Globally convergent B-semismooth Newton Methods for ℓ_1-Tikhonov regularization

Dissertation

zur Erlangung des Grades
Doktor der Naturwissenschaften

am Fachbereich Physik, Mathematik und Informatik
der Johannes Gutenberg-Universität
in Mainz

vorgelegt von

Esther Hans, geb. in Trier

Mainz, im Januar 2017

Datum der mündlichen Prüfung: 23.03.2017

D 77 Mainzer Dissertation

Abstract

In this thesis, we are concerned with the globalization of a semismooth Newton method for ℓ_1-Tikhonov regularization. This regularization strategy for inverse problems with sparsity constraints leads to a convex, nonsmooth minimization problem. Here, one assumes that solutions to the inverse problem can be represented in a basis by few basis vectors. Such problems appear in various applications from science and engineering. We consider discrete problems over the sequence space ℓ_2. The optimality conditions of the minimization problem considered in this work are equivalent to a root-finding problem for a nonlinear, locally Lipschitz continuous operator. This mapping is not Fréchet differentiable but it has a (nonunique) Newton derivative. Based on this generalized derivative concept, a locally superlinearly convergent, semismooth Newton method has been proposed in the literature. However, the convergence of local Newton methods is not guaranteed in general for an arbitrary initial guess. In order to globalize the algorithm, we consider a B(ouligand)-Newton method. We show that the mapping is Bouligand differentiable. That is, we verify that the nonlinearity is directionally differentiable and that the directional derivative in addition fulfills an approximation property. The B-Newton method is defined by using the Bouligand derivative. We discuss the feasibility of the B-Newton method. The resulting algorithm is called a B-semismooth Newton method because it can also be interpreted as a semismooth Newton method. At points where the nonlinearity is differentiable, the iterates of the local B-semismooth Newton method coincide with the iterates of the original semismooth Newton method from the literature. The algorithm converges locally superlinearly and is an active-set method. In other words, the Newton equations are finite-dimensional in the vicinity of a root. In particular, linear inverse problems can be solved within a finite number of iterations. The B-Newton directions satisfy a descent property with respect to the square norm of the residual. Due to this fact, we globalize the algorithm in a finite-dimensional setting by inexact line search. A drawback of this approach is that the sequence of iterates might begin to stagnate. Therefore, by a modification of the Newton equation, a globally convergent algorithm is proposed. The modified approach is less efficient in practice than the globalized B-semismooth Newton method. Therefore, we recommend a hybrid algorithm that combines both methods and is locally superlinearly and globally convergent. We present numerical results that demonstrate the efficiency of the methods. Moreover, we discuss the influence of different parameters on the computational cost of the algorithms.

Zusammenfassung

Diese Arbeit behandelt die Globalisierung eines halbglatten Newtonverfahrens für die ℓ_1-Tikhonovregularisierung. Dieses Regularisierungsverfahren für inverse Probleme mit dünnbesetzten Lösungen führt zu einem konvexen, nichtglatten Minimierungsproblem. Hierbei wird angenommen, dass Lösungen des inversen Problems in einer gegebenen Basis durch wenige Basisvektoren repräsentiert werden können. Solche Problemstellungen treten in zahlreichen Anwendungen aus Naturwissenschaft und Technik auf. Wir betrachten diskrete Probleme über dem Folgenraum ℓ_2. Die Optimalitätsbedingungen der in dieser Arbeit betrachteten Minimierungsaufgabe führen zu einem Nullstellenproblem einer nichtlinearen, lokal Lipschitz-stetigen Abbildung. Diese Abbildung ist zwar nicht Fréchet-differenzierbar, allerdings besitzt sie eine (nicht eindeutige) Newton-Ableitung. Basierend auf diesem verallgemeinerten Ableitungskonzept wurde in der Literatur ein lokal superlinear konvergentes, halbglattes Newtonverfahren vorgeschlagen. Die Konvergenz lokaler Newtonverfahren ist jedoch im Allgemeinen nicht für jede Startnäherung gesichert. Mit dem Ziel der Globalisierung des Verfahrens betrachten wir in dieser Arbeit ein B(ouligand)-Newtonverfahren. Wir zeigen, dass die Nichtlinearität Bouligand-differenzierbar ist. Das heißt, dass die Abbildung richtungsdifferenzierbar ist und die Richtungsableitung zusätzlich eine Approximationseigenschaft erfüllt. Das B-Newtonverfahren basiert auf dieser verallgemeinerten Ableitung. Wir diskutieren die Durchführbarkeit des B-Newtonverfahrens. Es stellt sich heraus, dass es sich hierbei ebenfalls um ein halbglattes Newtonverfahren, also um ein B-halbglattes Newtonverfahren, handelt. An Differenzierbarkeitsstellen der Nichtlinearität stimmen die Iterierten des B-halbglatten Newtonverfahrens mit denen des ursprünglichen, halbglatten Newtonverfahrens aus der Literatur überein. Der Algorithmus konvergiert lokal superlinear und ist eine Aktive-Mengen-Methode, das heißt, die Newtongleichungen sind in der Nähe einer Nullstelle endlich-dimensional. Insbesondere können lineare inverse Probleme in endlich vielen Schritten gelöst werden. Die B-Newtonrichtungen besitzen eine Abstiegseigenschaft bezüglich der Quadratnorm des Residuums. Mit Hilfe dieser Eigenschaft wird der Algorithmus im Endlichdimensionalen durch inexakte Liniensuche globalisiert. Ein Schwachpunkt des globalisierten Algorithmus ist, dass die Folge der Iterierten zu stagnieren beginnen kann. Durch eine Abwandlung der Newtongleichung wird ein modifiziertes Verfahren hergeleitet, das ohne Zusatzannahmen global konvergiert. Da dieser modifizierte Algorithmus in der Praxis aufwendiger als das globalisierte B-halbglatte Newtonverfahren ist, wird ein hybrides Verfahren vorgeschlagen. Dieser Algorithmus kombiniert beide Verfahren und ist lokal superlinear und global konvergent. Numerische Resultate demonstrieren die Effizienz der Verfahren. Außerdem wird der Einfluss verschiedener Parameter auf den Rechenaufwand der Algorithmen diskutiert.

ACKNOWLEDGMENTS

The acknowledgments are left out in this version due to data privacy reasons.

Contents

Introduction

In natural sciences and engineering, unknown quantities can often only be deduced from indirect measurements. Such problems are called *inverse problems* [38,74,76,96,126]. Numerous applications come for instance from image processing, medical imaging, nondestructive material testing, geophysics and financial mathematics. In this thesis, we are concerned with the efficient numerical treatment of a special class of discrete, possibly infinite-dimensional inverse problems where a solution can be represented in some basis by a finite number of basis vectors. This property is called *sparsity*. A well-known regularization method for inverse problems with sparsity constraints is ℓ_1-Tikhonov regularization. Here, the optimality conditions of a nonsmooth minimization problem lead to a zero-finding problem that is defined by a nonsmooth but locally Lipschitz continuous nonlinearity. Linear inverse problems with sparsity constraints can efficiently be solved with a *semismooth Newton method*, a generalized Newton method introduced by Herzog and Lorenz [58] as long as an initial guess is available that is sufficiently close to a solution. Muoi, Hào, Maass and Pidcock [103] generalized this method to nonlinear inverse problems and proposed a *quasi-Newton method*. Semismooth Newton methods are very efficient in practice and competitive with other state-of-the-art methods. However, they are not globally convergent in general. In other words, if the initial guess is poor, *local* semismooth Newton methods may fail to solve the problem. In this work, we address the *globalization* of the latter methods to ensure convergence for any initial guess.

Newton's method is a well-known iterative algorithm for nonlinear zero-finding problems. In mathematical programming, Newton's method can be applied to the optimality conditions of unconstrained, convex minimization problems [31,32,52]. The nonlinearity is linearized at the current iterate and the root of this linearization is defined as the new iterate. Hence, in each step, a linear system has to be solved that is called the Newton equation. Under appropriate assumptions, the sequence of iterates approaches a root of the nonlinear function. The linearizations are defined using the Fréchet derivative. Therefore, the nonlinearity has to be smooth for the application of Newton's method. Newton's method is competitive due to its locally quadratic convergence in a neighborhood of a root. To this end, one has to require some regularity assumptions on the nonlinearity [31,32,50,145].

Zero-finding problems for nonlinearities that are not Fréchet differentiable are common in applications. The research in this area was influenced by emerging applications from economy and engineering that can be formulated as nonlinear programs or variational inequalities [41,71]. In the case of nonsmooth nonlinearities, one may make use of some kind of *generalized derivative* to apply Newton's method at points where the function is not differentiable. Such *generalized Newton methods* for nonsmooth zero-finding problems have been studied extensively in the last decades [41,109,114]. For a review on the developments in the field, we refer to the preface of [40] and to [71,114], see also the

references therein. In the next paragraph, we will give a short overview of some concepts of generalized derivatives and related generalized Newton methods.

Kojima and Shindo [85] extended Newton's method to piecewise continuously differentiable functions between finite-dimensional spaces. A *successive approximation method* [113] was introduced by Qi and Chen. There, the nonsmooth nonlinearity is approximated at each Newton iterate by a smooth function. Some other methods are based on different types of generalized derivatives. The concept of *generalized Jacobians* for finite-dimensional, locally Lipschitz continuous mappings was introduced by Clarke [19] and Mifflin [98] for functionals and by Qi and Sun [115] for mappings, see also [20]. The generalized Newton equation in *semismooth Newton methods*, proposed by Kummer [87–89] and Qi and Sun [115], is defined by using generalized Jacobians. Note that this generalized derivative is set-valued at points where the mapping is not smooth. The definition of generalized Jacobians in finite-dimensional spaces is based on *Rademacher's theorem* [117] which states that every locally Lipschitz continuous mapping between finite-dimensional spaces is differentiable almost everywhere. Chen, Nashed and Qi [18] as well as Ulbrich [136, 137] extended the concepts of generalized derivatives and semismooth Newton methods to infinite-dimensional spaces. We also refer to the work [73] of Hintermüller, Ito and Kunisch who considered a semismooth Newton method for constrained optimal control problems. In [107], Pang introduced the *B(ouligand)-Newton method* for finite-dimensional mappings based on the *Bouligand derivative*. The Bouligand derivative is a generalized derivative concept that was introduced by Robinson [121]. Shapiro [128] showed that the Bouligand derivative of a locally Lipschitz continuous mapping between finite-dimensional spaces coincides with its directional derivative. Moreover, we refer to the related work [122] where a Newton-type method using *point-based approximations* was introduced in a finite-dimensional setting. Furthermore, some modifications of the B-Newton method were proposed by Pang [108] and by Han, Pang and Rangaraj [63], see also the work [77] of Ito and Kunisch. Qi [111] proved new convergence results for several generalized Newton methods for finite-dimensional zero-finding problems. Recently, Hoheisel, Kanzow, Mordukhovich and Phan [75] introduced a version of Newton's method based on *graphical derivatives*. There, in contrast to the aforementioned methods, the (finite-dimensional) nonlinearity is only required to be continuous.

A well-known drawback of Newton's method is the *cycling effect*. Here, a cycle of iterates arises that is repeated over and over again without approaching a root of the nonlinearity. Hence, the method may fail to converge, i.e., it does not converge globally in general. For that reason, globalization strategies have been developed in the literature to ensure convergence. Several approaches for classical Newton methods are treated in [31, 32]. Famous strategies for nonsmooth mappings are *(inexact) line search*, *path search* and *trust-region* methods [41]. The authors of the works [63, 77, 107, 108, 111, 113] mentioned in the last paragraph discussed the globalization of the respective methods by using line search strategies. For further applications of line search methods, we refer to [27, 28, 45, 54, 70, 72, 81, 100, 130, 143, 144, 147]. Path search globalization strategies are, e.g., treated in [14, 118]. Trust-region methods are considered, e.g., in [116, 132, 136].

In this thesis, we consider a nonsmooth zero-finding problem that appears when applying ℓ_1-Tikhonov regularization to inverse problems with sparsity constraints. In order to illustrate the term *inverse problems*, let us for the moment assume that data f are avail-

able from indirect measurements of a quantity of interest u. For instance, the data f could be a blurred image taken while moving the camera, and u could be the unknown sharp image. Additionally, we assume that the *forward model*, i.e. the mapping K that maps the unknown quantity u to the data f, is known and can be formulated as an operator equation

$$K(u) = f,$$

where K is called the *forward operator*. The operator K could, e.g., be a blurring operator that maps a sharp image u to a blurred image f. The *direct problem* is the calculation of f from the given u. The *inverse problem* is the computation of u from the given data f. In this case, the operator equation needs to be inverted to determine u. Operator equations that arise in inverse problems are often *ill-posed* in the sense that the inverse of K does not exist or that K is not continuously invertible. Hence, the unknown u does not continuously depend on the data f. In practice, the data f are corrupted by measurement errors. In other words, only noisy data $f^\delta \approx f$ are available. Here, the subscript δ indicates the noise that is here assumed to be deterministic. The direct inversion of such an operator equation may lead to large errors in the solution, even if the relative noise level $\|f - f^\delta\| / \|f\|$ is small, see [38, 96].

Regularization methods enable the stable numerical solution of inverse problems. In this thesis, we are concerned with a discrete, ill-posed operator equation, K mapping the sequence space $\ell_2 = \ell_2(\mathcal{N})$ into a separable Hilbert space H and data $f^\delta \in H$, where we have either $\mathcal{N} = \mathbb{N}$ or $\mathcal{N} = \{1, \dots, n\}$. A popular regularization strategy is *Tikhonov regularization* [38, 74, 76, 96, 126]. Here, a weighted sum of a *discrepancy term* and a *penalty term* is minimized. In ℓ_2-*Tikhonov regularization*, the discrepancy term is defined as the square norm of the residual, and solutions with small ℓ_2-norm are preferred by adding an ℓ_2-penalty term. Hence, one considers

$$\min_{\mathbf{u} \in \ell_2} \frac{1}{2} \|K(\mathbf{u}) - f^\delta\|_H^2 + w \sum_{k \in \mathcal{N}} |u_k|^2,$$

where $\mathbf{u} = (u_k)_{k \in \mathcal{N}}$ and $w > 0$ is a *regularization parameter* that balances the data fidelity and the size of the ℓ_2-norm of the minimizer. The discrepancy term describes the mismatch between a solution $\mathbf{u}$ and the data f^δ. The penalty term models some a priori assumption on $\mathbf{u}$. In the following, we assume that a solution $\mathbf{u}$ to the inverse problem is sparse. Therefore, we use a sparsity-inducing, weighted ℓ_1-penalty term. If the inverse problem has a sparse solution, convincing reconstructions are presented by using a sparsity-inducing penalty term, e.g., for parameter identification problems with piecewise constant solutions like electrical impedance tomography [51, 78, 79]. These reconstructions often have a higher quality than reconstructions using an ℓ_2-penalty term, see also [6] for results for an inverse heat conduction problem. We consider a generic discrepancy term g that depends on the forward operator K and the (noisy) data f^δ. This leads to the minimization problem of ℓ_1-*Tikhonov regularization*,

$$\min_{\mathbf{u} \in \ell_2} g(\mathbf{u}; K, f^\delta) + \sum_{k \in \mathcal{N}} w_k |u_k|.$$

Here, each entry of a minimizer is either regularized individually or the regularization sequence $w_k \equiv w$ is chosen equal. The sizes of the *regularization parameters* w_k control the

sparsity of a minimizer. The ℓ_1-penalty term ensures that a minimizer has only finitely many nonzero entries if the regularization parameters satisfy $w_k \geq w_0 > 0$ [58]. The larger the regularization parameters are chosen, the sparser is a minimizer $\mathbf{u}$. The functional g should be chosen according to the type of noise that is expected to be contained in the data, see, e.g., [21,76]. If the data are perturbed by Gaussian noise, one may use the square norm of the residual as the discrepancy term, see, e.g., [76]. Note that in this thesis, we assume g to be smooth and strictly convex. Other applications involving the considered minimization problem come from regression analysis and machine learning where the ℓ_1-penalty term enforces a sparse solution, see, e.g., [15,16,49,82,90,106,129,133,134].

ℓ_1-Tikhonov regularization was considered in many publications, see, e.g., [3, 26, 56, 57, 76,101], the topical review [79] as well as the references therein. For the examples considered in this thesis, the regularized solutions to the inverse problem under consideration converge to a solution to the operator equation $K(u) = f$ with unperturbed right-hand side f, as the noise level tends to zero. To show this, one usually assumes that the regularization parameters w_k are chosen according to some parameter choice strategy, that appropriate source conditions hold and that the true solution is sparse, see [56, 57, 102]. There exist a number of alternative penalization strategies. As stated above, an ℓ_2-penalty term can be used. Other penalty terms that are applied in practice are, for instance, ℓ_p-(quasi) norms with $p \in (0, 1)$, see, e.g., [148], ℓ_p-norms with $p \in (1, 2)$, see, e.g., [56, 119], *total variation* penalties [62, 104] or *elastic net* penalty terms, where a weighted sum of an ℓ_1- and an ℓ_2-norm is considered, see, e.g., [76]. The choice of the penalties depends on the a priori assumption on the unknown solution.

In contrast to ℓ_2-Tikhonov regularization, the objective function in the minimization problem of ℓ_1-Tikhonov regularization, i.e., the sum of the discrepancy and the ℓ_1-penalty term, is not differentiable because of the absolute values in the penalty term. This fact makes the numerical minimization more difficult. Nevertheless, the optimality conditions of the minimization problem from ℓ_1-Tikhonov regularization can, under some smoothness assumptions on the discrepancy term g, be formulated as a zero-finding problem

$$\mathbf{F}(\mathbf{u}) = \mathbf{0},$$

for a nonsmooth, locally Lipschitz continuous mapping $\mathbf{F}\colon \ell_2 \to \ell_2$, see, e.g., [58, 93, 99, 103]. Many solution algorithms are based on this formulation of the optimality conditions. There exists a whole range of algorithms for the numerical solution of the above nonsmooth minimization problem for ℓ_1-Tikhonov regularization or special cases thereof, see, e.g., [3, 5, 10, 12, 23, 35, 37, 44, 55, 61, 83, 93–95, 104–106, 120, 127, 129, 139, 142, 146]. A famous example is the *iterative soft-thresholding algorithm* introduced by Figueiredo and Nowak [43] and Daubechies, Defrise and De Mol [26]. Many iterative methods for ℓ_1-Tikhonov regularization are linearly convergent, see, e.g., [11,60]. The semismooth Newton methods from [58, 103] mentioned above are locally superlinearly convergent but depend on an initial guess which has to be chosen in the vicinity of a root to ensure convergence.

Globalized Newton-type methods for ℓ_1-penalized minimization in a finite-dimensional setting were developed in recent years. These algorithms follow different strategies. Milzarek and Ulbrich [99] use a *filter globalization* technique. There, in each iteration a semismooth Newton step is performed and if the new iterate is not accepted, one it-

eration of a globally convergent fixed-point method is executed. Byrd, Chin, Nocedal and Öztoprak [15] introduced a *family of second-order methods*. There, various choices of generalized Jacobians at points where the nonlinearity is not differentiable are proposed. The authors present a globalization procedure which is based on a reduction of the cardinality of the entries of the iterates that do not yet fulfill the optimality condition of the minimization problem. Lee, Sun and Saunders [90] and Byrd, Nocedal and Öztoprak [16] proposed *proximal Newton-type methods*. Here, the sum of a quadratic approximation of the discrepancy term around the current iterate and an ℓ_1-penalty term is minimized to obtain a proximal Newton direction. Hence, a piecewise quadratic functional is minimized in each proximal Newton step. These methods are globalized by line search strategies. Keskar, Nocedal, Öztoprak and Wächter [82] introduced a *orthant-based method with active set prediction* that is globalized using a line search procedure involving a reference point computed by a soft-thresholding iterate. In a different approach of Fountoulakis and Gondzio [49], the ℓ_1-norm is replaced by a smooth *pseudo-Huber function* and a *primal-dual Newton conjugate gradient method* is considered.

In this work, we focus on the globalization of the semismooth Newton methods from [58, 103] for ℓ_1-Tikhonov regularization. To this end, we have to show that the nonlinearity $\mathbf{F}$ under consideration is Bouligand differentiable on the entire sequence space $\ell_2 = \ell_2(\mathbb{N})$. That is, the nonlinearity $\mathbf{F}$ is directionally differentiable and satisfies an approximation property. For that reason, the discrepancy term g has to fulfill appropriate smoothness requirements. In order to develop a globally convergent Newton-type method, we start with a *local* B-Newton method based on the Bouligand derivative. It is shown that the resulting generalized Newton equation is uniquely solvable and is equivalent to a *mixed linear complementarity problem* [24,123]. However, the generalized Newton equation usually reduces to a linear system. It turns out that the local B-Newton method can be interpreted as a semismooth Newton method. Therefore, our algorithm is called *B-semismooth Newton method*. The iterates of the B-semismooth Newton algorithm coincide with the iterates of the semismooth Newton methods from [58, 103] at points where the nonlinearity $\mathbf{F}$ is smooth. By construction, the generalized Newton directions of B-Newton methods are descent directions with respect to the square norm of the residual, cf. [107]. Hence, the Newton directions of [58, 103] are descent directions if the nonlinearity $\mathbf{F}$ is smooth at the actual iterate. In a neighborhood of a root, the generalized Newton equations of the B-semismooth Newton method are finite-dimensional and their sizes usually decrease along the iteration. Hence, like the methods from [58, 103], the algorithm can be efficiently implemented as an *active-set method*. Based on the descent property of the generalized Newton directions, we propose an inexact line search strategy to globalize the B-semismooth Newton algorithm in a finite-dimensional setting, i.e. in the case $\mathcal{N} = \{1, \dots, n\}$, cf. [107]. In that way, sufficient descent is obtained by introducing suitable damping parameters. Global convergence is ensured as long as the damping parameters are bounded away from zero. Otherwise, a technical assumption on the a priori unknown accumulation point of the sequence of iterates is necessary. That is why we propose a modification of the Newton directions. This modification is motivated by a similar work [108] of Pang who introduced a modified B-Newton method for nonlinear complementarity problems, nonlinear programs and variational inequalities. We take advantage of similarities of the nonlinearity $\mathbf{F}$ considered in this thesis with a

min-*formulation* for nonlinear complementarity problems. Inspired by loc. cit., we modify the generalized Newton equation in such a way that the resulting algorithm is globally convergent. The *modified B-semismooth Newton method* also fits into the framework of [63] which is a generalization of [108]. The generalized Newton equation of our modified algorithm usually includes a linear complementarity problem. Therefore, the computational cost of the modified algorithm is higher than for the (unmodified) B-semismooth Newton method. For that reason, we propose a hybrid algorithm that combines both methods to an efficient, globally convergent method.

The outline of this thesis is as follows. In Chapter 1, we introduce our notation and we review some concepts of derivatives for smooth and nonsmooth functions. Using tools from convex analysis, we recall the derivation of the nonsmooth zero-finding problem considered in this thesis. Moreover, we shortly review some concepts of generalized Newton methods: the semismooth Newton method, the B-Newton method and a modified B-Newton method. Chapter 2 deals with a local and a global B-semismooth Newton method for ℓ_1-Tikhonov regularization. The properties of the considered nonlinearity are studied. Moreover, convergence results of the proposed algorithms are shown. To overcome the theoretical drawback that the method from Chapter 2 may begin to stagnate, a modified method is proposed in Chapter 3. Global convergence and locally quadratic convergence are verified. A hybrid method is proposed by combining the algorithms from Chapters 2 and 3. Numerical results for the proposed algorithms are presented in Chapter 4. We consider the linear inverse problem of inverse integration and a nonlinear parameter identification problem in an elliptic partial differential equation. The computational cost of the different methods as well as the influence of parameters on the convergence properties of the algorithms are discussed.

The results of this thesis are based on joint work with Prof. Dr. Thorsten Raasch. For a similar setting, the results from Chapter 2 were published in [66]. Chapter 3 is a part of our preprint [67]. The forward code for a parameter identification problem used in Chapter 4 was implemented by Dr. Fabrice Delbary.

Chapter 1

Preliminaries

In this chapter, we review different concepts of derivatives for smooth and nonsmooth mappings as well as generalized Newton methods. All results stated in this chapter are well-known. Let us first introduce some notation.

1.1 Sequence and function spaces

In this section, we recall well-known sequence spaces as well as function spaces that are considered in this thesis using the textbooks [2, 39]. Let X, Y be real Banach spaces. $L(X,Y)$ denotes the space of linear and bounded operators mapping X to Y. In the case $X = Y$, we shortly write $L(X) := L(X,X)$. For $1 \leq p \leq \infty$, we recall the sequence space $\ell_p = \ell_p(\mathbb{N}) := \{\mathbf{u} = (u_k)_{k\in\mathbb{N}} : \|\mathbf{u}\|_{\ell_p} < \infty\}$, where

$$\|\mathbf{u}\|_{\ell_p} := \begin{cases} \Big(\sum\limits_{k=1}^{\infty} |u_k|^p\Big)^{1/p}, & 1 \leq p < \infty, \\ \sup\limits_{k\in\mathbb{N}} |u_k|, & p = \infty. \end{cases}$$

In the finite-dimensional space $\mathbb{R}^n$, the norms $\|\cdot\|_p$, $1 \leq p \leq \infty$, are defined as

$$\|\mathbf{u}\|_p := \begin{cases} \Big(\sum\limits_{k=1}^{n} |u_k|^p\Big)^{1/p}, & 1 \leq p < \infty, \\ \max\limits_{1\leq k\leq n} |u_k|, & p = \infty. \end{cases}$$

In the following, we denote vectors $\mathbf{u} \in \mathbb{R}^n$ and matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ as well as sequences $\mathbf{u} \in \ell_2$ and operators $\mathbf{K} \in L(\ell_2)$ in boldface.

Let Ω be an open, bounded subset of $\mathbb{R}^n$ with smooth boundary $\partial\Omega$. Let $C(\Omega)$ denote the space of continuous functions and $C^k(\Omega)$ for $k = 1, \ldots, \infty$ denote the space of k-times continuously differentiable functions $u\colon \Omega \to \mathbb{R}$. The subspace $C_c^k(\Omega) \subset C^k(\Omega)$ denotes the subspace of all k-times continuously differentiable functions with compact support. For $1 \leq p \leq \infty$, we recall the spaces

$$L_p(\Omega) := \{u\colon \Omega \to \mathbb{R} : \|u\|_{L_p(\Omega)} < \infty\},$$

where

$$\|u\|_{L_p(\Omega)} := \begin{cases} (\int_\Omega |u|^p \,\mathrm{d}\mathbf{x})^{1/p}, & 1 \leq p < \infty, \\ \operatorname{ess\,sup}_\Omega |u| = \inf\{c \in \mathbb{R} : |\{u > c\}| = 0\}, & p = \infty. \end{cases}$$

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \geq 0$, $i = 1, \ldots, n$, be a multi-index. A function $v \in L_2(\Omega)$ is called the *$\boldsymbol{\alpha}$-th weak partial derivative* $D^{\boldsymbol{\alpha}} u$ of u if we have for all *test functions* $\Phi \in C_c^\infty(\Omega)$,

$$\int_\Omega u D^{\boldsymbol{\alpha}} \Phi \, \mathrm{d}\mathbf{x} = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega v \Phi \, \mathrm{d}\mathbf{x}.$$

We recall the *Sobolev space*

$$H^1(\Omega) := \{u \colon \Omega \to \mathbb{R} : \|u\|_{H^1(\Omega)} < \infty\},$$

where

$$\|u\|_{H^1(\Omega)} := \big(\|u\|^2_{L_2(\Omega)} + \sum_{|\boldsymbol{\alpha}|=1} \|D^{\boldsymbol{\alpha}} u\|^2_{L_2(\Omega)}\big)^{1/2}.$$

We cite the well-known *trace theorem* for the space $H^1(\Omega)$, see, e.g., [39, p. 258].

Theorem 1.1 *Let $\Omega \subset \mathbb{R}^n$ be open and bounded and let the boundary $\partial\Omega$ be continuously differentiable. Then, there exists a bounded linear operator $T \colon H^1(\Omega) \to L_2(\partial\Omega)$ and a constant $C > 0$ that depends only on Ω with*

$$\|Tu\|_{L_2(\partial\Omega)} \leq C \|u\|_{H^1(\Omega)},$$

for each $u \in H^1(\Omega)$ and $Tu = u|_{\partial\Omega}$ if $u \in H^1(\Omega) \cap C(\overline{\Omega})$.

With Theorem 1.1, we are in the position to define the subspace $H_0^1(\Omega) \subset H^1(\Omega)$ as

$$H_0^1(\Omega) := \{u \in H^1(\Omega) : u|_{\partial\Omega} = 0\}.$$

In the next section, we consider some classical concepts of derivatives.

1.2 Fréchet and directional derivatives

In the following, we mainly use the textbook [2] and the paper [128]. Concerning the Fréchet derivative of an operator $F \colon X \to Y$, we cite [2, Definition 2.45].

Definition 1.2 (Fréchet derivative) Let D be an open subset of X and $F \colon D \to Y$. F is *Fréchet differentiable* at $u \in D$, if there exists a linear operator $DF(u) \in L(X, Y)$ such that

$$\lim_{h \to 0} \frac{\|F(u+h) - F(u) - DF(u)h\|_Y}{\|h\|_X} = 0. \tag{1.1}$$

We call F *twice Fréchet differentiable* at $u \in D$, if there exists a linear operator $D^2F(u) \in L(X, L(X, Y))$ with

$$\lim_{h \to 0} \frac{\|DF(u+h) - DF(u) - D^2F(u)h\|_{L(X,Y)}}{\|h\|_X} = 0.$$

F is called (twice) Fréchet differentiable in D if F is (twice) Fréchet differentiable at every $u \in D$.

Let H be a real Hilbert space with scalar product $\langle \cdot, \cdot \rangle_H$. In the following, we identify the Fréchet derivative Dg of a functional $g\colon H \to \mathbb{R}$ with an element $z \in H$, i.e., $Dg(u)d = \langle z, d \rangle_H$, and denote $z = Dg(u)$. Analogously, we identify the second Fréchet derivative $D^2g(u)$ of g with an element $K \in L(H)$, i.e., $(D^2g(u)h)h' = \langle h, Kh' \rangle_H$ and denote $K = D^2g(u)$, cf. [2, Remark 2.44]. For finite-dimensional functionals $g\colon \mathbb{R}^n \to \mathbb{R}$, $\nabla g(\mathbf{u}) \in \mathbb{R}^n$ and $\nabla^2 g(\mathbf{u}) \in \mathbb{R}^{n \times n}$ denote the gradient and the Hessian of g at $\mathbf{u} \in \mathbb{R}^n$, respectively.

The following definition of directional differentiability can be found, e.g., in [128].

Definition 1.3 (Directional derivative) A mapping $F\colon X \to Y$ is called *directionally differentiable* at $u \in X$, if the limit

$$\lim_{t \searrow 0} \frac{F(u + th) - F(u)}{t} \tag{1.2}$$

exists for every $h \in X$. The directional derivative at u in the direction h is given by the limit (1.2) and is denoted by $F'(u, h)$. F is called directionally differentiable if F is directionally differentiable at every $u \in X$.

For the sake of completeness, we consider the well-known concept of Gâteaux derivatives, additionally requiring linearity of the directional derivative. We cite [2, Definition 2.43].

Definition 1.4 (Gâteaux derivative) Let D be an open subset of X. A mapping $F\colon D \to Y$ is called *Gâteaux differentiable* at $u \in D$, if F is directionally differentiable at u and if there exists a linear and bounded operator $DF(u) \in L(X, Y)$ with $F'(u, h) = DF(u)h$. F is called Gâteaux differentiable in D if F is Gâteaux differentiable at every $u \in D$.

The following concept of directional differentiability from [128], see also the references therein, is needed in order to prove a chain rule for the directional derivative from Definition 1.3.

Definition 1.5 (Hadamard directional derivative) A mapping $F\colon X \to Y$ is called *directionally differentiable in the sense of Hadamard* at $u \in X$, if the double limit

$$F'(u, h) = \lim_{t \searrow 0,\, h' \to h} \frac{F(u + th') - F(u)}{t} \tag{1.3}$$

exists for every $h \in X$. F is called directionally differentiable in the sense of Hadamard in an open subset $D \subseteq X$ if F is directionally differentiable in the sense of Hadamard at every $u \in D$.

A mapping $F\colon X \to Y$ is called *locally Lipschitz continuous* in $D \subseteq X$ if, for every $u \in D$, there exists a neighborhood U of u and a constant $L > 0$ such that $\|F(u) - F(v)\|_Y \leq L\|u - v\|_X$ for all $v \in U$, cf. [2, Definition 1.46]. Note that every locally Lipschitz continuous and directionally differentiable operator $F\colon X \to Y$ is directionally differentiable in the sense of Hadamard, see [128] and the references therein. Hence, the following chain rule also holds for locally Lipschitz continuous mappings. We cite the next lemma from [128, Proposition 3.6 (i)], see also [124, Theorem 3.1.1] or [53, Satz 5.33], for the chain rule in case of finite-dimensional mappings.

Lemma 1.6 (Chain rule for directional derivatives) *Let $F\colon \tilde{Y} \to Y$ be directionally differentiable in the sense of Hadamard and $T\colon X \to \tilde{Y}$ be directionally differentiable. Then, the composition $F \circ T\colon X \to Y$ is directionally differentiable and it holds the chain rule*

$$(F \circ T)'(u,h) = F'(T(u), T'(u,h)).$$

Proof. The claim follows from

$$\begin{aligned}\lim_{t\searrow 0} \frac{(F\circ T)(u+th) - (F\circ T)(u)}{t} &= \lim_{t\searrow 0} \frac{F(T(u) + t\frac{T(u+th)-T(u)}{t}) - F(T(u))}{t} \\ &= F'(T(u), T'(u,h)),\end{aligned}$$

using the Hadamard directional differentiability of F, see (1.3) with $h' = \frac{T(u+th)-T(u)}{t}$, and the directional differentiability of T,

$$\lim_{t\searrow 0} \frac{T(u+th)-T(u)}{t} = T'(u,h).$$

□

1.3 A nonsmooth zero-finding problem

In this section, we consider the optimality conditions for the minimization problem

$$\min_{\mathbf{u}\in\ell_2} g(\mathbf{u}) + \sum_{k=1}^{\infty} w_k |u_k|, \tag{1.4}$$

where $g\colon \ell_2 \to \mathbb{R}$ is continuously Fréchet differentiable and $w_k \geq w_0 > 0$. In the following definition, we cite elementary properties of mappings $J\colon X \to \mathbb{R} \cup \{-\infty, \infty\}$. The definitions are taken from [2, Definition 1.4, Definition 8.6, Definition 11.10], [9, Definition 6.7] and [149, pp. 507, 859-860].

Definition 1.7 Let $J\colon X \to \mathbb{R} \cup \{-\infty, \infty\}$.

(i) J is called *proper* if $\mathrm{dom}(J) := \{u \in X : J(u) < \infty\} \neq \emptyset$ and $-\infty \notin \mathrm{im}(J) := J(X)$.

(ii) J is called *lower semicontinuous* if it holds

$$\liminf_{j\to\infty} J(u^{(j)}) \geq J(u),$$

for every sequence $\{u^{(j)}\}_j$ and $u \in X$ with $u^{(j)} \to u$, $j \to \infty$.

(iii) J is called *coercive* if $\lim_{\|u\|_X\to\infty} J(u) = \infty$.

(iv) A proper function J is called *convex* if $J(\lambda u + (1-\lambda)v) \leq \lambda J(u) + (1-\lambda)J(v)$ for all $u, v \in X$ and $\lambda \in (0,1)$.

(v) A proper function J is called *strictly convex* if $J(\lambda u + (1-\lambda)v) < \lambda J(u) + (1-\lambda)J(v)$ for all $u, v \in X$, $u \neq v$ and $\lambda \in (0,1)$.

In the following proposition, we illustrate the above properties for the weighted ℓ_1-penalty term in the minimization problem (1.4).

Proposition 1.8 *The functional* $h\colon \ell_2 \to \mathbb{R} \cup \{\infty\}$,

$$h(\mathbf{u}) := \sum_{k=1}^{\infty} w_k |u_k|,$$

where $w_k \geq w_0 > 0$, *is lower semicontinuous, proper, coercive and convex.*

Proof. The function h is proper because it is bounded from below by 0 and $\mathbf{0} \in \operatorname{dom} h$. The convexity follows directly from an application of the triangle inequality. Let us now consider the lower semicontinuity of h. Let $\{\mathbf{u}^{(j)}\}_j$ be a sequence with $\|\mathbf{u}^{(j)} - \mathbf{u}\|_{\ell_2} \to 0$, $j \to \infty$. Then, it follows with Fatou's lemma, see, e.g., [9, Lemma 2.46],

$$\liminf_{j\to\infty} h(\mathbf{u}^{(j)}) = \liminf_{j\to\infty} \sum_{k=1}^{\infty} w_k |u_k^{(j)}| \geq \sum_{k=1}^{\infty} w_k \liminf_{j\to\infty} |u_k^{(j)}| = \sum_{k=1}^{\infty} w_k |u_k| = h(\mathbf{u}).$$

The coercivity of h directly follows from

$$h(\mathbf{u}) = \sum_{k=1}^{\infty} w_k |u_k| \geq w_0 \|\mathbf{u}\|_{\ell_1} \geq w_0 \|\mathbf{u}\|_{\ell_2},$$

finishing the proof. □

In the following, H denotes a (real) Hilbert space with scalar product $\langle \cdot, \cdot \rangle_H$. Second-order derivative information can be used to characterize convexity [2]. A twice Gâteaux differentiable functional $J\colon H \to \mathbb{R}$ is convex if and only if

$$\langle h, D^2 J(u) h \rangle_H \geq 0 \qquad \text{for all } h \in H,$$

see [2, Proposition 17.10]. Additionally, it was shown in [2, Proposition 17.13] that

$$\langle h, D^2 J(u) h \rangle_H > 0 \qquad \text{for all } h \in H, h \neq 0,$$

implies strict convexity of J.

The sum of two convex functions is convex [2, Proposition 8.15]. We say that $u \in H$ is a *local minimizer* of a proper function $J\colon H \to \mathbb{R} \cup \{\infty\}$ if there exists a neighborhood U of u with $J(u) \leq J(v)$ for all $v \in U$ and u is a *minimizer* of J if $J(u) \leq J(v)$ holds for all $v \in H$ [2, Definition 11.3]. If, in addition, J is convex, then every local minimizer of J is a minimizer of J [2, Proposition 11.4].

In order to derive optimality conditions for the nonsmooth minimization problem (1.4), we need to consider subgradients and the well-known Fermat's rule using the textbook [2]. We begin with citing the following definition from [2, Definition 16.1].

Definition 1.9 (Subdifferential) Let H be a Hilbert space with scalar product $\langle \cdot, \cdot \rangle_H$. For a proper functional $J\colon H \to (-\infty, \infty]$, the *subdifferential* of J at u is defined by the set

$$\partial J(u) := \{x \in H : \langle v - u, x \rangle_H + J(u) \leq J(v) \quad \text{for all } v \in H\}.$$

An element $x \in \partial J(u)$ is called a *subgradient* of J at u. We call J *subdifferentiable* at $u \in H$ if the set $\partial J(u)$ is not empty.

For two sets A, B, and a scalar $\gamma \in \mathbb{R}$, we define $A + \gamma B = \{a + \gamma b : a \in A, b \in B\}$. The subdifferential $\partial J \colon H \to 2^H$ is a set-valued operator which maps H to the *family of all subsets* of H, denoted by 2^H, cf. [2]. If $J \colon H \to (-\infty, \infty]$ is proper, convex and continuous at $u \in \operatorname{dom}(J)$, then $\partial J(u)$ is nonempty, see [2, Proposition 16.14(ii)]. If $J \colon H \to (-\infty, \infty]$ is proper, convex and Gâteaux differentiable at $u \in \operatorname{dom}(J)$, where the Gâteaux derivative is identified with an element $DJ(u) \in H$ as above, we have $\partial J(u) = \{DJ(u)\}$, see [2, Proposition 17.26]. We cite the following definition from [2, Section 1.2] which treats the inverse of set-valued operators.

Definition 1.10 (Inverse of set-valued operators) For a set-valued operator $\Psi \colon H \to 2^H$, we denote the *domain* $\operatorname{dom} \Psi := \{u \in H : \Psi(u) \neq \emptyset\}$ and the *image* $\operatorname{im} \Psi := \Psi(X)$. The *graph* of Ψ is defined by

$$\operatorname{graph}(\Psi) := \{(u, v) \in H \times H : v \in \Psi(u)\},$$

and Ψ^{-1} is characterized by its graph

$$\operatorname{graph}(\Psi^{-1}) := \{(v, u) \in H \times H : (u, v) \in \operatorname{graph}(\Psi)\},$$

i.e., we define $\Psi^{-1} \colon \operatorname{im}(\Psi) \to \operatorname{dom} \Psi$, $\Psi^{-1}(v) := \{u \in H : v \in \Psi(u)\}$.

The following proposition can, e.g., be found in [2, Theorem 16.2] or [149, Proposition 32.14], and it is an important tool of convex analysis.

Proposition 1.11 (Fermat's rule) *Let H be a Hilbert space and let $J \colon H \to (-\infty, \infty]$ be proper. Then, $u \in H$ is a minimizer of J if and only if $0 \in \partial J(u)$.*

Proof. We proceed as in the proof of [2, Theorem 16.2]. If u is a minimizer of J, we have $\langle v - u, 0 \rangle_H + J(u) = J(u) \leq J(v)$ for all $v \in H$. Hence, one has $0 \in \partial J(u)$. Conversely, if $0 \in \partial J(u)$, it directly follows from Definition 1.9 that $J(u) \leq J(v)$ for each $v \in H$. □

We next define the soft-thresholding operator which is needed in order to establish the optimality condition for the minimization problem (1.4). The following definition can be, e.g., found in [26] or [58, Definition 2.1].

Definition 1.12 (Soft-thresholding) Let $\mathbf{w} = (w_k)_k$ be a positive weight sequence with $w_k \geq w_0 > 0$. The *soft-thresholding* operator $\mathbf{S}_\mathbf{w} \colon \ell_2 \to \ell_2$ with respect to the weight sequence $\mathbf{w}$ is defined componentwise by

$$(\mathbf{S}_\mathbf{w}(\mathbf{u}))_k = S_{w_k}(u_k) := (|u_k| - w_k)_+ \operatorname{sgn}(u_k), \tag{1.5}$$

where

$$\operatorname{sgn}(x) := \begin{cases} \frac{x}{|x|}, & x \neq 0, \\ 0, & x = 0, \end{cases}$$

and $x_+ := \max\{x, 0\}$.

Note that it was shown in [26, Lemma 2.2], that the soft-thresholding operator is nonexpansive, i.e., it is Lipschitz continuous with Lipschitz constant 1. The following definition of the proximity operator can be found, e.g., in [23] and [2, Definition 12.23].

Definition 1.13 Let H be a Hilbert space and let $h\colon H \to \mathbb{R} \cup \{\infty\}$ be lower semicontinuous, proper and convex. The *proximity operator* $\operatorname{prox}_h\colon H \to H$ of h is defined as

$$\operatorname{prox}_h : u \mapsto \operatorname*{arg\,min}_{v \in H} \; h(v) + \frac{1}{2}\|u - v\|_H^2. \tag{1.6}$$

Note that the proximity operator is well-defined because the objective function in (1.6) has a unique minimizer, see [2, Corollary 11.16]. It was shown in [23, Lemma 2.4] and [2, Proposition 12.27] that proximity operators are nonexpansive, i.e. Lipschitz continuous with Lipschitz constant 1.

Remark 1.14 *The soft-thresholding operator from Definition 1.12 is the proximity operator of the weighted penalty*

$$h\colon \ell_2 \to \mathbb{R} \cup \{\infty\}, \quad h(\mathbf{u}) := \sum_{k=1}^{\infty} w_k |u_k|. \tag{1.7}$$

This fact was shown in [23, Example 2.20] using the proximity operator

$$\operatorname{prox}_{w|\cdot|}(u) = (|u| - w)_+ \operatorname{sgn}(u),$$

of the weighted absolute value $w|\cdot|$ with weight $w > 0$ and that the proximity operator of h can here be calculated componentwise, see [23, Example 2.19] and [2, Proposition 23.34], i.e.,

$$\operatorname{prox}_h \mathbf{u} = \big(\operatorname{prox}_{w_k|\cdot|}(u_k)\big)_{k\in\mathbb{N}} = \big((|u_k| - w_k)_+ \operatorname{sgn}(u_k)\big)_{k\in\mathbb{N}} = \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u}).$$

The following proposition treats resolvents of subdifferentials of lower semicontinuous, proper and convex functions. We use [58], [2, Proposition 16.34] and [149, Proposition 32.17, Corollary 32.30].

Proposition 1.15 *Let H be a Hilbert space and let $h\colon H \to \mathbb{R} \cup \{\infty\}$ be lower semicontinuous, proper and convex and $\gamma > 0$ an arbitrary parameter. Then, the resolvent $(I + \gamma\partial h)^{-1}$ of the subdifferential ∂h exists, is single-valued and*

$$\operatorname{prox}_{\gamma h} = (I + \gamma\partial h)^{-1}. \tag{1.8}$$

Now, we are in the position to derive the optimality conditions of the minimization problem (1.4) which result in a zero-finding problem. We cite the following theorem from [23, Proposition 3.1], [58, Proposition 2.3], [93, Lemma 3.2 (2)] and [103, Lemma 2.6], respectively.

Proposition 1.16 *Let $g\colon \ell_2 \to \mathbb{R}$ be Fréchet differentiable. Assume that the minimization problem* (1.4) *has a minimizer $\mathbf{u}^*$. Then, $\mathbf{u}^*$ is a zero of the mapping $\mathbf{F}$,*

$$\mathbf{F}(\mathbf{u}^*) = \mathbf{0}, \tag{1.9}$$

where

$$\mathbf{F}\colon \ell_2 \to \ell_2, \quad \mathbf{F}(\mathbf{u}) := \mathbf{u} - \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u} - \gamma Dg(\mathbf{u})), \tag{1.10}$$

for any parameter $\gamma > 0$. Moreover, if g is convex, then $\mathbf{u}^*$ is a zero of $\mathbf{F}$ if and only if $\mathbf{u}^*$ is a minimizer of (1.4).

Proof. In the following, we use ideas from [58, proof of Proposition 2.3]. Assume that the minimization problem (1.4) has a minimizer $\mathbf{u}^*$. Defining

$$J\colon \ell_2 \to \mathbb{R} \cup \{\infty\}, \quad J(\mathbf{u}) := g(\mathbf{u}) + h(\mathbf{u}),$$

with h from (1.7), we have for all $\mathbf{u} \in \ell_2$,

$$J(\mathbf{u}^*) \leq J(\mathbf{u}).$$

The functional h is real-valued at $\mathbf{u}^*$, i.e. $\mathbf{u}^* \in \operatorname{dom} h$, because $\mathbf{u}^*$ is a minimizer of J and because h is proper. The functional J is directionally differentiable at $\mathbf{u}^*$ due to the Fréchet differentiability of g and the convexity of h, cf. [2, Proposition 17.2]. Because $\mathbf{u}^*$ minimizes J, the directional derivative $J'(\mathbf{u}^*, \mathbf{d})$ of J at $\mathbf{u}^*$ in the direction $\mathbf{d}$ is nonnegative for all $\mathbf{d} \in \ell_2$,

$$J'(\mathbf{u}^*, \mathbf{d}) = \lim_{t \searrow 0} \frac{J(\mathbf{u}^* + t\mathbf{d}) - J(\mathbf{u}^*)}{t} \geq 0. \tag{1.11}$$

With the Fréchet derivative Dg of g, we get for all $\mathbf{d} \in \ell_2$,

$$-Dg(\mathbf{u}^*)\mathbf{d} \leq -Dg(\mathbf{u}^*)\mathbf{d} + J'(\mathbf{u}^*, \mathbf{d}) = h'(\mathbf{u}^*, \mathbf{d}).$$

Defining $\ell_2 \ni \tilde{\mathbf{d}} := \mathbf{d} + \mathbf{u}^*$, it follows with the convexity of h,

$$\frac{h(\mathbf{u}^* + t(\tilde{\mathbf{d}} - \mathbf{u}^*)) - h(\mathbf{u}^*)}{t} \leq \frac{t\, h(\tilde{\mathbf{d}}) + (1-t)h(\mathbf{u}^*) - h(\mathbf{u}^*)}{t} = h(\tilde{\mathbf{d}}) - h(\mathbf{u}^*),$$

for every $t > 0$, see also [2, Proposition 17.2 (iii)]. Therefore, we get for all $\tilde{\mathbf{d}} \in \ell_2$,

$$-Dg(\mathbf{u}^*)(\tilde{\mathbf{d}} - \mathbf{u}^*) \leq h'(\mathbf{u}^*, \tilde{\mathbf{d}} - \mathbf{u}^*) \leq h(\tilde{\mathbf{d}}) - h(\mathbf{u}^*).$$

Hence, we have with Definition 1.9 (see also [2, proof of Proposition 17.17]),

$$-Dg(\mathbf{u}^*) \in \partial h(\mathbf{u}^*),$$

or equivalent

$$\mathbf{u}^* - \gamma Dg(\mathbf{u}^*) \in \{\mathbf{u}^*\} + \gamma \partial h(\mathbf{u}^*) = (I + \gamma \partial h)(\mathbf{u}^*),$$

for any $\gamma > 0$. Using Proposition 1.8, Proposition 1.15 and Remark 1.14, we conclude

$$\mathbf{u}^* = (I + \gamma \partial h)^{-1}(\mathbf{u}^* - \gamma Dg(\mathbf{u}^*)) = \mathbf{S}_{\gamma \mathbf{w}}(\mathbf{u}^* - \gamma Dg(\mathbf{u}^*)).$$

Let us now consider the case that g is convex. Here, the sum rule for subdifferentials [2, Corollary 16.38] and Fermat's rule (Proposition 1.11) yields that $\mathbf{u}^*$ is a minimizer of J if and only if

$$0 \in \partial J(\mathbf{u}^*) = \{Dg(\mathbf{u}^*)\} + \partial h(\mathbf{u}^*). \tag{1.12}$$

As shown above, (1.12) is equivalent to Equation (1.9) for any $\gamma > 0$, see also [2, Theorem 26.2]. □

Pang and Qi [109, Section 2.7] mentioned zero-finding problems involving resolvents like (1.8) as an example of nonsmooth systems of equations defined by Lipschitz continuous mappings. An alternative proof of Proposition 1.16 can be found in [99]. In loc. cit., the authors used the directional derivative of J to show in a finite-dimensional setting that $\mathbf{u}^*$ is a stationary point of the objective function J in the minimization problem (1.4) if and only if (1.9) holds.

If g is bounded from below by a constant c, the objective function in (1.4) with $w_k \geq w_0 > 0$ is coercive because of

$$g(\mathbf{u}) + \sum_{k=1}^{\infty} w_k |u_k| \geq c + w_0 \|\mathbf{u}\|_{\ell_1} \geq c + w_0 \|\mathbf{u}\|_{\ell_2}.$$

Hence, if g is continuous, strictly convex and bounded from below, then there exists a unique minimizer $\mathbf{u}^*$ of (1.4), cf. Remark 2.1, [2, Corollary 11.15] and [23, Proposition 3.1]. This is true, e.g., in the case of a quadratic discrepancy term $g(\mathbf{u}) := \frac{1}{2}\|K\mathbf{u} - f\|_H^2$, where H is a separable Hilbert space, $K \in L(\ell_2, H)$ is injective and $f \in H$, see [58].

Remark 1.17 *For a zero $\mathbf{u}^* \in \ell_2$ of $\mathbf{F}$ from* (1.10) *it holds $\mathbf{u}^* - \gamma Dg(\mathbf{u}^*) \in \ell_2$ and there exists an integer $k_0 > 0$ with $u_k^* = S_{\gamma w_k}(u_k^* - \gamma Dg(\mathbf{u}^*)_k) = (|u_k^* - \gamma Dg(\mathbf{u}^*)_k| - \gamma w_k)_+ \operatorname{sgn}(u_k^* - \gamma Dg(\mathbf{u}^*)_k) = 0$ for all $k \geq k_0$. Hence, a zero $\mathbf{u}^*$ of $\mathbf{F}$ from* (1.10) *is a finitely supported sequence, see [58, Corollary 2.4].*

The operator $\mathbf{F}$ from (1.10) is not Fréchet differentiable because of the soft-thresholding operator. Nevertheless, if the Fréchet derivative Dg of g is locally Lipschitz continuous, then the mapping $\mathbf{F}$ is locally Lipschitz continuous, motivating why we elaborate on generalized derivatives for locally Lipschitz continuous operators in the following section.

1.4 Generalized derivatives

In this section, we recall some concepts of generalized derivatives in finite-dimensional spaces, see, e.g., [20, 40, 72, 98, 107, 109, 111, 115, 121, 124], and in infinite-dimensional spaces, see, e.g., [18, 58, 103, 128, 136, 137]. Historical overviews on the subject can be found, e.g., in [18, 40, 41, 115].

The finite-dimensional case

For locally Lipschitz continuous mappings $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^m$, Rademacher's theorem states that $\mathbf{F}$ is differentiable almost everywhere [117]. Based on this result, Clarke [19, 20] defined the *generalized Jacobian* for locally Lipschitz continuous functions, see also [40, Definition 4.6.2], [41, Definition 7.1.1], or [18, 115].

Definition 1.18 (Clarke's generalized Jacobian) Let D be an open subset of $\mathbb{R}^m$ and let $\mathbf{F}\colon D \to \mathbb{R}^n$ be a mapping that is locally Lipschitz continuous at $\mathbf{u} \in D$. Then, the *B-subdifferential* is defined by

$$\partial_B \mathbf{F}(\mathbf{u}) := \left\{ \mathbf{G} \in \mathbb{R}^{n \times m} : \mathbf{G} = \lim_{j \to \infty} \nabla \mathbf{F}(\mathbf{u}^{(j)}), \mathbf{u}^{(j)} \to \mathbf{u}, \{\mathbf{u}^{(j)}\}_j \subseteq \mathcal{D}_{\mathbf{F}} \right\},$$

where $\mathcal{D}_{\mathbf{F}}$ denotes the set of points where $\mathbf{F}$ is Fréchet differentiable and $\nabla\mathbf{F}(\mathbf{u}^{(j)})$ denotes the Jacobian of $\mathbf{F}$ at $\mathbf{u}^{(j)} \in \mathcal{D}_{\mathbf{F}}$. The *generalized Jacobian* is defined as the convex hull of the set $\partial_B\mathbf{F}(\mathbf{u})$,

$$\partial\mathbf{F}(\mathbf{u}) := \operatorname{conv}\{\partial_B\mathbf{F}(\mathbf{u})\}.$$

The generalized Jacobian is related to directional derivatives. We cite the following lemma from [115, Lemma 2.2] and [124, Theorem 3.1.2], respectively.

Lemma 1.19 *Let $\mathbf{F}\colon \mathbb{R}^m \to \mathbb{R}^n$ be locally Lipschitz continuous and directionally differentiable at $\mathbf{u} \in \mathbb{R}^m$ with Lipschitz constant $L_{\mathbf{F}}$ in a neighborhood of $\mathbf{u}$. Then, we have*

(i) $\mathbf{F}'(\mathbf{u}, \cdot)$ is globally Lipschitz continuous with Lipschitz constant $L_{\mathbf{F}}$,

(ii) for any $\mathbf{h} \in \mathbb{R}^m$, there exists a $\mathbf{G} \in \partial\mathbf{F}(\mathbf{u})$ with $\mathbf{F}'(\mathbf{u}, \mathbf{h}) = \mathbf{G}\mathbf{h}$.

Mifflin [98] introduced the concept of *semismoothness* for functionals and Qi and Sun [115] for mappings $\mathbf{F}\colon \mathbb{R}^m \to \mathbb{R}^n$. We cite the following definition from [115, p. 355].

Definition 1.20 A mapping $\mathbf{F}\colon \mathbb{R}^m \to \mathbb{R}^n$ is called *semismooth* at $\mathbf{u} \in \mathbb{R}^m$ if $\mathbf{F}$ is locally Lipschitz continuous at $\mathbf{u}$ and the limit

$$\lim_{\substack{\mathbf{V}\in\partial\mathbf{F}(\mathbf{u}+t\mathbf{h}'),\\ t\searrow 0,\ \mathbf{h}'\to\mathbf{h}}} \{\mathbf{V}\mathbf{h}'\}$$

exists for any $\mathbf{h} \in \mathbb{R}^m$.

Note that it was pointed out in [115] that if $\mathbf{F}$ is semismooth, then $\mathbf{F}$ is directionally differentiable in the sense of Hadamard,

$$\lim_{t\searrow 0,\ \mathbf{h}'\to\mathbf{h}} \frac{\mathbf{F}(\mathbf{u}+t\mathbf{h}') - \mathbf{F}(\mathbf{u})}{t} = \lim_{\substack{\mathbf{V}\in\partial\mathbf{F}(\mathbf{u}+t\mathbf{h}'),\\ t\searrow 0,\ \mathbf{h}'\to\mathbf{h}}} \{\mathbf{V}\mathbf{h}'\}.$$

The infinite-dimensional case

In the following, we introduce generalized derivative concepts for mappings between infinite-dimensional spaces. The above definitions, based on Rademacher's theorem, are not applicable in this case because Rademacher's theorem does not hold in infinite-dimensional spaces.

Originally, Robinson [121, Appendix] defined the *Bouligand derivative* for locally Lipschitz continuous mappings $\mathbf{F}\colon D \subset \mathbb{R}^m \to \mathbb{R}^n$ by a cone property, and proved an approximation property similar to the approximation property (1.1) of the Fréchet derivative. Pang [107] introduced Bouligand differentiability for finite-dimensional mappings. Here, we cite the following definition from [18] which is based on an approximation property. Similar definitions can, e.g., be found in [72, 107, 111, 115, 124, 128].

Definition 1.21 (Bouligand derivative) A mapping $F\colon X \to Y$ is called *Bouligand differentiable* at $u \in X$ if it is directionally differentiable at u and the directional derivative fulfills the approximation property

$$\lim_{h\to 0} \frac{F(u+h) - F(u) - F'(u,h)}{\|h\|_X} = 0.$$

F is called Bouligand differentiable in an open subset $D \subseteq X$ if F is Bouligand differentiable at every $u \in D$.

Shapiro [128] showed that for finite-dimensional, locally Lipschitz continuous mappings $\mathbf{F}\colon \mathbb{R}^m \to \mathbb{R}^n$, Bouligand differentiability and directional differentiability are equivalent and both derivatives coincide.

For mappings between infinite-dimensional function spaces, we consider the so-called *Newton derivative* which is in the literature also known as *generalized derivative* or *slanting function* [18,73,136,137]. We cite the following definition from [18, Definition 2.2] and [73, Definition 1].

Definition 1.22 (Newton derivative/slanting function) Let D be an open subset of X. A mapping $F\colon D \subset X \to Y$ is called *Newton differentiable* or *slanty differentiable* in D if there exists a family of mappings $G\colon D \to L(X,Y)$ which fulfill the approximation property

$$\lim_{h\to 0} \frac{\|F(u+h) - F(u) - G(u+h)h\|_Y}{\|h\|_X} = 0, \tag{1.13}$$

for every $u \in D$. Each mapping G is called a *slanting function* or *Newton derivative* of F in D.

Note that in contrast to the Fréchet derivative, a Newton derivative G is not unique and G is evaluated at $u+h$ instead of u in the limit property (1.13). Chen, Nashed and Qi [18, Definition 2.3] additionally introduced the *slant derivative* which is stated in the following definition.

Definition 1.23 (slant derivative) Let G be a Newton derivative of $F\colon D \subset X \to Y$ at $u \in D$. The *slant derivative* $\partial_S^G F(u)$ w.r.t. the Newton derivative G is defined as

$$\partial_S^G F(u) := \{\tilde{G} \in L(X,Y) : \tilde{G} = \lim_{j\to\infty} G(u^{(j)}), u^{(j)} \to u, \{u^{(j)}\} \subseteq D\}.$$

Chen, Nashed and Qi proved that a mapping $F\colon X \to Y$ is Newton differentiable at $u \in X$ if and only if F is Lipschitz continuous at u [18, Theorem 2.6], hence for locally Lipschitz continuous mappings, concepts of generalized derivatives are available. In Section 1.5, we consider generalized Newton methods based on these generalized derivatives.

As a generalization of the semismoothness property in the finite-dimensional case, Chen, Nashed an Qi [18] introduced the concept of *semismoothness* for mappings between infinite-dimensional spaces. We cite the following definition from [18, Definition 3.2].

Definition 1.24 A mapping $F\colon X \to Y$ is called *semismooth* at $u \in X$ if there exists a Newton derivative G of F in a neighborhood of u, such that it holds for G and for the slant derivative $\partial_S^G F$:

(i) the limit $\lim_{t\searrow 0} G(u+th)h$ exists for every $h \in X$ and fulfills

$$\lim_{h\to 0} \frac{\lim_{t\searrow 0} G(u+th)h - G(u+h)h}{\|h\|_X} = 0,$$

(ii) for all $V \in \partial_S^G F(u+h)$, it holds

$$G(u+h)h - Vh = o(\|h\|_X), \quad \|h\|_X \to 0.$$

Finally, we cite the following useful theorem from [18, Theorem 3.3].

Theorem 1.25 *Let $F\colon X \to Y$ be Newton differentiable in a neighborhood of u with a Newton derivative G. Then, the mapping F is semismooth at u if and only if F is Bouligand differentiable at u and if we have $F'(u,h) - Vh = o(\|h\|_X)$, $\|h\|_X \to 0$ for all $V \in \partial_S^G(u+h)$.*

1.5 Generalized Newton methods

In the following, we consider mappings $F\colon X \to Y$. The classical Newton method [31,32] is an efficient approach to solve smooth, nonlinear zero-finding problems

$$F(u) = 0,$$

where F is continuously Fréchet differentiable. Here, one chooses an initial guess $u^{(0)} \in X$ and computes a sequence of iterates $\{u^{(j)}\}_j$ approaching a zero u^* of F once the initial guess was close enough to u^*. Linearizing the nonlinearity F around the current iterate $u^{(j)}$,

$$0 = F(u^*) \approx F(u^{(j)}) + DF(u^{(j)})(u^* - u^{(j)}),$$

yields, if the inverse $DF(u^{(j)})^{-1}$ exists,

$$u^* \approx u^{(j)} - DF(u^{(j)})^{-1} F(u^{(j)}). \tag{1.14}$$

The classical (local) Newton method is therefore iteratively defined by

$$u^{(j+1)} := u^{(j)} + d^{(j)}, \quad DF(u^{(j)})d^{(j)} = -F(u^{(j)}), \quad j = 0, 1, \dots \tag{1.15}$$

assuming implicitly that the derivatives $DF(u^{(j)})$ are nonsingular in each iteration. For the computation of the Newton direction $d^{(j)}$ one has to solve only one system of linear equations. Locally quadratic convergence can be shown under the assumption that F is Lipschitz continuously differentiable, that the Fréchet derivatives $DF(u^{(j)})$ are invertible in a neighborhood of u^* and that the inverses $DF(u^{(j)})^{-1}$ are uniformly bounded in that neighborhood, see, e.g., [31, 32, 145]. Nevertheless, it is well-known that the Newton method is, in general, not globally convergent. For an extensive survey of Newton methods including proofs of the above stated well-known properties as well as for an historic overview of the development of classical Newton methods, we refer to the textbooks [31, 32].

Variants of the classical Newton methods are, e.g., the quasi-Newton or the inexact Newton method, see [29, 30, 36]. In the quasi-Newton method, the Fréchet derivative $DF(u^{(j)})$ is in the Newton equation (1.15) replaced by an approximation $D^{(j)} \approx DF(u^{(j)})$. Moreover, in inexact Newton methods, the Newton equation (1.15) is solved only inexactly, see also [31, 32].

For nonlinearities that are not Fréchet differentiable, generalized Newton methods exist which are based on the generalized derivatives stated above. An overview of existing

Table 1.1: Overview of generalized Newton methods for $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$

generalized Newton method	generalized Newton equation		descent w.r.t. Θ
semismooth Newton method	$\mathbf{G}(\mathbf{u})\mathbf{d} = -\mathbf{F}(\mathbf{u})$	linear	usually not
B(ouligand)-Newton method	$\mathbf{F}'(\mathbf{u}, \mathbf{d}) = -\mathbf{F}(\mathbf{u})$	usually nonlinear	✓
modified B-Newton method	$\tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d}) = -\mathbf{F}(\mathbf{u})$	usually nonlinear	✓

Newton-type methods for the numerical solution of nonsmooth zero-finding problems can, e.g., be found in [41, 109]. In the following, we give a short overview of the generalized Newton methods that are considered in Chapters 2 and 3. We are concerned with semismooth Newton methods [18, 73, 115, 136, 137], B(ouligand)-Newton methods [70,72,107] as well as modified B-Newton methods [63,77,108,111]. An overview of these generalized Newton methods is given in Table 1.1.

The semismooth Newton method

A semismooth Newton method [18, 73, 136, 137] for the solution of an operator equation $F(u) = 0$, $F\colon X \to Y$, can be defined by replacing the Fréchet derivative DF in the classical Newton method (1.15) by a slanting function G of F,

$$u^{(j+1)} := u^{(j)} + d^{(j)}, \qquad G(u^{(j)})d^{(j)} = -F(u^{(j)}), \qquad j = 0, 1, \dots \tag{1.16}$$

For finite-dimensional mappings $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$, one chooses an element $\mathbf{G}^{(j)}$ of the generalized Jacobian of $\mathbf{F}$ at $\mathbf{u}^{(j)}$ in each iteration j,

$$\begin{aligned} \mathbf{u}^{(j+1)} &:= \mathbf{u}^{(j)} + \mathbf{d}^{(j)}, \\ \mathbf{G}^{(j)}\mathbf{d}^{(j)} &= -\mathbf{F}(\mathbf{u}^{(j)}), \qquad \mathbf{G}^{(j)} \in \partial\mathbf{F}(\mathbf{u}^{(j)}), \qquad j = 0, 1, \dots, \end{aligned}$$

see [115]. We cite the following local convergence theorem for the semismooth Newton iteration (1.16) from [18, Theorem 3.4, Theorem 3.5], [73, Theorem 1.1]. Similar results hold for finite-dimensional mappings [111, 115].

Theorem 1.26 *Let $F\colon X \to Y$ be slanty differentiable in a neighborhood U of a zero u^* of F and let G be a slanting function in U. If $G(u)$ is invertible for all $u \in U$ and if the inverses are uniformly bounded in U, i.e., there exists a constant $C > 0$ with $\|G(u)^{-1}\|_{L(Y,X)} \leq C$, then the semismooth Newton iteration* (1.16) *is well-defined and is locally superlinearly convergent to u^*.*

The B(ouligand)-Newton method

For a zero-finding problem $F(u) = 0$, where $F\colon X \to Y$ is Bouligand differentiable, the B-Newton method [107] is defined by introducing the Bouligand derivative in the generalized Newton equation,

$$u^{(j+1)} := u^{(j)} + d^{(j)}, \qquad F'(u^{(j)}, d^{(j)}) = -F(u^{(j)}), \qquad j = 0, 1, \dots \tag{1.17}$$

On the one hand, in contrast to the classical Newton method (1.15) and the semismooth Newton method (1.16), the generalized Newton equation,

$$F'(u,d) = -F(u), \tag{1.18}$$

for the computation of the Newton direction d, if solvable, usually is a nonlinear system of equations because the Bouligand derivative is not linear in d. On the other hand, for mappings $\mathbf{F}\colon \ell_2 \to \ell_2$, the Newton direction $\mathbf{d}$, determined by the generalized Newton equation (1.18), is a descent direction with respect to the merit functional $\Theta\colon \ell_2 \to \mathbb{R}$, $\Theta(\mathbf{u}) := \|\mathbf{F}(\mathbf{u})\|_{\ell_2}^2$, see [107, Lemma 1]. Indeed, we have for $\mathbf{u} \in \ell_2$ with $\mathbf{F}(\mathbf{u}) \neq \mathbf{0}$,

$$\Theta(\mathbf{u}, \mathbf{d}) = -2\Theta(\mathbf{u}) < 0.$$

This property of the B-Newton method is used in Chapter 2 to apply a globalization strategy proposed in [107]. In the finite-dimensional case, according to Lemma 1.19, the B-Newton method is a semismooth Newton method with specially chosen generalized Jacobians in each step, cf. [115, Proposition 3.4].

The modified B-Newton method

Applying a line search damping strategy from [107] to the B-Newton method (1.17) yields a globalized B-Newton method for finite-dimensional zero-finding problems. In loc. cit., the globalization is based on the descent property of the Newton directions w.r.t. the merit functional $\Theta\colon \mathbb{R}^n \to \mathbb{R}$. The proof of global convergence hinges on a technical assumption on an a priori unknown accumulation point of the sequence of iterates [107].

To overcome this theoretical drawback, a modified B-Newton method was introduced by Han, Pang and Rangaraj [63] as a generalization of the previous work [108] of Pang. There, one assumes that the mapping $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz continuous and that a mapping $\tilde{\mathbf{G}}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is given such that $\tilde{\mathbf{G}}(\mathbf{u}, \cdot)$ is surjective for each $\mathbf{u} \in \mathbb{R}^n$ and such that for all $\mathbf{u}, \mathbf{d} \in \mathbb{R}^n$, it holds

$$2\langle \mathbf{F}(\mathbf{u}), \tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d})\rangle \geq \Theta'(\mathbf{u}, \mathbf{d}). \tag{1.19}$$

The modified B-Newton method [63] is then defined as

$$\mathbf{u}^{(j+1)} := \mathbf{u}^{(j)} + \mathbf{d}^{(j)}, \qquad \tilde{\mathbf{G}}(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = -\mathbf{F}(\mathbf{u}^{(j)}), \qquad j = 0, 1, \ldots \tag{1.20}$$

By (1.19) and the generalized Newton equation $\tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d}) = -\mathbf{F}(\mathbf{u})$, it follows $\Theta'(\mathbf{u}, \mathbf{d}) \leq -2\Theta(\mathbf{u})$. Hence the descent property w.r.t. the merit functional Θ is retained, see [63,108]. Nevertheless, the generalized Newton equation $\tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d}) = -\mathbf{F}(\mathbf{u})$ is nonlinear in general.

Ito and Kunisch [77, Definition 4.1] introduced a mapping similar to $\tilde{\mathbf{G}}$ from [63]. There, the authors called this mapping *quasi-directional derivative* and the resulting Newton direction $\mathbf{d}$ *generalized Bouligand direction*, respectively. In [77, 111], hybrid methods combining the B-Newton method (1.17) and the modified B-Newton method (1.20) were proposed.

Chapter 2

A B-semismooth Newton method

In this chapter, we introduce a globalized B-semismooth Newton method for the efficient numerical solution of the zero-finding problem

$$\mathbf{F}(\mathbf{u}) = \mathbf{0}$$

with $\mathbf{F}$ from (1.10). A short review of a local semismooth Newton method introduced by Herzog and Lorenz [58] for a quadratic discrepancy term g and by Muoi et al. [103] for general functionals g is given in Section 2.1. The algorithm and the feasibility of a local B-semismooth Newton method is treated in Section 2.2. In Section 2.3, we are concerned with the globalization of the local B-semismooth Newton method from Section 2.2.

This chapter is a part of our publication [66], the theory in its present form being reformulated for general discrepancy terms g. Moreover, Lemma 2.5, Corollary 2.6, Lemma 2.7, Remark 2.12, Theorem 2.15 and Remark 2.28 are new and Remark 2.14 is corrected and reformulated in an infinite-dimensional setting. Furthermore, a proof of Corollary 2.16 is added and the proof of Proposition 2.23 is shortened. Additional minor changes were made.

In this chapter, we denote $\ell_2 = \ell_2(\mathcal{N})$, where $\mathcal{N}$ is either $\mathbb{N}$ or $\{1, \dots, n\}$. The local generalized Newton methods from Section 2.1 and Section 2.2 are formulated in the infinite-dimensional setting (1.4), i.e. we have $\mathcal{N} = \mathbb{N}$. Here, we require the following assumptions, similar to [103, Assumption 3.1].

Assumption A *(A1) The zero-finding problem* (1.9) *has a solution* $\mathbf{u}^*$.

(A2) The functional $g\colon \ell_2 \to \mathbb{R}$ *is twice Fréchet differentiable and the Fréchet derivatives* Dg *and* D^2g *are locally Lipschitz continuous.*

(A3) The second Fréchet derivative is positive for all $\mathbf{u} \in \ell_2$ *in a neighborhood of the zero* $\mathbf{u}^*$ *of* $\mathbf{F}$*, i.e., we have* $D^2g(\mathbf{u})(\mathbf{h}, \mathbf{h}) > 0$ *for all* $\mathbf{h} \neq \mathbf{0}$, $\mathbf{h} \in \ell_2$.

(A4) For each fixed, finite index set $\mathcal{J}$*, there exists a neighborhood of the zero* $\mathbf{u}^*$ *of* $\mathbf{F}$ *in which the inverses* $(D^2g(\mathbf{u})_{\mathcal{J},\mathcal{J}})^{-1}$ *are uniformly bounded.*

Remark 2.1 *If Assumption (A3) is fulfilled in the whole sequence space* ℓ_2*, then Assumption (A3) implies strict convexity of* g *ensuring, with Assumption (A1), the unique solvability of the minimization problem* (1.4)*, see, e.g., [2, Corollary 11.8, Proposition 17.13]. A sufficient condition for Assumption (A1) is convexity, continuity and boundedness from below of* g*, see, e.g., [149, Theorem 25.E] or [2, Corollary 11.15], compare Section 1.3. The inverse of the finite submatrix* $D^2g(\mathbf{u})_{\mathcal{J},\mathcal{J}}$ *exits because of the symmetry of* D^2g *and (A3). If the solvability of the*

finite-dimensional linear complementarity problem (2.32)*, see Section 2.2.2, is guaranteed, Assumption (A3) may be relaxed. In particular, similar to (A4), it suffices to assume in (A3) that for each fixed, finite index set* $\mathcal{J}$*, the submatrices* $D^2g(\mathbf{u})_{\mathcal{J},\mathcal{J}}$ *are positive definite to ensure the solvability of the linear complementarity problem* (2.32)*. The local Lipschitz continuity of* Dg *is necessary to ensure the local Lipschitz continuity of* $\mathbf{F}$*. In particular, the local Lipschitz continuity of* Dg *is essential for the Newton differentiability and Bouligand differentiability of* $\mathbf{F}$*. The local Lipschitz continuity of* D^2g *is only needed to use a chain rule in the proof of Theorem 2.3 from [103, Theorem 2.9], to show the limit property* (2.20) *in the proof of Lemma 2.7 and in Theorem 2.15 in order to show the semismoothness of the nonlinearity* $\mathbf{F}$*. Furthermore, Assumption (A4) and the local Lipschitz continuity of* Dg *is required to ensure locally superlinear convergence of the local B-semismooth Newton method in Corollary 2.16, see also [103, Remark 3.2].*

In case of a quadratic discrepancy term $g(\mathbf{u}) = \frac{1}{2}\|K\mathbf{u} - \mathbf{f}\|_{\ell_2}^2$ *with* K *linear, bounded and injective, it is straightforward to show (A2) and (A3). Assumption (A4) was verified in [58, Section 3.4] for an injective operator* K*. As pointed out in [58, Remark 3.15] and [103, Example 3.3], it is sufficient to assume that* $K_{\mathcal{J},\mathcal{J}}$ *is injective for any finite index set* $\mathcal{J}$ *instead of the injectivity of the whole operator* K*, i.e., to assume the so-called* finite basis injectivity (FBI) *property, see [11].*

The globalized B-semismooth Newton method from Section 2.3 is considered in the finite-dimensional setting

$$\min_{\mathbf{u}\in\mathbb{R}^n} g(\mathbf{u}) + \sum_{k=1}^{n} w_k|u_k|, \tag{2.1}$$

i.e. we set $\mathcal{N} = \{1,\dots,n\}$. The globalization strategy is based on suitable descent with respect to the merit functional $\Theta_p\colon \mathbb{R}^n \to \mathbb{R}$,

$$\Theta_p(\mathbf{u}) := \big\|\mathbf{F}(\mathbf{u})\big\|_p^p = \sum_{k=1}^{n} \big|F_k(\mathbf{u})\big|^p, \tag{2.2}$$

with $p \in \{1,2\}$, cf. [66]. Here, we require the following assumptions similar to [103, Assumption 3.1, Example 3.4], cf. [67].

Assumption B *(B1) The minimization problem* (2.1) *has a solution.*

(B2) The level set $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)}) := \{\mathbf{u} \in \mathbb{R}^n : \Theta_p(\mathbf{u}) \le \Theta_p(\mathbf{u}^{(0)})\}$ *of* Θ_p *is compact.*

(B3) The function g *is twice Lipschitz continuously differentiable and the Hessian* $\nabla^2 g(\mathbf{u})$ *is positive definite for all* $\mathbf{u} \in \mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$*. Moreover, there exist constants* $0 < c_1, c_2 < \infty$ *with*

$$c_1\|\mathbf{h}\|_2^2 \le \langle \nabla^2 g(\mathbf{u})\mathbf{h}, \mathbf{h}\rangle \le c_2\|\mathbf{h}\|_2^2, \qquad \text{for all } \mathbf{h} \in \mathbb{R}^n, \tag{2.3}$$

uniformly for all $\mathbf{u} \in \mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$.

Remark 2.2 *According to Assumptions (B1) and (B3), the minimization problem* (2.1) *is uniquely solvable. The assumption of the positive definiteness of* $\nabla^2 g$ *ensures the solvability of the finite-dimensional linear complementarity problem* (2.32)*, see Section 2.2.2. In the finite-dimensional setting* (2.1)*, Assumption (B3) corresponds to (A2)–(A4) from the infinite-dimensional case. The compactness of the level set* $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$ *is necessary for the proof of global convergence, cf. Theorem 2.25.*

In Section 2.1 and Section 2.2, we consider the infinite-dimensional setting (1.4) and assume Assumption A to hold.

2.1 Review of a semismooth Newton method

In this section, we are going to review the semismooth Newton method from [58, 103], applied to the numerical solution to the shrinkage equation (1.10). In the following, $\mathcal{N}$ denotes either $\mathbb{N}$ or the finite set $\{1, \dots, n\}$.

2.1.1 The semismooth Newton method and its basic properties

One of the key observations of [58, 103] was that the nonlinearity $\mathbf{F}$ of the shrinkage equation is Newton differentiable on the full sequence space ℓ_2, cf. Definition 1.22. We cite the following theorem from [103, Theorem 2.9], see also [58, Proposition 3.7] for quadratic functionals g.

Theorem 2.3 *The mapping $\mathbf{F} : \ell_2 \to \ell_2$ from* (1.10) *is Newton differentiable in each $\mathbf{u} \in \ell_2$ with (nonunique) Newton derivative*

$$\begin{pmatrix} \mathbf{G}(\mathbf{u})_{\mathcal{A},\mathcal{A}} & \mathbf{G}(\mathbf{u})_{\mathcal{A},\mathcal{I}} \\ \mathbf{G}(\mathbf{u})_{\mathcal{I},\mathcal{A}} & \mathbf{G}(\mathbf{u})_{\mathcal{I},\mathcal{I}} \end{pmatrix} = \begin{pmatrix} \gamma D^2 g(\mathbf{u})_{\mathcal{A},\mathcal{A}} & \gamma D^2 g(\mathbf{u})_{\mathcal{A},\mathcal{I}} \\ \mathbf{0}_{\mathcal{I},\mathcal{A}} & \mathbf{I}_{\mathcal{I},\mathcal{I}} \end{pmatrix} \in L(\ell_2) \tag{2.4}$$

which is a block matrix with respect to the active set

$$\mathcal{A} = \mathcal{A}(\mathbf{u}) := \{k \in \mathcal{N} : |(\mathbf{u} - \gamma Dg(\mathbf{u}))_k| > \gamma w_k\} \tag{2.5}$$

and the inactive set $\mathcal{I} = \mathcal{I}(\mathbf{u}) := \mathcal{N} \setminus \mathcal{A}(\mathbf{u})$.

Note that by $w_k \geq w_0 > 0$ and $Dg(\mathbf{u}) \in \ell_2$, the active sets $\mathcal{A}(\mathbf{u})$ are finite for each $\mathbf{u} \in \ell_2$, even in the case $\mathcal{N} = \mathbb{N}$, cf. [58, Remark 3.8]. Therefore, $\mathbf{G}(\mathbf{u})$ is boundedly invertible for each $\mathbf{u} \in \ell_2$, cf. Assumption A. What is more, $\mathbf{G}(\mathbf{u})^{-1}$ is uniformly bounded in the operator norm in a sufficiently small neighborhood of the Tikhonov minimizer $\mathbf{u}^*$, see the following lemma from [58, Proposition 3.11] and [103, Proof of Theorem 3.7].

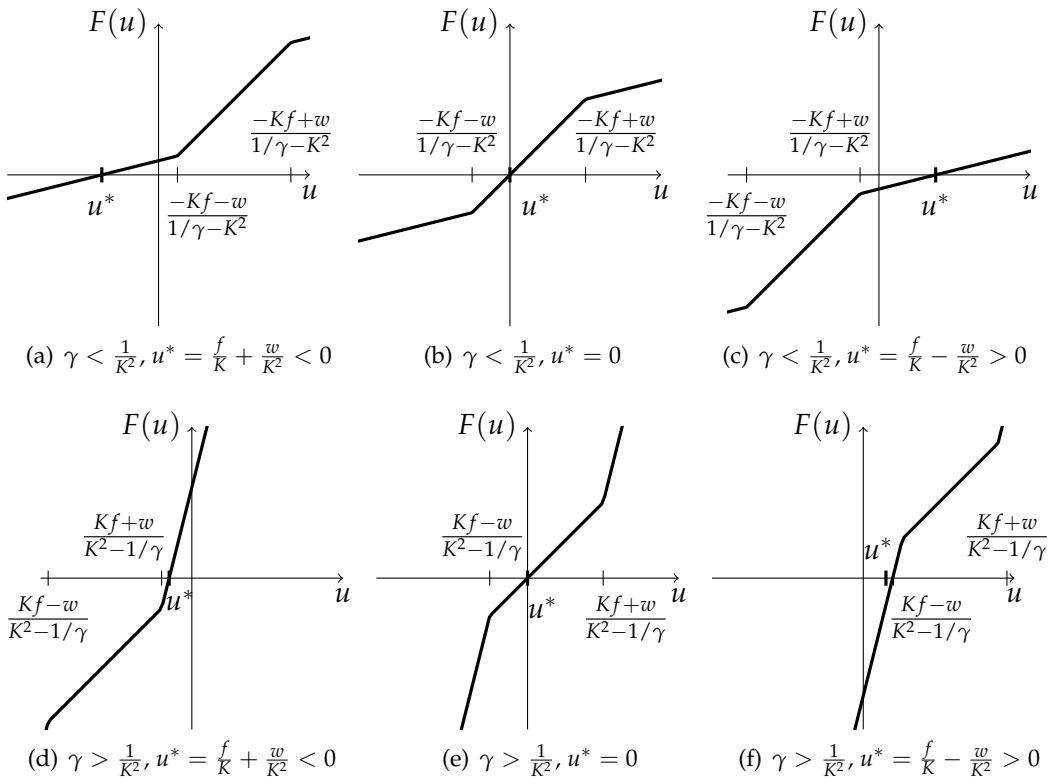
Lemma 2.4 *There exist $\rho, C > 0$ such that $\|\mathbf{G}(\mathbf{u})^{-1}\|_{L(\ell_2)} \leq C$ for all $\mathbf{u} \in \ell_2$ with $\|\mathbf{u} - \mathbf{u}^*\|_{\ell_2} \leq \rho$.*

The iteration (1.16) is well-defined by the bounded invertibility of $\mathbf{G}$ from (2.4) on ℓ_2. Moreover, by using the local uniform boundedness of $\mathbf{G}(\mathbf{u})^{-1}$ and by applying the generic results from [17, 18, 73], it was proved in [58, 103] that (1.16) is locally superlinearly convergent. Moreover, as pointed out in [58], in case of a quadratic discrepancy term g the proof of Theorem 2.3 reveals that the numerator of (1.13) for $\mathbf{F}$ from (1.10), i.e., $\mathbf{F}(\mathbf{u} + \mathbf{h}) - \mathbf{F}(\mathbf{u}) - \mathbf{G}(\mathbf{u} + \mathbf{h})\mathbf{h}$, already vanishes for sufficiently small $\|\mathbf{h}\|_{\ell_2}$, $\mathbf{u} \in \ell_2$ being fixed. Therefore, if the current iterate $\mathbf{u}^{(j)}$ is sufficiently close to the exact minimizer $\mathbf{u}^*$, we have

$$\mathbf{F}(\mathbf{u}^{(j)}) - \mathbf{F}(\mathbf{u}^*) - \mathbf{G}(\mathbf{u}^{(j)})(\mathbf{u}^{(j)} - \mathbf{u}^*) = \mathbf{F}(\mathbf{u}^{(j)}) - \mathbf{G}(\mathbf{u}^{(j)})(\mathbf{u}^{(j)} - \mathbf{u}^*) = \mathbf{0},$$

so that the next step of the semismooth Newton iteration (1.16) jumps into $\mathbf{u}^*$,

$$\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} - \mathbf{G}(\mathbf{u}^{(j)})^{-1}\mathbf{F}(\mathbf{u}^{(j)}) = \mathbf{u}^*.$$



(a) $\gamma < \frac{1}{K^2}$, $u^* = \frac{f}{K} + \frac{w}{K^2} < 0$ (b) $\gamma < \frac{1}{K^2}$, $u^* = 0$ (c) $\gamma < \frac{1}{K^2}$, $u^* = \frac{f}{K} - \frac{w}{K^2} > 0$

(d) $\gamma > \frac{1}{K^2}$, $u^* = \frac{f}{K} + \frac{w}{K^2} < 0$ (e) $\gamma > \frac{1}{K^2}$, $u^* = 0$ (f) $\gamma > \frac{1}{K^2}$, $u^* = \frac{f}{K} - \frac{w}{K^2} > 0$

Figure 2.1: Qualitative behavior of F and the location of u^* in the one-dimensional case.

2.1.2 Cycling effect

One of the well-known pitfalls of the classical Newton method is the *cycling effect*. Here, after a certain number of iterations, the iterates start to oscillate between two or more distinct points. We will show now that the semismooth Newton iteration (1.16) may run into the same situation, at least if the parameter γ is not chosen large enough.

One-dimensional case

Assume that $\mathcal{N} = \{1\}$, $f \in \mathbb{R}$, $K > 0$ and $w > 0$. The unique minimizer of the Tikhonov functional with quadratic discrepancy term $g(u) := \frac{1}{2}(Ku - f)^2$,

$$J(u) = g(u) + w|u|, \qquad u \in \mathbb{R} \tag{2.6}$$

is given by

$$u^* = S_{\frac{w}{K^2}}(\tfrac{f}{K}) = \begin{cases} \frac{f}{K} - \frac{w}{K^2}, & \frac{f}{K} > \frac{w}{K^2}, \\ 0, & |\frac{f}{K}| \le \frac{w}{K^2}, \\ \frac{f}{K} + \frac{w}{K^2}, & \frac{f}{K} < -\frac{w}{K^2}. \end{cases} \tag{2.7}$$

Let us analyze the convergence behavior of the semismooth Newton iteration (1.16), (2.4) in detail. Using the derivative $g'(u) = -K(f - Ku)$, the nonlinearity F from the shrinkage

Table 2.1: Behavior of the semismooth Newton iteration in the one-dimensional case.

	location of $u^{(j)}$		
u^*	$u^{(j)} - \gamma g'(u^{(j)}) > \gamma w$	$\lvert u^{(j)} - \gamma g'(u^{(j)})\rvert \le \gamma w$	$u^{(j)} - \gamma g'(u^{(j)}) < -\gamma w$
$\frac{f}{K} + \frac{w}{K^2} < 0$	$u^{(j+1)} = \frac{f}{K} - \frac{w}{K^2}$ $u^{(j+2)} = \begin{cases} u^*, & \frac{f}{K} - \frac{w}{K^2} < -2\gamma w \\ 0, & \frac{f}{K} - \frac{w}{K^2} \ge -2\gamma w \end{cases}$ $u^{(j+3)} = u^*$	$u^{(j+1)} = 0$ $u^{(j+2)} = u^*$	$u^{(j+1)} = u^*$
0	$u^{(j+1)} = \frac{f}{K} - \frac{w}{K^2}$ $u^{(j+2)} = \begin{cases} \frac{f}{K} + \frac{w}{K^2}, & \frac{f}{K} - \frac{w}{K^2} < -2\gamma w \\ u^*, & \frac{f}{K} - \frac{w}{K^2} \ge -2\gamma w \end{cases}$	$u^{(j+1)} = u^*$	$u^{(j+1)} = \frac{f}{K} + \frac{w}{K^2}$ $u^{(j+2)} = \begin{cases} u^*, & \frac{f}{K} + \frac{w}{K^2} \le 2\gamma w \\ \frac{f}{K} - \frac{w}{K^2}, & \frac{f}{K} + \frac{w}{K^2} > 2\gamma w \end{cases}$
$\frac{f}{K} - \frac{w}{K^2} > 0$	$u^{(j+1)} = u^*$	$u^{(j+1)} = 0$ $u^{(j+2)} = u^*$	$u^{(j+1)} = \frac{f}{K} + \frac{w}{K^2}$ $u^{(j+2)} = \begin{cases} 0, & \frac{f}{K} + \frac{w}{K^2} \le 2\gamma w \\ u^*, & \frac{f}{K} + \frac{w}{K^2} > 2\gamma w \end{cases}$ $u^{(j+3)} = u^*$

equation (1.10) has the explicit form

$$F(u) = \begin{cases} \gamma g'(u) + \gamma w, & u - \gamma g'(u) > \gamma w, \\ u, & \big|u - \gamma g'(u)\big| \le \gamma w, \\ \gamma g'(u) - \gamma w, & u - \gamma g'(u) < -\gamma w. \end{cases} \tag{2.8}$$

The shape of the graph of F depends on the size of the parameter $\gamma > 0$. In the outer region $|u - \gamma g'(u)| = |(1 - \gamma K^2)u + \gamma K f| > \gamma w$, the slope of F is given by γK^2. For small parameters $0 < \gamma < \frac{1}{K^2}$, the graph of F is hence changing from convex to concave as u is growing, and vice versa for large parameters $\gamma > \frac{1}{K^2}$. We refer to Figure 2.1 for the six possible cases concerning the location of the zero u^* of F.

The Newton derivative (2.4) of F at $u \in \mathbb{R}$ is

$$G(u) = \begin{cases} \gamma K^2, & \big|u - \gamma g'(u)\big| > \gamma w, \\ 1, & \big|u - \gamma g'(u)\big| \le \gamma w, \end{cases} \tag{2.9}$$

so that one semismooth Newton step maps the current iterate $u^{(j)} \in \mathbb{R}$ to

$$u^{(j+1)} = u^{(j)} - \frac{F(u^{(j)})}{G(u^{(j)})} = \begin{cases} \frac{f}{K} - \frac{w}{K^2}, & u^{(j)} - \gamma g'(u^{(j)}) > \gamma w, \\ 0, & \big|u^{(j)} - \gamma g'(u^{(j)})\big| \le \gamma w, \\ \frac{f}{K} + \frac{w}{K^2}, & u^{(j)} - \gamma g'(u^{(j)}) < -\gamma w. \end{cases} \tag{2.10}$$

By taking all possible configurations into account, the complete behavior of the semismooth Newton iteration (1.16), (2.4) in the one-dimensional case can be summarized as in Table 2.1, compare also Figure 2.1.

If $u^* \neq 0$ and for any $\gamma > 0$, the algorithm obviously terminates after at most three iterations with the exact minimizer u^*. In case that $u^* = 0$, however, the behavior of the semismooth Newton iteration strongly depends on the size of γ.

On the one hand, if $u^* = 0$ and $\gamma > \frac{1}{K^2}$, then $u^{(j+2)} = 0 = u^*$ regardless of the choice of $u^{(j)}$, because one of the cases for $u^{(j+2)}$ in the second row of Table 2.1 never

occurs. As an example, if $u^{(j)} - \gamma g'(u^{(j)}) > \gamma w$ and hence $u^{(j+1)} = \frac{f}{K} - \frac{w}{K^2}$, it follows that $u^{(j+1)} - \gamma g'(u^{(j+1)}) = \frac{f}{K} - \frac{w}{K^2} + \gamma w$. $u^* = 0$ and hence $|\frac{f}{K}| \leq \frac{w}{K^2}$ imply that $u^{(j+1)} - \gamma g'(u^{(j+1)}) \leq \gamma w$. Finally, $\gamma > \frac{1}{K^2}$ yields $u^{(j+1)} - \gamma g'(u^{(j+1)}) \geq -\frac{2w}{K^2} + \gamma w > -\gamma w$ and hence $u^{(j+2)} = 0 = u^*$. The same argument works in the case $u^{(j)} - \gamma g'(u^{(j)}) < -\gamma w$. Summarizing, we can conclude that the semismooth Newton iteration (1.16), (2.4) for a quadratic discrepancy term g converges in at most three steps if γ is chosen larger than $\frac{1}{K^2}$.

On the other hand, if $u^* = 0$, i.e., $|\frac{f}{K}| \leq \frac{w}{K^2}$, and $0 < \gamma < \frac{1}{K^2}$ is sufficiently small, the semismooth Newton iteration might start to oscillate between $\frac{f}{K} \pm \frac{w}{K^2}$. We obtain a cycle, e.g., if $|\frac{f}{K}| < \frac{w}{K^2}$, $0 < \gamma < \frac{1}{2}(\frac{1}{K^2} - |\frac{f}{wK}|)$ and $u^{(j)} = \frac{f}{K} + \frac{w}{K^2} \neq 0 = u^*$. Here $u^{(j)} - \gamma g'(u^{(j)}) = \frac{f}{K} + \frac{w}{K^2} - \gamma w \geq \frac{w}{K^2} - |\frac{f}{K}| - \gamma w > \gamma w$ and hence $u^{(j+1)} = \frac{f}{K} - \frac{w}{K^2}$. In the next step, $u^{(j+1)} - \gamma g'(u^{(j+1)}) = \frac{f}{K} - \frac{w}{K^2} + \gamma w \leq |\frac{f}{K}| - \frac{w}{K^2} + \gamma w < -\gamma w$ and hence $u^{(j+2)} = \frac{f}{K} + \frac{w}{K^2} = u^{(j)}$.

Finite-dimensional case

The cycling effect can be observed in higher dimensions as well. In the two-dimensional case, choosing

$$\mathbf{K} = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \gamma = \frac{3}{2}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} -6 \\ 12 \end{pmatrix},$$

the next iterates are

$$\mathbf{u}^{(1)} = \begin{pmatrix} 10 \\ -12 \end{pmatrix}, \quad \mathbf{u}^{(2)} = \begin{pmatrix} -6 \\ 12 \end{pmatrix} = \mathbf{u}^{(0)},$$

i.e., a cycle arises. Choosing $\gamma = 3$ instead, the zero $(0,0)^\top$ of $\mathbf{F}$ would be found within two steps.

In the four-dimensional case, the choice

$$\mathbf{K} = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \gamma = 2, \quad \mathbf{u}^{(0)} = \begin{pmatrix} 36 \\ -\frac{112}{3} \\ 0 \\ 0 \end{pmatrix}$$

leads to the iterates

$$\mathbf{u}^{(1)} = \begin{pmatrix} -28 \\ \frac{112}{3} \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{u}^{(2)} = \begin{pmatrix} 36 \\ -\frac{112}{3} \\ 0 \\ 0 \end{pmatrix} = \mathbf{u}^{(0)}.$$

These examples show that the semismooth Newton method (1.16), (2.4) is not globally convergent in general and may even lead to cycles, motivating the analysis of suitable globalization strategies.

2.2 The local B-semismooth Newton method

Instead of (1.16), we are going to study the B-semismooth Newton method (1.17) which is based on the Bouligand derivative $\mathbf{F}'$. The B-Newton directions $\mathbf{d}^{(j)}$ will be chosen by solving the generalized Newton equation (1.18), cf. Subsection 2.2.2.

2.2.1 Bouligand differentiability of the nonlinearity F

Let us first analyze the directional differentiability of the nonlinearity $\mathbf{F} = (F_k)_{k\in\mathcal{N}}$, i.e., the existence and concrete shape of $\mathbf{F}'(\mathbf{u},\mathbf{d}) = \lim_{h\searrow 0} \frac{\mathbf{F}(\mathbf{u}+h\mathbf{d})-\mathbf{F}(\mathbf{u})}{h}$, where $\mathbf{u},\mathbf{d} \in \ell_2$. To this end, we shall use the following straightforward componentwise representation of $\mathbf{F}$:

$$F_k(\mathbf{u}) = \begin{cases} \gamma\big(Dg(\mathbf{u})\big)_k + \gamma w_k, & k \in \mathcal{A}^+(\mathbf{u}), \\ \gamma\big(Dg(\mathbf{u})\big)_k - \gamma w_k, & k \in \mathcal{A}^-(\mathbf{u}), \\ u_k, & k \in \mathcal{I}(\mathbf{u}), \end{cases} \tag{2.11}$$

where we have split the active set into $\mathcal{A}(\mathbf{u}) = \mathcal{A}^+(\mathbf{u}) \cup \mathcal{A}^-(\mathbf{u})$ with

$$\mathcal{A}^+(\mathbf{u}) := \big\{k \in \mathcal{N} : (\mathbf{u} - \gamma Dg(\mathbf{u}))_k > \gamma w_k\big\}, \tag{2.12}$$
$$\mathcal{A}^-(\mathbf{u}) := \big\{k \in \mathcal{N} : (\mathbf{u} - \gamma Dg(\mathbf{u}))_k < -\gamma w_k\big\}. \tag{2.13}$$

Let us also split the inactive set into $\mathcal{I}(\mathbf{u}) = \mathcal{I}^+(\mathbf{u}) \cup \mathcal{I}^\circ(\mathbf{u}) \cup \mathcal{I}^-(\mathbf{u})$, where

$$\mathcal{I}^+(\mathbf{u}) := \big\{k \in \mathcal{N} : (\mathbf{u} - \gamma Dg(\mathbf{u}))_k = \gamma w_k\big\}, \tag{2.14}$$
$$\mathcal{I}^-(\mathbf{u}) := \big\{k \in \mathcal{N} : (\mathbf{u} - \gamma Dg(\mathbf{u}))_k = -\gamma w_k\big\}, \tag{2.15}$$
$$\mathcal{I}^\circ(\mathbf{u}) := \big\{k \in \mathcal{N} : |(\mathbf{u} - \gamma Dg(\mathbf{u}))_k| < \gamma w_k\big\}. \tag{2.16}$$

In the following, we drop the argument $\mathbf{u}$ of the active and inactive sets if there is no risk of confusion. With this notation at hand, we shall now prove the directional differentiability of $\mathbf{F}$. While the directional differentiability of a single component F_k and the corresponding formula for the directional derivative $F_k'(\mathbf{u},\mathbf{d}) = \lim_{h\searrow 0} \frac{F_k(\mathbf{u}+h\mathbf{d})-F_k(\mathbf{u})}{h}$ are almost obvious, the directional differentiability of the full operator $\mathbf{F}$ is nontrivial, at least in the case $\mathcal{N} = \mathbb{N}$. In the next lemma, we first prove the directional differentiability of the shrinkage operator.

Lemma 2.5 *Let $\boldsymbol{\alpha} = (\alpha_k)_k$ be a positive weight sequence with $\alpha_k \geq \alpha_0 > 0$. The soft-thresholding operator $\mathbf{S}_{\boldsymbol{\alpha}} \colon \ell_2 \to \ell_2$, $\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{v}) := (|\mathbf{v}| - \boldsymbol{\alpha})_+ \operatorname{sgn}(\mathbf{v})$ from Definition 1.12 is directionally differentiable and the directional derivative $\mathbf{S}'_{\boldsymbol{\alpha}}(\mathbf{v},\mathbf{d})$ at $\mathbf{u} \in \ell_2$ in the direction $\mathbf{d} \in \ell_2$ is given elementwise by*

$$S'_{\alpha_k}(\mathbf{v},\mathbf{d}) = \begin{cases} d_k, & k \in I_1(\mathbf{v},\mathbf{d}), \\ 0, & k \in I_2(\mathbf{v},\mathbf{d}), \end{cases} \tag{2.17}$$

where $I_1(\mathbf{v},\mathbf{d}) := \{k : |v_k| > \alpha_k\} \cup \{k : v_k = \alpha_k, d_k > 0\} \cup \{k : v_k = -\alpha_k, d_k < 0\}$ and $I_2(\mathbf{v},\mathbf{d}) := \{k : |v_k| < \alpha_k\} \cup \{k : v_k = \alpha_k, d_k \leq 0\} \cup \{k : v_k = -\alpha_k, d_k \geq 0\}$.

Proof. In this proof, we use arguments from the proof of [58, Proposition 3.3]. We consider the limit

$$\lim_{h \searrow 0} \frac{\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{v} + h\mathbf{d}) - \mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{v})}{h}.$$

We may suppose $h|d_k| < \alpha_0/2$ for all $k \in \mathbb{N}$. Because we have $\mathbf{v} \in \ell_2$, there exists an index $k_0 \in \mathbb{N}$ with $|v_k| < \alpha_0/2$ for all $k > k_0$, implying $|v_k + hd_k| \leq |v_k| + h|d_k| < \alpha_0 \leq \alpha_k$ for all $k > k_0$. Therefore, we have $S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = 0$ for all k sufficiently large. In the following, we can restrict ourselves to the indices $k \leq k_0$. We start considering the cases with $|v_k| \neq \alpha_k$. Here, we suppose

$$h|d_k| < \min\{\big||v_k| - \alpha_k\big| : k \leq k_0, |v_k| \neq \alpha_k\}.$$

Let us consider the case $|v_k| < \alpha_k$. We have $h|d_k| < \big||v_k| - \alpha_k\big| = \alpha_k - |v_k|$ and therefore it follows $|v_k + hd_k| \leq |v_k| + h|d_k| < \alpha_k$, implying

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = 0.$$

For indices k with $|v_k| > \alpha_k$, one has $h|d_k| < \big||v_k| - \alpha_k\big| = |v_k| - \alpha_k$. Thus, it follows $|v_k + hd_k| \geq |v_k| - h|d_k| > \alpha_k$. In the case $v_k > \alpha_k > 0$, we have $h|d_k| < v_k - \alpha_k$, implying $v_k + hd_k > \alpha_k$ independently of the sign of d_k. If $v_k < -\alpha_k < 0$, it follows $h|d_k| < -v_k - \alpha_k$, implying $v_k + hd_k < -\alpha_k$. We conclude

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = hd_k.$$

In the case $v_k = \alpha_k$, we have to distinguish $d_k < 0$, $d_k = 0$ and $d_k > 0$. If $d_k < 0$, we have $\alpha_k > v_k + hd_k = \alpha_k + hd_k > 0$ because $h|d_k| < \alpha_0/2 < \alpha_k$, implying $|v_k + hd_k| < \alpha_k$ and

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = 0.$$

In the case $d_k = 0$, we trivially have $S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = 0$. If $d_k > 0$, it follows $v_k + hd_k = \alpha_k + hd_k > \alpha_k$, implying

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = \alpha_k + hd_k - \alpha_k - 0 = hd_k.$$

Finally, we consider the case $v_k = -\alpha_k$. If we additionally have $d_k > 0$, it follows with $hd_k = h|d_k| < \alpha_0/2 < \alpha_k$ that $-\alpha_k < v_k + hd_k = -\alpha_k + hd_k < \alpha_k$, implying

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = 0.$$

The same equation holds true if $d_k = 0$. In the case $d_k < 0$, we have $v_k + hd_k = -\alpha_k + hd_k < -\alpha_k$ and therefore

$$S_{\alpha_k}(v_k + hd_k) - S_{\alpha_k}(v_k) = -\alpha_k + hd_k + \alpha_k - 0 = hd_k.$$

Altogether, we obtain

$$\begin{aligned}\|\mathbf{S}'_{\boldsymbol{\alpha}}(\mathbf{v}, \mathbf{d})\|_{\ell_2}^2 &= \sum_{k=1}^{\infty} \Big(\lim_{h \searrow 0} \frac{S_{\alpha_k}(\mathbf{v} + h\mathbf{d}) - S_{\alpha_k}(\mathbf{v})}{h} \Big)^2 \\ &= \sum_{k \in I_1(\mathbf{v},\mathbf{d})} d_k^2 + \sum_{k \in I_2(\mathbf{v},\mathbf{d})} 0^2 \leq \|\mathbf{d}\|_{\ell_2}^2 < \infty,\end{aligned}$$

i.e., $\mathbf{S}'_{\boldsymbol{\alpha}}(\mathbf{v}, \mathbf{d})$ given elementwise by (2.17) is an element of ℓ_2. □

As a consequence, we deduce the directional differentiability of $\mathbf{F}$ from (1.10) in the following corollary.

Corollary 2.6 *Let $\mathbf{w} = (w_k)_k$ be a positive weight sequence with $w_k \geq w_0 > 0$ and let $\gamma > 0$. The operator $\mathbf{F}\colon \ell_2 \to \ell_2$, $\mathbf{F}(\mathbf{u}) := \mathbf{u} - \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u} - \gamma Dg(\mathbf{u}))$ is directionally differentiable and the directional derivative $\mathbf{F}'(\mathbf{u}, \mathbf{d})$ of $\mathbf{F}$ at $\mathbf{u} \in \ell_2$ in the direction $\mathbf{d} \in \ell_2$ is given elementwise by*

$$F_k'(\mathbf{u}, \mathbf{d}) = \begin{cases} \gamma(D^2 g(\mathbf{u})\mathbf{d})_k, & k \in \mathcal{A}(\mathbf{u}), \\ d_k, & k \in \mathcal{I}^\circ(\mathbf{u}), \\ \min\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}, & k \in \mathcal{I}^+(\mathbf{u}), \\ \max\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}, & k \in \mathcal{I}^-(\mathbf{u}). \end{cases} \tag{2.18}$$

Proof. Because the mapping $\mathbf{u} \mapsto \mathbf{u} - \gamma Dg(\mathbf{u})$ is Fréchet differentiable by Assumption A and because the shrinkage operator $\mathbf{S}_{\gamma\mathbf{w}}$ is Lipschitz continuous (it was shown, e.g., in [26, Lemma 2.2] that $\mathbf{S}_{\gamma\mathbf{w}}$ is nonexpansive), it follows with the chain rule for directional derivatives from Lemma 1.6 that $\mathbf{F}$ is directionally differentiable and that the directional derivative of $\mathbf{F}$ at $\mathbf{u} \in \ell_2$ in the direction $\mathbf{d} \in \ell_2$ is given by

$$\mathbf{F}'(\mathbf{u}, \mathbf{d}) = \mathbf{d} - \mathbf{S}'_{\gamma\mathbf{w}}(\mathbf{u} - \gamma Dg(\mathbf{u}), \mathbf{d} - \gamma D^2 g(\mathbf{u})\mathbf{d}).$$

Componentwise, using Lemma 2.5, we have $F_k'(\mathbf{u}, \mathbf{d}) = d_k - (d_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k) = \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$ in the case $|u_k - \gamma(Dg(\mathbf{u}))_k| > \gamma w_k$ and $F_k'(\mathbf{u}, \mathbf{d}) = d_k - 0 = d_k$ if $|u_k - \gamma(Dg(\mathbf{u}))_k| < \gamma w_k$. If it holds $u_k - \gamma(Dg(\mathbf{u}))_k = \gamma w_k$ and $((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k > 0$ or if $u_k - \gamma(g(\mathbf{u}))_k = -\gamma w_k$ and $((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k < 0$, we get $F_k'(\mathbf{u}, \mathbf{d}) = d_k - (d_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k) = \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$. Finally, we have $F_k'(\mathbf{u}, \mathbf{d}) = d_k$ if $u_k - \gamma(Dg(\mathbf{u}))_k = \gamma w_k$ and $((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k \leq 0$ or if $u_k - \gamma(Dg(\mathbf{u}))_k = -\gamma w_k$ and $((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k \geq 0$, finishing the proof. □

The following lemma is a generalization of our result [66, Lemma 3.1], cf. Lemma 2.8.

Lemma 2.7 *The operator $\mathbf{F}\colon \ell_2 \to \ell_2$, $\mathbf{F}(\mathbf{u}) := \mathbf{u} - \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u} - \gamma Dg(\mathbf{u}))$ with positive weights $w_k \geq w_0 > 0$ and $\gamma > 0$ is Bouligand differentiable, i.e., the directional derivative $\mathbf{F}'$ fulfills the approximation property*

$$\|\mathbf{F}(\mathbf{u} + \mathbf{d}) - \mathbf{F}(\mathbf{u}) - \mathbf{F}'(\mathbf{u}, \mathbf{d})\|_{\ell_2} = o(\|\mathbf{d}\|_{\ell_2}), \quad \mathbf{d} \to \mathbf{0}. \tag{2.19}$$

Additionally, we have

$$\|\mathbf{G}(\mathbf{u} + \mathbf{d}, \mathbf{d}) - \mathbf{F}'(\mathbf{u}, \mathbf{d})\|_{\ell_2} = o(\|\mathbf{d}\|_{\ell_2}), \quad \mathbf{d} \to \mathbf{0}, \tag{2.20}$$

where $\mathbf{G}\colon \ell_2 \times \ell_2 \to \ell_2$ is a uniquely determined mapping with

$$(\mathbf{G}(\mathbf{u}, \mathbf{d}))_k = \begin{cases} d_k, & k \in \mathcal{I}^\circ(\mathbf{u}), \\ \gamma(D^2 g(\mathbf{u})\mathbf{d})_k, & k \in \mathcal{A}(\mathbf{u}), \\ h_k(\mathbf{u}, \mathbf{d}), & k \in \mathcal{I}^\pm(\mathbf{u}), \end{cases} \tag{2.21}$$

where either $h_k(\mathbf{u}, \mathbf{d}) := d_k$ or $h_k(\mathbf{u}, \mathbf{d}) := \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$ for $k \in \mathcal{I}^\pm(\mathbf{u})$. The limit property (2.20) *especially holds true for $\mathbf{G}(\mathbf{u}, \mathbf{d}) := \mathbf{F}'(\mathbf{u}, \mathbf{d})$.*

Proof. We generalize our proof of [66, Lemma 3.1], see Lemma 2.8. For $\mathbf{u}, \mathbf{d} \in \ell_2$, we have

$$\|\mathbf{F}(\mathbf{u}+\mathbf{d}) - \mathbf{F}(\mathbf{u}) - \mathbf{F}'(\mathbf{u},\mathbf{d})\|_{\ell_2}^2 = \sum_{k=1}^{\infty} \big(F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})\big)^2.$$

We suppose $0 < \|\mathbf{d}\|_{\ell_2} \leq d_0$ with

$$d_0 < \frac{\gamma w_0}{2(1+\gamma L_{Dg})},$$

where L_{Dg} denotes the Lipschitz constant of the Fréchet derivative Dg of g in a neighborhood of $\mathbf{u}$ and we are going to reduce d_0 further in the course of the proof.

For $k \in \mathcal{I}^\circ(\mathbf{u})$, we have $|u_k - \gamma(Dg(\mathbf{u}))_k| < \gamma w_k$. By defining

$$I_1(\mathbf{u}) := \{k \in \mathcal{I}^\circ(\mathbf{u}) : |u_k - \gamma(Dg(\mathbf{u}))_k| > \gamma w_0/2\},$$
$$I_2(\mathbf{u}) := \{k \in \mathcal{I}^\circ(\mathbf{u}) : |u_k - \gamma(Dg(\mathbf{u}))_k| \leq \gamma w_0/2\},$$

we have $\mathcal{I}^\circ(\mathbf{u}) = I_1(\mathbf{u}) \cup I_2(\mathbf{u})$ and $|I_1(\mathbf{u})| < \infty$ because $\mathbf{u} - \gamma Dg(\mathbf{u}) \in \ell_2$. On the one hand, using the fact that $I_1(\mathbf{u})$ is finite, we have $I_1(\mathbf{u}) \subset \mathcal{I}^\circ(\mathbf{u}+\mathbf{d})$ for all $\mathbf{d} \in \ell_2$ with $0 < \|\mathbf{d}\|_{\ell_2} \leq d_0$, where

$$d_0 \leq \min_{l \in I_1(\mathbf{u})} \inf\{\|\mathbf{d}\|_{\ell_2} > 0 : |u_l + d_l - \gamma(Dg(\mathbf{u}+\mathbf{d}))_l| < \gamma w_l\}.$$

On the other hand, we have for $k \in I_2(\mathbf{u})$

$$\begin{aligned} |u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k| &\leq |u_k - \gamma(Dg(\mathbf{u}))_k| + |d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k| \\ &\leq \frac{\gamma w_0}{2} + (1+\gamma L_{Dg})\|\mathbf{d}\|_{\ell_2} < \gamma w_0 \leq \gamma w_k, \end{aligned}$$

implying $I_2(\mathbf{u}) \subset \mathcal{I}^\circ(\mathbf{u}+\mathbf{d})$. Because we have shown that $\mathcal{I}^\circ(\mathbf{u}) \subset \mathcal{I}^\circ(\mathbf{u}+\mathbf{d})$ for $\|\mathbf{d}\|_{\ell_2}$ small enough, with (2.11), (2.18) and (2.21), it follows $F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d}) = 0$ and $G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d}) = d_k - d_k = 0$ for all $k \in \mathcal{I}^\circ(\mathbf{u})$.

Because of the finiteness of the active sets $\mathcal{A}^\pm(\mathbf{u})$, we may suppose $0 < \|\mathbf{d}\|_{\ell_2} < d_0$, where

$$d_0 \leq \min_{l \in \mathcal{A}^+(\mathbf{u})} \inf\{\|\mathbf{d}\|_{\ell_2} > 0 : u_l + d_l - \gamma(Dg(\mathbf{u}+\mathbf{d}))_l > \gamma w_l\},$$

as well as

$$d_0 \leq \min_{l \in \mathcal{A}^-(\mathbf{u})} \inf\{\|\mathbf{d}\|_{\ell_2} > 0 : u_l + d_l - \gamma(Dg(\mathbf{u}+\mathbf{d}))_l < -\gamma w_l\}.$$

Therefore, we have $\mathcal{A}^\pm(\mathbf{u}) \subset \mathcal{A}^\pm(\mathbf{u}+\mathbf{d})$ if $\|\mathbf{d}\|_{\ell_2}$ is sufficiently small. Because of Assumption A, it follows with (2.11) and (2.18),

$$|F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})| = |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|.$$

Additionally, we have with (2.21),

$$|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| = |\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|.$$

Let us now consider the finite index sets $\mathcal{I}^{\pm}(\mathbf{u})$. In the case $k \in \mathcal{I}^{+}(\mathbf{u})$, with (2.11) and (2.14), we have $F_k(\mathbf{u}) = u_k = \gamma(Dg(\mathbf{u}))_k + \gamma w_k$. We distinguish the cases $d_k > \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$ and $d_k \leq \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$. If we have $k \in \mathcal{I}^{+}(\mathbf{u})$ and $d_k > \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$, then it follows

$$u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k > u_k - \gamma(Dg(\mathbf{u}))_k = \gamma w_k,$$

hence $k \in \mathcal{A}^{+}(\mathbf{u}+\mathbf{d})$. Therefore, using (2.11) and (2.18),

$$\begin{aligned}
&|F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})| \\
&= |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \min\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\
&\leq |\gamma\big((Dg(\mathbf{u}+\mathbf{d}))_k - (Dg(\mathbf{u}))_k - (D^2g(\mathbf{u})\mathbf{d})_k\big)|,
\end{aligned}$$

because in the case $d_k < \gamma(D^2g(\mathbf{u}))_k$, we have

$$0 > \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - d_k > \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2g(\mathbf{u}))_k.$$

Moreover, with (2.18) and (2.21),

$$\begin{aligned}
&|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| \\
&= |\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \min\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\
&\leq \max\{|\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\
&\qquad |\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (D^2g(\mathbf{u})\mathbf{d})_k\big)|\} \\
&\leq\ 2\max\{|\gamma\big((D^2g(\mathbf{u})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\
&\qquad |\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (D^2g(\mathbf{u})\mathbf{d})_k\big)|\}
\end{aligned}$$

because in the case $d_k < \gamma(D^2g(\mathbf{u})\mathbf{d})_k$, we have

$$\begin{aligned}
\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (D^2g(\mathbf{u})\mathbf{d})_k\big) &< \gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - d_k \\
&< \gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big),
\end{aligned}$$

and by an application of the triangle inequality, it follows

$$\begin{aligned}
&|\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k| \\
&\leq |\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k| \\
&\quad + |\gamma(D^2g(\mathbf{u})\mathbf{d})_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k|.
\end{aligned} \tag{2.22}$$

If we have $k \in \mathcal{I}^{+}(\mathbf{u})$ and $d_k \leq \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$, then with

$$|\gamma(Dg(\mathbf{u}))_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + d_k| \leq (1+\gamma L_{Dg})\|\mathbf{d}\|_{\ell_2} \leq \frac{\gamma w_0}{2} \leq \frac{\gamma w_k}{2}, \tag{2.23}$$

we obtain

$$\gamma w_k \geq u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k = \gamma w_k + \gamma(Dg(\mathbf{u}))_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + d_k \geq \frac{\gamma w_k}{2}.$$

Thus, we have $k \in \mathcal{I}^\circ(\mathbf{u}+\mathbf{d}) \cup \mathcal{I}^+(\mathbf{u}+\mathbf{d})$. With (2.11) and (2.18), it follows

$$\begin{aligned}|F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})| &= |u_k + d_k - u_k - \min\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}| \\ &= |\max\{0, ((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k\}| \\ &\leq |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k|,\end{aligned}$$

because if $d_k \leq \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$, the maximum is equal to zero, and if $d_k > \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$, the claim follows with

$$0 < d_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k \leq \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k.$$

Additionally, with (2.18) and (2.21), we have for $k \in \mathcal{I}^\circ(\mathbf{u}+\mathbf{d})$

$$\begin{aligned}|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| &= |d_k - \min\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}| \\ &= |\max\{0, ((\mathbf{I} - \gamma D^2 g(\mathbf{u}))\mathbf{d})_k\}| \\ &\leq |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k|\end{aligned}$$

with the same arguments as above. If $k \in \mathcal{I}^+(\mathbf{u}+\mathbf{d})$, we have with (2.14)

$$u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k = \gamma w_k = u_k - \gamma(Dg(\mathbf{u}))_k.$$

Hence, we have $d_k = \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$. With (2.18) and (2.21), it follows

$$\begin{aligned}|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| =& |G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - \min\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}| \\ &\leq 2\max\{|\gamma\big((D^2 g(\mathbf{u})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\ &\qquad |\gamma(D^2 g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k|\},\end{aligned}$$

because if $G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) = d_k$, the difference $G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - \min\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}$ is either equal to 0 or equal to $\gamma((Dg(\mathbf{u}+\mathbf{d}))_k - (Dg(\mathbf{u}))_k - (D^2 g(\mathbf{u})\mathbf{d})_k)$. Otherwise, if we have $G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) = \gamma\big((D^2 g(\mathbf{u}+\mathbf{d})\mathbf{d})_k$, the difference is equal to $\gamma(D^2 g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$ or equal to $\gamma((D^2 g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k)$ and the estimate follows from (2.22).

In the case $k \in \mathcal{I}^-(\mathbf{u})$, we have $F_k(\mathbf{u}) = u_k = \gamma(Dg(\mathbf{u}))_k - \gamma w_k$ because of (2.11) and (2.15). First, we consider the case $d_k < \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$. Because of

$$u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k < u_k - \gamma(Dg(\mathbf{u}))_k = -\gamma w_k,$$

it follows $k \in \mathcal{A}^-(\mathbf{u}+\mathbf{d})$. Hence, using (2.11) and (2.18), we get

$$\begin{aligned}&|F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})| \\ &= |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \max\{d_k, \gamma(D^2 g(\mathbf{u})\mathbf{d})_k\}| \\ &\leq |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2 g(\mathbf{u})\mathbf{d})_k|,\end{aligned}$$

because if $d_k > \gamma(D^2 g(\mathbf{u})\mathbf{d})_k$, one has

$$\gamma(D^2 g(\mathbf{u})\mathbf{d})_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k < d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k < 0.$$

Moreover, with (2.18) and (2.21), we get

$$\begin{aligned}
&|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| \\
&= |\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \max\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\
&\le \max\{|\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\
&\qquad |\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (D^2g(\mathbf{u})\mathbf{d})_k\big)|\} \\
&\le 2\max\{|\gamma\big((D^2g(\mathbf{u})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\
&\qquad |\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - (D^2g(\mathbf{u})\mathbf{d})_k\big)|\}
\end{aligned}$$

because in the case $d_k > \gamma(D^2g(\mathbf{u})\mathbf{d})_k$, one has

$$\begin{aligned}
&\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k > \gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - d_k \\
&> \gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + \gamma(Dg(\mathbf{u}))_k
\end{aligned}$$

and the estimate follows as in the case $k \in \mathcal{I}^+(\mathbf{u}) \cap \mathcal{A}^+(\mathbf{u}+\mathbf{d})$ with (2.22).

Second, we consider the case $k \in \mathcal{I}^-(\mathbf{u})$ and $d_k \ge \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$. With (2.23), we have

$$\begin{aligned}
-\frac{\gamma w_k}{2} &\ge -\gamma w_k + \gamma(Dg(\mathbf{u}))_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k + d_k \\
&= u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k \\
&\ge u_k - \gamma(Dg(\mathbf{u}))_k = -\gamma w_k,
\end{aligned}$$

hence $k \in \mathcal{I}^\circ(\mathbf{u}+\mathbf{d}) \cup \mathcal{I}^-(\mathbf{u}+\mathbf{d})$. Now it follows with (2.11) and (2.18)

$$\begin{aligned}
|F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})| &= |u_k + d_k - u_k - \max\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\
&= |\min\{0, ((\mathbf{I} - \gamma D^2g(\mathbf{u}))\mathbf{d})_k\}| \\
&\le |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|,
\end{aligned}$$

because on the one hand, the minimum is equal to zero if $d_k \ge \gamma(D^2g(\mathbf{u})\mathbf{d})_k$. On the other hand, if $d_k < \gamma(D^2g(\mathbf{u})\mathbf{d})_k$, we have

$$\begin{aligned}
0 &> d_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k \\
&\ge \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k.
\end{aligned}$$

Furthermore, using (2.18) and (2.21), we have for $k \in \mathcal{I}^\circ(\mathbf{u}+\mathbf{d})$

$$\begin{aligned}
|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| &= |d_k - \max\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\
&= |\min\{0, ((\mathbf{I} - \gamma D^2g(\mathbf{u}))\mathbf{d})_k\}| \\
&\le |\gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|,
\end{aligned}$$

with the same arguments as above. In the case $k \in \mathcal{I}^-(\mathbf{u}+\mathbf{d})$, we have with (2.15)

$$u_k + d_k - \gamma(Dg(\mathbf{u}+\mathbf{d}))_k = -\gamma w_k = u_k - \gamma(Dg(\mathbf{u}))_k,$$

implying $d_k = \gamma(Dg(\mathbf{u}+\mathbf{d}))_k - \gamma(Dg(\mathbf{u}))_k$ and hence with (2.18) and (2.21)

$$\begin{aligned}|G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})| &= |G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - \max\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}| \\ &\leq 2\max\{|\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|, \\ &\qquad |\gamma\big((D^2g(\mathbf{u})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|\}.\end{aligned}$$

This estimate follows analogously to the case $k \in \mathcal{I}^+(\mathbf{u}) \cap \mathcal{I}^+(\mathbf{u}+\mathbf{d})$.

Altogether, we obtain with the smoothness assumption on g and the local Lipschitz continuity of the second derivative D^2g, cf. Assumption A,

$$\begin{aligned}\|\mathbf{F}(\mathbf{u}+\mathbf{d}) - \mathbf{F}(\mathbf{u}) - \mathbf{F}'(\mathbf{u},\mathbf{d})\|_{\ell_2}^2 &= \sum_{k=1}^{\infty} |F_k(\mathbf{u}+\mathbf{d}) - F_k(\mathbf{u}) - F_k'(\mathbf{u},\mathbf{d})|^2 \\ &\leq \sum_{k\in\mathcal{A}(\mathbf{u})\cup\mathcal{I}^{\pm}(\mathbf{u})} |\gamma(Dg(\mathbf{u}+\mathbf{d}) - Dg(\mathbf{u}) - D^2g(\mathbf{u})\mathbf{d})_k|^2 \\ &= o(\|\mathbf{d}\|_{\ell_2}^2), \quad \mathbf{d}\to\mathbf{0},\end{aligned}$$

and

$$\begin{aligned}&\|\mathbf{G}(\mathbf{u}+\mathbf{d},\mathbf{d}) - \mathbf{F}'(\mathbf{u},\mathbf{d})\|_{\ell_2}^2 \\ &= \sum_{k=1}^{\infty} |G_k(\mathbf{u}+\mathbf{d},\mathbf{d}) - F_k'(\mathbf{u},\mathbf{d})|^2 \\ &\leq \sum_{k\in\mathcal{A}(\mathbf{u})} |\gamma(D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma(D^2g(\mathbf{u})\mathbf{d})_k|^2 \\ &\quad + 4\sum_{k\in\mathcal{I}^{\pm}(\mathbf{u})} \max\{|\gamma\big((D^2g(\mathbf{u})\mathbf{d})_k - (Dg(\mathbf{u}+\mathbf{d}))_k + (Dg(\mathbf{u}))_k\big)|, \\ &\qquad\qquad |\gamma\big((D^2g(\mathbf{u}+\mathbf{d})\mathbf{d})_k - \gamma\big((D^2g(\mathbf{u})\mathbf{d})_k|\}^2 \\ &= o(\|\mathbf{d}\|_{\ell_2}^2), \quad \mathbf{d}\to\mathbf{0},\end{aligned}$$

showing (2.19) and (2.20), finishing the proof. □

The limit property (2.20) will be needed in Remark 2.14 in order to show that the B-Newton method for the nonlinearity $\mathbf{F}$ from (1.10) also is a semismooth Newton method as well as in Theorem 2.15 to verify the semismoothness of $\mathbf{F}$. Note that (2.20) still holds true for $h_k(\mathbf{u},\mathbf{d}) \in \operatorname{conv}\{d_k, \gamma(D^2g(\mathbf{u})\mathbf{d})_k\}$ for $k \in \mathcal{I}^{\pm}(\mathbf{u})$. In case of a quadratic functional $g(\mathbf{u}) = \frac{1}{2}\|K\mathbf{u} - f\|_H^2$, we will prove the stronger property that the residual $\mathbf{F}(\mathbf{u}+h\mathbf{d}) - \mathbf{F}(\mathbf{u}) - h\mathbf{F}'(\mathbf{u},\mathbf{d})$ is exactly equal to $\mathbf{0}$ for sufficiently small h.

Lemma 2.8 *Let $\mathbf{u},\mathbf{d} \in \ell_2$ be arbitrary. There exists $h_0 = h_0(\mathbf{u},\mathbf{d}) > 0$, so that*

$$\frac{F_k(\mathbf{u}+h\mathbf{d}) - F_k(\mathbf{u})}{h} = \begin{cases} \gamma(K^*K\mathbf{d})_k, & k \in \mathcal{A}(\mathbf{u}) \\ d_k, & k \in \mathcal{I}^{\circ}(\mathbf{u}) \\ \min\{d_k, \gamma(K^*K\mathbf{d})_k\}, & k \in \mathcal{I}^{+}(\mathbf{u}) \\ \max\{d_k, \gamma(K^*K\mathbf{d})_k\}, & k \in \mathcal{I}^{-}(\mathbf{u}) \end{cases}, \qquad \text{for all } 0 < h \leq h_0. \tag{2.24}$$

In particular, the directional derivative $\mathbf{F}'(\mathbf{u},\mathbf{d}) = (F_k'(\mathbf{u},\mathbf{d}))_{k\in\mathcal{N}} \in \ell_2$ is given componentwise by the right-hand side of (2.24).

Proof. Let $\mathbf{u}, \mathbf{d} \in \ell_2$, $k \in \mathcal{N}$ and $0 < h < h_0$, where $h_0 \|(\mathbf{I} - \gamma K^*K)\mathbf{d}\|_{\ell_2} \leq \gamma \frac{w_0}{2}$. The value h_0 will be reduced further in the course of the proof.

If $k \in \mathcal{I}^+(\mathbf{u})$, (2.11) and (2.14) tell us that

$$F_k(\mathbf{u}) = u_k = \gamma(K^*(K\mathbf{u} - f))_k + \gamma w_k.$$

If, additionally, $((\mathbf{I} - \gamma K^*K)\mathbf{d})_k > 0$, then $k \in \mathcal{A}^+(\mathbf{u} + h\mathbf{d})$, so that an application of (2.11) yields $F_k(\mathbf{u} + h\mathbf{d}) - F_k(\mathbf{u}) = h\gamma(K^*K\mathbf{d})_k$, showing (2.24). The same argument works for $k \in \mathcal{I}^-(\mathbf{u})$ and $((\mathbf{I} - \gamma K^*K)\mathbf{d})_k < 0$.

If $k \in \mathcal{I}^+(\mathbf{u})$ and $((\mathbf{I} - \gamma K^*K)\mathbf{d})_k \leq 0$, it follows that for all $0 < h \leq h_0$,

$$\big(\mathbf{u} + h\mathbf{d} + \gamma K^*(f - K(\mathbf{u} + h\mathbf{d}))\big)_k = \gamma w_k + h((\mathbf{I} - \gamma K^*K)\mathbf{d})_k \in (-\gamma w_k, \gamma w_k]$$

and hence $k \in \mathcal{I}(\mathbf{u} + h\mathbf{d})$. We conclude that $F_k(\mathbf{u} + h\mathbf{d}) - F_k(\mathbf{u}) = hd_k$ for all $0 < h \leq h_0$, showing (2.24). The same argument works for $k \in \mathcal{I}^-(\mathbf{u})$ and $((\mathbf{I} - \gamma K^*K)\mathbf{d})_k \geq 0$.

Moreover, due to the fact that $\mathcal{A}(\mathbf{u})$ is a finite set, we can find $h_0 > 0$ such that $\mathcal{A}(\mathbf{u}) \subset \mathcal{A}(\mathbf{u} + h\mathbf{d})$ for all $0 < h \leq h_0$ with

$$h_0 \leq \min_{l \in \mathcal{A}(\mathbf{u})} \inf \Big\{ h > 0 : \big|(\mathbf{u} + h\mathbf{d} + \gamma K^*(f - K(\mathbf{u} + h\mathbf{d})))_l\big| > \gamma w_l \Big\}.$$

For $0 < h \leq h_0$ and $k \in \mathcal{A}(\mathbf{u}) \subset \mathcal{A}(\mathbf{u} + h\mathbf{d})$, (2.11) yields $F_k(\mathbf{u} + h\mathbf{d}) - F_k(\mathbf{u}) = h\gamma(K^*K\mathbf{d})_k$ and hence (2.24).

Finally, in order to show (2.24) for $k \in \mathcal{I}^\circ(\mathbf{u})$, we decompose $\mathcal{I}^\circ(\mathbf{u}) = \mathcal{I}_1(\mathbf{u}) \cup \mathcal{I}_2(\mathbf{u})$, setting

$$\begin{aligned} \mathcal{I}_1(\mathbf{u}) &:= \big\{ k \in \mathcal{I}^\circ(\mathbf{u}) : |(\mathbf{u} + \gamma(K^*(f - K\mathbf{u})))_k| > \tfrac{\gamma w_0}{2} \big\}, \\ \mathcal{I}_2(\mathbf{u}) &:= \big\{ k \in \mathcal{I}^\circ(\mathbf{u}) : |(\mathbf{u} + \gamma(K^*(f - K\mathbf{u})))_k| \leq \tfrac{\gamma w_0}{2} \big\}. \end{aligned}$$

Note that $\mathcal{I}_1(\mathbf{u})$ is a finite set, due to $\mathbf{u} + \gamma K^*(f - K\mathbf{u}) \in \ell_2$. We may therefore choose $h_0 > 0$ even sufficiently small that $\mathcal{I}_1(\mathbf{u}) \subset \mathcal{I}(\mathbf{u} + h\mathbf{d})$ for all $0 < h \leq h_0$, e.g., by additionally requiring that

$$h_0 \leq \min_{l \in \mathcal{I}_1(\mathbf{u})} \inf \Big\{ h > 0 : \big|(\mathbf{u} + h\mathbf{d} + \gamma K^*(f - K(\mathbf{u} + h\mathbf{d})))_l\big| < \gamma w_l \Big\}.$$

We conclude that for all $0 < h \leq h_0$ and $k \in \mathcal{I}_2(\mathbf{u})$,

$$\begin{aligned} \big|(\mathbf{u} + h\mathbf{d} + \gamma K^*(f - K(\mathbf{u} + h\mathbf{d})))_k\big| &\leq \big|(\mathbf{u} + \gamma K^*(f - K\mathbf{u}))_k\big| + h\big\|(\mathbf{I} - \gamma K^*K)\mathbf{d}\big\|_{\ell_2} \\ &\leq \gamma w_0 \leq \gamma w_k, \end{aligned}$$

so that also $\mathcal{I}_2(\mathbf{u}) \subset \mathcal{I}(\mathbf{u} + h\mathbf{d})$ for all $0 < h \leq h_0$. (2.11) implies that $F_k(\mathbf{u} + h\mathbf{d}) - F_k(\mathbf{u}) = hd_k$ for all $k \in \mathcal{I}^\circ(\mathbf{u})$ and $0 < h \leq h_0$. Note that $h_0 = h_0(\mathbf{u}, \mathbf{d})$ is independent of k. □

As already mentioned, our globalization strategy in Section 2.3 will choose damping parameters t_j in such a way that sufficient descent of the merit functional Θ_p from (2.2) is guaranteed. To this end, we compute the directional derivatives of Θ_p.

Lemma 2.9 *Let $\mathbf{u}, \mathbf{d} \in \ell_2$ be arbitrary. Then for $p \in \{1, 2\}$, $\Theta_p := \|\mathbf{F}(\cdot)\|_{\ell_p}^p$ is directionally differentiable at $\mathbf{u}$ in the direction $\mathbf{d}$ with directional derivatives*

$$\Theta_p'(\mathbf{u}, \mathbf{d}) = \begin{cases} \sum\limits_{F_k(\mathbf{u})>0} F_k'(\mathbf{u}, \mathbf{d}) + \sum\limits_{F_k(\mathbf{u})=0} \left|F_k'(\mathbf{u}, \mathbf{d})\right| - \sum\limits_{F_k(\mathbf{u})<0} F_k'(\mathbf{u}, \mathbf{d}), & p = 1, \\ 2 \sum\limits_{k \in \mathcal{N}} F_k'(\mathbf{u}, \mathbf{d}) F_k(\mathbf{u}), & p = 2. \end{cases} \tag{2.25}$$

Proof. The identity (2.25) immediately follows by an application of the chain rule for directionally differentiable mappings, see Lemma 1.6, cf. [124, Theorem 3.1.1] or [128, Proposition 3.6 (i), Proposition 3.5]. □

2.2.2 The feasibility of the local B-semismooth Newton iteration

In the sequel, for a given $\mathbf{u} \in \ell_2$, we will discuss the existence and uniqueness of a solution $\mathbf{d} \in \ell_2$ to the nonlinear problem (1.18)

$$\mathbf{F}'(\mathbf{u}, \mathbf{d}) = -\mathbf{F}(\mathbf{u}).$$

Note that in view of Lemma 2.9, a solution $\mathbf{d}$ to (1.18) has the particularly interesting properties that

$$\begin{aligned} \Theta_1'(\mathbf{u}, \mathbf{d}) &= -\sum_{k \in \mathcal{N}} |F_k(\mathbf{u})| = -\Theta_1(\mathbf{u}), \\ \Theta_2'(\mathbf{u}, \mathbf{d}) &= 2\big\langle \mathbf{F}'(\mathbf{u}, \mathbf{d}), \mathbf{F}(\mathbf{u}) \big\rangle = -2\Theta_2(\mathbf{u}). \end{aligned} \tag{2.26}$$

It will turn out that a solution $\mathbf{d}$ to (1.18) coincides with the semismooth Newton direction from [58, 103] if $\mathcal{I}^{\pm}(\mathbf{u}) = \emptyset$, see Lemma 2.11. The following auxiliary results ensure the solvability of (1.18).

Lemma 2.10 *Let $\mathbf{u} \in \ell_2$, $\gamma > 0$, $\mathbf{M} := D^2 g(\mathbf{u})$ and*

$$\mathbf{N} := \gamma \begin{pmatrix} \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^+} - \mathbf{M}_{\mathcal{I}^+,\mathcal{A}} \mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1} \mathbf{M}_{\mathcal{A},\mathcal{I}^+} & \mathbf{M}_{\mathcal{I}^+,\mathcal{A}} \mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1} \mathbf{M}_{\mathcal{A},\mathcal{I}^-} - \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^-} \\ \mathbf{M}_{\mathcal{I}^-,\mathcal{A}} \mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1} \mathbf{M}_{\mathcal{A},\mathcal{I}^+} - \mathbf{M}_{\mathcal{I}^-,\mathcal{I}^+} & \mathbf{M}_{\mathcal{I}^-,\mathcal{I}^-} - \mathbf{M}_{\mathcal{I}^-,\mathcal{A}} \mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1} \mathbf{M}_{\mathcal{A},\mathcal{I}^-} \end{pmatrix}. \tag{2.27}$$

Then the finite-dimensional matrix $\mathbf{N}$ is symmetric and positive definite.

Proof. Without loss of generality, let $\gamma = 1$. $\mathbf{N}$ is finite-dimensional because of the finiteness of the index sets $\mathcal{I}^+, \mathcal{I}^-$ and $\mathcal{A}$. The symmetry of $\mathbf{N}$ readily follows from that of $\mathbf{M}$. Concerning the positive definiteness of $\mathbf{N}$, note that each real 3×3 block matrix with symmetric off-diagonal blocks and symmetric invertible diagonal block $\mathbf{F}$ can be decomposed as

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{B}^\top & \mathbf{D} & \mathbf{E} \\ \mathbf{C}^\top & \mathbf{E}^\top & \mathbf{F} \end{pmatrix} = \mathbf{X} \begin{pmatrix} \mathbf{A} - \mathbf{C}\mathbf{F}^{-1}\mathbf{C}^\top & \mathbf{C}\mathbf{F}^{-1}\mathbf{E}^\top - \mathbf{B} & \mathbf{0} \\ \mathbf{E}\mathbf{F}^{-1}\mathbf{C}^\top - \mathbf{B}^\top & \mathbf{D} - \mathbf{E}\mathbf{F}^{-1}\mathbf{E}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F} \end{pmatrix} \mathbf{X}^\top, \tag{2.28}$$

with

$$\mathbf{X} = \begin{pmatrix} -\mathbf{I} & \mathbf{0} & \mathbf{C}\mathbf{F}^{-1} \\ \mathbf{0} & \mathbf{I} & \mathbf{E}\mathbf{F}^{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

Setting $\mathbf{A} := \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^+}$, $\mathbf{B} := \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^-}$, $\mathbf{C} := \mathbf{M}_{\mathcal{I}^+,\mathcal{A}}$, $\mathbf{D} := \mathbf{M}_{\mathcal{I}^-,\mathcal{I}^-}$, $\mathbf{E} := \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}$ and $\mathbf{F} := \mathbf{M}_{\mathcal{A},\mathcal{A}}$, (2.28) tells us that $\mathbf{N}$ is the upper left 2×2 diagonal block of the matrix $\mathbf{X}^{-1}\mathbf{M}_{\mathcal{I}^+\cup\mathcal{I}^-\cup\mathcal{A},\mathcal{I}^+\cup\mathcal{I}^-\cup\mathcal{A}}\mathbf{X}^{-T}$ and hence with Sylvester's law of inertia positive definite, since $\mathbf{M}_{\mathcal{A},\mathcal{A}}$ and $\mathbf{M}_{\mathcal{I}^\pm,\mathcal{I}^\pm}$ are positive definite according to Assumption A. □

The solvability of (1.18) hinges on the solvability of a particular linear complementarity problem.

Lemma 2.11 *Let $\mathbf{u} \in \ell_2$. Then with $\mathbf{M} := D^2 g(\mathbf{u})$, $\mathbf{d} \in \ell_2$ solves* (1.18) *if and only if*

$$\gamma(\mathbf{M}\mathbf{d})_{\mathcal{A}} = -\mathbf{F}(\mathbf{u})_{\mathcal{A}}, \tag{2.29}$$

$$\mathbf{d}_{\mathcal{I}^\circ} = -\mathbf{u}_{\mathcal{I}^\circ}, \tag{2.30}$$

and

$$\mathbf{x} := \begin{pmatrix} \mathbf{d}_{\mathcal{I}^+} + \mathbf{u}_{\mathcal{I}^+} \\ -\mathbf{d}_{\mathcal{I}^-} - \mathbf{u}_{\mathcal{I}^-} \end{pmatrix}, \quad \mathbf{y} := \begin{pmatrix} \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^+} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ -\gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^-} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix}, \tag{2.31}$$

solve the linear complementarity problem

$$\mathbf{x}, \mathbf{y} \geq \mathbf{0}, \quad \mathbf{y} = \mathbf{N}\mathbf{x} + \mathbf{z}, \quad \langle \mathbf{x}, \mathbf{y} \rangle = 0, \tag{2.32}$$

where the data $\mathbf{N} = \mathbf{N}(\mathbf{u})$ and $\mathbf{z} = \mathbf{z}(\mathbf{u})$ are given by (2.27) *and*

$$\begin{aligned} \mathbf{z} = &\begin{pmatrix} \gamma(\mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^\circ})\mathbf{u}_{\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{F}(\mathbf{u})_{\mathcal{A}} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ \gamma(\mathbf{M}_{\mathcal{I}^-,\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^\circ})\mathbf{u}_{\mathcal{I}^\circ} + \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{F}(\mathbf{u})_{\mathcal{A}} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix} \\ &- \gamma \begin{pmatrix} \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^+} - \mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^+} & \mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^-} - \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^-} \\ \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^+} - \mathbf{M}_{\mathcal{I}^-,\mathcal{I}^+} & \mathbf{M}_{\mathcal{I}^-,\mathcal{I}^-} - \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^-} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{\mathcal{I}^+} \\ -\mathbf{u}_{\mathcal{I}^-} \end{pmatrix}. \end{aligned} \tag{2.33}$$

Proof. By componentwise application of (2.11) and (2.18) to the indices from $\mathcal{A}$, $\mathcal{I}^\circ$, $\mathcal{I}^+$ and $\mathcal{I}^-$, we observe first that (1.18) is equivalent to (2.29), (2.30) and the conditions

$$\begin{aligned} &\big(d_k + u_k \geq 0 \wedge \gamma(\mathbf{M}\mathbf{d})_k + F_k(\mathbf{u}) = 0\big) \vee \big(d_k + u_k = 0 \wedge \gamma(\mathbf{M}\mathbf{d})_k + F_k(\mathbf{u}) \geq 0\big), \quad k \in \mathcal{I}^+, \\ &\big(d_k + u_k \leq 0 \wedge \gamma(\mathbf{M}\mathbf{d})_k + F_k(\mathbf{u}) = 0\big) \vee \big(d_k + u_k = 0 \wedge \gamma(\mathbf{M}\mathbf{d})_k + F_k(\mathbf{u}) \leq 0\big), \quad k \in \mathcal{I}^-, \end{aligned}$$

respectively because $F_k(\mathbf{u}) = u_k$ for $k \in \mathcal{I}^\pm(\mathbf{u})$. The last two conditions are actually equivalent to

$$\mathbf{d}_{\mathcal{I}^+} + \mathbf{u}_{\mathcal{I}^+} \geq \mathbf{0}, \quad \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} + \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^+} \geq \mathbf{0}, \quad \big\langle \mathbf{d}_{\mathcal{I}^+} + \mathbf{u}_{\mathcal{I}^+}, \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} + \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^+} \big\rangle = 0 \tag{2.34}$$

and

$$\mathbf{d}_{\mathcal{I}^-} + \mathbf{u}_{\mathcal{I}^-} \leq \mathbf{0}, \quad \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} + \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^-} \leq \mathbf{0}, \quad \big\langle \mathbf{d}_{\mathcal{I}^-} + \mathbf{u}_{\mathcal{I}^-}, \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} + \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^-} \big\rangle = 0. \tag{2.35}$$

Now assume that $\mathbf{d} \in \ell_2$ solves (1.18) and hence (2.29), (2.30), (2.34) and (2.35) hold. By defining $\mathbf{x}$, $\mathbf{y}$ as in (2.31), (2.34) and (2.35) tell us that $\mathbf{x} \geq \mathbf{0}$, $\mathbf{y} \geq \mathbf{0}$ and $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{0}$. Moreover, (2.29) and (2.30) imply that

$$\mathbf{d}_{\mathcal{A}} = \tfrac{1}{\gamma}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\big(\gamma\mathbf{M}_{\mathcal{A},\mathcal{I}^\circ}\mathbf{u}_{\mathcal{I}^\circ} - \mathbf{F}(\mathbf{u})_{\mathcal{A}} - \gamma\mathbf{M}_{\mathcal{A},\mathcal{I}^+}\mathbf{d}_{\mathcal{I}^+} - \gamma\mathbf{M}_{\mathcal{A},\mathcal{I}^-}\mathbf{d}_{\mathcal{I}^-}\big). \tag{2.36}$$

Inserting (2.36) into the definition of $\mathbf{y}$ from (2.31), we obtain

$$\begin{aligned}
\mathbf{y} &= \begin{pmatrix} \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^+} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ -\gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^-} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix} \\
&= \begin{pmatrix} \gamma\mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{d}_{\mathcal{A}} + \gamma\mathbf{M}_{\mathcal{I}^+,\mathcal{I}^+}\mathbf{d}_{\mathcal{I}^+} + \gamma\mathbf{M}_{\mathcal{I}^+,\mathcal{I}^-}\mathbf{d}_{\mathcal{I}^-} - \gamma\mathbf{M}_{\mathcal{I}^+,\mathcal{I}^\circ}\mathbf{u}_{\mathcal{I}^\circ} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ -\gamma\mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{d}_{\mathcal{A}} - \gamma\mathbf{M}_{\mathcal{I}^-,\mathcal{I}^+}\mathbf{d}_{\mathcal{I}^+} - \gamma\mathbf{M}_{\mathcal{I}^-,\mathcal{I}^-}\mathbf{d}_{\mathcal{I}^-} + \gamma\mathbf{M}_{\mathcal{I}^-,\mathcal{I}^\circ}\mathbf{u}_{\mathcal{I}^\circ} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix} \\
&= \mathbf{N}\begin{pmatrix} \mathbf{d}_{\mathcal{I}^+} \\ -\mathbf{d}_{\mathcal{I}^-} \end{pmatrix} + \begin{pmatrix} \gamma(\mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^+,\mathcal{I}^\circ})\mathbf{u}_{\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^+,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{F}(\mathbf{u})_{\mathcal{A}} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ \gamma(\mathbf{M}_{\mathcal{I}^-,\mathcal{I}^\circ} - \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{M}_{\mathcal{A},\mathcal{I}^\circ})\mathbf{u}_{\mathcal{I}^\circ} + \mathbf{M}_{\mathcal{I}^-,\mathcal{A}}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\mathbf{F}(\mathbf{u})_{\mathcal{A}} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix} \\
&= \mathbf{N}\mathbf{x} + \mathbf{z},
\end{aligned}$$

with $\mathbf{N}$ from (2.27) and $\mathbf{z}$ from (2.33), showing (2.32).

Conversely, assume that $\mathbf{x}$, $\mathbf{y}$ solve (2.32) and define $\mathbf{d} \in \ell_2$ blockwise via

$$\begin{pmatrix} \mathbf{d}_{\mathcal{I}^+} \\ -\mathbf{d}_{\mathcal{I}^-} \end{pmatrix} := \mathbf{x} + \begin{pmatrix} -\mathbf{u}_{\mathcal{I}^+} \\ \mathbf{u}_{\mathcal{I}^-} \end{pmatrix}$$

and the system (2.29), (2.30) for the remaining entries from $\mathcal{A} \cup \mathcal{I}^\circ$. It remains to show that $\mathbf{d}$ solves (1.18). As we have already seen above, (2.29) and (2.30) are equivalent to (1.18) holding in the entries corresponding to $\mathcal{A} \cup \mathcal{I}^\circ$. Moreover, as above, (2.29) and (2.30) imply (2.36), which yields

$$\mathbf{y} = \mathbf{N}\mathbf{x} + \mathbf{z} = \begin{pmatrix} \gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^+} + \mathbf{F}(\mathbf{u})_{\mathcal{I}^+} \\ -\gamma(\mathbf{M}\mathbf{d})_{\mathcal{I}^-} - \mathbf{F}(\mathbf{u})_{\mathcal{I}^-} \end{pmatrix},$$

tracing back the previous computation. By consequence, (2.31) holds. (2.32) implies (2.34) and (2.35), and thus (1.18) holds in the remaining entries corresponding to $\mathcal{I}^+ \cup \mathcal{I}^-$. □

Remark 2.12 *The definition of the vector* $\mathbf{y}$ *from* (2.31) *and the vector* $\mathbf{z}$ *from* (2.33) *differs from the definition in our publication [66, Lemma 3.4]. Here, we write* $\mathbf{F}(\mathbf{u})_{\mathcal{I}^\pm}$ *instead of* $\mathbf{u}_{\mathcal{I}^\pm}$ *in each component of* $\mathbf{y}$ *and in the minuend of* $\mathbf{z}$*, respectively. Although we have* $\mathbf{F}(\mathbf{u})_{\mathcal{I}^\pm} = \mathbf{u}_{\mathcal{I}^\pm}$*, this detail is essential for establishing the modified B-semismooth Newton method in Chapter 3.*

We are now in the position to prove the following theorem on the solvability of (1.18).

Theorem 2.13 *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{|\mathcal{I}^\pm|}$ *be the unique solution to* (2.32) *for an iterate* $\mathbf{u}^{(j)} \in \ell_2$ *and let* $\mathcal{I} = \mathcal{I}^\circ \cup \mathcal{I}^+ \cup \mathcal{I}^-$ *and* $\mathbf{M} := D^2 g(\mathbf{u}^{(j)})$*. Then the local B-semismooth Newton iteration* (1.17) *is well-defined, with*

$$\begin{aligned}
\mathbf{d}_{\mathcal{I}^\circ}^{(j)} &= -\mathbf{u}_{\mathcal{I}^\circ}^{(j)}, \\
\mathbf{d}_{\mathcal{I}^+}^{(j)} &= \mathbf{x}_{\mathcal{I}^+} - \mathbf{u}_{\mathcal{I}^+}^{(j)}, \\
\mathbf{d}_{\mathcal{I}^-}^{(j)} &= -\mathbf{x}_{\mathcal{I}^-} - \mathbf{u}_{\mathcal{I}^-}^{(j)}, \\
\mathbf{d}_{\mathcal{A}}^{(j)} &= \frac{1}{\gamma}\mathbf{M}_{\mathcal{A},\mathcal{A}}^{-1}\big(-\gamma\mathbf{M}_{\mathcal{A},\mathcal{I}}\mathbf{d}_{\mathcal{I}}^{(j)} - \mathbf{F}(\mathbf{u}^{(j)})_{\mathcal{A}}\big).
\end{aligned} \tag{2.37}$$

Proof. The finite-dimensional linear complementarity problem (2.32) is uniquely solvable if $\mathbf{N}$ is a P-matrix, i.e., if all of its principal minors are positive [24, Definition 3.3.1, Theorem 3.3.7]. According to Lemma 2.10, $\mathbf{N} = \mathbf{N}(\mathbf{u}^{(j)})$ is symmetric and positive definite. Hence, $\mathbf{N}$ is a P-Matrix and (1.17), (2.37) are well-defined. □

There exist standard methods for the efficient numerical solution for the linear complementarity problem that appears in (2.37), e.g., Lemke's method or pivoting methods, see [24]. The CVX toolbox [25] or generalized damped Newton methods like [70] can be used as well.

Remark 2.14 *Analogously to our remark [66, Remark 4.11], we now justify the naming of the* B-semismooth Newton method (1.17)*, see also [115, Lemma 2.2, Proposition 3.4]. In matrix notation, the generalized Newton equation for the Newton direction* (2.37) *can be expressed by*

$$\mathbf{G}(\mathbf{u})\mathbf{d} = -\mathbf{F}(\mathbf{u}),$$

where $\mathbf{G}$ *is defined block by block*

$$\begin{pmatrix} \mathbf{G}(\mathbf{u})_{\mathcal{B},\mathcal{B}} & \mathbf{G}(\mathbf{u})_{\mathcal{B},\mathcal{C}} \\ \mathbf{G}(\mathbf{u})_{\mathcal{C},\mathcal{B}} & \mathbf{G}(\mathbf{u})_{\mathcal{C},\mathcal{C}} \end{pmatrix} = \begin{pmatrix} \gamma D^2 g(\mathbf{u})_{\mathcal{B},\mathcal{B}} & \gamma D^2 g(\mathbf{u})_{\mathcal{B},\mathcal{C}} \\ \mathbf{0}_{\mathcal{C},\mathcal{B}} & \mathbf{I}_{\mathcal{C},\mathcal{C}} \end{pmatrix}, \tag{2.38}$$

and

$$\begin{aligned} \mathcal{B} = \mathcal{B}(\mathbf{u}) &:= \mathcal{A}(\mathbf{u}) \cup \{k \in \mathcal{I}^{\pm}(\mathbf{u}) : x_k > 0\}, \\ \mathcal{C} = \mathcal{C}(\mathbf{u}) &:= \mathcal{I}^{\circ}(\mathbf{u}) \cup \{k \in \mathcal{I}^{\pm}(\mathbf{u}) : x_k = 0\} = \mathcal{N} \backslash \mathcal{B}, \end{aligned} \tag{2.39}$$

and $\mathbf{x} = (x_k)_k$ *denotes the solution to the linear complementarity problem* (2.32)*. With*

$$\begin{aligned} &\lim_{\mathbf{d}\to\mathbf{0}} \frac{\left\|\mathbf{F}(\mathbf{u}+\mathbf{d}) - \mathbf{F}(\mathbf{u}) - \mathbf{G}(\mathbf{u}+\mathbf{d})\mathbf{d}\right\|_{\ell_2}}{\|\mathbf{d}\|_{\ell_2}} \\ &\leq \lim_{\mathbf{d}\to\mathbf{0}} \frac{\left\|\mathbf{F}(\mathbf{u}+\mathbf{d}) - \mathbf{F}(\mathbf{u}) - \mathbf{F}'(\mathbf{u},\mathbf{d})\right\|_{\ell_2}}{\|\mathbf{d}\|_{\ell_2}} + \lim_{\mathbf{d}\to\mathbf{0}} \frac{\left\|\mathbf{F}'(\mathbf{u},\mathbf{d}) - \mathbf{G}(\mathbf{u}+\mathbf{d})\mathbf{d}\right\|_{\ell_2}}{\|\mathbf{d}\|_{\ell_2}} \end{aligned}$$

and Lemma 2.7, it follows that $\mathbf{G}$ *is indeed a Newton derivative of* $\mathbf{F}$*, justifying the naming of the method. The second limit in the sum is equal to zero because the limit property* (2.20) *holds true for every element of the slant derivative* $\partial_S^G \mathbf{F}(\mathbf{u}+\mathbf{d})$*, see Definition 1.23. Note that* $\mathbf{x} = \mathbf{x}(\mathbf{u})$ *as well as the matrix* $\mathbf{G}(\mathbf{u})$ *do not depend on the vector* $\mathbf{d}$*. Nevertheless, the generalized Newton equation is not only a system of linear equations because one has to solve the linear complementarity problem* (2.32) *to set up* $\mathbf{G}(\mathbf{u})$*. Moreover, the identity* $\mathbf{F}'(\mathbf{u},\mathbf{d}) = \mathbf{G}(\mathbf{u})\mathbf{d}$ *holds true only for the solution* $\mathbf{d} = \mathbf{d}(\mathbf{u})$ *to the generalized Newton equation* (1.18)*.*

Theorem 2.15 *Let* $\mathbf{w} = (w_k)_k$ *be a positive weight sequence with* $w_k \geq w_0 > 0$ *and let Assumption A be fulfilled. Then, the mapping* $\mathbf{F}\colon \ell_2 \to \ell_2$*,* $\mathbf{F}(\mathbf{u}) = \mathbf{u} - \gamma \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u} - \gamma Dg(\mathbf{u}))$ *is semismooth.*

Proof. The claim directly follows from Remark 2.14, Lemma 2.7 and [18, Theorem 3.3], see Theorem 1.25. □

Concerning the convergence properties of the local B-semismooth Newton method, we can rely on the following result from [58, Theorem 3.13] and [103, Theorem 3.7].

Corollary 2.16 *Let* $\mathbf{u}^*$ *be a solution to* $\mathbf{F}(\mathbf{u}) = \mathbf{0}$*. Then, the iterates of the local B-semismooth Newton iteration* (1.17) *converge to* $\mathbf{u}^*$ *superlinearly in a neighborhood of* $\mathbf{u}^*$*.*

Proof. We follow the outline of [58, Section 3.4] and [103, Section 3], respectively. First, we show that there exists a finite index $\kappa \in \mathcal{N}$ such that we have $\mathcal{B}(\mathbf{u}) \subset \{1, \dots, \kappa\}$, with $\mathcal{B}$ from (2.39), for all $\mathbf{u}$ in a neighborhood of the zero $\mathbf{u}^*$ of $\mathbf{F}$. Second, we will prove the uniform boundedness of $\mathbf{G}(\mathbf{u})^{-1}$ with $\mathbf{G}$ from (2.38) for all $\mathbf{u}$ sufficiently close to $\mathbf{u}^*$ and finally, we deduce the local superlinear convergence of the B-semismooth Newton method (1.17).

We determine whether an index k is an element of the index set $\mathcal{B}(\mathbf{u})$, i.e., whether we have $|u_k - \gamma(Dg(\mathbf{u}))_k| \geq \gamma w_k$ for an index k. Because $\mathbf{u}^* - \gamma Dg(\mathbf{u}^*) \in \ell_2$, there exists an index $\kappa > 0$ with $|u_k^* - \gamma(Dg(\mathbf{u}^*))_k| < \gamma w_0/2$ for all $k \geq \kappa$ and hence

$$\begin{aligned} |u_k - \gamma(Dg(\mathbf{u}))_k| &\leq |u_k^* - \gamma(Dg(\mathbf{u}^*))_k| + |u_k - u_k^* - \gamma(Dg(\mathbf{u}))_k + \gamma(Dg(\mathbf{u}^*))_k| \\ &< \frac{\gamma w_0}{2} + (1 + \gamma L_{Dg})\|\mathbf{u} - \mathbf{u}^*\|_{\ell_2} \\ &\leq \gamma w_0 \leq \gamma w_k, \end{aligned}$$

for all $\mathbf{u} \in \ell_2$ with $\|\mathbf{u} - \mathbf{u}^*\|_{\ell_2} \leq \frac{\gamma w_0}{2(1+\gamma L_{Dg})} =: r$ and $k \geq \kappa$, where L_{Dg} denotes the Lipschitz constant of Dg in a neighborhoof of $\mathbf{u}^*$. Thus, we have $k \notin \mathcal{B}(\mathbf{u})$ for all $k \geq \kappa$ if $\mathbf{u}$ is sufficiently close to $\mathbf{u}^*$ and κ only depends on $\mathbf{u}^*$, γ and w_0.

Because of Assumption A, the finite submatrix $D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}}$ is boundedly invertible for any fixed, finite index set $\mathcal{B}$ in a neighborhood of $\mathbf{u}^*$. Without loss of generality, we may assume that this assumption is fulfilled in $B_{\tilde{r}}(\mathbf{u}^*) := \{\mathbf{u} \in \ell_2 : \|\mathbf{u} - \mathbf{u}^*\|_{\ell_2} < \tilde{r}\}$ with $0 < \tilde{r} \leq r$. Hence, $\mathbf{G}$ is invertible at $\mathbf{u}$ and the inverse, blocked according to the index sets $\mathcal{B}$ and $\mathcal{C}$, is given by

$$\begin{pmatrix} (\mathbf{G}(\mathbf{u})^{-1})_{\mathcal{B},\mathcal{B}} & (\mathbf{G}(\mathbf{u})^{-1})_{\mathcal{B},\mathcal{C}} \\ (\mathbf{G}(\mathbf{u})^{-1})_{\mathcal{C},\mathcal{B}} & (\mathbf{G}(\mathbf{u})^{-1})_{\mathcal{C},\mathcal{C}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\gamma}(D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}})^{-1} & -(D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}})^{-1}D^2g(\mathbf{u})_{\mathcal{B},\mathcal{C}} \\ \mathbf{0}_{\mathcal{C},\mathcal{B}} & \mathbf{I}_{\mathcal{C},\mathcal{C}} \end{pmatrix}.$$

Therefore, there exist constants $c_0, c_1 > 0$ such that the inverse $\mathbf{G}(\mathbf{u})^{-1}$ is uniformly bounded for all $\mathbf{u} \in B_{\tilde{r}}(\mathbf{u}^*)$,

$$\begin{aligned} &\|\mathbf{G}(\mathbf{u})^{-1}\|_{L(\ell_2)} \\ &\leq \frac{1}{\gamma}\|(D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}})^{-1}\|_2 + \|(D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}})^{-1}\|_2\|D^2g(\mathbf{u})_{\mathcal{B},\mathcal{C}}\|_{L(\ell_2,\mathbb{R}^{|\mathcal{B}|})} + 1 \\ &\leq \Big(\frac{1}{\gamma} + \|D^2g(\mathbf{u})\|_{L(\ell_2)}\Big)\|(D^2g(\mathbf{u})_{\mathcal{B},\mathcal{B}})^{-1}\|_2 + 1 \\ &\leq \Big(\frac{1}{\gamma} + c_0\Big) \max_{\emptyset \neq \mathcal{J} \subset \{1,\dots,\kappa\}} \sup_{\mathbf{u} \in B_{\tilde{r}}(\mathbf{u}^*)} \|(D^2g(\mathbf{u})_{\mathcal{J},\mathcal{J}})^{-1}\|_2 + 1 \\ &\leq \Big(\frac{1}{\gamma} + c_0\Big)c_1 + 1 =: C, \end{aligned} \tag{2.40}$$

cf. [58, Proposition 3.11] and [103, Lemma 3.6]. Here, we used the fact that $\|D^2g(\mathbf{u})\|_{L(\ell_2)}$ is uniformly bounded in a neighborhood of $\mathbf{u}^*$ because of the local Lipschitz continuity of Dg, see Assumption A, cf. [103, proof of Theorem 3.7]. Using the idea of [73, proof of Theorem 1.1] (see also [17, Remark 2.7], [18, Theorem 3.4]), we can argue that for every

Algorithm 1 The damped B-semismooth Newton method `BSSN`

Choose a starting vector $\mathbf{u}^{(0)} \in \mathbb{R}^n$, parameters $p \in \{1,2\}$, $\beta \in (0,1), \sigma \in (0, \frac{1}{p})$ and a tolerance $tol > 0$ and set $j := 0$.
while $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2 \geq tol$ **do**
  Compute the Newton direction $\mathbf{d}^{(j)}$ according to (2.37)
  $t_j = 1$
  while $\Theta_p(\mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}) > (1 - p\sigma t_j)\Theta_p(\mathbf{u}^{(j)})$ **do**
    $t_j = t_j \beta$
  end while
  $\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}$
  $j = j + 1$
end while

$\xi \in (0,1]$ there exists a radius $\hat{r}$ with $\tilde{r} \geq \hat{r} > 0$ such that we have for all $\mathbf{u}^{(j)} \in B_{\hat{r}}(\mathbf{u}^*)$,

$$\begin{aligned}
\|\mathbf{u}^{(j+1)} - \mathbf{u}^*\|_{\ell_2} &= \|\mathbf{u}^{(j)} - \mathbf{G}(\mathbf{u}^{(j)})^{-1}\mathbf{F}(\mathbf{u}^{(j)}) - \mathbf{u}^*\|_{\ell_2} \\
&= \|\mathbf{u}^{(j)} - \mathbf{u}^* - \mathbf{G}(\mathbf{u}^{(j)})^{-1}(\mathbf{F}(\mathbf{u}^{(j)}) - \mathbf{F}(\mathbf{u}^*))\|_{\ell_2} \\
&\leq \|\mathbf{G}(\mathbf{u}^{(j)})^{-1}\|_{L(\ell_2)} \|\mathbf{G}(\mathbf{u}^{(j)})(\mathbf{u}^{(j)} - \mathbf{u}^*) - \mathbf{F}(\mathbf{u}^{(j)}) + \mathbf{F}(\mathbf{u}^*)\|_{\ell_2} \\
&\leq \|\mathbf{G}(\mathbf{u}^{(j)})^{-1}\|_{L(\ell_2)} \frac{\xi}{C} \|\mathbf{u}^{(j)} - \mathbf{u}^*\|_{\ell_2} \leq \xi \|\mathbf{u}^{(j)} - \mathbf{u}^*\|_{\ell_2} \leq \|\mathbf{u}^{(j)} - \mathbf{u}^*\|_{\ell_2},
\end{aligned}$$

because of the Newton differentiability of $\mathbf{F}$. Inductively, we have $\mathbf{u}^{(j)} \in B_{\hat{r}}(\mathbf{u}^*)$ for all $j \geq 1$ if $\mathbf{u}^{(0)} \in B_{\hat{r}}(\mathbf{u}^*)$. The claim follows from the fact that $\xi \in (0,1]$ was arbitrarily chosen. □

Remark 2.17 *Local B-semismooth Newton iterations may show a cycling effect, as can be shown by the very examples of Subsection 2.1.2. Indeed, the Newton directions of* (2.37) *and the Newton directions from [58] coincide in these examples. Any globalization strategy which relies on residual descent will avoid the cycling effect, see Section 2.3.*

2.3 The global B-semismooth Newton iteration

In this section, we will analyze the following damped B-semismooth Newton method for the computation of ℓ_1 Tikhonov minimizers

$$\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}, \qquad \mathbf{F}'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = -\mathbf{F}(\mathbf{u}^{(j)}), \qquad j = 0, 1, \ldots \tag{2.41}$$

and its global convergence properties in a finite-dimensional setting $\mathcal{N} = \{1, \ldots, n\}$. Inspired by the works [63, 77, 107, 111], we shall choose the damping parameters t_j by inexact line search. The overall algorithm can be found in Algorithm 1.

Let us first show that damping parameters t_j are well-defined.

Proposition 2.18 *Let $p \in \{1,2\}, \beta \in (0,1), \sigma \in (0, 1/p)$. Let $\mathbf{u}^{(j)}$ with $\Theta_p(\mathbf{u}^{(j)}) > 0$ be an iterate in Algorithm 1 and let $\mathbf{d}^{(j)} = \mathbf{d}(\mathbf{u}^{(j)})$ be chosen according to* (2.37)*. Then there exists an index $l \in \mathbb{N}$ with*

$$\Theta_p(\mathbf{u}^{(j)} + \beta^l \mathbf{d}^{(j)}) \leq (1 - p\sigma\beta^l)\Theta_p(\mathbf{u}^{(j)}). \tag{2.42}$$

Proof. According to (2.26), we have $\Theta_p'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = -p\Theta_p(\mathbf{u}^{(j)}) < 0$. Assuming that there exists no such index $l \in \mathbb{N}$, one has for all $l \in \mathbb{N}$

$$\frac{\Theta_p(\mathbf{u}^{(j)} + \beta^l \mathbf{d}^{(j)}) - \Theta_p(\mathbf{u}^{(j)})}{\beta^l} > -p\sigma\Theta_p(\mathbf{u}^{(j)}).$$

For $l \to \infty$, it follows that $\Theta_p'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) \geq -p\sigma\Theta_p(\mathbf{u}^{(j)})$, which is a contradiction to $\Theta_p'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = -p\Theta_p(\mathbf{u}^{(j)}) < -p\sigma\Theta_p(\mathbf{u}^{(j)})$. □

Remark 2.19 *For the particular B-Newton direction from* (2.37)*, inequality* (2.42) *is equivalent to the ordinary Armijo rule, because*

$$\begin{aligned}\Theta_p(\mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}) &\leq (1 - p\sigma t_j)\Theta_p(\mathbf{u}^{(j)}) = \Theta_p(\mathbf{u}^{(j)}) - p\sigma t_j \Theta_p(\mathbf{u}^{(j)}) \\ &= \Theta_p(\mathbf{u}^{(j)}) + \sigma t_j \Theta_p'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)})\end{aligned}$$

with $\sigma \in (0, \frac{1}{p})$.

As an important ingredient of our convergence analysis, we require that the level sets $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$ are compact, cf. Assumption B. The compactness of the level sets is ensured if $\|\mathbf{F}(\cdot)\|_p$ is coercive. In the case of a quadratic discrepancy term $g\colon \mathbb{R}^n \to \mathbb{R}$,

$$g(\mathbf{u}) := \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \mathbf{u}^\top \mathbf{b} + c, \tag{2.43}$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetric and nonsingular, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$ we will in the following show that $\|\mathbf{F}(\cdot)\|_p$, $p \in \{1, 2\}$ is coercive with respect to the Euclidean norm, i.e., $\lim_{\|\mathbf{u}\|_2\to\infty} \|\mathbf{F}(\mathbf{u})\|_p = \infty$. As a special case the claim follows for the discrepancy term

$$g(\mathbf{u}) := \frac{1}{2}\|\mathbf{K}\mathbf{u} - \mathbf{f}\|_2^2, \tag{2.44}$$

with $\mathbf{K} \in \mathbb{R}^{m\times n}$ injective and $\mathbf{A} = \mathbf{K}^\top\mathbf{K}$, $\mathbf{b} = -\mathbf{K}^\top\mathbf{f}$ and $c = \frac{1}{2}\|\mathbf{f}\|_2^2$. For the proof, we need the following estimates for soft-thresholding in finite-dimensional spaces.

Lemma 2.20 *Let* $\mathbf{0} \leq \mathbf{w} \in \mathbb{R}^n$*. Then for* $p \in [1, \infty]$ *with* $\frac{1}{\infty} := 0$ *it follows that*

$$\|\mathbf{u} - \mathbf{S}_\mathbf{w}(\mathbf{u})\|_p \leq n^{\frac{1}{p}}\|\mathbf{w}\|_\infty, \quad \textit{for all } \mathbf{u} \in \mathbb{R}^n. \tag{2.45}$$

Proof. For each $\mathbf{u} \in \mathbb{R}^n$, (2.45) for $1 \leq p < \infty$ follows from

$$\|\mathbf{u} - \mathbf{S}_\mathbf{w}(\mathbf{u})\|_p^p = \sum_{|u_k|\leq w_k} |u_k|^p + \sum_{|u_k|>w_k} w_k^p \leq \sum_{k=1}^n w_k^p \leq n\|\mathbf{w}\|_\infty^p.$$

For $p = \infty$, we have

$$\begin{aligned}\|\mathbf{u} - \mathbf{S}_\mathbf{w}(\mathbf{u})\|_\infty &= \max_{1\leq i\leq n} |(\mathbf{u} - \mathbf{S}_\mathbf{w}(\mathbf{u}))_i| = \max\Big\{ \max_{\{k:|u_k|\leq w_k\}} |u_k|, \max_{\{k:|u_k|>w_k\}} w_k \Big\} \\ &\leq \max_k w_k = \|\mathbf{w}\|_\infty.\end{aligned}$$

□

In case of a quadratic discrepancy term (2.43), the coercivity of $\|\mathbf{F}(\cdot)\|_p$ with respect to the Euclidean norm for each $\gamma > 0$ immediately follows from the following lemma, because $\mathbf{I} - \gamma\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric and 1 is not an eigenvalue of this matrix because $\mathbf{A}$ is nonsingular.

Lemma 2.21 *Let $p \in \{1,2\}$, $\mathbf{0} \leq \mathbf{w} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric, and assume that 1 is not an eigenvalue of $\mathbf{B}$. Then there exist $\alpha > 0$ and $\beta \in \mathbb{R}$ with*

$$\big\|\mathbf{u} - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u} + \mathbf{c})\big\|_p \geq \alpha\|\mathbf{u}\|_2 + \beta, \quad \textit{for all } \mathbf{u} \in \mathbb{R}^n. \tag{2.46}$$

Proof. By assumption on the spectral properties of $\mathbf{B}$, there exist $\mathbf{B}$-invariant subspaces U_1, U_2 and constants $0 < \rho < 1 < \eta$ with

$$\|\mathbf{B}\mathbf{u}_1\|_2 \leq \rho\|\mathbf{u}_1\|_2, \quad \text{for all } \mathbf{u}_1 \in U_1,$$

and

$$\|\mathbf{B}\mathbf{u}_2\|_2 \geq \eta\|\mathbf{u}_2\|_2, \quad \text{for all } \mathbf{u}_2 \in U_2,$$

such that the splitting $\mathbb{R}^n = U_1 \oplus U_2$ is orthogonal. For an arbitrary $\mathbf{u} \in \mathbb{R}^n$ and its orthogonal projections of $\mathbf{u}$ onto $U_j, 1 \leq j \leq 2$, a double application of Pythagoras' theorem implies that

$$\big\|(\mathbf{I} - \mathbf{B})\mathbf{u}\big\|_2 = \sqrt{\big\|(\mathbf{I} - \mathbf{B})\mathbf{u}_1\big\|_2^2 + \big\|(\mathbf{I} - \mathbf{B})\mathbf{u}_2\big\|_2^2} \geq \min\{1-\rho, \eta - 1\}\|\mathbf{u}\|_2. \tag{2.47}$$

By combining (2.47), (2.45) and the nonexpansivity of the soft shrinkage operator $\mathbf{S}_{\mathbf{w}}$, we obtain that

$$\begin{aligned}\big\|\mathbf{u} - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u} + \mathbf{c})\big\|_2 &\geq \big\|(\mathbf{I} - \mathbf{B})(\mathbf{u})\big\|_2 - \big\|\mathbf{B}\mathbf{u} - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u})\big\|_2 - \big\|\mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u}) - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u} + \mathbf{c})\big\|_2 \\ &\geq \min\{1-\rho, \eta - 1\}\|\mathbf{u}\|_2 - \sqrt{n}\|\mathbf{w}\|_\infty - \|\mathbf{c}\|_2,\end{aligned}$$

proving the claim for $p = 2$. The claim for $p = 1$ follows from

$$\big\|\mathbf{u} - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u} + \mathbf{c})\big\|_1 \geq \big\|\mathbf{u} - \mathbf{S}_{\mathbf{w}}(\mathbf{B}\mathbf{u} + \mathbf{c})\big\|_2.$$

□

We proceed by verifying the compactness of the level sets of $\|\mathbf{F}(\cdot)\|_p$ in case of a quadratic discrepancy term (2.43).

Proposition 2.22 *Let $\mathbf{u}^{(0)} \in \mathbb{R}^n$ be an arbitrary starting vector of Algorithm 1 and let g be a quadratic discrepancy term* (2.43) *with $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetric and nonsingular. Then the level set $\mathcal{L}(\mathbf{u}^{(0)}) := \{\mathbf{u} : \|\mathbf{F}(\mathbf{u})\|_p \leq \|\mathbf{F}(\mathbf{u}^{(0)})\|_p\}$, $p \in \{1,2\}$ is compact.*

Proof. This follows immediately from the coercivity (2.46) and the continuity of the function $\|\mathbf{F}(\cdot)\|_p$. □

In what follows, we will return to arbitrary discrepancy terms g. In the following proposition, we will show that the B-semismooth Newton directions from (2.37) are bounded with respect to the current residual norms.

Proposition 2.23 *Let* $\mathbf{u} \in \mathbb{R}^n, p \in \{1,2\}$, *and let* $\mathbf{d} = \mathbf{d}(\mathbf{u})$ *be chosen according to* (2.37). *There exists a constant* $C > 0$, *independent of* $\mathbf{u}$, *with*

$$\|\mathbf{d}\|_p \leq C\big\|\mathbf{F}(\mathbf{u})\big\|_p.$$

Proof. We proceed in a similar way as in the proof of (2.40). Because of the equivalence of the norms in $\mathbb{R}^n$, it suffices to consider the case $p = 2$. The claim immediately follows from the bounded invertibility of $\mathbf{G}(\mathbf{u})^{-1}$ in the level set $\mathcal{L}_{\Theta_2}(\mathbf{u}^{(0)})$. Using [58, Proposition 3.11] and [103, Lemma 3.6], we have

$$\begin{aligned}
&\|\mathbf{d}\|_2 \leq \|\mathbf{G}(\mathbf{u})^{-1}\|_2 \|\mathbf{F}(\mathbf{u})\|_2 \\
&\leq \Big(\Big(\frac{1}{\gamma} + \sup_{\mathbf{u}\in\mathcal{L}_{\Theta_2}(\mathbf{u}^{(0)})} \|\nabla^2 g(\mathbf{u})\|_2\Big) \max_{\varnothing\neq\mathcal{J}\subset\{1,\dots,n\}} \sup_{\mathbf{u}\in\mathcal{L}_{\Theta_2}(\mathbf{u}^{(0)})} \|(\nabla^2 g(\mathbf{u})_{\mathcal{J},\mathcal{J}})^{-1}\|_2 + 1\Big)\|\mathbf{F}(\mathbf{u})\|_2 \\
&\leq \Big(\Big(\frac{1}{\gamma} + c_2\Big)\frac{1}{c_1} + 1\Big)\|\mathbf{F}(\mathbf{u})\|_2
\end{aligned}$$

with $c_1, c_2 > 0$ from Assumption B, finishing the proof. □

In our convergence proof, we shall need the continuity of Θ_p'.

Lemma 2.24 *Let* $(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) \to (\mathbf{u}^*, \overline{\mathbf{d}})$, $j \to \infty$, $\Theta := \Theta_p$, $p \in \{1,2\}$. *If* Θ *from* (2.2) *fulfills the condition*

$$\lim_{(\mathbf{u},\mathbf{v})\to(\mathbf{u}^*,\mathbf{u}^*)} \frac{\Theta(\mathbf{u}) - \Theta(\mathbf{v}) - \Theta'(\mathbf{u}^*, \mathbf{u}-\mathbf{v})}{\|\mathbf{u}-\mathbf{v}\|_p} = 0, \tag{2.48}$$

then the directional derivative $\Theta'(\mathbf{u}, \mathbf{d})$ *is continuous at* $(\mathbf{u}^*, \overline{\mathbf{d}})$, *as a function of* $(\mathbf{u}, \mathbf{d})$, *i.e.,*

$$\lim_{j\to\infty} \Theta'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = \Theta'(\mathbf{u}^*, \overline{\mathbf{d}}).$$

Proof. Since Θ from (2.2) is locally Lipschitz continuous and directionally differentiable in the level set $\mathcal{L}_\Theta(\mathbf{u}^{(0)})$ from Assumption B, it follows for $\mathbf{u} \in \mathcal{L}_\Theta(\mathbf{u}^{(0)})$ that $\Theta'(\mathbf{u}, \cdot)$ is a globally Lipschitz continuous function with the same Lipschitz constant as Θ, see Lemma 1.19 or [124, Theorem 3.1.2], i.e.,

$$\lim_{j\to\infty} \Theta'(\mathbf{u}^*, \mathbf{d}^{(j)}) = \Theta'(\mathbf{u}^*, \overline{\mathbf{d}}).$$

Because of condition (2.48) of Θ at $\mathbf{u}^*$, the double limit

$$\lim_{j\to\infty, t\searrow 0} \frac{\Theta(\mathbf{u}^{(j)} + t\mathbf{d}^{(j)}) - \Theta(\mathbf{u}^{(j)})}{t}$$

exists and is equal to $\Theta'(\mathbf{u}^*, \overline{\mathbf{d}})$. Additionally, one has

$$\lim_{t\searrow 0} \frac{\Theta(\mathbf{u}^{(j)} + t\mathbf{d}^{(j)}) - \Theta(\mathbf{u}^{(j)})}{t} = \Theta'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)})$$

for every fixed $j \in \mathbb{N}$. Therefore, we conclude that

$$\begin{aligned}\lim_{j\to\infty} \Theta'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) &= \lim_{j\to\infty} \left(\lim_{t\searrow 0} \frac{\Theta(\mathbf{u}^{(j)} + t\mathbf{d}^{(j)}) - \Theta(\mathbf{u}^{(j)})}{t} \right) \\ &= \lim_{j\to\infty, t\searrow 0} \frac{\Theta(\mathbf{u}^{(j)} + t\mathbf{d}^{(j)}) - \Theta(\mathbf{u}^{(j)})}{t} \\ &= \Theta'(\mathbf{u}^*, \overline{\mathbf{d}}).\end{aligned}$$

□

We may now prove our main result on the global convergence of Algorithm 1, cf. [63, Theorem 1].

Theorem 2.25 *Consider the damped B-semismooth Newton method for the finite-dimensional shrinkage equation* (1.10)*. Let* $\{\mathbf{u}^{(j)}\}_j$ *be a sequence of iterates produced by Algorithm 1,* $\{t_j\}_j$ *the chosen step sizes and* $p \in \{1, 2\}$, $\Theta := \Theta_p$.

(i) If $\limsup_{j\to\infty} t_j > 0$, *then* $\mathbf{u}^{(j)} \to \mathbf{u}^*$, $j \to \infty$ *with* $\Theta(\mathbf{u}^*) = 0$.

(ii) If $\limsup_{j\to\infty} t_j = 0$ *and if* $\mathbf{u}^*$ *is an accumulation point of* $\{\mathbf{u}^{(j)}\}_j$ *where condition* (2.48) *holds at* $\mathbf{u}^*$, *then* $\mathbf{u}^{(j)} \to \mathbf{u}^*$, $j \to \infty$ *with* $\Theta(\mathbf{u}^*) = 0$.

Proof. We proceed as in [63, 77, 107, 111]. The B-Newton equation (1.18) has a unique solution in each step, according to Theorem 2.13. Moreover, Proposition 2.18 ensures that the Armijo rule outputs well-defined step sizes t_j. Due to the boundedness of the level set $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$, cf. Assumption B and Proposition 2.22 in case of a quadratic discrepancy term g, respectively, the sequence of iterates $\{\mathbf{u}^{(j)}\}_j$ is bounded and has an accumulation point $\mathbf{u}^*$. If t_j is chosen equal to 1 infinitely many times, then it follows with $1 - p\sigma < 1$ that $\mathbf{u}^{(j)} \to \mathbf{u}^*$, $j \to \infty$, with $\Theta(\mathbf{u}^*) = 0$.
Otherwise, $\{\Theta(\mathbf{u}^{(j)})\}_j$ is still strictly decreasing, due to the Armijo rule, and obviously bounded below by 0. Therefore, the sequence $\{\Theta(\mathbf{u}^{(j)})\}_j$ is convergent and the sequence of differences $\{\Theta(\mathbf{u}^{(j+1)}) - \Theta(\mathbf{u}^{(j)})\}_j$ tends to 0. The Armijo rule (2.42) implies

$$t_j \Theta(\mathbf{u}^{(j)}) \leq \frac{\Theta(\mathbf{u}^{(j)}) - \Theta(\mathbf{u}^{(j+1)})}{p\sigma} \to 0, \quad j \to \infty.$$

If the chosen step sizes are bounded away from 0, i.e., $\limsup_{j\to\infty} t_j > 0$, it follows that $\mathbf{u}^{(j)} \to \mathbf{u}^*$, $j \to \infty$ with $\Theta(\mathbf{u}^*) = 0$, proving (i).
Suppose now that $\limsup_{j\to\infty} t_j = 0$, which implies that $\lim_{j\to\infty} t_j = 0$. Let $\{\Theta(\mathbf{u}^{(j)})\}_{j\in J}$ be a subsequence converging to $\mathbf{u}^*$ with $\Theta(\mathbf{u}^*) > 0$. Let $t_j = \beta^{l_j}$ denote the chosen step sizes in Algorithm 1. We define $\tau_j := \beta^{l_j - 1}$. The Armijo rule (2.42) yields

$$\frac{\Theta(\mathbf{u}^{(j)} + \tau_j \mathbf{d}^{(j)}) - \Theta(\mathbf{u}^{(j)})}{\tau_j} > -p\sigma\Theta(\mathbf{u}^{(j)}), \quad j = 0, 1, \ldots \tag{2.49}$$

According to Proposition 2.23, the sequence $\{\mathbf{d}^{(j)}\}_j$ is bounded. Without loss of generality we may suppose that $\mathbf{d}^{(j)} \to \overline{\mathbf{d}}$, $j \to \infty$, $j \in J$. Passing to the limit in (2.49), assumption (2.48) of Θ at $\mathbf{u}^*$ yields with Lemma 2.24

$$\Theta'(\mathbf{u}^*, \overline{\mathbf{d}}) \geq -p\sigma\Theta(\mathbf{u}^*).$$

Due to (2.48), the directional derivative Θ' as a function of $(\mathbf{u}, \mathbf{v})$ is continuous at $(\mathbf{u}^*, \overline{\mathbf{d}})$, see Lemma 2.24. Therefore,

$$\Theta'(\mathbf{u}^*, \overline{\mathbf{d}}) = \lim_{j\to\infty, j\in J} \Theta'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = \lim_{j\to\infty, j\in J} -p\Theta(\mathbf{u}^{(j)}) = -p\Theta(\mathbf{u}^*).$$

Altogether, we obtain

$$-p\Theta(\mathbf{u}^*) \geq -p\sigma\Theta(\mathbf{u}^*),$$

which is a contradiction to $\sigma < \frac{1}{p} \leq 1$ and hence finishes the proof. □

Remark 2.26 *Assumption* (2.48) *on Θ is only needed in case (ii) of Theorem 2.25. Pang proved in [108, Proposition 1] that if the merit function Θ_2 has no strong Fréchet derivative at the accumulation point $\mathbf{u}^*$, which is a sufficient condition for* (2.48) *according to [107, Theorem 2], then $\mathbf{u}^*$ is not a zero of Θ_2. Therefore, this assumption in Theorem 2.25 is necessary, see also [63, Corollary 1].*

However, Assumption (2.48) *is not always fulfilled. For instance, if we have a quadratic discrepancy term $g\colon \mathbb{R} \to \mathbb{R}$, $g(u) = \frac{1}{2}(Ku - f)^2$, $Kf = w$ and $\gamma K^2 < 1$ in the one-dimensional case with $p = 1$, the zero of F will be given by $u^* = 0$. For the two sequences $u_n = n^{-1}$ and $v_n = -n^{-1}$, the limit in* (2.48) *is equal to $-\frac{1+\gamma K^2}{2} \neq 0$, so that* (2.48) *does not hold at the zero u^* of F.*

Concerning the speed of convergence of our B-semismooth Newton scheme, we cite the following corollary from [111, Theorem 4.3, Corollary 4.4].

Corollary 2.27 *Let $\{\mathbf{u}^{(j)}\}_j$ be any sequence generated by Algorithm 1 with $p = 2$ and $\mathbf{F}(\mathbf{u}^{(j)}) \neq \mathbf{0}$ for all j. Let $\mathbf{u}^*$ be an accumulation point of $\{\mathbf{u}^{(j)}\}$. Then, on the one hand, $\mathbf{u}^*$ is the zero of $\mathbf{F}$ if and only if the sequence of iterates $\{\mathbf{u}^{(j)}\}$ converges to $\mathbf{u}^*$ locally superlinearly and there exists an index j_0 with $t_j = 1$ for all $j \geq j_0$. On the other hand, $\mathbf{F}(\mathbf{u}^*) \neq \mathbf{0}$ if and only if either the sequence of iterates $\{\mathbf{u}^{(j)}\}_j$ diverges or the step sizes t_j tend to zero.*

Remark 2.28 *There are some issues left to prove the global convergence of Algorithm 1 in the infinite-dimensional setting* (1.4). *The uniform boundedness of the index sets $\mathcal{A}(\mathbf{u}) \cup \mathcal{I}^{\pm}(\mathbf{u}) \subset \{1, \ldots, \kappa\}$, cf. the proof of Corollary 2.16, can be expected to hold in the vicinity of the zero $\mathbf{u}^*$ of $\mathbf{F}$, but not in the level set $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)}) = \{\mathbf{u} \in \ell_2 : \Theta_p(\mathbf{u}) \leq \Theta_p(\mathbf{u}^{(0)})\}$. Therefore, it is not ensured that the inverse $\mathbf{G}(\mathbf{u})^{-1}$ stays uniformly bounded in the level set $\mathcal{L}_{\Theta_p}(\mathbf{u}^{(0)})$ and that the sequence $\{\mathbf{d}^{(j)}\}_j$ of Newton directions stays bounded. Recently, Gerdts, Horn and Kimmerle [54] have proven a result similar to Theorem 2.25 in an infinite-dimensional setting assuming that there exists a positive constant C with $\|\mathbf{G}(\mathbf{u}^{(j)})^{-1}\|_{L(\ell_2)} \leq C$ for all j. If this assumption is fulfilled for the sequence of iterates, the results of loc. cit. can be applied in our setting.*

Chapter 3

A modified and a hybrid B-semismooth Newton method

In this chapter, we introduce a modified B-semismooth Newton method that is globally convergent without any additional assumption on an a priori unknown accumulation point of the sequence of iterates like Condition (2.48). In the following, we only consider the finite-dimensional setting (2.1), i.e. $\mathcal{N} = \{1, \dots, n\}$, and the merit functional $\Theta = \Theta_2$ from (2.2), i.e., $\Theta\colon \mathbb{R}^n \to \mathbb{R}$,

$$\Theta(\mathbf{u}) := \|\mathbf{F}(\mathbf{u})\|_2^2, \tag{3.1}$$

and we require Assumption B to hold. Chapter 3 is, up to minor changes, a part of our preprint [67].

3.1 Algorithms and their feasibility

In this section, we present a modified version of the global B-semismooth Newton method from Section 2.3 and discuss its feasibility. Additionally, we suggest a hybrid method.

3.1.1 A modified B-semismooth Newton method

In the following, we introduce a modified B-semismooth Newton method for the numerical solution of (2.1). Below, the active and inactive sets from (2.12)–(2.16) are represented in a slightly different way. We recall the *active set* $\mathcal{A}(\mathbf{u}) := \mathcal{A}^+(\mathbf{u}) \cup \mathcal{A}^-(\mathbf{u})$, where

$$\mathcal{A}^+(\mathbf{u}) := \{k : \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k < u_k\}, \tag{3.2}$$
$$\mathcal{A}^-(\mathbf{u}) := \{k : u_k < \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k\}, \tag{3.3}$$

and the *inactive set* $\mathcal{I}(\mathbf{u}) := \mathcal{I}^\circ(\mathbf{u}) \cup \mathcal{I}^+(\mathbf{u}) \cup \mathcal{I}^-(\mathbf{u})$, where

$$\mathcal{I}^\circ(\mathbf{u}) := \{k : \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k < u_k < \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k\}, \tag{3.4}$$
$$\mathcal{I}^+(\mathbf{u}) := \{k : u_k = \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k\}, \tag{3.5}$$
$$\mathcal{I}^-(\mathbf{u}) := \{k : u_k = \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k\}. \tag{3.6}$$

Below, we drop the argument $\mathbf{u}$ if there is no risk of confusion.

For $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$ defined by (1.10), we have

$$F_k(\mathbf{u}) = \min\{\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k, u_k\}, \qquad k \in \mathcal{A}^+(\mathbf{u}) \cup \mathcal{I}^\circ(\mathbf{u}) \cup \mathcal{I}^+(\mathbf{u}), \tag{3.7}$$
$$F_k(\mathbf{u}) = \max\{\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k, u_k\}, \qquad k \in \mathcal{A}^-(\mathbf{u}) \cup \mathcal{I}^\circ(\mathbf{u}) \cup \mathcal{I}^-(\mathbf{u}). \tag{3.8}$$

By Assumption B, $\mathbf{F}$ is Lipschitz continuous and directionally differentiable. The directional derivative of $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$ was identified in Chapter 2. The directional derivative of $\mathbf{F}$ at $\mathbf{u} \in \mathbb{R}^n$ in the direction $\mathbf{d} \in \mathbb{R}^n$ is given elementwise by

$$F_k'(\mathbf{u},\mathbf{d}) = \begin{cases} \gamma(\nabla^2 g(\mathbf{u})\mathbf{d})_k, & k \in \mathcal{A}(\mathbf{u}), \\ d_k, & k \in \mathcal{I}^\circ(\mathbf{u}), \\ \min\{\gamma(\nabla^2 g(\mathbf{u})\mathbf{d})_k, d_k\}, & k \in \mathcal{I}^+(\mathbf{u}), \\ \max\{\gamma(\nabla^2 g(\mathbf{u})\mathbf{d})_k, d_k\}, & k \in \mathcal{I}^-(\mathbf{u}), \end{cases} \tag{3.9}$$

cf. Corollary 2.6. The directional derivative of the merit functional Θ from (3.1) at $\mathbf{u} \in \mathbb{R}^n$ in the direction $\mathbf{d} \in \mathbb{R}^n$ is given by $\Theta'(\mathbf{u},\mathbf{d}) = 2\langle \mathbf{F}'(\mathbf{u},\mathbf{d}), \mathbf{F}(\mathbf{u})\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product, cf. (2.25).

To introduce the modified semismooth Newton method, we define the subsets

$$\mathcal{A}_+^+(\mathbf{u}) := \{k : \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k < u_k < 0\}, \tag{3.10}$$
$$\mathcal{A}_-^-(\mathbf{u}) := \{k : 0 < u_k < \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k\}, \tag{3.11}$$
$$\mathcal{I}_+^\circ(\mathbf{u}) := \{k : \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k < u_k < \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k < 0\}, \tag{3.12}$$
$$\mathcal{I}_-^\circ(\mathbf{u}) := \{k : 0 < \gamma(\nabla g(\mathbf{u}))_k - \gamma w_k < u_k < \gamma(\nabla g(\mathbf{u}))_k + \gamma w_k\}. \tag{3.13}$$

Inspired by [108], we define the modified index sets

$$\overline{\mathcal{A}^+}(\mathbf{u}) := \mathcal{A}^+(\mathbf{u}) \setminus \mathcal{A}_+^+(\mathbf{u}), \tag{3.14}$$
$$\overline{\mathcal{A}^-}(\mathbf{u}) := \mathcal{A}^-(\mathbf{u}) \setminus \mathcal{A}_-^-(\mathbf{u}), \tag{3.15}$$
$$\overline{\mathcal{I}^\circ}(\mathbf{u}) := \mathcal{I}^\circ(\mathbf{u}) \setminus \big(\mathcal{I}_+^\circ(\mathbf{u}) \cup \mathcal{I}_-^\circ(\mathbf{u})\big), \tag{3.16}$$
$$\overline{\mathcal{I}^+}(\mathbf{u}) := \mathcal{I}^+(\mathbf{u}) \cup \mathcal{A}_+^+(\mathbf{u}) \cup \mathcal{I}_+^\circ(\mathbf{u}), \tag{3.17}$$
$$\overline{\mathcal{I}^-}(\mathbf{u}) := \mathcal{I}^-(\mathbf{u}) \cup \mathcal{A}_-^-(\mathbf{u}) \cup \mathcal{I}_-^\circ(\mathbf{u}). \tag{3.18}$$

We denote $\overline{\mathcal{A}}(\mathbf{u}) := \overline{\mathcal{A}^+}(\mathbf{u}) \cup \overline{\mathcal{A}^-}(\mathbf{u})$ and $\overline{\mathcal{I}}(\mathbf{u}) := \overline{\mathcal{I}^\circ}(\mathbf{u}) \cup \overline{\mathcal{I}^+}(\mathbf{u}) \cup \overline{\mathcal{I}^-}(\mathbf{u})$ respectively. The subsets (3.10)–(3.13) fulfill $\mathcal{A}_+^+(\mathbf{u}) = \emptyset$, $\mathcal{A}_-^-(\mathbf{u}) = \emptyset$, $\mathcal{I}_+^\circ(\mathbf{u}) = \emptyset$ and $\mathcal{I}_-^\circ(\mathbf{u}) = \emptyset$ if $\mathbf{F}(\mathbf{u}) = \mathbf{0}$.

In the following lemma, we consider a linear complementarity problem which is important for all further discussions, cf. Section 2.2.2.

Lemma 3.1 *Let $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{M} := \nabla^2 g(\mathbf{u})$. The linear complementarity problem*

$$\mathbf{x}, \mathbf{y} \geq \mathbf{0}, \quad \mathbf{y} = \mathbf{N}\mathbf{x} + \mathbf{z}, \quad \langle \mathbf{x}, \mathbf{y}\rangle = 0, \tag{3.19}$$

with

$$\begin{aligned} \mathbf{N} &= \mathbf{N}(\mathbf{u}) \\ &:= \gamma \begin{pmatrix} \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{I}^+}} - \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^+}} & \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^-}} - \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{I}^-}} \\ \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^+}} - \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{I}^+}} & \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{I}^-}} - \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^-}} \end{pmatrix} \end{aligned} \tag{3.20}$$

and

$$\begin{aligned}\mathbf{z} &= \mathbf{z}(\mathbf{u}) \\ &:= \begin{pmatrix} \gamma(\mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^\circ}} - \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{I}^\circ}})\mathbf{u}_{\overline{\mathcal{I}^\circ}} - \mathbf{M}_{\overline{\mathcal{I}^+},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{F}(\mathbf{u})_{\overline{\mathcal{A}}} + \mathbf{F}(\mathbf{u})_{\overline{\mathcal{I}^+}} \\ \gamma(\mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{I}^\circ}} - \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}^\circ}})\mathbf{u}_{\overline{\mathcal{I}^\circ}} + \mathbf{M}_{\overline{\mathcal{I}^-},\overline{\mathcal{A}}}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\mathbf{F}(\mathbf{u})_{\overline{\mathcal{A}}} - \mathbf{F}(\mathbf{u})_{\overline{\mathcal{I}^-}} \end{pmatrix} \\ &\quad - \mathbf{N}(\mathbf{u}) \begin{pmatrix} \mathbf{u}_{\overline{\mathcal{I}^+}} \\ -\mathbf{u}_{\overline{\mathcal{I}^-}} \end{pmatrix}\end{aligned} \tag{3.21}$$

has a unique solution.

Proof. By Assumption B, $\mathbf{M} = \nabla^2 g(\mathbf{u})$ is symmetric and positive definite. Therefore, $\mathbf{N}$ from (3.20) is symmetric and positive definite, see Lemma 2.10. Hence (3.19) is uniquely solvable, see [24, Theorem 3.3.7] and Theorem 2.13. □

Remark 3.2 *Note that the matrix* $\mathbf{N}$ *from* (3.20) *and the vector* $\mathbf{z}$ *from* (3.21) *differ from* (2.27) *and* (2.33)*, respectively, only in the choice of the index sets.*

Now we can define the generalized Newton equation for $\mathbf{F}$, cf. Remark 2.14. Let $\mathbf{u} \in \mathbb{R}^n$ and

$$\begin{aligned}\overline{\mathcal{B}} = \overline{\mathcal{B}}(\mathbf{u}) &:= \overline{\mathcal{A}}(\mathbf{u}) \cup \{k \in \overline{\mathcal{I}^+}(\mathbf{u}) \cup \overline{\mathcal{I}^-}(\mathbf{u}) : x_k > 0\}, \\ \overline{\mathcal{C}} = \overline{\mathcal{C}}(\mathbf{u}) &:= \overline{\mathcal{I}^\circ}(\mathbf{u}) \cup \{k \in \overline{\mathcal{I}^+}(\mathbf{u}) \cup \overline{\mathcal{I}^-}(\mathbf{u}) : x_k = 0\},\end{aligned} \tag{3.22}$$

where $\mathbf{x} = (x_k)_k$ is the unique solution to the linear complementarity problem (3.19). Then, by defining the generalized derivative blockwise

$$\begin{pmatrix} \mathbf{G}(\mathbf{u})_{\overline{\mathcal{B}},\overline{\mathcal{B}}} & \mathbf{G}(\mathbf{u})_{\overline{\mathcal{B}},\overline{\mathcal{C}}} \\ \mathbf{G}(\mathbf{u})_{\overline{\mathcal{C}},\overline{\mathcal{B}}} & \mathbf{G}(\mathbf{u})_{\overline{\mathcal{C}},\overline{\mathcal{C}}} \end{pmatrix} := \begin{pmatrix} \gamma(\nabla^2 g(\mathbf{u}))_{\overline{\mathcal{B}},\overline{\mathcal{B}}} & \gamma(\nabla^2 g(\mathbf{u}))_{\overline{\mathcal{B}},\overline{\mathcal{C}}} \\ \mathbf{0}_{\overline{\mathcal{C}},\overline{\mathcal{B}}} & \mathbf{I}_{\overline{\mathcal{C}},\overline{\mathcal{C}}} \end{pmatrix} \in \mathbb{R}^{n\times n}, \tag{3.23}$$

the modified B-semismooth Newton method is given by

$$\mathbf{G}(\mathbf{u}^{(j)})\mathbf{d}^{(j)} = -\mathbf{F}(\mathbf{u}^{(j)}), \tag{3.24}$$

$$\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}, \quad j = 0, 1, \ldots \tag{3.25}$$

with suitably chosen damping parameters $t_j > 0$.

Remark 3.3 *In [103], Muoi et al. chose the slanting function*

$$\begin{pmatrix} \mathbf{G}(\mathbf{u})_{\mathcal{A},\mathcal{A}} & \mathbf{G}(\mathbf{u})_{\mathcal{A},\mathcal{I}} \\ \mathbf{G}(\mathbf{u})_{\mathcal{I},\mathcal{A}} & \mathbf{G}(\mathbf{u})_{\mathcal{I},\mathcal{I}} \end{pmatrix} = \begin{pmatrix} \gamma(\nabla^2 g(\mathbf{u}))_{\mathcal{A},\mathcal{A}} & \gamma(\nabla^2 g(\mathbf{u}))_{\mathcal{A},\mathcal{I}} \\ \mathbf{0}_{\mathcal{I},\mathcal{A}} & \mathbf{I}_{\mathcal{I},\mathcal{I}} \end{pmatrix}, \tag{3.26}$$

blocked according to the active and inactive sets, to define the local semismooth Newton method (1.16)*. The key difference of* (3.23) *compared to* (3.26) *and to the slanting function* $\mathbf{G}$ *from* (2.38)*, respectively, is the modification of the index sets. Note that* $\mathbf{G}$ *from* (3.23) *is not a slanting function in general because in regions where* $\mathbf{F}$ *is smooth,* $\mathbf{G}$ *does not coincide with the Fréchet derivative of* $\mathbf{F}$*.*

Let $\mathbf{u}^{(j)} \in \mathbb{R}^n$ and $\mathbf{M} := \nabla^2 g(\mathbf{u}^{(j)})$. Then $\mathbf{d}^{(j)} \in \mathbb{R}^n$ solves (3.24) if and only if

$$\gamma(\mathbf{M}\mathbf{d}^{(j)})_{\overline{\mathcal{A}}} = -\mathbf{F}(\mathbf{u}^{(j)})_{\overline{\mathcal{A}}}, \tag{3.27}$$

$$\mathbf{d}^{(j)}_{\overline{\mathcal{I}^\circ}} = -\mathbf{u}^{(j)}_{\overline{\mathcal{I}^\circ}}, \tag{3.28}$$

and

$$\begin{aligned} \mathbf{x} &:= \begin{pmatrix} \mathbf{d}^{(j)}_{\overline{\mathcal{I}^+}} + \mathbf{u}^{(j)}_{\overline{\mathcal{I}^+}} \\ -\mathbf{d}^{(j)}_{\overline{\mathcal{I}^-}} - \mathbf{u}^{(j)}_{\overline{\mathcal{I}^-}} \end{pmatrix}, \\ \mathbf{y} &:= \mathbf{N}(\mathbf{u}^{(j)})\mathbf{x} + \mathbf{z}(\mathbf{u}^{(j)}) = \begin{pmatrix} \gamma(\mathbf{M}\mathbf{d}^{(j)})_{\overline{\mathcal{I}^+}} + (\mathbf{F}(\mathbf{u}^{(j)}))_{\overline{\mathcal{I}^+}} \\ -\gamma(\mathbf{M}\mathbf{d}^{(j)})_{\overline{\mathcal{I}^-}} - (\mathbf{F}(\mathbf{u}^{(j)}))_{\overline{\mathcal{I}^-}} \end{pmatrix}, \end{aligned} \tag{3.29}$$

where $\mathbf{x}$, $\mathbf{y}$ solve the linear complementarity problem (3.19), cf. Lemma 2.11.

We summarize the above observations in the following theorem, cf. Theorem 2.13.

Theorem 3.4 *Let* $\mathbf{x}$ *be the unique solution to* (3.19) *for an iterate* $\mathbf{u}^{(j)} \in \mathbb{R}^n$ *and* $\mathbf{M} := \nabla^2 g(\mathbf{u}^{(j)})$. *Then, the Newton update* $\mathbf{d}^{(j)}$ *from* (3.24) *is given by*

$$\begin{aligned} \mathbf{d}^{(j)}_{\overline{\mathcal{I}^\circ}} &= -\mathbf{u}^{(j)}_{\overline{\mathcal{I}^\circ}}, \\ \mathbf{d}^{(j)}_{\overline{\mathcal{I}^+}} &= \mathbf{x}_{\overline{\mathcal{I}^+}} - \mathbf{u}^{(j)}_{\overline{\mathcal{I}^+}}, \\ \mathbf{d}^{(j)}_{\overline{\mathcal{I}^-}} &= -\mathbf{x}_{\overline{\mathcal{I}^-}} - \mathbf{u}^{(j)}_{\overline{\mathcal{I}^-}}, \\ \mathbf{d}^{(j)}_{\overline{\mathcal{A}}} &= \frac{1}{\gamma}\mathbf{M}^{-1}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}\big(-\gamma\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{I}}}\mathbf{d}^{(j)}_{\overline{\mathcal{I}}} - \mathbf{F}(\mathbf{u}^{(j)})_{\overline{\mathcal{A}}}\big). \end{aligned} \tag{3.30}$$

Before proceeding, we prove some useful identities similar to [108, Lemma 2].

Lemma 3.5 *Let* $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{d} = \mathbf{d}(\mathbf{u})$ *the unique solution to* (3.30) *and* $\mathbf{M} = \nabla^2 g(\mathbf{u})$. *For* $k \in \{1, \dots, n\}$, *we have the following identities*

$$(\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k)\gamma(\mathbf{M}\mathbf{d})_k = -(\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k)^2, \qquad k \in \overline{\mathcal{A}^+}(\mathbf{u}), \tag{3.31}$$

$$(\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k)\gamma(\mathbf{M}\mathbf{d})_k = -(\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k)^2, \qquad k \in \overline{\mathcal{A}^-}(\mathbf{u}), \tag{3.32}$$

$$u_k d_k = -u_k^2, \qquad k \in \overline{\mathcal{I}^\circ}(\mathbf{u}). \tag{3.33}$$

Additionally, for $k \in \mathcal{A}^+_+(\mathbf{u}) \cup \mathcal{I}^\circ_+(\mathbf{u}) \cup \{k \in \mathcal{I}^+(\mathbf{u}) : F_k(\mathbf{u}) < 0\}$ *the inequality*

$$(\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k)\gamma(\mathbf{M}\mathbf{d})_k \le -(\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k)^2, \tag{3.34}$$

holds, for $k \in \mathcal{A}^-_-(\mathbf{u}) \cup \mathcal{I}^\circ_-(\mathbf{u}) \cup \{k \in \mathcal{I}^-(\mathbf{u}) : F_k(\mathbf{u}) > 0\}$ *we have*

$$(\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k)\gamma(\mathbf{M}\mathbf{d})_k \le -(\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k)^2, \tag{3.35}$$

and for $k \in \mathcal{A}^+_+(\mathbf{u}) \cup \mathcal{A}^-_-(\mathbf{u}) \cup \mathcal{I}^\circ_+(\mathbf{u}) \cup \mathcal{I}^\circ_-(\mathbf{u}) \cup \{k \in \mathcal{I}^+(\mathbf{u}) : F_k(\mathbf{u}) < 0\} \cup \{k \in \mathcal{I}^-(\mathbf{u}) : F_k(\mathbf{u}) > 0\}$ *we have*

$$u_k d_k \le -u_k^2. \tag{3.36}$$

Proof. Equations (3.31), (3.32) and (3.33) immediately follow from (3.27) and (3.28). For $k \in \mathcal{A}_+^+(\mathbf{u}) \cup \mathcal{I}_+^\circ(\mathbf{u}) \cup \{k \in \mathcal{I}^+(\mathbf{u}) : F_k(\mathbf{u}) < 0)\}$, we have by definition $\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k < 0$ and with (3.19) and (3.29) we have $\gamma(\mathbf{Md})_k \geq -F_k(\mathbf{u}) \geq -(\gamma(\nabla g(\mathbf{u}))_k + \gamma w_k)$ implying (3.34). For $k \in \mathcal{A}_-^-(\mathbf{u}) \cup \mathcal{I}_-^\circ(\mathbf{u}) \cup \{k \in \mathcal{I}^-(\mathbf{u}) : F_k(\mathbf{u}) > 0)\}$, we have by definition $\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k > 0$ and with (3.19) and (3.29) we have $\gamma(\mathbf{Md})_k \leq -F_k(\mathbf{u}) \leq -(\gamma(\nabla g(\mathbf{u}))_k - \gamma w_k)$, implying (3.35).

For $k \in \mathcal{A}_+^+(\mathbf{u}) \cup \mathcal{I}_+^\circ(\mathbf{u}) \cup \{k \in \mathcal{I}^+(\mathbf{u}) : F_k(\mathbf{u}) < 0\}$, we have $u_k < 0$ and $d_k \geq -u_k$ because of (3.19) and (3.29). If $k \in \mathcal{A}_-^-(\mathbf{u}) \cup \mathcal{I}_-^\circ(\mathbf{u}) \cup \{k \in \mathcal{I}^-(\mathbf{u}) : F_k(\mathbf{u}) > 0\}$, we have $u_k > 0$ and $d_k \leq -u_k$ because of (3.19) and (3.29). In both cases (3.36) follows. □

Now we verify that $\mathbf{d} = \mathbf{d}(\mathbf{u})$ from (3.30) is a descent direction of the merit functional Θ from (3.1) at $\mathbf{u}$.

Lemma 3.6 *Let $\mathbf{u} \in \mathbb{R}^n$ with $\Theta(\mathbf{u}) > 0$. Let $\mathbf{d} = \mathbf{d}(\mathbf{u}) \in \mathbb{R}^n$ be the solution to* (3.30). *Then, we have*

$$\Theta'(\mathbf{u}, \mathbf{d}) \leq -2\Theta(\mathbf{u}) < 0,$$

i.e., $\mathbf{d}$ is a descent direction of Θ at $\mathbf{u}$ in the direction $\mathbf{d}$.

Proof. The proof follows the idea of [108, Proof of Proposition 5]. We have

$$\Theta'(\mathbf{u}, \mathbf{d}) = 2\langle \mathbf{F}(\mathbf{u}), \mathbf{F}'(\mathbf{u}, \mathbf{d})\rangle = 2\sum_{i=1}^{8} T_i,$$

where we have with Equation (3.9) and $\mathbf{M} := \nabla^2 g(\mathbf{u})$,

$$\begin{aligned}
T_1 &:= \sum_{k \in \overline{\mathcal{A}}(\mathbf{u})} F_k(\mathbf{u})\gamma(\mathbf{Md})_k, & T_2 &:= \sum_{k \in \overline{\mathcal{I}^\circ}(\mathbf{u})} u_k d_k,\\
T_3 &:= \sum_{k \in \mathcal{I}^+(\mathbf{u})} F_k(\mathbf{u}) \min\{d_k, \gamma(\mathbf{Md})_k\}, & T_4 &:= \sum_{k \in \mathcal{I}^-(\mathbf{u})} F_k(\mathbf{u}) \max\{d_k, \gamma(\mathbf{Md})_k\},\\
T_5 &:= \sum_{k \in \mathcal{A}_+^+(\mathbf{u})} F_k(\mathbf{u})\gamma(\mathbf{Md})_k, & T_6 &:= \sum_{k \in \mathcal{A}_-^-(\mathbf{u})} F_k(\mathbf{u})\gamma(\mathbf{Md})_k,\\
T_7 &:= \sum_{k \in \mathcal{I}_+^\circ(\mathbf{u})} u_k d_k, & T_8 &:= \sum_{k \in \mathcal{I}_-^\circ(\mathbf{u})} u_k d_k.
\end{aligned}$$

For $k \in \mathcal{I}^+(\mathbf{u})$, we have

$$\min\{d_k, \gamma(\mathbf{Md})_k\} = \min\{d_k + F_k(\mathbf{u}), \gamma(\mathbf{Md})_k + F_k(\mathbf{u})\} - F_k(\mathbf{u}) = -F_k(\mathbf{u}),$$

because of (3.19) and (3.29). Similarly, for $k \in \mathcal{I}^-(\mathbf{u})$ we have

$$\max\{d_k, \gamma(\mathbf{Md})_k\} = \max\{d_k + F_k(\mathbf{u}), \gamma(\mathbf{Md})_k + F_k(\mathbf{u})\} - F_k(\mathbf{u}) = -F_k(\mathbf{u}).$$

With (3.31)–(3.36), we obtain

$$\Theta'(\mathbf{u}, \mathbf{d}) = 2\sum_{i=1}^{8} T_i \leq -2\sum_{k=1}^{n} F_k(\mathbf{u})^2 = -2\Theta(\mathbf{u}),$$

finishing the proof. □

Algorithm 2 The modified B-semismooth Newton method `modBSSN`

Choose a starting vector $\mathbf{u}^{(0)} \in \mathbb{R}^n$, parameters $\beta \in (0,1)$, $\sigma \in (0, \frac{1}{2})$, a tolerance $tol > 0$ and set $j := 0$.
while $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2 \geq tol$ **do**
  Compute the Newton direction $\mathbf{d}^{(j)}$ from (3.30).
  $t_j := 1$
  while $\Theta(\mathbf{u}^{(j)} + t_j\mathbf{d}^{(j)}) > (1 - 2\sigma t_j)\Theta(\mathbf{u}^{(j)})$ **do**
    $t_j := t_j\beta$
  end while
  $\mathbf{u}^{(j+1)} := \mathbf{u}^{(j)} + t_j\mathbf{d}^{(j)}$
  $j := j + 1$
end while

We choose the step sizes $t_j \in (0,1]$ in (3.25) by the well-known Armijo rule

$$t_j := \max\{\beta^l : \Theta(\mathbf{u}^{(j)} + \beta^l\mathbf{d}^{(j)}) \leq (1 - 2\sigma\beta^l)\Theta(\mathbf{u}^{(j)}), \quad l = 0, 1, \ldots\},$$

where $\beta \in (0,1)$ and $\sigma \in (0, \frac{1}{2})$, see also [63, 77, 107, 108, 111]. These step sizes can be computed in finitely many iterations, cf. Proposition 2.18.

Lemma 3.7 *Let $\beta \in (0,1)$, $\sigma \in (0, \frac{1}{2})$. Let $\mathbf{u}^{(j)} \in \mathbb{R}^n$ with $\Theta(\mathbf{u}^{(j)}) > 0$ and let $\mathbf{d}^{(j)} = \mathbf{d}(\mathbf{u}^{(j)})$ be computed by* (3.30)*. Then, there exists a finite index $l \in \mathbb{N}$ with*

$$\Theta(\mathbf{u}^{(j)} + \beta^l\mathbf{d}^{(j)}) \leq (1 - 2\sigma\beta^l)\Theta(\mathbf{u}^{(j)}). \tag{3.37}$$

Proof. According to Lemma 3.6, it holds $\Theta'(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) \leq -2\Theta(\mathbf{u}^{(j)}) < 0$. The remainder of the proof follows the proof of Proposition 2.18. □

The algorithm of the modified B-semismooth Newton method, in the following denoted by `modBSSN`, is stated in Algorithm 2. The feasibility of Algorithm `modBSSN` is guaranteed because of the lemmata stated above.

Remark 3.8 *Pang [108] introduced a modified B-Newton method for a nonlinear complementarity problem. Han, Pang and Rangaraj [63] interpreted this iteration as a generalized Newton method*

$$\mathbf{F}(\mathbf{u}^{(j)}) + \tilde{\mathbf{G}}(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = \mathbf{0}, \qquad \mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + t_j\mathbf{d}^{(j)}, \qquad j = 0, 1, \ldots,$$

where $\tilde{\mathbf{G}}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ fulfills the assumption that $\tilde{\mathbf{G}}(\mathbf{u}, \cdot)$ is surjective for each fixed $\mathbf{u} \in \mathbb{R}^n$, and

$$2\langle \mathbf{F}(\mathbf{u}), \tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d})\rangle \geq \Theta'(\mathbf{u}, \mathbf{d}) \tag{3.38}$$

for all $\mathbf{u}$, $\mathbf{d} \in \mathbb{R}^n$, see [63, Section 2.3]. As pointed out in [63], the surjectivity of $\tilde{\mathbf{G}}(\mathbf{u}, \cdot)$ ensures the solvability of the generalized Newton equation $\mathbf{F}(\mathbf{u}) + \tilde{\mathbf{G}}(\mathbf{u}, \mathbf{d}) = \mathbf{0}$ and (3.38) *implies descent of the resulting generalized Newton direction $\mathbf{d}$ w.r.t. the merit functional Θ. In the very same way, our Algorithm `modBSSN` can be interpreted as a generalized Newton method with $\tilde{\mathbf{G}}(\mathbf{u}^{(j)}, \mathbf{d}^{(j)}) = \mathbf{G}(\mathbf{u}^{(j)})\mathbf{d}^{(j)}$ and $\mathbf{G}$ from* (3.23)*, cf. Lemma 3.6.*

Algorithm 3 The hybrid B-semismooth Newton method `hybridBSSN`

Choose a starting vector $\mathbf{u}^{(0)} \in \mathbb{R}^n$, parameters $\beta \in (0,1)$, $\sigma \in (0, \frac{1}{2})$, a tolerance $tol > 0$, $j_{max} \in \mathbb{N}$ and $t_{min} > 0$ and set $j := 0$, $t_{-1} := 1$.
while $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2 \geq tol$ **do**
  if $j \leq j_{max}$ **and** $t_{j-1} \geq t_{min}$ **then**
    Compute the Newton direction $\mathbf{d}^{(j)}$ from (2.37).
  else
    Compute the Newton direction $\mathbf{d}^{(j)}$ from (3.30) and set $j_{max} := j$.
  end if
  $t_j := 1$
  while $\Theta(\mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}) > (1 - 2\sigma t_j)\Theta(\mathbf{u}^{(j)})$ **do**
    $t_j := t_j \beta$
  end while
  $\mathbf{u}^{(j+1)} := \mathbf{u}^{(j)} + t_j \mathbf{d}^{(j)}$
  $j := j + 1$
end while

Remark 3.9 *The B-semismooth Newton method* `BSSN` *from Chapter 2 is identical to Algorithm* `modBSSN` *replacing the modified index sets* (3.14)–(3.18) *by the original index sets* (3.2)–(3.6) *in* (3.19)–(3.21) *and* (3.30)*, cf. Algorithm 1 and Algorithm 2. The modification of the index sets in Algorithm* `modBSSN` *is needed to prove global convergence without any additional requirements, see Section 3.2. Let* $\mathbf{u}^*$ *be the unique zero of* $\mathbf{F}$ *and let* $\mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) = \emptyset$*, i.e.,* $\mathbf{F}$ *is smooth at* $\mathbf{u}^*$*. Then, there exists a neighborhood* U *of* $\mathbf{u}^*$ *where the index subsets* (3.10)–(3.13) *are empty for all* $\mathbf{u} \in U$*, i.e., the modified index sets* (3.14)–(3.18) *match the original index sets* (3.2)–(3.6)*. Therefore, Algorithm* `modBSSN` *locally coincides with Algorithm* `BSSN` *in a neighborhood of the zero* $\mathbf{u}^*$ *of* $\mathbf{F}$ *if* $\mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) = \emptyset$ *and hence is a semismooth Newton method there.*

3.1.2 A globally convergent hybrid method

The index sets $\mathcal{I}^{\pm}(\mathbf{u}^{(j)})$ in step j of the B-semismooth Newton method (`BSSN`) from Algorithm 1 are usually empty. Hence, the generalized Newton equation simplifies to a system of linear equations of the size $|\mathcal{A}(\mathbf{u}^{(j)})|$ and Algorithm 1 is efficient in practice. The size of the system of linear equations usually decreases along the iteration, cf. [58]. Nevertheless, the method may fail to converge, see Remark 3.9 and Theorem 2.25. However, the global convergence of Algorithm `modBSSN` from Algorithm 2 is ensured by Theorem 3.11, but here a linear complementarity problem and a system of linear equations have to be solved in each iteration, see (3.30). Additionally, in order to set up the matrix $\mathbf{N}$ and the vector $\mathbf{z}$ from (3.20) and (3.21), $|\overline{\mathcal{I}}(\mathbf{u}^{(j)})| + 1$ systems of linear equations of the size $|\overline{\mathcal{A}}(\mathbf{u}^{(j)}|$ with the same matrix have to be solved if $\overline{\mathcal{I}^{\pm}}(\mathbf{u}^{(j)}) \neq \emptyset$. Note that in (3.27) respectively (3.30), no additional system of linear equations has to be solved for the computation of $\mathbf{d}^{(j)}_{\mathcal{A}}$. Nevertheless, Algorithm `modBSSN` is usually less efficient than Algorithm `BSSN`.

We suggest a hybrid method by starting with Algorithm `BSSN` and switching to Algorithm `modBSSN` when Algorithm `BSSN` begins to stagnate, by replacing the index sets (3.2)–(3.6) by the modified index sets (3.14)–(3.18) in (3.19)–(3.21) and (3.30). In our numerical experiments, we switch to Algorithm `modBSSN` if the number of Newton steps exceeds a

limit $j_{max} \in \mathbb{N}$ or if the chosen step size is smaller than a threshold $t_{min} > 0$, i.e., if $j > j_{max}$ or $t_j < t_{min}$. In the sequel, this hybrid method is called `hybridBSSN`. The algorithm of the hybrid B-semismooth Newton method is given in Algorithm 3. Similar hybrid methods, combining the fast local convergence properties of a local semismooth Newton method with the globally convergent generalized Newton method from [63] were proposed by Qi [111] and Ito and Kunisch [77].

3.2 Global convergence and local convergence speed

In this section, we consider the convergence properties of the algorithms from Section 3.1.

3.2.1 Convergence of the modified B-semismooth Newton method

In the following, we address the global convergence of Algorithm `modBSSN` and its convergence speed in a neighborhood of the zero of $\mathbf{F}$. Concerning the boundedness of the sequence of Newton directions $\{\mathbf{d}^{(j)}\}_j$, we can rely on Proposition 2.23.

Lemma 3.10 *Let $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{d} = \mathbf{d}(\mathbf{u})$ be the solution to* (3.30)*. Then, there exists a constant $C > 0$ independent of $\mathbf{u}$, with*

$$\|\mathbf{d}\|_2 \leq C\|\mathbf{F}(\mathbf{u})\|_2.$$

Proof. The proof follows the proof of Proposition 2.23 by substituting the index sets $\mathcal{A}^{\pm}$, $\mathcal{I}^{\circ}$ and $\mathcal{I}^{\pm}$ by the modified index sets $\overline{\mathcal{A}^{\pm}}$, $\overline{\mathcal{I}^{\circ}}$ and $\overline{\mathcal{I}^{\pm}}$ respectively. □

In the following theorem, we present our main result on the global convergence of Algorithm `modBSSN`.

Theorem 3.11 *Let $\mathbf{u}^* \in \mathbb{R}^n$ be an accumulation point of the sequence of iterates $\{\mathbf{u}^{(j)}\}_j$ produced by Algorithm* `modBSSN`*. Then, we have $\Theta(\mathbf{u}^*) = 0$.*

Proof. We proceed analogously to the proof of [108, Theorem 1] and we also use the proof of [108, Proposition 1]. We suppose $\Theta(\mathbf{u}^{(j)}) > 0$ for all j, because otherwise the claim is proven. Because of the Armijo rule (3.37), the sequence $\{\Theta(\mathbf{u}^{(j)})\}_j$ strictly decreases and is bounded from below by 0, i.e., convergent. Let $t_j = \beta^{l_j}$ be the computed Armijo step size in step j. From the Armijo rule (3.37), it follows

$$0 < 2\sigma t_j \Theta(\mathbf{u}^{(j)}) \leq \Theta(\mathbf{u}^{(j)}) - \Theta(\mathbf{u}^{(j+1)}) \to 0, \quad j \to \infty.$$

Therefore, we have

$$\lim_{j\to\infty} t_j \Theta(\mathbf{u}^{(j)}) = 0.$$

The level set $\mathcal{L}_\Theta(\mathbf{u}^{(0)}) = \{\mathbf{u} \in \mathbb{R}^n : \Theta(\mathbf{u}) \leq \Theta(\mathbf{u}^{(0)})\}$ is bounded by Assumption B, implying that the sequence $\{\mathbf{u}^{(j)}\}_j$ is bounded and has an accumulation point $\mathbf{u}^*$. Let $\{\mathbf{u}^{(j)}\}_{j\in J}$ be a subsequence converging to $\mathbf{u}^*$. If the step sizes t_j are bounded away from zero, i.e., we have $\limsup_{j\to\infty, j\in J} t_j > 0$, it directly follows $\Theta(\mathbf{u}^*) = 0$, cf. the proof of Theorem 2.25.

Let us now consider the case $\limsup_{j\to\infty, j\in J} t_j = 0$. Without loss of generality, we suppose $\lim_{j\to\infty, j\in J} t_j = 0$. By the Armijo rule (3.37), we have for all $j \in J$

$$\Theta(\mathbf{u}^{(j)}) - \Theta(\mathbf{u}^{(j)} + \beta^{l_j-1}\mathbf{d}^{(j)}) < 2\sigma\beta^{l_j-1}\Theta(\mathbf{u}^{(j)}). \tag{3.39}$$

We define $\hat{\mathbf{u}}^{(j)} := \mathbf{u}^{(j)} + \beta^{l_j-1}\mathbf{d}^{(j)}$. The sequence $\{\mathbf{d}^{(j)}\}_j$ of Newton directions is bounded because of Lemma 3.10, implying that $\mathbf{u}^*$ is the limit of the subsequence $\{\hat{\mathbf{u}}^{(j)}\}_{j\in J}$. Therefore, without loss of generality we have

$$\begin{aligned}
\mathcal{A}_+^+(\mathbf{u}^*) &\subset \mathcal{A}_+^+(\mathbf{u}^{(j)}) \cap \mathcal{A}_+^+(\hat{\mathbf{u}}^{(j)}),\\
\mathcal{A}_-^-(\mathbf{u}^*) &\subset \mathcal{A}_-^-(\mathbf{u}^{(j)}) \cap \mathcal{A}_-^-(\hat{\mathbf{u}}^{(j)}),\\
\mathcal{I}_+^\circ(\mathbf{u}^*) &\subset \mathcal{I}_+^\circ(\mathbf{u}^{(j)}) \cap \mathcal{I}_+^\circ(\hat{\mathbf{u}}^{(j)}),\\
\mathcal{I}_-^\circ(\mathbf{u}^*) &\subset \mathcal{I}_-^\circ(\mathbf{u}^{(j)}) \cap \mathcal{I}_-^\circ(\hat{\mathbf{u}}^{(j)}),
\end{aligned}$$

for all $j \in J$ large enough. Now we consider

$$\Theta(\mathbf{u}^{(j)}) - \Theta(\hat{\mathbf{u}}^{(j)}) = \sum_{i=1}^{8} \tilde{T}_i, \tag{3.40}$$

where

$$\begin{aligned}
\tilde{T}_1 &:= \sum_{k\in\overline{\mathcal{A}}(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big), & \tilde{T}_2 &:= \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big),\\
\tilde{T}_3 &:= \sum_{k\in\mathcal{I}^+(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big), & \tilde{T}_4 &:= \sum_{k\in\mathcal{I}^-(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big),\\
\tilde{T}_5 &:= \sum_{k\in\mathcal{A}_+^+(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big), & \tilde{T}_6 &:= \sum_{k\in\mathcal{A}_-^-(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big),\\
\tilde{T}_7 &:= \sum_{k\in\mathcal{I}_+^\circ(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big), & \tilde{T}_8 &:= \sum_{k\in\mathcal{I}_-^\circ(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big).
\end{aligned}$$

In the following, we estimate each sum from below. Finally, we prove the claim by using (3.39) and by taking the limit $j \to \infty$, $j \in J$.

If $k \in \overline{\mathcal{A}}(\mathbf{u}^*)$, we have for $j \in J$ large enough that $k \in \overline{\mathcal{A}}(\mathbf{u}^{(j)})$, $k \in \mathcal{A}_+^+(\mathbf{u}^{(j)})$ or $k \in \mathcal{A}_-^-(\mathbf{u}^{(j)})$. Using (3.31), (3.32), (3.34) and (3.35), we obtain

$$\begin{aligned}
\tilde{T}_1 &= \sum_{k\in\overline{\mathcal{A}}(\mathbf{u}^*)} \big((\gamma(\nabla g(\mathbf{u}^{(j)}))_k \pm \gamma w_k)^2 - (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k \pm \gamma w_k)^2\big)\\
&= \sum_{k\in\overline{\mathcal{A}}(\mathbf{u}^*)} -2\beta^{l_j-1}(\gamma(\nabla g(\mathbf{u}^{(j)}))_k \pm \gamma w_k)\gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k\\
&\quad + o(\|\hat{\mathbf{u}}^{(j)} - \mathbf{u}^{(j)}\|_2)\\
&\geq 2\beta^{l_j-1} \sum_{k\in\overline{\mathcal{A}}(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j \to \infty, j \in J.
\end{aligned}$$

Analogously it follows with (3.34) and (3.35)

$$\tilde{T}_5 \geq 2\beta^{l_j-1} \sum_{k\in\mathcal{A}_+^+(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j \to \infty, j \in J,$$

and

$$\tilde{T}_6 \geq 2\beta^{l_j-1} \sum_{k\in\mathcal{A}^-(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j\to\infty, j\in J.$$

For $k \in \overline{I^\circ}(\mathbf{u}^*)$, we have to consider the cases $k \in \overline{\mathcal{I}^\circ}(\mathbf{u}^{(j)})$, $k \in \mathcal{I}^\circ_+(\mathbf{u}^{(j)})$ and $k \in \mathcal{I}^\circ_-(\mathbf{u}^{(j)})$. With (3.33) and (3.36), we have

$$\begin{aligned}\tilde{T}_2 &= \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} ((u_k^{(j)})^2 - (\hat{u}_k^{(j)})^2) = \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} \Big(-2\beta^{l_j-1}u_k^{(j)}d_k^{(j)} - (\beta^{l_j-1}d_k^{(j)})^2\Big)\\ &\geq 2\beta^{l_j-1} \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} (u_k^{(j)})^2 - \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} (\beta^{l_j-1}d_k^{(j)})^2\\ &= 2\beta^{l_j-1} \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 - \sum_{k\in\overline{\mathcal{I}^\circ}(\mathbf{u}^*)} (\beta^{l_j-1}d_k^{(j)})^2.\end{aligned}$$

Accordingly, it follows with (3.36)

$$\tilde{T}_7 \geq 2\beta^{l_j-1} \sum_{k\in\mathcal{I}^\circ_+(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 - \sum_{k\in\mathcal{I}^\circ_+(\mathbf{u}^*)} (\beta^{l_j-1}d_k^{(j)})^2$$

and

$$\tilde{T}_8 \geq 2\beta^{l_j-1} \sum_{k\in\mathcal{I}^\circ_-(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 - \sum_{k\in\mathcal{I}^\circ_-(\mathbf{u}^*)} (\beta^{l_j-1}d_k^{(j)})^2.$$

In the following, we treat the sum $\tilde{T}_3$. For $k \in \mathcal{I}^+(\mathbf{u}^*)$, we may assume without loss of generality

$$k \in \Big(\mathcal{I}^+(\mathbf{u}^{(j)}) \cup \mathcal{I}^\circ(\mathbf{u}^{(j)}) \cup \mathcal{A}^+(\mathbf{u}^{(j)})\Big) \cap \Big(\mathcal{I}^+(\hat{\mathbf{u}}^{(j)}) \cup \mathcal{I}^\circ(\hat{\mathbf{u}}^{(j)}) \cup \mathcal{A}^+(\hat{\mathbf{u}}^{(j)})\Big).$$

We split $\mathcal{I}^+(\mathbf{u}^*) = S_1(\mathbf{u}^*) \cup S_2(\mathbf{u}^*) \cup S_3(\mathbf{u}^*)$, where

$$\begin{aligned}S_1(\mathbf{u}^*) &:= \{k : u_k^* = \gamma(\nabla g(\mathbf{u}^*))_k + \gamma w_k > 0\},\\ S_2(\mathbf{u}^*) &:= \{k : u_k^* = \gamma(\nabla g(\mathbf{u}^*))_k + \gamma w_k = 0\},\\ S_3(\mathbf{u}^*) &:= \{k : u_k^* = \gamma(\nabla g(\mathbf{u}^*))_k + \gamma w_k < 0\}.\end{aligned}$$

For $k \in S_1(\mathbf{u}^*)$, we may assume with (3.7)

$$\begin{aligned}F_k(\mathbf{u}^{(j)}) &= \min\{u_k^{(j)}, \gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k\} > 0,\\ F_k(\hat{\mathbf{u}}^{(j)}) &= \min\{\hat{u}_k^{(j)}, \gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k\} > 0.\end{aligned}$$

Therefore, we have

$$\begin{aligned}F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 =& \min\{(u_k^{(j)})^2, (\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2\}\\ &- \min\{(\hat{u}_k^{(j)})^2, (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2\}.\end{aligned}$$

In the case $k \in \mathcal{I}^\circ(\mathbf{u}^{(j)})$, we have $k \in \overline{\mathcal{I}^\circ}(\mathbf{u}^{(j)})$ or $k \in \mathcal{I}^\circ_-(\mathbf{u}^{(j)})$ because $F_k(\mathbf{u}^{(j)}) > 0$. With (3.33) and (3.36), we have

$$\begin{aligned}F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 &\geq (u_k^{(j)})^2 - (\hat{u}_k^{(j)})^2 = -2\beta^{l_j-1}u_k^{(j)}d_k^{(j)} - (\beta^{l_j-1}d_k^{(j)})^2\\ &\geq 2\beta^{l_j-1}(u_k^{(j)})^2 - (\beta^{l_j-1}d_k^{(j)})^2.\end{aligned}$$

For $k \in \mathcal{A}^+(\mathbf{u}^{(j)})$, it follows $k \in \overline{\mathcal{A}^+}(\mathbf{u}^{(j)})$ because $F_k(\mathbf{u}^{(j)}) > 0$. Hence, one has with (3.31)

$$\begin{aligned}
&F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 \\
&\geq (\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2 - (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2 \\
&= -2\beta^{l_j-1}(\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)\gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k + o(\|\beta^{l_j-1}\mathbf{d}^{(j)}\|_2) \\
&= 2\beta^{l_j-1}F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j \to \infty, j \in J.
\end{aligned}$$

If $k \in \mathcal{I}^+(\mathbf{u}^{(j)})$, we have $F_k(\mathbf{u}^{(j)}) = u_k^{(j)} = \gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k$ and with (3.19) and (3.29) we have either $d_k^{(j)} = -u_k^{(j)}$ or $\gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k = -(\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)$. As in the cases $k \in \mathcal{I}^\circ(\mathbf{u}^{(j)})$ and $k \in \mathcal{A}^+(\mathbf{u}^{(j)})$, we conclude with (3.19) and (3.29)

$$F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 \geq 2\beta^{l_j-1}F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(k)}\|_2), \quad j \to \infty, j \in J.$$

Altogether, we get

$$\begin{aligned}
&\sum_{k \in S_1(\mathbf{u}^*)} \left(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\right) \\
&\geq 2\beta^{l_j-1} \sum_{k \in S_1(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j \to \infty, j \in J.
\end{aligned}$$

For $k \in S_2(\mathbf{u}^*)$, we have with the Lipschitz constant L of F_k

$$\begin{aligned}
F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 &= \left(F_k(\mathbf{u}^{(j)}) - F_k(\hat{\mathbf{u}}^{(j)})\right)\left(F_k(\mathbf{u}^{(j)}) + F_k(\hat{\mathbf{u}}^{(j)})\right) \\
&\leq L\|\hat{\mathbf{u}}^{(j)} - \mathbf{u}^{(j)}\|_2 \left|F_k(\mathbf{u}^{(j)}) + F_k(\hat{\mathbf{u}}^{(j)})\right| \\
&= L\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2 \left|F_k(\mathbf{u}^{(j)}) + F_k(\hat{\mathbf{u}}^{(j)})\right|.
\end{aligned}$$

It follows

$$\lim_{j \to \infty, j \in J} \sum_{k \in S_2(\mathbf{u}^*)} \frac{F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2}{\beta^{l_j-1}} = 0.$$

Let now $k \in S_3(\mathbf{u}^*)$. We may assume $u_k^{(j)} < 0$, $\hat{u}_k^{(j)} < 0$, $\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k < 0$ and $\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k < 0$. With (3.7), one has

$$\begin{aligned}
F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 = &\max\{(u_k^{(j)})^2, (\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2\} \\
&- \max\{(\hat{u}_k^{(j)})^2, (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2\}.
\end{aligned}$$

First, we treat the case $(\hat{u}_k^{(j)})^2 < (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2$. We have to consider the subcases $k \in \mathcal{I}_+^\circ(\mathbf{u}^{(j)})$, $k \in \mathcal{A}_+^+(\mathbf{u}^{(j)})$ and $k \in \{k \in \mathcal{I}^+(\mathbf{u}^{(j)}) : F_k(\mathbf{u}^{(j)}) < 0)\}$. With (3.34), we have

$$\begin{aligned}
&F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 \\
&\geq (\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2 - (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2 \\
&= -2\beta^{l_j-1}(\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)\gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2) \\
&\geq 2\beta^{l_j-1}(\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2 + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \quad j \to \infty, j \in J.
\end{aligned}$$

Second, we consider the case $(\hat{u}_k^{(j)})^2 \geq (\gamma(\nabla g(\hat{\mathbf{u}}^{(j)}))_k + \gamma w_k)^2$. With (3.36), we have analogously

$$\begin{aligned} F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2 &\geq (u_k^{(j)})^2 - (\hat{u}_k^{(j)})^2 = -2\beta^{l_j-1} u_k^{(j)} d_k^{(j)} - (\beta^{l_j-1} d_k^{(j)})^2 \\ &\geq 2\beta^{l_j-1}(u_k^{(j)})^2 - (\beta^{l_j-1} d_k^{(j)})^2. \end{aligned}$$

Altogether, we obtain

$$\begin{aligned} &\sum_{k \in S_3(\mathbf{u}^*)} \left(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\right) \\ &\geq 2\beta^{l_j-1} \sum_{k \in S_3(\mathbf{u}^*)} \min\{(u_k^{(j)})^2, (\gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k)^2\} \\ &\quad + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), j \to \infty, j \in J. \end{aligned}$$

By symmetry, we can treat the sum $\tilde{T}_4$ similarly. For $j \to \infty, j \in J$, we get

$$\begin{aligned} &\sum_{\{k \in \mathcal{I}^-(\mathbf{u}^*): u_k^* \neq 0\}} \left(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\right) \\ &\geq 2\beta^{l_j-1} \sum_{\{k \in \mathcal{I}^-(\mathbf{u}^*): u_k^* \neq 0\}} \min\{(u_k^{(j)})^2, (\gamma(\nabla g(\mathbf{u}^{(j)}))_k - \gamma w_k)^2\} \\ &\quad + o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2), \end{aligned}$$

and

$$\lim_{j \to \infty, j \in J} \sum_{\{k \in \mathcal{I}^+(\mathbf{u}^*): u_k^* = 0\}} \frac{F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2}{\beta^{l_j-1}} = 0.$$

Finally, we divide both sides of the inequality (3.39) by β^{l_j-1} and take the limit $j \to \infty$, $j \in J$, obtaining with (3.40) and the previous estimates

$$2\Theta(\mathbf{u}^*) \leq 2\sigma\Theta(\mathbf{u}^*).$$

Here, we have used the fact that the sequence $\{\mathbf{d}^{(j)}\}_j$ is bounded, implying

$$\lim_{j \to \infty, j \in J} \frac{o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2)}{\beta^{l_j-1}} = \lim_{j \to \infty, j \in J} \frac{o(\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2)}{\beta^{l_j-1}\|\mathbf{d}^{(j)}\|_2} \|\mathbf{d}^{(j)}\|_2 = 0.$$

The choice $\sigma < 1/2$ implies $\Theta(\mathbf{u}^*) = 0$, finishing the proof. □

As a consequence of the last theorem, we can argue that the step sizes in Algorithm `modBSSN` are eventually chosen equal to 1. In the following theorem, we additionally assume that g is more regular and that $\mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) = \emptyset$ holds at the unique zero $\mathbf{u}^*$ of $\mathbf{F}$.

Theorem 3.12 *Let g be three times continuously differentiable. Let $\{\mathbf{u}^{(j)}\}_j$ be a sequence produced by Algorithm* `modBSSN` *converging to a limit point $\mathbf{u}^*$ with $\mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) = \emptyset$. Then, there exists an index $j_0 \in \mathbb{N}$ such that $t_j = 1$ for all $j \geq j_0$.*

Proof. We proceed as in the proof of [108, Theorem 2]. Inspired by loc. cit., we show that for all j large enough, we have

$$\Theta(\mathbf{u}^{(j)}) - \Theta(\mathbf{u}^{(j)} + \mathbf{d}^{(j)}) \geq 2\sigma\Theta(\mathbf{u}^{(j)}).$$

We prove the claim by contradiction. Let the subsequence $\{\mathbf{u}^{(j)}\}_{j\in J}$ fulfill

$$\Theta(\mathbf{u}^{(j)}) - \Theta(\mathbf{u}^{(j)} + \mathbf{d}^{(j)}) < 2\sigma\Theta(\mathbf{u}^{(j)}) \tag{3.41}$$

for all $j \in J$ large enough. Because of Lemma 3.10, we have $\|\mathbf{d}^{(j)}\|_2 \leq C\|\mathbf{F}(\mathbf{u}^{(j)})\|_2$ with a constant $C > 0$. Therefore, with $\hat{\mathbf{u}}^{(j)} := \mathbf{u}^{(j)} + \mathbf{d}^{(j)}$, the sequence $\{\hat{\mathbf{u}}^{(j)}\}_{j\in J}$ has the limit $\mathbf{u}^*$. We consider

$$\Theta(\mathbf{u}^{(j)}) - \Theta(\hat{\mathbf{u}}^{(j)}) = \sum_{i=1}^{2} \hat{T}_i,$$

where

$$\begin{aligned}
\hat{T}_1 &:= \sum_{k\in\mathcal{A}(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big),\\
\hat{T}_2 &:= \sum_{k\in\mathcal{I}^\circ(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big).
\end{aligned}$$

Because of Theorem 3.11, we have

$$\begin{aligned}
\mathcal{A}^+(\mathbf{u}^*) &= \{k : 0 = \gamma(\nabla g(\mathbf{u}^*))_k + \gamma w_k < u_k^*\},\\
\mathcal{A}^-(\mathbf{u}^*) &= \{k : 0 = \gamma(\nabla g(\mathbf{u}^*))_k - \gamma w_k > u_k^*\},\\
\mathcal{I}^\circ(\mathbf{u}^*) &= \{k : \gamma(\nabla g(\mathbf{u}^*))_k - \gamma w_k < u_k^* = 0 < \gamma(\nabla g(\mathbf{u}^*))_k + \gamma w_k\}.
\end{aligned}$$

For all $j \in J$ large enough, we have

$$\begin{aligned}
\mathcal{A}^+(\mathbf{u}^*) &\subset \overline{\mathcal{A}^+}(\mathbf{u}^{(j)}) \cap \overline{\mathcal{A}^+}(\hat{\mathbf{u}}^{(j)}),\\
\mathcal{A}^-(\mathbf{u}^*) &\subset \overline{\mathcal{A}^-}(\mathbf{u}^{(j)}) \cap \overline{\mathcal{A}^-}(\hat{\mathbf{u}}^{(j)}),\\
\mathcal{I}^\circ(\mathbf{u}^*) &\subset \overline{\mathcal{I}^\circ}(\mathbf{u}^{(j)}) \cap \overline{\mathcal{I}^\circ}(\hat{\mathbf{u}}^{(j)}).
\end{aligned}$$

Lemma 3.10 implies the boundedness of the subsequence $\{\mathbf{d}^{(j)}/\|\mathbf{F}(\mathbf{u}^{(j)})\|_2\}_{j\in J}$ of quotients and without loss of generality, this subsequence has a limit $\tilde{\mathbf{d}} \in \mathbb{R}^n$, and the subsequence $\{\mathbf{F}(\mathbf{u}^{(j)})/\|\mathbf{F}(\mathbf{u}^{(j)})\|_2\}_{j\in J}$ of unit vectors converges to a unit vector $\tilde{\mathbf{F}} \in \mathbb{R}^n$.

Similar to the proof of Theorem 3.11, we consider the sums $\hat{T}_1$ and $\hat{T}_2$. First, we treat the sum $\hat{T}_1$. Because $k \in \mathcal{A}(\mathbf{u}^*) \subset \overline{\mathcal{A}}(\mathbf{u}^{(j)}) \cap \overline{\mathcal{A}}(\hat{\mathbf{u}}^{(j)})$, with (3.30), we have $F_k(\mathbf{u}^{(j)}) + \gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k = 0$. Dividing by $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2$ and taking the limit $j \to \infty, j \in J$, it follows $\tilde{\mathbf{F}}_k + \gamma(\nabla^2 g(\mathbf{u}^*)\tilde{\mathbf{d}})_k = 0$. Using Taylor's theorem and the mean-value theorem,

there exists a vector $\mathbf{v}$ on the line segment between $\mathbf{u}^{(j)}$ and $\hat{\mathbf{u}}^{(j)}$ with

$$\begin{aligned}\hat{T}_1 &= \sum_{k\in\mathcal{A}(\mathbf{u}^*)} \big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big)\\ &= \sum_{k\in\mathcal{A}(\mathbf{u}^*)} \Big(-2F_k(\mathbf{u}^{(j)})(\gamma\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k - (\gamma\nabla^2 g(\mathbf{v})\mathbf{d}^{(j)})_k^2\\ &\qquad\qquad - F_k(\mathbf{v})\gamma\sum_{l,m=1}^{n}\frac{\partial^3 g(\mathbf{v})}{\partial u_l\partial u_m\partial u_k}d_l^{(j)}d_m^{(j)}\Big)\\ &= \sum_{k\in\mathcal{A}(\mathbf{u}^*)} \Big(2F_k(\mathbf{u}^{(j)})^2 - (\gamma\nabla^2 g(\mathbf{v})\mathbf{d}^{(j)})_k^2 - F_k(\mathbf{v})\gamma\sum_{l,m=1}^{n}\frac{\partial^3 g(\mathbf{v})}{\partial u_l\partial u_m\partial u_k}d_l^{(j)}d_m^{(j)}\Big),\end{aligned}$$

where we have used (3.30) in the last equality. Dividing by $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2^2$ and taking the limit $j\to\infty$, $j\in J$, it follows

$$\lim_{j\to\infty, j\in J}\frac{\hat{T}_1}{\|\mathbf{F}(\mathbf{u}^{(j)})\|_2^2} = \sum_{k\in\mathcal{A}(\mathbf{u}^*)}\tilde{F}_k^2.$$

Now we consider the sum $\hat{T}_2$. We have $k\in\mathcal{I}^\circ(\mathbf{u}^*)\subset\overline{\mathcal{I}^\circ}(\mathbf{u}^{(j)})\cap\overline{\mathcal{I}^\circ}(\hat{\mathbf{u}}^{(j)})$ and with (3.30), we obtain

$$\begin{aligned}\hat{T}_2 &= \sum_{k\in\mathcal{I}^\circ(\mathbf{u}^*)}\big(F_k(\mathbf{u}^{(j)})^2 - F_k(\hat{\mathbf{u}}^{(j)})^2\big) = \sum_{k\in\mathcal{I}^\circ(\mathbf{u}^*)}\big((u_k^{(j)})^2 - (\hat{u}_k^{(j)})^2\big)\\ &= \sum_{k\in\mathcal{I}^\circ(\mathbf{u}^*)} F_k(\mathbf{u}^{(j)})^2.\end{aligned}$$

Therefore, we have

$$\lim_{j\to\infty, j\in J}\frac{\hat{T}_2}{\|\mathbf{F}(\mathbf{u}^{(j)})\|_2^2} = \sum_{k\in\mathcal{I}^\circ(\mathbf{u}^*)}\tilde{F}_k^2.$$

Finally, we divide both sides of the inequality (3.41) by $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2^2$ and take the limit $j\to\infty$, $j\in J$, obtaining

$$\|\tilde{\mathbf{F}}\|_2^2 \le 2\sigma\|\tilde{\mathbf{F}}\|_2^2$$

which is a contradiction to $\|\tilde{\mathbf{F}}\|_2 = 1$ and the choice $\sigma < 1/2$ in the Armijo rule (3.37), finishing the proof. □

Now we consider the locally quadratic convergence of Algorithm `modBSSN` in the case that the step sizes t_j are eventually chosen equal to 1, i.e., according to Theorem 3.12 especially in the case $\mathcal{I}^+(\mathbf{u}^*)\cup\mathcal{I}^-(\mathbf{u}^*) = \emptyset$. In the following theorem, we need the bounded invertibility of $\mathbf{G}(\mathbf{u})$ from (3.23) in a neighborhood of the zero $\mathbf{u}^*$ of $\mathbf{F}$. Because $\mathbf{M} := \nabla^2 g(\mathbf{u})$ is symmetric and positive definite, the inverse of $\mathbf{G}$ at $\mathbf{u}$ exists and is bounded by a constant $\tilde{C} > 0$

$$\|\mathbf{G}(\mathbf{u})^{-1}\|_2 \le \|\mathbf{M}_{\overline{\mathcal{B}},\overline{\mathcal{B}}}^{-1}\|_2\Big(\frac{1}{\gamma} + \|\mathbf{M}_{\overline{\mathcal{B}},\overline{\mathcal{C}}}\|_2\Big) + 1 \le \|\mathbf{M}^{-1}\|_2\Big(\frac{1}{\gamma} + \|\mathbf{M}\|_2\Big) + 1 \le \tilde{C}, \qquad (3.42)$$

see [58, Proposition 3.11] and [103, Lemma 3.6], cf. Corollary 2.16 and Proposition 2.23. The boundedness follows from Assumption B. For the following theorem, we need again the additional assumption that g is more regular.

Theorem 3.13 *Let g be three times continuously differentiable and let the step sizes t_j be chosen equal to 1 for all j large enough. Let $\{\mathbf{u}^{(j)}\}_j$ be a sequence produced by Algorithm* `modBSSN` *converging to $\mathbf{u}^*$. Then, there exists a constant $C > 0$ so that locally quadratic convergence is achieved, i.e., for all j large enough, we have*

$$\|\mathbf{u}^{(j+1)} - \mathbf{u}^*\|_2 \leq C\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2^2.$$

Proof. We follow the proof of [108, Theorem 3]. By assumption, we have $t_j = 1$, i.e., $\mathbf{u}^{(j+1)} = \mathbf{u}^{(j)} + \mathbf{d}^{(j)}$, for all j large enough. With $\overline{\mathcal{B}}(\mathbf{u}^{(j)})$, $\overline{\mathcal{C}}(\mathbf{u}^{(j)})$ from (3.22), we have

$$\begin{aligned} F_k(\mathbf{u}^{(j)}) + \gamma(\nabla^2 g(\mathbf{u}^{(j)})\mathbf{d}^{(j)})_k = 0, \qquad & \text{for } k \in \overline{\mathcal{B}}(\mathbf{u}^{(j)}), \\ u_k^{(j+1)} = 0, \qquad & \text{for } k \in \overline{\mathcal{C}}(\mathbf{u}^{(j)}). \end{aligned}$$

Because $\mathbf{u}^*$ is the limit of $\{\mathbf{u}^{(j)}\}_j$, we have with (3.14) and (3.15) for j large enough

$$\begin{aligned} \mathcal{A}(\mathbf{u}^*) &\subseteq \big(\{k : u_k^{(j)} > \gamma(\nabla g(\mathbf{u}^{(j)}))_k + \gamma w_k \wedge u_k^{(j)} > 0\} \\ &\quad \cup \{k : u_k^{(j)} < \gamma(\nabla g(\mathbf{u}^{(j)}))_k - \gamma w_k \wedge u_k^{(j)} < 0\}\big) \\ &\subseteq \overline{\mathcal{B}}(\mathbf{u}^{(j)}). \end{aligned}$$

This yields the inclusion $\overline{\mathcal{C}}(\mathbf{u}^{(j)}) \subseteq \mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) \cup \mathcal{I}^\circ(\mathbf{u}^*)$, implying $\mathbf{u}^*_{\overline{\mathcal{C}}(\mathbf{u}^{(j)})} = \mathbf{0}$. Analogously, with (3.16), we have for j large enough $\mathcal{I}^\circ(\mathbf{u}^*) \subseteq \overline{\mathcal{I}^\circ}(\mathbf{u}^{(j)}) \subseteq \overline{\mathcal{C}}(\mathbf{u}^{(j)})$. Consequently, we have $\overline{\mathcal{B}}(\mathbf{u}^{(j)}) \subseteq \mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) \cup \mathcal{A}(\mathbf{u}^*)$, implying $0 = F_k(\mathbf{u}^*) = \gamma \nabla g(\mathbf{u}^*)_k \pm \gamma w_k$, respectively, for all $k \in \overline{\mathcal{B}}(\mathbf{u}^{(j)})$.

Skipping the arguments $\overline{\mathcal{B}} = \overline{\mathcal{B}}(\mathbf{u}^{(j)})$, $\overline{\mathcal{C}} = \overline{\mathcal{C}}(\mathbf{u}^{(j)})$, we obtain with $\mathbf{u}^*_{\overline{\mathcal{C}}} = \mathbf{0}$, $\mathbf{F}(\mathbf{u}^*)_{\overline{\mathcal{B}}} = \mathbf{0}$ and the mean value theorem

$$\begin{aligned} &\begin{pmatrix} \big(\mathbf{G}(\mathbf{u}^{(j)})(\mathbf{u}^{(j+1)} - \mathbf{u}^*)\big)_{\overline{\mathcal{B}}} \\ \big(\mathbf{G}(\mathbf{u}^{(j)})(\mathbf{u}^{(j+1)} - \mathbf{u}^*)\big)_{\overline{\mathcal{C}}} \end{pmatrix} \\ &= \begin{pmatrix} \gamma(\nabla^2 g(\mathbf{u}^{(j)}))_{\overline{\mathcal{B}},\overline{\mathcal{B}}} & \gamma(\nabla^2 g(\mathbf{u}^{(j)}))_{\overline{\mathcal{B}},\overline{\mathcal{C}}} \\ \mathbf{0}_{\overline{\mathcal{C}},\overline{\mathcal{B}}} & \mathbf{I}_{\overline{\mathcal{C}},\overline{\mathcal{C}}} \end{pmatrix} \begin{pmatrix} (\mathbf{u}^{(j+1)} - \mathbf{u}^{(j)} + \mathbf{u}^{(j)} - \mathbf{u}^*)_{\overline{\mathcal{B}}} \\ (\mathbf{u}^{(j+1)} - \mathbf{u}^{(j)} + \mathbf{u}^{(j)} - \mathbf{u}^*)_{\overline{\mathcal{C}}} \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{F}(\mathbf{u}^{(j)})_{\overline{\mathcal{B}}} + \gamma\nabla^2 g(\mathbf{u}^{(j)})_{\overline{\mathcal{B}},\overline{\mathcal{B}}}(\mathbf{u}^{(j)} - \mathbf{u}^*)_{\overline{\mathcal{B}}} + \gamma\nabla^2 g(\mathbf{u}^{(j)})_{\overline{\mathcal{B}},\overline{\mathcal{C}}}(\mathbf{u}^{(j)} - \mathbf{u}^*)_{\overline{\mathcal{C}}} \\ -\mathbf{u}^{(j)}_{\overline{\mathcal{C}}} + (\mathbf{u}^{(j)} - \mathbf{u}^*)_{\overline{\mathcal{C}}} \end{pmatrix} + \begin{pmatrix} \mathbf{F}(\mathbf{u}^*)_{\overline{\mathcal{B}}} \\ \mathbf{u}^*_{\overline{\mathcal{C}}} \end{pmatrix} \\ &= \begin{pmatrix} \Big(\sum_{l,m=1}^{n} \frac{\gamma}{2} \frac{\partial^3 g(\mathbf{v})}{\partial u_l \partial u_m \partial u_k}(u_l^* - u_l^{(j)})(u_m^* - u_m^{(j)})\Big)_{k \in \overline{\mathcal{B}}} \\ \mathbf{0}_{\overline{\mathcal{C}}} \end{pmatrix}, \end{aligned}$$

where $\mathbf{v}$ is a vector on the line segment between $\mathbf{u}^{(j)}$ and $\mathbf{u}^*$. The matrix $\mathbf{G}(\mathbf{u}^{(j)})$ is boundedly invertible by Assumption B, cf. (3.42). Therefore, there exists a constant $C > 0$, depending only on $\mathbf{u}^*$, with

$$\|\mathbf{u}^{(j+1)} - \mathbf{u}^*\|_2 \leq C\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2^2,$$

for all j large enough, proving the claim. □

Note that in case of a quadratic functional $g(\mathbf{u}) = \frac{1}{2}\|\mathbf{K}\mathbf{u} - \mathbf{f}\|_2^2$ with $\mathbf{K}$ injective, $\mathbf{G}(\mathbf{u})^{-1}$ was shown to be uniformly bounded in a neighborhood of the zero $\mathbf{u}^*$ of $\mathbf{F}$ [58], otherwise the uniformly boundedness follows from Assumption B, see (3.42). Hence, in case of a quadratic functional g with $\mathcal{I}^+(\mathbf{u}^*) \cup \mathcal{I}^-(\mathbf{u}^*) = \emptyset$, the step sizes in Algorithm `modBSSN` are eventually chosen equal to 1, locally quadratic convergence is achieved and $\mathbf{u}^*$ is found within finitely many steps, see also Remark 3.9 and Chapter 2. The numerical results from Chapter 4 demonstrate these theoretical results.

3.2.2 Convergence of the hybrid method

The global convergence and the local convergence speed of Algorithm `hybridBSSN` from Section 3.1.2 directly follow from Theorem 3.11, Theorem 3.13 and Section 3.1.2, respectively. The method combines the efficiency of Algorithm `BSSN` and the stronger convergence properties of Algorithm `modBSSN`.

Chapter 4

Numerical results

In this chapter, we present numerical results illustrating the convergence properties and the computational complexity of Algorithm 2 (`modBSSN`) and Algorithm 3 (`hybridBSSN`). Note that the iterates produced by Algorithm 1 (`BSSN`) and Algorithm `hybridBSSN` coincide if Algorithm 1 does not begin to stagnate. In our numerical experiments, the index sets were in some test cases switched to the modified index sets in Algorithm `hybridBSSN`, cf. Section 3.1.2. It is unknown if the sequence of iterates in Algorithm 1 would have been converged in these cases. The numerical experiments were run on a desktop computer with INTEL® Xeon® CPU (W3530, 2.80 GHz) and the algorithms were implemented in MATLAB® 2015a. If not stated otherwise, the parameters from Table 4.1 are used in Algorithms `modBSSN` and `hybridBSSN`. The sequence $(w_k)_k$ of regularization parameters, see (1.4), was chosen to be constant in all experiments, i.e., we have $w_k \equiv w > 0$ for all k. In our numerical tests, we considered normally distributed noise computed with the MATLAB® subroutine `randn` and the CPU times have been measured with the subroutine `cputime`.

The linear inverse problem of inverse integration is considered in Section 4.1. Section 4.2 treats a nonlinear parameter identification problem. In the following, we first discuss the numerical treatment of the generalized Newton equations in Algorithm `hybridBSSN` and Algorithm `modBSSN`. Second, we make a comment on the chosen stopping criteria.

The remarks on the numerical solution of the generalized Newton equations were in a similar form made in our preprint [67]. The example of inverse integration was considered in our publication [66] for Algorithm 1 and is here considered for Algorithms `hybridBSSN` and `modBSSN`. Some of the experiments in Section 4.1 were performed for a deblurring problem in our preprint [67] but are new for the example of inverse integration.

Numerical solution of the generalized Newton equations

In each generalized Newton step of Algorithms `hybridBSSN` and `modBSSN`, the generalized Newton equation (2.37) or (3.30) has to be solved, respectively. This equation consists of a linear complementarity problem (2.32) or (3.19) of the size $|\mathcal{I}^{\pm}|$ or $|\overline{\mathcal{I}^{\pm}}|$ and one or more linear systems of the size $|\mathcal{A}^{\pm}|$ or $|\overline{\mathcal{A}^{\pm}}|$, respectively, see Section 3.1.2.

In our computations, the systems of linear equations were solved using the MATLAB® backslash subroutine. In case of ill-conditioned matrices one might need to precondition the linear systems. In the numerical experiments presented in this chapter, this was not necessary.

Table 4.1: Choice of parameters in the B-semismooth Newton methods.

parameter	domain	chosen value
$\mathbf{u}^{(0)}$	$\mathbb{R}^n$	$\mathbf{0}$
p	$\{1,2\}$	2
σ	$(0,1/p)$	0.01
β	$(0,1)$	0.5
tol	$\mathbb{R}_+$	10^{-7}
t_{min}	$(0,1]$	10^{-5}

As pointed out in our preprint [67], the linear complementarity problems appearing in Algorithms `hybridBSSN` and `modBSSN` are solved numerically up to machine precision. An inexact solution of the generalized Newton equations may effect an increased number of B-semismooth Newton steps until convergence. Additionally, no convergence theory has been developed for inexact B-semismooth Newton methods so far. Lemke's method [24, 91,123] is a popular algorithm that solves the linear complementarity problems (2.32) and (3.19) within finitely many iterations. In our computations, we used the *modified damped Newton method* from [70] for the numerical solution of the arising linear complementarity problems. This very method is a version of the damped B-Newton method from [107] for the special case of *linear* complementarity problems. If the sequence of iterates in this algorithm converges, a solution is found within a finite number of iterations, see [47]. The numerical results in [70] demonstrate that the modified B-Newton method often has a better performance in terms of computation time than Lemke's method. In each generalized Newton step of the modified B-Newton method from [70], there arises only one system of linear equations if the initial guess $\mathbf{x}^{(0)}$ fulfills a regularity assumption [70]. More precisely, the starting vector $\mathbf{x}^{(0)}$ has to satisfy

$$x_k^{(0)} \neq y_k^{(0)} := (\mathbf{N}\mathbf{x}^{(0)} + \mathbf{z})_k, \tag{4.1}$$

for all k, with the matrix $\mathbf{N}$ and the vector $\mathbf{z}$ from (2.27), (3.20) and (2.33), (3.21), respectively. In our computations, we started with the initial guess $\mathbf{x}^{(0)} = \mathbf{0}$ and stopped the iteration if the residual norm $\| \min\{\mathbf{N}\mathbf{x}^{(j)} + \mathbf{z}, \mathbf{x}^{(j)}\}\|_2$, for an iterate $\mathbf{x}^{(j)}$, $j \geq 0$, was smaller than the tolerance value 10^{-7}. If the regularity assumption (4.1) was violated, or if the number of Newton steps exceeded 50, we switched to Lemke's method. However, this is usually not the case. For Lemke's method we used an implementation from [141] that is included in the RPI MATLAB simulator. The code is available on the web page `http://code.google.com/p/rpi-matlab-simulator` (30 June 2015). There are several alternatives for the numerical treatment of linear complementarity problems, e.g., see [24, 34, 42, 45, 123].

A comment on the chosen stopping criteria

The stopping criterion used in Algorithms `hybridBSSN` and `modBSSN` is a residual norm smaller than a tolerance $tol = 10^{-7}$,

$$\|\mathbf{F}(\mathbf{u}^{(j)})\|_2 = \|\mathbf{u}^{(j)} - \mathbf{S}_{\gamma\mathbf{w}}(\mathbf{u}^{(j)} - \gamma\nabla g(\mathbf{u}^{(j)}))\|_2 < tol.$$

This stopping criterion depends on the choice of the parameter $\gamma > 0$. Alternatively, one may think of taking some reference parameter $\tilde{\gamma}$ to define a stopping criterion that is independent of the parameter γ used in the B-semismooth Newton algorithms. Unfortunately, the iterates of Algorithms `hybridBSSN` and `modBSSN` with parameter γ do usually not decrease with respect to the residual norm of the nonlinearity $\mathbf{F}$ depending on another parameter $\tilde{\gamma}$. Therefore, in the experiments presented in this chapter, the stopping criterion depended on the nonlinearity under consideration, i.e., the parameter γ.

The stopping criterion based on the residual norm worked well in our numerical tests. A local error estimate,

$$\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2 \leq C\|\mathbf{F}(\mathbf{u}^{(j)})\|_2,$$

where $C > 0$ is a constant, can be derived analogously to the smooth case under the assumption that the Newton derivative $\mathbf{G}(\mathbf{u}^*)$ at the unknown root $\mathbf{u}^*$ of $\mathbf{F}$ is invertible, see [138, Lemma 10.4]. Using the Newton differentiability of $\mathbf{F}$, one can prove that the above error estimate holds in a neighborhood of $\mathbf{u}^*$ with $C = 2\|\mathbf{G}(\mathbf{u}^*)^{-1}\|_2$. Nevertheless, this constant is unknown in practice. Note that in case of quadratic discrepancy terms, Algorithms `hybridBSSN` and `modBSSN` find the zero of $\mathbf{F}$ (up to machine precision) within a finite number of iterations, cf. Section 2.1. Other possible stopping criteria could, e.g., be defined by the distance of the iterates or of their Tikhonov functional values, but the latter are not strictly decreasing in Algorithms `hybridBSSN` and `modBSSN`.

Note that the computation time and the number of computed iterations depend on the chosen stopping criterion. For comparison of the B-semismooth Newton methods with other state-of-the-art algorithms, we stopped these algorithms once the Tikhonov functional value of an iterate was below a threshold. This threshold was equal to the Tikhonov functional value $J(\mathbf{u}^*)$ at an approximation $\mathbf{u}^*$ of the root of $\mathbf{F}$ plus a tolerance, see Figures 4.3 and 4.4. Here, an approximation $\mathbf{u}^*$ was computed using Algorithm `hybridBSSN` or Algorithm `modBSSN` with one fixed parameter γ.

4.1 Inverse integration

In this section, we are concerned with the linear inverse problem of *inverse integration* from [10, 58, 125]. The same example was considered for Algorithm 1 in [66]. Similar experiments for another linear inverse problem were presented in our preprint [67]. Our aim is to solve the operator equation $Ku(x) = f(x)$, where the forward operator K is given by

$$K\colon L_2([0,1]) \to L_2([0,1]), \quad Ku(x) = \int_0^x u(t)\mathrm{d}t.$$

Hence, we are concerned with a Volterra integral equation of the first kind. The operator K is compact, its singular values are $\sigma_k = 2/((2k-1)\pi)$, $k = 1, 2, \ldots$, and hence the operator equation is moderately ill-posed, see [74]. Such problems arise in practical applications like the computation of the velocity of an object from GPS data, see, e.g., [74, 126].

We consider the equidistant grid $\Delta = \{x_k = k/n : k = 1, \ldots, n\}$ on the interval $[0, 1]$. The discretization $\mathbf{K} = (k_{ij})_{ij} \in \mathbb{R}^{n\times n}$ of the forward operator K is a lower triangular

matrix with

$$k_{ij} = \begin{cases} 1/N, & i \geq j, \\ 0, & i < j. \end{cases}$$

The true solution $u \in L_2([0,1])$ is given by

$$u(x) := \begin{cases} 80, & 0.11 \leq x \leq 0.12, \\ -50, & 0.32 \leq x \leq 0.33, \\ 20, & 0.53 \leq x \leq 0.54, \\ 60, & 0.66 \leq x \leq 0.67, \\ -100, & 0.9 \leq x \leq 0.91, \\ 0, & \text{else.} \end{cases} \tag{4.2}$$

The data $\mathbf{f} = (f(x_k))_{k=1,\ldots,n}$ are equal to the functional values of f at the grid points $x_k \in \Delta$. To avoid *inverse crime*, see, e.g., [22, p. 154] and [68, Section 7.2], the data $\mathbf{f}$ were here computed on the grid $\tilde{\Delta} = \{\tilde{x}_k = k/\tilde{n} : k = 1, \ldots, \tilde{n}\}$, where $\tilde{n} = 2n + 1$, using Simpson's rule on $\tilde{\Delta}$, followed by a linear interpolation of the data to the grid Δ. In this example, we consider the quadratic discrepancy term

$$g(\mathbf{u}) := \frac{1}{2}\|\mathbf{K}\mathbf{u} - \mathbf{f}\|_2^2.$$

In practical applications, the right-hand side $\mathbf{f}$ contains measurement errors. In the following, we denote the noise level δ as the relative noise contained in the noisy right-hand side $\mathbf{f}^\delta$, i.e., $\|\mathbf{f} - \mathbf{f}^\delta\|_2 = \delta\|\mathbf{f}\|_2$. Let us assume that the unperturbed right-hand side $\mathbf{f}$ is known and let $\mathbf{f}^\delta$ denote the perturbed right-hand side. If not stated otherwise, the regularization parameter was, for the example of inverse integration, chosen a posteriori according to the *discrepancy principle* [1, 4, 38, 126]. Here, we chose $w^{(0)} := 0.9^{10} \approx 0.3487$, $q := 0.9$ and $\tau := 1.5$. For $l = 0, 1, \ldots$, we computed the minimizer $\mathbf{u}_{w^{(l)}}$ of (2.1) with $g(\mathbf{u}) = \frac{1}{2}\|\mathbf{K}\mathbf{u} - \mathbf{f}^\delta\|_2^2$ and regularization sequence $w_k \equiv w^{(l)} := w^{(0)}q^l$ iteratively, until the discrepancy $\|\mathbf{K}\mathbf{u}_{w^{(l)}} - \mathbf{f}^\delta\|_2$ was smaller than or equal to $\tau\|\mathbf{f} - \mathbf{f}^\delta\|_2$. We chose $w_k \equiv w^{(l)}$ as the largest such parameter, see also our previous works [66, 67].

Let us outline the following paragraphs. The noisy data and a reconstruction for the inverse integration example are presented in Figure 4.1. The history of one run of Algorithms `hybridBSSN` and `modBSSN` is shown in Tables 4.2 and 4.3, respectively. The influence of the starting vector on the number of iterations of the algorithms is treated in Table 4.4. Experiments on the influence of several parameters on the computational cost of the considered algorithms is presented in Tables 4.5 and 4.6. Here, the regularization parameters were chosen according to the discrepancy principle. Similar experiments are shown in Tables 4.7 and 4.8 without using a parameter choice strategy for the regularization parameter w. A comparison of Algorithms `hybridBSSN` and `modBSSN` with the local semismooth Newton method from [58] is presented in Figure 4.2. A comparison with some other state-of-the-art methods is shown in Figures 4.3 and 4.4, respectively.

Figure 4.1 shows the function u from (4.2), the 3% noisy data $\mathbf{f}^\delta$ and a reconstruction on the equidistant grid Δ with $n = 500$ nodes. The reconstruction was computed using Algorithm `hybridBSSN` with the parameters $\gamma = 10^5$, $k_{max} = 250$ and regularization parameter

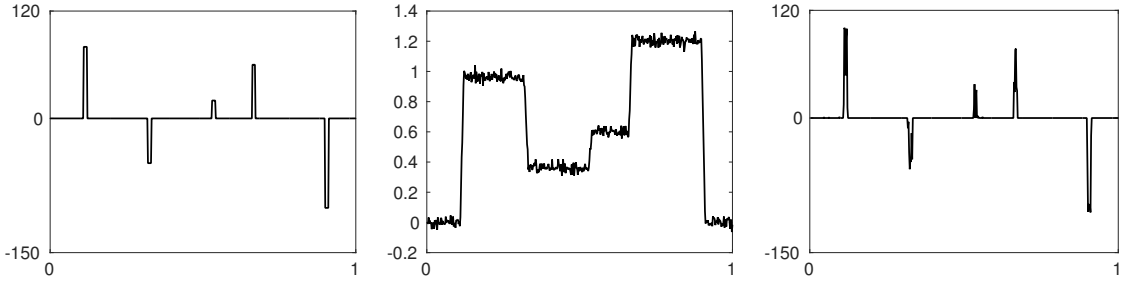


Figure 4.1: Example of inverse integration on a grid with $n = 500$ nodes. From left to right: true solution u, 3% noisy data, reconstruction.

Table 4.2: Convergence history of Algorithm `hybridBSSN` for the example from Figure 4.1 with $n = 500$, $\delta = 0.03$, $w = 0.9^{55} \approx 0.003$, $\gamma = 10^5$ and $k_{max} = 250$.

j	$\|\mathbf{F}(\mathbf{u}^{(j)})\|_2$	$\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2$	$\mathcal{A}$	$\mathcal{I}^\circ$	$\mathcal{I}^+$	$\mathcal{I}^-$	size LCP	size SLE	# SLE	cond$(\mathbf{M}_{\mathcal{A},\mathcal{A}})$	t_j
0	9.03e+05	353.26	452	48	0	0	-	-	-	-	-
1	5.29e+02	515.87	230	270	0	0	0	452	1	4.05e+05	1
2	4.03e+02	425.74	186	314	0	0	0	230	1	1.97e+05	0.5
3	3.78e+02	373.11	171	329	0	0	0	186	1	1.59e+05	0.25
4	2.73e+02	310.98	128	372	0	0	0	171	1	1.40e+05	0.5
5	2.04e+02	280.30	98	402	0	0	0	128	1	1.03e+05	0.5
6	1.88e+02	256.90	97	403	0	0	0	98	1	7.73e+04	0.125
7	1.52e+02	209.11	62	438	0	0	0	97	1	7.91e+04	1
8	8.85e+01	117.77	48	452	0	0	0	62	1	4.84e+04	1
9	7.08e+01	88.87	44	456	0	0	0	48	1	3.83e+04	1
10	2.70e+01	33.45	42	458	0	0	0	44	1	3.72e+04	1
11	1.71e+01	22.59	41	459	0	0	0	42	1	3.60e+04	1
12	1.49e+01	20.90	40	460	0	0	0	41	1	3.53e+04	1
13	7.24e-11	2.29e-10	40	460	0	0	0	40	1	3.46e+04	1

$w = 0.9^{55} \approx 0.003$. Table 4.2 and Table 4.3 present the history of Algorithm `hybridBSSN` and Algorithm `modBSSN` for this very example, both with $\gamma = 10^5$. Each table shows the history of the residual norms, the Euclidean distances of the iterates $\mathbf{u}^{(j)}$ to the reconstruction $\mathbf{u}^*$ (computed with Algorithm `modBSSN`), the sizes of the index sets, the sizes of the linear complementarity problems (LCP) and of the systems of linear equations (SLE), the numbers of linear systems, the condition numbers of the matrices $\mathbf{M}_{\mathcal{A},\mathcal{A}} = (\mathbf{K}^*\mathbf{K})_{\mathcal{A},\mathcal{A}}$ or $\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{A}}} = (\mathbf{K}^*\mathbf{K})_{\overline{\mathcal{A}},\overline{\mathcal{A}}}$ (computed using the MATLAB® subroutine `cond`) as well as the step sizes t_j for each iterate $\mathbf{u}^{(j)}$. The reconstructions $\mathbf{u}^*_{hybrid}$ and $\mathbf{u}^*_{mod}$, computed with Algorithm `hybridBSSN` and Algorithm `modBSSN`, respectively, vary only slightly. Here, we had $\|\mathbf{u}^*_{mod} - \mathbf{u}^*_{hybrid}\|_\infty \approx 8.1 \cdot 10^{-11}$. Therefore, it is negligible to choose the reference solution $\mathbf{u}^*$ as the reconstruction produced with Algorithm `hybridBSSN` or Algorithm `modBSSN`. The residual norms strictly decreased in both algorithms and the iterates jumped into the root $\mathbf{u}^*$ of $\mathbf{F}$ (up to machine precision) within one step, cf. Section 2.1.1. The sizes of the active sets usually decreased in the course of the iteration. In Algorithm `hybridBSSN`, the index sets $\mathcal{I}^\pm$ were empty. Therefore, there were no linear complementarity problems to

Table 4.3: Convergence history of Algorithm `modBSSN` for the example from Figure 4.1 with $n = 500$, $\delta = 0.03$, $w = 0.9^{55} \approx 0.003$ and $\gamma = 10^5$.

j	$\|\mathbf{F}(\mathbf{u}^{(j)})\|_2$	$\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2$	$\overline{\mathcal{A}}$	$\overline{\mathcal{I}^\circ}$	$\overline{\mathcal{I}^+}$	$\overline{\mathcal{I}^-}$	size LCP	size SLE	# SLE	cond($\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}$)	t_j
0	9.03e+05	353.26	452	48	0	0	-	-	-	-	-
1	5.29e+02	515.87	230	167	103	0	0	452	1	4.05e+05	1
2	4.42e+02	497.19	178	252	70	0	103	230	271	1.97e+05	0.5
3	4.06e+02	382.66	158	299	43	0	70	178	323	1.57e+05	0.25
4	3.14e+02	372.79	112	366	16	6	43	158	343	1.30e+05	0.5
5	2.19e+02	297.95	95	389	10	6	22	112	389	9.09e+04	0.5
6	1.76e+02	253.39	84	410	2	4	16	95	406	7.40e+04	0.25
7	1.73e+02	248.08	85	410	1	4	6	84	417	6.60e+04	0.031
8	1.52e+02	208.41	57	417	24	2	5	85	416	6.67e+04	1
9	9.00e+01	119.30	50	444	6	0	26	57	444	4.38e+04	1
10	7.60e+01	94.30	46	450	3	1	6	50	451	4.13e+04	0.5
11	4.43e+01	56.40	41	458	1	0	4	46	455	3.76e+04	1
12	1.97e+01	31.21	40	458	2	0	1	41	460	3.40e+04	1
13	1.94e+01	31.88	40	460	0	0	2	40	461	3.46e+04	1
14	5.54e+00	15.97	40	460	0	0	0	40	1	3.46e+04	1
15	9.19e-11	0	40	460	0	0	0	40	1	3.46e+04	1

Table 4.4: Influence of the starting vector on the number of iterations of the Algorithms `hybridBSSN` and `modBSSN` for the inverse integration example with $n = 2000$, $\delta = 0.05$, $\gamma = 10^5$ and $w = 0.9^{51} \approx 0.0046$.

$\|\mathbf{u}^{(0)} - \mathbf{u}^*\|_2$	`hybridBSSN`			`modifiedBSSN`		
	max. #iter.	min. #iter.	av. #iter.	max. #iter.	min. #iter.	av. #iter.
10^0	13	4	8.0	27	6	19.3
10^1	13	7	11.3	40	22	32.1
10^2	24	12	18.5	66	22	52.6
10^3	28	11	20.9	78	46	58.5
10^4	35	21	24.6	72	39	56.7
$\mathbf{u}^{(0)} = \mathbf{0}$	17	17	17	42	42	42

solve. The sizes of the linear systems are equal to the number of active indices and hence they also decreased along the iteration. The computational effort of Algorithm `modBSSN` was higher than the one of Algorithm `hybridBSSN` due to the linear complementarity problems and several linear systems that had to be solved in most of the steps in Algorithm `modBSSN`. The numbers of linear complementarity problems and the numbers of systems of linear equations were also decreasing along the iteration. The condition numbers of the system matrices $\mathbf{M}_{\mathcal{A},\mathcal{A}}$ and $\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{A}}}$, respectively, of the linear equations had comparable magnitudes in both algorithms. Additionally, the step sizes t_j were eventually equal to 1.

In Table 4.4, we consider the influence of the starting vector on the number of iterations until convergence in Algorithm `hybridBSSN` and in Algorithm `modBSSN`. Here, we chose $n = 2000$, $w = 0.9^{51} \approx 0.0046$, $\gamma = 10^5$ and $k_{max} = 250$. The noisy right-hand

Table 4.5: Influence of the number n of unknowns on the computational cost of Algorithms `hybridBSSN` and `modBSSN` for the example of inverse integration with $\delta = 0.05$ for different values of γ.

$\gamma = 10^4$, 5% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	50	0.07	14	7	0.19	18	4
500	50	0.19	21	6	0.65	36	3
1000	51	0.93	68	6	1.88	60	3
2500	51	20.05	289	2	32.66	337	3
5000	51	302.34	903	3	253.67	732	4

$\gamma = 10^5$, 5% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	50	0.06	13	9	0.13	11	8
500	50	0.13	15	10	0.24	15	8
1000	51	0.3	14	9	0.91	24	2
2500	51	3.4	26	11	13.1	67	5
5000	51	19.99*	57*	5*	102.62	178	4

$\gamma = 10^6$, 5% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	50	0.08	17	5	0.16	17	5
500	50	0.17	20	5	0.31	20	5
1000	51	0.45	21	2	0.72	20	4
2500	51	2.96	20	4	6.58	29	3
5000	51	18.05	28	6	69.79	63	3

* Algorithm `hybridBSSN` switched to the modified index sets in step $j = 42$.

side $\mathbf{f}^\delta$ was fixed, i.e., we used one noise realization, and $\mathbf{f}^\delta$ contained 5% of noise. An approximation $\mathbf{u}^*$ of the zero of $\mathbf{F}$ was here computed with Algorithm `hybridBSSN` using $\gamma = 10^5$ and initial guess $\mathbf{u}^{(0)} = \mathbf{0}$. We considered 20 normalized, standard normally distributed random vectors $\mathbf{n}^{(1)}, \ldots, \mathbf{n}^{(20)}$, i.e., we had $\|\mathbf{n}^{(i)}\|_2 = 1$ for all i. For each of the radii $r = 10^0, \ldots, 10^4$, we solved the zero-finding problem $\mathbf{F}(\mathbf{u}) = \mathbf{0}$ for one each of the 20 different starting vectors of the form $\mathbf{u}_i^{(0)} := \mathbf{u}^* + r\mathbf{n}^{(i)}$, $i = 1, \ldots, 20$. For both algorithms, we list the maximal, minimal and average numbers of iterations for these 20 runs depending on the radius r. The number of Newton steps that had to be computed until the residual norm was smaller than $tol = 10^{-7}$ was usually larger when using starting vectors far away from $\mathbf{u}^*$, i.e., for larger r. Additionally, the number of iterations of both algorithms with the starting vector $\mathbf{u}^{(0)} = \mathbf{0}$ is listed. Here, the Euclidean distance to $\mathbf{u}^*$ was $\|\mathbf{0} - \mathbf{u}^*\|_2 \approx 827.1$.

The average CPU time of five runs, the number of iterations until convergence as well as the number $\#\{j : t_j = 1\}$ of step sizes equal to 1 in Algorithms `hybridBSSN` and `modBSSN`

Table 4.6: Influence of the number n of unknowns on the computational cost of Algorithms `hybridBSSN` and `modBSSN` for the example of inverse integration with $\delta = 0.01$ for different values of γ.

$\gamma = 10^4$, 1% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	66	0.11	16	5	0.27	18	6
500	65	0.19	18	6	0.67	34	5
1000	66	1.17	76	5	2.89	91	4
2500	67	19.55	230	3	19.8	131	3
5000	67	135.84*	316*	3*	195.58	416	2

$\gamma = 10^5$, 1% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	66	0.06	11	9	0.12	11	9
500	65	0.12	15	13	0.42	18	10
1000	66	0.33	15	9	0.88	23	9
2500	67	3.37	29	10	13.58	61	5
5000	67	26.24	63	6	140.1	145	6

$\gamma = 10^6$, 1% of noise

		`hybridBSSN`			`modifiedBSSN`		
n	k	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$	av. cputime(s)	# iter.	# $\{j : t_j = 1\}$
250	66	0.11	15	7	0.25	16	7
500	65	0.2	19	13	0.37	18	13
1000	66	0.58	20	4	1.17	20	4
2500	67	3.74	22	10	10.22	25	8
5000	67	21.56	26	9	75.92	51	7

* Algorithm `hybridBSSN` switched to the modified index sets in step $j = 305$.

for increasing numbers n of unknowns, for different values of the parameter $\gamma = 10^4, 10^5, 10^6$ and the noise levels $\delta = 0.05, 0.01$ is presented in Table 4.5 and Table 4.6. The integers k denote the exponents of the regularization parameters $w = 0.9^k$ chosen according to the discrepancy principle. The regularization parameters in case of 1% noisy data were smaller than those for 5% of noise. To ensure comparability, we used one noise vector $\mathbf{n} \in \mathbb{R}^{5000}$ in all computations. If n was smaller than 5000, we used the vector of every $n/5000$-th entry of $\mathbf{n}$ as the noise vector. The parameter k_{max} in Algorithm `hybridBSSN` was chosen as $k_{max} = 1000$. For each fixed γ and δ, the average CPU time as well as the number of iterations usually increased for larger n. The amount of step sizes t_j that were equal to 1, the number of iterations as well as the average CPU time depended strongly on the choice of the parameter γ in both algorithms. Usually, Algorithm `hybridBSSN` was faster than Algorithm `modBSSN` in terms of computation time. Nevertheless, in the case $n = 5000$, $\gamma = 10^4$ and 5% noisy data, 903 iterations were computed in Algorithm `hybridBSSN` and the average CPU time of Algorithm `hybridBSSN` was larger than the

Table 4.7: Influence of the system size n, of the parameter γ and of the regularization parameter w on the computational cost of Algorithms `hybridBSSN` and `modBSSN` with $k_{max} = 250$ and 5% noisy data for the inverse integration example.

		$n = 500$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	0.2	0.4	0.3	0.1	0.1	0.1	0.2	0.1	0.1
	# iter.	18	43	31	11	10	14	22	15	9
	$\#\{j : t_j = 1\}$	7	5	4	9	7	6	4	5	7
	$\#\{k : u_k \neq 0\}/n$	0.07	0.10	0.26	0.07	0.10	0.26	0.07	0.10	0.26
modBSSN	av. cputime(s)	0.6	0.5	0.5	0.3	0.2	0.3	0.3	0.2	0.2
	# iter.	30	30	30	21	13	19	22	15	9
	$\#\{j : t_j = 1\}$	3	3	3	6	6	8	4	5	7
	$\#\{k : u_k \neq 0\}/n$	0.07	0.10	0.26	0.07	0.10	0.26	0.07	0.10	0.26

		$n = 1000$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	0.6	2.3	1.7	0.3	0.4	0.7	0.4	0.4	0.3
	# iter.	40	126	75	16	13	30	19	13	11
	$\#\{j : t_j = 1\}$	5	5	2	9	8	6	5	8	8
	$\#\{k : u_k \neq 0\}/n$	0.04	0.06	0.12	0.04	0.06	0.12	0.04	0.06	0.12
modBSSN	av. cputime(s)	2.6	2.6	3.3	1.4	1.0	1.7	0.9	0.6	0.5
	# iter.	81	84	101	38	27	37	25	14	11
	$\#\{j : t_j = 1\}$	4	4	3	4	7	5	4	9	8
	$\#\{k : u_k \neq 0\}/n$	0.04	0.06	0.12	0.04	0.06	0.12	0.04	0.06	0.12

		$n = 2000$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	4.5	13.9*	16.4	1.9	2.5	9.4	1.8	1.5	2.3
	# iter.	143	265*	165	25	52	89	25	17	23
	$\#\{j : t_j = 1\}$	4	3*	3	5	9	2	11	11	8
	$\#\{k : u_k \neq 0\}/n$	0.03	0.04	0.06	0.03	0.04	0.06	0.03	0.04	0.06
modBSSN	av. cputime(s)	12.7	14.3	17.7	8.3	7.2	13.8	3.6	3.8	4.5
	# iter.	182	205	149	56	48	85	26	28	21
	$\#\{j : t_j = 1\}$	4	2	2	2	4	4	2	5	3
	$\#\{k : u_k \neq 0\}/n$	0.03	0.04	0.06	0.03	0.04	0.06	0.03	0.04	0.06

* Algorithm `hybridBSSN` switched to the modified index sets in step $j = 245$.

average CPU time of Algorithm `modBSSN`. Here, a smaller choice of k_{max} could have been reasonable. Algorithm `hybridBSSN` switched to the modified index sets in step $j = 42$ in the case $n = 5000$, $\gamma = 10^5$ and $\delta = 0.05$ as well as in step $j = 305$ in the case $n = 5000$, $\gamma = 10^4$ and $\delta = 0.01$. For $\gamma = 10^6$, the numbers of iterations did increase less for increasing n than for smaller choices of γ.

Table 4.7 and Table 4.8 demonstrate the influence of the number n of unknowns, of

Table 4.8: Influence of the system size n, of the parameter γ and of the regularization parameter w on the computational cost of Algorithms `hybridBSSN` and `modBSSN` with $k_{max} = 250$ and 1% noisy data for the inverse integration example.

		$n = 500$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	0.2	0.2	0.3	0.2	0.1	0.1	0.2	0.1	0.1
	# iter.	23	17	36	18	11	9	24	15	9
	$\#\{j : t_j = 1\}$	3	7	6	8	9	8	3	8	8
	$\#\{k : u_k \neq 0\}/n$	0.07	0.09	0.11	0.07	0.09	0.11	0.07	0.09	0.11
modBSSN	av. cputime(s)	0.5	0.5	0.6	0.3	0.2	0.2	0.4	0.3	0.2
	# iter.	30	29	37	20	15	14	24	15	9
	$\#\{j : t_j = 1\}$	3	8	2	3	11	5	3	8	8
	$\#\{k : u_k \neq 0\}/n$	0.07	0.09	0.11	0.07	0.09	0.11	0.07	0.09	0.11
		$n = 1000$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	0.7	1.4	3.5	0.4	0.4	0.6	0.6	0.5	0.4
	# iter.	36	104	148	19	16	28	27	19	13
	$\#\{j : t_j = 1\}$	3	7	2	9	7	7	3	3	10
	$\#\{k : u_k \neq 0\}/n$	0.05	0.07	0.08	0.05	0.07	0.08	0.05	0.07	0.08
modBSSN	av. cputime(s)	2.5	1.8	5.0	1.2	1.2	1.5	0.7	1.1	0.6
	# iter.	69	48	166	30	20	25	22	19	13
	$\#\{j : t_j = 1\}$	4	4	2	3	3	5	9	3	10
	$\#\{k : u_k \neq 0\}/n$	0.05	0.07	0.08	0.05	0.07	0.08	0.05	0.07	0.08
		$n = 2000$								
		$\gamma = 10^4$			$\gamma = 10^5$			$\gamma = 10^6$		
	w	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
hybridBSSN	av. cputime(s)	3.0	9.8	14.3	2.2	2.0	3.5	2.1	2.6	2.1
	# iter.	62	196	224	31	26	70	31	24	15
	$\#\{j : t_j = 1\}$	4	4	3	2	11	6	2	8	10
	$\#\{k : u_k \neq 0\}/n$	0.04	0.06	0.07	0.04	0.06	0.07	0.04	0.06	0.07
modBSSN	av. cputime(s)	12.1	20.0	26.5	8.8	6.8	6.6	5.2	5.7	4.5
	# iter.	113	305	326	64	46	52	44	28	19
	$\#\{j : t_j = 1\}$	3	3	3	3	9	7	2	7	9
	$\#\{k : u_k \neq 0\}/n$	0.04	0.06	0.07	0.04	0.06	0.07	0.04	0.06	0.07

the parameter γ and of the regularization parameter w on the average CPU time of five runs, on the number of iterations, on the number of step sizes that were equal to 1 and on the relative sparsity $\#\{k : u_k \neq 0\}/n$ of the reconstructions computed with Algorithms `hybridBSSN` and `modBSSN`. Here, in contrast to Table 4.5 and Table 4.6, the regularization parameters were not chosen according to the discrepancy principle. The results in Tables 4.7 and 4.8 were computed with 5% and 1% noisy data, respectively. As in Tables 4.5 and

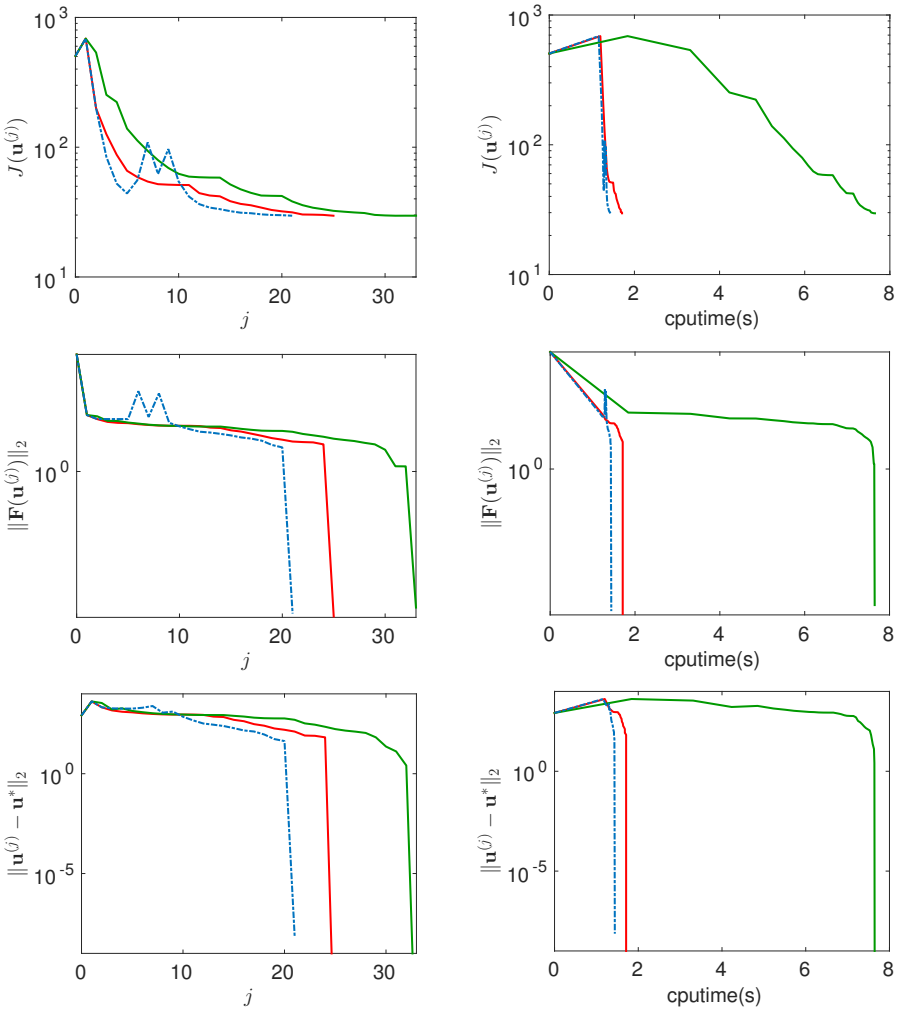


Figure 4.2: Convergence history of the Tikhonov functional values, the residual norms and the error norms of Algorithm `localBSSN` (blue dashed line), Algorithm `hybridBSSN` (red line) and Algorithm `modBSSN` (green line) versus the iteration index j and the CPU time for the example of inverse integration with $n = 2000$ and $\delta = 0.05$. The parameters used in the algorithms were $w = 0.9^{51} \approx 0.0046$ and $\gamma = 10^6$.

4.6, we used one single noise vector in all computations. The computed reconstructions $\mathbf{u}$ were less sparse for smaller w. Moreover, the relative sparsity $\#\{k : u_k \neq 0\}/n$ of a reconstruction $\mathbf{u}$ decreased for increasing n using a fixed regularization parameter w. As in the tables above, the influence of the choice of the parameter γ on the computational cost of both Algorithms `hybridBSSN` and `modBSSN` was obvious. Algorithm `hybridBSSN` switched to the modified index sets in the case $n = 2000$, $\gamma = 10^4$, $w = 10^{-3}$ and 5% of noise. Here, the parameter k_{max} was chosen equal to 250.

Figure 4.2 presents the convergence history of Algorithm `hybridBSSN` (red), of Algorithm `modBSSN` (green) and of the local B-semismooth Newton method `localBSSN` (blue dashed), i.e., Algorithm 1 with constant step sizes $t_j \equiv 1$. In this test case, we chose $n = 2000$. Here, the iterates of the local B-semismooth Newton method `localBSSN` co-

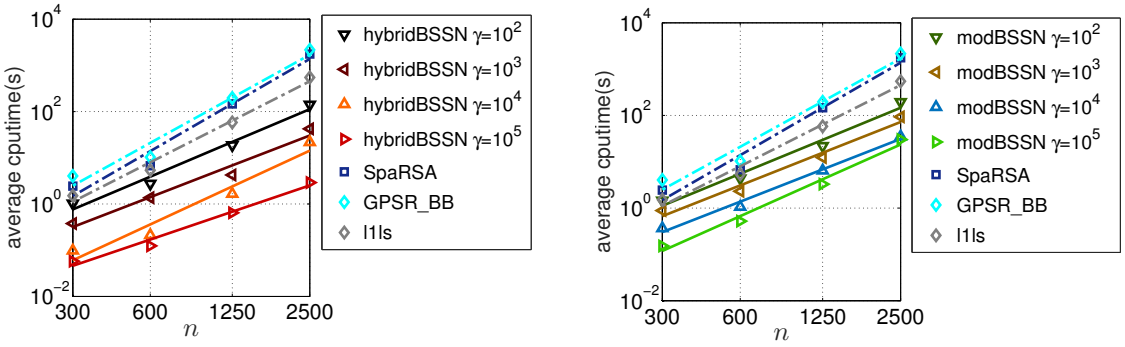


Figure 4.3: CPU time comparison of different algorithms for increasing numbers n of unknowns for the inverse integration example with $\delta = 0.05$. Left: comparison of Algorithm `hybridBSSN` using different values of γ with state-of-the art methods. Right: comparison of Algorithm `modBSSN` using different values of γ with state-of-the art methods.

incided with the local semismooth Newton method of Herzog and Lorenz [58] because the index sets $\mathcal{I}^{\pm}(\mathbf{u}^{(j)})$ were empty. The regularization parameter $w = 0.9^{51} \approx 0.0046$ was computed using the discrepancy principle. The parameter γ was in all methods chosen equal to 10^6. The starting vector in all three algorithms was the zero vector and the parameter k_{max} in Algorithm `hybridBSSN` was $k_{max} = 250$. The data contained 5% of noise and $\mathbf{u}^*$ here was the reconstruction computed with Algorithm `modBSSN`. The history of the Tikhonov functional values $J(\mathbf{u}^{(j)})$, the residual norms $\|\mathbf{F}(\mathbf{u}^{(j)})\|_2$ and the errors $\|\mathbf{u}^{(j)} - \mathbf{u}^*\|_2$ in the Euclidean norm of one run are presented depending on the iteration number j as well as the CPU time in seconds. In contrast to Algorithm `localBSSN`, the residual norms of Algorithm `hybridBSSN` and Algorithm `modBSSN` were strictly decreasing. The Tikhonov functional values of the three algorithms usually do not strictly decrease. The local superlinear convergence of the sequence of iterates in each algorithm is demonstrated by the plots of the error norms. The CPU time in the first step was large in comparison to the subsequent iterations because the matrix product $\mathbf{K}^*\mathbf{K}$ was precomputed before starting the main iteration and the CPU time for the computation of the matrix product was included in the first step. In this test case, the local B-semismooth Newton method `localBSSN` needed the fewest iterations and CPU time followed by Algorithms `hybridBSSN` and `modBSSN`.

A CPU time comparison of Algorithms `hybridBSSN` and `modBSSN` using different values of the parameter γ with an interior-point method (`l1ls`) [83], with sparse reconstruction for separable approximation (`SpaRSA`) [142] and with a Barzilai-Borwein gradient projection method (`GPSR_BB`) [44] for increasing numbers $n = 300, 600, 1250, 2500$ of unknowns is presented in Figure 4.3. We used MATLAB® implementations of these three algorithms from the web page `www.lx.it.pt/~mtf/SpaRSA` (30 June 2015). For each n, one noise vector was used. The data contained 5% of noise, i.e., $\delta = 0.05$, and the starting vector was $\mathbf{u}^{(0)} = \mathbf{0}$ in all algorithms. The CPU time results depend strongly on the expected accuracy of the reconstruction. The stopping criterion for all B-semismooth Newton methods was again a residual norm smaller than $tol = 10^{-7}$. Algorithms `l1ls`, `SpaRSA` and `GPSR_BB` were stopped once the Tikhonov functional value $J(\mathbf{u}^{(j)})$ of the iterates was smaller than

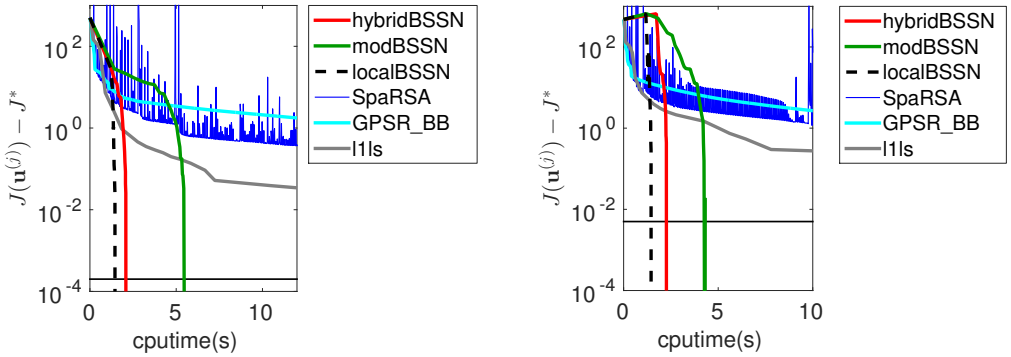


Figure 4.4: Convergence history of the difference of the Tikhonov functional values at the iterates to the Tikhonov functional value J^* at convergence for different algorithms for the inverse integration example with $n = 2000$. Left: $\delta = 0.01$ and $w = 0.9^{66} \approx 9.55 \cdot 10^{-4}$. Right: $\delta = 0.05$ and $w = 0.9^{51} \approx 0.0046$.

or equal to the threshold $J^* + 2\delta^2$, where J^* denotes the Tikhonov functional value of the approximation computed with Algorithm `hybridBSSN` with $\gamma = 10^5$, see also Figure 4.4. This stopping criterion was also used in our preprint [67]. It is motivated by the fact, that ℓ_1-Tikhonov regularization has a linear convergence rate if the regularization parameters are chosen by the discrepancy principle [57]. Linear convergence means that there exists a positive constant c with $\|\mathbf{u}^*_{w,\delta} - \mathbf{u}^\dagger\| \leq c\delta\|\mathbf{f}\|$, where $\mathbf{u}^*_{w,\delta}$ denotes the ℓ_1-Tikhonov regularized solution with regularization parameter w and relative noise level δ in the perturbed right-hand side $\mathbf{f}^\delta$, and $\mathbf{u}^\dagger$ denotes the solution to the operator equation $\mathbf{K}\mathbf{u} = \mathbf{f}$ with unperturbed right-hand side $\mathbf{f}$. Because no higher accuracy of the regularized solution is expected in case of measurement errors, it suffices to solve the minimization problem (2.1) up to an accuracy of $\mathcal{O}(\delta^2)$, $\delta \to 0$, motivating the choice of the above threshold, see also Figure 4.4. The chosen regularization parameters were $w = 0.9^{52} \approx 0.0042$ in the cases $n = 300$, 600, 1250 and $w = 0.9^{51} \approx 0.0046$ in the case $n = 2500$. The size of the parameter γ had a strong influence on the computation time of the B-semismooth Newton methods. For small values of γ, the index sets in Algorithm `hybridBSSN` (here with $k_{max} = 400$) were switched to the modified index sets in the case $n = 1250$ choosing $\gamma = 100$ (in step $j = 401$) and in the case $n = 2500$ choosing $\gamma = 100$ (in step $j = 222$) and choosing $\gamma = 1000$ (in step $j = 401$). A suitable choice of the parameter k_{max} depends on the number of unknowns, on the parameter γ, on the regularization parameter w and on the considered example. For sufficiently large parameters γ, Algorithm `hybridBSSN` was usually efficient without switching to the modified index sets in the experiments presented in this chapter. For small values of γ, the considered B-semismooth Newton methods had a larger computation time than the considered state-of-the-art, linearly convergent methods. Note that the computation times depend on the considered example. In our preprint [67], we discussed the example of horizontal motion blur. There, the CPU time of Algorithm `modBSSN` was larger than the CPU time of `SpaRSA` and `GPSR_BB`.

Figure 4.4 illustrates the history of the sequences $\{J(\mathbf{u}^{(j)}) - J^*\}_j$ in the algorithms considered in Figure 4.3 for 1% as well as for 5% noisy data. Here, J^* denotes the Tikhonov functional value of the reconstruction computed with Algorithm `hybridBSSN`. The pa-

rameter γ in the B-semismooth Newton methods `localBSSN` (cf. Figure 4.2), `hybridBSSN` and `modBSSN` was chosen equal to 10^6 and the number of unknowns was $n = 2000$. The threshold $2\delta^2$ used in the computations of Figure 4.3 is marked by a black thin line in each plot. Once again, Algorithm `localBSSN` was faster than Algorithm `hybridBSSN` and Algorithm `modBSSN` with respect to CPU time. Nevertheless, Algorithm `localBSSN` is not ensured to converge globally.

4.2 Nonlinear parameter identification

In this section, we deal with a nonlinear parameter identification problem considered, e.g., in [93, 101, 103, 104]. Applications of the considered inverse problem come from stationary heat conduction or groundwater flow, see, e.g., [38,64]. This problem has been intensively studied in the literature, see [64,69,84,93,102,150] and the references therein. We follow the outline of the problem description from [93,101,103].

In the following, we are concerned with the elliptic boundary value problem

$$\begin{aligned} -\operatorname{div}(a\nabla u) &= f, \quad \text{in } \Omega := \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 < 1\}, \\ u &= 0, \qquad\qquad\qquad\quad \text{on } \partial\Omega, \end{aligned} \tag{4.3}$$

with a fixed right-hand side $f \in L_2(\Omega)$. The forward problem here is solving the partial differential equation (4.3), i.e., computing u, for a given parameter $a \in L_\infty(\Omega)$ and the right-hand side f. The *parameter-to-solution map* K is given by

$$K\colon A \subset L_\infty(\Omega) \to H_0^1(\Omega), \quad a \mapsto u,$$

where the *admissible set* A is defined as

$$A := \{a \in L_\infty(\Omega) : 0 < \underline{a} \leq a \leq \overline{a} \text{ a.e. in } \Omega, \quad \operatorname{supp}(a - a_0) \subsetneq \Omega\},$$

with constants $\overline{a} \geq \underline{a} > 0$, see, e.g., [69,93,101–103]. Here, one assumes that the parameters $a \in A$ differ from a constant function $a_0 \in L_\infty(\Omega)$ only on a small strict subset of Ω. It is well-known that the boundary value problem (4.3) has a unique weak solution $u \in H_0^1(\Omega)$ for each fixed $a \in A$ [39,65], i.e., u solves the variational problem

$$\int_\Omega a\nabla u \cdot \nabla\varphi \,\mathrm{d}\mathbf{x} = \int_\Omega f\varphi \,\mathrm{d}\mathbf{x}, \tag{4.4}$$

for all test functions $\varphi \in H_0^1(\Omega)$.

Our aim is to solve the inverse problem, i.e., to reconstruct the parameter a from (noisy) measurements u^δ of u, given in the whole domain Ω. We can formulate the inverse problem as solving the nonlinear (ill-posed) operator equation

$$K(a) = u^\delta, \tag{4.5}$$

see [64,69,93,103]. In the one-dimensional case with $\Omega = (0,1)$,

$$-(a(x)u'(x))' = f(x), \quad u(0) = u(1) = 0,$$

where a is, e.g., piecewise constant, there exists an explicit solution formula for the coefficient a,

$$a(x) = \frac{1}{u'(x)}\Big(a(0)u'(0) - \int_0^x f(t)\mathrm{d}t\Big),$$

if $u'(x) \neq 0$ for all $x \in [0,1]$, see [38]. As pointed out in loc. cit., one has to differentiate the measurement data u and to divide by the differentiated data u' in order to reconstruct the coefficient a from the data u. Hence, the inverse problem is nonlinear and ill-posed. Moreover, if u' is equal to zero at a point x, we are not able to reconstruct the coefficient a at x from the measured data u.

In order to apply Tikhonov regularization, we choose a discrepancy term g that was proposed, e.g., in [69, 84, 93, 101–103, 150],

$$g\colon A \to \mathbb{R}, \quad g(a) := \int_\Omega a\|\nabla K(a) - \nabla u^\delta\|_2^2 \,\mathrm{d}\mathbf{x}. \tag{4.6}$$

The choice of this functional g was motivated by *energy functional approaches*. It was shown in [69, 84, 101, 102] that g is twice Fréchet differentiable with a Lipschitz continuous first Fréchet derivative. The first and second Fréchet derivatives of g are given by

$$\begin{aligned} Dg(a)h &= -\int_\Omega h\big(\|\nabla K(a)\|_2^2 - \|\nabla u^\delta\|_2^2\big)\,\mathrm{d}\mathbf{x}, \\ D^2g(a)(h,h') &= 2\int_\Omega a\nabla DK(a)h \cdot \nabla DK(a)h'\,\mathrm{d}\mathbf{x}, \end{aligned}$$

see loc. cit. It is well-known that the Fréchet derivative $DK(a) \in L(L_\infty(\Omega), H_0^1(\Omega))$ of the solution operator K from (4.5) is the solution operator to a boundary value problem with a modified right-hand side compared to (4.3), see, e.g., [69, Lemma 2.2], [102, Lemma 4] and [64]. More precisely, for $a \in A$ and $h \in L_\infty(\Omega)$, $v := DK(a)h \in H_0^1(\Omega)$ is the weak solution to the boundary value problem

$$\begin{aligned} -\operatorname{div}(a\nabla v) &= \operatorname{div}(h\nabla K(a)), && \text{in } \Omega, \\ v &= 0, && \text{on } \partial\Omega. \end{aligned}$$

Hence, using integration by parts, v fulfills the variational problem

$$\int_\Omega a\nabla v \cdot \nabla\varphi\,\mathrm{d}\mathbf{x} = -\int_\Omega h\nabla K(a)\cdot\nabla\varphi\,\mathrm{d}\mathbf{x},$$

for all test functions $\varphi \in H_0^1(\Omega)$. It was shown in [102, Lemma 6], [69, Lemma 2.3] and [84] that the second Fréchet derivative of g is uniformly bounded on A in the operator norm and that, for $h \in L_\infty(\Omega)$, the second Fréchet derivative of g is nonnegative,

$$D^2g(a)(h,h) = 2\int_\Omega a\|\nabla DK(a)h\|_2^2\,\mathrm{d}\mathbf{x} \geq 0.$$

Hence, it was shown that g is convex on A.

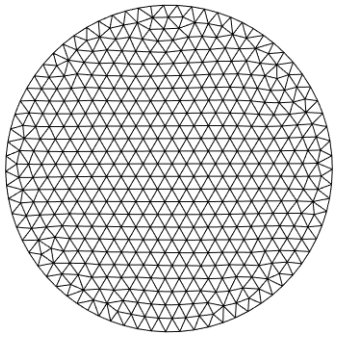


Figure 4.5: Uniform mesh on the unit circle with $n = 500$ nodes.

Using a suitable penalty term $\Phi_{\mathbf{w}}$, where $\mathbf{w} = (w_k)_k$ is a sequence of regularization parameters with $w_k \geq w_0 > 0$, we consider the minimization problem

$$\min_{a \in A} g(a) + \Phi_{\mathbf{w}}(a - a_0).$$

We have in mind a weighted norm as penalty term $\Phi_{\mathbf{w}}$. After a variable transformation $\tilde{a} = a - a_0$, and defining

$$J \colon L_2(\Omega) \to \mathbb{R}, \quad J(\tilde{a}) := \begin{cases} g(\tilde{a} + a_0) + \Phi_{\mathbf{w}}(\tilde{a}), & \tilde{a} \in \tilde{A} \cap \operatorname{dom} \Phi_{\mathbf{w}}, \\ \infty, & \text{else}, \end{cases}$$

where

$$\tilde{A} := \{\tilde{a} \in L_\infty(\Omega) : 0 < \underline{a} \leq \tilde{a} + a_0 \leq \overline{a} \text{ a.e. in } \Omega, \quad \operatorname{supp}(\tilde{a}) \subsetneq \Omega\},$$

we are concerned with the minimization problem

$$\min_{\tilde{a} \in L_2(\Omega)} J(\tilde{a}), \tag{4.7}$$

cf. [93, 101, 103]. If $\tilde{a}$ is a minimizer of the minimization problem (4.7), then $\tilde{a} + a_0$ is a regularized solution to the inverse problem (4.5). Note that we now consider the space $L_2(\Omega)$ instead of $L_\infty(\Omega)$. It was shown in [102, Lemma 6] that g is continuous on A with respect to the L_2-norm. To our best knowledge, it is up to now unknown if Assumption A holds with respect to the L_2-norm.

The boundary value problem (4.3) is discretized using the *finite element method*, see, e.g., [65, 150]. We used piecewise linear finite elements for a, u and f, respectively. The triangulations $\mathcal{T}$ on the unit ball Ω were generated using the mesh generator `distmesh` [110] that is available on the web page `http://persson.berkeley.edu/distmesh` (17 May 2016). An example of a uniform mesh with $n = 500$ nodes is shown in Figure 4.5. Moreover, we used an implementation of Dr. Fabrice Delbary (University of Mainz, Institute of mathematics) to solve the forward problem. In the following, we derive a discretization of the minimization problem (4.7). We proceed similarly to [150]. Let $V_h^{\mathcal{T}} := \operatorname{span}\{\Lambda_1, \ldots, \Lambda_n\}$ denote the space of piecewise linear ansatz functions on a fixed triangulation $\mathcal{T}$ with n nodes $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $V_h^{\mathcal{T},0} := V_h^{\mathcal{T}} \cap H_0^1(\Omega)$. The linear interpolant of the data $a_k \in \mathbb{R}$, $k = 1 \ldots, n$, given at the nodes $\mathbf{x}_k$ is denoted by $a_h = \sum_{k=1}^n a_k \Lambda_k \in V_h^{\mathcal{T}}$.

With the synthesis operator $T\colon \mathbb{R}^n \to V_h^{\mathcal{T}}$, $T\mathbf{a} := \sum_{k=1}^n a_k \Lambda_k$, the discretized minimization problem is then given by

$$\min_{\tilde{\mathbf{a}} \in \mathbb{R}^n} J_h(\tilde{\mathbf{a}}),$$

where

$$J_h\colon \mathbb{R}^n \to \mathbb{R}, \quad J_h(\tilde{\mathbf{a}}) := \begin{cases} (g_h \circ T)(\tilde{\mathbf{a}} + a_0 \mathbb{1}) + \|\tilde{\mathbf{a}}\|_{1,\mathbf{w}}, & \tilde{\mathbf{a}} > -a_0 \mathbb{1}, \\ \infty, & \text{else}, \end{cases}$$

the weighted norm $\|\tilde{\mathbf{a}}\|_{1,\mathbf{w}} := \sum_{k=1}^n w_k |\tilde{a}_k|$, the vector of ones $\mathbb{1} := (1, \dots, 1)^\top \in \mathbb{R}^n$ and

$$g_h\colon V_h^{\mathcal{T}} \to \mathbb{R}, \quad g_h(a_h) := \int_\Omega a_h \|\nabla u_h - \nabla u_h^\delta\|_2^2 \, \mathrm{d}\mathbf{x}.$$

Here, $u_h, u_h^\delta \in V_h^{\mathcal{T},0}$ denote the Galerkin approximations of $u = K(a_h)$ and u^δ, respectively. Hence, we consider the finite-dimensional zero-finding problem

$$\mathbf{F}(\tilde{\mathbf{a}}) = \mathbf{0}, \tag{4.8}$$

where $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{F}(\tilde{\mathbf{a}}) := \tilde{\mathbf{a}} - \mathbf{S}_{\gamma \mathbf{w}}(\tilde{\mathbf{a}} - \gamma D g_h(T(\tilde{\mathbf{a}} + a_0 \mathbb{1})))$, Dg_h is the discretization of Dg and $\tilde{\mathbf{a}} = (\tilde{a}_k)_k$ denotes the sequence of basis coefficients of $\tilde{a}_h = T\tilde{\mathbf{a}}$. In the following, we choose a constant sequence of regularization parameters $w_k \equiv w$.

As proposed in [93,101,103], the noisy data $u_h^\delta \in H_0^1(\Omega)$ were, in our numerical experiments, computed as the numerical solution to the weak formulation (4.4) of the boundary value problem (4.3) with unperturbed parameter a and noisy right-hand side $f^{\tilde{\delta}}$, i.e., we had $\|f - f^{\tilde{\delta}}\|_{L_2(\Omega)} = \tilde{\delta} \|f\|_{L_2(\Omega)}$. Here, if not stated otherwise, we chose $\tilde{\delta} = 0.2$. Then, the relative noise level δ of u_h^δ was given by

$$\delta = \frac{\|u_h^\delta - u_h\|_{H^1(\Omega)}}{\|u_h\|_{H^1(\Omega)}}.$$

Note that for the numerical solution of the inverse problem, i.e., the nonlinear operator equation (4.5), we used the unperturbed right-hand side f in (4.3). The noisy data $u_h^\delta \in V_h^{\mathcal{T},0}$ were computed on a slightly finer mesh than the mesh in the inverse problem. Nevertheless, we used the same model for the inverse problem as for the forward problem. It is necessary to test the considered algorithms also for data u_h^δ that were not synthetically produced with the same model as the inverse problem is solved with, see [22, p. 154] and [68, Section 7.2]. Hence, up to now, we cannot exclude an *inverse crime*. The study of the influence of model mismatches on the results of the considered algorithms is up to future research.

For the B-semismooth Newton methods considered below, it is not ensured that the iterates $\tilde{\mathbf{a}}^{(j)}$ stay componentwise larger than $-a_0$ and thus that $\tilde{\mathbf{a}}^{(j)} + a_0 \mathbb{1}$ stays positive, even if starting with an initial guess $\tilde{\mathbf{a}}^{(0)}$ that is componentwise bounded away from $-a_0$. Therefore, the solvability of the forward problem is not ensured in this case. We decided to choose a_0 large enough, here $a_0 = 3$. For this choice, the sequence of iterates in the considered algorithms usually converged if choosing the zero vector as initial guess. Nevertheless, the sequence of iterates in the algorithms did not always converge, especially if

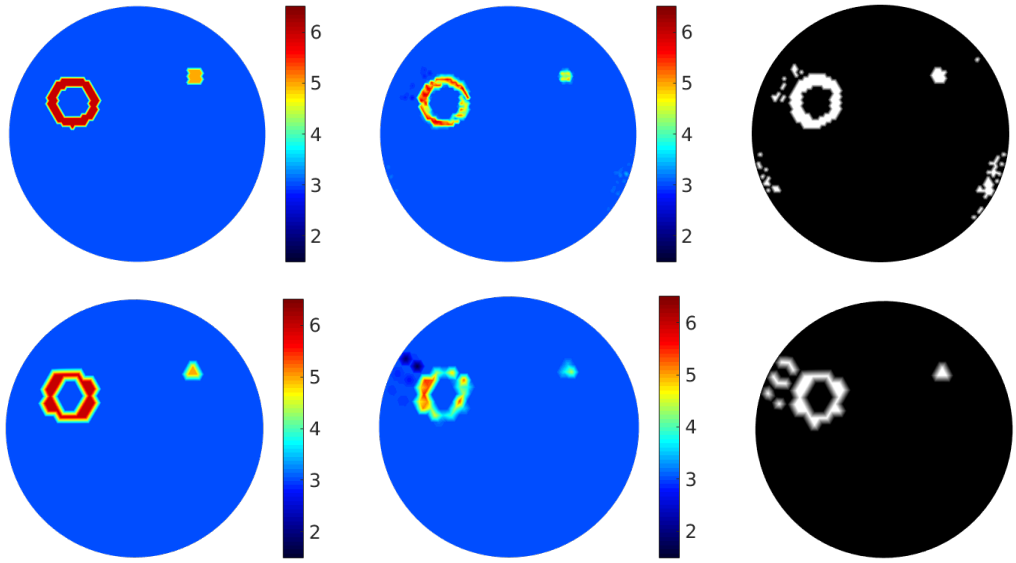


Figure 4.6: The true parameter a and reconstructions from 1.3% noisy data on different meshes. Upper row: mesh with $n = 5000$ nodes. Lower Row: mesh with $n = 1000$ nodes. From left to right: true parameter a, reconstruction, sparsity pattern of the reconstruction (black: $a \equiv a_0$, white: $a \neq a_0$).

the starting vectors were not close to the zero of $\mathbf{F}$. In some experiments presented here, the iterates contained some components smaller than $-a_0$, at least for early iterates. We only present results where the sequence of iterates in the algorithms did converge, even though some entries possibly were smaller than $-a_0$ in some iterates. Up to now, feasibility of the considered algorithms is not always guaranteed in this example. Feasibility might be achieved by a modification of the Armijo rule (2.42) and (3.37), respectively, but this is part of future work. In the results presented in this section, we always chose $a_0 = 3$ and $\mathbf{u}^{(0)} = \mathbf{0}$.

Let us give an overview of the subsequent paragraphs. The true right-hand side, some noisy data and reconstructions are presented in Figures 4.6, 4.7 and 4.8. The convergence history of one run of Algorithms `hybridBSSN` and `modBSSN` is shown in Tables 4.9 and 4.10, respectively. Table 4.11 treats the influence of the parameter γ on the computational cost of the considered algorithms. The influence of the choice of the regularization parameter on the sparsity of the computed reconstructions as well as on the computational cost of Algorithms `hybridBSSN` and `modBSSN` is treated in Figure 4.9.

The true parameter a, a reconstruction from noisy data as well as the sparsity pattern of the reconstruction on a mesh with $n = 5000$ and $n = 1000$ nodes, respectively, is presented in Figure 4.6. In the plots on the right-hand side, the unit circle is black in regions where the reconstruction a is equal to $a_0 = 3$, whereas it is colored white in regions where the reconstruction differs from a_0. Here, the true parameter was

$$a(x_1, x_2) = \begin{cases} 6, & \frac{1}{8} \leq \|\mathbf{x} - (-\frac{1}{2}, \frac{1}{4})^\top\|_2 \leq \frac{1}{5}, \\ 5, & \frac{2}{5} \leq x_1 \leq \frac{1}{2}, \frac{2}{5} \leq x_2 \leq \frac{1}{2}, \\ 3, & \text{else.} \end{cases}$$

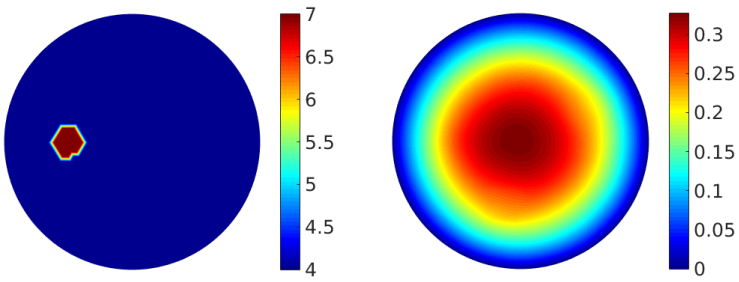


Figure 4.7: Left: true right-hand side f on a mesh with $n = 2000$ nodes. Right: noisy data u^δ with $\delta \approx 0.0181$.

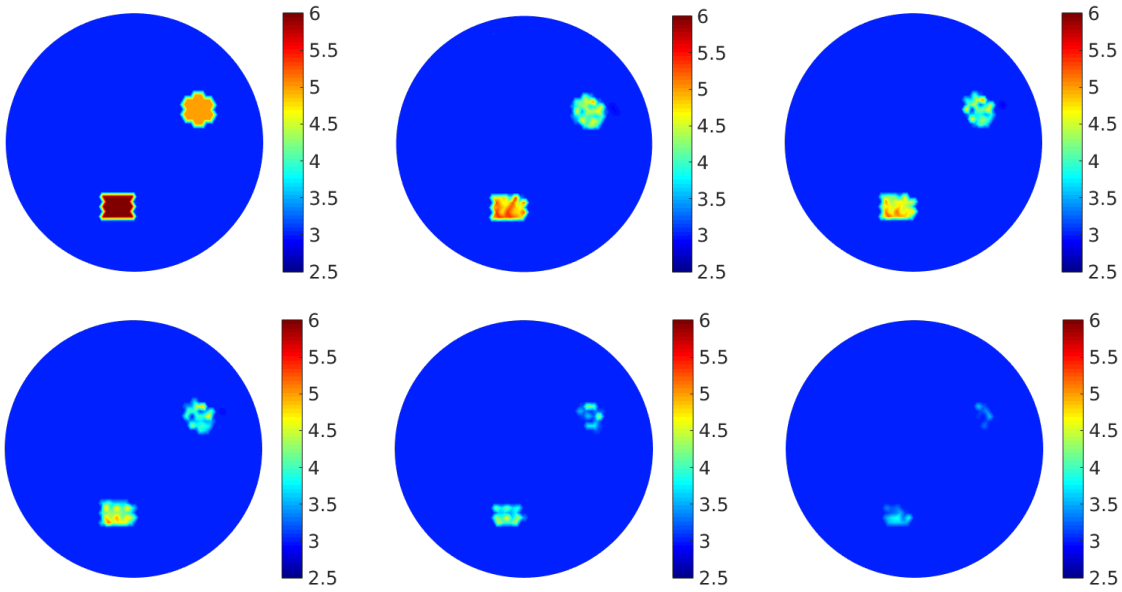


Figure 4.8: From left to right and top to bottom: true parameter a on a mesh with $n = 2000$ nodes, reconstructions computed with regularization parameters $w = 3 \cdot 10^{-5}$, $4 \cdot 10^{-5}$, $5 \cdot 10^{-5}$, $7.5 \cdot 10^{-5}$, 10^{-4} from 1.81% noisy data u^δ.

The right-hand side f was chosen constant, $f(x_1, x_2) \equiv 4$. On the finer mesh, the true parameter a can be approximated in more detail. For instance, the rectangle is better represented on a mesh with $n = 5000$ nodes than on the coarse mesh with $n = 1000$ nodes. A larger number of nodes increases the resolution. The relative noise contained in the data was $\delta \approx 1.3\%$ in both cases. The parameters $\tilde{\delta}$ were chosen as 0.2 in the case $n = 5000$ and as 0.1 in the case $n = 1000$. Note that the quality of the reconstructions is expected to be lower if the relative noise in the data is larger. Both reconstructions were computed with Algorithm `hybridBSSN` with parameter $\gamma = 10^5$. The regularization parameters were chosen as $w = 10^{-5}$ in the case $n = 5000$ and $w = 5 \cdot 10^{-5}$ in the case $n = 1000$, respectively. The images on the right-hand side of Figure 4.6 show that the reconstructions of the parameter a here also differed from the background value $a_0 = 3$ near the ring and near the boundary of the unit circle.

Table 4.9: Convergence history of Algorithm `hybridBSSN` for the nonlinear parameter identification problem using a mesh with $n = 1000$ nodes, 2.6% noisy data, regularization parameter $w = 10^{-4}$, $k_{max} = 100$ and $\gamma = 10^5$.

j	$\|\mathbf{F}(\tilde{\mathbf{a}}^{(j)})\|_2$	$J(\tilde{\mathbf{a}}^{(j)}) - J^*$	$\|\tilde{\mathbf{a}}^{(j)} - \tilde{\mathbf{a}}^*\|_2$	$\mathcal{A}$	$\mathcal{I}^\circ$	$\mathcal{I}^+$	$\mathcal{I}^-$	sizeLCP	size SLE	#SLE	$\text{cond}(\mathbf{M}_{\mathcal{A},\mathcal{A}})$	t_j
0	4.95e+01	9.96e-04	5.07e+00	39	961	0	0	-	-	-	-	-
1	1.04e+01	2.05e-03	5.41e+00	31	969	0	0	0	39	1	1.79e+02	1
2	2.31e+00	1.15e-03	2.87e+00	20	980	0	0	0	31	1	1.42e+02	1
3	7.55e-02	7.45e-09	2.84e-02	20	980	0	0	0	20	1	7.58e+01	1
4	6.92e-05	5.88e-15	2.65e-05	20	980	0	0	0	20	1	7.52e+01	1
5	7.81e-11	-7.98e-17	3.27e-11	20	980	0	0	0	20	1	7.53e+01	1

In the results presented in the rest of this section, we considered the true parameter a,

$$a(x_1, x_2) = \begin{cases} 5, & \|\mathbf{x} - (\frac{1}{2}, \frac{1}{4})^\top\|_2 \leq \frac{1}{8}, \\ 6, & -\frac{1}{4} \leq x_1 \leq 0, -\frac{7}{12} \leq x_2 \leq -\frac{5}{12}, \\ 3, & \text{else}, \end{cases} \tag{4.9}$$

and the right-hand side f,

$$f(x_1, x_2) = \begin{cases} 7, & \|\mathbf{x} - (-\frac{1}{2}, 0)^\top\|_2 \leq \frac{1}{8}, \\ 4, & \text{else}. \end{cases} \tag{4.10}$$

Figure 4.7 shows the unperturbed right-hand side f on a mesh with $n = 2000$ nodes as well as the noisy data u^δ on the same mesh. Here, we chose $\tilde{\delta} = 0.2$ and the relative noise in u^δ was $\delta \approx 0.0181$. The true parameter a and five reconstructions, computed with the noisy data from Figure 4.7 using five different regularization parameters w, are presented in Figure 4.8. The reconstructions were produced with Algorithm `hybridBSSN` choosing $\gamma = 1000$ and $k_{max} = 100$. The relative number of nonzero basis coefficients of the true parameter a in the linear finite element basis on the mesh with $n = 2000$ nodes was $\#\{k : a_k \neq 3\}/2000 \approx 0.0295$. The reconstructions were computed with the regularization parameters $w = 10^{-4}$, $7.5 \cdot 10^{-5}$, $5 \cdot 10^{-5}$, $4 \cdot 10^{-5}$ and $3 \cdot 10^{-5}$. Here, the relative numbers of nonzero entries of the reconstructions were equal to 0.012, 0.0205, 0.0275, 0.030 and 0.034, respectively. For smaller regularization parameters w, the reconstructions were more densely populated, compare also Figure 4.9.

Table 4.9 and Table 4.10 show the convergence history of Algorithms `hybridBSSN` and `modBSSN`, respectively. Here, a mesh with $n = 1000$ nodes and noisy data u^δ with $\delta \approx 0.0261$ ($\tilde{\delta} = 0.2$) were used. In both algorithms, we chose the parameter $\gamma = 10^5$ and in Algorithm `hybridBSSN` we set $k_{max} = 100$. The regularization parameter was $w = 10^{-4}$. The relative number of entries different from $a_0 = 3$ of the true parameter was approximately 0.027. The computed reconstruction contained approximately 2% of entries that did not coincide with a_0. Table 4.9 and Table 4.10 present the same quantities as Tables 4.2 and 4.3 from Section 4.1. The condition numbers of the system matrices were here computed with the MATLAB® subroutine `condest`. The sequences of iterates in the algorithms converged within 5 and 6 steps, respectively, and all step sizes t_j were equal to

Table 4.10: Convergence history of Algorithm `modBSSN` for the nonlinear parameter identification problem using a mesh with $n = 1000$ nodes, 2.6% noisy data, regularization parameter $w = 10^{-4}$ and $\gamma = 10^5$.

j	$\|\mathbf{F}(\tilde{\mathbf{a}}^{(j)})\|_2$	$J(\tilde{\mathbf{a}}^{(j)}) - J^*$	$\|\tilde{\mathbf{a}}^{(j)} - \tilde{\mathbf{a}}^*\|_2$	$\overline{\mathcal{A}}$	$\overline{\mathcal{I}^\circ}$	$\overline{\mathcal{I}^+}$	$\overline{\mathcal{I}^-}$	sizeLCP	size SLE	#SLE	$\mathrm{cond}(\mathbf{M}_{\overline{\mathcal{A}},\overline{\mathcal{A}}})$	t_j
0	4.95e+01	9.96e-04	5.07e+00	39	961	0	0	-	-	-	-	-
1	1.04e+01	2.05e-03	5.41e+00	24	961	11	4	0	39	1	1.95e+02	1
2	1.33e+00	4.14e-04	1.55e+00	20	975	5	0	15	24	977	9.56e+01	1
3	4.03e-01	2.13e-06	5.31e-01	20	980	0	0	5	20	981	7.47e+01	1
4	9.10e-03	4.92e-11	1.22e-03	20	980	0	0	0	20	1	7.52e+01	1
5	6.84e-07	1.02e-16	1.06e-07	20	980	0	0	0	20	1	7.53e+01	1
6	2.72e-12	0	0	20	980	0	0	0	20	1	7.53e+01	1

1. The residual norms decreased strictly and the sequences if iterates in the algorithms were locally superlinearly convergent, cf. the decrease of the norms $\|\tilde{\mathbf{a}}^{(j)} - \tilde{\mathbf{a}}^*\|_2$. Here $\tilde{\mathbf{a}}^*$ denotes the approximation of the root of $\mathbf{F}$ computed with Algorithm `modBSSN`. The sizes of the linear systems decreased in the course of the iteration. As opposed to Algorithm `hybridBSSN`, some linear complementarity problems had to be solved in Algorithm `modBSSN`. However, these systems as well as the linear systems were small compared to the system size $n = 1000$.

The computational cost of the B-semismooth Newton methods depends on the choice of the parameter γ, cf. the example from Section 4.1. In Table 4.11, the average CPU time of five runs, the number of iterations as well as the number of step sizes equal to 1 are listed for different values of γ, both Algorithms `hybridBSSN` and `modBSSN` and two different regularization parameters $w = 10^{-4}$ and $w = 5 \cdot 10^{-5}$. The number of unknowns was $n = 2000$, the parameter k_{max} in Algorithm `hybridBSSN` was chosen equal to 100 and the data contained 1.79% of noise ($\tilde{\delta} = 0.2$). The computational cost of both algorithms seriously depended on the choice of the parameter γ. In this example, the algorithms performed best for the choices $\gamma = 10^5$ and $\gamma = 10^6$. With the choice $\gamma = 10^6$, here, both methods coincided with the local method because all step sizes were equal to 1. For small values of γ, the methods were slower in terms of computation time, especially when choosing the smaller regularization parameter $w = 5 \cdot 10^{-5}$. In some cases, Algorithm `modBSSN` needed less computation time than Algorithm `hybridBSSN` with the same parameter γ. This can be the case if either less Newton steps are computed in Algorithm `modBSSN` than in Algorithm `hybridBSSN`, if the sizes of the active sets in Algorithm `modBSSN` are smaller than in Algorithm `hybridBSSN` or if the damping parameters in Algorithm `modBSSN` are larger on average than the step sizes in Algorithm `hybridBSSN`. The step sizes are chosen according to the Armijo rule (2.42) and (3.37), respectively. In each trial of the Armijo rule, one function evaluation of the nonlinearity $\mathbf{F}$ is necessary. This involves the computation of the gradient of the discrepancy term including one solution of the forward problem. For small damping parameters, many trials of the Armijo rule are performed. When choosing $\gamma = 10$, the index sets of the iterates in Algorithm `hybridBSSN` were switched to the modified index sets in both cases $w = 10^{-4}$ and $w = 5 \cdot 10^{-5}$. Some early iterates in Algorithm `hybridBSSN` contained some entries smaller than $-a_0$ in the cases $w = 5 \cdot 10^{-5}$, $\gamma = 10^2$ or $\gamma = 10^4$, respectively. Nevertheless, the sequences of

Table 4.11: Influence of the parameter γ on the computational cost of Algorithm `hybridBSSN` and Algorithm `modBSSN` for the nonlinear parameter identification problem with $n = 2000$ unknowns and $\delta = 0.0179$.

$w = 10^{-4}$

	`hybridBSSN`			`modBSSN`		
γ	av. cputime(s)	# iter.	$\#\{j : t_j = 1\}$	av. cputime(s)	# iter.	$\#\{j : t_j = 1\}$
10^1	21.34 *1	15 *1	3 *1	21.39	15	3
10^2	27.73	17	3	22.54	18	2
10^3	7.37	7	4	19.87	17	3
10^4	4.37	5	4	18.35	18	4
10^5	4.25	5	5	6.21	7	5
10^6	5.12	6	6	4.32	5	5

$w = 5 \cdot 10^{-5}$

	`hybridBSSN`			`modBSSN`		
γ	av. cputime(s)	# iter.	$\#\{j : t_j = 1\}$	av. cputime(s)	# iter.	$\#\{j : t_j = 1\}$
10^1	189.77 *2	111 *2	3 *2	181.93	110	3
10^2	89.88	63	3	74.63	54	3
10^3	29.88	26	4	46.02	38	3
10^4	12.16	12	5	18.46	17	4
10^5	6.07	7	7	9.13	10	7
10^6	6.91	8	8	6.31	7	7

*1 Algorithm `hybridBSSN` switched to the modified index sets in step $j = 1$.
*2 Algorithm `hybridBSSN` switched to the modified index sets in step $j = 2$.

iterates in both methods converged in all presented experiments.

The blue axis in Figure 4.9 shows the dependence of the relative number $\#\{k : \tilde{a}_k \neq 0\}/n$ of nonzero entries of the approximation $\tilde{\mathbf{a}}$ of the zero of $\mathbf{F}$ computed with Algorithm `hybridBSSN` and Algorithm `modBSSN`, respectively, on the regularization parameter w. Additionally, on the second axis, the average CPU time of five runs as well as the number of iterations until convergence are presented for the Algorithms `hybridBSSN` and `modBSSN` with different choices of the parameter γ. As before, we chose $n = 2000$, $k_{max} = 100$, $\tilde{\delta} = 0.2$ and the noise level was $\delta = 0.0179$. We considered $w = \frac{1}{4} \cdot 1.25^l \cdot 10^{-4}$, $l = 1, \cdots, 14$. Using regularization parameters larger than $w \approx 1.4901 \cdot 10^{-4}$, the computed reconstructions were equal to $a_0 \cdot \mathbb{1} = 3 \cdot \mathbb{1}$ (i.e., $\tilde{\mathbf{a}} = \mathbf{0}$), so that the starting vector $\tilde{\mathbf{a}}^{(0)} = \mathbf{0}$ of the algorithms already coincided with the zero of $\mathbf{F}$. The reconstructions were less sparse when choosing smaller regularization parameters w. As remarked in our preprint [67], a *sparsity-based* parameter choice strategy for the computation of the regularization parameter proposed by Kolehmainen et al. [86] could be applied here. In this very strategy, one assumes that the relative sparsity of the unknown solution is known. The authors of loc. cit. propose to compute the relative sparsity of several reconstructions using different regularization parameters w and to choose w according to the expected sparsity of the unknown solution, see also [62] for the *S-curve method*. In our experiments, the true parameter a contained 2.95% of entries that differed from $a_0 = 3$. Therefore, $w \approx 4.8828 \cdot 10^{-5} = \frac{1}{4} \cdot 1.25^3 \cdot 10^{-4}$ could be an appropriate candidate for a suitable regularization parameter. In [86], it was pointed out that, in practice, it could be

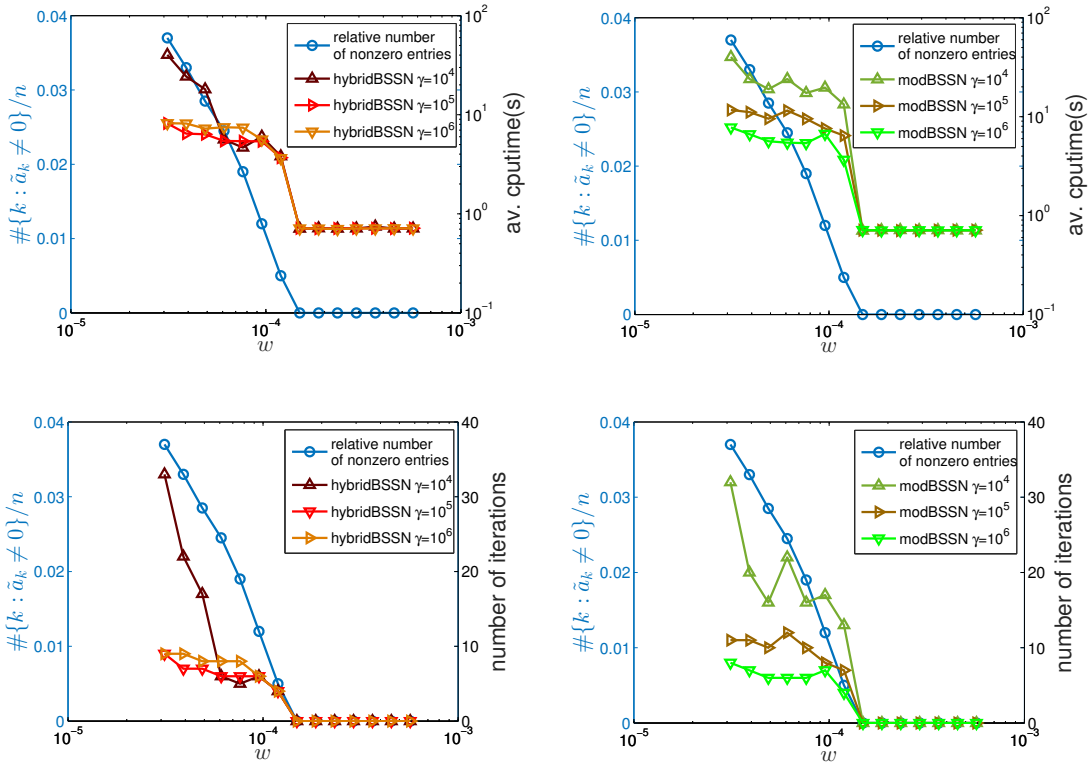


Figure 4.9: Relative sparsity of the approximation $\tilde{\mathbf{a}}$ of the zero of $\mathbf{F}$, average CPU time and number of iterations of Algorithms `hybridBSSN` and `modBSSN` for the nonlinear parameter identification problem with $n = 2000$ unknowns and 1.79% noisy data.

advantageous to consider $\#\{k : |\tilde{a}_k| > \epsilon\}/n$ for a small but positive parameter ϵ instead of the relative sparsity $\#\{k : \tilde{a}_k \neq 0\}/n$ in order to choose the regularization parameter w. In the experiments shown in Figure 4.9, some entries of the iterates were smaller than $-a_0$. This was the case in all methods choosing $w = 3.125 \cdot 10^{-5}$, in all methods beside Algorithm `modBSSN` with $\gamma = 10^4$ choosing $w = 3.9063 \cdot 10^{-5}$, and for $w = 4.8828 \cdot 10^{-5}$ only some iterates in Algorithm `hybridBSSN` with $\gamma = 10^4$ contained some entries smaller than $-a_0$. Algorithm `hybridBSSN` switched to the modified index sets in the case $\gamma = 10^4$ and $w \approx 3.1250 \cdot 10^{-5}$ (in step $j = 18$). The sizes of the support sets of the reconstruction increased for smaller regularization parameters w. Figure 4.9 demonstrates once again the influence of the parameter γ on the computational cost of Algorithms `hybridBSSN` and `modBSSN`. In particular, the choice of γ is important in the case of a small regularization parameter w.

Chapter 5

Conclusion and outlook

In this thesis, we addressed the efficient numerical solution of a nonsmooth minimization problem. This problem appears in ℓ_1-Tikhonov regularization for inverse problems with sparsity constraints. The objective function in the minimization problem is the sum of a smooth, strictly convex functional and a weighted ℓ_1-norm. The optimality conditions of the minimization problem lead to a root-finding problem for a nonlinear mapping. This nonlinearity is not Fréchet differentiable, but it is still Lipschitz continuous. Therefore, we can apply generalized Newton methods to solve the root-finding problem under consideration. The goal of this work was the globalization of the locally superlinearly convergent, semismooth Newton methods from [58, 103] for ℓ_1-Tikhonov regularization.

To this end, we proved, under appropriate assumptions, the Bouligand differentiability of the considered mapping in the sequence space ℓ_2. Furthermore, we discussed the feasibility of a local B-Newton algorithm that turned out to be a special semismooth Newton approach. In contrast to the original algorithms that were introduced by Herzog ans Lorenz [58] and by Muoi, Hào, Maass and Pidcock [103], the generalized Newton equations of the B-semismooth Newton method include (uniquely solvable) linear complementarity problems. Nevertheless, in our numerical experiments, the B-Newton equations were usually linear systems because the linear complementarity problems did not appear. The iterates of the B-semismooth Newton method differ from the iterates of the original methods from loc. cit. only at points where the nonlinearity is not smooth. Therefore, the B-Newton approach inherits advantageous properties from the original semismooth Newton methods. For instance, the algorithm is locally superlinearly convergent, the Newton equations are finite-dimensional in a neighborhood of the (unique) root and a solution to a *linear* inverse problem is found within a finite number of iterations if the initial guess is sufficiently close to the root.

However, local Newton methods may have cycles. In order to derive a globally convergent Newton-type method, we benefited from a descent property due to the B-Newton directions. We proposed a globalized B-semismooth Newton method in a finite-dimensional setting. Here, global convergence is ensured if either the damping parameters stay bounded from below by a positive constant or if the nonlinearity satisfies a smoothness assumption at an accumulation point of the sequence of iterates. A modification of the Newton directions ensured global convergence and enabled us to overcome the drawback that the globalized B-semismooth Newton method may begin to stagnate. Unfortunately, the modified method is less efficient in practice. Here, in most cases, the modified, generalized Newton equation consists of several linear systems and a linear complementarity problem. We proposed a globally convergent, hybrid algorithm that combines both methods and that is efficient in practice. The goal of this work was achieved by the de-

velopment of a globally convergent Newton-type method for ℓ_1-Tikhonov regularization. Numerical results for a linear and a nonlinear inverse problem were discussed. Here, we observed that the choice of the parameter γ that defines the nonlinearity has a crucial impact on the computational cost of the proposed methods.

There are several possibilities for future research. For instance, the proof of global convergence in an infinite-dimensional setting may be addressed, cf. Remark 2.28. Furthermore, the feasibility of the nonlinear parameter identification problem from Section 4.2 could be addressed. That is, the coefficient a in the elliptic boundary value problem (4.3) should be ensured to stay bounded from below by a positive constant, i.e. the iterates $\tilde{\mathbf{a}}^{(j)}$ should be bounded componentwise from below by $-a_0$, where a_0 denotes the background coefficient. Moreover, this nonlinear parameter identification example needs to be analyzed in case that the synthetic data are produced with a forward operator that differs from the forward operator that is used in the numerical solution of the inverse problem, cf. Section 4.2. Furthermore, test case generators for ℓ_1-regularized least-squares, i.e., for linear inverse problems, have been developed in [48, 92]. There, the authors discussed how to control the number of unknowns, the sparsity of the true solution and the conditioning of the matrices for random test examples. These tools may be used to test the proposed algorithms on further problems.

In our numerical experiments, the Hessians of the discrepancy term have been evaluated at each iterate. In order to reduce the computational cost for large-scale problems, one may examine a globally convergent quasi-Newton method by using approximations of the Hessian at the iterates. Quasi-Newton-type methods for nonsmooth problems were proposed, e.g., in [17, 85, 103, 112, 131, 140], see also the references therein. Additionally, one could think of an inexact (matrix-free) solution of the generalized Newton equations. Hence, one may consider a (globally convergent) inexact Newton-type method as proposed for other nonsmooth zero-finding problems, see, e.g., [7, 28, 33, 80, 97] and the references therein.

In case of nonlinear complementarity problems, alternative but equivalent formulations of the optimality conditions turned out to be advantageous for the application of Newton-type methods, see, e.g., [27,28,40,45,46,72,80,81,100] and the references therein. To this end, one can use *NCP-functions*. A function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ is called a NCP-function if φ satisfies the property

$$\varphi(x,y) = 0 \quad \Leftrightarrow \quad x \geq 0, y \geq 0, xy = 0.$$

Famous examples are the *Fischer-Burmeister function* $\varphi(x,y) := \sqrt{x^2+y^2} - x - y$ or the min-function $\varphi(x,y) := \min\{x,y\}$. Using the componentwise representation (3.7), (3.8) of $\mathbf{F}$, the root-finding problem $\mathbf{F}(\mathbf{u}) = \mathbf{0}$ might be reformulated by using another NCP-function. Different representations of zero-finding problems have advantages and disadvantages for associated generalized Newton methods. There exist some reformulations for nonlinear complementarity problems with a merit functional $\boldsymbol{\Theta}(\mathbf{u}) := \|\mathbf{F}(\mathbf{u})\|_2^2$ that is continuously differentiable, see loc. cit. This is an advantage for the development of globally convergent algorithms. Hence, a restatement of the nonlinearity considered in this work may also be a starting point for further research. Nevertheless, the active-set property of a resulting Newton-type method and the finite convergence property for linear inverse problems may be lost. Besides, an adaptation of the parameter γ along

the iteration could be addressed because the performance of the proposed Newton-type methods seriously depends on the choice of γ. A variable choice of the parameter γ was also proposed in [15].

In the following paragraphs, we shortly present a promising acceleration strategy for the B-semismooth Newton methods: the combination of the algorithms proposed in this thesis with an algebraic multilevel method that was introduced in the literature. Recently, Bosse [8] has proposed another nonsmooth multi-grid method. Such multi-grid techniques might also be a starting point for further work.

Acceleration by an algebraic multilevel approach

In *classical multi-grid methods* [13, 59] for the solution of linear systems that for instance appear in the numerical solution of boundary value problems in differential equations, a hierarchy of grids and data transfer operations are introduced. *Restrictions* transfer the data from one grid to a coarser grid and *prolongations* map an approximation given on a coarse grid to a finer grid, e.g., by interpolation. By smoothing iterations, high-frequency error components are reduced. On the coarsest grid, an equation with less unknowns is solved exactly. The *coarse grid correction* reduces low-frequency error components. The computational cost of classical multi-grid iterations can be shown to be proportional to the problem size. Applying multi-grid methods to nonlinear systems, one may either first linearize the nonlinear equation by Newton's method and then solve the linear systems appearing in the Newton equations using linear multi-grid techniques (*Newton-multi-grid method*), or a *nonlinear multi-grid method* is applied to the nonlinear system and then a small nonlinear problem is solved exactly on the coarsest grid using Newton's method [59]. Moreover, the hierarchy of grids can either be defined with respect to the geometry of the problem or can be defined algebraically.

In further research, we are going to combine the globalized B-semismooth Newton methods with a nonlinear algebraic multilevel method that was introduced by Treister, Turek and Yavneh [134, 135]. This algorithm takes advantage of special properties of the considered root-finding problem but it has a different concept than classical multi-grid methods. The idea of this very method is based on the sparsity structure of the unknown solution of the finite-dimensional root-finding problem,

$$\mathbf{F}(\mathbf{u}) = \mathbf{0},$$

with $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$ defined in (1.10), see also (2.11). A hierarchy of levels is defined by algebraic equations that are motivated by the optimality conditions of the nonsmooth minimization problem from ℓ_1-Tikhonov regularization, see Proposition 1.16. The considered levels are independent of the geometry of the problem so that the multilevel method can be used as black box algorithm on any kind of discretization. The idea of this multilevel methods might be transferred to other (nonsmooth) zero-finding problems. The multilevel methods that are presented in this section were introduced by Treister, Turek and Yavneh [134, 135]. Except for the exact solution of a root-finding problem on the coarsest level with Algorithm `hybridBSSN` and the stopping criterion, the methods considered in this section and the algorithms from loc. cit. are almost identical.

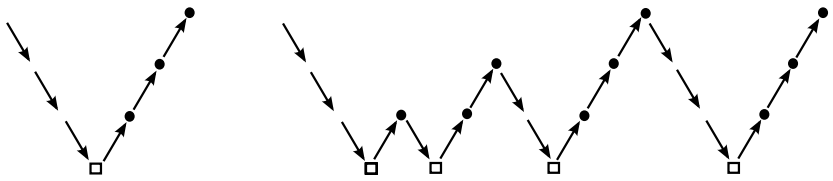


Figure 5.1: Schematic representation of the V-cycle and the F-cycle multilevel method (*modified from [135, Fig. 1]*): $\searrow$ = restriction, $\nearrow$ = prolongation, $\square$ = exact solution on the coarsest level, $\bullet$ = application of one relaxation. Left: one multilevel V-cycle iteration. Right: one multilevel F-cycle iteration.

In [134,135], a level means a set of indices. Treister, Turek and Yavneh defined the finest level 0 as the index set $I_0 := \{1, \ldots, n\}$. The next coarser level includes only those indices that are expected to be contained in the support of the unknown root $\mathbf{u}^*$, i.e., the indices of interest. It follows from Equations (2.12)–(2.16) that $\mathbf{u}^* \in \mathbb{R}^n$ is a zero of $\mathbf{F}$ if and only if we have

$$\begin{aligned} u_k^* &= 0, & k &\in \mathcal{I}(\mathbf{u}^*) = \{k : |(\nabla g(\mathbf{u}^*))_k| \leq w_k\}, \\ (\nabla g(\mathbf{u}^*))_k &= -w_k \operatorname{sgn}(u_k^*), & k &\in \mathcal{A}(\mathbf{u}^*) = \{k : u_k^* \neq 0\}. \end{aligned}$$

Hence, one is only interested in those indices that are included in the active set $\mathcal{A}(\mathbf{u}^*) = \{k : u_k^* \neq 0\}$ of the zero $\mathbf{u}^*$ of $\mathbf{F}$ because it holds $u_k^* = 0$ for all $k \in \mathcal{I}(\mathbf{u}^*)$. Nevertheless, the active set of the root $\mathbf{u}^*$ of $\mathbf{F}$ is unknown. In the first iteration, if starting with an initial guess $\mathbf{u}^{(0)} \in \mathbb{R}^n$, those indices k for which the k-th component of the gradient $\nabla g(\mathbf{u}^{(0)})$ has a relatively large absolute value are presumably contained in $\mathcal{A}(\mathbf{u}^*)$, see [134, 135]. For that reason, on the level i, where $i \geq 1$, the authors of loc. cit. defined the index subset $I_i^{(0)} = I_i(\mathbf{u}^{(0)}) \subset I_0$ that contains all indices k with $u_k^{(0)} \neq 0$ and additionally the $\max\{0, \lceil n/2^i \rceil - |\{k : u_k^{(0)} \neq 0\}|\}$ largest indices k of the vector $|\nabla g(\mathbf{u}^{(0)})|$ for which we have $u_k^{(0)} = 0$. Here, $|\cdot|$ denotes the componentwise absolute value of a vector. A maximal number $i_{max} \in \mathbb{N}$ of levels should be chosen according to the expected relative sparsity of the root $\mathbf{u}^*$. If the support of the current approximation is large, then less than i_{max} levels are chosen. The initial guess needs to be sparse. In what follows, we choose $\mathbf{u}^{(0)} = \mathbf{0} \in \mathbb{R}^n$.

Schematic representations of one V-cycle and one F-cycle multilevel iteration proposed in [134,135] are given in Figure 5.1, respectively. To perform one multilevel V-cycle iteration with initial guess $\mathbf{u}^{(0)} \in \mathbb{R}^n$, one starts with the computation of the index sets

$$I_i^{(0)} = I_i(\mathbf{u}^{(0)}), \quad i \geq 1.$$

As data transfer operations, the restriction operator from level $i-1$ ($i \geq 1$) to level i is defined as the restriction $\mathbf{u}_{I_i} \in \mathbb{R}^{|I_i|}$ of the current approximation $\mathbf{u}_{I_{i-1}} \in \mathbb{R}^{|I_{i-1}|}$ on level $i-1$ to the indices in the index set I_i. The prolongation operator is the zero padding operator, see [134, 135]. That is, one prolongates a vector $\mathbf{u}_{I_{i-1}} \in \mathbb{R}^{|I_{i-1}|}$ to the next finer level i by setting $u_k = 0$ for all additional indices $k \in I_{i-1} \setminus I_i$. In one V-cycle iteration, the initial guess is restricted to the indices in the index set $I_i^{(0)}$ that defines the coarsest level. Then, the zero-finding problem restricted to the coarsest level, i.e. with a reduced

number of unknowns, is solved exactly. As opposed to [134, 135], we solve the zero-finding problem on the coarsest level with the hybrid B-semismooth Newton method (Algorithm `hybridBSSN`) in order to accelerate this algorithm and because the algorithm is efficient in terms of computation time. This zero-finding problem on the coarsest level reads as $\mathbf{F}^{(0)}(\mathbf{u}) = \mathbf{0}$ with

$$\mathbf{F}^{(0)} \colon \mathbb{R}^{|I_i^{(0)}|} \to \mathbb{R}^{|I_i^{(0)}|}, \quad \mathbf{F}^{(0)}(\mathbf{u}) := \mathbf{u} - \mathbf{S}_{\gamma \mathbf{w}_{I_i^{(0)}}}(\mathbf{u} - \gamma(\nabla g(\tilde{\mathbf{P}}_i^{(0)}\mathbf{u}))_{I_i^{(0)}}), \tag{5.1}$$

where $I_i^{(0)}$ here denotes the index sets of the coarsest level, $\tilde{\mathbf{P}}_i^{(0)}$ denotes the composite prolongation of a vector defined on the coarsest level i to the finest level 0 and $\mathbf{w}_{I_i^{(0)}}$ denotes the weight sequence $\mathbf{w}$ restricted to the indices in level i. Finally, the approximation $\tilde{\mathbf{u}} \in \mathbb{R}^{|I_i^{(0)}|}$ of the root of $\mathbf{F}^{(0)}$ from (5.1) is prolongated until the finest level I_0 is reached. If the active set $\mathcal{A}(\mathbf{u}^*)$ of the zero $\mathbf{u}^*$ of $\mathbf{F}$ is contained in the coarsest level and if the low-level problem (5.1) is solved exactly, one can show that the multilevel method terminates with $\mathbf{u}^*$ after that V-cycle, see [134,135]. Therefore, the overall goal of the multilevel method is to include the active set of the root $\mathbf{u}^*$ into the coarsest level. To this end, Treister, Turek and Yavneh proposed to perform iterative soft-thresholding iterations on the appropriate level i after each prolongation,

$$\mathcal{S}_i^{(0)} \colon \mathbb{R}^{|I_i^{(0)}|} \to \mathbb{R}^{|I_i^{(0)}|}, \quad \mathbf{v} \mapsto \mathbf{S}_{\alpha \mathbf{w}_{I_i^{(0)}}}(\mathbf{v} - \alpha(\nabla g(\tilde{\mathbf{P}}_i^{(0)}\mathbf{v}))_{I_i^{(0)}}), \tag{5.2}$$

with the soft-thresholding operator from (1.5) with respect to the sequence $\alpha \mathbf{w}_{I_i^{(0)}}$, the composite prolongation $\tilde{\mathbf{P}}_i^{(0)}$ from level i to level 0, and $\alpha > 0$ is some parameter. Treister, Turek and Yavneh called (5.2) the *relaxation*. Here, the parameter α should be chosen small enough such that the soft-thresholding iteration converges, cf. [26, 61, 134, 135]. We perform only *one* soft-thresholding iteration after each prolongation.

In, [134,135], a sequence $\{\mathbf{u}^{(\nu)}\}_{\nu=0,\ldots,\nu_{max}}$, where $\nu_{max} \in \mathbb{N}$ denotes the maximal number of V-cycle iterations, is produced by performing several multilevel V-cycles with initial guess $\mathbf{u}^{(\nu)}$, $\nu = 0, 1, \ldots$, respectively, until the residual norm $\|\mathbf{F}(\mathbf{u}^{(\nu)})\|_2$ is smaller than the tolerance 10^{-7} or the maximal number ν_{max} of V-cycle iterations is reached. Note that the index sets

$$I_i^{(\nu)} = I_i(\mathbf{u}^{(\nu)}), \quad i = 0, 1, \ldots,$$

as well as the prolongation and restriction operators change in each V-cycle iteration. Using the results from [37] and requiring appropriate assumptions, it was shown in [134, 135] that the iterative soft-thresholding iterations on the different levels ensure the convergence of the sequence of iterates $\{\mathbf{u}^{(\nu)}\}_\nu$ in the multilevel method to the root $\mathbf{u}^*$ of $\mathbf{F}$. The computational effort of these iterations consists mainly of one computation of the gradient of g. Other possible choices of the relaxation method than the soft-thresholding iteration (5.2) are proposed in loc. cit. If the residual norm of the iterate $\mathbf{u}^{(\nu_{max})}$ in the V-cycle method is not smaller than the tolerance 10^{-7} after ν_{max} V-cycle iterations, we solve the zero-finding problem $\mathbf{F}(\mathbf{u}) = \mathbf{0}$ on the finest level 0 with Algorithm `hybridBSSN` and initial guess $\mathbf{u}^{(\nu_{max})}$. The authors of [134, 135] also propose an *F-cycle method* (full multilevel cycle), where additionally one V-cycle iteration is performed after each prolongation

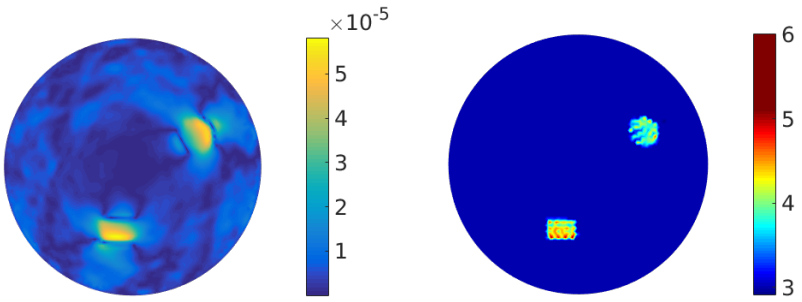


Figure 5.2: Motivation for the definition of the level sets in the multilevel V-cycle method. Left: gradient $\nabla g(\mathbf{u}^{(0)})$ of the discrepancy term g at the starting vector $\mathbf{u}^{(0)} = \mathbf{0}$ on a mesh with $n = 5000$ nodes. Right: reconstruction computed with a three-level V-cycle method with parameters $w_k \equiv 2 \cdot 10^{-5}$, $\alpha = 10^{-4}$ and $\gamma = 10^5$. The data contained approximately 1.12% of noise.

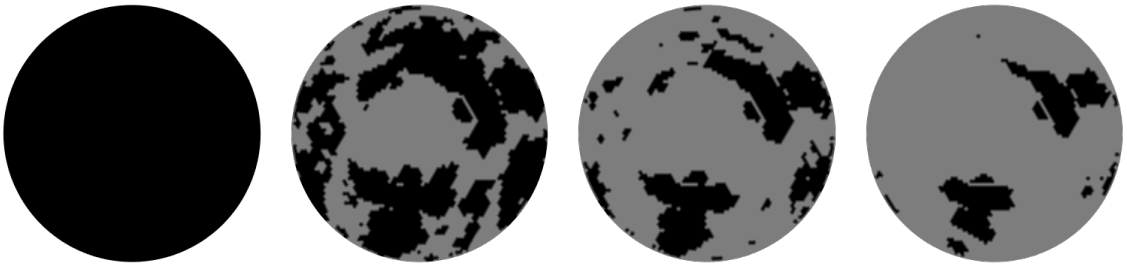


Figure 5.3: Levels used in one three-level V-cycle iteration on a mesh with $n = 5000$ nodes (nodes in black areas are included in the level, nodes in gray areas are not contained in the level). From left to right: level 0, level 1, level 2, level 3.

on the appropriate level, cf. Figure 5.1. The results of our first numerical experiments indicate that this method may also be a promising acceleration method for Algorithm `hybridBSSN`.

Numerical results for the examples of inverse integration and of nonlinear parameter identification from Chapter 4 demonstrate that the globalized B-semismooth Newton methods from Chapters 2 and 3 may be accelerated by the algebraic multilevel V-cycle method described above. In the following, we illustrate some first numerical results for the nonlinear parameter identification example from Section 4.2. In the following, we consider the coefficient a from (4.9) and the right-hand side f from (4.10).

The left-hand plot in Figure 5.2 shows the componentwise absolute value of the (discrete) gradient $\nabla g(\mathbf{u}^{(0)})$, where $\mathbf{u}^{(0)} = \mathbf{0} \in \mathbb{R}^{5000}$. The size of the gradient at the nodes indicates the possible position of areas where the coefficient a differs from the background value $a_0 = 3$. Note that the gradient $\nabla g(\mathbf{u}^{(0)})$ can only be computed approximately and that the data contained about 1% of noise. A reconstruction from the noisy data that was computed with a three-level V-cycle method ($i_{max} = 3$) is shown on the right-hand side in Figure 5.2. The regularization parameters were $w_k \equiv 2 \cdot 10^{-5}$, the parameter in the soft-thresholding iterations was $\alpha = 10^{-4}$, and we chose the parameter $\gamma = 10^5$ in Algorithm `hybridBSSN` for the exact solution on the coarsest level. The V-cycle method terminated with a residual norm of $1.6826 \cdot 10^{-10}$ after only one V-cycle iteration. The level sets $I_0, \ldots, I_3$ that were computed in this V-cycle are illustrated in Figure 5.3. The

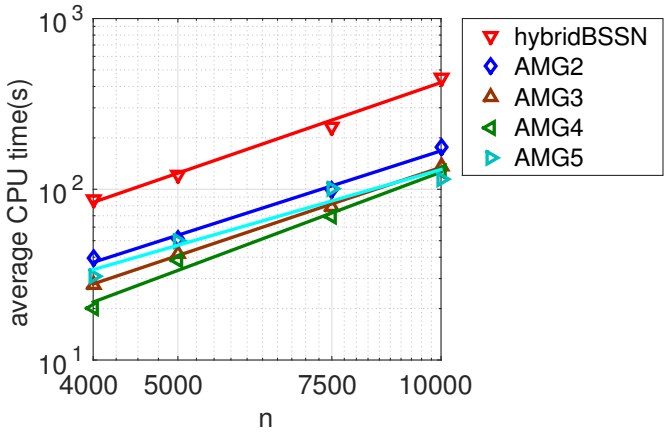


Figure 5.4: CPU time comparison of different multilevel V-cycle methods with Algorithm `hybridBSSN` for increasing numbers n of unknowns. The parameters used in the computations were $w_k \equiv 2 \cdot 10^{-5}$, $\gamma = 10^5$, $\alpha = 10^{-4}$.

black-colored areas were included in the appropriate level and the gray-colored areas were not contained in the level. Level 0 contains the entire unit square. Here, the active set of the reconstruction was included in the coarsest level 3. Therefore, the stopping criterion in the multilevel method was satisfied after one V-cycle in this test case.

Figure 5.4 shows a CPU time comparison of Algorithm `hybridBSSN` with different algebraic multilevel V-cycle methods `AMG`i_{max} with $i_{max} = 2, \ldots, 5$ for increasing numbers $n = 4000, 5000, 7500, 10000$ of unknowns. The parameter γ was equal to 10^5 and we chose $k_{max} = 10000$ in Algorithm `hybridBSSN`. The regularization parameters were $w_k \equiv 2.5 \cdot 10^{-5}$ and we had $\alpha = 10^{-4}$. We computed 0.84% noisy data on the triangulation with $n = 10000$ nodes (we chose $\tilde{\delta} = 0.2$, cf. Section 4.2) and interpolated the noisy data to the grids with less nodes. The resulting relative noise on the grid with $n = 4000$ nodes was 0.91%. On the grid with $n = 5000$ nodes, we had 0.95% of relative noise and the relative noise on the grid with $n = 7500$ nodes was 1%. In each of the two-, three- and four-level V-cycle methods (`AMG2`, `AMG3`, `AMG4`), the stopping criterion was satisfied after one V-cycle. In the five-level method `AMG5`, two V-cycles had to be performed in the cases $n = 4000$, $n = 5000$ as well as $n = 7500$ and in the case $n = 10000$, the residual norm of the iterate in Algorithm `AMG5` was smaller than the tolerance 10^{-7} after one V-cycle. Algorithm `hybridBSSN` needed 9 iterations in the case $n = 4000$, 8 iterations in the case $n = 5000$ and 7 iterations each in the cases $n = 7500$ and $n = 10000$ until the stopping criterion was satisfied. The CPU time results of this experiment confirm that Algorithm `hybridBSSN` can be accelerated significantly by a combination with algebraic multilevel methods. Parameters like the relative sparsity of the root $\mathbf{u}^*$, the maximal number i_{max} of levels, the regularization parameters w_k, the parameter α in the soft-thresholding iterations, the choice of the parameter γ and the amount of relative noise in the data may affect the efficiency of the multilevel methods. The influence of different parameters on the number of computed V-cycles and the computational cost of the multilevel methods need to be investigated in detail in a future work.

Summarizing, we can conclude that we developed a globally convergent hybrid B-semi-smooth Newton method for ℓ_1-Tikhonov regularization. This algorithm was efficient in our numerical experiments. There are several areas for further research. A promising subject is the acceleration of the methods by combining them with algebraic multilevel methods that have been proposed recently in the literature.

List of Figures

List of Tables

Bibliography

[1] S. W. Anzengruber and R. Ramlau. Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26(2):025001, 2010.

[2] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. New York, Springer, 2011.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] T. Bonesky. Morozov's discrepancy principle and Tikhonov-type functionals. *Inverse Problems*, 25(1):015015, 2009.

[5] T. Bonesky, K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method for nonlinear operator equations with sparsity constraints. *Inverse Problems*, 23(5):2041–2058, 2007.

[6] T. Bonesky, S. Dahlke, P. Maass, and T. Raasch. Adaptive wavelet methods and sparsity reconstruction for inverse heat conduction problems. *Adv. Comput. Math.*, 33(4):385–411, 2010.

[7] S. Bonettini and F. Tinti. A nonmonotone semismooth inexact Newton method. *Optim. Methods Softw.*, 22(4):637–657, 2007.

[8] T. Bosse. Multigrid method for nonsmooth problems. *Preprint*, 2015.

[9] K. Bredies and D. Lorenz. *Mathematische Bildverarbeitung. Einführung in Grundlagen und moderne Theorie.* Wiesbaden, Vieweg+Teubner, 2011.

[10] K. Bredies and D. A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM J. Sci. Comput.*, 30(2):657–683, 2008.

[11] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *J. Fourier Anal. Appl.*, 14(5-6):813–837, 2008.

[12] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Comput. Optim. Appl.*, 42(2):173–193, 2009.

[13] W. L. Briggs. A multigrid tutorial. Philadelphia, SIAM, 1987.

[14] S. Bütikofer. Globalizing a nonsmooth Newton method via nonmonotone path search. *Math. Methods Oper. Res.*, 68(2):235–256, 2008.

[15] R. H. Byrd, G. M. Chin, J. Nocedal, and F. Oztoprak. A family of second-order methods for convex ℓ_1-regularized optimization. *Math. Program., Ser. A*, 159(1):435–467, 2016.

[16] R. H. Byrd, J. Nocedal, and F. Oztoprak. An inexact successive quadratic approximation method for L-1 regularized optimization. *Math. Program., Ser. B*, 157(2):375–396, 2016.

[17] X. Chen. Superlinear convergence of smoothing quasi-Newton methods for nonsmooth equations. *J. Comput. Appl. Math.*, 80(1):105–126, 1997.

[18] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216, 2000.

[19] F. H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975.

[20] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Philadelphia, SIAM, 1990.

[21] C. Clason. *Numerical solution of optimal control and inverse problems in non-reflexive Banach spaces*. Habilitation thesis, University of Graz, 2012.

[22] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*. New York, Springer, 2013.

[23] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[24] R. W. Cottle, J.-S. Pang, and R. E. Stone. *The linear complementarity problem*. Philadelphia, SIAM, 2009.

[25] CVX Research, Inc. *CVX: Matlab software for disciplined convex programming, version 2.0.* http://cvxr.com/cvx (03.07.14).

[26] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[27] T. De Luca, F. Facchinei, and C. Kanzow. A semismooth equation approach to the solution of nonlinear complementarity problems. *Math. Program.*, 75(3):407–439, 1996.

[28] T. De Luca, F. Facchinei, and C. Kanzow. A theoretical and numerical comparison of some semismooth algorithms for complementarity problems. *Comput. Optim. Appl.*, 16(2):173–205, 2000.

[29] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 19(2):400–408, 1982.

[30] J. E. Dennis, Jr. and J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.*, 28(126):549–560, 1974.

[31] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations.* Philadelphia, SIAM, 1996.

[32] P. Deuflhard. *Newton methods for nonlinear problems. Affine invariance and adaptive algorithms.* Berlin, Springer, 2004.

[33] P. Dingguo and T. Weiwen. Globally convergent inexact generalized Newton's methods for nonsmooth equations. *J. Comput. Appl. Math.*, 138(1):37 – 49, 2002.

[34] S. P. Dirkse and M. C. Ferris. The path solver: a nonmonotone stabilization scheme for mixed complementarity problems. *Optim. Methods Softw.*, 5(2):123–156, 1995.

[35] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

[36] S. C. Eisenstat and H. F. Walker. Globally convergent inexact Newton methods. *SIAM J. Optim.*, 4(2):393–422, 1994.

[37] M. Elad, B. Matalon, and M. Zibulevsky. Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *Appl. Comput. Harmon. Anal.*, 23(3):346 – 367, 2007.

[38] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Dordrecht, Kluwer Academic Publishers, 1996.

[39] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. Providence, Amer. Math. Soc., 2002.

[40] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems. Vol. I.* New York, Springer, 2003.

[41] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems. Vol. II.* New York, Springer, 2003.

[42] M. C. Ferris and T. S. Munson. Interfaces to PATH 3.0: design, implementation and usage. *Comput. Optim. Appl.*, 12(1):207–227, 1999.

[43] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, 2003.

[44] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Process.*, 1(4):586–597, 2007.

[45] A. Fischer. A Newton-type method for positive-semidefinite linear complementarity problems. *J. Optim. Theory Appl.*, 86(3):585–608, 1995.

[46] A. Fischer and H. Jiang. Merit functions for complementarity and related problems: a survey. *Comput. Optim. Appl.*, 17(2):159–182, 2000.

[47] A. Fischer and C. Kanzow. On finite termination of an iterative method for linear complementarity problems. *Math. Program.*, 74(3):279–292, 1996.

[48] K. Fountoulakis and J. Gondzio. Performance of first- and second-order methods for ℓ_1-regularized least squares problems. *Comput. Optim. Appl.*, 65(3):605–635, 2016.

[49] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Math. Program., Ser. A*, 156(1):189–219, 2016.

[50] A. Galántai. The theory of Newton's method. *J. Comput. Appl. Math.*, 124:25–44, 2000.

[51] M. Gehre, T. Kluth, A. Lipponen, B. Jin, A. Seppänen, J. P. Kaipio, and P. Maass. Sparsity reconstruction in electrical impedance tomography: an experimental evaluation. *J. Comput. Appl. Math.*, 236(8):2126–2136, 2012.

[52] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Berlin, Springer, 1999.

[53] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben.* Berlin, Springer, 2002.

[54] M. Gerdts, S. Horn, and S.-J. Kimmerle. Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with applications in optimal control. *J. Ind. Manag. Optim.*, 13(1):47–62, 2017.

[55] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2(2):323–343, 2009.

[56] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with l^q penalty term. *Inverse Problems*, 24(5):055020, 2008.

[57] M. Grasmair, M. Haltmeier, and O. Scherzer. Necessary and sufficient conditions for linear convergence of ℓ_1-regularization. *Comm. Pure Appl. Math.*, 64(2):161–182, 2011.

[58] R. Griesse and D. A. Lorenz. A semismooth Newton method for Tikhonov functionals with sparsity constraints. *Inverse Problems*, 24(3):035007, 2008.

[59] W. Hackbusch. *Multi-grid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Berlin, Springer, 1985.

[60] W. W. Hager, D. T. Phan, and H. Zhang. Gradient-based methods for sparse recovery. *SIAM J. Imaging Sci.*, 4(1):146–165, 2011.

[61] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008.

[62] K. Hämäläinen, A. Kallonen, V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparse tomography. *SIAM J. Sci. Comput.*, 35(3):B644–B665, 2013.

[63] S.-P. Han, J.-S. Pang, and N. Rangaraj. Globally convergent Newton methods for nonsmooth equations. *Math. Oper. Res.*, 17(3):586–607, 1992.

[64] M. Hanke. A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Problems*, 13(1):79–95, 1997.

[65] M. Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens.* Wiesbaden, Vieweg+Teubner, 2009.

[66] E. Hans and T. Raasch. Global convergence of damped semismooth Newton methods for ℓ_1 Tikhonov regularization. *Inverse Problems*, 31(2):025005, 2015. ©IOP Publishing. Reproduced with permission. All rights reserved. doi:10.1088/0266-5611/31/2/025005.

[67] E. Hans and T. Raasch. A globally convergent and locally quadratically convergent modified B-semismooth Newton method for ℓ_1-penalized minimization. *arXiv:1508.03448v3 [math.OC]*, 2016.

[68] P. C. Hansen. *Discrete inverse problems. Insight and algorithms.* Philadelphia, SIAM, 2010.

[69] D. N. Hào and T. N. T. Quyen. Convergence rates for Tikhonov regularization of coefficient identification problems in Laplace-type equations. *Inverse Problems*, 26(12):125014, 2010.

[70] P. T. Harker and J.-S. Pang. A damped-Newton method for the linear complementarity problem. In E. L. Allgower and K. Georg, editors, *Computational solution of nonlinear systems of equations*, volume 26 of *Lectures in Applied Mathematics*, pages 265–284. Providence, Amer. Math. Soc., 1990.

[71] P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Math. Program.*, 48(1):161–220, 1990.

[72] P. T. Harker and B. Xiao. Newton's method for the nonlinear complementarity problem: a B-differentiable equation approach. *Math. Program.*, 48:339–357, 1990.

[73] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.

[74] B. Hofmann. *Mathematik inverser Probleme.* Stuttgart, Teubner, 1999.

[75] T. Hoheisel, C. Kanzow, B. S. Mordukhovich, and H. Phan. Generalized Newton's method based on graphical derivatives. *Nonlinear Anal.*, 75(3):1324 – 1340, 2012.

[76] K. Ito and B. Jin. *Inverse problems. Tikhonov theory and algorithms.*, volume 22 of *Series on Applied Mathematics*. Hackensack, World Scientific, 2015.

[77] K. Ito and K. Kunisch. On a semi-smooth Newton method and its globalization. *Math. Program., Ser. A*, 118(2):347–370, 2009.

[78] B. Jin, T. Khan, and P. Maass. A reconstruction algorithm for electrical impedance tomography based on sparsity regularization. *Int. J. Numer. Meth. Eng.*, 89(3):337–353, 2012.

[79] B. Jin and P. Maass. Sparsity regularization for parameter identification problems. *Inverse Problems*, 28(12):123001, 2012.

[80] C. Kanzow. Inexact semismooth Newton methods for large-scale complementarity problems. *Optim. Methods Softw.*, 19(3-4):309–325, 2004.

[81] C. Kanzow and H. Kleinmichel. A new class of semismooth Newton-type methods for nonlinear complementarity problems. *Comput. Optim. Appl.*, 11(3):227–251, 1998.

[82] N. Keskar, J. Nocedal, F. Öztoprak, and A. Wächter. A second-order method for convex l1-regularized optimization with active-set prediction. *Optim. Methods Softw.*, 31(3):605–621, 2016.

[83] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE J. Sel. Topics Signal Process.*, 1(4):606–617, 2007.

[84] I. Knowles. Parameter identification for elliptic problems. *J. Comput. Appl. Math.*, 131(1-2):175–194, 2001.

[85] M. Kojima and S. Shindo. Extension of Newton and quasi-Newton methods to systems of PC^1 equations. *J. Oper. Res. Soc. Japan*, 29:352–375, 1986.

[86] V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparsity-promoting Bayesian inversion. *Inverse Problems*, 28(2):025005, 2012.

[87] B. Kummer. Newton's method for non-differentiable functions. In J. Guddat, B. Bank, H. Hollatz, P. Knall, D. Klatte, B. Kummer, K. Lommatzsch, K. Tammer, M. Vlach, and K. Zimmermann, editors, *Advances in Mathematical Optimization*, Math. Res. 45, pages 114–125. Berlin, Akademie-Verlag, 1988.

[88] B. Kummer. Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis. In W. Oettli and D. Pallaschke, editors, *Advances in Optimization*, volume 382 of *Lecture Notes in Economics and Mathematical Systems*, pages 171–194. Berlin, Springer, 1992.

[89] B. Kummer. On stability and Newton-type methods for Lipschitzian equations with applications to optimization problems. In L. D. Davisson, A. G. J. MacFarlane, H. Kwakernaak, J. L. Massey, Y. Z. Tsypkin, A. J. Viterbi, and P. Kall, editors, *System Modelling and Optimization*, volume 180 of *Lecture Notes in Control and Information Sciences*, pages 3–16. Berlin, Springer, 1992.

[90] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.*, 24(3):1420–1443, 2014.

[91] C. E. Lemke. Bimatrix equilibrium points and mathematical programming. *Management Science*, 11(7):681–689, 1965.

[92] D. A. Lorenz. Constructing test instances for basis pursuit denoising. *IEEE Trans. Signal Process.*, 61(5):1210–1214, 2013.

[93] D. A. Lorenz, P. Maass, and P. Q. Muoi. Gradient descent for Tikhonov functionals with sparsity constraints: theory and numerical comparison of step size rules. *Electron. Trans. Numer. Anal.*, 39:437–463, 2012.

[94] I. Loris. On the performance of algorithms for the minimization of ℓ_1-penalized functionals. *Inverse Problems*, 25(3):035008, 2009.

[95] I. Loris, M. Bertero, C. De Mol, R. Zanella, and L. Zanni. Accelerating gradient projection methods for ℓ_1-constrained signal recovery by steplength selection rules. *Appl. Comput. Harmon. Anal.*, 27(2):247–254, 2009.

[96] A. K. Louis. *Inverse und schlecht gestellte Probleme.* Stuttgart, Teubner, 1989.

[97] J. M. Martínez and L. Qi. Inexact Newton methods for solving nonsmooth equations. *J. Comput. Appl. Math.*, 60(1):127 – 145, 1995.

[98] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, 15(6):959–972, 1977.

[99] A. Milzarek and M. Ulbrich. A semismooth Newton method with multidimensional filter globalization for ℓ_1-optimization. *SIAM J. Optim.*, 24(1):298–333, 2014.

[100] T. S. Munson, F. Facchinei, M. C. Ferris, A. Fischer, and C. Kanzow. The semismooth algorithm for large scale complementarity problems. *INFORMS J. Comput.*, 13(4):294–311, 2001.

[101] P. Q. Muoi. *Sparsity constraints and regularization for nonlinear inverse problems*. PhD thesis, University of Bremen, 2012.

[102] P. Q. Muoi. Sparsity regularization of the diffusion coefficient identification problem: well-posedness and convergence rates. *Bull. Malays. Math. Sci. Soc.*, 39(3):1145–1164, 2016.

[103] P. Q. Muoi, D. N. Hào, P. Maass, and M. Pidcock. Semismooth Newton and quasi-Newton methods in weighted ℓ^1-regularization. *J. Inverse Ill-Posed Probl.*, 21(5):665–693, 2013.

[104] P. Q. Muoi, D. N. Hào, P. Maass, and M. Pidcock. Descent gradient methods for nonsmooth minimization problems in ill-posed problems. *J. Comput. Appl. Math.*, 298:105–122, 2016.

[105] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program., Ser. B*, 140(1):125–161, 2013.

[106] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–404, 2000.

[107] J.-S. Pang. Newton's method for B-differentiable equations. *Math. Oper. Res.*, 15(2):311–341, 1990.

[108] J.-S. Pang. A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems. *Math. Program.*, 51(1):101–131, 1991.

[109] J.-S. Pang and L. Qi. Nonsmooth equations: motivation and algorithms. *SIAM J. Optim.*, 3(3):443–465, 1993.

[110] P.-O. Persson and G. Strang. A simple mesh generator in MATLAB. *SIAM Rev.*, 46(2):329–345, 2004.

[111] L. Qi. Convergence analysis of some algorithms for solving nonsmooth equations. *Math. Oper. Res.*, 18(1):227–244, 1993.

[112] L. Qi. On superlinear convergence of quasi-Newton methods for nonsmooth equations. *Oper. Res. Lett.*, 20(5):223 – 228, 1997.

[113] L. Qi and X. Chen. A globally convergent successive approximation method for severely nonsmooth equations. *SIAM J. Control Optim.*, 33(2):402–418, 1995.

[114] L. Qi and D. Sun. A survey of some nonsmooth equations and smoothing Newton methods. In A. Eberhard, B. M. Glover, R. Hill, and D. Ralph, editors, *Progress in optimization*, volume 30 of *Applied Optimization*, pages 121–146. Dordrecht, Kluwer Academic Publishers, 1999.

[115] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Program.*, 58(3):353–367, 1993.

[116] L. Qi and J. Sun. A trust region algorithm for minimization of locally Lipschitzian functions. *Math. Program.*, 66(1):25–43, 1994.

[117] H. Rademacher. Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale. *Math. Ann.*, 79(4):340–359, 1919.

[118] D. Ralph. Global convergence of damped Newton's method for nonsmooth equations via the path search. *Math. Oper. Res.*, 19(2):352–389, 1994.

[119] R. Ramlau and E. Resmerita. Convergence rates for regularization with sparsity constraints. *Electron. Trans. Numer. Anal.*, 37:87–104, 2010.

[120] R. Ramlau and G. Teschke. A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. *Numer. Math.*, 104(2):177–203, 2006.

[121] S. M. Robinson. Local structure of feasible sets in nonlinear programming, part III: stability and sensitivity. *Math. Program. Study*, 30:45–66, 1987.

[122] S. M. Robinson. Newton's method for a class of nonsmooth functions. *Set-Valued Anal.*, 2(1):291–305, 1994.

[123] U. Schäfer. *Das lineare Komplementaritätsproblem. Eine Einführung*. Berlin, Springer, 2008.

[124] S. Scholtes. *Introduction to piecewise differentiable equations*. New York, Springer, 2012.

[125] F. Schöpfer, A. K. Louis, and T. Schuster. Nonlinear iterative methods for linear ill-posed problems in Banach spaces. *Inverse Problems*, 22(1):311–329, 2006.

[126] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski. *Regularization methods in Banach spaces*, volume 10 of *Radon Series on Computational and Applied Mathematics*. Berlin, de Gruyter, 2012.

[127] S. Setzer. Operator splittings, Bregman methods and frame shrinkage in image processing. *Int. J. Comput. Vis.*, 92(3):265–280, 2011.

[128] A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66(3):477–487, 1990.

[129] S. Solntsev, J. Nocedal, and R. H. Byrd. An algorithm for quadratic ℓ_1-regularized optimization with a flexible active-set strategy. *Optim. Methods Softw.*, 30(6):1213–1237, 2015.

[130] M. V. Solodov and B. F. Svaiter. A truly globally convergent Newton-type method for the monotone nonlinear complementarity problem. *SIAM J. Optim.*, 10(2):605–625, 2000.

[131] D. Sun and J. Han. Newton and quasi-Newton methods for a class of nonsmooth equations and related problems. *SIAM J. Optim.*, 7(2):463–480, 1997.

[132] J. Sun. On piecewise quadratic Newton and trust region problems. *Math. Program.*, 76(3):451–467, 1997.

[133] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58(1):267–288, 1996.

[134] E. Treister, J. S. Turek, and I. Yavneh. A multilevel framework for sparse optimization with application to inverse covariance estimation and logistic regression. *SIAM J. Sci. Comput.*, 38(5):S566–S592, 2016.

[135] E. Treister and I. Yavneh. A multilevel iterated-shrinkage approach to l_1 penalized least-squares minimization. *IEEE Trans. Signal Process.*, 60(12):6319–6329, 2012.

[136] M. Ulbrich. *Nonsmooth Newton-like methods for variational inequalities and constrained optimization problems in function spaces*. Habilitation thesis, Technical University Munich, 2002.

[137] M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–841, 2003.

[138] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung.* Basel, Birkhäuser, 2012.

[139] W. Wang, S. W. Anzengruber, R. Ramlau, and B. Han. A global minimization algorithm for Tikhonov functionals with sparsity constraints. *Appl. Anal.*, 94(3):580–611, 2015.

[140] X. Wang, C. Ma, and M. Li. A globally and superlinearly convergent quasi-Newton method for general box constrained variational inequalities without smoothing approximation. *J. Global Optim.*, 50(4):675–694, 2011.

[141] J. Williams, Y. Lu, S. Niebe, M. Andersen, K. Erleben, and J. C. Trinkle. RPI-MATLAB-Simulator: a tool for efficient research and practical teaching in multibody dynamics. In J. Bender, J. Dequidt, C. Duriez, and G. Zachmann, editors, *Workshop on virtual reality interaction and physical simulation*, pages 71–80. The Eurographics Association, 2013.

[142] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

[143] B. Xiao and P. T. Harker. A nonsmooth Newton method for variational inequalities, I: theory. *Math. Program.*, 65:151–194, 1994.

[144] B. Xiao and P. T. Harker. A nonsmooth Newton method for variational inequalities, II: numerical results. *Math. Program.*, 65:195–216, 1994.

[145] T. Yamamoto. Historical developments in convergence analysis for Newton's and Newton-like methods. *J. Comput. Appl. Math.*, 124:1 – 23, 2000.

[146] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM J. Sci. Comput.*, 33(1):250–278, 2011.

[147] X.-T. Yuan and S. Yan. A finite Newton algorithm for non-degenerate piecewise linear systems. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 841–854, 2011.

[148] C. A. Zarzer. On Tikhonov regularization with non-convex sparsity constraints. *Inverse Problems*, 25(2):025006, 2009.

[149] E. Zeidler. *Nonlinear functional analysis and its applications. II/B: Nonlinear monotone operators.* New York, Springer, 1990.

[150] J. Zou. Numerical methods for elliptic inverse problems. *Int. J. Comput. Math.*, 70(2):211–232, 1998.