

# Highlights

## **Quasi-Newton iterative solution approaches for nonsmooth elliptic operators with applications to elasto-plasticity**

János Karátson, Stanislav Sysala<sup>1</sup>, Michal Béréš

- Extension of the quasi-Newton/variable preconditioning method for non-smooth operators
- Convergence analysis of the method and its application to elasto-plasticity
- Combination of the method with the deflated conjugate gradient and aggregation-based algebraic multigrid methods
- Extension of well-documented and available Matlab codes on elasto-plasticity
- Comparison of the quasi-Newton and Newton methods on 3D numerical examples

---

<sup>1</sup>Corresponding author: [stanislav.sysala@ugn.cas.cz](mailto:stanislav.sysala@ugn.cas.cz)

# Quasi-Newton iterative solution approaches for nonsmooth elliptic operators with applications to elasto-plasticity

János Karátson<sup>a,b</sup>, Stanislav Sysala<sup>1c</sup>, Michal Béréšč<sup>c,d</sup>

<sup>a</sup>*Department of Applied Analysis & HUN-REN-ELTE Numerical Analysis and Large Networks Research Group, Eötvös Loránd University, Budapest, Hungary*

<sup>b</sup>*Department of Analysis and Operations Research, Budapest University of Technology and Economics, Budapest, Hungary*

<sup>c</sup>*Department of Applied Mathematics and Computational Sciences, Institute of Geonics of the Czech Academy of Sciences, Ostrava, Czech Republic*

<sup>d</sup>*Department of Applied Mathematics, VSB - Technical University of Ostrava, Ostrava, Czech Republic*

---

## Abstract

This paper is devoted to the extension of a quasi-Newton/variable preconditioning (QNVP) method to non-smooth problems, motivated by elasto-plastic models. Two approaches are discussed: the first one is carried out via regularized approximations of the nonsmooth problem, and the second one gives an extension to nonsmooth operators in order to be applied directly. Convergence analysis is presented for both variants. Then these abstract methods are applied to elasto-plasticity where two different variants of QNVP are investigated and combined with the deflated conjugate gradient and aggregation-based algebraic multigrid methods. The convergence results are illustrated on numerical examples in 3D inspired by real-life problems, and they demonstrate that the suggested QNVP methods are competitive with the standard Newton method. Well-documented Matlab codes on elasto-plasticity are used and enriched by the suggested methods.

## Keywords:

nonlinear operator equation, non-smoothness, quasi-Newton/variable preconditioning, elasto-plasticity, large-scale systems

2000 MSC: 35J60, 35Q74, 47J05, 65J15, 65N12, 74C05, 74S05

---

## 1. Introduction

This paper is focused on the iterative solution of nonlinear operator equations in the form

$$F(\mathbf{u}^*) = 0 \tag{1.1}$$

where  $F: H \rightarrow H$  is a semismooth operator defined in a Hilbert space  $H$ . Such elliptic equations appear in certain physical problems, and their proper discretizations lead to large-scale nonlinear systems. These system are often solved by the semismooth Newton method, see e.g., [1–4]. Since we encounter large-scale problems, the linearized systems of equations appearing in each Newton’s iteration have to be solved by preconditioned iterative methods and thus we arrive at inexact variants of the semismooth Newton method, see e.g. [5–7]. Even in this case a next difficulty is the fact that the generalized derivative (Jacobian) of  $F$  has usually a complicated structure, which is costly to update and it is not easy to find convenient

---

<sup>1</sup>Corresponding author: stanislav.sysala@ugn.cas.cz

preconditioners for them. To overcome this drawback, approximations of the Jacobians with simpler structure are chosen and thus we arrive at quasi-Newton (QN) methods.

The family of QN methods is very broad. For an overview, we refer to, for example, [8]. The most known are BFGS, Broyden, DFP or SR1 methods, where the approximations of Jacobians are constructed recursively with proper algebraic techniques. The construction and convergence analysis of some QN methods for semismooth operators can be found, for example, in [3,9].

We focus on the so-called quasi-Newton/variable preconditioning (QNPV) method elaborated earlier in [10] and the monograph [11] and applied further in the authors' papers [12,13] for general elliptic problems. Unlike other QN methods, the QNPV method involves spectrally equivalent approximations of the Jacobians, and for this aim it exploits the structure of the underlying PDE in the iterative solution of the problem discretized by finite elements. Therefore, in particular, the method is convenient for large-scale problems. It has been recently used for smooth problems in nonlinear elasticity in [14] and proved to be competitive with full Newton linearizations regarding overall computational cost. It is recommended there to combine the QNPV method with the deflated conjugate gradient method [15].

Up to now, the QNPV method has been used only for smooth operators. Therefore, the first aim of this paper is to extend its convergence analysis for nonsmooth operators within the abstract setting of the problem. In Section 2, two different approaches are presented together with local linear convergence analysis:

- First, we consider regularized approximations of the nonsmooth operator, such that the smooth versions of QNPV iterations can be applied. We show that the convergence is bounded independently of the regularization parameter  $\epsilon$ .
- Secondly, we define QNPV iterations for the original problem directly, under the assumption of the semismoothness of  $F$ . We consider two theoretical variants and prove their local linear convergence.

The next aim of this paper is to apply the abstract framework to elasto-plasticity, see Sections 3 and 4. Elasto-plastic models are described, for example, in [6,16–18], and the arising discretized elasto-plastic operators, defined in terms of displacements, are usually semismooth as was shown, for example, in [19,20]. For convergence analysis of Newton-like methods in elasto-plasticity, we refer to [5,6,20–23].

In Section 3, we present a von Mises type elasto-plastic operator, its smooth regularization and space discretization by the finite element method. The assumptions on the abstract operator  $F$ , specified in Section 2, are verified in this elasto-plastic case.

In Section 4, we introduce two preconditioners approximating the generalized Jacobian of the investigated elasto-plastic operator. They are in the form of the elastic stiffness matrix with fixed and variable material parameters. We derive spectral bounds for these preconditioners and specify the convergence results.

Section 5 is devoted to preconditioned iterative solvers for the linearized systems corresponding to the semismooth Newton and QNPV methods. As in [14], the deflated conjugate gradient (DCG) method is considered. Unlike the incomplete Cholesky factor preconditioner used in [14], we combine DCG with the displacement decomposition [24–26] and the aggregation-based algebraic multigrid (AGMG) preconditioner [27–29], leading to much better numerical results.

Section 6 reports in detail on numerical experiments in 3D, carried out for two geotechnical problems involving homogeneous and heterogeneous materials, respectively. The first problem is related to the analysis of the bearing capacity of a soil foundation, whereas the second

one concerns a composite sample comprising coal and polyurethane, derived from computer tomography. The performance of the proposed methods is tested and compared to the exact Newton linearizations, showing that one can achieve lower overall runtimes with the suggested QNVP methods, hence the latter are competitive with Newton's method.

The results of this paper are also supported by the extension of recently developed and well-documented codes on elasto-plasticity (see [30]) to include the QNVP, DCG and AGMG methods. The extended codes are available for download at [31].

## 2. The QNVP method: abstract background

In this section, we consider the abstract operator equation (1.1) and introduce the QNVP method. First we summarize the existing background for smooth operators, then we extend it for nonsmooth operators. To this end we use two different approaches: a regularization approach and a direct nonsmooth approach.

Let  $H$  denote a real Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and induced norm  $\|\cdot\|$ . For simplicity, the norm of linear operators are also denoted by  $\|\cdot\|$ .

### 2.1. QNVP approaches based on smooth operators

#### 2.1.1. Smooth operators

First, let  $F$  be a smooth operator. In fact, it suffices if it has a Gâteaux derivative denoted as  $F'$ , further, if  $F'$  is continuous only in the weak sense of **bihemicontinuity**, see [11], which means that for any  $u, v, w, z \in H$  the mapping  $(s, t) \mapsto F'(u + tv + sw)z$  is continuous from  $\mathbf{R}^2$  to  $H$ .

We briefly state our earlier related result, for more details we refer to the monograph [11] and the authors' recent papers [12,13], see also [10].

We formulate the following properties for  $F'$ :

#### Assumptions 2.1.1

- (i) (Symmetry.) For any  $u \in H$  the operator  $F'(u)$  is self-adjoint.
- (ii) (Ellipticity.) There exist constants  $\beta_1 \geq \beta_2 > 0$  such that

$$\beta_2 \|h\|^2 \leq \langle F'(u)h, h \rangle \leq \beta_1 \|h\|^2 \quad (\forall u, h \in H). \quad (2.1)$$

- (iii) (Lipschitz continuity.) There exists a constant  $L > 0$  such that

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

Conditions (i)–(ii) imply that equation (1.1) has a unique solution  $u^* \in H$ , see, e.g., [11]. The formulation of the following theorem uses the energy  $*$ -norm  $\|v\|_* := \langle F'(u^*)^{-1}v, v \rangle^{1/2}$ .

**Theorem 1.** *Let Assumptions 2.1.1 hold. Let  $u_0$  be in a properly small neighborhood of  $u^*$ , and let the sequence  $\{u_n\}$  be defined by*

$$u_{n+1} := u_n - \frac{2}{M_n + m_n} B_n^{-1} F(u_n) \quad (n \in \mathbf{N}), \quad (2.2)$$

where  $0 < m_n \leq M_n$  and the self-adjoint linear operators  $B_n : H \rightarrow H$  satisfy

$$m_n \langle B_n h, h \rangle \leq \langle F'(u_n)h, h \rangle \leq M_n \langle B_n h, h \rangle \quad (n \in \mathbf{N}, h \in H). \quad (2.3)$$

We require  $\{m_n\}$  to be positively bounded from below and  $\{M_n\}$  bounded from above. Then  $\{u_n\}$  converges to  $u^*$ , that is,  $\|u_n - u^*\| \leq \frac{1}{\beta_2} \|F(u_n)\| \rightarrow 0$  and

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

*Proof.* It follows from [11], see also [13].  $\square$

**Remark 1.** The above local convergence result can be made global by proper damping, which is typical for Newton type methods. See [11] and [12] for details.

**Remark 2.** It follows from (2.1) that the operator  $F$  itself satisfies the uniform monotonicity

$$\beta_2 \|u - v\|^2 \leq \langle F(u) - F(v), u - v \rangle \quad (\forall u, v \in H), \quad (2.4)$$

and thus also

$$\beta_2 \|u - v\| \leq \|F(u) - F(v)\| \quad (\forall u, v \in H). \quad (2.5)$$

### 2.1.2. Regularized nonsmooth operators

Now we consider equation (1.1) when  $F$  is nonsmooth (non-differentiable). We assume that it can be approximated with smooth operators of the type considered in Section 3.1. Namely, for any  $\epsilon > 0$  there exists an operator

$$F_\epsilon : H \rightarrow H$$

with the following properties:

#### Assumptions 2.1.2

(i) There exists  $K > 0$  such that for any  $\epsilon > 0$ :

$$\|F(u) - F_\epsilon(u)\| \leq K\epsilon \|u\| \quad (\forall u \in H). \quad (2.6)$$

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

(iii) There exist constants  $\beta_1 \geq \beta_2 > 0$ , independently of  $\epsilon$ , such that

$$\beta_2 \|h\|^2 \leq \langle F'_\epsilon(u)h, h \rangle \leq \beta_1 \|h\|^2 \quad (\forall u, h \in H). \quad (2.7)$$

(iv) (Lipschitz continuity.) For any  $\epsilon > 0$  there exists  $L(\epsilon) > 0$  such that

$$\|F'_\epsilon(u) - F'_\epsilon(v)\| \leq L(\epsilon) \|u - v\| \quad (\forall u, h \in H). \quad (2.8)$$

(Here we may allow that  $\lim_{\epsilon \rightarrow 0} L(\epsilon) = +\infty$ .)

**Theorem 2.** *Let Assumptions 2.1.2 hold.*

(1) *If  $u^*$  and  $u_\epsilon^*$  are the solutions of equations  $F(u) = 0$  and  $F_\epsilon(u_\epsilon) = 0$ , respectively, then*

$$\|u^* - u_\epsilon^*\| \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0_+.$$

(2) *Let  $\epsilon > 0$  be fixed. Let  $u_0$  be in a properly small neighborhood of  $u_\epsilon$ , and let the sequence  $\{u_n\}$  be defined by*

$$u_{n+1} := u_n - \frac{2}{M_n^{(\epsilon)} + m_n^{(\epsilon)}} (B_n^{(\epsilon)})^{-1} F_\epsilon(u_n) \quad (n \in \mathbf{N}), \quad (2.9)$$

*where  $0 < m_n^{(\epsilon)} \leq M_n^{(\epsilon)}$ , and the self-adjoint linear operators  $B_n^{(\epsilon)} : H \rightarrow H$  satisfy*

$$m_n^{(\epsilon)} \langle B_n^{(\epsilon)} h, h \rangle \leq \langle F'_\epsilon(u_n) h, h \rangle \leq M_n^{(\epsilon)} \langle B_n^{(\epsilon)} h, h \rangle \quad (n \in \mathbf{N}, h \in H). \quad (2.10)$$

Assume that there exist constants  $\overline{M} \geq \underline{m} > 0$ , independently of  $\epsilon$ , such that

$$\underline{m} \leq m_n^{(\epsilon)} \leq M_n^{(\epsilon)} \leq \overline{M} \quad (\forall n \in \mathbf{N}). \quad (2.11)$$

Then  $\{u_n\}$  converges to  $u_\epsilon^*$  as  $n \rightarrow \infty$ , that is,  $\|u_n - u_\epsilon^*\| \leq \frac{1}{\beta_1} \|F_\epsilon(u_n)\| \rightarrow 0$  and

$$\limsup_{n \rightarrow +\infty} \frac{\|F_\epsilon(u_{n+1})\|_*}{\|F_\epsilon(u_n)\|_*} \leq \limsup_{n \rightarrow +\infty} \frac{M_n^{(\epsilon)} - m_n^{(\epsilon)}}{M_n^{(\epsilon)} + m_n^{(\epsilon)}} \leq Q, \quad (2.12)$$

$$\text{where} \quad Q := \frac{\overline{M} - \underline{m}}{\overline{M} + \underline{m}} < 1 \quad \text{is independent of } \epsilon. \quad (2.13)$$

*Proof.* (1) Comparing the two equations, we have  $F(u^*) = F_\epsilon(u_\epsilon^*) = 0$ . Hence,

$$F_\epsilon(u^*) - F(u^*) = F_\epsilon(u^*) - F_\epsilon(u_\epsilon^*).$$

Assumption (iii) and Remark 2 imply

$$\beta_2 \|u^* - u_\epsilon^*\| \leq \|F_\epsilon(u^*) - F_\epsilon(u_\epsilon^*)\|.$$

From this relation and assumption (i), we obtain

$$\beta_2 \|u^* - u_\epsilon^*\| \leq \|F_\epsilon(u^*) - F(u^*)\| \leq K\epsilon \|u^*\| \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0.$$

(2) We may apply Theorem 1 to the operator  $F_\epsilon$  with fixed  $\epsilon$ , since it satisfies Assumptions 2.1.1. This yields (2.12). The bounds in (2.11), which are independent of  $\epsilon$ , imply that (2.13) holds as well. Note that  $Q$  does not involve the Lipschitz constant  $L(\epsilon)$ .  $\square$

## 2.2. The QNVP method for nonsmooth operators

Now our goal is to study the QNVP method for an in general nonsmooth nonlinear operator  $F : H \rightarrow H$  directly, without regularized approximations. In this case, proper analogues of Assumptions 2.1.1 are imposed on the generalized derivatives of  $F$ , denoted by  $F^o(u)$ . Altogether, we assume that  $F$  satisfies the following properties:

### Assumptions 2.2

- (A1)  $F$  is Lipschitz continuous and uniformly monotone in the sense of (2.4).
- (A2) For any  $u \in H$  there exists a bounded self-adjoint linear operator  $F^o(u) : H \rightarrow H$ , representing a generalized derivative of  $F$  at  $u$ , and satisfying the assumptions below.
- (A3) There exist constants  $\beta_1 \geq \beta_2 > 0$  such that

$$\beta_2 \|h\|^2 \leq \langle F^o(u)h, h \rangle \leq \beta_1 \|h\|^2 \quad (\forall u, h \in H). \quad (2.14)$$

- (A4) For any  $u \in H$ , there exist constants  $\delta_u > 0$  and  $L_u > 0$  such that

$$\|F(v) - F(u) - F^o(u)(v - u)\| \leq L_u \|u - v\|^2 \quad (\forall v \in V(u, \delta_u)), \quad (2.15)$$

where  $V(u, \delta_u) = \{v \in H \mid \|u - v\| \leq \delta_u\}$ .

**Remark 3.** If  $H$  is finite-dimensional, Assumption (A4) means that the function  $F$  is strongly semismooth in  $H$ . For an infinite-dimensional  $H$ , the operator  $F^o$  may represent the so-called slant derivative. Assumption (A1) implies that there exists a unique solution  $u^* \in H$  of (1.1), see, e.g., [32].

In order to solve problem (1.1), we use the following QNVP method, which is the nonsmooth analogue of (2.2)–(2.3):

$$u_{n+1} := u_n - \frac{2}{M_n + m_n} B_n^{-1} F(u_n) \quad (n \in \mathbf{N}), \quad (2.16)$$

where:

(A5)  $0 < m_{\min} \leq m_n \leq M_n \leq M_{\max}$  and  $B_n : H \rightarrow H \times H$  are self-adjoint operators satisfying

$$m_n \langle B_n h, h \rangle \leq \langle F^o(u_n) h, h \rangle \leq M_n \langle B_n h, h \rangle \quad (n \in \mathbf{N}, h \in H). \quad (2.17)$$

We present two results on local linear convergence of Algorithm (2.16)–(2.17) under different restrictions. The first one is relevant for **mild nonlinearities** (that is, when the bounds  $\beta_1$  and  $\beta_2$  in (2.14) are relatively close to each other, see (2.19) below) and is expressed with the standard norm in  $H$ . The second one may cover **stronger** nonlinearities, using proper energy norms, provided that the latter satisfy extra stability type conditions near the solution.

### 2.2.1. *Mild nonlinearities*

**Theorem 3.** (Local linear convergence in  $\|\cdot\|$  norm.) *Let Assumptions (A1)–(A5) hold and  $u_n$  be sufficiently close to  $u^*$ . Then*

$$\|u^* - u_{n+1}\| \leq \frac{\beta_1}{\beta_2} \frac{M_n - m_n}{M_n + m_n} \|u^* - u_n\| + O(\|u^* - u_n\|^2). \quad (2.18)$$

Consequently, if

$$q := \frac{\beta_1}{\beta_2} \frac{M_{\max} - m_{\min}}{M_{\max} + m_{\min}} < 1, \quad (2.19)$$

then Algorithm (2.16)–(2.17) provides at least a linear rate of convergence.

*Proof.* Using (2.16) and (1.1), we consequently derive:

$$\begin{aligned} u^* - u_{n+1} &= u^* - u_n + \frac{2}{M_n + m_n} B_n^{-1} (F(u_n) - F(u^*)) = \\ &= \left( I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) \right) (u^* - u_n) + \frac{2}{M_n + m_n} B_n^{-1} (F(u_n) - F(u^*) - F^o(u_n)(u_n - u^*)). \end{aligned} \quad (2.20)$$

If  $u_n$  is sufficiently close to  $u^*$ , then assumptions (A3) and (A4) imply

$$\|u^* - u_{n+1}\| \leq \left\| I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) \right\| \|u_n - u^*\| + O(\|u_n - u^*\|^2). \quad (2.21)$$

Further, we have:

$$\left\| I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) \right\| \leq \left\| F^o(u_n)^{-1} - \frac{2}{M_n + m_n} B_n^{-1} \right\| \cdot \|F^o(u_n)\|, \quad (2.22)$$

where assumption (A3) yields

$$\|F^o(u_n)\| = \sup_{\|v\|=1} |\langle F^o(u_n) v, v \rangle| \leq \beta_1. \quad (2.23)$$

Applying [10] to (2.17) and (2.14), respectively, we have for any  $n \in \mathbf{N}$  and  $v \in H$ :

$$M_n^{-1} \langle B_n^{-1} v, v \rangle \leq \langle F^o(u_n)^{-1} v, v \rangle \leq m_n^{-1} \langle B_n^{-1} v, v \rangle, \quad \beta_1^{-1} \|v\|^2 \leq \langle F^o(u_n)^{-1} v, v \rangle \leq \beta_2^{-1} \|v\|^2.$$

Hence,

$$-\frac{M_n - m_n}{M_n + m_n} \langle F^o(u_n)^{-1} v, v \rangle \leq \left\langle \left( F^o(u_n)^{-1} - \frac{2}{M_n + m_n} B_n^{-1} \right) v, v \right\rangle \leq \frac{M_n - m_n}{M_n + m_n} \langle F^o(u_n)^{-1} v, v \rangle, \quad (2.24)$$

$$\begin{aligned} \left\| F^o(u_n)^{-1} - \frac{2}{M_n + m_n} B_n^{-1} \right\| &= \sup_{\|v\|=1} \left| \left\langle \left( F^o(u_n)^{-1} - \frac{2}{M_n + m_n} B_n^{-1} \right) v, v \right\rangle \right| \\ &\leq \frac{M_n - m_n}{M_n + m_n} \sup_{\|v\|=1} \langle F^o(u_n)^{-1} v, v \rangle \leq \frac{M_n - m_n}{M_n + m_n} \frac{1}{\beta_2}. \end{aligned} \quad (2.25)$$

Inserting (2.22)–(2.25) into (2.21), we arrive at (2.18).  $\square$

If, in particular, we choose  $B_n = F^o(u_n)$ , then we arrive at the standard semismooth Newton method, in addition,  $m_n = M_n = 1$  and (2.18) implies the local quadratic convergence of Algorithm (2.16).

However, if  $B_n \neq F^o(u_n)$ , then possibly  $\beta_1/\beta_2 \gg 1$  may occur for strongly nonlinear problems and thus the convergence criterion  $q < 1$  in the theorem can be too restrictive. To **overcome** this drawback, we study the local linear convergence in an appropriate energy norm, see the next subsection.

### 2.2.2. *Stronger nonlinearities*

Let us introduce the following norms induced by the operators  $F^o(u^*)$  and  $F^o(u_n)$ , respectively:

$$\|v\|_{u^*} = \sqrt{\langle F^o(u^*)v, v \rangle}, \quad \|v\|_{u_n} = \sqrt{\langle F^o(u_n)v, v \rangle} \quad (u, v \in H).$$

From Assumption (A3), it follows that there exist positive constants  $\gamma_1(n), \gamma_2(n), \bar{\gamma}_1, \bar{\gamma}_2, n \in \mathbf{N}$ , such that  $0 < \bar{\gamma}_2 \leq \gamma_2(n) \leq \gamma_1(n) \leq \bar{\gamma}_1$  and

$$\gamma_2(n) \|v\|_{u^*}^2 \leq \|v\|_{u_n}^2 \leq \gamma_1(n) \|v\|_{u^*}^2 \quad (\forall u, v \in H, \forall n \in \mathbf{N}). \quad (2.26)$$

The main advantage of the bound (2.26) against (2.14) is the fact that the constants  $\gamma_1(n)$  and  $\gamma_2(n)$  can be close to one if  $u_n$  tends to  $u^*$ . For example, it occurs if  $F$  has locally Lipschitz continuous derivative in vicinity of  $u^*$ . Therefore, our aim is to improve the estimate (2.18) using the norm  $\|\cdot\|_{u^*}$  and the constants  $\gamma_1(n)$  and  $\gamma_2(n)$ . To this end, we need the following auxiliary result. Let  $A, S$  be linear self-adjoint operators in  $H$  such that  $A$  is also positive definite, and let

$$\langle u, v \rangle_A = \langle Au, v \rangle, \quad \|v\|_A = \sqrt{\langle Av, v \rangle} \quad \forall u, v \in H,$$

define the corresponding inner product and norm in  $H$ . Then  $A^{-1}S$  is  $A$ -selfadjoint: for any  $z, h \in H$  we have  $\langle A^{-1}Sh, z \rangle_A = \langle Sh, z \rangle = \langle h, Sz \rangle = \langle h, A^{-1}Sz \rangle_A$ . Hence,

$$\|A^{-1}S\|_A \left( := \sup_{h \neq 0} \frac{\|A^{-1}Sh\|_A}{\|h\|_A} \right) = \sup_{h \neq 0} \frac{|\langle A^{-1}Sh, h \rangle_A|}{\|h\|_A^2} = \sup_{h \neq 0} \frac{|\langle Sh, h \rangle|}{\langle Ah, h \rangle}. \quad (2.27)$$

**Theorem 4.** (Local linear convergence in  $\|\cdot\|_{u^*}$  norm.) *Let Assumptions (A1)–(A5) hold and  $u_n$  be sufficiently close to  $u^*$ . Then*

$$\|u^* - u_{n+1}\|_{u^*} \leq \sqrt{\frac{\gamma_1(n)}{\gamma_2(n)} \frac{M_n - m_n}{M_n + m_n}} \|u^* - u_n\|_{u^*} + O(\|u^* - u_n\|_{u^*}^2). \quad (2.28)$$

*Consequently, if there exists  $q < 1$  such that  $\sqrt{\frac{\gamma_1(n)}{\gamma_2(n)} \frac{M_n - m_n}{M_n + m_n}} \leq q$  for all  $n \in \mathbf{N}$ , then Algorithm (2.16) provides at least a linear rate of the convergence.*

*Proof.* The identity (2.20) can be derived as in the proof of Theorem 1. If  $u_n$  is sufficiently close to  $u^*$ , then the assumptions (A3) and (A4) imply

$$\|u^* - u_{n+1}\|_{u_n} \leq \left\| I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) \right\|_{u_n} \|u_n - u^*\|_{u_n} + O(\|u_n - u^*\|_{u_n}^2). \quad (2.29)$$

Since

$$I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) = F^o(u_n)^{-1} \left( F^o(u_n) - \frac{2}{M_n + m_n} F^o(u_n) B_n^{-1} F^o(u_n) \right),$$

the identity (2.27) can be applied with  $A = F^o(u_n)$  and  $S = F^o(u_n) - \frac{2}{M_n + m_n} F^o(u_n) B_n^{-1} F^o(u_n)$ . Namely, for any  $h \in H$ ,

$$|\langle Sh, h \rangle| = \left| \left\langle \left( F^o(u_n) - \frac{2}{M_n + m_n} F^o(u_n) B_n^{-1} F^o(u_n) \right) h, h \right\rangle \right| \quad (2.30)$$

$$= \left| \left\langle \left( F^o(u_n)^{-1} - \frac{2}{M_n + m_n} B_n^{-1} \right) F^o(u_n) h, F^o(u_n) h \right\rangle \right| \quad (2.31)$$

$$\leq \frac{M_n - m_n}{M_n + m_n} \langle h, F^o(u_n) h \rangle = \frac{M_n - m_n}{M_n + m_n} \langle Ah, h \rangle, \quad (2.32)$$

where (2.24) was used for  $v := F^o(u_n)h$ . Hence, (2.27) yields

$$\left\| I - \frac{2}{M_n + m_n} B_n^{-1} F^o(u_n) \right\|_{u_n} = \sup_{h \neq 0} \frac{|\langle Sh, h \rangle|}{\langle Ah, h \rangle} \leq \frac{M_n - m_n}{M_n + m_n}. \quad (2.33)$$

Further, (2.26) implies

$$\|u^* - u_{n+1}\|_{u_n} \geq \sqrt{\gamma_2(n)} \|u^* - u_{n+1}\|_{u^*}, \quad \|u^* - u_n\|_{u_n} \leq \sqrt{\gamma_1(n)} \|u^* - u_n\|_{u^*}. \quad (2.34)$$

Inserting (2.33)–(2.34) into (2.29), we arrive at (2.28).  $\square$

Based on Theorem 4, one can introduce the following assumption on  $F$  to ensure local linear convergence of Algorithm (2.16):

(A6) There exist positive constants  $\delta$ ,  $\gamma_{\min}$  and  $\gamma_{\max}$  such that

$$\gamma_{\min} \|v\|_{u^*}^2 \leq \langle F^o(u_n) v, v \rangle \leq \gamma_{\max} \|v\|_{u^*}^2 \quad (\forall n \in \mathbf{N}, \|u^* - u_n\|_{u^*} \leq \delta), \quad (2.35)$$

$$\sqrt{\frac{\gamma_{\max}}{\gamma_{\min}}} \frac{M_{\max} - m_{\min}}{M_{\max} + m_{\min}} < 1, \quad (2.36)$$

where  $M_{\max}$  and  $m_{\min}$  are defined within Assumption (A5).

**Corollary 1.** *Let Assumptions (A1)–(A6) hold. Then Algorithm (2.16) provides local linear convergence.*

**Remark 4.** The above convergence is ensured if  $\|F^o(u_n) - F^o(u^*)\| \rightarrow 0$  as  $n \rightarrow \infty$ , which guarantees that  $\gamma_{\max}$  and  $\gamma_{\min} \rightarrow 1$  as  $n \rightarrow \infty$ . It is satisfied, for example, if  $F$  has a locally Lipschitz continuous derivative in a vicinity of  $u^*$ .

### 3. The elasto-plastic problem and its smooth approximation

In this section, we specify the operators  $F$ ,  $F^o$  and  $F_\epsilon$  from Section 2 for a selected elasto-plastic problem and verify that the assumptions are satisfied for these operators. In particular, we consider a simplified static version of the elasto-plastic model with the von Mises yield criterion and isotropic hardening. It is well-known that a quite similar problem arises from time discretization of the related quasi-static problem. We follow [14,18,33].

#### 3.1. The nonlinear PDE system and its finite element discretization

We consider a body occupying a three-dimensional bounded domain  $\Omega \subset \mathbf{R}^3$  with Lipschitz continuous boundary denoted as  $\partial\Omega$ . The investigated problem can be described via the following quantities: the displacement vector  $u : \Omega \rightarrow \mathbf{R}^3$ , the strain tensor  $\varepsilon : \Omega \rightarrow \mathbf{R}^{3 \times 3}$  and the stress tensor  $\sigma : \Omega \rightarrow \mathbf{R}^{3 \times 3}$ . The basic system of equations is as follows:

$$-\operatorname{div} \sigma_i = \varphi_i(x) \quad \text{in } \Omega, \quad \sigma_i \cdot \nu = \gamma_i(x) \quad \text{on } \Gamma_N, \quad u_i = 0 \quad \text{on } \Gamma_D \quad (i = 1, 2, 3). \quad (3.1)$$

Here  $x \in \Omega$  is the space variable,  $\nu$  is the outward unit normal to  $\partial\Omega$ ,  $\sigma_i = (\sigma_{i1}, \sigma_{i2}, \sigma_{i3})$  ( $i = 1, 2, 3$ ) denotes the  $i$ th row of the matrix  $\sigma$ . The functions  $\varphi : \Omega \rightarrow \mathbf{R}^3$  and  $\gamma : \Gamma_N \rightarrow \mathbf{R}^3$  describe the body and boundary force vectors, respectively. The body is clamped on  $\Gamma_D$ , where  $\partial\Omega = \overline{\Gamma_N} \cup \overline{\Gamma_D}$  is a disjoint measurable subdivision and  $\Gamma_D$  is nonempty and open in  $\partial\Omega$ . The strain tensor is determined from the displacement, that is,  $\varepsilon = \varepsilon(u)$ , where

$$\varepsilon(u) = \frac{1}{2} (\nabla u + \nabla u^\top). \quad (3.2)$$

(Here  $\nabla u^\top(x)$  denotes the transpose of the matrix  $\nabla u(x) \in \mathbf{R}^{3 \times 3}$ .) This means that in this model the strains are assumed to be small enough such that geometric nonlinearity can be neglected. Finally, the following general form of nonlinear stress-strain operator is considered:

$$\sigma(x) = T(x, \varepsilon(u(x))) \quad (3.3)$$

The matrix-valued mapping  $T : \Omega \times \mathbf{R}^{3 \times 3} \rightarrow \mathbf{R}^{3 \times 3}$  is given and will be specified in Section 3.2 together with its generalized derivative  $T^o$  and its smooth approximation  $T_\epsilon$ .

Inserting (3.2) and (3.3) into (3.1), we arrive at a second order system of partial differential equations in terms of the displacement  $u = (u_1, u_2, u_3)$ . Its weak formulation is defined in a standard way in the Sobolev space

$$H_D^1(\Omega)^3, \quad \text{where} \quad H_D^1(\Omega) := \{u \in H^1(\Omega) : u|_{\Gamma_D} = 0\},$$

endowed with the inner product and the corresponding norm

$$\langle u, v \rangle = \int_{\Omega} \nabla u \cdot \nabla v, \quad \|u\|^2 = \int_{\Omega} |\nabla u|^2 \quad (u, v \in H_D^1(\Omega)^3). \quad (3.4)$$

The weak problem reads as follows. Find  $u = (u_1, u_2, u_3) \in H_D^1(\Omega)^3$  such that

$$\int_{\Omega} T(x, \varepsilon(u)) \cdot \varepsilon(v) - \int_{\Omega} \varphi \cdot v - \int_{\Gamma_N} \gamma \cdot v = 0 \quad (v \in H_D^1(\Omega)^3). \quad (3.5)$$

The nonlinear problem (3.5) is properly discretized and then solved numerically by some iterative method. For the discretization we use finite elements (FEM). In general, let  $V_h \subset H_D^1(\Omega)$  be a fixed FEM subspace. To be in accordance with the notation from Section 2, we

set  $H := V_h^3$  for a fixed discretization parameter  $h$ . Then the FEM problem is the discrete analogue of the weak form (3.5): find  $u^* \in H$  such that

$$\int_{\Omega} T(x, \varepsilon(u^*)) : \varepsilon(v) - \int_{\Omega} \varphi \cdot v - \int_{\Gamma_N} \gamma \cdot v = 0 \quad (v \in V_h^3). \quad (3.6)$$

It leads to the following weak operator equation

$$\langle F(u^*), v \rangle = 0 \quad (v \in H). \quad (3.7)$$

or just

$$F(u^*) = 0 \text{ in } H. \quad (3.8)$$

Analogously, one can introduce smoothing operators  $F_\epsilon$  by using smooth approximations  $T_\epsilon$  instead of  $T$ , detailed in the next subsection. It is evident that the properties of  $F$  and  $F_\epsilon$  depend mainly on the properties of the nonlinear mappings  $T$  and  $T_\epsilon$ .

### 3.2. Constitutive operators and their properties

We investigate the constitutive operators  $T$  and  $T_\epsilon$  for a fixed material point  $x$  within this section. Therefore, we do not highlight the dependence of these operators on  $x$ , for the sake of brevity.

Let  $I$  denote the identity matrix, and for any matrix  $A \in \mathbf{R}^{3 \times 3}$ , let  $\text{tr } A = \sum_{i=1}^3 A_{ii}$  be the trace of  $A$ . We denote

$$\text{vol } A = \frac{1}{3}(\text{tr } A)I, \quad \text{dev } A = A - \text{vol } A. \quad (3.9)$$

Let  $:$  and  $|\cdot|$  denote the elementwise matrix product and the corresponding Frobenius norm defined by

$$A : B = \sum_{i,k=1}^3 A_{ik}B_{ik}, \quad |A|^2 = A : A = \sum_{i,k=1}^3 A_{ik}^2 \quad (A, B \in \mathbf{R}^{3 \times 3}). \quad (3.10)$$

The stress-strain operator  $T$  representing the von Mises yield criterion and isotropic hardening has the following definition (see e.g. [22]):

$$T(A) = 3k \text{vol } A + 2\mu(|\text{dev } A|^2) \text{dev } A \quad (A \in \mathbf{R}^{3 \times 3}). \quad (3.11)$$

Here  $k > 0$  is the bulk modulus and  $\mu(z)$  is the nonlinear Lamé coefficient defined by

$$\mu(z) := \begin{cases} \mu_0, & \text{if } 2\mu_0\sqrt{z} \leq Y, \\ (1 - \alpha)\mu_0 + \frac{\alpha}{2\sqrt{z}}Y & \text{if } 2\mu_0\sqrt{z} \geq Y, \end{cases} \quad (3.12)$$

where  $\mu_0 > 0$  is the linear Lamé coefficient,  $\alpha \in (0, 1)$  represents the isotropic hardening parameter and  $Y > 0$  is the initial yield stress. It is easy to see that the function  $\mu$  is Lipschitz continuous and piecewise smooth. Such functions are also strongly semismooth. Further, smooth functions with locally Lipschitz derivatives are strongly semismooth. In addition, finite sums, products or compositions of strongly semismooth functions are also strongly semismooth. Therefore, the function  $T$  defined by (3.11) is strongly semismooth in  $\mathbf{R}^{3 \times 3}$ , see also [20].

The generalized derivative  $T^\circ$  can be defined by

$$T^\circ(A)B = 3k \text{vol } B + 2\mu(|\text{dev } A|^2) \text{dev } B + 4\mu^\circ(|\text{dev } A|^2) (\text{vol } A : \text{vol } B) \text{vol } A \quad (\forall B \in \mathbf{R}^{3 \times 3}), \quad (3.13)$$

where

$$\mu^\circ(z) := \begin{cases} \mu'(z) = 0 & \text{if } 2\mu_0\sqrt{z} < Y, \\ 0 & \text{if } 2\mu_0\sqrt{z} = Y, \\ \mu'(z) = -\frac{\alpha}{4z^{3/2}}Y & \text{if } 2\mu_0\sqrt{z} > Y. \end{cases} \quad (3.14)$$

In order to relate the operator  $T^\circ$  with its simplified approximations, arising later in the preconditioners in Section 4, we introduce the following fourth-order tensors:

$$\mathbb{T}_{el} B := 3k \operatorname{vol} B + 2\mu_0 \operatorname{dev} B \quad \forall B \in \mathbf{R}^{3 \times 3}, \quad (3.15)$$

$$\overline{\mathbb{T}}(A)B := 3k \operatorname{vol} B + 2\mu(|\operatorname{dev} A|^2) \operatorname{dev} B \quad \forall B \in \mathbf{R}^{3 \times 3}, \quad (3.16)$$

$$\underline{\mathbb{T}}(A)B := 3k \operatorname{vol} B + 2\underline{\mu}(|\operatorname{dev} A|^2) \operatorname{dev} B \quad \forall B \in \mathbf{R}^{3 \times 3}, \quad (3.17)$$

where

$$\underline{\mu}(z) := \mu(z) + 2\mu^\circ(z)z, \quad (1 - \alpha)\mu_0 \leq \underline{\mu}(z) \leq \mu(z) \leq \mu_0 \quad \forall z \in \mathbf{R}^+. \quad (3.18)$$

Notice that  $\mathbb{T}_{el}$  represents Hooke's law for purely linear elastic response. The following estimates hold for any  $B \in \mathbf{R}^{3 \times 3}$ :

$$(1 - \alpha)\mathbb{T}_{el} B : B \leq \underline{\mathbb{T}}(A)B : B \leq T^\circ(A)B : B \leq \overline{\mathbb{T}}(A)B : B \leq \mathbb{T}_{el} B : B. \quad (3.19)$$

Similar bounds were derived, for example, in [14] or [20].

To define the smoothing operator  $T_\epsilon$  to  $T$ , it suffices to replace the function  $\mu$  in (3.11) with its smooth approximation for given  $\epsilon \in (0, 1)$ :

$$\mu_\epsilon(z) := \begin{cases} \mu_0, & 2\mu_0\sqrt{z} \leq Y(1 - \epsilon), \\ (1 - \alpha)\mu_0 + \frac{\alpha}{2\sqrt{z}}[Y - \frac{1}{4\epsilon Y}(2\mu_0\sqrt{z} - Y - \epsilon)^2], & 2\mu_0\sqrt{z} \in [Y(1 - \epsilon), Y(1 + \epsilon)], \\ (1 - \alpha)\mu_0 + \frac{\alpha}{2\sqrt{z}}Y, & 2\mu_0\sqrt{z} \geq Y(1 + \epsilon). \end{cases} \quad (3.20)$$

This function was introduced in [14] and the following bounds were derived there:

$$(1 - \alpha)\mathbb{T}_{el} B : B \leq \underline{\mathbb{T}}_\epsilon(A)B : B \leq T'_\epsilon(A)B : B \leq \overline{\mathbb{T}}_\epsilon(A)B : B \leq \mathbb{T}_{el} B : B, \quad (3.21)$$

where

$$\overline{\mathbb{T}}_\epsilon(A)B := 3k \operatorname{vol} B + 2\mu_\epsilon(|\operatorname{dev} A|^2) \operatorname{dev} B \quad \forall B \in \mathbf{R}^{3 \times 3}, \quad (3.22)$$

$$\underline{\mathbb{T}}_\epsilon(A)B := 3k \operatorname{vol} B + 2\underline{\mu}_\epsilon(|\operatorname{dev} A|^2) \operatorname{dev} B \quad \forall B \in \mathbf{R}^{3 \times 3}, \quad (3.23)$$

$$\underline{\mu}_\epsilon(z) := \mu_\epsilon(z) + 2\underline{\mu}'_\epsilon(z)z, \quad (1 - \alpha)\mu_0 \leq \underline{\mu}_\epsilon(z) \leq \mu_\epsilon(z) \leq \mu_0 \quad \forall z \in \mathbf{R}^+. \quad (3.24)$$

We also refer to [22] for the relationships between  $T$  and  $T_\epsilon$  which are important for the verification of Assumptions 2.1.2.

### 3.3. Properties of the elasto-plastic operators $F$ , $F^\circ$ and $F_\epsilon$

In general, we suppose that the investigated body consists of a heterogeneous material and assume that the material functions  $k$ ,  $\mu_0$ ,  $Y$ ,  $\alpha$  belong to  $L^\infty(\Omega)$ . In addition, we assume that  $k$ ,  $\mu_0$ ,  $Y$  are uniformly bounded from below by the positive constants  $\underline{k}$ ,  $\underline{\mu}_0$ ,  $\underline{Y}$ , respectively, and that  $\alpha$  is bounded from above by  $\bar{\alpha} < 1$ .

The operators  $F$ ,  $F^o$  and  $F_\epsilon$  are defined as follows:

$$\langle F(u), v \rangle = \int_{\Omega} T(x, \varepsilon(u)) : \varepsilon(v) - \int_{\Omega} \varphi \cdot v - \int_{\Gamma_N} \gamma \cdot v \quad (u, v \in H), \quad (3.25)$$

$$\langle F_\epsilon(u), v \rangle = \int_{\Omega} T_\epsilon(x, \varepsilon(u)) : \varepsilon(v) - \int_{\Omega} \varphi \cdot v - \int_{\Gamma_N} \gamma \cdot v \quad (u, v \in H), \quad (3.26)$$

$$\langle F^o(u)v, v \rangle = \int_{\Omega} T^o(x, \varepsilon(u))\varepsilon(v) : \varepsilon(v) \quad (u, v \in H). \quad (3.27)$$

**Remark 5.** If higher-order finite elements are considered, the integrals introduced within the definitions  $F$ ,  $F^o$  and  $F_\epsilon$  are usually computed numerically. It means that the operator  $F$  is defined as a finite sum of terms containing the operator  $T$  at integration points.

Similarly to  $F^o$ , one can also introduce the operators  $F_{el}$ ,  $\underline{F}^o$ ,  $\underline{F}'_\epsilon$ ,  $\overline{F}^o$ , and  $\overline{F}'_\epsilon$  defined by  $\mathbb{T}_{el}$ ,  $\underline{\mathbb{T}}$ ,  $\underline{\mathbb{T}}_\epsilon$ ,  $\overline{\mathbb{T}}$ , and  $\overline{\mathbb{T}}_\epsilon$ , respectively. It is well-known that one can apply the Korn inequality [18] to prove that the elastic operator  $F_{el}$  induces an inner product in  $H$  which is equivalent to the standard inner product  $\langle \cdot, \cdot \rangle$ . Therefore, it suffices to investigate Assumptions 2.1.2 and 2.2 with respect to the following elastic inner product and the corresponding norm:

$$\langle u, v \rangle_{el} := \langle F_{el}u, v \rangle, \quad \|u\|_{el}^2 := \langle F_{el}u, u \rangle \quad (u, v \in H). \quad (3.28)$$

From (3.19) and (3.21), the following estimates can be derived for any  $u, v \in H$ :

$$(1 - \bar{\alpha})\|v\|_{el}^2 \leq \langle \underline{F}^o(u)v, v \rangle \leq \langle F^o(u)v, v \rangle \leq \langle \overline{F}^o(u)v, v \rangle \leq \|v\|_{el}^2, \quad (3.29)$$

$$(1 - \bar{\alpha})\|v\|_{el}^2 \leq \langle \underline{F}'_\epsilon(u)v, v \rangle \leq \langle F'_\epsilon(u)v, v \rangle \leq \langle \overline{F}'_\epsilon(u)v, v \rangle \leq \|v\|_{el}^2. \quad (3.30)$$

These bounds imply the assumptions (2.7) and (2.14) and can be easily extended to the infinite-dimensional space  $(H_D^1(\Omega))^3$ . In addition, properties (2.6) and (2.8) follow from [22] or [23]. The assumption (2.15) was proven for the investigated elasto-plastic model in [20]. It implies that the operator  $F$  is strongly semismooth in  $H$ . To the best of our knowledge, this property has not been extended to  $(H_D^1(\Omega))^3$ , yet.

Altogether, we arrive at the following result:

**Corollary 2.** *The operators  $F$ ,  $F^o$  and  $F_\epsilon$  defined by (3.25)–(3.27) satisfy Assumptions 2.1.2 and 2.2.*

In particular, it follows that problem (3.5) has a unique weak solution and that the FE approximations converge to it for admissible families of FE subspaces.

#### 4. The QNVP method for the elasto-plastic system

In this section, we suggest preconditioners  $B_n^{(\epsilon)}$  and  $B_n$  for Algorithms (2.9) and (2.16), respectively, to be convenient for the investigated elasto-plastic problem. We follow on our recent work [14] and extend its results, especially for the semismooth operator  $F$ . In particular, we investigate two types of preconditioners, with variable and fixed elastic stiffness matrices, and calculate their corresponding bounds used in Algorithms (2.9) and (2.16). Finally, we derive the corresponding convergence results for all preconditioners.

#### 4.1. Preconditioner 1 – the elastic stiffness matrix with variable parameters

For a given iterate  $u_n \in V_h^3$ , let us choose weight functions  $\mu_{n,\epsilon}, \mu_n \in L^\infty(\Omega)$  such that the following inequalities hold almost everywhere in  $\Omega$ :

$$\underline{\mu}_\epsilon(|\operatorname{dev} \varepsilon(u_n)|^2) \leq \mu_{n,\epsilon} \leq \mu_\epsilon(|\operatorname{dev} \varepsilon(u_n)|^2), \quad (4.1)$$

$$\underline{\mu}(|\operatorname{dev} \varepsilon(u_n)|^2) \leq \mu_n \leq \mu(|\operatorname{dev} \varepsilon(u_n)|^2), \quad (4.2)$$

where the functions  $\mu, \mu_\epsilon, \underline{\mu}$  and  $\underline{\mu}_\epsilon$  were defined in Section 3.2. The preconditioners  $B_n^{(\epsilon)} := B_n^{(1,\epsilon)}$  and  $B_n := B_n^{(1)}$  corresponding to these weight functions are defined by

$$\langle B_n^{(1,\epsilon)} h, v \rangle := \int_\Omega \left( 3k \operatorname{vol} \varepsilon(h) : \operatorname{vol} \varepsilon(v) + 2\mu_{n,\epsilon} \operatorname{dev} \varepsilon(h) : \operatorname{dev} \varepsilon(v) \right), \quad (4.3)$$

$$\langle B_n^{(1)} h, v \rangle := \int_\Omega \left( 3k \operatorname{vol} \varepsilon(h) : \operatorname{vol} \varepsilon(v) + 2\mu_n \operatorname{dev} \varepsilon(h) : \operatorname{dev} \varepsilon(v) \right), \quad (4.4)$$

for all  $h, v \in H$ .

**Remark 6.** Within Preconditioner 1, we approximate the dev part using only a scalar coefficient, in contrast to the full Newton linearization where the tensor  $\operatorname{dev} \varepsilon(u_n) \operatorname{dev} \varepsilon(u_n)^\top$  is also required in the second terms. The above difference in the complexity is reflected as well in the fact that the assembly of the FE matrix requires less operations for  $B_n^{(1,\epsilon)}$  than for  $F'(u_n)$ .

The following estimates hold for the arising spectral bounds.

**Proposition 1.** *The operators  $B_n^{(1,\epsilon)}$  and  $B_n^{(1)}$  satisfy*

$$m_n^{(1,\epsilon)} \langle B_n^{(1,\epsilon)} h, h \rangle \leq \langle F'_\epsilon(u_n) h, h \rangle \leq M_n^{(1,\epsilon)} \langle B_n^{(1,\epsilon)} h, h \rangle \quad (\forall h \in H), \quad (4.5)$$

$$m_n^{(1)} \langle B_n^{(1)} h, h \rangle \leq \langle F^o(u_n) h, h \rangle \leq M_n^{(1)} \langle B_n^{(1)} h, h \rangle \quad (\forall h \in H), \quad (4.6)$$

where

$$m_n^{(1,\epsilon)} = \inf_\Omega \frac{\underline{\mu}_\epsilon(|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_{n,\epsilon}} \in [1 - \bar{\alpha}, 1], \quad M_n^{(1,\epsilon)} = \sup_\Omega \frac{\mu_\epsilon(|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_{n,\epsilon}} \in [1, \frac{1}{1 - \bar{\alpha}}], \quad (4.7)$$

$$m_n^{(1)} = \inf_\Omega \frac{\underline{\mu}(|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_n} \in [1 - \bar{\alpha}, 1], \quad M_n^{(1)} = \sup_\Omega \frac{\mu(|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_n} \in [1, \frac{1}{1 - \bar{\alpha}}]. \quad (4.8)$$

*Proof.* The bounds in (4.7) and (4.8) follow from (3.18), (3.24), (4.1) and (4.2). Further, using (3.29) and (4.4), we have

$$m_n^{(1)} \langle B_n^{(1)} h, h \rangle \leq \langle \underline{F}^o(u_n) h, h \rangle \leq \langle F^o(u_n) h, h \rangle \leq \langle \overline{F}^o(u_n) h, h \rangle \leq M_n^{(1)} \langle B_n^{(1)} h, h \rangle,$$

implying (4.6). Similarly, one can prove (4.5).  $\square$

In practice it is natural to let  $\mu_n$  be a convex combination of  $\mu$  and  $\underline{\mu}$ , e.g. the arithmetic mean as in our numerical examples presented in Section 6. The choice  $\mu_n = \mu$  leads to the preconditioner corresponding to the well-known Picard's/Kachanov's iterations applied to the investigated elasto-plastic problem.

In the numerical examples, we denote the QNVP method with Preconditioner 1 as *QNVP-1a*. Since the assembly of  $B_n^{(1)}$  or  $B_n^{(1,\epsilon)}$  in each Newton's iteration is computationally costly, we introduce the following modification [14]:

$$\left. \begin{array}{l} \text{if } |\mu_n - \mu_{n-1}|/|\mu_n| \leq 0.1 \text{ for any } x \in \Omega \text{ then set } B_n^{(1)} = B_{n-1}^{(1)}, \\ \text{if } |\mu_{n,\epsilon} - \mu_{n-1,\epsilon}|/|\mu_{n,\epsilon}| \leq 0.1 \text{ for any } x \in \Omega \text{ then set } B_n^{(1,\epsilon)} = B_{n-1}^{(1,\epsilon)}. \end{array} \right\} \quad (4.9)$$

We term the corresponding variant as *QNVP-1b*.

#### 4.2. Preconditioner 2 – the elastic stiffness matrix with fixed parameters

The operators (4.3) and (4.4) can be simplified if the scalar coefficients are kept the same throughout the iteration. Then, we arrive at the “purely elastic fixed preconditioner”,  $B^{(2)} = B_n^{(2)} = B_n^{(2,\epsilon)} = F_{el}$  where  $F_{el}$  was introduced in Section 3.3. Thus,

$$\langle B^{(2)}h, v \rangle = \langle h, v \rangle_{el} = \int_{\Omega} \left( 3k \operatorname{vol} \varepsilon(h) : \operatorname{vol} \varepsilon(v) + 2\mu_0 \operatorname{dev} \varepsilon(h) : \operatorname{dev} \varepsilon(v) \right) \quad (4.10)$$

(for all  $h, v \in V_h^3$ ). A fixed (i.e. constant) preconditioner is favourable, since it does not need updating, moreover, one may use efficient CG solutions for the corresponding sequence of systems with a fixed matrix using a deflated CG method as described in [14]. We also note that such a globally spectrally equivalent fixed preconditioner provides global convergence without damping, see, e.g., [11].

**Proposition 2.** *The operator (4.10) satisfies*

$$m_n^{(2,\epsilon)} \langle B^{(2)}h, h \rangle \leq \langle F'_\epsilon(u_n)h, h \rangle \leq M_n^{(2,\epsilon)} \langle B^{(2)}h, h \rangle \quad (\forall h \in H), \quad (4.11)$$

$$m_n^{(2)} \langle B^{(2)}h, h \rangle \leq \langle F^o(u_n)h, h \rangle \leq M_n^{(2)} \langle B^{(2)}h, h \rangle \quad (\forall h \in H), \quad (4.12)$$

where

$$m_n^{(2,\epsilon)} = \inf_{\Omega} \frac{\mu_\epsilon (|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_0} \in [1 - \bar{\alpha}, 1], \quad M_n^{(2,\epsilon)} = 1, \quad (4.13)$$

$$m_n^{(2)} = \inf_{\Omega} \frac{\mu (|\operatorname{dev} \varepsilon(u_n)|^2)}{\mu_0} \in [1 - \bar{\alpha}, 1], \quad M_n^{(2)} = 1. \quad (4.14)$$

*Proof.* This result follows from the bounds (3.29), (3.30) and the definitions of  $\underline{F}^o$ ,  $\underline{F}_\epsilon^o$ ,  $\overline{F}^o$ , and  $\overline{F}_\epsilon^o$ .  $\square$

In the numerical examples presented in Section 6, we denote the QNVP method with Preconditioner 2 as *QNVP-2*.

#### 4.3. Convergence of the QNVP iterations

We analyze the convergence properties of the quasi-Newton iterative solver using both Preconditioners 1 and 2, relying on the abstract results of Section 2.

Owing to Propositions 1 and 2 and the properties of  $F$ ,  $F^o$  and  $F_\epsilon$  derived in Section 3.3, we see that the assumptions of Theorems 2, 3 and 4 are satisfied and therefore, these theorems can be used to investigate the local linear convergence of Algorithm (2.9) with operators  $B_n^{(1,\epsilon)}$  and  $B^{(2)}$  and of Algorithm (2.16) with operators  $B_n^{(1)}$  and  $B^{(2)}$ . We summarize the results below in three corollaries.

First, let us consider the regularized versions. The  $\epsilon$ -independent bounds in (4.7) and (4.13) imply that the condition (2.11) of Theorem 2 is satisfied with both preconditioners. Together with the validity of the other assumptions of Theorem 2, we obtain:

**Corollary 3.** *Let  $u_0$  be sufficiently close to  $u^*$  and  $B_n^{(1,\epsilon)}$ ,  $B^{(2)} = B_n^{(2,\epsilon)}$  be defined by (4.3) and (4.10), respectively. Then Algorithm (2.9) with the preconditioner  $B_n^{(1,\epsilon)}$  or  $B^{(2)}$  exhibits at least a linear rate of convergence and the convergence rates in (2.12) are bounded uniformly in  $\epsilon$ .*

Now let us turn to the nonsmooth versions of the iterations without regularization and apply Theorems 3 and 4. In this case, we involve additional sufficient conditions on the problem. First, when applying Theorem 3, we bound the isotropic hardening parameter.

**Corollary 4.** *Let  $u_0$  be sufficiently close to  $u^*$ .*

- (1) *If  $\bar{\alpha} < 0.456$ , then Algorithm (2.16) with preconditioner  $B_n^{(1)}$  provides a linear rate of convergence.*
- (2) *If  $\bar{\alpha} < 0.585$ , then Algorithm (2.16) with preconditioner  $B^{(2)}$  provides a linear rate of convergence.*

*Proof.* We apply Theorem 3. Let us find the values of  $\beta_1, \beta_2, M_{max}, m_{min}$ . To find  $\beta_1, \beta_2$ , we use the elastic norm  $\|\cdot\|_{el}$  instead of  $\|\cdot\|$ . Then one can set  $\beta_1 = 1$  and  $\beta_2 = (1 - \bar{\alpha})$ , which follows from (3.30).

(1) From (4.8), it follows that  $m_{min} = 1 - \bar{\alpha}$  and  $M_{max} = 1/(1 - \bar{\alpha})$ . Inserting these constants and  $\beta_1, \beta_2$  into (2.19), we obtain

$$q = \frac{1}{1 - \bar{\alpha}} \cdot \frac{1 - (1 - \bar{\alpha})^2}{1 + (1 - \bar{\alpha})^2} < 1.$$

This condition is satisfied for  $\bar{\alpha} < 0.456$ .

(2) From (4.14), it follows that  $m_{min} = 1 - \bar{\alpha}$  and  $M_{max} = 1$ . Inserting these constants and  $\beta_1, \beta_2$  into (2.19), we obtain

$$q = \frac{1}{1 - \bar{\alpha}} \cdot \frac{\bar{\alpha}}{2 - \bar{\alpha}} < 1.$$

This condition is satisfied for  $\bar{\alpha} < 2 - \sqrt{2} \approx 0.585$ . □

Finally, Theorem 4 and Corollary 1 imply the local linear convergence of Algorithm (2.16) if the extra assumption (A6) holds, since we have seen that the other assumptions (A1)-(A5) hold for the operator and preconditioners.

**Corollary 5.** *Let  $u_0$  be sufficiently close to  $u^*$ , assumption (A6) hold and  $B_n^{(1)}, B^{(2)}$  be defined by (4.4) and (4.10), respectively. Then Algorithm (2.16) with the preconditioner  $B_n^{(1)}$  or  $B^{(2)}$  exhibits at least a linear rate of convergence.*

**Remark 7.** Corollary 5 involves no restriction on  $\alpha$ , in contrast to Corollary 4. In turn, it relies on the extra assumption (A6). This assumption is readily satisfied if the function  $F$  is differentiable at the solution  $u^*$ . To justify this statement, it suffices to show that the existence of  $F'(u^*)$  implies that  $F$  has a locally Lipschitz continuous derivative in vicinity of  $u^*$ , see Remark 4. Indeed, it follows from the fact that  $F$  can be defined as a finite sum of terms containing the operator  $T$  at integration points, see Remark 5. Further, from the definition of  $T$ , it is easy to see that if  $T$  is differentiable at  $A$ , then  $T$  is also differentiable in vicinity of  $A$  and is Lipschitz continuous there.

In addition, we know that the function  $F$  is almost everywhere differentiable. Therefore, one can expect that the assumption (A6) is satisfied despite the fact that the existence of  $F'(u^*)$  cannot be a priori checked. The numerical tests in Section 6 demonstrate that the convergence holds indeed similarly to the regularized case.

## 5. Iterative solvers for the linearized equations

During the solution process using either the Newton or QNVP method, we encounter a series of linear systems of equations. The individual systems in the series share similar structures and properties. In our previous work [14], we demonstrated that the QNVP method can be competitive with the Newton method if an appropriate iterative solver for the linearized

systems is employed. This finding arises from the fact that the linear systems encountered in QNVP methods are slightly better conditioned than those in the Newton method, as well as the observation that the QNVP method requires less accurate solutions of linear systems than the Newton method to maintain its convergence rate.

In this section, we present our solution concept, which slightly improves the one from [14] and consists of the deflated conjugate gradient method and an appropriate preconditioner for the linearized systems. We use the unifying notation  $C_n s_n = b_n$  for the linearized systems, where the matrix  $C_n$  represents either  $F^o(u_n)$ ,  $F'_\epsilon(u_n)$  (within the Newton method) or  $B_n^{(1)}$ ,  $B_n^{(1,\epsilon)}$ ,  $B^{(2)}$  (in case of QNVP-1 and QNVP-2),  $b_n = -F(u_n)$ , and  $s_n$  denotes the step direction.

### 5.1. The deflated conjugate gradient (DCG) method

The deflated conjugate gradient (DCG) method is described, for example, in [15]. In this work, we utilize the DCG method as an extension of the preconditioned conjugate gradient (PCG) method. This extension is based on the deflation basis  $W_n$ , which serves as an additional input parameter in the solution process. The solution is additively split into two parts:  $s_n = s_n^1 + s_n^2$ :

- The first part,  $s_n^1$ , consists of the projection of the solution of the system  $C_n s_n = b_n$  onto the deflation space  $W_n$ , which can be constructed as

$$s_n^1 = W_n (W_n^\top C_n W_n)^{-1} W_n^\top b_n.$$

- The second part,  $s_n^2$ , is obtained by solving the system using the PCG method on the orthogonal complement of the deflation space,  $W_n^\perp$ . This is done by projecting the residual during the PCG iterations with the projector

$$P_n = I - W_n (W_n^\top C_n W_n)^{-1} W_n^\top C_n.$$

The DCG method is applicable to any linearly independent basis  $W_n$ . However, to achieve faster convergence of the PCG method on  $W_n^\perp$  (compared to the standard PCG applied to the original system), the deflation space must either capture a substantial portion of the solution  $s_n$  or enhance the spectral properties of  $C_n$  on  $W_n^\perp$ . The DCG approach adds a small overhead to the computational time of PCG. This overhead is proportional to the size of the deflation basis and it is negligible in comparison with the cost of the preconditioner discussed in Section 5.2. As a result, we observe cost savings that are proportional to the number of saved PCG iterations.

The aforementioned similarity of individual systems of linearized equations allows us to construct a deflation space  $W_n$  that reduces PCG iterations at a very low computational cost. The idea is to construct the deflation space iteratively:

- In the first Newton or QNVP iteration, the corresponding linearized system is solved with an empty deflation space  $W_1 = \emptyset$ , i.e., by the PCG method.
- In all subsequent iterations, the deflation basis consists of the solutions of previous systems in the series,  $W_n = [s_1, \dots, s_{n-1}]$ .

The efficiency of such a construction of the deflation basis depends on the similarity of the linearized systems, which differs between the Newton and QNVP methods. Based on our numerical examples presented in the following section or in [14], we observe that for the Newton method, the reduction in the number of PCG iterations between the first and last linearized

systems is typically 10-30 %, while for the QNVP-1a,b and QNVP-2 methods, it is usually up to 50 %.

In the case of QNVP-2, the series of linear systems share the same system matrix  $B^{(2)}$  in each iteration. This allows for a different and more efficient modification of the DCG method, which consists of two key aspects:

1. Instead of collecting just a solution from each linear system, we collect the entire Krylov subspace  $K_n$ . Consequently, the resulting series of deflation bases consist of  $W_n = [K_1, \dots, K_{n-1}]$ .
2. The deflation space collected in this manner is  $B^{(2)}$ -orthogonal, meaning  $(W_n^\top B^{(2)} W_n)^{-1}$  is a diagonal matrix. Moreover, the values on the diagonal are computed and collected during the PCG iteration process as the inverses of the  $B^{(2)}$ -norms of the conjugate directions.

In this way, the most computationally expensive part of the deflation process is eliminated, leaving only the additional system matrix multiplication and  $W_n^\top$  multiplication per PCG iteration. This means that even with much larger deflation spaces, the overhead remains minimal.

### 5.2. Preconditioner for the linear systems

The systems for QNVP methods are essentially linear elasticity systems with specified heterogeneous material, whereas the systems for Newton's method are similar, but cannot be exactly interpreted as linear elasticity. In our previous study [14], we employed an incomplete Cholesky factor preconditioner for Newton's and QNVP methods. However, this preconditioner is computationally expensive and its cost does not scale linearly with the discretization size. In this work, we present an alternative approach that alleviates these deficiencies. It combines separate displacement decomposition (developed in [24–26]) with aggregation-based algebraic multigrid (AGMG) (see [27–29]). For comparison purposes, we aim to use the same type of preconditioner across all methods.

As described in [24–26], the separate displacement decomposition approach involves separating the degrees of freedom for each nodal displacement direction. Since our problem from Section 3 is defined in 3D, we can express the system matrix in block form as

$$C_n = \begin{bmatrix} C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,1} & C_{3,2} & C_{3,3} \end{bmatrix},$$

where  $C_{1,1}, C_{2,2}, C_{3,3}$  represent blocks corresponding to individual displacement components and depend on  $n$ . This form may be virtual, achieved through permutation matrices within the context of the preconditioner. The displacement decomposition preconditioner employs the block diagonal as a preconditioner:

$$M_n = \text{diag}(C_{1,1}, C_{2,2}, C_{3,3}),$$

whose spectral equivalence to the original matrix relies on Korn's inequality. Each block's solution can then be achieved using CG with an AGMG preconditioner. Here, we apply only one iteration of the multigrid  $V$ -cycle for each  $C_{1,1}, C_{2,2}, C_{3,3}$  system, i.e., we utilize the preconditioner

$$\tilde{M}_n = \text{diag}(\tilde{C}_{1,1}, \tilde{C}_{2,2}, \tilde{C}_{3,3}),$$

where  $\tilde{C}_{1,1}, \tilde{C}_{2,2}, \tilde{C}_{3,3}$  denote the AGMG preconditioners for each  $C_{1,1}, C_{2,2}, C_{3,3}$ . For our numerical experiments, we employ the AGMG implementation [29].

For the Newton method,  $\tilde{M}_n$  must be set up for any  $n$ . In contrast, the QNVP-1a,b and QNVP-2 methods can utilize a constant preconditioner ( $\tilde{M}_n := \tilde{M}$ ) corresponding to the elastic matrix  $C_n = B^{(2)}$ . In the case of QNVP-1a and QNVP-1b, such a choice of  $\tilde{M}$  is feasible even though the matrix  $B_n^{(1)}$  depends on Newton's step  $n$  and differs from  $B^{(2)}$ . This approach is justified because the time required to set up the preconditioner outweighs the additional time spent in an increased number of iterations due to a sub-optimal preconditioner.

It is worth noting that an AGMG preconditioner could also be applied directly to the entire system, provided a variant of AGMG that utilizes information about the near-null space of the matrix (here consisting of rigid body modes) is available (see, e.g., [34]).

### 5.3. Accuracy of the inner solver

The DCG method is equipped with a stopping criterion based on the relative residuum. Let us denote  $\eta$  as the prescribed tolerance that defines the accuracy of DCG method. The influence of  $\eta$  on the convergence of Newton's solvers in elasto-plasticity was analyzed, for example, in [5,6]. It was observed that the number of Newton's iterations is not particularly sensitive to  $\eta$ , and a relatively rough value of  $\eta = 10^{-1}$  was sufficient. In [14], the total runtimes of the Newton and QNVP solvers were compared as a function of  $\eta$ . It was observed that the optimal value of  $\eta$  was  $10^{-4}$  for the Newton method, while for the QNVP method,  $\eta = 10^{-1}$  or even larger values were sufficient. This difference can be explained by the fact that the main advantage of the Newton method is its superlinear convergence. However, with rough accuracy in the inner solver, this superlinear convergence can be lost, unlike the linear convergence of the QNVP method.

The fact that the QNVP method is less sensitive to  $\eta$  than the Newton method is also used in the numerical experiments presented below. This characteristic is one of the most notable advantages of the QNVP methods, making them competitive with the Newton method.

## 6. Numerical experiments

In this section, we demonstrate the validity of our theoretical results and also compare the suggested QNVP-1a, QNVP-1b and QNVP-2 methods with the Newton method on two distinct numerical examples involving homogeneous and heterogeneous materials, see Subsections 6.1 and 6.2, respectively. We also investigate the influence of mesh density and the parameters  $\alpha$  and  $\epsilon$  on the convergence results.

For the FEM solution, we employ P1 elements. The stopping criterion for Newton-like solvers is defined as follows:

$$u_0 := 0, \quad \|F(u_n)\| \leq 10^{-K} \|F(u_0)\| \quad \text{with } K = 10. \quad (6.1)$$

It is worth noting that the chosen higher tolerance with  $K = 10$  is more favourable for the Newton method (having a local quadratic convergence rate) than for the QNVP methods (having only a local linear convergence rate). Further, we set the tolerance  $\eta$  for the inner solver of linearized systems of equations (see Subsection 5.3) at  $10^{-1}$  for the quasi-Newton methods and  $10^{-4}$  for Newton's method.

For the comparison of the methods, we record the resulting number of Newton (QNVP) iterations, cumulative CG iterations, and runtimes for each test and method. This comprehensive approach aids in distinguishing different types of factors. For example, if the number of Newton's iterations remains constant but the cumulative number of CG iterations increases, it indicates that the linearized problems are more difficult to solve (indicative of worse conditioning). Conversely, if the cumulative number of CG iterations remains stable but the runtime

increases, the additional computational difficulty likely lies in the assembly of the systems and possibly in the application of the AGCM preconditioner from Section 5.

We have utilized in-house Matlab codes, available at [31], for the numerical examples presented below. These codes are derived from recently developed and well-documented codes on elasto-plasticity (see [30]) such that the QNVP, DCG and AGMG methods have been added.

### 6.1. Numerical example with a homogeneous material

The first numerical example is inspired by the well-known *strip-footing problem*, where the bearing capacity of a soil foundation is analyzed. The geometry of this problem is illustrated in Figure 1. We consider a cubic body with dimensions of 10 meters on each side. Zero normal displacements,  $u \cdot \nu = 0$ , are prescribed on all vertical faces and the bottom.

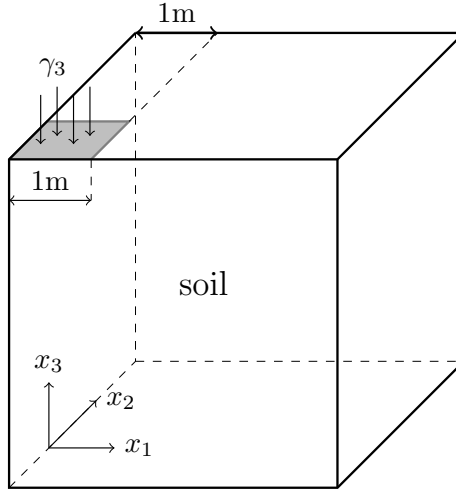


Figure 1: Geometry of the strip-footing problem.

The strip-footing area, located on the top domain, is a one-meter square. The positioning of this area at the corner is justified by the anticipated symmetry with respect to the  $x_1$  and  $x_2$  axes. A constant pressure of  $\gamma_3 = -450$  Pa is applied in the  $x_3$  direction. Volume forces are not considered in this model.

Regarding the material properties, we set the elastic parameters as follows: Young's modulus  $E = 10$  MPa and Poisson's ratio  $\nu_0 = 0.48$ . Using these values, the aforementioned elastic parameters  $k$  and  $\mu_0$  are derived from established formulas. Additionally, the initial yield stress is set at  $Y = 100$  Pa.

In our analysis, we vary both the hardening parameter  $\alpha$  and the regularization parameter  $\epsilon$  to assess their effects on the results. We investigate a range of mesh densities, originating from gradual uniform refinement, which encompasses configurations from 25,578 (mesh level 2) up to 5,241,478 (mesh level 12) unknowns (degrees of freedom).

The first experiment focuses on the effect of the hardening parameter  $\alpha$  on the convergence of the investigated algorithms. It is crucial to note that higher values of  $\alpha$  induce stronger nonlinearity as follows from the bounds presented in Sections 3 and 4. Specifically, we analyze cases with  $\alpha$  set at 0.3, 0.5, 0.9, and  $\epsilon$  at 0, 0.1, using the finest mesh. The results of the experiments are presented in Table 1. Additionally, the runtimes are graphically represented in Figure 2.

It is noteworthy that the standard Newton method exhibits minimal sensitivity to changes in  $\alpha$ . In contrast, both QNVP methods demonstrate a strong dependence on  $\alpha$ , especially when comparing the number of iterations in the table. We observe that the suggested methods

QNVP-1b and QNVP-2 are faster than the Newton method for smaller values of  $\alpha$ , indicating their better efficiency in problems with **milder** nonlinearity. Conversely, the Newton method is the fastest for  $\alpha = 0.9$ . It is also noted in these and subsequent experiments that QNVP-1a and QNVP-1b have the same number of Newton's and CG iterations but different runtimes. This discrepancy is attributed to the freezing criterion (4.9) implemented in QNVP-1b. Due to this criterion, which depends on the smoothness of  $\mu_{n,\epsilon}$ , the runtimes of QNVP-1b are also influenced by  $\epsilon$ , unlike the other methods investigated (as further detailed in the next experiment). Specifically, the matrix  $B_n^{(1)}$  is updated more frequently than the matrix  $B_n^{(1,\epsilon)}$  with  $\epsilon = 0.1$ .

| Setup            | $\alpha$   | Newton              | QNVP-1a      | QNVP-1b              | QNVP-2        |
|------------------|------------|---------------------|--------------|----------------------|---------------|
| $\epsilon = 0$   | <b>0.3</b> | 6/305/194.3         | 12/195/195.9 | 12/195/ <b>141.0</b> | 16/192/146.6  |
|                  | <b>0.5</b> | 6/313/192.5         | 15/222/238.7 | 15/222/ <b>174.1</b> | 22/221/175.3  |
|                  | <b>0.9</b> | 7/460/ <b>257.8</b> | 49/385/774.2 | 49/385/442.9         | 115/531/511.3 |
| $\epsilon = 0.1$ | <b>0.3</b> | 5/249/155.4         | 12/196/197.2 | 12/196/ <b>114.8</b> | 16/192/148.2  |
|                  | <b>0.5</b> | 6/313/193.3         | 15/224/238.2 | 15/225/ <b>140.0</b> | 22/221/176.4  |
|                  | <b>0.9</b> | 7/469/ <b>256.5</b> | 47/393/731.4 | 47/394/388.0         | 115/521/521.9 |

Table 1: Homogeneous material: Dependence on  $\alpha$ , finest mesh (level 12). Values in the table are Newton's iterations/Cumulative CG iterations/Runtime [s]. **The best timing results are highlighted in blue.**

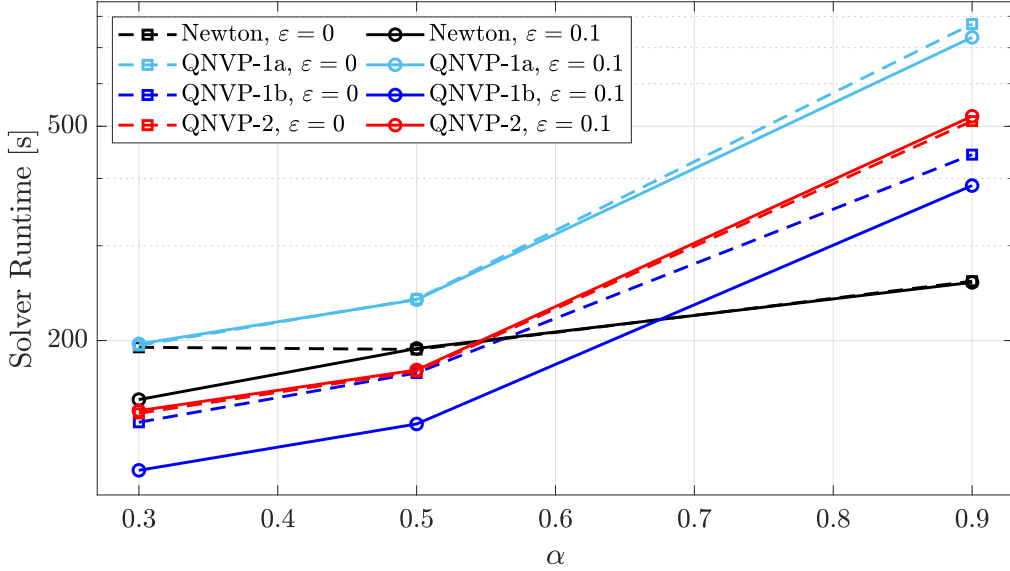


Figure 2: Homogeneous material: Dependence on  $\alpha$ , finest mesh (level 12).

The second experiment focuses on the dependence on the regularization parameter  $\epsilon$ . Results for the smoothest grid, with two variants of  $\alpha = 0.5, 0.9$  and a series of  $\epsilon$  values, are recorded in Table 2 and visualized in Figure 3. Let us note that the **values on  $x$ -axis** of the plot are not to scale and is for illustrative purposes only.

Here, the primary observation is that both Newton and QNVP-2 remain unaffected by the choice of  $\epsilon$ . The QNVP-1b exhibits the same behavior as in the previous example. **Further**, QNVP-1a and QNVP-1b converge approximately 20% faster for  $\epsilon = 0.5$  than for  $\epsilon = 0$ . This accelerated convergence is likely due to the increased smoothness of  $\mu_{n,\epsilon}(z)$  for  $\epsilon = 0.5$ . Finally,

it is worth mentioning that the observed convergence of QNVP-1a and QNVP-2 is uniform for  $0 < \epsilon \leq 0.1$ , which is in accordance with Corollary 3. Additionally, the convergence does not degenerate for the limit case  $\epsilon = 0$ .

| Setup          | $\epsilon$   | Newton              | QNVP-1a      | QNVP-1b              | QNVP-2        |
|----------------|--------------|---------------------|--------------|----------------------|---------------|
| $\alpha = 0.5$ | <b>0</b>     | 6/313/192.5         | 15/222/238.7 | 15/222/ <b>174.1</b> | 22/221/175.3  |
|                | <b>0.001</b> | 6/317/193.7         | 15/222/238.1 | 15/222/ <b>157.0</b> | 22/221/176.1  |
|                | <b>0.01</b>  | 6/311/192.2         | 15/223/240.1 | 15/222/ <b>149.0</b> | 22/221/180.3  |
|                | <b>0.1</b>   | 6/313/193.3         | 15/224/238.2 | 15/225/ <b>140.0</b> | 22/221/176.4  |
|                | <b>0.5</b>   | 6/313/191.8         | 14/215/223.9 | 14/210/ <b>122.4</b> | 22/227/181.9  |
| $\alpha = 0.9$ | <b>0</b>     | 7/460/ <b>257.8</b> | 49/385/774.2 | 49/385/442.9         | 115/531/511.3 |
|                | <b>0.001</b> | 7/461/ <b>248.8</b> | 49/387/764.1 | 49/387/430.8         | 114/532/521.1 |
|                | <b>0.01</b>  | 7/459/ <b>250.2</b> | 49/393/757.5 | 49/393/415.7         | 115/542/523.1 |
|                | <b>0.1</b>   | 7/469/ <b>256.5</b> | 47/393/731.4 | 47/394/388.0         | 115/521/521.9 |
|                | <b>0.5</b>   | 7/473/ <b>257.4</b> | 38/376/594.1 | 38/376/320.5         | 113/530/511.7 |

Table 2: Homogeneous material: Dependence on  $\epsilon$ , finest mesh (level 12). Values in the table are Newton’s iterations/Cumulative CG iterations/Runtime [s]. The best timing results are highlighted in blue.

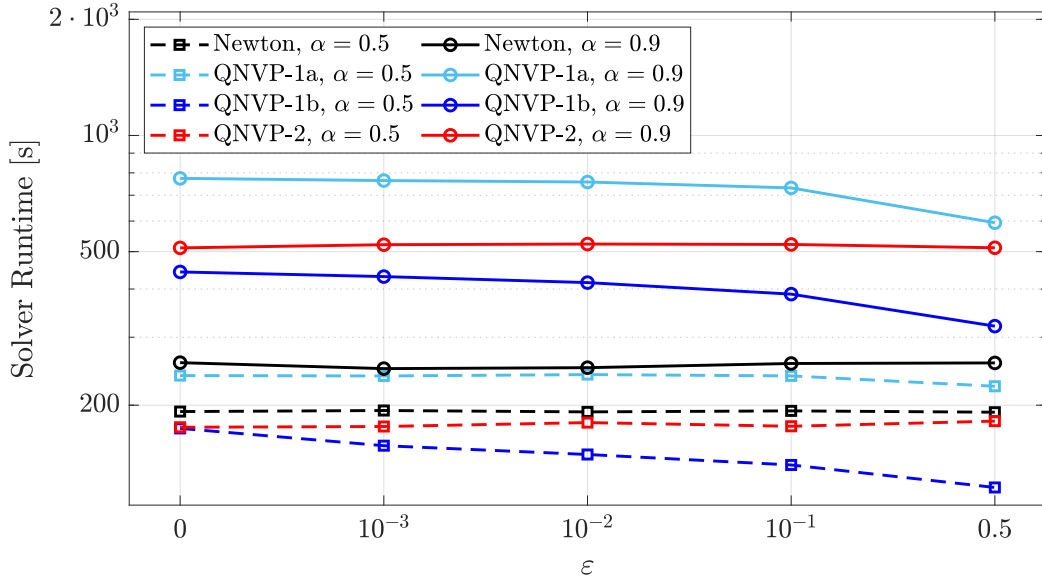


Figure 3: Homogeneous material: Dependence on  $\epsilon$ , finest mesh (level 12). The values on  $x$ -axis are not to scale.

The final experiment, which addresses the homogeneous problem, investigates the dependence on mesh density. We explore mesh levels 2, 4, 6, 8, 10, and 12, examining two distinct configurations:  $(\alpha = 0.3, \epsilon = 0.5)$  and  $(\alpha = 0.9, \epsilon = 0)$ . The results are recorded in Table 3 and visualized in Figure 4.

We see that the number of Newton’s iterations is not overly sensitive to the mesh density. Only for  $\alpha = 0.9$  and QNVP-1a,b, the iteration number is moderately dependent on the mesh size. Additionally, QNVP-1b remains the fastest method for  $\alpha = 0.3$  and all the investigated meshes.

Regarding scaling relative to the increasing problem size, for the parameters  $(\alpha = 0.3, \epsilon = 0.5)$ , it is observed to be roughly linear  $\mathcal{O}(n^1)$  for all methods, see Figure 4. This linearity

| Setup                              | dofs             | Newton              | QNVP-1a      | QNVP-1b              | QNVP-2        |
|------------------------------------|------------------|---------------------|--------------|----------------------|---------------|
| $\alpha = 0.3$<br>$\epsilon = 0.5$ | <b>25,578</b>    | 5/185/2.1           | 11/128/1.7   | 11/122/ <b>1.4</b>   | 16/136/1.7    |
|                                    | <b>198,358</b>   | 5/215/7.2           | 11/146/7.3   | 12/158/ <b>4.8</b>   | 16/162/5.7    |
|                                    | <b>662,338</b>   | 5/230/19.8          | 11/158/22.2  | 12/172/ <b>13.1</b>  | 15/160/15.5   |
|                                    | <b>1,561,518</b> | 5/238/45.8          | 11/167/51.6  | 12/178/ <b>29.6</b>  | 16/184/40.5   |
|                                    | <b>3,039,898</b> | 5/245/87.8          | 11/171/99.8  | 12/185/ <b>57.2</b>  | 16/185/78.5   |
|                                    | <b>5,241,478</b> | 5/251/155.7         | 11/175/176.7 | 12/192/ <b>102.8</b> | 16/188/142.1  |
| $\alpha = 0.9$<br>$\epsilon = 0$   | <b>25,578</b>    | 6/234/2.7           | 22/172/2.7   | 22/169/ <b>2.1</b>   | 113/237/3.2   |
|                                    | <b>198,358</b>   | 7/331/10.6          | 30/234/17.1  | 29/230/ <b>10.3</b>  | 112/284/13.4  |
|                                    | <b>662,338</b>   | 7/372/ <b>30.2</b>  | 36/283/67.4  | 36/283/39.4          | 113/309/45.8  |
|                                    | <b>1,561,518</b> | 7/397/ <b>71.0</b>  | 43/325/196.8 | 43/325/115.0         | 114/413/133.3 |
|                                    | <b>3,039,898</b> | 7/426/ <b>149.1</b> | 47/362/423.9 | 47/362/263.3         | 113/469/265.7 |
|                                    | <b>5,241,478</b> | 7/460/ <b>257.8</b> | 49/385/774.2 | 49/385/442.9         | 115/531/511.3 |

Table 3: Homogeneous material: Dependence on the discretization (number of degrees of freedom). Values in the table are Newton’s iterations/Cumulative CG iterations/Runtime [s]. **The best timing results are highlighted in blue.**

persists for the Newton method even with parameters ( $\alpha = 0.9, \epsilon = 0$ ). However, for these parameters, all QNVP methods demonstrate less favourable scaling, with an approximate complexity of  $\mathcal{O}(n^{1.2})$ .

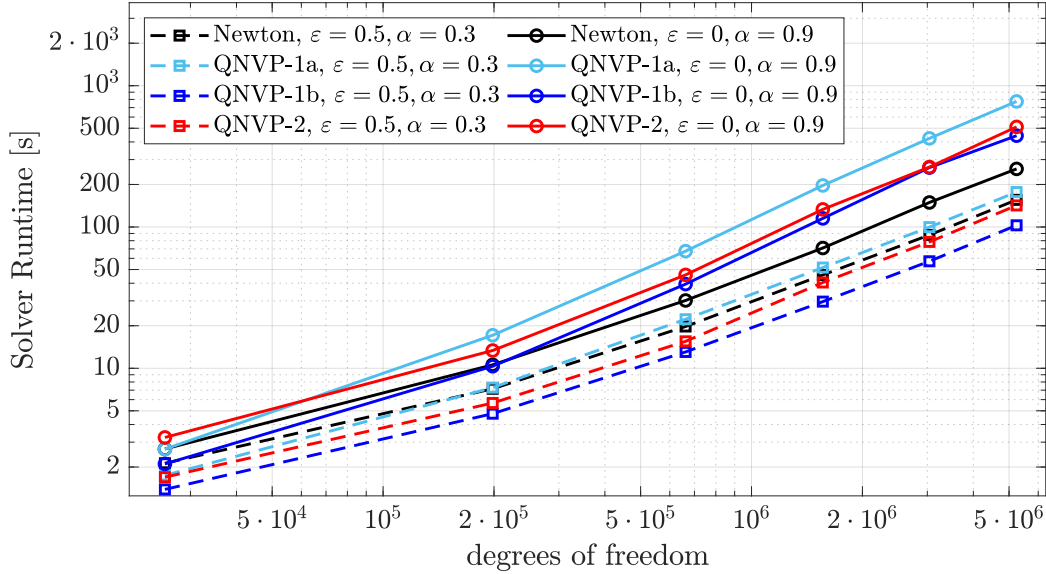


Figure 4: Homogeneous material: Dependence on the discretization (number of degrees of freedom).

## 6.2. Numerical example with a heterogeneous material

We consider a cubic composite sample comprising coal (in black) and polyurethane (in yellow), as shown in Figure 5. The heterogeneity of this sample was derived in [35] by computer tomography and the corresponding mechanical properties were analyzed in [35] and [36].

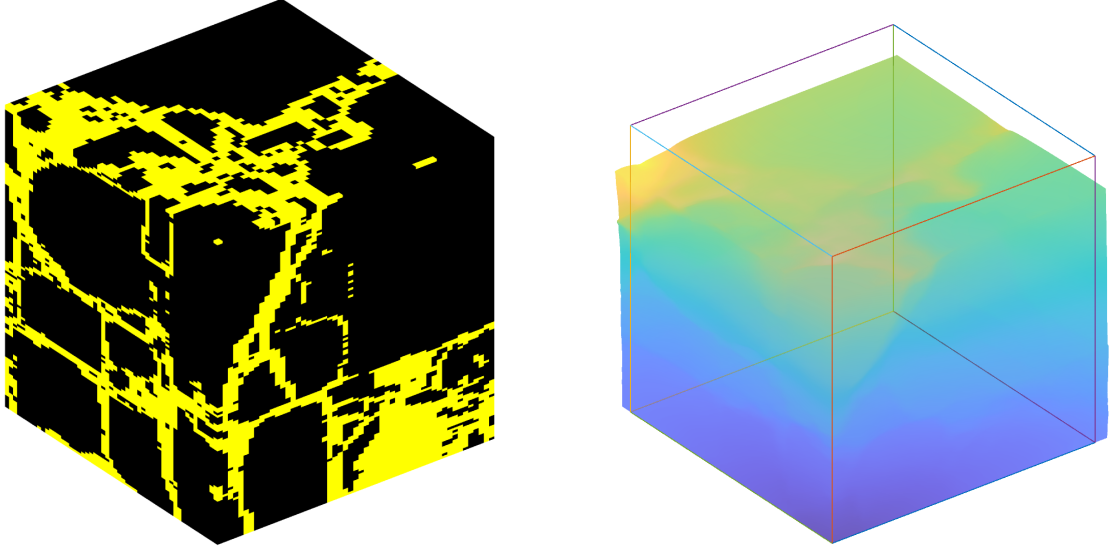


Figure 5: Left – heterogeneous composite material, right – displacement field and deformed shape w.r.t the applied forces.

We prescribe zero normal displacements at the bottom and on two vertical faces at the front, as depicted in Figure 5. A constant pressure of  $\gamma_3 = 20$  MPa is applied to the entire top face. Additionally, we set the following material parameters: for coal,  $E = 2$  GPa,  $\nu_0 = 0.3$ ,  $Y = 30$  MPa,  $\alpha = 0.5$ ; for polyurethane,  $E = 0.3$  GPa,  $\nu_0 = 0.4$ ,  $Y = 10$  MPa,  $\alpha = 0.5$ .

We keep the regularization parameter  $\epsilon$  constant throughout the entire cube and investigate a series of  $\epsilon$  values. Additionally, we explore two different discretizations: mesh level 1, consisting of 1,094,364 degrees of freedom, and mesh level 2, consisting of 4,404,888 degrees of freedom, in the experiments described below. The deformed shape of the composite caused by the applied force is depicted in Figure 5. This output is not overly sensitive to  $\epsilon$  and the mesh density. Further, we record the resulting number of Newton iterations, cumulative CG iterations, and runtimes in Table 4. The runtimes are also visualized in Figures 6 and 7.

The main observations are very similar to those in the homogeneous case, i.e., QNVP-1a, QNVP-1b, and QNVP-2 remain competitive with the Newton method. Additionally, we observe that QNVP-1b and QNVP-2 are significantly faster than the Newton method for the investigated values of parameters. Unlike the example from Section 6.1, we find that QNVP-2 is now faster than QNVP-1b. Further, from Figure 6, it is evident that both Newton and QNVP-2 methods remain unaffected by the choice of  $\epsilon$ , whereas QNVP-1a and QNVP-1b respond favourably to increasing  $\epsilon$ . A notable new observation here is that the Newton method experiences significantly worse runtimes for  $\epsilon = 0.5$ , despite maintaining the same number of iterations. This is attributed to the more costly assembly and CG preconditioner setup and application, resulting from a larger number of non-zero elements during matrix assembly due to an increased volume in the plastic zone of  $\mu$ . Lastly, Figure 7 shows that the scaling obtained from the two mesh sizes is linear, denoted as  $\mathcal{O}(n^1)$ , for all methods.

| Setup        | $\epsilon$ | Newton      | QNVP-1a     | QNVP-1b     | QNVP-2             |
|--------------|------------|-------------|-------------|-------------|--------------------|
| Mesh level 1 | 0          | 6/124/31.4  | 16/83/42.3  | 16/83/28.0  | 22/80/ <b>14.9</b> |
|              | 0.001      | 6/115/32.7  | 16/83/44.1  | 16/83/24.2  | 22/80/ <b>15.6</b> |
|              | 0.01       | 6/117/31.9  | 16/83/44.4  | 16/83/22.2  | 22/80/ <b>15.4</b> |
|              | 0.1        | 6/120/32.4  | 15/81/40.8  | 15/81/19.0  | 22/80/ <b>15.6</b> |
|              | 0.5        | 6/120/43.9  | 13/79/38.0  | 13/80/18.8  | 21/82/ <b>17.1</b> |
| Mesh level 2 | 0          | 6/113/126.6 | 16/79/181.3 | 16/79/117.1 | 22/69/ <b>59.3</b> |
|              | 0.001      | 6/113/128.7 | 16/79/180.3 | 16/79/103.9 | 22/69/ <b>58.4</b> |
|              | 0.01       | 6/115/126.5 | 16/79/179.6 | 16/79/95.6  | 22/69/ <b>58.0</b> |
|              | 0.1        | 6/118/129.9 | 15/77/169.0 | 15/78/77.7  | 22/71/ <b>59.3</b> |
|              | 0.5        | 6/117/182.0 | 14/77/168.2 | 14/76/81.9  | 21/70/ <b>70.1</b> |

Table 4: Heterogeneous material: Values in the table are Newton’s iterations/Cumulative CG iterations/Run-time [s]. The best timing results are highlighted in blue.

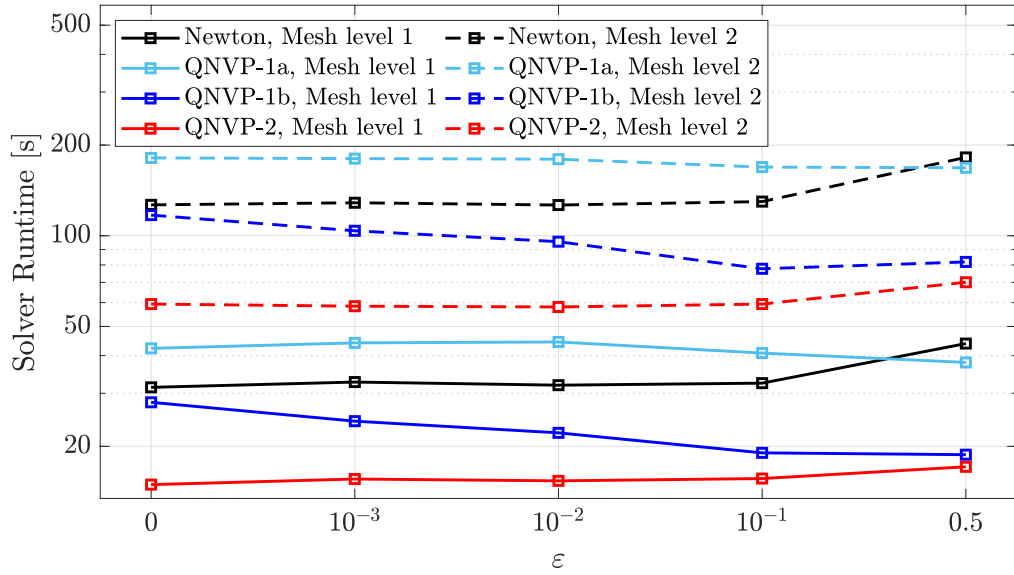


Figure 6: Heterogeneous material: Dependence of runtimes on  $\epsilon$ . The values on  $x$ -axis are not to scale.

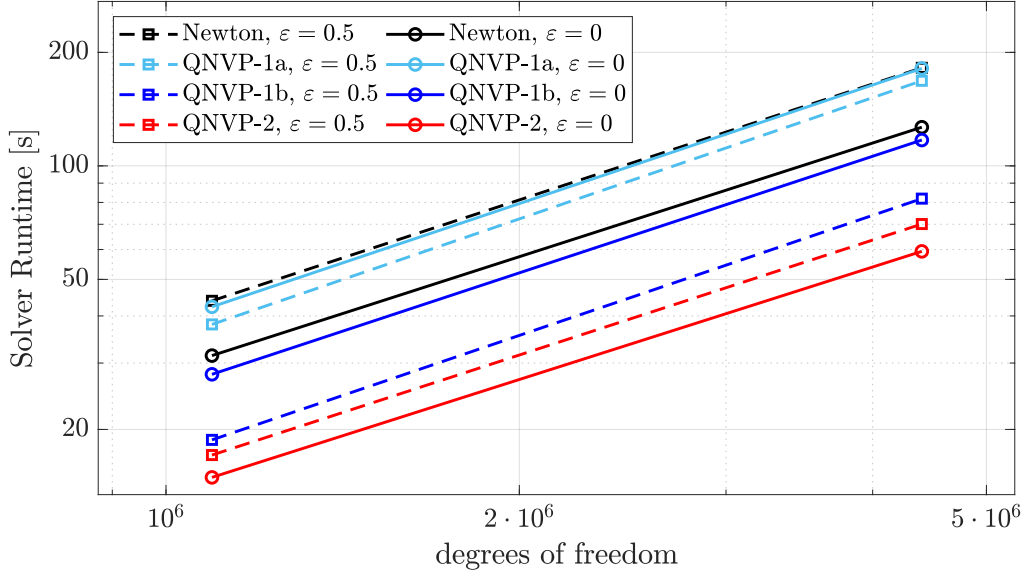


Figure 7: Heterogeneous material: Dependence of runtimes on grid size.

## 7. Conclusion

We have extended the QNVP method for non-smooth operators and derived related convergence results dealing with local linear convergence. The developed abstract framework has been successfully applied to elasto-plasticity and tested on numerical examples in 3D where comprehensive numerical analysis has been done. We have compared the suggested variants of the QNVP method with the standard Newton method and developed in house codes in Matlab.

It was shown that the QNVP method is faster than the Newton method under the following two conditions:

1. Convenient iterative solvers are considered for linearized systems of equations appearing in each Newton's iteration. In particular, we have suggested to use the deflated conjugate gradient method with the AGMG preconditioner.
2. Applied external forces do not cause a strongly non-linear material response. In particular, the suggested QNVP method is rather convenient for plasticity with hardening than for perfect plasticity. It follows from the fact that the convergence of the QNVP method depends on the hardening parameter.

Finally, it is important to note that some more advanced elasto-plastic models than the investigated one also appear in literature, see e.g. [17]. The related constitutive operators are often in implicit forms and it is difficult to construct their derivative for Newton's method. Therefore, the QNVP method or its modifications could be also useful for this advanced treatment in future work.

**Acknowledgements.** The research of J. Karátson is supported by the Hungarian National Research, Development and Innovation Fund (NKFIH) at the Ministry of Innovation and Technology, under the funding scheme ELTE TKP 2021-NKTA-62 and the grant no. K137699. The work of S. Sysala is supported by the Ministry of Industry and Trade of the Czech Republic under the OP TAK project No. CZ.01.01.01/01/22\_002/0000621 co-financed by EU. The research of M. Béréš was cofunded by the financial support of the European Union under the REFRESH - Research Excellence For REgion Sustainability and High-tech Industries project

number CZ.10.03.01/00/22\_003/0000048 via the Operational Programme Just Transition. The authors would also like to thank Dr. Alexej Kolcun who provided us with the heterogeneity for the second numerical example.

**Declaration of interest:** none.

## References

- [1] M. Ulbrich, Semismooth Newton Methods for Operator Equations in Function Spaces, *SIAM J. Optim.* 13 (2002) 805–841.<https://doi.org/10.1137/S1052623400371569>.
- [2] L. Qi, J. Sun, A nonsmooth version of Newton’s method, *Math. Program.* 58 (1993) 353–367.<https://doi.org/10.1007/BF01581275>.
- [3] D. Sun, J. Han, Newton and Quasi-Newton Methods for a Class of Nonsmooth Equations and Related Problems, *SIAM J. Optim.* 7 (1997) 463–480.<https://doi.org/10.1137/S1052623494274970>.
- [4] X. Chen, Z. Nashed, L. Qi, Smoothing methods and semismooth methods for nondifferentiable operator equations, *SIAM J. Numer. Anal.* 38 (2000) 1200–1216.<https://doi.org/10.1137/S0036142999356719>.
- [5] R. Blaheta, O. Axelsson, Convergence of inexact Newton-like iterations in incremental finite element analysis of elasto-plastic problems, *Comput. Methods Appl. Mech. Eng.* 141 (1997) 281–295.[https://doi.org/10.1016/S0045-7825\(96\)01119-X](https://doi.org/10.1016/S0045-7825(96)01119-X).
- [6] R. Blaheta, Numerical methods in elasto-plasticity, PERES Publishers, Prague, 1999.
- [7] N.A. Barnafi, L.F. Pavarino, S. Scacchi, Parallel inexact Newton-Krylov and quasi-Newton solvers for nonlinear elasticity, *Comput. Methods Appl. Mech. Eng.* 400 (2022) 115557.<https://doi.org/10.1016/j.cma.2022.115557>.
- [8] J. Nocedal, S.J. Wright, Numerical optimization, Springer, New York, 2006.
- [9] A. Lewis, M. Overton, Nonsmooth optimization via quasi-Newton methods, *Math. Program., Ser. A* 141 (2013) 135–163.
- [10] J. Karátson, I. Faragó, Variable preconditioning via quasi-Newton methods for nonlinear problems in Hilbert space, *SIAM J. Numer. Anal.* 41 (2003) 1242–1262.<https://doi.org/10.1137/S0036142901384277>.
- [11] I. Faragó, J. Karátson, Numerical solution of nonlinear elliptic problems via preconditioning operators: Theory and application, NOVA Science Publishers, New York, 2002.
- [12] B. Borsos, J. Karátson, Quasi-Newton variable preconditioning for nonlinear nonuniformly monotone elliptic problems posed in Banach spaces, *IMA J. Numer. Anal.* 42 (2022) 1806–1830.<https://doi.org/10.1093/imanum/drab024>.
- [13] B. Borsos, J. Karátson, Robust iterative solvers for Gao-type nonlinear beam models in elasticity, *Comput. Methods Appl. Math.* 22 (2022) 1–13.<https://doi.org/10.1515/cmam-2020-0133>.
- [14] J. Karátson, S. Sysala, M. Béréš, Quasi-Newton variable preconditioning for nonlinear elasticity systems in 3D, *Numer. Linear Algebra Appl.* 31 (2024) e2537.<https://doi.org/10.1002/nla.2537>.
- [15] Y. Saad, M. Yeung, J. Erhel, F. Guyomarc’h, A Deflated Version of the Conjugate Gradient Algorithm, *SIAM J. Sci. Comput.* 21 (2000) 1909–1926.<https://doi.org/10.1137/S1064829598339761>.
- [16] P.G. Ciarlet, Mathematical elasticity. vol. i. three-dimensional elasticity, Reprint of the 1988 edition, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2022.<https://doi.org/10.1137/1.9781611976786>.

- [17] E.A. de Souza Neto, D. Peric, D.R.J. Owen, Computational methods for plasticity: Theory and applications, Wiley-Blackwell, 2008.
- [18] J. Nečas, I. Hlaváček, Mathematical theory of elastic and elasto-plastic bodies: an introduction, Elsevier Pub. Co. ; Distributors for the U.S. and Canada, Elsevier/North Holland, Amsterdam ; New York : New York, 1981.
- [19] P.G. Gruber, J. Valdmán, Solution of one-time step problems in elastoplasticity by a slant newton method, SIAM J. Sci. Comput. 31 (2009) 1558–1580.<https://doi.org/10.1137/070690079>.
- [20] S. Sysala, Application of a modified semismooth Newton method to some elasto-plastic problems, Math Comput Simul 82 (2012) 2004–2021.<https://doi.org/10.1016/j.matcom.2012.03.012>.
- [21] O. Axelsson, A. Padiy, On a two-level Newton-type procedure applied for solving non-linear elasticity problems, Internat. J. Numer. Methods Engrg. 49 (2000) 1479–1493.[https://doi.org/10.1002/1097-0207\(20001230\)49:12%3C1479::AID-NME4%3E3.0.CO;2-4](https://doi.org/10.1002/1097-0207(20001230)49:12%3C1479::AID-NME4%3E3.0.CO;2-4).
- [22] O. Axelsson, S. Sysala, Continuation Newton methods, Comput. Math. Appl. 70 (2015) 2621–2637.<https://doi.org/10.1016/j.camwa.2015.07.024>.
- [23] S. Sysala, Continuation Newton Methods with Applications to Plasticity, in: Proceedings of the 14th International Conference on Large-Scale Scientific Computations - LSSC 2023, Sozopol, Bulgaria, 2023.
- [24] O. Axelsson, On iterative solvers in structural mechanics; separate displacement orderings and mixed variable methods, Math. Comput. Simulation 50 (1999) 11–30.[https://doi.org/10.1016/S0378-4754\(99\)00058-0](https://doi.org/10.1016/S0378-4754(99)00058-0).
- [25] O. Axelsson, J. Karátson, Conditioning analysis of separate displacement preconditioners for some nonlinear elasticity systems, Math. Comput. Simul. 64 (2004) 649–668.<https://doi.org/10.1016/j.matcom.2003.11.017>.
- [26] R. Blaheta, Displacement decomposition – incomplete factorization preconditioning techniques for linearly elasticity problems, Numer. Linear Algebra Appl. 1 (1994) 107–128.<https://doi.org/10.1002/nla.1680010203>.
- [27] A. Napov, Y. Notay, An Algebraic Multigrid Method with Guaranteed Convergence Rate, SIAM J. Sci. Comput. 34 (2012) A1079–A1109.<https://doi.org/10.1137/100818509>.
- [28] Y. Notay, An aggregation-based algebraic multigrid method, Electron. Trans. Numer. Anal. 37 (2010) 123–146.
- [29] Y. Notay, AGMG software and documentation, (2023). <http://agmg.eu>.
- [30] M. Čermák, S. Sysala, J. Valdmán, Efficient and flexible MATLAB implementation of 2D and 3D elastoplastic problems, Appl. Math. Comput. 355 (2019) 595–614.<https://doi.org/10.1016/j.amc.2019.02.054>.
- [31] S. Sysala, M. Béréš, MATLAB implementation of Quasi-Newton Iterative Solutions for Non-linear Elasto-Plastic Problems, (n.d.). [https://github.com/sysala/Matlab\\_nonlinear\\_elasticity\\_3D\\_FEM\\_quasi-Newton\\_DCG](https://github.com/sysala/Matlab_nonlinear_elasticity_3D_FEM_quasi-Newton_DCG).
- [32] E. Zeidler, Nonlinear Functional Analysis and Its Applications, Springer New York, New York, NY, 1990.<https://doi.org/10.1007/978-1-4612-0985-0>.
- [33] R. Blaheta, Multilevel Newton methods for nonlinear problems with applications to elasticity, Copernicus, 1997. [http://www.math.ru.nl/Copernicus/Reports/IGAS\\_R\\_07.ps.gz](http://www.math.ru.nl/Copernicus/Reports/IGAS_R_07.ps.gz).
- [34] A.H. Baker, T.V. Kolev, U.M. Yang, Improving algebraic multigrid interpolation operators for linear elasticity problems, Numer. Linear Algebra Appl. 17 (2010) 495–517.<https://doi.org/10.1002/nla.688>.
- [35] R. Blaheta, R. Kohut, A. Kolcun, K. Souček, L. Staš, L. Vavro, Digital image based numerical micromechanics of geocomposites with application to chemical grouting, Int. J.

- Rock Mech. Min. Sci. 77 (2015) 77–88.<https://doi.org/10.1016/j.ijrmms.2015.03.012>.
- [36] S. Sysala, R. Blaheta, A. Kolcun, J. Ščučka, K. Souček, P.Z. Pan, Computation of Composite Strengths by Limit Analysis, Key Eng. Mater. 810 (2019) 137–142.<https://doi.org/10.4028/www.scientific.net/KEM.810.137>.