Variable preconditioning for strongly nonlinear elliptic problems

by B. Borsos[1], J. Karátson[2]

Abstract

Variable preconditioning has earlier been developed as a realization of quasi-Newton methods for elliptic problems with uniformly bounded nonlinearities. This paper presents a generalization of this approach to strongly nonlinear problems, first on an operator level, then for elliptic problems allowing power order growth of nonlinearities. Numerical tests reinforce the convergence results.

Key words. Variable preconditioning, quasi-Newton methods, iterative methods, nonlinear elliptic problems

AMS subject classifications. 65N30, 47N20.

1 Introduction

Nonlinear elliptic problems arise in various applications for models that describe stationary states. We may mention, for instance, elastoplasticity, magnetic potential equations, and flow problems in physics and other fields, see, e.g., [4, 6, 7, 10] and the references there. 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 [1, 5, 11]. A general approach to construct quasi-Newton methods has been given in [9], where approximate Jacobians are defined via spectral equivalence, and hence they can be regarded as variable preconditioners. This method, as well as much of the mentioned theory, is applicable to problems with uniformly bounded nonlinearities that allow well-posedness in the underlying Hilbert function space.

The goal of this paper is to extend the above approach of variable preconditioning to problems with stronger nonlinearities without a 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 [9] to a class of unbounded nonlinearities in Section 2. Then the result is applied to a class of elliptic problems with power order nonlinearities in Section 3. Numerical tests reinforce the theoretical convergence results.

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

$$F(u) = 0$$

[1]Department of Analysis, Technical University; Budapest, Hungary

[2]Department of Applied Analysis & MTA-ELTE Numerical Analysis and Large Networks Research Group, ELTE University; Department of Analysis, Technical University; Budapest, Hungary

for a given nonlinear operator $F : H \to H$. Our goal is to extend Theorem 3.1 in [9] to nonlinearities without a 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 [9, Section 3], 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.

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

$$\lambda \|h\|^2 \le \langle F'(u)h, h\rangle \le \Lambda(\|u\|)\, \|h\|^2 \qquad (\forall u, h \in H). \tag{2.1}$$

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

$$\|F'(u) - F'(v)\| \le L(\max\{\|u\|, \|v\|\})\, \|u - v\| \qquad (\forall u,\, h \in H). \tag{2.2}$$

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

$$m\langle B_n h, h\rangle \le \langle F'(u_n)h, h\rangle \le M\langle B_n h, h\rangle \qquad (\forall h \in H). \tag{2.3}$$

Then the sequence, defined by

$$u_{n+1} := u_n - \frac{2}{M+m} B_n^{-1} F(u_n) \qquad (\forall n \in \mathbb{N}), \tag{2.4}$$

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

$$\|u_n - u^*\| \le C \left(\frac{M-m}{M+m}\right)^n \qquad (\forall n \in \mathbb{N}). \tag{2.5}$$

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 $\|\cdot\|_A$. The following properties are known for spectrally equivalent operators:

Lemma 2.1. [9] *Let A and B be bounded self-adjoint linear operators on H with positive lower bound, and let there exist constants $M \ge m > 0$ such that*

$$m\langle Bh, h\rangle \le \langle Ah, h\rangle \le M\langle Bh, h\rangle \qquad (\forall h \in H). \tag{2.6}$$

Then

$$m\langle A^{-1}h,h\rangle \le \langle B^{-1}h,h\rangle \le M\langle A^{-1}h,h\rangle \qquad (\forall h\in H) \tag{2.7}$$

and

$$\left\| I - \frac{2}{M+m}AB^{-1}\right\|_{A^{-1}} \le \frac{M-m}{M+m}\,. \tag{2.8}$$

Lemma 2.2. *The preconditioning operators in* (2.3) *have uniformly bounded inverses, namely,*

$$\|B_n^{-1}\| \le \lambda^{-1}M \qquad (\forall n\in\mathbb{N}). \tag{2.9}$$

Proof. The upper and lower estimates in (2.3) and (2.1), respectively, yield:

$$\lambda\|h\|^2 \le \langle F'(u_n)h,h\rangle \le M\langle B_nh,h\rangle \le M\|B_nh\|\|h\| \qquad (\forall h\in H). \tag{2.10}$$

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

Remark 2.1. 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 (2.2) implies the upper bound in (2.1): for any $u,h\in H$

$$\langle F'(u)h,h\rangle = \langle (F'(u)-F'(0))h,h\rangle + \langle F'(0)h,h\rangle \le \left(L(\|u\|)\|u\| + \|F'(0)\|\right)\|h\|^2\,,$$

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

$$\tilde\Lambda_* := \Lambda(\|u^*\|), \tag{2.11}$$

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

$$\tilde L_{n,n+1} := L(\max\{\|u_n\|,\|u_{n+1}\|\}), \qquad \tilde L_{n,*} := L(\max\{\|u_n\|,\|u_*\|\})\,. \tag{2.12}$$

Furthermore, we introduce the following energy norms:

$$\|h\|_u := \langle F'(u)^{-1}h,h\rangle^{1/2} \quad (\text{for given } u\in H), \qquad \|\cdot\|_* := \|\cdot\|_{u^*}\,, \qquad \|\cdot\|_n := \|\cdot\|_{u_n} \tag{2.13}$$

(for given $n\in\mathbb{N}$). It follows readily (with Lemma 2.1) that for fixed u the norms $\|\cdot\|_u$ and $\|\cdot\|$ are equivalent, namely:

$$\lambda^{1/2}\|h\|_u \le \|h\| \le \Lambda^{1/2}(\|u\|)\|h\|_u \qquad (\forall h\in H). \tag{2.14}$$

Two important special cases are

$$\|h\|_n \le \lambda^{-1/2}\|h\|, \qquad \|h\| \le \tilde\Lambda_*^{1/2}\|h\|_* \qquad (\forall h\in H). \tag{2.15}$$

The norms $\|\cdot\|_*$ and $\|\cdot\|_n$ are related by the following non-uniform extension of [9, Cor. 3.4]:

Lemma 2.3. *For all $h \in H$ we have*

$$\frac{1}{1+\mu_n(u_n)} \le \frac{\|h\|_*^2}{\|h\|_n^2} \le 1+\mu_n(u_n)\,, \tag{2.16}$$

where

$$\mu_n(u_n) := \tilde{L}_{n,*}\tilde{\Lambda}_*^{1/2}\lambda^{-2}\|F(u_n)\|_*\ . \tag{2.17}$$

Proof. The lower bound in (2.1) implies a corresponding lower estimate for the variation of F:

$$\|F(u)-F(v)\| \ge \lambda\|u-v\| \qquad (\forall u,v\in H). \tag{2.18}$$

This, together with the assumptions (2.1)–(2.2) on F', implies

$$\langle F'(u)h,\ h\rangle \le \langle F'(v)h,\ h\rangle(1+L(\max\{\|u\|,\ \|v\|\})\lambda^{-2}\|F(u)-F(v)\|) \qquad (\forall u,v,h\in H). \tag{2.19}$$

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 (2.15) and reversing the role of u^* and u_n, we obtain

$$\frac{1}{1+\mu_n(u_n)} \le \frac{\langle F'(u^*)h,h\rangle}{\langle F'(u_n)h,h\rangle} \le 1+\mu_n(u_n) \qquad (\forall h\in H). \tag{2.20}$$

From this, Lemma 2.1 yields the desired estimate. ∎

Proof of Theorem 2.1. 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., [6]. 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 (2.18) will yield that $\lim_{n\to\infty}u_n=u^*$ as well.

(2) The proof can be started in a similar way as in [9], but even in this first part we must follow more carefully the non-uniform constants. For given $n\in\mathbb{N}$ in the iteration,

$$F(u_{n+1}) = F(u_n)+F'(u_n)(u_{n+1}-u_n)+R(u_n). \tag{2.21}$$

In this situation, analogously to inequalities (18)–(19) in [9], one can estimate the linear part (from Lemma 2.1) and the remainder as follows, using that (owing to (2.1)) F' is Lipschitz continuous in the ball $B(0,\max\{\|u_n\|,\|u_{n+1}\|\})$ with a corresponding constant $\tilde{L}_{n,n+1}$:

$$\|F(u_n)+F'(u_n)(u_{n+1}-u_n)\|_n \le \frac{M-m}{M+m}\|F(u_n)\|_n\,, \tag{2.22}$$

$$\|R(u_n)\|_n \le \frac{2\tilde{L}_{n,n+1}}{\lambda^{1/2}(M+m)^2}\|B_n^{-1}F(u_n)\|^2. \tag{2.23}$$

Here a further estimation can be given: using that (2.9) and (2.15) yield

$$\|B_n^{-1}F(u_n)\| \le \lambda^{-1}M\|F(u_n)\| \le \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_* \tag{2.24}$$

and letting $K := \frac{2M^2\tilde{\Lambda}_*}{\lambda^{5/2}(M+m)^2}$, (2.23) implies

$$\|R(u_n)\|_n \le K\tilde{L}_{n,n+1}\|F(u_n)\|_*^2, \tag{2.25}$$

Let us estimate the $\|\cdot\|_n$-norm of $F(u_{n+1})$ in (2.21), using (2.22) and (2.25):

$$\|F(u_{n+1})\|_n \le \frac{M-m}{M+m}\|F(u_n)\|_n + K\tilde{L}_{n,n+1}\|F(u_n)\|_*^2, \tag{2.26}$$

hence Lemma 2.3 yields

$$\|F(u_{n+1})\|_* \le (1+\mu_n(u_n))^{1/2}\left(\frac{M-m}{M+m}(1+\mu_n(u_n))^{1/2} + K\tilde{L}_{n,n+1}\|F(u_n)\|_*\right)\|F(u_n)\|_*. \tag{2.27}$$

(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

$$\|F(u_{n+1})\|_* \le (1+\mu_n(u_n))\left(\frac{M-m}{M+m} + K\tilde{L}_{n,n+1}\|F(u_n)\|_*\right)\|F(u_n)\|_*. \tag{2.28}$$

By substituting the definition (2.17) of $\mu_n(u_n)$ and defining the real function

$$\varphi_n(t) := (1+\tilde{L}_{n,*}\,\tilde{\Lambda}_*^{1/2}\lambda^{-2}t)\left(Q + K\tilde{L}_{n,n+1}t\right), \qquad \text{where} \quad Q := \frac{M-m}{M+m}, \tag{2.29}$$

the estimate (2.28) can be reformulated as

$$\|F(u_{n+1})\|_* \le \varphi_n\left(\|F(u_n)\|_*\right)\|F(u_n)\|_*. \tag{2.30}$$

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 $\tilde{L}_{n,*}$. 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 (2.4) yields

$$\|u_{n+1}\| \le \|u_n\| + \frac{2}{M+m}\|B_n^{-1}F(u_n)\| \qquad (\forall u \in [u_n,\ u_{n+1}]), \tag{2.31}$$

hence a bound for $\max\{\|u_n\|, \|u_{n+1}\|\}$ can be obtained as a bound for the above r.h.s. Here, first, (2.18) yields $\|F(u_n) - F(0)\| \ge \lambda\|u_n\|$, hence, also using (2.15),

$$\|u_n\| \le \lambda^{-1}\left(\|F(u_n)\| + \|F(0)\|\right) \le \lambda^{-1}\left(\Lambda_*^{1/2}\|F(u_n)\|_* + \|F(0)\|\right). \tag{2.32}$$

Further, to estimate $\|B_n^{-1}F(u_n)\|$, we use Lemma 2.2 and (2.15) to derive for all $h \in H$ that $\|B_n^{-1}h\| \le \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|h\|_*$, hence

$$\|B_n^{-1}F(u_n)\| \le \lambda^{-1}M\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_*\,. \tag{2.33}$$

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

$$f(t) := \lambda^{-1}\tilde{\Lambda}_*^{1/2}\big(1 + \frac{2M}{M+m}\big)\, t \; + \lambda^{-1}\|F(0)\|\,. \tag{2.34}$$

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

$$\tilde{L}_{n,n+1} := L(\max\{\|u_n\|, \|u_{n+1}\|\}) \le L_f(\|F(u_n)\|_*), \tag{2.35}$$

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 (2.18),

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

further, $\|u_n\|$ has the larger bound (2.32), 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

$$\tilde{L}_{n,*} := L(\max\{\|u_n\|, \|u_*\|\}) \le L_g(\|F(u_n)\|_*) \tag{2.36}$$

using the increasing continuous functions

$$g(t) := \lambda^{-1}\tilde{\Lambda}_*^{1/2}t + \lambda^{-1}\|F(0)\|, \qquad L_g(t) := L(g(t)). \tag{2.37}$$

(6) Altogether, using (2.35) and (2.36), the function (2.29) can be estimated as

$$\varphi_n(t) \le (1 + L_g(t)\tilde{\Lambda}_*^{1/2}\lambda^{-2}t)\,(Q + KL_f(t)t) =: \varphi(t), \tag{2.38}$$

and accordingly, inequalities (2.30) and (2.38) result in

$$\|F(u_{n+1})\|_* \le \varphi\left(\|F(u_n)\|_*\right)\|F(u_n)\|_*\,, \tag{2.39}$$

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

(7) Now it can be shown, just as in [9], that if the initial guess satisfies

$$\varphi(\|F(u_0)\|_*) < 1, \tag{2.40}$$

then $\lim_{n\to\infty}\|F(u_n)\|_* = 0$. In fact, using notation $r := \varphi\left(\|F(u_0)\|_*\right)$, estimate (2.39) and the monotonicity of φ, one can derive by induction that

$$\|F(u_n)\|_* \le r^n\|F(u_0)\|_* \to 0 \qquad (\text{as } n \to \infty). \tag{2.41}$$

Here (2.15) implies $\lim\limits_{n\to\infty}\|F(u_n)\| = 0$ in the original norm too, and then, as mentioned in item

(1) of the proof, (2.18) yields that $\lim_{n\to\infty} u_n = u^*$ as well.

(8) It remains to show the estimate (2.5), 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

$$e_n := \|F(u_n)\|_*, \tag{2.42}$$

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

$$\lim_{n\to\infty} \varphi(e_n) = \varphi(0) = Q. \tag{2.43}$$

Further, by (2.39), the errors satisfy

$$e_n \le \left(\prod_{k=0}^{n-1} \varphi(e_k)\right) e_0 = \left(\prod_{k=0}^{n-1} \frac{\varphi(e_k)}{Q}\right) Q^n e_0 \qquad (\forall n \in \mathbb{N}). \tag{2.44}$$

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 (2.38) and introducing the notations $d_1 := \tilde{\Lambda}_*^{1/2}\lambda^{-2}L_g(e_0), \quad d_2 := KL_f(e_0)$, we obtain

$$\varphi(e_k) \le (1 + d_1 e_k)(Q + d_2 e_k). \tag{2.45}$$

This and (2.41) imply

$$\frac{\varphi(e_k)}{Q} \le (1 + d_1 e_k)(1 + d_3 e_k) = 1 + d_4 e_k + d_5 e_k^2 \le 1 + d_4 e_0 r^k + d_5 e_0^2 r^{2k}, \tag{2.46}$$

with constants $d_3 = \frac{d_2}{Q}, \quad d_4 = d_1 + d_3, \quad d_5 = d_1 d_3$. From here we can follow [9] to deduce the following upper estimate from (2.44):

$$e_n \le \exp\left(\frac{d_4 e_0}{1-r} + \frac{d_5 e_0^2}{1-r^2}\right) e_0 Q^n \equiv E e_0 Q^n. \tag{2.47}$$

This, together with (2.15) and (2.18), yields

$$\|u_n - u^*\| \le \lambda^{-1}\|F(u_n)\| \le \lambda^{-1}\tilde{\Lambda}_*^{1/2}\|F(u_n)\|_* =: \lambda^{-1}\tilde{\Lambda}_*^{1/2} e_n \le \lambda^{-1}\tilde{\Lambda}_*^{1/2} e_0 E Q^n, \tag{2.48}$$

hence (with constant $C := \lambda^{-1}\tilde{\Lambda}_*^{1/2} e_0 E$) we obtain (2.5). ∎

3 Application to power order nonlinear elliptic problems

In this section 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 $p \ge 3$ and $k_1, k_2 > 0$ be given constants, $g \in L^2(\Omega)$ a given function, and consider

the following boundary value problem:

$$\left\{ \begin{array}{rcl} -\mathrm{div}\big((k_1 + k_2|\nabla u|^{p-2})\ \nabla u\big) & = & g, \\ u_{|\partial\Omega} & = & 0, \end{array} \right. \tag{3.1}$$

Such a nonlinear operator, which is of regularized p-Laplacian type, arises, e.g., in electrorheological fluid models, where $p = 4$, see [2]. Problem (3.1) has a unique weak solution in the Sobolev space $W_0^{1,p}(\Omega)$, see, e.g., [13].

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

$$\int_\Omega (k_1 + k_2|\nabla u|^{p-2})\nabla u \cdot \nabla v = \int_\Omega gv \qquad (\forall v \in V_h). \tag{3.2}$$

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

3.1 Construction of the iteration

First we cast the problem into the setting of section 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

$$\langle u,\ v\rangle := \int_\Omega \nabla u \cdot \nabla v, \qquad \|u\|_{H_0^1} := \|\nabla u\|_{L^2}, \tag{3.3}$$

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

$$\|\nabla u\|_{L^p} \le \hat{c}\,\|\nabla u\|_{L^2} \qquad (\forall u \in V_h). \tag{3.4}$$

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

$$\langle F(u), v\rangle \equiv \int_\Omega (k_1 + k_2|\nabla u|^{p-2})\nabla u \cdot \nabla v - \int_\Omega gv \qquad (\forall u, v \in V_h). \tag{3.5}$$

Then the FEM problem (3.2) 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 2.1, 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 (3.5) can be written as

$$\langle F(u), v\rangle = \int_\Omega f(\nabla u) \cdot \nabla v - \int_\Omega gv, \tag{3.6}$$

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

$$f(\eta) := (k_1 + k_2|\eta|^{p-2})\eta \qquad (\eta \in \mathbb{R}^N). \tag{3.7}$$

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

$$\partial_\eta f(\eta) = k_2(p-2)|\eta|^{p-4}(\eta \cdot \eta^T) + (k_1 + k_2|\eta|^{p-2})I, \tag{3.8}$$

where $\eta \cdot \eta^T$ denotes the diadic matrix with entries $\eta_i\,\eta_j \quad (i,j = 1,...,N)$. Following [3, 6], the Gâteaux derivative $F'(u)$ satisfies

$$\langle F'(u)h, v\rangle = \int_\Omega \partial_\eta f(\nabla u)\nabla h \cdot \nabla v, \tag{3.9}$$

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 (3.8) 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

$$\langle B_n h, v\rangle \equiv \int_\Omega \left(k_1 + k_2|\nabla u_n|^{p-2}\right)\nabla h \cdot \nabla v \qquad (\forall h, v \in V_h). \tag{3.10}$$

Then we obtain the following sequence from (2.4). 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:

$$\begin{cases} \text{solve} \quad B_n z_n = F(u_n); \\ \text{let} \quad u_{n+1} := u_n - \frac{2}{M+m} z_n\,. \end{cases} \tag{3.11}$$

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

$$\int_\Omega \left(k_1 + k_2|\nabla u_n|^{p-2}\right)\nabla z_n \cdot \nabla v = \ell_n v \qquad (v \in V_h). \tag{3.12}$$

3.2 Convergence of the iteration

Proposition 3.1. *The nonlinear operator F, defined by (3.5), and the linear operators B_n, defined by* (3.10)*, satisfy the conditions of Theorem 2.1.*

Proof.

(1) The Jacobians (3.8) are symmetric, hence (3.9) is self-adjoint. First we check that F

satisfies (2.1) and (2.2). As mentioned in Remark 2.1, the upper bound in (2.1) can be omitted, since it follows from (2.2). The lower bound is straightforward with $\lambda := k_1$, namely, (3.8) and (3.9) with $v = h$ yield

$$\langle F'(u)h, h\rangle = \int_\Omega k_2(p-2)|\nabla u|^{p-4}(\nabla u \cdot \nabla h)^2 + \int_\Omega (k_1 + k_2|\nabla u|^{p-2})|\nabla h|^2 \geq k_1\|h\|^2_{H^1_0}. \quad (3.13)$$

Now we have to prove the local Lipschitz property (2.2) 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:

$$\|F'(u) - F'(v)\| = \sup_{\|h\|_{H^1_0}=1} |\langle (F'(u) - F'(v))h,\ h\rangle| = \sup_{\|h\|_{H^1_0}=1} \left| \int_\Omega (\partial_\eta f(\nabla u) - \partial_\eta f(\nabla v))\nabla h \cdot \nabla h \right|. \quad (3.14)$$

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

$$\left\| \frac{\partial^2 f(\eta)}{\partial \eta^2} \right\| = \sup_{|h|=1} \left| \frac{\partial^2 f(\eta)}{\partial \eta^2}(h,\ h,\ h) \right| \quad (3.15)$$

owing to their symmetry [12]. Such tensors are discussed in [8] including general nonlinearities of the following form:

$$f(\eta) = a(|\eta|^2)\eta, \quad (3.16)$$

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 [8] for (3.16) 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, (3.7) 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 (3.15) gives

$$\left\| \frac{\partial^2 f(\eta)}{\partial \eta^2} \right\| \leq c_2(p) \sup_{|h|=1} |\eta|^{p-3}|h|^3 = c_2(p)\,|\eta|^{p-3}. \quad (3.17)$$

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

$$\|\partial_\eta f(\eta_1) - \partial_\eta f(\eta_2)\| \le \sup_{\tilde\eta\in[\eta_1,\ \eta_2]} \left\|\frac{\partial^2 f}{\partial\eta^2}(\tilde\eta)\right\| \cdot |\eta_1-\eta_2| \le c_2(p)\max\{|\eta_1|^{p-3},|\eta_2|^{p-3}\}|\eta_1-\eta_2|. \tag{3.18}$$

Combining this with (3.14), we obtain

$$\|F'(u)-F'(v)\| \le \sup_{\|h\|_{H_0^1}=1} \int_\Omega \|\partial_\eta f(\nabla u) - \partial_\eta f(\nabla v)\|\ |\nabla h|^2 \tag{3.19}$$

$$\begin{aligned} &\le c_2(p) \sup_{\|h\|_{H_0^1}=1} \int_\Omega \max\{|\nabla u|^{p-3},|\nabla v|^{p-3}\}|\nabla u-\nabla v||\nabla h|^2 \le \\ &\le c_2(p) \sup_{\|h\|_{H_0^1}=1} \left(\int_\Omega |\nabla u|^{p-3}|\nabla u-\nabla v||\nabla h|^2 + \int_\Omega |\nabla v|^{p-3}|\nabla u-\nabla v||\nabla h|^2\right). \end{aligned} \tag{3.20}$$

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

$$\int_\Omega |f|^{p-3}|g_1\,g_2\,g_3| \le \|f\|_{L^p}^{p-3}\|g_1\|_{L^p}\|g_2\|_{L^p}\|g_3\|_{L^p} \qquad (\forall f,\ g_1,\ g_2,\ g_3\in L^p(\Omega)), \tag{3.21}$$

which yields

$$\|F'(u)-F'(v)\| \le c_2(p)\left(\|\nabla u\|_{L^p}^{p-3} + \|\nabla v\|_{L^p}^{p-3}\right)\|\nabla u-\nabla v\|_{L^p} \sup_{\|h\|_{H_0^1}=1} \|\nabla h\|_{L^p}^2. \tag{3.22}$$

Owing to (3.4), we can apply the estimate $\|\nabla z\|_{L^p} \le \hat c\,\|z\|_{H_0^1}$ in each norm above, thus we get

$$\begin{aligned} \|F'(u)-F'(v)\| &\le c_2(p)\,\hat c^p\left(\|u\|_{H_0^1}^{p-3} + \|v\|_{H_0^1}^{p-3}\right)\|u-v\|_{H_0^1} \sup_{\|h\|=1}\|h\|_{H_0^1}^2 \le \\ &\le 2c_2(p)\,\hat c^p\left(\max\{\|u\|_{H_0^1},\|v\|_{H_0^1}\}\right)^{p-3}\|u-v\|_{H_0^1}, \end{aligned} \tag{3.23}$$

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

$$\|F'(u)-F'(v)\| \le L(\max\{\|u\|_{H_0^1},\|v\|_{H_0^1}\})\,\|u-v\|_{H_0^1}, \qquad L(t) = c_p\,t^{p-3} \tag{3.24}$$

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

(2) We prove that the operator B_n in (3.10) satisfies (2.3) with proper uniform constants M and m. Lower estimation of (3.13) gives

$$\langle F'(u_n)h,h\rangle \ge \int_\Omega \left(k_1 + k_2|\nabla u_n|^{p-2}\right)|\nabla h|^2 = \langle B_n h,h\rangle, \tag{3.25}$$

and upper estimation of (3.13) with Cauchy-Schwarz inequality yields

$$\langle F'(u_n)h, h\rangle \le \int_\Omega k_2(p-2)|\nabla u_n|^{p-4}|\nabla u_n|^2|\nabla h|^2 + \int_\Omega (k_1 + k_2|\nabla u_n|^{p-2})|\nabla h|^2$$
$$= \int_\Omega (k_1 + k_2(p-1)|\nabla u_n|^{p-2})|\nabla h|^2 \le (p-1)\int_\Omega (k_1 + k_2|\nabla u_n|^{p-2})|\nabla h|^2 = (p-1)\langle B_n h, h\rangle\,. \tag{3.26}$$

Hence (2.3) holds with lower and upper bounds

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

Now we can readily formulate

Theorem 3.1. *The iteration* (3.11)*, defined in Subsection 3.1, converges locally according to the estimate*

$$\|u_n - u^*\|_{H_0^1} \le C\Big(1 - \frac{2}{p}\Big)^n \qquad (\forall n \in \mathbb{N}). \tag{3.27}$$

Proof. By Proposition 3.1, the conditions of Theorem 2.1 are satisfied, further, iteration (2.4) coincides with (3.11) in our situation. Hence (2.5) holds locally, and for the obtained bounds $m = 1$ and $M = p-1$ the convergence factor is

$$\frac{M-m}{M+m} = 1 - \frac{2}{p}. \qquad \blacksquare \tag{3.28}$$

Remark 3.1. Alternatively to (3.10), the preconditioner may be defined with a different constant $\tilde{k}_2 > 0$ instead of k_2:

$$\langle B_n z, v\rangle := \int_\Omega \Big(k_1 + \tilde{k}_2|\nabla u_n|^{p-2}\Big)\nabla z \cdot \nabla v, \tag{3.29}$$

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 3.1, 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 3.1.

Remark 3.2. The obtained result provides linear convergence. Based on the techniques of [9], it can be expected that one may obtain convergence of higher order up to 2 if a sharper approximation of the derivatives is available. This and other extensions are the subject of forthcoming research.

3.3 Numerical experiments

Consider the following boundary value problem:

$$\begin{cases} -\mathrm{div}\big((\chi_1 + \chi_2|\nabla u|^2)\ \nabla u\big) &= g, \\ u_{|\partial\Omega} &= 0, \end{cases} \tag{3.30}$$

where $\chi_1, \chi_2 > 0$ are given constants. Such a nonlinear operator arises, e.g., in electrorheological fluid models, see [2]. This problem, which describes a stationary fluid, is a special case of (3.1) 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 (3.11) with preconditioning operators (3.29):

$$\langle B_n h, v\rangle \equiv \int_\Omega \big(\chi_1 + \tilde\chi_2|\nabla u_n|^2\big)\,\nabla h\cdot\nabla v \qquad (\forall h, v \in V_h). \tag{3.31}$$

Since $p = 4$, here we let $\chi_2 \le \tilde\chi_2 \le 3\chi_2$ as suggested in Remark 3.1, and we obtain from (3.28) that the theoretical convergence factor is

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

independently of the constants χ_1, χ_2.

We have run the iteration (3.11)–(3.12) 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 1. 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.

		$\chi_2 = 10$			$\chi_2 = 100$			$\chi_2 = 1000$		
	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 1: Number of iterations to achieve $\varepsilon_n < 10^{-6}$ and ratio with the expected relative residual error.

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.

Acknowledgement. This research was supported by the Hungarian Scientific Research Fund OTKA, No. K112157 and SNN125119.

References

[1] Axelsson, O. and Maubach, J., On the updating and assembly of the Hessian matrix in finite element methods, *Comput. Methods Appl. Mech. Engrg.*, 71 (1988), pp. 41–67.

[2] 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.

[3] Ciarlet, Ph., *The finite element method for elliptic problems,* North-Holland, Amsterdam, 1978

[4] Ciarlet, Ph., *Linear and nonlinear functional analysis with applications,* SIAM, Philadelphia, 2013.

[5] Deuflhard, P. and Weiser, M., Global inexact Newton multilevel FEM for nonlinear elliptic problems, in Multigrid Methods V, *Lect. Notes Comput. Sci. Eng.* 3, Springer, Berlin, 1998, pp. 71–89.

[6] 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.

[7] Glowinski, R., *Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems*, SIAM, Philadelphia, 2015.

[8] Karátson J. , On the Lipschitz continuity of derivatives for some scalar nonlinearities, *J. Math. Anal. Appl.* 346 (2008) 170-176.

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

[10] Křižek, M. and Neittaanmäki, P. , *Mathematical and Numerical Modeling in Electrical Engineering: Theory and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[11] Rossi, T., Toivanen, J., Parallel fictitious domain method for a non-linear elliptic Neumann boundary value problem, *Numer. Linear Algebra Appl.* 6 (1999), no. 1, 51–60.

[12] Waterhouse, W. C., The absolute-value estimate for symmetric multilinear forms, *Linear Algebra Appl.* 128 (1990), 97–105.

[13] Zeidler, E., *Nonlinear functional analysis and its applications,* Springer, 1986