i$. For all $k = 1, \dots, s$, using (3.4.42), $\max\limits_{\overline{\Omega}} u_k^h = \max\limits_{\overline{\Omega}} \sum\limits_{p=1}^{\bar{n}} c_p^{(k)} \varphi_p \le c_{p^*}^{(k^*)} \sum\limits_{p=1}^{\bar{n}} \varphi_p = c_{p^*}^{(k^*)}$, further, using (3.4.43), $u_{(k^*)}^h(B_{p^*}) = \sum\limits_{p=1}^{\bar{n}} c_p^{(k^*)} \varphi_p(B_{p^*}) = \sum\limits_{p=1}^{\bar{n}} c_p^{(k^*)} \delta_{p,p^*} = c_{p^*}^{(k^*)}$. These together mean that $\max\limits_{k=1,\dots,s} \max\limits_{\overline{\Omega}} u_k^h = u_{(k^*)}^h(B_{p^*})$. By (3.4.68), either $c_{p^*}^{(k^*)} \le 0$ or $p^* \in \{n_0 + 1, \dots, \bar{n}\}$ (i.e. p^* is a 'boundary index', for which $B_{p^*} \in \Gamma_D$). In the first case

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h = u_{(k^*)}^h(B_{p^*}) = c_{p^*}^{(k^*)} \le 0,$$

and in the second case

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h = u_{(k^*)}^h(B_{p^*}) = \max_{\Gamma_D} u_{(k^*)}^h = \max_{k=1,\dots,s} \max_{\Gamma_D} u_k^h = \max_{k=1,\dots,s} \max_{\Gamma_D} g_k^h\ .$$

These two relations just mean that (3.4.69) holds. ■

Thus we obtain the *discrete maximum principle* for system (3.4.32):

Theorem 3.4.9 *Let the assumptions of Theorem 3.4.7 hold and let*

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

Let the basis functions satisfy (3.4.42)–(3.4.43). Then for sufficiently small h, if $u^h = (u_1^h, \dots, u_s^h)^T$ *is the FEM solution of system (3.4.32), then*

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \le \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} g_k^h\}. \tag{3.4.70}$$

Remark 3.4.4 (i) Let $f_k \le 0$, $\gamma_k \le 0$ for all k. The result (3.4.70) can be divided in two cases, both of which are remarkable: if at least one of the functions g_k^h has positive values on Γ_D then

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h = \max_{k=1,\dots,s} \max_{\Gamma_D} g_k^h \tag{3.4.71}$$

(which can be called more directly a discrete maximum principle than (3.4.70)), and if $g_k \le 0$ on Γ_D for all k, then we obtain the nonpositivity property

$$u_k^h \le 0 \quad \text{on } \Omega \text{ for all } k\,. \tag{3.4.72}$$

(ii) Analogously, if $f_k \ge 0$, $\gamma_k \ge 0$ for all k, then (by reversing signs) we can derive the corresponding discrete minimum principles instead of (3.4.70) and (3.4.71), or the corresponding nonnegativity property instead of (3.4.72).

Remark 3.4.5 The key assumption for the meshes in the above results is property (3.4.61). A simple but stronger sufficient condition to satisfy (3.4.61) is that for any $p = 1, ..., \bar{n}_0$, $t = 1, ..., \bar{n}$ $(p \ne t)$, (3.4.15) should hold, and in addition, if the family of meshes is quasi-regular according to Definition 3.4.1, then (3.4.61) is satisfied. For simplicial FEM, assumption (3.4.15) corresponds to acute triangulations. These properties and less strong assumptions to satisfy (3.4.61) will be addressed in (3.4.119) and the discussion afterwards.

(b) Systems with general reaction terms of sublinear growth

It is somewhat restrictive in (3.4.32) that both the principal and lower-order parts of the equations are given as containing products of coefficients with ∇u_k and u_l, respectively. Whereas this is widespread in real models for the principal part (and often the coefficient of ∇u_k depends only on x, or x and $|\nabla u|$), on the contrary, the lower order terms are usually not given in such a coefficient form. Now we consider problems where the dependence on the lower order terms is given as general functions of x and u. In this section these functions are allowed to grow at most linearly, in which case one can reduce the problem to the previous one (3.4.32) directly. (Superlinear growth of q_k will be dealt with in the next section.) Accordingly, let us now consider the system

$$\left.\begin{aligned} -\mathrm{div}\Big(b_k(x,u,\nabla u)\,\nabla u_k\Big) + \ q_k(x,u_1,\dots,u_s) \ &= \ f_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x,u,\nabla u)\tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \\ u_k \ &= \ g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k = 1, \dots, s) \tag{3.4.73}$$

under the following assumptions:

Assumptions 3.4.2.

(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 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 $q_k \in W^{1,\infty}(\Omega \times \mathbf{R}^s)$.

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

(iv) (Cooperativity.) We have

$$\frac{\partial q_k}{\partial \xi_l}(x, \xi) \leq 0 \qquad (k, l = 1, \dots, s,\ k \neq l;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.74}$$

(v) (Weak diagonal dominance for the Jacobians.) We have

$$\sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x, \xi) \geq 0 \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.75}$$

(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 $g^* \in H^1(\Omega)$.

Remark 3.4.6 Similarly to Remark 3.4.3, assumptions (3.4.74)–(3.4.75) now imply

$$\frac{\partial q_k}{\partial \xi_k}(x, \xi) \geq 0 \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.76}$$

The basic idea to deal with problem (3.4.73) is to reduce it to (3.4.32) via suitably defined functions $V_{kl} : \Omega \times \mathbf{R}^s \to \mathbf{R}$. Namely, let

$$V_{kl}(x, \xi) := \int_0^1 \frac{\partial q_k}{\partial \xi_l}(x, t\xi)\, dt \qquad (k, l = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.77}$$

Then the Newton-Leibniz formula yields

$$q_k(x, \xi) = q_k(x, 0) + \sum_{l=1}^{s} V_{kl}(x, \xi)\, \xi_l \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.78}$$

Defining

$$\hat{f}_k(x) := f_k(x) - q_k(x, 0) \qquad (k = 1, \dots, s), \tag{3.4.79}$$

problem (3.4.73) then becomes

$$\left.\begin{aligned} -\mathrm{div}\Big(b_k(x, u, \nabla u)\, \nabla u_k\Big) + \sum_{l=1}^{s} V_{kl}(x, u)\, u_l &= \hat{f}_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x, u, \nabla u) \tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \\ u_k &= g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k = 1, \dots, s), \tag{3.4.80}$$

which is a special case of (3.4.32). Here the assumption $q_k \in W^{1,\infty}(\Omega \times \mathbf{R}^s)$ yields that $V_{kl} \in L^\infty(\Omega \times \mathbf{R}^s)$ $(k,l = 1,\dots,s)$. Clearly, assumptions (3.4.74) and (3.4.75) imply that the functions V_{kl} defined in (3.4.77) satisfy (3.4.33) and (3.4.34), respectively. The remaining items of Assumptions 3.4.7 and 3.4.2 coincide, therefore system (3.4.80) satisfies Assumptions 3.4.2.

Consequently, all our results obtained for (3.4.32) can be applied to (3.4.73) too. First, Propositions 3.4.1–3.4.2 yield corresponding *continuous maximum principles.* Further, for a finite element discretization developed as for the system before, Theorem 3.4.8 yields the *discrete maximum principle* (3.4.69) for suitable discretizations of (3.4.80), provided $\hat{f}_k \le 0$ and $\gamma_k \le 0$ $(k = 1,\dots,s)$. For the original system (3.4.73), we thus obtain

Corollary 3.4.2 *Let problem (3.4.73) satisfy Assumptions 3.4.2, and let its FEM discretization satisfy the corresponding conditions of Theorem 3.4.7. If*

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

and $u^h = (u_1^h,\dots,u_s^h)^T$ is the FEM solution of system (3.4.73), then for sufficiently small h,

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \le \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} g_k^h\}. \tag{3.4.81}$$

(c) Systems with general reaction terms of superlinear growth

In the previous section we have required the functions q_k to grow at most linearly via the condition $q_k \in W^{1,\infty}(\Omega \times \mathbf{R}^s)$. However, this is a strong restriction and is not satisfied even by nonlinear polynomials of u_k that often arise in reaction-diffusion problems. In this section we extend the previous results to problems where the functions q_k may grow polynomially. This generalization, however, needs stronger assumptions in other parts of the problem, because we now need the monotonicity of the corresponding operator in the proof of the DMP. For this to hold, the row-diagonal dominance for the Jacobians in assumption 3.4.2 (v) must be strengthened to diagonal dominance w.r.t. both rows and columns. (In addition, the principal part must be more specific too, but this is not so much restrictive since in practice it is usually even linear.)

Accordingly, let us now consider the system

$$\left.\begin{aligned} -\mathrm{div}\Big(b_k(x,\nabla u_k)\,\nabla u_k\Big) + \; q_k(x,u_1,\dots,u_s) \;&=\; f_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x,\nabla u_k)\tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \\ u_k \;&=\; g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k = 1,\dots,s) \tag{3.4.82}$$

under the following assumptions:

Assumptions 3.4.10.

(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

$$2 \le p < p^*, \quad \text{where } p^* := \tfrac{2d}{d-2} \text{ if } d \ge 3 \text{ and } p^* := +\infty \text{ if } d = 2; \tag{3.4.83}$$

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

$$\left| \frac{\partial q_k}{\partial \xi_l}(x, \xi) \right| \le \beta_1 + \beta_2 |\xi|^{p-2} \qquad (k, l = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.4.84}$$

(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 (3.4.74).

(v) (Weak diagonal dominance for the Jacobians w.r.t. rows and columns.) We have for all $k = 1, \dots, s,\ \ x \in \Omega,\ \xi \in \mathbf{R}^s$

$$\sum_{l=1}^{s} \frac{\partial q_k}{\partial \xi_l}(x, \xi) \ge 0, \qquad \sum_{l=1}^{s} \frac{\partial q_l}{\partial \xi_k}(x, \xi) \ge 0. \tag{3.4.85}$$

(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 $g^* \in H^1(\Omega)$.

Remark 3.4.7 Similarly to Remark 3.4.3, the assumptions imply the nonnegativity (3.4.76).

To handle system (3.4.82), we start as in the previous subsection by reducing it to a system with nonlinear coefficients: if the functions V_{kl} and $\hat{f}_k$ $(k, l = 1, \dots, s)$ are defined as in (3.4.77) and (3.4.79), respectively, then (3.4.82) takes a form similar to (3.4.80):

$$\left.\begin{aligned} -\mathrm{div}\left(b_k(x, \nabla u)\, \nabla u_k\right) + \sum_{l=1}^{s} V_{kl}(x, u)\, u_l &= \hat{f}_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x, u, \nabla u) \tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \\ u_k &= g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k = 1, \dots, s). \tag{3.4.86}$$

The difference compared to the previous subsection is the superlinear growth allowed in (3.4.84), which does not let us apply Theorem 3.4.8 directly as we did for system (3.4.73). Instead, we must reprove Theorem 3.4.7 under Assumptions 3.4.10. (We note in contrast that a continuous maximum principle holds as in paragraph (b), since Proposition 3.4.2 does not require boundedness of the V_{kl}.)

First, when considering a finite element discretization developed as before, we need a strengthened assumption for the quasi-regularity of the mesh.

Definition 3.4.2 Let $\Omega \subset \mathbf{R}^d$ and let us consider a family of FEM subspaces $\mathcal{V} = \{V_h\}_{h \to 0}$ constructed as in paragraph (a). 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 mesh will be called *quasi-regular* w.r.t. problem (3.4.82) if

$$c_1 h^\gamma \le meas(\operatorname{supp} \varphi_p) \le c_2 h^d \,, \tag{3.4.87}$$

where the positive real number γ satisfies

$$d \le \gamma < \gamma_d^*(p) := 2d - \frac{(d-2)p}{2} \tag{3.4.88}$$

with p from Assumption 3.4.10 (ii).

Remark 3.4.8 Assumption (3.4.88) makes sense for γ since by (3.4.83),

$$d < d + d(1 - \tfrac{p}{p^*}) = \gamma_d^*(p) \,. \tag{3.4.89}$$

Note on the other hand that $\gamma_d^*(p) \le \gamma_d^*(2) = d+2$, which is in accordance with (3.4.56). Further, we have, in particular, in 2D: $\gamma_2^*(p) \equiv 4$ for all $2 \le p < \infty$, and in 3D: $\gamma_3^*(p) = 6 - (p/2)$ (where $2 \le p \le 6$, and accordingly $3 \le \gamma_3^*(p) \le 5$).

Next, as an analogue of Lemma 3.4.1, we need a technical result for problem (3.4.82):

Lemma 3.4.2 *Let Assumptions 3.4.10 hold. Analogously to (3.4.57), for any $u \in H^1(\Omega)^s$ let us define the operators $B(u)$ and $R(u)$ via*

$$\langle B(u)w, v\rangle = \int_\Omega \sum_{k=1}^s b_k(x, \nabla u)\, \nabla w_k \cdot \nabla v_k, \quad \langle R(u)w, v\rangle = \int_\Omega \sum_{k,l=1}^s V_{kl}(x,u)\, w_l\, v_k$$

($w \in H^1(\Omega)^s$, $v \in H^1_D(\Omega)^s$). Together with $A(u) := B(u)u + R(u)u$, the operators $B(u)$ and $R(u)$ satisfy Assumptions 3.3.1-3.3.2.

Proof. First, we must verify Assumptions 3.3.1. The stronger growth (3.4.84) causes a difference only in proving Assumption 3.3.1 (iv), i.e. to fulfil (3.3.3). Hence we only verify this property, the proof of the other items of Assumption 3.3.1 is the same as in Lemma 3.4.1.

Consider p^* as defined in (3.4.83). Then by [1] we have the Sobolev embedding estimate

$$\|z\|_{L^{p^*}(\Omega)} \le k_1 \|z\|_{H^1} \quad (z \in H^1(\Omega)) \tag{3.4.90}$$

with a constant $k_1 > 0$, where $\|z\|^2_{H^1} := \displaystyle\int_\Omega |\nabla z|^2 + \int_{\Gamma_D} |z|^2$. This is inherited for $v \in H^1(\Omega)^s$ too under the product norm $\|.\|$ on $H^1(\Omega)^s$ defined in (3.4.58). Here, by (3.4.77) and (3.4.84),

$$|\langle R(u)w, v\rangle| = |\int_\Omega \sum_{k,l=1}^s V_{kl}(x,u)\, w_l\, v_k| \le \int_\Omega \sum_{k,l=1}^s \left(\beta_1 + \beta_2 |u|^{p-2}\right) |w_l|\, |v_k| \tag{3.4.91}$$

for all $u, v, w \in H^1(\Omega)^s$. Letting $|v|^2 := \sum\limits_{k=1}^{s} v_k^2 \quad (v \in H^1(\Omega)^s)$, we have $\sum\limits_{k,l=1}^{s} |w_l|\,|v_k| \le s|w|\,|v|$, hence

$$|\langle R(u)w, v\rangle| \le s \int_\Omega \bigl(\beta_1 + \beta_2 |u|^{p-2}\bigr)\,|w|\,|v|\,. \tag{3.4.92}$$

For vector functions $v \in L^p(\Omega)^s$, we define $\|v\|_{L^p_s} := \Bigl\| |v| \Bigr\|_{L^p(\Omega)}$ with $|v|$ defined as above. Let us now fix a real number r satisfying

$$1 < r \le \frac{p^*}{p-2}\,. \tag{3.4.93}$$

If $q > 1$ is chosen to have $\frac{1}{r} + \frac{1}{q} = 1$, then Hölder's inequality implies

$$\int\limits_\Omega |u|^{p-2}\,|w|\,|v| \;\le \Bigl\| |u|^{p-2} \Bigr\|_{L^r(\Omega)} \|w\|_{L^{2q}_s} \|v\|_{L^{2q}_s} = \|u\|^{p-2}_{L_s^{(p-2)r}} \|w\|_{L^{2q}_s} \|v\|_{L^{2q}_s}. \tag{3.4.94}$$

Here $(p-2)r \le p^*$ and (3.4.90) imply

$$\|u\|^{p-2}_{L^{(p-2)r}(\Omega)} \le k_2 \|u\|^{p-2}_{L^{p^*}(\Omega)} \le k_3 \|u\|^{p-2} \tag{3.4.95}$$

with some constants $k_2, k_3 > 0$. Setting $u \equiv 1$ in (3.4.94) and using (3.4.95), we obtain

$$\int\limits_\Omega |w|\,|v| \;\le k_4\, \|w\|_{L^{2q}_s} \|v\|_{L^{2q}_s} \tag{3.4.96}$$

with some constant $k_4 > 0$. Then (3.4.92), (3.4.96) and (3.4.94) imply

$$|\langle R(u)w, v\rangle| \le s\left(\beta_1\, k_4 + \beta_2\, k_3\, \|u\|^{p-2}\right) \|w\|_{L^{2q}_s} \|v\|_{L^{2q}_s}. \tag{3.4.97}$$

That is, if we define the new norm $|||\,.\,|||$ as $|||v||| := \|v\|_{L^{2q}_s}$ $(v \in H^1(\Omega)^s)$, then (3.3.3) holds with

$$M_R(t) := s(\beta_1\, k_4 + \beta_2\, k_3\, t^{p-2}) \qquad (t \ge 0). \tag{3.4.98}$$

Now we have to verify Assumptions 3.3.2. Note first that we have

$$\langle A(u), v\rangle = \int_\Omega \Bigl(\sum_{k=1}^{s} b_k(x, \nabla u)\, \nabla u_k \cdot \nabla v_k + \sum_{k,l=1}^{s} V_{kl}(x,u)\, u_l\, v_k \Bigr) \tag{3.4.99}$$

$(u \in H^1(\Omega)^s,\ v \in H^1_D(\Omega)^s)$. Using the notation a_k from Assumption 3.4.2 (iii) and relation (3.4.78), we obtain $\langle A(u), v\rangle = \langle F(u), v\rangle + \langle G(u), v\rangle$ where

$$\langle F(u), v\rangle = \int_\Omega \sum_{k=1}^{s} a_k(x, \nabla u)\cdot \nabla v_k, \qquad \langle G(u), v\rangle = \int_\Omega \Bigl(\sum_{k=1}^{s} q_k(x,u)\; v_k - \sum_{k=1}^{s} q_k(x,0)\, v_k \Bigr) \tag{3.4.100}$$

$(u \in H^1(\Omega)^s,\ v \in H^1_D(\Omega)^s)$. Here, by Assumption 3.4.2 (iii), there exist constants $M \ge m > 0$ such that

$$\frac{\partial a_k}{\partial \eta}(x,\eta)\; \xi\cdot\xi \ge\; m|\xi|^2, \qquad \frac{\partial a_k}{\partial \eta}(x,\eta)\; \xi\cdot\zeta \le\; M|\xi|\,|\zeta| \tag{3.4.101}$$

$(x \in \Omega,\ \eta, \xi, \zeta \in \mathbf{R}^d)$. We can now check properties (i)–(iv) of Assumptions 3.3.2.

(i) Under Assumptions 3.4.2, it follows e.g. from [55, Theorem 6.2] that the operators F, G in (3.4.100) are Gateaux differentiable, further, that F' and G' are bihemicontinuous. In fact, the latter have the form

$$\langle F'(u)w, v\rangle = \int_\Omega \sum_{k=1}^s \frac{\partial a_k}{\partial \eta}(x, \nabla u)\, \nabla w_k \cdot \nabla v_k, \quad \langle G'(u)w, v\rangle = \int_\Omega \sum_{k,l=1}^s \frac{\partial q_k}{\partial \xi_l}(x, u)\, w_l\, v_k\,. \tag{3.4.102}$$

(ii) Let $u \in H^1(\Omega)^s$, $w, v \in H^1_D(\Omega)^s$. We obtain from (3.4.101) and (3.4.102) that

$$\langle F'(u)w, v\rangle \le M\|w\|\,\|v\| \tag{3.4.103}$$

where $\|h\|^2 := \sum\limits_{k=1}^s \int_\Omega |\nabla h_k|^2$ is the product norm $\|.\|$ on $H^1_D(\Omega)^s$. Further, by (3.4.102) and (3.4.84),

$$|\langle G'(u)w, v\rangle| \le \int_\Omega \sum_{k,l=1}^s \left(\beta_1 + \beta_2 |u|^{p-2}\right) |w_l|\, |v_k|\,. \tag{3.4.104}$$

This means that $G'(u)$ has the same bound as $R(u)$ in (3.4.91), but the latter has been estimated above by (3.4.97), hence $G'(u)$ also has the bound (3.4.97). If we now choose $r = \frac{p}{p-2}$ in (3.4.93), then condition $\frac{1}{r} + \frac{1}{q} = 1$ yields $q = \frac{p}{2}$, and setting the latter in the bound in (3.4.97) thus gives

$$|\langle G'(u)w, v\rangle| \le \left(\beta_1\, k_4 + \beta_2\, k_3\, \|u\|^{p-2}\right) \|w\|_{L^p_s}\, \|v\|_{L^p_s}\,. \tag{3.4.105}$$

Using (3.4.90) and that $p < p^*$, we obtain $\|w\|_{L^p_s} \le k_5 \|w\|_{L^{p^*}_s} \le k_6 \|w\|$ on $H^1_D(\Omega)^s$, hence

$$|\langle G'(u)w, v\rangle| \le k_6\big(\beta_1\, k_4 + \beta_2\, k_3\, \|u\|^{p-2}\big)\, \|w\|\, \|v\|\,. \tag{3.4.106}$$

Finally, from $A'(u) = F'(u) + G'(u)$, using (3.4.103) and (3.4.106), we obtain

$$|\langle A'(u)w, v\rangle| \le \Big(M + k_6\big(\beta_1\, k_4 + \beta_2\, k_3\, \|u\|^{p-2}\big)\Big)\, \|w\|\, \|v\|\,,$$

i.e. the required estimate (3.3.4) with $M_A(t) := M + k_6\big(\beta_1\, k_4 + \beta_2\, k_3\, t^{p-2}\big) \quad (t \ge 0)$.

(iii) We obtain immediately from (3.4.101) and (3.4.102) that $\langle F'(u)v, v\rangle \ge m\|v\|^2 \;\; (u \in H^1(\Omega)^s,\; v \in H^1_D(\Omega)^s)$.

(iv) By Assumptions 3.4.2 (iv)–(v), for all $x \in \Omega$ and $\xi \in \mathbf{R}^s$ the Jacobians $\frac{\partial q_k}{\partial \xi_l}(x, \xi)$ are M-matrices and weakly diagonally dominant w.r.t. both rows and columns. It is well-known that such matrices are positive semidefinite. Therefore

$$\langle G'(u)v, v\rangle = \int_\Omega \sum_{k,l=1}^s \frac{\partial q_k}{\partial \xi_l}(x, u)\, v_l\, v_k \ge\; 0 \qquad (u \in H,\; v \in H_0). \qquad \blacksquare \tag{3.4.107}$$

Now we can prove the desired nonnegativity result for the stiffness matrix, i.e. the analogue of Theorem 3.4.7 for system (3.4.82). Here the entries of $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ are

$$a_{ij}(\bar{\mathbf{c}}) = \int_\Omega \Big(\sum_{k=1}^{s} b_k(x, \nabla u^h)\, (\nabla \phi_j)_k \cdot (\nabla \phi_i)_k + \sum_{k,l=1}^{s} V_{kl}(x, u^h)\, (\phi_j)_l\, (\phi_i)_k \Big), \tag{3.4.108}$$

where by (3.4.77),

$$V_{kl}(x, u^h(x)) = \int_0^1 \frac{\partial q_k}{\partial \xi_l}(x, t u^h(x))\, dt \qquad (k, l = 1, \dots, s;\ x \in \Omega). \tag{3.4.109}$$

Theorem 3.4.10 *Let problem (3.4.82) satisfy Assumptions 3.4.10. Let us consider a family of finite element subspaces V_h $(h \to 0)$ satisfying the following property: there exists a real number γ satisfying (3.4.88) such that for any indices $p = 1, ..., \bar{n}_0,\ t = 1, ..., \bar{n}$ $(p \neq t)$, if $meas(\operatorname{supp} \varphi_p \cap \operatorname{supp} \varphi_t) > 0$ then*

$$\nabla \varphi_t \cdot \nabla \varphi_p \leq 0 \quad \text{on } \Omega \quad \text{and} \quad \int_\Omega \nabla \varphi_t \cdot \nabla \varphi_p \leq -K_0\, h^{\gamma - 2} \tag{3.4.110}$$

with some constant $K_0 > 0$ independent of p, t and h. Further, let the family of meshes be quasi-regular, according to Definition 3.4.2.

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.4.108) is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.

Proof. We follow the proof of Theorem 3.4.7 and wish to apply Theorem 3.3.1. Most of the arguments are identical, corresponding to the conditions that coincide in Assumptions 3.4.7 and 3.4.10. We will concentrate on the different parts. Since Assumptions 3.3.1 hold by Lemma 3.4.1, we are left to check assumptions (a)–(e) of Theorem 3.3.1.

(a) Let $\phi_i \in V_h^0$, $\phi_j \in V_h$, and let ϕ_i have φ_p at its k-th entry and ϕ_j have φ_t at its l-th entry. We obtain similarly as in the proof of Theorem 3.4.7 that (3.3.17) holds if either $k \neq l$, or $k = l$ and $meas(\Omega_{pt}) = 0$, where $\Omega_{pt} := \operatorname{supp} \varphi_p \cap \operatorname{supp} \varphi_t$. The stronger growth (3.4.84) causes a difference only in verifying (3.3.18)–(3.3.20) in the case $k = l$ and $meas(\Omega_{pt}) > 0$. Here, in the same way as in (3.4.64), we obtain

$$\langle B(u^h)\phi_j, \phi_i \rangle \leq -\hat{c}_1\, h^{\gamma-2} =: -M_B(h) \tag{3.4.111}$$

and we must check (3.3.20). Let us now choose a real number r satisfying

$$\frac{d}{2 + d - \gamma} < r \leq \frac{p^*}{p - 2}\,. \tag{3.4.112}$$

Here $\gamma \geq 2$ implies $d/(2 + d - \gamma) \geq 1$, hence (3.4.112) is a special case of (3.4.93). Such an r exists for the following reason. If $d = 2$ then $p^* = +\infty$, hence there is nothing to prove. If $d \geq 3$ then we first observe that the fact $p \geq 2$ and (3.4.88) imply

$$\gamma < 2d - \tfrac{(d-2)p}{2} = d + 2 - \tfrac{(d-2)(p-2)}{2} \leq d + 2\,, \tag{3.4.113}$$

hence the denominator of $d/(2+d-\gamma)$ is positive. Hence the reciprocal of the l.h.s. of (3.4.112) must be greater than its r.h.s. The first inequality of (3.4.113) yields

$$2(2+d-\gamma) > (d-2)(p-2),$$

and by the definition $p^* := \frac{2d}{d-2}$ we obtain the desired inequality. Now let $q > 1$ be chosen to satisfy $\frac{1}{r} + \frac{1}{q} = 1$, and let us define the corresponding norm via

$$|||v|||^2 := \|v\|^2_{L^{2q}_s} = \Big\|\sum_{k=1}^{s} v_k^2\Big\|_{L^q(\Omega)} \qquad (v \in H^1(\Omega)^s). \tag{3.4.114}$$

For this, as seen in Lemma 3.4.2, estimate (3.3.3) holds. Hence we obtain the following estimate, where ϕ_j has φ_t at its l-th entry as before, and we use (3.4.16) and that (3.4.42) implies $\varphi_t \le 1$:

$$|||\phi_j|||^2 = \Big\| \, |\varphi_t|^2 \Big\|_{L^q(\Omega)} = \|\,\varphi_t\|^2_{L^q(\Omega)} \le \Big(\int\limits_{\operatorname{supp}\varphi_t} 1 \Big)^{1/q} = meas(\operatorname{supp}\varphi_t)^{1/q} \le c_2 h^{d/q}\,, \tag{3.4.115}$$

hence (3.3.19) gives $T(h)^2 \le h^{d/q}$. Here $\frac{1}{r} + \frac{1}{q} = 1$ and (3.4.112) imply $\gamma - 2 - (d/q) = \gamma - 2 - d + (d/r) < 0$. From this, using (3.4.111) we obtain

$$\lim_{h\to 0} \frac{M_B(h)}{T(h)^2} \ge \frac{\hat{c}_1}{c_2} \lim_{h\to 0} h^{\gamma-2-(d/q)} = +\infty. \tag{3.4.116}$$

(b) This assumption is proved identically to that in Theorem 3.4.7, using the same definition of neighbouring basis vectors.

(c) We must verify that $M_R(\|u^h\|) = s(\beta_1\, k_4 + \beta_2\, k_3\, \|u^h\|^{p-2})$ is bounded as $h \to 0$. Note that Assumptions 3.3.2 hold by Lemma 3.4.1, and the functions $g_h \in V_h$ in (3.4.49) (that are the V_h-interpolants of g on Γ_D) are bounded in $H^1(\Omega)^s$-norm as $h \to 0$. From these two properties, as pointed out in Remark 3.3.1, it follows that $\|u^h\|$ is bounded as $h \to 0$, and then obviously $M_R(\|u^h\|)$ is bounded too.

(d)–(e) These assumptions are independent of the growth conditions on q_k, and are proved identically to those in Theorem 3.4.7. ■

Similarly as in Corollary 3.4.2, using Theorem 3.4.10, Corollary 3.3.1 and Theorem 3.4.8, respectively, we obtain the *discrete maximum principle* for system (3.4.82):

Corollary 3.4.3 *Let problem (3.4.82) satisfy Assumptions 3.4.10, and let its FEM discretization satisfy the conditions of Theorem 3.4.10. If*

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

then for sufficiently small h, the FEM solution $u^h = (u_1^h,\dots,u_s^h)$ of system (3.4.82) satisfies

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h \le \max_{k=1,\dots,s} \max\{0, \max_{\Gamma_D} g_k^h\}. \tag{3.4.117}$$

Remark 3.4.9 As pointed out in Remark 3.4.4, the result (3.4.117) can be divided in two cases: a 'more direct' DMP (3.4.71) or the nonpositivity property (3.4.72). Further, if $f_k \geq q_k(x,0)$, $\gamma_k \geq 0$ for all k, then (by reversing signs) one can derive the corresponding discrete minimum principle or nonnegativity property. We formulate the latter below for its practical importance.

Corollary 3.4.4 *Let problem (3.4.82) satisfy Assumptions 3.4.10, and let its FEM discretization satisfy the conditions of Theorem 3.4.10. If*

$$f_k \geq q_k(x,0), \quad \gamma_k \geq 0, \quad g_k \geq 0 \qquad (k = 1, \dots, s)$$

then for sufficiently small h, the FEM solution $u^h = (u_1^h, \dots, u_s^h)^T$ of system (3.4.82) satisfies

$$u_k^h \geq 0 \quad on\ \Omega \qquad (k = 1, \dots, s). \tag{3.4.118}$$

Sufficient conditions and their geometric meaning. The key assumption for the FEM subspaces V_h and the associated meshes in the above results has been the following property, see (3.4.61) in Theorem 3.4.7 and (3.4.110) in Theorem 3.4.10. There exists a real number γ satisfying (3.4.56) or (3.4.88), respectively, such that for any indices $p = 1, ..., \bar{n}_0,\ t = 1, ..., \bar{n}\ (p \neq t)$, if $\ meas(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$ then

$$\nabla\varphi_t \cdot \nabla\varphi_p \leq 0 \ \text{ on } \Omega \quad \text{and} \quad \int_\Omega \nabla\varphi_t \cdot \nabla\varphi_p \leq -K_0\, h^{\gamma-2} \tag{3.4.119}$$

with some constant $K_0 > 0$ independent of p, t and h. (The family of meshes must also be regular from above as in (3.4.54), but that requirement obviously holds for the usual definition of the mesh parameter h as the maximal diameter of elements.)

A classical way to satisfy such conditions is a pointwise inequality like (3.4.15) together with suitable mesh regularity, see Remark 3.4.5. However, one can ensure (3.4.119) with less strong conditions as well. We summarize some possibilities below.

Proposition 3.4.3 *Let the family of FEM discretizations $\mathcal{V} = \{V_h\}_{h\to 0}$ satisfy either of the following conditions, where φ_t, φ_p are arbitrary basis functions such that $p = 1, ..., \bar{n}_0,\ t = 1, ..., \bar{n},\ p \neq t$, we let*

$$\Omega_{pt} := \operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t\,,$$

further, let $\ \sigma > 0$ and $c_1, c_2, c_3 > 0\ $ denote constants independent of the indices p, t and the mesh parameter h, and finally, d is the space dimension and γ satisfies (3.4.88).

(i) Let there exist $0 < \varepsilon \leq \gamma - d$ such that the basis functions satisfy

$$\nabla\varphi_t \cdot \nabla\varphi_p \leq -\frac{\sigma}{h^{2-\varepsilon}} < 0 \quad on\ \Omega_{pt}, \tag{3.4.120}$$

but let the quasi-regularity (3.4.87) of the family of meshes be now strengthened to

$$c_1 h^{\gamma-\varepsilon} \leq meas(\operatorname{supp}\varphi_p) \leq c_2 h^d\,. \tag{3.4.121}$$

(ii) Let 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

$$\nabla\varphi_t \cdot \nabla\varphi_p \le -\frac{\sigma}{h^2} < 0 \quad on \;\; \Omega_{pt}^+, \qquad \nabla\varphi_t \cdot \nabla\varphi_p \le 0 \quad on \;\; \Omega_{pt} \setminus \Omega_{pt}^+ \tag{3.4.122}$$

further, let the family of meshes be quasi-regular as in (3.4.87).

Then (3.4.119) holds.

The proof of this proposition is obvious. The weaker conditions (3.4.120) and (3.4.122) allow in theory easier refinement procedures as the property of (strict) acuteness is often hard to preserve in refinement procedures, e.g. by bisection algorithms.

First, (3.4.120) may allow the acute mesh angles to deteriorate (i.e. tend to 90°) as $h \to 0$. Namely, if a family of simplicial meshes is regular then $|\nabla\varphi_t| = O(h^{-1})$ for all linear basis functions: hence, considering two basis functions φ_p, φ_t and letting α denote the angle of their gradients on a given simplex, the sufficient condition $\cos\alpha \le -\sigma h^\varepsilon$ (with some constant $\sigma > 0$ independent of h) implies $\nabla\varphi_t \cdot \nabla\varphi_p = |\nabla\varphi_t|\,|\nabla\varphi_p| \cos\alpha \le -\frac{\sigma\, h^\varepsilon}{h^2}$, i.e. (3.4.120) holds. Clearly, if $h \to 0$ then this allows $\cos\alpha \to 0$, i.e. $\alpha \to 90°$, for the angle of gradients, in which case the corresponding mesh angle also tends to 90°. (In particular, for problem (3.4.32), when (3.4.88) coincides with $d \le \gamma < d+2$ as in (3.4.56), then $\gamma - d$ can be chosen arbitrarily close to 2. Hence the exponent $2-\varepsilon$ in (3.4.120) can be arbitrarily close to 0, i.e. the decay of mesh angles to 90° may be fast as $h \to 0$.)

Second, (3.4.122) means that one can allow some right mesh angles, but each Ω_{pt}, which consists of a finite number of elements, must contain some elements with acute mesh angles and the measure of these must not asymptotically vanish.

3.4.3 Nonlinear systems including first order terms

(a) Nonsymmetric systems with linear convection coefficients

Finally we consider systems including first order terms [94]. 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 (3.4.32), where $k = 1, \dots, s$:

$$-\mathrm{div}\left(b_k(x,u,\nabla u)\,\nabla u_k\right) + \mathbf{w}_k(x)\cdot\nabla u_k + \sum_{l=1}^{s} V_{kl}(x,u,\nabla u)\,u_l \;=\; f_k(x). \tag{3.4.123}$$

Assumptions 3.4.11. 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 3.4.7.

A continuous maximum principle holds in the same form as in Proposition 3.4.2, since the first-order terms do not destroy the positivity used in the proof.

When considering a FEM discretization developed as in subsection 3.4.2, we need again a strengthened assumption for the quasi-regularity of the mesh such that (3.4.56) for γ is now replaced by

$$d \le \gamma < \frac{d(d+2)}{d+1}\,. \tag{3.4.124}$$

Theorem 3.4.11 *Let problem (3.4.123) satisfy Assumptions 3.4.11, and let assumptions of Theorem 3.4.7 hold except that the mesh quasi-regularity is understood with γ satisfying (3.4.124).*

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 3.2.1.

Hence, if $f_k \le 0$, $\gamma_k \le 0$ $(k = 1,\dots,s)$ and the basis functions satisfy (3.4.42)–(3.4.43), then for sufficiently small h the FEM solution of system (3.4.123), satisfes (3.4.70).

Proof. One has to modify the proof of Theorem 3.4.7 in an obvious way, since the assumptions on **w** are standard ones used to preserve the coercivity properties of the remaining terms. The underlying Theorem 3.3.1 has to be modified according to Remark 3.3.2. ■

(b) 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 (3.4.126) and the homogeneity of the Dirichlet data. The applicability of these conditions will be illustrated in the example in subsection 3.5.2.

Let us consider the following system, where $k = 1,\dots,s$:

$$\left.\begin{aligned} -\mathrm{div}\Big(b_k(x,\nabla u)\,\nabla u_k\Big) + \mathbf{w}_k(x,u)\cdot\nabla u_k + q_k(x,u_1,...,u_s) &= f_k(x) \quad \text{a.e. in } \Omega,\\ b_k(x,\nabla u)\tfrac{\partial u_k}{\partial\nu} &= \gamma_k(x) \quad \text{a.e. on } \Gamma_N,\\ u_k &= 0 \qquad \text{a.e. on } \Gamma_D \end{aligned}\right\} \tag{3.4.125}$$

Assumptions 3.4.12. The convection coefficients satisfy $\mathbf{w}_k \in L^\infty(\Omega\times\mathbf{R})$. The domain Ω and the other coefficients satisfy Assumptions 3.4.10, except that item (v) in the latter is strengthened as follows: there exists $\mu > 0$ such that

$$\sum_{l=1}^{s}\frac{\partial q_k}{\partial\xi_l}(x,\xi)\ge\mu, \qquad \sum_{l=1}^{s}\frac{\partial q_l}{\partial\xi_k}(x,\xi)\ge\mu \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s), \tag{3.4.126}$$

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

We proceed similarly as in the previous subsection, system (3.4.125) is reduced to a system with nonlinear coefficients as before via the functions $V_{kl} : \Omega\times\mathbf{R}^s \to \mathbf{R}$ and $\hat{f}_k$ from (3.4.77) and (3.4.79), respectively. The difference is the nonlinear convection term. Taking this into account, we must reprove Theorem 3.4.10 under Assumptions 3.4.12, but only those parts are addressed where the convection term is involved. (The same process implies again a continuous maximum principle too, which we do not detail here.)

The operator corresponding to our problem is

$$\langle A(u),v\rangle = \int_\Omega\Big(\sum_{k=1}^{s} b_k(x,\nabla u)\,\nabla u_k\cdot\nabla v_k + \sum_{k=1}^{s}(\mathbf{w}_k(x,u)\cdot\nabla u_k)\,v_k + \sum_{k,l=1}^{s} V_{kl}(x,u)\,u_l\,v_k\Big) \tag{3.4.127}$$

($u \in H^1(\Omega)^s,\ v \in H^1_D(\Omega)^s$). First we properly modify Lemma 3.4.1, where the main point is to compensate for the presence of the convection term in the positivity of the operator without a coercivity condition on $\mathbf{w}_k$. We define the operators

$$\langle B(u)z, v\rangle = \int_\Omega \sum_{k=1}^s \Big(b_k(x,\nabla u)\,\nabla z_k\cdot\nabla v_k + \mu z_k v_k\Big),\ \ \langle N(u)z, v\rangle = \int_\Omega \sum_{k=1}^s (\mathbf{w}_k(x,u)\cdot\nabla z_k)\,v_k$$
$$\langle R(u)z, v\rangle = \int_\Omega \Big(\sum_{k,l=1}^s V_{kl}(x,u)\,z_l\,v_k - \mu\sum_{k=1}^s z_k v_k\Big) \tag{3.4.128}$$

($z \in H^1(\Omega)^s$, $v \in H^1_D(\Omega)^s$). We note that (3.4.77) and (3.4.126) yield

$$\sum_{l=1}^s V_{kl}(x,\xi) \ge \mu \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s), \tag{3.4.129}$$

and hence, since $V_{kl}(x,\xi)\le 0$ for $k\ne l$ by Assumption 3.4.12 (v), we also have

$$V_{kk}(x,\xi) \ge \mu \qquad (k=1,\dots,s;\ x\in\Omega,\ \xi\in\mathbf{R}^s). \tag{3.4.130}$$

Lemma 3.4.3 *Let Assumptions 3.4.12 hold. For any $u \in H^1(\Omega)^s$, the operators $B(u)$, $N(u)$ and $R(u)$, together with the operator $A(u)$ in (3.4.127), satisfy Assumptions 3.3.1, modified according to Remark 3.3.2, in the spaces $H = H^1(\Omega)^s$ and $H_0 = H^1_D(\Omega)^s$.*

Proof. We must reprove those parts of Lemma 3.4.1 that involve the convection term or the modifications of $B(u)$ and $R(u)$ with the term containing μ.

(i) It is obvious from (3.4.127) and (3.4.128) that $A(u) = B(u)u + N(u)u + R(u)u$.

(ii) We must prove property (b) in Assumptions 3.3.1. Here for all $u \in H^1(\Omega)^s$ and $v \in H^1_D(\Omega)^s$,

$$\Big\langle \Big(B(u)+N(u)\Big)v, v\Big\rangle = \int_\Omega \sum_{k=1}^s \Big(b_k(x,\nabla u)\,|\nabla v_k|^2 + \mu v_k^2\Big) + \int_\Omega \sum_{k=1}^s (\mathbf{w}_k(x,u)\cdot\nabla v_k)\,v_k$$
$$\ge m\|\nabla v\|^2_{L^2(\Omega)^s} + \mu\|v\|^2_{L^2(\Omega)^s} - \omega\|\nabla v\|_{L^2(\Omega)^s}\,\|v\|_{L^2(\Omega)^s} \tag{3.4.131}$$

where $\omega := \|\mathbf{w}\|_{L^\infty(\Omega)^s}$. Using the basic inequality $xy \le \frac12\Big(\varepsilon x^2 + \frac1\varepsilon y^2\Big)$ ($\varepsilon > 0$, $x, y \in \mathbf{R}$) for the last two factors, we obtain

$$\Big\langle \Big(B(u)+N(u)\Big)v, v\Big\rangle \ge \Big(m - \frac{\omega\varepsilon}{2}\Big)\|\nabla v\|^2_{L^2(\Omega)^s} + \Big(\mu - \frac{\omega}{2\varepsilon}\Big)\|v\|^2_{L^2(\Omega)^s}\,.$$

Choosing $\varepsilon := \frac{\omega}{2\mu}$, we have $\Big\langle \Big(B(u)+N(u)\Big)v, v\Big\rangle \ge \hat m\,\|\nabla v\|^2_{L^2(\Omega)^s} \equiv \hat m\,\|v\|^2$ where $\hat m := m - \frac{\omega^2}{4\mu} > 0$ by assumption.

(iii) Let us consider the sets P and D, defined in paragraph (iii) of the proof of Lemma 3.4.1. That is, $v \in D$ iff $v = (0,\dots,0,g,0,\dots,0)^T$ with g at the k-th entry for some $1\le k\le s$ and $g \in H^1(\Omega)$, $g \ge 0$. Further, $v \in P$ iff $v = (y,\dots,y)$ for some function

$y \in H^1(\Omega)$, $y \geq 0$. We must prove that for any $u \in H^1(\Omega)^s$ and $v \in D$, we have

$$\langle R(u)z, v\rangle \geq 0 \tag{3.4.132}$$

provided that either $z \in P$ or $z = v \in D$. If $z \in P$, then

$$\langle R(u)z, v\rangle = \int_\Omega \Big(\sum_{l=1}^s V_{kl}(x,u) - \mu\Big)\, yg \geq 0$$

by (3.4.129) and that $y, g \geq 0$. If $z = v \in D$, then by (3.4.130)

$$\langle R(u)v, v\rangle = \int_\Omega \Big(V_{kk}(x,u) - \mu\Big)\, g^2 \geq 0.$$

(iv) This follows in the same way as in Lemma 3.4.1. For $N(u)$, we can similarly factor out $\|\mathbf{w}\|_{L^\infty(\Omega)^s}$. For $R(u)$, the new norms can remain $|||v|||^2_{R_1} = |||v|||^2_{R_2} = \|v\|^2_{L^{2q}(\Omega)^s}$ as in (3.4.114), since the additional term in (3.4.128) can be bounded by the product L^2-norm $\|.\|_{L^2(\Omega)^s}$, which is (up to a constant factor) not larger than the norm $\|.\|^2_{L^{2q}(\Omega)^s}$ owing to the Sobolev inequality. ■

Now we can derive the nonnegativity of the stiffness matrix. Here the entries of $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ are, for any $\bar{\mathbf{c}} = (c_1, ..., c_n)^T \in \mathbf{R}^n$ and $i = 1, ..., n_0$, $j = 1, ..., n$,

$$\begin{aligned} a_{ij}(\bar{\mathbf{c}}) := \int_\Omega \Bigg(\sum_{k=1}^s b_k(x, \nabla u^h)\, (\nabla\phi_j)_k \cdot (\nabla\phi_i)_k + \sum_{k=1}^s \Big(\mathbf{w}_k(x,u^h)\cdot(\nabla\phi_j)_k\Big)\, (\phi_i)_k \\ + \sum_{k,l=1}^s V_{kl}(x,u^h)\, (\phi_j)_l\, (\phi_i)_k \Bigg) \end{aligned} \tag{3.4.133}$$

where $V_{kl}(x, u^h)$ is as in (3.4.109).

Theorem 3.4.12 *Let problem (3.4.125) satisfy Assumptions 3.4.12. Let us consider a family of finite element subspaces $\mathcal{V} = \{V_h\}_{h\to 0}$, such that the corresponding family of meshes is quasi-regular according to Definition 3.4.2, further, for any $p = 1, ..., \bar{n}_0$, $t = 1, ..., \bar{n}$ $(p \neq t)$, if meas$(\operatorname{supp}\varphi_p \cap \operatorname{supp}\varphi_t) > 0$ then (3.4.119) holds, where γ is from (3.4.88) and $K_0 > 0$ is a constant independent of p, t and h.*

Then for sufficiently small h, the matrix $\bar{\mathbf{A}}(\bar{\mathbf{c}})$ defined in (3.4.133) is of generalized nonnegative type with irreducible blocks in the sense of Definition 3.2.1.

Proof. The proof is similar to that of Theorem 3.4.10, with a few differences. First, the proof for assumption (a) relies on (3.4.64), where by (3.4.128), now $\langle B(u^h)\phi_j, \phi_i\rangle$ contains the additional term $\int_\Omega \mu \sum_{k=1}^s (\phi_j)_k\, (\phi_i)_k$. However, this integrand is bounded by μs, hence altogether (3.4.64) is preserved with another constant instead of $\hat{c}_1$ and still tends to $-\infty$. In the other parts of the proof we only need the sum of $B(u)$ and $R(u)$, in which the additional terms vanish by definition.

Finally, Theorem 3.4.10 contains the boundedness of $M_R(\|u^h\|)$, see the end of its proof. It was used to have uniform monotonicity of A in order to prove that

$$\langle A(u^h) - A(g_h), u^h - g_h\rangle \geq m\,\|u^h - g_h\|^2\,, \tag{3.4.134}$$

since this implies the boundedness of $\|u^h\|$ if we assume the boundedness of $\|g_h\|$ (as $h \to 0$). Now we have $g_h = 0$ by the homogeneous Dirichlet data in (3.4.125), hence we only need (3.4.134) for the special case $g_h = 0$. Therefore, to prove our theorem, it suffices instead to verify

$$\langle A(u^h), u^h\rangle \geq \tilde{m}\,\|u^h\|^2 \qquad (h > 0) \tag{3.4.135}$$

for some constant $\tilde{m} > 0$, independent of the FEM solution u^h of our problem.

Since $u^h = 0$ on Γ_D, we can substitute $u = v = u^h$ in (3.4.127):

$$\langle A(u^h), u^h\rangle = \int_\Omega \Big(\sum_{k=1}^s b_k(x, \nabla u^h)\,|\nabla u_k^h|^2 + \mu|u_k^h|^2 + \sum_{k=1}^s (\mathbf{w}_k(x,u^h)\cdot\nabla u_k^h)\,u_k^h\Big) \tag{3.4.136}$$

$$+ \int_\Omega \sum_{k,l=1}^s \Big(V_{kl}(x,u^h)\,u_l^h\,u_k^h - \mu|u_k^h|^2\Big). \tag{3.4.137}$$

We can estimate (3.4.136) in the same way as in (3.4.131), and obtain the lower bound $\hat{m}\,\|u^h\|^2$ where $\hat{m} := m - \frac{\omega^2}{4\mu} > 0$. For (3.4.137), note that (3.4.126) and (3.4.77) imply that μ is a lower uniform spectral bound for the matrices $V(x,\xi)$, i.e.

$$V(x,\xi)\,\zeta\cdot\zeta \equiv \sum_{k,l=1}^s V_{kl}(x,\xi)\,\zeta_l\,\zeta_k \geq \mu|\zeta|^2 \tag{3.4.138}$$

(for all $(x,\xi) \in \Omega \times \mathbf{R}^s$ and $\zeta \in \mathbf{R}^s$), which yields that the expression in (3.4.137) is nonnegative. Altogether, (3.4.135) holds with $\tilde{m} := \hat{m}$. ■

As before, we can derive the corresponding DMP under the conditions of Theorem 3.4.12. Since now $g = 0$, this becomes the discrete nonpositivity property $u_k^h \leq 0$. By reversing signs, one similarly obtains the discrete nonnegativity property, which is more noteworthy to formulate here:

Corollary 3.4.5 *Let problem (3.4.125) satisfy Assumptions 3.4.12, and let its FEM discretization satisfy the corresponding conditions of Theorem 3.4.12. If $f_k \geq q_k(x,0)$ and $\gamma_k \geq 0$ $(k = 1,\dots,s)$, then for sufficiently small h, the FEM solution $u^h = (u_1^h,\dots,u_s^h)^T$ of system (3.4.125) satisfies*

$$u_k^h \geq 0 \quad \textit{on } \Omega \qquad (k = 1,\dots,s). \tag{3.4.139}$$

3.5 Some applications

3.5.1 DMP for model equations

(a) Nonnegativity properties for semilinear reaction-diffusion equations

In various model problems the solution has to satisfy the sign condition $u \geq 0$ to have a physical meaning. Therefore, the same is required for the discretized problem. Since the nonlinearity in such problems is only defined for nonnegative arguments u, one has to extend it for $u \le 0$. In the case $q(x,0) = 0$, this is done, e.g., by the formula $q(x,-u) := -q(x,u)$. Then q is increasing in u, and Theorem 3.4.5 yields the fact that the discrete solution satisfies $\min_{\overline{\Omega}} u_h \ge 0$. That is, u_h is the solution of the problem with the original nonlinearity and preserves the physical meaning.

We give three examples when such a procedure is valid. These problems are semilinear, i.e. they have a linear principal part, further, the examples involve both Dirichlet and Robin boundary conditions. The formulations of these problems can be found in [41, 99].

(i) Autocatalytic chemical reactions. The problem

$$\begin{cases} -\Delta u + u^p = 0 & \text{in } \Omega, \\ \qquad\quad u = 1 & \text{on } \partial\Omega \end{cases} \tag{3.5.1}$$

in a planar domain $\Omega \subset \mathbf{R}^2$ with some $p \ge 1$ describes a chemical reaction-diffusion process where the reaction is autocatalytic, i.e. the growth of the concentration $u \ge 0$ speeds up the rate of the reaction.

(ii) Diffusion-kinetic enzyme problems. The steady-state concentration $u \ge 0$ of the substrate in a cell $\Omega \subset \mathbf{R}^3$ satisfies

$$\begin{cases} -\mathrm{div}\,(d(x)\,\nabla u) + \dfrac{1}{\varepsilon}\dfrac{u}{u+k} = 0 & \text{in } \Omega, \\ d(x)\frac{\partial u}{\partial \nu} + h(x)\,(u - u_0(x)) = 0 & \text{on } \partial\Omega\,, \end{cases} \tag{3.5.2}$$

where $d(x) > 0$ is the molecular diffusion coefficient, $k > 0$ is the Michaelis constant and $\varepsilon > 0$, $h(x) > 0$ is the permeability of the membrane and $u_0(x) > 0$ is the external concentration. The nonlinearity describes the rate of the enzyme-substrate reaction by the Michaelis-Menten rule.

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

$$\begin{cases} -\mathrm{div}\,(\kappa(x)\,\nabla u) + \sigma(x)u^4 = 0 & \text{in } \Omega, \\ \kappa(x)\frac{\partial u}{\partial \nu} + \alpha(x)\,(u - \tilde{u}(x)) = 0 & \text{on } \partial\Omega\,, \end{cases} \tag{3.5.3}$$

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.

For each of the above problems it is easy to check that the coefficients satisfy the conditions of Theorem 3.4.3 and hence those of Theorem 3.4.5. Therefore, for any FEM discretization with the acuteness property (3.4.15) given there, Theorem 3.4.5 provides the physically meaningful numerical solution. That is:

Corollary 3.5.1 *Let u_h be the FEM solution to one of the problems (3.5.1)–(3.5.3) under a FEM discretization with the acuteness property (3.4.15). If h is sufficiently small then*

$$\min_{\overline{\Omega}} u_h \ge 0.$$

(b) Subsonic potential flow

A typical example of nonlinear elliptic problem with both Dirichlet and Neumann boundaries is related to the subsonic potential flow equation. The behaviour of potential flows has been studied in several works, see e.g. [21] and the references therein. The subsonic potential flow in a wind tunnel section $\Omega \subset \mathbf{R}^2$ is described by the boundary value problem

$$\left\{ \begin{aligned} -\text{div } (\varrho(|\nabla u|^2)\, \nabla u) &= 0 && \text{in } \Omega\,, \\ \varrho(|\nabla u|^2)\, \tfrac{\partial u}{\partial \nu} &= \gamma(x) && \text{on } \Gamma_N\,, \\ u &= \omega(x) && \text{on } \Gamma_D \end{aligned} \right. \tag{3.5.4}$$

with the nonlinearity $\varrho(|\nabla u|^2) = \varrho_0 \left(1+\frac{1}{5}(M^2-|\nabla u|^2)\right)^{5/2}$, where M and ϱ_0 are the Mach number and the air density at infinity, respectively. In the case of the subsonic flow there holds $\sup_\Omega |\nabla u| < 1$. The boundary portion Γ_N consists of disjoint subparts $\Gamma_N^{(0)}$ and $\Gamma_N^{(1)}$ (the sides and the end of the wind tunnel section, respectively) such that $\gamma = 0$ on $\Gamma_N^{(0)}$ and $\gamma = c_\infty$ on $\Gamma_N^{(1)}$ where the constant $c_\infty > 0$ is the wind outlet velocity. On Γ_D the function ω describes the wind inblow. Then the minimum of ω is a lower bound for u, i.e., $\min_{\overline{\Omega}} u = \min_{\Gamma_D} \omega$.

This minimization property is preserved by appropriate FEM discretizations due to statement (2) of Theorem 3.4.6. In fact, owing to the special form of the problem, it suffices here to have the nonobtuseness property (3.4.31) instead of (3.4.15):

Corollary 3.5.2 *Let u_h be the FEM solution of problem (3.5.4) under a FEM discretization with the nonobtuseness property (3.4.31). Then*

$$\min_{\overline{\Omega}} u_h = \min_{\Gamma_D} \omega_h\ .$$

3.5.2 Discrete nonnegativity for systems

(a) Reaction-diffusion systems in chemistry

The steady states of certain reaction-diffusion processes in chemistry are described by systems of the following form:

$$\left. \begin{aligned} -b_k\Delta u_k + \ P_k(x,u_1,\dots,u_s) &= f_k(x) && \text{in } \Omega, \\ b_k \tfrac{\partial u_k}{\partial \nu} &= \gamma_k(x) && \text{on } \Gamma_N, \\ u_k &= g_k(x) && \text{on } \Gamma_D \end{aligned} \right\} \qquad (k=1,\dots,s). \tag{3.5.5}$$

Here, for all k, the quantity u_k describes the concentration of the kth species, and P_k is a polynomial which characterizes the rate of the reactions involving the k-th species. A common way to describe such reactions is the so-called mass action type kinetics [72], which implies that P_k has no constant term for any k, in other words, $P_k(x,0) \equiv 0$ on Ω for all k. The reaction between different species is often proportional to the product of their concentration. The function $f_k \geq 0$ describes a source independent of concentrations.

We consider system (3.5.5) under the following conditions, such that it becomes a special case of system (3.4.82). As pointed out later, such chemical models describe processes with cross-catalysis and strong autoinhibiton.

Assumptions 3.5.2.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain, where $d = 2$ or 3, and 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 $k, l = 1, \dots, s$, the functions P_k are polynomials of arbitrary degree if $d = 2$ and of degree at most 4 if $d = 3$, further, $P_k(x, 0) \equiv 0$ on Ω.

(iii) (Ellipticity.) $b_k > 0$ $(k = 1, \dots, s)$ are given numbers.

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

(v) (Weak diagonal dominance for the Jacobians w.r.t. rows and columns.) We have

$$\sum_{l=1}^{s} \frac{\partial P_k}{\partial \xi_l}(x, \xi) \ge 0, \qquad \sum_{l=1}^{s} \frac{\partial P_l}{\partial \xi_k}(x, \xi) \ge 0 \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \quad (3.5.6)$$

(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 $g^* \in H^1(\Omega)$.

Similarly to (3.4.76), assumptions (iv)–(v) now imply

$$\frac{\partial P_k}{\partial \xi_k}(x, \xi) \ge 0 \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \quad (3.5.7)$$

Returning to the model described by system (3.5.5), the chemical meaning of the cooperativity is cross-catalysis, whereas (3.5.7) means autoinhibiton. Cross-catalysis arises e.g. in gradient systems [145]. Condition (3.5.6) means that autoinhibition is strong enough to ensure both weak diagonal dominances.

By definition, the concentrations u_k are nonnegative, therefore a proper numerical model must produce such numerical solutions. We can use Corollary 3.4.4 to obtain the required property:

Corollary 3.5.3 *Let problem (3.5.5) satisfy Assumptions 3.5.2, and let its FEM discretization satisfy the conditions of Theorem 3.4.10. If* $f_k \ge 0$, $\gamma_k \ge 0$, $g_k \ge 0$ $(k = 1, \dots, s)$ *then for sufficiently small h, the FEM solution* $u^h = (u_1^h, \dots, u_s^h)^T$ *of system (3.5.5) satisfies*

$$u_k^h \ge 0 \quad \text{on } \Omega \qquad (k = 1, \dots, s). \quad (3.5.8)$$

(b) Linear elliptic systems

Maximum principles or nonnegativity preservation for linear elliptic systems have attracted great interest, as mentioned in the introduction. Hence it is worthwile to derive the corresponding DMPs from the previous results. Let us therefore consider linear elliptic systems of the form

$$\left.\begin{aligned} -\mathrm{div}\,(b_k(x)\,\nabla u_k) + \sum_{l=1}^{s} V_{kl}(x)\,u_l \;&=\; f_k(x) \quad \text{a.e. in } \Omega, \\ b_k(x)\tfrac{\partial u_k}{\partial \nu} \;&=\; \gamma_k(x) \quad \text{a.e. on } \Gamma_N, \\ u_k \;&=\; g_k(x) \quad \text{a.e. on } \Gamma_D \end{aligned}\right\} \qquad (k = 1, \dots, s) \qquad (3.5.9)$$

where for all $k, l = 1, \dots, s$ we have $b_k \in W^{1,\infty}(\Omega)$ and $V_{kl} \in L^\infty(\Omega)$.

Let Assumptions 3.4.7 hold. Then (3.5.9) is a special case of (3.4.32), hence Corollary 3.4.9 holds, as well as the analogous results mentioned in Remark 3.4.4. Here we formulate two of these:

Corollary 3.5.4 *Let problem (3.5.9) satisfy Assumptions 3.4.7, let its FEM discretization satisfy the conditions of Theorem 3.4.7 and let h be sufficiently small. If* $u^h = (u_1^h, \dots, u_s^h)^T$ *is the FEM solution of system (3.5.9), then the following properties hold.*

(1) If $f_k \le 0$, $\gamma_k \le 0$ *(*$k = 1, \dots, s$*) and* $\max\limits_{k=1,\dots,s} \max\limits_{\Gamma_D} g_k^h > 0$, *then*

$$\max_{k=1,\dots,s} \max_{\overline{\Omega}} u_k^h = \max_{k=1,\dots,s} \max_{\Gamma_D} g_k^h . \tag{3.5.10}$$

(2) If $f_k \ge 0$, $\gamma_k \ge 0$ *and* $g_k \ge 0$ *(*$k = 1, \dots, s$*), then*

$$u_k^h \ge 0 \quad \text{on } \Omega \qquad (k = 1, \dots, s). \tag{3.5.11}$$

(c) Nonsymmetric transport systems

The description of nonlinear transport processes for certain agents (pollutants), involving diffusion, convection and reaction, often leads to systems of the form

$$\left.\begin{array}{rcll} -b_k \Delta u_k + \mathbf{w}_k(x,u) \cdot \nabla u_k + P_k(x, u_1, ..., u_s) &=& f_k(x) & \text{a.e. in } \Omega, \\ b_k \frac{\partial u_k}{\partial \nu} &=& \gamma_k(x) & \text{a.e. on } \Gamma_N, \\ u_k &=& 0 & \text{a.e. on } \Gamma_D \end{array}\right\} \tag{3.5.12}$$

($k = 1, \dots, s$). We consider diffusion-dominated processes, i.e. when the fixed numbers $b_k > 0$ are comparable to the magnitude of the coefficients $\mathbf{w}_k$. Here $u_k \ge 0$ are the concentrations of the agents. One expects any numerical solution method to reproduce the nonnegativity of the solution.

Assumptions 3.5.2.

(i) The numbers b_k and functions P_k, f_k and γ_k satisfy Assumptions 3.5.2.

(ii) We have $\mathbf{w}_k \in L^\infty(\Omega \times \mathbf{R})$ $(k = 1, \dots, s)$.

(iii) There exists $\mu > 0$ such that

$$\sum_{l=1}^{s} \frac{\partial P_k}{\partial \xi_l}(x, \xi) \ge \mu, \qquad \sum_{l=1}^{s} \frac{\partial P_l}{\partial \xi_k}(x, \xi) \ge \mu \qquad (k = 1, \dots, s;\ x \in \Omega,\ \xi \in \mathbf{R}^s). \tag{3.5.13}$$

Moreover,

$$\mu > \frac{\|\mathbf{w}\|^2_{L^\infty(\Omega)^s}}{4m} \tag{3.5.14}$$

where $\|\mathbf{w}\|_{L^\infty(\Omega)^s} := \sup\limits_{\substack{k=1,\dots,s \\ (x,\xi)\in\Omega\times\mathbf{R}^s}} |\mathbf{w}_k(x, \xi)|$ and $m := \min_k b_k > 0$.

Systems of the form (3.5.12) typically arise from the time discretization of the time-dependent transport system

$$\frac{\partial u_k}{\partial t} - b_k \Delta u_k + \mathbf{w}_k(x,u) \cdot \nabla u_k + R_k(x, u_1, ..., u_s) \;=\; g_k(x,t) \tag{3.5.15}$$

with the boundary conditions of (3.5.12) and an initial condition $u_k(x,0) = u_0(x)$ $(x \in \Omega)$. Here $\mathbf{w}_k(x,u)$ is the convective term, e.g. wind, and R_k is a polynomial which characterizes the rate of the reactions involving the k-th species. Here the R_k do not satisfy a condition like (3.5.13), this will come instead from the numerical process below.

The standard numerical solution first uses a time discretization, resulting in the following equations, where u_k^i denotes the solution on the ith time level t_i:

$$\frac{u_k^i - u_k^{i-1}}{\tau} - b_k \Delta u_k^i + \mathbf{w}_k(x,u^i) \cdot \nabla u_k^i + R_k(x, u_1^i, ..., u_s^i) \;=\; g_k^i(x)\,.$$

Rearranging this as

$$-b_k \Delta u_k^i + \mathbf{w}_k(x,u^i) \cdot \nabla u_k^i + \Big(R_k(x, u_1^i, ..., u_s^i) \,+ \frac{1}{\tau}\, u_k^i \Big) \;=\; g_k^i(x) + \frac{1}{\tau}\, u_k^{i-1},$$

we obtain a system for the unknown function u_k^i in the form (3.5.12) with coefficients

$$P_k(x, \xi_1, ..., \xi_s) := R_k(x, \xi_1, ..., \xi_s) \,+ \frac{1}{\tau}\, \xi_k \tag{3.5.16}$$

and $f_k(x) := g_k^i(x) + \frac{1}{\tau}\, u_k^{i-1}(x)$. Then the strong uniform diagonal dominance (3.5.13)–(3.5.14) can be ensured as follows. Assume that we have an estimate

$$\inf_{\substack{k=1,\dots,s \\ (x,\xi)\in\Omega\times\mathbf{R}^s}} \sum_{l=1}^{s} \frac{\partial R_k}{\partial \xi_l}(x,\xi) \geq -\mu_0, \qquad \inf_{\substack{k=1,\dots,s \\ (x,\xi)\in\Omega\times\mathbf{R}^s}} \sum_{l=1}^{s} \frac{\partial R_l}{\partial \xi_k}(x,\xi) \geq -\mu_0$$

for some $\mu_0 \geq 0$, and let μ be a number satisfying (3.5.14). Then we can choose the time-step τ to be small enough, namely, $\tau \leq \frac{1}{\mu_0 + \mu}$. In this case, using (3.5.16), we obtain

$$\sum_{l=1}^{s} \frac{\partial P_k}{\partial \xi_l}(x,\xi) \geq -\mu_0 + \frac{1}{\tau} \geq -\mu_0 + (\mu_0 + \mu) = \mu,$$

and similarly for the other sum in (3.5.13).

Under the above conditions, system (3.5.12) is a special case of system (3.4.125), hence we can apply Corollary 3.4.5. Here, as mentioned in subsection 3.5.2, $P_k(x,0) \equiv 0$ on Ω for all k, further, we have homogeneous Dirichlet boundary conditions. Hence the result has the following form:

Corollary 3.5.5 *Let problem (3.5.12) satisfy Assumptions 3.5.2, and let its FEM discretization satisfy the corresponding conditions of Theorem 3.4.12. If $f_k \geq 0$ and $\gamma_k \geq 0$ $(k = 1, \dots, s)$, then for sufficiently small h, the FEM solution $u^h = (u_1^h, \dots, u_s^h)^T$ of system (3.5.12) satisfies*

$$u_k^h \geq 0 \quad \text{on } \Omega \qquad (k = 1, \dots, s). \tag{3.5.17}$$

Chapter 4

A posteriori error estimates

The reliability of the computer solution of a (linear or nonlinear) elliptic problem requires a verification of the accuracy for the computed approximations. This leads to a posteriori error estimation methods. Several approaches have been suggested for this (see e.g. [2, 126, 138] and the references there), which use the fact that the computed solutions are true finite element (FE) approximations. A different approach, based on functional analysis background, has been developed in [122], see also the references therein. Hereby the estimation is developed independently of the applied numerical method. One can thus obtain sharp estimates for linear problems and for certain nonlinear problems; however, for nonlinear problems in general, these estimates may fail to ensure the best upper bound [122]. This method has been further developed for nonsymmetric linear problems [109].

In this chapter we follow the second approach and extend it to give sharp a posteriori error estimates for nonlinear variational operator equations. Then we apply the developed framework to various classes of elliptic problems. To allow more generality, as long as no Hilbert space structure needs to be used, the general background will be given in Banach spaces. However, as we will see, the practical realization will exploit that the base space is a Hilbert space.

4.1 Basic properties

(a) Some elementary definitions and properties. Let V be a given Banach space with norm $\|.\|_V$. Then its dual space V^* consists of all bounded linear functionals $l : V \to \mathbf{R}$ on V. If $l \in V^*$ and $u \in V$, then the value of l at u is denoted by $\langle l, u\rangle$, where $\langle ., .\rangle$ is the duality pairing.

We consider operator equations of the form

$$F(u) + l = 0 \tag{4.1.1}$$

in a Banach space V with a given nonlinear operator $F : V \to V^*$ and a given bounded linear functional $l \in V^*$. We will assume certain monotonicity properties of F that both ensure well-posedness for (4.1.1) and allow a suitable measuring of the error. It is well-known [158] that certain properties, used earlier in this thesis in a Hilbert space, can be defined in the same way for operators $F : V \to V^*$. In particular, one can define (uniformly) monotone operators, bihemicontinuous Gateaux derivatives, and the following results hold (see, e.g., [158]):

Proposition 4.1.1 *Let the operator $F : V \to V^*$ have a bihemicontinuous Gateaux derivative.*

(1) F is a potential operator if and only if $F'(u)$ is symmetric for any $u \in V$.

(2) If the above holds and there exists a constant $m > 0$ such that

$$\langle F'(u)v, v\rangle \geq m\|v\|_V^2 \qquad (u, v \in V),$$

then for any $l \in V^*$ the operator equation (4.1.1) *has a unique solution $u^* \in V$.*

The solution u^* is the unique minimizer of the functional $J(u) := \phi(u) + \langle l, u\rangle$, where ϕ is a potential.

(b) Error functionals for monotone operators. Let us assume that the operator equation (4.1.1) has a unique solution $u^* \in V$. (For sufficient conditions, see Proposition 4.1.1 or later in Section 4.2.) We consider some approximate solution $u \in V$ of equation (4.1.1), i.e. $u \approx u^*$ where u^* is the exact solution. 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 (4.1.1):

$$E(u) := \langle F(u) + l,\, u - u^*\rangle \qquad (u \in V) \tag{4.1.2}$$

or in other form

$$E(u) = \langle F(u) - F(u^*),\, u - u^*\rangle \qquad (u \in V). \tag{4.1.3}$$

The following facts obviously hold. If F is monotone then $E(u) \geq 0 = E(u^*) \quad (u \in V)$. If F is also strictly monotone then $E(u) = 0$ if and only if $u = u^*$. If F is also uniformly monotone then

$$E(u) \geq m\|u - u^*\|_V^2 \qquad (u \in V). \tag{4.1.4}$$

(c) Integral mean operators. Let Y be a Banach space and $A : Y \to Y^*$ an operator having a bihemicontinuous symmetric Gateaux derivative.

Definition 4.1.1 For any vectors $y, z \in Y$, we define $A'_{[y,z]} \in B(Y, Y^*)$, that is, a bounded linear operator $A'_{[y,z]} : Y \to Y^*$, by the formula

$$A'_{[y,z]} := \int_0^1 A'(y + t(z - y))\, dt\ . \tag{4.1.5}$$

This is an integral of a family of operators, understood via the corresponding bilinear forms:

$$\langle A'_{[y,z]}p, q\rangle = \int_0^1 \langle A'(y + t(z - y))p, q\rangle\, dt \qquad (p, q \in Y). \tag{4.1.6}$$

The unique existence of $A'_{[y,z]}$ (i.e., that this definition is correct) is ensured by the fact that

$$\int_0^1 \langle A'(y + t(z - y))p, q\rangle\, dt\ \leq\ \Big(\max_{t\in[0,1]} \|A'(y + t(z - y))\|\Big)\, \|p\|_Y\, \|q\|_Y \tag{4.1.7}$$

(where the maximum exists by the continuity of the mapping $t \mapsto A'(y + t(z - y))$ and of the operator norm), which means that the r.h.s. of (4.1.6) is a bounded bilinear form in p and q. Then we obtain by the definition of Y^* that this bilinear form can be represented as the bilinear form of a bounded linear operator from Y to Y^*.

Proposition 4.1.2 *For any $y, z \in Y$*

(i) the operator $A'_{[y,z]}$ is symmetric, i.e.,

$$\langle A'_{[y,z]}p, q\rangle = \langle A'_{[y,z]}q, p\rangle \qquad (p, q \in Y). \tag{4.1.8}$$

(ii) $A'_{[y,z]} = A'_{[z,y]}$.

(iii) $A(z) - A(y) = A'_{[y,z]}(z - y)$.

The proof follows directly from the definition. Due to the last equality, the operator $A'_{[y,z]}$ is a so-called divided difference.

4.2 A sharp global error estimate in Banach space

In what follows, our goal is to find upper bounds for $E(u)$. Following the setting of [101, 122], let $u \in V$ be arbitrary and look for a bound involving some other vector parameters.

Let $J : V \to \mathbf{R}$ be a functional of the form

$$J(u) := G(\Lambda u) + \langle l, u\rangle \qquad (u \in V) \tag{4.2.1}$$

under the following conditions:

Assumptions 4.2.1.

(i) Y is another Banach space and $\Lambda : V \to Y$ is a linear operator for which

$$\|\Lambda u\|_Y = \|u\|_V \qquad (u \in V), \tag{4.2.2}$$

(ii) $G : Y \to \mathbf{R}$ is a functional having a bihemicontinuous symmetric second Gateaux derivative (according to Definition 2.2.1),

(iii) there exists a constant $m > 0$ such that $\langle G''(y)p, p\rangle \geq m\,\|p\|_Y^2 \qquad (y, p \in Y)$.

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

$$\langle F(u), v\rangle = \langle G'(\Lambda u), \Lambda v\rangle \qquad (u, v \in V). \tag{4.2.3}$$

Proposition 4.2.1 *Under Assumptions 4.2.1, for any $l \in V^*$ the operator equation* (4.1.1) *has a unique solution $u^* \in V$.*

Proof. The assumptions yield that F has a bihemicontinuous symmetric Gateaux derivative that satisfies

$$\langle F'(u)v, v\rangle = \langle G''(\Lambda u)\Lambda v, \Lambda v\rangle \geq m\,\|\Lambda v\|_Y^2 = m\|v\|_V^2 \qquad (u, v \in V). \tag{4.2.4}$$

Then Proposition 4.1.1 implies well-posedness for (4.1.1). ■

We note that the solution u^* of (4.1.1) is the unique minimizer of J. However, from now on, our calculations will involve the operator G' in (4.2.3) rather than the functional G. Hence we study below the solution of equation (4.1.1) directly, instead of using the corresponding minimization problem.

Now we will replace the minimization problem for (4.2.1) by the corresponding operator equation, which is a more detailed form of (4.1.1) for this case. For this purpose, we introduce the operator

$$A := G'. \tag{4.2.5}$$

Then Assumptions 4.2.1 are equivalent to

Assumptions 4.2.2.

(i) Y is another Banach space and $\Lambda : V \to Y$ is a linear operator for which

$$\|\Lambda u\|_Y = \|u\|_V \qquad (u \in V); \tag{4.2.6}$$

(ii) the operator $A : Y \to Y^*$ has a bihemicontinuous symmetric Gateaux derivative (according to Definition 2.2.1);

(iii) there exists a constant $m > 0$ such that

$$\langle A'(y)p, p\rangle \geq m\,\|p\|_Y^2 \qquad (y, p \in Y); \tag{4.2.7}$$

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

$$\langle F(u), v\rangle = \langle A(\Lambda u), \Lambda v\rangle \qquad (u, v \in V). \tag{4.2.8}$$

Assumptions (ii)-(iii) imply in particular that A is bijective, i.e. $A^{-1} : Y^* \to Y$ exists. By (4.2.8), equation (4.1.1) can be written as

$$\langle A(\Lambda u), \Lambda v\rangle + \langle l, v\rangle = 0 \qquad (v \in V) \tag{4.2.9}$$

which has a unique solution $u^* \in V$ for any $l \in V^*$ by Proposition 4.2.1.

We will need some further related properties. First, Proposition 4.1.2 (i) and (4.2.7) imply

Proposition 4.2.2 *Under Assumptions 4.2.2, for any $y, z \in Y$ the mapping $p, q \mapsto \langle A'_{[y,z]}p, q\rangle$ is an inner product on Y.*

Proposition 4.2.3 *Under Assumptions 4.2.2, the following properties hold:*

(i) $E(u) = \langle A'_{[\Lambda u^*,\Lambda u]}\,\Lambda(u-u^*),\ \Lambda(u-u^*)\rangle \qquad (u\in V).$

(ii) $E(u) \geq m\,\|u-u^*\|_V^2 = m\,\|\Lambda(u-u^*)\|_Y^2 \qquad (u\in V).$

(iii) $\|A(z)-A(y)\|_{Y^*} \geq m\,\|z-y\|_Y \qquad (y,z\in Y).$

Proof. (i) Using (4.2.8) and Proposition 4.1.2 (iii) for $z=\Lambda u$ and $y=\Lambda u^*$,

$$E(u) = \langle F(u)-F(u^*),\, u-u^*\rangle = \langle A(\Lambda u)-A(\Lambda u^*),\, \Lambda(u-u^*)\rangle \tag{4.2.10}$$
$$= \langle A'_{[\Lambda u^*,\Lambda u]}\,\Lambda(u-u^*),\ \Lambda(u-u^*)\rangle\,.$$

(ii) Estimate (4.2.4) implies that F is uniformly monotone, hence (4.1.4) and (4.2.6) yield the required statement.

(iii) Estimate (4.2.7) implies

$$\langle A(z)-A(y), z-y\rangle \geq m\,\|z-y\|_Y^2 \qquad (y,z\in Y), \tag{4.2.11}$$

whence we obtain the required statement by definition. ■

For the V^*-norm of a linear functional $l\in V^*$, we introduce the notation of [122]:

$$|\,l\,| := \|\,l\,\|_{V^*} \qquad (l\in V^*). \tag{4.2.12}$$

Here (4.2.6) yields

$$|l| = \sup_{w\in V}\frac{\langle l,\,w\rangle}{\|w\|_V} = \sup_{w\in V}\frac{\langle l,\,w\rangle}{\|\Lambda w\|_Y}\,. \tag{4.2.13}$$

Now we let $y^*\in Y^*$ be an arbitrary vector. We give a preliminary estimate, which is a starting point for our study.

Lemma 4.2.1 *Let Assumptions 4.2.2 hold and $u^*\in V$ be the solution of* (4.1.1). *Let $u\in V$ and $y^*\in Y^*$ be arbitrary, let $z^* := A^{-1}(y^*)$. Then*

$$E(u) \leq\ |\Lambda^* y^* + l|\, m^{-1/2}\, E(u)^{1/2}\ +\ \langle A'_{[z^*,\Lambda u]}\,(\Lambda u - z^*),\ \Lambda(u-u^*)\rangle\,. \tag{4.2.14}$$

Proof. We have

$$E(u) = \langle F(u)+l,\, u-u^*\rangle = \langle \Lambda^* y^* + l, u-u^*\rangle + \langle F(u)-\Lambda^* y^*, u-u^*\rangle\,. \tag{4.2.15}$$

For the first term, we use (4.2.12) and Proposition 4.2.3 (ii) to obtain

$$|\langle \Lambda^* y^* + l, u-u^*\rangle| \leq |\Lambda^* y^* + l|\ \|u-u^*\|_V \leq |\Lambda^* y^* + l|\, m^{-1/2}\, E(u)^{1/2}\,. \tag{4.2.16}$$

The second term equals

$$\langle F(u)-\Lambda^* y^*,\, u-u^*\rangle = \langle A(\Lambda u)-y^*,\, \Lambda(u-u^*)\rangle = \langle A(\Lambda u)-A(z^*),\, \Lambda(u-u^*)\rangle$$
$$= \langle A'_{[z^*,\Lambda u]}\,(\Lambda u - z^*),\ \Lambda(u-u^*)\rangle \tag{4.2.17}$$

where (4.2.8) and Proposition 4.1.2 (iii) have been used. ■

The r.h.s. of (4.2.14) becomes computable if the second term is further estimated. A sharp estimation requires a further assumption on the Lipschitz continuity of the derivative of the nonlinear operator, hence we complete Assumptions 4.2.2 by additional conditions:

Assumptions 4.2.3.

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

(ii) There exists a constant $M > 0$ such that

$$\langle A'(y)p, p\rangle \le M\,\|p\|_Y^2 \qquad (y, p \in Y). \tag{4.2.18}$$

By Assumption 4.2.3-(i), there exists a constant $L > 0$ such that

$$|\langle (A'(z) - A'(y))\,w, p\rangle| \le L\,\|z-y\|_Y\|w\|_W\|p\|_Y \qquad (y, z, p \in Y,\ w \in W). \tag{4.2.19}$$

Lemma 4.2.2 *Let Assumption 4.2.3-(i) hold. Then the operators defined in* (4.1.5) *satisfy for all $y, v, z \in Y$*

$$\|A'_{[z,v]} - A'_{[y,v]}\|_{B(W,Y^*)} \le \tfrac{L}{2}\,\|z-y\|_Y. \tag{4.2.20}$$

Proof. We have

$$\|A'_{[z,v]} - A'_{[y,v]}\|_{B(W,Y^*)} \le \int_0^1 \|A'(z+t(v-z)) - A'(y+t(v-y))\|_{B(W,Y^*)}\,dt$$

$$\le L\,\textstyle\int_0^1 (1-t)\,\|z-y\|_Y\,dt \;=\; \tfrac{L}{2}\,\|z-y\|_Y. \qquad \blacksquare$$

In more detailed form (as in (4.2.19)), property (4.2.20) means that

$$|\langle (A'_{[z,v]} - A'_{[y,v]})w, p\rangle| \le \tfrac{L}{2}\,\|z-y\|_Y\|w\|_W\|p\|_Y \qquad (y, v, z, p \in Y,\ w \in W). \tag{4.2.21}$$

Assumption 4.2.3-(ii) implies that the upper analogue of Proposition 4.2.3 (iii) holds:

$$\|A(z) - A(y)\|_{Y^*} \le M\,\|z-y\|_Y \qquad (y, z \in Y). \tag{4.2.22}$$

Further, we will need the following inequality:

Lemma 4.2.3 *Let Assumptions 4.2.2-4.2.3 hold and $u^* \in V$ be the solution of* (4.1.1). *Let $y^* \in Y^*$ be arbitrary and $z^* := A^{-1}(y^*)$. Then for any $h \in V$*

$$\|z^* - \Lambda u^*\|_Y \le \tfrac{M}{m}\,\|z^* - \Lambda h\|_Y + \tfrac{1}{m}\,|\Lambda^* y^* + l|\,. \tag{4.2.23}$$

Proof. Let $w^* \in V$ satisfy $F(w^*) = \Lambda^* y^*$. By (4.2.8), w^* is the solution of equation

$$\langle A(\Lambda w^*), \Lambda v\rangle = \langle \Lambda^* y^*, v\rangle \qquad (v \in V). \tag{4.2.24}$$

We have

$$\|z^* - \Lambda u^*\|_Y \le \|z^* - \Lambda w^*\|_Y + \|\Lambda(w^* - u^*)\|_Y\,. \tag{4.2.25}$$

Here (4.2.24) implies

$$\langle A(\Lambda w^*),\, \Lambda v\rangle = \langle y^*,\, \Lambda v\rangle = \langle A(z^*),\, \Lambda v\rangle \qquad (v \in V),$$

that is

$$\langle A(z^*) - A(\Lambda w^*),\, \Lambda v\rangle = 0 \qquad (v \in V). \tag{4.2.26}$$

Using (4.2.11), (4.2.26) and (4.2.22), respectively, we obtain for any $h \in V$ that

$$m\,\|z^* - \Lambda w^*\|_Y^2 \le \langle A(z^*) - A(\Lambda w^*),\, z^* - \Lambda w^*\rangle = \langle A(z^*) - A(\Lambda w^*),\, z^* - \Lambda h\rangle$$
$$\le\; M\,\|z^* - \Lambda w^*\|_Y\,\|z^* - \Lambda h\|_Y\,,$$

that is,

$$\|z^* - \Lambda w^*\|_Y \le \tfrac{M}{m}\,\|z^* - \Lambda h\|_Y\,. \tag{4.2.27}$$

Further, using (4.2.6), (4.2.11), (4.2.24) and that u^* solves (4.2.9),

$$m\,\|w^* - u^*\|_V^2 = m\,\|\Lambda(w^* - u^*)\|_Y^2 \le \langle A(\Lambda w^*) - A(\Lambda u^*),\, \Lambda w^* - \Lambda u^*\rangle = \langle \Lambda^* y^* + l,\, w^* - u^*\rangle$$
$$\le |\Lambda^* y^* + l|\,\|w^* - u^*\|_V\,,$$

hence

$$\|w^* - u^*\|_V \le \tfrac{1}{m}\,|\Lambda^* y^* + l|\,. \tag{4.2.28}$$

Then (4.2.25), (4.2.27) and (4.2.28) give the desired estimate. ■

Now we can prove our main result.

Theorem 4.2.1 *Let Assumptions 4.2.2-4.2.3 hold and $u^* \in V$ be the solution of* (4.1.1). *Let $u \in V$ be an approximation of u^* such that $\Lambda u \in W$. Then for arbitrary $y^* \in Y^*$ such that $z^* := A^{-1}(y^*) \in W$ and for arbitrary $h \in V$,*

$$E(u) \le\; EST(u; y^*, h) := \Big(m^{-1/2}\,|\Lambda^* y^* + l| \;+\; \tfrac{L}{2}\,m^{-3/2}\,D(u; y^*, h) \tag{4.2.29}$$
$$+\Big(\langle A(\Lambda u) - y^*,\; \Lambda u - A^{-1}(y^*)\rangle \;+\; \tfrac{L}{2m}\,D(u; y^*, h)\,\|\Lambda u - A^{-1}(y^*)\|_Y\Big)^{1/2}\Big)^2\,,$$

where

$$D(u; y^*, h) := \Big(M\,\|A^{-1}(y^*) - \Lambda h\|_Y + |\Lambda^* y^* + l|\Big)\,\|\Lambda u - A^{-1}(y^*)\|_W\,. \tag{4.2.30}$$

Proof. Lemma 4.2.1 provides

$$E(u) \le\; |\Lambda^* y^* + l|\,m^{-1/2}\,E(u)^{1/2} \;+\; \langle A'_{[z^*,\Lambda u]}\,(\Lambda u - z^*),\; \Lambda(u - u^*)\rangle\,, \tag{4.2.31}$$

and our goal is to estimate the second term accurately. First, we observe that

$$\langle A'_{[z^*,\Lambda u]}\,(\Lambda u - z^*),\; \Lambda(u - u^*)\rangle$$
$$= \Big\langle \Big(A'_{[z^*,\Lambda u]} - A'_{[\Lambda u^*,\Lambda u]}\Big)\,(\Lambda u - z^*),\; \Lambda(u - u^*)\Big\rangle + \langle A'_{[\Lambda u^*,\Lambda u]}\,(\Lambda u - z^*),\; \Lambda(u - u^*)\rangle\,. \tag{4.2.32}$$

Using (4.2.21), the first term of (4.2.32) satisfies

$$\Big\langle \Big(A'_{[z^*,\Lambda u]} - A'_{[\Lambda u^*,\Lambda u]} \Big) (\Lambda u - z^*),\ \Lambda(u-u^*) \Big\rangle \ \le\ \tfrac{L}{2} \|z^* - \Lambda u^*\|_Y \, \|\Lambda u - z^*\|_W \, \|\Lambda(u-u^*)\|_Y \tag{4.2.33}$$

where $\|z^* - \Lambda u^*\|_Y$ fulfils (4.2.23) and $\|\Lambda(u-u^*)\|_Y \le m^{-1/2} \, E(u)^{1/2}$ by Proposition 4.2.3 (ii), hence

$$\begin{aligned} &\Big\langle \Big(A'_{[z^*,\Lambda u]} - A'_{[\Lambda u^*,\Lambda u]} \Big) (\Lambda u - z^*),\ \Lambda(u-u^*) \Big\rangle \\ &\le\ \tfrac{L}{2} \, m^{-3/2} \Big(M \, \|z^* - \Lambda h\|_Y + |\Lambda^* y^* + l| \Big) \, \|\Lambda u - z^*\|_W \, E(u)^{1/2} \,. \end{aligned} \tag{4.2.34}$$

The second term of (4.2.32) can be estimated with the Cauchy-Schwarz inequality:

$$\begin{aligned} &\langle A'_{[\Lambda u^*,\Lambda u]} (\Lambda u - z^*),\ \Lambda(u-u^*) \rangle \\ &\le\ \langle A'_{[\Lambda u^*,\Lambda u]} (\Lambda u - z^*),\ \Lambda u - z^* \rangle^{1/2} \, \langle A'_{[\Lambda u^*,\Lambda u]} \Lambda(u-u^*),\ \Lambda(u-u^*) \rangle^{1/2} \,. \end{aligned} \tag{4.2.35}$$

Proposition 4.2.3 (i) states that the second factor of (4.2.35) equals $E(u)^{1/2}$. For the first factor,

$$\begin{aligned} &\langle A'_{[\Lambda u^*,\Lambda u]} (\Lambda u - z^*),\ \Lambda u - z^* \rangle \\ &= \langle A'_{[z^*,\Lambda u]} (\Lambda u - z^*),\ \Lambda u - z^* \rangle \ +\ \langle (A'_{[\Lambda u^*,\Lambda u]} - A'_{[z^*,\Lambda u]}) (\Lambda u - z^*),\ \Lambda u - z^* \rangle \,. \end{aligned} \tag{4.2.36}$$

Here Proposition 4.1.2 (iii) yields

$$\begin{aligned} \langle A'_{[z^*,\Lambda u]} (\Lambda u - z^*),\ \Lambda u - z^* \rangle \ &= \langle A(\Lambda u) - A(z^*),\ \Lambda u - z^* \rangle \\ &= \langle A(\Lambda u) - y^*,\ \Lambda u - A^{-1}(y^*) \rangle \end{aligned} \tag{4.2.37}$$

and (4.2.21) and (4.2.23) imply

$$\begin{aligned} \langle (A'_{[\Lambda u^*,\Lambda u]} - A'_{[z^*,\Lambda u]}) (\Lambda u - z^*),\ \Lambda u - z^* \rangle \ &\le\ \tfrac{L}{2} \, \|\Lambda u^* - z^*\|_Y \, \|\Lambda u - z^*\|_W \, \|\Lambda u - z^*\|_Y \\ &\le \tfrac{L}{2m} \Big(M \, \|z^* - \Lambda h\|_Y + |\Lambda^* y^* + l| \Big) \, \|\Lambda u - z^*\|_W \, \|\Lambda u - z^*\|_Y \,. \end{aligned} \tag{4.2.38}$$

Summing up, (4.2.31), (4.2.32), (4.2.34), (4.2.36), (4.2.37) and (4.2.38) yield

$$\begin{aligned} E(u)^{1/2} \le \ \ & m^{-1/2} \, |\Lambda^* y^* + l| \ +\ \tfrac{L}{2} \, m^{-3/2} \Big(M \, \|z^* - \Lambda h\|_Y + |\Lambda^* y^* + l| \Big) \, \|\Lambda u - z^*\|_W \\ &+ \Big(\langle A(\Lambda u) - y^*,\ \Lambda u - A^{-1}(y^*) \rangle \ +\ \tfrac{L}{2m} \Big(M \, \|z^* - \Lambda h\|_Y + |\Lambda^* y^* + l| \Big) \, \|\Lambda u - z^*\|_W \, \|\Lambda u - z^*\|_Y \Big)^{1/2} \\ = \ \ & m^{-1/2} \, |\Lambda^* y^* + l| \ +\ \tfrac{L}{2} \, m^{-3/2} \, D(u; y^*, h) \\ &+ \big(\langle A(\Lambda u) - y^*,\ \Lambda u - A^{-1}(y^*) \rangle \ +\ \tfrac{L}{2m} \, D(u; y^*, h) \, \|\Lambda u - z^*\|_Y \big)^{1/2} \,. \qquad \blacksquare \end{aligned}$$

Remark 4.2.1 It is convenient to reformulate Theorem 4.2.1 for $z^* = A^{-1}(y^*)$ in order to avoid A^{-1}. Then for arbitrary $z^* \in W$ and for arbitrary $h \in V$,

$$E(u) \leq \; E\tilde{S}T(u; z^*, h) := \Big(m^{-1/2} \, |\Lambda^* A(z^*) + l| \, + \, \tfrac{L}{2} \, m^{-3/2} \, \tilde{D}(u; z^*, h) \tag{4.2.39}$$

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

where

$$\tilde{D}(u; z^*, h) := \Big(M \, \|z^* - \Lambda h\|_Y + \, |\Lambda^* A(z^*) + l| \Big) \, \|\Lambda u - z^*\|_W \, . \tag{4.2.40}$$

Now we can turn to the problem of sharpness.

Proposition 4.2.4 *Estimate*(4.2.29) *is sharp in the following sense: assuming* $\Lambda u^* \in W$*, and denoting* $A(W) := \{A(v) : \; v \in W\}$*, we have*

$$\min_{\substack{y^* \in A(W), \\ h \in V}} EST(u; y^*, h) = E(u).$$

Proof. Let us choose

$$y^* := A(\Lambda u^*) \quad \text{and} \quad h := u^*. \tag{4.2.41}$$

Then $z^* = A^{-1}(y^*) = \Lambda u^* \in W$, hence this y^* satisfies the assumption of Theorem 4.2.1. Here $y^* = A(\Lambda u^*)$ satisfies $\Lambda^* y^* + l = 0$, similarly to the linear case. Hence the first term in $EST(u; y^*, h)$ is zero in this case, further, $A^{-1}(y^*) - \Lambda h = \Lambda u^* - \Lambda u^* = 0$, therefore $D(u; A(\Lambda u^*), \Lambda u^*) = 0$ and thus the terms containing $D(u; y^*, h)$ are also zero in this case. That is,

$$EST(u; A(\Lambda u^*), \Lambda u^*) = \langle A(\Lambda u) - A(\Lambda u^*), \; \Lambda u - \Lambda u^* \rangle = E(u)$$

where (4.2.10) has been used. ■

Remark 4.2.2 (Finding the optimal h in a Hilbert space.) In practice, y^* is obtained as an approximation of the optimal unknown value $A(\Lambda u^*)$ (cf. (4.2.41)). For given y^*, one can determine the optimal h via projection when Y is a Hilbert space. (In this case $\langle ., . \rangle$ means inner product.) This is achieved as follows. Let $z^* := A^{-1}(y^*)$ and let h_{opt} be the solution of the problem

$$\langle \Lambda h_{opt}, \Lambda v \rangle = \langle z^*, \Lambda v \rangle \qquad (v \in V), \tag{4.2.42}$$

i.e., h_{opt} is the orthogonal projection of z^* on the range of Λ. Then for all $h \in V$

$$z^* - \Lambda h = (z^* - \Lambda h_{opt}) + (\Lambda h_{opt} - \Lambda h),$$

where (4.2.42) for $v := h_{opt} - h$ shows that the terms on the right are orthogonal. Therefore

$$\|z^* - \Lambda h_{opt}\|_Y \leq \|z^* - \Lambda h\|_Y \, .$$

That is, h_{opt} provides the smallest value of $\|z^* - \Lambda h\|_Y$ in (4.2.40).

Remark 4.2.3 (The Lipschitz condition for scalar nonlinearities.) The following class of operators A is an important example of the type discussed above, which occurs in many practical models (see Section 4.3) and has the Lipschitz property from Assumption 4.2.3-(i).

Let $\mathcal{E}$ be a finite dimensional Euclidean space with scalar product $[.,.]$, and let Y be the function space $L^2(\Omega, \mathcal{E})$, i.e.,

$$Y := \{p : \Omega \to \mathcal{E} : \text{ the function } [p,p] \in L^2(\Omega)\}\,.$$

Then Y is a Hilbert space with inner product $\langle p,q\rangle = \int_\Omega [p,q]$, hence Y is a Banach space as well and $Y^* = Y$. Then we define the operator $A : Y \to Y$ as $A(p) := a([p,p])p$, or equivalently (in a test function form)

$$\langle A(p), q\rangle = \int_\Omega \Big(a([p,p])\, [p,q] \Big) \qquad (p,q \in Y), \tag{4.2.43}$$

where $a : \mathbf{R}^+ \to \mathbf{R}^+$ is a scalar C^2 function with the following properties: there exist constants $M \geq m > 0$ such that

$$0 < m \leq a(t) \leq M, \qquad 0 < m \leq \tfrac{d}{dt}\Big(a(t^2)t\Big) \leq M \qquad (t \geq 0), \tag{4.2.44}$$

further, there exists a constant $L_1 > 0$ such that

$$\Big| \tfrac{d^2}{dt^2}\Big(a(t^2)t\Big) \Big| \leq L_1 \qquad (t \geq 0). \tag{4.2.45}$$

Let

$$L := \max\{L_1, 3L_2\}, \quad \text{ where } \quad L_2 := \sup_{t \geq 0} \tfrac{d}{dt}(a(t^2))\,. \tag{4.2.46}$$

Then (4.2.44) implies that A has a bihemicontinuous symmetric Gateaux derivative satisfying

$$m\, \|p\|_Y^2 \leq \langle A'(y)p,p\rangle \leq\ M\, \|p\|_Y^2 \qquad (y,p \in Y) \tag{4.2.47}$$

(similarly to subsection 2.4.2, paragraph (a)), that is, Assumptions 4.2.2 (ii)-(iii) and Assumption 4.2.3 (ii) hold. Further, let

$$W := \{p \in Y : \ [p,p] \in L^\infty(\Omega)\}\,, \qquad \|p\|_W := \|\, |p|_{\mathcal{E}}\, \|_{L^\infty(\Omega)}\,,$$

where $|x|_{\mathcal{E}} := [x,x]^{1/2}$ $(x \in \mathcal{E})$. Then, as proved in [86], A' is Lipschitz continuous as an operator from Y to $B(W, Y^*)$, with Lipschitz constant L from (4.2.46). That is, for all $p,q,s \in Y$, $r \in W$

$$|\langle (A'(p) - A'(q))r, s\rangle| \leq\ L\, \|p-q\|_Y \|r\|_W \|s\|_Y\,, \tag{4.2.48}$$

which is (4.2.19), that is, Assumption 4.2.3-(i) holds as well.

We underline that (4.2.45) is a natural property for functions satisfying (4.2.44) (it almost follows except for some pathological counterexamples.) In particular, if $\frac{d^2}{dt^2}\Big(a(t^2)t\Big)$ is monotone for sufficiently large t, then it is elementary to verify that (4.2.44) implies (4.2.45).

The above results (4.2.47) and (4.2.48) obviously remain valid under natural generalizations of the conditions (4.2.44)–(4.2.45). First, one can allow dependence on x: we let $a : \Omega \times \mathbf{R}^+ \to \mathbf{R}^+$ be a scalar-valued function that is measurable and bounded w.r. to the variable $x \in \Omega$ and C^2 in the variable $t \in \mathbf{R}$, and satisfies

$$0 < m \leq a(x,t) \leq M, \qquad 0 < m \leq \tfrac{\partial}{\partial t}\Big(a(x,t^2)t\Big) \leq M \qquad (x \in \Omega,\ t \geq 0), \tag{4.2.49}$$

$$\Big|\tfrac{\partial^2}{\partial t^2}\Big(a(x,t^2)t\Big)\Big| \leq L \qquad (x \in \Omega,\ t \geq 0). \tag{4.2.50}$$

The operator A, where $a([p,p])$ in (4.2.43) is replaced by $a(x,[p,p])$, then satisfies (4.2.47) and (4.2.48). Further, the sum of such operators also inherits this property. For instance, the results hold for

$$\langle A(p),q\rangle = \int_\Omega \left(a(x,[p,p])\,[p,q] + b(x,\{p,p\})\,\{p,q\}\right) \qquad (p,q \in Y) \tag{4.2.51}$$

where $[.,.]$ and $\{.,.\}$ are two different semi-scalar products on $\mathcal{E}$, such that the sum $[x,y] + \{x,y\}$ for $x,y \in \mathcal{E}$ is already a scalar product on $\mathcal{E}$, further, a and b are functions each satisfying (4.2.49)–(4.2.50). Finally, it is enough to require a to be C^2 except for finitely many points.

4.3 Sharp global error estimates for nonlinear elliptic problems

The previous results can be applied to various concrete types of nonlinear elliptic problems, including second order problems with both Dirichlet and mixed boundary conditions, fourth order problems and second order systems. The restrictions are that they are in divergence form and consist of principal part only: however, as will be pointed out, we thus cover many important real-life models. To avoid extra length, we only detail the exposition for second order Dirichlet problems and sketch the analogous results for the other problems.

4.3.1 Second order Dirichlet problems

We consider the problem

$$\begin{cases} -\mathrm{div}\, f(\nabla u) = g \\ \ \ u_{|\partial\Omega} = 0. \end{cases} \tag{4.3.1}$$

Assumptions 4.3.1.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded domain with piecewise C^2 boundary, locally convex at the corners.

(ii) $f \in C^1(\mathbf{R}^d,\mathbf{R}^d)$, the Jacobians $f'(\eta) := \frac{\partial f(\eta)}{\partial \eta}$ are symmetric and there exist constants $M \geq m > 0$ such that

$$m|\xi|^2 \leq f'(\eta)\ \xi \cdot \xi \leq M|\xi|^2 \qquad (\eta,\xi \in \mathbf{R}^d). \tag{4.3.2}$$

(iii) $f' : \mathbf{R}^d \to \mathbf{R}^{d\times d}$ is Lipschitz continuous with Lipschitz constant L.

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

Let $H_0^1(\Omega)$ denote the usual Sobolev space with inner product

$$\langle u, v\rangle_{H_0^1} := \int_\Omega \nabla u \cdot \nabla v, \tag{4.3.3}$$

further, let

$$H(\mathrm{div}\,) := \{y \in L^2(\Omega)^d : \ \mathrm{div}\, y \in L^2(\Omega)\}.$$

We will also use the space $L^2(\Omega)^d$ with the usual inner product $\langle y, z\rangle_{L^2(\Omega)^d} := \int_\Omega y \cdot z$.

Assumptions (ii) and (iv) imply that problem (4.3.1) has a unique weak solution $u^* \in H_0^1(\Omega)$, i.e., that satisfies

$$\int_\Omega f(\nabla u^*) \cdot \nabla v - \int_\Omega gv = 0 \qquad (v \in H_0^1(\Omega)). \tag{4.3.4}$$

We consider an approximate solution $u \in H_0^1(\Omega)$ and measure the error by the functional

$$E(u) := \int_\Omega (f(\nabla u) - f(\nabla u^*)) \cdot (\nabla u - \nabla u^*) = \int_\Omega f(\nabla u) \cdot (\nabla u - \nabla u^*) - \int_\Omega g(u - u^*). \tag{4.3.5}$$

We note that by (4.1.4),

$$\|u - u^*\|^2_{H_0^1} \le m^{-1}\, E(u).$$

(a) The error estimation

Now we formulate and prove our main result on the error estimation for (4.3.1) for the approximate solution u.

Theorem 4.3.1 *Let $u \in W^{1,\infty}(\Omega)$. Then for arbitrary $y^* \in H(\mathrm{div}\,) \cap L^\infty(\Omega)^d$ and arbitrary $h \in H_0^1(\Omega)$,*

$$E(u) \le \ EST(u; y^*, h) := \big(m^{-1/2}\, C_\Omega\, \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)} \ + \ \tfrac{L}{2}\, m^{-3/2}\, D(u; y^*, h) \tag{4.3.6}$$

$$+ \big(\langle f(\nabla u) - y^*, \ \nabla u - f^{-1}(y^*)\rangle_{L^2(\Omega)^d} \ + \ \tfrac{L}{2m}\, D(u; y^*, h)\, \|\nabla u - f^{-1}(y^*)\|_{L^2(\Omega)^d}\big)^{1/2}\Big)^2 ,$$

where

$$D(u; y^*, h) := \Big(M\, \|f^{-1}(y^*) - \nabla h\|_{L^2(\Omega)^d} + C_\Omega\, \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)}\Big)\, \|\nabla u - f^{-1}(y^*)\|_{L^\infty(\Omega)^d}\, . \tag{4.3.7}$$

Proof. Let $V := H_0^1(\Omega)$ and $Y := L^2(\Omega)^d$. We will use Theorem 4.2.1, to which end we must verify that Assumptions 4.2.2-4.2.3 hold for the corresponding spaces and operators.

First, Assumption 4.2.2 (i) is valid for the operator $\Lambda := \nabla$, since (4.3.3) just yields that (4.2.6) holds. Now let $A : L^2(\Omega)^d \to L^2(\Omega)^d$ be defined by

$$A(y) := f(y) \quad \text{(or, more precisely, } f \circ y\text{)}, \tag{4.3.8}$$

that is, outer composition with f. Such an operator is often called a Nemyczki operator (see, e.g., [158]), and it follows in a standard way [55, 158] from our condition $f \in C^1(\mathbf{R}^d, \mathbf{R}^d)$ and from the assumed symmetry of the Jacobians that A has a bihemicontinuous symmetric Gateaux derivative according to Definition 2.2.1, i.e., Assumption 4.2.2 (ii) holds. The Gateaux derivative of A satisfies

$$\langle A'(y)p, q\rangle_{L^2(\Omega)^d} = \int_\Omega f'(y)\, p \cdot q \qquad (y, p, q \in L^2(\Omega)^d), \tag{4.3.9}$$

hence by (4.3.2) we have

$$m\,\|p\|^2_{L^2(\Omega)^d} \le \langle A'(y)p, p\rangle_{L^2(\Omega)^d} \le\ M\,\|p\|^2_{L^2(\Omega)^d} \qquad (y, p \in L^2(\Omega)^d). \tag{4.3.10}$$

The left-hand side of (4.3.10) coincides with Assumption 4.2.2 (iii). Finally, defining the operator $F : H^1_0(\Omega) \to H^{-1}(\Omega)$ via

$$\langle F(u), v\rangle \equiv \int_\Omega f(\nabla u) \cdot \nabla v \qquad (u, v \in H^1_0(\Omega)), \tag{4.3.11}$$

we obtain the equality (4.2.8), required for Assumption 4.2.2 (iv) to hold.

To verify Assumption 4.2.3 (i), let us define $W := L^\infty(\Omega)^d$ with the standard norm $\|y\|_{L^\infty(\Omega)^d} := \operatorname{ess\,sup}_\Omega |y|$. For the required Lipschitz continuity of A' from $L^2(\Omega)^d$ to $B(L^\infty(\Omega)^d, L^2(\Omega)^d)$, we must prove (4.2.19) for (4.3.8). In fact, we have imposed in Assumption 4.3.1 (iii) the Lipschitz continuity of f' with constant $L > 0$, i.e.,

$$\|f'(\xi) - f'(\eta)\| \le L|\xi - \eta| \qquad (\xi, \eta \in \mathbf{R}^d). \tag{4.3.12}$$

Therefore

$$|\langle (A'(z) - A'(y))\, w, p\rangle| = |\int_\Omega (f'(z) - f'(y))\, w \cdot p|$$

$$\le L \int_\Omega |z-y|\,|w|\,|p| \ \le\ L\,\|z-y\|_{L^2(\Omega)^d}\|w\|_{L^\infty(\Omega)^d}\|p\|_{L^2(\Omega)^d} \qquad (y, z, p \in L^2(\Omega)^d, w \in L^\infty(\Omega)^d), \tag{4.3.13}$$

which is the desired estimate. Assumption 4.2.3 (ii) for (4.3.8) coincides with the right-hand side of (4.3.10).

It is left to check the remaining assumptions of Theorem 4.2.1. Defining the linear functional $l : H^1_0(\Omega) \to \mathbf{R}$ as

$$\langle l, v\rangle \equiv -\int_\Omega gv \qquad (v \in H^1_0(\Omega)) \tag{4.3.14}$$

and using (4.3.11), the weak formulation (4.3.4) of our problem becomes

$$\langle F(u^*), v\rangle + \langle l, v\rangle = 0,$$

i.e. u^* is the solution of (4.1.1) indeed. We have chosen u to satisfy $u \in W^{1,\infty}(\Omega)$, hence $u \in V = H_0^1(\Omega)$ and $\Lambda u = \nabla u \in W = L^\infty(\Omega)^d$. Further, we have assumed $y^* \in W = L^\infty(\Omega)^d$, and the left-hand side of (4.3.2) implies trivially that f^{-1} carries bounded sets into bounded sets (since it grows at most linearly with factor $1/m$), therefore $z^* := A^{-1}(y^*) = f^{-1}(y^*) \in L^\infty(\Omega)^d = W$. Finally, $h \in H_0^1(\Omega) = V$. That is, all the assumptions of Theorem 4.2.1 hold, therefore (4.2.29) is valid for our problem.

It remains to show that the general estimate (4.2.29) for our problem becomes estimate (4.3.6). Here, using $y^* \in H(\mathrm{div}\,)$,

$$\langle \Lambda^* y^*, v\rangle = \langle y^*, \Lambda v\rangle = \int_\Omega y^* \cdot \nabla v = -\int_\Omega (\mathrm{div}\, y^*) v \qquad (v \in H_0^1(\Omega)),$$

hence $\Lambda^* y^* = -\mathrm{div}\, y^*$. Then, by (4.2.13),

$$\begin{aligned}|\Lambda^* y^* + l| &= \sup_{\|v\|_{H_0^1}=1} |\langle \Lambda^* y^* + l, v\rangle| = \sup_{\|v\|_{H_0^1}=1} \Big| -\int_\Omega (\mathrm{div}\, y^* + g) v \Big| \\ &\le \sup_{\|v\|_{H_0^1}=1} \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \le C_\Omega \, \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)},\end{aligned}$$

(where $C_\Omega > 0$ comes from the Friedrichs inequality), see also [122]. Plugging the latter into (4.2.29) and (4.2.30), and replacing V, Y, W, Λ and A by $H_0^1(\Omega)$, $L^2(\Omega)^d$, $L^\infty(\Omega)^d$, ∇ and f, respectively, we obtain (4.3.6). ■

Remark 4.3.1 Following Remark 4.2.1, it is convenient to reformulate Theorem 4.3.1 for $z^* := f^{-1}(y^*)$ in order to avoid the computation of f^{-1}. Then, letting $u \in W^{1,\infty}(\Omega)$ be any approximate solution, for arbitrary $z^* \in L^\infty(\Omega)^d$ such that $f(z^*) \in H(\mathrm{div}\,)$, and for arbitrary $h \in H_0^1(\Omega)$,

$$\begin{aligned}E(u) \le \; \tilde{EST}(u; z^*, h) := \Big(& m^{-1/2}\, C_\Omega \, \|\mathrm{div}\, f(z^*) + g\|_{L^2(\Omega)} \; + \; \tfrac{L}{2}\, m^{-3/2}\, \tilde{D}(u; z^*, h) \\ & + \Big(\langle f(\nabla u) - f(z^*), \; \nabla u - z^* \rangle_{L^2(\Omega)^d} \; + \; \tfrac{L}{2m}\, \tilde{D}(u; z^*, h)\, \|\nabla u - z^*\|_{L^2(\Omega)^d} \Big)^{1/2} \Big)^2 ,\end{aligned} \tag{4.3.15}$$

where

$$\tilde{D}(u; z^*, h) := \Big(M\, \|z^* - \nabla h\|_{L^2(\Omega)^d} + C_\Omega \, \|\mathrm{div}\, f(z^*) + g\|_{L^2(\Omega)} \Big) \, \|\nabla u - z^*\|_{L^\infty(\Omega)^d} \, . \tag{4.3.16}$$

Now we state the sharpness of the estimate:

Proposition 4.3.1 *Estimate* (4.3.6) *is sharp, that is,*

$$\min_{\substack{y^* \in H(\mathrm{div}\,)\cap L^\infty(\Omega)^d, \\ h \in H_0^1(\Omega)}} EST(u; y^*, h) = E(u).$$

Proof. By [117], the weak solution of (4.3.1) satisfies $u^* \in C^{1,\alpha}(\overline{\Omega})$ with some $0 < \alpha < 1$, hence $\nabla u^* \in L^\infty(\Omega)^d = W$. Therefore we can apply Proposition 4.2.4 to obtain the desired statement. ■

We note that by (4.2.41), the optimal values for "free" parameters in the estimate are

$$y^* := f(\nabla u^*) \quad \text{and} \quad h := u^*. \tag{4.3.17}$$

The practical approximations of these will be discussed below.

(b) Practical considerations

Finite element solution. A most important practical case is when finite element approximation is used. In general, let V_h be a given FEM subspace and $u_h \in V_h$ be the corresponding FEM approximation of the exact solution u^*. Then our error measure is

$$E(u_h) = \langle F(u_h) - F(u^*),\, u_h - u^* \rangle. \tag{4.3.18}$$

Here u_h is a continuous piecewise polynomial, hence condition $u_h \in W^{1,\infty}(\Omega)$ in Theorem 4.3.1 is satisfied. If we choose y^* to be any continuous piecewise polynomial function, e.g. a function from another FEM subspace, and arbitrary $w \in H_0^1(\Omega)$, then $y^* \in H(\mathrm{div}\,) \cap L^\infty(\Omega)^d$, hence Theorem 4.3.1 can be applied and yields

$$E(u_h) \leq \ EST(u_h; y^*, w). \tag{4.3.19}$$

We note that it is useful to replace $f^{-1}(y^*)$ by z^* as in (4.3.15)–(4.3.16) to avoid the computation of f^{-1}. The obtained expressions are directly computable integrals.

Determining the optimal y^* and w in $EST(u_h; y^*, w)$. Following (4.3.17), the optimal value of the parameter y^* should be a sufficiently accurate approximation of $f(\nabla u^*)$. For finite element solutions, a common and "computationally cheap" way to achieve this goal is to use an averaging procedure, i.e., to replace the unknown function ∇u^* (the gradient of the exact solution) by $G_h(\nabla u_h)$, where G_h is some averaging operator. For the case of linear finite elements, $G_h(\nabla u_h)$ is closer to ∇u^* than is ∇u_h by an order of magnitude, namely, the original approximation order $\|\nabla u^* - \nabla u_h\|_{L^2} = O(h)$ can be thus improved to $\|\nabla u^* - G_h(\nabla u_h)\|_{L^2} = O(h^2)$ if u^* is sufficiently smooth, see [73, Part I] for details. Accordingly, we can define

$$y^* := f\Big(G_h(\nabla u_h)\Big), \quad z^* = f^{-1}(y^*) = G_h(\nabla u_h) \tag{4.3.20}$$

as a first candidate for the parameter y^* (or z^*). (If this still gives a too rough bound, then one executes a minimization process for y^*, see [109] for more details.)

Next, using Remark 4.2.2, the optimal w for this z^* is given as the solution of the following linear auxiliary problem: find $w_{opt} \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla w_{opt} \cdot \nabla v = \int_\Omega z^* \cdot \nabla v \qquad (v \in H_0^1(\Omega)), \tag{4.3.21}$$

that is, the weak solution of the Poisson problem

$$\left\{ \begin{array}{l} -\Delta w_{opt} = -\mathrm{div}\, z^* \\ \ \ w_{opt\,|\partial\Omega} = 0. \end{array} \right. \tag{4.3.22}$$

This means that for given y^*, the optimal estimate for the second parameter w is found by solving a kind of adjoint or auxiliary equation; however, the latter is linear, hence its numerical solution costs much less than for the original one. For piecewise linear FEM, if (4.3.22) is solved numerically on the same mesh as used for u_h, then its right-hand side $-\mathrm{div}\, z^* = -\mathrm{div}\, G_h(\nabla u_h)$ is constant on each element, hence it requires minimal numerical integration and is therefore a cheap auxiliary problem. On the other hand, using a finer (or just different) mesh for (4.3.22) than the one used for u_h may considerably increase the accuracy of the estimate, similarly as for adjoint problems for linear equations [100], with low extra cost due to the linearity of (4.3.22).

Calculating the required constants. The constants used in estimate (4.3.6) are C_Ω, m, M and L. The only one depending on the domain is C_Ω, which can be easily estimated from above, as mentioned in Remark 1.2.2. Further, the three remaining constants m, M and L come from the given nonlinearity, see Assumptions 4.3.1 (ii)-(iii), where we note that a crucial point in our sharp estimates is the existence of L, i.e., the condition of Lipschitz continuity of the derivative of f. Based on Remark 4.2.3, one can see that this Lipschitz condition usually means no restriction in practice, since it is satisfied for most real problems. Namely, problems of the type (4.3.1) in real models are generally of the following special form, involving a scalar nonlinearity:

$$\begin{cases} -\mathrm{div}\,\Big(a(|\nabla u|^2)\,\nabla u\Big) = g \\ u_{|\partial\Omega} = 0 \end{cases} \tag{4.3.23}$$

(which corresponds to $f(\eta) = a(|\eta|^2)\,\eta$ in (4.3.1)), where $a : \mathbf{R}^+ \to \mathbf{R}^+$ is a scalar C^2 function with properties (4.2.44)–(4.2.45). Such nonlinearities form the main examples for (4.3.1), arising, e.g., in elasto-plastic torsion [77], or in electromagneticity, see the presentation from nonlinear Maxwell equations in [106] and for nonlinear magnetostatic field in [35]. One may even have explicit formulae for the function a, such as (2.6.3) which characterizes the reluctance of stator sheets, or a similar formula which describes magnetostatic field; the constants in these formulas are given positive characteristic physical values. Using Remark 4.2.3, condition (4.2.45) implies the Lipschitz continuity for f. It has also been pointed out in Remark 4.2.3 that condition (4.2.45) follows from the standard ellipticity property (4.2.44) except for some unrealistic special cases.

Summing up, it follows that the bounds m and M and the Lipschitz constant L, needed to calculate $E(u)$, can be determined from lower or upper bounds, respectively, for the scalar functions in (4.2.44)–(4.2.45). These only require an elementary numerical calculation. Moreover, if the parameters y^* and w are close to the optimal choice, then (using Proposition 4.2.4) all terms containing these constants (as well as C_Ω) in $EST(u_h; y^*, w)$ are close to zero, hence the global constants need not be estimated from above much accurately.

4.3.2 Other elliptic problems

We sketch the results for some other nonlinear elliptic problems as analogues to the above.

(a) Second order mixed problems

Let us first consider second order problems with mixed boundary conditions. Here we also allow dependence of the nonlinearity f on x, which was not included in (4.3.1) for simplicity. That is,

$$\begin{cases} -\mathrm{div}\, f(x,\nabla u) = g \\ u_{|\Gamma_D} = 0 \\ f(x,\nabla u)\cdot \nu_{\ |\Gamma_N} = \gamma \end{cases} \tag{4.3.24}$$

(where ν denotes the outer normal unit vector). Here Assumptions 4.3.1 are completed with the following conditions: Γ_D, Γ_N are disjoint open subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \neq \emptyset$, further, $\gamma \in L^2(\Gamma_N)$; finally, in assumption (ii), the conditions on $f'(\eta)$ are replaced in an obvious way with that for $f'(x,\eta) := \frac{\partial f(x,\eta)}{\partial \eta}$.

The treatment of this problem uses the Sobolev space $H^1_D(\Omega)$ with inner product $\langle u,v\rangle_{H^1_D} := \int_\Omega \nabla u \cdot \nabla v$, further, let

$$H(\mathrm{div}\,, \Gamma_N) := \{y \in L^2(\Omega)^d : \ \mathrm{div}\, y \in L^2(\Omega), \ \ y\cdot\nu \in L^2(\Gamma_N)\}.$$

We now use the estimates

$$\|v\|_{L^2(\Omega)} \le C'_\Omega \, \|\nabla v\|_{L^2(\Omega)^d}, \qquad \|v\|_{L^2(\Gamma_N)} \le C_{\Gamma_N} \, \|\nabla v\|_{L^2(\Omega)^d} \qquad (v \in H^1_D(\Omega)) \tag{4.3.25}$$

for some suitable constants $C'_\Omega, C_{\Gamma_N} > 0$.

To formulate the main result, we note that by Assumption 4.3.2 (ii), for all fixed $x \in \Omega$, the function $f(x,.)$ is invertible on $\mathbf{R}^d$ w.r.t. η. We will denote by f^{-1} the inverse w.r.t. η, i.e.

$$f(x,\eta) = \xi \quad \Rightarrow \quad f^{-1}(x,\xi) := \eta. \tag{4.3.26}$$

Then one can prove the main results similarly as before:

Theorem 4.3.2 *Let $u \in W^{1,\infty}(\Omega)$. Then for arbitrary $y^* \in H(\mathrm{div}\,, \Gamma_N) \cap L^\infty(\Omega)^d$ and arbitrary $h \in H^1_D(\Omega)$,*

$$E(u) \le \ EST(u; y^*, h) := \tag{4.3.27}$$

$$\big(m^{-1/2}\, C'_\Omega \, \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)} \ + \ m^{-1/2}\, C_{\Gamma_N} \|y^*\cdot\nu - \gamma\|_{L^2(\Gamma_N)} + \ \tfrac{L}{2}\, m^{-3/2}\, D(u; y^*, h)$$

$$+ \big(\langle f(x,\nabla u) - y^*, \ \nabla u - f^{-1}(x,y^*)\rangle_{L^2(\Omega)^d} \ + \ \tfrac{L}{2m}\, D(u; y^*, h)\, \|\nabla u - f^{-1}(x,y^*)\|_{L^2(\Omega)^d}\big)^{1/2}\Big)^2$$

where

$$D(u; y^*, h) := \Big(M\, \|f^{-1}(x,y^*) - \nabla h\|_{L^2(\Omega)^d} + \ C'_\Omega \, \|\mathrm{div}\, y^* + g\|_{L^2(\Omega)} \tag{4.3.28}$$

$$+ C_{\Gamma_N} \|y^*\cdot\nu - \gamma\|_{L^2(\Gamma_N)}\Big) \, \|\nabla u - f^{-1}(x,y^*)\|_{L^\infty(\Omega)^d} \, .$$

Proof. It follows from Theorem 4.2.1 in the given spaces. ■

Turning to the sharpness problem, Proposition 4.2.4 yields

Proposition 4.3.2 *Estimate* (4.3.27) *is sharp, that is,*

$$\min_{\substack{y^* \in H(\mathrm{div}\,, \Gamma_N) \cap L^\infty(\Omega)^d, \\ h \in H^1_0(\Omega)}} EST(u; y^*, h) = E(u), \tag{4.3.29}$$

provided the exact solution satisfies $u^* \in W^{1,\infty}(\Omega)$.

Remark 4.3.2 The analogues of Theorem 4.3.2 can be proved similarly if (4.3.24) is replaced by one of the following problems:

(a) Neumann problem. Allowing $\Gamma_D = \emptyset$ in Assumption 4.3.2. (i), we have

$$\begin{cases} -\mathrm{div}\, f(x, \nabla u) = g \\ f(x, \nabla u) \cdot \nu_{\,|\partial\Omega} = \gamma. \end{cases} \tag{4.3.30}$$

Then Theorem 4.3.2 remains true if we substitute the factorized space $V := \dot{H}^1(\Omega) := \{u \in H^1(\Omega) : \int_\Omega u = 0\}$ instead of $H^1_D(\Omega)$ and replace Γ_N by $\partial\Omega$ in the formulas. In particular, the resulting constant $C_{\partial\Omega}$ to satisfy the second inequality in (4.3.25) for all $v \in \dot{H}^1(\Omega)$ is the smallest positive eigenvalue of $-\Delta$ with Neumann boundary conditions.

(b) Interface problems. Let Γ_{int} be a piecewise smooth surface lying in the interior of Ω, and let us consider the problem

$$\begin{cases} -\mathrm{div}\, f(x, \nabla u) = g \\ u_{|\Gamma_D} = 0, \quad f(x, \nabla u) \cdot \nu_{\,|\Gamma_N} = \gamma_N\,, \quad f(x, \nabla u) \cdot \nu_{\,|\Gamma_{int}} = \gamma_{int}, \end{cases} \tag{4.3.31}$$

where the assumptions for the mixed problem are modified such that $\gamma_N \in L^2(\Gamma_N)$ and $\gamma_{int} \in L^2(\Gamma_{int})$. The weak form of this problem is the same as for the mixed problem if Γ_N is replaced by $\Gamma := \Gamma_N \cup \Gamma_{int}$, see [93] for a related setting. Defining $\gamma \in L^2(\Gamma)$ such that its restrictions to Γ_N and Γ_{int} are γ_N and γ_{int}, respectively, Theorem 4.3.2 remains true if we replace Γ_N by Γ in the formulas.

In practice, to determine suitable y^* and w in $EST(u_h; y^*, w)$, first y^* should be some approximation of $f(x, \nabla u^*)$. For finite element solutions, using averaging as in (4.3.20), we can first let

$$y^* := f\Big(x, G_h(\nabla u_h)\Big), \quad z^* = f^{-1}(x, y^*) = G_h(\nabla u_h), \tag{4.3.32}$$

where G_h is some averaging operator and f^{-1} is understood w.r.t. η as in (4.3.26). Averaging for mixed boundary conditions is discussed, e.g., in [73, Part II]. More accurate error bounds can be obtained by suitable minimization as mentioned before.

Then by Remark 4.2.2, the optimal w for this z^* to set in $EST(u_h; y^*, w)$ is given as the solution of a linear auxiliary problem, which is the modification of (4.3.21) for mixed boundary conditions. This can be solved on a suitably chosen mesh, either the same as used for u_h or a finer/different mesh, as discussed in Section 4.3.1.

The constants used can be obtained easily for most of the practical cases, using a scalar form of the nonlinearity as in (4.3.23). Some examples are the x-dependent nonlinearity (2.6.2) in magnetic potential [63, 106], or that describing air density in a subsonic potential flow, see, e.g., [21], which we have already described after (3.5.4). In the corresponding mixed problem, Γ_D is the wind inblow part and Γ_N consists of the other sides of the wind tunnel section. Altogether, the constants can be therefore determined by elementary numerical calculation.

(b) Fourth order problems

In this subsection we study 4th order Dirichlet problems. The concise presentation requires some basic notations: let D^2u denote the Hessian of a function $u : \Omega \to \mathbf{R}$ if $u \in H^2(\Omega)$, we define the elementwise matrix product and the corresponding Frobenius norm in the standard way

$$P : Q := \sum_{i,k=1}^{d} P_{ik}\, Q_{ik}, \qquad |P|_F := (P : P)^{1/2} \qquad (P, Q \in \mathbf{R}^{d\times d}), \tag{4.3.33}$$

further, for a matrix-valued function $P : \Omega \to \mathbf{R}^{d\times d}$ we let $\operatorname{div}^2 P := \sum\limits_{i,k=1}^{d} \frac{\partial^2 P_{ik}}{\partial x_i \partial x_k}$, provided that these derivatives exist.

Now we can formulate the problems considered, defined via a matrix-valued nonlinearity B, in the form

$$\begin{cases} \operatorname{div}^2 B(x, D^2u) \ = \ g \\ u_{|\partial\Omega} = \frac{\partial u}{\partial \nu}\big|_{\partial\Omega} = 0\,, \end{cases} \tag{4.3.34}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$ with a piecewise C^1 boundary, with $g \in L^2(\Omega)$ as before, under the following assumptions on the nonlinearity B:

(i) The matrix-valued function $B : \Omega \times \mathbf{R}^{d\times d} \to \mathbf{R}^{d\times d}$ is measurable and bounded w.r. to the variable $x \in \Omega$ and C^2 in the matrix variable $\Theta \in \mathbf{R}^{d\times d}$. The Jacobian arrays

$$B'(x, \Theta) := \frac{\partial B(x, \Theta)}{\partial \Theta} \ = \ \left\{ \frac{\partial B_{rs}(x, \Theta)}{\partial \Theta_{ik}} \right\}_{i,k,r,s=1}^{d} \in \mathbf{R}^{(d\times d)^2}$$

are symmetric, i.e. $\partial B_{rs}/\partial\Theta_{ik} = \partial B_{ik}/\partial\Theta_{rs}$ for all i, k, r, s, and there exist constants $M \geq m > 0$ such that

$$m|\Phi|_F^2 \leq B'(x, \Theta)\Phi : \Phi \leq M|\Phi|_F^2 \qquad (x \in \Omega; \ \ \Theta, \Phi \in \mathbf{R}^{d\times d}). \tag{4.3.35}$$

(ii) $B' : \Omega \times \mathbf{R}^{d\times d} \to \mathbf{R}^{(d\times d)^2}$ is Lipschitz continuous in the matrix variable $\Theta \in \mathbf{R}^{d\times d}$, with Lipschitz constant L.

In the treatment of this problem we follow the previous sections. Now we use the Lebesgue space

$$L^2(\Omega)^{d\times d} := \{P : \Omega \to \mathbf{R}^{d\times d} : \ \ P_{ik} \in L^2(\Omega) \ \text{ for all } i, k = 1, \ldots, d\} \tag{4.3.36}$$

with inner product $\langle P, Q\rangle_{L^2(\Omega)^{d\times d}} := \int_\Omega P : Q$, and the Sobolev space

$$H_0^2(\Omega) := \{u \in H^2(\Omega) : \ u_{|\partial\Omega} = \tfrac{\partial u}{\partial \nu}\big|_{\partial\Omega} = 0 \text{ in trace sense}\} \tag{4.3.37}$$

with inner product $\langle u, v\rangle_{H_0^2} := \langle D^2u, D^2v\rangle_{L^2(\Omega)^{d\times d}} = \int_\Omega D^2u : D^2v$. Further, let

$$H(\operatorname{div}^2) := \{P \in L^2(\Omega)^{d\times d} : \ \operatorname{div}^2 P \in L^2(\Omega)\}.$$

The actual counterpart of the Friedrichs inequality is as follows:

$$\|v\|_{L^2(\Omega)} \leq \tilde{C}_\Omega \, \|D^2 v\|_{L^2(\Omega)^{d\times d}} \qquad (v \in H^2_0(\Omega)) \tag{4.3.38}$$

for some suitable constant $\tilde{C}_\Omega > 0$. Analogously to (4.3.26), we will denote by B^{-1} the inverse w.r.t. Θ, i.e.

$$B(x,\Theta) = \Phi \quad \Rightarrow \quad B^{-1}(x,\Phi) := \Theta, \tag{4.3.39}$$

where B^{-1} exists by the assumptions on B. Then one can prove the main results similarly as before:

Theorem 4.3.3 *Let $u \in W^{2,\infty}(\Omega)$. Then for arbitrary $Y^* \in H(\operatorname{div}^2) \cap L^\infty(\Omega)^{d\times d}$ and arbitrary $h \in H^2_0(\Omega)$,*

$$\begin{aligned} E(u) \leq \; EST(u;Y^*,h) := \Big(& m^{-1/2}\, \tilde{C}_\Omega \, \|\operatorname{div}^2 Y^* - g\|_{L^2(\Omega)} \; + \; \tfrac{L}{2}\, m^{-3/2}\, D(u;Y^*,h) \\ & + \big(\langle B(x,D^2u) - Y^*, \; D^2u - B^{-1}(x,Y^*) \rangle_{L^2(\Omega)^{d\times d}} \\ & + \tfrac{L}{2m}\, D(u;Y^*,h)\, \|D^2u - B^{-1}(x,Y^*)\|_{L^2(\Omega)^{d\times d}} \big)^{1/2} \Big)^2 \end{aligned} \tag{4.3.40}$$

where

$$\begin{aligned} D(u;Y^*,h) := \Big(& M \, \|B^{-1}(x,Y^*) - D^2h\|_{L^2(\Omega)^{d\times d}} + \tilde{C}_\Omega \, \|\operatorname{div}^2 Y^* - g\|_{L^2(\Omega)} \Big) \times \\ & \times \|D^2u - B^{-1}(x,Y^*)\|_{L^\infty(\Omega)^{d\times d}} \,. \end{aligned} \tag{4.3.41}$$

Proof. It follows from Theorem 4.2.1 in the given spaces. ■

Remark 4.3.3 Following [122, Chap. 6.6], the term $\tilde{C}_\Omega \, \|\operatorname{div}^2 Y^* - g\|_{L^2(\Omega)}$ in (4.3.40) can be replaced by

$$\hat{C}_\Omega \, \|\operatorname{div} Y^* - \eta^*\|_{L^2(\Omega)^{d\times d}} + \tilde{C}_\Omega \, \|\operatorname{div} \eta^* - g\|_{L^2(\Omega)}$$

for some new parameter function $\eta^* \in H(\operatorname{div})$. In this case the requirement $Y^* \in H(\operatorname{div}^2)$ can be weakened to $Y^* \in H(\operatorname{div})$ (understood row-wise).

Note that our result is a direct extension of earlier sharp error estimates obtained for linear fourth order problems [121]. In our case, Proposition 4.2.4 yields

Proposition 4.3.3 *Estimate* (4.3.40) *is sharp, that is,*

$$\min_{\substack{Y^* \in H(\operatorname{div}^2) \cap L^\infty(\Omega)^{d\times d}, \\ h \in H^2_0(\Omega)}} EST(u;Y^*,h) = E(u), \tag{4.3.42}$$

provided that the exact solution satisfies $u^* \in W^{2,\infty}(\Omega)$.

In practice for FEM, in order to have an approximate solution $u_h \in H_0^2(\Omega)$, one uses C^1-elements (i.e. $u_h \in C^1$ and u_h is piecewise polynomial), see, e.g., [34]. In this case we automatically have $u \in W^{2,\infty}(\Omega)$, which was required for Theorem 4.3.3 to hold. (Another common FEM approach is to use mixed variables to have less smoothness for u_h. In this case one may expect to reformulate the terms containing D^2u in (4.3.40) via the mixed variables in a similar vein as in Remark 4.3.3, which we do not consider here.) Next, following (4.2.41), Y^* should be some approximation of $B(x, D^2u^*)$. For finite element solutions, using averaging as before, we can first let

$$Y^* := B\Big(x, G_h(D^2u_h)\Big), \quad Z^* = B^{-1}(x, Y^*) = G_h(D^2u_h), \tag{4.3.43}$$

where G_h is some averaging operator that defines a C^1-approximation of D^2u_h, and B^{-1} is understood w.r.t. Θ as in (4.3.39). Then by Remark 4.2.2, the optimal w for this Z^* to set in $EST(u_h; Y^*, w)$ is the solution of a corresponding linear biharmonic auxiliary problem with r.h.s. $\text{div}^2 Z^*$. Note that Z^* need not be in $H(\text{div}^2)$ to pose the latter: in general $\text{div}^2 Z^*$ can be understood in a distributional sense, which exactly means that we need to use the weak form, and thus the weaker condition $Y^* \in H(\text{div})$ (or equivalently $Z^* \in H(\text{div}))$ can be used. Altogether, one can define w as the numerical solution of the biharmonic auxiliary problem on a suitably chosen mesh, either the same as used for u_h or a finer mesh, as discussed in Section 4.3.1.

The most important real-life model that uses fourth order equations like (4.3.34) describes the elasto-plastic bending of a clamped thin plane plate $\Omega \subset \mathbf{R}^2$, see, e.g., [55] and subsection 2.4.2. This problem is as follows:

$$\begin{cases} \text{div}^2 \Big(\overline{g}(E(D^2u))\ \tilde{D}^2u\Big) \ = \ \alpha \\ u_{|\partial\Omega} = \frac{\partial u}{\partial \nu}\big|_{\partial\Omega} = 0 \end{cases} \tag{4.3.44}$$

where

$$\tilde{D}^2u := \tfrac{1}{2}\left(D^2u + \Delta u \cdot I\right), \qquad E(D^2u) := \tfrac{1}{2}\left(|D^2u|_F^2 + (\Delta u)^2\right)$$

and $\overline{g}$ is a scalar material function satisfying (4.2.44)-(4.2.45) (with $\overline{g}$ substituted for a). This problem leads to an operator like (4.2.43), see more details in [55].

(c) Second order elasticity systems

Symmetric second order systems arise in the description of the elastic behaviour of a body. We follow the description in subsection 2.6.5, based on [120]. We impose as an additional condition that k and μ are also piecewise C^2 (i.e. C^2 except for finitely many isolated points, which in practice typically separate the domain of linear and nonlinear behaviour), further, that there exists a constant $L > 0$ such that

$$\left|\tfrac{\partial^2}{\partial t^2}\Big(k(x,t^2)t\Big)\right| \le L, \qquad \left|\tfrac{\partial^2}{\partial t^2}\Big(\mu(x,t^2)t\Big)\right| \le L \qquad (x \in \Omega,\ t \ge 0). \tag{4.3.45}$$

We note that some concrete measurements or explicit expressions on k and μ are given, e.g., in [120, 122], and k is often considered as constant. With the notations of (4.3.33) and (4.3.35), we obtain the analogue of (4.3.35):

$$m|\Phi|_F^2 \le T'(x,\Theta)\Phi : \Phi \le M|\Phi|_F^2 \qquad (x \in \Omega; \quad \Theta, \Phi \in \mathbf{R}^{3\times 3}). \tag{4.3.46}$$

This property implies well-posedness in $H^1_D(\Omega)^3$ in view the famous Korn's inequality

$$\kappa \int_\Omega |\nabla u|^2 \leq \int_\Omega |\varepsilon(u)|^2 \leq \int_\Omega |\nabla u|^2 \qquad (u \in H^1_D(\Omega)^3) \tag{4.3.47}$$

(where $\kappa > 0$), see more details, e.g., in [120].

In the treatment of error estimation for the elasticity problem, we follow the previous sections. Now we use the Lebesgue space

$$L^2(\Omega)^{3\times 3}_{symm} := \{P : \Omega \to \mathbf{R}^{3\times 3} : \ P_{ik} = P_{ki} \in L^2(\Omega) \text{ for all } i,k = 1,\dots,3\} \tag{4.3.48}$$

with inner product $\langle P, Q\rangle_{L^2(\Omega)^{3\times 3}} := \int_\Omega P : Q$, using notation (4.3.33), Further, we endow the space $H^1_D(\Omega)^3$ with inner product

$$\langle u, v\rangle_\varepsilon := \langle \varepsilon(u), \varepsilon(v)\rangle_{L^2(\Omega)^{3\times 3}} = \int_\Omega \varepsilon(u) : \varepsilon(v), \tag{4.3.49}$$

which is equivalent to the standard inner product owing to (4.3.47). Inequalities (4.3.25) and (4.3.47) then imply

$$\|v\|_{L^2(\Omega)^3} \leq \kappa^{-1/2}\, C'_\Omega \,\|v\|_\varepsilon, \qquad \|v\|_{L^2(\Gamma_N)^3} \leq \kappa^{-1/2}\, C_{\Gamma_N}\, \|v\|_\varepsilon \qquad (v \in H^1_D(\Omega)^3). \tag{4.3.50}$$

We define $L^\infty(\Omega)^{3\times 3}_{symm}$ analogously to (4.3.48), and finally let

$$H(\mathrm{div}\,, \mathbf{R}^3;\, \Gamma_N) := \{P \in L^2(\Omega)^{3\times 3}_{symm} : \ \mathrm{div}\, P \in L^2(\Omega)^3, \ P\cdot\nu \in L^2(\Gamma_N)^3\}.$$

We will use notation T^{-1} in the sense of (4.3.39).

Theorem 4.3.4 *Let $u \in W^{1,\infty}(\Omega)^3$. Then for arbitrary $Y^* \in H(\mathrm{div}\,, \mathbf{R}^3;\, \Gamma_N) \cap L^\infty(\Omega)^{3\times 3}_{symm}$ and arbitrary $h \in H^1_D(\Omega)^3$,*

$$\begin{aligned} E(u) \leq \ & EST(u; Y^*, h) := \Big((\kappa m)^{-1/2}\, C'_\Omega \,\|\mathrm{div}\, Y^* + \varphi\|_{L^2(\Omega)^3} + (\kappa m)^{-1/2}\, C_{\Gamma_N} \|Y^*\cdot\nu - \tau\|_{L^2(\Gamma_N)^3} \\ & + \tfrac{L}{2}\, m^{-3/2}\, D(u; Y^*, h) + \big(\langle T(x, \varepsilon(u)) - Y^*, \ \varepsilon(u) - T^{-1}(x, Y^*)\rangle_{L^2(\Omega)^{3\times 3}} \\ & \qquad + \tfrac{L}{2m}\, D(u; Y^*, h)\, \|\varepsilon(u) - T^{-1}(x, Y^*)\|_{L^2(\Omega)^{3\times 3}}\big)^{1/2}\Big)^2 , \end{aligned} \tag{4.3.51}$$

where

$$\begin{aligned} D(u; Y^*, h) := & \Big(M\, \|T^{-1}(x, Y^*) - \varepsilon(h)\|_{L^2(\Omega)^{3\times 3}} + \kappa^{-1/2}\, C'_\Omega\, \|\mathrm{div}\, Y^* + \varphi\|_{L^2(\Omega)^3} \\ & + \kappa^{-1/2}\, C_{\Gamma_N} \|Y^*\cdot\nu - \tau\|_{L^2(\Gamma_N)^3}\Big) \,\|\varepsilon(u) - T^{-1}(x, Y^*)\|_{L^\infty(\Omega)^{3\times 3}} . \end{aligned} \tag{4.3.52}$$

PROOF. It follows from Theorem 4.2.1 in the given spaces. ■

Our result is a direct extension of earlier sharp error estimates obtained for linear elasticity problems [122]. Now Proposition 4.2.4 yields

Proposition 4.3.4 *Estimate* (4.3.51) *is sharp, that is,*

$$\min_{\substack{Y^* \in H(\mathrm{div}\,,\mathbf{R}^3;\Gamma_N)\cap L^\infty(\Omega)^{3\times 3}_{symm},\\ h\in H^1_D(\Omega)^3}} EST(u;Y^*,h) = E(u), \qquad (4.3.53)$$

provided that the exact solution satisfies $u^* \in W^{1,\infty}(\Omega)^3$.

In practice, for finite element solutions, all three coordinate functions of the FEM approximation $u_h \in V_h \subset H^1_D(\Omega)^3$ are continuous piecewise polynomials, hence condition $u_h \in W^{1,\infty}(\Omega)^3$ in Theorem 4.3.4 is satisfied. If we choose Y^* to be a symmetric matrix function whose entries are also continuous piecewise polynomial functions, e.g. , functions from another FEM subspace, and arbitrary $w \in H^1_D(\Omega)^3$, then $Y^* \in H(\mathrm{div}\,,\mathbf{R}^3;\Gamma_N) \cap L^\infty(\Omega)^{3\times 3}_{symm}$, hence Theorem 4.3.4 can be applied. Next, following (4.2.41), Y^* should be some approximation of $T(x,\varepsilon(u^*))$. For finite element solutions, using averaging as before yields

$$Y^* := T\Big(x, G_h(\varepsilon(u_h))\Big), \quad Z^* = T^{-1}(x,Y^*) = G_h(\varepsilon(u_h)), \qquad (4.3.54)$$

where G_h is some averaging operator, based on [73] where averaging is discussed in the context of elasticity problems, further, T^{-1} is understood w.r.t. Θ. Then by Remark 4.2.2, the optimal w for this Z^* to set in $EST(u_h;Y^*,w)$ is the solution of the following linear auxiliary problem: find $w_{opt} \in H^1_D(\Omega)^3$ such that

$$\int_\Omega \varepsilon(w_{opt}) : \varepsilon(v) = \int_\Omega Z^* : \varepsilon(v) \qquad (v \in H^1_D(\Omega)^3). \qquad (4.3.55)$$

Hence one can define w as the numerical solution of (4.3.55) on a suitable mesh (either the same as used for u_h or a finer mesh, as discussed in Section 4.3.1). Regarding the required constants, estimates for C'_Ω and C_{Γ_N} can be done similarly to [134], see also Remark 1.2.2. Several explicit values and estimates for Korn's constant κ are given in [75], finally, as pointed out at the end of Remark 4.2.3, the bounds m and M and the Lipschitz constant L can be calculated numerically from (2.6.25) and (4.3.45).

Bibliography

[1] Adams, R.A., *Sobolev Spaces*, Academic Press, 1975.

[2] Ainsworth, M., Oden, J. T., *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, Inc., 2000.

[3] 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.

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

[5] 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.

[6] Ashby, S. F., Manteuffel, T. A., Saylor, P. E., A taxonomy for conjugate gradient methods, *SIAM J. Numer. Anal.* 27 (1990), no. 6, 1542–1568.

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

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

[9] 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.

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

[11] Axelsson, O., Faragó I., Karátson J., On the application of preconditioning operators for nonlinear elliptic problems, in: *Conjugate Gradient Algorithms and Finite Element Methods*, pp. 247-261, Springer, 2004.

[12] 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.

[13] Axelsson, O., Kaporin, I., On the sublinear and superlinear rate of convergence of conjugate gradient methods. Mathematical journey through analysis, matrix theory and scientific computation (Kent, OH, 1999), *Numer. Algorithms* 25 (2000), no. 1-4, 1–22.

[14] Axelsson, O., Karátson J., Symmetric part preconditioning for the conjugate gradient method in Hilbert space, *Numer. Funct. Anal.* 24 (2003), No. 5-6, 455-474.

[15] Axelsson, O., Karátson J., Conditioning analysis of separate displacement preconditioners for some nonlinear elasticity systems, *Math. Comput. Simul.* 64 (2004), No.6, pp. 649-668.

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

[17] Axelsson, O., Karátson J., Symmetric part preconditioning of the CGM for Stokes type saddle-point systems, *Numer. Funct. Anal.* 28 (2007), 9-10, pp. 1027-1049

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

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

[20] Axelsson, O., Layton, W., A two-level discretization of nonlinear boundary value problems, *SIAM J. Numer. Anal.* 33 (1996), no. 6, 2359–2374.

[21] Axelsson, O., Maubach, J., On the updating and assembly of the Hessian matrix in finite element methods, *Comp. Meth. Appl. Mech. Engrg.*, 71 (1988), pp. 41-67.

[22] Axelsson, O., Neytcheva, M., Scalable algorithms for the solution of Navier's equations of elasticity, *J. Comput. Appl. Math.* 63 (1995), no. 1-3, 149–178.

[23] Bjorstad, P. E., Tjostheim, B. P., Efficient algorithms for solving a fourth-order equation with the spectral-Galerkin method, *SIAM J. Sci. Comput.* 18 (1997), no. 2, 621–632.

[24] Benzi, M., Golub, G. H., Liesen, J., Numerical solution of saddle point problems, *Acta Numer.* 14 (2005), 1–137.

[25] Bertaccini D., Golub G. H., Serra S., Spectral analysis of a preconditioned iterative method for the convection-diffusion equation, *SIAM J. Matr. Anal. Appl.* 29-1, 2007, pp. 260–278.

[26] Blaheta, R., Multilevel Newton methods for nonlinear problems with applications to elasticity, Copernicus 940820, Technical report.

[27] Brandts, J., Korotov, S., Křížek, M., Dissection of the path-simplex in $\mathbf{R^n}$ into n path-subsimplices, *Linear Algebra Appl.* 421 (2007), no. 2-3, 382–393.

[28] Brezzi, F., Fortin, M., *Mixed and hybrid finite element methods,* Springer Series in Computational Mathematics, 15, Springer-Verlag, New York, 1991.

[29] Brown, P.N., Vassilevski, P.S., Woodward, C.S., On mesh-independent convergence of an inexact Newton-multigrid algorithm, *SIAM J. Sci. Comput.* 25 (2003), no. 2, 570–590.

[30] Carey, G.F., Jiang, B.-N., Nonlinear preconditioned conjugate gradient and least-squares finite elements, *Comp. Meth. Appl. Mech. Engrg.*, 62 (1987), pp. 145-154.

[31] Caristi, G., Mitidieri, E., Further results on maximum principles for noncooperative elliptic systems, *Nonlinear Anal.* 17 (1991), no. 6, 547–558.

[32] Chicco, M., A maximum principle for mixed boundary value problems for elliptic equations in non-divergence form, *Boll. Unione Mat. Ital.*, VII. Ser., B 11, No.3, 531-538 (1997).

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

[34] Ciarlet, P. G., *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978.

[35] Concus, P., Numerical solution of the nonlinear magnetostatic field equation in two dimensions, *J. Comput. Phys.* 1 (1967), 330-342.

[36] Concus, P., Golub, G.H., Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations, *SIAM J. Numer. Anal.* 10 (1973), 1103–1120.

[37] Concus, P., Golub, G.H., A generalized conjugate method for non-symmetric systems of linear equations, in: *Lect. Notes Math. Syst.* 134 (eds. Glowinski, R., Lions, J.-L.), pp. 56-65, Springer, 1976.

[38] Courant, R, Hilbert, D., *Methods of Mathematical Physics II.*, Wiley Classics Library, J. Wiley & Sons, 1989.

[39] Czách, L., *The steepest descent method for elliptic differential equations* (in Russian), C.Sc. thesis, 1955.

[40] de Figueiredo, D. G., Mitidieri, E., Maximum principles for cooperative elliptic systems, *C. R. Acad. Sci. Paris Sér. I Math.* 310 (1990), no. 2, 49–52.

[41] Díaz, J. I., Applications of symmetric rearrangement to certain nonlinear elliptic equations with a free boundary, *Nonlinear differential equations* (Granada, 1984), 155–181, Res. Notes in Math., 132, Pitman, 1985.

[42] Draganescu, A., Dupont, T. F., Scott, L. R., Failure of the discrete maximum principle for an elliptic finite element problem, *Math. Comp.* 74 (2005), no. 249, 1–23.

[43] D'yakonov, E. G., On an iterative method for the solution of finite difference equations (in Russian), *Dokl. Akad. Nauk SSSR* 138 (1961), 522–525.

[44] D'yakonov, E. G., The construction of iterative methods based on the use of spectrally equivalent operators, *USSR Comput. Math. and Math. Phys.*, 6 (1965), pp. 14-46.

[45] Eisenstat, S.C., Elman, H.C., Schultz. M.H., Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.* 20 (1983), no. 2, 345–357.

[46] Elman, H.C., Preconditioning for the steady-state Navier-Stokes equations with low viscosity, *SIAM J. Sci. Comput.,* 20 (1999), No.4, pp. 1299-1316.

[47] Elman, H.C., Golub, G. H., Inexact and preconditioned Uzawa algorithms for saddle point problems, *SIAM J. Numer. Anal.* 31, No.6, 1645-1661 (1994).

[48] Elman, H. C., Silvester, D. J., Wathen, A. J., *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[49] Elman, H.C., Schultz. M.H., Preconditioning by fast direct methods for non-selfadjoint nonseparable elliptic equations, *SIAM J. Numer. Anal.,* 23 (1986), 44-57.

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

[51] Faber, V., Manteuffel, T., Necessary and sufficient conditions for the existence of a conjugate gradient method, *SIAM J. Numer. Anal.* 21 (1984), no. 2, 352–362.

[52] 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.

[53] Faragó, I., Karátson, J., The gradient–finite element method for elliptic problems, *Comput. Math. Appl.* 42 (2001), 1043-1053.

[54] Faragó, I., Karátson, J., Gradient–finite element method for nonlinear Neumann problems, *J. Appl. Anal.* 7 (2001) No. 2, 257-269.

[55] 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.

[56] Faragó, I., Karátson, J., Preconditioning operators and Sobolev gradients for nonlinear elliptic problems, *Comput. Math. Appl.* 50 (2005), pp. 1077-1092.

[57] Faragó, I., Karátson, J., Sobolev gradient type preconditioning for the Saint-Venant model of elasto-plastic torsion, *Int. J. Numer. Anal. Modelling* Vol.5, No. 2, pp. 206-222 (2008)

[58] Friedrichs, K. O., On the boundary-value problems of the theory of elasticity and Korn's inequality. *Ann. of Math.* (2) 48, (1947). 441–471.

[59] Gajewski, H., Gröger, K., Zacharias, K., *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen,* Akademie-Verlag, Berlin, 1974

[60] Gidas, B., Ni, W.N., Nirenberg, L., Symmetry and related properties via the maximum principle, *Commun. Math. Phys.* **68** (1979), 209-243.

[61] Gilbarg, D., Trudinger, N. S., *Elliptic partial differential equations of second order* (2nd edition), Grundlehren der Mathematischen Wissenschaften 224, Springer, 1983.

[62] Glowinski, R., *Numerical methods for nonlinear variational problems.* Springer Series in Computational Physics. Springer-Verlag, New York, 1984.

[63] Glowinski, R., Marrocco, A., Analyse numérique du champ magnétique d'un alternateur par éléments finis et sur-relaxation ponctuelle non linéaire, *Comput. Methods Appl. Mech. Engrg.* 3 (1974), no. 1, 55–85.

[64] Gohberg, I., Goldberg, S., *Basic Operator Theory*, Birkhäuser, Boston, Mass., 1981.

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

[66] 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.

[67] Greenbaum, A., Diagonal scalings of the Laplacian as preconditioners for other elliptic differential operators, *SIAM J. Matrix Anal. Appl.,* 13 (1992), 826-846.

[68] Gunn, J. E., The numerical solution of $\nabla \cdot a\nabla u = f$ by a semi-explicit alternating direction iterative method, *Numer. Math.* 6 (1964), 181-184.

[69] Hackbusch, W., *Multigrid methods and applications*, Springer Series in Computational Mathematics 4, Springer, Berlin, 1985.

[70] Hackbusch, W., *Elliptic differential equations. Theory and numerical treatment*, Springer Series in Computational Mathematics 18, Springer, Berlin, 1992.

[71] Hannukainen, A., Korotov, S., Vejchodský, T., Discrete maximum principles for FE solutions of the diffusion-reaction problem on prismatic meshes, *J. Comput. Appl. Math.* 226 (2009), 275–287.

[72] Hárs, V., Tóth, J., On the inverse problem of reaction kinetics, In: *Qualitative Theory of Differential Equations* (Szeged, Hungary, 1979), Coll. Math. Soc. János Bolyai 30, ed. M. Farkas, North-Holland - János Bolyai Mathematical Society, Budapest, 1981, pp. 363-379.

[73] Hlaváček, I., Křížek, M., On a superconvergent finite element scheme for elliptic systems, *Apl. Mat.* 32 (1987). I. Dirichlet boundary condition, no. 2, 131–154; II. Boundary conditions of Newton's or Neumann's type, no. 3, 200–213; III. Optimal interior estimates, no. 4, 276–289.

[74] Hlaváček, I., Krížek, M., Malý, J., On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type, *J. Math. Anal. Appl.* 184 (1994), no. 1, 168–189.

[75] Horgan, C. O., Korn's inequalities and their applications in continuum mechanics, *SIAM Rev.* 37 (1995), no. 4, 491–511.

[76] Juncu, Gh., Preconditioning by approximations of the discrete Laplacian for 2-D nonlinear free convection elliptic equations, *Int. J. Numer. Methods Heat Fluid Flow* 9, No.5, 586-600 (1999).

[77] Kachanov, L.M., *Foundations of the theory of plasticity,* North-Holland, 1971

[78] Kadlec, J., On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a convex open set, *Czechosl. Math. J.*, 14 (89), (1964), pp. 386-393.

[79] Kantorovich, L.V., Akilov, G.P., *Functional Analysis,* Pergamon Press, 1984.

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

[81] Karátson J., Sobolev space preconditioning of strongly nonlinear fourth order problems, in: *Numerical Analysis and Its Applications*, eds. L. Vulkov, J. Wasniewski, P. Yalamov, pp. 459-466, *Lecture Notes Comp. Sci.* Vol. 1988, Springer, 2001.

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

[83] Karátson J., Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators *Appl. Math.* (Prague) 50 (2005), No. 3, 277-290.

[84] Karátson J., On the superlinear convergence rate of the preconditioned CGM for some nonsymmetric elliptic problems, *Numer. Funct. Anal.* 28 (2007), 9-10, pp. 1153-1164.

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

[86] Karátson J., On the Lipschitz continuity of derivatives for some scalar nonlinearities, *J. Math. Anal. Appl.* 346 (2008), pp. 170–176.

[87] Karátson J., Operator preconditioning with efficient applications for nonlinear elliptic problems (a survey), to appear in *Central Eur. J. Math.*, 2011.

[88] Karátson J., Characterizing mesh independent quadratic convergence of Newton's method for a class of elliptic problems, submitted (*SIAM J. Math. Anal.*, under revision); http://www.cs.elte.hu/applanal/preprints/Kar_new_mesh_ind.pdf

[89] Karátson J., Faragó I., Sobolev space preconditioning for nonlinear mixed boundary value problems, in: *Large-scale Scientific Computing,* eds. S. Margenov, J. Wasniewski, P. Yalamov, pp. 104-112, *Lecture Notes Comp. Sci.* Vol. 2179, Springer, 2001.

[90] 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.

[91] Karátson, J., Korotov, S., Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, *Numer. Math.* 99 (2005), 669–698.

[92] Karátson, J., Korotov, S., A discrete maximum principle in Hilbert space with applications to nonlinear cooperative elliptic systems, *SIAM J. Numer. Anal.* 47 (2009), No. 4., pp. 2518-2549.

[93] Karátson, J., Korotov, S., Discrete maximum principles for FEM solutions of some nonlinear elliptic interface problems, *Int. J. Numer. Anal. Modelling.* Vol.6, No. 1, pp. 1-16 (2009).

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

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

[96] 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.

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

[98] Karátson J., Neuberger, J. W., Newton's method in the context of gradients, *Electron. J. Diff. Eqns.* Vol. 2007(2007), No. 124, pp. 1-13.

[99] Keller, H. B., Elliptic boundary value problems suggested by nonlinear diffusion processes, *Arch. Rational Mech. Anal.* 35 (1969), 363–381.

[100] Korotov, S., A posteriori error estimation of goal-oriented quantities for elliptic type BVPs, *J. Comput. Appl. Math.* 191 (2006), pp. 216–227.

[101] Korotov, S., Two-sided a posteriori error estimates for linear elliptic problems with mixed boundary conditions, *Appl. Math.* 52 (2007), 235–249.

[102] Korotov, S., Křížek, M., Acute type refinements of tetrahedral partitions of polyhedral domains, *SIAM J. Numer. Anal.* 39 (2001), 724–733.

[103] Korotov, S., Křížek, M., Tetrahedral partitions and their refinements, In: *Proc. Conf. Finite Element Methods: Three-dimensional Problems, Univ. of Jyväskylä, GAKUTO Internat. Ser. Math. Sci. Appl.,* vol. 15, Gakkotosho, Tokyo, 2001, 118–134.

[104] Korotov, S., Křížek, M., Neittaanmäki, P., Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, *Math. Comp.* 70 (2001), 107–119.

[105] Kovács B., A comparison of some efficient numerical methods for a nonlinear elliptic problem, to appear in *Central Eur. J. Math.*, 2011.

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

[107] Křížek, M., Lin Qun, On diagonal dominance of stiffness matrices in 3D, *East-West J. Numer. Math.* 3 (1995), 59–69.

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

[109] Kuzmin, D., Hannukainen, A., Korotov, S., A new a posteriori error estimate for convection-reaction-diffusion problems, *J. Comput. Appl. Math.* 218 (2008), pp. 70–78.

[110] López-Gómez, J., Molina-Meyer, M., The maximum principle for cooperative weakly coupled elliptic systems and some applications, *Diff. Int. Equations* 7 (1994), no. 2, 383–398.

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

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

[113] Martinsson, P.G., A fast direct solver for a class of elliptic partial differential equations, *J. Sci. Comput.* (2009) 38: 316-330.

[114] Mayo, A., The fast solution of Poisson's and the biharmonic equations on irregular regions, *SIAM J. Numer. Anal.* 21 (1984), no. 2, 285–299.

[115] McCormick, S.F. (ed.), *Multigrid methods,* Frontiers in Applied Mathematics 3, SIAM, Philadelphia, 1987.

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

[117] Miersemann, E., Zur Regularität verallgemeinerter Lösungen von quasilinearen elliptischen Differentialgleichungen zweiter Ordnung in Gebieten mit Ecken, *Z. Anal. Anw.* 1 (1982), no. 4, 59–71.

[118] Mikhlin, S.G., *Constants in some inequalities of analysis* (translated from the Russian by R. Lehmann), John Wiley and Sons, Ltd., Chichester, 1986.

[119] Mitidieri, E., Sweers, G., Weakly coupled elliptic systems and positivity, *Math. Nachr.* 173 (1995), 259–286.

[120] Nečas, J., Hlaváček, I., *Mathematical Theory of Elastic and Elasto-Plastic Bodies: an Introduction,* Studies in Applied Mechanics 3, Elsevier Scientific Publishing Co., Amsterdam-New York, 1980.

[121] Neittaanmäki, P., Repin, S., A posteriori error estimates for boundary-value problems related to the biharmonic operator, *East-West J. Numer. Math.* 9 (2001), pp. 157–178.

[122] Neittaanmäki, P., Repin, S., *Reliable Methods for Computer Simulation. Error Control and A Posteriori Estimates,* Studies in Mathematics and its Applications, 33. Elsevier Science B.V., Amsterdam, 2004.

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

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

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

[126] Neumaier, A., Certified error bounds for uncertain elliptic equations, *J. Comput. Appl. Math.* 218 (2008), pp. 125–136.

[127] Nittka R., Sauter M., Sobolev gradients for differential algebraic equations, *Electron. J. Diff. Eqns.,* Vol. 2008(2008), No. 42, pp. 1-31.

[128] Pinkus, A., *n-widths in approximation theory,* Springer, 1985.

[129] Popa, C., Mesh independence of the condition number of discrete Galerkin systems by preconditioning, *Int. J. Comput. Math.* 51 (1994), p. 127.

[130] Popa, C., Mesh independence principle for nonlinear equations on Hilbert spaces by preconditioning, *Int. J. Comput. Math.* 69 (1998), no. 3-4, 295–318.

[131] Protas B., Adjoint-based optimization of PDE systems with alternative gradients, *J. Comput. Phys.* 227 (2008), 6490-6510.

[132] Protter, M. H., Maximum principles, in: Maximum principles and eigenvalue problems in partial differential equations, Proc. Conf., Knoxville/Tenn. 1987, Pitman Res. Notes Math. Ser. 175, 1-14 (1988).

[133] Protter, M. H., Weinberger, H. F., *Maximum principles in differential equations*, Springer-Verlag, New York, 1984.

[134] Repin, S., Sauter, S., Smolianski, A., A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions, *J. Comput. Appl. Math.* 164/165 (2004), 601–612.

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

[136] Riesz F., Sz.-Nagy B., *Vorlesungen über Funktionalanalysis*, Verlag H. Deutsch, 1982.

[137] Rossi, T., Toivanen, J., A parallel fast direct solver for block tridiagonal systems with separable matrices of arbitrary dimension, *SIAM J. Sci. Comput.* 20 (1999), no. 5, 1778–1796 (electronic).

[138] Segeth, K., A review of some a posteriori error estimates for adaptive finite element methods, *Math. Comput. Simul.* 80 (2010), no. 8, 1589-1600.

[139] Sirakov, B., Some estimates and maximum principles for weakly coupled systems of elliptic PDE, *Nonlinear Anal., Theory Methods Appl.* 70, No. 8, A, 3039-3046 (2009).

[140] Stoyan, G., On a maximum principle for matrices, and on conservation of monotonicity. With applications to discretization methods, *Z. Angew. Math. Mech.* 62, 375-381 (1982).

[141] Stoyan, G., Iterative Stokes solvers in the harmonic Velte subspace, *Computing* 67, No.1, 13-33 (2001).

[142] Stynes, M., Steady-state convection-diffusion problems, *Acta Numer.* 14 (2005), 445–508.

[143] Swarztrauber, P. N., A direct method for the discrete solution of separable elliptic equations, *SIAM J. Numer. Anal.* 11 (1974), 1136–1150.

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

[145] Tóth, J., Gradient systems are cross-catalytic, *Reaction Kinetics and Catalysis Letters* 12 (3) (1979), 253–257.

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

[147] Varga, R., *Matrix iterative analysis*, Prentice Hall, New Jersey, 1962.

[148] Vladimirov, V. S., *Equations of Mathematical Physics* (translated from the Russian by E. Yankovsky), Mir, Moscow, 1984.

[149] Vozovoi L., Israeli M. and Averbuch A., A fast Poisson solver of arbitrary order accuracy in rectangular regions, *SIAM J. Sci. Comput.* 19 (1998), No.3, pp. 933–952.

[150] Vörös, G. (ELTE, Institute of Physics), personal communication.

[151] Vörös, I., Stability properties of non-negative solutions of semilinear symmetric cooperative systems, *Electron J. Diff. Eqns.* 2004 (2004), No. 105, pp. 1-6.

[152] Waterhouse, W. C., The absolute-value estimate for symmetric multilinear forms, *Linear Algebra Appl.* 128 (1990), 97–105.

[153] 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.

[154] Widlund, O., On the use of fast methods for separable finite difference equations for the solution of general elliptic problems, in *Sparse Matrices and Applications*, D.J. Rose and R.A. Willoughby (eds.), Plenum Press, N.Y. 1972, pp. 121–134.

[155] Widlund, O., A Lanczos method for a class of non-symmetric systems of linear equations, *SIAM J. Numer. Anal.,* 15 (1978), 801-812.

[156] Winter, R., Some superlinear convergence results for the conjugate gradient method, *SIAM J. Numer. Anal.*, 17 (1980), 14-17.

[157] Xu, J., Zikatanov, L., A monotone finite element scheme for convection-diffusion equations, *Math. Comp.* 68 (1999), 1429–1446.

[158] Zeidler, E., *Nonlinear functional analysis and its applications,* Springer, 1986

[159] Ženíšek, A., *Nonlinear elliptic and evolution problems and their finite element approximations.* Computational Mathematics and Applications, Academic Press, Inc., London, 1990

[160] Zlatev, Z., *Computer treatment of large air pollution models*, Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.