Quasi-Newton type iterative solution of nonlinear elliptic PDEs with non-uniform monotonicity conditions

PhD Dissertation

Benjámin Borsos

Department of Analysis

Institute of Mathematics

Budapest University of Technology and Economics

Supervisor: János Karátson

2021

“Don’t be sorry, be better!” Ghost of Sparta

Acknowledgement.
I am grateful to my supervisor indeed, János Karátson, who introduced me to this mellow field of mathematics. He supported my work largely. I am especially grateful to and for my family for providing the emotional background for the years of work.

This research was supported by the Hungarian Scientific Research Fund OTKA, No. K112157 and SNN125119, and further, it was supported by the National Research, Development and Innovation Fund (TUDFO/51757/2019-ITM, Thematic Excellence Program). Additionally, this research has been supported by the BME NC TKP2020 grant of NKFIH Hungary.

Contents

1 Introduction

Nonlinear elliptic problems arise in various applications for models that describe stationary states, see, e.g., [23, 27, 35, 43, 57] and the references there. We may mention, for instance, elasticity, glaciology, heat radiation, flow problems in physics and other fields, see, e.g., [14, 26, 29, 59, 62]. As shown by such works as well, a widespread way to solve such problems is to use finite element discretization (FEM) and then to apply a Newton-like iteration, see also [31, 43, 71].

To construct quasi-Newton methods, a general approach to has been given in [52], where approximate Jacobians are defined via spectral equivalence, and hence they can be regarded as variable preconditioners. Thus variable preconditioning provides a transition between fixed preconditioning and Newton's method. With fixed preconditioning, one can define simple iteration that is often able to yield favorable speed of global convergence if supplied with suitable preconditioning, and in these cases its usage can be justified versus Newton's method owing to the extra work of forming the Jacobians (see, e.g., [5, 9] for early work in this direction, further [64] and the references therein for later applications.). However, it might be insufficient for strong nonlinearities. These can favor Newton's method, which is more complicated and costly for one iteration step, but provides better convergence altogether for the strong nonlinearities. Quasi-Newton methods, which can also be regarded as variable preconditioners, may combine the advantages of these methods. Alternatively, one can use Newton's method for outer iteration, and exploit the preconditioner in case of the inner iteration, as follows.

The Newton-type method yields linear problems that can be solved by direct or approximate methods, depending on the scale of the problem. A widely used approximate approach is to use conjugate gradient method (CGM) in the inner iterations. The construction of such inner-outer iterations can be found in [30, 70], their framework for uniformly monotone elliptic problems has been presented in [4, 71], see also [60, 79] for recent applications. Preconditioned CGM can be readily formulated with the help of the variable preconditioners discussed above.

The aim of this dissertation is to extend earlier results that provide theory only for the assumption of uniform ellipticity, which does not hold for many real-life problems, for example, in non-Newtonian flows, nonlinear optics, minimal surface problems, glacier modelling, etc.

We make use of tools of functional analysis to address this task, since Sobolev spaces are the natural underlying spaces of the boundary value problems that are the subjects of our investigation. Given a nonlinear boundary value problem, one can usually formulate an operator equation of the form $F(u) = 0$, where F maps either a Hilbert or Banach space to its dual. Depending on the original nonlinearity, the resulting operator F may or may not exhibit ellipticity with uniform lower and upper bounds.

The structure of the dissertation is the following.

In Section 2, the results prior to the work presented here are discussed, detailing the convergence rates of the quasi-Newton method for problems with uniform lower and upper bound in the ellipticity condition, for well-posed problems in Hilbert function space, for both local and global convergence. Furthermore, a brief insight is given into the related boundary value problems and the place of quasi-Newton methods among Newton-type methods. This section is based on [52].

Sections 3 and 4 are devoted to quasi-Newton methods for problems with stronger nonlinearities.

In Section 3, the uniform upper bound in the ellipticity condition is relaxed, and the deriva-

tion of the resulting convergence rates are discussed for local and global convergence. The Lipschitz condition is also found to be relaxed as a result of the more general assumptions. An example problem is discussed in detail using an equation from nonlinear optics, and numerical results are presented.

In Section 4, Banach space setting is employed for operators with relaxed both lower and upper bounds. Details of a model classification is given, and various examples are shown for equations, based on real-life models, that satisfy our conditions and for efficient variable preconditioners.

Section 5 presents the results for inner-outer iterations in case of inexact Newton methods. The results are illustrated using a nonlinear fluid flow model.

Section 6 studies a one-dimensional fourth-order engineering model, with detailed theoretical and practical comparison of the three methods (simple iteration, quasi-Newton and full Newton).

Section 7 presents a problem with a Stefan-Boltzmann heat radiation boundary nonlinearity in 3D. Nonnegativity result is obtained, finite element approximation is discussed, and the applicability of the quasi-Newton methods is shown.

Sections 3, 4, 5, 6 and 7 are based on the author's papers [17], [18], [19], [20] and [21], respectively. The numerical investigations have been carried out using Matlab.

Section 8 is entitled to give a short summary for developers on the apparent findings they might find interesting. In Section 9, a brief general summary of the presented results can be found.

2 Theoretical background

This section is entitled to provide results of work prior (see [52]) to this dissertation for the sake of convenience and to provide reference material for later sections.

Let H be a real Hilbert space with inner product $\langle ., . \rangle$ and corresponding norm $\| \cdot \|$. For the operator $F : H \to H$, the following operator equation is under investigation

$$F(u) = 0. \tag{2.1}$$

2.1 Basic concept and motivation

In this subsection, we discuss the basic concept of quasi-Newton methods, and their place as compared to other methods. Algorithms used for local convergence are discussed, but the conceptual illustration holds for more complicated requirements.

Nonlinear boundary value problems can often be written in the form of (2.1). If this has a unique solution u^*, we may approach it with a sequence $(u_n) \subset H$ using the following algorithm:

$$u_{n+1} := u_n + p_n \qquad (\forall n \in \mathbb{N}),$$

$$\text{where :} \qquad B_n p_n = -F(u_n),$$

with an auxiliary linear operator $B_n : \ H \to H$ and sequence $(p_n) \subset H$. Different formulations of B_n correspond to different methods.

For the Sobolev gradient method, $B_n = \text{const.} \cdot I$, hence the algorithm reads simply as $u_{n+1} = u_n - \text{const.} \cdot F(u_n)$, allowing swift computation for each n, however, the required number of iterations might be large.

In contrast, Newton's method uses $B_n = F'(u_n)$, which may be hard to compute for each n, but it provides convergence of quadratic order.

The idea of quasi-Newton methods is to choose auxiliary operators B_n between these two in a sophisticated manner, so that the advantages of these methods can be combined. The "sophisticated manner", in which B_n is obtained, is similar to the general twofold idea of preconditioning:

(i) a symbolic simplification of the operator $F'(u_n)$ by intuition, so that we obtain a significantly simpler operator, moreover,

(ii) a spectral equivalence relation can be established between B_n and $F'(u_n)$ which provides a favorable convergence result.

The term variable preconditioning can be used, because the auxiliary operator B_n is allowed to be different for each n, otherwise, it could be called constant preconditioning.

For a comparison in case of a one-dimensional example, see Section 6. For further details regarding the comparison of the three methods, see Section 8.

In the next two subsections, we summarize the corresponding results in literature prior to this work.

2.2 Linear convergence by constant preconditioning

Let us recall the following theorem, which provides a linear convergence result [37].

Theorem 2.1. *Let H be a real Hilbert space and $F : H \to H$ a nonlinear operator. Let F have a Gâteaux derivative that satisfies the following properties:*

(i) For any $u \in H$ the operator $F'(u)$ is self-adjoint.

(ii) There exist constants $\Lambda \geq \lambda > 0$ satisfying:

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

Denote by $u^* \in H$ the unique solution of $F(u) = 0$. Let $M \geq m > 0$ be given constants and $B : H \to H$ a bounded self-adjoint linear operator such that:

$$m\langle Bh, h\rangle \leq \langle F'(u)h, h\rangle \leq M\langle Bh, h\rangle \qquad (\forall u,\ h \in H). \tag{2.2}$$

Then the sequence, defined by

$$u_{n+1} := u_n - \frac{2}{M+m} B^{-1} F(u_n) \qquad (\forall n \in \mathbb{N}),$$

converges linearly to u^*, namely, for given $u_0 \in H$ there exists a constant $C > 0$ such that

$$\|u_n - u^*\| \leq C \left(\frac{M-m}{M+m}\right)^n \qquad (\forall n \in \mathbb{N}).$$

Proof. See [37]. ∎

Here, (ii) is called the ellipticity condition. This condition is too strong for various applications (see details later in e.g. Subsection 3.3), the main aim of this work is to show that similar results can be obtained for certain more general conditions.

In (2.2), the chain of inequalities is called the spectral equivalence of operators $F'(u)$ and B, while m and M can be called the spectral equivalence constants, or spectral bounds.

The following two lemmas, see [35], are keys in the proofs of this work, showing inequalities for energy norms of the inverses of operators which are connected by spectral equivalence. The following energy inner product is used for positive definite self-adjoint operator A:

$$\langle u, v\rangle_A := \langle Au, v\rangle \qquad (\forall u,\ v \in H),$$

with corresponding norm $\|\cdot\|_A$. The energy norm for a linear operator in H, self-adjoint w.r.t. the A^{-1}-inner product, is

$$\|B\|_{A^{-1}} = \sup_{h \neq 0} \frac{|\langle Bh, h\rangle_{A^{-1}}|}{\|h\|^2_{A^{-1}}}.$$

Assumptions 2.2. *Let A and B be self-adjoint linear operators in H with positive lower bound, and there exist constants $M \geq m > 0$ such that*

$$m\langle Bh, h\rangle \leq \langle Ah, h\rangle \leq M\langle Bh, h\rangle \qquad (\forall h \in H).$$

Lemma 2.3. *Let operators A and B satisfy Assumptions 2.2, then*

$$m\langle A^{-1}h,h\rangle \le \langle B^{-1}h,h\rangle \le M\langle A^{-1}h,h\rangle \qquad (\forall h\in H).$$

Lemma 2.4. *Let operators A and B satisfy Assumptions 2.2, then*

$$\left\| I-\frac{2}{M+m}AB^{-1}\right\|_{A^{-1}} \le \frac{M-m}{M+m}.$$

2.3 Linear convergence by variable preconditioning

The next theorem shows that using a different auxiliary operator for every step n (i. e. variable preconditioning), with the same fixed spectral bounds as in Theorem 2.1, entails local linear convergence for Lipschitz-continuous operators. This variable preconditioning is vital for a useful algorithm, and the convergence can be made global again by damping, as it can be seen below. For the sake of convenience, the unchanged parts of the previous theorem are repeated. This result was presented in [52].

Assumptions 2.5. *Let H be a real Hilbert space and $F:H\to H$ a nonlinear operator. Let F have a Gâteaux derivative that satisfies the following properties:*

(i) For any $u\in H$ the operator $F'(u)$ is self-adjoint.

(ii) There exist constants $\Lambda\ge\lambda>0$ satisfying:

$$\lambda\|h\|^2 \le \langle F'(u)h,h\rangle \le \Lambda\,\|h\|^2 \qquad (\forall u,h\in H). \tag{2.3}$$

(iii) There exists $L>0$ such that

$$\|F'(u)-F'(v)\| \le L\|u-v\| \qquad (u,\ v\in H). \tag{2.4}$$

Denote by $u^*\in H$ the unique solution of $F(u)=0$. Let $M\ge m>0$ be given constants, and for any $n\in\mathbb{N}$ let us choose a bounded self-adjoint linear operator $B_n:H\to H$ such that

$$m\langle B_nh,h\rangle \le \langle F'(u_n)h,h\rangle \le M\langle B_nh,h\rangle \qquad (\forall h\in H). \tag{2.5}$$

Algorithm 2.6. *With Assumptions 2.5, starting from a $u_0\in H$, we define a sequence from the following formula:*

$$u_{n+1} := u_n - \frac{2}{M+m}B_n^{-1}F(u_n) \qquad (\forall n\in\mathbb{N}). \tag{2.6}$$

Theorem 2.7. *With Assumptions 2.5, the sequence generated by Algorithm 2.6 converges locally linearly to u^*, namely, there exists a neighbourhood U of u^* and for given $u_0\in U$ there exists a constant $C>0$ such that*

$$\|u_n-u^*\| \le C\left(\frac{M-m}{M+m}\right)^n \qquad (\forall n\in\mathbb{N}). \tag{2.7}$$

We introduce the following energy norms:

$$\|h\|_u := \langle F'(u)^{-1}h,h\rangle^{1/2} \quad (\text{for given } u\in H), \qquad \|\cdot\|_* := \|\cdot\|_{u^*}, \qquad \|\cdot\|_n := \|\cdot\|_{u_n} \tag{2.8}$$

(for given $n \in \mathbb{N}$).

It follows readily (with Lemma 2.3) that for fixed u the norms $\|\cdot\|_u$ and $\|\cdot\|$ are equivalent, namely:

$$\lambda^{1/2}\|h\|_u \le \|h\| \le \Lambda^{1/2}\|h\|_u \qquad (\forall h \in H). \tag{2.9}$$

Lemma 2.8. *For all $h \in H$ we have*

$$\frac{1}{1+\mu(u_n)} \le \frac{\|h\|_*^2}{\|h\|_n^2} \le 1+\mu(u_n)\,,$$

where

$$\mu(u_n) := L\lambda^{-2}\|F(u_n)\|\ .$$

Proof. The lower bound in (2.3) implies a corresponding lower estimate for the variation of F:

$$\|F(u)-F(v)\| \ge \lambda\|u-v\| \qquad (\forall u, v \in H). \tag{2.10}$$

This, together with the assumptions (2.3)–(2.4) on F', implies

$$\langle F'(u)h,\ h\rangle \le \langle F'(v)h,\ h\rangle(1+L\lambda^{-2}\|F(u)-F(v)\|) \qquad (\forall u, v, h \in H).$$

For the case $u = u^*$ and $v = u_n$, this gives $\langle F'(u^*)h,\ h\rangle \le \langle F'(u_n)h,\ h\rangle(1+L\,\lambda^{-2}\|F(u_n)\|)$. Reversing the role of u^* and u_n, we obtain

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

From this, Lemma 2.3 yields the desired estimate. ∎

Lemma 2.9. *With (2.4), the following formulation can be made:*

$$\begin{aligned} &F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1}-u_n) + R(u_n),\\ &\text{where}: \|R(u_n)\| \le \frac{L}{2}\|u_{n+1}-u_n\|^2 \qquad (\forall n \in \mathbb{N}). \end{aligned} \tag{2.11}$$

Proof. This formulation can be written due to (iii) of Assumptions 2.5. ∎

Lemma 2.10. *With Assumption 2.5 (i) and the sequence definition of Algorithm 2.6, we obtain*

$$\|F(u_n)+F'(u_n)(u_{n+1}-u_n)\|_n \le \frac{M-m}{M+m}\|F(u_n)\|_n \qquad (\forall n \in \mathbb{N}).$$

Proof. Let $n \in \mathbb{N}$ be fixed. Using (2.6), one can write

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

taking the $\|\cdot\|_n$ norm of both sides, we get:

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

and Lemma 2.4 yields the result. ∎

Lemma 2.11. *With Assumptions 2.5, and the sequence definition of Algorithm 2.6, we obtain*

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

Proof. Due to (ii) and (iii) of Assumptions 2.5 and (2.5), we have $\lambda M^{-1}\|h\|^2 \leq \langle B_n h, h\rangle$ for all $h \in H$, this yields the result. ∎

Lemma 2.12. *With Assumptions 2.5, and the sequence definition of Algorithm 2.6, we obtain*

$$\|R(u_n)\|_n \leq \frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2. \tag{2.12}$$

Proof. Applying (2.9) and using (2.6) on (2.11) yields the result. ∎

Proof of Theorem 2.7.

We use Lemma 2.11, (2.5) and Lemma 2.3 on (2.12), this yields

$$\begin{aligned}\|B_n^{-1}F(u_n)\|^2 &\leq \|B_n^{-1/2}\|^2\|B_n^{-1/2}F(u_n)\|^2 \leq M\lambda^{-1}\langle B_n^{-1}F(u_n), F(u_n)\rangle \\ &\leq M^2\lambda^{-1}\langle F'(u_n)^{-1}F(u_n), F(u_n)\rangle = M^2\lambda^{-1}\|F(u_n)\|_n^2.\end{aligned} \tag{2.13}$$

Hence

$$R(u_n)_n \leq \frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2. \tag{2.14}$$

Combining Lemma 2.9 with Lemma 2.10 and (2.14) entails

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

Finally, using Lemma 2.8, we obtain

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

where $\mu^*(u_n) = L\Lambda^{1/2}\lambda^{-2}\|F(u_n)\|_*$ ($\mu^*(u_n)$ gives an upper estimate for $\mu(u_n)$ in $\|\cdot\|_*$). That is,

$$\|F(u_{n+1})\|_* \leq \varphi(\|F(u_n)\|_*)\|F(u_n)\|_* \tag{2.15}$$

where

$$\varphi(t) = (1+\beta\Lambda^{1/2}t)\left(Q + M^2\beta\alpha^{-2}\lambda^{1/2}(t/2)\left(1+\beta\Lambda^{1/2}t\right)^{1/2}\right) \tag{2.16}$$

and the notations

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

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

Estimate (2.15) puts us in the position to prove the required convergence estimate (2.7), provided that the assumption

$$r := \varphi(\|F(u_0)\|_*) < 1$$

is satisfied for the initial guess.

First, we obtain by induction that

$$\|F(u_{n+1})\|_* \le r\|F(u_n)\|_* \quad (n \in \mathbb{N}). \tag{2.17}$$

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

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

Inequality (2.17) implies that $\|F(u_n)\|_* \le r^n\|F(u_0)\|_* \to 0$, $\quad \varphi(\|F(u_n)\|_*) \to Q$, and hence

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

From now on we use the notation for the remainder of the subsection

$$e_n = \|F(u_n)\|_*.$$

Then (2.17) implies that

$$e_n \le \left(\prod_{k=0}^{n-1} \varphi(e_k)\right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q}\right) Q^n e_0 \quad (n \in \mathbb{N}). \tag{2.18}$$

Using (2.16) and the notations $c = \beta\Lambda^{1/2}$, $d = (M^2\beta\alpha^{-2}\lambda^{1/2})/2$, we have

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

Here

$$\begin{aligned}
\frac{\varphi(e_k)}{Q} &= (1+ce_k)\left(1 + \frac{d}{Q}e_k(1+ce_k)^{1/2}\right) \\
&\le (1+ce_k)\left(1 + \frac{d}{Q}e_k\left(1+\frac{c}{2}e_k\right)\right) = 1 + \left(c+\frac{d}{Q}\right)e_k + \frac{cd}{Q}e_k^2 + \frac{c^2d}{2Q}e_k^3 \\
&\le 1 + \left(c+\frac{d}{Q}\right)e_0 r^k + \frac{cd}{Q}e_0^2 r^{2k} + \frac{c^2d}{2Q}e_0^3 r^{3k}.
\end{aligned}$$

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

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

Therefore (2.18) yields

$$e_n \;\le\; e_0 E \cdot Q^n \qquad (n \in \mathbb{N}).$$

Finally, using (2.10) and (2.9), this implies

$$\|u_n - u^*\| \le \lambda^{-1}\|F(u_n)\| \le\; \lambda^{-1}\Lambda^{1/2} e_0 E \cdot Q^n \qquad (n \in \mathbb{N}),$$

which coincides with the required convergence estimate with $C = \lambda^{-1}\Lambda^{1/2}e_0 E$.

Remark 2.13. The convergence has been proved under the sufficient condition

$$\varphi(\|F(u_0)\|_*) < 1 \tag{2.19}$$

for the initial guess, with φ defined in (2.16). In connection with this we note the following:

(a) The condition (2.19) is satisfied if

$$K\frac{L}{\lambda^2}\|F(u_0)\|_* < 1\,,$$

where $K = \Lambda^{1/2}(M/m)\max\{1,\, 2(M-m)^{-1}(\lambda/\Lambda)^{1/2}\}$. Relating this to the well-known sufficient condition $\frac{L}{2\lambda^2}\|F(u_0)\| < 1$ of the exact Newton iteration, we observe that the order is similar (although K is obviously somewhat larger than $1/2$).

(b) The sufficient condition of convergence can be given using the original norm as follows. Since the theoretical norm $\|\,.\,\|_*$ satisfies $\|F(u_0)\|_* \le \lambda^{-1/2}\|F(u_0)\|$ by (2.9), and φ increases, therefore we obtain the condition

$$\varphi(\lambda^{-1/2}\|F(u_0)\|) < 1$$

to be checked for u_0.

2.4 Damped quasi-Newton method as variable preconditioning

We recall the following definitions of norms (see (2.8)), where (u_n) is an iterative sequence and u^* is the solution of $F(u) = 0$:

$$\|h\|_n = \langle F'(u_n)^{-1}h, h\rangle^{1/2} \quad (n \in \mathbb{N}), \qquad \|h\|_* = \langle F'(u^*)^{-1}h, h\rangle^{1/2}. \tag{2.20}$$

The following theorem generalizes Theorem 2.7. Using damped iteration and variable spectral bound preconditioning, it provides global convergence up to second order. This is another result of [52].

Assumptions 2.14. *Let H be a real Hilbert space and $F : H \to H$ a nonlinear operator. Let F have a Gâteaux derivative that satisfies the properties (i)-(iii) of Assumptions 2.5.*
Denote u^* the unique solution of equation $F(u) = 0$.
Furthermore, the following conditions hold:

(iv) *For each $n \in \mathbb{N}$, let $M_n \ge m_n > 0$ and let us choose a bounded self-adjoint linear operator $B_n : H \to H$ such that*

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

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

(v) We define

$$\tau_n = \min\left\{1, \frac{1-Q_n}{2\rho_n}\right\}, \tag{2.21}$$

where $Q_n = \frac{M_n - m_n}{M_n + m_n}(1+\mu(u_n))$, $\rho_n = 2LM_n^2\lambda^{-3/2}(M_n+m_n)^{-2}\|F(u_n)\|_n(1+\mu(u_n))^{1/2}$, $\mu(u_n)$ *is as in condition (iv), and* $\|\,.\,\|_n$ *is defined in* (2.20). *(This value of* τ_n *ensures optimal contractivity in the* n*th step in the* $\|\,.\,\|_*$*-norm.)*

Algorithm 2.15. *With Assumptions 2.14, for arbitrary* $u_0 \in H$ *let* (u_n) *be the sequence defined by*

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} B_n^{-1} F(u_n) \qquad (n \in \mathbb{N}). \tag{2.22}$$

Theorem 2.16. *With Assumptions 2.14, the sequence generated by Algorithm 2.15 converges globally linearly to* u^**, namely,*

$$\|u_n - u^*\| \le \lambda^{-1}\|F(u_n)\| \;\to\; 0; \tag{2.23}$$

namely,

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup \frac{M_n - m_n}{M_n + m_n} < 1\,. \tag{2.24}$$

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

$$\|F(u_{n+1})\|_* \le d_1\|F(u_n)\|_*^{1+\gamma} \qquad (n \in \mathbb{N}) \tag{2.25}$$

with some constant $d_1 > 0$*.*

Owing to the equivalence of the norms $\|\,.\,\|$ and $\|\,.\,\|_*$, the orders of convergence corresponding to the estimates (2.24) and (2.25) can be formulated with the original norm.

Corollary 2.17. *(rate of convergence in the original norm)*

(a) *If* $\limsup M_n/m_n = K > 1$*, then*

$$\|u_n - u^*\| \le \lambda^{-1}\|F(u_n)\| \le \text{const.} \cdot \rho^n$$

with $\rho = (K-1)/(K+1)$*.*

(b) *In the case* $M_n/m_n \le 1 + c_1\|F(u_n)\|^\gamma$ *(with constants* $c_1 > 0$, $0 < \gamma \le 1$*) there holds*

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

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

The proof of Theorem 2.16 is carried out as a sequence of lemmas for later use below.

Lemma 2.18. *With Assumptions 2.14, and the sequence definition of Algorithm 2.15, the following inequality holds:*

$$\|F(u_{n+1})\|_* \le \left(1 - \tau_n(1-Q_n) + \tau_n^2\rho_n\right)\|F(u_n)\|_*\,, \tag{2.26}$$

where Q_n *and* ρ_n *are as in condition (v) of Assumptions 2.14.*

PROOF. Using Lemma 2.9 and (2.22), we obtain

$$F(u_{n+1}) = (1-\tau_n)F(u_n) - \tau_n\left(F(u_n) - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}F(u_n)\right) + R(u_n).$$

Hence

$$\|F(u_{n+1})\|_* \le (1-\tau_n)\|F(u_n)\|_* + \tau_n\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* + \|R(u_n)\|_*.$$

Here, using Lemma 2.8 and Lemma 2.4,

$$\begin{aligned}\left\|\left(I - \frac{2}{M_n+m_n}F'(u_n)B_n^{-1}\right)F(u_n)\right\|_* &\le (1+\mu(u_n))^{1/2}\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_n \\ &\le (1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_*,\end{aligned}$$

where $\mu(u_n) = L\lambda^{-2}\|F(u_n)\|$. Further, from Lemma 2.9 and (2.9) there follows

$$\|R(u_n)\|_* \le \frac{L}{2\lambda^{1/2}}\|u_{n+1}-u_n\|^2 = \tau_n^2\frac{2L}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2,$$

hence, using the estimate (2.13) and then Lemma 2.8, we obtain

$$\begin{aligned}\|R(u_n)\|_* &\le \tau_n^2\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n^2 \\ &\le \tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\|F(u_n)\|_*.\end{aligned}$$

Summing up, we obtain

$$\begin{aligned}\|F(u_{n+1})\|_* \le \Bigg(&1-\tau_n+\tau_n(1+\mu(u_n))\frac{M_n-m_n}{M_n+m_n} \\ &+\tau_n^2(1+\mu(u_n))^{1/2}\frac{2LM^2}{\lambda^{3/2}(M+m)^2}\|F(u_n)\|_n\Bigg)\|F(u_n)\|_*,\end{aligned}$$

where the definitions of Q_n and ρ_n in condition *(v)* of Assumptions 2.14 yield the result. ∎

Lemma 2.19. *With Assumptions 2.14, there exists $\tilde{Q} < 1$, such that*

$$Q_n \le \tilde{Q} \qquad (n \in \mathbb{N}).$$

PROOF. The assumption $M_n/m_n \le 1 + 2/(\varepsilon + K\mu(u_n))$ with $K > 1$ and $\varepsilon > 0$ implies that

$$1+\varepsilon+K\mu(u_n) \le 1 + \frac{2}{(M_n/m_n)-1} = \frac{M_n+m_n}{M_n-m_n};$$

hence

$$1+\mu(u_n) \le \tilde{Q}\,\frac{M_n+m_n}{M_n-m_n}$$

with $\tilde{Q} := \max\{1/K, 1/(1+\varepsilon)\} < 1$. ∎

Lemma 2.20. *We take Assumptions 2.14, and the sequence definition of Algorithm 2.15, furthermore, let us introduce the function* $p : [0,1] \to \mathbb{R}$, $p(t) := 1 - (1 - Q_n)t + \rho_n t^2$. *Then,*

$$\|F(u_{n+1})\|_* \le p(\tau_n)\|F(u_n)\|_* < \|F(u_n)\|_*\,, \tag{2.27}$$

and (2.23)-(2.24) hold.

Proof. Since $p'(0) = -(1 - Q_n) < 0$, we can readily obtain that $p'(t) = -(1 - Q_n) + 2\rho_n t$ yields that τ_n defined in (2.21) satisfies

$$p(\tau_n) = \min_{t\in[0,1]} p(t) < 1.$$

Hence from (2.26), we have (2.27).

Moreover, if $\tau_n = 1$ (i.e., when $1 \le (1 - Q_n)/2\rho_n$), then

$$p(\tau_n) = Q_n + \rho_n \le Q_n + (1 - Q_n)/2 = (1 + Q_n)/2 \le (1 + \tilde{Q})/2 < 1. \tag{2.28}$$

In the case $\tau_n = (1 - Q_n)/2\rho_n$ we have

$$p(\tau_n) = 1 - (1 - Q_n)^2/(4\rho_n) \le 1 - (1 - \tilde{Q})^2/(4\sup_n \rho_n) =: Q' < 1. \tag{2.29}$$

The latter holds since by (2.27) $\|F(u_n)\|_*$ is bounded, and hence

$$\rho_n = \text{const.} \cdot \|F(u_n)\|_n \left(1 + \text{const.} \cdot \|F(u_n)\|\right)^{1/2} \tag{2.30}$$

is bounded, the three norms being equivalent. Altogether, from (2.27) we obtain

$$\|F(u_n)\|_* \le \text{const.} \cdot r^n \to 0,$$

where $r = \max\{(1 + \tilde{Q})/2, Q'\}$. This also implies that $\rho_n \to 0$ and $\mu(u_n) = L\lambda^{-2}\|F(u_n)\| \to 0$, which yields (2.23) by (2.10). A brief calculation gives

$$p(\tau_n) = Q_n + \rho_n \left(1 - (1 - \tau_n)^2\right) \tag{2.31}$$

(for both $\tau_n = 1$ and $\tau_n < 1$); hence (2.27) yields

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup Q_n = \limsup \frac{M_n - m_n}{M_n + m_n}\,. \qquad \blacksquare$$

Remark 2.21. *The bound* $M_n/m_n \le 1 + 2/\varepsilon$ *in (iv) of Assumptions 2.14 implies that*

$$\limsup \frac{M_n - m_n}{M_n + m_n} \le \frac{1}{1 + \varepsilon} < 1\,.$$

Lemma 2.22. *We take Assumptions 2.14, and the sequence definition of Algorithm 2.15. Additionally, let* $M_n/m_n \le 1 + c_1\|F(u_n)\|^\gamma$ *with constants* $c_1 > 0$, $0 < \gamma \le 1$. *Then (2.25) holds with some constant* $d_1 > 0$.

PROOF. Here, $M_n/m_n \le 1 + c_2\|F(u_n)\|_*^\gamma$ with $c_2 = c_1\Lambda^{1/2}$; hence

$$\frac{M_n - m_n}{M_n + m_n} < \frac{M_n - m_n}{m_n} \le c_2\|F(u_n)\|_*^\gamma\,,$$

and therefore

$$Q_n \le c_3\|F(u_n)\|_*^\gamma$$

with $c_3 = c_2(1 + \sup_n \mu(u_n))$. Also,

$$\rho_n \le c_4\|F(u_n)\|_*$$

with some $c_4 > 0$ since $\|F(u_n)\|_*$ is bounded (cf. (2.30)). With the use of notation $e_n = \|F(u_n)\|_*$, we obtain from (2.27) and (2.31) that

$$\begin{aligned} e_{n+1} &\le (Q_n + \rho_n)\, e_n \le (Q_n + c_4 e_n)\, e_n \le \left(c_3 e_n^\gamma + c_4 e_0 \frac{e_n}{e_0}\right) e_n \\ &\le \left(c_3 e_n^\gamma + c_4 e_0 \left(\frac{e_n}{e_0}\right)^\gamma\right) e_n = d_1 e_n^{1+\gamma} \end{aligned}$$

with $d_1 = c_3 + c_4 e_0^{1-\gamma}$. ∎

Proof of Theorem 2.16. Lemmas 2.18-2.22 yield the result. ∎

Remark 2.23. (a) It is worth mentioning that Theorems 2.7 and 2.16 use descent methods, similarly to the simple iteration. Namely, conditions (i)–(iii) of Assumptions 2.5 imply the existence of a potential $\Phi : H \to \mathbb{R}$, i.e., $\Phi'(u) = F(u) \quad (u \in H)$. Then the directions $-B_n^{-1}F(u_n)$ are descent directions, since their angle is acute with the steepest descent direction $-F(u_n)$ owing to

$$\langle B_n^{-1}F(u_n), F(u_n)\rangle > 0\,.$$

We also note that the residuals $\Phi(u_n) - \Phi(u^*)$ are equivalent to $\|u_n - u^*\|^2$ owing to the ellipticity condition (ii) of Assumptions 2.5.

(b) The relaxation of conditions of Theorems 2.7 and 2.16 is the main aim of this work. Further, the proofs can be repeated with obvious modification if (iii) is replaced by Hölder continuity only, see Section 4. These propositions have been previously made in [52] (see also [6]).

(c) The value of τ_n need not necessarily be maximized by 1 as in (2.21), but suitable overrelaxation is also feasible which may accelerate the convergence.

2.5 Examples of nonlinear problems

In this subsection we give an example for problems fitting into the setting of this section, namely, uniform ellipticity. Examples in more detail can be found in Subsection 4.2 for the setting of Section 4. One can find problems there, which do not fit into the current setting.

Let us consider the BVP

$$\begin{cases} -\operatorname{div} f(x, \nabla u) = g(x), \\ u_{|\partial\Omega} = 0 \end{cases} \tag{2.32}$$

with the following conditions: $\Omega \subset \mathbb{R}^N$ is a bounded domain, $g \in L^2(\Omega)$, f is measurable and $f(x,.) \in C^1(\mathbb{R}^N, \mathbb{R}^N)$ for all $x \in \Omega$. Furthermore, the Jacobians $\frac{\partial f(x,\eta)}{\partial \eta}$ are Lipschitz continuous in η, symmetric, and satisfy

$$\lambda|\xi|^2 \leq \frac{\partial f(x,\eta)}{\partial \eta}\xi \cdot \xi \leq \Lambda|\xi|^2, \qquad (x,\eta) \in \Omega \times \mathbb{R}^N, \xi \in \mathbb{R}^N,$$

with constants $\Lambda \geq \lambda > 0$ independent of (x,η).

An important special case of f is of the form $f(x,\eta) = a(|\eta|)\eta$, where $0 < \lambda \leq a(r) \leq (ra(r))' \leq \Lambda \quad (r > 0)$. Then (2.32) becomes

$$\begin{cases} -\operatorname{div} a(|\nabla u|)\nabla u = g(x), \\ u_{|\partial\Omega} = 0. \end{cases}$$

This kind of operator arises, e.g., in plasticity theory or in connection with magnetic potential (see, e.g., [57, 61]).

Owing to uniform ellipticity, problem (2.32) has a unique weak solution $u^* \in H_0^1(\Omega)$, and the finite element approximations converge to u^*.

Let $V_h \subset H_0^1(\Omega)$ be a finite element subspace, and denote $u_h \in V_h$ the solution of the discretized problem

$$\int_\Omega f(x, \nabla u_h) \cdot \nabla v = \int_\Omega gv \qquad (v \in V_h). \tag{2.33}$$

As shown in [52], one can find preconditioners for the iterative solution of (2.33) based on Theorem 2.16.

3 Quasi-Newton variable preconditioning under strong upper growth conditions in Hilbert spaces

3.1 Preliminaries

The goal of this section is to extend the approach of variable preconditioning, quoted in the previous section, to problems with stronger nonlinearities without an upper uniform boundedness assumption. This situation covers power order growth of nonlinearities, which also appears in various physical models. First we generalize the Hilbert space method of Section 2 to a class of unbounded nonlinearities in Section 3.2. Then the result is applied to a class of elliptic problems with power order nonlinearities in Section 3.3. Numerical tests reinforce the theoretical convergence results.

The results of this section are based on [17].

3.2 Variable preconditioning for strongly nonlinear operator equations

Let H be a real Hilbert space with inner product $\langle ., . \rangle$ and corresponding norm $\| \cdot \|$. We study operator equations

$$F(u) = 0$$

for a given nonlinear operator $F : H \to H$. Our goal is to extend Theorem 2.7 to nonlinearities without an upper uniform boundedness assumption, and thereby to prove the convergence of a proper iteration with variable preconditioning.

The allowed strong nonlinearity of the operator means that both the upper spectral bounds and the Lipschitz constants of the Gâteaux derivatives are allowed to grow up to infinity along with the norms of the arguments. The setting is based on the results recalled in Section 2, however, its technique has to be essentially redone to follow and eliminate the effect of the non-uniform nonlinearities. This is done in such a way that the variable bounds are incorporated in a modified recursive estimation of the residuals. Thus one can ensure that the overall convergence is not spoiled by the growth of nonlinearities. The method and its convergence are formulated as follows.

Assumptions 3.1. *Let H be a real Hilbert space and $F : H \to H$ a nonlinear operator. Let F have a Gâteaux derivative that satisfies the following properties:*

(i) For any $u \in H$ the operator $F'(u)$ is self-adjoint.

(ii) There exists a constant $\lambda > 0$ and a continuous increasing function $\Lambda : \mathbb{R}^+ \to \mathbb{R}^+$ such that the following condition is satisfied:

$$\lambda \|h\|^2 \leq \langle F'(u)h, h \rangle \leq \Lambda(\|u\|)\, \|h\|^2 \qquad (\forall u, h \in H). \tag{3.1}$$

(iii) There exists a continuous increasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying

$$\|F'(u) - F'(v)\| \leq L(\max\{\|u\|, \|v\|\})\, \|u - v\| \qquad (\forall u,\, h \in H). \tag{3.2}$$

Denote by $u^* \in H$ the unique solution of $F(u) = 0$. Let $M \geq m > 0$ be given constants, and for any $n \in \mathbb{N}$ let us choose a bounded self-adjoint linear operator $B_n : H \to H$ such that

$$m\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M\langle B_n h, h\rangle \qquad (\forall h \in H). \tag{3.3}$$

The following algorithm is identical to Algorithm 2.6, however, it is repeated here for the sake of convenience:

Algorithm 3.2. *With Assumptions 3.1, starting from a $u_0 \in H$, we define a sequence from the following formula:*

$$u_{n+1} := u_n - \frac{2}{M+m} B_n^{-1} F(u_n) \qquad (\forall n \in \mathbb{N}). \tag{3.4}$$

Theorem 3.3. *With Assumptions 3.1, the sequence generated by Algorithm 3.2 converges locally linearly to u^*, namely, there exists a neighbourhood U of u^* and for given $u_0 \in U$ there exists a constant $C > 0$ such that*

$$\|u_n - u^*\| \leq C \left(\frac{M-m}{M+m}\right)^n \qquad (\forall n \in \mathbb{N}). \tag{3.5}$$

The proof will be preceded by suitable lemmas. First, for a given bounded self-adjoint strictly positive operator A, the following notation stands for the energy inner product: $\langle u, v\rangle_A := \langle Au, v\rangle$, and the corresponding norm is $\|\cdot\|_A$.

Remark 3.4. The main assumption in the theorem is the local Lipschitz continuity of F', so we will derive some of its consequences below. In fact, it is easy to see that (3.2) implies the upper bound in (3.1): for any $u, h \in H$

$$\begin{aligned}\langle F'(u)h, h\rangle &= \langle (F'(u) - F'(0))h, h\rangle + \langle F'(0)h, h\rangle \\ &\leq \left(L(\|u\|)\|u\| + \|F'(0)\|\right)\|h\|^2,\end{aligned}$$

i.e. we have a bound of the form $\Lambda(\|u\|)\,\|h\|^2$ with the real function $\Lambda(t) := L(t)t + \|F'(0)\|$. This upper assumption in the theorem is only present in order to indicate the analogy with the cited earlier result.

Notations. In what follows, the functions Λ and L will be often evaluated on balls, in particular when we follow the iteration steps from u_n to u_{n+1}. Hence the following notations will be used: let

$$\tilde{\Lambda}_* := \Lambda(\|u^*\|),$$

and for fixed $n \in \mathbb{N}$ let

$$\tilde{L}_{n,n+1} := L(\max\{\|u_n\|, \|u_{n+1}\|\}), \qquad \tilde{L}_{n,*} := L(\max\{\|u_n\|, \|u_*\|\})\,. \tag{3.6}$$

Furthermore, we repeat the definition of energy norms of (2.8):

$$\|h\|_u := \langle F'(u)^{-1}h, h\rangle^{1/2} \quad (\text{for given } u \in H), \qquad \|\cdot\|_* := \|\cdot\|_{u^*}\,, \qquad \|\cdot\|_n := \|\cdot\|_{u_n}$$

(for given $n \in \mathbb{N}$). It follows readily (with Lemma 2.3) that for fixed u the norms $\|\cdot\|_u$ and

$\|\cdot\|$ are equivalent, namely:

$$\lambda^{1/2}\|h\|_u \le \|h\| \le \Lambda^{1/2}(\|u\|)\|h\|_u \qquad (\forall h \in H). \tag{3.7}$$

Two important special cases are

$$\|h\|_n \le \lambda^{-1/2}\|h\|, \qquad \|h\| \le \tilde{\Lambda}_*^{1/2}\|h\|_* \qquad (\forall h \in H). \tag{3.8}$$

Most of the proof presented in [17] is divided into several lemmas for clear understanding of the difference between the proof of Theorems 2.7 and 3.3. The following lemmas require minor modification of earlier results, but even in this first part we must follow more carefully the non-uniform constants.

Firstly, the norms $\|\cdot\|_*$ and $\|\cdot\|_n$ are related by the following non-uniform extension of Lemma 2.8.

Lemma 3.5. *For all $h \in H$ we have*

$$\frac{1}{1+\mu_n(u_n)} \le \frac{\|h\|_*^2}{\|h\|_n^2} \le 1+\mu_n(u_n)\,,$$

where

$$\mu_n(u_n) := \tilde{L}_{n,*}\tilde{\Lambda}_*^{1/2}\lambda^{-2}\|F(u_n)\|_*\,. \tag{3.9}$$

Proof. The lower bound in (3.1) implies a corresponding lower estimate for the variation of F identical to (2.10), repeated here:

$$\|F(u)-F(v)\| \ge \lambda\|u-v\| \qquad (\forall u, v \in H). \tag{3.10}$$

This, together with the assumptions (3.1)–(3.2) on F', implies

$$\langle F'(u)h,\ h\rangle \le \langle F'(v)h,\ h\rangle(1+L(\max\{\|u\|,\ \|v\|\})\lambda^{-2}\|F(u)-F(v)\|) \qquad (\forall u, v, h \in H).$$

For the case $u = u^*$ and $v = u_n$, this gives $\langle F'(u^*)h,\ h\rangle \le \langle F'(u_n)h,\ h\rangle(1+\tilde{L}_{n,*}\,\lambda^{-2}\|F(u_n)\|)$. Using (3.8) and reversing the role of u^* and u_n, we obtain

$$\frac{1}{1+\mu_n(u_n)} \le \frac{\langle F'(u^*)h, h\rangle}{\langle F'(u_n)h, h\rangle} \le 1+\mu_n(u_n) \qquad (\forall h \in H).$$

From this, Lemma 2.3 yields the desired estimate. ∎

Lemma 2.9 and Lemma 2.12 are substituted by the two lemmas below, respectively, while Lemma 2.10 and its proof applies here.

Lemma 3.6. *With (3.2), the following formulation can be made:*

$$\begin{aligned} &F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1}-u_n) + R(u_n), \\ &\text{where : } \|R(u_n)\| \le \frac{\tilde{L}_{n,n+1}}{2}\|u_{n+1}-u_n\|^2 \qquad (\forall n \in \mathbb{N}). \end{aligned} \tag{3.11}$$

Proof. This formulation can be written due to (iii) of Assumptions 3.1, more precisely, since (owing to (3.1)) F' is Lipschitz continuous in the ball $B(0, \max\{\|u_n\|, \|u_{n+1}\|\})$ with a corresponding constant $\tilde{L}_{n,n+1}$. ∎

Lemma 3.7. *With Assumptions 3.1, and the sequence definition of Algorithm 3.2, we obtain*

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

PROOF. Applying (3.7) and using (3.4) on (3.11) yields the result.

Lemma 3.8. *With Assumptions 3.1, and the sequence definition of Algorithm 3.2, we obtain that the preconditioning operators in* (3.3) *have uniformly bounded inverses, namely,*

$$\|B_n^{-1}\| \leq \lambda^{-1}M \qquad (\forall n \in \mathbb{N}).$$

PROOF. The upper and lower estimates in (2.5) and (2.3), respectively, yield:

$$\lambda\|h\|^2 \leq \langle F'(u_n)h, h\rangle \leq M\langle B_n h, h\rangle \leq M\|B_n h\|\|h\| \qquad (\forall h \in H).$$

Dividing by $\lambda\|h\|$ and using that $B_n : H \to H$ is bijection, we obtain the desired bound. ∎

Lemma 3.9. *With Assumptions 3.1, and the sequence definition of Algorithm 3.2, we obtain*

$$\|R(u_n)\|_n \leq K\tilde{L}_{n,n+1}\|F(u_n)\|_*^2,$$

where $K := \frac{2M^2\tilde{\Lambda}_*}{\lambda^{5/2}(M+m)^2}$.

PROOF. Lemma 3.7 and Lemma 3.8 apply here, this and (3.8) leads to the following estimate

$$\|B_n^{-1}F(u_n)\| \leq \lambda^{-1}M\|F(u_n)\| \leq \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_*.$$

By Lemma 3.7, this yields the result. ∎

The following lemma is a generalization of a phenomenon used earlier.

Lemma 3.10. *Let us consider a non-negative series* $(e_n)_{n\in\mathbb{N}} \subset \mathbb{R}^+$ *and a nondecreasing function* $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$. *We have the following assumptions:*

(i) For any $n \in \mathbb{N}$, *we have* $e_{n+1} \leq \varphi(e_n)e_n$,

(ii) $r := \varphi(e_0) < 1$.

Then

$$e_{n+1} \leq r^n e_0$$

holds for all $n \in \mathbb{N}$.

If we additionally assume, that $\frac{\varphi(e_k)}{Q} \leq 1 + \delta_k$ *holds for all* $k \in \mathbb{N}$, *we get*

$$e_n \leq e_0 Q^n \exp\Big(\sum_{k\in\mathbb{N}} \delta_k\Big).$$

PROOF. See the proof of Theorem 2.7, where a more specific case was investigated. ∎

Since the proofs of Theorems 2.7 and 3.3 have diverged, the parts right below are not divided into lemmas. Below we use the lemmas of this section above and those of the previous section that can be directly applied for the generalized Assumptions 3.1.

Proof of Theorem 3.3. The proof is carried out in several steps.

(1) The existence and uniqueness of the solution $u^* \in H$ is well-known for such potential problems, see, e.g., [35]. The major part of the proof will be the derivation of the local convergence of the residuals, i.e. $\lim_{n\to\infty} \|F(u_n)\| = 0$. Then (3.10) will yield that $\lim_{n\to\infty} u_n = u^*$ as well.

(2) As mentioned before, we can use Lemma 2.10, since its Assumption 2.5 (i) coincides with our present Assumption 3.1 (i). Also using Lemma 3.9 on Lemma 3.6, we get

$$\|F(u_{n+1})\|_n \le \frac{M-m}{M+m}\|F(u_n)\|_n + K\tilde{L}_{n,n+1}\|F(u_n)\|_*^2,$$

hence Lemma 3.5 yields

$$\|F(u_{n+1})\|_* \le (1+\mu_n(u_n))^{1/2}\left(\frac{M-m}{M+m}(1+\mu_n(u_n))^{1/2} + K\tilde{L}_{n,n+1}\|F(u_n)\|_*\right)\|F(u_n)\|_*.$$

(3) We formulate a recurrence for the sequence $\|F(u_n)\|_*$ as follows. By definition $\mu_n(u_n) \ge 0$, thus $(1+\mu_n(u_n))^{1/2} \ge 1$, consequently

$$\|F(u_{n+1})\|_* \le (1+\mu_n(u_n))\left(\frac{M-m}{M+m} + K\tilde{L}_{n,n+1}\|F(u_n)\|_*\right)\|F(u_n)\|_*. \tag{3.12}$$

By substituting the definition (3.9) of $\mu_n(u_n)$ and defining the real function

$$\varphi_n(t) := (1+\tilde{L}_{n,*}\tilde{\Lambda}_*^{1/2}\lambda^{-2}t)\left(Q + K\tilde{L}_{n,n+1}t\right), \qquad \text{where} \quad Q := \frac{M-m}{M+m}, \tag{3.13}$$

the estimate (3.12) can be reformulated as

$$\|F(u_{n+1})\|_* \le \varphi_n\left(\|F(u_n)\|_*\right)\|F(u_n)\|_*. \tag{3.14}$$

However, this recurrence cannot be directly used to derive convergence, since φ_n is a stepwise varying function containing $\tilde{L}_{n,n+1}$ and $\tilde{L}_{n,*}$, which is a new phenomenon compared to the proof of Theorem 2.7. Below we will show that these constants can be estimated as a function of $\|F(u_n)\|_*$, so that finally φ_n can be estimated independently of n.

(4) For the estimation of the constant $\tilde{L}_{n,n+1}$ we need to bound $\max\{\|u_n\|, \|u_{n+1}\|\}$ in terms of $\|F(u_n)\|_*$ and fixed constants. Here the definition (3.4) yields

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

hence a bound for $\max\{\|u_n\|, \|u_{n+1}\|\}$ can be obtained as a bound for the above r.h.s. Here, first, (3.10) yields $\|F(u_n) - F(0)\| \ge \lambda\|u_n\|$, hence, also using (3.8),

$$\|u_n\| \le \lambda^{-1}\left(\|F(u_n)\| + \|F(0)\|\right) \le \lambda^{-1}\left(\Lambda_*^{1/2}\|F(u_n)\|_* + \|F(0)\|\right). \tag{3.15}$$

Further, to estimate $\|B_n^{-1}F(u_n)\|$, we use Lemma 3.8 and (3.8) to derive for all $h \in H$ that $\|B_n^{-1}h\| \le \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|h\|_*$, hence

$$\|B_n^{-1}F(u_n)\| \le \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_*\,.$$

Thus, summing up, we have

$$\max\{\|u_n\|, \|u_{n+1}\|\} \le \lambda^{-1}\tilde{\Lambda}_*^{1/2}\big(1+\frac{2M}{M+m}\big)\|F(u_n)\|_* + \lambda^{-1}\|F(0)\| \;=:\; f(\|F(u_n)\|_*),$$

where the real function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is defined as

$$f(t) := \lambda^{-1}\tilde{\Lambda}_*^{1/2}\big(1+\frac{2M}{M+m}\big)\, t \;+ \lambda^{-1}\|F(0)\|\,.$$

Hence the estimation of the constant $\tilde{L}_{n,n+1}$ can be given as

$$\tilde{L}_{n,n+1} := L(\max\{\|u_n\|, \|u_{n+1}\|\}) \le L_f(\|F(u_n)\|_*), \tag{3.16}$$

where

$$L_f(t) := L(f(t))$$

is an increasing continuous function (since both L and f have this property), and L_f is already independent of n.

(5) Similarly, for the estimation of the constant $\tilde{L}_{n,*}$ we need to bound $\max\{\|u_n\|, \|u^*\|\}$ in terms of $\|F(u_n)\|_*$ and fixed constants. Now, using (3.10),

$$\|u^*\| \le \lambda^{-1}\|F(0)\|,$$

further, $\|u_n\|$ has the larger bound (3.15), hence the latter is also a bound for their maximum:

$$\max\{\|u_n\|, \|u^*\|\} \le \lambda^{-1}\left(\Lambda_*^{1/2}\|F(u_n)\|_* + \|F(0)\|\right).$$

Hence

$$\tilde{L}_{n,*} := L(\max\{\|u_n\|, \|u_*\|\}) \le L_g(\|F(u_n)\|_*) \tag{3.17}$$

using the increasing continuous functions

$$g(t) := \lambda^{-1}\tilde{\Lambda}_*^{1/2}t + \lambda^{-1}\|F(0)\|, \qquad L_g(t) := L(g(t)).$$

(6) Altogether, using (3.16) and (3.17), the function (3.13) can be estimated as

$$\varphi_n(t) \le (1 + L_g(t)\tilde{\Lambda}_*^{1/2}\lambda^{-2}t)\,(Q + KL_f(t)t) =: \varphi(t), \tag{3.18}$$

and accordingly, inequalities (3.14) and (3.18) result in

$$\|F(u_{n+1})\|_* \le \varphi\left(\|F(u_n)\|_*\right)\|F(u_n)\|_*\,, \tag{3.19}$$

where φ is an increasing continuous real function and is independent of n.

(7) Now, by Lemma 3.10, if the initial guess satisfies

$$\varphi(\|F(u_0)\|_*) < 1,$$

then $\lim_{n\to\infty}\|F(u_n)\|_* = 0$. In fact, using notation $r := \varphi\left(\|F(u_0)\|_*\right)$, estimate (3.19) and the

monotonicity of φ, one can derive by induction that

$$\|F(u_n)\|_* \le r^n \|F(u_0)\|_* \to 0 \qquad (\text{as } n \to \infty). \tag{3.20}$$

Here (3.8) implies $\lim\limits_{n\to\infty} \|F(u_n)\| = 0$ in the original norm too, and then, as mentioned in the beginning of the proof, (3.10) yields that $\lim_{n\to\infty} u_n = u^*$ as well.

(8) It remains to show the estimate (3.5), which means that the convergence factor can be improved to

$$Q := \frac{M-m}{M+m}$$

independently of u_0. The main point is to derive this rate for the weigthed residual errors

$$e_n := \|F(u_n)\|_*.$$

This part is similar to the corresponding part of the proof of Theorem 2.7, however, the used expressions are somewhat different.

First observe that the continuity of φ and $e_n \to 0$ imply

$$\lim_{n\to\infty} \varphi(e_n) = \varphi(0) = Q.$$

Further, by (3.19), the errors satisfy

$$e_n \le \left(\prod_{k=0}^{n-1} \varphi(e_k)\right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q}\right) Q^n e_0 \qquad (\forall n \in \mathbb{N}). \tag{3.21}$$

For all k we have $e_k \le e_0$, thus we have $L_g(e_k) \le L_g(e_0), \quad L_f(e_k) \le L_f(e_0) \quad (\forall k \in \mathbb{N})$. Using (3.18) and introducing the notations $d_1 := \tilde{\Lambda}_*^{1/2} \lambda^{-2} L_g(e_0), \quad d_2 := K L_f(e_0)$, we obtain

$$\varphi(e_k) \le (1 + d_1 e_k)(Q + d_2 e_k).$$

This and (3.20) imply

$$\frac{\varphi(e_k)}{Q} \le (1 + d_1 e_k)(1 + d_3 e_k) = 1 + d_4 e_k + d_5 e_k^2 \le 1 + d_4 e_0 r^k + d_5 e_0^2 r^{2k},$$

with constants $d_3 = \frac{d_2}{Q}, \quad d_4 = d_1 + d_3, \quad d_5 = d_1 d_3$. From here, by Lemma 3.10, we deduce the following upper estimate from (3.21):

$$e_n \le \exp\left(\frac{d_4 e_0}{1-r} + \frac{d_5 e_0^2}{1-r^2}\right) e_0 Q^n \equiv E e_0 Q^n.$$

This, together with (3.8) and (3.10), yields

$$\|u_n - u^*\| \le \lambda^{-1} \|F(u_n)\| \le \lambda^{-1} \tilde{\Lambda}_*^{1/2} \|F(u_n)\|_* =: \lambda^{-1} \tilde{\Lambda}_*^{1/2} e_n \le \lambda^{-1} \tilde{\Lambda}_*^{1/2} e_0 E Q^n, \tag{3.22}$$

hence (with constant $C := \lambda^{-1} \tilde{\Lambda}_*^{1/2} e_0 E$) we obtain (3.5). ∎

3.3 Application to power order nonlinear elliptic problems

In this subsection we apply the obtained iterative method to the finite element discretization of a strongly nonlinear elliptic problem with power order nonlinearity. Let $\Omega \subset \mathbb{R}^N$ be a bounded domain, let $p \geq 3$ and $k_1, k_2 > 0$ be given constants, $g \in L^2(\Omega)$ a given function, and consider the following boundary value problem:

$$\begin{cases} -\mathrm{div}\big((k_1 + k_2|\nabla u|^{p-2})\ \nabla u\big) &= g, \\ u_{|\partial\Omega} &= 0. \end{cases} \tag{3.23}$$

Such a nonlinear operator, which is of regularized p-Laplacian type, arises, e.g., in electrorheological fluid models, see Subsection 3.3.3. Problem (3.23) has a unique weak solution in the Sobolev space $W_0^{1,p}(\Omega)$, see, e.g., [81].

We apply the finite element method (FEM) for the discretization of the problem. Let V_h be a given FE subpace of certain continuous piecewise polynomial functions, then we look for $u \in V_h$ such that

$$\int_\Omega (k_1 + k_2|\nabla u|^{p-2})\nabla u \cdot \nabla v = \int_\Omega gv \qquad (\forall v \in V_h). \tag{3.24}$$

Our goal is to define the corresponding iterative method for this problem and to prove its convergence.

3.3.1 Construction of the iteration

First we cast the problem into the setting of Section 3.2. Our Hilbert space H will be the finite dimensional space V_h, endowed with the H_0^1 Sobolev inner product and induced norm

$$\langle u,\ v\rangle := \int_\Omega \nabla u \cdot \nabla v, \qquad \|u\|_{H_0^1} := \|\nabla u\|_{L^2},$$

respectively. Note that, owing to $p > 2$, we have $V_h \subset W_0^{1,p}(\Omega) \subset H_0^1(\Omega)$, further (since V_h is finite dimensional), $\|\nabla u\|_{L^p}$ and $\|\nabla u\|_{L^2}$ define equivalent norms on V_h, in particular, there exists a constant $\hat{c} > 0$ such that

$$\|\nabla u\|_{L^p} \leq \hat{c}\,\|\nabla u\|_{L^2} \qquad (\forall u \in V_h). \tag{3.25}$$

This shows that although the original BVP is posed in the Banach space $W_0^{1,p}(\Omega)$, the Hilbert space structure on V_h is a proper choice.

The operator $F : V_h \to V_h$, corresponding to our problem, is defined in a weak form as

$$\langle F(u), v\rangle \equiv \int_\Omega (k_1 + k_2|\nabla u|^{p-2})\nabla u \cdot \nabla v - \int_\Omega gv \qquad (\forall u, v \in V_h). \tag{3.26}$$

Then the FEM problem (3.24) is equivalent to finding $u \in V_h$ such that $\langle F(u), v\rangle = 0$ $(\forall v \in V_h)$, or simply

$$F(u) = 0 \qquad \text{in } V_h\,.$$

We want to apply the iteration, defined in Theorem 3.3, with properly chosen operators B_n that

approximate the Gâteaux derivatives $F'(u_n)$. For this, we first have to determine the operators $F'(u_n)$. Here (3.26) can be written as

$$\langle F(u), v\rangle = \int_\Omega f(\nabla u)\cdot\nabla v - \int_\Omega gv,$$

using the notation $f:\ \mathbb{R}^N \to \mathbb{R}^N$ defined by

$$f(\eta) := (k_1 + k_2|\eta|^{p-2})\eta \qquad (\eta \in \mathbb{R}^N). \tag{3.27}$$

The derivative of f at some $\eta \in \mathbb{R}^N$ is the Jacobian matrix

$$\partial_\eta f(\eta) = k_2(p-2)|\eta|^{p-4}(\eta\cdot\eta^T) + (k_1 + k_2|\eta|^{p-2})I, \tag{3.28}$$

where $\eta\cdot\eta^T$ denotes the diadic matrix with entries $\eta_i\,\eta_j$ $(i,j = 1,...,N)$. Following [28, 35], the Gâteaux derivative $F'(u)$ satisfies

$$\langle F'(u)h, v\rangle = \int_\Omega \partial_\eta f(\nabla u)\nabla h\cdot\nabla v, \tag{3.29}$$

which means that the diffusion coefficient in the operator $F'(u_n)$ is the full matrix $\partial_\eta f(\nabla u)$.

In order to significantly simplify this operator, one can propose to omit the diadic matrices, and to include only the second term in (3.28) to obtain an operator with diagonal diffusion coefficient. Therefore we introduce the operators $B_n : V_h \to V_h$ defined by the following weak forms: for given $u_n \in V_h$ in the iteration, let

$$\langle B_n h, v\rangle \equiv \int_\Omega \left(k_1 + k_2|\nabla u_n|^{p-2}\right)\nabla h\cdot\nabla v \qquad (\forall h, v \in V_h). \tag{3.30}$$

Then we obtain the following sequence from (3.4). Let $u_0 \in V_h$ be given and assume that $u_n \in V_h$ is constructed. Then u_{n+1} is found as follows:

$$\begin{cases} \text{solve} \quad B_n z_n = F(u_n), \\ \text{let} \quad u_{n+1} := u_n - \frac{2}{M+m} z_n\,. \end{cases} \tag{3.31}$$

In particular, the auxiliary equation $B_n z_n = F(u_n)$ can be written in weak form as

$$\langle B_n z_n, v\rangle = \langle F(u_n), v\rangle \qquad (\forall v \in V_h).$$

That is, introducing the linear functional

$$\ell_n v := \langle F(u_n), v\rangle \equiv \int_\Omega (k_1 + k_2|\nabla u_n|^{p-2})\nabla u_n\cdot\nabla v - \int_\Omega gv \qquad (v \in V_h),$$

the update $z_n \in V_h$ is the solution of the linear elliptic FEM problem

$$\int_\Omega \left(k_1 + k_2|\nabla u_n|^{p-2}\right)\nabla z_n\cdot\nabla v = \ell_n v \qquad (v \in V_h). \tag{3.32}$$

3.3.2 Convergence of the iteration

Proposition 3.11. *The nonlinear operator F, defined by (3.26), and the linear operators B_n, defined by* (3.30)*, satisfy the conditions of Theorem 3.3.*

Proof.

(1) The Jacobians (3.28) are symmetric, hence (3.29) is self-adjoint. First we check that F satisfies (3.1) and (3.2). As mentioned in Remark 3.4, the upper bound in (3.1) can be omitted, since it follows from (3.2). The lower bound is straightforward with $\lambda := k_1$, namely, (3.28) and (3.29) with $v = h$ yield

$$\langle F'(u)h, h\rangle = \int_\Omega k_2(p-2)|\nabla u|^{p-4}(\nabla u\cdot\nabla h)^2 + \int_\Omega (k_1 + k_2|\nabla u|^{p-2})|\nabla h|^2 \geq k_1\|h\|^2_{H^1_0}. \quad (3.33)$$

Now we have to prove the local Lipschitz property (3.2) in H^1_0-norm. Since the Gâteaux derivatives of F are symmetric for all $u \in H^1_0$, therefore $F'(u) - F'(v)$ is also symmetric, thus its operator norm can be calculated using its quadratic form:

$$\begin{aligned}\|F'(u) - F'(v)\| &= \sup_{\|h\|_{H^1_0}=1} |\langle (F'(u) - F'(v))h,\ h\rangle| \\ &= \sup_{\|h\|_{H^1_0}=1} \left|\int_\Omega (\partial_\eta f(\nabla u) - \partial_\eta f(\nabla v))\nabla h\cdot\nabla h\right|.\end{aligned} \quad (3.34)$$

To estimate the integrand, we first study the norms of the tensors $\frac{\partial^2 f(\eta)}{\partial\eta^2}$, which satisfy

$$\left\|\frac{\partial^2 f(\eta)}{\partial\eta^2}\right\| = \sup_{|h|=1}\left|\frac{\partial^2 f(\eta)}{\partial\eta^2}(h,\ h,\ h)\right| \quad (3.35)$$

owing to their symmetry [78]. Such tensors are discussed in [51] including general nonlinearities of the following form:

$$f(\eta) = a(|\eta|^2)\eta, \quad (3.36)$$

where $r \mapsto a(r)$ is a smooth scalar function. Using the notations $a'_r(r)$, $a''_r(r)$ for the first two derivatives of a, the formula in [51] for (3.36) implies

$$\frac{\partial^2 f(\eta)}{\partial\eta^2}(h,\ h,\ h) = 6a'_r(|\eta|^2)(\eta\cdot h)|h|^2 + 4a''_r(|\eta|^2)(\eta\cdot h)^3\,.$$

In our case, (3.27) is defined by the scalar nonlinearity $a(r) := k_1 + k_2 r^{\frac{p-2}{2}}$, for which we have

$$a'_r(r) = k_2\,\frac{p-2}{2}\,r^{\frac{p-4}{2}}, \quad a''_r(r) = k_2\,\frac{(p-2)(p-4)}{4}\,r^{\frac{p-6}{2}}\,.$$

Substitution and Cauchy-Schwarz inequalities yield

$$\frac{\partial^2 f(\eta)}{\partial\eta^2}(h,\ h,\ h) = 3k_2(p-2)|\eta|^{p-4}(\eta\cdot h)|h|^2 + k_2(p-2)(p-4)|\eta|^{p-6}(\eta\cdot h)^3\,,$$

$$\left|\frac{\partial^2 f(\eta)}{\partial\eta^2}(h,\ h,\ h)\right| \leq k_2(p-2)(3+|p-4|)\,|\eta|^{p-3}|h|^3 \;=:\; c_2(p)\,|\eta|^{p-3}|h|^3\,,$$

where $c_2(p) := k_2(p-2)(3+|p-4|)$. Hence (3.35) gives

$$\left\| \frac{\partial^2 f(\eta)}{\partial \eta^2} \right\| \le c_2(p) \sup_{|h|=1} |\eta|^{p-3} |h|^3 = c_2(p)\, |\eta|^{p-3}. \tag{3.37}$$

Now, with the application of the mean value theorem on the derivative function $\partial_\eta f$ in an arbitrary segment $[\eta_1,\ \eta_2]$, we get

$$\begin{aligned} \|\partial_\eta f(\eta_1) - \partial_\eta f(\eta_2)\| &\le \sup_{\tilde\eta \in [\eta_1,\ \eta_2]} \left\| \frac{\partial^2 f}{\partial \eta^2}(\tilde\eta) \right\| \cdot |\eta_1 - \eta_2| \\ &\le c_2(p) \max\{|\eta_1|^{p-3}, |\eta_2|^{p-3}\} |\eta_1 - \eta_2|. \end{aligned}$$

Combining this with (3.34), we obtain

$$\|F'(u) - F'(v)\| \le \sup_{\|h\|_{H_0^1}=1} \int_\Omega \|\partial_\eta f(\nabla u) - \partial_\eta f(\nabla v)\|\ |\nabla h|^2$$

$$\le c_2(p) \sup_{\|h\|_{H_0^1}=1} \int_\Omega \max\{|\nabla u|^{p-3}, |\nabla v|^{p-3}\} |\nabla u - \nabla v| \nabla h|^2$$

$$\le c_2(p) \sup_{\|h\|_{H_0^1}=1} \left(\int_\Omega |\nabla u|^{p-3} |\nabla u - \nabla v| |\nabla h|^2 + \int_\Omega |\nabla v|^{p-3} |\nabla u - \nabla v| |\nabla h|^2 \right).$$

In this expression we can apply Hölder's inequality of the following four-term form:

$$\int_\Omega |f|^{p-3} |g_1\, g_2\, g_3| \le \|f\|_{L^p}^{p-3} \|g_1\|_{L^p} \|g_2\|_{L^p} \|g_3\|_{L^p} \qquad (\forall f,\ g_1,\ g_2,\ g_3 \in L^p(\Omega)),$$

which yields

$$\|F'(u) - F'(v)\| \le c_2(p) \left(\|\nabla u\|_{L^p}^{p-3} + \|\nabla v\|_{L^p}^{p-3} \right) \|\nabla u - \nabla v\|_{L^p} \sup_{\|h\|_{H_0^1}=1} \|\nabla h\|_{L^p}^2 .$$

Owing to (3.25), we can apply the estimate $\|\nabla z\|_{L^p} \le \hat c\, \|z\|_{H_0^1}$ in each norm above, thus we get

$$\begin{aligned} \|F'(u) - F'(v)\| &\le c_2(p)\, \hat c^p \left(\|u\|_{H_0^1}^{p-3} + \|v\|_{H_0^1}^{p-3} \right) \|u - v\|_{H_0^1} \sup_{\|h\|=1} \|h\|_{H_0^1}^2 \\ &\le 2c_2(p)\, \hat c^p \left(\max\left\{ \|u\|_{H_0^1}, \|v\|_{H_0^1} \right\} \right)^{p-3} \|u - v\|_{H_0^1}, \end{aligned}$$

hence F' is locally Lipschitz continuous with the Lipschitz coefficient function $L : \mathbb{R}_0^+ \to \mathbb{R}_0^+$:

$$\|F'(u) - F'(v)\| \le L(\max\left\{ \|u\|_{H_0^1}, \|v\|_{H_0^1} \right\}) \|u - v\|_{H_0^1}, \qquad L(t) = c_p\, t^{p-3}$$

where $c_p := 2c_2(p)\, \hat c^p$.

(2) We prove that the operator B_n in (3.30) satisfies (3.3) with proper uniform constants M and m. Lower estimation of (3.33) gives

$$\langle F'(u_n)h, h\rangle \ge \int_\Omega \left(k_1 + k_2 |\nabla u_n|^{p-2} \right) |\nabla h|^2 = \langle B_n h, h\rangle,$$

and upper estimation of (3.33) with Cauchy-Schwarz inequality yields

$$\begin{aligned}\langle F'(u_n)h,h\rangle &\le \int_\Omega k_2(p-2)|\nabla u_n|^{p-4}|\nabla u_n|^2|\nabla h|^2 + \int_\Omega (k_1+k_2|\nabla u_n|^{p-2})|\nabla h|^2\\ &= \int_\Omega (k_1+k_2(p-1)|\nabla u_n|^{p-2})|\nabla h|^2\\ &\le (p-1)\int_\Omega (k_1+k_2|\nabla u_n|^{p-2})|\nabla h|^2 \;= (p-1)\langle B_n h,h\rangle\,.\end{aligned}$$

Hence (3.3) holds with lower and upper bounds

$$m := 1, \qquad M := p-1. \qquad \blacksquare$$

Now we can readily formulate

Theorem 3.12. *The iteration* (3.31)*, defined in Subsection 3.3.1, converges locally according to the estimate*

$$\|u_n - u^*\|_{H_0^1} \le C\Big(1-\frac{2}{p}\Big)^n \qquad (\forall n\in\mathbb{N}).$$

Proof. By Proposition 3.11, the conditions of Theorem 3.3 are satisfied, further, iteration (3.4) coincides with (3.31) in our situation. Hence (3.5) holds locally, and for the obtained bounds $m=1$ and $M=p-1$ the convergence factor is

$$\frac{M-m}{M+m} = 1-\frac{2}{p}. \qquad \blacksquare \tag{3.38}$$

Remark 3.13. Alternatively to (3.30), the preconditioner may be defined with a different constant $\tilde{k}_2>0$ instead of k_2:

$$\langle B_n z,v\rangle := \int_\Omega \Big(k_1+\tilde{k}_2|\nabla u_n|^{p-2}\Big)\,\nabla z\cdot\nabla v, \tag{3.39}$$

in order to try to balance between k_1 and k_2. A simple calculation shows that the modified bounds then become $m=\min\{1,\ \frac{k_2}{\tilde{k}_2}\}$, $M=\max\{1,\ \frac{k_2}{\tilde{k}_2}(p-1)\}$. Then it is easy to see that the estimation of the convergence factor cannot be improved in this way: that is, if $k_2\le\tilde{k}_2\le k_2(p-1)$ then we recover $\frac{M-m}{M+m}=1-\frac{2}{p}$ just as in Theorem 3.12, whereas for values of $\tilde{k}_2$ outside the interval $[k_2,\ k_2(p-1)]$ we even obtain larger convergence factors. One may expect to make a reasonable choice by either defining the constant to be in the middle of the interval $[k_2,\ k_2(p-1)]$ (that is, $\tilde{k}_2 := k_2\,\frac{p}{2}$ as a formal balance between the two endpoints) or leaving $\tilde{k}_2 := k_2$. Both choices ensure convergence with the speed as in Theorem 3.12.

3.3.3 Numerical experiments

Consider the following boundary value problem:

$$\begin{cases} -\mathrm{div}\big((\chi_1+\chi_2|\nabla u|^2)\ \nabla u\big) &= g,\\ u_{|\partial\Omega} &= 0,\end{cases}$$

where $\chi_1, \chi_2 > 0$ are given constants. Such a nonlinear operator arises, e.g., in electrorheological fluid models, see [24], where χ_1, χ_2 are susceptibility coefficients. This problem, which describes a stationary fluid, is a special case of (3.23) with $p = 4$. Our test domain is the unit square $\Omega := [0,1]^2$, and we use piecewise linear finite elements.

We apply the iteration (3.31) with preconditioning operators (3.39):

$$\langle B_n h, v\rangle \equiv \int_\Omega \left(\chi_1 + \tilde{\chi}_2 |\nabla u_n|^2\right) \nabla h \cdot \nabla v \qquad (\forall h, v \in V_h).$$

Since $p = 4$, here we let $\chi_2 \le \tilde{\chi}_2 \le 3\chi_2$ as suggested in Remark 3.13, and we obtain from (3.38) that the theoretical convergence factor is

$$\frac{M-m}{M+m} = \frac{1}{2}$$

independently of the constants χ_1, χ_2. As (3.22) and the preceding inequalities show, one has the same convergence factor for the residual errors, moreover, the ultimate estimate is a consequence of this.

We have run the iteration (3.31)–(3.32) with the following variation of parameters. A uniform mesh was used with $N = 10, 20, \ldots, 50$ node points in each direction. Since the equation can be scaled, we let $\chi_1 = 1$ and we varied χ_2 using the values $10, 100, 1000$. Similarly, we defined g as a constant with values $10, 100, 1000$. The initial guess u_0 was the solution of the Poisson equation with r.h.s. g. We measured the relative residual error

$$\varepsilon_n := \frac{\|F(u_n)\|_{H_0^1}}{\|F(u_0)\|_{H_0^1}}$$

throughout the iteration.

The results with the choice $\tilde{\chi}_2 := \chi_2$ are given in Table 3.1.

		$\chi_2 = 10$			$\chi_2 = 100$			$\chi_2 = 1000$		
	N	$g = 10$	$g = 100$	$g = 1000$	$g = 10$	$g = 100$	$g = 1000$	$g = 10$	$g = 100$	$g = 1000$
n	10	14	16	15	15	16	13	16	15	12
	20	14	16	15	15	16	13	16	15	12
	30	14	16	15	15	15	13	16	15	12
	40	14	16	15	15	15	13	16	15	12
	50	14	16	15	15	15	13	16	15	12
$\varepsilon_n\, 2^n$	10	0.837	0.984	1.052	0.835	1.047	1.014	0.984	1.052	0.934
	20	0.926	0.977	1.041	0.830	1.038	1.014	0.977	1.041	0.901
	30	0.914	0.976	1.039	0.830	1.036	1.014	0.976	1.039	0.895
	40	0.907	0.975	1.038	0.830	1.036	1.013	0.975	1.038	0.894
	50	0.904	0.975	1.038	0.830	1.035	1.013	0.975	1.038	0.895

Table 3.1: Number of iterations to achieve $\varepsilon_n < 10^{-6}$ and ratio with the expected relative residual error.

The upper part contains the number n of iterations to achieve accuracy $\varepsilon_n < 10^{-6}$. The lower part contains the values of $\varepsilon_n\, 2^n$, i.e. the ratio of ε_n with the expected relative residual

error $1/2^n$. (We have repeated the tests with $\tilde{\chi}_2 := 2\chi_2$, then we obtained very similar but slightly worse results.)

We may observe that the actual convergence follows very closely the expected theoretical error. Further, both the number of iterations and the relative residual errors behave in a robust way w.r.t. the variation of all parameters.

3.3.4 Conclusions

We have generalized the variable preconditioning quasi-Newton approach to strongly nonlinear elliptic problems and derived its convergence. Numerical tests reinforce the theoretical results, moreover, the method exhibits robust convergence w.r.t. the variation of the coefficients and the mesh size.

4 Quasi-Newton variable preconditioning under non-uniform monotonicity conditions in Banach spaces

In contrast to previous sections, where Hilbert space was used, the theoretical results are proven in this section in Banach space level, since the latter is a more natural underlying space for the corresponding problems. Non-uniform lower bounds are also allowed in the ellipticity condition, in addition to the strong upper growth. The results of this section are based on [18].

The main theoretical results are studied in Subsection 4.1. A classification of corresponding models are discussed by Subsection 4.2. Several examples are also given there for the non-uniform boundedness assumption, giving motivation for the work. Numerical results are presented in Subsection 4.3.

4.1 The abstract iterations in Banach spaces

The main theoretical results are presented in two stages. First a simpler version is developed with fixed spectral bounds and without damping, that generalizes the results of Theorem 3.3. The main point is that, in addition to the strong upper growth, non-uniform lower bounds are also allowed in the ellipticity condition. This generalization is motivated by various real models. Then a general version is presented, which generalizes Theorem 2.16 similarly.

Our results involve a complete rewriting of proofs of Sections 2-3: besides going to a Banach space setting, careful sequences of recursive estimates are needed to avoid the uniform lower bound that was exploited throughout the previous proofs.

4.1.1 The quasi-Newton method with fixed spectral bounds

Let X be a real Banach space with dual space X'. The action of a $v \in X'$ on $u \in X$ will be denoted by $\langle v, u \rangle$. The norm sign $\|\cdot\|$ will be used both in X and X', this never makes a confusion thanks to the context. We study an operator equation

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

where $F: X \to X'$ is a nonlinear operator.

Remark 4.1. *Several lemmas of previous sections hold here because of an important feature of the Banach space used in this section and its dual. Namely, proper energy norms on these spaces can be induced by inner products corresponding to a Hilbert-space structure. For example, see Corollary 4.9 below.*

The lemmas that hold for this reason and are used below: Lemma 2.10 and Lemma 3.6.

Lemma 3.10 holds trivially, however, Lemma 3.9 does not hold due to the non-uniform lower bound. Additionally, Lemmas 2.3 and 2.4 are also reformulated below for the current setting.

Assumptions 4.2. *Let X be a real Banach space and $F: X \to X'$ a nonlinear operator. Let F have a bihemicontinuous Gâteaux derivative that satisfies the following properties:*

(i) For any $u \in X$ the operator $F'(u)$ is symmetric.

(ii) There exists a continuous nonincreasing function $\lambda : \mathbb{R}_0^+ \to \mathbb{R}^+$ such that

$$\int_0^{+\infty} \lambda(t)\, dt \;=\; +\infty \tag{4.2}$$

and

$$\langle F'(u)h, h\rangle \geq \lambda(\|u\|)\, \|h\|^2 \qquad (\forall u, h \in X). \tag{4.3}$$

(iii) There exists a continuous nondecreasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|F'(u) - F'(h)\| \leq L(\max\{\|u\|,\, \|h\|\})\, \|u - h\| \qquad (\forall u,\, h \in X). \tag{4.4}$$

Denote by $u^* \in X$ the unique solution of (4.1)*. Let $M \geq m > 0$ be given constants, and for any $n \in \mathbb{N}$ let us choose a bounded symmetric linear operator $B_n : X \to X'$ such that*

$$m\langle B_n h, h\rangle \leq \langle F'(u_n)h, h\rangle \leq M\langle B_n h, h\rangle \qquad (\forall h \in X). \tag{4.5}$$

Remark 4.3. *If the function λ in (4.3) can be chosen constant, then the operator is uniformly monotone. However, we allow $\inf\limits_{t\in\mathbb{R}^+} \lambda(t) = 0$, which means non-uniform monotonicity.*

Algorithm 4.4. *With Assumptions 4.2, starting from a $u_0 \in H$, we obtain a sequence from the following formula:*

$$u_{n+1} := u_n - \frac{2}{M+m} B_n^{-1} F(u_n) \qquad (\forall n \in \mathbb{N}).$$

Theorem 4.5. *With Assumptions 4.2, the sequence generated by Algorithm 4.4 converges locally linearly to u^*, namely, there exists a neighbourhood U of u^* and for given $u_0 \in U$ there exists a constant $C > 0$ such that*

$$\|u_n - u^*\| \leq C \left(\frac{M-m}{M+m}\right)^n \qquad (\forall n \in \mathbb{N}).$$

We give some lemmas needed for the proof. The first one confirms well-posedness. The second one has been proved for operators $H \to H$ in Hilbert spaces H (see Lemmas 2.3 and 2.4), but we will see now that it extends to our Banach space setting as well. The third one is related to the special norms used, which are defined so that these fit well to our current Banach space setting.

Lemma 4.6. *Equation* (4.1) *has a unique solution $u^* \in X$.*

Proof. Condition (4.3) implies $\lambda(\|u\|)||h||^2 \leq \|F'(u)h\|\, \|h\|$ for all $h \in X$, thus

$$\|F'(u)^{-1}\| = \sup_{\|h\|\neq 0} \frac{\|h\|}{\|F'(u)h\|} \leq \frac{1}{\lambda(\|u\|)} \quad (\forall u, h \in X), \tag{4.6}$$

$$\int_0^\infty \inf_{\|u\|\leq s} \frac{1}{\|(F'(u))^{-1}\|} \mathrm{d}s \geq \int_0^\infty \inf_{\|u\|\leq s} \lambda(\|u\|) \mathrm{d}s = \int_0^\infty \lambda(s) \mathrm{d}s = \infty,$$

hence [66, Thm. 3.2] yields that $F : X \to X'$ is a homeomorphism. ∎

Lemma 4.7. *Let $A, B : X \to X'$ be bounded, symmetric, coercive linear operators, and let there exist constants $M \geq m > 0$ such that*

$$m\langle Bh, h\rangle \leq \langle Ah, h\rangle \leq M\langle Bh, h\rangle \qquad (\forall h \in X). \tag{4.7}$$

Then

$$m\langle v, A^{-1}v\rangle \leq \langle v, B^{-1}v\rangle \leq M\langle v, A^{-1}v\rangle \qquad (\forall v \in X') \tag{4.8}$$

and

$$\left\| I - \frac{2}{M+m} AB^{-1} \right\|_{A^{-1}} \leq \frac{M-m}{M+m}. \tag{4.9}$$

Proof. The main point is that we can still define an (energy) inner product on X by $\langle x, y\rangle_B := \langle Bx, y\rangle$. The symmetry and positivity of B ensures that this is indeed an inner product. Then (4.7) reads as

$$m\|h\|_B^2 \leq \langle B^{-1}Ah, h\rangle_B \leq M\|h\|_B^2 \qquad (\forall h \in X),$$

where $B^{-1}A$ already maps from X to X itself. Applying the original statement Lemma 2.3 to this situation on the space X for the operators $B^{-1}A$ and I (see Remark 4.1), we obtain

$$m\langle \left(B^{-1}A\right)^{-1}h, h\rangle_B \leq \|h\|_B^2 \leq M\langle \left(B^{-1}A\right)^{-1}h, h\rangle_B \qquad (\forall h \in X)$$

$$\text{or} \qquad m\langle A^{-1}Bh, Bh\rangle \leq \langle Bh, h\rangle \leq M\langle A^{-1}Bh, Bh\rangle \qquad (\forall h \in X).$$

Setting $v := Bh$ (which can be arbitrary in X'), we obtain (4.8). Finally, with the above idea, (4.9) follows exactly as in Lemma 2.4 (for the proof, see [52, Lemma 2.3]) since we can use the inner product generated by A^{-1} on X'. ∎

To replace the previous energy norm formulation of (2.8) with one more convenient in our setting, in the sequel, we will use the following energy norms in X':

$$\|v\|_u := \langle v,\, F'(u)^{-1}v\rangle^{1/2} \quad (\text{for given } u \in X), \quad \|\cdot\|_* := \|\cdot\|_{u^*}, \quad \|\cdot\|_n := \|\cdot\|_{u_n} \tag{4.10}$$

(for given $n \in \mathbb{N}$), where u^* is the solution of (4.1) and the u_n (for $n \in \mathbb{N}$) are from Algorithm 4.4. We let

$$\Lambda(t) := L(t)t + \|F'(0)\| \qquad (t \geq 0). \tag{4.11}$$

For fixed $u \in X$, the norms $\|\cdot\|_u$ and $\|\cdot\|$ are equivalent, namely the following lemma holds.

Lemma 4.8. *Denoting $n(u) := \frac{\lambda(\|u\|)}{\Lambda^{1/2}(\|u\|)}$, $N(u) := \frac{\Lambda(\|u\|)}{\lambda^{1/2}(\|u\|)}$, we have*

$$n(u)\|v\|_u \leq \|v\| \leq N(u)\|v\|_u \qquad (\forall v \in X').$$

Proof. First, with (4.11), note that (4.4) implies the upper bound

$$\langle F'(u)h, h\rangle \leq \|F'(u)\|\,\|h\|^2 \leq (\|F'(u) - F'(0)\| + \|F'(0)\|)\,\|h\|^2 \leq \Lambda(\|u\|)\,\|h\|^2. \tag{4.12}$$

Let $v := F'(u)h$ (which can be arbitrary in X'). Using (4.12), (4.6) and (4.3), respectively,

$$\|v\|_u^2 = \langle v,\, F'(u)^{-1}v\rangle = \langle F'(u)h, h\rangle \leq \Lambda(\|u\|)\,\|F'(u)^{-1}v\|^2 \leq \frac{\Lambda(\|u\|)}{\lambda^2(\|u\|)}\,\|v\|^2,$$

$$\|v\|_u^2 = \langle F'(u)h, h\rangle \geq \lambda(\|u\|)\,\|F'(u)^{-1}v\|^2 \geq \frac{\lambda(\|u\|)}{\|F'(u)\|^2}\,\|v\|^2 \geq \frac{\lambda(\|u\|)}{\Lambda^2(\|u\|)}\,\|v\|^2.$$ ∎

Using notations

$$\tilde{\lambda}_* := \frac{\lambda^2(\|u^*\|)}{\Lambda(\|u^*\|)}, \qquad \tilde{\Lambda}_* := \frac{\Lambda^2(\|u^*\|)}{\lambda(\|u^*\|)},$$

it follows readily with Lemma 4.7 that for any $n \in \mathbb{N}$ and $v \in X'$

$$\tilde{\lambda}_*^{1/2}\|v\|_* \le \|v\| \le \tilde{\Lambda}_*^{1/2}\|v\|_*, \qquad \|v\|_n \le \frac{\Lambda^{1/2}(\|u_n\|)}{\lambda(\|u_n\|)}\,\|v\|\,. \tag{4.13}$$

Corollary 4.9. *The norm of the Banach space $(X',\ \|\cdot\|_n)$ is induced by a scalar product corresponding to a Hilbert space and the norms $\|\cdot\|_n$ are equivalent with each other and with the original norm $\|\cdot\|$.*

Lemma 4.10. *With Assumptions 4.2, and the sequence definition of Algorithm 4.4, we obtain that the preconditioning operators in* (4.5) *have uniformly bounded inverses, namely,*

$$\|B_n^{-1}\| \le \frac{M}{\lambda(\|u_n\|)}.$$

Proof. By (4.3) and (4.5), we have

$$\lambda(\|u_n\|)\|h\|^2 \le \langle F'(u_n)h, h\rangle \le M\langle B_n h, h\rangle \le M\|B_n h\|\|h\|.$$

Dividing by $\lambda(\|u_n\|)\|h\|$ since $B_n : X \to X'$ is bijection, we obtain the desired bound. ∎

Proof of Theorem 4.5. We have to rewrite the proof of Theorem 3.3 such that all the previous uniform bounds are replaced by careful estimates in terms of the proper energy norm of the residual: $\|F(u_n)\|_*$. This will enable us to derive a suitable recurrence. An important role will be played by the strictly increasing real function

$$\lambda_1(t) := \int_0^t \lambda(s)\mathrm{d}s\,.$$

Since $\lambda_1(0) = 0$, $\lim_{t\to\infty} \lambda_1(t) = \infty$, there exists the inverse function $\lambda_1^{-1} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$. Here we have

$$\|F(u) - F(0)\| \ge \lambda_1(\|u\|) \quad (\forall u \in X),$$

since (4.3) implies

$$\begin{aligned}\lambda_1(\|u\|)\|u\| &= \int_0^{\|u\|} \lambda(s)\mathrm{d}s\|u\| = \int_0^1 \lambda(t\|u\|)\|u\|^2\mathrm{d}t \le \int_0^1 \langle F'(tu)u, u\rangle\mathrm{d}t \\ &= [\langle F(tu), u\rangle]_0^1 = \langle F(u) - F(0), u\rangle \le \|F(u) - F(0)\|\|u\|\,.\end{aligned}$$

Similarly to (2.10), a less sharp but useful estimate with λ is

$$\|F(u) - F(v)\| \ge \lambda(\max\{\|u\|, \|v\|\})\|u - v\| \quad (\forall u, v \in X). \tag{4.14}$$

Here (4.1.1) yields

$$\|u\| \le \lambda_1^{-1}(\|F(u) - F(0)\|), \tag{4.15}$$

in particular (also using (4.13))

$$\|u^*\| \le \lambda_1^{-1}(\|F(0)\|),$$

$$\|u_n\| \le \lambda_1^{-1}(\|F(u_n) - F(0)\|) \le \lambda_1^{-1}(\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_* + \|F(0)\|), \tag{4.16}$$

and altogether

$$\|u_n\| \le \max\{\|u_n\|, \|u^*\|\} \le G_*(\|F(u_n)\|_*), \tag{4.17}$$

$$\text{where} \qquad G_*(t) := \lambda_1^{-1}(\,\tilde{\Lambda}_*^{1/2} t + \|F(0)\|) \qquad (t \ge 0) \tag{4.18}$$

and G_* is a strictly increasing function on $\mathbb{R}_0^+$.

In this section, we will often use the relative Lipschitz constants on the segments of the iteration formulated exactly as before (see (3.6)): for fixed $n \in \mathbb{N}$ let

$$\tilde{L}_{n,n+1} := L(\max\{\|u_n\|, \|u_{n+1}\|\}). \tag{4.19}$$

As in the proof of Theorem 3.3, we can write the expansion

$$F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1} - u_n) + R(u_n) \tag{4.20}$$

with the remainder $R(u_n)$, furthermore (see Remark 4.1), Lemma 2.10 and Lemma 3.6 hold, allowing us to give estimations to the terms in the r. h. s. of this expression. However, Lemma 3.9 does not hold due to the non-uniform lower bound. First, with (4.19), using the Lipschitz continuity of F' on $[u_n,\, u_{n+1}]$ and Algorithm 4.4, we have

$$\|R(u_n)\| \le \frac{\tilde{L}_{n,n+1}}{2}\|u_{n+1} - u_n\|^2 \le \frac{2\tilde{L}_{n,n+1}}{(M+m)^2}\|B_n^{-1}F(u_n)\|^2. \tag{4.21}$$

Although this expression holds in the case of Theorem 3.3, the proofs diverge again here due to the non-uniform lower bound. Using Lemma 4.10, also using (4.16)–(4.18),

$$\|B_n^{-1}\| \le \frac{M}{\lambda_*(\|F(u_n)\|_*)}, \qquad \text{where} \quad \lambda_* := \lambda \circ G^* \tag{4.22}$$

with λ_* being a nonincreasing function on $\mathbb{R}_0^+$. Then

$$\|B_n^{-1}F(u_n)\| \le \frac{M}{\lambda_*(\|F(u_n)\|_*)}\,\|F(u_n)\|\,. \tag{4.23}$$

Similarly to (4.22), from (4.13), for any $v \in X'$

$$\|v\|_n \le \frac{\Lambda^{1/2}(\|u_n\|)}{\lambda(\|u_n\|)}\|v\| \le \frac{\Lambda_*^{1/2}(\|F(u_n)\|_*)}{\lambda_*(\|F(u_n)\|_*)}\|v\| \quad \text{where} \;\; \Lambda_* := \Lambda \circ G^*\,.$$

Using this, (4.21)-(4.22) and (4.13), we can estimate $R(u_n)$ with $\|F(u_n)\|_*$ as

$$\|R(u_n)\|_n \le \tilde{L}_{n,n+1}\, K^*(\|F(u_n)\|_*)\, \|F(u_n)\|_*^2, \tag{4.24}$$

$$\text{where} \quad K^*(t) := \frac{K_2^*\, \Lambda_*^{1/2}(t)}{\lambda_*^3(t)} \qquad \text{with} \quad K_2^* := \frac{2M^2\tilde{\Lambda}_*}{(M+m)^2},$$

and K^* is a strictly increasing function on $\mathbb{R}^+$.

On the other hand, since Lemma 2.10 holds (since the necessary relations hold, as expressed

in Lemma 4.7), this yields:

$$\|F(u_n) + F'(u_n)(u_{n+1} - u_n)\|_n \le \frac{M-m}{M+m}\|F(u_n)\|_n\,. \tag{4.25}$$

Thus (4.20), (4.25) and (4.24) yield

$$||F(u_{n+1})||_n \le \frac{M-m}{M+m}||F(u_n)||_n + \tilde{L}_{n,n+1}\,K^*(\|F(u_n)\|_*)\,\|F(u_n)\|_*^2. \tag{4.26}$$

To estimate the Lipschitz constants (4.19): from Algorithm 4.4, (4.17) and (4.23),

$$\max\{\|u_n\|, \|u_{n+1}\|\} \le \|u_n\| + \frac{2}{M+m}\|B_n^{-1}F(u_n)\| \le \Phi^*(\|F(u_n)\|_*),$$

where $\Phi^*(t) := G^*(t) + \frac{2M\tilde{\Lambda}_*^{1/2}t}{(M+m)\lambda_*(t)}$ is strictly increasing on $\mathbb{R}^+$, hence

$$\tilde{L}_{n,n+1} \le L(\Phi^*(\|F(u_n)\|_*))\,. \tag{4.27}$$

We need the constants in the equivalence of the norms $\|\,.\,\|_*$ and $\|\,.\,\|_n$. Using (4.3), (4.4), (4.14), (4.17) and (4.13),

$$\begin{aligned}\frac{\langle F'(u^*)h,h\rangle}{\langle F'(u_n)h,h\rangle} &\le 1 + \frac{L(\max\{\|u_n\|,\|u^*\|\})}{\lambda^2(\max\{\|u_n\|,\|u^*\|\})}\|F(u_n) - F(u^*)\| \\ &\le 1 + \frac{L(G^*(\|F(u_n)\|_*))}{\lambda^2(G^*(\|F(u_n)\|_*))}\|F(u_n)\| \le 1 + R^*(\|F(u_n)\|_*), \\ \text{where} \qquad R^*(t) &:= \frac{\tilde{\Lambda}_*^{1/2}L(G^*(t))}{\lambda^2(G^*(t))}\,t \qquad (t \ge 0)\end{aligned} \tag{4.28}$$

is strictly increasing on $\mathbb{R}^+$. The same follows by reversing roles, hence Lemma 4.7 yields

$$\frac{1}{1 + R^*(\|F(u_n)\|_*)} \le \frac{\|v\|_*^2}{\|v\|_n^2} \le 1 + R^*(\|F(u_n)\|_*) \qquad (v \in X'), \tag{4.29}$$

that is,

$$\|v\|_n \le \big(1 + R^*(\|F(u_n)\|_*)\big)^{1/2}\|v\|_*, \qquad \|v\|_* \le \big(1 + R^*(\|F(u_n)\|_*)\big)^{1/2}\|v\|_n$$

for all $v \in X'$. Combining this with (4.26)–(4.27), and using that $R^*(\|F(u_n)\|_*) \ge 0$, we obtain

$$\|F(u_{n+1})\|_* \le \varphi\left(\|F(u_n)\|_*\right)\,\|F(u_n)\|_*\,, \tag{4.30}$$

where

$$\varphi(t) := \big(1 + R^*(t)\big)\Big(\frac{M-m}{M+m} + L(\Phi^*(t))\,K^*(t)\,t\Big) \qquad (t \ge 0) \tag{4.31}$$

is a strictly increasing continuous real function on $\mathbb{R}^+$.

Having the recurrence (4.30), it now follows by Lemma 3.10 that if $\varphi\left(\|F(u_0)\|_*\right) =: r < 1$ then $\|F(u_n)\|_* \le r^n\|F(u_0)\|_* \to 0$. Since φ is continuous, hence $\varphi\left(\|F(u_0)\|_*\right) \to \varphi(0) = \frac{M-m}{M+m}$,

thus from (4.30)

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \lim \varphi(\|F(u_n)\|_*) = \frac{M-m}{M+m} =: Q.$$

Using notation $e_n := \|F(u_n)\|_*$, (4.28) implies $R^*(e_n) \le \frac{\tilde\Lambda_*^{1/2} L(G^*(e_0))}{\lambda^2(G^*(e_0))} e_n =: R_0 e_n$ and $L(\Phi^*(e_n))\, K^*(e_n) \le L(\Phi^*(e_0))\, K^*(e_0) =: P_0$, hence (4.31) implies $\varphi(e_n) \le \big(1+R_0 e_n\big)\big(Q+ P_0 e_n\big) \le \big(1+R_0 e_0 r^n\big)\big(Q+ P_0 e_0 r^n\big)$. Now, using (4.30), we can conclude by Lemma 3.10 that

$$\|F(u_n)\|_* = e_n \le \left(\prod_{k=0}^{n-1} \varphi(e_k)\right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q}\right) Q^n e_0 \le E_0 e_0\, Q^n, \quad \text{where}$$

$$\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q} \le \prod_{k=0}^{n-1} \big(1 + K_1 r^n\big)\big(1 + K_2 r^n\big) \le \exp\Big(\sum_{k=0}^{\infty} (K_1+K_2) r^n\Big) =: E_0\,.$$

Here (4.13) yields $\|F(u_n)\| \le \tilde\Lambda_*^{1/2} \|F(u_n)\|_* \le \tilde\Lambda_*^{1/2} E_0 e_0\, Q^n$. Finally, from (4.14), (4.17), (4.22) and using $\|F(u_n)\|_* \le \|F(u_0)\|_*$,

$$\|u_n - u^*\| \le \|F(u_n)\| \,/\, \lambda(\max\{\|u_n\|, \|u^*\|\}) \le C_0\, \|F(u_n)\|, \tag{4.32}$$

with $C_0 := 1/\,\lambda_*(\|F(u_0)\|_*)$, hence, letting $C := \tilde\Lambda_*^{1/2} E_0 e_0 \,/\, \lambda_*(e_0)$,

$$\|u_n - u^*\| \le C\, Q^n\,. \qquad \blacksquare$$

4.1.2 Damped quasi-Newton method with variable spectral bounds

In this subsection, the previous result is extended to a damped version that exhibits global convergence, further, variable spectral bounds are allowed which yield convergence up to quadratic order.

The proof of Theorem 2.16 is adapted here. The result below is also an extension of that theorem. However, we have to replace all the uniform bounds by proper estimates with $\|F(u_n)\|_*$.

Both proofs are given as a sequence of lemmas to highlight the differences and clarify the new contributions corresponding to this section. Furthermore, Lemmas 2.18–2.22 are coupled with Lemmas 4.14–4.17, respectively. Although the reader might find it cumbersome to check the new proofs, we intend to compare the results of Subsections 2.4 and 4.1.2 here.

Assumptions 4.11. *Let X be a real Banach space and $F : X \to X'$ a nonlinear operator. Let F have a bihemicontinuous Gâteaux derivative that satisfies conditions (i)–(iii) of Assumptions 4.2.*

Denote by $u^* \in X$ the unique solution of equation $F(u) = 0$.

Furthermore, the following conditions hold:

(iv) $M_n \ge m_n > 0$ and the symmetric linear operators $B_n : X \to X'$ satisfy

$$m_n \langle B_n h, h\rangle \le \langle F'(u_n)h, h\rangle \le M_n \langle B_n h, h\rangle \qquad (n \in \mathbb{N},\ h \in X); \tag{4.33}$$

further, using notations (4.10) *and* (4.28)*, there exist constants* $K > 1$ *and* $\varepsilon > 0$ *such that* $M_n/m_n \le 1 + 2/(\varepsilon + KR^*(\|F(u_n)\|_*))$.

(v) We define

$$\tau_n := \min\Big\{1, \frac{1-Q_n}{2\rho_n}\Big\}, \tag{4.34}$$

where $Q_n := \frac{M_n - m_n}{M_n + m_n}(1 + R^*(\|F(u_n)\|_*))$, $\rho_n := H^*(\|F(u_n)\|_*)$ *using notation* (4.42) *below.*

Algorithm 4.12. *With Assumptions 4.11, for arbitrary* $u_0 \in X$*, let* $(u_n) \subset X$ *be the sequence defined by*

$$u_{n+1} = u_n - \frac{2\tau_n}{M_n + m_n} B_n^{-1} F(u_n) \qquad (n \in \mathbb{N}). \tag{4.35}$$

Theorem 4.13. *We take Assumptions 4.11 and the sequence definition of Algorithm 4.12. Then there holds*

$$\|u_n - u^*\| \le C_0 \|F(u_n)\| \to 0 \tag{4.36}$$

with C_0 *from* (4.32)*, and with the* $*$*-norm from* (4.10)*:*

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup \frac{M_n - m_n}{M_n + m_n} < 1\,. \tag{4.37}$$

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

$$\|F(u_{n+1})\|_* \le d_1 \|F(u_n)\|_*^{1+\gamma} \qquad (n \in \mathbb{N}) \tag{4.38}$$

with some constant $d_1 > 0$.

Lemma 2.18 has to be reformulated for the new constant definitions of *(v)* of Assumption 4.11 with attention to the non-uniform bounds. In another perspective, the important expression of (2.26) for achieving the desired convergence rates can be achieved with different coefficients here, owing to the more general setting. The proof of the lemma is totally different, as it can be seen below.

Lemma 4.14. *With Assumptions 4.11, and the sequence definition of Algorithm 4.12, the following inequality holds:*

$$\|F(u_{n+1})\|_* \le \big(1 - \tau_n(1 - Q_n) + \tau_n^2 \rho_n\big)\ \|F(u_n)\|_*\,, \tag{4.39}$$

where Q_n *and* ρ_n *are as in condition (v) of Assumptions 4.11.*

Proof. Using (4.20) and (4.35), we obtain

$$F(u_{n+1}) = (1 - \tau_n)F(u_n) + \tau_n\Big(F(u_n) - \frac{2}{M_n + m_n} F'(u_n) B_n^{-1} F(u_n)\Big) + R(u_n).$$

Here (4.9) implies

$$\Big\|F(u_n) - \frac{2}{M_n + m_n} F'(u_n) B_n^{-1} F(u_n)\Big\|_n \le \frac{M_n - m_n}{M_n + m_n} \|F(u_n)\|_n,$$

hence, also using (4.29) twice,

$$\|F(u_{n+1})\|_* \le (1-\tau_n)\|F(u_n)\|_* + \tau_n\Big(1+R^*(\|F(u_n)\|_*)\Big)\tfrac{M_n-m_n}{M_n+m_n}\|F(u_n)\|_* + \|R(u_n)\|_*. \quad (4.40)$$

Here we can estimate $\|R(u_n)\|_*$ with $\|F(u_n)\|_*$ as follows. Using (4.13), (4.21), (4.35), (4.23), (4.27) and again (4.13), respectively, and since $M_n \le M_n + m_n$,

$$\begin{aligned}\|R(u_n)\|_* &\le \tilde\lambda_*^{-1/2}\|R(u_n)\| \le \frac{2\tilde L_{n,n+1}\tilde\lambda_*^{-1/2}}{(M_n+m_n)^2}\,\tau_n^2\,\|B_n^{-1}F(u_n)\|^2\\ &\le \frac{2\tilde L_{n,n+1}\tilde\lambda_*^{-1/2}M_n^2}{\lambda_*^2(\|F(u_n)\|_*)(M_n+m_n)^2}\,\tau_n^2\,\|F(u_n)\|^2\\ &\le \frac{2L(\Phi^*(\|F(u_n)\|_*))\tilde\lambda_*^{-1/2}\tilde\Lambda_*}{\lambda_*^2(\|F(u_n)\|_*)}\,\tau_n^2\,\|F(u_n)\|_*^2\,.\end{aligned}$$

$$\text{That is,}\qquad \|R(u_n)\|_* \le \tau_n^2\,H^*(\|F(u_n)\|_*)\;\|F(u_n)\|_*, \quad (4.41)$$

where

$$H^*(t) := \frac{2\tilde\lambda_*^{-1/2}\tilde\Lambda_* L(\Phi^*(t))}{\lambda_*^2(t)}\,t \qquad (t\ge 0) \quad (4.42)$$

defines a strictly increasing function on $\mathbb{R}^+$. Altogether, with definitions

$$Q_n := (1+R^*(\|F(u_n)\|_*))\,\frac{M_n-m_n}{M_n+m_n}, \qquad \rho_n := H^*(\|F(u_n)\|_*)\,,$$

by (4.40)–(4.41) we have (4.39). ∎

Lemma 2.19 holds for the current setting. Although the proof has to be updated for the new setting, it can be adapted readily from there, as we can see below.

Lemma 4.15. *With Assumptions 4.11, there exists $\tilde Q < 1$, such that*

$$Q_n \le \tilde Q \qquad (n\in\mathbb{N}).$$

Proof. We replace the symbols $\mu(u_n)$ in the proof of Lemma 2.19 with symbols $R^*(\|F(u_n)\|_*)$ to obtain the proof. ∎

Lemma 2.20 and its proof can be adapted here by replacing the formally identical results of the earlier proof with the new ones, where the symbols have different meaning, and paying attention to the effects of the non-uniform bounds. The proof is formally similar, but not identical, as we can see below.

Lemma 4.16. *We take Assumptions 4.13, and the sequence definition of Algorithm 4.13, furthermore, let us introduce the function $p:[0,1]\to\mathbb{R}$, $p(t) := 1-(1-Q_n)t+\rho_n t^2$. Then,*

$$\|F(u_{n+1})\|_* \le p(\tau_n)\|F(u_n)\|_* < \|F(u_n)\|_*\,, \quad (4.43)$$

and (4.36)-(4.37) hold.

PROOF. To conclude (4.43), we use the first paragraph of the proof of Lemma 2.20, replacing the references of the equations, namely, (2.21), (2.26), and (2.27) by (4.34), (4.39), and (4.43), respectively.

Inequalities (2.28)–(2.29) hold here. However, the derivation of the latter uses (2.30) to conclude the boundedness of ρ_n. In contrast, here, the boundedness of $\|F(u_n)\|_*$ (resulting from (4.43) instead of (2.27)), readily yields the boundedness of $\rho_n = H^*(\|F(u_n)\|_*)$, owing to the strictly increasing property of H^*. Then (4.43) implies $\|F(u_n)\|_* \leq \|F(u_0)\|_* r^n \to 0$, where $r = \max\{\frac{1+\bar{Q}}{2},\ Q'\}$. Then (4.32) yields (4.36).

The form of (2.31) is sustained here:

$$p(\tau_n) = Q_n + \rho_n \left(1 - (1-\tau_n)^2\right), \tag{4.44}$$

by (4.43), this yields (4.37). ∎

The last lemma holds for the new setting, and the proof can be readily adapted from Lemma 2.22 as we can see below.

Lemma 4.17. *We take Assumptions 4.11, and the sequence definition of Algorithm 4.12. Additionally, let* $M_n/m_n \leq 1 + c_1\|F(u_n)\|^\gamma$ *with constants* $c_1 > 0$, $0 < \gamma \leq 1$. *Then (4.38) holds with some constant* $d_1 > 0$.

PROOF. In the proof of Lemma 2.22, we replace symbol $\mu(u_n)$ with symbol $R^*(u_n)$, we do not need (2.30).

Additionally, we replace (2.27) and (2.31) with (4.43) and (4.44), respectively. ∎

Proof of Theorem 4.13. Lemmas 4.14-4.17 yield the result. ∎

Remark 4.18. If we need local convergence only, Theorem 4.13 can be simplified, namely, below (4.33), the inequality for M_n/m_n is not necessary, furthermore, *(v)* can be omitted, and we can use $\tau_n = 1$ instead. These simplified assumptions yield the local convergence.

Remark 4.19. If F' is only assumed to be Hölder continuous with exponent $0 < \nu < 1$, then the quadratic estimate of $\|R(u_n)\|_*$ in the proof is replaced by the order $1 + \nu$, hence we can only obtain convergence up to order $1 + \nu$.

4.2 Applications to elliptic problems

The ellipticity framework (4.2)–(4.4) covers problems arising in various applications, such as non-Newtonian flow models with Carreau type laws, rheology, nonlinear optics, minimal surfaces or subsonic flow models. A proper choice of variable preconditioners B_n, satisfying the given spectral equivalences, can significantly reduce the cost of the iteration. We are interested in finite element solution of these problems, hence the iterations are constructed in the considered FEM subspace.

This subsection first presents some general classes of equations and then gives a list of various concrete models where our results are applicable. The main point is that for all of these problems the convergence results of Subsection 3.2 are valid. This will also be illustrated with numerical tests at the end.

4.2.1 General nonlinearities

4.2.1.1 A class of nonlinear problems in $W^{1,p}(\Omega)$. An important wide class of elliptic problems arises in the divergence form

$$\left\{ \begin{array}{rcl} -\mathrm{div}\ f(x,\,\nabla u) & = & \omega, \\ u_{|\partial\Omega} & = & g \end{array} \right. \tag{4.45}$$

with a nonlinear vector field $f \in C^1(\Omega \times \mathbb{R}^n,\ \mathbb{R}^n)$ (that is, $f : \Omega \times \mathbb{R}^n \to \mathbb{R}^n$ and f is C^1) which has symmetric Jacobians w.r.t. η and satisfies

$$c_1 \left(k_0 + |\eta|^2\right)^{\frac{p-2}{2}} |\xi|^2 \le \frac{\partial f}{\partial \eta}(x,\eta)\,\xi \cdot \xi \le \widetilde{c}_1 \left(k_0 + |\eta|^2\right)^{\frac{p-2}{2}} |\xi|^2, \tag{4.46}$$

$$\left\| \frac{\partial f}{\partial \eta}(x,\eta_1) - \frac{\partial f}{\partial \eta}(x,\eta_2) \right\| \le d_1 \max_{\eta \in [\eta_1,\eta_2]} \left\{ \left(k_0 + |\eta|^2\right)^{\frac{p-3}{2}} \right\} |\eta_1 - \eta_2| \tag{4.47}$$

$(\forall x \in \Omega,\ \xi, \eta, \eta_1, \eta_2 \in \mathbb{R}^n)$ for some proper constants

$$1 < p < \infty, \quad \widetilde{c}_1 \ge c_1 > 0, \quad k_0 > 0\,.$$

We assume $\omega \in L^{p'}(\Omega)$ and that g has a Dirichlet lift $\widetilde{g} \in W^{1,p}(\Omega)$.

The growth condition (4.46) allows us to pose the problem in the Sobolev space $W^{1,p}(\Omega)$. This covers most of the practical situations for non-uniform diffusion coefficients, as will be illustrated by various examples in this subsection. (Nonlinearities with stronger increase or decrease, as e.g. exponential Arrhenius laws, typically arise for lower order 'reaction type' terms and are not relevant in the principal part.) The well-posedness of problem (4.45) follows from the theory of monotone operators, discussed in [72, 81], see, in particular, [81, Thm. 40.3].

We apply the finite element method (FEM) for the discretization of our problem. Related convergence results on the FEM are found, e.g., in [28, 48, 76], see, in particular, [81, Thms. 5.3.2 and 5.3.5]. One usually uses linear elements due to the low regularity. In our study of the iterative solution, we only exploit the general assumption that $V_h \subset W^{1,\infty}(\Omega) \subset W^{1,p}(\Omega)$ is a finite element subspace of conforming piecewise polynomials.

(a) Homogeneous boundary conditions. First we study problem (4.45) with $g = 0$. We consider a FEM subspace $V_h \subset W_0^{1,p}(\Omega)$ as a Banach space itself, endowed with the norm

$$\|u\|_{W_0^{1,p}} := \Big(\int_\Omega |\nabla u|^p \Big)^{1/p}.$$

Since V_h is finite-dimensional, there exist constants $\sigma_{\infty,p}$, $\sigma_{p,2} > 0$ such that

$$\frac{1}{\sigma_{\infty,p}} \|\nabla u\|_\infty \le \|u\|_{W_0^{1,p}} \le \sigma_{p,2} \|u\|_{W_0^{1,2}} \qquad (\forall u \in V_h). \tag{4.48}$$

The constants $\sigma_{\infty,p}$, $\sigma_{p,2}$ depend on h, however, they will not appear in the convergence bounds below in Corollary 4.21. The weak form of the operator, corresponding to problem (4.45), satisfies

$$\langle F(u), v \rangle = \int_\Omega f(x,\,\nabla u) \cdot \nabla v - \int_\Omega \omega v \tag{4.49}$$

with Gateaux derivative

$$\langle F'(u)h, v\rangle = \int_\Omega \frac{\partial f}{\partial \eta}(x,\, \nabla u)\nabla h \cdot \nabla v \qquad (u, h, v \in V_h). \tag{4.50}$$

Proposition 4.20. *The operator* (4.49) *satisfies conditions (i)–(iii) of Assumptions 4.2 in the space* $X := V_h$.

Proof. (i) This follows from (4.50) and the assumption that f has symmetric Jacobians w.r.t. η.

(ii) By (4.46), we have

$$\frac{\partial f}{\partial \eta}(x,\eta)\,\xi\cdot\xi \geq c_1\, \min\Big\{k_0{}^{\frac{p-2}{2}},\ \big(k_0+|\eta|^2\big)^{\frac{p-2}{2}}\Big\}\,|\xi|^2$$

where the minimum equals $k_0{}^{\frac{p-2}{2}}$ if $p \geq 2$ and equals $(k_0+|\eta|^2)^{\frac{p-2}{2}}$ if $p < 2$ since in that case the latter decreases w.r.t. $|\eta|^2$. Hence, from (4.50) and (4.48),

$$\langle F'(u)h, h\rangle \geq c_1\, \min\Big\{k_0{}^{\frac{p-2}{2}},\ \big(k_0+\|\nabla u\|_\infty^2\big)^{\frac{p-2}{2}}\Big\}\int_\Omega |\nabla h|^2 \geq \lambda(\|u\|_{W_0^{1,p}})\,\|h\|_{W_0^{1,p}}^2$$

with the nonincreasing function $\lambda(t) := \frac{c_1}{\sigma_{p,2}^2}\min\big\{k_0{}^{\frac{p-2}{2}},\ \big(k_0+\sigma_{\infty,p}^2\,t^2\big)^{\frac{p-2}{2}}\big\}$, that is, (4.3) holds. Since $\lambda(t) \geq \text{const.}\cdot t^{p-2}$ for $t \geq t_0 > 0$, where $p-2 > -1$, (4.2) is also satisfied.

(iii) From (4.50), letting $B_1 := \{h \in W_0^{1,p}(\Omega):\ \|h\|_{W_0^{1,p}} = 1\}$, we have

$$\|F'(u)-F'(v)\| \leq \sup_{h,k\in B_1}\int_\Omega \Big\|\frac{\partial f}{\partial \eta}(x,\nabla u) - \frac{\partial f}{\partial \eta}(x,\nabla v)\Big\|\ |\nabla h|\ |\nabla k|\,.$$

Using (4.47) and standard Hölder inequalities, we obtain in the case $p \geq 3$ that

$$\|F'(u)-F'(v)\| \leq \tilde{d}_1\Big(k_0+\max\{\|u\|_{W_0^{1,p}},\|v\|_{W_0^{1,p}}\}^2\Big)^{\frac{p-3}{2}}\,\|u-v\|_{W_0^{1,p}},$$

and for $p \leq 3$ (since $t \mapsto t^{\frac{p-3}{2}}$ then decreases) that

$$\|F'(u)-F'(v)\| \leq \tilde{d}_1\, k_0^{\frac{p-3}{2}}\,\|u-v\|_{W_0^{1,p}}.$$

Defining the nondecreasing function $L(t) := \tilde{d}_1\,\max\big\{k_0{}^{\frac{p-3}{2}},\ \big(k_0+t^2\big)^{\frac{p-3}{2}}\big\}$, we altogether obtain that (4.4) holds. ∎

(b) Preconditioning operators and robust convergence. A straightforward choice for preconditioning operators B_n can be defined by the formula

$$\langle B_n h, v\rangle \equiv \int_\Omega \big(k_0+|\nabla u_n|^2\big)^{\frac{p-2}{2}}\ \nabla h\cdot\nabla v \qquad (\forall h, v \in V_h). \tag{4.51}$$

Then (4.50) and condition (4.46) trivially yield

Corollary 4.21. *The operators* B_n *satisfy the spectral equivalence* (4.5) *with constants* $m = c_1$ *and* $M = \widetilde{c}_1$ *(which are independent of* h*).*

Consequently, the corresponding iteration in V_h converges as in Theorem 3.3.

Further, one may improve the above values to some variable bounds m_n, M_n depending on the actual relative spectral bounds of the Jacobian $\frac{\partial f}{\partial \eta}(x,\, \nabla u_n)$, such that Theorem 4.13 can also be applied. In addition, for special forms of f one may define more sophisticated preconditioners and derive more concrete variable bounds. This will be discussed later in Subsection 4.2.1.4.

(c) Nonhomogeneous boundary conditions. This problem can be reduced to the homogeneous case in the standard way. Then the solution of (4.45) is $u = z + \widetilde{g}$, where z solves a homogeneous problem with nonlinearity $\widetilde{f}(x,\eta) := f(x,\, \eta + \nabla\widetilde{g}(x))$. Here $\frac{\partial \widetilde{f}}{\partial \eta}$ inherits the estimates (4.46)–(4.47) from $\frac{\partial f}{\partial \eta}$ since the translation with $\nabla\widetilde{g}(x)$ is independent of η, hence our previous results hold. Then the coefficients of the operators B_n also contain translated functions $u_n = z_n + \tilde{g}$.

4.2.1.2 Scalar nonlinearities.

Problem (4.45) usually arises in the following special form:

$$\left\{ \begin{array}{rcl} -\mathrm{div}\big(a(x,\, |\nabla u|^2)\ \nabla u\big) &=& \omega, \\ u_{|\partial\Omega} &=& g, \end{array} \right. \tag{4.52}$$

i.e., the vector field $f \in C^1(\Omega \times \mathbb{R}^n,\ \mathbb{R}^n)$ is defined via a scalar nonlinearity:

$$f(x,\, \eta) := a(x,\, |\eta|^2)\ \eta \qquad \text{for } (x,\, \eta) \in \Omega \times \mathbb{R}^n \tag{4.53}$$

where $a \in C^1(\Omega \times \mathbb{R}^+,\ \mathbb{R}^+)$ is a given scalar-valued function. Such problems arise in various models such as gas dynamics, non-Newtonian flows, etc., see later in Subsections 4.2.2–4.2.3, and often a depends only on $|\eta|^2$ and not on x, yielding a so-called problem with radial structure [16]. The Jacobians that appear in the operator (4.50) are now

$$\frac{\partial f}{\partial \eta}(x,\, \nabla u) = 2\ \frac{\partial a}{\partial r}(x,\, |\nabla u|^2)\,(\nabla u \cdot \nabla u^T) + a(x,\, |\nabla u|^2)\, I\,. \tag{4.54}$$

Bounds for the Jacobians of such functions f can be reduced to bounds involving a, see e.g. [35, Remark 6.1] and [51], where we can assume that $a \in C^2$ w.r.t. r. Let

$$\underline{a}(x, r^2) := \min\Big\{ a(x, r^2),\ \frac{\partial}{\partial r}\big(a(x, r^2)\, r\big) \Big\} \tag{4.55}$$

and we define $\overline{a}$ by replacing *min* with *max* here.

We require that the following inequalities hold for some $d_2,\ d_3 > 0$

$$\underline{a}(x, r^2) \geq c_1 \left(k_0 + r^2\right)^{\frac{p-2}{2}}, \quad \overline{a}(x, r^2) \leq \tilde{c}_1 (k_0 + r^2)^{\frac{p-2}{2}}, \tag{4.56}$$

$$\frac{\partial a}{\partial r}(x, r^2) \leq d_2 (k_0 + r^2)^{\frac{p-4}{2}}, \qquad \frac{\partial^2 a}{\partial r^2}(x, r^2) \leq d_3 (k_0 + r^2)^{\frac{p-6}{2}} \tag{4.57}$$

(for all $x \in \Omega$, $r \in \mathbb{R}^+$).

Proposition 4.22. *(4.56)–(4.57) imply that the vector field* (4.53) *satisfies conditions* (4.46)–(4.47).

Proof. By the proof of [51, Prop. 2.1.], we have

$$\underline{a}(x, |\eta|^2)|\xi|^2 \leq \frac{\partial f}{\partial \eta}(x, \eta)\, \xi \cdot \xi \leq \overline{a}(x, |\eta|^2)|\xi|^2,$$

then (4.56) yields (4.46). The proof of [51, Prop. 3.1.] readily yields:

$$\left|\frac{\partial^2 f(x,\ p)}{\partial p^2}(h,\ h,\ h)\right| \leq |h|^3 \max\left\{6\left|\frac{\partial a}{\partial r}(x,|p|^2)\right||p|,\ \left|6\frac{\partial a}{\partial r}(x,|p|^2)|p| + 4\frac{\partial^2 a}{\partial r^2}(x,|p|^2)|p|^3\right|\right\},$$

here, we use (4.57) and the inequality $|p| \leq (k_0 + |p|^2)^{1/2}$ to obtain that

$$\left\|\frac{\partial^2 f(x,\ p)}{\partial p^2}\right\| \leq (6d_2 + 4d_3)\left(k_0 + |p|^2\right)^{\frac{p-3}{2}},$$

which yields (4.47). ∎

4.2.1.3 Nonlinearities with polinomial growth. Let us consider (4.45) and the assumptions below:

$$(c_1^{\rm pol} + c_2^{\rm pol}|\eta|^{p-2})|\xi|^2 \leq \partial_\eta f(x,\eta)\xi\cdot\xi \leq (\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol}|\eta|^{p-2})|\xi|^2, \quad \text{where}: p>2, \tag{4.58}$$

and

$$\left\|\frac{\partial f}{\partial \eta}(x,\eta_1) - \frac{\partial f}{\partial \eta}(x,\eta_2)\right\| \leq \left(d_1^{\rm pol} + d_2^{\rm pol}\max\{|\eta_1|,|\eta_2|\}^{p-3}\right)|\eta_1-\eta_2|, \quad \text{if } p>3, \tag{4.59}$$

$$\left\|\frac{\partial f}{\partial \eta}(x,\eta_1) - \frac{\partial f}{\partial \eta}(x,\eta_2)\right\| \leq d_2^{\rm pol}|\eta_1-\eta_2|^{p-2}, \quad \text{if } 2<p\leq 3. \tag{4.60}$$

Proposition 4.23. *The condition (4.58) implies (4.46) for $p>2$, condition (4.59) implies (4.47) for $p>3$ with the following constants:*

$$c_1 := 2^{\frac{2}{p-2}}c_2^{\rm pol}, \quad k_0 := \frac{1}{2}\left(\frac{c_1^{\rm pol}}{c_1}\right)^{\frac{2}{p-2}},$$

$$\tilde{c}_1 := (\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol})\max\left\{1,\ k_0^{-\frac{p-2}{2}}\right\}, \quad d_1 := \max\{1,\ k_0^{-\frac{p-3}{2}}\}(d_1^{\rm pol} + d_2^{\rm pol}).$$

Proof. First, we establish the lower, then the upper bounds of (4.46) for $p>2$. If η is such that $k_0 \leq |\eta|^2$,

$$c_1^{\rm pol} + c_2^{\rm pol}|\eta|^{p-2} \geq c_2^{\rm pol}|\eta|^{p-2} = 2^{\frac{2}{p-2}}c_2^{\rm pol}(2|\eta|^2)^{\frac{p-2}{2}} \geq c_1(k_0+|\eta|^2)^{\frac{p-2}{2}},$$

on the other hand, if η satisfies $0 \leq |\eta|^2 < k_0$, then

$$c_1^{\rm pol} + c_2^{\rm pol}|\eta|^{p-2} \geq c_1^{\rm pol} = (2\cdot k_0)^{\frac{p-2}{2}}c_1 \geq c_1(k_0+|\eta|^2)^{\frac{p-2}{2}}.$$

If $|\eta| \geq 1$, then

$$\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol}|\eta|^{p-2} \leq (\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol})|\eta|^{p-2} \leq (\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol})(k_0+|\eta|^2)^{\frac{p-2}{2}} \leq \tilde{c}_1(k_0+|\eta|^2)^{\frac{p-2}{2}},$$

on the other hand, if $0 \leq |\eta| < 1$, then

$$\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol}|\eta|^{p-2} \leq \tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol} \leq k_0^{-\frac{p-2}{2}}(\tilde{c}_1^{\rm pol} + \tilde{c}_2^{\rm pol})(k_0+|\eta|^2)^{\frac{p-2}{2}} \leq \tilde{c}_1(k_0+|\eta|^2)^{\frac{p-2}{2}}.$$

Inequality (4.47) can be obtained for $p > 3$ as follows. If $\max\{|\eta_1|,\ |\eta_2|\} \geq 1$, then

$$d_1^{\rm pol} + d_2^{\rm pol} \max\{|\eta_1|, |\eta_2|\}^{p-3} \leq (d_1^{\rm pol} + d_2^{\rm pol}) \max\{|\eta_1|, |\eta_2|\}^{p-3}$$
$$\leq (d_1^{\rm pol} + d_2^{\rm pol}) \left(k_0 + \max\{|\eta_1|, |\eta_2|\}^2\right)^{\frac{p-3}{2}} \leq d_1 \left(k_0 + \max\{|\eta_1|, |\eta_2|\}^2\right)^{\frac{p-3}{2}},$$

if $\max\{|\eta_1|,\ |\eta_2|\} < 1$, we have

$$d_1^{\rm pol} + d_2^{\rm pol} \max\{|\eta_1|, |\eta_2|\}^{p-3} \leq d_1^{\rm pol} + d_2^{\rm pol}$$
$$\leq k_0^{-\frac{p-3}{2}} (d_1^{\rm pol} + d_2^{\rm pol}) k_0^{\frac{p-3}{2}} \leq d_1 \left(k_0 + \max\{|\eta_1|, |\eta_2|\}^2\right)^{\frac{p-3}{2}},$$

this establishes (4.47). ∎

Remark 4.24. *Since we deal with (4.45), the corresponding operator F can be defined exactly as in (4.49). The condition (4.60) implies the Hölder continuity of F' with exponent $p-2$.*

4.2.1.4 Further choices of preconditioners. Extending the concrete choice in (4.51), the operators B_n can be defined with a general, properly chosen scalar coefficient function β via

$$\langle B_n h, v\rangle = \int_\Omega \beta(x,\ |\nabla u_n|^2)\ \nabla h \cdot \nabla v \qquad (\forall h, v \in V_h). \tag{4.61}$$

The function β shall satisfy

$$\underline{a}(x, r^2) \leq \beta(x,\ r^2) \leq \overline{a}(x, r^2), \quad \text{e.g.} \quad \beta(x,\ r^2) := \frac{1}{2}\big(\underline{a}(x,\ r^2) + \overline{a}(x,\ r^2)\big),$$

for which it is easy to see that (4.33) holds with

$$\frac{1}{m_n} = M_n = \max_{\overline{\Omega}} \frac{\overline{a}(x,|\nabla u_n|^2)}{\underline{a}(x,|\nabla u_n|^2)}.$$

For instance, by expanding the derivative in (4.55), we obtain that $\overline{a}(x, r^2) = \frac{\partial}{\partial r}\big(a(x, r^2)\, r\big) = a(x, r^2) + 2r^2 \frac{\partial a}{\partial r}(x, r^2)$, hence if $\frac{\partial a}{\partial r} \geq 0$, then β shall satisfy

$$a(x, r^2) \leq \beta(x, r^2) \leq a(x, r^2) + 2r^2 \frac{\partial a}{\partial r}(x, r^2). \tag{4.62}$$

Further, the operators (4.61) can be simplified using piecewise constant coefficients instead of β, following the process, developed in [52, Sec. 5.2] for uniformly elliptic problems.

Finally, we note that preconditioners might also be defined beyond the type (4.61) since one only needs the spectral equivalences (3.3) or (4.33). Extensions to Schwarz or multigrid preconditioners in this context may be the subject of further research.

4.2.1.5 Some extensions (a) *Hölder continuity.* If the Lipschitz condition (4.47) is replaced by Hölder continuity with some exponent $0 < \nu < 1$, then F' will also be only Hölder continuous, and by Remark 4.19 we can only obtain convergence up to order $1 + \nu$.

(b) *Mixed boundary conditions.* Problem (4.45) with mixed boundary conditions

$$u_{\,|\Gamma_{\rm D}} = g, \qquad f(x,\ \nabla u) \cdot \mathbf{n}_{\,|\Gamma_{\rm N}} = \gamma$$

can be dealt with in a similar way. Then one only homogenizes the Dirichlet boundary condition, and the operators shall be defined on a FEM subspace $V_h \subset W^{1,p}_{\rm D}(\Omega) := \{u \in W^{1,p}(\Omega) :\ u_{|\Gamma_{\rm D}} = 0\}$. In case (4.52) the Neumann condition is $a(x, |\nabla u|^2)\, \frac{\partial u}{\partial \mathbf{n}} = \gamma$, and when $\gamma = 0$ then this reduces to $\frac{\partial u}{\partial \mathbf{n}} = 0$ (since $a > 0$).

4.2.1.6 The complexity of the preconditioning operators. One can spare a significant cost with the proposed quasi-Newton approach, i.e. by replacing the full derivative operators $F'(u_n)$ with the preconditioning operators B_n. We can demonstrate this with the scalar nonlinearity (4.52), which covers most of the real-life cases. Namely, using the original Newton method, the stiffness matrix corresponding to the full derivative tensor (4.54) has entries

$$\int_\Omega \Big(2\ \frac{\partial a}{\partial r}(x,\ |\nabla u_n|^2)\, (\nabla u_n \cdot \nabla \varphi_i)\, (\nabla u_n \cdot \nabla \varphi_j) + a(x,\ |\nabla u_n|^2)\, \nabla \varphi_i \cdot \nabla \varphi_j \Big)\,.$$

In contrast, using (4.61), the arising stiffness matrix has entries

$$\int_\Omega \beta(x,\ |\nabla u_n|^2)\ \nabla \varphi_i \cdot \nabla \varphi_j \tag{4.63}$$

in which only a scalar coefficient is present. In particular, using linear FEM, the entries of the stiffness matrix (4.63) are assembled from the constant terms

$$c_{n,\,T}\ \mathrm{meas}(T)\ \nabla \varphi_i \cdot \nabla \varphi_j, \quad \text{where } \ c_{n,\,T} := \beta(x,\ |\nabla (u_{n|T})|^2)$$

(on each simplex T), which differ from the case of a Poisson equation only with the constant scaling multipliers $c_{n,\,T}$. Hence in the iteration we only have to solve properly scaled Poisson equations.

4.2.2 Some second order nonlinear model problems

Here we present a list of various concrete models that belong to the classes of problems given in Subsection 4.2.1.1 and for which our results are therefore applicable.

4.2.2.1 Regularized p-Laplacians. Let us first consider problems of the type

$$\left\{ \begin{array}{rcl} -\mathrm{div}\big((k_1 + k_2 |\nabla u|^{p-2})\ \nabla u\big) & = & \omega, \\ u_{|\partial\Omega} & = & g \end{array} \right. \tag{4.64}$$

where $p \geq 2$ and $k_1, k_2 > 0$. Such regularized forms of the p-Laplacian often arise in numerical computations, see, e.g., [14] and the references therein. Problem (4.64) was studied in Subsection 3.3 with homogeneous boundary conditions for $p \geq 3$.

Based on Subsection 4.2.1, our method applies for any $p \geq 2$ and nonhomogeneous boundary conditions are also allowed. A direct calculation yields that (4.46)–(4.47) hold for all $p > 1$. Whereas (3.37) yields local Lipschitz continuity only for $p \geq 3$ indeed, we obtain Hölder continuity by simple integration for $2 < p < 3$ with exponent $\nu = p - 2$, hence, as mentioned in item (a) of Subsection 4.2.1.1, the proper iteration converges with order $p - 1$. For $1 < p < 2$, however, Hölder continuity is violated around points where $\nabla u = 0$, such singular problems will be mentioned in Subsection 4.2.4.3.

4.2.2.2 An equation with cubic nonlinearity. A special case of equation (4.64) with $\omega = 0$ and $p = 4$ is

$$-\mathrm{div}\big((\chi_1 + \chi_2|\nabla u|^2)\ \nabla u\big) = 0, \tag{4.65}$$

where $\chi_1, \chi_2 > 0$ are given constants. This equation was used in [24] for modelling an electrorheologic flow in incompressible Newtonian fluids under an electric field E with potential u, with given boundary data g. Such quadratic susceptibility coefficients $\chi_1 + \chi_2|E|^2$ also arise in nonlinear optics when describing waves of given large frequency in a solid medium, see [22, 34]. Singular problems with $p = 4$ have also been considered in [2]. Equation (4.65) is a special case of that in (4.64), hence our results can be applied. Numerical tests will be presented in Subsection 4.3.

4.2.2.3 A non-Newtonian fluid flow model. A parallel sided slab model of stationary motion in glaciology provides the equation

$$-\mathrm{div}\Big(\frac{2}{T_0+\sqrt{T_0^2+|\nabla u|}}\ \nabla u\Big) = P,$$

see [45], where Ω is the domain occupied by the glacier and P is the hydrostatic pressure. The arising Dirichlet problem is a special case of (4.52): the coefficient is commensurable with $\big(k_0 + |\nabla u|^2\big)^{-1/4}$, hence (4.46)–(4.47) hold with $p = 3/2$.

4.2.3 Power law nonlinearities (Carreau's model)

A wide class of nonlinear elliptic problems contains so-called power law nonlinearities (or Carreau's model). We first discuss such second-order diffusion-type problems, shown to be a special case of the class (4.45). Then some other models are given with a similar structure of nonlinearity: our results from Subsection 4.2.1.1 can be adapted with obvious modifications to these situations, with product FEM subspaces $\underline{V}_h \subset W_0^{1,p}(\Omega)^d$ in the second-order vector case and with $V_h \subset W_0^{2,p}(\Omega)$ in the fourth-order case.

4.2.3.1 Power law diffusion coefficients. Second order elliptic problems with a power law nonlinearity in the diffusion coefficient are relevant in various applications, e.g., in non-Newtonian flows, glaciology, porous media and fluid-structure interactions, see [29, 33, 46, 49, 65, 76] and the references therein. The arising equations have the form

$$-\mu_0\ \mathrm{div}\Big(\big(k_0 + |\nabla u|^2\big)^{\frac{p-2}{2}}\ \nabla u\Big) = \omega \tag{4.66}$$

where $1 < p < \infty$. The variant $\mu_\infty + (\mu_0 - \mu_\infty)(1 + |\nabla u|^2)^{\frac{p-2}{2}}$ of the nonlinearity is also frequently used. Here the non-Newtonian flow is pseudoplastic if $p < 2$ and dilatant if $p > 2$. Some typical values for a pseudoplastic flow range between $p = 1.15$ and 1.86, see [38], and for porous media between 1.3 and 1.9, see [46]. For a blood flow $p = 1.36$ was measured [49]. For dilatant values, the case $p = 4$ coincides with (4.65), and $p = 3$ was used in [44, 76]. Finally we note that the value of p may be even a concentration-dependent variable exponent in some chemically reacting fluids [55], however, we restrict ourselves to the case of constant p.

Problem (4.66) is a special case of (4.45), namely, it is easy to see that (4.46)–(4.47) are satisfied since the nonlinearity in (4.66) equals the comparison function of (4.46). In particular, the bounds in (4.46) are now realized as follows. Since

$$\frac{\partial f}{\partial \eta}(\eta)\,\xi\cdot\xi = (p-2)\,(k_0+|\eta|^2)^{\frac{p-4}{2}}\,(\eta\cdot\xi)^2 + (k_0+|\eta|^2)^{\frac{p-2}{2}}|\xi|^2, \tag{4.67}$$

we obtain

$$c_p(k_0+|\eta|^2)^{\frac{p-2}{2}}|\xi|^2 \le \frac{\partial f}{\partial \eta}(\eta)\,\xi\cdot\xi \le C_p(k_0+|\eta|^2)^{\frac{p-2}{2}}|\xi|^2, \tag{4.68}$$

where $c_p := \min\{1, p-1\}$, $C_p := \max\{1, p-1\}$ and the estimation $k_0 + |\eta|^2 \ge |\eta|^2$ has been used. Accordingly, by Corollary 4.21 and Theorem 4.5, the convergence factor is uniformly bounded by

$$Q := \frac{C_p - c_p}{C_p + c_p} = \left|1 - \frac{2}{p}\right|. \tag{4.69}$$

4.2.3.2 Power-law stress tensors. Problems with a similar structure as in (4.66) arise for steady flows of incompressible homogeneous fluids, involving the rate of deformation tensor $\mathcal{E}\mathbf{u} := \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$ of the velocity vector function $\mathbf{u}$, see [48]:

$$-\mu_0\ \mathrm{div}\Big(\big(\varepsilon_0^2 + |\mathcal{E}\mathbf{u}|^2\big)^{\frac{p-2}{2}}\ \mathcal{E}\mathbf{u}\Big) = \mathbf{d},$$

where $\mathbf{d} = \mathbf{f} - \nabla P$ comes from the external force $\mathbf{f}$ and kinematic pressure P.

4.2.3.3 Fourth order problems for elastic bending. Fourth-order boundary value problems arise in the bending of elastic beams, where some materials satisfy a nonlinear power-like stress-strain law [13]. With Carreau's law, the corresponding equations in one 1D (beam) or its analogue in 2D (plate), respectively, have the form

$$\Big(\big(c_1 + |u''|^2\big)^{\frac{p-2}{2}}\ u''\Big)'' = \omega\,, \qquad \Delta\Big(\big(c_1 + |\Delta u|^2\big)^{\frac{p-2}{2}}\ \Delta u\Big) = \omega\,.$$

4.2.4 Problems with coefficients in limiting case

Finally, we discuss some problems where our theory is not directly applicable since some parameter is in a limiting case of the admissible values. Although we cannot rigorously cover these cases yet, we indicate that our method might work with some modifications in practical situations. This expectation will be confirmed in the numerical tests in Section 4.3.

4.2.4.1 Minimal surface equation. The minimal surface over a domain Ω is described by the problem

$$-\mathrm{div}\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0, \qquad u_{|\partial\Omega} = g \tag{4.70}$$

see, e.g., [28, Chap. 5.2] including FEM convergence as well. A similar equation with r.h.s. nH arises in the mean curvature problem. Vector-valued analogues of such equations appear for limiting strain models [16].

Problem (4.70) is the limiting case $p = 1$ of (4.66), for which the conditions (4.46) do not hold any more: as seen in Subsection 4.2.3.1, the lower bound $C_p := \min\{1, p-1\}$ deteriorates. If we

do not neglect the term k_0 in the intermediate estimation therein, then a weaker lower bound $(1+|\eta|^2)^{-3/2}|\xi|^2$ can be obtained, but condition (4.2) is still violated since $\int_0^{+\infty}(1+t^2)^{-3/2}\,dt = 1$. However, noting that

$$\|F(u)-F(0)\| = \|F(u)\| = \sup_{\|v\|_{W_0^{1,1}}=1} \int_\Omega \frac{\nabla u}{\sqrt{1+|\nabla u|^2}} \cdot \nabla v \le \sup_{\|v\|_{W_0^{1,1}}=1} \int_\Omega |\nabla v| = 1,$$

(4.2) can be omitted since (4.15) needs to hold only on the range of $\|F(u)-F(0)\|$.

4.2.4.2 Subsonic flow equation. The subsonic potential flow in a wind tunnel section $\Omega \subset \mathbb{R}^2$ is described as

$$\begin{cases} -\mathrm{div}\ (\varrho(|\nabla u|^2)\,\nabla u) = 0 & \text{in } \Omega \\ \varrho(|\nabla u|^2)\ \frac{\partial u}{\partial \mathbf{n}} = \gamma \quad \text{on } \Gamma_{\mathrm{N}}\,, & u = 0 \quad \text{on } \Gamma_{\mathrm{D}} \end{cases} \tag{4.71}$$

with the power type nonlinearity $\varrho(|\nabla u|^2) = \varrho_\infty \left(1 + \frac{1}{5}(M_\infty^2 - |\nabla u|^2)\right)^{5/2}$, where $M_\infty > 0$ is the Mach number at infinity, Γ_{N} consists of the sides and the end of the wind tunnel section, further, $\gamma = v_\infty > 0$ is a parameter describing relative outflow velocity at the end of the tunnel section and $\gamma = 0$ on the other subparts of Γ_{N}. This problem (see, e.g., [12]) is only elliptic as long as $|\nabla u|$ is pointwise below the subsonic limit, hence our operator cannot be defined on a whole function space. One can still expect that the proposed iteration converges properly when the solution is (and the iterative sequence runs) in the subsonic regime.

We only deal with the case involving zero Dirichlet boundary condition without the loss of generality, since u is a potential (one may observe that (4.71) only contains derivatives of u except for the constant Dirichlet boundary condition).

4.2.4.3 Singular p-Laplace problems. The p-Laplace equation, which also arises in some previously mentioned models such as non-Newtonian flows, corresponds to (4.64) with $k_1 = 0$:

$$-\mathrm{div}\big(|\nabla u|^{p-2}\ \nabla u\big) = \omega\,, \tag{4.72}$$

which becomes singular where $\nabla u = 0$, hence our theory is not valid. In particular, if $p > 2$ then ellipticity is lost, and if $p < 2$ then the coefficient may become ∞. In the latter case one may formally achieve the lower bound in (4.3): setting k_0 in (4.68) and using (4.48) and (4.50), we obtain

$$\langle F'(u)h, h\rangle \ge C_p \int_\Omega |\nabla u|^{p-2}\,|\nabla h|^2 \ge \frac{C_p}{\|\nabla u\|_\infty^{2-p}} \int_\Omega |\nabla h|^2 \ge \lambda(\|u\|_{W_0^{1,p}})\,\|h\|_{W_0^{1,p}}^2$$

where $\lambda(t) := C_p \sigma_{\infty,p}^{p-2} \sigma_{p,2}^{-2}\, t^{p-2}$. However, λ becomes unbounded around points where $\nabla u = 0$, hence one cannot formulate proper upper (Lipschitz or Hölder) bounds.

On the other hand, this may only cause a breakdown in the special case when the current FEM solution is constant on a subdomain. Excluding this case, note that the theoretical bounds in Corollary 4.21 indicate a uniform behaviour independent of the coefficients.

4.3 Numerical tests

We have solved four test problems using Courant elements on various meshes. We reduced each problem to the homogeneous one and then applied zero initial guess. The preconditioning operators were based on (4.51) or its extensions. Damping was defined adaptively and was not necessary in most runs, we only used a factor 0.7 for the p-Laplace problem.

The tests firstly demonstrate that our quasi-Newton iteration converges as expected, moreover, it performs just as well in the limiting cases mentioned in Subsection 4.2.4. Secondly, in most situations the computational cost of our overall method is less than for a full Newton iteration.

4.3.1 The test problems

We consider the following boundary value problems.

(a) Cubic nonlinearity. We have run tests for problem (4.65) on the unit square with the normed coefficient $\chi_1 = 1$ and with $\chi_2 = 3.78$ $(\mathrm{m/TV})^2$, which is a typical susceptibility coefficient (see [22, Chap. 1]). The boundary value was $g(x, y) = 1 + xy$. The preconditioning operators were defined via (4.51).

Besides varying the mesh size h, we also replaced the above value of χ_2 by 37.8 and 0.378 to see the dependence on the ratio of χ_2 and χ_1. As we will see in the tests, the iteration numbers are not sensitive to the ratio of χ_2 and χ_1. This is in accordance with the theoretical bounds in Corollary 4.21, which indicates a uniform behaviour independent of the coefficients.

(b) Minimal surface. We have solved problem (4.70) on the domain $\Omega = [-1, 1]^2$ with a cylindrical boundary function g. The maximum $M_g > 0$ of g was varied. Namely, $g(x, \pm 1) = M_g\sqrt{1 - x^2}$ and $g(\pm 1, y) = 0$ for x or $y \in [-1, 1]$. The preconditioning operators were defined via (4.51).

This equation is a special one in the set of problems, since its exponent $p = 1$ violates the condition $p > 1$ of our theoretical estimates, and this character will be reflected in the test results.

(c) Subsonic flow. We have run tests for problem (4.71) in a wind tunnel given in [12] for a range of parameters $0 < v_\infty < 1$. The preconditioning operators were based on (4.61) and (4.62):

$$\langle B_n h, v\rangle \equiv \int_\Omega \left(|\nabla u_n|^2 + \varrho'(|\nabla u_n|^2)\,|\nabla u_n|^2\right)\,\nabla h \cdot \nabla v \qquad (\forall h, v \in V_h).$$

The results showed that the iteration runs in the subsonic regime, hence the convergence properties can be directly compared to the previous tests.

(d) p-Laplacian in a glacier. We have solved the singular problem

$$-\mathrm{div}\big(|\nabla u|^{-2/3}\,\nabla u\big) = \alpha\,, \qquad u = 0 \quad \text{on } \Gamma_{\mathrm{D}}, \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_{\mathrm{N}}$$

which is a shallow ice model for the motion of a glacier in a valley [36, Chap. 10]. The domain Ω is the planar profile of the glacier, and $\alpha = A^{1/3}$, where $A = 0.2$ $\mathrm{bar}^{-3}\mathrm{y}^{-1}$ is the rate factor. The exponent $-2/3$ is a special case of Glen's law $-(n-1)/n$ with the standard constant $n = 3$. Then the operator is a p-Laplacian for $p = 4/3$, see (4.72). The preconditioning operators were the analogues of (4.51) with $k_0 = 0$, that is, the coefficients of the B_n are simply $|\nabla u_n|^{-2/3}$.

This problem is singular, since the integrand becomes unbounded around points where $\nabla u = 0$. (A latter point always occurs for any u in the function space, e.g. at least where u has a local maximum.) On the other hand, as discussed in Subsection 4.2.4.3, this fact might only cause a breakdown for the numerical iteration in the special case when the current FEM solution is constant on a subdomain. The latter is essentially excludable by the nonsymmetry of the considered real-life domains.

Moreover, the theory predicts (see Corollary 4.21 for any $p > 1$) that the iteration numbers for regularized p-Laplacians are insensitive to the ratio of k_2 and k_1, and the tests for problem (a) for $p = 4$ confirm this expectation. This suggests that a similar convergence can be expected for the singular limit case $k_1 = 0$ as well. Indeed, our tests will show such a convergence behaviour for the glacier problem.

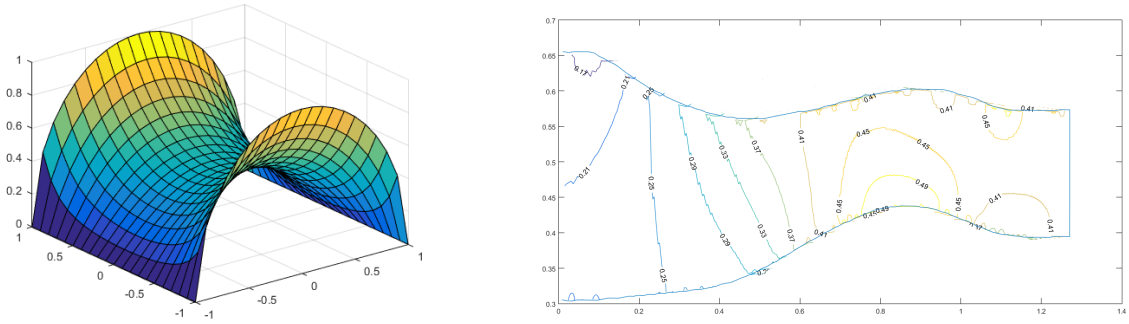

Figure 4.1: A numerical minimal surface; the contours of the modulus of velocity in the wind tunnel

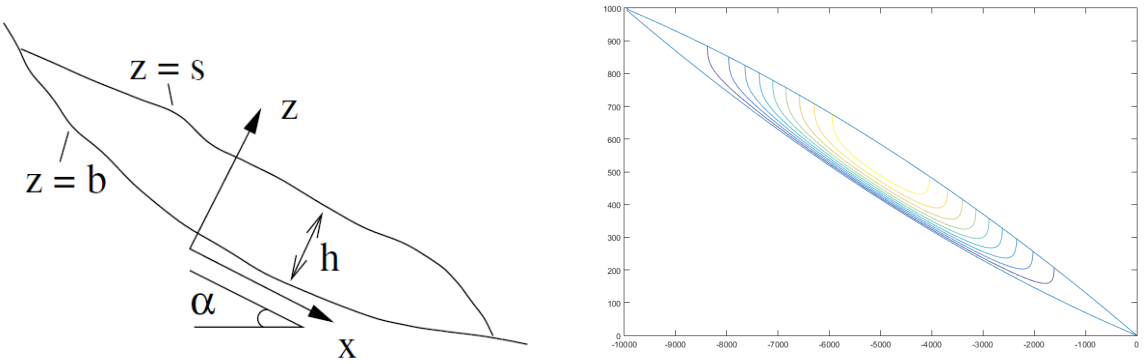

Figure 4.2: A compressed profile scheme of the glacier [36]; contours of velocity potential

As an illustration, the obtained numerical solutions for problems *(b)–(d)* are shown in Figures 4.1–4.2.

4.3.2 Outline of the main test results

The robustness of iterations is illustrated in Table 4.1. For each test problem we used up to five different meshes, denoted by $\tau_1, \ldots, \tau_5$, where the finest mesh has DoFs between 10^4 and 10^5. Namely, for problems *(a)–(b)* we created uniform grids with 2^k $(k = 4, ..., 8)$ subintervals on the sides of the square, that is, with 225, 961, 3969, 16129 and 65025 DoFs. For problems *(c)–(d)*

the meshes were generated by ANSYS in a quasi-uniform way. The DoFs were 243, 884, 3432, 53664 for the wind tunnel and 3394, 12845, 52040 for the glacier profile, where we note that the elongated geometry of the domain excludes reasonable coarse meshes for the start.

The tolerance was $\varepsilon = 10^{-4}$. The behaviour of the iterations was similar for other parameters not included in the table.

One may observe that three of the model problems produce robust behaviour (that is, the number of iterations is bounded uniformly w.r.t the mesh size h), whereas the minimal surface equation shows a slight sensitivity to h. This is in accordance with the theoretical estimates. Namely, the three 'robust' equations have an exponent $p > 1$, in which case our theory ensures a mesh independent convergence bound, as shown by Corollary 4.21. In turn, the minimal surface equation has the exponent $p = 1$, when, as shown in Subsection 4.2.4.1, the conditions (4.46) do not hold any more, so neither does Corollary 4.21.

	Cubic nonlinearity		Subsonic flow		Minimal surface		Glacier
	$\chi_2 = 3.78$	$\chi_2 = 37.8$	$v_\infty = 0.4$	$v_\infty = 0.6$	$M_g = 0.1$	$M_g = 0.25$	$A = 0.2$
τ_1	10	11	—	—	4	5	—
τ_2	10	11	4	4	5	6	—
τ_3	10	11	4	4	6	7	21
τ_4	11	12	4	4	7	7	21
τ_5	9	9	4	4	8	8	21

Table 4.1: The number of iterations for different model problems and meshes for $\varepsilon = 10^{-4}$.

The convergence history is illustrated in Table 4.2 for two test problems, here we show the sequence of residual errors in the case of the cubic nonlinearity and the minimal surface equation.

First, for the cubic nonlinearity the obtained convergence factor is around $\frac{1}{4}$. This is lower than the theoretical uniform bound $Q = \frac{C_p - c_p}{C_p + c_p} = \frac{3-1}{3+1} = \frac{1}{2}$ arising from (4.69) for $p = 4$. We note that (4.69) is pessimistic, and only estimates the first term in (4.67) by zero. If the corresponding bound $c_p = 1$ in an actual iteration takes instead some values between 1.5 and 2, then Q will be between $\frac{1}{3}$ and $\frac{1}{5}$.

Secondly, the residuals for the minimal surface equation show a sensitivity to the boundary function. In the case $M_g = 0.1$ the first few residuals decrease rather slowly, which is in accordance with the fact that there is no uniform bound in this case, due to $p = 1$. In turn, it is interesting to see that the superlinear phase of the iteration is reached. This might be related to the fact that our quasi-Newton linearized operators correspond to neglecting the first term in the Jacobians (4.67). If $p = 1$, then the neglected term is smaller and has an opposite sign compared to the previous example $p = 4$, and possibly this may result in a better approximation of the true Jacobians after fewer steps.

Comparison with a full Newton method. Finally, for the above runs we also compared the runtimes of our quasi-Newton iterations with those of the full Newton method where exact Jacobians are used, hence less but more costly steps are taken. Table 4.3 shows the ratio of runtimes of the quasi-Newton and full Newton method, denoted by t_{qN} and t_{N}, respectively.

As a conclusion, we may state the following observations:

- The iteration numbers were bounded uniformly w.r.t. h for the three test problems when

	Cubic nonlinearity		Minimal surface	
	$\chi_2 = 3.78$	$\chi_2 = 37.8$	$M_g = 0.1$	$M_g = 0.25$
1	0.25	0.25	0.695	0.5884
2	0.0625	0.0625	0.509	0.3173
3	0.0156	0.0156	0.2827	0.0866
4	0.0039	0.0039	0.1074	0.0132
5	$9.9 \cdot 10^{-4}$	$9.86 \cdot 10^{-4}$	0.0067	$9.57 \cdot 10^{-4}$
6	$2.57 \cdot 10^{-4}$	$2.53 \cdot 10^{-4}$	$4.84 \cdot 10^{-6}$	$9.33 \cdot 10^{-5}$
7	$7.32 \cdot 10^{-5}$	$7.04 \cdot 10^{-5}$	—	—

Table 4.2: The sequence of residuals for two model problems.

	Cubic nonlinearity		Subsonic flow		Minimal surface		Glacier
	$\chi_2 = 3.78$	$\chi_2 = 37.8$	$v_\infty = 0.4$	$v_\infty = 0.6$	$M_g = 0.1$	$M_g = 0.25$	$A = 0.2$
τ_1	0.497	0.610	—	—	0.972	0.713	—
τ_2	0.329	0.564	0.925	0.679	0.640	0.609	—
τ_3	0.418	0.522	0.915	0.688	0.470	0.814	0.365
τ_4	0.354	0.445	0.920	0.697	0.250	0.435	0.382
τ_5	0.162	0.390	1.284	0.883	0.183	0.236	0.426

Table 4.3: Comparison of quasi-Newton and full Newton runtimes: the values of $t_{\mathrm{qN}}/t_{\mathrm{N}}$.

$p > 1$. For the minimal surface problem (when $p = 1$ and the theory does not fully apply) there was still convergence, but with a certain increase of iteration numbers.

- The quasi-Newton steps allowed considerably simpler coding in accordance with Subsection 4.2.1.6.
- In most of the tests the total computational time of our quasi-Newton method was less than for a full Newton method, that is, the cheaper linear systems in the iteration steps were able to compensate for the larger number of iterations and yielded less overall computational work.

Altogether, the numerical tests confirm the convergence and efficiency of the method, moreover, they indicate wider applicability for problems not yet fully covered by the theory.

5 Outer-inner iterations: inexact Newton method coupled with preconditioned CG

5.1 Introduction

Besides the quasi-Newton methods seen in the previous sections, a widely used approach to solve a discretized nonlinear elliptic problem is to apply a Newton-type iterative solver with conjugate gradient method used in inner iteration. The construction of such inner-outer iterations can be found in [30, 70], their framework for uniformly monotone elliptic problems has been presented in [4, 71], see also [60, 79] for recent applications. The present section extends these methods for non-uniformly monotone elliptic operators. Although these algorithms are more compound than a quasi-Newton iteration, by varying the number of inner steps, this approach may give more flexibility and a lower number of outer linearizations.

The use of similar preconditioners for elliptic problems can be found in the previous sections in the context of quasi-Newton methods.

This section, based on [19], provides an inexact Newton method, coupled with preconditioned conjugate gradient method in inner iterations, for non-uniformly elliptic problems based on the setting of the previous sections. The preconditioners are based on spectrally equivalent operators. Additionally, the results of a numerical experiment for a subsonic flow model (see [12]) are provided as an example.

Section 5.2 contains the convergence result, Section 5.3 presents models that fall under our assumptions, while Section 5.4 shows results of the numerical experiment.

5.2 Abstract inner-outer iteration in Banach spaces

The theorems below show convergence results for inner-outer iteration in Banach space.

5.2.1 Convergence of the inexact Newton's method

The following assumptions repeat exactly the setting of Section 4, particularly Assumptions 4.2.

Assumptions 5.1. *We make the following assumptions:*

(i) Let X be a real Banach space with norm $\|\cdot\|$, and X' its dual, with usual notation $\langle v,\ u\rangle := vu$ (where $v \in X',\ u \in X$). The norm in X' is also denoted by $\|\cdot\|$.

(ii) We study operator equation

$$F(u) = 0, \tag{5.1}$$

where $F : X \to X'$ is a nonlinear operator with bihemicontinuous Gâteaux derivative. The latter is denoted by $F'(u)$ at given $u \in X$. The unique solution of equation (5.1) is denoted by u^*.

(iii) For any $u \in X$ the operator $F'(u)$ is symmetric.

(iv) There exists a continuous nonincreasing function $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ *such that*

$$\int_0^{+\infty} \lambda(t)\, dt \;=\; +\infty, \qquad \text{and}$$

$$\langle F'(u)h, h\rangle \geq \lambda(\|u\|)\, \|h\|^2 \qquad (\forall u, h \in X). \tag{5.2}$$

(v) There exists a continuous nondecreasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$ *such that*

$$\|F'(u) - F'(h)\| \leq L(\max\{\|u\|,\, \|h\|\})\, \|u - h\| \qquad (\forall u,\, h \in X).$$

Algorithm 5.2. *For arbitrary* $u_0 \in X$ *let* $(u_n) \subset X$ *be the sequence defined by*

$$u_{n+1} = u_n + \, p_n \qquad (n \in \mathbb{N}), \tag{5.3}$$

where p_n *satisfies*

$$\|F'(u_n)\, p_n + F(u_n)\|_n \leq \delta_n \|F(u_n)\|_n, \qquad (0 < \delta_n \leq \delta_0 < 1), \tag{5.4}$$

where the energy norm $\|\cdot\|_n$ *is defined below in (5.6) and*

$$\exists c_\gamma > 0,\ 0 < \gamma \leq 1, \text{ such that } \delta_n \leq c_\gamma \|F(u_n)\|_n^\gamma. \tag{5.5}$$

Theorem 5.3. *Let Assumptions 5.1 be satisfied. Then the sequence defined by Algorithm 5.2 converges locally to* u^* *with order* $(1+\gamma)$*, namely, there exists a neighbourhood* U *of* u^* *that for a given* $u_0 \in U$ *there exists constants* $C > 0$ *and* $0 < Q < 1$ *such that*

$$\|u_n - u^*\| \leq C Q^{(1+\gamma)^n} \quad (n \in \mathbb{N}).$$

Below, we collect the lemmas of Section 4 that are needed for the proof for the sake of convenience. The previous proofs for the lemmas apply here, if not shown otherwise below. This way this section can be read mostly individually.

Lemma 5.4. *Equation* (5.1) *has a unique solution* $u^* \in X$*.*

We define the following energy norms in X':

$$\begin{aligned} \|v\|_u &:= \langle v,\, F'(u)^{-1} v\rangle^{1/2} \quad (\text{for given } u \in X), \\ \|\cdot\|_* := \|\cdot\|_{u^*}\,, &\quad \|\cdot\|_n := \|\cdot\|_{u_n} \quad (\text{for given } n \in \mathbb{N})), \end{aligned} \tag{5.6}$$

and strictly increasing function $\Lambda :\ \mathbb{R}^+ \to \mathbb{R}^+,\ t \mapsto L(t)t + \|F'(0)\|$. For fixed $u \in X$, the norms $\|\cdot\|_u$ and $\|\cdot\|$ are equivalent, namely:

Lemma 5.5. *Denoting* $n(u) := \frac{\lambda(\|u\|)}{\Lambda^{1/2}(\|u\|)}$, $N(u) := \frac{\Lambda(\|u\|)}{\lambda^{1/2}(\|u\|)}$, *we have*

$$n(u)\|v\|_u \leq \|v\| \leq N(u)\|v\|_u \qquad (\forall v \in X').$$

Specifically, the following formulations can be made:

$$\tilde{\lambda}_*^{1/2}\|v\|_* \le \|v\| \le \tilde{\Lambda}_*^{1/2}\|v\|_* \quad (\forall v \in X'),$$
$$\text{where } \tilde{\lambda}_* := \frac{\lambda^2(\|u^*\|)}{\Lambda(\|u^*\|)}, \quad \tilde{\Lambda}_* := \frac{\Lambda^2(\|u^*\|)}{\lambda(\|u^*\|)} \quad \text{and} \tag{5.7}$$

$$\tilde{\lambda}_n^{1/2}\|v\|_n \le \|v\| \le \tilde{\Lambda}_n^{1/2}\|v\|_n \quad (\forall v \in X'),$$
$$\text{where } \tilde{\lambda}_n := \frac{\lambda^2(\|u_n\|)}{\Lambda(\|u_n\|)}, \quad \tilde{\Lambda}_n := \frac{\Lambda^2(\|u_n\|)}{\lambda(\|u_n\|)}. \tag{5.8}$$

Lemma 5.6. *There exists a strictly increasing function $R^* : \mathbb{R}^+ \to \mathbb{R}^+$, such that*

$$\frac{1}{1 + R^*(\|F(u_n)\|_*)} \le \frac{\|v\|_*^2}{\|v\|_n^2} \le 1 + R^*(\|F(u_n)\|_*) \qquad (v \in X').$$

The investigation of norms of elements of X in certain segments leads to the following observation.

Lemma 5.7. *There exists a nonincreasing function $\lambda_* : \mathbb{R}^+ \to \mathbb{R}^+$ such that*

$$\lambda(\|u\|) \ge \lambda_*(\|F(u)\|_*) \quad (u \in X). \tag{5.9}$$

Lemma 5.8. *The following inequality holds*

$$\|F'(u_n)^{-1}\| \le \frac{1}{\lambda_*(\|F(u_n)\|_*)}. \tag{5.10}$$

Proof. (5.2) entails:

$$\lambda(\|u_n\|)\|h\|^2 \le \langle F'(u_n)h, h\rangle \le \|F'(u_n)h\|\|h\| \quad (\forall h \in X),$$

owing to $F'(u_n)$ being a bijection, by (5.9), we have (5.10). ∎

Lemma 5.9. *There exists a strictly increasing function $\Phi^* : \mathbb{R}^+ \to \mathbb{R}^+$ such that*

$$\tilde{L}_{n,n+1} := L(\max\{\|u_n\|, \|u_{n+1}\|\}) \le L(\Phi^*(\|F(u_n)\|_*)). \tag{5.11}$$

Proof. The following result can be readily obtained from (4.13) and (4.15), similarly to (4.16). There exists a strictly increasing function $G^* : \mathbb{R}^+ \to \mathbb{R}^+$, such that

$$\|u\| \le G^*(\|F(u)\|_*) \qquad (u \in X). \tag{5.12}$$

Let us define $p_n^* := -F'(u_n)^{-1}F(u_n)$, which is the Newton step, and write expansion

$$u_{n+1} = (u_n + p_n^*) + (p_n - p_n^*), \tag{5.13}$$

where, due to (5.7), (5.10) and (5.12), the following estimation holds for the first term

$$\|u_n + p_n^*\| \le \|u_n\| + \|F'(u_n)^{-1}\|\|F(u_n)\| \le G^*(\|F(u_n)\|_*) + \frac{\tilde{\Lambda}_*^{1/2}}{\lambda_*(\|F(u_n)\|_*)}\|F(u_n)\|_*\,.$$

On the other hand, one can write

$$p_n - p_n^* = F'(u_n)^{-1}F'(u_n)(p_n - p_n^*),$$

using (5.4), (5.5), (5.8), Lemma 5.6 and (5.10) results in the following estimation of the second term of (5.13)

$$\begin{aligned}\|p_n - p_n^*\| &= \|F'(u_n)^{-1}\|\|F'(u_n)(p_n - p_n^*)\| \\ &\leq \frac{\tilde{\Lambda}_n^{1/2}}{\lambda_*(\|F(u_n)\|_*)} c_\gamma (1 + R^*(\|F(u_n)\|_*))^{\frac{1+\gamma}{2}} \|F(u_n)\|_*^{1+\gamma},\end{aligned}$$

the result follows. ∎

Lemma 5.10. *The following estimate holds for all $u, v \in X$:*

$$\|F(u) - F(v)\| \geq \lambda(\max\{\|u\|, \|v\|\})\|u - v\|,$$

in particular:

$$\|u_n - u^*\| \leq \|F(u_n)\|/\lambda(\max\{\|u_n\|, \|u^*\|\}). \tag{5.14}$$

Proof of Theorem 5.3. For given $n \in \mathbb{N}$, one can write expansion

$$\begin{gathered}F(u_{n+1}) = F(u_n) + F'(u_n)(u_{n+1} - u_n) + R(u_n), \\ \text{where } \|R(u_n)\| \leq \frac{\tilde{L}_{n,n+1}}{2}\|u_{n+1} - u_n\|^2,\end{gathered}$$

by (5.3) and (5.4) we obtain

$$\|F(u_{n+1})\|_n \leq \delta_n \|F(u_n)\|_n + \|R(u_n)\|_n,$$

applying (5.8) and (5.11) entail

$$\|F(u_{n+1})\|_n \leq \delta_n \|F(u_n)\|_n + \frac{L(\Phi^*(\|F(u_n)\|_*))}{2\tilde{\lambda}_n^{1/2}}\|p_n\|^2. \tag{5.15}$$

Here, (5.8), (5.4), Lemma 5.8 and Lemma 5.6 imply

$$\begin{aligned}\|p_n\| &\leq \tilde{\Lambda}_n^{1/2}\|F'(u_n)^{-1}\|\|F'(u_n)p_n\|_n \\ &\leq \tilde{\Lambda}_n^{1/2}\|F'(u_n)^{-1}\|(\|F(u_n) + F'(u_n)p_n\|_n + \|F(u_n)\|_n) \\ &\leq \frac{\tilde{\Lambda}_n^{1/2}}{\lambda_*(\|F(u_n)\|_*)}\|F(u_n)\|_n(1 + \delta_n) \\ &\leq \frac{\tilde{\Lambda}_n^{1/2}}{\lambda_*(\|F(u_n)\|_*)}(1 + \delta_n)(1 + R^*(\|F(u_n)\|_*))^{1/2}\|F(u_n)\|_*.\end{aligned}$$

Combining this and (5.15), then using (5.5), and applying Lemma 5.6 again

$$\begin{aligned}\|F(u_{n+1})\|_* \le (1+R^*(\|F(u_n)\|_*))^{\frac{3}{2}+\gamma}\Big(c_\gamma\|F(u_n)\|_*^{1+\gamma}+\\ +\frac{L(\Phi^*(\|F(u_n)\|_*))\tilde{\Lambda}_n}{2\tilde{\lambda}_n^{1/2}\lambda_*^2(\|F(u_n)\|_*)}(1+c_\gamma\|F(u_n)\|_*^\gamma)^2\|F(u_n)\|_*^2\Big).\end{aligned} \tag{5.16}$$

To conclude local convergence by induction, we need a slightly different approach compared to Sections 2-4, due to inner iteration. From (5.16), we obtain

$$\|F(u_{n+1})\|_* \le \varphi(\|F(u_n)\|_*)\|F(u_n)\|_*^{1+\gamma}, \tag{5.17}$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is strictly increasing function

$$\varphi(\|F(u_n)\|_*) := a(\|F(u_n)\|_*)\Big(c_\gamma + b(\|F(u_n)\|_*)(1+c_\gamma\|F(u_n)\|_*^\gamma)^2\|F(u_n)\|_*^{1-\gamma}\Big),$$

with strictly increasing and nondecreasing functions $a,\ b : \mathbb{R}^+ \to \mathbb{R}^+$, respectively:

$$a(t) := (1+R^*(t))^{\frac{3}{2}+\gamma}, \quad b(t) := \frac{L(\Phi^*(t))\tilde{\Lambda}_n}{2\tilde{\lambda}_n^{1/2}\lambda_*^2(t)}.$$

Similarly, we can construct a function $\varphi_1 : \mathbb{R}^+ \to \mathbb{R}^+$, that is also strictly increasing, and

$$\|F(u_{n+1})\|_* \le \varphi_1(\|F(u_n)\|_*)\|F(u_n)\|_*$$

can be written, namely: $\varphi_1(t) := \varphi(t)t^\gamma$.

If $\varphi_1(\|F(u_0)\|_*) < 1$, then, due to φ_1 being increasing, Lemma 3.10 yields $\|F(u_n)\|_* < \|F(u_0)\|_*$ for all $n \in \mathbb{N}$. Let $r := \varphi(\|F(u_0)\|_*)$, then (5.17) yields by induction that

$$\begin{gathered}\|F(u_n)\|_* \le r^{\frac{(1+\gamma)^n-1}{\gamma}}\|F(u_0)\|_*^{(1+\gamma)^n} \le dQ^{(1+\gamma)^n},\\ \text{where } d := r^{-1/\gamma},\ Q := r^{1/\gamma}\|F(u_0)\|_* = \varphi_1^{1/\gamma}(\|F(u_0)\|_*) < 1,\end{gathered} \tag{5.18}$$

thus $\lim\limits_{n\to\infty}\|F(u_n)\|_* = 0$, and $\lim\limits_{n\to\infty}\varphi_1(\|F(u_n)\|_*) = 0$.

Using (5.9) on (5.14), then (5.7) gives

$$\|u_n - u^*\| \le \tilde{\Lambda}_*^{1/2}\|F(u_n)\|_*/\lambda_*(\|F(u_n)\|_*).$$

Combining this with (5.18), and using definition $C := d\tilde{\Lambda}_*^{1/2}/\lambda_*(\|F(u_0)\|_*)$, we get:

$$\|u_n - u^*\| \le CQ^{(1+\gamma)^n}.$$ ∎

Remark 5.11. *One can obtain, using Section 4, the following inequalities for the convergence for $\gamma < 1$:*

$$\limsup\frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} = 0, \quad \limsup\frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*^{1+\gamma}} \le c_\gamma.$$

5.2.2 Inner-outer iteration

In what follows, the applied inner iteration is specified, i. e., for given $n \in \mathbb{N}$, the method of obtaining approximate solution $p_n \in X$ to auxiliary equation

$$F'(u_n)p_n^* = -F(u_n).$$

Let us introduce the energy inner product on X as $\langle x, y\rangle_B = \langle Bx, y\rangle$.

For fixed $n \in \mathbb{N}$, $F'(u_n)$ is a uniformly positive bounded linear symmetric operator. Let B_n be a uniformly positive bounded linear symmetric operator, for which

$$m_n\langle B_n h, h\rangle \le \langle F'(u_n)h, h\rangle \le M_n\langle B_n h, h\rangle \quad (\forall h \in X), \tag{5.19}$$

holds ($M_n \ge m_n > 0$). The following algorithm defines the sequence $\left(p_n^{(k)}\right) \subset X$, where $k \in \mathbb{N}$ is the index corresponding to the inner iteration.

Algorithm 5.12. *We can apply the preconditioned conjugate gradient method to obtain the p_n above from the sequence $\left(p_n^{(k)}\right) \subset X$. We set $\left(p_n^{(0)}\right) := 0$. Furthermore, let us denote the error and the residual error $e_n^{(k)} := p_n^{(k)} - p_n^*$, and $r_n^{(k)} := F'(u_n)e_n^{(k)} = F'(u_n)p_n^{(k)} + F(u_n)$, respectively. The step is defined as follows, where s_n denotes the conjugate directions:*

$$p_n^{(k+1)} := p_n^{(k)} + \alpha_n^{(k)} s_n^{(k)}, \quad r_n^{(k+1)} := r_n^{(k)} + \alpha_n^{(k)} z_n^{(k)}, \quad \textit{where}:$$

$$B_n z_n^{(k)} := F'(u_n)s_n^{(k)}, \quad \textit{and } \alpha_n^{(k)} := -\frac{\|r_n^{(k)}\|_{B_n}^2}{\langle F'(u_n)s_n^{(k)}, s_n^{(k)}\rangle},$$

$$s_n^{(k+1)} := r_n^{(k+1)} + \beta_n^{(k)} s_n^{(k)}, \quad \textit{where}: \beta_n^{(k)} := \frac{\|r_n^{(k+1)}\|_{B_n}^2}{\|r_n^{(k)}\|_{B_n}^2}.$$

By (5.19), we have

$$m_n\|h\|_{B_n}^2 \le \langle B_n^{-1}F'(u_n)h, h\rangle_{B_n} \le M_n\|h\|_{B_n}^2 \quad (\forall h \in X),$$

therefore, the conjugate gradient method can be applied [8] in the energy space corresponding to operator B_n.

We choose outer iteration step

$$p_n := p_n^{(k)}, \quad \text{for some } k \ge k_{n,min}, \quad \text{where } k_{n,min} = \left\lceil \frac{\ln(\delta_n/2)}{\ln(Q_n)} \right\rceil \tag{5.20}$$

is the minimum number of iterations, and

$$Q_n := \frac{\sqrt{M_n} - \sqrt{m_n}}{\sqrt{M_n} + \sqrt{m_n}}. \tag{5.21}$$

Theorem 5.13. *Let Assumptions 5.1 be satisfied. For the sequence generated by Algorithm 5.2, let us apply Algorithm 5.12 in the inner iteration in each step. If the two iterations are connected by (5.20)-(5.21), then we have*

$$\|F'(u_n)p_n^{(k)} + F(u_n)\|_n \le 2Q_n^k\|F(u_n)\|_n \quad (n, k \in \mathbb{N}) \tag{5.22}$$

and (5.4) holds for all $n \in \mathbb{N}$.

PROOF. For the conjugate gradient method in the energy space corresponding to operator B_n, the following is known

$$\frac{\|e_n^{(k)}\|_{F'(u_n)}}{\|e_n^{(0)}\|_{F'(u_n)}} \leq 2Q_n^k,$$

on the other hand

$$\|e_n^{(k)}\|^2_{F'(u_n)} = \langle F'(u_n)e_n^{(k)}, e_n^{(k)}\rangle = \langle r_n^{(k)}, F'(u_n)^{-1}r_n^{(k)}\rangle = \|r_n^{(k)}\|^2_{F'(u_n)^{-1}}.$$

Combining these and using $\|r_n^{(k)}\|_{F'(u_n)^{-1}} = \|r_n^{(k)}\|_n$ yields

$$\|r_n^{(k)}\|_n \leq 2Q_n^k\|r_n^{(0)}\|_n,$$

since $r_n^{(0)} = F(u_n)$, (5.22) follows.

Therefore, inequality

$$2Q_n^k \leq \delta_n$$

is sufficient for (5.4) to hold, and this follows from assumption (5.20). ∎

5.3 Elliptic models

By Subsection 4.2, the following boundary value problems posed in $W^{1,p}(\Omega)$ fall under Assumptions 5.1. Such BVPs arise e.g. in non-Newtonian fluids [48], bending of elastic beams [13], etc.

Firstly, as a general nonlinearity, let us consider the class of models described by

$$\left\{\begin{array}{rcl} -\text{div } f(x,\, \nabla u) & = & \omega, \\ u_{|\partial\Omega} & = & 0, \end{array}\right. \tag{5.23}$$

where $f : \Omega \times \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 nonlinear vector field. It is assumed to have symmetric Jacobians $\frac{\partial f(x,\, \eta)}{\partial \eta}$, that satisfy:

$$c_1 \left(k_0 + |\eta|^2\right)^{\frac{p-2}{2}} |\xi|^2 \leq \frac{\partial f}{\partial \eta}(x,\eta)\, \xi \cdot \xi \leq \widetilde{c}_1 \left(k_0 + |\eta|^2\right)^{\frac{p-2}{2}} |\xi|^2, \tag{5.24}$$

$$\left\| \frac{\partial f}{\partial \eta}(x,\eta_1) - \frac{\partial f}{\partial \eta}(x,\eta_2) \right\| \leq d_1 \max_{\eta\in[\eta_1,\eta_2]} \left\{ (k_0 + |\eta|^2)^{\frac{p-3}{2}} \right\} |\eta_1 - \eta_2| \tag{5.25}$$

$(\forall x \in \Omega,\ \xi, \eta, \eta_1, \eta_2 \in \mathbb{R}^n)$ for some constants $1 < p < \infty$, $\widetilde{c}_1 \geq c_1 > 0$, $k_0 > 0$, and we assume $\omega \in L^{p'}(\Omega)$.

One can also use mixed boundary conditions in (5.23)

$$u_{\,|\Gamma_{\rm D}} = 0, \qquad f(x,\, \nabla u) \cdot \mathbf{n}_{\,|\Gamma_{\rm N}} = \gamma, \tag{5.26}$$

where $\Gamma_{\rm D} \cup \Gamma_{\rm N} = \partial\Omega$. Then the solution is looked for in the subspace $\{u \in W^{1,p}(\Omega) :\ u_{\,|\Gamma_{\rm D}} = 0\}$.

In particular, we may have a given scalar nonlinearity (4.53), where $a : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$ is a C^1 scalar-valued function. Then problem (5.23) with mixed boundary conditions (5.26)

becomes

$$\begin{cases} -\mathrm{div}\big(a(x,\,|\nabla u|^2)\;\nabla u\big) &=& \omega, \quad \text{in } \Omega\,, \\ u &=& 0, \quad \text{on } \Gamma_{\mathrm{D}}\,, \\ a(x,|\nabla u|^2)\,\frac{\partial u}{\partial \mathbf{n}} &=& \gamma, \quad \text{on } \Gamma_{\mathrm{N}}\,. \end{cases}$$

If we assume that for all $(x,\ r) \in \Omega \times \mathbb{R}^+$, $a(x, r^2) \in C^2$ w.r.t. r, furthermore, (4.56)–(4.57) hold, then the vector field (4.53) satisfies (5.24)–(5.25).

In accordance with Subsection 4.2, the definition of operators B_n can be given with scalar coefficients

$$\langle B_n h, v\rangle = \int_\Omega \beta(x,\,|\nabla u_n|^2)\;\nabla h \cdot \nabla v \qquad (\forall h, v \in V_h).$$

If the function β satisfies

$$\underline{a}(x, r^2) \le \beta(x,\,r^2) \le \overline{a}(x, r^2), \quad \text{e.g.} \quad \beta(x,\,r^2) := \frac{1}{2}\big(\underline{a}(x,\,r^2) + \overline{a}(x,\,r^2)\big)\,,$$

then, for all n, spectral equivalence of B_n and $F'(u_n)$ can be obtained readily, in other words, (5.19) holds.

The Newton equation must be discretized which adds to the inexactness. This can be based on a combination of a coarse and a fine mesh which can save computer time (see [7, 10, 80], etc.). Future work in the topic might include corresponding investigations.

5.4 Numerical experiment

5.4.1 Subsonic flow example

The following boundary value problem describing potential flow in a wind tunnel section $\Omega \subset \mathbb{R}^2$ has been presented in (4.71) in Subsection 4.2.4.2. The scalar nonlinearity is $\varrho(|\nabla u|^2) = \varrho_\infty \left(1 + \frac{1}{5}(M_\infty^2 - |\nabla u|^2)\right)^{5/2}$, where $\varrho_\infty > 0$ is the air density at infinity, and u is the velocity potential. $M_\infty > 0$ denotes the Mach number at infinity, Γ_{D} stands for the Dirichlet boundary, while $\Gamma_{\mathrm{N}} := \partial\Omega \setminus \Gamma_{\mathrm{D}}$ is the Neumann boundary. The range of γ is $\{0,\ v_\infty\}$, where $v_\infty > 0$ is a parameter describing outflow velocity.

By involving problem (4.71), our goal is to test that our method may work even beyond the limitations posed by our previous theoretical assumptions. Here, the operator cannot be defined on a whole function space. However, as addressed in Section 4, one can expect that the method of present section converges properly while the solution and the utilized part of the iterative sequence satisfy the subsonic limit condition.

We apply the results of Section 5.2, and use the finite element method (FEM) for the discretization of the problem, namely, Courant elements. Hence the Banach space X is a finite dimensional space consisting of piecewise linear functions for which $u_{|\Gamma_{\mathrm{D}}} = 0$. Therefore, all norms are equivalent.

Thus one can define the operator describing (4.71) in weak form as

$$\langle F(u),\ v\rangle \equiv \int_\Omega \rho(|\nabla u|^2)\nabla u \cdot \nabla v - \int_{\Gamma_{\mathrm{N}}} \gamma v \quad (\forall u,\ v \in V_h).$$

Consequently, the FEM problem becomes the task of finding $u \in V_h$, such that

$$\langle F(u),\ v\rangle = 0\ (\forall v \in V_h).$$

This can be written shortly in the form of (5.1) as

$$F(u) = 0 \quad \text{in } V_h.$$

The Gâteaux derivative of the operator can be obtained readily in weak form

$$\langle F'(u)h,\ v\rangle = \int_\Omega \left(\rho(|\nabla u|^2)\nabla h\cdot\nabla v + 2\rho'(|\nabla u|^2)(\nabla u\cdot\nabla h)(\nabla u\cdot\nabla v)\right) \quad (\forall h,\ v \in V_h).$$

The applied preconditioner in the nth outer step is

$$\langle B_n h,\ v\rangle = \int_\Omega \left(\rho(|\nabla u_n|^2) + \rho'(|\nabla u_n|^2)|\nabla u_n|^2\right)\nabla h\cdot\nabla v \quad (\forall h,\ v \in V_h).$$

It provides a substantial simplification of the Gâteaux derivative of the operator.

5.4.2 Numerical results

The results of the above experiment with five different meshes and two different v_∞ values are presented below.

Let symbol DoF stand for the degrees of freedom of the FEM model. Denote $n_1,\ n_2$ the number of outer iteration steps necessary to achieve smaller relative residual error than 10^{-4} and 10^{-6}, respectively. Let k denote the number of inner iteration steps required for relative residual error to be smaller than 10^{-4} for an individual outer iteration step. The numerical results are summarized in Table 5.1, where the outer step number, which the given value of k corresponds to, can be identified in the header.

	$\tilde{v}_\infty = 0.4$						$\tilde{v}_\infty = 0.6$						
DoF	n_1	n_2	k				n_1	n_2	k				
			1	2	3	4			1	2	3	4	5
243	3	4	1	3	3	3	4	5	1	3	4	4	4
884	3	4	1	3	3	3	4	5	1	3	3	4	4
3432	3	4	1	2	3	3	4	5	1	3	4	4	4
13520	3	4	1	2	3	3	4	5	1	3	4	4	4
53664	3	4	1	2	3	3	4	5	1	3	4	4	4

Table 5.1: The number of required outer (n_1, n_2) and inner (k) iteration steps.

As a conclusion, we state the following observations. Firstly, we readily find robustness of the method developed here for the subsonic flow example. Robustness for a quasi-Newton method with the same preconditioner was found in Section 4.

Secondly, (5.20) states for the subsonic model that at most 8 inner iterations are sufficient to reach relative tolerance 10^{-4}. Comparing this to Table 5.1 shows that in fact far less iterations can be sufficient as well.

6 Robust iterative solvers for nonlinear Gao beam models in elasticity

6.1 Introduction

The numerical study of the deformation of thin beams and plates is a widespread problem in elasticity theory and engineering practice, since such elastic structures regularly appear in several real applications, see, e.g., [3, 47, 69, 74, 75]. These models generally lead to fourth order equations. The linear models, which are basic in engineering applications, can be used only for small deformations. The most popular linear beam model is the Euler–Bernoulli beam: if the elastic modulus E and the moment of inertia I are constant, then one obtains a fourth order ODE with constant coefficient. The analogous two-dimensional model for a thin plate involves the biharmonic operator. The obtained simple models read as

$$Du^{IV} = q \qquad \text{and} \qquad D\Delta^2 u = q \tag{6.1}$$

where $D = EI$ is the flexural rigidity and q is the distributed load. However, such linear models are no more valid unless the deformation can be considered as infinitesimal. In more realistic models of engineering structures involving larger deformations, nonlinear behaviour has to be taken into account. Some of these models have been introduced in [39, 40], leading to so-called Gao beam models, see [42, 59] for more recent applications. Gao beam theory respects the Euler-Bernoulli hypothesis, but involves a geometrical type of nonlinearity [63, 77]. The involved nonlinearity requires the application of efficient solvers for the nonlinear system arising for the finite element approximation (FEM).

This section presents the results of [20]. The goal here is to present various types of such iterative solvers in the setting of the finite element method (FEM), and, in particular, to show the robust behaviour of these methods, i.e. convergence independently of the mesh parameters. The presentation of these methods, based on a Hilbert space framework, includes the proper forms of the Sobolev gradient iteration and of Newton's method adapted to the given beam problem. Further, based on Section 2, a quasi-Newton/variable preconditioning method is presented as an intermediate version between the above two. Here the balance between speed and cost is achieved with auxiliary operators chosen via spectral equivalence.

The paper first summarizes the corresponding theory in Subsection 6.2. Subsection 6.3 contains the main results. After the weak formulation of the problem, the necessary properties of the corresponding operators are proved in the used finite element subspaces, and robust convergence is derived. Then the results of numerical experiments are presented. A brief conclusion is given in Subsection 6.4.

6.2 Preliminaries

6.2.1 Nonlinear Gao beam models

For the description of the bending of a beam resting on a foundation, to treat deformations beyond the infinitesimal region, models have been developed in [40, 42] derived from the presence of lateral stress. The following model considers a beam with classical Winkler foundation,

which is a widespread concept in civil engineering, also with a profound effect on the field of adhesion mechanics, see [32]. Here the deflection u of the beam is described by the following equation:

$$EIu^{IV} - E\alpha(u')^2u'' + k_F u = f \qquad \text{in } J := [0, b] \tag{6.2}$$

with the following constants: $E > 0$ is the elastic modulus, $I > 0$ is the moment of inertia for the beam's cross-section, $\alpha = 3h(1 - \nu^2)$ where h is thickness measured from the axis and $\nu > 0$ denotes the Poisson ratio, and $k_F > 0$ is the foundation stiffness coefficient. Further, the transverse distributed load function is denoted by q, and $f = (1 - \nu^2)q$.

A slightly modified version of the above equation, involved e.g. in contact problems, is

$$EIu^{IV} - E\alpha(u')^2u'' + P\mu u'' = f \qquad \text{in } J := [0, b]\,, \tag{6.3}$$

where P is the axial force, assumed to be below the Euler critical load P^E_{cr}, and $\mu = (1+\nu)(1-\nu^2)$.

6.2.2 Some iterative methods

Here we summarize some iterative methods in a Hilbert space framework. Since we will apply them to our given beam equation, we formulate these theorems under a common set of conditions that can be verified for the beam problem. For more details we refer to the monograph [35] and Section 2. In this section let H be a real Hilbert space and $F : H \to H$ a nonlinear operator which has a bihemicontinuous Gâteaux derivative. We formulate the following properties for F':

Assumptions 6.2.2 We formulate the following properties for F'.

(i) (Symmetry.) For any $u \in H$ the operator $F'(u)$ is self-adjoint.

(ii) (Regularity.) There exists a constant $\lambda > 0$ such that

$$\lambda\|h\|^2 \le \langle F'(u)h, h\rangle \qquad (\forall u, h \in H). \tag{6.4}$$

(iii) (Upper growth.) There exists a continuous increasing function $\Lambda : \mathbb{R}^+ \to \mathbb{R}^+$ such that

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

(iv) (Local Lipschitz.) There exists a continuous increasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|F'(u) - F'(v)\| \le L(\max\{\|u\|, \|v\|\})\,\|u - v\| \qquad (\forall u,\, h \in H).$$

We wish to solve the operator equation

$$F(u) = 0.$$

Conditions (i)-(ii) imply that there is a unique solution $u_* \in H$, see, e.g., [35]. We study three types of iterative methods.

Theorem 6.1. (Simple iteration/gradient method.) *Let Assumptions 6.2.2 (i)–(iii) hold. Let $u_0 \in H$ be arbitrary and $\Lambda_0 := \Lambda\left(\|u_0\| + \frac{1}{\lambda}\|F(u_0)\|\right)$. Then the sequence, defined by*

$$u_{n+1} := u_n - \frac{2}{\Lambda_0 + \lambda} F(u_n) \qquad (\forall n \in \mathbb{N})$$

converges to u_*, namely,

$$\|u_n - u_*\| \le \frac{1}{\lambda}\|F(u_0)\| \left(\frac{\Lambda_0 - \lambda}{\Lambda_0 + \lambda}\right)^n \qquad (\forall n \in \mathbb{N})\,.$$

Moreover, in fact $\|u_n - u_*\| \le \frac{1}{\lambda}\|F(u_n)\|$ and

$$\frac{\|F(u_{n+1})\|}{\|F(u_n)\|} \le \frac{\Lambda_0 - \lambda}{\Lambda_0 + \lambda}\,.$$

Proof. See [35, Thm. 5.5]. ∎

Theorem 6.2. (Newton's method.) *Let Assumptions 6.2.2 (i), (ii) and (iv) hold. Let u_0 be in a properly small neighbourhood of u_* and $L_0 := L\left(\|u_0\| + \frac{2}{\lambda}\|F(u_0)\|\right)$. Then the sequence, defined by*

$$u_{n+1} := u_n - F'(u_n)^{-1}F(u_n) \qquad (\forall n \in \mathbb{N}),$$

converges to u_*, that is, $\|u_n - u_*\| \le \frac{1}{\lambda}\|F(u_n)\| \to 0$ and

$$\|F(u_{n+1})\| \le \frac{L_0}{2\lambda^2}\|F(u_n)\|^2 \qquad (n \in \mathbb{N}).$$

Proof. This follows from [35, Thm. 5.9 and Remark 5.17]. ∎

The formulation of the third theorem uses the energy $*$-norm $\|v\|_* := \langle F'(u^*)^{-1}v,\, v\rangle^{1/2}$.

Theorem 6.3. (Quasi-Newton/Variable Preconditioning Method.) *Let Assumptions 6.2.2 (i)–(iv) hold. Let u_0 be in a properly small neighbourhood of u_*. Let the sequence (u_n) be defined by*

$$u_{n+1} := u_n - \frac{2}{M_n + m_n} B_n^{-1}F(u_n) \qquad (n \in \mathbb{N}),$$

where $0 < m_n \le M_n$ and the symmetric linear operators $B_n : H \to H$ satisfy

$$m_n\langle B_n h, h\rangle \le \langle F'(u_n)h, h\rangle \le M_n\langle B_n h, h\rangle \qquad (n \in \mathbb{N},\, h \in H). \tag{6.5}$$

We require (m_n) to be positively bounded from below and (M_n) bounded from above. Then (u_n) converges to u_*, that is, $\|u_n - u_*\| \le \frac{1}{\lambda}\|F(u_n)\| \to 0$ and

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup \frac{M_n - m_n}{M_n + m_n} < 1. \tag{6.6}$$

Proof. This follows from Theorem 3.3 and Remark 4.18. ∎

We note that a reasonable requirement is to outperform the simple iteration, for which one should choose m_n and M_n closer than the original spectral bounds of $F'(u_n)$.

Remark 6.4. (i) (Efficiency.) Obviously, as is well-known, Newton's method is the fastest and the simple iteration is the slowest of the above methods (quadratic speed vs linear speed), but Newton's method is more costly. The quasi-Newton method is an intermediate version where the variable preconditioners B_n may be cheaper than the full linearized operators, still its convergence can be superlinear when the lim sup in (6.6) equals 0.

(ii) (Damped versions, global convergence.) The local convergence of the above versions of the Newton and quasi-Newton methods can be made global by damping with proper parameters $\tau_n \leq 1$. See, e.g., Theorem 4.13.

6.2.3 Estimates in Sobolev spaces

Here let $\Omega \subset \mathbb{R}^d$ be a bounded domain. In most of this section we will let $d = 1$ and $\Omega = J$ (the studied interval). The following statements, which are well-known (see, e.g., [1]), will be useful later:

Proposition 6.5. (Generalized Hölder inequality.) *If the numbers $p_i \geq 1$ satisfy $\sum_{i=1}^s \frac{1}{p_i} = 1$, then for any functions $u_i \in L^{p_i}(\Omega)$ we have $\int_\Omega \prod_{i=1}^s |u_i| \leq \prod_{i=1}^s \|u_i\|_{L^{p_i}}$.*

Proposition 6.6. (Sobolev embedding.) *Let $1 \leq p$ (if $d \leq 2$) or $1 \leq p \leq \frac{2d}{d-2}$ (if $d > 2$). Then*

$$H_0^1(\Omega) \subset L^p(\Omega), \qquad \|v\|_{L^p} \leq C_p \|\nabla v\|_{L^2} \quad (\forall v \in H_0^1(\Omega)) \tag{6.7}$$

for some constant $C_p > 0$ independent of v.

6.3 Numerical solution of the beam problem

6.3.1 Weak formulation and finite elements

We rewrite (6.2) by dividing with EI and letting

$$\beta := \frac{\alpha}{3I}, \quad k := \frac{k_F}{EI}, \quad g := \frac{f}{EI},$$

further, we impose clamped boundary conditions. Then our problem becomes

$$u^{IV} - \beta\big((u')^3\big)' + ku = g\,, \qquad u(0) = u'(0) = u(b) = u'(b) = 0. \tag{6.8}$$

The weak formulation uses the Sobolev space $H^2(J)$, moreover, using the boundary conditions, we work in the space

$$H_0^2(J) = \{u \in H^2(J) :\ u(0) = u'(0) = u(b) = u'(b) = 0\},$$

which has the standard inner product and corresponding norm

$$\langle u, v\rangle_{H_0^2} := \int_0^b u''v'', \qquad \|u\|_{H_0^2}^2 = \int_0^b (u'')^2\,. \tag{6.9}$$

Then the weak solution of problem (6.8) is defined as $u_* \in H_0^2(J)$ satisfying

$$\int_0^b \big(u_*'' v'' + \beta (u_*')^3 v' + k u_* v\big) = \int_0^b gv \qquad (\forall v \in H_0^2(J)). \tag{6.10}$$

The well-posedness of the problem will be discussed in the next subsection. The weak solution minimizes the energy

$$E(u) := \int_0^b \big(\tfrac{1}{2}(u'')^2 + \tfrac{\beta}{4}(u')^4 + \tfrac{k}{2}u^2 - gu\big)$$

in $H_0^2(J)$.

The finite element method (FEM) can be used to solve (6.10) numerically. Let $V_h \subset H_0^2(J)$ be a given FEM subspace. Then we seek for $u_h \in V_h$ satisfying

$$\int_0^b \big(u_h'' v'' + \beta (u_h')^3 v' + k u_h v\big) = \int_0^b gv \qquad (\forall v \in V_h). \tag{6.11}$$

The formulation and properties of the problem will be given for a general FEM subspace $V_h \subset H_0^2(J)$. Later for the realization we will specify V_h to consist of suitable spline functions.

Rearranging (6.10) and using Riesz representation, one can define an operator $F : V_h \to V_h$ satisfying

$$\langle F(u), v\rangle_{H_0^2} = \int_0^b \big(u'' v'' + \beta (u')^3 v' + kuv - gv\big) \qquad (\forall v \in V_h) \tag{6.12}$$

so that (6.10) simply becomes an equation

$$F(u) = 0 \tag{6.13}$$

in the space V_h (also equipped with the H_0^2 inner product). Then, using standard calculations, see, e.g., [35, Chap. 6.1], one can derive that F has a Gâteaux derivative satisfying

$$\langle F'(u)h, v\rangle_{H_0^2} = \int_0^b \big(h'' v'' + 3\beta (u')^2 h' v' + khv\big) \qquad (\forall u, h, v \in V_h), \tag{6.14}$$

further, that F' is bihemicontinuous. In addition, the role of h and v is interchangeable in formula (6.14), which readily implies the following corollary.

Corollary 6.7. *For any $u \in V_h$ the operator $F'(u)$ is self-adjoint.*

6.3.2 Properties of the linearized operator

The following ellipticity properties hold:

Proposition 6.8. *There exists a continuous increasing function $\Lambda : \mathbb{R}^+ \to \mathbb{R}^+$, independently of h, such that*

$$\|h\|_{H_0^2}^2 \le \langle F'(u)h, h\rangle_{H_0^2} \le \Lambda(\|u\|_{H_0^2})\, \|h\|_{H_0^2}^2 \qquad (\forall u, h \in V_h). \tag{6.15}$$

Namely, $\Lambda(t) = 1 + kC_2^4 + 3\beta C_4^4 t^2$.

PROOF. Owing to (6.14), the quadratic form reads as:

$$\langle F'(u)h, h\rangle_{H_0^2} = \int_0^b \left((h'')^2 + 3\beta(u')^2(h')^2 + kh^2\right) \qquad (\forall u, h \in V_h).$$

Since $\beta,\ k \geq 0$, the first inequality of (6.15) holds.
Furthermore, (6.7) implies that $\|z'\|_{L^4} \leq C_4\|z''\|_{L^2} = C_4\|z\|_{H_0^2}$. Thus we have

$$\int_0^b 3\beta(u')^2(h')^2 \leq 3\beta\|(u')^2\|_{L^2}\|(h')^2\|_{L^2} = 3\beta\|u'\|_{L^4}^2\|h'\|_{L^4}^2 \leq 3\beta C_4^4\|u\|_{H_0^2}^2\|h\|_{H_0^2}^2 \qquad (\forall u, h \in V_h),$$

and similarly,

$$\int_0^b kh^2 = k\|h\|_{L^2}^2 \leq kC_2^2\|h'\|_{L^2}^2 \leq kC_2^4\|h''\|_{L^2}^2 = kC_2^4\|h\|_{H_0^2}^2 \qquad (\forall u, h \in V_h),$$

hence the result follows. ∎

The above shows that Assumptions 6.2.2 (i)-(ii) hold for problem (6.13). As seen in Subsection 6.2.2, this implies well-posedness:

Corollary 6.9. *Problem* (6.10) *has a unique solution* $u^* \in H_0^2(J)$*, further, for any FEM subspace* $V_h \subset H_0^2(J)$*, problem* (6.11) *has a unique solution* $u_h \in V_h$*.*

One can also establish local Lipschitz continuity:

Proposition 6.10. *There exists a continuous increasing function* $L : \mathbb{R}^+ \to \mathbb{R}^+$*, independently of* h*, such that*

$$\|F'(u) - F'(v)\| \leq L(\max\{\|u\|_{H_0^2},\ \|v\|_{H_0^2}\})\,\|u - v\|_{H_0^2} \qquad (\forall u,\ h \in V_h).$$

Namely, $L(t) = 6C_4^4\beta t$*.*

PROOF. Corollary 6.7 implies the symmetry of $F'(u) - F'(v)$ for each $u,\ v \in V_h$, therefore it is possible to obtain its norm using its quadratic form:

$$\|F'(u) - F'(v)\| = \sup_{\|h\|_{H_0^2}=1} |\langle (F'(u) - F'(v))h, h\rangle| = \sup_{\|h\|_{H_0^2}=1} \left|\int_0^b 3\beta((u')^2 - (v')^2)(h')^2\right|.$$

As in the proof of Proposition 6.8 above, we use Proposition 6.5 and the fact that $\|z^2\|_{L^2} = \|z\|_{L^4}^2$. With Proposition 6.6, these yield

$$\begin{aligned} \|F'(u) - F'(v)\| &\leq \sup_{\|h\|_{H_0^2}=1} 3\beta\|(u')^2 - (v')^2\|_{L^2}\|(h')^2\|_{L^2} \\ &\leq \sup_{\|h\|_{H_0^2}=1} 3\beta\|(u')^2 - (v')^2\|_{L^2}C_4^2\|h''\|_{L^2}^2. \end{aligned} \tag{6.16}$$

Since $\|z''\|_{L^2} = \|z\|_{H_0^2}$, this readily entails

$$\|F'(u) - F'(v)\| \leq 3C_4^2\beta\|(u')^2 - (v')^2\|_{L^2} = 3C_4^2\beta\sqrt{\|(u' - v')^2(u' + v')^2\|_{L^1}}. \tag{6.17}$$

Repeating the technique used in (6.16) gives

$$\begin{aligned}\sqrt{\|(u'-v')^2(u'+v')^2\|_{L^1}} &\le \sqrt{\|(u'-v')^2\|_{L^2}\|(u'+v')^2\|_{L^2}} \\ = \|u'-v'\|_{L^4}\|u'+v'\|_{L^4} &\le C_4^2\|u''-v''\|_{L^2}\|u''+v''\|_{L^2}\,.\end{aligned} \tag{6.18}$$

Combining (6.17)-(6.18) yields

$$\|F'(u)-F'(v)\| \le 3C_4^4\beta\|u-v\|_{H_0^2}\|u+v\|_{H_0^2}\,.$$

Finally, since $\|u+v\|_{H_0^2} \le \|u\|_{H_0^2} + \|v\|_{H_0^2} \le 2\max(\|u\|_{H_0^2}, \|v\|_{H_0^2})$, this yields

$$\|F'(u)-F'(v)\| \le 6C_4^4\beta\max(\|u\|_{H_0^2}, \|v\|_{H_0^2})\|u-v\|_{H_0^2}. \qquad \blacksquare$$

6.3.3 Finite elements using splines

The straightforward way to implement finite elements for fourth-order beam problems is to use piecewise cubic splines, that is, Hermitian elements with four degrees of freedom, which satisfy C^1-continuity at the node points. These functions are in $H^2(J)$, and accordingly, the integrals in (6.9) are finite. Alternatively, one could use quadratic B-splines and still achieve C^1-continuity.

We apply a uniform mesh, where the piecewise cubic basis functions are constructed below for mesh parameter h. Two such functions are obtained for each interior node. These nodes are denoted n_k, where $k \in K := \{1,\ 2,\ \cdots,\ b/h-1\}$, and n_k is at location $x_k = hk$.

Let us consider piecewise cubic functions $f_1,\ f_2 : [-1,1] \to \mathbb{R}$, which are defined as

$$f_i(x) = \begin{cases} f_i^*(x) & x \in [0,\ 1] \\ (-1)^{(i-1)} f_i^*(-x) & x \in [-1,\ 0) \end{cases} \qquad (i = 1,\ 2),$$

where $f_1^*(x) = 2x^3 - 3x^2 + 1$, $f_2^*(x) = x^3 - 2x^2 + x$, see Figure 6.1.

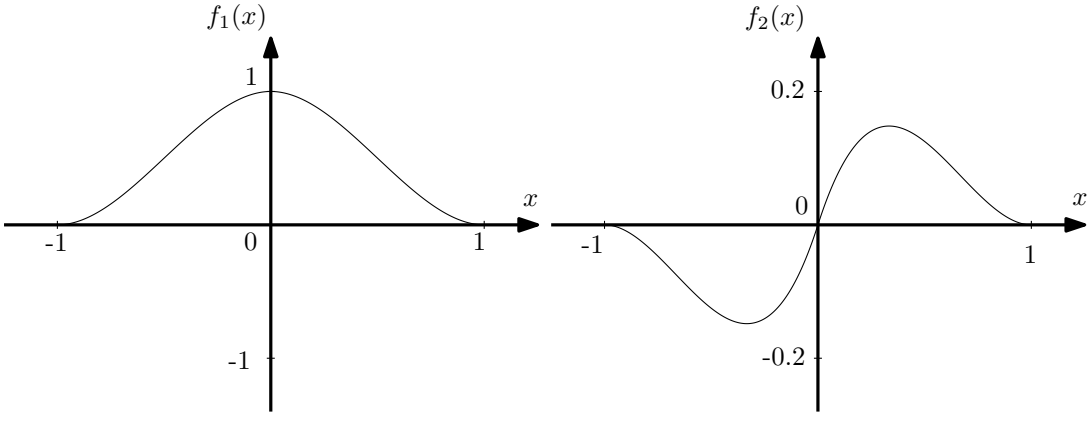

Figure 6.1: Piecewise cubic functions on $[-1,\ 1]$

The basis functions are obtained via affine transformations L_k $(k \in K)$, such that the domain of $L_k(f_i)$ is $[h(k-1), h(k+1)]$ for $i = 1,\ 2$.

6.3.4 The iterative solvers: construction, convergence and mesh-independence

Now we come to the main point of the numerical method, which is the iterative solution of the FEM problem (6.12)–(6.13). We consider the iterative methods of Subsection 6.2.2. For notational simplicity we do not indicate in which subspace V_h our sequences $(u_n)_{n\in\mathbb{N}}$ run in. In fact, we think of V_h as being fixed.

Construction 6.3.4. The three recurrences can be summarized as follows:

$$u_{n+1} := u_n - \sigma z_n\,, \tag{6.19}$$

where the function $z_n \in V_h$ and the constant $\sigma > 0$ are defined in the way described below. We note that for the Newton and quasi-Newton methods one formally has to solve an operator equation, further, the quasi-Newton method will be uniquely defined if we choose proper operators B_n.

- *Simple iteration/gradient method:* $z_n := F(u_n), \qquad \sigma := \frac{2}{\Lambda_0+\lambda}$.

  To determine z_n, let us write the equality with test functions:

  $$\langle z_n, v\rangle_{H_0^2} = \langle F(u_n), v\rangle_{H_0^2} \qquad (\forall v \in V_h).$$

  Here and henceforth let us define the r.h.s. as a functional ℓ_n, that is, by (6.12),

  $$\ell_n v := \langle F(u_n), v\rangle_{H_0^2} \equiv \int_0^b \big(u_n'' v'' + \beta (u_n')^3 v' + k u_n v - g v\big) \qquad (\forall v \in V_h). \tag{6.20}$$

  Then the $z_n \in V_h$ is the solution of the auxiliary FEM problem

  $$\int_0^b z_n'' v'' = \ell_n v \qquad (\forall v \in V_h).$$

- *Newton's method:* $F'(u_n) z_n = F(u_n), \qquad \sigma := 1$.

  To determine z_n, let us again involve test functions:

  $$\langle F'(u_n) z_n, v\rangle_{H_0^2} = \langle F(u_n), v\rangle_{H_0^2} \qquad (\forall v \in V_h).$$

  Using (6.14) and (6.20), $z_n \in V_h$ is the solution of the auxiliary FEM problem

  $$\int_0^b \big(z_n'' v'' + 3\beta (u_n')^2 z_n' v' + k z_n v\big) = \ell_n v \qquad (\forall v \in V_h).$$

- *Quasi-Newton/variable preconditioning:* $B_n z_n = F(u_n), \qquad \sigma := \frac{2}{M_n+m_n}$.

  Now the equation for z_n with test functions is:

  $$\langle B_n z_n, v\rangle_{H_0^2} = \langle F(u_n), v\rangle_{H_0^2} \qquad (\forall v \in V_h).$$

Here we propose to choose the operator B_n as an approximation of $F'(u_n)$ such that the term $3\beta(u_n')^2$ is replaced by proper constants $w_n > 0$: let

$$\langle B_n h, v\rangle_{H_0^2} := \int_0^b \big(h''v'' + w_n h'v' + khv\big) \qquad (\forall v \in V_h). \tag{6.21}$$

That is, by (6.20), $z_n \in V_h$ is the solution of the auxiliary FEM problem

$$\int_0^b \big(z_n''v'' + w_n z_n'v' + kz_nv\big) = \ell_n v \qquad (\forall v \in V_h).$$

It is informative to summarize also the strong versions of the auxiliary problems, which are formal equations involving fourth derivatives. Namely, let us denote

$$r_n := u_n^{IV} - \beta\big((u_n')^3\big)' + ku_n - g\,.$$

It is readily seen by integration that (for smooth u_n)

$$\ell_n v = \int_0^b r_n v \qquad (\forall v \in V_h).$$

Using similar formal integration for the l.h.s., the auxiliary equations become the following, where in each problem the boundary conditions are $z_n(0) = z_n'(0) = z_n(b) = z_n'(b) = 0$:

- *Simple iteration/gradient method:* $z_n^{IV} = r_n$
- *Newton's method:* $z_n^{IV} - 3\beta\big((u_n')^2 z_n'\big)' + kz_n = r_n$
- *Quasi-Newton/variable preconditioning:* $z_n^{IV} - w_n z_n'' + kz_n = r_n$

That is, the auxiliary problems correspond to the solution of proper linear fourth-order ODEs. In reality, of course, we consider the FEM solution of the weak versions of these problems, where u_n is only an H^2 function.

Now we derive the convergence of the iterations, based on the properties in Subsection 6.3.2. In addition, we need the spectral equivalence property (6.5) for the above defined operators B_n and $F'(u_n)$. A reasonable choice of w_n is in the range of the function $3\beta \max_{\overline{\Omega}}(u_n')^2$ that it approximates:

$$0 \le w_n \le 3\beta \max(u_n')^2\,. \tag{6.22}$$

A convenient choice is the arithmetic mean

$$w_n := \frac{3\beta}{2} \max(u_n')^2. \tag{6.23}$$

Proposition 6.11. *For given $u_n \in V_h$, let the constant w_n satisfy* (6.22)*. Then*

$$m_n \langle B_n h, h\rangle_{H_0^2} \le \langle F'(u_n)h, h\rangle_{H_0^2} \le M_n \langle B_n h, h\rangle_{H_0^2} \qquad (\forall h \in V_h),$$

where

$$m_n = \frac{1}{1 + w_n C_2^2}, \qquad M_n = \frac{3\beta \max(u_n')^2}{w_n}\,. \tag{6.24}$$

PROOF. Firstly, to obtain the upper bound, by (6.14), one can write

$$\begin{aligned}\langle F'(u_n)h,h\rangle_{H_0^2} &= \int_0^b \big((h'')^2 + 3\beta(u_n')^2(h')^2 + kh^2\big)\\ &\le \int_0^b \big((h'')^2 + 3\beta \max_{\overline{\Omega}}\{(u_n')^2\}(h')^2 + kh^2\big) \qquad (\forall h \in V_h),\end{aligned}$$

on the other hand, owing to (6.22), we have $1 \le M_n := \frac{3\beta \max_{\overline{\Omega}}(u_n')^2}{w_n}$ for each $n \in \mathbb{N}$, hence

$$\langle F'(u_n)h,h\rangle_{H_0^2} \le M_n \int_0^b \big((h'')^2 + w_n(h')^2 + kh^2\big) = M_n\langle B_n h,h\rangle_{H_0^2} \qquad (\forall h \in V_h).$$

To obtain the lower bound, by (6.21) and Proposition 6.6, we have

$$\langle B_n h,h\rangle_{H_0^2} = \|h''\|_{L^2}^2 + w_n\|h'\|_{L^2}^2 + k\|h\|_{L^2}^2 \le (1 + w_n C_2^2)\|h''\|_{L^2}^2 + k\|h\|_{L^2}^2 \qquad (\forall h \in V_h),$$

writing this back to integral form and adding the term $3\beta(u_n')^2(h')^2 \ge 0$ yields

$$\langle B_n h,h\rangle_{H_0^2} \le \int_0^b \big((1 + w_n C_2^2)(h'')^2 + 3\beta(u_n')^2(h')^2 + kh^2\big) \qquad (\forall h \in V_h).$$

This readily entails

$$\langle B_n h,h\rangle_{H_0^2} \le (1 + w_n C_2^2)\langle F'(u_n)h,h\rangle_{H_0^2}. \qquad \blacksquare$$

Now we can formulate the convergence results:

Theorem 6.12. *The iterative methods, defined in Construction 6.3.4, provide the following convergence estimates:*

- *Simple iteration/gradient method:*

$$\frac{\|F(u_{n+1})\|_{H_0^2}}{\|F(u_n)\|_{H_0^2}} \le \frac{\Lambda_0 - \lambda}{\Lambda_0 + \lambda}, \quad \text{where } \Lambda_0 = 1 + kC_2^4 + 3\beta C_4^4\left(\|u_0\| + \frac{1}{\lambda}\|F(u_0)\|\right)^2.$$

- *Newton's method:*

$$\frac{\|F(u_{n+1})\|_{H_0^2}}{\|F(u_n)\|_{H_0^2}^2} \le \frac{L_0}{2\lambda^2}, \quad \text{where } L_0 = 6C_4^4\beta\left(\|u_0\| + \frac{2}{\lambda}\|F(u_0)\|\right).$$

- *Quasi-Newton/variable preconditioning: if* (6.22) *holds then*

$$\limsup \frac{\|F(u_{n+1})\|_*}{\|F(u_n)\|_*} \le \limsup \frac{M_n - m_n}{M_n + m_n}$$

with the constants in (6.24).

These hold globally for the simple iteration and locally for the Newton and quasi-Newton methods.

PROOF. It follows from Theorems 6.1–6.3, Propositions 6.8–6.10 and Proposition 6.11. ∎

Remark 6.13. (a) The estimates in Theorem 6.12 are uniform, i.e. the constant on the r.h.s. of the inequalities are *mesh-independent*.

(b) Global convergence for the Newton and quasi-Newton methods can be achieved via damped versions (see, e.g., [18, 35] for the abstract theorems), which are not detailed here. In this case the above estimates are ultimate, i.e. they hold in lim sup sense also for Newton's method.

(c) In the above we considered the iterations on a fixed mesh. The methods might be generalized to a multilevel setting to increase the efficiency of preconditioning, see related work in [11, 31], but such extensions are beyond the scope of the present section.

6.3.5 Generalizations

6.3.5.1 Other boundary conditions. Instead of the rigidly clamped beam in (6.8), one can consider a freely supported beam. Then the first derivatives at the endpoints are replaced by second derivatives, i.e., the problem becomes

$$u^{IV} - \beta\big((u')^3\big)' + ku = g\,, \qquad u(0) = u''(0) = u(b) = u''(b) = 0. \tag{6.25}$$

This problem can be posed in the Sobolev space $H^2(J) \cap H_0^1(J)$, which is endowed with the same inner product (6.9) as used in $H_0^2(J)$. Hence the calculations can be repeated and the same convergence results hold as in Subsection 6.3.4. The auxiliary problems are obviously solved with the freely supported boundary conditions as in (6.25).

6.3.5.2 A modified equation. The above results can also be reproduced for the modified version (6.3). Owing to the fact that the axial force P is below the Euler critical load P_{cr}^E, it follows that the energy functional is uniformly convex, see, e.g., [59], which implies that the uniform monotonicity (6.4) holds. The other conditions remain unchanged, hence the calculations can be repeated to obtain the same convergence results.

6.3.5.3 Extension to plane problems. For the 2D plate problem in (6.1), the analogue of equation (6.8) is

$$\Delta^2 u - \beta\,\mathrm{div}\big(|\nabla u|^2\ \nabla u\big) + ku = g \quad \text{in } \Omega, \qquad u_{|\partial\Omega} = \tfrac{\partial u}{\partial \nu}\big|_{\partial\Omega} = 0$$

for a thin plate $\Omega \subset \mathbb{R}^2$. The weak solution minimizes the energy

$$E(u) := \int_\Omega \big(\tfrac{1}{2}|D^2 u|^2 + \tfrac{\beta}{4}|\nabla u|^4 + \tfrac{k}{2}u^2 - gu\big)$$

in the Sobolev space $H_0^2(\Omega)$. It is easy to see that our 1D results can be readily extended to this situation. The problem is posed in $H_0^2(\Omega)$ endowed with the inner product $\langle u, v\rangle_{H_0^2} := \int_\Omega D^2 u : D^2 v$. The main analytic point is that the required Sobolev embedding $H_0^1(\Omega) \subset L^4(\Omega)$ also holds in 2D owing to (6.7). The other used techniques are independent of the dimension of the domain. In the case of the freely supported plate, the boundary conditions become $u_{|\partial\Omega} = \frac{\partial^2 u}{\partial \nu^2}\big|_{\partial\Omega} = 0$ and the problem is posed in the Sobolev space $H^2(\Omega) \cap H_0^1(\Omega)$. Altogether, the calculations can be repeated to obtain the same convergence results as before.

6.3.6 Numerical experiments

The model described by (6.8) was used for simulation. The results are presented with the original physical parameters used in (6.2) for the sake of convenience. The physical and mesh parameters where chosen with the help of [41, 42].

The investigated problems included steel and concrete beams, with length $L = 2$ m, and we applied contact stiffness $k = 3 \cdot 10^8 \ \frac{\mathrm{N}}{\mathrm{m}^2}$. The steel and concrete beams have modulus of elasticity $E_1 = 2.1 \cdot 10^{11}$ Pa and $E_2 = 3 \cdot 10^{10}$ Pa, respectively, and Poisson's ratio $\nu_1 = 0.3$ and $\nu_2 = 0.2$, respectively. As second moment of area, $I = 2/3 \cdot 10^{-3}$ m^4 was used. This results from a rectangular cross section, namely, the beam height is $h = 0.1$ m, and the beam width is considered as a unit. The total vertical loadings are

$$F_1 = -1.5 \cdot 10^8 \text{ N}, \quad F_2 = -3 \cdot 10^8 \text{ N}, \quad F_3 = -5 \cdot 10^8 \text{ N}$$

for the steel beam and

$$F_4 = -1 \cdot 10^7 \text{ N}, \quad F_5 = -4 \cdot 10^7 \text{ N}, \quad F_6 = -8 \cdot 10^7 \text{ N}$$

for the concrete beam. These loads are distributed uniformly along the beam, and the distributed force is $q = \frac{F}{L}$ (that is, $q_i = \frac{F_i}{L}$ in the distinct experiments). The number of finite elements is denoted by NE.

For all simulations, the initial guess was constant 0, and the stopping criterion was the relative residual decreased below 10^{-4}. A direct linear solver was used in case of the obtained linear problems in all iteration steps. For the quasi-Newton method, the suggested choice (6.23) was used for the preconditioner.

It has been found that all the three methods converge and are robust.

The constant σ in (6.19) was replaced by different values in an attempt to achieve faster convergence. The investigation showed that $\sigma = 1$ is suitable for all methods, hence all three methods were considered with this constant. Damping was not necessary for these methods.

Tables 6.1, 6.2 and 6.3 show the number of iterations with the quasi-Newton method, the Sobolev gradient method and the full Newton method, respectively, to illustrate the robustness of the methods.

	$E = E_1$, $\nu = \nu_1$			$E = E_2$, $\nu = \nu_2$		
NE	$q = q_1$	$q = q_2$	$q = q_3$	$q = q_4$	$q = q_5$	$q = q_6$
8	3	4	5	3	4	5
80	3	4	5	3	4	5
800	3	4	5	3	4	5
8000	4	4	5	3	4	5

Table 6.1: Number of iterations using the quasi-Newton method.

The total runtimes (i.e. the runtimes of the whole simulation, not only an individual iteration step) for the quasi-Newton method, the Sobolev gradient method and the Newton method are denoted by t_{qN}, t_{g}, and t_{N}, respectively. These values were obtained by averaging the total runtimes of multiple simulations for each parameter combination. Namely, for the cases NE = 8, 80, 800, 8000, the number of simulations used to measure the total runtimes

	$E = E_1$, $\nu = \nu_1$			$E = E_2$, $\nu = \nu_2$		
NE	$q = q_1$	$q = q_2$	$q = q_3$	$q = q_4$	$q = q_5$	$q = q_6$
8	5	6	9	14	16	26
80	5	6	9	14	16	26
800	5	6	9	14	16	26
8000	5	6	9	14	16	26

Table 6.2: Number of iterations using the Sobolev gradient method.

	$E = E_1$, $\nu = \nu_1$			$E = E_2$, $\nu = \nu_2$		
NE	$q = q_1$	$q = q_2$	$q = q_3$	$q = q_4$	$q = q_5$	$q = q_6$
8	3	3	4	3	3	4
80	3	3	4	3	3	4
800	3	3	4	3	3	4
8000	3	3	4	3	4	4

Table 6.3: Number of iterations using the full Newton method.

were 50000, 5000, 500, 50, respectively, and the averages of the measured runtimes were used. We have compared Newton's method (considered as a standard nonlinear solver) with the two other ones in Tables 6.4 and 6.5.

In Table 6.4, all of the values $t_{\mathrm{qN}}/t_{\mathrm{N}}$ are smaller than 1 for the investigated problems, i.e. quasi-Newton method is more efficient than full Newton method for the investigated problems with respect to computational cost. Furthermore, increasing mesh density further improves the relative performance of the quasi-Newton method, owing to the increasing benefit from the simplification of the stiffness matrix.

In Table 6.5, one can observe that (especially in the case of coarse meshes, e.g. 8 and 80 elements) the Sobolev gradient method may underperform the full Newton method in runtimes, e.g. for the concrete beam, though it is apparently faster for the steel beam. The relative computational cost of Sobolev gradient method also improves with increasing mesh density. Altogether, the Sobolev gradient method often performs well, due to no assembling required in the steps.

In 10 of the cases under investigation, the quasi-Newton method is the most effective with respect to computational cost, while the Sobolev gradient method is favored in the remaining 14 cases.

	$E = E_1$, $\nu = \nu_1$			$E = E_2$, $\nu = \nu_2$		
NE	$q = q_1$	$q = q_2$	$q = q_3$	$q = q_4$	$q = q_5$	$q = q_6$
8	0.656	0.814	0.739	0.733	0.831	0.757
80	0.570	0.679	0.603	0.654	0.722	0.648
800	0.508	0.586	0.500	0.571	0.611	0.527
8000	0.458	0.454	0.370	0.451	0.362	0.372

Table 6.4: Comparison of quasi-Newton and full Newton runtimes: the values of $t_{\mathrm{qN}}/t_{\mathrm{N}}$.

For one particular case, namely, $E = E_2$, $\nu = \nu_2$, $q = q_6$, NE$= 8000$, for the quasi-Newton

	$E = E_1$, $\nu = \nu_1$			$E = E_2$, $\nu = \nu_2$		
NE	$q = q_1$	$q = q_2$	$q = q_3$	$q = q_4$	$q = q_5$	$q = q_6$
8	0.567	0.662	0.735	1.377	1.541	1.945
80	0.466	0.541	0.599	1.165	1.328	1.664
800	0.357	0.417	0.440	0.890	0.921	1.114
8000	0.197	0.210	0.196	0.317	0.267	0.456

Table 6.5: Comparison of Sobolev gradient and full Newton runtimes: the values of $t_{\mathrm{g}}/t_{\mathrm{N}}$.

method, the apparent fulfillment of Theorem 6.12 is illustrated for each iteration step n in Table 6.6.

n	1	2	3	4
$\|F(u_{n+1})\|_*/\|F(u_n)\|_*$	0.055	0.052	0.056	0.056
$(M_n - m_n)/(M_n + m_n)$	0.333	0.526	0.511	0.512

Table 6.6: One quasi-Newton iteration.

Additional experiments have been carried out to determine whether a change in Poisson's ratio ν affects these results. For this task, Poisson's ratio of the concrete beam was changed to $\nu = 0.3$ and $\nu = 0.1$, while other parameters were left unchanged. For all methods, it has been found that the robustness result holds for the new parameter combinations as well, moreover, the corresponding iteration numbers remain almost unchanged. The relative total runtimes also qualitatively coincided with the original results.

Other experiments included replacing the contact stiffness k of the concrete beam with lower values, $k = 2 \cdot 10^7 \ \frac{\mathrm{N}}{\mathrm{m}^2}$ and $k = 0 \ \frac{\mathrm{N}}{\mathrm{m}^2}$, in the latter case the model effectively lacking contact stiffness. For the large deformation of these soft models, Gao beam model (6.2) might not hold due to yielding of the material, however, the simulations can be carried out. The robustness result holds for these simulations. The supremacy of the quasi-Newton method over the full Newton method is also sustained, though the Sobolev gradient method exhibits relative improvement.

One can conclude that all three examined methods are robust, and quasi-Newton method can replace full Newton method for this nonlinear model. The Sobolev gradient method is very efficient for thousands of elements and more, otherwise, one should use quasi-Newton method. These coarse meshes appear to be of significance, as [41] states that 32 elements already suffice for accurate computations.

6.4 Conclusions

The present section provides a detailed description of three iterative methods for nonlinear Gao beam models using finite elements. The description includes the Sobolev gradient method, Newton's method and a quasi-Newton method with a recently developed framework for elliptic problems with non-uniformly monotone bounds. The results of both theoretical and practical work are shown. It has been found that all three methods are robust for the problems, and a comparison has been made between their behaviour under varying parameters.

7 Stefan-Boltzmann heat radiation problems in 3D

7.1 Introduction

In this section we consider a stationary heat conduction problem, involving nonlinear Stefan-Boltzmann radiation boundary conditions, on a bounded domain in $\mathbb{R}^3$. We present here the results of [21], which was motivated by the paper [58], where this problem was treated carefully, and our goal is to extend their results in two directions.

The problem consists of the elliptic heat conduction equation

$$-\mathrm{div}\big(A\nabla u\big) = f \quad \text{in } \Omega \tag{7.1}$$

equipped with mixed Dirichlet and nonlinear Stefan-Boltzmann radiation boundary conditions

$$u_{|\partial\Omega} = \overline{u} \quad \text{on } \Gamma_{\mathrm{D}}, \tag{7.2}$$

$$\alpha u + \nu^{\mathrm{T}} A\nabla u + \beta u^4 = g \quad \text{on } \Gamma_{\mathrm{N}}, \tag{7.3}$$

where Ω is a bounded domain in $\mathbb{R}^3$ with Lipschitz continuous boundary $\partial\Omega = \overline{\Gamma}_{\mathrm{D}} \cup \overline{\Gamma}_{\mathrm{N}}$, where Γ_{D} and Γ_{N} have positive 2-dimensional (surface) measure. The 3×3 matrix A of heat conductivities is symmetric and uniformly positive definite, $f \geq 0$ is the density of body heat sources, $u \geq 0$ is the absolute temperature, $\bar{u} \geq 0$ is the prescribed temperature. The heat transfer coefficient is denoted by $\alpha \geq 0$, ν is the outward unit normal to Γ_{N}, and $\beta = \sigma f_{\mathrm{em}}$ with the Stefan-Boltzmann constant $\sigma = 5.669 \cdot 10^{-8}$ $\mathrm{Wm^{-2}K^{-4}}$ and the relative emissivity function $0 \leq f_{\mathrm{em}} \leq 1$.

Nonlinear heat radiation problems have been widely studied in several situations owing to their importance, see, e.g., [26, 56, 62, 67, 68, 73], including problem (7.1)–(7.3) on 2D domains and axially symmetric 3D domains. The general 3D case was then clarified in the motivating paper [58], where the proper function space was set up, and the convergence of the finite element approximation and of Newton's method was derived. However, their results have certain limitations. First, they only cover the case of diagonal matrix A, and formulate as an open problem to treat general anisotropic materials where the matrix A of heat conductivities is a non-diagonal full matrix. Moreover, they only consider the exact Newton method, but most often it is much useful from practical aspect to involve quasi-Newton type methods where the Jacobians are approximated, and hence one can spare a significant computational work.

The goal of this section is accordingly twofold. First, we solve the open problem and show that a non-diagonal full matrix can be allowed in the results of [58]. The main point here is to prove the nonnegativity of the solution. This enables one to use a rewritten form of the problem, which allows the use of a proper operator formulation. Second, we develop a quasi-Newton method for this problem in the form of variable preconditioning, based on the previous sections. We prove the convergence of the quasi-Newton method using proper preconditioning operators.

This section is organized as follows. The problem of nonnegativity of the solution for non-diagonal matrices is treated in Subsection 7.2. The consequent results on finite element approximation and exact Newton method are summarized briefly in Subsection 7.3. The pro-

posed quasi-Newton method is established in Subsection 7.4, the construction and convergence are presented. Finally, in Subsection 7.5, numerical experiments illustrate the robustness of the methods and the reduced computational cost of the quasi-Newton method.

7.2 The modified problem and the nonnegativity of the solution

A proper approach to study a nonlinear elliptic problem is to write its weak form as an operator equation in a Banach or Hilbert space [35, 81], often related to the minimization of a potential. This approach is taken in [58] as well. However, problem (7.1)–(7.3) in its original form does not allow to involve a convex potential, since the term $u \mapsto u^4$ is not monotone. Consider now the problem where this term is replaced with $|u|^3u$. If one can show that the solution of the modified form of the problem is nonnegative, then $|u|^3u$ and u^4 coincide, hence the original and the modified BVPs are equivalent.

The paper [58] proves the desired nonnegativity under the restriction that A is a diagonal matrix. In this subsection we prove nonnegativity for the full matrix case. This requires a different technique from that of [58], namely, we adapt the idea of the proof of [54, Theorem 5].

The modified form of BVP (7.1)–(7.3) is the following:

$$-\mathrm{div}\big(A\nabla u\big) = f \quad \text{in } \Omega, \tag{7.4}$$

$$u_{|\Gamma_\mathrm{D}} = \overline{u} \quad \text{on } \Gamma_\mathrm{D}, \tag{7.5}$$

$$\alpha u + \nu^\mathrm{T} A\nabla u + \beta|u|^3u = g \quad \text{on } \Gamma_\mathrm{N}. \tag{7.6}$$

For the proper formulation of the problem, besides the sign conditions posed for (7.1)–(7.3), we assume that the entries a_{ij} of the function-valued matrix A are in $L^\infty(\Omega)$, further, $f \in L^2(\Omega)$, $g \in L^2(\Gamma_2)$, $\bar{u} \in H^1(\Omega)$, $\bar{u}|_\Gamma \in L^5(\Gamma_\mathrm{N})$, $\alpha,\ \beta \in L^\infty(\Gamma_\mathrm{N})$, and $\exists\beta_0 > 0$ such that $\beta \geq \beta_0$ a. e. Furthermore, $A(x)$ is not only symmetric and positive definite for any $x \in \Omega$, but we assume that there exist constants $\mu_0,\ \mu_1 > 0$ such that, for all $x \in \Omega$ and vector $v \in \mathbb{R}^3$,

$$\mu_0|v|^2 \leq A(x)v \cdot v \leq \mu_1|v|^2\,. \tag{7.7}$$

The weak solution is looked for within the space $H^1(\Omega)$, and the test functions within the space

$$H^1_\mathrm{D}(\Omega) := \{v \in H^1(\Omega) :\ v_{|\Gamma_\mathrm{D}} = 0 \ \text{ in trace sense}\},$$

which has the natural norm

$$\|v\|_{H^1_\mathrm{D}} := \|\nabla v\|_{L^2(\Omega)}.$$

However, as pointed out in [58], the traces of the variational solution should belong to the Lebesgue space $L^5(\Gamma_\mathrm{N})$, which is true in 2D but no longer true in 3D for a general function in $H^1(\Omega)$. (The need for 5th power integrals on Γ_N will be seen in (7.9).) Hence the proper function space for problem (7.4)–(7.6) in 3D is

$$V := \left\{v \in H^1(\Omega) :\ v|_{\Gamma_\mathrm{N}} \in L^5(\Gamma_\mathrm{N})\right\},$$

which is a Banach space equipped with the norm

$$\|v\|_V := \|v\|_{H^1(\Omega)} + \|v\|_{L^5(\Gamma_N)}.$$

Moreover, we will use the Banach space

$$V_D := V \cap H^1_D(\Omega)$$

with the norm

$$\|v\|_{V_D} := \|v\|_{H^1_D(\Omega)} + \|v\|_{L^5(\Gamma_N)}, \tag{7.8}$$

to serve as the proper space for the test functions and the solution of the homogenized problem, which vanish on Γ_D.

Then the weak form of problem (7.4)–(7.6) can be written as follows: we look for $u \in V$, satisfying $u - \bar{u} \in V_D$, such that

$$\int_\Omega A\nabla u \cdot \nabla v + \int_{\Gamma_N} (\alpha + \beta|u|^3)uv = \int_\Omega fv + \int_{\Gamma_N} gv \qquad (\forall v \in V_D). \tag{7.9}$$

Theorem 7.1. *If u is the weak solution of (7.4)–(7.6), then u is nonnegative.*

Proof. Let u satisfy (7.9), and let us use the specific test function

$$v := \min\{u,\ 0\}. \tag{7.10}$$

We must check that $v \in V_D$. Indeed, $u \in H^1(\Omega)$ implies $v \in H^1(\Omega)$, see [58], and obviously $u|_{\Gamma_N} \in L^5(\Gamma_N)$ implies $v|_{\Gamma_N} \in L^5(\Gamma_N)$; finally, $v|_{\Gamma_D} = \min\{\bar{u},\ 0\} = 0$.

By definition, we have $v \le 0$. In order to prove that $u \ge 0$, we must show that $v = 0$ on Ω.

Let us substitute this v in (7.9), and rewrite the latter with the following decomposition of Ω:

$$\Omega_+ := \{x \in \Omega : \quad u(x) > 0\} \quad \text{and} \quad \Omega_- := \{x \in \Omega : \quad u(x) \le 0\}.$$

Since $v = 0$ and $\nabla v = 0$ in Ω_+, thus the integrals used in (7.9) are zero on the subdomain Ω_+, that is, it suffices to integrate on Ω_-, and, similarly, on $\Gamma_{N^-} := \{x \in \Gamma_N : \quad u(x) \le 0\}$, respectively. In turn, $u = v$ on $\Omega_- \cup \Gamma_{N^-}$ and $\nabla u = \nabla v$ in Ω_-, hence we obtain

$$\int_{\Omega_-} A\nabla v \cdot \nabla v + \int_{\Gamma_{N^-}} (\alpha + \beta|v|^3)v^2 = \int_{\Omega_-} fv + \int_{\Gamma_{N^-}} gv\,.$$

Now we can add zero $\big($the integrals on Ω_+ and on $\Gamma_{N^+} := \Gamma_N \setminus \Gamma_{N^-}\big)$ to both sides, thus we again integrate on the whole domain Ω and Γ_N:

$$\int_\Omega A\nabla v \cdot \nabla v + \int_{\Gamma_N} (\alpha + \beta|v|^3)v^2 = \int_\Omega fv + \int_{\Gamma_N} gv. \tag{7.11}$$

Here, owing to the positive definiteness of A and the sign conditions $\alpha, \beta \ge 0$, we have

$$0 \le \int_\Omega A\nabla v \cdot \nabla v, \qquad 0 \le \int_{\Gamma_N} (\alpha + \beta|v|^3)v^2,$$

this implies that the l. h. s. of (7.11) is nonnegative. On the other hand, the r. h. s. of (7.11) is nonpositive due to $v \le 0$ and the sign conditions $f, g \ge 0$. Therefore both sides of (7.11) are zero, and in particular,

$$\int_\Omega A\nabla v \cdot \nabla v = 0.$$

Then the positive definiteness of A yields that $\nabla v = 0$, so $v = c$ is constant. Moreover, we know that $v|_{\Gamma_\mathrm{D}} = 0$, hence $c = 0$, that is, $v = 0$, which we wanted to prove. ∎

7.3 Well-posedness, finite element approximation and Newton iteration

In this subsection we collect those results of [58] which are directly applicable here to (7.4)–(7.6), and, knowing now the nonnegativity property, to the original problem (7.1)–(7.3) as well. The main point is that [58] does not explicitly exploit the diagonality assumption, the latter is only used therein to obtain the nonnegativity property. The following theorems of [58] only use the property

$$a(v,v) := \int_\Omega A\nabla v \cdot \nabla v + \int_{\Gamma_\mathrm{N}} \alpha v^2 \ge c\|v\|^2_{H^1(\Omega)} \qquad (\forall v \in V_\mathrm{D})$$

for some constant $c > 0$. This ellipticity property remain naturally valid in our full matrix case, owing to the uniformly positivity condition (7.7):

$$a(v,v) \ge \mu_0 \int |\nabla v|^2 = \mu_0 \|v\|^2_{H^1_\mathrm{D}(\Omega)} \ge \mu_0 c_\mathrm{D} \|v\|^2_{H^1(\Omega)}\,, \tag{7.12}$$

where $c_\mathrm{D} > 0$ is a suitable constant, using that $\|v\|_{H^1_\mathrm{D}(\Omega)}$ and $\|v\|_{H^1(\Omega)}$ are equivalent norms, thanks to the property that Γ_D has positive surface measure.

Altogether, the four theorems below follow from (7.12) and [58, Theorems 2.1, 3.1, 3.2, 3.3], respectively.

Theorem 7.2. *Problem (7.4)–(7.6) has a unique weak solution.*

The family of finite element spaces $V_h^\mathrm{D} \subset V_\mathrm{D}$ $(h > 0)$ is chosen to satisfy the standard approximation property:

$$\text{for any } v \in V_\mathrm{D}, \quad \mathrm{dist}(v, V_h^\mathrm{D}) \to 0 \quad \text{as } h \to 0. \tag{7.13}$$

Theorem 7.3. *Assume that subspaces $\{V_h^\mathrm{D}\}_{h\to 0}$ satisfy hypothesis* (7.13)*. Let u and u_h be the exact and the FEM solutions, respectively. Then*

$$\|u - u_h\|_V \to 0 \quad \text{as } h \to 0.$$

The convergence order can be estimated under the following condition on the interpolant $I_h u$:

$$\text{for some } s \in [1,\, 7/6],\ c > 0 \text{ and integer } k \ge 2, \qquad \|u - I_h u\|_s \le c h^{k-s}. \tag{7.14}$$

Theorem 7.4. *Assume that $\bar u \in H^1(\Omega)$ and $\bar u|_{\Gamma_\mathrm{N}} \in L^6(\Gamma_\mathrm{N})$. Let u and u_h be the exact and the FEM solutions, respectively. If $u \in H^k(\Omega)$, and* (7.14) *holds, then there exists a positive*

constant c independent of h such that

$$\|u - u_h\|_V \le ch^{2(k-1)/5} \quad \text{as } h \to 0.$$

Finally, the convergence of the Newton iteration $(u_{h,n})_{n\in\mathbb{N}^+} \subset V_h^{\mathrm{D}}$ to the FEM solution u_h is formulated with the error $e_n := \|u_{h,n} - u_h\|_V$:

Theorem 7.5. *The Newton iteration is well-defined, and there exists positive constants δ and c independent of n such that for every $e_0 < \delta$, we have*

$$e_{n+1} \le ce_n^2.$$

7.4 The quasi-Newton method (variable preconditioning)

In this subsection we formulate our quasi-Newton method where the Jacobians are approximated based on spectral equivalence in the function space, that is, we define a kind of variable preconditioning using proper preconditioning operators. Such an approach has been used in previous sections in other situations, now we adapt it to the case of boundary nonlinearity in the given function space.

7.4.1 Background in Banach space

We will use an abstract result based on Sections 4 and 6. The theorem is presented in a form that will fit the situation of the studied radiation problem. Recall that the formulation involves the "energy $*$-norm" $\|v\|_* := \langle v,\, F'(z^*)^{-1}v\rangle^{1/2}$, which is equivalent to the original one.

The theorem below is a special case of Theorem 4.13, where a non-uniform lower bound was allowed and damping was used to extend the convergence domain. The latter might be applied here as well, but is not included in the theorem for simplicity (see Remark 4.18). It is also worth mentioning that Theorem 6.3 is the Hilbert space predecessor of the theorem below in just the same form.

Theorem 7.6. *Let X be a real Banach space, $F : X \to X'$ a nonlinear operator, and let us consider the operator equation*

$$F(z) = 0. \tag{7.15}$$

Let F have a bihemicontinuous Gâteaux derivative that satisfies the following properties:

(i) For any $z \in X$ the operator $F'(z)$ is symmetric.

(ii) There exists a constant $\lambda > 0$ such that

$$\lambda\|h\|^2 \le \langle F'(z)h, h\rangle \qquad (\forall z, h \in X).$$

(iii) There exists a continuous nondecreasing function $\Lambda :\ \mathbb{R} \to \mathbb{R}$ such that

$$\langle F'(z)h, h\rangle \le \Lambda(\|z\|)\|h\|^2 \qquad (\forall z,\ h \in X).$$

(iv) There exists a continuous nondecreasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|F'(z) - F'(w)\| \le L(\max\{\|z\|, \|w\|\})\,\|z - w\| \qquad (\forall z,\, w \in X).$$

Denote by $z^* \in X$ the unique solution of (7.15). Let z_0 be in a sufficiently small neighbourhood of z^*, and let the sequence (z_n) be defined by

$$z_{n+1} := z_n - \frac{2}{M_n + m_n}\, B_n^{-1} F(z_n) \qquad (n \in \mathbb{N}), \tag{7.16}$$

where $0 < m_n \le M_n$ and for any $n \in \mathbb{N}$ the bounded symmetric linear operator $B_n : X \to X'$ is chosen such that

$$m_n \langle B_n h, h\rangle \le \langle F'(z_n)h, h\rangle \le M_n \langle B_n h, h\rangle \qquad (\forall h \in X). \tag{7.17}$$

We require (m_n) to be positively bounded from below and (M_n) bounded from above. Then (z_n) converges to z^*, that is,

$$\|z_n - z^*\| \le \frac{1}{\lambda}\|F(z_n)\| \to 0,$$

moreover,

$$\limsup \frac{\|F(z_{n+1})\|_*}{\|F(z_n)\|_*} \le \limsup \frac{M_n - m_n}{M_n + m_n} < 1. \tag{7.18}$$

Proof. See Theorem 4.13. ∎

Remark 7.7. *Condition (iii) is redundant since the existence of Λ is actually a consequence of condition (iv), however, in practice a direct estimation will give sharper values of Λ.*

7.4.2 Convergence of the quasi-Newton method for the radiation problem

The application of Theorem 7.6 to our elliptic problem will require that the solution and the test functions be in the same Banach space, hence we homogenize BVP (7.4)–(7.6) by letting $z := u - \bar{u}$. This yields the following problem:

$$-\mathrm{div}\big(A\nabla z\big) = \tilde{f} \quad \text{in } \Omega,$$

$$z_{|\partial\Omega} = 0 \quad \text{on } \Gamma_{\mathrm{D}},$$

$$\alpha(z + \bar{u}) + \nu^{\mathrm{T}} A \nabla z + \beta |z + \bar{u}|^3 (z + \bar{u}) = \tilde{g} \quad \text{on } \Gamma_{\mathrm{N}},$$

where $\tilde{f} := f + \mathrm{div}\,\big(A\nabla\bar{u}\big)$, $\tilde{g} := g - \nu^{\mathrm{T}} A \nabla \bar{u}$. In weak form, we then look for $z \in V_{\mathrm{D}}$ such that

$$\int_\Omega A\nabla z \cdot \nabla v + \int_{\Gamma_{\mathrm{N}}} (\alpha + \beta|z + \bar{u}|^3)(z + \bar{u})v = \int_\Omega \tilde{f} v + \int_{\Gamma_{\mathrm{N}}} \tilde{g} v \qquad (\forall v \in V_{\mathrm{D}}). \tag{7.19}$$

In what follows, we apply the finite element method (FEM) for (7.19) in some FEM subspace $V_h \subset V_{\mathrm{D}}$, endowed with the same norm $\|.\|_{V_{\mathrm{D}}}$ as in (7.8). (The sole restriction is that, when refining the mesh, property (7.13) should hold for the family, then Theorem 7.3 ensures the convergence of the FEM.) In a given FEM subspace $V_h \subset V_{\mathrm{D}}$, the problem to solve is obtained by replacing V_{D} by V_h in (7.19).

Our goal is to solve the FEM problem with a quasi-Newton iteration, that is, we must show the applicability of Theorem 7.6 where $X = V_h$ and the corresponding operator $F : V_h \to V_h'$ is defined via the weak form:

$$\langle F(z),\ v\rangle = \int_\Omega A\nabla z\cdot\nabla v + \int_{\Gamma_{\mathrm{N}}} (\alpha+\beta|z+\bar u|^3)(z+\bar u)v - \int_\Omega \tilde f v - \int_{\Gamma_{\mathrm{N}}} \tilde g v \quad (\forall z, v\in V_h). \tag{7.20}$$

We note that elements of V_h might be denoted in a usual way by z_h, v_h etc., but we can omit such subscripts for simplicity, since from now on we only work in V_h.

Theorem 7.8. *Assumptions (i)-(iv) of Theorem 7.6 hold for the operator defined by* (7.20).

Proof. The derivative of the real function $t \mapsto (\alpha+\beta|t+\bar u|^3)(t+\bar u)$ is $t\mapsto \alpha + 4\beta|t+\bar u|^3$. Then, following standard techniques (see [35]), we can differentiate in the integral and obtain

$$\langle F'(z)h,\ v\rangle = \int_\Omega A\nabla h\cdot\nabla v + \int_{\Gamma_{\mathrm{N}}} (\alpha+4\beta|z+\bar u|^3)hv \qquad (\forall z,h,v\in V_h). \tag{7.21}$$

This shows that condition (i) holds trivially. To check conditions (ii)-(iii), set $v = h$:

$$\langle F'(z)h,\ h\rangle = \int_\Omega A\nabla h\cdot\nabla h + \int_{\Gamma_{\mathrm{N}}} (\alpha+4\beta|z+\bar u|^3)h^2. \tag{7.22}$$

Owing to (7.7) and $\alpha,\beta\ge 0$, we have

$$\langle F'(z)h,\ h\rangle \ge \int_\Omega A\nabla h\cdot\nabla h \ge \mu_0\|\nabla h\|^2_{L^2(\Omega)} = \mu_0\|h\|^2_{H^1_{\mathrm{D}}}.$$

Since $V_h\subset H^1(\Omega)$ is a finite dimensional subspace, we have $\|h\|_{H^1_{\mathrm{D}}}\ge c_1\|h\|_{V_{\mathrm{D}}}$ for some constant $c_1>0$, which yields

$$\langle F'(z)h,\ h\rangle \ge \mu_0 c_1^2\|h\|^2_{V_{\mathrm{D}}},$$

this yields condition (ii). Now let

$$\alpha_\infty := \|\alpha\|_{L^\infty(\Gamma_{\mathrm{N}})}(\mathrm{meas}(\Gamma_{\mathrm{N}}))^{3/5}, \quad \beta_\infty := 4\|\beta\|_{L^\infty(\Gamma_{\mathrm{N}})}.$$

We will use the following form of Hölder's inequality: with $\frac35+\frac25=1$,

$$\int_{\Gamma_{\mathrm{N}}} |v|^3h^2 \le \|v^3\|_{L^{5/3}(\Gamma_{\mathrm{N}})}\,\|h^2\|_{L^{5/2}(\Gamma_{\mathrm{N}})} = \|v\|^3_{L^5(\Gamma_{\mathrm{N}})}\,\|h\|^2_{L^5(\Gamma_{\mathrm{N}})} \qquad (\forall v,h\in L^5(\Gamma_{\mathrm{N}})).$$

Applying this to the second term of (7.22) yields

$$\begin{aligned}\int_{\Gamma_{\mathrm{N}}} (\alpha+4\beta|z+\bar u|^3)h^2 &\le \|\alpha\|_{L^\infty(\Gamma_{\mathrm{N}})}\textstyle\int_{\Gamma_{\mathrm{N}}} 1^3h^2 + 4\|\beta\|_{L^\infty(\Gamma_{\mathrm{N}})}\int_{\Gamma_{\mathrm{N}}}|z+\bar u|^3h^2\\ &\le \left(\alpha_\infty+\beta_\infty\|z+\bar u\|^3_{L^5(\Gamma_{\mathrm{N}})}\right)\|h\|^2_{L^5(\Gamma_{\mathrm{N}})}\\ &\le \left(\alpha_\infty+\beta_\infty\left(\|z\|_{L^5(\Gamma_{\mathrm{N}})}+\|\bar u\|_{L^5(\Gamma_{\mathrm{N}})}\right)^3\right)\|h\|^2_{L^5(\Gamma_{\mathrm{N}})}.\end{aligned}$$

Using this together with (7.7) and (7.8), we can estimate (7.22) as

$$\langle F'(z)h,\ h\rangle \le \mu_1\|h\|^2_{H^1_{\mathrm{D}}(\Omega)} + \Big(\alpha_\infty + \beta_\infty\left(\|z\|_{L^5(\Gamma_{\mathrm{N}})} + \|\bar u\|_{L^5(\Gamma_{\mathrm{N}})}\right)^3\Big)\|h\|^2_{L^5(\Gamma_{\mathrm{N}})} \le \Lambda(\|z\|_{V_{\mathrm{D}}})\|h\|^2_{V_{\mathrm{D}}},$$

where

$$\Lambda(t) := \max\left\{\mu_1,\ \alpha_\infty + \beta_\infty\left(t + \|\bar u\|_{L^5(\Gamma_{\mathrm{N}})}\right)^3\right\}.$$

This establishes condition (iii). Finally, to obtain condition (iv), first observe that for each $z,\ w \in H^1(\Omega)$ the symmetry of $F'(z) - F'(w)$ makes it possible to obtain its norm using the quadratic form:

$$\|F'(z) - F'(w)\| = \sup_{\|h\|_{V_{\mathrm{D}}}=1} |\langle (F'(z) - F'(w))h, h\rangle|.$$

Applying (7.22) yields

$$\|F'(z)-F'(w)\| = 4\sup_{\|h\|_{V_{\mathrm{D}}}=1}\left|\int_{\Gamma_{\mathrm{N}}}\beta(|z+\bar u|^3 - |w+\bar u|^3)h^2\right| \le \beta_\infty \sup_{\|h\|_{V_{\mathrm{D}}}=1}\int_{\Gamma_{\mathrm{N}}}\left||z+\bar u|^3 - |w+\bar u|^3\right| h^2\,.$$

Here we have

$$\left||z+\bar u|^3 - |w+\bar u|^3\right| = \left||(z+\bar u)^3| - |(w+\bar u)^3|\right| \le |(z+\bar u)^3 - (w+\bar u)^3|,$$

therefore,

$$\|F'(z) - F'(w)\| \le \beta_\infty \sup_{\|h\|_{V_{\mathrm{D}}}=1}\int_{\Gamma_{\mathrm{N}}} |z-w|\left|(z+\bar u)^2 + (z+\bar u)(w+\bar u) + (w+\bar u)^2\right| h^2.$$

Using the same form of general Hölder inequality as above implies

$$\|F'(z) - F'(w)\| \le \beta_\infty\|z-w\|_{L^5(\Gamma_{\mathrm{N}})}\Big(\|z+\bar u\|^2_{L^5(\Gamma_{\mathrm{N}})}$$
$$+\|z+\bar u\|_{L^5(\Gamma_{\mathrm{N}})}\|w+\bar u\|_{L^5(\Gamma_{\mathrm{N}})} + \|w+\bar u\|^2_{L^5(\Gamma_{\mathrm{N}})}\Big)\sup_{\|h\|_{V_{\mathrm{D}}}=1}\|h\|^2_{L^5(\Gamma_{\mathrm{N}})}.$$

Using $\|\cdot\|_{L^5(\Gamma_{\mathrm{N}})} \le \|\cdot\|_{V_{\mathrm{D}}}$, this yields

$$\|F'(z) - F'(w)\| \le L(\max\{\|z\|_{V_{\mathrm{D}}}, \|w\|_{V_{\mathrm{D}}}\})\|z-w\|_{V_{\mathrm{D}}},$$

where

$$L(t) := 3\beta_\infty(t + \|\bar u\|_{L^5(\Gamma_{\mathrm{N}})})^2,$$

hence (iv) is satisfied. ∎

From the obtained properties we can draw the general conclusion:

Corollary 7.9. *If bounded symmetric linear operators $B_n : V_h \to V_h'$ satisfy condition* (7.17)*, then the quasi-Newton iteration* (7.16) *for the weak nonlinear elliptic operator* (7.20) *converges according to* (7.18)*.*

Clearly, the application of the result needs a suitable definition of the operators B_n. We give a reasonable choice in the next subsection.

7.4.3 Preconditioning operators

In this subsection we propose choices for the auxiliary operators B_n. We must fulfill a double goal: B_n should be a good approximation of $F'(z_n)$, but essentially simpler to realize.

The general idea is to compose B_n from precomputed parts, so that the stepwise updating needs a minimal computational task. In general, we can approximate the principal part using an arbitrary but spectrally equivalent matrix coefficient: let G be a function-valued matrix with entries $g_{ij} \in L^\infty(\Omega)$, such that there exists constants $\lambda_0,\ \lambda_1 > 0$ for which

$$\lambda_0\ G(x)v \cdot v \le A(x)v \cdot v \le \lambda_1\ G(x)v \cdot v \tag{7.23}$$

for all $x \in \Omega$ and vector $v \in \mathbb{R}^3$. For given $n \in \mathbb{N}$, using the already computed iterate z_n, let B_n be defined via the weak form

$$\langle B_n h,\ v\rangle := \int_\Omega G(x)\ \nabla h \cdot \nabla v\ + (\alpha_0 + w_n) \int_{\Gamma_{\rm N}} hv \qquad (\forall v, h \in V_h), \tag{7.24}$$

where, for some fixed $0 < \varrho, \tau \le 1$,

$$\alpha_0 := \varrho\, \max \alpha, \qquad w_n := 4\tau \max\{\beta |z_n + \bar{u}|^3\}.$$

Then the discretization matrix corresponding to B_n has the form

$$\mathbf{B}_n = \mathbf{G} + (\alpha_0 + w_n)\mathbf{M},$$

where $\mathbf{G}$ is the stiffness matrix weighted with G and $\mathbf{M}$ is the boundary mass matrix on $\Gamma_{\rm N}$, both precomputable.

We will make use of the following Sobolev embedding estimate: there exists a constant $C_2 > 0$ such that

$$\|v\|_{L^2(\Gamma_{\rm N})} \le C_2 \|\nabla v\|_{L^2(\Omega)} \qquad (\forall v \in H^1_{\rm D}(\Omega)). \tag{7.25}$$

This is due to the property $v_{|\Gamma_{\rm D}} = 0$. Since $V_h \subset V_{\rm D} \subset H^1_{\rm D}(\Omega)$, we can apply (7.25) in our FEM subspace.

Now we verify that the B_n satisfy the corresponding condition in Theorem 7.6.

Theorem 7.10. *The spectral equivalence (7.17) holds for operators (7.21) and (7.24) with constants*

$$m_n = 1\Big/\Big(\frac{1}{\lambda_0} + \frac{(\alpha_0 + w_n)C_2^2}{\mu_0}\Big) \quad \textit{and} \quad M_n \equiv M := \max\Big\{\lambda_1,\ \frac{1}{\varrho},\ \frac{1}{\tau}\Big\}.$$

Proof. We have

$$\langle B_n h,\ h\rangle := \int_\Omega G(x)\ \nabla h \cdot \nabla h\ + (\alpha_0 + w_n) \int_{\Gamma_{\rm N}} h^2 \qquad (\forall h \in V_h).$$

We can use (7.7) with (7.22) and get

$$\begin{aligned}\langle F'(z)h,\ h\rangle &= \int_\Omega A\nabla h\cdot\nabla h + \int_{\Gamma_{\mathrm{N}}}(\alpha+4\beta|z+\bar u|^3)h^2\\ &\le \lambda_1\int_\Omega G(x)\ \nabla h\cdot\nabla h + \big(\max\alpha + 4\max\{\beta|z_n+\bar u|^3\}\big)\int_{\Gamma_{\mathrm{N}}}h^2\\ &\le \max\Big\{\lambda_1,\ \frac{1}{\varrho},\ \frac{1}{\tau}\Big\}\Big(\int_\Omega G(x)\ \nabla h\cdot\nabla h\ + (\alpha_0+w_n)\int_{\Gamma_{\mathrm{N}}}h^2\Big)\\ &= M\langle B_n h,\ h\rangle,\end{aligned}$$

where $M := \max\Big\{\lambda_1,\ \frac{1}{\varrho},\ \frac{1}{\tau}\Big\}$. Note that now $M_n \equiv M$ is independent of n.

For the other direction, we first note that from (7.23),

$$\int_\Omega G(x)\ \nabla h\cdot\nabla h \le \frac{1}{\lambda_0}\int_\Omega A(x)\ \nabla h\cdot\nabla h\,,$$

further, the Sobolev estimate (7.25) and (7.7) yield

$$\begin{aligned}(\alpha_0+w_n)\int_{\Gamma_{\mathrm{N}}}h^2 &= (\alpha_0+w_n)\,\|h\|^2_{L^2(\Gamma_{\mathrm{N}})}(\alpha_0+w_n)C_2^2\,\|\nabla h\|^2_{L^2(\Omega)}\\ &\le \frac{(\alpha_0+w_n)C_2^2}{\mu_0}\int_\Omega A(x)\ \nabla h\cdot\nabla h\,.\end{aligned}$$

Adding these up, we obtain

$$\begin{aligned}\langle B_n h,\ h\rangle &\le \Big(\frac{1}{\lambda_0}+\frac{(\alpha_0+w_n)C_2^2}{\mu_0}\Big)\int_\Omega A(x)\ \nabla h\cdot\nabla h\\ &\le \Big(\frac{1}{\lambda_0}+\frac{(\alpha_0+w_n)C_2^2}{\mu_0}\Big)\ \Big(\int_\Omega A\nabla h\cdot\nabla h + \int_{\Gamma_{\mathrm{N}}}(\alpha+4\beta|z+\bar u|^3)h^2\Big)\\ &= \Big(\frac{1}{\lambda_0}+\frac{(\alpha_0+w_n)C_2^2}{\mu_0}\Big)\ \langle F'(z)h,\ h\rangle\end{aligned}$$

since, from $\alpha,\beta\ge 0$, the integral on Γ_{N} is nonnegative. That is, we have

$$m_n\langle B_n h,\ h\rangle \le \langle F'(z)h,\ h\rangle,$$

where

$$m_n := 1\Big/\Big(\frac{1}{\lambda_0}+\frac{(\alpha_0+w_n)C_2^2}{\mu_0}\Big)\,.$$

■

Theorem 7.11. *Using the operators (7.24), the quasi-Newton iteration* (7.16) *for the weak nonlinear elliptic operator* (7.20) *converges according to* (7.18).

Proof. The linear operators B_n in (7.24) are bounded and symmetric, and they satisfy condition (7.17) by Theorem 7.10. Hence Corollary 7.9 yields that the iteration (7.16) converges according to (7.18). ■

Remark 7.12. (Special cases.)

(i) If the matrix A has a simple structure, then there is no need to replace it, hence we can let $G := A$. In this case
$$\lambda_0 = \lambda_1 = 1 \, .$$

(ii) If the matrix A has a large variation, and/or the domain has symmetries, then one may approximate the operator with a constant times Laplacian. In this case we can let $G := \tilde{\mu}\, I$, where, using (7.7), $\tilde{\mu} := \frac{\mu_0+\mu_1}{2}$, and I is the identity matrix. This suggestion replaces the coefficients in $F'(z_n)$ by constant scalars. Then B_n corresponds to the FEM discretization of linear elliptic Poisson problems with mixed boundary conditions:
$$-\tilde{\mu}\,\Delta h = r \qquad \text{in } \Omega$$
$$h_{|\partial\Omega} = 0 \quad \text{on } \Gamma_{\mathrm{D}}, \qquad \frac{\partial h}{\partial \nu} + (\alpha_0 + w_n)h = \gamma \quad \text{on } \Gamma_{\mathrm{N}} \, .$$

Remark 7.13. (2D analogues.) In this section we focus on the practically realistic 3D situation. The 2D case may also be of interest, e.g. on thin domains or on cross-sections. We note that our proposed method and Theorem 7.11 is valid in 2D as well, moreover, then the situation is simpler, since we can just use $H^1_{\mathrm{D}}(\Omega)$ as underlying function space.

7.5 Numerical experiments

In this subsection we present the results of numerical experiments. Our goal is to reinforce the robust convergence provided by the theoretical results, and to compare the performance of the quasi-Newton method with the exact Newton method.

Following [58], we solve BVP (7.1)-(7.3) on the unit cube $\Omega = (0,\, 1)^3$ with boundary portions
$$\Gamma_{\mathrm{N}} := \{(x,\, y,\, z) \in \overline{\Omega} : \quad z = 1\}, \quad \Gamma_{\mathrm{D}} = \partial\Omega \setminus \Gamma_{\mathrm{N}}$$
and using the data of [58, Section 4]. However, to address the setting of the present paper, we include nondiagonal coefficients in the heat conductivity matrix, and let
$$A := 60 \cdot \begin{pmatrix} 1 & \mu & 0 \\ \mu & 1 & \mu \\ 0 & \mu & 1 \end{pmatrix},$$
where $0 < \mu < 1/\sqrt{2}$ (the upper bound is required for uniform positivity). The parameters are $\alpha = 90,\ \beta = 0.75 \cdot 5.669 \cdot 10^{-8}$. The heat sources in Ω and on Γ_{N} are
$$f(x, y, z) = 36000\pi^2 z \sin \pi x \sin \pi y$$
and
$$g(x, y, z) = 27000 + 45000 \sin \pi x \sin \pi y + 344.39175(1 + \sin \pi x \sin \pi y)^4,$$
respectively, while $\bar{u}(x, y, z) = \bar{u}$ constant on Γ_{D}. (These data in [58] give rise to an exact solution, which we reproduced for $\mu = 0$ and $\bar{u} = 300$. However, we are now interested in the more general case.) We wish to study the numerical behaviour with varying μ and $\bar{u}$, that is, how sensitive the method is to the measure of non-diagonality and to the prescribed

temperature. The latter is important since higher temperatures yield an overall greater role of the nonlinear term, due to the used fourth power.

We applied trilinear finite elements. To obtain the mesh, we defined values k_i and then applied a uniform mesh with mesh parameters $h_i = \frac{1}{k_i+1}$, thus the number of degrees of freedom (DoF) can be calculated as $k_i^2(k_i+1)$. We chose four different meshes, corresponding to k_i values 14, 20, 30 and 40, resulting in DoF 2940, 8400, 27900 and, 65600, respectively. The integration was done with the midpoint rule, and the auxiliary equation was solved by a direct solver. The stopping criterion was the standard Sobolev norm going below 10^{-6}. The initial condition was the constant 0 function.

For each iteration step n, the preconditioner used for the quasi-Newton method is the following:

$$\langle B_n h,\ h\rangle = \int_\Omega A\,\nabla h\cdot\nabla h + \alpha\int_{\Gamma_N} h^2 + \max\{3\beta|z_n+\bar{u}|^3\}\int_{\Gamma_N} h^2$$

(that is, $G = A$, $\varrho = 1$ and $\tau = 3/4$). Here the first two terms and the integral in the third term do not depend on n. Thus, for every n, the stiffness matrix can be assembled from 2 precomputed matrices just using a linear combination.

The iteration numbers of (the exact) Newton method and the quasi-Newton method can be seen in Tables 7.1 and 7.2, respectively, for certain values of μ and $\bar{u}$ and for different mesh sizes. Apparently both methods are robust w.r.t. the mesh size, and slightly sensitive to the other parameters.

	$\mu = 0.2$			$\mu = 0.4$		
DoF	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$
2940	3	3	4	3	3	4
8400	3	3	4	3	3	4
27900	3	3	4	3	3	4
65600	3	3	4	3	3	4

Table 7.1: Number of iterations using Newton's method.

	$\mu = 0.2$			$\mu = 0.4$		
DoF	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$
2940	3	4	4	3	4	4
8400	3	3	4	3	3	4
27900	3	3	4	3	3	4
65600	3	3	4	3	3	4

Table 7.2: Number of iterations using the quasi-Newton method.

The ratios of the total runtimes (that is, not for just an individual iteration step but for the whole iteration) can be seen in Table 7.3. In particular, ratios of runtimes of assembling the stiffness matrix are shown in Table 7.4. We may observe that in most cases the quasi-Newton method consumed less overall runtime, and this is mostly due to the significantly cheaper assembly of stiffness matrices, thanks to their simplified structure.

	$\mu = 0.2$			$\mu = 0.4$		
DoF	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$
2940	0.9009	1.1895	0.8843	0.8884	1.1850	0.8881
8400	0.8863	0.8819	0.8855	0.8923	0.8728	0.8830
27900	0.9095	0.9082	0.9056	0.9048	0.9199	0.9029
65600	0.9086	0.9083	0.9103	0.9068	0.9126	0.9108

Table 7.3: Ratio of total runtimes: $t_{\mathrm{qN}}/t_{\mathrm{N}}$.

	$\mu = 0.2$			$\mu = 0.4$		
DoF	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$
2940	0.0272	0.0268	0.0268	0.0273	0.0272	0.0295
8400	0.0126	0.0123	0.0127	0.0131	0.0127	0.0129
27900	0.0049	0.0048	0.0047	0.0045	0.0045	0.0046
65600	0.0023	0.0022	0.0024	0.0023	0.0023	0.0024

Table 7.4: Ratio of average runtimes of assembling the stiffness matrix for an individual iteration step: $t_{\mathrm{as,qN}}/t_{\mathrm{as,N}}$.

Finally, the obtained numerical solutions allow us to examine the effects of some features of the problem. First, Table 7.5 reflects the effect of the nonlinearity in the Stefan–Boltzmann condition. Here (for the same parameters μ and $\bar{u}$ as above) we give the maximal and minimal values of the homogenized solution z_n for the "linear" problem, in which the fourth power term is neglected (that is, we set $\beta = 0$) and for the "nonlinear" problem, which is the same as throughout this subsection (that is, we set $\beta = 0.75 \cdot 5.669 \cdot 10^{-8}$). The function z_n here corresponds to our finest mesh with DoF 65600.

	$\mu = 0.2$			$\mu = 0.4$		
	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$	$\bar{u} = 300$	$\bar{u} = 600$	$\bar{u} = 1500$
$nonlin \max z_n$	253.57	147.89	58.94	196.44	107.85	49.83
$lin \max z_n$	262.38	178.61	82.25	201.35	130.32	61.83
$nonlin \min z_n$	0	-24.38	-360.02	0	-25.02	-357.75
$lin \min z_n$	0	-19.13	-152.66	0	-19.98	-158.26

Table 7.5: Maximal and minimal values of the homogenized solution z_n for linear/nonlinear problems for DoF = 65600.

Further, Figures 7.1 and 7.2 show colourmaps of the numerical solution. On Figure 7.1 we visualize the solution obtained for some specific parameters, whereas Figure 7.2 illustrates how the solution varies on the radiating upper subsurface as we modify the anisotropic parameter μ.

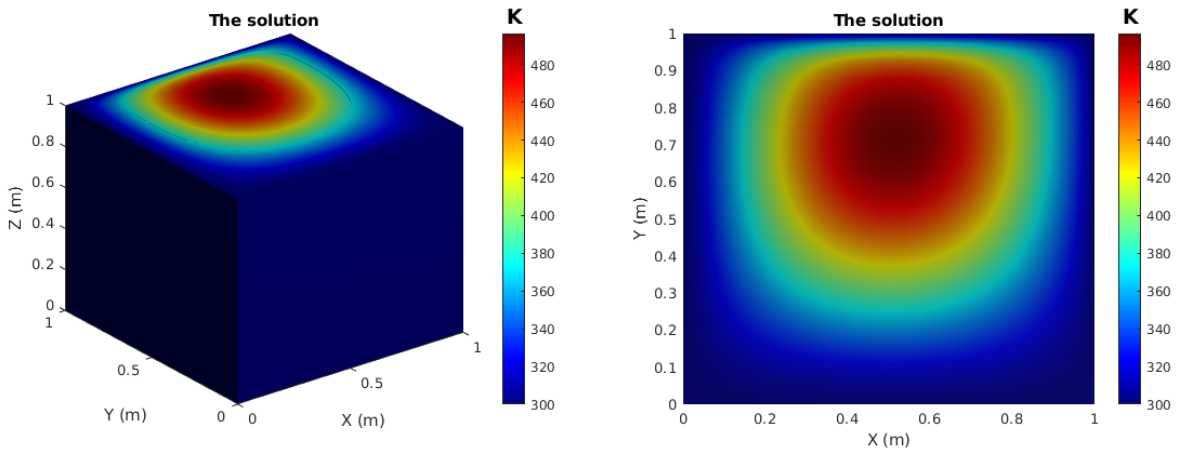

Figure 7.1: Heat colourmap of the numerical solution $u_n = z_n + \bar{u}$ on the whole cube and on the subsurface Γ_{N}, respectively, for DoF $= 65600$, $\bar{u} = 300$, and $\mu = 0.4$.

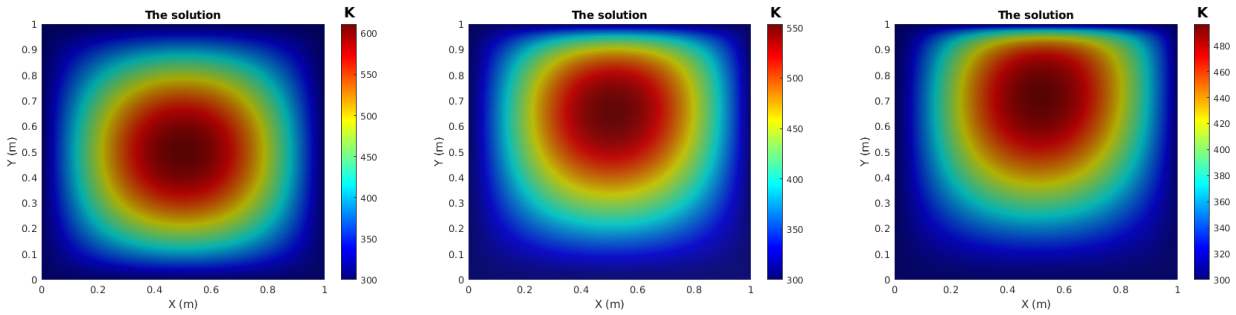

Figure 7.2: Heat colourmap of the numerical solution $u_n = z_n + \bar{u}$ on the subsurface Γ_{N} for DoF $= 65600$, $\bar{u} = 300$, for varied values of μ: respectively, $\mu = 0$, $\mu = 0.2$, $\mu = 0.4$.

8 Some practical suggestions

If one

(i) is developing a FEM software, and is interested in nonlinear problems (has already optimized a general nonlinear solver),

(ii) has a nonlinear problem, which is used by client(s) year after year, and reaction time is imporant, furthermore,

(iii) the parameters of the model and/or the boundary conditions are not expected to suffer great changes,

(iv) one can arrange the derivative of the operator be obtained,

we may encourage one to consider developing a quasi-Newton method, possibly with a manual option for the user to control a steplength.

Remark 8.1. *Exclusion criterion: problems close to singular blow-up.*

Remark 8.2. *One may use quasi-Newton for the first few steps, and Newton for the rest, as the latter is favored close to the solution for its quadratic convergence.*

Remark 8.3. *Varying steplength might seem the simplest way for further development for any method, provided that (i)-(iv) hold, even if an adaptive algorithm is used.*

Remark 8.4. *To obtain a quasi-Newton auxiliary operator from the derivative, see e.g. Subsection 3.3.1, where (3.29), including (3.28) is substituted by much simpler (3.30).*

9 Conclusion

In this dissertation, it is shown that previous results for the convergence of quasi-Newton methods can be extended to much more general settings, with relaxed ellipticity and Lipschitz conditions, and replacing the Hilbert space with a Banach space setting. Namely, similar favorable results hold as for the strict conditions.

Firstly, the upper ellipticity condition is relaxed, together with the relaxation of the Lipschitz condition in Hilbert space.

Then, a Banach space setting is employed, and the lower bound is relaxed. the results for both local and global convergence hold. A detailed classification is presented for the models falling under the assumptions used.

Additionally, inner-outer iterations are investigated in Banach space with preconditioned conjugate gradient method used in inner iterations without damping, yielding a favorable local convergence result.

A one-dimensional fourth-order nonlinear beam model is studied with a detailed presentation of the usage of Sobolev gradient method, quasi-Newton method, and full Newton's method.

A boundary nonlinearity is presented with results of nonnegativity and finite element approximation. Additionally, the applicability of quasi-Newton methods is shown and corresponding preconditioners are suggested.

The theoretical achievements are supported with several simulation results.

One can observe that the studied quasi-Newton methods perform perfect robustness for essentially all of the investigated nonlinearities.

Experimental damping coefficients are crucial to such work. With these coefficients involved in the methods under investigation, quasi-Newton methods can be faster than full-Newton method regarding overall runtime, i.e. more efficient with respect to computational cost.

References

[1] Adams, R.A., Fournier, J. F., *Sobolev Spaces*, Second edition, Pure and Applied Mathematics 140, Elsevier/Academic Press, Amsterdam, 2003.

[2] Andreianov, B., Boyer, F., Hubert, F., Finite volume schemes for the p-Laplacian on Cartesian meshes, *M2AN,* vol. 38 (2004), no. 6, pp. 931–959.

[3] Andrews, K.T. et al.,, Analysis and simulations of a nonlinear dynamic beam, *Z. Angew. Math. Phys.,* vol. 63 (2012), no. 6, pp. 1005–1019.

[4] Antal, I., Karátson, J., Mesh independent superlinear convergence of an inner-outer iterative method for semilinear elliptic interface problems, *J. Comput. Appl. Math.,* vol. 226 (2009), no. 2, pp. 190–196.

[5] Axelsson, O., *A mixed variable finite element method for the efficient solution of nonlinear diffusion and potential flow equations*, in Advances in Multi-grid Methods, Notes Numer. Fluid Mech. 11, D. Braess, W. Hackbusch, and U. Trottenberg, eds., Viewig, Braunschweig, 1985, pp. 1–11.

[6] Axelsson, O., *On global convergence of iterative methods*, in Iterative Solution of Nonlinear Systems of Equations, Lecture Notes in Math. 953, Springer, Berlin, 1982, pp. 1—19.

[7] Axelsson, O., On Mesh Independence and Newton-Type Methods, *Appl. Math.*, vol. 38 (1993), no. 4-5, pp. 249–265.

[8] Axelsson, O., Barker, V. A., *Finite Element Solution of Boundary Value Problems: Theory and Computation,* 0898714990, Society for Industrial and Applied Mathematics, 2001.

[9] Axelsson, O., Gustafsson, I., *An efficient finite element method for nonlinear diffusion problems*, Bull. Greek Math. Soc., 32 (1991), pp. 45–61.

[10] Axelsson, O., Layton, W., A two-level method for the discretization of non-linear boundary value problems, *SIAM J. Numer. Anal.,* vol. 33 (1996), no. 6, pp. 2359–2374.

[11] Axelsson, O., Margenov, S., On multilevel preconditioners which are optimal with respect to both problem and discretization parameters, *Comput. Methods Appl. Math.,* vol. 3 (2003), no. 1, 6–22.

[12] Axelsson, O., Maubach, J., On the updating and assembly of the Hessian matrix in finite element methods, *Comput. Methods Appl. Mech. Engrg.*, vol. 71 (1988), pp. 41–67.

[13] Bai, Zh., Huang, B., Ge, W., The iterative solutions for some fourth-order p-Laplace equation boundary value problems, *Appl. Math. Letters,* vol. 19 (2006), no. 1, pp. 8–14.

[14] Banz, L., Lamichhane, B. P., Stephan, E. P., Higher order mixed FEM for the obstacle problem of the p-Laplace equation using biorthogonal systems, *Comput. Methods Appl. Math.,* vol. 19 (2018), no. 2, pp. 169–188.

[15] Barrett, J.W., Liu, J.G., Quasi-norm error bounds for the finite element approximation of a non-Newtonian flow, *Numer. Math.,* vol. 68 (1994), no. 4, pp. 437–456.

[16] Beck, L., Bulíček, M., Málek, J., Süli, E., On the existence of integrable solutions to nonlinear elliptic systems and variational problems with linear growth, *Arch. Rational Mech. Anal.,* vol. 225 (2017), no. 2, pp. 717–769.

[17] Borsos, B., Karátson, J., Variable preconditioning for strongly nonlinear elliptic problems, *J. Comput. Appl. Math.,* vol. 350 (2019), pp. 155–164.

[18] Borsos, B., Karátson, J., Quasi-Newton Variable Preconditioning for non-uniformly monotone elliptic problems posed in Banach Spaces, *IMA J. Numer. Anal.*, 2021, https://doi.org/10.1093/imanum/drab024

[19] Borsos, B., An inexact Newton method with inner preconditioned CG for non-uniformly monotone elliptic problems, to appear in *Mathematical Modelling and Analysis*, 2021.

[20] Borsos, B., Karátson, J., Robust iterative solvers for nonlinear Gao beam models in elasticity, *Computational Methods in Applied Mathematics*, 2021, https://doi.org/10.1515/cmam-2020-0133.

[21] Borsos, B., Karátson, J., Numerical solution of nonlinear anisotropic elliptic Stefan-Boltzmann heat radiation problems in 3D using Newton type methods, *Computers and Mathematics with Applications*, submitted, 2021.

[22] Boyd, R., *Nonlinear Optics*, Academic Press, 2008.

[23] Böhmer, K., *Numerical methods for nonlinear elliptic differential equations,* Oxford University Press, Oxford, 2010.

[24] Busuioc, V., Cioranescu, D., On a class of electrorheological fluids. Contributions in honor of the memory of Ennio De Giorgi, *Ricerche Mat.,* 49 (2000), suppl., 29–60.

[25] Cai, X.-C., Keyes, D., Marcinkowski, L., Nonlinear additive Schwarz preconditioners and applications in computational fluid dynamics, *Internat. J. Numer. Meth. Fluid Mech.,* 40 (2002), 1463–1470.

[26] Chebotarev, A. Yu., Kovtanyuk, A. E., Botkin, N, D., Problem of radiation heat exchange with boundary conditions of the Cauchy type, *Commun. Nonlinear Sci. Numer. Simul.,* vol. 75 (2019), 262–269.

[27] Ciarlet, Ph., *Linear and nonlinear functional analysis with applications,* SIAM, Philadelphia, 2013.

[28] Ciarlet, Ph., *The finite element method for elliptic problems,* North-Holland, Amsterdam, 1978.

[29] Cioranescu, D., Girault, V., Rajagopal, K. R., *Mechanics and mathematics of fluids of the differential type*, Advances in Mechanics and Mathematics 35, Springer, 2016.

[30] Deuflhard, P., Global inexact Newton methods for very large scale nonlinear problems, *Impact Comput. Sci. Engrg.,* vol. 3, no. 4: 366-393, 1991.

[31] Deuflhard, P., Weiser, M., Global inexact Newton multilevel FEM for nonlinear elliptic problems, in Multigrid Methods V, *Lect. Notes Comput. Sci. Eng.,* vol. 3 (1998), pp. 71–89, Springer, Berlin.

[32] Dillard, D. A., Mukherjee, B., Karnal, P., Batra, R. C., Frechette, J., A review of Winkler's foundation and its profound influence on adhesion and soft matter applications, *Soft Matter*, vol. 14 (2018), 3669–3683.

[33] Endtmayer, B., Langer, U., Neitzel, I., Wick, T., Wollner, W., Mesh adaptivity and error estimates applied to a regularized p-Laplacian constrained optimal control problem for multiple quantities of interest, *Proc. Appl. Math. Mech. (PAMM)*, vol. 19 (2019), no. 1, e201900231.

[34] Eringen, A. C., Maugin, G. A., *Electrodynamics of Continua I,* Springer, 1990.

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

[36] Fowler, A., *Mathematical Geoscience*, Springer, 2011.

[37] Gajewski, H., Gröger, K., Zacharias, K., *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[38] Gan, Y.X. (ed.), *Continuum Mechanics: Progress in Fundamentals and Engineering Applications*, IntechOpen, 2012.

[39] Gao, D.Y., Finite deformation beam models and triality theory in dynamical post-buckling analysis, *Int. J. Non-Linear Mech.,* vol. 35 (2000), no. 1, pp. 103–131.

[40] Gao, D.Y., Nonlinear elastic beam theory with application in contact problems and variational approaches, *Mech. Res. Commun.,* vol. 23 (1996), pp. 11–17.

[41] Gao, D.Y., Machalová, J., Solution of contact problems for Gao beam and elastic foundation, *Mathematics and Mechanics of Solids.,* vol. 23, no. 3, pp. 473–488.

[42] Gao, D.Y., Machalová, J., Netuka. H., Mixed finite element solutions to contact problems of nonlinear Gao beam on elastic foundation, *Nonlin. Anal. Real World Appl.,* vol. 22 (2015), pp. 537–550.

[43] Glowinski, R., *Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems*, SIAM, Philadelphia, 2015.

[44] Glowinski, R., Marrocco, A., Sur l'approximation par éléments finis d'ordre 1, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires, *C. R. Acad. Sci. Paris Sér. A,* vol. 9 (1975), pp. 41–76.

[45] Glowinski, R., Rappaz, J., Approximation of a nonlinear elliptic problem arising in a non-Newtonian fluid flow model in glaciology, *ESAIM: Mathematical Modelling and Numerical Analysis,* vol. 37 (2003), no. 1, pp. 175–186.

[46] Götz, T., Parhusip, H.A., On an asymptotic expansion for Carreau fluids in porous media, *J. Engrg Math.,* vol. 51 (2005), no. 4, pp. 351–365.

[47] Haslinger, J., Hlaváček, I., Nečas, J., Numerical methods for unilateral problems in solid mechanics, *Handbook of Numerical Analysis,* vol. 4 (1996), North-Holland, Amsterdam, 1996, pp. 313–485.

[48] Hirn, A., Finite element approximation of singular power-law systems, *Math. Comp.,* vol. 82 (2013), pp. 1247–1268.

[49] Janela, J., Moura, A., Sequeira, A., A 3D non-Newtonian fluid-structure interaction model for blood flow in arteries, *J. Comput. Appl. Math.,* vol. 234 (2010), no. 9, pp. 2783–2791.

[50] Karátson, J., Characterizing mesh independent quadratic convergence of Newton's method for a class of elliptic problems, *SIAM J. Math. Anal.,* vol. 44 (2012), no. 3, pp. 1279–1303.

[51] Karátson, J., On the Lipschitz continuity of derivatives for some scalar nonlinearities, *J. Math. Anal. Appl.,* vol. 346 (2008), no. 1, pp. 170–176.

[52] Karátson J., Faragó I., Variable preconditioning via quasi-Newton methods for nonlinear problems in Hilbert space, *SIAM J. Numer. Anal.*, vol. 41 (2004), no. 4, pp. 1242-1262.

[53] Karátson, J., Kovács, B., Variable preconditioning in complex Hilbert space and its application to the nonlinear Schrődinger equation, *Comput. Math. Appl.,* vol. 65 (2013), no. 3, 449–459.

[54] Karátson, J., Korotov, S., Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, *Numer. Math.,* vol. 99 (2005), no. 4, pp. 669–698.

[55] Ko, S., Süli, E., Finite element approximation of steady flows of generalized Newtonian fluids with concentration-dependent power-law index, *Math. Comp.,* vol. 88 (2019), no. 317, pp. 1061–1090.

[56] Křížek, M., Liu, L., Neittaanmäki, P., Finite element analysis of a nonlinear elliptic problem with a pure radiation condition, *Appl. Nonlin. Anal.* (Sequeira et. al, Ed.), Kluwer Academic, New York, pp. 271–280 (1999).

[57] Křižek, M., Neittaanmäki, P., *Mathematical and Numerical Modeling in Electrical Engineering: Theory and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[58] Liu, L., Huang, M., Yuan, K., Křížek, M., Numerical Approximation of a Nonlinear 3D Heat Radiation Problem, *Adv. Appl. Math. Mech.*, vol. 1 (2009), no. 1, pp. 125–139.

[59] Machalová, J., Netuka. H., Solution of contact problems for Gao beam and elastic foundation, *Math. Mech. Solids,* vol. 23 (2018), no. 3, pp. 473–488.

[60] Mang, A., Biros, G., A semi-Lagrangian two-level preconditioned Newton-Krylov solver for constrained diffeomorphic image registration, *SIAM J. Sci. Comput.,* vol. 39 (2017), no. 6, B1064-B1101.

[61] Mikhlin, S.G., *The Numerical Performance of Variational Methods*, Walters-Noordhoff, Groningen, The Netherlands, 1971.

[62] Mustafa, M., Mushtaq, A., Hayat, T., Alsaedi, A., Model to study the non-linear radiation heat transfer in the stagnation-point flow of power-law fluid, *Internat. J. Numer. Methods Heat Fluid Flow,* vol. 25 (2015), no. 5, pp. 1107–1119.

[63] Nayfeh, A., Pai, P., *Linear and Nonlinear Structural Mechanics*, John Wiley & Sons, 2004.

[64] Neuberger, J. W., *Sobolev gradients and differential equations,* Second edition, Lecture Notes in Mathematics, 1670. Springer-Verlag, Berlin, 2010.

[65] Picasso M., Rappaz J., Reist A., Funk M., Blatter H., Numerical simulation of the motion of a two dimensional glacier, *Int. J. Numer. Methods Eng.,* vol. 60 (2004), pp. 995–1009.

[66] Plastock, R., Homeomorphisms between Banach spaces, *Trans. Amer. Math. Soc.,* vol. 200 (1974), pp. 169–183.

[67] Qatanani, N., Alzeer, I., A new approach for the computation of the visibility function for heat radiation problem, *Int. J. Math. Comput. Sci.,* vol. 2 (2007), no. 1, pp. 49–64.

[68] Qatanani, N., Alzeer, I., On the fast matrix computation for the heat radiation integral equation, *Int. J. Math. Comput. Sci.,* vol. 1, no. 4, pp. 461–472.

[69] Reddy, J.N., *An Introduction to Nonlinear Finite Element Analysis*, Oxford Univ. Press, Oxford, 2004.

[70] Rieder, A., Inexact Newton regularization using conjugate gradients as inner iteration, *SIAM J. Numer. Anal.,* vol. 43 (2005), no. 2, pp. 604-622.

[71] Rossi, T., Toivanen, J., Parallel fictitious domain method for a non-linear elliptic Neumann boundary value problem, *Numer. Linear Algebra Appl.,* vol. 6 (1999), no. 1, pp. 51–60.

[72] Roubíček, T., *Nonlinear Partial Differential Equations with Applications*, Volume 153 of ISNM, Birkhauser Verlag, 2005.

[73] Seaïd, M., EL-Amrani, M., Finite element P1 solution of unsteady thermal flow past a circular cyliner with radiation, *Int. J. Comput. Math.,* vol. 85 (2008), no. 3-4, pp. 641–656.

[74] Stoykov, S., Hofreither, C., Margenov, S., Isogeometric analysis for nonlinear dynamics of Timoshenko beams, *Numerical methods and applications*, 138–146, Lecture Notes in Comput. Sci. 8962, Springer, 2015.

[75] Strang, G., *Computational Science and Engineering*, Wellesley-Cambridge Press, MA, 2007.

[76] Toulopoulos, I., Wick, T., Numerical methods for power-law diffusion problems, *SIAM J. Sci. Comput.,* vol. 39 (2017), no. 3, pp. A681-A710.

[77] Wang, C., Reddy, J., Lee, K., *Shear deformable beams and plates*, Elsevier, 2000.

[78] Waterhouse, W. C., The absolute-value estimate for symmetric multilinear forms, *Linear Algebra Appl.,* vol. 128 (1990), 97–105.

[79] Xie, F., Wu, Q.-B., Dai, P.-F., Modified Newton-SHSS method for a class of systems of nonlinear equations, *Comput. Appl. Math.,* vol. 38 (2019), no. 1, paper no. 19.

[80] Xu, J., A novel two-grid method for semilinear elliptic equations, *SIAM J. Sci. Comput.,* vol. 15 (1994), no. 1, pp. 231-237.

[81] Zeidler, E., *Nonlinear functional analysis and its applications,* Springer, 1986