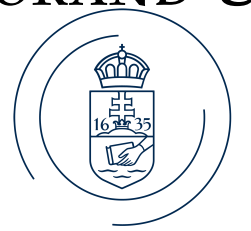
Eötvös Loránd University



ELTE
EÖTVÖS LORÁND
TUDOMÁNYEGYETEM

Institute of Mathematics

Master's thesis project

Newton-Krylov methods for nonlinear elliptic systems

Author:
Sebastian Josué Castillo Jaramillo

Adviser:
Dr. János Karátson

Budapest - May 26, 2023

Acknowledgments

First and foremost, I would like to thank God for presenting me with all the opportunities that allowed me to pursue my dream to study mathematics.

I would also like to thank my supervisor Dr. Karátson János for introducing me to this topic, and for his guidance and kindness. Without your advice and explanations, I could not have completed this project.

Finally, I would like to thank my friends and family for their love and support. No matter the distance, you have always been there for me.

Abstract

We consider numerical solutions of linear and non-linear elliptic systems of PDEs and their finite element discretizations. We obtain eigenvalue-based estimations of the rate of superlinear convergence of some preconditioned conjugate gradient-type methods in the case of symmetric and non-symmetric linear elliptic systems. These results are used in the nonlinear case to provide mesh independence estimations of the rate of superlinear convergence for inner iterations of a Newton-Krylov method in terms of the growth power property of the non-linearities. Numerical examples are implemented to verify our findings.

Keywords: Superlinear convergence, preconditioned conjugate gradient type methods, Sobolev spaces, elliptic partial differential equations, nonlinear elliptic transport systems, iterative methods.

Contents

Introduction

Linear and non-linear elliptic partial differential equations (PDEs) describe a large range of physical models, from diffusion phenomena to modeling the transport of air pollutants. Unfortunately, finding an analytical solution for such kinds of problems is often impossible. Another approach is to construct a finite element or finite difference discretization to solve our PDE approximately. The key issue is to solve the arising systems using some iterative processes. In this setting, there are many advantages to using iterative methods, such as memory cost efficiency and preserving the sparsity of the system. Some examples of widely used iterative solvers are the preconditioned conjugate gradient method (PCGM) for the symmetric linear case and the damped inexact Newton (DIN) method for the nonlinear case. Furthermore, for non-symmetric linear problems, several CG-type methods have been developed such as the generalized minimal residual algorithm (GMRES), [17]. For the PCG and GMRES algorithms, preconditioning is a fundamental part of the iterative process as it speeds up convergence. Moreover, sometimes it allows us to obtain, under certain conditions, mesh-independent superlinear convergence, [3]. When solving large systems of equations, by choosing a proper preconditioner we can transform the system into a simpler one to solve, see Remark 2.6. One approach for constructing such a preconditioner is to find an approximation of the original elliptic operator and choose as a preconditioner its discretization. This technique involves the theory of equivalent and compact-equivalent operators on Hilbert spaces. For a survey on this subject, see [4].

The mesh-independent superlinear convergence feature ensures that the rate of convergence of the method does not deteriorate as the mesh is refined and that the number of steps necessary to obtain a new correct digit in the approximate solution will be decreasing over the course of the iteration, [4]. There are many instances where this property comes in handy. For instance, when PCG or GMRES is used as an inner iteration of an outer process, such as the DIN algorithm. Here, for each step of the method, we require to solve some linear problem and therefore we seek optimal solvers in order not to spoil the overall convergence speed of the method.

In this master thesis, we consider different kinds of preconditioned second-order elliptic systems and their finite element discretization. We study the mesh-independent superlinear convergence of the PCGM applied to symmetric linear elliptic PDEs, and GMRES for the non-symmetric case. The main goal is to find an eigenvalue-based estimation of the concrete rate of superlinear convergence for such methods and show that a similar estimate can be obtained in the case of systems of PDEs. This extends previous results of [13] to the case of unbounded reaction coefficients in some Lebesgue spaces. Additionally, we analyze inner-outer iterations of the *Damped Inexact Newton*

(DIN) method applied to the finite element discretization of some non-linear systems of PDEs. Our goal is to give more explicit estimates for the rate of superlinear convergence obtained in [1], where the DIN plus CGN technique is used. We achieve this by replacing the CGN method with GMRES and then applying our results obtained for linear non-symmetric systems.

This work is organized as follows. In Chapter 1, we list some definitions and results from functional and numerical analysis that are essential for our work.

In Chapter 2, we prove our estimations of the rate of superlinear convergence for different linear problems: first for single equations, then for symmetric, and finally for non-symmetric systems.

In Chapter 3, we consider a family of nonlinear elliptic transport systems and their finite element discretization and approximate its solution using the DIN method. We realize this method as an inner-outer process and use GMRES to solve the linearized problem at each step of DIN. Then, we apply the results from Chapter 2 to obtain estimations on the rate of superlinear convergence of the inner iterations.

Finally, in Chapter 4, we present our conclusions and recommendations.

Chapter 1

Theoretical background: some basic results and definitions

This chapter presents various concepts and results of functional and numerical analysis, which are fundamental in this work. We start with some basic definitions and results on Lebesgue and Sobolev spaces. Then we briefly introduce various kinds of operators and some well-posedness theorems for linear and nonlinear operator equations. Finally, we discuss some results in numerical analysis. The terminology and notation used in this work are standard. We shall mostly work over the field $\mathbb{R}$. The main references are [7], [10], [11], [16].

1.1 Functional analysis

1.1.1 Sobolev and Lebesgue spaces

The study of the properties of Sobolev and Lebesgue spaces is a key part of the theory of PDEs. In this subsection, we present some of its most important results and definitions. We shall assume $\Omega \subset R^d$ for an arbitrary positive integer $d \geq 2$.

Lebesgue spaces

Let $1 \leq p < \infty$. For $f \in C_0^\infty(\Omega)$ we use the notation

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx \right)^{1/p}, \quad \|f\|_{L^\infty(\Omega)} = \sup_{x \in \Omega} |f(x)|. \tag{1.1}$$

This represents a norm in $C_0^\infty(\Omega)$ and is called the L^p − norm. The completion

$$L^p(\Omega) = \overline{\left(C_0^\infty(\Omega), \|\cdot\|_{L^p(\Omega)} \right)}$$

is the *Lebesgue space* $\mathrm{L}^p(\Omega)$ and its elements are equivalence classes given by the equivalence relation

$$f \sim g \iff \int_\Omega |f(x) - g(x)|dx = 0.$$

The next result provides a useful inequality in Lebesgue spaces.

Theorem 1.1 (Hölder's inequality)**.** *For any $f \in \mathrm{L}^p(\Omega)$ and $g \in \mathrm{L}^{p'}(\Omega)$, fg is in $\mathrm{L}^1(\Omega)$ and*

$$\|fg\|_{\mathrm{L}^1(\Omega)} \leq \|f\|_{\mathrm{L}^p(\Omega)} \|g\|_{\mathrm{L}^{p'}(\Omega)}$$

where $\frac{1}{p} + \frac{1}{p'} = 1$.

Remark 1.2. *In the next chapters, we shall use a more generalized version of Hölder's inequality, see e.g [7, Remark 2, Chapter 4.2]. Assume that $f_1, \ldots, f_k \in \mathrm{L}^{p_i}(\Omega)$ such that $\frac{1}{p_1} + \cdots + \frac{1}{p_k} = \frac{1}{p} \leq 1$. Then*

$$\|f\|_{\mathrm{L}^p(\Omega)} \leq \|f_1\|_{\mathrm{L}^{p_1}(\Omega)} \cdots \|f_k\|_{\mathrm{L}^{p_k}(\Omega)}.$$

We finish this topic by defining the space of *locally integrable functions*. We use the notation χ_K for the characteristic function of a set K:

$$\chi_K(x) = \begin{cases} 1 & , \text{ if } x \in K, \\ 0 & , \text{ if } x \notin K. \end{cases}$$

We say that a function $f : \Omega \to \mathbb{R}$ belongs to $L^p_{loc}(\Omega)$ if $f\chi_K \in L^p(\Omega)$ for every compact set K contained in Ω.

Sobolev spaces

We begin with a brief introduction to weak derivatives since Sobolev spaces are built using this notion. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a multiindex of order $|\alpha| = k$ and $u \in \mathrm{C}^k(\Omega)$. We use the notation

$$D^\alpha u(x) = \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n} u(x).$$

If $v \in \mathrm{C}_0^\infty(\Omega)$, then integration by parts yields

$$\int_\Omega u D^\alpha v = (-1)^{|\alpha|} \int_\Omega v D^\alpha u.$$

We can weaken this notion by allowing u to belong to a broader space. This motivates the definition of the weak derivative.

Definition 1.3. *Let $u \in L^p_{loc}(\Omega)$. If α is a multiindex, the α^{th}-weak partial derivative of u is a function $w \in L^1_{loc}(\Omega)$ such that*

$$\int_\Omega u D^\alpha v = (-1)^{|\alpha|} \int_\Omega uw$$

for all $v \in \mathrm{C}_0^\infty(\Omega)$. We write $w = D^\alpha u$.

Fix $1 \leq p \leq \infty$ and let k be a nonnegative integer. The *Sobolev space* $\mathrm{W}^{k,p}(\Omega)$ consists of all functions $u \in \mathrm{L}^1_{loc}(\Omega)$ such that for each multi-index α with $|\alpha| \leq k$, $D^\alpha u$ exists in the weak sense and belongs to $L^p(\Omega)$, i.e.,

$$\mathrm{W}^{k,p}(\Omega) = \{u \in \mathrm{L}^1_{loc}(\Omega) / \forall |\alpha| \leq k : D^\alpha u \in \mathrm{L}^p(\Omega)\}.$$

We shall focus on the special case when $p = 2$ and $k = 2$. Here, we write

$$\mathrm{H}^1(\Omega) = \mathrm{W}^{1,2}(\Omega).$$

This is a Hilbert space under the inner product

$$\langle f, g \rangle_{\mathrm{H}^1(\Omega)} = \int_\Omega \nabla f \cdot \nabla g + \int_\Omega fg \quad (f, g \in \mathrm{H}^1(\Omega)).$$

The completion of $\mathrm{C}_0^\infty(\Omega)$ under this inner product is denoted as $\mathrm{H}_0^1(\Omega)$. Furthermore,

$$\langle f, g \rangle_{\mathrm{H}_0^1(\Omega)} = \int_\Omega \nabla f \cdot \nabla g \quad (f, g \in \mathrm{H}_0^1(\Omega))$$

is an equivalent inner product in $\mathrm{H}_0^1(\Omega)$.

We finish this subsection by giving an important inequality for functions in $\mathrm{H}_0^1(\Omega)$. This result is called *Poincaré's inequality*.

Theorem 1.4. *Assume that Ω is a bounded open subset of $\mathbb{R}^d$. Then, for any $p \leq 2^* := \frac{2d}{d-2}$ and $u \in \mathrm{H}_0^1(\Omega)$*

$$\|u\|_{\mathrm{L}^p(\Omega)} \leq C_p \|u\|_{\mathrm{H}_0^1(\Omega)},$$

for some constant $C_p = C(p, d, \Omega) > 0$.

A proof of this theorem can be found in [10, Ch.5, Th. 3].

1.1.2 Operator theory

Let X, Y be Banach spaces. We denote by $B(X, Y)$ the space of all bounded linear operators $L : X \to Y$, i.e.,

$$\|Lx\|_Y \leq c\|x\|_X \quad (x \in X), \tag{1.2}$$

for some $c > 0$ independent of x. Furthermore, B(X,Y) is a Banach space with the *operator norm* defined by

$$\|L\| = \sup_{x \neq 0} \frac{\|Lx\|_Y}{\|x\|_X}.$$

Note that $\|L\|$ is the smallest constant such that (1.2) holds.

Example 1.5. *If $Y = \mathbb{R}$, then $B(X, \mathbb{R}) =: X^*$ is called the dual space of X. Here, we shall use the usual pairing notation between X^* and X. That is, if $\phi \in X^*$, then*

$$\langle \phi, v \rangle_{X^*, X} = \phi v \quad (v \in X).$$

The above notation is motivated by the following theorem. Furthermore, this result characterizes the dual space of a Hilbert space.

Theorem 1.6 (Riesz-Fréchet theorem)**.** *Let H be a Hilbert space. For each $\psi \in H^*$, there exists a unique $y_\psi \in H$ such that*

$$\psi x = \langle x, y_\psi \rangle, \text{ and } \|\psi\| = \|y_\psi\| \quad (x \in H).$$

A proof of this theorem can be found in [14, Th. II.4].

The following class of operators plays a major role in our work and they are very useful since they behave in some way similarly to operators in finite-dimensional spaces.

Definition 1.7. *Let X, Y be a pair of Banach spaces. A linear operator $L : X \to Y$ is called* compact *if for any bounded sequence $(x_n)_{n\in\mathbb{N}} \subset X$*

$$(Lx_n)_{n\in\mathbb{N}} \text{ has a convergent subsequence.}$$

Proposition 1.8. *Let X, Y, Z be Banach spaces. Let $S \in B(Y, Z)$ and $T : X \to Y$ or $T : Y \to X$. Assume that T or S is a compact operator. Then their product ST and TS are compact.*

Example 1.9 (Sobolev embeddings)**.** *Let $p < \frac{2d}{d-2}$. Then the operator $\mathcal{I} \colon \mathrm{H}_0^1(\Omega) \mapsto \mathrm{L}^p(\Omega)$ defined by $\mathcal{I}(u) = u$ is a compact operator. We say that* $\mathrm{H}_0^1(\Omega)$ is compactly embedded *in* $\mathrm{L}^p(\Omega)$*. Indeed, from Theorem 1.4 we know that for any $u \in \mathrm{H}_0^1(\Omega)$*

$$\|u\|_{\mathrm{L}^p(\Omega)} \leq C_p \|u\|_{\mathrm{H}_0^1(\Omega)},$$

for some constant $C_p = C(p, d, \Omega) > 0$. Hence $\mathcal{I}$ is well defined. Further, it can be proved that if $(u_m)_{m\in\mathbb{N}} \subset \mathrm{H}_0^1(\Omega)$ is bounded, then there exists a subsequence $(u_{m_j})_{j\in\mathbb{N}}$ which converges in $\mathrm{L}^p(\Omega)$.

This result is a special case of Rellich-Kondrachov theorem*, see e.g. [10].*

In addition to bounded and compact linear operators, there are other families of operators whose properties we shall use in the next chapters. In the following, we summarize them in a list. Let H be a Hilbert space. We say that a linear operator $L : H \to \mathrm{D}(L) \subset H$ is:

- **Unbounded** if $\sup_{u\neq 0} \frac{\|Lu\|}{\|u\|} = +\infty$.
- **Symmetric** if $\langle Lx, y\rangle = \langle x, Ly\rangle$ for any $x, y \in \mathrm{D}(L)$.
- **Positive/strictly positive** if for any $x \in \mathrm{D}(L)$: $\langle Lx, x\rangle \geq 0$, $\langle Lu, u\rangle > 0$ for $u \neq 0$, respectively.
- **Uniformly positive** if for any $x \in \mathrm{D}(L)$: $\langle Lx, x\rangle \geq m\|x\|^2$, for some $m > 0$ independent of x.

Let L be strictly positive and symmetric. Then the *energy inner product* of L is defined by

$$\langle u, v\rangle_L = \langle Lu, v\rangle.$$

Further, the induced energy norm is denoted by

$$\|u\|_L = \langle Lu, u\rangle^{\frac{1}{2}}.$$

Definition 1.10. *The energy space H_L is defined as the completion of* $\mathrm{D}(L)$ *under the energy norm, i.e.,*

$$H_L = \overline{(\mathrm{D}(L), \langle\cdot,\cdot\rangle_L)}.$$

Then H_L is a Hilbert space.

Note that if H is a complex Hilbert space, then the strict positivity of L implies symmetry, and thus assuming L is strictly positive is enough.

Let $L : H \to H$ be symmetric and uniformly positive and $g \in H$. We can define the weak form of the equation $Lu = g$ as follows

$$\langle u, v\rangle_L = \langle g, v\rangle \quad (v \in H_L).$$

It is well known that there exists a unique $u \in H_L$ such that the equation above holds. We call u the weak solution of the operator equation $Lu = g$.

If L is not symmetric, then we introduce an auxiliary operator $S : H \to H$ which is symmetric and uniformly positive. Then, we look for weak solutions of $Lu = g$ in H_S. For this, we require some conditions for L:

(i) $\mathrm{D}(L) \subset H_S$ and it is dense w.r.t $\|\cdot\|_S$.

(ii) **S-boundedness**: there exists $M > 0$ such that $|\langle Lu, v\rangle| \leq M\|u\|_S\|v\|_S$ for all $u, v \in \mathrm{D}(L)$.

(iii) **S-coercivity**: there exists $m > 0$ such that $\langle Lu, u\rangle \geq m\|u\|_S^2$ for all $u, v \in \mathrm{D}(L)$.

Under these conditions, it can be proved that there exists a unique bounded linear operator $L_S : H_S \to H_S$ such that $\langle L_S u, v\rangle_S = \langle Lu, v\rangle$. Hence, the weak formulation of the operator equation $Lu = g$ becomes

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \quad (v \in H_S).$$

Further, it can be proved that $u \in H_S$ exists and is unique.

Let us now study the nonlinear case. Here, in order to prove the existence and uniqueness of weak solutions we require some extra conditions on the operator.

Definition 1.11. *Let X, Y be normed spaces. We say that $F : X \to Y$ is* Gateaux differentiable (GD) *at $u \in X$ if*

(i) The following limit exists for any $v \in X$:

$$\partial_v F(u) = \lim_{t\to 0} \frac{F(u + tv) - F(u)}{t}.$$

(ii) The mapping $F'(u) : v \mapsto \partial_v F(u)$ is a linear bounded operator.

Remark 1.12. *Any bounded linear operator $L : X \to Y$ is GD. In fact, for any $u, v \in X$ we get*

$$\lim_{t \to 0} \frac{L(u + tv) - Lu}{t} = Lv$$

and $L'(u)v = Lv$.

The following theorem is a classical result on Gateaux differentiable functionals.

Theorem 1.13 (Lagrange mean value theorem)**.** *Let X be a normed space and $u, v \in X$ arbitrary. Let $\phi : X \to \mathbb{R}$ be GD. Then there exists $\xi \in [u, v] := \{u + t(v - u) : t \in [0, 1]\}$ such that*

$$\phi(u) - \phi(v) = \langle \phi'(\xi), v - u \rangle_{X^*, X}.$$

We finish this section with a well-posedness theorem for Gateaux differentiable non-linear operators.

Theorem 1.14. *Let $H \to H$ be a Hilbert space, $A : H \to H$ be a GD operator satisfying the following conditions:*

(i) ***Uniform monotonicity****: there exists $m > 0$ such that $\langle A'(u)h, h \rangle \geq m\|h\|^2$, for any $u, h \in H$.*

(ii) ***Boundedness****: there exists $M > 0$ such that $|\langle A'(u)h, v \rangle \leq M\|h\|\|v\|$, for any $u, h, v \in H$.*

Then, for any $b \in H$, there exists a unique $u \in H$ such that $A(u) = b$.

Remark 1.15. *We can allow M in the second condition on Theorem 1.14 to depend on $\|u\|$, see e.g. [11]. That is, we replace (ii) by the weaker condition*

$$|\langle A'(u)h, v \rangle| \leq M(\|u\|) \cdot \|h\|\|v\|,$$

for all $u, v, h \in H$.

1.2 Numerical analysis

Let us start by introducing the following well-posedness theorem on the variational problem.

Theorem 1.16 (Lax-Milgram theorem)**.** *Let H be a Hilbert space and $a : H \times H \to \mathbb{R}$ be a bilinear form which is*

(i) ***Bounded****, i.e., there exists* $M > 0$ *such that* $|a(u,v)| \leq M\|u\|\|v\|$ *for any* $u, v \in H$.

(ii) ***Coercive****, i.e., there exists* $m > 0$ *such that* $a(u,u) \geq m\|u\|^2$ *for any* $u \in H$.

Then for any bounded linear functional $\ell : H \to \mathbb{R}$, *the* variational problem

$$a(u,v) = \ell v \quad (v \in H)$$

has a unique solution $u \in H$.

Next, we introduce a class of methods for approximating a solution to the variational problem. These are called *Galerkin-type methods*.

Let H be a Hilbert space and $a : H \times H \to \mathbb{R}$ a bounded, coercive, symmetric bilinear form. Let V_H be a finite N dimensional subspace of H and $\{\psi_1, \dots \psi_N\}$ a basis in V_h. We look for solutions of the *projected equation*

$$a(u_h, v_h) = \ell v_h \quad (v_h \in V_h).$$

Indeed, by substituting $u_h = \sum_{j=1}^N c_j \psi_j$ in the projected equation and letting $v_h = \psi_i$, we get

$$\sum_{j=1}^N a(\psi_j, \psi_i) c_j = \ell \psi_i \quad (i = 1, \dots, N).$$

Hence, by denoting $A_h = \{a(\psi_i, \psi_j)\}_{i,j=1}^N$, $b_h = \{\ell\psi_i\}_{i=1}^N$ and $c = \{c_i\}_{i=1}^N$ we obtain the algebraic linear system $A_h c = b$. This system has a unique solution $c \in \mathbb{R}^N$ since u_h is unique by the Lax-Milgram theorem.

It is well known, see e.g [11], that if $\{V_h\}_{h>0}$ be a family of finite-dimensional subspaces of H such that

$$\lim_{h\to 0} \operatorname{dist}(u, V_h) = 0 \quad (u \in H)$$

then the Galerkin method converges in the sense that

$$\lim_{h\to 0} \|u_h - u^*\| = 0,$$

where u^* denotes the exact solution of the variational problem.

Chapter 2

Linear elliptic problems

The preconditioned conjugate gradient method (PCGM) is a widespread way to find the solution of discretized elliptic partial differential equations iteratively. Furthermore, the preconditioned CGM can be competitive with multigrid methods and, under certain conditions, operator preconditioning can provide mesh-independent superlinear convergence. In this chapter, we consider a self-adjoint second-order elliptic boundary value problem with variable zeroth order coefficient and its finite element discretization. We study the mesh-independent superlinear convergence of the preconditioned CGM for this type of problem see e.g [13], [4], and extend previous results of [13] to the case of unbounded reaction coefficients in some Lebesgue spaces. Our goal is to find an eigenvalue-based estimation of the rate of superlinear convergence and to show that a similar estimate can be obtained in the case of systems of PDEs.

2.1 General framework

2.1.1 The linear operator equation and its Galerkin discretization

Let H be a real Hilbert space and let us consider a linear operator equation

$$Au = g \tag{2.1}$$

with some $g \in H$, under the following

Assumptions 2.1.1:

(i) The operator A is decomposed as

$$A = S + Q \tag{2.2}$$

where S is a symmetric operator in H with dense domain D and Q is a compact self-adjoint operator defined on the domain H.

(ii) There exists $k > 0$ such that $\langle Su, u\rangle \geq k\|u\|^2$ $(\forall u \in D)$.

(iii) $\langle Qu, u\rangle \geq 0 \ \ (\forall u \in H)$.

We recall that the energy space H_S is the completion of D under the *energy inner product*

$$\langle u, v\rangle_S = \langle Su, v\rangle, \tag{2.3}$$

and the corresponding norm is denoted by $\|\cdot\|_S$. Assumption (ii) implies $H_S \subset H$. Then there exists a unique bounded linear operator, denoted by $Q_S : H_S \mapsto H_s$, such that

$$\langle Q_S u, v\rangle_S = \langle Qu, v\rangle \qquad (\forall u, v \in H_S).$$

We replace equation (2.1) by its formally preconditioned form

$$Bu \equiv S^{-1}Au = S^{-1}g,$$

that is, $(I + S^{-1}Q)u = S^{-1}g$ in H_S. This gives the weak formulation

$$\langle (I + Q_S)u, v\rangle_S = \langle g, v\rangle \qquad (\forall v \in H_s). \tag{2.4}$$

Since by assumption (iii) the bilinear form on the left is coercive on H_S, by the *Lax-Milgram theorem*, there exists a unique solution $u \in H_S$ of (2.4).

Now equation (2.4) is solved numerically using a *Galerkin discretization*. Consider a given finite-dimensional subspace $V = \text{span}\{\varphi_1, \ldots, \varphi_n\} \subset H_S$, and let

$$\mathbf{S}_h = \{\langle \varphi_i, \varphi_j\rangle_S\}_{i,j=1}^n \quad \text{and } \mathbf{Q}_h = \{\langle Q\varphi_i, \varphi_j\rangle\}_{i,j=1}^n$$

the *Gram matrices* corresponding to S and Q. We look for the numerical solution $u_V \in V$ of equation (2.4) in V, i.e., for which

$$\langle (I + Q_S)u_v, v\rangle_S = \langle g, v\rangle \qquad (\forall v \in V). \tag{2.5}$$

Then $u_V = \sum_{i,j=1}^n c_j\varphi_j$, where $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ is the solution of the system

$$(\mathbf{S}_h + \mathbf{Q}_h)\mathbf{c} = \mathbf{b} \tag{2.6}$$

with $\mathbf{b} = \{\langle g, \varphi_j\rangle\}_{j=1}^n$. The matrix $\mathbf{A}_h := \mathbf{S}_h + \mathbf{Q}_h$ is SPD.

By using matrix $\mathbf{S}_h$ as the preconditioner for the system (2.6), we shall work with the preconditioned system

$$(\mathbf{I} + \mathbf{S}_h^{-1}\mathbf{Q}_h)\mathbf{c} = \tilde{\mathbf{b}}, \tag{2.7}$$

where $\mathbf{I}$ is the identity matrix in $\mathbb{R}^n$ and $\tilde{\mathbf{b}} = \mathbf{S}_h^{-1}\mathbf{b}$. We apply the CGM for the solution of this system.

2.1.2 The preconditioned conjugate gradient method and superlinear convergence

Let us consider a general linear system $\mathbf{A}u = \mathbf{g}$ and its preconditioned form

$$\mathbf{B}u = \tilde{\mathbf{g}}, \tag{2.8}$$

where $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}$ and $\tilde{\mathbf{g}} = \mathbf{S}^{-1}\mathbf{g}$. The preconditioner $\mathbf{S}_h$ induces the energy inner product $\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbf{S}_h} := \mathbf{S}_h\, \mathbf{c} \cdot \mathbf{d}$.

Then the PCG method is given by the following algorithm. Let u_0 be arbitrary, $\rho_0 = \mathbf{A}u_0 - \mathbf{g}$, $\mathbf{S}p_0 = \rho_0$, $r_0 = \rho_0$ and for $k \in \mathbb{N}$

$$\begin{cases} u_{k+1} = u_k + \alpha_k p_k, \\ r_{k+1} = r_k + \alpha_k \mathbf{S}^{-1}\mathbf{A}p_k, \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases}$$

with

$$\alpha_k = \frac{-\|r_k\|_{\mathbf{S}}^2}{\langle \mathbf{A}p_k, p_k \rangle}, \quad \beta_k = \frac{\|r_{k+1}\|_{\mathbf{S}}^2}{\|r_k\|_{\mathbf{S}}^2}.$$

In fact, the vector $z_k := \mathbf{S}^{-1}\mathbf{A}p_k$ is computed by solving the auxiliary problem

$$\mathbf{S}z_k = \mathbf{A}p_k\,.$$

Moreover, setting $w_k = z_k - p_k$, this problem is equivalent to

$$\begin{cases} \mathbf{S}w_k = \mathbf{Q}p_k, \\ z_k = w_k + p_k\,. \end{cases} \tag{2.9}$$

We are interested in the superlinear convergence rates for the CGM, and now recall the corresponding well-known estimation. Let $\mathbf{A} = \mathbf{S} + \mathbf{Q}$. Then $\mathbf{B}$ in (2.8) has the compact perturbation form $\mathbf{B} = \mathbf{I} + \mathbf{E}$ with $\mathbf{E} := \mathbf{S}^{-1}\mathbf{Q}$. Let us order the eigenvalues of the latter according to $|\lambda_1(\mathbf{S}^{-1}\mathbf{Q})| \geq |\lambda_2(\mathbf{S}^{-1}\mathbf{Q})| \geq \cdots \geq |\lambda_n(\mathbf{S}^{-1}\mathbf{Q})|$. Then the *error vectors* $e_k := c_k - c$ are measured by $\langle \mathbf{B}e_k, e_k \rangle_{\mathbf{S}}^{1/2} = \left\langle \mathbf{S}^{-1}\mathbf{A}e_k, e_k \right\rangle_{\mathbf{S}}^{1/2} = \langle \mathbf{A}e_k, e_k \rangle^{1/2} = \|e_k\|_{\mathbf{A}}$, and they are known to satisfy

$$\left(\frac{\|e_k\|_{\mathbf{A}}}{\|e_0\|_{\mathbf{A}}} \right)^{1/k} \leq \frac{2\|\mathbf{B}^{-1}\|_{\mathbf{S}}}{k} \sum_{j=1}^{k} |\lambda_j(\mathbf{S}^{-1}\mathbf{Q})| \qquad (k = 1, 2, \ldots, n), \tag{2.10}$$

see, e.g., [2].

For the discretized problem described in subsection 2.1.1, the following result allows us to give a convergence rate for the upper bound of (2.10) through the eigenvalues of the operator Q_S. This is a modification of Theorem 1 in [13] where the square of eigenvalues was considered.

Lemma 2.1. *Let Assumptions 2.1.1 hold. Then for any $k = 1, 2, \ldots, n$*

$$\sum_{j=1}^{k} |\lambda_j(\mathbf{S}_h^{-1}\mathbf{Q}_h)| \leq \sum_{j=1}^{k} \lambda_j(Q_S). \tag{2.11}$$

Proof. We have in fact

$$\sum_{j=1}^{k} \sigma_j(\mathbf{S}_h^{-1}\mathbf{Q}_h) \leq \sum_{j=1}^{k} \sigma_j(Q_S), \tag{2.12}$$

where the σ_j denote the singular values of the given matrix or operator, see [6]. Now both the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ (w.r.t. the $\mathbf{S}_h$-inner product) and the operator Q_S (in H_S) are self-adjoint, hence their singular values coincide with the modulus of the eigenvalues. Since Q_S is a positive operator from assumption (iii), the modulus can be omitted. □

An immediate consequence of this lemma is the following mesh-independent bound.

Corollary 2.2. *For any* $k = 1, 2, \ldots, n$

$$\left(\frac{\|e_k\|_{\mathbf{A}_h}}{\|e_0\|_{\mathbf{A}_h}}\right)^{1/k} \leq \frac{2\|B^{-1}\|_S}{k} \sum_{j=1}^{k} \lambda_j(Q_S) \qquad (k = 1, 2, \ldots, n). \tag{2.13}$$

Proof. By [3, Prop. 4.1], we are able to estimate $\|\mathbf{B}^{-1}\|_{\mathbf{S}} \leq \|B^{-1}\|_S$. This, together with (2.10) and (2.11), completes the proof. □

Since $|\lambda_1(Q_S)| \geq |\lambda_2(Q_S)| \geq \cdots \geq 0$ and the eigenvalues tend to 0, the convergence factor is less than 1 for k sufficiently large. Hence the upper bound decreases as $k \to \infty$ and we obtain superlinear convergence rate.

2.2 Estimation of the rate of superlinear convergence

In this section, we apply the previous abstract setting to different kinds of second-order elliptic boundary value problems with a general variable coefficient η in some Lebesgue space $\mathrm{L}^q(\Omega)$. First, we develop the results in detail for single equations. Afterward, we extend the estimates for symmetric systems of PDEs and we show that these estimates also work for the nonsymmetric case by applying an adequate CG-type algorithm. This situation shows the real strength of the idea of preconditioning operators since one can reduce large coupled systems of PDEs to independent single PDEs, hence the numerical solution of the latter can be parallelized. In each case, we provide an estimation of the rate of mesh-independent superlinear convergence such that the dependence of the rate on the integrability exponent of the reaction coefficient is determined.

2.2.1 Single elliptic equations

Let $d \geq 2$, $p > 2$ and $\Omega \subset \mathbb{R}^N$ be a bounded domain. We consider the elliptic problem

$$\begin{cases} -\mathrm{div}(G\nabla u) + \eta u = g, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{2.14}$$

under the standard assumptions listed below. We shall focus on the case when the principal part has constant or separable coefficients, i.e.,

$$G(x) \equiv G \in \mathbb{R}^N \times \mathbb{R}^N \quad \text{or } G(x) \equiv \operatorname{diag}\{G_i(x_i)\}_{i=1}^N$$

whereas $\eta = \eta(x)$ is a general variable (i.e. nonconstant) coefficient.

Assumptions 2.2.1:

(i) The symmetric matrix-valued function $G \in \mathrm{L}^\infty(\overline{\Omega}, \mathbb{R}^N \times \mathbb{R}^N)$ satisfies

$$G(x)\xi \cdot \xi \geq m|\xi|^2$$

for all $\xi \in \mathbb{R}^N$, $m > 0$ independent of ξ.

(ii) $\eta \in \mathrm{L}^{p/(p-2)}(\Omega)$ and $\eta \geq 0$.

(iii) $\partial\Omega$ is a Lipschitz boundary.

(iv) $g \in \mathrm{L}^2(\Omega)$.

Then problem (2.14) has a unique weak solution in $\mathrm{H}_0^1(\Omega)$.

Let $V_h \subset \mathrm{H}_0^1(\Omega)$ be a given FEM subspace. We look for the numerical solution u_h of (2.14) in V_h:

$$\int_\Omega (G\nabla u_h \cdot \nabla v + \eta u_h v) = \int_\Omega gv, \quad v \in V_h. \tag{2.15}$$

The corresponding linear algebraic system has the form

$$(\mathbf{G}_h + \mathbf{D}_h)\mathbf{c} = \mathbf{g}_h,$$

where $\mathbf{G}_h$ and $\mathbf{D}_h$ are the corresponding stiffness and mass matrices, respectively. We apply the matrix $\mathbf{G}_h$ as preconditioner, thus the preconditioned form of (2.15) is given by

$$(\mathbf{I}_h + \mathbf{G}_h^{-1}\mathbf{D}_h)\mathbf{c} = \tilde{\mathbf{g}}_h \tag{2.16}$$

with $\tilde{\mathbf{g}}_h = \mathbf{G}_h^{-1}\mathbf{g}_h$. Then we apply the CGM to (2.16) and the auxiliary systems with $\mathbf{G}_h$ can be solved efficiently with fast solvers.

Theorem 2.3. *Let Assumptions 2.2.1 hold. Then there exists $\boldsymbol{C} > 0$ such that for all $k \in \mathbb{N}$*

$$\left(\frac{\|e_k\|_{\mathbf{A}_h}}{\|e_0\|_{\mathbf{A}_h}}\right)^{\frac{1}{k}} \leq \boldsymbol{C}k^{-\alpha}, \tag{2.17}$$

where $\alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}$.

Proof. Let us consider the Hilbert space $\mathrm{L}^2(\Omega)$ endowed with the usual inner product. Let $D = \{u \in \mathrm{H}_0^1(\Omega) \cap \mathrm{H}^2(\Omega) \,/\, G\nabla u \in \mathrm{H}^1(\Omega)^N\}$. We define the operators

$$Su \equiv -\operatorname{div}(G\nabla u), \quad u \in D \quad \text{and} \quad Qu \equiv \eta u, \quad u \in \mathrm{H}_0^1(\Omega) \tag{2.18}$$

and since $p < 2^* = \frac{2N}{N-2}$, the embedding $\mathcal{I} : \mathrm{H}_0^1(\Omega) \to \mathrm{L}^p(\Omega)$ is compact, see Example 1.9, in particular, there exists $C_p > 0$ such that for all $u \in \mathrm{H}_0^1(\Omega)$

$$\|u\|_{\mathrm{L}^p(\Omega)} \leq C_p \|u\|_{\mathrm{H}_0^1(\Omega)}. \tag{2.19}$$

Then

$$\langle Su, u\rangle \geq m \int_\Omega |\nabla u|^2 \geq m\nu \int_\Omega u^2, \qquad u \in D,$$

where ν is the Sobolev constant. Hence, the energy space H_S is a well-defined Hilbert space with $\langle u, v\rangle_S = \int_\Omega G\nabla u \cdot \nabla v$. It is easy to see that $H_S = \mathrm{H}_0^1(\Omega)$ and that the following inequality

$$\sqrt{m}\|u\|_{\mathrm{H}_0^1(\Omega)} \leq \|u\|_{H_S} \tag{2.20}$$

holds for all $u \in H_S$. Furthermore,

$$\begin{aligned}
\|Q_S v\|_{H_S} = \sup_{\|u\|_{H_S}=1} |\langle Q_S v, u\rangle_S| &= \sup_{\|u\|_{H_S}=1} \langle Qv, u\rangle \\
&= \sup_{\|u\|_{H_S}=1} \int_\Omega \eta v u \\
&\leq \sup_{\|u\|_{H_S}=1} \left(\int_\Omega |\eta|^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}} \left(\int_\Omega |v|^p\right)^{\frac{1}{p}} \left(\int_\Omega |u|^p\right)^{\frac{1}{p}} \\
&\leq C_p \sup_{\|u\|_{H_S}=1} \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)} \|v\|_{\mathrm{L}^p(\Omega)} \|u\|_{\mathrm{H}_0^1(\Omega)} \\
&\leq \frac{C_p M_\eta}{\sqrt{m}} \|v\|_{\mathrm{L}^p(\Omega)} \sup_{\|u\|_{H_S}=1} \|u\|_{H_S} \\
&= C_Q \|v\|_{\mathrm{L}^p(\Omega)},
\end{aligned} \tag{2.21}$$

where $M_\eta = \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)}$ and $C_Q = \frac{C_p M_\eta}{\sqrt{m}}$. Here we applied the extension of Hölder's inequality ([7, Th. 4.6]) with

$$1 = \frac{1}{p} + \frac{1}{p} + \left(\frac{p-2}{p}\right).$$

Hence Q_S is compact and self-adjoint in H_S.

Let $\lambda_n = \lambda_n(Q_S)$. Since Q_S is a compact self-adjoint operator in H_S, by [12, Ch.6, Th.1.5] we have the following characterization of the eigenvalues of Q_S:

$$\forall n \in \mathbb{N}: \quad \lambda_n(Q_S) = \min\{\|Q_S - L_{n-1}\| \,/\, L_{n-1} \in \mathcal{L}(H_S), \mathrm{rank}(L_{n-1}) \leq n-1\}. \tag{2.22}$$

By taking the minimum over a smaller subset of finite rank operators, we obtain

$$\lambda_n(Q_S) \leq \min\{\|Q_S - Q_S L_{n-1}\| \,/\, L_{n-1} \in \mathcal{L}(H_S), \mathrm{rank}(L_{n-1}) \leq n-1\}. \tag{2.23}$$

Now, by (2.21) and (2.20) we get

$$\begin{aligned}\|Q_S - Q_S L_{n-1}\| &= \sup_{u\in H_S} \frac{\|(Q_S - Q_S L_{n-1})u\|_{H_S}}{\|u\|_{H_S}} \\ &= \sup_{u\in H_S} \frac{\|Q_S(u - L_{n-1}u)\|_{H_S}}{\|u\|_{H_S}} \\ &\le C_Q \sup_{u\in H_S} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)}}{\|u\|_{H_S}} \\ &\le \frac{C_Q}{\sqrt{m}} \sup_{u\in \mathrm{H}_0^1(\Omega)} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)}}{\|u\|_{\mathrm{H}_0^1(\Omega)}}.\end{aligned}$$

This, together with (2.23) yields

$$\begin{aligned}\lambda_n(Q_S) &\le \frac{C_Q}{\sqrt{m}} \min\{\|\mathcal{I} - L_{n-1}\| \,/\, L_{n-1} \in \mathcal{L}(\mathrm{H}_0^1(\Omega), \mathrm{L}^p(\Omega)), \mathrm{rank}(L_{n-1}) \le n-1\} \\ &:= \frac{C_Q}{\sqrt{m}} a_n(\mathcal{I}),\end{aligned} \tag{2.24}$$

where $a_n(\mathcal{I})$ denotes the approximation numbers of the compact embedding $\mathcal{I}\colon \mathrm{H}_0^1(\Omega) \mapsto \mathrm{L}^p(\Omega)$, [18]. Furthermore, we have the estimation [9]

$$a_n(\mathcal{I}) \le C_{app} n^{-\alpha}, \quad \alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p},$$

for some constant $C_{app} > 0$. Therefore, we arrive at the inequality

$$\lambda_n(Q_S) \le \frac{C_{app} C_Q}{\sqrt{m}} n^{-\alpha}.$$

Now, taking the arithmetic mean on both sides and estimating the sum from above by an integral we obtain

$$\frac{1}{k}\sum_{n=1}^{k} \lambda_n(Q_S) \le \frac{C_{app} C_Q}{\sqrt{m}} \frac{1}{k} \left(1 + \int_1^k \frac{1}{x^\alpha}\right) \le \frac{C_{app} C_Q}{\sqrt{m}(1-\alpha)} \frac{1}{k^\alpha}. \tag{2.25}$$

Then, by (2.13), we conclude. □

Remark 2.4. *The auxiliary problem $\mathbf{S}w_k = \mathbf{Q}p_k$ for the PCGM can be solved easily with fast solvers due to the special structure of S, [15], [8].*

2.2.2 Symmetric elliptic systems

In this section, we prove that the previous results can be extended to certain systems of elliptic PDEs. For simplicity and also due to practical occurrence, we only include

Laplacian principal parts. Nonetheless, the results remain similar when the principal parts have the form (2.18).

First, let us consider systems of the form

$$\begin{cases} -\Delta u_i + \eta_{i1}u_1 + \dots \eta_{is}u_s = g_i, \\ u_i|_{\partial\Omega} = 0, \quad (i = 1, \dots, s), \end{cases} \tag{2.26}$$

where $\boldsymbol{H} = \{\eta_{ij}\}_{i,j=1}^{s}$ is a symmetric positive semidefinite variable coefficient matrix such that

$$\forall i, j \in \{1, \dots, s\}: \quad \eta_{ij} \in \mathrm{L}^{p/(p-2)}(\Omega).$$

We work with the space $\mathrm{L}^p(\Omega)^s$ with the norm

$$\|u\|_{\mathrm{L}^p(\Omega)^s} = \left(\sum_{j=1}^{s} \|u_j\|^2_{\mathrm{L}^p(\Omega)} \right)^{1/2}, \quad u = (u_1, \dots, u_s) \in \mathrm{L}^p(\Omega)^s.$$

Let $H = \mathrm{L}^2(\Omega)^s$. Let $u = (u_1 \dots u_s) \in D = (\mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega))^s$, we define the operators

$$Su = \begin{pmatrix} -\Delta u_1 \\ . \\ . \\ . \\ -\Delta u_s \end{pmatrix}, \qquad Qu = \boldsymbol{H}u, \qquad u \in \mathrm{H}^1_0(\Omega)^s. \tag{2.27}$$

Clearly, S is a uniformly positive symmetric operator in H. In fact, by Poincare's inequality

$$\langle Su, u \rangle \geq \frac{1}{\nu^2} \sum_{i=1}^{s} \|u_i\|^2_{\mathrm{L}^2(\Omega)} = \frac{1}{\nu^2} \|u\|^2_H, \tag{2.28}$$

where ν is the Sobolev constant. Then, the energy space H_S is well defined with

$$\langle u, v \rangle_S = \sum_{i=1}^{s} \int_\Omega \nabla u_i \nabla v_i, \quad \|u\|^2_{H_S} = \sum_{i=1}^{s} \int_\Omega |\nabla u_i|^2$$

and so $H_S = \mathrm{H}^1_0(\Omega)^s$. Furthermore, by (2.19) we have that

$$\|u\|^2_{H_S} \geq \frac{1}{C_p^2} \sum_{i=1}^{s} \|u_i\|^2_{\mathrm{L}^p(\Omega)} = \frac{1}{C_p^2} \|u\|^2_{\mathrm{L}^p(\Omega)^s}. \tag{2.29}$$

Then there exists a unique operator $Q_S \colon \mathrm{H}^1_0(\Omega)^s \to \mathrm{L}^2(\Omega)^s$ such that

$$\langle Q_S u, v \rangle_S = \int_\Omega \sum_{i,j=1}^{s} \eta_{ij} u_j v_i. \tag{2.30}$$

It is easy to see that Q_S is self-adjoint in H_S. Analogous to (2.21), by (2.29), (2.28) and Hölder's inequality we get

$$\begin{aligned}
\|Q_S v\|_{H_S} &= \sup_{\|u\|_S=1} |\langle Q_S v, u\rangle_S| \\
&\le \sup_{\|u\|_{H_S}=1} \sum_{i,j=1}^{s} \int_\Omega |\eta_{ij}||v_j||u_i| \\
&\le \sup_{\|u\|_{H_S}=1} \sum_{i,j=1}^{s} \|\eta_{ij}\|_{\mathrm{L}^{p/(p-2)}(\Omega)} \|v_j\|_{\mathrm{L}^p(\Omega)} \|u_i\|_{\mathrm{L}^p(\Omega)} \\
&\le M_\eta \sup_{\|u\|_{H_S}=1} \sum_{j=1}^{s} \|v_j\|_{\mathrm{L}^p(\Omega)} \sum_{i=1}^{s} \|u_i\|_{\mathrm{L}^p(\Omega)} \\
&\le M_\eta \sup_{\|u\|_{H_S}=1} \sqrt{s} \left(\sum_{j=1}^{s} \|v_j\|^2_{\mathrm{L}^p(\Omega)} \right)^{1/2} \sqrt{s} \left(\sum_{i=1}^{s} \|u_i\|^2_{\mathrm{L}^p(\Omega)} \right)^{1/2} \\
&= s M_\eta \sup_{\|u\|_{H_S}=1} \|v\|_{\mathrm{L}^p(\Omega)^s} \|u\|_{\mathrm{L}^p(\Omega)^s} \\
&\le s M_\eta C_p \|v\|_{\mathrm{L}^p(\Omega)^s} \\
&= C_Q \|v\|_{\mathrm{L}^p(\Omega)^s}.
\end{aligned} \tag{2.31}$$

where $M_\eta = \max_{i,j} \|\eta_{ij}\|_{\mathrm{L}^{p/(p-2)}(\Omega)}$ and $C_Q = s M_\eta C_p$. Hence, we have proved that Q_S is a compact self-adjoint operator in H_S. Then, the characterization (2.22) of the eigenvalues of Q_S holds. The rest of the proof follows by modifying the scalar case. In this case, we take the minimum over a smaller subset of finite rank operators to obtain

$$\lambda_n(Q_S) \le \min\{\|Q_S - Q_S L_{n-1}\| \ / \ L_{n-1} \in \mathcal{L}_{\mathrm{diag}}(H_S), \mathrm{rank}(L_{n-1}) \le n-1\},$$

with $L_{n-1} \in \mathcal{L}_{\mathrm{diag}}(H_S)$ if and only if

$$L_{n-1} u = \begin{pmatrix} L^s_{n-1} u_1 \\ . \\ . \\ . \\ L^s_{n-1} u_s \end{pmatrix}, \text{ such that } L^s_{n-1} \in \mathcal{L}(\mathrm{H}^1_0(\Omega)) \text{ and } \mathrm{rank}(L^s_{n-1}) \le \left[\frac{n-1}{s} \right],$$

where $[\cdot]$ denotes the lower integer part. Furthermore, we shall use the approximation numbers

$$a_{\left[\frac{n-1}{s}\right]} = \min \left\{ \|I - T_{n-1}\| \ / \ T_{n-1} \in \mathcal{L}(\mathrm{H}^1_0(\Omega), \mathrm{L}^p(\Omega)), \mathrm{rank}(T_{n-1}) \le \left[\frac{n-1}{s} \right] \right\}.$$

Note that if $n \le s$, then we can use $\lambda_n(Q_S) \le \|Q_S\|$, and for $n \ge s+1$ the above numbers are estimated by

$$a_{\left[\frac{n-1}{s}\right]} \le C_{app} \left[\frac{n-1}{s} \right]^{-\alpha}, \tag{2.32}$$

with $\alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}$. Then

$$\begin{aligned}
\|Q_S - Q_S L_{n-1}\| &= \sup_{u \in H_S} \frac{\|(Q_S - Q_S L_{n-1})u\|_{H_S}}{\|u\|_{H_S}} \\
&= \sup_{u \in H_S} \frac{\|Q_S(u - L_{n-1}u)\|_{H_S}}{\|u\|_{H_S}} \\
&\le C_Q \sup_{u \in H_S} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)^s}}{\|u\|_{H_S}} \\
&= C_Q \sup_{u \in H_S} \frac{\left(\sum_{j=1}^s \|u_i - L_{n-1}^s u_i\|^2_{\mathrm{L}^p(\Omega)}\right)^{1/2}}{\left(\sum_{j=1}^s \|u_i\|^2_{\mathrm{H}_0^1(\Omega)}\right)^{1/2}} \\
&\le C_Q \sup_{u \in H_S} \frac{\left(\|I - L_{n-1}^s\|^2_{\mathcal{L}(\mathrm{H}_0^1(\Omega),\mathrm{L}^p(\Omega))} \sum_{j=1}^s \|u_i\|^2_{\mathrm{H}_0^1(\Omega)}\right)^{1/2}}{\left(\sum_{j=1}^s \|u_i\|^2_{\mathrm{H}_0^1(\Omega)}\right)^{1/2}} \\
&= C_Q \|I - L_{n-1}^s\|_{\mathcal{L}\left(\mathrm{H}_0^1(\Omega),\mathrm{L}^p(\Omega)\right)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\lambda_n(Q_S) &\le C_Q \min\left\{ \|I - L_{n-1}^s\|_{\mathcal{L}\left(\mathrm{H}_0^1(\Omega),\mathrm{L}^p(\Omega)\right)} \;/\; L_{n-1}^s \in \mathcal{L}(\mathrm{H}_0^1(\Omega),\mathrm{L}^p(\Omega)), \mathrm{rank}(L_{n-1}^s) \le \left[\frac{n-1}{s}\right] \right\} \\
&= C_Q a_{\left[\frac{n-1}{s}\right]}.
\end{aligned}$$

Hence, by (2.32) we obtain the estimation

$$\lambda_n(Q_S) \le C_Q C_{app} \left[\frac{n-1}{s}\right]^{-\alpha}, \quad n \ge s+1. \tag{2.33}$$

$$\lambda_n(Q_S) \le \|Q_S\| \le C_Q \quad n \le s. \tag{2.34}$$

Note that

$$0.5 \le \frac{[x]}{x} \le 1, \quad \forall x > 1.$$

Thus, for $n \ge s+1$

$$\begin{aligned}
\left[\frac{n-1}{s}\right]^{-\alpha} &\le \frac{1}{0.5^\alpha} \frac{s^\alpha}{(n-1)^\alpha} \\
&= (2s)^\alpha \left(\frac{n^\alpha}{(n-1)^\alpha}\right) \frac{1}{n^\alpha} \\
&\le 2^\alpha (s+1)^\alpha \frac{1}{n^\alpha}.
\end{aligned}$$

Hence, (2.33) becomes

$$\lambda_n(Q_S) \leq C_Q C_{app} 2^\alpha (s+1)^\alpha \frac{1}{n^\alpha} := R_{s,\alpha} C_Q \frac{1}{n^\alpha}$$

with $R_{s,\alpha} = C_{app} 2^\alpha (s+1)^\alpha$. Then, by taking arithmetic meaning on both sides and splitting the sum we get

$$\begin{aligned}
\frac{1}{k}\sum_{n=1}^{k} \lambda_n(Q_S) &\leq \frac{1}{k}\left(s\|Q_S\| + \sum_{n=s+1}^{k} \lambda_n(Q_S) \right) \\
&\leq \frac{1}{k}\left(s\|Q_S\| + R_{s,\alpha} C_Q \sum_{n=s+1}^{k} \frac{1}{n^\alpha} \right) \\
&\leq \frac{1}{k}\left(s\|Q_S\| + R_{s,\alpha} C_Q \int_s^k \frac{1}{x^\alpha} \right) \\
&\leq \frac{s}{k}\|Q_S\| + \frac{R_{s,\alpha} C_Q}{1-\alpha}\frac{1}{k^\alpha} \\
&\leq C\frac{1}{k^\alpha},
\end{aligned} \tag{2.35}$$

where $C = \max\{s\|Q_S\|, R_{s,\alpha} C_Q (1-\alpha)^{-1}\}$. Finally, by Corollary 2.2, we have proved there exists $\boldsymbol{C} > 0$ such that for all $k \in \mathbb{N}$

$$\left(\frac{\|e_k\|_{\mathbf{A}_h}}{\|e_0\|_{\mathbf{A}_h}} \right)^{\frac{1}{k}} \leq \boldsymbol{C} k^{-\frac{1}{\alpha}}, \tag{2.36}$$

where $\boldsymbol{C} = C(s, p, \alpha, \mathbf{H})$.

2.2.3 Extension to non-symmetric systems

Consider the iterative solutions of the non-symmetric linear problem $\mathbf{B}u = \mathbf{g}$. Assume that $\mathbf{B} + \mathbf{B}^* > 0$, where $\mathbf{B}^*$ denotes the adjoint of $\mathbf{B}$ with respect to the inner product. Further, consider the decomposition $\mathbf{B} = \mathbf{I} + \mathbf{E}$. We apply the *generalized minimal residual (GMRES) method* to solve this system. This method is an extension of the CG method to non-symmetric systems, [17]. For simplicity, we show only the algorithm of the *Generalized Conjugate Residual (CGR) method*, since the latter is mathematically equivalent to GMRES, see e.g [16]. Let u_0 be arbitrary, $r_0 = \mathbf{S}g - \mathbf{B}u_0$, $p_0 = r_0$ and for $k \in \mathbb{N}$

$$\begin{cases}
x_{k+1} = x_k + \alpha_k p_k \\
r_{k+1} = r_k - \alpha_k \mathbf{B} p_k \\
p_{k+1} = r_{k+1} + \sum_{i=0}^{k} \beta_{ik} p_i
\end{cases}$$

with

$$\alpha_k = \frac{\langle r_k, \mathbf{B}p_k\rangle}{\langle \mathbf{B}p_k.\mathbf{B}p_k\rangle}, \quad \beta_{ik} = -\frac{\langle \mathbf{B}r_{k+1}, \mathbf{B}p_i\rangle}{\langle \mathbf{B}p_i.\mathbf{B}p_i\rangle} \quad (i = 0, 1, \dots, k).$$

Remark 2.5. *The main reason one does not use CGR directly is the amount of storage requirement and the number of steps CGR needs for convergence compared to GMRES. From the equations above we realize that we require to double the memory used in the algorithm compared to GMREs, since both the set of p_i's and $\mathbf{B}p_i'$s need to be saved. Moreover, the number of arithmetic operations per step is also around 50% higher than GMRES, [16]. Hence, the implementation of the method CGR is simpler but less efficient than GMRES.*

An advantage of GMRES is that superlinear convergence is not lost. In fact, if $\mathbf{B} \in B(H)$ and H is a Hilbert space, then we get the estimate

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \frac{\|\mathbf{B}^{-1}\|}{k} \sum_{j=1}^{k} s_j(\mathbf{E}) \quad (k = 1, 2, \dots)$$

where $(s_j(\mathbf{E}))_j$ are the singular values of $\mathbf{E}$.

Let us now go back to (2.26) for $\boldsymbol{H} = \{\eta_{i,j}\}_{i,j=1}^{s}$ non-symmetric. We apply GMRES to the corresponding discretized system. By [5], we have an analogue of Corollary 1 when A is non-Hermitian. In this case, the GMRES method provides superlinear convergence estimates for the residuals r_k, and (2.11) is replaced by the more general case (2.12). Altogether, we have

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \frac{\|B^{-1}\|_S}{k} \sum_{j=1}^{k} s_j(Q_S), \quad \forall k = 1, 2, \dots, n. \tag{2.37}$$

To show that Theorem 2.3 still holds in this case, we follow the same steps as we did previously. We define the operators S, Q, Q_S as before, (2.27), (2.30). Here, Q_S is no longer self-adjoint and its eigenvalues do not coincide with its singular values. Nonetheless, by [12, Ch.6, Th.1.5] we have the following characterization of the singular values of Q_S:

$$\forall n \in \mathbb{N}: \quad s_n(Q_S) = \min\{\|Q_S - L_{n-1}\| \ / \ L_{n-1} \in \mathcal{L}(H_S), \operatorname{rank}(L_{n-1}) \leq n-1\}. \tag{2.38}$$

Then, similarly to the proof for symmetric systems, we can show that

$$\frac{1}{k} \sum_{n=1}^{k} s_n(Q_S) \leq C \frac{1}{k^\alpha}, \quad \alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}, \tag{2.39}$$

where $C > 0$ is defined as in (2.35). Therefore, by (2.37), we obtain that there exists $\boldsymbol{C} > 0$ such that

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \boldsymbol{C} \frac{1}{k^\alpha}, \tag{2.40}$$

where $C = C(s, p, \alpha, \mathbf{H})$. Finally, note that $r_k = \mathbf{A}_h e_k$. Then $\|e_k\|_{\mathbf{A}_h} \leq \|\mathbf{A}_h^{-1}\|\|r_k\|$ and $\|r_0\| \leq \|\mathbf{A}_h\|\|e_0\|_{\mathbf{A}_h}$. Hence

$$\left(\frac{\|e_k\|_{\mathbf{A}_h}}{\|e_0\|_{\mathbf{A}_h}}\right)^{1/k} \leq C\frac{1}{k^\alpha}\mathrm{cond}(A)^{1/k} \leq C\frac{1}{k^\alpha}.$$

where $\mathrm{cond}(A) = \|A\|\|A^{-1}\| < 1$ denotes the conditioning number of A.

Remark 2.6. *For elliptic symmetric systems, the auxiliary problem $\mathbf{S}w_k = \mathbf{Q}p_k$ for the PCGM becomes*

$$\begin{cases} -\Delta(w_k)_1 &= \sum_{j=1}^s \eta_{1j}(p_k)_j, \\ -\Delta(w_k)_2 &= \sum_{j=1}^s \eta_{2j}(p_k)_j, \\ & \cdot \\ & \cdot \\ & \cdot \\ -\Delta(w_k)_s &= \sum_{j=1}^s \eta_{sj}(p_k)_j, \\ (w_i)|_{\partial\Omega} &= 0, \quad \forall i = 1, \ldots, s. \end{cases}$$

Note that these equations are independent of one another. Hence, they can be solved in parallel. Furthermore, in practice, these types of systems can be large, e.g in [19], long-range transport of air pollution models are described by a system of PDEs with $s = 30$. That is, $\mathbf{S}$ is considerably simpler than $\mathbf{B}$.

2.3 A numerical example

Let us solve the following PDEs numerically

$$\begin{cases} -\Delta u + \eta u = f_i, & \text{in } \Omega = [0,1]^2, \\ u|_{\partial\Omega} = 0 \end{cases} \tag{E_i}$$

with $i = 1, 2$. Here $\eta \in \mathrm{L}^{\frac{p}{p-2}}(\Omega)$ is defined as

$$\eta(x, y) = (x^2 + y^2)^{-\beta}, \qquad 0 < \beta < \frac{p-2}{p}$$

and

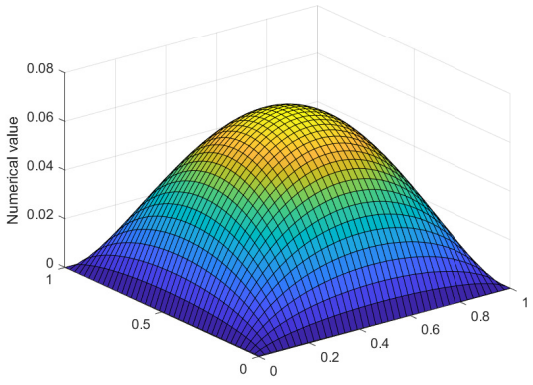
$$f_1(x, y) = 1,$$
$$f_2(x, y) = 1 - x - y,$$

$$f_3(x,y) = \frac{1}{10}(1+x+y).$$

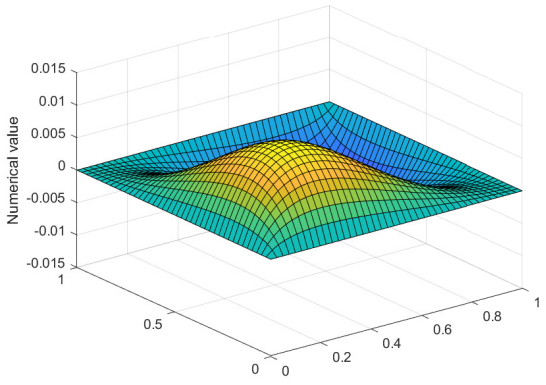
Applying FEM with Courant elements to (E_i) with stepsize $h = 1/(N+1)$ we obtain the algebraic system

$$(\mathbf{G}_h + \mathbf{D}_h)\mathbf{c}_i = \mathbf{g}_h^i, \quad i = 1,2. \tag{E_i'}$$

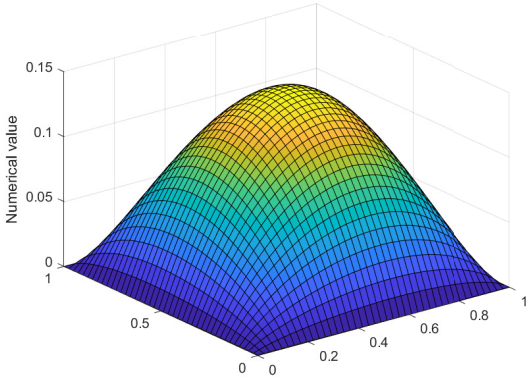
Then, we apply $\mathbf{G}_h$ as a preconditioner and we solve the preconditioned system using the CGM.



(a) Numerical solution of E_1



(b) Numerical solution of E_2



(c) Numerical solution of E_3

Figure 2.1: Graphs of the numerical solutions of (E_i) with $i = 1,2,3$, $N = 40$ and $\beta = 1/4$.

To measure the error of the PCGM, we use the energy norm

$$\|e\|_{\mathbf{A}_h} = \langle \mathbf{A}_h e, e\rangle^{\frac{1}{2}} \quad e \in \mathbb{R}^N,$$

where $\mathbf{A}_h = \mathbf{G}_h + \mathbf{D}_h$. Table 1 shows the residual error obtained at each iteration k of the method applied to (E_i') for $i = 1,2,3$ respectively. We see that it takes 7 steps to reach a $\mathcal{O}(10^{-14})$ residual error.

To test Theorem 2.3, note that $d = 2$ and so $\alpha = \frac{1}{p}$. Furthermore, recall that

$$\eta \in \mathrm{L}^{\frac{p}{p-2}}(\Omega) \quad \text{if } \beta < \frac{p-2}{p} = 1 - 2\alpha.$$

That is, if $p > \frac{2}{1-\beta}$, we get that the theorem holds when $\alpha < \frac{1-\beta}{2}$. Table 2 shows the values of

$$\hat{\delta}_k = \left(\frac{\|r_k\|_{\mathbf{G}_h}}{\|r_0\|_{\mathbf{G}_h}} \right)^{\frac{1}{k}} k^{\alpha}$$

for each problem (E_i') and for different choices of β (and hence of α) while fixing a mesh size. The value of $\hat{\delta}_k^i$ corresponds to the system (E_i'). This demonstrates that (2.17) holds in these cases since the values of $\hat{\delta}_k$ are bounded by a constant.

Finally, Table 3 shows the values of $\hat{\delta}_k$ for different mesh sizes while fixing the values of β. Here we verify that the results of Theorem 2.3 are not sensitive to the size of the mesh.

	$\|r_k^1\|_{\mathbf{G}_h}$	$\|r_k^2\|_{\mathbf{G}_h}$	$\|r_k^3\|_{\mathbf{G}_h}$
1	0.1872869890826060000000	0.0438591951304650000000	0.0377132992611370000000
2	0.0021778212752603100000	0.0003744621215674900000	0.0005135054728180300000
3	0.0000134272507943374000	0.0000089983886838118200	0.0000041947047929654800
4	0.0000001224317125796750	0.0000000614690100912966	0.0000000303294910379450
5	0.0000000044417617185916	0.0000000030075987138478	0.0000000001113030628806
6	0.0000000000021058757996	0.0000000000011266031196	0.0000000000005252511482
7	0.0000000000000082093367	0.0000000000000035087716	0.0000000000000020265486
8	0.0000000000000003001190	0.0000000000000000110731	0.0000000000000000071527
9	0.0000000000000000000816	0.0000000000000000000864	0.0000000000000000000238
10	0.0000000000000000000006	0.0000000000000000000005	0.0000000000000000000002

Table 2.1: Norm of residual error r_k^i at each iteration of PCGM applied to system (E_i'). Here $N = 40$ and $\beta = 1/4$.

	$\beta = 2/3, \alpha = 0.15$			$\beta = 3/4, \alpha = 0.12$			$\beta = 1/4, \alpha = 0.374$			$\beta = 1/2, \alpha = 0.24$		
	$\hat{\delta}_k^1$	$\hat{\delta}_k^2$	$\hat{\delta}_k^3$	$\hat{\delta}_k^1$	$\hat{\delta}_k^2$	$\hat{\delta}_k^3$	$\hat{\delta}_k^1$	$\hat{\delta}_k^2$	$\hat{\delta}_k^3$	$\hat{\delta}_k^1$	$\hat{\delta}_k^2$	$\hat{\delta}_k^3$
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1786	0.1904	0.1887	0.1921	0.2098	0.2006	0.1397	0.1197	0.1512	0.1592	0.1593	0.1715
3	0.0925	0.1087	0.0973	0.1027	0.1127	0.1070	0.0627	0.0889	0.0725	0.0790	0.1026	0.0853
4	0.0621	0.0744	0.0650	0.0702	0.0792	0.0731	0.0478	0.0578	0.0503	0.0537	0.0670	0.0562
5	0.0476	0.0548	0.0492	0.0542	0.0611	0.0559	0.0344	0.0427	0.0359	0.0406	0.0476	0.0422
6	0.0372	0.0434	0.0386	0.0432	0.0490	0.0447	0.0293	0.0336	0.0303	0.0324	0.0376	0.0331
7	0.0313	0.0352	0.0321	0.0362	0.0406	0.0370	0.0256	0.0279	0.0263	0.0260	0.0304	0.0265
8	0.0264	0.0297	0.0269	0.0300	0.0340	0.0309	0.0231	0.0244	0.0236	0.0225	0.0254	0.0223
9	0.0227	0.0253	0.0229	0.0261	0.0287	0.0268	0.0207	0.0245	0.0216	0.0205	0.0225	0.0195
10	0.0203	0.0221	0.0202	0.0232	0.0250	0.0236	0.0213	0.0242	0.0224	0.0191	0.0208	0.0189

Table 2.2: Values of $\hat{\delta}_k$ for different α's and β's, with a fixed mesh size. Here $N = 40$.

	$\hat{\delta}_k^1$			$\hat{\delta}_k^2$			$\hat{\delta}_k^3$		
	$N = 20$	$N = 40$	$N = 80$	$N = 20$	$N = 40$	$N = 80$	$N = 20$	$N = 40$	$N = 80$
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1910	0.1921	0.1924	0.2080	0.2098	0.2103	0.1997	0.2006	0.2009
3	0.1013	0.1027	0.1031	0.1123	0.1127	0.1128	0.1057	0.1070	0.1073
4	0.0683	0.0702	0.0707	0.0785	0.0792	0.0794	0.0713	0.0731	0.0736
5	0.0519	0.0542	0.0549	0.0594	0.0611	0.0616	0.0536	0.0559	0.0566
6	0.0403	0.0432	0.0443	0.0466	0.0490	0.0499	0.0418	0.0447	0.0457
7	0.0333	0.0362	0.0373	0.0375	0.0406	0.0418	0.0341	0.0370	0.0382
8	0.0274	0.0300	0.0316	0.0310	0.0340	0.0353	0.0279	0.0309	0.0325
9	0.0234	0.0260	0.0279	0.0279	0.0287	0.0305	0.0236	0.0268	0.0284
10	0.0223	0.0237	0.0245	0.0245	0.0250	0.0268	0.0227	0.0236	0.0248

Table 2.3: Values of $\hat{\delta}_k$ for different mesh sizes with $\beta = 3/4, \alpha = 0.12$.

Chapter 3

Non-linear elliptic systems

In this chapter, we consider inner-outer iterations of the *Damped Inexact Newton (DIN) method* applied to the finite element discretization of some non-linear systems of PDEs. Here, we use GMRES to solve the linearized problem that arises from the DIN process at each step. Thus, mesh-independent superlinear convergence of the inner iterations can be proved. This chapter includes a detailed version of the results found in [1], where the DIN plus CGN technique is used. Our main goal is to demonstrate that by modifying the method used for the inner iterations and applying the results from the previous chapter, we can give more explicit estimates for the superlinear convergence rate.

3.1 The problem

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with $d = 2, 3$. We consider the nonlinear elliptic transport system of the form

$$\begin{cases} -\operatorname{div}(K_i \nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + f_i(x, u_1, \ldots, u_l) & = g_i; \\ u_i|_{\partial\Omega} & = 0, \end{cases} \tag{3.1}$$

where $i = 1, \ldots, l$. To ease the notation, this can be written as

$$\begin{cases} -\operatorname{div}(\mathbf{K}\nabla \mathbf{u}) + \mathbf{b} \cdot \nabla \mathbf{u} + f(x, \mathbf{u}) & = \mathbf{g}; \\ \mathbf{u}|_{\partial\Omega} & = \mathbf{0}. \end{cases}$$

Furthermore, we follow the same assumptions as in [1]:

Assumption 3.1

(i) *Smoothness conditions*: $K_i \in \mathrm{L}^\infty(\Omega)$, $\mathbf{b}_i \in \mathrm{C}^1(\overline{\Omega})^d$ and $g_i \in \mathrm{L}^2(\Omega)$ with $i = 1, \dots, l$). Moreover, the function $f = (f_1, \dots, f_l) : \Omega \times \mathbb{R}^l \mapsto \mathbb{R}^l$ is measurable and bounded w.r.t the variable $x \in \Omega$ and C^1 in the variable $\xi \in \mathbb{R}^l$.

(ii) *Coercivity condition*: there is $m > 0$ such that $K_i \geq m$ for all $i = 1, \dots, l$. Furthermore,

$$\forall (x, \xi) \in \Omega \times \mathbb{R}^l, \forall \eta \in \mathbb{R}^l : \quad f'_\xi(x, \xi)\eta \cdot \eta - \frac{1}{2}\left(\max_i \mathrm{div}\mathbf{b}_i(x)\right)|\eta|^2 \geq 0, \tag{3.2}$$

where $f'_\xi(x, \xi) = \frac{\partial f(x,\xi)}{\partial \xi}$.

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

$$\|f'_\xi(x, \xi_1) - f'_\xi(x, \xi_2)\| \leq \left(c_1 + c_2 \max\{|\xi_1|, |\xi_2|\}^{p-3}\right)|\xi_1 - \xi_2|. \tag{3.3}$$

Remark 3.1. *Assumption (iii) implies the estimates*

$$\|f'_\xi(x, \xi)\| \leq c_3 + c_4|\xi|^{p-2}, \quad |f(x, \xi)| \leq c_5 + c_6|\xi|^{p-1} \tag{3.4}$$

for any $(x, \xi) \in \Omega \times \mathbb{R}^l$. *Furthermore, we get*

$$|(f'_\xi(x, \xi_1) - f'_\xi(x, \xi_2))\eta \cdot \zeta| \leq (c_1 + c_2(\max\{|\xi_1|, |\xi_2|\})^{p-3})|\xi_1 - \xi_2||\eta||\zeta| \tag{3.5}$$

for any $(x, \xi_1), (x, \xi_2) \in \Omega \times \mathbb{R}^l$ *and* $\eta, \zeta \in \mathbb{R}^l$.

Remark 3.2. *We recall Theorem 1.4 and Example 1.9. Denote* $p^* = \infty$ *if* $d = 2$ *or* $p^* = 6$ *if* $d = 3$. *Then for all* $p \leq p^*$ *the following Sobolev embedding holds:*

$$\mathrm{H}^1_0(\Omega) \subset \mathrm{L}^p(\Omega).$$

In addition, we have the estimate

$$\|v\|_{\mathrm{L}^p(\Omega)} \leq C_p\|v\|_{\mathrm{H}^1_0(\Omega)} \quad (v \in \mathrm{H}^1_0(\Omega)). \tag{3.6}$$

3.2 Well-posedness of the elliptic problem

We show that there exists a unique weak solution to Problem (3.1). For this, we denote

$$\begin{aligned} \langle F(\mathbf{u}), \mathbf{v}\rangle_{\mathrm{H}^1_0(\Omega)} &= \int_\Omega \sum_{i=1}^{l} K_i \nabla u_i \cdot \nabla v_i + (\mathbf{b}_i \cdot \nabla u_i)v_i + f_i(x, \mathbf{u})v_i) \\ &\equiv \int_\Omega (\mathbf{K}\nabla\mathbf{u} \cdot \nabla\mathbf{v} + (\mathbf{b} \cdot \nabla\mathbf{u}) \cdot \mathbf{v} + f(x, \mathbf{u}) \cdot \mathbf{v}) \end{aligned} \tag{3.7}$$

for any $\mathbf{u}, \mathbf{v} \in \mathrm{H}_0^1(\Omega)^l$. We claim that this defines an operator $F : \mathrm{H}_0^1(\Omega)^l \to \mathrm{H}_0^1(\Omega)^l$. Indeed, for any $\mathbf{u} \in \mathrm{H}_0^1(\Omega)^l$ we define the linear functional $\psi_{\mathbf{u}}$ by

$$\psi_{\mathbf{u}}(\mathbf{v}) = \int_\Omega (\mathbf{K}\nabla\mathbf{u} \cdot \nabla\mathbf{v} + (\mathbf{b} \cdot \nabla\mathbf{u}) \cdot \mathbf{v} + f(x, \mathbf{u}) \cdot \mathbf{v}), \quad (\mathbf{v} \in \mathrm{H}_0^1(\Omega)^l).$$

Then, Hölder inequality with $\frac{p-1}{p} + \frac{1}{p} = 1$ and Remarks 3.1, 3.2 yield

$$\begin{aligned}
|\psi_{\mathbf{u}}(\mathbf{v})| &\le \sum_{i=1}^{l} \int_\Omega \|K_i\|_{\mathrm{L}^\infty(\Omega)}(|\nabla u_i||\nabla v_i| + \|\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}|\nabla u_i||v_i| + |f_i(x,\mathbf{u})|v_i|) \\
&\le \max\left\{\max_i \|K_i\|_{\mathrm{L}^\infty(\Omega)}, \max_i \|\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}, 1\right\} \sum_{i=1}^{l} (\|u_i\|_{\mathrm{H}_0^1(\Omega)}\|v_i\|_{\mathrm{H}_0^1(\Omega)} + \|u_i\|_{\mathrm{H}_0^1(\Omega)}\|v_i\|_{\mathrm{L}^2(\Omega)} + \dots \\
&\cdots + \int_\Omega (c_5 + c_6|\mathbf{u}|^{p-1})|v_i|)) \\
&\le \mathrm{const} \cdot \sum_{i=1}^{l} (\|u_i\|_{\mathrm{H}_0^1(\Omega)}\|v_i\|_{\mathrm{H}_0^1(\Omega)} + C_2\|u_i\|_{\mathrm{H}_0^1(\Omega)}\|v_i\|_{\mathrm{H}_0^1(\Omega)} + c_5C_1\|v_i\|_{\mathrm{H}_0^1(\Omega)} + \dots \\
&\cdots + c_6\|\mathbf{u}\|_{\mathrm{L}^p(\Omega)}^{p-1}\|v_i\|_{\mathrm{L}^p(\Omega)}) \\
&\le \mathrm{const} \cdot \sum_{i=1}^{l} (\|u_i\|_{\mathrm{H}_0^1(\Omega)}\|v_i\|_{\mathrm{H}_0^1(\Omega)}) + c_5C_1 \sum_{i=1}^{l} \|v_i\|_{\mathrm{H}_0^1(\Omega)} + c_6C_p^p\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-1} \sum_{i=1}^{l} \|v_i\|_{\mathrm{H}_0^1(\Omega)} \\
&\le \mathrm{const} \cdot \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} + \mathrm{const}\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} + \mathrm{const} \cdot \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-1}\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \\
&\le \mathrm{const}\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}.
\end{aligned}$$

This shows that $\psi_{\mathbf{u}} : \mathrm{H}_0^1(\Omega)^l \to \mathbb{R}$ is a bounded linear functional. Hence, by the Riesz representation theorem, there exists a unique $F(\mathbf{u})$ such that

$$\psi_{\mathbf{u}}(\mathbf{v}) = \langle F(\mathbf{u}), \mathbf{v}\rangle_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{v} \in \mathrm{H}_0^1(\Omega)).$$

Similarly, we can show that there exists a unique $\overline{\mathbf{g}} \in \mathrm{H}_0^1(\Omega)^l$ such that

$$\int_\Omega \mathbf{g}\mathbf{v} = \langle \overline{\mathbf{g}}, v\rangle_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{v} \in \mathrm{H}_0^1(\Omega)).$$

Thus, proving the well-posedness of (3.1) is equivalent to showing that the equation $F(\mathbf{u}) = \overline{\mathbf{g}}$ has a unique solution in $\mathrm{H}_0^1(\Omega)^l$. The following proposition shows that the coercivity and continuity properties of f'_ξ are inherited by F'. The proof given here of this statement is a detailed version of the one found in [1].

Proposition 3.3. *The operator $F : \mathrm{H}_0^1(\Omega)^l \to \mathrm{H}_0^1(\Omega)^l$ is Gateaux differentiable and satisfies*

$$\langle F'(\mathbf{u})\mathbf{h}, \mathbf{h}\rangle_{\mathrm{H}_0^1(\Omega)} \ge m\|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)}^2 \quad (\mathbf{u}, \mathbf{h} \in \mathrm{H}_0^1(\Omega)^l). \tag{3.8}$$

Furthermore, F' is locally Lipschitz continuous. *That is,*

$$\|F'(\mathbf{u}) - F'(\mathbf{v})\| \le L(r)\|\mathbf{u} - \mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \tag{3.9}$$

for all $\mathbf{u}, \mathbf{v} \in \mathrm{H}_0^1(\Omega)^l$ *with* $\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)} \le r$, $\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \le r$, *where*

$$L(r) = c_1 C^3 + c_2 C_p^p r^{p-3} \quad (r > 0). \tag{3.10}$$

Proof. First, we show that F is Gateaux differentiable. Formally,

$$\begin{aligned}
\langle \partial_{\mathbf{h}} F(\mathbf{u}), \mathbf{v} \rangle_{\mathrm{H}_0^1(\Omega)} &= \int_\Omega \sum_{i=1}^{l} \lim_{t\to 0} \frac{1}{t} (k_i \nabla(u_i + t h_i) \cdot \nabla v_i + \mathbf{b}_i \cdot \nabla(u_i + t h_i) v_i + f_i(x, \mathbf{u} + t\mathbf{h}) v_i \dots \\
&\quad \dots - (k_i \nabla u_i \cdot \nabla v_i + (\mathbf{b}_i \cdot \nabla u_i) v_i + f_i(x, \mathbf{u}) v_i)) \\
&= \int_\Omega \sum_{i=1}^{l} (k_i \nabla h_i \cdot \nabla v_i + (\mathbf{b}_i \cdot \nabla v_i) v_i) + \lim_{t\to 0} \frac{1}{t} (f_i(x, \mathbf{u} + t\mathbf{h}) - f_i(x, \mathbf{u})) v_i \\
&\equiv \int_\Omega \mathbf{K} \nabla \mathbf{h} \cdot \nabla \mathbf{v} + (\mathbf{b} \cdot \nabla \mathbf{h}) \cdot \mathbf{v} + f_\xi'(x, \mathbf{u}) \mathbf{h} \cdot \mathbf{v} \\
&:= \langle D(\mathbf{u}, \mathbf{h}), \mathbf{v} \rangle_{\mathrm{H}_0^1(\Omega)}.
\end{aligned}$$

Using the Riesz representation theorem, we can prove that $D(\mathbf{u}, \mathbf{h}) \in \mathrm{H}_0^1(\Omega)^l$ for any $\mathbf{u}, \mathbf{v} \in \mathrm{H}_0^1(\Omega)^l$. Indeed,

$$\mathbf{v} \mapsto \int_\Omega \mathbf{K} \nabla \mathbf{h} \cdot \nabla \mathbf{v} + (\mathbf{b} \cdot \nabla \mathbf{h}) \cdot \mathbf{v} + f_\xi'(x, \mathbf{u}) \mathbf{h} \cdot \mathbf{v}$$

is a linear bounded functional:

$$\begin{aligned}
\left| \int_\Omega \dots \right| &\le \|\mathbf{K}\|_{\mathrm{L}^\infty(\Omega)} \int_\Omega |\nabla \mathbf{h}| |\nabla \mathbf{v}| + \|\mathbf{b}\|_{\max} \int_\Omega |\nabla \mathbf{h}| |\mathbf{v}| + \int_\Omega (c_3 + c_4 |\mathbf{u}|^{p-2}) |\mathbf{h}| |\mathbf{v}| \\
&\le \|\mathbf{K}\|_{\mathrm{L}^\infty(\Omega)} \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{b}\|_{\max} C_2 \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} + \dots \\
&\quad \dots + c_3 C_2^2 \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} + c_4 \|\mathbf{u}\|_{\mathrm{L}^p(\Omega)}^{p-2} \|\mathbf{v}\|_{\mathrm{L}^p(\Omega)} \|\mathbf{h}\|_{\mathrm{L}^p(\Omega)} \\
&\le \mathrm{const} \cdot \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} + c_4 C_p^p \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-2} \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \\
&\le \mathrm{const} \cdot \|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)}.
\end{aligned} \tag{3.11}$$

Here we use Remarks 3.1, 3.2 and Hölder inequality with $\frac{p-2}{p} + \frac{1}{p} + \frac{1}{p} = 1$. Hence, $D(\mathbf{u}, \mathbf{h}) \in \mathrm{H}_0^1(\Omega)^l$ is the Riesz representative of this functional. Since Gateaux differentiability is known for any bounded linear operator, see Remark 1.12, to prove Gateaux differentiability of F is enough to show that the nonlinear part of $\frac{1}{t}(F(\mathbf{u} + t\mathbf{h}) - F(\mathbf{u})) - D(\mathbf{u}, \mathbf{h})$ tends to 0 in $\mathrm{H}_0^1(\Omega)$, i.e.,

$$\lim_{t\to 0} \sup_{\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}=1} \int_\Omega \left(\frac{1}{t} (f(x, \mathbf{u} + t\mathbf{h}) - f(x, \mathbf{u})) \cdot \mathbf{v} - f_\xi'(x, \mathbf{u}) \mathbf{h} \cdot \mathbf{v} \right) = 0. \tag{3.12}$$

Using *Lagrange mean value theorem* we get $f(x,\mathbf{u}+t\mathbf{h})-f(x,\mathbf{u})=f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})t\mathbf{h}$, for some $\theta=\theta(x,t,\mathbf{u},\mathbf{h})\in[0,1]$. Then, the l.h.s limit above is equal to

$$\begin{aligned}
&\lim_{t\to 0}\sup_{\|\mathbf{v}\|_{\mathrm{H}^1_0(\Omega)}=1}\int_\Omega\Big((f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})\mathbf{h})\cdot\mathbf{v}-f'_\xi(x,\mathbf{u})\mathbf{h}\cdot\mathbf{v}\Big)\\
&\le\lim_{t\to 0}\sup_{\|\mathbf{v}\|_{\mathrm{H}^1_0(\Omega)}=1}\int_\Omega|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}\cdot\mathbf{v}|\\
&\le\lim_{t\to 0}\sup_{\|\mathbf{v}\|_{\mathrm{H}^1_0(\Omega)}=1}\|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}\|_{\mathrm{L}^{\frac{p}{p-1}}(\Omega)}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\\
&\le\lim_{t\to 0}\left(\int_\Omega|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}
\end{aligned}\tag{3.13}$$

Furthermore, since $f\in\mathrm{C}^1$ in the variable $\xi\in\mathbb{R}^l$,

$$\lim_{t\to 0}|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}|^{\frac{p}{p-1}}=0\quad\text{pointwise a.e.}\tag{3.14}$$

Therefore, it remains to prove that $|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}|^{\frac{p}{p-1}}$ is bounded by a function in $\mathrm{L}^1(\Omega)^s$ and we can apply *Lebesgue's dominated convergence theorem* to show that (3.13) tends to 0 as t tends to 0. In fact, we may assume $t\le 1$ and by (3.3) we obtain

$$\begin{aligned}
|(f'_\xi(x,\mathbf{u}+\theta t\mathbf{h})-f'_\xi(x,\mathbf{u}))\mathbf{h}|^{\frac{p}{p-1}}&\le(c_1+c_2\max\{|\mathbf{u}+\theta t\mathbf{h}|,|\mathbf{u}|\}^{p-3})^{\frac{p}{p-1}}(|t\theta\mathbf{h}||\mathbf{h}|)^{\frac{p}{p-1}}\\
&\le\text{const}\cdot(c_1+c_2\max\{|\mathbf{u}+\mathbf{h}|,|\mathbf{u}|\}^{\frac{p(p-3)}{p-1}})|\mathbf{h}|^{\frac{2p}{p-1}}\\
&\le\text{const}\cdot|\mathbf{h}|^{\frac{2p}{p-1}}+\text{const}\cdot(|\mathbf{u}|+|\mathbf{h}|)^{\frac{p(p-3)}{p-1}}(|\mathbf{u}|+|\mathbf{h}|)^{\frac{2p}{p-1}}\\
&\le\text{const}\cdot|\mathbf{h}|^{\frac{2p}{p-1}}+\text{const}\cdot(|\mathbf{u}|+|\mathbf{h}|)^p.
\end{aligned}\tag{3.15}$$

Notice that the first term in the RHS of (3.15) is in $\mathrm{L}^1(\Omega)$ since, by Hölder inequality with $\frac{2}{p-1}+\frac{p-3}{p-1}=1$ and the Sobolev embedding $\mathrm{H}^1_0(\Omega)\subset\mathrm{L}^p(\Omega)$ we get

$$\begin{aligned}
\int_\Omega|\mathbf{h}|^{\frac{2p}{p-1}}&\le\|\mathbf{h}\|^{\frac{2p}{p-1}}_{\mathrm{L}^p(\Omega)}|\Omega|^{\frac{p-3}{p-1}}\\
&\le C_p^{\frac{2p}{p-1}}|\Omega|^{\frac{p-3}{p-1}}\|\mathbf{h}\|^{\frac{2p}{p-1}}_{\mathrm{H}^1_0(\Omega)}<\infty.
\end{aligned}$$

Further,

$$\int_\Omega(|\mathbf{u}|+|\mathbf{h}|)^p\le C_p^p\||\mathbf{u}|+|\mathbf{h}|\|^p_{\mathrm{H}^1_0(\Omega)}<\infty.$$

Then, by (3.13), (3.14), and (3.15) we have proved (3.12). Hence, F is Gateaux differentiable with

$$\langle F'(\mathbf{u})\mathbf{h},\mathbf{v}\rangle_{\mathrm{H}_0^1(\Omega)} = \int_\Omega \mathbf{K}\nabla\mathbf{h}\cdot\nabla\mathbf{v} + (\mathbf{b}\cdot\nabla\mathbf{h})\cdot\mathbf{v} + f'_\xi(x,\mathbf{u})\mathbf{h}\cdot\mathbf{v}. \tag{3.16}$$

Notice that, by the divergence theorem,

$$\begin{aligned}\int_\Omega (\mathbf{b}_i\cdot\nabla h_i)h_i &= \int_\Omega \operatorname{div}(\mathbf{b}_i h_i^2) - \int_\Omega h_i(\mathbf{b}_i\cdot\nabla h_i) - \int_\Omega \operatorname{div}(\mathbf{b}_i)h_i^2\\ &= \int_{\partial\Omega}(\mathbf{b}_i h_i^2)\cdot\nu - \int_\Omega h_i(\mathbf{b}_i\cdot\nabla h_i) - \int_\Omega \operatorname{div}(\mathbf{b}_i)h_i^2\\ &= -\int_\Omega h_i(\mathbf{b}_i\cdot\nabla h_i) - \int_\Omega \operatorname{div}(\mathbf{b}_i)h_i^2.\end{aligned}$$

Hence, (3.16) with $\mathbf{v}=\mathbf{h}$ becomes

$$\langle F'(\mathbf{u})\mathbf{h},\mathbf{h}\rangle_{\mathrm{H}_0^1(\Omega)} = \int_\Omega \left(\mathbf{K}|\nabla\mathbf{h}|^2 + f'_\xi(x,\mathbf{u})\mathbf{h}\cdot\mathbf{h} - \frac{1}{2}\sum_{i=1}^{l}\operatorname{div}(\mathbf{b}_i)h_i^2\right). \tag{3.17}$$

The above and assumption (ii) imply

$$\begin{aligned}\langle F'(\mathbf{u})\mathbf{h},\mathbf{h}\rangle_{\mathrm{H}_0^1(\Omega)} &\geq \int_\Omega \left(m|\nabla\mathbf{h}|^2 + \frac{1}{2}\max_i \operatorname{div}(\mathbf{b}_i(x))|\mathbf{h}|^2 - \frac{1}{2}\max_i \operatorname{div}(\mathbf{b}_i(x))\sum_{i=1}^{l} h_i^2\right)\\ &= m\int_\Omega |\nabla\mathbf{h}|^2.\end{aligned}$$

This proves (3.8).

Let us show F' is locally Lipschitz continuous. By (3.5) in Remark 3.1 we have that, for any $\mathbf{u},\mathbf{v},\mathbf{h},\mathbf{z}\in\mathrm{H}_0^1(\Omega)^l$:

$$\begin{aligned}|\langle (F'(\mathbf{u})-F'(\mathbf{v}))\mathbf{h},\mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}| &= \left|\int_\Omega (f'_\xi(x,\mathbf{u})\mathbf{h}\cdot\mathbf{z} - f'_\xi(x,\mathbf{v})\mathbf{h}\cdot\mathbf{z})\right|\\ &\leq \int_\Omega (c_1 + c_2(\max\{|\mathbf{u}|,|\mathbf{v}|\})^{p-3})|\mathbf{u}-\mathbf{v}||\mathbf{h}||\mathbf{z}|\\ &\leq c_1\|\mathbf{u}-\mathbf{v}\|_{\mathrm{L}^3(\Omega)}\|\mathbf{h}\|_{\mathrm{L}^3(\Omega)}\|\mathbf{z}\|_{\mathrm{L}^3(\Omega)} + \dots\\ &\quad\cdots + c_2(\max\{\|\mathbf{u}\|_{\mathrm{L}^p(\Omega)},\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\})^{p-3}\|\mathbf{u}-\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{h}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{z}\|_{\mathrm{L}^p(\Omega)}\\ &\leq c_1C_3^3\|\mathbf{u}-\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}\|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)}\|\mathbf{z}\|_{\mathrm{H}_0^1(\Omega)} + \dots\\ &\quad\cdots + c_2C_p^p(\max\{\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)},\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}\})^{p-3}\|\mathbf{u}-\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}\|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)}\|\mathbf{z}\|_{\mathrm{H}_0^1(\Omega)}.\end{aligned} \tag{3.18}$$

Here we used the Sobolev embeddings $\mathrm{H}_0^1(\Omega)\subset\mathrm{L}^3(\Omega)$, $\mathrm{H}_0^1(\Omega)\subset\mathrm{L}^p(\Omega)$, and Hölder inequality twice with $\frac{1}{3}+\frac{1}{3}+\frac{1}{3}=1$ and $\frac{p-3}{p}+\frac{1}{p}+\frac{1}{p}+\frac{1}{p}=1$, respectively. Now, (3.18) yields

$$\begin{aligned}\|F'(\mathbf{u})-F'(\mathbf{v})\| &= \sup_{\|h\|_{\mathrm{H}_0^1(\Omega)}=\|z\|_{\mathrm{H}_0^1(\Omega)}=1} |\langle (F'(\mathbf{u})-F'(\mathbf{v}))\mathbf{h},\mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}|\\ &\leq \left(c_1C_3^3 + c_2C_p^p(\max\{\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)},\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}\})^{p-3}\right)\|\mathbf{u}-\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)}.\end{aligned} \tag{3.19}$$

By assuming $\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)} \leq r$, $\|\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \leq r$ and denoting $L(r) = c_1 C_3^3 + c_2 C_p^p r^{p-3}$ we get (3.9). □

Proposition 3.4. *There exists a unique solution in* $\mathrm{H}_0^1(\Omega)^l$ *for the equation* $F(\mathbf{u}) = \overline{\mathbf{g}}$.

Proof. We claim that $|\langle F'(\mathbf{u})\mathbf{h}, \mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}| \leq R(\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}) \cdot \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{z}\|_{\mathrm{H}_0^1(\Omega)}$. Indeed, by (3.18) with $\mathbf{v} \equiv 0$:

$$\left|\langle F'(\mathbf{u})\mathbf{h}, \mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}\right| \leq \mathrm{const} \cdot \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{z}\|_{\mathrm{H}_0^1(\Omega)} (\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-2} + \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}) + \left|\langle F'(\mathbf{0})\mathbf{h}, \mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}\right|.$$

Furthermore, using (3.11) we deduce that

$$|\langle F'(\mathbf{u})\mathbf{h}, \mathbf{z}\rangle_{\mathrm{H}_0^1(\Omega)}| \leq \mathrm{const} \cdot \|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \|\mathbf{z}\|_{\mathrm{H}_0^1(\Omega)} (\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-2} + \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)} + 1). \tag{3.20}$$

Thus, by denoting $R(\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}) = \mathrm{const} \cdot (\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^{p-2} + \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)} + 1)$, we proved our claim. Furthermore, since $F'(\mathbf{u})$ is coercive, by Theorem 1.14 and Remark 1.15, we conclude. □

3.3 FEM discretization and Newton iteration

Let $V_h \in \mathrm{H}_0^1(\Omega)^l$ be a N-dimensional subspace and $\psi_1, \ldots, \psi_N \in V_h$ be a basis. We look for $\mathbf{u}_h = \sum_{j=1}^N c_j \psi_j \in V_h$ such that

$$\langle F(\mathbf{u}_h), \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} = \langle \mathbf{f}, \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{v}_h \in V_h).$$

Denote $F_h : V_h \to V_h$ the operator given by

$$\langle F_h(\mathbf{u}_h), \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} = \langle F(\mathbf{u}_h), \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{v}_h \in V_h),$$

and $\mathbf{g}_h \in V_h$ by

$$\langle \mathbf{g}_h, \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} = \langle \mathbf{f}, \mathbf{v}_h\rangle_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{v}_h \in V_h).$$

This allows us to rewrite the problem as

$$F_h(\mathbf{u}_h) = \mathbf{g}_h \quad \text{in } V_h. \tag{3.21}$$

Let $\mathscr{A} : \mathbb{R}^N \to \mathbb{R}^N$ such that $\mathscr{A} = (\mathscr{A}_1, \ldots, \mathscr{A}_N)$ and $\mathscr{A}_i(\mathbf{c}) = \langle F_h(\mathbf{u}_h), \psi_i\rangle_{\mathrm{H}_0^1(\Omega)}$. Then $\mathbf{c} \in \mathbb{R}^N$ is the unique solution of the nonlinear algebraic system:

$$\mathscr{A}(\mathbf{c}) = \mathscr{G}, \tag{3.22}$$

where $\mathscr{G} = (\langle \mathbf{g}_h, \psi_1\rangle_{\mathrm{H}_0^1(\Omega)}, \ldots, \langle \mathbf{g}_h, \psi_N\rangle_{\mathrm{H}_0^1(\Omega)})$. To solve this system numerically, we apply the *Damped inexact Newton method (DIN)*.

Construction of the DIN iteration

Let $\mathbf{u}_0 \in V_h$ arbitrary. The DIN iteration defines a sequence $(\mathbf{u}_n) \subset V_h$ constructed recursively as

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau_n \mathbf{p}_n \quad (n \in \mathbb{N}),$$

where $\mathbf{p}_n \in V_h$ is the approximate solution of the linear auxiliary problem

$$\langle F_h'(\mathbf{u}_n)\mathbf{p}_n, \mathbf{v}_h \rangle_{\mathrm{H}_0^1(\Omega)} = -\langle F_h(\mathbf{u}_n) - \mathbf{g}_h, \mathbf{v}_h \rangle_{\mathrm{H}_0^1(\Omega)} \tag{3.23}$$

in the sense that

$$\|F_h'(\mathbf{u}_n)\mathbf{p}_n + (F_h(\mathbf{u}_n) - \mathbf{g}_h)\|_{\mathrm{H}_0^1(\Omega)} \leq \delta_n \|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} \quad 0 < \delta_n \leq \delta_0 < 1$$

and

$$\tau_n = \min\left\{1, \frac{1-\delta_n}{(1+\delta_n)^2} \frac{m^2}{L(R_0)\|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)}}\right\}.$$

Here $R_0 = \frac{2}{m}\|F_h(\mathbf{u}_0) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{u}_0\|_{\mathrm{H}_0^1(\Omega)}$ and $L(R_0)$ is defined as in (3.10).

The convergence of the method is given by the following theorem

Theorem 3.5. *Let Assumptions (i)-(iii) hold. Then*

$$\|\mathbf{u}_n - \mathbf{u}_h\|_{\mathrm{H}_0^1(\Omega)} \leq \frac{1}{m}\|F_h(\mathbf{u}_h) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} \to 0 \text{ monotonically.}$$

In particular, if

$$\delta_n \leq \text{const} \cdot \|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)}^{\gamma} \text{ with some } 0 < \gamma \leq 1,$$

then the convergence is local of order $1+\gamma$. That is, the convergence is linear for n_0 steps until $\|F_h(\mathbf{u}_h) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} \leq \epsilon$, where $\epsilon \leq (1-\delta_0)\frac{m^2}{2L(R_0)}$, and further on (as $\tau_n \equiv 1$)

$$\|\mathbf{u}_n - \mathbf{u}_h\|_{\mathrm{H}_0^1(\Omega)} \leq d_1 q^{(1+\gamma)^{n-n_0}}$$

for some $d_1 > 0$ and $0 < q < 1$.

Proof. First, we notice that for any $n \in \mathbb{N}$, $\|\mathbf{u}_n\|_{\mathrm{H}_0^1(\Omega)}$ satisfies the following a priori estimate:

$$\begin{aligned}
\|\mathbf{u}_n\|_{\mathrm{H}_0^1(\Omega)} &\leq \|\mathbf{u}_n - \mathbf{u}_h\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{u}_h - \mathbf{u}_0\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{u}_0\|_{\mathrm{H}_0^1(\Omega)} \\
&\leq \frac{2}{m}\|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{u}_0\|_{\mathrm{H}_0^1(\Omega)} \\
&\leq \frac{2}{m}\|F_h(\mathbf{u}_0) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)} + \|\mathbf{u}_0\|_{\mathrm{H}_0^1(\Omega)} \\
&= R_0,
\end{aligned} \tag{3.24}$$

where in the last step we used the fact that the sequence $(\|F_h(\mathbf{u}_n) - \mathbf{g}_h\|_{\mathrm{H}_0^1(\Omega)})_{n\in\mathbb{N}}$ is decreasing. Further, F_h is Gateaux differentiable and satisfies

$$\|F_h'(\mathbf{u})\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)} \geq m\|\mathbf{h}\|_{\mathrm{H}_0^1(\Omega)},$$

$$\|F_h'(\mathbf{u}) - F_h'(\mathbf{v})\| \leq L(r)\|\mathbf{u}-\mathbf{v}\|_{\mathrm{H}_0^1(\Omega)} \quad (\mathbf{u},\mathbf{v}\in B(0,r)\subset V_h).$$

Indeed, these properties are inherited from F', which were proven in Proposition 3.3. Then, by (3.24) F_h' has Lipschitz constant $L(r) = L(R_0)$ on the ball $B(0,R_0)$. Finally, we conclude by [11, Theorem 5.12, Remark 5.17]. □

3.4 Solution of the linearized problems: inner GMRES iterations

We can realize the previous method as an inner-outer iteration method, where we use inner iterations to numerically solve the linearized problem (3.23). We proceed as follows. Let $\mathbf{u}_n$ be constructed in the DIN iteration and consider the linearized problem (3.23), written by

$$F_h'(\mathbf{u}_n)\mathbf{p}_h = \mathbf{r}_h, \tag{3.25}$$

where $\mathbf{r}_h = \mathbf{g}_h - F_h(\mathbf{u}_n)$. This is equivalent to the FEM solution in V_h of the linear elliptic problem

$$\begin{cases} -\mathrm{div}(K_i\nabla p_i) + \mathbf{b}_i\cdot\nabla p_i + \sum_{j=1}^{l}\partial_j f_i(x,\mathbf{u}_n)p_j = r_i \\ p_i|_{\partial\Omega} = 0 \end{cases} \quad (i=1,\dots,l), \tag{3.26}$$

where $r_i = g_i + \mathrm{div}(K_i\nabla u_{n,i}) - \mathbf{b}_i\cdot\nabla u_{n,i} - f_i(x,\mathbf{u}_n)$. Denote by $\mathbf{L}_h^{(n)}$ the stiffness matrix corresponding to (3.26). We look for the solution to the system

$$\mathbf{L}_h^{(n)}\mathbf{c} = \mathbf{d}, \tag{3.27}$$

where $\mathbf{c}$ and $\mathbf{d}$ are the coefficient vectors of $\mathbf{p}_h$ and $\mathbf{r}_h$, respectively. To solve we system, we use a PCG-type algorithm. The preconditioner is constructed as follows: for any $u_i|_{\partial\Omega} = 0$ let

$$S_iu_i = -\mathrm{div}(K_i\nabla u_i) + h_iu_i \quad (i=1,\dots,l),$$

where $h_i\in\mathrm{L}^\infty(\Omega)$ and $h_i\geq 0$, then we define the equivalent operator

$$S\mathbf{u} = (S_1u_1,\dots S_lu_l). \tag{3.28}$$

Denote by $\mathbf{S}_h$ the stiffness matrix of S in the same FEM subspace V_h. Then the preconditioned system of (3.27) is given by

$$\mathbf{S}_h^{-1}\mathbf{L}_h^{(n)}\mathbf{c} = \mathbf{f} := \mathbf{S}_h^{-1}\mathbf{d}. \tag{3.29}$$

This results in the solution of non-symmetric auxiliary linear elliptic systems of the form

$$-\mathrm{div}(K_i\nabla u_i) + h_i u_i = f_i \quad (i = 1, \dots, l).$$

Such systems were studied in the previous chapter. Further, we already recognized that $\mathbf{S}_h$ is an efficient preconditioner, see Remark 2.4 and Remark 2.6.

Remark 3.6. *During the construction of the preconditioner S, we required $h_i \in \mathrm{L}^\infty(\Omega)$, but $h_i \in \mathrm{L}^{\frac{p}{p-2}}(\Omega)$ is enough. However, this could diminish the practicality of S as a preconditioner of system 3.27.*

Our goal is to solve (3.29) by applying a suitable CG-type iteration that preserves the superlinear convergence of the outer iteration.

3.4.1 Convergence analysis of GMRES for preconditioned non-symmetric linear problems

In the previous chapter, we briefly introduced the GMRES method and proved an estimation for the rate of superlinear convergence of the method when applied to nonlinear elliptic systems. In this section, we study the use of this method for solving the FEM discretization of the following Dirichlet problem

$$\begin{cases} L_i u_i \equiv -\mathrm{div}(K_i\nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + \sum_{j=1}^{l} V_{ij} u_i = g_i \\ u_i|_{\partial\Omega} = 0 \end{cases} \quad (i = 1, \dots, l) \tag{3.30}$$

on a bounded domain $\Omega \subset \mathbb{R}^d$. This problem is well-posed on $\mathrm{H}_0^1(\Omega)^l$ under the following conditions: K_i is as in Assumption 3.1, $g_i \in \mathrm{L}^2(\Omega)$, $\mathbf{b}_i \in \mathrm{C}^1(\overline{\Omega})^d$, $V_{ij} \in \mathrm{L}^{\frac{p}{p-2}}(\Omega)$ and the matrix $V = \{V_{ij}\}_{i,j=1}^l$ together with $\mathbf{b}_i$ satisfy the coercivity property

$$\lambda_{\min}(V + V^T) - \max_i \mathrm{div}\mathbf{b}_i \geq 0 \tag{3.31}$$

pointwise on Ω, where $\lambda_{\min}$ denotes the smallest eigenvalue. As before, we choose a FEM subspace $V_h \subset \mathrm{H}_0^1(\Omega)^l$ and look for the solutions of the corresponding algebraic system $\mathbf{L}_h\mathbf{c} = \mathbf{b}$.

We shall use as preconditioner the stiffness matrix $\mathbf{S}_h$ in V_h of the operator S defined in (3.28) and the corresponding inner product on $\mathrm{H}_0^1(\Omega)^l$

$$\langle \mathbf{u}, \mathbf{v} \rangle_S = \int_\Omega \sum_{i=1}^{l} (K_i \nabla u_i \cdot \nabla v_i + h_i u_i v_i).$$

This is well-defined since S is uniformly positive and symmetric w.r.t the usual inner product on $\mathrm{H}_0^1(\Omega)^l$. Moreover, the corresponding norm is equivalent to the standard norm in $\mathrm{H}_0^1(\Omega)$ and in particular

$$\|\mathbf{u}\|_S^2 = \int_\Omega \sum_{i=1}^{l} (K_i |\nabla u_i|^2 + h_i |u|_i^2) \geq m \|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}^2. \tag{3.32}$$

Then, we apply GMRES with the $\mathbf{S}_h-$inner product to the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c} = \mathbf{S}_h^{-1}\mathbf{b}$. First, notice that we have the decomposition

$$\mathbf{S}_h^{-1}\mathbf{L}_h = \mathbf{I} + \mathbf{Q}_{\mathbf{S}h},$$

where $\mathbf{Q}_{\mathbf{S}h}$ is the corresponding Gram matrix in V_h of the operator Q_S defined implicitly as follows:

$$\begin{aligned} \langle Q_S \mathbf{u}, \mathbf{v} \rangle_S &= \sum_{i=1}^{l} \int_\Omega \left((\mathbf{b}_i \cdot \nabla u_i) v_i + \left(\sum_{j=1}^{l} V_{ij} u_j - h_i u_i \right) v_i \right) \\ &\equiv \int_\Omega ((\mathbf{b} \cdot \nabla \mathbf{u}) \cdot \mathbf{v} + (V - \mathbf{h}\mathbf{I})\mathbf{u} \cdot \mathbf{v}) \quad (\mathbf{u}, \mathbf{v} \in \mathrm{H}_0^1(\Omega)^l). \end{aligned} \tag{3.33}$$

We claim that $Q_S : \mathrm{H}_0^1(\Omega)^l \to \mathrm{H}_0^1(\Omega)^l$ is a compact operator w.r.t the $S-$inner product. Since $\mathcal{I} : \mathrm{L}^p(\Omega) \to \mathrm{H}_0^1(\Omega)$ is compact, it is enough to prove that there exists $C_Q > 0$ such that

$$\|Q_S \mathbf{v}\|_S \leq C_Q \|\mathbf{v}\|_{\mathrm{L}^p(\Omega)} \quad (\mathbf{v} \in \mathrm{H}_0^1(\Omega)^l). \tag{3.34}$$

In fact, we can divide our problem by observing that $Q_S = Q_{S,1} + Q_{S,2}$ where

$$\langle Q_{S,1}\mathbf{u}, \mathbf{v} \rangle_S = \int_\Omega ((\mathbf{b} \cdot \nabla \mathbf{u}) \cdot \mathbf{v}, \quad \langle Q_{S,2}\mathbf{u}, \mathbf{v} \rangle_S = \int_\Omega (V - \mathbf{h}\mathbf{I})\mathbf{u} \cdot \mathbf{v}) \quad (\mathbf{u}, \mathbf{v} \in \mathrm{H}_0^1(\Omega)^l).$$

Here $Q_{S,2}$ is of the form (2.30), which was studied in Section 2.2.2, and for this case we proved (3.34) with $C_Q = l \frac{C_p}{\sqrt{m}} \max_{i,j} \|V_{ij} - h_i\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)}$. Then, we only need to show (3.34) for $Q_{S,1}$. This follows from the divergence theorem, inequality (3.32), Hölder

inequality with $\frac{2}{p}+\frac{p-2}{p}=1$ and the Sobolev embedding $\mathrm{H}_0^1(\Omega)\subset \mathrm{L}^2(\Omega)$:

$$\begin{aligned}
\|Q_{S,1}\mathbf{v}\|_S &= \sup_{\|\mathbf{u}\|_S=1} |\langle Q_{S,1}\mathbf{v},\mathbf{u}\rangle_S| \\
&= \sup_{\|\mathbf{u}\|_S=1} \left|\sum_{i=1}^{l}\int_\Omega |(\mathbf{b}_i\cdot\nabla v_i)u_i\right| \\
&\le \sup_{\|\mathbf{u}\|_S=1} \sum_{i=1}^{l}\int_\Omega |-v_i(\mathbf{b}_i\cdot\nabla u_i)-\mathrm{div}\mathbf{b}_i v_i u_i| \\
&\le \sup_{\|\mathbf{u}\|_S=1} (\|\mathbf{b}\|_{\max}\|\mathbf{v}\|_{\mathrm{L}^2(\Omega)}\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)} + \max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}\|\mathbf{v}\|_{\mathrm{L}^2(\Omega)}\|\mathbf{u}\|_{\mathrm{L}^2(\Omega)}) \\
&\le \sup_{\|\mathbf{u}\|_S=1} (\frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}\|\mathbf{b}\|_{\max}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{u}\|_S + \max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}\|\mathbf{v}\|_{\mathrm{L}^2(\Omega)}\|\mathbf{u}\|_{\mathrm{L}^2(\Omega)}) \\
&\le \sup_{\|\mathbf{u}\|_S=1} (\frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}\|\mathbf{b}\|_{\max}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{u}\|_S + \max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}|\Omega|^{\frac{p-2}{2p}}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}C_2\|\mathbf{u}\|_{\mathrm{H}_0^1(\Omega)}) \\
&\le \sup_{\|\mathbf{u}\|_S=1} (\frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}\|\mathbf{b}\|_{\max}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{u}\|_S + \frac{C_2|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}\max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}\|\mathbf{u}\|_S) \\
&= \frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}(\|\mathbf{b}\|_{\max} + C_2\max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)})\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)}
\end{aligned}$$

Altogether, we proved (3.34) with

$$C_Q = \frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}(\|\mathbf{b}\|_{\max} + C_2\max_i \|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}) + \frac{C_p l}{\sqrt{m}}\max_{i,j}\|V_{ij}-h_i\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)}. \tag{3.35}$$

Hence, using the results from the previous chapter: mainly inequalities (2.37) and (2.40) in Section 2.2.3, we obtain the following superlinear convergence theorem

Theorem 3.7. *The GMRES algorithm with $\mathbf{S}_h$-inner product, applied for the $N\times N$ preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c}=\mathbf{S}_h^{-1}\mathbf{b}$, yields*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{\|B^{-1}\|_S}{k}\sum_{j=1}^{k} s_j(Q_S) \quad (k=1,2,\dots,N), \tag{3.36}$$

where $B=I+Q_S$. Furthermore, there exists $C>0$ such that

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le C\frac{1}{k^\alpha}, \quad \text{where } \alpha=\frac{1}{d}-\frac{1}{2}+\frac{1}{p}. \tag{3.37}$$

Specifically,

$$C=\max\{l\|Q_S\|, R_{l,\alpha}\cdot C_Q(1-\alpha)^{-1}\}.$$

Thus $C=C(l,\alpha,\mathbf{b},\mathbf{h},V)$. Therefore, the superlinear convergence rate estimate (3.37) *is independent of V_h and n.*

3.4.2 Uniform superlinear convergence of the inner PGMRES iteration

In this subsection, we solve the preconditioned system (3.29) using GMRES with $\mathbf{S}_h-$inner product. By Theorem 3.7, our method verifies the superlinear convergence property. However, in this case, GMRES is being applied at each step of an outer Newton iteration. Indeed, in the construction of the operator Q_S, the matrix V is now replaced by the Jacobian $f'_\xi(x,\mathbf{u}_n)$. That is, $Q_S = Q_S^{(n)}$ given by

$$\begin{aligned}\langle Q_S^{(n)}\mathbf{v},\mathbf{z}\rangle_S &= \sum_{i=1}^{l}\int_\Omega \left((\mathbf{b}_i\cdot\nabla v_i)z_i + \left(\sum_{j=1}^{l}\partial_j f_i(x,\mathbf{u}_n)v_j - h_i v_i\right)z_i\right)\\ &\equiv \int_\Omega((\mathbf{b}\cdot\nabla\mathbf{v})\cdot\mathbf{z} + (f'_\xi(x,\mathbf{u}_n)-\mathbf{hI})\mathbf{v}\cdot\mathbf{z}) \quad (\mathbf{v},\mathbf{z}\in \mathrm{H}_0^1(\Omega)^l).\end{aligned}$$

Hence, it is not clear that the estimation of the superlinear convergence rate is independent of the outer Newton iterate $\mathbf{u}_n$ or V_h since in this context the constant C in Theorem 3.7 depends on $V = f'_\xi(x,\mathbf{u}_n)$, which depends on $\mathbf{u}_n$. Nonetheless, from the previous section, we know that this can be fixed by finding a proper upper bound for $\|Q_S\|_S$.

Theorem 3.8. *The GMRES algorithm with $\mathbf{S}_h$-inner product, applied for the $N\times N$ preconditioned system* (3.29)*, yields*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le C\frac{1}{k^\alpha}, \quad \alpha = \frac{1}{d}-\frac{1}{2}+\frac{1}{p} \quad (k=1,2,\dots,N), \tag{3.38}$$

where $C>0$ is independent of V_h and $\mathbf{u}_n$.

Proof. As before, consider the decomposition $Q_S^{(n)} = Q_{S,1}^{(n)} + Q_{S,2}^{(n)}$:

$$\langle Q_{S,1}^{(n)}\mathbf{v},\mathbf{z}\rangle_S = \int_\Omega((\mathbf{b}\cdot\nabla\mathbf{v})\cdot\mathbf{z}, \quad \langle Q_{S,2}^{(n)}\mathbf{v},\mathbf{z}\rangle_S = \int_\Omega(f'_\xi(x,\mathbf{u}_n)-\mathbf{hI})\mathbf{v}\cdot\mathbf{z}) \quad (\mathbf{v},\mathbf{z}\in\mathrm{H}_0^1(\Omega)^l).$$

Then, using the same calculations from the previous subsection, we can prove that

$$\|Q_S^{(n)}\mathbf{v}\|_S \le C_Q\|\mathbf{v}\|_{\mathrm{L}^p(\Omega)} \quad (\mathbf{v}\in\mathrm{H}_0^1(\Omega)^l), \tag{3.39}$$

where

$$C_Q = \frac{|\Omega|^{\frac{p-2}{2p}}}{\sqrt{m}}(\|\mathbf{b}\|_{\max} + C_2\max_i\|\mathrm{div}\mathbf{b}_i\|_{\mathrm{L}^\infty(\Omega)}) + \frac{C_p l}{\sqrt{m}}\max_{i,j}\|\partial_j f_i(x,\mathbf{u}_n)-h_i\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)}. \tag{3.40}$$

We shall focus on the second term above. Notice that

$$\begin{aligned}\max_{i,j}\|\partial_j f_i(x,\mathbf{u}_n)\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)} &\le \left(\int_\Omega \|f'_{\tilde{\xi}}(x,\mathbf{u}_n)\|^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}}\\ &\le \left(\int_\Omega (c_3+c_4|\mathbf{u}_n|^{p-2})^{\frac{p}{p-2}}\right)^{\frac{p}{p-2}}\\ &\le \text{const}\cdot\left(c_3|\Omega|+c_4\|\mathbf{u_n}\|^p_{\mathrm{L}^p(\Omega)}\right)^{\frac{p}{p-2}}\\ &\le \text{const}\cdot\left(c_3|\Omega|+c_4C_p^p\|\mathbf{u_n}\|^p_{\mathrm{H}^1_0(\Omega)}\right)^{\frac{p}{p-2}}.\end{aligned}$$

Recall that $\mathbf{u}_n$ satisfies the a priori estimate (3.24). Then

$$\max_{i,j}\|\partial_j f_i(x,\mathbf{u}_n)-h_i\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)} \le \text{const}\cdot\left(c_3|\Omega|+c_4C_p^pR_0^p\right)^{\frac{p}{p-2}}+\max_i\|h_i\|_{\mathrm{L}^{\frac{p}{p-2}}(\Omega)}.$$

Thus, we found another constant $\overline{C}_Q$ such that (3.39) holds. Furthermore, $\overline{C}_Q$ is independent of the step size of the mesh h and the outer Newton iterate $\mathbf{u}_n$. Therefore, in this case, the constant C appearing in Theorem (3.7) can be replaced by

$$C=\text{const}\cdot\max\{l\overline{C}_Q, R_{l,\alpha}\cdot\overline{C}_Q(1-\alpha)^{-1}\},$$

which does not depend on V_h nor $\mathbf{u}_n$. □

3.5 A numerical example

Let us solve the following PDEs numerically

$$\begin{cases} -\Delta u+\eta u^3=f, & \text{in } \Omega=[0,1]^2,\\ u|_{\partial\Omega}=0, \end{cases} \tag{3.41}$$

where η and f are defined by

$$\eta(x,y)=(x^2+y^2)^{-\frac{1}{4}},\quad f(x,y)=100\quad (x,y)\in\Omega.$$

We apply FEM with Courant elements to (3.41) with stepsize $h=1/(N+1)$ to obtain a nonlinear algebraic system as in (3.22). We look for approximate solutions using the DIN plus PCGM technique. That is, given u_n from the DIN iteration, we solve the linearized problem

$$F'(u_n)p_n=-(F(u_n)-f),$$

i.e.,

$$-\Delta p_n + 3\eta u_n^2 p_n = \Delta u_n - \eta u_n^3 + f.$$

This PDE is of the form (2.14). Hence, we find an approximate solution to p_n by using the same steps as in Section 2.3. That is, we consider the system

$$(\mathbf{G}_h + \mathbf{D}_h^n)\mathbf{p}_n = \mathbf{d}_n, \tag{3.42}$$

where $\mathbf{d}_n := -\mathbf{G}_h\mathbf{u}_n + h^2(\mathbf{f} - \eta\mathbf{u}_n^3)$ and $\mathbf{D}_h^n$ changes at each step of the outer iteration. Next, we apply $\mathbf{G}_h$ as a preconditioner and solve the preconditioned system using the CGM.

We measure the residual error of the PCGM (inner process) as follows

$$\|r_k\|_{\mathbf{G}_h} = \langle \mathbf{G}_h r_k, r_k \rangle^{\frac{1}{2}} \quad r_k \in \mathbb{R}^N.$$

To demonstrate Theorem 3.8, we proceed similarly to the previous example in Section 2.3. First, we define the numbers

$$\delta_k^n = \left(\frac{\|r_k^n\|_{\mathbf{G}_h}}{\|r_0^n\|_{\mathbf{G}_h}} \right)^{\frac{1}{k}} k^\alpha$$

where k denotes the inner steps of the process and n the outer ones. Then, we perform our algorithm for different values of α and check the behavior of the sequence $(\delta_k^n)_k$ for each outer iteration. Indeed, in Tables 3.1, 3.2, and 3.3 we can observe how these sequences of numbers are uniformly bounded. Further, we verify that this bound is independent of $\mathbf{u}_n$.

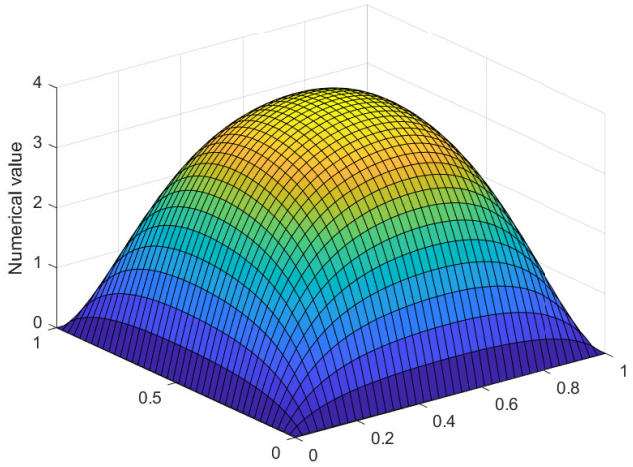


Figure 3.1: Graph of the numerical solution. Here $N = 40$.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$	$\mathbf{u}_7$	$\mathbf{u}_8$	$\mathbf{u}_9$	$\mathbf{u}_{10}$	$\mathbf{u}_{11}$	$\mathbf{u}_{12}$
δ_1^n	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
δ_2^n	0.0018	0.2629	0.4522	0.8505	0.1577	0.0136	0.0139	0.0139	0.0139	0.0139	0.0139	0.0139
δ_3^n	0.0005	0.3155	0.1811	0.2221	0.0699	0.0423	0.0492	0.0492	0.0492	0.0492	0.0492	0.0492
δ_4^n	0.0003	0.3395	0.1874	0.1736	0.0867	0.0645	0.0640	0.0640	0.0640	0.0640	0.0640	0.0640
δ_5^n	0.0002	0.3216	0.1862	0.1540	0.1115	0.0788	0.0785	0.0785	0.0785	0.0785	0.0785	0.0785
δ_6^n	0.0002	0.3403	0.2344	0.1883	0.1039	0.0859	0.0857	0.0857	0.0857	0.0857	0.0857	0.0857
δ_7^n	0.0002	0.3001	0.1857	0.1819	0.1258	0.0938	0.1009	0.1009	0.1009	0.1009	0.1009	0.1009
δ_8^n	0.0002	0.2872	0.1871	0.1764	0.1264	0.1172	0.1181	0.1181	0.1181	0.1181	0.1181	0.1181
δ_9^n	0.0002	0.3291	0.1896	0.1648	0.1335	0.1306	0.1304	0.1304	0.1304	0.1304	0.1304	0.1304
δ_{10}^n	0.0002	0.3072	0.2107	0.1849	0.1431	0.1272	0.1267	0.1267	0.1267	0.1267	0.1267	0.1267

Table 3.1: Values of δ_k^n at each step of the inner iterations. Here $N = 40$ and $\alpha = 0.1$.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$	$\mathbf{u}_7$	$\mathbf{u}_8$	$\mathbf{u}_9$	$\mathbf{u}_{10}$	$\mathbf{u}_{11}$	$\mathbf{u}_{12}$
δ_1^n	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
δ_2^n	0.0019	0.2817	0.4846	0.9115	0.1690	0.0146	0.0149	0.0149	0.0149	0.0149	0.0149	0.0149
δ_3^n	0.0006	0.3521	0.2022	0.2479	0.0780	0.0472	0.0549	0.0549	0.0549	0.0549	0.0549	0.0549
δ_4^n	0.0003	0.3900	0.2153	0.1994	0.0996	0.0741	0.0736	0.0736	0.0736	0.0736	0.0736	0.0736
δ_5^n	0.0003	0.3778	0.2187	0.1809	0.1310	0.0926	0.0923	0.0923	0.0923	0.0923	0.0923	0.0923
δ_6^n	0.0002	0.4071	0.2804	0.2252	0.1243	0.1027	0.1026	0.1026	0.1026	0.1026	0.1026	0.1026
δ_7^n	0.0002	0.3646	0.2256	0.2209	0.1528	0.1139	0.1226	0.1226	0.1226	0.1226	0.1226	0.1226
δ_8^n	0.0002	0.3536	0.2304	0.2172	0.1556	0.1443	0.1454	0.1454	0.1454	0.1454	0.1454	0.1454
δ_9^n	0.0002	0.4099	0.2362	0.2053	0.1663	0.1627	0.1624	0.1624	0.1624	0.1624	0.1624	0.1624
δ_{10}^n	0.0002	0.3868	0.2653	0.2328	0.1801	0.1601	0.1595	0.1595	0.1595	0.1595	0.1595	0.1595

Table 3.2: Values of δ_k^n at each step of the inner iterations. Here $N = 40$ and $\alpha = 0.2$.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$	$\mathbf{u}_7$	$\mathbf{u}_8$	$\mathbf{u}_9$	$\mathbf{u}_{10}$	$\mathbf{u}_{11}$	$\mathbf{u}_{12}$
δ_1^n	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
δ_2^n	0.0020	0.3020	0.5194	0.9769	0.1811	0.0157	0.0159	0.0159	0.0159	0.0159	0.0159	0.0159
δ_3^n	0.0007	0.3930	0.2257	0.2766	0.0871	0.0527	0.0613	0.0613	0.0613	0.0613	0.0613	0.0613
δ_4^n	0.0004	0.4480	0.2473	0.2290	0.1144	0.0851	0.0845	0.0845	0.0845	0.0845	0.0845	0.0845
δ_5^n	0.0003	0.4437	0.2569	0.2125	0.1539	0.1088	0.1084	0.1084	0.1084	0.1084	0.1084	0.1084
δ_6^n	0.0003	0.4870	0.3355	0.2694	0.1486	0.1229	0.1227	0.1227	0.1227	0.1227	0.1227	0.1227
δ_7^n	0.0003	0.4429	0.2741	0.2684	0.1856	0.1384	0.1490	0.1489	0.1489	0.1489	0.1489	0.1489
δ_8^n	0.0003	0.4353	0.2837	0.2674	0.1916	0.1777	0.1790	0.1790	0.1790	0.1790	0.1790	0.1790
δ_9^n	0.0003	0.5106	0.2943	0.2557	0.2072	0.2026	0.2023	0.2023	0.2023	0.2023	0.2023	0.2023
δ_{10}^n	0.0003	0.4869	0.3339	0.2930	0.2268	0.2015	0.2008	0.2008	0.2008	0.2008	0.2008	0.2008

Table 3.3: Values of δ_k^n at each step of the inner iterations. Here $N = 40$ and $\alpha = 0.3$.

Chapter 4

Conclusions

We have considered different kinds of preconditioned second-order elliptic systems and their finite element discretizations. First, we have obtained robust estimations of the rate of superlinear convergence of the PCGM and GMRES applied to linear systems. Under appropriate conditions, we have proved the following estimation when PCGM is applied to single equations or symmetric systems:

$$\left(\frac{\|e_k\|_{\mathbf{A}_h}}{\|e_0\|_{\mathbf{A}_h}}\right)^{\frac{1}{k}} \leq Ck^{-\alpha}, \quad \alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}.$$

Moreover, we have shown that this also holds for the non-symmetric case when replacing PCGM with GMRES. This extends previous results of [13] to the case of unbounded reaction coefficients in some Lebesgue spaces. Additionally, we have tested this result for the single equation case and verified our results. In fact, in Tables 2.2 and 2.3 we have observed the desired behavior of boundedness and mesh independence.

Finally, we have analyzed inner-outer iterations of the *Damped Inexact Newton (DIN) method* applied to the finite element discretization of some non-linear systems of PDEs. We have obtained more explicit estimates for the rate of superlinear convergence than the ones showed by [1], where the DIN plus CGN technique is used. We achieve this by replacing the CGN method with GMRES and then applying our previous results from linear non-symmetric systems. We have demonstrated our results with a numerical example, see Tables 3.1, 3.2, and 3.3. In these tables, we have observed that the sequences of residual errors of the inner iterations are uniformly bounded and independent of the outer process.

Bibliography

[1] I. Antal and J. Karátson, *A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems*, Computers & Mathematics with Applications, 55 (2008), pp. 2185–2196.

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

[3] O. Axelsson and J. Karátson, *Mesh independent superlinear PCG rates via compact-equivalent operators*, SIAM Journal on Numerical Analysis, 45 (2007), pp. 1495–1516.

[4] O. Axelsson and J. Karátson, *Equivalent operator preconditioning for elliptic problems*, Numerical Algorithms, 50 (2009), pp. 297–380.

[5] O. Axelsson, J. Karátson, and F. Magoules, *Robust superlinear krylov convergence for complex non-coercive compact-equivalent operator preconditioners*, SIAM J. Numer. Anal., (2023, to appear).

[6] O. Axelsson and J. Karátson, *Superlinear convergence of the gmres for pde-constrained optimization problems*, Numer. Funct. Anal. Optim., 39 (2018), p. 921–936.

[7] H. Brézis, *Functional analysis, Sobolev spaces and partial differential equations*, vol. 2, Springer, 2011.

[8] G. Chávez, G. Turkiyyah, S. Zampini, H. Ltaief, and D. Keyes, *Accelerated cyclic reduction: A distributed-memory fast solver for structured linear systems*, Parallel Computing, 74 (2018), pp. 65–83.

[9] D. E. Edmunds and H. Triebel, *Entropy numbers and approximation numbers in function spacess*, Proceedings of the London Mathematical Society, 3 (1989), pp. 137–152.

[10] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, 2010.

[11] I. Faragó and J. Karátson, *Numerical solution of nonlinear elliptic problems via preconditioning operators: Theory and applications*, vol. 11, Nova Publishers, 2002.

[12] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Operator theory: Advances and applications*, Classes of Linear Operators, 49 (1992).

[13] J. Karátson, *Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators*, Applications of Mathematics, 50 (2005), pp. 277–290.

[14] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, vol. I,Functional Analysis, Academic Press, 1980.

[15] T. Rossi and J. Toivanen, *A parallel fast direct solver for the discrete solution of separable elliptic equations.*, in PPSC, Citeseer, 1997.

[16] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, 2003.

[17] Y. Saad and M. H. Schultz, *Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on scientific and statistical computing, 7 (1986), pp. 856–869.

[18] J. Vybíral, *Widths of embeddings in function spaces*, Journal of Complexity, 24 (2008), pp. 545–570.

[19] Z. Zlatev, *Numerical treatment of large air pollution models*, in Computer Treatment of Large Air Pollution Models, Springer, 1995, pp. 69–109.